Modern Statistical Methods — Summary

Lucas Riedstra

November 16, 2020

Contents

Classical models rely on so-called "large n asymptotics" (where n is the sample size). This course focuses on the scenario where p, the number of variables, is larger or about as large as n. In this case, the classical theory breaks down, so we need new methods.

1 Kernel machines

We represent data are pairs $(Y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$ $(i = 1, \dots, n)$. The random variables Y_i are called the *responses*, and the (fixed) variables x_i are called *predictors*.

Recap 1.1. Let $X = (X_1, \dots, X_n)^\top$ be a multivariate random variable. Its distribution function is given by

$$F_X \colon \mathbb{R}^n \to [0,1] \colon \mathbf{x} \mapsto \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n).$$

Its expected value is given by

$$\mathbb{E}[X] \coloneqq (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])^\top \in \mathbb{R}^n.$$

Its covariance matrix is given by

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] = \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top.$$

The matrix $\mathrm{Var}[X]$ is symmetric positive semidefinite and satisfies $(\mathrm{Var}[X])_{ij} = \mathrm{Cov}(X_i, X_j)$.

Definition 1.2. In a *linear model*, we assume that

$$Y_i = x_i^\top \beta^0 + \varepsilon_i \quad (i = 1, \dots, n).$$

where $\beta \in \mathbb{R}^p$ is unknown and the multivariate random variable $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ satisfies $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I$.

Definition 1.3. For an estimator $\tilde{\beta}$ of β^0, its *mean squared error* (MSE) is given by

$$\mathbb{E}_{\beta^0, \sigma^2}\left[(\tilde{\beta} - \beta^0)(\tilde{\beta} - \beta^0)^\top\right] = \mathrm{Var}(\tilde{\beta}) + \left[\mathbb{E}(\tilde{\beta} - \beta^0)\right]\left[\mathbb{E}(\tilde{\beta} - \beta^0)\right]^\top.$$

Note that if $\tilde{\beta}$ is unbiased, the second term will disappear and the MSE is simply the variance.

Recap 1.4. The maximum likelihood estimator (MLE) in this model is the ordinary least squares (OLS) estimator $\hat{\beta}^{\mathrm{OLS}} = (X^\top X)^{-1} X^\top Y$, where the *design matrix* $X \in \mathbb{R}^{n \times p}$ is the matrix whose rows are the vectors x_i. This estimator only exists if X has full column rank, so in particular, it is required that $p \leq n$.

The Cramér-Rao lower bown states that, out of all unbiased estimators, the MLE has the optimal variance *asymptotically* (i.e., for $n \to \infty$).

1.1 Ridge regression

Definition 1.5. Let $\lambda \geq 0$, and let $\mathbf{1} \in \mathbb{R}^n$ be the all-ones vector. Then we define the *ridge regression* estimators

$$(\hat{\mu}_\lambda^{\mathrm{R}}, \hat{\beta}_\lambda^{\mathrm{R}}) \coloneqq \operatorname*{arg\,min}_{(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p} \left\{ \|Y - \mu\mathbf{1} - X\beta\|^2 + \lambda \|\beta\|^2 \right\},$$

where the used norm is the 2-norm. The parameter λ is called the *regularisation parameter*.

The parameter λ represents a penalty for large coefficients in the design matrix. The intercept is not penalised — this is because a shift in units should not affect the fitted values. However, $X\hat{\beta}$ is not invariant under scale transformations, so it is common practice to centre the columns of X to have mean 0, and then scale them to have ℓ_2-norm $\sqrt{n}$.

After that, we can compute $\hat{\mu}_\lambda^{\mathrm{R}}$ by taking the derivative:

$$\|Y - \mu\mathbf{1} - X\beta\|^2 = \sum_i (Y_i - \mu - \sum_j X_{ij}\beta_j)^2.$$

$$\frac{\partial}{\partial \mu}\|Y - \mu\mathbf{1} - X\beta\|^2 = -2\sum_i \left(Y_i - \mu - \sum_j X_{ij}\beta_j\right).$$

Setting this derivative equal to 0 yields

$$\begin{aligned} -2\sum_i \left(Y_i - \mu - \sum_j X_{ij}\beta_j\right) &= 0 \\ \sum_i Y_i - n\mu - \sum_j \beta_j(\sum_i X_{ij}) &= 0 \\ \sum_i Y_i - n\mu &= 0 \\ \mu = \frac{1}{n}\sum_i Y_i &= \bar{Y}. \end{aligned}$$

Therefore we conclude $\hat{\mu}_\lambda^{\mathrm{R}} = \bar{Y}$. After centering the responses (i.e. replacing Y_i by $Y_i - \bar{Y}$), the problem can be reduced to

$$\hat{\beta}_\lambda^{\mathrm{R}} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} Q(\beta) \coloneqq \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \left[\|Y - X\beta\|^2 + \lambda\|\beta\|^2\right].$$

Since $Q(\beta)$ is convex quadratic, there is a unique root, and to find it we compute

$$\nabla_\beta Q(\beta) = 2X^\top(Y - X\beta) + 2\lambda\beta = 0 \iff \beta = (X^\top X + \lambda I)^{-1}X^\top Y.$$

We conclude that $\hat{\beta}_\lambda^{\mathrm{R}} = (X^\top X + \lambda I)^{-1}X^\top Y$. Note that, even if X does not have full column rank, this estimator exists for all $\lambda > 0$. In fact, for λ sufficiently small, the ridge estimator outperforms the MLE in terms of mean squared error:

Theorem 1.6. *Fix β^0, σ^2, and assume that $\hat{\beta}^{\mathrm{OLS}}$ exists (i.e., X has full column rank). For some $\lambda > 0$ sufficiently small, it holds that the MSE of $\hat{\beta}^{\mathrm{OLS}}$ minus the MSE of $\hat{\beta}_\lambda^{\mathrm{R}}$ is positive definite.*

Proof. This is simply writing out the MSE's. In the end, we find that the result holds for $0 < \lambda < 2\sigma^2/\left\|\beta^0\right\|^2$. □

1.1.1 The SVD and PCA

Recap 1.7. Recall that any $X \in \mathbb{R}^{n\times p}$ can be factorised as $X = UDV^\top$, where U, V are $n \times n$ and $p \times p$ orthogonal matrices respectively, and $D \in \mathbb{R}^{n\times p}$ satisfies $D_{11} \geq \cdots \geq D_{mm} \geq 0$ where $m \coloneqq \min(n, p)$, and all other entries of D are 0. This is called the *singular value decomposition* or SVD of X.

If $n > p$, we can replace U by its first p columns and D by its first p rows to produce the so-called *thin SVD* of X. Then $U \in \mathbb{R}^{n\times p}$ has orthogo nal columns (so $U^\top U = I$) and $D \in \mathbb{R}^{p\times p}$

is square and diagonal.

Suppose $n \geq p$ and let $X = UDV^\top$ be the thin SVD of our design matrix X. Then we can write the fitted values from the ridge regression as follows:

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

Note that for OLS ($\lambda = 0$), this is simply the projection of Y onto the column space of X (if X has full column rank). If $\lambda > 0$, Y is still projected onto the column space of X, but the projection is shrunk in the directions of the left singular vectors, and the lower the corresponding singular value, the higher the shrinkage.

Principal component analysis Consider $v \in \mathbb{R}^p$ with norm 1, then since the columns of X have been centered, the sample mean of Xv is 0, and the sample variance is therefore

$$\frac{1}{n}\sum_i (Xv)_i^2 = \frac{1}{n}(Xv)^\top Xv = \frac{1}{n} v^\top X^\top Xv = \frac{1}{n} v^\top VD^2V^\top v.$$

Writing $a = V^\top v$ (with $\|a\| = 1$), we find

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

Therefore, we see that the above is maximised if $a = \pm e_1$, or equivalently $v = \pm V_1$. Therefore, V_1 determines which combination of columns of X has the largest variance (subject to having norm 1), and $XV_1 = D_{11}U_1$ is known as the *first principal component* of X. Analogously, it can be shown that $D_{22}U_2, \ldots, D_{pp}U_p$ have maximum variance D_{jj}^2/n, subject to being orthonormal to all earlier principal components.

We see that ridge regression shrinks Y most in the smaller principal components of X. Therefore it will work well if most of the information is in the larger principal components of X.

A comment on computation By analogous calculations as before, one cam compute $\hat{\beta}_\lambda^{\mathrm{R}} = V(D^2 + \lambda I)^{-1} DU^\top Y$. Since calculating the inverse of a diagonal matrix is trivial, we see that the complexity of computing $\hat{\beta}_\lambda^{\mathrm{R}}$ for any λ lies in $O(np)$. Of course, this is after computation of the SVD of X, which lies in $O(np\min(n, p))$.

1.2 v-fold cross validation

Of course, we are still left with the problem of choosing λ in ridge regression. We consider one possible way of doing so, namely v-fold cross validation, which is a general way of selection a good regression method from several competing methods. Here, we assume that our predictors are random, so that we have i.i.d. data pairs (x_i, Y_i) $(i = 1, \ldots, n)$. Suppose (x^*, Y^*) is a new data pair, independent of

(X, Y) and identically distributed. Ideally, we want to pick λ which minimises the prediction error (averaged over Y^* and x^*)

$$\mathbb{E}\left[\left(Y^* - (x^*)^\top \hat{\beta}^{\mathrm{R}}_\lambda(X, Y)\right)^2 \mid X, Y\right],$$

where the dependence of $\hat{\beta}^{\mathrm{R}}_\lambda$ on the training data (X, Y) is made explicit by denoting it $\hat{\beta}^{\mathrm{R}}_\lambda(X, Y)$.

This is impossible to minimise, but it may be possible to minimise the expected prediction error (averaged over the training data)

$$\mathbb{E}\left\{\mathbb{E}\left[\left(Y^* - (x^*)^\top \hat{\beta}^{\mathrm{R}}_\lambda(X, Y)\right)^2 \mid X, Y\right]\right\}. \tag{1}$$

This is still not possible to compute directly, but we estimate it using v-fold cross validation. Split the data into v groups or *folds* of roughly equal size $(X^{(1)}, Y^{(1)}), \ldots, (X^{(v)}, Y^{(v)})$ and let $(X^{(-k)}, Y^{(-k)})$ denote all data except that in the k-th fold. Then we define

$$\mathrm{CV}(\lambda) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - x_i^\top \hat{\beta}^{\mathrm{R}}_\lambda(X^{(-\kappa(i))}, Y^{(-\kappa(i))})\right]^2,$$

and choose the value of λ that minimises $\mathrm{CV}(\lambda)$.

The function $\mathrm{CV}(\lambda)$ is called the *out-of-sample error*, since the training data does not include x_i.

Recap 1.8. The *tower rule* states that for random variables X, Y we have $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$.

Note that for each i, we have

$$\mathbb{E}\left[\left\{Y_i - x_i^\top \hat{\beta}^{\mathrm{R}}_\lambda(X^{(-\kappa(i))}, Y^{(-\kappa(i))})\right\}^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\{Y_i - x_i^\top \hat{\beta}^{\mathrm{R}}_\lambda(X^{(-\kappa(i))}, Y^{(-\kappa(i))})\right\}^2 \mid X^{-\kappa(i)}, Y^{-\kappa(i)}\right]\right].$$

This equals the expected prediction error in eq. (1), except that the training data X, Y are replaced with a smaller data set.

We now have a bias-variance tradeoff in the size of the folds: if $v = n$ (known as "leave-one-out" cross-validation), the estimation will be almost unbiased, but the averaged quantities in $\mathrm{CV}(\lambda)$ will be highly correlated which leads to high variance. Typical choices of v are 5 or 10.

