1 Kernel Machines

Consider a linear model

$$Y_i = x_i^T \beta^0 + \varepsilon_i,\ i = 1, \dots, n,\ x_i \in \mathbb{R}^p \text{ fixed}$$

where $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$. We have

$$\begin{aligned}\hat{\beta}^{\text{ols}} &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - x_i^T \beta)^2 \\ &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \\ &= (X^T X)^{-1} X^T Y.\end{aligned}$$

Classical theory:

- $\hat{\beta}^{\text{ols}}$ unbiased,

  $$\mathrm{Var}(\hat{\beta}^{\text{ols}}) = \sigma^2 (X^T X)^{-1} = i^{-1}(\beta^0)$$

  Where i is the Fisher information.

- Cramér-Rao lower bound: if an estimator $\tilde{\beta}$ is unbiased then

  $$\mathrm{Var}(\tilde{\beta}) - i^{-1}(\beta^0) \underbrace{\geq}_{\text{positive semi-definite}} 0.$$

- If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then $\hat{\beta}^{\text{ols}}$ is the MLE of β^0. Furthermore $\sqrt{n}(\hat{\beta}^{\text{ols}} - \beta^0) \sim \mathcal{N}(0, n\sigma^2 (X^T X)^{-1})$. From this we can derive confidence intervals, hypothesis test, etc.

In a general model with parameter $\theta \in \mathbb{R}^p$, n independent observations, under regularity, we have asymptotic normality, i.e $\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta^0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta^0))$ (with p fixed).

Question: what happens when p is large relative to n?

- If $p > n$, $\hat{\beta}^{\text{ols}}$ is not even defined.
- If $p \approx n$, $\mathrm{Var}(\hat{\beta}^{\text{ols}})$ explodes since $X^T X$ is near singular.
- More generally, if $p, n \to \infty$ then asymptotic normality can break down.

Recall the bias-variance decomposition:

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

We introduce bias to reduce the variance.

1.1 Ridge regression

Define

$$(\hat{\mu}^R_\lambda, \hat{\beta}^R_\lambda) = \mathrm{argmin}_{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}\left[||Y - \mu\mathbf{1} - X\beta||^2 + \underbrace{\lambda||\beta||^2}_{\text{penalty for large } \beta} \right].$$

λ is called a *regularisation* or *tuning* parameter. We shall assume the columns of X have been standardised (mean 0, variance 1).

After standardisation, we can show that

$$\hat{\mu}^R_\lambda = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}.$$

Hence, if we replace Y with $Y - \mathbf{1}\bar{Y}$ we can write

$$\begin{aligned}\hat{\beta}^R_\lambda &= \mathrm{argmin}_{\beta\in\mathbb{R}^p}\left[\|Y - X\beta\|^2 + \lambda||\beta||^2\right]\\ &= \underbrace{(X^TX + \lambda I_p)^{-1}}_{\text{always invertible}} X^TY.\end{aligned}$$

Theorem 1.1. *For $\lambda > 0$ sufficiently small,*

$$\mathrm{mse}(\hat{\beta}^{ols}) - \mathrm{mse}(\hat{\beta}^R_\lambda) = \mathbb{E}\|\hat{\beta}^{ols} - \beta^0\|^2 - \mathbb{E}\|\hat{\beta}^R_\lambda - \beta^0\|^2 > 0. \qquad (*)$$

Proof. We have

$$Y = X\beta^0 + \varepsilon.$$

The bias of $\hat{\beta}^R_\lambda$ is

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

While we have variance

$$\begin{aligned}\mathrm{Var}(\hat{\beta}^R_\lambda) &= \mathbb{E}\left\|(X^TX + \lambda I)^{-1}X^T\varepsilon\right\|^2\\ &= \sigma^2\left[(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}\right].\end{aligned}$$

Then $(*)$ becomes

$$\begin{aligned}&\mathbb{E}\|\hat{\beta}^{\mathrm{ols}} - \beta^0\|^2 - \mathbb{E}\|\hat{\beta}^R_\lambda - \beta^0\|^2\\ &= \sigma^2(X^TX)^{-1} - \sigma^2(X^TX + \lambda I)X^TX(X^TX + \lambda I)^{-1}\\ &\quad - \lambda^2(X^TX + \lambda I)^{-1}\beta^0(\beta^0)^T(X^TX + \lambda I)^{-1}\\ &= \;\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{use SVD } X = UDV^T)\\ &= \lambda(X^TX + \lambda I)^{-1}\left[\sigma^2\left\{2I_p + \lambda(X^TX)^{-1}\right\} - \lambda\beta^0(\beta^0)^T\right](X^TX + \lambda I)^{-1}.\end{aligned}$$

We want to show this is positive definite. This is equivalent to

$$
\begin{aligned}
&\sigma^2 \left[2I + \lambda(X^TX)^{-1}\right] - \lambda\beta^0(\beta^0)^T > 0 \\
&\iff 2\sigma^2 I - \lambda\beta^0(\beta^0)^T > 0 \\
&\iff 2\sigma^2||z||^2 - \lambda(z^T\beta^0)^2 > 0 \quad \forall z \in \mathbb{R}^p. \qquad (\dagger)
\end{aligned}
$$

We also have $(z^T\beta^0)^2 \leq ||z||^2||\beta^0||^2$ by Cauchy-Schwarz. Hence ($\dagger$) holds for all $\lambda < \frac{2\sigma^2}{||\beta^0||^2}$. $\square$

Singular value decomposition

Suppose $n \geq p$, so we can always write $X \in \mathbb{R}^{n \times p}$ as

$$X = UDV^T \qquad (\text{"thin SVD"})$$

where $U \in \mathbb{R}^{n \times p}, V \in \mathbb{P}^{p \times p}$, with orthonormal columns, $D \in \mathbb{R}^{p \times p}$ diagonal with $D_{11} \geq D_{22} \geq \ldots \geq D_{pp} \geq 0$.

The fitted values in ridge regression are

$$\begin{aligned}
\hat{Y}_\lambda^R = X\hat{\beta}_\lambda^R &= X(X^TX + \lambda I)^{-1}X^TY \\
&= UDV^T(VD^2V^T + \lambda I)^{-1}VDU^TY \quad (\text{using } VV^T = V^TV = I) \\
&= UD(D^2 + \lambda I)^{-1}DU^TY \\
&= \sum_{j=1}^{p} U_j \frac{D_{jj}^2}{D_{jj}^2 + \lambda} U_j^T Y
\end{aligned}$$

where U_j denotes the jth column of U. For reference, in OLS regression

$$\hat{Y}^{ols} = X\hat{\beta}^{\text{ols}} = X(X^TX)^{-1}X^TY = \sum_{j=1}^{p} U_j U_j^T Y.$$

So ridge "projects" onto columns of U, but it shrinks jth component by a factor

$$\frac{D_{jj}^2}{D_{jj}^2 + \lambda}.$$

Hence it shrinks small singular values to 0 rapidly.

Note. The matrix $X(X^TX)^{-1}X^TY$ is known as the "hat matrix" and it represents an orthogonal projection onto the column space of X.

The SVD of X is related to principal component analysis.

Definition. The kth *principal component* $U^{(k)}$ of X and *principal direction* $v^{(k)}$ of X are defined recursively by

$$v^{(k)} = \text{argmax}_{v \in \mathbb{R}^p} \, ||Xv||^2 \text{ subject to } ||v|| = 1, \ (v^{(j)})^T X^T X v = 0 \ \forall j < k$$

and

$$u^{(k)} = Xv^{(k)}.$$

Lemma 1.2. *If $D_{jj} > 0$ for all $j \in \{1, \ldots, p\}$ then $v^{(k)} = V_k$, $u^{(k)} = D_{kk}U_k$.*

Message: ridge is good when the signal (β^0) is large for the top principal components of X.

Computation: we can compute $\hat{Y}_\lambda^R$ for any value of λ quickly after doing an SVD, which has cost $\mathcal{O}(np^2)$.

1.2 v-fold cross-validation

We assume that (x_i, Y_i), $i = 1, \dots, n$ is iid from some distribution (random design matrix). Let (x^*, Y^*) be another independent observation from this distribution. We may wish to pick λ minimising the mean-squared prediction error (MSPE) conditional on (X, Y):

$$\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}^R_\lambda(X, Y))^2 | (X, Y)\}.$$

A less ambitious goal is to minimise the MSPE

$$\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}^R_\lambda(X, Y))^2\} = \mathbb{E}\left[\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}^R_\lambda(X, Y))^2 | (X, Y)\}\right]. \qquad (\ddagger)$$

We can try to estimate this quantity for different values of λ, using data splitting.

- Let $(X^{(1)}, Y^{(1)}), \dots, (X^{(v)}, Y^{(v)})$ be groups of data points of roughly equal size. These are called *folds*.
- Let $(X^{(-k)}, Y^{(-k)})$ denote all the folds except the kth.
- Let $\kappa(i)$ be the fold to which sample i (i.e (X_i, Y_i)) belongs.

Our estimator of ($\ddagger$) is

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ Y_i - x_i^T \underbrace{\hat{\beta}^R_\lambda(X^{(-\kappa(i))}, Y^{(-\kappa(i))})}_{\substack{\text{using all folds except} \\ \text{the ones containing } (x_i, Y_i)}} \right\}^2.$$

Then define

$$\lambda_{\mathrm{CV}} = \operatorname{argmin}_{\lambda \in \{l_1, \dots, l_m\}} \mathrm{CV}(\lambda).$$

We use the estimator

$$\hat{\beta}^R_{\lambda_{\mathrm{CV}}}(X, Y).$$

How to choose v?

Note.

- The expectation of each summand in $\mathrm{CV}(\lambda)$ is almost the same as $\ddagger$, which is what we want to estimate. The only difference is the size of the training set. Hence the bias of $\mathrm{CV}(\lambda)$ is small when v is large [the extreme of this is $v = n$, called "leave one out" cross-validation].
- When v is large, the estimator $\hat{\beta}^R_\lambda(X^{(-k)}, Y^{(-k)})$ is similar for different values of k, which leads to positively correlated summands in $\mathrm{CV}(\lambda)$, leading to high variance.
- A common choice is $v = 5$ or $v = 10$.

1.3 Kernel trick

We have

$$\hat{Y}^R_\lambda = X(X^TX + \lambda I)^{-1}X^TY.$$

Note that

$$\begin{aligned}
&X^T(XX^T + \lambda I) = (X^TX + \lambda I)X^T \\
&\implies (X^TX + \lambda I)^{-1}X^T = X^T(XX^T + \lambda I)^{-1} \\
&\implies X(\underbrace{X^TX + \lambda I}_{p\times p})^{-1}X^TY = XX^T(\underbrace{XX^T + \lambda I}_{n\times n})^{-1}Y.
\end{aligned}$$

The computation cost of the LHS is $\mathcal{O}(np^2 + p^3)$ while the RHS is $\mathcal{O}(pn^2 + n^3)$.

- When $p \gg n$, the 2nd expression is cheaper to compute;
- The fitted values in ridge regression only depend on X through the "Gram matrix" $K = XX^T$, with entries $K_{ij} = \langle x_i, x_j \rangle$.

Suppose we wish to fit a quadratic model:

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

This can be done with a linear model where we replace the predictors $x_i \in \mathbb{R}^p$ with a new "feature" vector:

$$\phi(x_i) = (x_{i1}, x_{i2}, \ldots, x_{ip}, x_{i1}x_{i1}, x_{i1}x_{i2}, \ldots, x_{ip}x_{ip}) \in \mathbb{R}^{p+p^2}.$$

We call ϕ a "feature map". Now we have $\mathcal{O}(p^2)$ predictors. If $p^2 \gg n$, to compute ridge fitted values, we want to use the 2nd expression, with cost $\mathcal{O}(p^2n^2 + n^3)$.

However, the part that scales as $\mathcal{O}(p^2n^2)$ is just the computation of the Gram matrix with entries $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$.

The *kernel trick* offers a shortcut for computing K.

Idea:

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

The LHS can be computed in $\mathcal{O}(p)$ iterations, so we can obtain K in $\mathcal{O}(n^2p)$ iterations, and we can compute the fitted values in ridge regression in $\mathcal{O}(n^2p + n^3)$, which is not worse than the linear model!

Notes.

- For many feature maps ϕ, there are similar shortcuts.
- Instead of focusing on ϕ, we can directly think of the function $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ as a measure of "similarity" between inputs x_i, x_j.

