Function Estimation and Gaussian Sequence Models

Iain Johnstone

Dept. of Statistics, Stanford University, Stanford CA 94305, USA
E-mail address: imj@stat.stanford.edu
URL: http://www-stat.stanford.edu/~imj
Contents

Not-the-Preface: Remarks & Acknowledgements viii

Chapter 1. Prologue: Wavelets and function estimation 3
  1.1. Introduction 3
  1.2. A simple model for sparsity 4
  1.3. The “Signal in Gaussian white noise” model 13
  1.4. Optimality in the white noise model 15
  1.5. Concluding Remarks 26

Part 1. SEQUENCE MODELS 27

Chapter 2. The Multivariate Normal Distribution 28
  2.1. Linear Regression and sequence models 28
  2.2. Penalized Least Squares and thresholding 30
  2.3. Priors, posteriors and Bayes estimates 30
  2.4. Stein’s Unbiased Risk formula 30
  2.5. The James Stein estimate 30
  2.6. Mean squared error properties 30
  2.7. Differentiation and noise amplification 30
  2.8. Bibliographic Notes 30
  2.9. Problems 30

Chapter 3. The Gaussian Sequence model 32
  3.1. The Basic Model 32
  3.2. Non-white Gaussian sequence models 35
  3.3. Example: Integrated Wiener process priors 36
  3.4. Expected Errors of Estimation 39
  3.5. Example: Linear Estimators with Squared Error Loss 41
  3.6. Models with Gaussian limits* 43
  3.7. Problems 45

Chapter 4. Interlude: Spline Smoothing 47
  4.1. Cubic Splines and Periodic Splines 47
  4.2. The Equivalent Kernel for Spline smoothing 50
  4.3. The Variance-Bias Lemma 52
  4.4. Spline Estimates over Sobolev Ellipsoids 53
  4.5. Integrated Wiener Priors and Spline Smoothing 56
  4.6. Problems 61
Chapter 5. Some Linear Inverse Problems

5.1. Some examples

5.2. Singular value decomposition

5.3. Regularization schemes

5.4. Wahba’s generalized spline model

5.5. Prediction error and Estimation error

Part 2. SOME GAUSSIAN DECISION THEORY

Chapter 6. Bayes Risks and Minimaxity

6.1. Bayes Estimators

6.2. The Minimax Theorem

6.3. Minimaxity on Products.

6.4. The Bayes Minimax Method*

6.5. Appendix: Further details.

Exercises

Chapter 7. Intervals, Hyperrectangles and Quadratic Convexity

7.1. Single Bounded Normal Mean

7.2. Hyperrectangles

7.3. Orthosymmetry and Hardest rectangular subproblems

7.4. Appendix: Further details.

Exercises

Part 3. LINEAR ESTIMATION

Chapter 8. Pinsker’s Theorem and Asymptotics

8.1. Exact evaluation of linear minimax risk.

8.2. Pinsker’s Theorem

8.3. Interlude: Compactness and Consistency

Chapter 9. James-Stein Estimation and Adaptive Minimaxity over Ellipsoids

9.1. Stein’s Unbiased Risk Formula

9.2. The James Stein Estimate

9.3. Adaptive minimaxity

9.4. Interlude: Superefficiency

9.5. Discussion

Exercises

Part 4. Interlude:

Chapter 10. A Primer on Estimation by Wavelet Shrinkage

10.1. The Cascade algorithm for the Discrete Wavelet Transform

10.2. Wavelet shrinkage estimation
10.3. Choice of threshold. 138
10.4. Concluding remarks 141

Part 5. MONORESOLUTION MODELS 145

Chapter 11. Thresholding and Oracle inequalities 146
11.1. A crude MSE bound for hard thresholding. 146
11.2. Properties of Thresholding Estimators 148
11.3. Oracle Inequalities 153
11.4. Optimality of risk bounds. 156
11.5. Appendix: Further details 158
Exercises 159

Chapter 12. The optimal recovery approach to thresholding. 161
12.1. An Optimal Recovery Model 161
12.2. Evaluation of Moduli of Continuity 163
12.3. The Statistical Sequence Model 164
12.4. Further Details 167

Chapter 13. Model Selection, Penalization and Oracle Inequalities 169
13.1. A Gaussian concentration inequality 169
13.2. All subsets regression and complexity penalized least squares 170
13.3. Oracle inequalities for $2k \log(p/k)$ penalties 173
13.4. Proof of Oracle inequality 175
13.5. Aside: Stepwise methods vs. complexity penalization. 179
13.6. A version for orthogonal regression 180
13.7. False discovery rate estimation 183

Chapter 14. [Unrevised] Minimax estimation on $\ell_p$ balls 186
14.1. Linear Estimators. 191
14.2. Univariate Bayes Minimax Problem 193
14.3. Minimax Bayes Risk for $n$-dimensional data. 200
14.4. Minimax Risk in the Highly Sparse Case 204
14.5. Appendix: Further details 207

Part 6. MULTIRESOLUTION MODELS 209

Chapter 15. Sparsity, Weak $\ell_p$ and Thresholding 210
15.1. Approximation, Ideal Risk and Weak $\ell_p$ 210
15.2. A Risk Lower Bound via Embedding of hypercubes. 213
15.3. Quasi-norm equivalences 214
15.4. Besov Bodies and weak $\ell_p$ 215
15.5. Adaptive minimaxity for $\sqrt{2\log n}$ thresholding 216

Chapter 16. [To revise] Optimal Recovery and Wavelet Thresholding 220
16.1. Unconditional Bases 221
16.2. Optimal Recovery Model - Finite Data 222
Not-the-Preface: Remarks & Acknowledgements

These notes (version of January 2001) are a still rough and partial record of some series of lectures:

(1) at the École Normale Supérieure in Paris, February through May, 1998.

(2) For a course, Statistics 322 at Stanford, in Winter 1999 and Fall 2000.

Comments are welcome, as some further revision is likely, possibly in the direction of a set of lecture notes, or monograph, for publication. Corrections, bibliographic references and oversights are particularly sought!

I am very grateful to Lucien Birgé, Olivier Catoni and Pascal Massart for the invitation to give the lectures in Paris, and for their scientific leadership of the seminar “Statistique à L’École Normale”. I am also very grateful for the scientific and personal welcome extended by them and by Gérard Kerkyacharian, Dominique Picard and Alexander Tsybakov.