Instead of finding the single best λ, we can also aim to find the best weighted combination of λ's. For example, suppose λ is restricted to a grid $\lambda_1 > \cdots > \lambda_L$. Then we can use any nonnegative least-squares optimization algorithm to minimise

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \sum_{\ell=1}^{L} w_\ell x_i^\top \hat{\beta}^{\mathrm{R}}_{\lambda_\ell}(X^{(-\kappa(i))}, Y^{(-\kappa(i))})\right]^2,$$

over all $w \in \mathbb{R}^L_{\geq 0}$. This procedure is known as *stacking* and often outperforms cross-validation.

1.3 The kernel trick

We note that

$$X^\top(XX^\top + \lambda I) = (X^\top X + \lambda I)X^\top,$$

and multiplying from the left with $(X^\top X + \lambda I)^{-1}$ and from the right with $(XX^\top + \lambda I)^{-1}$ gives

$$(X^\top X + \lambda I)^{-1}X^\top = X^\top(XX^\top + \lambda I)^{-1}.$$

Using this, we see that we can rewrite the fitted values from ridge regression as follows:

$$X\hat{\beta}^R_\lambda = X(X^\top X + \lambda I)^{-1}X^\top Y = XX^\top(XX^\top + \lambda I)^{-1}Y.$$

Two important remarks:

1. Computing the LHS of this equation takes roughly $O(np^2 + p^3)$ operations, while computing the RHS takes $O(n^2p + n^3)$ operations (this is because in the LHS we invert an $p \times p$ matrix, while in the RHS we invert a $n \times n$ matrix). Therefore, if $p \gg n$, the RHS can be much cheaper to compute.

2. The LHS depends only on the matrix $K = XX^\top$ (this matrix is called the *kernel matrix*). Intuitively, since $K_{ij} = \langle x_i, x_j \rangle$, the entries of the kernel matrix show how 'similar' the corresponding predictors are.

Example 1.9. Suppose we have data $(Y_i, z_i)_{i=1,\dots,n}$ with $z_i = (z_{i1}, \dots, z_{id})^\top$, and we believe the following quadratic relation holds:

$$Y_i = \sum_k \sqrt{2}\gamma_k z_{ik} + \sum_{k,\ell} \vartheta_{k\ell} z_{ik} z_{i\ell} + \varepsilon_i.$$

To compute fitted values using ridge regression, we can rewrite this as a linear model $Y = X\beta + \varepsilon$ where

$$\beta = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_d \\ \vartheta_{11} \\ \vartheta_{12} \\ \vdots \\ \vartheta_{dd} \end{pmatrix}, \quad x_i = \begin{pmatrix} \sqrt{2}z_{i1} \\ \vdots \\ \sqrt{2}z_{id} \\ z_{i1}z_{i1} \\ z_{i1}z_{i2} \\ \vdots \\ z_{id}z_{id} \end{pmatrix}.$$

In this case, we have $p = d^2 + d$ variables, which means computing $(X^\top X + \lambda I)^{-1}$ takes $O(d^6)$ operations. In this case, computing $(XX^\top + \lambda I)^{-1}$ is probably easier.

We are still left with the problem of computing $K \coloneqq XX^\top$, which can take $O(n^2 p) = O(n^2 d^2)$ operations if done naively. However, observe that

$$K_{ij} = x_i^\top x_j = 2\sum_k z_{ik} z_{jk} + \sum_{k,\ell} z_{ik} z_{i\ell} z_{jk} z_{j\ell} = \left(1 + \sum_k z_{ik} z_{jk}\right)^2 - 1 = (1 + z_i^\top z_j) - 1.$$

This quantity can be computed in $O(d)$, and therefore K can be computed in $O(n^2 d)$ operations: we have a factor d improvement.

The general point of the previous example is that we can bypass the features x_i entirely and instead think directly of $K = XX^\top$ where an entry K_{ij} represents similarity between the inputs of the i-th and j-th samples. This leads to the notion of a kernel in general.

1.4 Kernels

We will assume our inputs $x_1, \dots, x_n$ live in an abstract space $\mathcal{X}$.

Definition 1.10. A (*positive-definite*) *kernel* is a symmetric map $k\colon \mathcal{X}^2 \to \mathbb{R}$ such that for all $n \in \mathbb{N}$ and all $x_1, \dots, x_n \in \mathcal{X}$, the matrix $K \in \mathbb{R}^{n \times n}$ with $K_{ij} = k(x_i, x_j)$ is positive semi-definite.

Proposition 1.11 (Cauchy-Schwarz for kernels)**.** *Let k be a kernel and $x, x' \in \mathcal{X}$, then*

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

Proof. The matrix $\left(\begin{smallmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{smallmatrix}\right)$ must be positive semi-definite so its determinant must be non-negative. □

In our old models, the data points x_i were vectors in $\mathbb{R}^p$. Now we try to think of them as points in an abstract space with an associated *feature map* $\varphi\colon \mathcal{X} \to \mathcal{H}$ (with $\mathcal{H}$ an inner product space), and a kernel $k(x, x')$ gives a measure of similarity between $\varphi(x)$ and $\varphi(x')$. In this case, we have the following:

Proposition 1.12. *Let $\mathcal{H}$ be an inner product space, $\varphi\colon \mathcal{X} \to \mathcal{H}$ and define $k(x, x') \coloneqq \langle \varphi(x), \varphi(x') \rangle$. Then k is a kernel.*

Proof. We have, for all $x_1, \dots, x_n \in \mathcal{X}$ and $\alpha \in \mathbb{R}^n$ that

$$\alpha^\top K \alpha = \sum_{i,j} K_{ij} \alpha_i \alpha_j = \sum_{i,j} \langle \varphi(x_i), \varphi(x_j) \rangle \alpha_i \alpha_j = \left\| \sum_i \alpha_i \varphi(x_i) \right\|^2 \geq 0.$$

□

1.4.1 Examples of kernels

The following proposition shows how to make new kernels from old:

Proposition 1.13. *Suppose $k_1, k_2, \dots$ are kernels. Then:*

1. *If $\alpha_1, \alpha_2 \geq 0$ then $\alpha_1 k_1 + \alpha_2 k_2$ is a kernel.*
2. *The pointwise limit of a sequence of kernels is a kernel (if it exists).*
3. *The pointwise product $k_1 k_2$ is a kernel.*

Proof. See Example Sheet 1. □

Example 1.14. Let us consider some examples of kernels:

1. For $\mathcal{X} = \mathbb{R}^p$ we have already seen the *linear kernel* $k(x, x') = x^\top x'$.
2. For $\mathcal{X} = \mathbb{R}^p$, the *polynomial kernel* is defined as $k(x, x') = (1 + x^\top x')^d$. This is a kernel since it is a power of a sum of two kernels.
3. For $\mathcal{X} = \mathbb{R}^p$, the *Gaussian kernel* is defined by

   $$k(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2} \right).$$

   To show this is a kernel, write k as the pointwise product $k_1 k_2$ where

   $$k_1(x, x') = \exp\left(-\frac{\|x\|^2}{2\sigma^2} \right) \exp\left(-\frac{\|x'\|^2}{2\sigma^2} \right), \quad k_2(x, x') = \exp\left(\frac{x^\top x'}{\sigma^2} \right).$$

   Clearly k_1 is the kernel induced by the feature map $\varphi(x) = \exp\Big(-\|x\|^2/(2\sigma^2)\Big)$, while k_2 can be seen to be a kernel by using the Taylor expansion, which shows that k_2 is a limit of nonnegative linear combinations of kernels.
4. For $\mathcal{X} = [0, 1]$, define the *Sobolev kernel* $k(x, x') = \min(x, x')$. The proof that this is a kernel is on example sheet 1.
5. For $\mathcal{X} = \mathcal{P}(\{1, \dots, p\})$, define the *Jaccard kernel*

   $$k(x, x') = \frac{|x \cap x'|}{|x \cup x'|} \quad \text{where } 0/0 \coloneqq 1.$$

   The proof that this is a kernel is on example sheet 1.

1.4.2 Reproducing kernel Hilbert spaces

By proposition 1.12, we see that every feature map $\varphi\colon \mathcal{X} \to \mathcal{H}$ gives rise to a kernel. In the next (important!) theorem, we will see that every kernel is in fact induced by a feature map.

Theorem 1.15. *Let k be a kernel, then there exists an inner product space $\mathcal{H}$ and a feature map $\varphi\colon \mathcal{X} \to \mathcal{H}$ such that*

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle \quad \textit{for all } x, x' \in \mathcal{X}.$$

Proof. We will construct $\mathcal{H}$ and φ explicitly. First we define the function space

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

Let $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{n} \beta_j k(\cdot, x'_j)$, then the inner product on $\mathcal{H}$ is given by

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

We define $\varphi\colon \mathcal{X} \to \mathcal{H}$ as $\varphi(x) = k(\cdot, x)$.

We must check that the inner product does not depend on the choice of representation of f and g. For this, note that

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

which holds by symmetry of the kernel. Since $\sum_i \alpha_i g(x_i)$ is independent of the representation of g, while $\sum_j \beta_j f(x'_j)$ is independent of the representation of f, we conclude that the entire expression is independent of both representations.

Secondly, we must verify that the formula $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ indeed holds. For any $f \in \mathcal{H}, x \in \mathcal{X}$ we have

$$\langle k(\cdot, x), f \rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x), \tag{2}$$

i.e., evaluation of a function is a linear functional in $\mathcal{H}$.

In particular, we have

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

Finally, we must check that $\langle \cdot, \cdot \rangle$ is indeed an inner product. Symmetry and bilinearity are clear. Furthermore, we have

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

by the fact that k is a kernel. We must now only show that $f \neq 0 \implies \langle f, f \rangle > 0$. For this, note that $\langle \cdot, \cdot \rangle$ is a kernel on $\mathcal{H}$, so by proposition 1.11 (Cauchy-Schwarz) we have

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

and therefore if f is nonzero anywhere, $\langle f, f \rangle$ must also be nonzero. □

While $\mathcal{H}$ constructed in the proof is an inner product space, it is not necessarily a Hilbert space. Let $(f_n) \subseteq \mathcal{H}$ be Cauchy, then by Cauchy-Schwarz for kernels we find

$$f_m(x) - f_n(x) = (f_m - f_n)(x) = \langle k(\cdot, x), f_m - f_n \rangle \leq \sqrt{k(x, x)} \| f_n - f_m \|.$$

We can do an analogous computation for $f_n - f_m$ to conclude that $|f_m(x) - f_n(x)| \leq \sqrt{k(x, x)} \| f_n - f_m \|$, and therefore, if (f_n) is Cauchy, then it converges pointwise to some $f^*\colon \mathcal{X} \to \mathbb{R}$. We will not prove the following theorem:

Theorem 1.16. *The inner product space $\mathcal{H}$ constructed in the proof of theorem 1.15 can be extended to a Hilbert space by adding all pointwise limits f^* of Cauchy sequences $(f_n) \subseteq \mathcal{H}$.*

The completion of $\mathcal{H}$ is a special type of Hilbert space:

Definition 1.17. A Hilbert space $\mathcal{B}$ of functions $f\colon \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel Hilbert space* (RKHS) if for all $x \in \mathcal{X}$, there exists $k_x \in \mathcal{B}$ such that

$$f(x) = \langle k_x, f \rangle,$$

i.e., evaluation of functions is a linear functional.

The function $k(x, x') = \langle k_x, k_{x'} \rangle$ is known as the *reproducing kernel* of $\mathcal{B}$ (induced by the feature map $\varphi(x) = k_x$).

If we start with a kernel k, construct the corresponding RKHS $\mathcal{B}$, then it is easily checked that k is indeed the reproducing kernel of $\mathcal{B}$.