Question: for which similarities k is there a feature map ϕ such that

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle?$$

1.4 Kernels

Definition. An *inner product space* is a real vector space $\mathcal{H}$ endowed with a map $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ satisfying:

(i) Symmetry: for all $u, v \in \mathcal{H}$ we have $\langle u, v \rangle = \langle v, u \rangle$;

(ii) Bilinearity: for all $a, b \in \mathbb{R}$ and all $u, v, w \in \mathcal{H}$ we have

$$\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle.$$

(iii) Positive-definiteness: we have $\langle u, u \rangle \geq 0$ for all $u \in \mathcal{H}$, with equality if and only if $u = 0$.

Suppose that regression inputs $x_1, \ldots, x_n$ take values in an abstract set $\mathcal{X}$ (so far we've had $\mathcal{X} = \mathbb{R}^p$, but the x_i's could be functions; images; graphs; etc.).

Goal: characterise similarity functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there is an inner product space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad \forall x, x' \in \mathcal{X}.$$

Definition. A *(positive-definite) kernel* k is a symmetric map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that 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.

Remark. A kernel is not an inner product on $\mathcal{X}$ in general. Indeed, $\mathcal{X}$ does not even need to be a vector space, and k need not be bilinear. However, we do have a version of the Cauchy-Schwarz inequality for kernels.

Proposition 1.3. *Let k be a kernel on $\mathcal{X}$. Then*

$$k(x, x')^2 \leq k(x, x)k(x', x') \quad \forall x, x' \in \mathcal{X}.$$

Proof. Since k is a kernel,

$$\begin{pmatrix} k(x, x) & k(x, x') \\ k(x', x) & k(x', x') \end{pmatrix} \geq 0.$$

Hence this has non-negative determinant and $k(x, x)k(x', x') - k(x, x')^2 \geq 0$. $\square$

Proposition 1.4. *Any similarity k defined by*

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad \forall x, x' \in \mathcal{X}$$

is a kernel.

Proof. Symmetry of k is clear. Let $x_1, \ldots, x_n \in \mathcal{X}$ be arbitrary and take any vector $\alpha \in \mathbb{R}^n$. We need to show $\alpha^T K \alpha \geq 0$. Indeed

$$\begin{aligned}
\alpha^T K \alpha &= \sum_{i,j} \alpha_i K_{ij} \alpha_j \\
&= \sum_{i,j} \alpha_i \langle \phi(x_i), \phi(x_j) \rangle \alpha_j \\
&= \langle \sum_{i=1}^n \alpha \phi(x_i), \sum_{j=1}^n \alpha_j \phi(x_j) \rangle && \text{(linearity of } \langle \cdot, \cdot \rangle\text{)} \\
&\geq 0. && \text{(positive-definiteness of } \langle \cdot, \cdot \rangle\text{)}
\end{aligned}$$

□

Examples of kernels

Proposition 1.5 (Closure property). *Suppose $k_1, k_2, \ldots$ are kernels on $\mathcal{X}$. Then*

(i) If $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1 k_1 + \alpha_2 k_2$ is a kernel. If $k(x, x') := \lim_{m\to\infty} k_m(x, x')$ exists for all $x, x' \in \mathcal{X}$, then k is a kernel.

(ii) The pointwise product $k(x, x') = k_1(x, x')k_2(x, x')$ is a kernel.

Proof. Example Sheet 1. □

Some examples of kernels are:

- Linear kernel: $k(x, x') = x^T x'$ (for $\mathcal{X} = \mathbb{R}^p$);
- Polynomial kernel: $k(x, x') = (1+x^T x')^d$, $d \in \mathbb{N}$ ($\mathcal{X} = \mathbb{R}^p$). Note $(x, x') \mapsto 1$ is a kernel so this is a kernel by the previous proposition;
- Gaussian kernel: $k(x, x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$, $\sigma^2 > 0$ the *bandwidth* of the kernel. Indeed note

$$\exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right) = \underbrace{\exp\left(-\frac{||x||^2}{2\sigma^2}\right)\exp\left(-\frac{||x'||^2}{2\sigma^2}\right)}_{:=k_1(x,x')}\underbrace{\exp\left(\frac{x^T x'}{\sigma^2}\right)}_{:=k_2(x,x')}.$$

  It suffices to show k_1, k_2 are kernels. For k_1 we have $k_1(x, x') = \langle \phi(x), \phi(x')\rangle$ where $\phi : \mathbb{R}^p \to \mathbb{R}$ is defined by

$$\phi(x) = \exp\left(-\frac{||x||^2}{2\sigma^2}\right).$$

  For k_2 we have that $(x, x') \mapsto x^T x'$ is a kernel and k_2 can be Taylor expanded so is the limit of kernels;
- Sobolev kernel: let $\mathcal{X} = [0, 1]$ and set $k(x, x') = \min(x, x') = \mathrm{Cov}(Wx, Wx')$ where $(W_t)_{t\geq 0}$ a Brownian motion (positive definite as a covariance);
- Jaccard similarity kernel: let $\mathcal{X} = \mathcal{P}(\{1, \ldots, p\})$ and set

$$k(x, x') = \begin{cases} \frac{|x\cap x'|}{|x\cup x'|} & \text{if } x \cup x' \neq \emptyset \\ 0 & \text{otherwise} \end{cases}.$$

  (For proof this is a kernel see Example Sheet 1.)

Remark. There is no finite-dimensional feature map $\phi : \mathbb{R}^p \to \mathbb{R}^m$ respresenting the Gaussian kernel.

Theorem 1.6 (Moore-Aronzajn Theorem). *For every kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a feature map ϕ taking values in some inner product space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x')\rangle$ for all $x, x' \in \mathcal{X}$.*

Proof. Take $\mathcal{H}$ to be the vector space of functions from $\mathcal{X}$ to $\mathbb{R}$ of the form

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}.$$

In other words, $\mathcal{H}$ is the linear span of functions of the form $f(\cdot, x)$ for $x \in \mathcal{X}$. Our feature map $\phi : \mathcal{X} \to \mathcal{H}$ will be $\phi(x) = k(\cdot, x)$. We now define the inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{H}$. Let

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}$$

and

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

Then define

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

In particular, the final two expressions show $\langle \cdot, \cdot \rangle$ is well-defined (it doesn't matter how we represent f, g as these linear combinations).

We observe directly from the definition that $\langle \phi(x), \phi(x') \rangle = \langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x')$ as required. We must show $\langle \cdot, \cdot \rangle$ is indeed an inner product. It is certainly bilinear and symmetric. So we show it is positive-definite. Note that

$$\langle f, f \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i k(x_i, x_j) \alpha_j \geq 0 \tag{‡}$$

since k is a kernel. It remains to show $\langle f, f \rangle$ implies $f(x) = 0$ for all $x \in \mathcal{X}$.

Note that $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is a kernel. Indeed, given functions $f_1, \ldots, f_m$ and $\gamma_1, \ldots, \gamma_n \in \mathbb{R}$ we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_i \langle f_i, f_j \rangle \gamma_j = \langle \sum_{i=1}^{n} \gamma_i f_i, \sum_{j=1}^{n} \gamma_j f_j \rangle \geq 0$$

by (‡).

Now note that

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

by the Cauchy-Schwarz property for kernels. Hence $\langle f, f \rangle = 0$ implies $f(x) = 0$ for all $x \in \mathcal{X}$. $\square$

Remark. The space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ constructed in the proof has the property that

$$f(x) = \langle f, \underbrace{k(\cdot, x)}_{\phi(x)} \rangle.$$

As a consequence

$$|f(x) - g(x)| = |\langle f - g, k(\cdot, x) \rangle| \leq \|f - g\|_{\mathcal{H}} \ k(x, x)^{1/2}.$$

Hence convergence in $(\mathcal{H}, \|\cdot\|)$ implies pointwise convergence.

Lemma 1.7. *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{V} \subseteq \mathcal{H}$ a closed subspace. Then $\mathcal{H} = \mathcal{V} \oplus \mathcal{V}^{\perp}$, i.e for any $f \in \mathcal{H}$ we have $f = u + v$ where $u \in \mathcal{V}$ and $v \in \mathcal{V}^{\perp}$ and u, v are unique.*

Proof. See Part II Linear Analysis. □

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

The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $(x, x') \mapsto \langle k_x, k_{x'} \rangle = k_x(x')$ is known as the reproducing kernel of $\mathcal{H}$.

Remark. By the Riesz Representation Theorem, it is equivalent to define an RKHS as a Hilbert space where the evaluation operator $E_x : f \mapsto f(x)$ is a continuous linear operator.

The Moore-Aronzajn Theorem says that whenever k is a kernel, there is an inner product space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ where $f(x) = \langle f, k(\cdot, x) \rangle$ and thus $k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle$.

This implies that if $(f_n)_{n \geq 1}$ is Cauchy in $\mathcal{H}$,

$$|f_n(x) - f_m(x)| \leq \sqrt{k(x,x)} ||f_n - f_m||_{\mathcal{H}} \to 0.$$

Hence $(f_n)_{n \geq 1}$ has a pointwise limit $f^* : \mathcal{X} \to \mathbb{R}$ by completeness of $\mathbb{R}$. So we can complete $\mathcal{H}$ by including all limits of Cauchy sequences (Hausdorff completion) to obtain a Hilbert space $\overline{\mathcal{H}}$. By construction, $\overline{\mathcal{H}}$ is a RKHS with reproducing kernel k.

Proposition 1.8. *If $\mathcal{G}$ is a RKHS of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\mathcal{G} \supseteq \mathcal{H}$, then $\overline{\mathcal{H}} = \mathcal{G}$.*

Proof. Example Sheet 1. □

Notation: from now on the RKHS is $\mathcal{H}$ (i.e $\mathcal{H} = \overline{\mathcal{H}}$).

Examples.

- Linear kernel: $k(x, x') = x^T x'$. Then $\mathcal{H} = \{f : f(x) = x^T \beta,\ \beta \in \mathbb{R}^p\}$. If $f(x) = x^T \beta$ then $||f||^2_{\mathcal{H}} = ||\beta||$.
- Sobolev kernel: $k(x, x') = \min(x, x')$ with $\mathcal{X} = [0, 1]$. Then $\mathcal{H}$ is the space of continuous functions $f : [0, 1] \to \mathbb{R}$ with $f(0) = 0$, for which

  $$\int_0^1 |f'(x)|^2 \mathrm{d}x < \infty$$

  where f' is the weak derivative.

The Representer Theorem

If $\mathcal{H}$ is the RKHS of the linear kernel, we can express ridge regression as

$$\hat{f} = \text{argmin}_{f \in \mathcal{H}} \left\{ \sum_{i=1}^{n} (Y_i - \underbrace{f(x_i)}_{x_i^T\beta})^2 + \lambda \underbrace{||f||_{\mathcal{H}}^2}_{\|\beta\|^2} \right\}.$$

In *kernel ridge regression*, we solve this problem in a more general RKHS with kernel k, e.g the Gaussian kernel.

Theorem 1.9 (Representer Theorem)**.** *Let:*

- $c : \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \to \mathbb{R}$ *be an arbitrary loss;*
- $J : [0, \infty) \to \mathbb{R}$ *be strictly increasing;*
- $x_1, \ldots, x_n \in \mathcal{X}$, $Y \in \mathbb{R}^n$*;*
- $\mathcal{H}$ *an RKHS with representing kernel* k*;*
- $K_{ij} = k(x_i, x_j)$, $i, j \in [n]$.

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

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

Example. In kernel ridge regression we just need to solve the quadratic program

$$\hat{\alpha} = \text{argmin}_{\alpha \in \mathbb{R}^n} \|Y - K\alpha\|^2 + \lambda \alpha^T K \alpha = (K + I\lambda)^{-1}.$$

Then the fitted values are given by $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$.

Intuition: to make a prediction at "test" point x^* the terms in $\hat{f}(x^*)$ that contribute the most are those for training points x_i with similarity $k(x^*, x_i)$ large.

Proof of the Representer Theorem. Note $V = \text{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$ is a closed (as its finite dimensional) subspace of $\mathcal{H}$. Hence any $f \in \mathcal{H}$ can be written as $f = u + v$ for $u \in \mathcal{V}$ and $v \in \mathcal{V}^\perp$.