I’d like to thank my collaborators: particularly Dave Donoho, with whom much of the wavelets-in-statistics work began, and also Gérard Kerkyacharian, Dominique Picard and Bernard Silverman.

This project has a prehistory which I gratefully acknowledge. In March 1995, Dave Donoho and I gave a ten lecture DMV seminar on “wavelets in statistics” at Oberwolfach. In June 1996, at the kind invitation of Piet Groeneboom, I gave a related week long course at Kasteel de Berkt in the Netherlands. Both these presentations were based on overhead transparencies, and the transition to blackboard lecture courses led to considerable reorganization.

Thanks also to the John Simon Guggenheim Memorial Foundation for a Fellowship during which the first draft was written, and to the National Science Foundation, which has supported much of my own research and writing. And to the Australian National University which provided space and time for writing.

For encouragement and/or thoughtful comments on the manuscript, I’m greatly indebted to Félix Abramovich, Peter Bickel, Larry Brown, Emmanuel Candès, Shingchang Kou, Yi Lin, Arthur Lu, Michael Nussbaum,
and Stéphane Mallat, as well as to students in Stat322 – Ery Arias Castro, Arnab Chakraborty, Jiashun Jin, Charles Mathis, Debashis Paul, Hualin Wang)
Introduction

This is a book about some of the theory of nonparametric function estimation. The premise is that much insight can be gained even if attention is confined to a Gaussian sequence model

\[(0.1) \quad y_i = \theta_i + \epsilon z_i, \quad i \in I,\]

where \(I\) is finite or countable, \(\{\theta_i\}\) is fixed and unknown, \(\{z_i\}\) are i.i.d. \(N(0,1)\) noise variables and \(\epsilon\) is a known noise level. If \(I\) is finite, this is an old friend, the multivariate normal means model, with independent coordinates and known variance. It is the centerpiece of parametric statistics, with many important, beautiful, and even surprising results whose influence extends well beyond the formal model into the practical, approximate world of data analysis.

It is perhaps not so obvious that the infinite sequence model could play a corresponding role in nonparametric statistics. For example, typical problems of nonparametric regression, density estimation and classification are typically formulated in terms of unknown functions, rather than sequences of parameters. Secondly the additive white Gaussian noise assumption may seem rather remote.

There are various responses to this objection. First, the model captures many of the conceptual issues associated with non-parametric estimation, with a minimum of technical complication. For example, non-parametrics must grapple with the apparent impossibility of trying to estimate an infinite-dimensional object – a function – on the basis of a finite amount \(n\) of noisy data. With a calibration \(\epsilon = 1/\sqrt{n}\), this challenge is plain to see in model (0.1). Of course, the broad strategy is to apply various methods that one understands in the multivariate normal model to finite submodels, and to argue that not too much is lost by ignoring the (many!) remaining parameters.

Second, models and theory are always an idealisation of practical reality. Advances in size of datasets and computing power have enormously increased the complexity of both what we attempt to do in data analysis and the algorithms that we invent to carry out our goals. If the goal of theory is to provide clearly formulated, generalizable insights that might inform and improve our computational efforts, then we may need to accept a greater degree of idealization in our models than was necessary when developing theory for the estimation of one, two or three parameters.

Thirdly, it turns out that model (0.1) is often a reasonable approximation in large samples. In parametric statistics, the central limit theorem and asymptotic normality of estimators extends the influence of multivariate normal results to generalized linear models and beyond. In nonparametric estimation, it has long been observed that similar results are often found in spectrum, density and regression estimation. Relatively recently, results
have appeared connecting these problems to model (0.1) and thereby providing some formal support for these observations.

Model (0.1) and its justifications have been used and understood for decades by Russian theoretical statisticians, led by I. A. Ibragimov and R. Z. Khasminskii. It is perhaps surprising that only relatively recently has the model received some acceptance in the West. Indeed it is still derided as irrelevant by many of a sternly practical bent.

Chapter 1 concentrates on an extended example – the interaction between wavelets and the theory of function estimation. [This is a modified and expanded version of Johnstone (1999).] The story, to be developed in more detail throughout the book, bolsters the case for the sequence model by showing how ideas from harmonic analysis and approximation theory lead to quite practical insights in signal processing and statistics.
CHAPTER 1

Prologue: Wavelets and function estimation

This chapter, which is much the same as Johnstone (1999), is an overview of many of the themes that will be developed in detail in this manuscript.

Nonparametric function estimation aims to estimate or recover or de-noise a function of interest, perhaps a signal, spectrum or image, that is observed in noise and possibly indirectly after some transformation, as in deconvolution. “Nonparametric” signifies that no a priori limit is placed on the number of unknown parameters used to model the signal. Such theories of estimation are necessarily quite different from traditional statistical models with a small number of parameters specified in advance.

Before wavelets, the theory was dominated by linear estimators and the exploitation of assumed smoothness in the unknown function to describe optimal methods. Wavelets provide a set of tools that make it natural to assert, in plausible theoretical models, that the sparsity of representation is a more basic notion than smoothness, and that non-linear thresholding can be a powerful competitor to traditional linear methods. We survey some of this story, showing how sparsity emerges from an optimality analysis via the game-theoretic notion of a least favorable distribution.

1.1. Introduction

Within statistics, the first applications of wavelets were to theory. While the potential for statistical application to a variety of problems was also apparent, and is now being realized (as surveyed in other articles in this volume), it was in the theory of nonparametric function estimation that progress was initially fastest. The goal of this article is to survey some of this work in retrospective.

Why should developments in the theory of statistics interest a wider scientific community? Primarily, perhaps, because theory attempts to isolate concepts of broad generality that clarify in what circumstances and under what assumptions particular data analytic methods can be expected to perform well, or not. As a classical example, the most widely used statistical tools - regression, hypothesis tests, confidence intervals - are typically associated with parametric models, that is, probability models for observed data that depend on at most a (small) finite number of unknown parameters. The true scope, versatility and applicability of these tools was clarified
by the development of underlying theoretical notions such as likelihood, sufficiency, unbiasedness, Cramér-Rao bounds, power, and so forth. Many of these concepts have passed into the general toolkit of scientific data analysis.

A further key point is that theory promotes portability of methods between scientific domains: thus, Fisher's analysis of variance was initially developed for agricultural field trials, but he also created theoretical support with such effect that most uses of analysis of variance now have nothing to do with agriculture.