Example 1.18 (Linear kernel)**.** Let $X = \mathbb{R}^p$ and $k(x, x') = x^\top x'$. Then we have

$$\mathcal{H} = \left\{ x \mapsto \sum_{i=1}^n \alpha_i x^\top x_i = x^\top \left(\sum_i \alpha_i x_i \right) \mid \alpha_i \in \mathbb{R}, x_i \in \mathbb{R}^p \right\} = \left\{ x \mapsto x^\top \beta \mid \beta \in \mathbb{R}^p \right\},$$

and if $f(x) = x^\top \beta$, $g(x) = x^\top \beta'$, then

$$\langle f, g \rangle = k(\beta, \beta') = \beta^\top \beta' \quad \text{so } \|f\|_{\mathcal{H}} = \|\beta\|_2.$$

1.4.3 The representer theorem

We started with data (Y_i, x_i) in a model where our fitted values depend only on $XX^\top = K$, where K is the kernel matrix derived from the kernel $k(x_i, x_j) = \langle x_i, x_j \rangle = x_i^\top x_j$. Suppose we change the kernel to some $\tilde{k}$, and replace K by $\tilde{K}$ accordingly. Then by theorem 1.15 this is equivalent by replacing our predictors x_i by $\varphi(x_i)$ for some feature map φ.

Recall that the ridge regression objective function is given by

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - x_i^\top \beta)^2 + \lambda \|\beta\|^2.$$

Let $\mathcal{H}$ be the RKHS of the linear kernel, then the above is equivalent to

$$\min_{f \in \mathcal{H}} \sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda \|f\|^2. \tag{3}$$

We will show that in general, if we change our predictors x_i to $\varphi(x_i)$, and update the corresponding RKHS in eq. (3), then eq. (3) solves our new problem. Note that there are no references to the feature map φ in eq. (3).

Theorem 1.19 (Representer theorem)**.** *Let:*

- $(Y_i, x_i)_{i=1}^n$ *be our data with* $x_i \in \mathcal{X}$*;*
- $c\colon \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \to \mathbb{R}$ *be an arbitrary loss function;*
- $J\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ *strictly increasing;*
- k *a kernel with RKHS* $\mathcal{H}$ *and kernel matrix* $K_{ij} = k(x_i, x_j)$*.*

Then $\hat{f} \in \mathcal{H}$ minimises

$$Q_1(f) \coloneqq c(Y, x_1, \dots, x_n, f(x_1), \dots, f(x_n))$$

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

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

Proof. Let $f \in \mathcal{H}$ and $U = \operatorname{Span}\{k(\cdot, x_1), \dots, k(\cdot, x_n)\} \subseteq \mathcal{H}$, then U is closed since it is finite-dimensional so we can write $f = u + v$ with $u \in U, v \in U^\perp$. Note that

$$f(x_i) = \langle k(\cdot, x_i), u + v\rangle = \langle k(\cdot, x_i), u\rangle = u(x_i) \quad (i = 1, \dots, n),$$

so minimisation of Q_1 is equivalent to minimisation of

$$c(Y, x_1, \dots, x_n, u(x_1), \dots, u(x_n)) + J(\|u\|^2 + \|v\|^2) \tag{4}$$

w.r.t. $u \in U$, $v \in U^\perp$. Now, since J is strictly increasing, any minimiser (u, v) of eq. (4) will satisfy $v = 0$.

Write $u(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i) \in U$, then clearly

$$\begin{bmatrix} u(x_1) \\ \vdots \\ u(x_n) \end{bmatrix} = \begin{bmatrix} \sum_i \alpha_i k(x_1, x_i) \\ \vdots \\ \sum_i \alpha_i k(x_n, x_i) \end{bmatrix} = K\alpha \quad \text{and} \quad \|u\|^2 = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \alpha^\top K\alpha.$$

This shows that minimisation of eq. (4) is equivalent to minimisation of

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

which completes the proof. □

Example 1.20. In ridge regression, the representer theorem tells us that minimising eq. (3) is equivalent to minimising $\|Y - K\alpha\|^2 + \lambda\alpha^\top K\alpha$, and indeed it can be shown (example sheet 1) that the minimiser satisfies $K\hat{\alpha} = K(K + \lambda I)^{-1}Y = X\hat{\beta}_\lambda^R$. Since $\mathcal{H}$ may be infinite-dimensional, this gives a way to rewrite an infinite-dimensional optimization problem to a finite-dimensional optimisation problem.

1.5 Kernel ridge regression

We have now defined kernel ridge regression and shown how the estimator may be computed, but we have yet to assess its predictive performance. We consider the model

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

where we assume $f^0 \in \mathcal{H}$ where $\mathcal{H}$ is an RKHS with reproducing kernel k. By scaling the equation on both sides we may assume $\|f^0\| \leq 1$ (note that this changes $\operatorname{Var}[\varepsilon]$). Let K be the kernel matrix with eigenvalues $d_1 \geq \cdots \geq d_n \geq 0$, and let $\hat{f} = \hat{f}_\lambda$ be the estimated regression function, so

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

Theorem 1.21. *The mean squared prediction error (MSPE) satisfies*

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

Proof. The representer theorem tells us that $(\hat{f}(x_1), \dots, \hat{f}(x_n))^\top = K(K+\lambda I)^{-1}Y$. By projecting f^0 onto $\operatorname{Span} k(\cdot, x_1), \dots, k(\cdot, x_n)$ it is easily seen that there exists $\alpha \in \mathbb{R}^n$ such that

$$(f^0(x_1), \dots, f^0(x_n))^\top = K\alpha \quad \text{and} \quad \|f^0\|^2 \geq \alpha^\top K \alpha.$$

Write $K = UDU^\top$ where $D_{ii} = d_i$ and define $\vartheta \coloneqq U^\top K\alpha = DU^\top \alpha$ (note $U\vartheta = K\alpha$ and note that $d_i = 0 \implies \vartheta_i = 0$). Then we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^n \left(f^0(x_i) - \hat{f}_\lambda(x_i)\right)^2\right] &= \mathbb{E}\|K\alpha - K(K+\lambda I)^{-1}Y\|^2 \\
&= \mathbb{E}\|K(K+\lambda I)^{-1}(U\vartheta + \varepsilon) - U\vartheta\|^2 \\
&= \mathbb{E}\|DU^\top(UDU^\top + \lambda I)^{-1}(U\vartheta + \varepsilon) - \vartheta\|^2 \\
&= \mathbb{E}\|D(D+\lambda I)^{-1}(\vartheta + U^\top\varepsilon) - \vartheta\|^2 \\
&= \mathbb{E}\|\left[D(D+\lambda I)^{-1} - I\right]\vartheta + D(D+\lambda I)^{-1}U^\top\varepsilon\|^2 \\
&= \|\left[D(D+\lambda I)^{-1} - I\right]\vartheta\|^2 + \mathbb{E}\|D(D+\lambda I)^{-1}U^\top\varepsilon\|^2.
\end{aligned}$$

Note that the cross-term in the final equality disappears since it is the expectation of a linear combination of ε, which is 0.

The first term is a deterministic quantity, which equals

$$\|\left[D(D+\lambda I)^{-1} - I\right]\vartheta\|^2 = \sum_{i=1}^n \left(\left(\frac{d_i}{d_i+\lambda} - 1\right)\vartheta\right)^2 = \sum_{i=1}^n \frac{\lambda^2\vartheta_i^2}{(d_i+\lambda)^2}.$$

Let D^+ be the diagonal matrix with $D^+_{ii} = d_i^{-1}$ if $d_i \neq 0$ and 0 else. Then we have

$$\sum_{i:d_i>0} \frac{\vartheta_i^2}{d_i} = \vartheta^\top D^+ \vartheta = \alpha^\top KUD^+U^\top K\alpha = \alpha^\top UDD^+DU^\top\alpha = \alpha^\top UDU^\top\alpha = \alpha^\top K\alpha \geq 1,$$

Using this, we can bound the first term by

$$\sum_{i=1}^n \frac{\lambda^2\vartheta_i^2}{(d_i+\lambda)^2} = \sum_{i:d_i\neq 0} \frac{\vartheta_i^2}{d_i}\frac{d_i\lambda^2}{(d_i+\lambda)^2} \leq \max_i \frac{d_i\lambda^2}{(d_i+\lambda)^2} \leq \max_i \frac{d_i\lambda^2}{4\lambda d_i} \leq \frac{\lambda}{4}.$$

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

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

This gives our first inequality MSPE $\leq \frac{\sigma^2}{n}\sum_{i=1}^n \frac{d_i^2}{(d_i+\lambda)^2} + \frac{\lambda}{4n}$. For the second inequality, note that $\frac{d_i^2}{(d_i+\lambda)^2} \leq 1$ and also $\frac{d_i^2}{(d_i+\lambda)^2} \leq \frac{d_i^2}{4d_i\lambda} = \frac{d_i}{4\lambda}$. Therefore, we can write $\frac{d_i^2}{(d_i+\lambda)^2} \leq \frac{1}{\lambda}\min(\lambda, \frac{d_i}{4})$ which proves the second inequality. □

We can now ask ourselves if kernel ridge regression is optimal in any sense. To this end, we define $\hat{\mu}_i \coloneqq d_i/n$ and $\lambda_n \coloneqq \lambda/n$. Then we can rewrite the upper bound from our precious theorem as

$$\text{MSPE} \le \frac{\sigma^2}{n\lambda_n} \sum_{i=1}^{n} \min(\frac{\hat{\mu}_i}{4}, \lambda_n) + \frac{\lambda_n}{4} =: \delta_n(\lambda_n).$$

Now, instead of taking $x_1, \dots, x_n \in \mathcal{X}$ to be fixed, assume that they are i.i.d. random variables (then K is random, so the $\hat{\mu}_i$ are random as well), and taking an expectation on both sides yields

$$\mathbb{E}[\text{MSPE}] \le \mathbb{E}[\delta_n(\lambda_n)].$$

We want to bound $\delta_n(\lambda_n)$ by some function of n. For this we will use Mercer's theorem:

Theorem 1.22 (Mercer)**.** *Let $\mathcal{X} = [a, b]$. If k is a continuous kernel on $\mathcal{X}$, then there exists an orthonormal basis (e_i) of $L^2[a, b]$ such that*

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j e_j(x) e_j(x')$$

as well as a sequence of nonnegative eigenvalues (μ_j) such that

$$\int_a^b k(x, x') e_j(x) \, \mathrm{d}x = \mu_j e_j(x).$$

Applying Mercer's theorem to k, it can be proved that for some $C > 0$ we have

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \min(\frac{\hat{\mu}_i}{4}, \lambda_n)\right] \le \frac{C}{n} \sum_{i=1}^{\infty} \min(\frac{\mu_i}{4}, \lambda_n).$$

Example 1.23. Let k be the Sobolev kernel. Then the eigenvalues satisfy

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

and with some calculations it can be shown that $\sum_{i=1}^{\infty} \min(\frac{\mu_i}{4}, \lambda_n) = O(\sqrt{\lambda_n})$ as $\lambda_n \to 0$. Combining this with the rest of the bound shows

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

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

It has been shown that this is in fact the optimal rate for estimating a function $f^0 \in \mathcal{H}$ (up to multiplicative constants).

1.6 Other kernel machines

1.6.1 The support vector machine

Consider a *classification problem* with data $x_1, \dots, x_n \in \mathbb{R}^p$ and binary response $Y_i \in \{\pm 1\}$. Furthermore assume that $x_1, \dots, x_n$ are separated by a hyperplane through the origin, that is, for some $\beta \in \mathbb{R}^p$ we have $Y_i x_i^\top \beta > 0$ for all i.