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

$$||f||_{\mathcal{H}}^2 = ||v||_{\mathcal{H}}^2 + ||u||_{\mathcal{H}}^2.$$

In the expression for Q_1, the first term only depends on u, and the second term is $J(||f||_{\mathcal{H}}^2) \geq J(||u||_{\mathcal{H}}^2)$ with equality if and only if $v = 0$. Hence any minimiser of Q_1 is contained in $\mathcal{V}$.

So write $f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ for the minimiser. Now note

$$(f(x_1), \ldots, f(x_n)) = K\alpha$$

$$||f||_{\mathcal{H}}^2 = \sum_{i,j=1}^n \alpha_i k(x_i, x_j)\alpha_j = \alpha^T K\alpha$$

so therefore for any $f \in \mathcal{V}$, $Q_1(f) = Q_2(\alpha)$. Hence $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$ minimises Q_1 if and only if $\hat{\alpha}$ minimises Q_2. $\square$

Now we will assume that

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

where $||f^0||_{\mathcal{H}} \leq 1$.

Note. This is equivalent to $cY_i = cf^0(x_i) + c\varepsilon_i$ so $||cf^0||_{\mathcal{H}} = c||f^0||_{\mathcal{H}}$, $\mathrm{Var}(c\varepsilon_i) = \sigma^2 c^2$. So the "signal-to-noise ratio" is

$$\frac{\mathrm{Var}(c\varepsilon_i)}{||cf^0||^2_{\mathcal{H}}} = \frac{\mathrm{Var}(\varepsilon_i)}{||f^0||_{\mathcal{H}}} \geq \sigma^2.$$

Theorem 1.10. *Let K have eigenvalues $d_1 \geq d_2 \geq \ldots \geq d_n \geq 0$. Then*

$$\begin{aligned}
\mathrm{MSPE}(\hat{f}_n) &= \frac{1}{n}\mathbb{E}\left\{\sum_{i=1}^n (f^0(x_i) - \hat{f}_n(x_i))^2\right\} \\
&\leq \frac{\sigma^2}{n}\sum_{i=1}^n \frac{d_i^2}{(d_i+\lambda)^2} + \frac{\lambda}{4n} \\
&\leq \frac{\sigma^2}{n}\frac{1}{\lambda}\sum_{i=1}^n \min\left(\frac{d_i}{4}, \lambda\right) + \frac{\lambda}{4n}.
\end{aligned}$$

Proof. From the Representer Theorem $(\hat{f}_n(x_1), \ldots, \hat{f}_n(x_n))^T = K(K+\lambda I)^{-1}Y$. As $f^0 \in \mathcal{H}$ we have $(f^0(x_1), \ldots, f^0(x_n))^T = K\alpha$ for some $\alpha \in \mathbb{R}^n$ (see Example Sheet). Moreover, $||f^0||^2_{\mathcal{H}} \geq \alpha^T K \alpha$. Let the UDU^T be the eigen-decomposition of K, with $D_{ii} = d_i$. Define $\Theta = U^T K \alpha$. Then

$$\begin{aligned}
n\mathrm{MSPE}(\hat{f}_n) &= \mathbb{E}\|K(K+\lambda I)^{-1}\underbrace{(U\Theta + \varepsilon)}_{Y} - \underbrace{U\Theta}_{(f^0(x_1),\ldots,f^0(x_n))^T}\|^2 \\
&= \mathbb{E}\|UDU^T(UDU^T + \lambda I)^{-1}(U\Theta + \varepsilon) - U\Theta\|^2 \\
&= \mathbb{E}\|D(D+\lambda I)^{-1}(\Theta + U^T\varepsilon) - \Theta\|^2 && (U^TU = I) \\
&= \underbrace{\|\{D(D+\lambda I)^{-1} - I\}\varepsilon\|^2}_{:=(1)} + \underbrace{\mathbb{E}\|D(D+\lambda I)^{-1}U^T\varepsilon\|^2}_{:=(2)}. && (\mathbb{E}\varepsilon = 0)
\end{aligned}$$

So

$$\begin{aligned}
(2) &= \mathbb{E}[\{D(D+\lambda I)^{-1}U^T\varepsilon\}^T\{D(D+\lambda I)^{-1}U^T\varepsilon\}] \\
&= \mathbb{E}[\mathrm{tr}(\{D(D+\lambda I)^{-1}U^T\varepsilon\}^T\{D(D+\lambda I)^{-1}U^T\varepsilon\})] \\
&= \mathbb{E}[\mathrm{tr}(D(D+\lambda I)^{1}\varepsilon\varepsilon^T D(D+\lambda I)^{-1})] && \text{(circular property of tr)} \\
&= \mathrm{tr}(D(D+\lambda I)^{-1}\sigma^2 I D(D+\lambda I)^{-1}) \\
&= \sigma^2 \sum_{i=1}^n \frac{d_i^2}{(d_i+\lambda)^2}.
\end{aligned}$$

Also

$$(1) = \sum_{i=1}^{n} \frac{\lambda^2 \Theta_i^2}{(d_i + \lambda)^2}.$$

Since $\Theta = DU^T\alpha$, so if $d_i = 0$ then $\Theta_i = 0$. So let D^+ be a diagonal matrix with $D^+_{ii} = \begin{cases} d_i^{-1} & \text{if } d_i \neq 0 \\ 0 & \text{otherwise} \end{cases}$.

Then,

$$\begin{aligned} \sum_{i:d_i>0} \frac{\Theta_i^2}{d_i} = \|\sqrt{D^+}\Theta\|^2 &= \alpha^T K U D^+ U^T K \alpha \\ &= \alpha^T U D D^+ D U^T \alpha \\ &= \alpha^T U D U^T \alpha && (DD^+D = D) \\ &= \alpha^T K \alpha \leq 1. \end{aligned}$$

Then

$$\begin{aligned} (1) = \sum_{i:d_i>0} \frac{\Theta_i^2}{d_i} \frac{d_i \lambda^2}{(d_i+\lambda)^2} &\leq \max_{1\leq i\leq n} \frac{d_i\lambda^2}{(d_i+\lambda)^2} \sum_{i:d_i>0} \frac{\Theta_i^2}{d_i} \\ &\leq \max_{1\leq i\leq n} \frac{d_i\lambda^2}{(d_i+\lambda)^2} \\ &\leq \frac{\lambda}{4}. && ((a+b)^2 \geq 4ab) \end{aligned}$$

Combining the bounds for (1) and (2) gives the first inequality. Finally, for the final inequality we note that

$$\frac{d_i^2}{(d_i+\lambda)^2} \leq \min\left\{1, \frac{d_i^2}{4d_i\lambda}\right\} = \frac{1}{\lambda}\min\left\{\lambda, \frac{d_i}{4}\right\}.$$

$\square$

Question: when is the upper bound good?

Random design

Let $(\mathcal{X}, \mathcal{B}, \mathbb{P})$ be a probability space, where $\mathcal{X}$ is a metric space, $\mathcal{B}$ is the Borel σ-algebra on $\mathcal{X}$. Assume that $x_1, \ldots, x_n \sim^{\text{iid}} \mathbb{P}$.

Theorem 1.11 (Mercer's Theorem)**.** *Under mild assumptions on $k, \mathbb{P}$, there is an orthonormal basis (e_i) of $\mathcal{L}^2(\mathbb{P})$, i.e*

$$\int_{\mathcal{X}} e_l(x) e_j(x) \mathrm{d}\mathbb{P}(x) = \mathbb{1}\{l = j\}$$

and eigenvalues (μ_i) *with* $\sum_{i=1}^n \mu_i < \infty$ *such that*

$$\mu_j e_j(x') = \int_{\mathcal{X}} k(x,x') e_j(x) \mathrm{d}\mathbb{P}(x).$$

Furthermore

$$k(x,x') = \sum_{l=1}^{\infty} \mu_l e_l(x) e_l(x')$$

and this series is absolutely convergent.

Proof. Not given. □

Let $\hat{\mu}_1, \dots, \hat{\mu}_n$ be (random) eigenvalues of K/n. As it turns out, when n is large $\hat{\mu}_i \approx \mu_i$. Let $\gamma = \lambda/n$, then a previous theorem gives

$$\mathrm{MSPE}(\hat{f}_{\gamma n}) \leq \frac{\sigma^2}{\gamma} \frac{1}{n} \sum_{i=1}^{n} \min\left(\frac{\hat{\mu}_i}{4}, \gamma\right) + \frac{\gamma}{4}.$$

Then the MSPE is a random variable depending on $x_1, \dots, x_n$.

Lemma 1.12.

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \min\left(\frac{\hat{\mu}_i}{4}, \gamma\right)\right) \leq \frac{1}{n}\sum_{i=1}^{\infty} \min\left(\frac{\mu_i}{4}, \gamma\right).$$

Proof. Not given. □

This lemma means we can bound

$$\underbrace{\mathbb{E}[\mathrm{MSPE}(\hat{f}_{n\gamma})]}_{\text{over } Y \text{ and } x_1,\dots,x_n} \leq \frac{\sigma^2}{\gamma} \frac{1}{n} \sum_{i=1}^{\infty} \min\left(\frac{\mu_i}{4}, \gamma\right) + \frac{\gamma}{4}. \qquad (*)$$

Theorem 1.13. *Under the assumptions of Mercer's Theorem, there is a sequence $(\gamma_n)_{n\geq 1}$ such that for fixed $\sigma^2 > 0$,*

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

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

$$\phi(\gamma) = \sum_{j=1}^{\infty} \min\left(\frac{\mu_j}{4}, \gamma\right).$$

Note ϕ is increasing, and as $\sum_{j=1}^{\infty}\mu_j < \infty$, $\lim_{\gamma\downarrow 0}\phi(\gamma) = 0$. Define $\gamma_n = n^{-1/2}\sqrt{\phi(n^{-1/2})}$, so $\gamma_n = o(n^{-1/2})$. Thus for n large enough, $\phi(\gamma_n) \leq \phi(n^{-1/2})$ and the upper bound in $(*)$ is

$$\sigma^2\frac{\phi(\gamma_n)}{n\gamma_n} + \frac{\gamma_n}{4} \leq \frac{\sigma^2\phi(n^{-1/2})}{n^{1/2}\sqrt{\phi(n^{-1/2})}} + o(n^{-1/2}) = o(n^{-1/2}).$$

□

When we know (μ_j), in some cases we can get a better bound on the MSPE.

Example. If k is the Sobolev kernel and $\mathbb{P}$ is the Lebesgue measure on $[0, 1]$, one can show that

$$\frac{\mu_i}{4} = \frac{1}{\pi^2(2i-1)^2}.$$

Then for any integer j,

$$\sum_{i=1}^{\infty} \min\left(\frac{\mu_i}{4}, \gamma_n\right) \leq \gamma_n j + \sum_{i=j+1}^{\infty}\frac{1}{\pi^2(2i-1)^2}.$$

So if we take $j = \frac{(\pi^2\gamma_n)^{-1/2}+1}{2}$ we get upper bound

$$\begin{aligned}&\frac{\gamma_n}{2}\left(\frac{1}{\sqrt{\pi^2\gamma_n}} + 1\right) + \frac{1}{\pi^2}\int_{\frac{(\pi^2\gamma_n)^{-1/2}+1}{2}}^{\infty}\frac{1}{(2x-1)^2}\mathrm{d}x \\ &= \mathcal{O}(\gamma_n^{1/2}) + \mathcal{O}(\gamma_n) = \mathcal{O}(\sqrt{\gamma_n}).\end{aligned}$$

By $(*)$ we have

$$\mathbb{E}(\text{MSPE}(\hat{f}_{\gamma_n,n})) \leq \mathcal{O}\left(\frac{\sigma^2}{n\gamma_n}\sqrt{\gamma_n} + \gamma_n\right).$$

Picking $\gamma_n \sim \left(\frac{\sigma^2}{n}\right)^{2/3}$ gives an error of at most $\mathcal{O}\left(\left(\frac{\sigma^2}{n}\right)^{2/3}\right)$.

Support Vector Machines

Suppose we have data $(x_i, Y_i)_{i\in[n]}$ where $x_i \in \mathbb{R}^p$, $Y_i \in \{-1, 1\}$. Suppose the two response classes can be separated by a hyperplane through the origin. Let β be a unit vector which is normal to the hyperplane.