What is meant by the term nonparametric function estimation? The advent of larger, and often instrumentally acquired, datasets and a desire for more flexible models, has stimulated the study of nonparametric models, in which there is no a priori bound on the number of parameters used to describe the observed data. For example, when fitting a curve to time-varying data, instead of an a priori restriction to, say, a cubic polynomial, one might allow polynomials of arbitrarily high degree or, more stably, a linear combination of splines of local support.

Despite prolific theoretical study of such infinite-dimensional models in recent decades, the conclusions have not dispersed as widely as those for the parametric theory. While this is partly because nonparametric theory is more recent, it is certainly also partly due to the greater nuance and complexity of its results, and a relative paucity of unifying principles.

The arrival of wavelet bases has improved the situation. Wavelets and related notions have highlighted sparsity of representation as an important principle in estimation and testing. Through the dyadic Gaussian sequence model, they have bridged parametric and nonparametric statistics and reinvigorated the study of estimation for multivariate Gaussian distributions of finite (but large) dimension. Wavelets have also been the vehicle for an influx of ideas – unconditional bases, fast algorithms, new function spaces – from computational harmonic analysis into statistics, a trend that seems likely to continue to grow in future.

1.2. A simple model for sparsity

We begin with an apparently naive discussion of sparsity in a “mono-resolution” model. This name is chosen to evoke what is going on in a single level of the discrete orthogonal wavelet transform. However, it is nothing other than the multivariate normal mean model with unknown location.

Suppose, then, that we observe an $n$–dimensional data vector $y$ consisting of an unknown signal $\theta$, which we wish to estimate, contaminated by additive Gaussian white noise of scale $\epsilon_n$. If the model is represented in terms of its coefficients in a particular orthonormal basis $B$, we obtain $(y_k^B), (\theta_k^B)$ etc., though the dependence on $B$ will usually be suppressed. Thus, in terms of basis coefficients,

$$ y_k = \theta_k + \epsilon_n z_k \quad k = 1, \ldots, n $$
and \( \{z_k\} \) are independently and identically distributed \( N(0, 1) \) random variables. Here, we emphasize that \( \theta = (\theta_k) \) is in general regarded as fixed and unknown. This model might be reasonable, for example, if we were viewing data as Fourier coefficients, and looking in a particular frequency band where the signal and noise spectrum are each about constant.

If, in addition, it is assumed that \( \{\theta_k\} \) are random, being drawn from a Gaussian distribution with \( \text{Var}(\theta_k) = \tau_n^2 \), then the linear Bayes estimator, also known as the Wiener filter, would involve \textit{linear shrinkage} by a constant factor:

\[
\hat{\theta}_k = \frac{\rho}{\rho + 1} y_k, \quad \rho = \frac{\tau_n^2}{\epsilon_n^2}.
\]

The ratio \( \tau_n^2 / \epsilon_n^2 \) (or some function of it) is usually called the \textit{signal-to-noise} ratio. Figure 1 is a simple illustration.

![Figure 1](image)

\textbf{Figure 1.} Left panel: Simulated realization from model (1.1): observed data are the dark points, being signal (pale points) with independent Gaussian errors added. Each point is independent of the next. Right panel: “red” points show the Wiener filter, or Bayes linear estimate - a clear improvement over the raw data as an estimate of the signal.

This traditional analysis has two key features. First, the assumption of a Gaussian prior distribution produces an optimal estimator which is linear. Second, the estimator does not depend on the choice of basis: both the model (1.1) and the Gaussian prior are invariant under orthogonal changes of basis, and so the optimal rule has the same linear shrinkage in all coordinate systems.

\textbf{Sparsity.} In contrast, sparsity has everything to do with the choice of bases. Informally, “sparsity” conveys the idea that most of the signal strength is concentrated in a few of the coefficients. Thus a ‘spike’ signal \( \gamma(1, 0, \ldots, 0) \) is much sparser than a ‘comb’ vector \( \gamma(n^{-1/2}, \ldots, n^{-1/2}) \) even though both have the same energy, or \( \ell_2 \) norm: indeed these could be representations of the same vector in two different bases. In contrast, noise, almost by definition, is not sparse in any basis. Thus, among representations
of signals in various bases, it is the ones that are sparse that will be most easily “denoised”.

**Remark.** Of course, in general terms, sparsity is a familiar notion in statistics and beyond: think of parsimonious model choice, “Occam’s razor” and so forth. It is the motivation for the principal components analysis of Hotelling (1933), suitable for high dimensional, approximately Gaussian data. However, in the specific domain of nonparametric function estimation, prior to the advent of wavelets, the role of sparsity was perhaps somewhat obscured by the focus on the related, although somewhat more special, notion of smoothness.

Figure 2 shows part of an actual signal represented in two different bases: panel a) is a subset of $2^7$ wavelet coefficients $\theta^W_k$, while panel b) is a subset of $2^7$ Fourier coefficients $\theta^F_k$. Evidently $\theta^W$ has a much sparser representation than does $\theta^F$.

![Figure 2](image-url)

**Figure 2.** Panel (a): $\theta^W_k$ =level 7 of estimated NMR reconstruction $g$ of Figure 10, while in panel (b): $\theta^F_k$ = Fourier coefficients of $g$ at frequencies 65...128, both real and imaginary parts shown. While these do not represent exactly the same projections of $f$, the two overlap and $\|\theta^F\|_2 = 25.3 \approx 23.1 = \|\theta^W\|_2$.

The sparsity of the coefficients in a given basis may be quantified using $\ell_p$ norms

$$\|\theta\|_p = \left(\sum_{k=1}^{n} |\theta_k|^p\right)^{1/p},$$

which track sparsity for $p < 2$, with smaller $p$ giving more stringent measures. Thus, while the $\ell_2$ norms of our two representations are roughly equal:

$$\|\theta^F\|_2 = 25.3 \approx 23.1 = \|\theta^W\|_2,$$

the $\ell_1$ norms differ by a factor of 6.5:

$$\|\theta^F\|_1 = 246.5 \gg 37.9 = \|\theta^W\|_1.$$
Figure 3. Contours of $\ell_p$ balls

Figure 3 shows that the sets

$$\{ \theta : \sum_{1}^{n} |\theta_k|^p \leq C^p \}$$

become progressively smaller and clustered around the co-ordinate axes as $p$ decreases. Thus, the only way for a signal in an $\ell_p$ ball to have large energy (i.e. $\ell_2$ norm) is for it to consist of a few large components, as opposed to many small components of roughly equal magnitude. Put another way, among all signals with a given energy, the sparse ones are precisely those with small $\ell_p$ norm.