To choose between different planes that separate the classes, we pick the hyperplane with the highest margin: that is, we compute

$$\max_{\beta \in \mathbb{R}^p, M \ge 0} M \qquad \text{such that } \frac{Y_i x_i^\top \beta}{\|\beta\|} \ge M \quad (i = 1, \dots, n).$$

When the classes are not separable, we can penalise according to the distance of a point "over the margin". The penalty should be 0 if x is on the correct side and should equal the distance over the boundary otherwise, measured in units of M. In this case, by considering minimisation of $1/M^2$ instead of maximisation of M we find

$$\arg\min_{M\geq 0,\beta\in\mathbb{R}^p} \frac{1}{M^2} + \lambda \sum_{i=1}^{n} \left(1 - \frac{Y_i x_i^\top \beta}{\|\beta\| M}\right)_+.$$

Since the above equation is independent of the norm of β, we may set $\|\beta\| = 1/M$ and remove M from the equation entirely, thus obtaining

$$\arg\min_{\beta\in\mathbb{R}^p} \|\beta\|^2 + \lambda \sum_{i=1}^{n} (1 - Y_i x_i^\top \beta)_+.$$

Replacing λ by $\frac{1}{\lambda}$ and multiplying the equation with λ we obtain

$$\arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} (1 - Y_i x_i^\top \beta)_+ + \lambda\|\beta\|^2.$$

Finally, if we lift the restriction that the hyperplane must go through the origin, we find the objective function

$$(\hat{\mu}, \hat{\beta}) = \arg\min_{\mu\in\mathbb{R},\beta\in\mathbb{R}^p} \sum_{i=1}^{n} (1 - Y_i(x_i^\top \beta + \mu))_+ + \lambda\|\beta\|^2. \tag{5}$$

The solution to this problem is known as the *support vector classifier*.

Introducing kernels Now we introduce kernels to the problem. Letting k be the linear kernel and $\mathcal{H}$ the corresponding RKHS, we can rewrite eq. (5) as

$$(\hat{\mu}, \hat{f}) = \arg\min_{\mu\in\mathbb{R}, f\in\mathcal{H}} \sum_{i=1}^{n} (1 - Y_i^\top (f(x_i) + \mu))_+ + \lambda\|f\|^2.$$

Suppose we change $\mathcal{H}$ to a different RKHS with a different kernel k. By a variant of the representer theorem (Example Sheet 1, question 9) the solution is equivalent to

$$(\hat{\mu}, \hat{\alpha}) = \arg\min_{\mu\in\mathbb{R},\alpha\in\mathbb{R}^n} \sum_{i=1}^{n} (1 - Y_i(K_i^\top \alpha + \mu))_+ + \lambda\alpha^\top K\alpha,$$

and the solution to the above problem is called the *support vector machine*. Predictions at a new point x^* are given by

$$\operatorname{sgn}(\hat{\mu} + \sum_{i=1}^{n} \hat{\alpha}_i k(x^*, x_i)) \in \pm 1.$$

Remark. Note that the support vector machine can correspond to a nonlinear boundary: the boundary is a hyperplane *in the corresponding RKHS*, but can be nonlinear in the space $\mathcal{X}$ where the data points live. This also makes it easy to overfit.

1.6.2 Logistic regression

Recap 1.24. The *standard logistic regression* model is motivated by assuming

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

The maximum-likelihood estimator $\hat{\beta}$ is given by (Example sheet 2, question 4)

$$\hat{\beta} \coloneqq \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \log\left(1 + \exp\left(-Y_i x_i^\top \beta\right)\right).$$

Just like in ridge regression, we will try to lower the variance of $\hat{\beta}$ by penalising large values, i.e.,

$$\begin{aligned}\hat{\beta}_\lambda &\coloneqq \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \log\left(1 + \exp\left(-Y_i x_i^\top \beta\right)\right) + \lambda \|\beta\|^2 \\ &= \operatorname*{arg\,min}_{f \in \mathcal{H}} \sum_{i=1}^{n} \log(1 + \exp(-Y_i f(x_i))) + \lambda \|f\|^2,\end{aligned}$$

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

Note that this is exactly the same as the support vector machine, but with a different loss function. The loss function in the support vector machine is called "hinge loss", while the loss function in logistic regression is called "logistic loss". The main differences are:

1. The logistic loss function increases much quicker for incorrectly classified data.
2. Even when a point is correctly classified, the logistic loss function decreases the further the datapoint is from the boundary. The hinge loss function will always be 0 if a point is correctly classified and lies outside of the margin.

The second point also explains the name *support vector machine*: the only points that contribute to the loss function are those that either violate the boundary of are within the margin (the *support vectors*), and perturbing any of the other points does not change the boundary. Because of this property, SVM's tend to be more stable.

1.7 Large scale kernel machines

We have seen that kernels give very flexible regression and classification estimators, and that the parameter λ helps us with the bias-variance tradeoff. Also, when $p \gg n$, using the kernel matrix saves computational effort.

We now turn to the case $n \gg p$. In this case, both kernel ridge regression and the SVM are very expensive to compute: the former has cost $O(n^3)$, while the latter can be computed using an iterative algorithm where every iteration has cost $O(n^2)$.

One approach to speed this process up is that of *random feature maps*: we develop a random map $\hat{\varphi} \colon \mathcal{X} \to \mathbb{R}^b$ (with b small) such that $\mathbb{E}[\hat{\varphi}(x)^\top \hat{\varphi}(x')] = k(x, x')$ for all $x, x' \in \mathcal{X}$. Then we consider L i.i.d. copies of $\hat{\varphi}$ (denotes $\hat{\varphi}_1, \dots, \hat{\varphi}_n$), and we consider the feature map

$$\psi(x) = L^{-1/2}(\hat{\varphi}_1(x), \dots, \hat{\varphi}_L(x))^\top \in \mathbb{R}^{Lb}, \text{ so that } \psi(x)^\top \psi(x') = \frac{1}{L} \sum_{i=1}^{L} \hat{\varphi}_i(x)^\top \hat{\varphi}_i(x').$$

Letting Φ be the matrix with rows $\psi(x_1)^\top, \dots, \psi(x_n)^\top$, we find that $\mathbb{E}(\Phi\Phi^\top)_{ij} = k(x_i, k_j)$, and that $\operatorname{Var}(\Phi\Phi^\top)_{ij}$ decreases with speed L^{-1}. So $\Phi\Phi^\top$ is a good approximation for K when L is moderately large.

One specific example for shift-invariant kernels ($k(x, x') = h(x - x')$ for all $x, x' \in \mathcal{X}$) is based on the work of Rahimi and Recht.

Theorem 1.25 (Bochner)**.** *Let $k \colon \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a continuous kernel. Then k is shift-invariant if and only if, for some $c > 0$ and some distribution F on $\mathbb{R}^p$, it holds for all $W \sim F$ that*

$$k(x, x') = c\mathbb{E}\left[\exp\left(i(x - x')^\top W\right)\right] = c\mathbb{E}\left[\cos\left((x - x')^\top W\right)\right].$$

To make use of this theorem, we let $u \sim U[-\pi, \pi]$, and it can then be computed (using $u \overset{d}{=} -u$ and $\cos(u) \overset{d}{=} \sin(u)$ that

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

Therefore, given a shift-invariant kernel k with associated distribution F, we define for $W \sim F$ the feature map

$$\hat{\varphi}(x) = \sqrt{2c}\cos\big(W^\top x + u\big) \in \mathbb{R},$$

so that

$$\mathbb{E}\hat{\varphi}(x)\hat{\varphi}(x') = 2c\mathbb{E}\big[\mathbb{E}[\cos\big(W^\top x + u\big)\cos\big(W^\top x' + u\big) \mid W]\big] = c\mathbb{E}\cos\big((x - x^\top)W\big) = k(x, x').$$

Example 1.26. Let k be the Gaussian kernel $k(x, x') = e^{-\|x-x'\|/(2\sigma^2)}$. Let $W \sim N(0, \sigma^{-2}I)$, then using the characteristic function of a normal distribution we find that

$$\mathbb{E}\Big(e^{i(x-x')^\top W}\Big) = \exp\Big(-\|x - x'\|^2/(2\sigma^2)\Big) = k(x, x'),$$

so we can set $c = 1$, i.e., define the feature map $\hat{\varphi}(x) = \sqrt{2}\cos\big(W^\top x + u\big)$.

2 The Lasso

2.1 Model selection

We go back to the linear model $Y = X\beta^0 + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Using the trace trick, one can easily compute that the MSPE of the OLS estimator is given by

$$\frac{1}{n}\mathbb{E}\left\|X\beta^0 - X\hat{\beta}^{\mathrm{OLS}}\right\|_2^2 = \frac{\sigma^2 p}{n}.$$

Defining $S = \left\{k \mid (\beta^0)_k \neq 0\right\}$, there is often reason to assume that S is small, i.e., $s \coloneqq |S| \ll p$. If we could fit a model using only the variables in S, the MSPE would be much $\frac{\sigma^2 s}{n} \ll \frac{\sigma^2 p}{n}$.

Best subset selection A natural way to find S is to consider all possible subsets of $\{1, \ldots, p\}$, and pick the best regression procedure using, for example, cross-validation. However, this can become computationally infeasible for moderately large p (say $p \approx 10$).

Forward selection This is a greedy way of performing best subset regression. Given a target model size m, we first compute the intercept-only model M_0, and then one-by-one add the predictor variable that reduces the residual sum of squares the most, until we have a model with m variables.

2.2 Lasso estimator

The *Least absolute shrinkage and selection operator* or *Lasso* is given by

$$(\hat{\mu}_\lambda^{\mathrm{L}}, \hat{\beta}_\lambda^{\mathrm{L}}) \coloneqq \underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \frac{1}{2n}\|Y - \mu\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

As with ridge regression, we usually centre and scale the matrix X, as well as centre the responses Y, in which case we find

$$\hat{\beta}_\lambda^{\mathrm{L}} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

The main difference between the lasso estimator and the ridge regression estimator is that it is likely that the lasso estimator has some zero components. This means that the lasso estimator also estimates which variables are relevant.

