Chapter 3

Stein’s phenomenon, regularization

3.1 The James-Stein theorem

Consider the normal means model where we observe $Y \sim N_n(\theta, \sigma^2 I)$, for some $n \in \mathbb{N}$ and $\sigma > 0$, and the aim is to estimate the mean $\theta$ (see Example 1.2.2). In the case $n = 1$ it seems rather clear that if we do not know anything about the parameter $\theta$, we can not do much better than estimating it by the observation $Y$. Proving this rigorously is actually not completely trivial, see Exercise 3.5. For larger $n$ it is in fact also not immediately clear whether if we assume no further structure on $\theta$, we can do better than simply using the maximum likelihood estimator for $\hat{\theta}_{\text{MLE}} = Y$. It turns out however that it is possible to perform strictly better, in the sense of expected quadratic error.

To get a first indication of this fact, note that for any estimator $\hat{\theta}$ with a finite covariance we have the bias-variance decomposition

$$E_d \| \hat{\theta} - \theta \|^2 = \| E_d \hat{\theta} - \theta \|^2 + \text{tr} \text{Cov}_d \hat{\theta}$$

(check!). Here $\| \cdot \|$ is the Euclidean norm, $\text{Cov}_X = E(X - EX)(X - EX)^T$ is the covariance matrix of a random vector $X$ and $\text{tr} A$ is the trace of the matrix $A$. If we apply this to $\hat{\theta}_c = cY$ we find that $E_d \| cY - \theta \|^2 = (c - 1)^2 \| \theta \|^2 + c^2 \sigma^2 n$, which, for given $\theta$, is minimal for $c$ equal to

$$c_0 = \frac{\| \theta \|^2}{\| \theta \|^2 + \sigma^2 n},$$

and the minimal value is

$$E_d \| \hat{\theta}_{c_0} - \theta \|^2 = \frac{\sigma^2 n \| \theta \|^2}{\| \theta \|^2 + \sigma^2 n},$$

(3.1)

Since $c_0 < 1$, this indicates that it might be advantageous to shrink the estimator $Y$ towards 0, that is, to multiply it by a factor strictly smaller than 1.
Multiplying by \( c \) does not lead to a valid estimator however, since \( c \) depends on the unknown parameter \( \theta \). It turns out that for \( n \geq 3 \), shrinking by an appropriate data-dependent constant leads to an estimator with an expected squared error that is strictly smaller than that of the MLE.

**Theorem 3.1.1** (James-Stein). Define

\[
\hat{\theta}_{JS} = \left(1 - \frac{\sigma^2(n-2)}{||Y||^2}\right)Y.
\]

For \( n \geq 3 \), we have \( E_\theta ||\hat{\theta}_{JS} - \theta||^2 < E_\theta ||\hat{\theta}_{MLE} - \theta||^2 \) for all \( \theta \in \mathbb{R}^n \).

*Proof.* For the bias and variance of the \( i \)th component of the JS estimator we have

\[
E_\theta \hat{\theta}_{JS,i} - \theta_i = -\sigma^2(n-2)E_\theta \frac{Y_i}{||Y||^2}
\]

and

\[
\text{Var}_\theta \hat{\theta}_{JS,i} = \sigma^2 + \sigma^4(n-2)^2\text{Var}_\theta \frac{Y_i}{||Y||^2} - 2\sigma^2(n-2)\left(E_\theta \frac{Y_i^2}{||Y||^2} - E_\theta \frac{\theta_i Y_i}{||Y||^2}\right),
\]

respectively. (Note that since \( E_\theta 1/||Y||^p \) is finite if and only if \( n > p \), all expectations here are finite for \( n \geq 3 \). See Exercise 3.1.) It follows that the mean squared error of the estimator is given by

\[
\sigma^2n + \sigma^4(n-2)^2E_\theta \frac{1}{||Y||^2} - 2\sigma^2(n-2)\left(\sum E_\theta \frac{Y_i(Y_i - \theta_i)}{||Y||^2}\right)
\]

(check!). By Lemma 3.1.2 below,

\[
E_\theta \frac{Y_i(Y_i - \theta_i)}{||Y||^2} = E_\theta \frac{\sigma^2}{||Y||^2} - 2E_\theta \frac{\sigma^2 Y_i^2}{||Y||^4}.
\]

Hence, the mean squared error (MSE) \( E_\theta ||\hat{\theta}_{JS} - \theta||^2 \) equals

\[
\sigma^2n - \sigma^2(n-2)^2E_\theta \frac{1}{||Y||^2}.
\]

(3.2)

Since the MSE of the MLE \( Y \) equals \( n \), this completes the proof. \( \square \)

**Lemma 3.1.2.** Let \( Y \sim N_n(\theta, I) \) and let \( f : \mathbb{R}^n \to \mathbb{R} \) be an almost everywhere differentiable function such that \( E_\theta |(\partial f/\partial x_i)(Y)| < \infty \) for \( i = 1, \ldots, n \). Then for \( i = 1, \ldots, n \),

\[
E_\theta (Y_i - \theta_i)f(Y) = E_\theta \frac{\partial f}{\partial x_i}(Y).
\]
Proof. Integration by parts, see Exercise 3.2.

The James-Stein theorem gives a number of very interesting insights in statistics for high-dimensional models. It shows that by shrinking the MLE towards zero, thereby reducing the variance at the cost of increasing the bias, we obtain an estimator with a strictly better risk \( E_k \| \hat{\theta} - \theta \|^2 \). Moreover, although the observed \( Y_i \) are independent by assumption, the shrinkage factor depends on all the observations. Hence, to estimate the \( i \)-th component \( \theta_i \), we do not only use the information in \( Y_i \), but we also borrow strength from the other observations.

One argument that Stein (1956) used to intuitively justify the concept of shrinkage is the observation that if \( Y \sim N_n(\theta, I) \), then by the law of large numbers it holds for large \( n \) that \( \| Y \|^2 \approx \| \theta \|^2 + n \). So the norm of the MLE \( Y \) is typically substantially larger than the norm of the parameter \( \theta \) it is supposed to estimate. Therefore, it may be beneficial to shrink the vector \( Y \) so that the norm is reduced.

Alternatively, we may argue that shrinking reduces the contributions of outliers, i.e. relatively large observations \( Y_i \), on the squared estimation error. This possibly comes at the cost of increasing the error made in the other coordinates, but the net effect is that shrinking improves the total squared error \( \| \hat{\theta}_{JS} - \theta \|^2 \) of the estimator on average. Observe that this reasoning indicates that it is essential that we assess the quality of the estimator using a norm that simultaneously takes all coordinates of \( \theta \) into account. This allows us to trade off gains in one coordinate with losses in others.

The James-Stein theorem can be generalized in many directions, for instance away from the normal distribution with unit variance, using other norms, other statistical models, et cetera. The precise form of the shrinking is not crucial either. Shrinking towards a fixed point \( v \in \mathbb{R}^d \) other than 0 works just as well for instance (see Exercise 3.3). The general message is always that in high-dimensional settings it is typically advantageous to somehow reduce the variance by shrinking, or otherwise regularizing. We explore this further in the next section.

Theorem 3.1.1 shows that for \( n \geq 3 \), the MLE \( \hat{\theta}_{MLE} = Y \) is inadmissible in the model \( Y \sim N_n(\theta, I) \), with respect to the squared Euclidean risk. By definition this means that there exists another estimator \( \hat{\theta} \) such that \( E_\theta \| \hat{\theta} - \theta \|^2 \leq E_\theta \| \hat{\theta}_{MLE} - \theta \|^2 \) for all \( \theta \in \mathbb{R}^n \), with strict inequality for at least one \( \theta \in \mathbb{R}^n \). The theorem asserts that the James-Stein estimator is such an estimator. It can be shown however that the James-Stein estimator itself is inadmissible as well. For example the positive part Stein estimator

\[
\hat{\theta}_{JS+} = \left(1 - \frac{n - 2}{\| Y \|^2}\right) Y
\]

is an estimator with strictly smaller risk for all \( \theta \in \mathbb{R}^n \). See for instance Section 3.4 of Tsybakov (2009). Unfortunately, \( \hat{\theta}_{JS+} \) is not admissible either.
Admissible estimators exist, but have more complicated forms. See for instance Section 5.5 of Lehmann and Casella (1998).