Thus, we will use sets $\{ \| \theta \|_p \leq C \}$ as models for a priori constraints that the signal $\theta$ has a sparse representation in the given basis. Assume, for simplicity here, that $\epsilon_n = 1/\sqrt{n}$ and that $p = C = 1$: it is thus supposed that $\sum_{1}^{n} |\theta_k| \leq 1$. Other situations can be handled by developing the theory for general $(p, C, \epsilon_n)$, see Chapter 14 and Donoho & Johnstone (1994b). How to exploit this sparsity information in order better to estimate $\theta$; in other words, can we estimate $\theta^W$ better than $\theta^F$? We quantify the quality of estimation using $\hat{\theta}(y)$ using Mean Squared Error (MSE):

$$E\| \hat{\theta} - \theta \|^2 = \sum_{i=1}^{n} E(\hat{\theta}_i - \theta_i)^2.$$  

(1.3)

Figure 4 shows an idealized case in which all $\theta_k$ are zero except for two spikes, each of size $1/2$. Two extreme examples of linear estimators are $\hat{\theta}_1(y) \equiv y$, which leaves the data unadjusted, and $\hat{\theta}_2(y) \equiv 0$, which sets every coordinate to zero. The first, an estimator with no bias and only variance, has $MSE = E(y_i - \hat{\theta}_1)^2 = \epsilon_n^2 = 1/n$ in each of the $n$ coordinates, for a total $MSE = 1$. The second, $\hat{\theta}_0$, an estimator with no variance and only bias, is exactly correct on all but the two spikes, where it suffers a total $MSE = 2 \times (1/2)^2 = 1/2$. Given the symmetry of the prior knowledge and the statistical independence of the observations, the only other plausible choices for a linear estimator have the form $cy$, for a constant $c$, $0 \leq c \leq 1$. It can be shown (Chapter 14) that such estimators are effectively a combination
of the two extremes, and in particular do not have noticeably better MSE performance. In a like vein, panels (b) and (c) use vertical lines to indicate the absolute, or $\ell_1$ error $\sum |\hat{\theta}_i - \hat{\theta}_i|$, which is likewise relatively large for the linear estimates.

In the situation of Figure 4, thresholding is natural. Define the hard threshold estimator by its action on coordinates:

$$
\hat{\theta}_{0,k}(y) = \begin{cases} 
y_k & \text{if } |y_k| \geq \lambda, \\
0 & \text{otherwise}. 
\end{cases}
$$

and Figure 4(a) shows a threshold of $\lambda = 2.4 \epsilon_n = 0.3$. For the particular configuration of true means $\theta_k$ shown there, the data from the two spikes pass the threshold unchanged, and so are essentially unbiased estimators. Meanwhile, in all other coordinates, the threshold correctly sets all data to zero except for the small fraction of noise that exceeds the threshold. Thus, it can be directly verified that

$$
MSE(\hat{\theta}_n, \theta) \approx 2\epsilon_n^2 + n\epsilon_n^2 E\{Z^2, Z^2 > \log n\} \approx 2n^{-1} + 2\sqrt{\log n} n,
$$

where $Z$ is a standard Gaussian variate. This mean squared error is of course much better than for any of the linear estimators.
Statistical Games and the Minimax theorem. The skeptic will object that the configuration of Figure 4 was chosen to highlight the advantages of thresholding, and indeed it was! It is precisely to avoid the possibility of being misled by such reasoning from constructed cases that the tools of game theory have been adapted for use in statistics. A sterner and fairer test of an estimator is obtained by creating a statistical two person zero sum game, or statistical decision problem. In our setting, this has the following rules:

(i) Player I (“the Statistician”) is allowed to choose any estimator \( \hat{\theta}(y) \), linear, threshold or of more complicated type.

(ii) Player II (“Nature”) may choose \( \theta \in \mathbb{R}^n \) at random and may choose a probability distribution \( \pi \) for \( \theta \) subject only to the sparsity constraint that \( E_\pi \| \theta \|_1 \leq 1 \).

(iii) The payoff is calculated as the expected mean squared error of \( \hat{\theta}(y) \) when \( \theta \) is chosen according to \( \pi \) and then the data \( y \) is drawn from model (1.1): \( y = \theta + \epsilon_n z \) for \( z \sim N_n(0, I) \). Thus the payoff now averages over both \( \theta \) and \( y \):

\[
B(\hat{\theta}, \pi) = E_\pi E_{\theta|y}(\| \hat{\theta}(y) - \theta \|_2^2).
\]

Of course, the Statistician tries to minimize the payoff and Nature to maximize it.

Classical work in statistical decision theory (Wald 1950, Le Cam 1986) shows that the minimax theorem of von Neumann can be adapted to apply here, and that the game has a well defined value, the minimax risk:

\[
(1.5) \quad R_n = \inf_{\hat{\theta}} \sup_{\pi} B(\hat{\theta}, \pi) = \sup_{\pi} \inf_{\hat{\theta}} B(\hat{\theta}, \pi).
\]

An estimator \( \hat{\theta}^* \) attaining the left hand infimum in (1.5) is called a minimax strategy or estimator for player I, while a prior distribution \( \pi^* \) attaining the right hand supremum is called least favorable and is an optimal strategy for player II. Schematically, the pair of optimal strategies \( (\hat{\theta}^*, \pi^*) \) forms a saddlepoint, Figure 5: if Nature uses \( \pi^* \), the best the Statistician can do is to use \( \hat{\theta}^* \). Conversely, if the Statistician uses \( \hat{\theta}^* \), the optimal strategy for Nature is to choose \( \pi^* \).

It is the structure of these optimal strategies, and their effect on the minimax risk \( R_n \) that is of chief statistical interest.