There could be many separating hyperplanes. One way of choosing a single one of these is to maximise an empty margin, i.e

$$\max_{\substack{M>0 \\ \beta\in\mathbb{S}^{p-1}}} M \text{ subject to } Y_i x_i^T \beta \geq M \text{ for all } i \in [n].$$

Reparameterising by $\beta \to \beta/M$, this problem becomes

$$\max_{\beta\in\mathbb{R}^p} \frac{1}{\|\beta\|} \text{ subject to } Y_i x_i^T \beta \geq 1 \text{ for all } i \in [n]$$

or equivalently

$$\min_{\beta\in\mathbb{R}^p} \|\beta\|^2 \text{ subject to } Y_i x_i^T \beta \geq 1 \text{ for all } i \in [n].$$

Instead, what if just a few samples fall on the wrong side of the margin? A different estimator, known as a *support vector classifier* replaces the constraint $Y_i x_i^T \beta \geq 1$ with a penalty $\left[(1 - Y_i)x_i^T \beta\right]_+$.

Remark. This works even if there is no separating hyperplane.

So our problem is

$$\min_{\beta\in\mathbb{R}^p} \left[\lambda\|\beta\|^2 + \sum_{i=1}^{n}(1 - Y_i x_i^T \beta)_+\right].$$

λ is a tuning parameter which balances "maximum margin" objective and penalty.

In general, we may want to estimate a hyperplane which does not pass through the origin; $x^T\beta + \mu = 0$. We can define a similar optimisation:

$$\min_{\substack{\beta\in\mathbb{R}^p \\ \mu\in\mathbb{R}}} \left[\lambda\|\beta\|^2 + \sum_{i=1}^{n}(1 - Y_i(x_i^T \beta + \mu))_+\right].$$

If $\mathcal{H}$ is the RKHS for the linear kernel, this problem can be written as

$$(\hat{\mu}, \hat{f}) = \text{argmin}_{(\mu,f)\in\mathbb{R}\times\mathcal{H}} \left[\sum_{i=1}^{n}(1 - Y_i(f(x_i) + \mu))_+ + \lambda\|f\|_{\mathcal{H}}^2\right]$$

where $\hat{f}(x) = x^T\hat{\beta}$.

A *support vector machine* is defined by this optimisation with a generic RKHS $\mathcal{H}$ with reproducing kernel k.

Prediction: given $(\hat{\mu}, \hat{f})$ and a new input x^* we predict $\hat{Y}^* = \text{sgn}(\hat{f}(x^*) + \hat{\mu})$.

Note. In $\mathcal{X}$ the separating 'hyperplane' is not necessarily linear, but upon mapping (via ϕ) to $\mathcal{H}$ (i.e $x \mapsto \phi(x) = k(\cdot, x)$) it becomes a hyperplane since the class boundary $\{x \in \mathcal{X} : f(x) + \mu = 0\}$ is mapped to $\{k(\cdot, x) : \langle k(\cdot, x), f\rangle_{\mathcal{H}} + \mu = 0\}$.

Using a slight generalisation of the Representer Theorem (see Example Sheet 1), we can show that

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

where

$$(\hat{\alpha}, \hat{\mu}) = \text{argmin}_{(\alpha,\mu) \in \mathbb{R}^n \times \mathbb{R}} \sum_{i=1}^{n} (1 - Y_i(K_i^T \alpha + \mu))_+ + \lambda \alpha^T K \alpha$$

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

Remark. We can have $\hat{\alpha}_i = 0$ for some i, so we do not use the corresponding x_i at all in the estimator.

Kernel Logistic Regression

We have standard logistic regression

$$\log \frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = -1)} = x_i^T \beta.$$

Maximising the likelihood with (x_i, Y_i), $i \in [n]$ is equivalent to solving

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \log \left(1 + \exp(-Y_i x_i^T \beta)\right).$$

As in ridge regression, we may wish to penalise $\|\beta\|^2$:

$$\min_{\beta \in \mathbb{R}^p} \left[\sum_{i=1}^{n} \log \left(1 + \exp(-Y_i x_i^T \beta)\right) + \lambda \|\beta\|_2^2 \right].$$

This is the same as

$$\min_{f \in \mathcal{H}} \left[\sum_{i=1}^{n} \log \left(1 + \exp(-Y_i f(x_i))\right) + \lambda \|f\|_{\mathcal{H}}^2 \right]$$

where $\mathcal{H}$ is the linear RKHS.

In kernel logistic regression we build the class boundary $\hat{f}(\cdot)$ by solving this problem with an arbitrary RHKS.

Question: how does this compare with the Support Vector Machine?

In each case, the objective is

$$\sum_{i=1}^{n} l(Y_i f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2$$

with $l(z) = (1-z)_+$ and $l(z) = \log(1 + e^{-z})$ for the SVM and logistic regression respectively.

1.5 Large-scale Kernel Machines

Suppose for a kernel k, there is a feature map $\phi : \mathcal{X} \to \mathbb{R}^q$. Let $K_{ij} = k(x_i, x_j)$ and $\Phi = \begin{pmatrix} \phi(x_1) \\ \vdots \\ \phi(x_n) \end{pmatrix} \in \mathbb{R}^{n \times q}$ so that $K = \Phi\Phi^T$.

Consider kernel ridge regression. There are two ways of computing the fitted values:

$$K(\underbrace{K + I\lambda}_{n \times n})^{-1} Y \text{ or;}$$

$$\Phi(\underbrace{\Phi^T\Phi + \lambda I}_{q\times q})^{-1}\Phi^T Y.$$

These have costs $\mathcal{O}(n^3)$ and $\mathcal{O}(q^3 + nq^2)$ respectively. So when n is much larger than q, we want to use the latter expression.

In other kernel machines, it is helpful to have a low rank kernel matrix $K = \Phi\Phi^T$.

Example. Consider the optimisation problem resulting from the representer theorem

$$\min_{\alpha\in\mathbb{R}^n} \left[c(Y, x_1, \ldots, x_n, K\alpha) + \lambda\alpha^T K\alpha\right].$$

The gradient of the penalty term is $2\lambda K\alpha$. Computing this has cost $\mathcal{O}(n^2)$ (since K is $n \times n$), but if $K = \Phi\Phi^T$ $(q < n)$ we can compute $2\lambda\Phi\Phi^T\alpha$ in $\mathcal{O}(nq)$ iterations.

Problem: what if there is no feature map ϕ onto $\mathbb{R}^q$ with $q \ll n$? For example, the Gaussian kernel.

Idea: find an approximation $\hat{\Phi}$ such that $K \approx \hat{\Phi}\hat{\Phi}^T$. Our approach will be to develop a random feature map $\hat{\Psi} : \mathcal{X} \to \mathbb{R}^b$ satisfying

$$\mathbb{E}[\hat{\Psi}(x)^T\hat{\Psi}(x')] = k(x, x') \text{ for all } x, x' \in \mathcal{X}.$$

Then, we can let $\hat{\Psi}_i$, $i \in [L]$ be iid copies of $\hat{\Psi}$; define the approximate feature map

$$\hat{\phi} : x \mapsto \frac{1}{\sqrt{L}}(\hat{\Psi}_1(x), \ldots, \hat{\Psi}_L(x)) \in \mathbb{R}^{b\times L}.$$

Then $\hat{\phi}(x)^T\hat{\phi}(x') = \frac{1}{L}\sum_{i=1}^L \hat{\Psi}_i(x)^T\hat{\Psi}_i(x')$. In particular $\mathbb{E}[\hat{\phi}(x)^T\hat{\phi}(x)] = k(x, x')$ and $\mathrm{Var}[\hat{\phi}(x)^T\hat{\phi}(x)] = \mathcal{O}(L^{-1})$.

Then approximate $K \approx \hat{\Phi}\hat{\Phi}^T$ where $\hat{\Phi} = \begin{pmatrix}\hat{\phi}(x_1)\\ \vdots \\ \hat{\phi}(x_n)\end{pmatrix}$. In some cases the error

$$\|K - \hat{\Phi}^T\hat{\Phi}\|$$

is small with $Lb \ll n$.

Random Fourier Feature

Theorem 1.14 (Bochner)**.** *Let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a continuous kernel. Then k is shift-invariant (there exists h such that $k(x, x') = h(x - x')$) if and only if there exists $c > 0$ and some distribution F in $\mathbb{R}^p$ such that if $W \sim F$, then*

$$k(x, x') = c\mathbb{E}[e^{i(x-x')^T W}] = c\mathbb{E}[\cos((x - x')^T W)].$$

Proof. Not given. $\square$

Example. If $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ is the Gaussian kernel, then we have the representation in the theorem with $W \sim \mathcal{N}(0, \sigma^{-2} I)$.

We can use the theorem to construct a random feature map

$$\hat{\Psi}(x) = \sqrt{2c}\cos(W^T x + U) \in \mathbb{R}$$

where $W \sim F$, $U \sim \text{Unif}(-\pi, \pi)$ are independent.

Lemma 1.15.

$$\mathbb{E}(\hat{\Psi}(x)\hat{\Psi}(y)) = k(x, y) \textit{ for all } x, y \in \mathbb{R}^p.$$

Proof. The LHS is

$$\begin{aligned}
&2c\mathbb{E}[\cos(W^T x + U)\cos(W^T y + U)] \\
&= 2c\mathbb{E}\big[\big[\cos(W^T x)\cos U - \sin(W^T x)\sin U\big] \times \big[\cos(W^T y)\cos U - \sin(W^T y)\sin U\big]\big] \\
&= c\mathbb{E}[\cos(W^T x)\cos(W^T y) + \sin(W^T x)\sin(W^T y)] \qquad \text{(since } \mathbb{E}[\cos u \sin u] = 0\text{)} \\
&= c\mathbb{E}[\cos(W^T(x - y))] \\
&= k(x, y). \qquad \text{(Bochner's Theorem)}
\end{aligned}$$

$\square$

2 The Lasso & Beyond

Consider the standard linear model

$$Y = X\beta^0 + \varepsilon, \quad \mathbb{E}\varepsilon = 0, \operatorname{Var}\varepsilon = \sigma^2 I.$$

Then

$$\begin{aligned}
\mathrm{MSPE}(\hat{\beta}^{\mathrm{ols}}) &= \frac{1}{n}\mathbb{E}\|X\beta^0 - X\hat{\beta}^{\mathrm{ols}}\| \\
&= \frac{1}{n}\mathbb{E}[\mathrm{tr}((\beta^0 - \hat{\beta}^{\mathrm{ols}})(\beta^0 - \hat{\beta}^{\mathrm{ols}})X^T X^)] \\
&= \frac{1}{n}\,\mathrm{tr}\left(\underbrace{\mathbb{E}[(\beta^0 - \hat{\beta}^{\mathrm{ols}})(\beta^0 - \hat{\beta}^{\mathrm{ols}})^T]}_{\mathrm{Var}(\hat{\beta}^{\mathrm{ols}})} X^T X\right) \\
&= \frac{1}{n}\,\mathrm{tr}(\sigma^2 (X^X)^{-1} X^T X) \\
&= \frac{1}{n}\,\mathrm{tr}(\sigma^2 I_p) = \frac{\sigma^2 p}{n}.
\end{aligned}$$

Let $S\{k : \beta_k^0 \neq 0\}$ be the "relevant" predictors.

Question: what if $s := |S| \ll p$?

The model $Y = X_s\beta_s^0 + \varepsilon$, where X_s is the matrix with columns which are columns of X with index in S, and β_s^0 are the coefficients for predictors in S. So if we fit a model with design matrix X_s instead of X, we get MSPE $= \frac{\sigma^2 s}{n} \ll \frac{\sigma^2 p}{n}$.

In practice, we don't know S, but we can try to estimate it (variable selection).

Best subset regression

Fit every model with a subset $M \subseteq \{1, \ldots, p\}$ of the predictors. Then choose the best M by cross-validation.

Problem: there are 2^p possibilities, which is too large even for relatively small p.

Forward selection

This is a greedy way of approximating best subset regression.

1. Start by fitting intercept-only model;
2. Add to the model the predictor that decreases the sum-of-squares residuals the most;
3. Repeat step 2 until we have m predictors.

We treat m as a tuning parameter, chosen by cross-validation.