3.2 Regularization in the white noise model

The main message that we should pick up from the James-Stein theorem is that if we are estimating a high-dimensional or infinite-dimensional parameter and we want good performance relative to a global norm on the parameter that somehow takes all dimensions into account simultaneously, then we should use some form of shrinking. In Theorem 3.1.1 the shrinking is done quite literally. The MLE is multiplied by a factor smaller than 1, which shrinks the estimator towards 0.

In this section we explore similar methods in the context of the white noise model. It is convenient to work in the sequence formulation (2.4) in which we want to infer the sequence of coefficients \( (f_j) \) in \( \ell^2 \) from the observations \( Y_j \) satisfying

\[
Y_j = f_j + \frac{1}{\sqrt{n}} Z_j, \quad j = 1, 2, \ldots,
\]

where \( Z_1, Z_2, \ldots \) are independent standard Gaussians. We assess the quality of an estimator \( \hat{f} \) by the mean squared error, or squared \( \ell^2 \)-risk \( E_f \| \hat{f} - f \|_2^2 = E_f \sum (f_j - \hat{f}_j)^2 \). The MLE in this model is simply the sequence of observations \( Y_j \). This is clearly not a very useful estimator. It a.s. does not belong to \( \ell^2 \) (why?) and in particular its mean squared error is equal to \( \infty \). As in the preceding section, we can improve the situation by using one of several possible methods to “shrink” the MLE. In this case simply multiplying by a single factor less than 1 will not work, since this would lead to an estimator which still does not belong to \( \ell^2 \). There are various other methods however, which can all be viewed as some form of shrinking.

The key in every method is to optimize the risk by changing the bias-variance trade-off. The risk of an estimator \( \hat{f} \) can in the present situation be decomposed as

\[
E_f \| \hat{f} - f \|_2^2 = \sum (E_f \hat{f}_j - f_j)^2 + \sum \text{Var}_f \hat{f}_j.
\]

As usual, we call the first term the squared bias term and the second one the variance term. In the white noise model the MLE \( Y \) has squared bias 0 and variance \( \infty \). The examples below give methods that reduces the variance, at the cost of introducing some bias. The net effect can then be that the risk of the estimator improves.

Example 3.2.1 (Simple projection estimators). A straightforward method for shrinking the MLE towards 0 is to choose a “dimension” \( J \in \mathbb{N} \) and to truncate all observations above that dimension. In other words, we define the estimator \( \hat{f} \) by setting

\[
\hat{f}_j = Y_j 1_{j \leq J}, \quad j = 1, 2, \ldots
\]
This is called a projection estimator, because it projects the MLE $Y$ onto a finite-dimensional subspace of $\mathbb{R}^\infty$. The risk of the estimator is given by

$$E_f\|\hat{f} - f\|_2^2 = \sum_{j > J} f_j^2 + \frac{J}{n},$$

where the first term is the squared bias and the second one is the variance. For every choice of $J$ the projection estimator beats the MLE in terms of risk. We see from the expression for the risk that the optimal choice of $J$ depends on the rate of decay of the squared coefficients $f_j^2$. □

Example 3.2.2 (Smoothing). Recall that the $f_j$ in the sequence formulation of the white noise model can be interpreted as the Fourier coefficients of the signal $f$ in the model (2.1) relative to some orthonormal basis $e_j$ of $L^2[0, 1]$. In the case of the ordinary Fourier basis, the rate of decay of the Fourier coefficients describes the “smoothness”, or “regularity” of the signal. Roughly speaking: the faster the rate of decay, the less high frequencies the signal contains, the smoother it is (see Exercise 3.6). Viewed from this perspective, the projection estimator can be seen as performing some kind of smoothing. Instead of the MLE, which corresponds to a very rough signal containing all frequencies at the same energy level, it returns a signal in which all Fourier coefficients except the first $J$ are 0. In other words, it projects the MLE onto a space of smoother signals.

We can generalize this idea and consider other methods that somehow project onto spaces of “smooth” signals. For instance, we can fix a sequence of numbers $c_j \in \ell^2$ and consider the estimator $\hat{f}$ defined by

$$\hat{f}_j = Y_j c_j.$$

If $c_j = 1_{j \leq J}$ we recover the simple projection estimator. In other cases it is not exactly equal to the simple projection estimator, but it reduces the Fourier coefficients in a similar way and can be seen as a form of smoothing as well. The risk of the estimator is given by

$$E_f\|\hat{f} - f\|_2^2 = \sum (1 - c_j)^2 f_j^2 + \frac{\|c\|^2_2}{n}. \quad (3.3)$$

Again, the size of the risk is determined by the combination of the choice of $c$ and the rate of decay of the Fourier coefficients. □

Example 3.2.3 (Restricting the parameter space). We can view the shrinkage estimators we have just seen as restricted maximum likelihood estimators, where we maximize the likelihood over a particular subset of (Fourier coefficients of) signals. Indeed, in the sequence model the likelihood (relative to an infinite product of $N(0, 1/n)$-distributions) is proportional to

$$f \mapsto \prod e^{-\frac{1}{2}(Y_j - f_j)^2}.$$
Maximizing this over all possible sequences $f$ gives the MLE $Y$. But if instead we fix $J \in \mathbb{N}$ and consider the set $\mathcal{F}_J = \{ f \in \mathbb{R}^\infty : f_j = 0 \text{ for all } j > J \}$ and maximize the likelihood over $\mathcal{F}_J$, then we obtain as maximizer the simple projection estimator of Example 3.2.1.

More generally, we can define an estimator by choosing some set of sequences $\mathcal{F} \subset \ell^2$ and defining $\hat{f}$ to be the (or a) point in $\mathcal{F}$ where the likelihood is maximal, provided that it exists. Since maximizing the likelihood is equivalent to minimizing the function $f \mapsto \sum (Y_j - f_j)^2$, this estimator is given by

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum (Y_j - f_j)^2.$$ (3.4)

Different choices for $\mathcal{F}$ correspond to different ways of shrinking or smoothing. In the next example we will see for instance that if $\mathcal{F}$ is an ellipsoid of the form $\mathcal{F} = \{ f \in \ell^2 : \sum a_i^2 f_i^2 \leq r^2 \}$, then the restricted MLE coincides with a smoothing estimator as considered in Example 3.2.2. The size of the set $\mathcal{F}$ determines the degree of shrinking in some sense. The smaller the $\ell^2$-diameter of the set $\mathcal{F}$ for instance, the more the MLE is shrunk towards 0.

Example 3.2.4 (Penalization). Consider again the estimator $\hat{f}$ defined as the maximizer of the likelihood over a restricted set of (Fourier coefficients of) signals $\mathcal{F}$. In many situations the minimization problem (3.4) has a dual formulation that can be useful and that gives yet another perspective on shrinkage. To indicate how this works, suppose for concreteness that the set is of the form $\mathcal{F} = \{ f \in \ell^2 : p(f) \leq c \}$ for some function $p : \ell^2 \to \mathbb{R}$ and $c \in \mathbb{R}$. So the restricted MLE is given by

$$\hat{f} = \arg\min_{f \in \ell^2, p(f) \leq c} \sum (Y_j - f_j)^2.$$ (3.5)

Then setting $h = p - c$ this constrained minimization problem is equivalent to the global minimization problem

$$\hat{f} = \arg\min_{f \in \ell^2} \sum (Y_j - f_j)^2 + \infty \cdot 1_{h(f) > 0}.$$  

Now the idea is to replace the function $x \mapsto \infty \cdot 1_{x > 0}$ by the linear function $x \mapsto \lambda x$ for some $\lambda \geq 0$ and to consider instead the minimization of the Lagrangian

$$L(f, \lambda) = \sum (Y_j - f_j)^2 + \lambda h(f).$$

Observe that for every $f \in \ell^2$ we have

$$\sup_{\lambda \geq 0} L(f, \lambda) = \sum (Y_j - f_j)^2 + \infty \cdot 1_{h(f) > 0},$$

hence finding $\hat{f}$ means solving the optimization problem $\inf_{f \in \ell^2} \sup_{\lambda \geq 0} L(f, \lambda)$.  