While these optimal strategies cannot be exactly evaluated for finite \( n \), informative asymptotic approximations are available (Donoho & Johnstone 1994b), with the consequence that under our unit norm sparsity constraint,

\[
R_n \sim \sqrt{\log n / n}
\]

as \( n \to \infty \). Indeed, an approximately least favorable distribution is given by drawing the individual coordinates \( \theta_k \) independently from a two point
Figure 5. Left to right axis: strategies $\pi$ for Nature. Front to back axis: strategies $\hat{\theta}$ for the Statistician. Vertical axis: payoff $B(\hat{\theta}, \pi)$ from the Statistician to Nature. The saddle-point indicates a pair $(\hat{\theta}^*, \pi^*)$ of optimal strategies.

distribution with

\[
\theta_k = \begin{cases} 
\epsilon_n \sqrt{\log n} & \text{with probability } \epsilon_n = 1/\sqrt{n \log n} \\
0 & \text{otherwise.}
\end{cases}
\]

This amounts to repeated tossing of a coin highly biased towards zero. Thus, in $n$ draws, we expect to see a relatively small number, namely \( n \epsilon_n = \sqrt{n/\log n} \) of non-zero components. The size of these non-zero values is such that they are hard to distinguish from the larger values among the more numerous remaining $n - \sqrt{n/\log n}$ observations that are pure noise. Of course, what makes this distribution difficult for Player I is that the locations of the non-zero components are random as well.

It can also be shown (Chapter 14) that an approximately minimax estimator for this setting is given by the hard thresholding rule described earlier, but with threshold given (at least approximately) by $\lambda_n = \sqrt{\log(n \log n)/n}$. This estimate asymptotically achieves the minimax value $R_n \sim \sqrt{\log n/n}$ for MSE. It can also be verified that no linear estimator can achieve a payoff of better than $1/2$ if Nature chooses a suitably uncooperative probability distribution for $\theta$.

It is perhaps the qualitative features of this solution that most deserve comment. Had we worked with simply a signal to noise constraint, $E_{\theta} \| \theta \|_2 \leq 1$, say, we would have obtained a Gaussian prior distribution as being approximately least favorable and the linear Wiener filter (1.2) with $\epsilon_n^2 = \tau_n^2 = 1/n$ as an approximately minimax estimator. The imposition of a sparsity constraint $E_{\theta} \| \theta \|_1 \leq 1$ yields great improvements in the quality of possible estimation, and produces optimal strategies that take us far away from Gaussian priors and linear methods.
<table>
<thead>
<tr>
<th>Prior Constraint</th>
<th>traditional ($\ell_2$)</th>
<th>sparsity ($\ell_1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimax estimator</td>
<td>linear</td>
<td>thresholding</td>
</tr>
<tr>
<td>least favorable $\pi$</td>
<td>Gaussian</td>
<td>sparse</td>
</tr>
<tr>
<td>minimax M.S.E.</td>
<td>$= 1/2$</td>
<td>$\sim \sqrt{\frac{\log n}{n}}$</td>
</tr>
</tbody>
</table>

Table 1. Comparison of structure of optimal strategies in the monoresolution game under traditional and sparsity assumptions.

Sparsity and Improved MSE. There is an alternative way to suggest how sparsity of representation affects the mean squared error of estimation using thresholding. Return to model (1.1): as we just saw, the MSE of $\hat{\theta}_1(y_i) = y_i$ for estimating $\theta_1$ is $c_n^2$, while the MSE of $\hat{\theta}_0(y_i) = 0$ is $\theta_1^2$. Given a choice, an omniscient “oracle” would choose the estimator that yields the smaller of the two MSE’s. Repeating this for each coordinate leads to a notion of ideal risk:

$$R(\theta, \varepsilon) = \sum_i \min(\theta_i^2, \varepsilon^2).$$

(1.7)

Suppose that the coefficients are rearranged in decreasing order:

$$\theta_1^2 \geq \theta_2^2 \geq \cdots \geq \theta_n^2.$$

The notion of “compressibility” captures the idea that the number of large coefficients, $N_r(\theta) = \#\{\theta_i : |\theta_i| \geq \varepsilon\}$ is small, and also that there is little energy in the remaining tail sums

$$c_k^2 = \sum_{i > k} \theta_i^2.$$

Then (1.7) takes the form

$$R(\theta, \varepsilon) = N_r(\theta) \varepsilon^2 + c_k^2(\theta),$$

which says that good compressibility is actually equivalent to small ideal risk (compare Figure 6).

Ideal risk cannot be attained by any estimator, which as a function of $y$ alone lacks access to the oracle. However, thresholding comes relatively close to mimicking ideal risk. To state a simple result, we turn here to soft thresholding, so-called because it is continuous in the data:

$$\hat{\theta}_{\lambda,k}^S(y) = \begin{cases} 
  y_k - \lambda & y_k \geq \lambda \\
  0 & |y_k| < \lambda \\
  y_k + \lambda & y_k < -\lambda .
\end{cases}$$

(1.8)
Figure 6. The decreasing rearrangement of the squared signal components $\theta^2$ is shown. A sparsely represented signal will have a few large coordinates, $N_s$ - which contribute $\epsilon^2$ to the ideal risk - and many small ones, contributing a total of $\epsilon^2 N_s$ to the ideal risk.

For soft thresholding at $\lambda_n = \epsilon \sqrt{2 \log n}$, Chapter 11 and Donoho & Johnstone (1994a) show for all $n$, and for all $\theta \in \mathbb{R}^n$, that

$$E\|\theta_{ST} - \theta\|^2 \leq (2 \log n + 1)\epsilon^2 + \mathcal{R}(\theta, \epsilon).$$

Thus, sparsity implies good compressibility, which in turn implies the possibility of good estimation, in the sense of relatively small MSE.

Remarks. 1. The scope of model (1.1) is broader than may first appear. Suppose that the observed data satisfies

\begin{equation}
(1.9) \quad Y = A\theta + \epsilon
\end{equation}

where $Y$ is an $N_n \times 1$ data vector and $A$ is an $N_n \times n$ orthogonal design matrix ($A' A = mI_n$) and $\epsilon$ has independent Gaussian components. Model (1.1) is recovered by premultiplying (1.9) by $m^{-1} A'$. Thus, $A$ might be a change of basis, and so our analysis covers situations where there is some known basis in which the signal is thought to be sparse. Indeed, this is how sparsity in wavelet bases is employed, with $A$ being (part of the inverse of) a wavelet transform.

2. The sparsity analysis, motivated by wavelet methods below, considers a sequence of models of increasing dimensionality - indeed, the index $n$ is precisely the number of variables. This is in sharp contrast with traditional parametric statistical theory, in which the number of unknown parameters is held fixed as the sample size increases. In practice, however, larger quantities of data $N_n$ typically permit or even require the estimation of richer models with more parameters. Model (1.9) allows $N_n$ to grow with $n$. Thus wavelet considerations promote a style of asymptotics whose importance has long been recognized. (Huber 1981, Sec 7.4).