2.1 The Lasso

$$(\hat{\mu}^L_\lambda, \hat{\beta}^L_\lambda) \in \text{argmin}_{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p} \left[\frac{1}{2n}\|Y - \mu\mathbf{1} - X\beta\|^2 + \lambda\|\beta\|_1\right]$$

where $\|\beta\|_1 = \sum_{k=1}^p |\beta_k|$. As we did for ridge regression, we can remove μ by standardising the columns of X and centering the response Y:

$$\hat{\beta}^L_\lambda \in \text{argmin}_{\beta\in\mathbb{R}^p} \left[\frac{1}{2n}\|Y - X\beta\|^2 + \lambda\|\beta\|_1\right].$$

Note that $\hat{\beta}^L_\lambda$ minimises

$$\|Y - X\beta\|^2 \text{ subject to } \|\beta\|_1 \leq \|\hat{\beta}^L_\lambda\|_1.$$

Similarly, $\hat{\beta}^R_\lambda$ minimises

$$\|Y - X\beta\|^2 \text{ subject to } \|\beta\| \leq \|\hat{\beta}^R_\lambda\|.$$

Fact: in general, $\hat{\beta}^R_\lambda$ has all non-zero entries, whereas $\hat{\beta}^L_\lambda$ can have many entries equal to zero.

Prediction error of the Lasso (slow rate)

Assume the columns of X are standardised, and Y is the centred response

$$Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1}.$$

Further assume $\varepsilon \in \mathcal{N}(0, \sigma^2 I)$.

Theorem 2.1. *Let $\hat{\beta}$ be any Lasso solution with $\lambda = A\sigma\sqrt{\log(p)/n}$. Then with probability $\geq 1 - 2p^{-(A^2/2-1)}$,*

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

Remarks.

- Instead of bounding the MSPE (the expectation of the LHS) we bound the SPE with high-probability;
- This is called the "slow rate" with respect to n, since we know the MSPE usually decreases as $\mathcal{O}(n^{-1})$ (e.g MSPE of $\hat{\beta}^{\text{ols}}$).
- However, we trade the factor of p in the numerator with $\sqrt{\log p}\|\beta^0\|_1$, which can be much smaller than in OLS in general.
- We make no assumptions about X!

Lemma 2.2. *Let* $\|X^T\varepsilon\|_\infty = \max_k |(\varepsilon^T X)_k|$ *and let* $\Omega = \left\{\frac{\|X^T\varepsilon\|_\infty}{n} \leq \lambda\right\}$ *then*

$$\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/2-1)}.$$

Proof of slow rate. By the definition of $\hat{\beta}$

$$\frac{1}{2n}\|\underbrace{Y - X\hat{\beta}}_{X(\beta^0-\hat{\beta})+\varepsilon-\mathbf{1}\bar{\varepsilon}}\|^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|\underbrace{Y - X\beta^0}_{\varepsilon-\mathbf{1}\bar{\varepsilon}}\|^2 + \lambda\|\beta^0\|_1.$$

Since $X^T\mathbf{1} = 0$, rearranging terms (and using the previous lemma) gives

$$\begin{aligned}
\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|^2 &\leq \frac{1}{n}\varepsilon^T X(\beta^0 - \hat{\beta}) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1 \\
&\leq \|\varepsilon^T X\|_\infty\|\beta^0 - \hat{\beta}\|_1 + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1 \\
&\leq \lambda\left[\|\beta^0 - \hat{\beta}\|_1 + \|\beta^0\|_1 - \|\hat{\beta}\|_1\right]. \qquad \text{(On } \Omega\text{)}
\end{aligned}$$

Thus by the triangle inequality

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|^2 \leq 4\lambda\|\beta^0\|_1.$$

□

Concentration Inequalities

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

$$\mathbb{E}[e^{\alpha(W-\mathbb{E}W)}] \leq e^{\frac{\alpha^2\sigma^2}{2}}.$$

Proposition 2.3. *If W is σ-sub-Gaussian, then*

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

Proof. Apply Markov's inequality to $bbP(W-\mathbb{E}W \geq t) = \mathbb{P}(\exp(\alpha(W-\mathbb{E}W)) \geq \exp(\alpha t))$ and minimise over α (Chernoff bound). $\square$

All bounded random variables are sub-Gaussian.

Lemma 2.4. *If W is a random variable taking values in $[a, b]$, then it is $\left(\frac{b-a}{2}\right)$-sub-Gaussian.*

Proof. See Part III Topics in Statistical Theory. $\square$

Proposition 2.5. *Let $W_1, \ldots, W_n$ be independent random variables where W_i is σ_i-sub-Gaussian. Let $\gamma \in \mathbb{R}^n$. Then $\gamma^T W = \sum_{i=1}^n \gamma_i W_i$ is sub-Gaussian with parameter $\sqrt{\sum_{i=1}^n \gamma_i^2\sigma_i^2}$.*

Proof. Without loss of generality, assume $\mathbb{E}W_i = 0$ for all $i \in [n]$. Then

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

$\square$

Recall:

Lemma 2.6. *Let $\|X^T\varepsilon\|_\infty = \max_k |(\varepsilon^T X)_k|$ and let $\Omega = \left\{\frac{\|X^T\varepsilon\|_\infty}{n} \leq \lambda\right\}$ then*

$$\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/2-1)}.$$

We will prove a stronger result:

Lemma. *Suppose $(\varepsilon_i)_{i=1}^n$ are independent mean zero random variables, and are sub-Gaussian with common parameter σ. Let $\lambda = A\sigma\sqrt{\frac{\log p}{n}}$. Then*

$$\mathbb{P}\left(\frac{\|X^T\varepsilon\|}{n} \leq \lambda\right) \geq 1 - 2p^{-(A^2/2-1)}.$$

Proof. We have

$$\mathbb{P}\left(\frac{\|X^T\varepsilon\|_\infty}{n} > \lambda\right) \leq \sum_{j=1}^{p} \mathbb{P}\left(\frac{|X_j^T\varepsilon|}{n} > \lambda\right)$$
$$= \sum_{j=1}^{p} \left[\mathbb{P}\left(\frac{X_j^T}{n} > \lambda\right) + \mathbb{P}\left(-\frac{X_j^T\varepsilon}{n} > \lambda\right)\right].$$

By the previous proposition, $\pm\frac{X_j^T\varepsilon}{n}$ is mean zero sub-Gaussian with parameter $\left(\frac{\sigma^2\|X_j\|^2}{n}\right)^{1/2} = \frac{\sigma}{\sqrt{n}}$. Hence, the above expression is bounded above by

$$2p\exp\left(-\frac{\lambda^2}{(2\frac{\sigma^2}{n})}\right) = 2p\exp\left(-A^2\frac{\log p}{2}\right) = 2p^{1-A^2/2}.$$

□

Now we recall some facts from complex analysis.

Proposition 2.7. *Let $C \subseteq \mathbb{R}^d$ be convex.*

(i) Let $f_1, \ldots, f_m : C \to \mathbb{R}$ be convex and $c_1, \ldots, c_m \geq 0$. Then $c_1 f_1 + \ldots + c_m f_m$ is convex.

(ii) Let $A : \mathbb{R}^m \to \mathbb{R}^d$ be affine ($A(x) = Mx + b$). Let $D = A^{-1}(C) = \{x : A(x) \in C\}$ and let $f : C \to \mathbb{R}$ be convex. Then D is convex and the composition $f \circ A : D \to \mathbb{R}$, $x \mapsto f(A(x))$ is a convex function.

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

(a) f is convex if and only if its Hessian $H(x)$ is positive semi-definite for all $x \in C$;

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

Lagrangian method

Consider a problem of the form

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

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

$$: (x, \theta) = f(x) + \theta^T g(x), \ \ \theta \in \mathbb{R}^b.$$

Note that for all θ,

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

The Lagrangian method involves finding θ^* such that the minimiser x^* of the LHS in the above has $g(x^*) = 0$, in which case this is a minimiser of the original problem.

Subgradient

Definition. Given a convex $C \subseteq \mathbb{R}^d$, convex $f : C \to \mathbb{R}$, define the *subdifferential of f at x* $\partial f(x) \subseteq \mathbb{R}^d$ defined by

$$\partial f(x) = \{v \in \mathbb{R}^d : f(y) \geq f(x) + v^T(y - x) \quad \forall y \in C\}.$$

An element $v \in \partial f(x)$ is called a *subgradient* of f at x.

Proposition 2.8. *If $f : C \to \mathbb{R}$ is convex and differentiable at $x \in \operatorname{int}(C)$, then*

$$\partial f(x) = \{\nabla f(x)\}.$$

Proposition 2.9. *If $f, g : C \to \mathbb{R}$ are convex with $\operatorname{int}(C) \neq \emptyset$, then*

$$\partial(\alpha f)(x) = \alpha \partial f(x) = \{\alpha v : v \in \partial f(x)\}$$

for all $\alpha \in \mathbb{R}$. Also

$$\partial(f + g)(x) = \partial f(x) + \partial g(x) = \{v + w : v \in \partial f(x), w \in \partial g(x)\}.$$

Karush-Kuhn-Tucker (KKT) conditions

Proposition 2.10. *Given $f : C \to \mathbb{R}$ convex, $x^* \in \operatorname{argmin}_{x \in C} f(x)$ if and only if $0 \in \partial f(x^*)$.*

Proof. Trivial. □

Subdifferenital of $\|\cdot\|_1$

By the triangle inequality,

$$\|tx + (1-t)y\|_1 \le t\|x\|_1 + (1-t)\|y\|_1 \ \forall t \in (0,1)$$

so $\|cdot\|_1$ is convex. Let $A = \{k_1, \ldots, k_m\} \subseteq \{1, \ldots, d\}$. For $x \in \mathbb{R}^d$ we write x_A for the vector $(x_{k_1}, \ldots, x_{k_m}) \in \mathbb{R}^m$. For $X \in \mathbb{R}^{n \times d}$ we write X_A for the matrix with columns $(X_{k_1}, \ldots, X_{k_m})$. We also write x_{-j} and x_{-jk} for $x_{\{j\}^c}$ and $x_{\{j,k\}^c}$ respectively. Also X_A^T and X_A^{-1} denote $(X_A)^T$ and $(X_A)^{-1}$ respectively.

We also define the sgn function by

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

for $x_i \in \mathbb{R}$. For $x \in \mathbb{R}^d$ we define $\operatorname{sgn}(x) == (\operatorname{sgn}(x_1), \ldots, \operatorname{sgn}(x_d))$.

Proposition 2.11. *For $x \in \mathbb{R}^d$, let $A = \{j : x_j \neq 0\}$. Then*

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

Proof. Let $g_j : \mathbb{R}^d \to \mathbb{R}$ be defined by $x \mapsto |x_j|$. Define $g = \sum_{j=1}^d g_j$ so $g(x) = \|x\|_1$ for all $x \in \mathbb{R}^d$. Then by a previous proposition, $\partial g(x) = \sum_{j=1}^d \partial g_j(x)$. When $x_j \neq 0$, g_j is differentiable at x so $\partial g_j(x) = \{\partial g_j(x)\} = \{\operatorname{sgn}(x_j)e_j\}$, where $e_j \in \mathbb{R}^d$ has all entries 0 except jth entry 1.

When $x_j = 0$,

$$\begin{aligned} v \in \partial g(x_j) &\iff g_j(y) \ge g_j(x) + v^T(y-x) \ \forall y \in \mathbb{R}^d \\ &\iff |y_j| \ge |x_j| + v^T(y-x) \ \forall y \in \mathbb{R}^d \\ &\iff |y_j| \ge v^T(y-x) \ \forall y \in \mathbb{R}^d. \end{aligned}$$

We claim this holds if and only if $v_j \in [-1,1]$ and $v_{-j} = 0$. Indeed if $v_{-j} = 0$ the above becomes $|y_j| \ge v_j y_j$ which holds as long as $v_j \in [-1,1]$. Conversely, taking $y \in \mathbb{R}^d$ with $y_j = 0$ and $y_{-j} = x_{-j} + v_{-j}$ gives $0 = |y_j| \ge |v_{-j}|^2$, implying $v_{-j} = 0$. Taking $y \in \mathbb{R}^d$ with $y_{-j} = 0$ and $y_j = \operatorname{sgn}(v_j)$ gives $1 \ge |\operatorname{sgn}(v_j)| = |y_j| \ge v_j \operatorname{sgn}(v_j) = |v_j|$.

The proposition now follows from $\partial g(x) = \sum_{j=1}^d \partial g_j(x)$. □

Lasso solutions

We have

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

KKT says that $0 \in \partial Q_\lambda(\hat{\beta}_\lambda^L)$ for any solution $\hat{\beta}_\lambda^L$. This is equivalent to

$$-\frac{1}{n}X^T(Y - X\hat{\beta}_\lambda^L) + \lambda\hat{v} = 0$$

for $\hat{v} \in \partial\|\hat{\beta}_\lambda^L\|_1$, i.e $\|\hat{v}\|_\infty \leq 1$ and $\hat{v}_{\hat{S}_\lambda} = \operatorname{sgn}(\hat{\beta}^L_{\lambda,\hat{S}_\lambda})$, where $\hat{S}_\lambda = \{k : \hat{\beta}^L_{\lambda,k} \neq 0\}$.