We might hope that we can interchange the inf and the sup in this problem and that it is equivalent to solve the dual problem
\[ \sup_{\lambda \geq 0} \inf_{f \in \ell^2} L(f, \lambda). \]
This is indeed the case under some conditions. It is easy to see that we always have the inequality
\[ \sup_{\lambda \geq 0} \inf_{f \in \ell^2} L(f, \lambda) \leq \inf_{f \in \ell^2} \sup_{\lambda \geq 0} L(f, \lambda), \tag{3.6} \]
which is called weak duality (see Exercise 3.7). If \( h \) is convex, then the function \( L \) is convex in \( f \) and linear in \( \lambda \). An appropriate version of the minimax theorem (see for instance Section 45 of Strasser (1985)) then implies that for all compact \( \Lambda \subset [0, \infty) \),
\[ \sup_{\lambda \in \Lambda} \inf_{f \in \ell^2} L(f, \lambda) = \inf_{f \in \ell^2} \sup_{\lambda \in \Lambda} L(f, \lambda). \]
Thus we have the desired strong duality if
\[ \inf_{f \in \ell^2} \sup_{\lambda \in [0, \infty]} L(f, \lambda) \to \inf_{f \in \ell^2} \sup_{\lambda \geq 0} L(f, \lambda). \]
This is true for instance if the so-called Slater condition holds, which postulates that there exists an \( f_0 \in \ell^2 \) such that \( h(f_0) < 0 \). In our case this corresponds to the condition that there exists an \( f_0 \in \ell^2 \) such that \( p(f_0) < c \), which is typically very mild. See for instance Section 5.3.2 of Boyd and Vandenberghe (2004).

If we indeed have strong duality and the estimator \( \hat{f} \) can be obtained by solving the dual optimization problem, then for some \( \lambda > 0 \), called the Lagrange multiplier, we have
\[ \hat{f} = \arg\min_{f \in \ell^2} \sum_{j} (Y_j - f_j)^2 + \lambda p(f). \tag{3.7} \]
The multiplier \( \lambda \) will obviously depend on the constant \( c \) in the primal formulation (3.5) of the problem.

An estimator like (3.7) is called a penalized maximum likelihood estimator and the term \( \lambda p(f) \) that is added to the likelihood is called a penalty term. The idea is that functions \( f \) for which the penalty \( p(f) \) is large are considered to be undesirable as estimators. Minimizing the penalized least squared criterion on the right of (3.7) means balancing the likelihood term and the penalty term, i.e. finding functions \( f \) for which the the likelihood is large, but at the same time the penalty is small. The tuning parameter \( \lambda \) controls the relative importance of the likelihood and the penalty. For \( \lambda = 0 \) the penalty disappears and we just have the MLE. As \( \lambda \) becomes larger, the penalty becomes more important, which corresponds to more and more regularization or shrinking.
Penalties that are routinely used include various kinds of norms on the function $f$ or its derivatives.

As a concrete example, consider a penalty of the form

$$p(f) = \sum a_j^2 f_j^2$$

for some sequence $a_j$. (If $a_j = j^\beta$ for some $\beta > 0$, then we can think of $p(f)$ as the squared $L^2$-norm of the $\beta$th derivative of the function $f$ with Fourier coefficients $f_k$, cf. Exercise 3.6.) Then the minimization in (3.7) can be done coordinatewise and we find that the penalized MLE $\hat{f}$ is in this case given by

$$\hat{f}_j = \frac{1}{1 + \lambda a_j^2} Y_j$$

(check!). So we see that the smoothing estimators considered in Example 3.2.2 can also be viewed as particular penalized MLEs. Note that the tuning parameter $\lambda$ indeed controls the degree of shrinkage.

$\blacksquare$

**Example 3.2.5 (Bayesian methods).** The expression (3.7) for the penalized MLE points at a connection to Bayesian methods. Doing Bayesian inference in the white noise model involves putting some prior distribution on $f$ and computing the corresponding posterior. As noted in Example 3.2.3 the likelihood is a multiple of $\prod \exp(-\frac{1}{2}(Y_j - f_j)^2)$. Hence if the prior has density $\pi$ relative to some dominating measure, then the posterior has density proportional to

$$f \mapsto \pi(f) \prod e^{-\frac{1}{2}(Y_j - f_j)^2}.$$

In particular, we see that the posterior mode, or maximum a-posteriori (MAP) estimator, is given by

$$\hat{f}_{\text{MAP}} = \arg\min_f \sum (Y_j - f_j)^2 - \frac{2}{n} \log \pi(f).$$

(3.8)

This shows that the MAP can be viewed as a penalized MLE which shrinks the estimator towards the mode of the prior. The smaller the prior density $\pi$ at a function $f$, the more $f$ is penalized.

As a concrete example, suppose we put a product of normals as prior on $f_j$ with under the prior $f_j \sim N(0, \tau_j^2)$ for some sequence of variances $\tau_j^2$. Then since both the likelihood and the prior factorize, the posterior factorizes as well and we can compute the the posterior coordinatewise. The posterior density for $f_j$ is proportional to

$$e^{-\frac{1}{2}(Y_j - f_j)^2} e^{-\frac{1}{2}f_j^2/\tau_j^2}.$$  

Completing the square we find that the posterior is a product of normals

$$\propto N\left(\frac{n \tau_j^2}{1 + n \tau_j^2} Y_j, \frac{\tau_j^2}{1 + n \tau_j^2}\right).$$
(check!). In particular, the posterior mean is the estimator \( \hat{f} \) given by

\[
\hat{f}_j = \frac{n\tau_j^2}{1 + n\tau_j^2} Y_j, \quad j = 1, 2, \ldots
\] (3.9)

Hence, the posterior mean is a shrinkage, or smoothing estimator and the prior variances \( \tau_j^2 \) determine the degree of smoothing. We note that since the posterior is Gaussian in this case, its mean is in fact equal to its mode. Indeed, in this example it can be verified directly that the estimator defined by (3.8) coincides with the estimator \( \hat{f} \) given above (Exercise 3.8).

### 3.3 Regularization in other models

In these notes we mostly focus on the analysis of the canonical signal-in-white-noise model. The idea of regularization is general however and plays a role in some form in every high-dimensional or nonparametric problem. In this section we give a few more well-known examples.

#### Example 3.3.1 (Kernel density estimation)

Consider the density estimation problem of Example 1.2.3. So we have an i.i.d. sample \( X_1, \ldots, X_n \) from a positive density \( p \) on the line. In this model too, it is necessary to use some form of regularization. The likelihood is given by

\[ p \rightarrow \prod p(X_j). \]

By considering a sequence of densities with higher and higher spikes at the observations it is seen that the maximum over all densities is \( \infty \), hence the unrestricted MLE does not exist.

Although strictly speaking it does not exist as a function, the MLE can be thought of as a sum of delta functions at the observations. The classical kernel density estimator can be seen as a way of regularizing this “estimator”. Instead of putting a delta function at each observation we distribute mass \( 1/n \) over the line by putting a smooth “bump” of mass \( 1/n \) at each observation. To do so, we fix a kernel function \( K \), which is a function that integrates to 1 and has mean 0 (usually a probability density) and a so-called bandwidth parameter \( h > 0 \). Then at every observation \( X_j \) we put the “bump” \( x \rightarrow (nh)^{-1} K((x - X_j)/h) \), which is a function that integrates to \( 1/n \). The kernel estimator \( \hat{p}_h \) for \( p \) is thus defined as

\[
\hat{p}_h(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right), \quad x \in \mathbb{R}.
\]

Note that \( h \) controls the smoothness of the estimator. If \( h \) is very small each bump looks like a delta function and \( \hat{p}_h \) is very spiky. If \( h \) becomes very large the bumps look more and more constant and \( \hat{p}_h \) “shrinks” towards a “uniform” distribution.