2.2.1 Slow prediction error rate

We have the following bound on the prediction error:

Theorem 2.1 (Slow rate)**.** *Assume X has centred and scaled columns, and assume that Y has been centred, so $Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1}$. Let $A > 0$ and suppose*

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

Let $\hat{\beta} = \hat{\beta}_\lambda^{\mathrm{L}}$, then with probability at least $1 - 2p^{-(A^2/2-1)}$ we have that the MSPE satisfies

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

Proof. By definition we have

$$\frac{1}{2n}\left\|Y - X\hat{\beta}\right\|_2^2 + \lambda\left\|\hat{\beta}\right\|_1 \leq \frac{1}{2n}\|Y - X\beta^0\|_2^2 + \lambda\left\|\beta^0\right\|_1,$$

and rearranging the terms gives

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

By Hölder's inequality we have $\left|\varepsilon^\top X(\beta^0-\hat{\beta})\right| \le \left\|X^\top\varepsilon\right\|_\infty \left\|\hat{\beta}-\beta^0\right\|_1$. Define the event $\Omega = \left\{\left\|X^\top\varepsilon\right\|_\infty/n \le \lambda\right\}$, then conditional on Ω we find

$$\frac{1}{n}\left\|X(\beta^0-\hat{\beta})\right\|_2^2 \le 2\lambda\Big(\left\|\beta^0-\hat{\beta}\right\|_1+\|\beta^0\|_1-\left\|\hat{\beta}\right\|_1\Big) \le 4\lambda\|\beta^0\|_1.$$

In lemma 2.5, we will show that $\mathbb{P}(\Omega) \ge 1-2p^{-(A^2/2-1)}$, which completes the proof. □

2.2.2 Concentration inequalities

Let W be any random variable and $\varphi\colon \mathbb{R}\to[0,\infty)$ strictly increasing. Then by Markov's inequality we have

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

Plugging in $\varphi(x)=e^{\alpha x}$ (for some $\alpha>0$), we get

$$\mathbb{P}(W\ge t) \le e^{-\alpha t}\mathbb{E}[e^{\alpha W}] = e^{-\alpha t}M_W(\alpha).$$

Now we can take the infimum over all α on the right-hand side, and we get what is called the *Chernoff bound*:

$$\mathbb{P}(W\ge t) \le \inf_{\alpha>0} e^{-\alpha t}M_W(\alpha).$$

Definition 2.2. A random variable W with mean μ is called *sub-Gaussian with parameter* $\sigma>0$ or σ-sub-Gaussian if

$$M_{W-\mu} \le M_{N(0,\sigma^2)} \quad \text{or equivalently} \quad \mathbb{E}[e^{\alpha(W-\mu)}] \le e^{\alpha^2\sigma^2/2} \text{ for all } \alpha\in\mathbb{R}.$$

We need the following lemma, which characterises an important class of sub-Gaussian random variables:

Lemma 2.3 (Hoeffding)**.** *If W is a mean-zero random variable which takes values in $[a,b]$, then W is sub-Gaussian with parameter $(b-a)/2$.*

By Chernoff bounding, we obtain for a σ-sub-Gaussian random variable W that

$$\mathbb{P}(W-\mu\ge t) \le e^{-t^2/(2\sigma^2)}.$$

Proposition 2.4. *Let $W_1,\dots,W_n$ be independent, mean-zero random variables where W_i is σ_i-sub-Gaussian. For any $\gamma\in\mathbb{R}^n$, the random variable $\gamma^\top W$ is sub-Gaussian with parameter $\left(\sum_i\sigma_i^2\gamma_i^2\right)^{1/2}$.*

Proof. Since the W_i are independent we have for all $\alpha\in\mathbb{R}$ that

$$\mathbb{E}[e^{\alpha\sum_i\gamma_iW_i}] = \prod_{i=1}^n \mathbb{E}[e^{\alpha\gamma_iW_i}] \le \prod_{i=1}^n e^{\alpha^2\gamma_i^2\sigma_i^2/2} = e^{\alpha^2\sum_i\gamma_i^2\sigma_i^2/2}.$$

□

Lemma 2.5. *Suppose $\varepsilon_1,\dots,\varepsilon_n$ are independent mean-zero σ-sub-Gaussian random variables and let $\lambda \coloneqq A\sigma\sqrt{\log(p)/n}$. Then*

$$\mathbb{P}\left(\frac{\left\|X^\top\varepsilon\right\|_\infty}{n}\le\lambda\right) \ge 1-2p\exp\left(\frac{-\lambda^2}{2(\sigma/\sqrt{n})^2}\right) = 1-2p^{-(A^2/2-1)}.$$

Proof. We have

$$\mathbb{P}\left(\frac{\left\|X^\top\varepsilon\right\|_\infty}{n} > \lambda\right) = \mathbb{P}\left(\bigcup_i \frac{\left|X_i^\top\varepsilon\right|}{n} > \lambda\right) \leq \sum_i \mathbb{P}\left(\frac{\left|X_i^\top\varepsilon\right|}{n} > \lambda\right).$$

By the previous proposition, both $X_i^\top\varepsilon$ and $-X_i^\top\varepsilon$ are sub-Gaussian with parameter $\sigma/\sqrt{n}$, so we have

$$\sum_i \mathbb{P}\left(\frac{\left|X_i^\top\varepsilon\right|}{n} > \lambda\right) \leq 2p\exp\left(\frac{-\lambda^2}{2(\sigma/\sqrt{n})^2}\right) = 2p\exp\left(-A^2\log(p)/2\right) = 2p^{-(A^2/2-1)}.$$

□

We now move to finding tail bounds for products of sub-Gaussian random variables. To this end, we consider Bernstein's in equality.

Definition 2.6. A random variable W with mean μ is said to satisfy *Bernstein's condition* with parameters $\sigma > 0, b > 0$ if

$$\mathbb{E}\left(|W-\mu|^k\right) \leq \frac{1}{2}k!\sigma^2 b^{k-2} \qquad k = 2, 3, \ldots$$

Proposition 2.7 (Bernstein's inequality)**.** *Let $W_1, W_2, \ldots$ be independent random variables with mean μ, where each W_i satisfies Bernstein's condition with parameter (σ, b). Then*

1. $\mathbb{E}[\exp(\alpha(W_i-\mu))] \leq \exp\left(\frac{\alpha^2\sigma^2}{2(1-b|\alpha|)}\right)$ *for all* $|\alpha| < b^{-1}$;
2. $\mathbb{P}\left(\overline{W}-\mu \geq t\right) \leq \exp\left(\frac{-nt^2}{2(\sigma^2+bt)}\right)$ *for all* $t \geq 0$.

Proof. See notes from topics in statistical theory. □

Lemma 2.8. *Let W, Z be mean-zero, sub-Gaussian random variables with parameters σ_W, σ_Z. Then WZ satisfies the Bernstein condition with parameters $(8\sigma_W\sigma_Z, 4\sigma_W\sigma_Z)$.*

Proof. Using the tail probability formula for expectations one can prove that any mean-zero sub-Gaussian random variable X satisfies $\mathbb{E}[X^{2k}] \leq 2^{k+1}\sigma_X^{2k}k!$.

Furthermore for any random variable Y we have

$$\begin{aligned}
\mathbb{E}|Y-\mathbb{E}Y|^k &\leq \mathbb{E}(|Y|+|\mathbb{E}Y|)^k \\
&= \mathbb{E}\sum_{t=0}^k \binom{k}{t}|Y|^t|\mathbb{E}Y|^{k-t} \\
&\overset{*}{\leq} \sum_{t=0}^k \binom{k}{t}\mathbb{E}(|Y|^t)\mathbb{E}(|Y|^{k-t}) \\
&\overset{*}{\leq} \sum_{t=0}^k \binom{k}{t}\mathbb{E}(|Y|^k)^{t/k}\mathbb{E}(|Y|^k)^{k-t/k} \\
&= \mathbb{E}(|Y|^k)\sum_{t=0}^k \binom{k}{t} = 2^k\mathbb{E}|Y|^k,
\end{aligned}$$

where both equations $*$ are applications of Jensen's inequality.

From the above bound, we obtian

$$\begin{aligned}
\mathbb{E}|WZ-\mathbb{E}WZ|^k &\leq 2^k\mathbb{E}(|W|^k|Z|^k) \overset{\text{CS}}{\leq} 2^k(\mathbb{E}W^{2k})^{1/2}(\mathbb{E}Z^{2k})^{1/2} \leq 2^{2k+1}\sigma_W^k\sigma_Z^k k! \\
&= \frac{k!}{2}(8\sigma_W\sigma_Z)^2(4\sigma_W\sigma_Z)^{k-2},
\end{aligned}$$

which proves the claim. □

2.2.3 Optimisation theory and convex analysis

Lagrangian optimisation

Definition 2.9. For the optimisation problem

$$\inf\{f(x) \mid g(x) = 0, h(x) \leq 0\}, \tag{6}$$

its *Lagrangian* is given by

$$L(x, \lambda, \nu) \coloneqq f(x) + \lambda h(x) + \nu g(x),$$

and its *Lagrange dual function* is

$$\tilde{f}(\lambda, \nu) = \inf_x L(x, \lambda, \nu). \tag{7}$$

Note that if p^* is the minimum value of (6), then for any feasible x (i.e., x with $g(x) = 0$, $h(x) \leq 0$) we have for $\lambda \geq 0$ that

$$f(x) \geq L(x, \lambda, \nu) \geq \tilde{f}(\lambda, \nu).$$

By taking the infimum over x in the left-hand side and the supremum over λ, ν in the right-hand side we find that

$$p^* \geq \sup_{\lambda \geq 0, \nu} \tilde{f}(\lambda, \nu),$$

Sometimes we actually have equality here, and in this case we say that the *duality gap* is zero. Note that $\tilde{f}$ is a concave function, and is therefore easy to maximise.

A sufficient condition for a zero duality gap is that $L(\cdot, \lambda, \nu)$ is convex and that the problem is feasible.

Convex analysis

Let $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, and define for any $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ its (*effective*) *domain* $\operatorname{dom}(f) \coloneqq \{x \mid f(x) < \infty\}$.

Proposition 2.10. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. Then f is convex (resp. strictly convex) if and only if its Hessian is positive semi-definite (resp. positive definite) for all $x \in \mathbb{R}^d$.*

Definition 2.11. Let $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex and $x \in \mathbb{R}^d$. Then $v \in \mathbb{R}^d$ is called a *subgradient* of f at x if

$$f(y) \geq f(x) + v^\top (y - x) \qquad \forall y \in \mathbb{R}^d.$$

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

Proposition 2.12. *If f is convex and differentiable at $x \in \operatorname{Int}(\operatorname{dom} f)$, then $\partial f(x) = \{\nabla f(x)\}$.*

Proposition 2.13. *Let $f, g\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex with $\operatorname{Int}(\operatorname{dom} f) \cap \operatorname{Int}(\operatorname{dom} g) \neq \varnothing$, $x \in \mathbb{R}^d$ and $\alpha > 0$. Then $\partial(\alpha f)(x) = \alpha \partial f(x)$ and $\partial(f + g)(x) = \partial f(x) + \partial g(x)$.*

Proposition 2.14 (KKT conditions)**.** *Let $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ be convex. Then x^* is a global minimiser of f if and only if $0 \in \partial f(x^*)$.*

Proof. Clearly x^* is a global minimiser of f if and only if $f(y) \geq f(x^*) + 0^\top (y - x)$ for all $y \in \mathbb{R}^d$. □

We define for $x \in \mathbb{R}$

$$\operatorname{sgn}(x) \coloneqq \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0, \end{cases}$$

and for $x \in \mathbb{R}^d$ we define $\operatorname{sgn}(x) \coloneqq (\operatorname{sgn}(x_1), \ldots, \operatorname{sgn}(x_n))$.