It turns out the Lasso fitted values are unique!

Proposition 2.12. *Fix $\lambda > 0$, suppose $\beta^{(1)}, \beta^{(2)}$ are two lasso solutions. Then $X\beta^{(1)} = X\beta^{(2)}$.*

Proof. We have $Q_\lambda(\beta^{(1)}) = Q_\lambda(\beta^(2)) = c^*$, the optimal value of Q_λ. By strict convexity of $\|\cdot\|_2^2$,

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

with equality if and only if $X\beta^{(1)} = X\beta^{(2)}$. We'll show this is an equality. We construct a chain of inequalities

$$\begin{aligned}
c^* &\leq Q_\lambda\left(\frac{\beta^{(1)} + \beta^{(2)}}{2}\right) \\
&= \frac{1}{2n}\|Y - \frac{X\beta^{(1)}}{2} - \frac{X\beta^{(2)}}{2}\|_2^2 + \lambda\|\frac{\beta^{(1)}}{2} + \frac{\beta^{(2)}}{2}\|_1 \\
&\leq \frac{1}{4n}\|Y - X\beta^{(1)}\|_2^2 + \frac{1}{4n}\|Y - X\beta^{(2)}\|_2^2 + \lambda\|\frac{\beta^{(1)} + \beta^{(2)}}{2}\|_1 \qquad (*) \\
&\leq \frac{1}{4n}\|Y - X\beta^{(1)}\|_2^2 + \frac{1}{4n}\|Y - X\beta^{(2)}\|_2^2 + \lambda\|\frac{\beta^{(1)}}{2}\|_1 + \|\frac{\beta^{(2)}}{2}\|_1 \\
&= \frac{Q_\lambda(\beta^{(1)}) + Q_\lambda(\beta^{(2)})}{2} = c^*.
\end{aligned}$$

Hence all of these inequalities are in fact equalities. The equality at $(*)$ implies $X\beta^{(1)} = X\beta^{(2)}$. □

Define the *equicorrelation set*

$$\hat{E}_\lambda = \{k : \frac{1}{n}|X_k^T(Y - X\hat{\beta}_\lambda^L)| = \lambda\}$$

which is well defined, since $X\hat{\beta}_\lambda^L$ does not depend on the choice of $\hat{\beta}_\lambda^L$. The KKT conditions tell us that $\hat{E}_\lambda$ contains all non-zero entries in $\hat{\beta}_\lambda^L$, i.e $\hat{S}_\lambda$.

Proposition 2.13. *If $\operatorname{rank}(X_{\hat{E}_\lambda}) = |\hat{E}_\lambda|$, then $\hat{\beta}_\lambda^L$ is unique.*

Proof. Say $\beta^{(1)}, \beta^{(2)}$ are lasso solutions. We know $\beta^{(1)}_{-\hat{E}_\lambda} = \beta^{(2)}_{-\hat{E}_\lambda} = 0$ so

$$X(\beta^{(1)} - \beta^{(2)}) = X_{\hat{E}_\lambda}(\beta^{(1)}_{\hat{E}_\lambda} - \beta^{(2)}_{\hat{E}_\lambda}) = 0$$

by uniqueness of the fitted values. If $X_{\hat{E}_\lambda}$ has full column rank, the only solution is $\beta^{(1)}_{\hat{E}_\lambda} - \beta^{(2)}_{\hat{E}_\lambda} = 0$. □

Variable selection

How good is the lasso at recovering the non-zero entries of β^0? When do we have $\text{sgn}(\hat{\beta}^L_\lambda) = \text{sgn}(\beta^0)$?

In the lectures we'll deal with the noiseless case $Y = X\beta^0$ (see Example Sheet for more general case).

Define $S = \{k : \beta^0_k \neq 0\}$ and $N = S^c$. Assume without loss of generality that $S = \{1, \ldots, s\}$. Assume $\text{rank}(X_s) = s$. Define $\Delta = X_N^T X_S (X_S^T X_S)^{-1} \text{sgn}(\beta^0_S)$ so $\Delta_k = ((X_S^T X_S)^{-1} X_S^T X_k) \,\text{sgn}(\beta^0_S)$, for $k > s$. Note that $(X_S^T X_S)^{-1} X_S^T X_k$ represents the coefficients in regression of non-significant variables X_k onto significant variables X_S.

Fix $\lambda > 0$ and define conditions

(A) $\|\Delta\|_\infty \leq 1$ (irrepresentable condition);

(B) $|\beta_k^0| > \lambda |\operatorname{sgn}(\beta_S^0)^T (X_S^T X_S)_k^{-1}|$ for all $k \in S$ (strong signal);

(C) There exists a lasso solution $\hat{\beta}_\lambda^L$ with $\operatorname{sgn}(\hat{\beta}_\lambda^L) = \operatorname{sgn}(\beta^0)$.

Theorem 2.14. *A & B imply C, and C implies A.*

Proof. Write $\hat{\beta} = \hat{\beta}_\lambda^L$, $\hat{S} = \{k : \hat{\beta} \neq 0\}$. The KKT conditions for the lasso say

$$\frac{1}{n} X^T X(\beta^0 - \hat{\beta})$$

where $\|\hat{v}\|_\infty \leq 1$ and $v_{\hat{S}} = \operatorname{sgn}(\hat{\beta}_S)$. Blockwise, this is

$$\frac{1}{n}\begin{pmatrix} X_S^T X_S & X_S^T X_N \\ X_N^T X_S & X_N^T X_N \end{pmatrix}\begin{pmatrix} \beta_S^0 - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} = \lambda \begin{pmatrix} \hat{v}_S \\ \hat{v}_N \end{pmatrix}. \tag{(I)}$$

First we show C implies A: if $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^0)$ then $\hat{v}_S = \operatorname{sgn}(\beta_S^0)$ and $\hat{\beta}_N = 0$. The top block of (I) gives

$$\frac{1}{n} X_S^T X_S(\beta_S^0 - \hat{\beta}_S) = \lambda \operatorname{sgn}(\beta_S^0).$$

Since X_S has full rank, $X_S^T X_S$ is invertible and $\beta_S^0 - \hat{\beta}_S = n\lambda (X_S^T X_S)^{-1} \operatorname{sgn}(\beta_S^0)$. Plugging this into the bottom block of (I) gives

$$\lambda \frac{1}{n} X_N T X_S(n(X_S^T X_S)^{-1} \operatorname{sgn}(\beta_S^0)) = \lambda \hat{v}_N$$

so

$$\Delta = X_N^T X_S (X_S^T X_S)^{-1} \operatorname{sgn}(\beta_S^0) = \hat{v}_N.$$

Since $\|v\|_\infty \leq 1$ we have $\|\Delta\|_\infty \leq 1$, i.e (A).

Now we show A & B imply C: it is enough to exhibit $(\hat{\beta}, \hat{v})$ satisfying (I) with $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^0)$. Take

$$\begin{aligned} (\hat{\beta}_S, \hat{\beta}_N) &= (\beta_S^0 - \lambda n (X_S^T X_S)^{-1} \operatorname{sgn}(\beta_S^0), 0) \\ (\hat{v}_S, \hat{v}_N) &= (\operatorname{sgn}(\beta_S^0), \Delta). \end{aligned}$$

It is easy to verify this solves (I). By condition (A), $\|\Delta\|_\infty \leq 1$ so $\|\hat{v}\|_\infty \leq 1$. Now we just check $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^0)$. Condition (B) gives that $\operatorname{sgn}(\beta_S^0) = \operatorname{sgn}(\hat{\beta}_S)$, and $\hat{\beta}_N = 0 = \beta_N^0$. Finally, we need $\hat{v}_{\hat{S}} = \operatorname{sgn}(\hat{\beta}_{\hat{S}})$. But by construction, $\hat{S} = S$ and we showed that $\hat{v}_S := \operatorname{sgn}(\beta_S^0) = \operatorname{sgn}(\hat{\beta}_S)$. □

Prediction & Estimation (fast rates)

Return to a noisy linear model

$$Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1}$$

where ε_i are independent and σ-sub-Gaussian random variables.

Definition. Given $X \in \mathbb{R}^{n\times p}$ and $S \subseteq \{1,\dots,p\}$ non-empty, define the *compatibility factor*

$$\phi^2 = \inf_{\substack{\delta\in\mathbb{R}^p \\ \delta_S\neq 0 \\ \|\delta_N\|_1\leq 3\|\delta_S\|_1}} \frac{\frac{1}{n}\|X\delta\|_2^2}{\frac{1}{s}\|\delta_S\|_1^2}$$

for $s := |S|$.

Remark. This is similar to a variational characterisation of the smallest eigenvalue $c_{\min}$ of $\frac{1}{n}X^TX$:

$$c_{\min} = \inf_{\substack{\delta\in\mathbb{R}^p \\ \delta\neq 0}} \frac{\delta^T(\frac{1}{n}X^TX)\delta}{\delta^T\delta} = \inf_{\substack{\delta\in\mathbb{R}^p \\ \delta\neq 0}} \frac{\frac{1}{n}\|X\delta\|_2^2}{\|\delta\|_2^2}.$$

ϕ^2 is sometimes called a “restricted eigenvalue”. Note that

$$\|\delta_S\|_1 = \operatorname{sgn}(\delta_S)^T\delta_S \leq \sqrt{s}\|\delta_S\|_2 \leq \sqrt{s}\|\delta\|_2.$$

Hence $\phi^2 \geq c_{\min}$.

The compatibility condition says $\phi^2 > 0$. In high-dimensions ($p > n$) we have $c_{\min} = 0$, but we can have $\phi^2 > 0$ for some S.

Theorem 2.15. *Suppose $\phi^2 > 0$, let $\lambda^* = A\sigma\sqrt{\frac{\log p}{n}}$ with $A > 2\sqrt{2}$. Then with probability $\geq 1 - 2p^{-(A^2/8-1)}$ we have for all $\lambda \geq \lambda^*$ that*

$$\underbrace{\frac{1}{n}\|X(\beta^0 - \hat{\beta}^L_\lambda)\|_2^2}_{\textit{prediction}} + \underbrace{\lambda\|\beta^0 - \hat{\beta}^L_\lambda\|_1}_{\textit{estimation}} \leq \frac{16\lambda^2 s}{\phi^2}.$$

Corollary 2.16. *If $\lambda = \lambda^*$,*

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta}^L_{\lambda^*})\|_2^2 \leq \frac{16A^2\log p}{\phi^2}\frac{\sigma^2 s}{n}$$

and

$$\|\beta^0 - \hat{\beta}^L_{\lambda^*}\|_1 \leq \frac{16A\sigma s}{\phi^2}\sqrt{\frac{\log p}{n}}.$$

Proof of theorem. Let $\hat{\beta} = \hat{\beta}^L_\lambda$. We have from a previous result that $\Omega = \{\frac{2\|X^T\varepsilon\|_\infty}{n} \leq \lambda\}$ has $\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/8-1)}$ and on Ω

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

Let $a = \|X(\hat{\beta} - \beta^0)\|^2/(n\lambda)$. Dividing (II) through by λ gives

$$a + 2(\|\hat{\beta}_S\|_1 + \|\hat{\beta}_N\|_1) \leq \|\hat{\beta}_S - \beta^0_S\|_1 + \|\hat{\beta}_N\|_1 + 2\|\beta^0_S\|_1.$$

Hence

$$\begin{aligned} a + \|\hat{\beta}_N\|_1 &\leq \|\hat{\beta}_S - \beta^0_S\|_1 + 2\|\beta^0_S\|_1 - 2\|\hat{\beta}_S\|_1 \\ &\leq \|\hat{\beta}_S - \beta^0_S\|_1 + 2\|\hat{\beta}_S - \beta^0_S\|_1 \\ &= 3\|\hat{\beta}_S - \beta^0_S\|_1 \end{aligned}$$

so adding $\|\hat{\beta}_S - \beta^0_S\|_1$ to both sides gives

$$a + \|\hat{\beta} - \beta^0\|_1 \leq 4\|\hat{\beta}_S - \beta^0_S\|_1. \tag{(III)}$$

Let $\delta = \hat{\beta} - \beta^0$, so we have $\|\delta_N\|_1 \leq 3\|\delta_S\|_1$ and hence

$$\phi^2 \leq \frac{\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2}{\frac{1}{s}\|\hat{\beta}_S - \beta^0_S\|_1} \implies \|\hat{\beta}_S - \beta^0_S\|_1 \leq \sqrt{\frac{s}{n}}\frac{1}{\phi}\|X(\beta^0 - \hat{\beta})\|_2.$$

Plugging this into (III):

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\beta^0 - \hat{\beta}\|_1 \leq \frac{4\lambda}{\phi}\sqrt{\frac{s}{n}}\|X(\hat{\beta} - \beta^0)\|_2. \tag{(IV)}$$

This implies

$$\frac{1}{\sqrt{n}}\|X(\hat{\beta} - \beta^0)\|_2 \leq \frac{4\lambda s}{\phi}$$

so plugging into (IV) gives the desired inequality. $\square$

The compatibility condition

- $\phi^2 > 0$ is a weaker assumption that the smallest eigenvalue of $\frac{1}{n}X^TX$, $c_{\min}$, is positive;
- If $p > n$, $c_{\min} = 0$ but we can have $\phi^2 > 0$;
- We want to prove that if the rows x_i of X are centered iid random variables, under certain conditions on covariance $\Sigma^0 = \mathbb{E}[x_1 x_1^T] = \mathbb{E}[\frac{1}{n}XX^T]$, and certain assumptions on the tails of x_1, we can have $\phi^2 > 0$ for *all* subsets S of a given size, with high probability.