In Exercise 3.9 you are asked to explore kernel density estimation in R. ■
**Example 3.3.2** (Estimating a monotone density). Kernel density estimators provide smooth estimators of the unknown density, the degree of smoothness being controlled in some sense by the bandwidth. If the true density itself is believed to be smooth, this kind of regularization makes perfect sense. There are other methods of regularizing as well, for instance by restricting shape instead of smoothness.

To illustrate this, consider an i.i.d. sample $X_1, \ldots, X_n$ from a density $p$ on $[0, 1]$. If the true density is believed to be decreasing, it makes sense to consider the class $P_{\text{dec}}$ of decreasing densities on $[0, 1)$ and to consider the restricted MLE over that class, i.e.

$$
\hat{p} = \arg\max_{p \in P_{\text{dec}}} \prod p(X_i).
$$

It can be seen that this criterion is maximized at densities that are left-continuous, locally constant, with jumps only at the observations (check!). For such a maximizer the constant value between 0 and $X_1$ can be at most $1/X_1$, since the density must integrate to 1. Hence, if we denote the function values of a potential maximizer from left to right by $h_1, \ldots, h_n$, then $h = (h_1, \ldots, h_n)$ must belong to the set

$$
\mathcal{H} = \{h \in \mathbb{R}^n : 0 \leq h_1 \leq \cdots \leq h_n \leq 1/X_1; \sum h_j(X_{(j)} - X_{(j-1)}) = 1\}.
$$

Maximizing the likelihood over $P_{\text{dec}}$ then boils down to maximizing the continuous function $h \mapsto \prod h_j$ over the compact set $\mathcal{H}$. This shows that the restricted MLE $\hat{p}$ exists.

It can be shown that the solution is unique and that it is the left derivative of the so-called *least concave majorant* of the empirical distribution function. As the name suggests, the latter is the smallest concave function that lies pointwise above the empirical distribution function. This is a piecewise linear function with kinks at the observations, hence its derivative is indeed a piecewise constant function with jumps at the observations. This monotone density estimator is known as the *Grenander estimator*.

In Exercise 3.10 you are asked to compare the Grenander estimator to kernel density estimation on simulated data. $\blacksquare$

**Example 3.3.3** (Ridge regression). Consider a linear model $Y = X\theta + \varepsilon$, with $X$ a fixed $n \times p$ matrix, $\theta \in \mathbb{R}^p$ and $\varepsilon \sim N_n(0, \sigma^2 I)$. The MLE for $\theta$ is the least squares estimator, which minimizes the residual sum of squares $\theta \mapsto \|Y - X\theta\|^2$. Equivalently, it is the solution of the linear system $X^T X \theta = X^T Y$. In the high-dimensional setting $p > n$ this system does not have a unique solution however, since $X^T X$ is a $p \times p$ matrix with rank at most $n < p$. To remedy the situation a form of regularization can be employed, for instance by adding a penalty term to the least squares criterion. One method that is often used is adding a penalty proportional to the squared Euclidean norm of $\theta$. This is called the *ridge penalty* and the resulting procedure is called *ridge regression*. Specifically,
the ridge regression estimator is defined by choosing a tuning parameter \( \lambda > 0 \) and defining
\[
\hat{\theta} = \arg\min_{\theta} \|Y - X\theta\|^2 + \lambda \|\theta\|^2.
\]
A simple calculation shows that the minimizer satisfies the equation \( X^T Y - X^T X \hat{\theta} = \lambda \hat{\theta} \), i.e. the ridge regression estimator is given by
\[
\hat{\theta} = \left( X^T X + \lambda I \right)^{-1} X^T Y
\]
(check!). We see that adding the ridge penalty indeed results in a degree of shrinkage towards 0, controlled by the tuning parameter \( \lambda \).

**Example 3.3.4 (The lasso).** Consider again the high-dimensional linear model as in the preceding example. In ridge regression the penalty is the Euclidean, or \( \ell^2 \)-norm of the parameter \( \theta \). Another very popular method is adding an \( \ell^1 \)-penalty. That is, we define \( \|\theta\|_1 = \sum |\theta_j| \), choose some tuning parameter \( \lambda > 0 \) and define the estimator for \( \theta \) by
\[
\hat{\theta} = \arg\min_{\theta} \|Y - X\theta\|^2 + \lambda \|\theta\|_1.
\]
This estimator is called the lasso (least absolute shrinkage and selection operator) estimator. Contrary to the ridge regression estimator, the lasso can in general not be expressed explicitly. Computing it is still a convex optimization problem however and there exist efficient algorithms to compute it numerically. An important feature of the lasso is that since it minimizes a convex function over an \( \ell^1 \)-ball (see the dual perspective explained in Example 3.2.4), it typically produces estimators with many 0 coordinates. This is a desirable feature in many applications in which the true parameter is believed to be sparse, i.e. to have many 0 coordinates.

This is most easily seen in the case \( X = I \), i.e. in the normal means model. It can be verified that in that case the lasso defined above is given explicitly by
\[
\hat{\theta}_i = h_\lambda(Y_i),
\]
where \( h_\lambda \) is the function defined by
\[
h_\lambda(y) = \begin{cases} 
  y + \lambda/2 & \text{if } y < -\lambda/2, \\
  0 & \text{if } -\lambda/2 \leq y \leq \lambda/2, \\
  y - \lambda/2 & \text{if } y > \lambda/2
\end{cases}
\]
(see Exercise 3.11). The lasso is therefore called a soft thresholding procedure: if \( |Y_i| \) is below the threshold \( \lambda/2 \), then the estimator for \( \theta_i \) is 0. For larger values the estimator moves away from 0 linearly. For \( |Y_i| \) very large the impact of \( \lambda \) becomes negligible and we have \( \hat{\theta}_i \approx Y_i \).
Example 3.3.5 ($\ell^0$-penalization). Another penalty that is used in the high-dimensional linear model setting is the $\ell^0$-norm. The $\ell^0$-norm $\|\theta\|_0$ of a vector in $\mathbb{R}^n$ is defined as the number of non-zero coordinates: $\|\theta\|_0 = |\{i : \theta_i \neq 0\}|$. This penalty penalizes non-sparse vectors even more clearly than the $\ell^1$-penalty of the lasso. The corresponding penalized MLE for $\theta$ is defined by

$$\hat{\theta} = \arg\min_{\theta} \|Y - X\theta\|^2 + \lambda \|\theta\|_0,$$

(3.10)

for some tuning parameter $\lambda > 0$. Computing this estimator is in general computationally hard if $p$ is large, since it is not a convex optimization problem. The computation essentially involves searching over all possible sparsity patterns. The lasso can be viewed as a convex relaxation of this problem. It still has much of the same flavour, but is computationally much more feasible.

Also in this case we can compute the estimator explicitly in the normal means model, i.e. when $X = I$. It can be seen that the $\ell^0$-penalized estimator is then given by

$$\hat{\theta}_j = Y_j 1_{|Y_j| > \sqrt{\lambda}}$$

(see Exercise 3.12). For obvious reasons it is therefore called a hard thresholding estimator.

The examples we gave in this section treat some important classes of regularization methods in the context of several classical statistical models. It should be clear that this list is by no means complete and that many variations on this theme exist. See exercises 3.13 and 3.14 for some more examples.

3.4 The bias-variance trade-off and convergence rates

We have seen in the James-Stein theorem and in the examples in this chapter that in high-dimensional or nonparametric models, naively applying methods like maximum likelihood is either simply not possible, or leads to procedures with sub-optimal performance in terms of mean squared error, for instance. Regularization methods improve the situation by changing the bias-variance trade-off. They yield estimators with reduced variance, at the cost of introducing some bias. The net effect is generally an estimator with better overall performance.