3. A common criticism of the use of minimax analyses in statistics holds that it is unreasonable to cast “Nature” as a malicious opponent, and that to do so risks throwing up as “worst cases” parameter or prior configurations that are irrelevant to normal use. This challenge would be most pertinent if one were to propose an estimator on the basis of a single decision problem.
Our perspective is different: we analyse families of statistical games, hoping to discover the common structure of optimal strategies - both estimators and least favorable distributions. If an estimator class, such as thresholding, emerges from many such analyses, then it has a certain robustness of validity that a single minimax analysis lacks. This point of view is explored further, with a different example, in Chapter 9.

1.3. The “Signal in Gaussian white noise” model

The multivariate Gaussian distribution \( N_p(\theta, \epsilon^2 I) \) with mean \( \theta \) and \( p \) independent coordinates of standard deviation \( \epsilon \) is the central model of parametric statistical inference, arising as the large sample limit of other \( p \)-parameter models, as well as in its own right.

The natural extension of this multivariate mean model to a countable index set \( I \) yields the Gaussian sequence model

\[
y_I = \theta_I + \epsilon z_I \quad I \in I,
\]

in which \( \theta = (\theta_I) \in \ell_2 \) is assumed to be unknown, but square summable, and \( z = (z_I) \) consists of independent and identically distributed (i.i.d.) standard Gaussian variates.

The most direct connection to nonparametric statistics comes via the “signal in Gaussian white noise” model, which turns out to be equivalent to (1.10). In this model, the observation process \( \{Y(s), 0 \leq s \leq 1\} \) is assumed to satisfy

\[
Y(t) = \int_0^t f(s) ds + \epsilon W(t),
\]

where \( f \) is a square integrable function on \([0, 1]\) and \( W \) is a standard Brownian, or Wiener process starting at \( W(0) = 0 \). In infinitesimal form, this becomes

\[
dY(t) = f(t) dt + dW(t),
\]

suggesting that the observations are built up from data on \( f(t) \) corrupted by independent white noise increments \( dW(t) \). The unknown parameter is now \( f \), and it is desired to estimate or test \( f \) or various functionals of \( f \), such as point values or integrals, on the basis of \( Y \).

As leaders of the Russian school of nonparametric function estimation, Ibragimov & Khas'minskii (1971) gave a central place to model (1.11) in their book, arguing that all of its challenges are conceptual and not merely technical. As in the finite dimensional case, (1.11) arises as an appropriate large sample or low noise limit of certain other nonparametric models, such as probability density estimation, regression and spectrum estimation – see, for example, Brown & Low (1996) and Nussbaum (1996). It extends to images or other objects if one replaces \( t \in [0, 1] \) by a multiparameter index \( t \in D \subseteq \mathbb{R}^d \) and \( W \) by a Brownian sheet.
Let \( \{\psi_I, I \in I\} \) be an orthonormal basis for \( L_2([0,1]) \), the square integrable functions on the unit interval. The equivalence of (1.10) and (1.11) is seen by forming the coefficients of (1.11) in this basis: let
\[
y_I = \int \psi_I dY, 
\]
and similarly
\[
\theta_I = \int \psi_I \quad \text{and} \quad z_I = \int \psi_I dW, 
\]
the latter being a Wiener-Itô stochastic integral. By the elementary properties of stochastic integrals, \( z_I \) are independent and identically distributed (i.i.d.) standard Gaussian variates, and we obtain (1.11).

**Example.** Figure 7 shows a sample nuclear magnetic resonance spectrum (NMR). It is a discrete data set, observed at \( n = 1024 \) points. Nevertheless, one might think of its cumulative sums as a potential example of model (1.11): thus \( Y(i/n) \) would correspond to the sum of the first \( i \) observations in panel (a) starting from the left. Panels (b) and (c) present the data in terms of its coefficients in Fourier and (a particular) orthonormal wavelet basis respectively. Of course, there are only a finite number (1024) of empirical coefficients, but we contend that it is helpful to think of this as an approximation to, or more precisely an “initial segment” of the sequence model (1.10).

![NMR Spectrum and Fourier Coefficients](image)

**Figure 7.** (a) Sample NMR spectrum provided by A. Maudsley and C. Raphael, \( n = 1024 \). (b) Fourier coefficients of the sample noisy spectrum shown with frequency increasing from left to right. (c) Wavelet coefficients (using the Daubechies “Symmlet” of order 6) organized by location \( k \) within resolution level \( j \).
While (1.10) looks like a straightforward infinite dimensional extension of the Euclidean $N_p(\theta, \epsilon^2 I)$ model, there are significant difficulties. For example, the sequence $\{y_t\}$ is with probability one not square summable, because the noise $\{z_t\}$ is i.i.d. Similarly, there is no probability distribution supported on square summable sequences that is invariant under all orthogonal transformations, or even simply under permutation of the coordinates. If the index set $\mathcal{I}$ is linearly ordered, as for the Fourier basis, one must typically work with sequences of weights, often polynomially decreasing, which lack simiplifying invariance properties.

The multiresolution character of wavelet bases is helpful here. If $\{\psi_I\}$ is now an orthonormal wavelet basis for $L^2[0,1]$, such as those of Cohen et al. (1993a), then the index $I = (j, k)$ becomes bivariate, corresponding to level $j$ and location $k2^{-j}$ within each level. The index set $\mathcal{I}$ becomes $\bigcup_{j \geq 0} \mathcal{I}_j \cup \mathcal{I}_{-1}$ with $|\mathcal{I}_j| = 2^j$ counting the possible values of $k$, and $\mathcal{I}_{-1}$ an exceptional set for the scaling function $\phi_0$.

Collect the data coefficients $y_{jk}$ in (1.10) observed at level $j$ into a vector $y_j$; this has a finite dimensional $N_{2^j}(\theta_j, \epsilon^2 I)$ distribution. For many theoretical and practical purposes, it is effective to work with each of these levelwise distributions separately. Since they are of finite (but growing!) dimension, it is possible, and often a scientifically reasonable simplification, to give $\theta_j$ an orthogonally or permutation invariant probability distribution. Indeed, the sparsity results of Section 1.2, derived for Gaussian distributions of large, finite dimension, had precisely this permutation invariance character, and can be applied to each level in the dyadic sequence model.