Define

$$\phi_\Sigma^2(S) = \inf_{\substack{\delta:\delta_S \neq 0 \\ \|\delta_N\|_1 \leq 3\|\delta_S\|_1}} \frac{\delta^T \Sigma \delta}{\frac{1}{|S|}\|\delta_S\|_1^2}.$$

Lemma 2.17. *Suppose* $\phi_\Theta^2(S) > 0$ *and* Σ *is such that* $\max_{j,k} |\Theta_{jk} - \Sigma_{jk}| \leq \frac{\phi_\Theta^2(S)}{32|S|}$. *Then* $\phi_\Sigma^2(S) \geq \frac{\phi_\Theta^2(S)}{2}$.

Proof. For simplicity, we will neglect dependence on S. Write $s = |S|$ and define $B = \{\delta \in \mathbb{R}^p : \|\delta_S\|_1 = 1, \|\delta_N\|_1 \leq 3\}$. Note

$$\phi_\Theta^2(S) = |S| \inf_{\delta \in B} \delta^T \Theta \delta \ \forall \Theta.$$

Now for $\delta \in B$

$$\begin{aligned} s\delta^T \Sigma \delta &= s\delta^T \Theta \delta - s\delta^T(\Theta - \Sigma)\delta \\ &\geq \phi_\Theta^2 - s|\delta^T(\Theta - \Sigma)\delta|. \end{aligned}$$

Note

$$\begin{aligned} |\delta^T(\Theta - \Sigma)\delta| &\leq \|\delta\|_1 \|(\Theta - \Sigma)\delta\|_\infty && \text{(Hölder)} \\ &\leq \|\delta\|_1^2 \max_{j,k} |\Theta_{jk} - \Sigma_{jk}| && \text{(Hölder)} \\ &\leq \|\delta\|_1^2 \frac{\phi_\Theta^2}{32s}. \end{aligned}$$

As $\delta \in B$ we have $\|\delta\|_1 = \|\delta_S\|_1 + \|\delta_N\|_1 \leq 4$. Hence,

$$s\delta^T \Sigma \delta \geq \phi_\Theta^2 - s\frac{4^2 \phi_\Theta^2}{32s} = \frac{\phi_\Theta^2}{2}$$

for any $\delta \in B$, so taking the infimum over $\delta \in B$ gives the result. □

Plan:

- If the rows of X are iid with covariance Σ^0 then $\hat{\Sigma} = \frac{1}{n}X^TX$ is an estimate of Σ^0;

- Use concentration inequalities to find high probability bounds for $\max_{j,k}\left|\Sigma^0_{jk} - \hat{\Sigma}_{jk}\right|$;
- Apply the previous lemma to argue $\phi^2_{\hat{\Sigma}} \geq \frac{1}{2}\phi^2_{\Sigma^0}$.

Question: when can we assume $\phi^2_{\Sigma^0} > 0$?

Example. Take $\Sigma^0 = I$ (predictors uncorrelated). Then $\phi^2_{\Sigma^0}(S) > c_{\min}(\Sigma^0) = 1$ for any subset $S \subseteq \{1, \ldots, p\}$.

Concentration Inequalities Continued

Goal: obtain tail bounds on products of sub-Gaussian random variables.

Definition. We say a random variable W satisfies the *Bernstein condition* with parameters (σ, b), where $\sigma, b > 0$ if

$$\mathbb{E}(|W - \mathbb{E}W|^k) \leq \frac{1}{2}k!\sigma^2 b^{k-2} \text{ for } k = 2, 3, \ldots.$$

Proposition 2.18 (Bernstein's inequality)**.** *Let $W_1, \ldots, W_n$ be independent random variables with mean μ. Suppose that W_i is Bernstein(σ, b) for all $i \in [n]$. Then*

$$\mathbb{E}[e^{\alpha(W_i - \mu)}] \leq \exp\left(\frac{\frac{\alpha^2\sigma^2}{2}}{1 - b|\alpha|}\right) \textit{ for all } |\alpha| < 1/b$$

and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n W_i - \mu \geq t\right) \leq \exp\left(-\frac{nt^2}{2(\sigma^2 + bt)}\right) \ \forall t > 0.$$

Proof. Fix i, let $W = W_i$. Then for $|\alpha| < 1/b$

$$\begin{aligned}
\mathbb{E}(e^{\alpha(W-\mu)}) &= \mathbb{E}\left(1 + \alpha(W - \mu) + \sum_{k=2}^{\infty}\frac{\alpha^k(W-\mu)^k}{k!}\right) \\
&\leq \mathbb{E}\left(1 + \sum_{k=2}^{\infty}\frac{|\alpha|^k|W-\mu|^k}{k!}\right) \\
&= 1 + \sum_{k=2}^{\infty}\frac{|\alpha|^k\mathbb{E}(|W-\mu|^k)}{k!} && \text{(Fubini)} \\
&\leq 1 + \frac{\sigma^2\alpha^2}{2}\sum_{k=2}^{\infty}|\alpha|^{k-2}b^{k-2} && \text{(Bernstein condition)} \\
&= 1 + \frac{\sigma^2\alpha^2}{2}\frac{1}{1 - |\alpha|b} && (|\alpha| < 1/b) \\
&\leq \exp\left(\frac{\alpha^2\sigma^2/2}{1 - |\alpha|b}\right).
\end{aligned}$$

For the tail bound we have

$$\begin{aligned}\mathbb{E}\left(e^{\alpha\sum_{i=1}^n \frac{W_i-\mu}{n}}\right) &= \prod_{i=1}^n \mathbb{E}\exp\left(\frac{\alpha(W_i-\mu)}{n}\right)\\ &\le \exp\left(\frac{n(\alpha/n)^2\sigma^2/2}{1-b|\alpha/n|}\right)\ \forall|\alpha/n|<1/b\end{aligned}$$

so by Markov's inequality

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n W_i-\mu\ge t\right) &= \mathbb{P}\left(\exp\left(\frac{\alpha}{n}\sum_{i=1}^n W_i-\mu\right)\ge e^{\alpha t}\right)\\ &\le \exp\left(n\frac{(\alpha/n)^2\sigma^2/2}{1-b|\alpha/n|}-\alpha t\right)\ \forall|\alpha/n|<1/b.\end{aligned}$$

Taking $\alpha/n=\frac{t}{bt+\sigma^2}\in(0,1/b)$ gives the result. $\square$

Lemma 2.19. *Let W, Z be sub-Gaussian with parameters σ_W, σ_Z respectively. Then WZ is Bernstein$(8\sigma_W\sigma_Z, 4\sigma_W\sigma_Z)$ [W, Z need not be independent].*

Proof. We have

$$\begin{aligned}\mathbb{E}(W^{2k}) = \mathbb{E}\left[\int_0^\infty \mathbb{1}\{y < w^{2k}\}\mathrm{d}y\right] &= \int_0^\infty \mathbb{P}(y < W^{2k})\mathrm{d}y \\ &= 2k\int_0^\infty t^{2k-1}\mathbb{P}(|W| > t)\mathrm{d}t \\ &\leq 4k\int_0^\infty t^{2k-1}\exp\left(-\frac{t^2}{2\sigma_W^2}\right)\mathrm{d}t \\ &= 2^{k+1}\sigma_W^{2k}k!\end{aligned}$$

by Fubini's Theorem. Also, for any random variable Y we have $\mathbb{E}|Y - \mathbb{E}Y|^k \leq 2^k\mathbb{E}[|Y|^k]$ by the binomial theorem and Jensen. Hence

$$\begin{aligned}\mathbb{E}[|WZ - \mathbb{E}[WZ]|^k] \leq 2^k\mathbb{E}[|WZ|^k] &\leq 2^k\sqrt{\mathbb{E}[W^{2k}]\mathbb{E}[Z^{2k}]} \\ &\leq k!2^{2k+1}(\sigma_W\sigma_Z)^k \\ &= \frac{k!}{2}(8\sigma_W\sigma_Z)^2(4\sigma_W\sigma_Z)^{k-2}.\end{aligned}$$

□

Random Design

Theorem 2.20. *Suppose the rows of X are iid, and each entry of X is v-sub-Gaussian. Let $\hat{\Sigma} = \frac{1}{n}X^TX$ and $\Sigma^0 = \mathbb{E}\hat{\Sigma}$. Define "worst case" compatibility factors:*

$$\phi^2_{\hat{\Sigma},s} = \min_{S:|S|=s}\phi^2_{\hat{\Sigma}}(S), \quad \phi^2_{\Sigma^0,s} = \min_{S:|S|=s}\phi^2_{\Sigma^0}(S).$$

Suppose $\phi^2_{\Sigma^0,s} > c > 0$. Then

$$\mathbb{P}(\sigma^2_{\hat{\Sigma},s} \geq \frac{\phi^2_{\Sigma^0,s}}{2}) \geq 1 - 2p^{\left(2 - M\frac{n}{s^2\log p}\right)}$$

for some constant $M > 0$ independent of n, s and p.

Remarks.

- This theorem says the compatibility condition holds uniformly over all S of size s with high probability if $\frac{n}{s^2\log p}$ is large, i.e $s \ll \sqrt{n/\log p}$.
- The condition $\phi^2_{\Sigma^0,s} > 0$ holds for c being the smallest eigenvalue of Σ^0.
- The theorem does <u>not</u> require X_{ij}, X_{ik} for $k \neq j$ to be independent or even uncorrelated.