Proposition 2.15. *Let $x \in \mathbb{R}^d$ and $A \coloneqq \{j \mid x_j \neq 0\}$. Then*

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

Proof. Define $g_j(x) = |x_j|$ so that $\|\cdot\|_1 = \sum_j g_j$ and $\partial\|x\|_1 = \sum_j \partial g_j(x)$. Now, if $x_j \neq 0$, then g_j is differentiable at x so $\partial g_j(x) = \{\operatorname{sgn}(x_j)e_j\}$. If $x_j = 0$, then it can be shown that

$$\partial g_j(x) = \{ce_j \mid c \in [-1, 1]\}.$$

Combining the above facts proves the claim. □

2.2.4 Lasso solutions

With the previous proposition, we can write characterise solutions to the Lasso: namely we have

$$\hat{\beta}^L_\lambda = \arg\min_\beta \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 = \arg\min_\beta Q(\lambda)$$

if and only if $0 \in \partial Q(\hat{\beta}^L_\lambda)$. A simple computation gives, with $A = \{k \mid \beta_k \neq 0\}$,

$$\begin{aligned}\partial Q(\beta) &= \partial(\frac{1}{2n}\|Y - X\beta\|_2^2) + \partial(\lambda\|\beta\|_1) \\ &= \left\{-\frac{1}{n}X^\top(Y - X\beta)\right\} + \{\lambda\nu : \|\nu\|_\infty \leq 1, \nu_A = \operatorname{sgn}(\beta_A)\}.\end{aligned}$$

We conclude that $\hat{\beta}$ is a Lasso solution if and only if there exists a vector $\hat{\nu}$ with $\|\hat{\nu}\|_\infty \leq 1$ and $\hat{\nu}_A = \operatorname{sgn}(\hat{\beta}_A)$ such that

$$\frac{1}{n}X^\top(Y - X\hat{\beta}) = \lambda\hat{\nu}.$$

Note that $\hat{\beta}^L_\lambda$ need not be unique, however, the fitted values are unique:

Proposition 2.16. *$X\hat{\beta}^L_\lambda$ is unique.*

Proof. Write $\frac{1}{2n}\|Y - X\beta\|_2^2 = Q_1(X\beta)$ and $\|\beta\|_1 = Q_2(\beta)$. Note that Q_1 is strictly convex. Now, for any $t \in (0, 1)$ and any two lasso solutions $\hat{\beta}, \hat{\gamma}$ with $Q(\hat{\beta}) = Q(\hat{\gamma}) = c^*$ we have

$$\begin{aligned}c^* &\leq Q(t\hat{\beta} + (1-t)\hat{\gamma}) = Q_1(tX\hat{\beta} + (1-t)X\hat{\gamma}) + Q_2(t\hat{\beta} + (1-t)\hat{\gamma}) \\ &\overset{\star}{\leq} tQ_1(X\hat{\beta}) + (1-t)Q_2(X\hat{\gamma}) + Q_2(t\hat{\beta} + (1-t)\hat{\gamma}) \\ &\leq tQ(\hat{\beta}) + (1-t)Q(\hat{\gamma}) = c^*,\end{aligned}$$

so all inequalities must be equalities. In particular, by $\star$ we find $Q_1(tX\hat{\beta} + (1-t)X\hat{\gamma}) = tQ_1(X\hat{\beta}) + (1-t)Q_1(X\hat{\gamma})$, and since Q_1 is strictly convex this implies $X\hat{\beta} = X\hat{\gamma}$. □

2.2.5 Variable selection

We consider the noiseless model $Y = X\beta^0$, where no randomness is involved. Let $S = \{k \mid \beta_k \neq 0\}$, $N = \{1, \ldots, p\} \setminus S$, and assume that $S = \{1, \ldots, s\}$ (so $\beta_1, \ldots, \beta_s$ are nonzero and $\beta_{s+1}, \ldots, \beta_p$ are zero).

Furthermore, write $X = \begin{bmatrix} X_S & X_N \end{bmatrix}$ and assume that $\operatorname{rank}(X_S) = s$, so that $X_S^\top X_S$ is invertible (NB in particular this implies $n \geq s$).

Theorem 2.17. *Let $\lambda > 0$ and define $\Delta = X_N^\top X_S (X_S^\top X_S)^{-1} \operatorname{sgn}(\beta^0_S)$.*

1. *If* $\|\Delta\|_\infty \leq 1$ *and for* $k = 1, \ldots, s$ *we have*

$$|\beta_k^0| > \lambda \left| \operatorname{sgn}(\beta_S^0)^\top \left[\frac{1}{n} X_S^\top X_S \right]_k^{-1} \right|,$$

then there exists a Lasso solution $\hat{\beta}_\lambda^L$ *with* $\operatorname{sgn}(\hat{\beta}_\lambda^L) = \operatorname{sgn}(\beta^0)$.

2. *If there exists a Lasso solution* $\hat{\beta}_\lambda^L$ *with* $\operatorname{sgn}(\hat{\beta}_\lambda^L) = \operatorname{sgn}(\beta^0)$, *then* $\|\Delta\|_\infty = 1$.

Remark. The condition $\|\Delta\|_\infty \leq 1$ is called the *irrepresentable condition*, and it roughly states that none of the insignificant predictors align "too well" with the response. TODO: Elaborate on this (lecture 12).

Proof. Let $\hat{\beta}$ be any lasso solution and write $\hat{S} = \{k \mid \beta_k \neq 0\}$. Using the fact that $\beta_N^0 = 0$, the KKT conditions for the Lasso can be expanded as

$$\frac{1}{n} \begin{pmatrix} X_S^\top X_S & X_S^\top X_N \\ X_N^\top X_S & X_N^\top X_N \end{pmatrix} \begin{pmatrix} \beta_S^0 - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} = \lambda \begin{pmatrix} \hat{\nu}_S \\ \hat{\nu}_N \end{pmatrix} \tag{8}$$

for some $\hat{\nu}$ with $\|\hat{\nu}\|_\infty \leq 1$ and $\hat{\nu}_{\hat{S}} = \operatorname{sgn}(\hat{\beta}_{\hat{S}})$.

1. It is easily checked that taking

$$(\hat{\beta}_S, \hat{\beta}_N) = \left(\beta_S^0 - \lambda \left(\frac{1}{n} X_S^\top X_S \right)^{-1} \operatorname{sgn}(\beta_S^0), 0 \right)$$
$$(\hat{\nu}_S, \hat{\nu}_N) = (\operatorname{sgn}(\beta_S^0), \Delta)$$

is a Lasso solution with $\operatorname{sgn}(\hat{\beta}_S^0) = \operatorname{sgn}(\hat{\beta}_S)$.

2. If $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^0)$, then $\hat{S} = S$, $\hat{\nu}_S = \operatorname{sgn}(\beta_S^0)$ and $\hat{\beta}_N = 0$. The top block of eq. (8) can be written as

$$\frac{1}{n} X_S^\top X_S (\beta_S^0 - \hat{\beta}_S) = \lambda \operatorname{sgn}(\beta_S^0) \implies \beta_S^0 - \hat{\beta}_S = \lambda \left[\frac{1}{n} X_S^\top X_S \right]^{-1} \operatorname{sgn}(\beta_S^0).$$

Plugging that into the bottom block of the same equation, we get

$$\lambda \frac{1}{n} X_N^\top X_S \left[\frac{1}{n} X_S^\top X_S \right]^{-1} \operatorname{sgn}(\beta_S^0) = \lambda \hat{\nu}_N \implies \|\Delta\|_\infty \leq 1.$$

□

2.2.6 Prediction and estimation

Definition 2.18. Let $X \in \mathbb{R}^{n \times p}$ be a design matrix and $S \subseteq \{1, \ldots, p\}$ where $s \coloneqq |S| > 0$, then the *compatibility constant* $\varphi^2 = \varphi^2(X, S)$ is defined as

$$\varphi^2 \coloneqq \inf \left\{ \frac{\frac{1}{n}\|X\beta\|_2^2}{\frac{1}{s}\|\beta_S\|_1^2} : \beta \in \mathbb{R}^p, \beta_S \neq 0, \|\beta_N\|_1 \leq 3\|\beta_S\|_1 \right\}.$$

The *compatibility condition* is that $\varphi^2 > 0$.

Note that if X_S is rank-deficient, then there exists $\beta_S \neq 0$ with $X_S \beta_S = 0$, and by setting $\beta_N = 0$ we find that $\varphi^2 = 0$. Therefore, X_S must have full column rank for the compatibility condition to be satisfied, and in particular we must have $n \geq s$.

Proposition 2.19. *If $X^\top X/n$ has minimal eigenvalue $c_{\min}$ then $\varphi^2 \geq c_{\min}$. In particular, if X has full column rank (i.e., $c_{\min} > 0$), the compatibility condition is always satisfied.*

Proof. We have

$$\|\beta_S\|_1 = \operatorname{sgn}(\beta_S)^\top \beta_S \leq \|\operatorname{sgn}(\beta_S)\|_2 \|\beta_S\|_2 = \sqrt{s}\|\beta\|_2.$$

Therefore, we find

$$\varphi^2 \geq \inf_{\beta \neq 0} \frac{\frac{1}{n}\|X\beta\|_2^2}{\frac{1}{s}\|\beta\|_1^2} \geq \frac{1}{n} \inf_{\beta \neq 0} \frac{\|X\beta\|_2^2}{\|\beta\|_2^2} = \inf_{\|\beta\|=1} \beta^\top \left(\frac{X^\top X}{n}\right)\beta = c_{\min}.$$

□

Note that the converse of this proposition does not necessarily hold: X does not need to have full column rank to satisfy the compatibility condition. In particular, it does not need to be the case that $n < p$.

Now we present the "fast" convergence rate of the lasso (compare with theorem 2.1):

Theorem 2.20 (Fast rate)**.** *Let the conditions of theorem 2.1 be satisfied and assume furthermore that the compatibility condition holds. Then with probability at least $1 - 2p^{-(A^2/8-1)}$ we have*

$$\frac{1}{n}\left\|X(\beta^0 - \hat{\beta})\right\|_2^2 + \lambda\left\|\hat{\beta} - \beta^0\right\|_1 \leq \frac{16\lambda^2 s}{\varphi^2} = \frac{16A^2 \log(p)\sigma^2 s}{\varphi^2 n}.$$

Proof. As in the proof of theorem 2.1, we have

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

Define the event $\Omega = \left\{2\left\|X^\top \varepsilon\right\|_\infty / n \leq \lambda\right\}$, then by Hölder's inequality we get, conditional on Ω,

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

By lemma 2.5, we find that $\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/8-1)}$.

Dividing by λ, we have

$$\begin{aligned}
\frac{1}{\lambda n}\left\|X(\hat{\beta} - \beta^0)\right\|_2^2 + 2\left\|\hat{\beta}\right\|_1 &\leq \left\|\hat{\beta} - \beta^0\right\|_1 + 2\|\beta^0\|_1 \\
\frac{1}{\lambda n}\left\|X(\hat{\beta} - \beta^0)\right\|_2^2 + 2(\left\|\hat{\beta}_S\right\|_1 + \left\|\hat{\beta}_N\right\|_1) &\leq \left\|\hat{\beta}_S - \beta_S^0\right\|_1 + \left\|\hat{\beta}_N\right\|_1 + 2\|\beta_S^0\|_1 \\
\frac{1}{\lambda n}\left\|X(\hat{\beta} - \beta^0)\right\|_2^2 + \left\|\hat{\beta}_N\right\|_1 &\leq \left\|\hat{\beta}_S - \beta_S^0\right\|_1 - 2\left\|\hat{\beta}_S\right\|_1 + 2\|\beta_S^0\|_1 \\
&\leq 3\left\|\hat{\beta}_S - \beta_S^0\right\|_1.
\end{aligned}$$

Now we note that $\left\|\hat{\beta}_N\right\|_1 = \left\|\hat{\beta}_N + \beta_N^0\right\|_1$, and we add $\left\|\hat{\beta}_S - \beta_S^0\right\|_1$ to both sides of the inequality to get

$$\frac{1}{\lambda n}\left\|X(\hat{\beta} - \beta^0)\right\|_2^2 + \left\|\hat{\beta} - \beta^0\right\|_1 \leq 4\|\beta_S - \beta_S^0\|_1. \tag{9}$$

Assume that $\left\|\hat{\beta} - \beta^0\right\| \neq 0$ (otherwise the theorem is definitely true), and note that $\left\|(\hat{\beta} - \beta^0)_N\right\|_1 \leq 3\left\|\hat{\beta}\right\|_1 \leq 3\left\|(\hat{\beta} - \beta^0)_S\right\|$ by our previous inequality, we find that the compatibility constant satisfies

$$\varphi^2 \leq \frac{\frac{1}{n}\left\|X(\hat{\beta} - \beta^0)\right\|_2^2}{\frac{1}{s}\left\|\hat{\beta} - \beta^0\right\|_1^2} \implies \frac{1}{\varphi} \geq \sqrt{\frac{n}{s}} \frac{\left\|\hat{\beta} - \beta^0\right\|_1}{\left\|X(\hat{\beta} - \beta^0)\right\|_2},$$

and plugging this into eq. (9) we get

$$\frac{1}{\lambda n}\left\|X(\hat{\beta}-\beta^0)\right\|_2^2+\left\|\hat{\beta}-\beta^0\right\|_1\le 4\|\beta_S-\beta_S^0\|_1\le\frac{4}{\varphi}\sqrt{\frac{s}{n}}\left\|X(\hat{\beta}-\beta^0)\right\|_2, \tag{10}$$

or equivalently

$$\frac{1}{\lambda n}\left\|X(\hat{\beta}-\beta^0)\right\|_2^2-\frac{4}{\varphi}\sqrt{\frac{s}{n}}\left\|X(\hat{\beta}-\beta^0)\right\|_2+\left\|\hat{\beta}-\beta^0\right\|_1\le 0.$$

This is a quadratic inequality in $\left\|X(\hat{\beta}-\beta^0)\right\|_2$, and solving it yields

$$\frac{1}{\sqrt{n}}\left\|X(\hat{\beta}-\beta^0)\right\|_2\le\frac{4\lambda\sqrt{s}}{\varphi}.$$

Plugging that into the right hand side of eq. (10) yields the required result. □

The compatibility condition and random design

We will show that if X is a random matrix, then the compatibility condition holds with high probability.