The variance of an estimator quantifies how much the estimator depends on the particular dataset. Of course the estimator should depend on the dataset, but if the variance is too large it will depend too much on the particular realisation that has been observed and will not provide a stable estimator of the underlying truth. On the other hand if the variance is too small then the bias will be relatively large which results in an estimator which is quite stable, but which makes a large structural error. If the variance is too large, we also say that the estimator is overfitting, reflecting the idea that it follows the fluctuations in the particular dataset too closely. If the variance is too small, hence the bias too large, we say we are underfitting.
Example 3.4.1 (over- and undersmoothing in kernel density estimation). To illustrate the issue of over- and underfitting we simulate i.i.d. data from a known density and try to recover it using a kernel density estimator as in Example 3.3.1. Figure 3.1 shows kernel density estimates for 6 different samples from the same density, namely a mixture of two normals \( N(\pm 1, 0.5) \), with equal weights 1/2.

The performance depends crucially on the bandwidth parameter \( h \) that is chosen. In the top row \( h \) was chosen too small on purpose. The estimator then closely follows the particular data that is observed. As a result the variance of the estimator becomes too large, i.e. we are overfitting. In this context this is also called \textit{undersmoothing}, meaning that we do not regularize enough, the estimator is too “rough” compared to the truth. In the lower row the bandwidth is chosen too large. The estimator is very stable, but we introduce a large bias and the estimator remains far away from the truth. This is called \textit{oversmoothing}: the estimator is too “smooth” relative to the truth. Choosing of a good bandwidth means finding a comprise between these two extremes.
As we have seen either explicitly or implicitly in the examples, all regularization methods have some tuning mechanism to control the bias-variance trade-off. For the simple projection estimator it is the dimension $J$ at which the MLE is truncated, for the penalization methods it is the tuning parameter $\lambda$ which multiplies the penalty, for the restricted MLE it is the size of the set over which the likelihood is maximized, for the kernel density estimator it is the bandwidth $h$, et cetera. What we can also see from the examples, for instance from expressions for the risk like (3.3) or from numerical examples like Example 3.4.1, is that the optimal choice of the tuning parameters will depend on properties of the true parameter. Unfortunately, these are usually properties like some form of “smoothness”, or another measure of the “complexity” of the true parameter that we don’t have direct access to, since the true parameter is unknown.

It is useful however to understand how the tuning parameters of a given procedure should be set if properties like smoothness were in fact known. This optimal setting is sometimes called the oracle setting of a tuning parameter, reflecting the idea that an oracle that knows aspects of the truth that are inaccessible to us would know how to set it. Knowing this oracle choice of a tuning parameter gives insight in the best performance that can in principle be achieved with a given method.

One way to assess the performance of a statistical procedure is to take an asymptotic perspective and study how the risk of the estimator behaves as $n \to \infty$, where $n$ denotes the “amount” of data that is available. In the case of i.i.d. observations $n$ is the sample size, in the case of the white noise model it is the signal-to-noise ratio $n$ in (2.1). Typically the risk of an estimator will tend to 0 as $n \to \infty$. Of course the rate at which this happens depends on the particular risk that it is considered. In these notes we often consider risks of the form $\mathbb{E}|\hat{\theta}_n - \theta|^p$, for some norm $\| \cdot \|$ on the parameter of interest and $p = 1$ or $p = 2$. If the risk tends to 0 as $n \to \infty$, then by Markov’s inequality the distance $\|\hat{\theta}_n - \theta\|$ of the estimator to the true parameter is of stochastic order $(\mathbb{E}|\hat{\theta}_n - \theta|^p)^{1/p}$. This is then called the rate of convergence of the estimator (relative to the norm $\| \cdot \|$). More generally, if $\|\hat{\theta}_n - \theta\| = O_{\text{p}}(r_n)$ for some $r_n \to 0$, then $r_n$ is called a rate of convergence of the estimator $\hat{\theta}_n$.

As discussed in Section 1.1, estimators in smooth parametric models typically have rate of convergence $1/\sqrt{n}$ (relative to the Euclidean norm). Rates in nonparametric or high-dimensional problems are usually slower.

**Example 3.4.2** (Simple projection estimator, ellipsoids). Consider the situation of Example 3.2.1 again. We can optimize the risk by tuning the truncation point $J$ to the rate of decay of the true sequence $f_j$. Indeed, suppose that for some $\beta, r > 0$, $f$ belongs to the ellipsoid

$$
\mathcal{F} = \{ f : \sum j^{2\beta} f_j^2 \leq r^2 \}.
$$
For $f \in \mathcal{F}$ the squared bias satisfies
\[ \sum_{j > J} f_j^2 = \sum_{j > J} j^{-2\beta} j^{2\beta} f_j^2 \leq J^{-2\beta} r^2. \]
This is balanced with the variance term $J/n$ for $J \sim n^{1/(1+2\beta)}$. The risk of the estimator is then of the order $n^{-2\beta/(1+2\beta)}$, hence the rate of convergence is of the order $n^{-\beta/(1+2\beta)}$.

Recall that the sequence formulation (2.4) of the white noise model can be seen as the “frequency domain” view, which corresponds to the “time domain” view (2.1). In the time domain, the assumption that the sequence of Fourier coefficients belongs to $\mathcal{F}$ can be viewed as a “smoothness” condition on the function in (2.1). As explained in Exercise 3.6, the condition that the Fourier coefficients of a function $f$ belong to the ellipsoid $\mathcal{F}$ corresponds to the smoothness condition that the signal $f$ has $\beta$ square integrable derivatives, with $L^2$-norms bounded by a common constant. Such a space of functions is called a Sobolev ball of order $\beta$ (a precise definition is given ahead).

Using this terminology, we see in this example that if the signal $f$ has Sobolev smoothness $\beta$, then the projection estimator with $J \sim n^{1/(1+2\beta)}$ converges in $\ell^2$ to the true $f$ at the rate $n^{-\beta/(1+2\beta)}$. For very small $\beta$, i.e. very rough signals, this rate is very slow. For $\beta \rightarrow \infty$ the rate comes closer and closer to the parametric rate $1/n$. We will see in Chapter 5 that the rate $n^{-\beta/(1+2\beta)}$ is in some sense the best we can get if we want to do well for all signals in a Sobolev ball of order $\beta$.

For kernel density estimation, as considered in Examples 3.3.1 and 3.4.1, results similar those for the projection estimator can be obtained both for the pointwise risk $E_p(\hat{p}_h(x) - p(x))^2$, with $x$ a fixed point, and for the integrated risk $E_p \int (\hat{p}_h(x) - p(x))^2 \, dx$. A good bias-variance trade-off can be achieved by appropriately tuning the bandwidth parameter $h$ to the smoothness of the unknown density $p$, see Exercise 3.15.

The following two examples consider kernel smoothing in the context of the white noise model. We investigate in particular how the rate changes if we consider uniform risk instead of pointwise risk. In this setting the natural notion of smoothness to consider is Hölder regularity. We say a function $f$ on $[0,1]$ has Hölder regularity $\beta > 0$ if, for $\beta$ the largest integer strictly smaller than $\beta$, it has $\beta$ continuous derivatives and the highest derivative $f^{(\beta)}$ satisfies
\[ |f^{(\beta)}(t) - f^{(\beta)}(s)| \leq C|t - s|^{(\beta - \beta)} \]
for a constant $C > 0$ and all $s, t \in [0,1]$. If this holds we write $f \in C^\beta[0,1]$.

Example 3.4.3 (Kernel smoothing in the white noise model, pointwise risk). Consider the white noise model (2.1). To estimate the signal $f$ we fix a symmetric, smooth, compactly supported probability density $K$ and a bandwidth
We define the estimator $\hat{f}_h$ by

$$\hat{f}_h(t) = \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) dX_s, \quad t \in [0,1],$$

(3.11)

where the integral is a Wiener integral as defined in Section 2.2.