Proof. By a previous lemma,

$$\begin{aligned}\mathbb{P}\left(\phi^2_{\hat{\Sigma},s} \geq \frac{\phi^2_{\Sigma^0,s}}{2}\right) &\geq \mathbb{P}\left(\max_{j,k}|\hat{\Sigma}_{jk} - \Sigma^0_{jk}| \leq \frac{\phi^2_{\Sigma^0,s}}{32s}\right)\\ &\geq 1 - p^2 \min_{j,k}\mathbb{P}\left(|\hat{\Sigma}_{jk} - \Sigma^0_{jk}| \geq \frac{\phi^2_{\Sigma^0,s}}{32s}\right)\end{aligned}$$

so it suffices to show

$$\min_{j,k}\mathbb{P}\left(|\hat{\Sigma}_{jk} - \Sigma^0_{jk}| \geq \frac{\phi^2_{\Sigma^0,s}}{32s}\right) \leq 2p^{-M\frac{n}{s^2\log p}}$$

for some $M > 0$. Indeed note

$$\mathbb{P}\left(|\hat{\Sigma}_{jk} - \Sigma^0_{jk}| \geq \frac{\Sigma^2_{\Sigma^0,s}}{32s}\right) = \mathbb{P}\left(\left|\sum_{i=1}^n \frac{X_{ij}X_{ik}}{n} - \Sigma^0_{jn}\right| \geq \frac{\phi^2_{\Sigma^0,s}}{32s}\right).$$

Since $X_{ij}X_{ik}$ is a product of independent v-sub-Gaussian random variables, it is Bernstein$(8v^2, 4v^2)$. By Bernstein's inequality we then have

$$\mathbb{P}\left(\left|\sum_{i=1}^n \frac{X_{ij}X_{ik}}{n} - \Sigma^0_{jn}\right| \geq \frac{\phi^2_{\Sigma^0,s}}{32s}\right) \leq 2\exp\left(-\frac{n\left(\frac{\phi^2_{\Sigma^0,s}}{32s}\right)^2}{2\left(64v^4 + 4v^2\left(\frac{\phi^2_{\Sigma^0,s}}{32s}\right)\right)}\right)$$

and

$$-\frac{n\left(\frac{\phi^2_{\Sigma^0,s}}{32s}\right)^2}{2\left(64v^4 + 4v^2\left(\frac{\phi^2_{\Sigma^0,s}}{32s}\right)\right)} = -\frac{n}{s^2}C'\frac{(\phi^2_{\Sigma^0,s})^2}{C'' + \frac{\phi^2_{\Sigma^0,s}}{s}}$$

and

$$\begin{aligned}C'\frac{(\phi^2_{\Sigma^0,s})^2}{C'' + \frac{\phi^2_{\Sigma^0,s}}{s}} &\geq \frac{(\phi^2_{\Sigma^0,s})^2}{C'' + \phi^2_{\Sigma^0,s}}\\ &\geq \frac{C^2}{C''+C}\end{aligned}$$

where the last inequality follows since the penultimate term is increasing in $\phi^2_{\Sigma^0,s}$ and $\phi^2_{\Sigma^0,s} > 0$. Hence the desired inequality holds with $M = C'\frac{C^2}{C''+C}$. □

Computing $\hat{\beta}^L_\lambda$

The lasso objective is of the form

$$f(x) = g(x) + \sum_{j=1}^d h_j(x_j), \quad x \in \mathbb{R}^d$$

whee g is convex and differentiable, h_j is convex.

Coordinate descent

Initialise at $x^{(0)} \in \mathbb{R}^d$. For each $m = 1, \ldots, M$ set

$$\begin{aligned}
x_1^{(m)} &= \operatorname{argmin}_{x_1 \in \mathbb{R}} f(x_1, x_2^{(m-1)}, x_3^{(m-1)}, \ldots, x_d^{(m-1)}) \\
x_2^{(m)} &= \operatorname{argmin}_{x_2 \in \mathbb{R}} f(x_1^{(m)}, x_2, x_3^{(m-1)}, \ldots, x_d^{(m-1)}) \\
&\vdots \\
x_d^{(m)} &= \operatorname{argmin}_{x_d \in \mathbb{R}} f(x_1^{(m)}, x_2^{(m)}, x_3^{(m)}, \ldots, x_d).
\end{aligned}$$

Lemma 2.21. *Suppose $A_0 = \{x \in \mathbb{R}^d : f(x) \leq f(x^{(0)})\}$ is compact. Then*

(i) $f(x^{(m)}) \to f(x^*)$ as $m \to \infty$, where $x^* \in \operatorname{argmin}_{x \in A_0} f(x)$;

(ii) If x^* is the unique minimiser, then $x^{(m)} \to x^*$.

Proof. Not given. □

For the lasso, each coordinate descent step is

$$\hat{\beta}_k^{(m)} = \operatorname{argmin}_{\beta \in \mathbb{R}} \left\{ \frac{1}{2n} \|R - X_k \beta\|^2 + \lambda |\beta| \right\}$$

where

$$R := Y - \sum_{j=1}^{k-1} X_j \hat{\beta}_j^{(m)} - \sum_{j=k+1}^{p} X_j \hat{\beta}_j^{(m-1)}.$$

This objective is strictly convex, hence it has a unique solution, which satisfies the following KKT condition:

$$-\frac{1}{n} X_k^T R + \hat{\beta}_k^{(m)} + \lambda \hat{v} = 0$$

where $\hat{v} \in [-1, 1]$, and if $\hat{\beta}_k^{(m)} \neq 0$, $\hat{v} = \operatorname{sgn}(\hat{\beta}_k^{(m)})$.

The solution is given by

$$\hat{\beta}_k^{(m)} = S_\lambda(X_k^T R / n)$$

where

$$S_\lambda(u) := \operatorname{sgn}(u)(|u| - \lambda)_+$$

is the "soft-thresholding" function.

For fixed λ, we can find $\hat{\beta}^L_\lambda$ by coordinate descent. Suppose we want to compute $\hat{\beta}^L_\lambda$ for λ in $\lambda_0 > \ldots > \lambda_L$.
Warm starts:

- Find $\hat{\beta}^L_{\lambda_0}$
- For $\ell = 1, \ldots, L$, find $\hat{\beta}^L_{\lambda_\ell}$ by CD initialised at $\hat{\beta}_{\lambda_{\ell-1}}$.

Active set strategy:

1. Initialise $A_\ell = \{k : \hat{\beta}^L_{\lambda_{\ell-1},k} \neq 0\}$
2. Perform coordinate descent only on coordinates in A_ℓ to obtain $\hat{\beta}$ (with $\hat{\beta}_j = 0$ for all $j \notin A_\ell$)
3. Let $V = \{k : |X_k^T(Y - X\hat{\beta}) > \lambda_\ell\}$ be the coordinates where the KKT conditions are violated.
4. If V is empty, set $\hat{\beta}_{\lambda_\ell} = \hat{\beta}$. Otherwise, update $A_\ell \to A_\ell \cup V$ and go back to step 2.

Extensions of the Lasso

The square-root Lasso

Prediction and estimation error bounds for the Lasso required setting $\lambda = A\sigma\sqrt{\log p/n}$. Although to do this, we need to know σ! Instead we can try to estimate σ. Define

$$\hat{\sigma}^L_\lambda = \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}^L_\lambda\|.$$

If $\hat{\beta}^L_\lambda$ is accurate, then $\hat{\sigma}^L_\lambda$ should be accurate.

Idea: find optimal λ by setting $\lambda = \lambda_0$, then iterating

- Find $\hat{\beta}$ by solving Lasso(λ)
- Set $\hat{\sigma} = \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}\|$
- Set $\lambda = A\hat{\sigma}\sqrt{\log p/n}$.

This is equivalent to alternating minimising with respect to β and σ, the objective

$$Q^{\mathrm{sq}}_\gamma(\beta, \sigma) = \frac{1}{2n\sigma}\|Y - X\beta\|^2 + \frac{\sigma}{2} + \gamma\|\beta\|_1$$

where $\gamma = A\sqrt{\log n/p}$, $\lambda = \gamma\sigma$.

A more direct route to minimising $Q_\gamma^{\mathrm{sq}}(\beta, \sigma)$ is to minimise

$$\min_{\sigma>0} Q_\gamma^{\mathrm{sq}}(\beta, \sigma) = \frac{1}{\sqrt{n}} \|Y - X\beta\| + \gamma \|\beta\|_1 := Q_\gamma^{\mathrm{sq}}(\beta)$$

with respect to β, provided $Y \neq X\beta$. A minimiser of $Q_\gamma^{\mathrm{sq}}(\beta)$ is called the *square-root Lasso estimator*, written $\hat{\beta}_\gamma^{\mathrm{sq}}$. Contrast the loss with

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

Define $\hat{\sigma}_\gamma^{\mathrm{sq}} = \frac{1}{\sqrt{n}} \|Y - X\hat{\beta}_\gamma^{\mathrm{sq}}\|$. From this definition, we can see that if $\lambda = \hat{\sigma}_\gamma^{\mathrm{sq}} \gamma$, then

$$\hat{\beta} \in \operatorname{argmin}_\beta Q_\gamma^{\mathrm{sq}}(\beta) \iff \hat{\beta} \in \operatorname{argmin}_\beta Q_\lambda^2(\beta).$$

Conclusion: square-root Lasso is simply a reparameterisation of the regular path of the Lasso.

Theorem 2.22. *Consider a model $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let $\hat{\beta}$ be a square-root Lasso estimator with $\gamma = B\sqrt{\log p/n}$, $B > 2\sqrt{2}$. Consider an asymptotic regime: $n, p \to \infty$ with $\frac{s \log p}{n} \to 0$ and compatibility factor $\phi^2 > c > 0$. Then with probability tending to 1,*

$$\frac{1}{n} \|X(\beta^0 - \hat{\beta})\|^2 \leq \frac{17B^2 \log p}{\phi^2} \frac{s\sigma^2}{n} \quad \textit{and}$$
$$\|\beta^0 - \hat{\beta}\|_1 \leq \frac{17B\sigma s}{\phi^2} \sqrt{\log p/n}.$$

Proof. Let $\hat{\sigma} := \hat{\sigma}_\gamma^{\mathrm{sq}} = \frac{1}{\sqrt{n}} \|Y - X\hat{\beta}\|$. So $\hat{\beta}$ is a Lasso estimator with parameter $\lambda = \hat{\sigma}\gamma$. Take $\lambda_j = \sigma A_j \sqrt{\log p/n}$ for $j = 1, 2, \ldots$, where $2\sqrt{2} < A_1 < B < A_2$ with $16A^2 \leq 17B^2$ (this implies $16A_2 < 17B$).

Strategy:

- Show that $\lambda_1 \leq \hat{\sigma}\gamma \leq \lambda_2$ with high probability.
- By a previous theorem, $\hat{\beta}_{\lambda_1}^L, \hat{\beta}_{\lambda_2}^L$ are "accurate".
- Deduce from this that $\hat{\beta} = \hat{\beta}_{\hat{\sigma}\gamma}^L$ is "accurate".

In the Example Sheet we'll show there exists a sequence $a_n \to 0$ such that on a sequence of events $\Omega_n^{(1)}$ with $\mathbb{P}(\Omega_n^{(1)}) \to 1$, we have

$$1 - a_n \leq \frac{\sigma}{\hat{\sigma}_{\lambda_j}^L} \leq 1 - a_n \text{ for } j = 1, 2.$$

Hence on $\Omega_n^{(1)}$, for n large enough,

$$\begin{aligned}\gamma_1 := \frac{\lambda_1}{\hat{\sigma}^L_{\lambda_1}} = \frac{\sigma}{\hat{\sigma}^L_{\lambda_1}} A_1 \sqrt{\frac{\log p}{n}} &\leq (1+a_n) A_1 \sqrt{\frac{\log p}{n}} \\ &\leq B\sqrt{\frac{\log p}{n}} \\ &= \gamma.\end{aligned}$$

And similarly,

$$\gamma \leq \gamma_2 := \frac{\lambda_2}{\hat{\sigma}^L_{\lambda_2}}.$$

Note $\hat{\sigma}^L_{\lambda_j} = \hat{\sigma}^{\text{sq}}_{\gamma_j}$ for $j = 1, 2$. We'll show $\gamma_1 \leq \gamma$ implies $\hat{\sigma}^{\text{sq}}_{\gamma_1} \leq \hat{\sigma}^{\text{sq}}_{\gamma}$. By optimality,

$$\frac{1}{\sqrt{n}}\|Y - X\hat{\beta}^{\text{sq}}_{\gamma_1}\| + \gamma_1\|\hat{\beta}^{\text{sq}}_{\gamma_1}\| \leq \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}^{\text{sq}}_{\gamma}\| + \gamma_1\|\hat{\beta}^{\text{sq}}_{\gamma}\|_1 \tag{(I)}$$

$$\frac{1}{\sqrt{n}}\|Y - X\hat{\beta}^{\text{sq}}_{\gamma}\| + \gamma_1\|\hat{\beta}^{\text{sq}}_{\gamma}\| \leq \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}^{\text{sq}}_{\gamma_1}\| + \gamma_1\|\hat{\beta}^{\text{sq}}_{\gamma_1}\|_1.$$

Adding these and rearranging,

$$(\underbrace{\gamma - \gamma_1}_{\geq 0})(\|\hat{\beta}^{\text{sq}}_{\gamma_1}\|_1 - \|\hat{\beta}^{\text{sq}}_{\gamma}\|_1) \geq 0$$

$$\implies \|\hat{\beta}^{\text{sq}}_{\gamma_1}\| \geq \|\hat{\beta}^{\text{sq}}_{\gamma}\|_1.$$

Plugging this into (I) gives

$$\hat{\sigma}^{\text{sq}}_{\gamma_1} = \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}^{\text{sq}}_{\gamma_1}\| \leq \frac{1}{\sqrt{n}}\|Y - X\hat{\beta}^{\text{sq}}_{\gamma}\| = \hat{\sigma}^{\text{sq}}_{\gamma}.$$

□