The full nonparametric estimation conclusions are obtained by combining results across resolution levels. However, it often turns out, especially for minimax analyses, that for a given noise level $\epsilon$, the “least favorable” behaviour occurs at a single resolution level $j = j(\epsilon)$, so that conclusions from the $j(\epsilon)$th permutation invariant Gaussian model provide the key to the non-parametric situation. As the noise level $\epsilon$ decreases, the critical level $j(\epsilon)$ increases, but in a controllable fashion. Thus wavelet-inspired dyadic sequence models allow comparatively simple finite dimensional Gaussian calculations to reveal the essence of non-parametric estimation theory.

In this sense wavelet bases have rehabilitated the finite dimensional multivariate Gaussian distribution as a tool for nonparametric theory, establishing in the process a bridge between parametric and nonparametric models.

### 1.4. Optimality in the white noise model

Our discussion of optimality in the white noise model illustrates the truisim that the available tools, conceptual and mathematical, influence the theory that can be created at a given time. Prior to the advent of wavelet bases, formulations emphasising good properties of linear estimators were the norm; subsequently theoretical conclusions became possible that were more in accord with recent practical experience with algorithms and data.
As a framework for comparing estimators, we again use an appropriate notion of mean squared error. In parallel with the earlier discussion, after (1.1) and at (1.3), we initially regard the unknown function, and its basis coefficients \( \theta(f) \), as fixed. Thus with observations from the white noise model (1.11) (or equivalently, the sequence form (1.10)), an arbitrary estimator \( \hat{\theta} \leftrightarrow \hat{f}(t) \) is evaluated by

\[
r(\hat{\theta}, \theta; \epsilon) = E \sum_I (\hat{\theta}_I - \theta_I)^2 = E \int_0^1 (\hat{f} - f)^2.
\]

If \( f \) is assumed to belong to a class of functions \( \mathcal{F} \subset L_2[0,1] \), corresponding to coefficient set \( \Theta = \Theta(\mathcal{F}) \), then the minimax mean squared error is defined as

\[
R(\Theta; \epsilon) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} r(\hat{\theta}, \theta; \epsilon).
\]

It is necessary to restrict \( \mathcal{F} \) to be a compact subset of \( L_2[0,1] \), for otherwise the minimax risk does not even decrease to zero in the low noise limit \( (\epsilon \to 0) \); in other words, even consistency cannot be guaranteed without restricting \( \mathcal{F} \) (Chapter 8). The restrictions usually imposed have been on smoothness, requiring that \( f \) have \( \alpha \) derivatives with bounded size in some norm. In the 1970’s and 80’s, the norms chosen were typically either Hölder, requiring uniform smoothness, or Hilbert-Sobolev, requiring smoothness in a mean square sense.

We continue to use statistical games and the minimax principle. In the sequence model (1.10), a strategy for Player I, the Statistician, is a sequence of estimator coefficients \( \hat{\theta}(y) = (\hat{\theta}_I) \), which in terms of functions becomes

\[
\hat{f}(t) = \sum_I \hat{\theta}_I \psi_I(t).
\]

A strategy for player II, Nature, is a prior distribution \( \pi \) on \( \theta \), subject to a constraint that \( \pi \in \mathcal{P} \). In function terms, this corresponds to choosing a random process model for \( \{f(t), 0 \leq t \leq 1\} \). The payoff function from the Statistician to Nature is

\[
B(\hat{\theta}, \pi) = E_\pi \|\hat{\theta}(Y) - \theta\|^2 = E_\pi \int (\hat{f} - f)^2.
\]

The constraint set \( \mathcal{P} = \mathcal{P}(\Theta) \) usually requires, in some average sense usually defined by moments, that \( \pi \) concentrates on the set \( \Theta = \Theta(\mathcal{F}) \).

### 1.4.1. Linear Estimators

To describe the historical background, we start with linear methods. Estimators that are linear functions of observed data arise in a number of guises in application: they are natural because they are simple to compute and study and already offer considerable flexibility. In the single time parameter model (1.11) - (1.10), time-shift invariance is also natural, in the absence of specific prior information to the contrary. Thus, in what follows, we switch freely between time domain \( \hat{f} \) and Fourier-coefficient
domain $\hat{\theta} = (\hat{\theta}_k)$. It then turns out that all shift invariant estimators have
similar structure:

i) **Weighted Fourier series.** Using the Fourier series form (1.10) for the
data model,

$$\hat{\theta}_k = \hat{\kappa}(hk)y_k,$$

where the shrinkage function $\hat{\kappa}$ is decreasing, corresponding to a downweight-
ing of signals at higher frequencies. The ‘bandwidth’ parameter $h$ controls
the actual location of the ‘cutoff’ frequency band.

ii) **Kernel estimators.** In the time domain, the estimator involves convolu-
tion with a window function $K$, scaled to have ‘window width’ $h$:

$$\hat{f}(t) = \int \frac{1}{h} K\left(\frac{t - s}{h}\right)dY(s).$$

The representation (1.12) follows after taking Fourier coefficients.

iii) **Smoothing splines.** The estimator $\hat{\theta}$ minimizes

$$\sum (y_k - \theta_k)^2 + \lambda^{2m} \sum k^{2m} \theta_k^2,$$

where the roughness penalty term takes the mean square form

$$c \int_0^1 (D^mf)^2,$$

in the time domain for some positive integer $r$. In this case, calculus shows
that $\hat{\theta}_k$ again has the form (1.12) with

$$\hat{\kappa}(\lambda k) = [1 + (\lambda k)^{2m}]^{-1}.$$

Each of these forms was studied by numerous authors, either in the white
noise model, or in asymptotically similar models - regression, density estima-
tion - usually over Hölder or Hilbert-Sobolev $\mathcal{F}$. A crowning result of Pinsker
(1980) showed that linear estimators of the form (1.12) were asymptotically
minimax among all estimators over ellipsoidal function classes. More specifi-
cally, suppose that $\mathcal{F}$ may be represented in the sequence space model (1.10)
in terms of an ellipsoid with semi-axes determined by a sequence $\{a_k\}$: thus

$$\Theta(\mathcal{F}) = \{\theta : \sum_k a_k^2 \theta_k^2 \leq C^2\}.$$ 

For example, if $\mathcal{F}$ corresponds to functions with $\alpha$ mean squared derivatives

$$\int (D^nf)^2 \leq L^2,$$

then

$$a_{2k-1} = a_{2k} = (2k)^{\alpha} \quad \text{and} \quad C^2 = L^2/\pi^{2\alpha}.$$ 

We denote the resulting space $\Theta = \Theta^\alpha(C)$. Pinsker constructed a family of
linear shrinkage estimators $\hat{f}_K \leftrightarrow \hat{\theta}_C$ of the form (1.12) with $\hat{\kappa} = \hat{\kappa}_C$ (and also
depending on $\alpha$ and $C)$ so that the worse case MSE of $\hat{\theta}_k$ over $\Theta$ was best possible in the small noise limit:

$$\sup_{\theta \in \Theta} r(\hat{\theta}_k, \theta) \sim R(\Theta; \epsilon) \quad \epsilon \to 0.$$ 

Furthermore, Pinsker showed that an asymptotically least favorable sequence of prior distributions could be described by assigning each $\theta_k$ an independent Gaussian distribution with mean zero and appropriate variance $\bar{\sigma}_k^2(\epsilon, \alpha, C)$. For an explicit example, see Figure 8.