Now fix a point $t \in (0,1)$. Using the SDE for $X$ we see that

$$\hat{f}_h(t) = \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) f(s) ds + \frac{1}{\sqrt{n}} \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) dW_s,$$

Hence, it follows from the properties of the Wiener integral that the bias and the variance of the estimator $\hat{f}_h(t)$ are given by

$$E_f \hat{f}_h(t) - f(t) = \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) f(s) ds - f(t),$$

$$\text{Var}_f \hat{f}_h(t) = \frac{1}{n} \int_0^1 \frac{1}{h^2} K^2\left(\frac{t-s}{h}\right) ds,$$

respectively. Note that the variance does not depend on the true signal $f$. A change of variables shows it is of the order $1/(nh)$. To bound the bias, first note that by a change of variables and using that the support of $K$ is contained in a compact interval $[-a,a]$ for some $a > 0$, we have, for $h$ small enough,

$$E_f \hat{f}_h(t) - f(t) = \int_{-a}^a K(u)(f(t-hu) - f(t)) du$$

(check!). Now suppose that $f \in C^\beta[0,1]$ for some $\beta \in (0,2]$. Let $\beta$ be the largest integer strictly smaller than $\beta$. Then by Taylor’s formula,

$$f(t-hu) = \sum_{k \leq \beta} f^{(k)}(t)(-hu)^k + r,$$

where $|r| \leq \text{const} \times |hu|^{\beta}$. Since $K$ integrates to 1 and its first moment is 0, we conclude that

$$\left|E_f \hat{f}_h(t) - f(t)\right| \leq \text{const} \times h^\beta \int_{-a}^a |K(u)||u|^{\beta} du.$$

So the squared bias is bounded by a constant times $h^{2\beta}$ if $f \in C^\beta$. Taking the bandwidth $h$ of the order $n^{-1/(1+2\beta)}$ then balances squared bias and variance and leads to a mean squared error, or pointwise risk $E_f (\hat{f}_h(t) - f(t))^2$ of the order $n^{-2\beta/(1+2\beta)}$. Hence, the pointwise rate of convergence is of the order $n^{-\beta/(1+2\beta)}$.

It is also possible to get the rate $n^{-\beta/(1+2\beta)}$ if $f \in C^\beta[0,1]$ for $\beta > 2$, but then the probability kernel $K$ has to be replaced by a so-called higher order kernel, which has 0 moments of sufficiently high order.
In the preceding example we only examined the pointwise risk for points \( t \) strictly inside \((0, 1)\). This is because for \( t = 0 \) or \( t = 1 \) some boundary issues arise with the kernel estimator \((3.11)\). The variance is still of the same order (but the constants are different), but the behaviour of the bias changes. Indeed, for \( t = 0 \) for instance, we get that the bias is

\[
E_f \hat{f}_h(0) - f(0) = \int_0^a K(u)(f(-hu) - f(0)) du
\]

for \( h \) small enough (check!). For \( \beta \in (0, 1] \) this is still of the order \( h^\beta \), but for \( \beta \in (1, 2) \) this is no longer necessarily true. The intuitive reason for the different behaviour of the estimator at the boundaries is that for \( t \) in the interior, \((3.11)\) in some sense uses observations \( X_s \) both to the left and to the right of \( t \) in a symmetric manner. At the boundaries, one could say the estimator somehow uses only half the information it uses in the interior.

In the next example we investigate how the rate of the kernel estimator changes if we replace the pointwise risk by uniform risk. To avoid the technical boundary issues just indicated, we consider the uniform distance over a compact subinterval of \((0, 1)\). Since uniform convergence is stronger than pointwise convergence, we can expect the rate for the uniform risk to be worse than the one obtained in the preceding example. It turns out that the difference in the rate is only a logarithmic factor.

**Example 3.4.4** (Kernel smoothing in the white noise model, uniform risk). We consider the same estimator \( \hat{f}_h \) as in Example 3.4.3, but now we study how the uniform risk \( E_f \sup_{u \leq t \leq v} |\hat{f}_h(t) - f(t)| \) behaves for fixed \( 0 < u < v < 1 \). We assume again that \( f \in C^\beta[0, 1] \) for \( \beta \in (0, 2] \).

For every \( t \in [0, 1] \) we have

\[
|\hat{f}_h(t) - f(t)| \leq \left| \int_0^1 \frac{1}{h} K\left( \frac{t - s}{h} \right) f(s) ds - f(t) \right| + \frac{1}{\sqrt{n}} \left| \int_0^1 \frac{1}{h} K\left( \frac{t - s}{h} \right) dW_s \right|.
\]

By the bias considerations in Example 3.4.3 the first term is bounded by a constant times \( h^\beta \), uniformly for \( t \in [u, v] \). Hence, it remains to investigate the quantity

\[
E_f \sup_{u \leq t \leq v} \left| \int_0^1 \frac{1}{h} K\left( \frac{t - s}{h} \right) dW_s \right|.
\]

This is the expectation of the supremum of a centered Gaussian process and we can use some general bounds that are available for such quantities. For instance, Dudley’s maximal inequality for (sub-)Gaussian processes asserts that for a centered Gaussian process \( Z \) indexed by a set \( T \) and any \( t_0 \in T \) it holds that

\[
E \sup_{t \in T} |Z_t| \leq E|Z_{t_0}| + C \int_0^{\text{diam}(T)} \sqrt{\log N(\varepsilon, T, d)} d\varepsilon,
\]
where $C > 0$ is a universal constant, $d$ is the semimetric on $T$ given by $d^2(s, t) = E(Z_t - Z_s)^2$ and $N(ε,T,d)$ is the minimal number of balls of $d$-radius $ε$ that is needed to cover $T$ (see for instance Van der Vaart and Wellner (1996)). We apply this inequality with $Z_t = Z_{t_0}^1 h K_{\frac{t}{h}} ds.

$T = [u, v]$ and $t_0 = u$. In Example 3.4.3 we already saw that the variance of $Z_{t_0}$ is bounded by $1/h$. By Gaussianity, this implies that $E|Z_{t_0}|$ is of the order $1/p h$. These semimetric $d$ is in this case given by

$$d^2(t_1, t_2) = \int_{t_1}^{t_2} \frac{1}{h} K(\frac{t_1 - s}{h}) - \frac{1}{h} K(\frac{t_2 - s}{h}) ds.$$

If we assume that $K$ has support $[-1, 1]$ for concreteness, we see from this expression that two points $t_1, t_2 \in T$ such that $|t_1 - t_2| > 2h$ are at $d$-distance of the order $1/\sqrt{h}$. It follows that the diameter of $(T,d)$ is of the order $1/\sqrt{h}$ as well. Next, we have

$$K(\frac{t_1 - s}{h}) - K(\frac{t_2 - s}{h}) = \int_{t_1}^{t_2} \frac{1}{h} K'(\frac{x - s}{h}) dx.$$

By Jensen’s inequality and Fubini, it follows that

$$d^2(t_1, t_2) \leq \frac{t_2 - t_1}{h^3} \int_{t_1}^{t_2} \left( \int_{0}^{1} \frac{1}{h} (K')^2(\frac{x - s}{h}) ds \right) dx.$$

Since $K$ has compact support the inner integral is bounded by a constant, hence $d^2(t_1, t_2) \leq C(t_1 - t_2)^2/h^3$ for a constant $C > 0$. This implies that the covering number $N(ε,T,d)$ is bounded by a constant times $1/(εh^{3/2})$ and therefore

$$\int_0^{diam(T)} \sqrt{\log N(ε,T,d)} dε = \frac{1}{h^{3/2}} \int_0^{h} \sqrt{\log \frac{1}{x}} dx \leq \frac{\log 1/h}{\sqrt{h}}.$$

All together, we conclude that

$$E_{\tilde{f}} \sup_{u \leq t \leq v} |\tilde{f}_h(t) - f(t)| \lesssim h^\beta + \frac{\sqrt{\log 1/h}}{\sqrt{n h}}.$$

The two terms on the right are balanced for $h$ of the order $(n/\log n)^{-1/(1+2\beta)}$ and the resulting convergence rate relative to the uniform norm is of the order $(n/\log n)^{-\beta/(1+2\beta)}$.