Define, for $\Sigma\in\mathbb{R}^{p\times p}$ and $S\subseteq\{1,\dots,p\}$, the quantity

$$\varphi_\Sigma^2(S)=\inf_{\beta_S\neq 0,\|\beta_N\|_1\le 3\|\beta_S\|_1}\frac{\beta^\top\Sigma\beta}{\frac{1}{|S|}\|\beta_S\|_1^2},$$

so that the compatiblity constant satisfies $\varphi^2=\varphi_{\hat{\Sigma}}^2(S)$ with $\hat{\Sigma}\coloneqq X^\top X/n$ and S the support of β^0.

To prove our main result, we first need a lemma:

Lemma 2.21. *Suppose* $\varphi_{\check{\Sigma}}^2(S)>0$ *and* $\max_{jk}\left|\hat{\Sigma}_{jk}-\check{\Sigma}_{jk}\right|\le\varphi_{\check{\Sigma}}^2(S)/(32|S|)$. *Then* $\varphi_{\hat{\Sigma}}^2(S)\ge\varphi_{\check{\Sigma}}^2(S)$.

Proof. Let $s=|S|$ and $t\coloneqq\varphi_{\check{\Sigma}}^2/(32s)$. Then by applying Hölder's inequality we obtain

$$\left|\beta^\top(\check{\Sigma}-\hat{\Sigma})\beta\right|\le\|\beta\|_1\left\|(\check{\Sigma}-\hat{\Sigma})\beta\right\|_\infty\le\|\beta\|_1^2\max_{j,k}\left|\hat{\Sigma}_{jk}-\check{\Sigma}_{jk}\right|\le\|\beta\|_1^2 t.$$

Furthermore, if $\|\beta_N\|_1\le 3\|\beta_S\|_1$, then

$$\|\beta\|_1=\|\beta_N\|_1+\|\beta_S\|_1\le 4\|\beta_S\|_1\le\frac{4\sqrt{\beta^\top\check{\Sigma}\beta}}{\varphi_{\check{\Sigma}}/s}\implies\|\beta_1\|^2\le\frac{16\beta^\top\check{\Sigma}\beta}{\varphi_{\check{\Sigma}}^2/s^2},$$

and therefore

$$\frac{1}{2}\beta^\top\check{\Sigma}\beta=\beta^\top\check{\Sigma}\beta-\frac{\varphi_{\check{\Sigma}}^2}{32s}\cdot\frac{16\beta^\top\check{\Sigma}\beta}{\varphi_{\check{\Sigma}}^2/s^2}\le\beta^\top\check{\Sigma}\beta-t\|\beta\|_1^2\le\beta^\top\check{\Sigma}\beta-\beta^\top(\check{\Sigma}-\hat{\Sigma})\beta=\beta^\top\hat{\Sigma}\beta.$$

Taking the infimum over all β with $\|\beta_N\|_1\le 3\|\beta_S\|_1$ and $\beta_S\neq 0$, we obtain the required result. □

Now we can state the main theorem. We consider a sequence of models with sample size n (where $n\to\infty$), and with p and s increasing to ∞ as functions of n.

Theorem 2.22. *Suppose the rows of* X *are i.i.d. and each entry is a mean-zero sub-Gaussian random variable with parameter* v. *Let* $\Sigma^0\coloneqq\mathbb{E}[X^\top X/n]$, *and suppose that* $s\sqrt{\log(p)/n}\to 0$ *as* $n\to\infty$. *Define*

$$\varphi_{\hat{\Sigma},s}^2\coloneqq\min_{|S|=s}\varphi_{\hat{\Sigma}}^2(S),\quad\varphi_{\Sigma^0,s}^2\coloneqq\min_{|S|=s}\varphi_{\Sigma^0}^2(S).$$

If $\varphi_{\Sigma^0,s}^2>c>0$ *(for all* n*), then* $\mathbb{P}(\varphi_{\hat{\Sigma},s}^2\ge\varphi_{\Sigma^0,s}^2/2)\to 1$ *as* $n\to\infty$.

Proof. By the previous lemma, it suffices to show that

$$(*) = \mathbb{P}(\max_{jk} \left|\hat{\Sigma}_{jk} - \Sigma^0_{jk}\right| \geq \varphi^2_{\Sigma^0,s}/(32s)) \to 0 \quad \text{as } n \to \infty.$$

Letting $t \coloneqq \varphi^2_{\Sigma^0,s}/(32s)$, then by a union bound and Bernstein's theorem, we have that

$$\begin{aligned}(*) &\leq p^2 \max_{j,k} \mathbb{P}\left(\left|\sum_{i=1}^n X_{ij}X_{ik}/n - \Sigma^0_{jk}\right| \geq t\right) \\ &\overset{\star}{\leq} p^2 2\exp\left(-\frac{nt^2}{2(64v^4 + 4v^2 t)}\right) \\ &\leq c_1 \exp\Big(-c_2\frac{n}{s^2} + c_3\log(p)\Big) \to 0 \quad \text{if } s\sqrt{\log(p)/n} \to 0 \text{ as } n \to \infty.\end{aligned}$$

where $\star$ follows from the fact that $X_{ij}X_{ik}$ satisfies Bernstein's condition with parameters $(8v^2, 4v^2)$. □

What does the previous theorem tell us? We have a weak growth condition: for example, the theorem holds if $p = n^d$ and $s = (\sqrt{n})^{1-\alpha}$ for any $d, \alpha > 0$.

Example 2.23. Suppose the rows of X are taken form a $N_p(0, \Sigma^0)$ distribution, where Σ^0 has bounded diagonal entries and the smallest eigenvalue of Σ^0 is $c_{\min} > c > 0$ (for all n). Then $\mathbb{P}(\varphi^2_{\hat{\Sigma},s} \geq \frac{c_{\min}}{2}) \to 1$ as $n \to \infty$ if the growth condition is satisfied. This holds for example if $\Sigma^0 = I_p$.

Example 2.24. Suppose X has i.i.d. entries $\mathbb{P}(X_{ij} = 1) = \frac{1}{2} = \mathbb{P}(X_{ij} = -1)$. Then by Hoeffding's lemma, X_{ij} is sub-Gaussian, and therefore the theorem holds under appropriate conditions.

Example 2.25. Let W be a matrix whose rows are vectors in a basis, and let X have rows which are random subsets of W. This does not satisfy the conditions of theorem 2.22, yet the lasso still works very well in this case.

Generally, the Lasso demonstrates a very low estimation if n/p is much larger than s/p, but a very high estimation error if s/p is much larger than n/p. Usually, there is a sharp "phase transition" here.

2.2.7 Computation

The *coordinate descent* algorithm is an algorithm for minimising functions of the form $f\colon \mathbb{R}^d \to \mathbb{R}$, and works particularly well for functions of the form

$$f(x) = g(x) + \sum_{j=1}^d h_j(x),$$

where g is convex and differentiable and the $h_j\colon \mathbb{R} \to \mathbb{R}$ are convex and continuous. Note that the Lasso objective function is clearly of this form. We start with an initial guess $x^{(0)}$ for a minimiser, and then for $m = 1, 2, \ldots$ we choose, for $i = 1, \ldots, d$,

$$x_i^{(m)} = \arg\min_{x_i \in \mathbb{R}} f\Big(x_1^{(m)}, \ldots, x_{i-1}^{(m)}, x_i, x_{i+1}^{(m-1)}, \ldots, x_d^{(m-1)}\Big).$$

It can be shown that for the Lasso objective function, each step of the coordinate descent algorithm admits a closed-form solution.

We have the following:

Theorem 2.26 (Tseng)**.** *If $A_0 \coloneqq \{x \mid f(x) \leq f(x^0)\}$ is compact, then every converging subsequence of $(x^{(m)})$ converges to a minimiser of f.*

Using this theorem, we can see the following:

Corollary 2.27. *Suppose A^0 is compact. Then*

1. *There exists a mimiser x^* of f, and $f(x^{(m)}) \to f(x^*)$;*
2. *If x^* is the unique minimiser of f, then $x^{(m)} \to x^*$.*

Proof. Since f is continuous, it attains a minimum $f(x^*)$ on A_0. Suppose $f(x^{(m)}) \not\to f(x^*)$. Note that $f(x^{(m)})$ is a monotone decreasing sequence, so it must converge. Furthermore, $(x^{(m)}) \subseteq A_0$, so there exists a subsequence of $(f(x^{(m)}))$ converging to $f(x^*)$ by theorem 2.26. This proves that $f(x^{(m)}) \to f(x^*)$.

Suppose that x^* is the unique minimiser of f, then by theorem 2.26 we know that every convergent subsequence of $x^{(m)}$ converges to x^*. Since $x^{(m)}$ is bounded, this proves that $x^{(m)} \to x^*$ as well. □

Often, we want to solve the Lasso on a grid of values $\lambda_0 > \cdots > \lambda_L$. To speed up the coordinate descent procedure, we can use the minimiser for $\lambda_{\ell-1}$ as a starting point for coordinate descent on λ_ℓ. An "active set" strategy can speed up computation even further:

1. Initialise $A_\ell \coloneqq \left\{k \mid \beta^L_{\ell-1,k} \neq 0\right\}$;
2. Perform coordinate descent with λ_ℓ only on the coordinates in A_ℓ, call this result $\hat{\beta}$.
3. Let V be the set of coordinates which violate the KKT conditions;
4. If $V = \varnothing$, set $\hat{\beta}^L_{\lambda_\ell} = \hat{\beta}$. Else, update $A_\ell = A_\ell \cup V$ and return to step 2.

Remark. There are multiple possible algorithms for computing Lasso solutions, coordinate descent is simply one option.

2.3 Extensions of the Lasso

One of the ways to extend the Lasso is to change the loss function (for example, the logistic loss function can be used if $Y_i \in \{-1, 1\}$). There are other extensions.

2.3.1 Structural penalties

Example 2.28 (Group lasso)**.** Let $G_1, \ldots, G_q$ be a partition of $\{1, \ldots, p\}$, then instead of the Lasso penalty we can use

$$\lambda \sum_{j=1}^{q} m_j \left\|\beta_{G_j}\right\|_2,$$

where m_j is typically chosen as $\sqrt{|G_j|}$. This estimator will generally ensure that either the entire group G_j is shrunk to 0 or not at all.