The examples we have seen show that in the white noise model we have an estimator for the unknown signal $f$ that attains the rate $n^{-\beta/(1+2\beta)}$ relative to
the $L^2$-norm if $f$ belongs to a Sobolev space of order $\beta$, we have an estimator that attains the same rate pointwise if $f$ belongs to a Hölder space of order $\beta$ and under the same Hölder condition we have an estimator that attains the slightly slower rate $(n/\log n)^{-\beta/(1+2\beta)}$ relative to the uniform norm. Closer inspection of the examples shows that the bounds on the different risks actually hold uniformly for $f$ in Sobolev or Hölder balls of order $\beta$.

Concretely, for $f \in C^\beta[0,1]$, we define the Hölder norm of order $\beta$ by

$$
\|f\|_{C^\beta} = \max_{k \leq 2} \|f^{(k)}\|_\infty + \sup_{s \neq t} \frac{|f^{(2)}(t) - f^{(2)}(s)|}{|t-s|^{\beta}}.
$$

The Hölder ball of order $\beta$ and radius $R$ is then defined as $C^\beta_R[0,1] = \{ f \in C^\beta[0,1] : \|f\|_{C^\beta} \leq R \}$. Inspection of Example 3.4.3 shows that for the appropriate choice of the bandwidth $h$, the kernel estimator $\hat{f}_h$ satisfies

$$
\sup_{f \in C^\beta_1[0,1]} E_f(\hat{f}_h(t) - f(t))^2 \leq \text{const.} \times n^{-2\beta/(1+2\beta)} \quad (3.12)
$$

for all $t \in (0,1)$. Similarly, Example 3.4.4 in fact shows that for an appropriate (different!) choice of the bandwidth, we have

$$
\sup_{f \in C^\beta_2[0,1]} E_f \sup_{u \leq t \leq v} |\hat{f}_h(t) - f(t)| \leq \text{const.} \times \left( \frac{n}{\log n} \right)^{-\beta/(1+2\beta)} \quad (3.13)
$$

for all $0 < u < v < 1$. The result of Example 3.4.2 is uniform over balls as well. For $\beta > 0$ and $e_j$ an orthonormal basis of $L^2[0,1]$, we define the associated Sobolev space of order $\beta$ by

$$
H^\beta[0,1] = \{ f = \sum f_j e_j : \sum j^{2\beta} f_j^2 < \infty \}.
$$

The Sobolev norm of a function $f = \sum f_j e_j$ in $H^\beta[0,1]$ is defined by

$$
\|f\|_{H^\beta}^2 = \sum j^{2\beta} f_j^2.
$$

(Recall that if $e_j$ is the standard Fourier basis and $\beta$ is an integer, then this is essentially the $L^2$-norm of the $\beta$th derivative of $f$, see Exercise 3.6.) The Sobolev ball of order $\beta$ and radius $R$ is $H^\beta_R[0,1] = \{ f \in H^\beta[0,1] : \|f\|_{H^\beta} \leq R \}$. The result of Example 3.4.2 is that for the appropriate truncation level $J$, the projection estimator $\hat{f}_J$ satisfies

$$
\sup_{f \in H^\beta_1[0,1]} E_f \|\hat{f}_J - f\|_2^2 \leq \text{const.} \times n^{-2\beta/(1+2\beta)}. \quad (3.14)
$$

We will see in Chapter 5 that the rates (3.12)–(3.14) we found are in fact the best possible, in the sense that no other estimators can achieve better risks uniformly over these balls.

The examples considered so far give convergence rates of estimators under smoothness conditions. We end the section with an example of a risk bound under a sparsity condition.
Example 3.4.5 ($\ell^0$-penalization, sparse signals). Consider the normal means model $Y = \theta + \varepsilon$, with $\varepsilon \sim \mathcal{N}_n(0, I)$. Then the $\ell^0$-penalized estimator (3.10) is given by $\hat{\theta}_j = Y_j^\dagger 1_{|Y_j| \geq \sqrt{\lambda}}$, see Example 3.3.5. In this case the bias-variance trade-off can be optimized by tuning $\lambda$ to the sparseness of the true signal $\theta$. Let $\theta$ be $s$-sparse, mean that it has $s$ non-zero components. Denote its support set, which is the set of indices of the non-zero components, by $S$.

Note that $\hat{\theta}_j$ is symmetric and hence unbiased if $j \notin S$. If $j \in S$, then

$$(E_\theta \hat{\theta}_j - \theta_j)^2 = (E_\theta Y_j - \theta_j + E_\theta Y_j^\dagger 1_{|Y_j| < \sqrt{\lambda}})^2 \leq \lambda.$$ 

Hence, the squared bias term in the risk $E_\theta ||\hat{\theta} - \theta||^2$ is bounded by $s\lambda$. As for the variance, note that for $j \notin S$ we have

$$\text{Var}_\theta \hat{\theta}_j = E_\theta Y_j^2 1_{|Y_j| \geq \sqrt{\lambda}} \lesssim \lambda e^{-\lambda/2}$$

for $\lambda \geq 1$ (see Exercise 3.17). For $j \in S$,

$$\text{Var}_\theta \hat{\theta}_j \leq E_\theta (\hat{\theta}_j - \theta_j)^2 = E_\theta ((Y_j - \theta_j) + Y_j^\dagger 1_{|Y_j| < \sqrt{\lambda}})^2 \leq 2E_\theta (Y_j - \theta_j)^2 + 2E_\theta Y_j^2 1_{|Y_j| < \sqrt{\lambda}} \leq 2 + 2\lambda.$$ 

Together, we find that

$$E_\theta ||\hat{\theta} - \theta||^2 \lesssim s(1 + \lambda) + (n - s)\lambda e^{-\lambda/2}.$$ 

For $\lambda = 2 \log(n/s)$ we see that the risk is bounded by a multiple of $s \log(n/s)$, provided $s = o(n)$.

Note that through the tuning parameter the estimator uses the knowledge of the number $s$ of non-zero coefficients, but not their locations. If we know which $s$ of the $\theta_j$’s are nonzero, we can estimate them by $Y_j$, while we estimate all others by $0$. The resulting estimator is unbiased, and has variance $s$. Hence, the risk of that estimator is of the order $s$. The additional factor $\log(n/s)$ is the price that we pay for not knowing the support set of the signal. The rate $s \log(n/s)$ is in fact the best possible rate that can be attained uniformly over the set so-called nearly black vectors, i.e. the set of $s$-sparse vectors with $s = o(n)$, cf. Donoho et al. (1992). 

In Exercise 3.18 you are asked to work out the rate of convergence of the lasso in the normal means model.
3.5 Exercises

Exercise 3.1 (Negative moments of the multivariate Gaussian). Let $Z \sim N_n(0, I)$. Show that $E(1/\|Z\|^p) < \infty$ if and only if $n > p$.

Exercise 3.2 (Proof of the James-Stein lemma). Prove Lemma 3.1.2.

Exercise 3.3 (Shrinking towards another point). Let $Y \sim N_n(\theta, I)$ and $v \in \mathbb{R}^n$. Define the estimator $\tilde{\theta}_{JS} = v + \left(1 - \frac{n - 2}{\|Y - v\|^2}\right)(Y - v)$. Prove that for $n \geq 3$, this estimator also satisfies $E_{\theta}\|\tilde{\theta}_{JS} - \theta\|^2 < E_{\theta}\|\hat{\theta}_{MLE} - \theta\|^2$ for all $\theta \in \mathbb{R}^n$.

Exercise 3.4 (Oracle version of James-Stein). Use the expression (3.2) for the risk of the James-Stein estimator to prove that if $Y \sim N(\theta, \sigma^2 I)$, then for every $\theta \in \mathbb{R}^n$ and $n \geq 3$,

$$E_{\theta}\|\tilde{\theta}_{JS} - \theta\|^2 \leq 4\sigma^2 + \inf_{c \in \mathbb{R}} E_{\theta}\|cY - \theta\|^2.$$ 

This is a so-called oracle inequality that asserts that up to a constant, the risk of the James-Stein estimator is as good as the risk that could be achieved by an oracle that may use its knowledge of the true parameter $\theta$ to choose the degree of shrinking.

Exercise 3.5 (Admissibility of the MLE in the normal mean model for $n = 1$). For $n = 1$ and $n = 2$ the MLE $\hat{Y}$ is admissible in the model $Y \sim N_n(\theta, I)$. This exercise deals with the case $n = 1$. So we assume that $Y \sim N(\theta, 1)$. The goal is to prove that there exists no other estimator $\hat{\theta}$ such that $E_{\theta}(\hat{\theta} - \theta)^2 \leq E_{\theta}(Y - \theta)^2$ for all $\theta \in \mathbb{R}$, with strict inequality for some $\theta \in \mathbb{R}$.

For $\tau > 0$, consider the $N(0, \tau)$ prior on the parameter $\theta$. Denote the corresponding prior density by $\pi_\tau$.

(i) Show that if an estimator $\hat{\theta}$ as described above would exist, then there would exist an $\varepsilon > 0$ and $\theta_0 < \theta_1$ such that

$$1 - \int_{\theta_0}^{\theta_1} E_{\theta}(\hat{\theta} - \theta)^2 \pi_\tau(\theta) d\theta \geq \varepsilon \int_{\theta_0}^{\theta_1} \pi_\tau(\theta) d\theta.$$
(ii) Let \( \hat{\theta}_\tau \) be the posterior mean corresponding to the prior \( \pi_\tau \). Compute the corresponding Bayes risk
\[
\int E_\theta (\hat{\theta}_\tau - \theta)^2 \pi_\tau (\theta) \, d\theta.
\]

(iii) Using the results of (i) and (ii), show that if an estimator \( \hat{\theta} \) as described above would exist, then
\[
\frac{1 - \int E_\theta (\hat{\theta} - \theta)^2 \pi_\tau (\theta) \, d\theta}{1 - \int E_\theta (\hat{\theta}_\tau - \theta)^2 \pi_\tau (\theta) \, d\theta} \to \infty
\]
as \( \tau \to \infty \). Derive a contradiction.

Admissibility of the MLE in the case \( n = 2 \) can also be proved using this approach via the Bayes risk. The analysis is more involved however, since using conjugate Gaussian priors as in the case \( n = 1 \) does not work. See Problem 4.5 on p. 398 of Lehmann and Casella (1998).

Exercise 3.6 (Smoothness and decay of Fourier coefficients). Let \( f \) be a square integrable function on \([0, 1]\) satisfying \( f(0) = f(1) \). Let \( e_k, k \in \mathbb{Z} \), be the standard (complex) Fourier basis of \( L^2[0, 1] \), defined by \( e_k(x) = \exp(i2\pi kx) \). Show that \( f \) has \( \beta \) square integrable, almost everywhere defined derivatives if and only if its Fourier coefficients \( f_k \) satisfy \( \sum k^{2\beta} |f_k|^2 < \infty \).

Exercise 3.7 (Weak duality). In Example 3.2.4, show that the weak duality (3.6) holds.

Exercise 3.8 (MAP estimator versus posterior mean). Show by direct computation that for the prior \( \pi \) considered in the last part of Example 3.2.5, the MAP estimator (3.8) coincides with the posterior mean (3.9).

Exercise 3.9 (Kernel density estimation). Using the R function \texttt{density}, fit a kernel density estimator to the galaxy dataset of Figure 1.2 (which can be found in the package \texttt{MASS}). Plot the result over a histogram of the data. Experiment with different bandwidths and kernels. Do the same with the dataset \texttt{faithful$eruptions}.
Exercise 3.10 (Grenander estimator). The Grenander estimator is implemented for instance in the R function `grenander` in the package `fdrtool`. Try it out on a simulated sample of size 200 from the standard exponential distribution. Compare to kernel density estimates for the same dataset.

Exercise 3.11 (The lasso). Verify the expression for the lasso estimator in the normal means model, given in Example 3.3.4. Plot the soft thresholding function $h_\lambda$.

Exercise 3.12 ($\ell^0$-penalization). Verify the expression for the $\ell^0$-penalized estimator in the normal means model, given in Example 3.3.5. Plot the corresponding hard thresholding function.

Exercise 3.13 (Data reduction as regularization). Consider a general statistical experiment with an observation of the form $(X, Y)$ in some Euclidean space $\mathbb{R}^k \times \mathbb{R}^m$ (for simplicity). Suppose we have a parametrized statistical model $(P_\theta : \theta \in \Theta)$ for some parameter space $\Theta$ and assume that every $P_\theta$ admits a Lebesgue density. Suppose that under each $P_\theta$ both the distribution of $Y$ and the conditional distribution of $X$ given $Y$ have densities on, respectively, $\mathbb{R}^m$ and $\mathbb{R}^k$ as well.

Show that the MLE for $\theta$ based on the reduced data $Y$ can be viewed as a penalized MLE for the full data $(X, Y)$. Show that the penalization is non-trivial if and only if $Y$ is not a sufficient statistic for $\theta$.

Exercise 3.14 (MNIST digit prediction using penalized logistic regression). In Exercise 1.3 we considered the problem of predicting the correct label of images of handwritten 0’s and 1’s from the MNIST dataset. We fitted two types of logistic regression models to the training dataset shown in Figure 1.3. Then we used these models to predict the labels of the remaining 2115 images in the dataset. We compared logistic regression using the whole images and using just the $2 \times 2$ sub-images in the center. It turned out that the latter works better.

By Exercise 3.13 we can view using the sub-images instead of the full images as a form of regularization. We can also combine logistic regression with other penalties, for instance ridge, i.e. $\ell^2$, or lasso-type, i.e. $\ell^1$, penalties. The function `glmnet` in the R package with the same name implements logistic regression with these penalties. Use this to fit an $\ell^1$-penalized logistic regression model to the training data of Figure 1.3, using the default choice of the tuning parameter. Use this model to predict the labels of the remaining 2115 images in the dataset. How does the prediction performance compare to that of the earlier methods?
Exercise 3.15 (Rates for kernel density estimation). Mimic the approach of Example 3.4.3 to derive the best rate of convergence for the kernel density estimator under the condition that the true density belongs to the Hölder space $C^\beta[0,1]$ for $\beta \in (0, 2]$. Do it both for the pointwise risk and for the $L^2$-risk, or mean integrated squared error (MISE)

$$E_p \int_0^1 (\hat{p}_h(x) - p(x))^2 \, dx.$$ 

What is the optimal choice of the bandwidth?

Exercise 3.16 (Rates for Bayesian procedures). Consider the Bayes procedure of Example 3.2.5, with prior variances given by $\tau_j^2 = j^{-1-2\alpha}$ for some $\alpha > 0$. Show that for this choice, the squared $L^2$-risk of the posterior mean (3.9) satisfies, for $\beta, R > 0$,

$$\sup_{f \in \mathcal{H}_R^2[0,1]} E_f \|\hat{f} - f\|_2^2 \lesssim$$

Which choice of $\alpha$ gives the best upper bound?

Exercise 3.17 (Gaussian tail bound). Let $Z$ be a standard normal random variable. Show that for all $p > 0$ there exists a constant $C_p > 0$ such that for all $a \geq 1$,

$$E Z^p 1_{Z > a} \leq C_p a^{p-1} e^{-\frac{1}{2} a^2}.$$ 

Exercise 3.18 (Rate of convergence for the lasso). Investigate which rate the lasso can achieve for $s$-sparse vectors in the normal means model. What is the appropriate oracle choice of the tuning parameter?