Example 2.29 (Fused lasso)**.** Suppose the (β_j) are ordered, for example if j fixes time. If we expect (β_j) to be piecewise constant (or at least that β^0_j is close to β^0_{j+1}), we can use the penalty

$$\lambda_1 \sum_{j=1}^{p-1} |\beta_j - \beta_{j+1}| + \lambda_2 \|\beta\|_1,$$

where the term $\lambda_2\|\beta\|_1$ can be omitted if shrinkage is not desired.

2.3.2 Debiasing the Lasso

A useful property of the Lasso is that non-significant coefficients can be shrunk towards 0, but this same effect also causes many significant coefficients to be shrunk towards 0. A possible solution is to only use the Lasso to estimate which coefficients are 0 and then use a different estimator (i.e., OLS) for the other coefficients.

Another option is to use a non-convex penalty function which does not shrink large values of $\hat{\beta}$ as much.

3 High-dimensional covariance estimation and PCA

Until now, we've focused on regression. In this chapter, we will assume the samples $x_1, \dots, x_n$ are i.i.d. from some distribution in $\mathbb{R}^d$. We are interested in estimating the covariance matrix of x_i. There are two reasons for this:

1. Covariance can reveal underlying structure relating the variables.
2. When d is large, interpreting the covariance matrix directly is hard. However, its largest eigenvalues and corresponding eigenvectors can reveal the principal modes of variation in the data.

3.1 Covariance estimation

3.1.1 Maximum likelihood in multivariate normal model

Let $x_1, \dots, x_n$ be i.i.d. samples from a $N_d(\mu, \Sigma)$ distribution. We will assume that Σ is invertible with inverse Ω (this inverse is sometimes called the *precision matrix*). Furthermore, we define the *sample covariance*

$$\hat{\Sigma} \coloneqq \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{X})^\top.$$

It can be shown that the maximum likelihood estimate for μ is $\bar{X} \coloneqq \frac{1}{n}\sum_{i=1}^n x_i$, and that the maximum likelihood estimate Ω is

$$\Omega = \min_{\Omega \succ 0} \Big(-\log(\det \Omega) + \operatorname{tr}\Big(\hat{\Sigma}\Omega\Big)\Big).$$

Since this is the minimisation of a convex function over a convex set, we know that Ω is a minimiser if and only if the derivative w.r.t. each element of Ω is 0. From this, we can compute that $(\hat{\Omega}^{\mathrm{ML}})^{-1} = \hat{\Sigma}$. Therefore, if X has full column rank so that $\hat{\Sigma}$ is invertible, we find $\hat{\Omega}^{\mathrm{ML}} = \hat{\Sigma}^{-1}$, and since matrix inversion is a bijective map on the set of positive definite matrices, we conclude $\hat{\Sigma}^{\mathrm{ML}} = \hat{\Sigma}$.

3.1.2 Non-asymptotic error bounds

In the classical setting, we know that the maximum likelihood estimator is asymptotically optimal if we let $n \to \infty$ and keep all other parameters fixed. However, we are interested in *non-asymptotic error bounds*, which do not rely on n going to ∞. From these bounds, we can infer much more about the $n \to \infty$ case, namely, how much the other parameters are allowed to grow with n.

We will assume for simplicity that $\mu = 0$ and that $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$. Recall that the largest (resp. smallest) eigenvalue of a symmetric matrix $A \in \mathbb{R}^{n\times n}$ is given by the maximum (resp. minimum) of $v^\top A v$ over all $v \in S^{n-1}$. It follows that the operator norm of A is given by $\|A\|_{\mathrm{op}} = \max_{v\in S^{n-1}} \left|v^\top A v\right|$.

Theorem 3.1. *Let $X \in \mathbb{R}^{n\times d}$ have i.i.d. rows $x_1, \dots, x_n$ with mean 0 and covariance Σ, so that*

$$\mathbb{E}\Big[e^{\lambda v^\top x_1}\Big] \le e^{\lambda^2\sigma^2/2} \quad \text{for all } v \in S^{d-1}.$$

Then there are universal constants c, C such that $\hat{\Sigma}$ satisfies

$$\mathbb{E}\left[e^{\lambda\left\|\hat{\Sigma}-\Sigma\right\|_{\mathrm{op}}}\right] \le \exp\left(c\left[\frac{\lambda^2\sigma^4}{n} + d\right]\right) \quad \text{for all } |\lambda| \le \frac{n}{16\sigma^2},$$

and for all $\delta > 0$, we have

$$\mathbb{P}\left(\frac{\left\|\hat{\Sigma}-\Sigma\right\|_{\mathrm{op}}}{\sigma^2} \le C\left(\frac{d+\delta}{n} \vee \sqrt{\frac{d+\delta}{n}}\right)\right) \ge 1 - e^{-\delta}.$$

This is a non-asymptotic error bound. Note that, as long as $d/n \to 0$, the estimator $\hat{\Sigma}$ will be consistent.

Proof. The tail probability bound follows from Chernoff bounding, see lecture 17 for the details.

For the proof of (i), cover S^{d-1} with balls of radius $1/8$ centred at points $v_1, \dots, v_N \in S^{d-1}$. It is a known geometric fact that this is possible with $N \le e^{4d}/2$. We will prove that for any symmetric Q we have $\|Q\|_{\mathrm{op}} \le 2 \max_i |v_i^\top Q v_i|$. For this, let $v \in S^{d-1}$ and choose i such that $\|v - v_i\| \le \frac{1}{8}$, then

$$\begin{aligned}
|v^\top Q v| &= |(v - v_i)^\top Q (v - v_i) + 2 v_i^\top Q (v - v_i) - v_i^\top Q v_i| \\
&\le |(v - v_i)^\top Q (v - v_i)| + 2|v_i^\top Q (v - v_i)| + |v_i^\top Q v_i| \\
&\le \|v - v_i\|^2 \|Q\|_{\mathrm{op}} + 2\|v_i\| \|Q\|_{\mathrm{op}} \|v - v_i\| + |v_i^\top Q v_i| \\
&\le \left(\frac{2}{8} + \frac{1}{64}\right) \|Q\|_{\mathrm{op}} + |v_i^\top Q v_i| \le \frac{1}{2} \|Q\|_{\mathrm{op}} + |v_i^\top Q v_i| \\
&\le \frac{1}{2} \|Q\|_{\mathrm{op}} + \max_i |v_i^\top Q v_i|.
\end{aligned}$$

Since the right-hand side is independent of v we can maximise the left-hand side over v and obtain

$$\|Q\|_{\mathrm{op}} \le \frac{1}{2} \|Q\|_{\mathrm{op}} + \max_i |v_i^\top Q v_i| \implies \|Q\|_{\mathrm{op}} \le 2 \max_i |v_i^\top Q v_i|.$$

Now, we apply this to $Q \coloneqq \hat{\Sigma} - \Sigma$. Then we have

$$\mathbb{E}[e^{\lambda \|Q\|_{\mathrm{op}}}] \le \mathbb{E}[e^{2\lambda \max_i |v_i^\top Q v_i|}] \le \sum_{i=1}^N \mathbb{E}[e^{2\lambda v_i^\top Q v_i}] + \mathbb{E}[e^{-2\lambda v_i^\top Q v_i}], \tag{11}$$

where for the final step we use that $\max_i |v_i^\top Q v_i| \in \{\pm v_i^\top Q v_i \mid i = 1, \dots, n\}$.

Next, we claim that for fixed $u \in S^{d-1}$, we have $\mathbb{E}[e^{t u^\top Q u}] \le \exp\left(64 \frac{t^2 \sigma^4}{n}\right)$ for $|t| \le \frac{n}{8\sigma^2}$. For this, observe that

$$u^\top Q u = \frac{1}{n} \sum_{i=1}^n (u^\top x_i)^2 - u^\top \Sigma u,$$

and note that $u^\top \Sigma u = \mathrm{Var}(u^\top x_i) = \mathbb{E}[(u^\top x_i)^2]$ since the x_i have mean zero. Using the fact that the x_i are independent we have

$$\mathbb{E}[\exp(t u^\top Q u)] = \mathbb{E}\left[\exp\left(\frac{t}{n}\left[(u^\top x_1)^2 - u^\top \Sigma u\right]\right)\right]^n = \mathbb{E}[\exp(Z^2 - \mathbb{E}[Z^2])]^n,$$

where $Z \coloneqq \sqrt{\frac{t}{n}} u^\top x_1$ is sub-Gaussian with parameter $\sqrt{|t|\sigma^2/n}$. Therefore, Z^2 satisfies the Bernstein condition with parmeters $(8|t|\sigma^2/n, 4|t|\sigma^2/n)$, and recalling that $|t| \le \frac{n}{8\sigma^2}$ we obtain

$$\begin{aligned}
\mathbb{E}[\exp(Z^2 - \mathbb{E}[Z^2])] &\le 1 + \sum_{k=2}^\infty \frac{\mathbb{E}[(Z^2 - \mathbb{E}Z^2)^k]}{k!} \\
&\overset{\star}{\le} 1 + \sum_{k=2}^\infty 2^{2k+1} \sigma^{2k} \left(\frac{|t|}{n}\right)^k \\
&\overset{\star\star}{\le} 1 + \frac{2^5 \sigma^4 t^2}{n} \frac{1}{1 - 4\sigma^2 |t|/n} \\
&\le 1 + \frac{64 \sigma^2 t^2}{n} \le e^{64\sigma^4 t^2/n},
\end{aligned}$$

where $\star$ follows from Bernstein's inequality and $\star\star$ from a geometric series, which proves the claim.

Substituting this back into eq. (11) using $t = \pm 2\lambda$ shows that for $|\lambda| < \frac{n}{16\sigma^2}$ we have

$$\mathbb{E}[e^{\lambda \|Q\|_{\mathrm{op}}}] \leq 2N e^{256\lambda^2\sigma^4/n} \leq e^{256\lambda^2\sigma^4/n + 4d} \leq e^{256(\lambda^2\sigma^4/n + d)}.$$

□

The upper bound in this theorem is tight when $\Sigma = I$. However, note that our bound in eq. (11) is quite crude, and we can obtain much better rates if we assume that the x_i are normally distributed (in this case, more is known about the supremum).

Furthermore, when the spectrum of Σ is concentrated on a few directions, the rate can be improved significantly. For this, we need a definition.

Definition 3.2. The *effective rank* $r(\Sigma)$ of a nonzero symmetric positive semi-definite matrix Σ with eigenvalues $0 \leq \gamma_d(\Sigma) \leq \cdots \leq \gamma_1(\Sigma)$ is given by

$$r(\Sigma) = \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}} = \frac{\sum_{i=1}^{d} \gamma_i(\Sigma)}{\gamma_1(\Sigma)}.$$

It is easily seen that $r(\Sigma) \leq \mathrm{rank}(\Sigma)$, with equality if and only if all nonzero eigenvalues of Σ are equal. The following theorem (which we will not prove) gives a tighter bound if we can assume that the rows of X are normally distributed and that $r(\Sigma)$ is constrained:

Theorem 3.3. *Suppose X has i.i.d. $N(0, \Sigma)$ rows with $r(\Sigma) \leq cn$ (??). Then there exists C such that for all $\delta > 0$ we have*

$$\mathbb{P}\left(\frac{\left\|\hat{\Sigma} - \Sigma\right\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}} \leq C\left(\frac{r(\Sigma) + \delta}{n} \vee \sqrt{\frac{r(\Sigma) + \delta}{n}} \right) \right) \geq 1 - e^{-\delta}.$$

In comparison with the previous bound, the dimension of the samples d is replaced by the effective rank $r(\Sigma)$ (in fact, d does not occur at all in this theorem!). This means that d can grow rapidly with n, as long as $r(\Sigma)/n \to 0$. This can even be generalised to an infinite-dimensional setting with appropriate definitions.