![Least Favorable Prior: Standard Devs](image-url)

**FIGURE 8.** Example of Pinsker’s estimator, in the case of twice differentiable functions ($\alpha = 2$) with $\epsilon^2 = .05$. Panel (b) shows the weights $w_k$ in the optimal linear estimator $\hat{\theta}_{e,k}(y) = w_k y_k$: higher frequencies are downweighted, and discarded for $k > 85$. Panel (a) shows the standard deviations $\eta_k$ of the $k$th co-ordinate in the (approximately) least favorable Gaussian distribution, and panel (c) shows a representative sample path from this distribution.

This result would seem to give definitive justification for the use of linear methods: the least favorable distributions for ellipsoids are approximately Gaussian, and for Gaussian processes, the optimal (Bayes) estimators are linear.

At about this time, however, some cracks began to appear in this pleasant linear/Gaussian picture. In the theoretical domain, Nemirovskii (1985) and Nemirovskii et al. (1985) showed, for certain function classes $\mathcal{F}$ in which smoothness was measured in a mean absolute error ($L_1$) sense, that linear estimators were no longer minimax, and indeed had suboptimal rates of convergence of error to zero as $\epsilon \to 0$. 

18
Meanwhile, methodological and applied statistical research harnessed computing power to develop smoothing algorithms that used different, and \textit{data-determined}, window widths at differing times \( t \). This is in clear contrast with the fixed global bandwidth \( h \) implied by the kernel representation (1.13). For example, Cleveland (1979) investigates local smoothing, and Friedman & Stuetze (1981), in describing the univariate smoother they constructed for projection pursuit regression say explicitly

\begin{quote}
“the actual bandwidth used for local averaging at a particular value of (the predictor) can be larger or smaller than the average bandwidth. Larger bandwidths are used in regions of high local variability of the response.”
\end{quote}

This amounts to an implicit rejection of the ellipsoid model. The algorithms of Friedman/Stuetze and others were iterative, involving multiple passes over the data, and were thus beyond theoretical analysis.

\textit{Example continued.} Returning to the NMR signal, from Figure 9 it seems intuitively reasonable to use averaging kernels of different window width \( h \) at different parts of the spectrum. On the other hand, a translation invariant linear rule of the form (1.12) - (1.13) forces a single width throughout, with unsatisfactory results away from the peaks.

\begin{figure}
\includegraphics[width=0.5\textwidth]{figure9a.png}
\includegraphics[width=0.5\textwidth]{figure9b.png}
\caption{Panel (a): a small averaging window width \( h \) seems appropriate near the central peak, while a larger one would average out the baseline noise at the right hand end of the spectrum. Panel (b): Estimate that corresponds (see Chapter ??) to a single window width \( h \) - this \( h \) must be kept small to preserve the peak, and hence insufficient noise averaging occurs on the baseline.}
\end{figure}
1.4.2. Wavelet bases and thresholding. The appearance of wavelet bases enabled a reconciliation of the Gaussian-linear theory with these divergent trends. Informally, this might be explained by 'Mallat’s heuristic', quoted by Donoho (1993):

“Bases of smooth wavelets are the best bases for representing objects composed of singularities, when there may be an arbitrary number of singularities, which may be located in all possible spatial positions.”
This captures the notion that a function with spatially varying smoothness — transients at some points, very smooth elsewhere — might be sparsely represented in a smooth wavelet basis and hence well estimated.

Indeed, Panel (c) of Figure 10 illustrates the improvement yielded by wavelet thresholding on a noisy NMR signal. Figure 11 shows in detail what happens to the wavelet coefficients at level 7. The left panel shows the 128 raw wavelet coefficients of the NMR signal. Hard thresholding, as at (1.4), is applied with threshold \( \lambda = \hat{\epsilon} \sqrt{2 \log n} \), yielding the “denoised” coefficients in the right panel.

The final wavelet thresholding reconstruction, panel (c), should be compared with panel (e), which shows an arguably best near linear estimator in the spirit of Pinsker’s theorem (for further details, see Donoho & Johnstone (1995) and Chapter 9 here.) Clearly the Pinsker type estimator fails to adjust the (implied) window width in (1.13) so as both to capture the sharp peaks and to average out noise elsewhere.

![Figure 11. Caption here](image.png)

We turn now to describe some of the theory that underlies these reconstructions and Mallat’s heuristic. More technically, wavelets form an unconditional basis simultaneously for a vast menagerie of function spaces, allowing more flexible measures of smoothness than the Hölder and Hilbert-Sobolev spaces hitherto used in statistics. An unconditional basis for a Banach space \( B \) with norm \( \| \cdot \| \) is defined by a countable family \( \{ \psi_I \} \subset B \) with two key properties:

- (i) any element \( f \in B \) has a unique representation:

\[
f = \sum_{I} \theta_I \psi_I
\]

in terms of coefficients \( \theta_I \in \mathbb{C} \), and
• (ii) shrinkage: there is an absolute constant $C$ such that if $|\theta_I| \leq |\theta_I|$ for all $I$, then

$$\left\| \sum \theta_I \psi_I \right\| \leq C \left\| \sum \theta_I \psi_I \right\|.$$ 

The corresponding statistical significance of these two properties is firstly that functions $f \in B$ may be described in terms of coefficient sequences $\{\theta_I\}$, and secondly that the basic statistical operation of shrinkage on these sequences, whether linear or via thresholding, is stable in $B$, in that the norms cannot be badly inflated. Notably, this property is not shared by the Fourier basis, (Kahane et al. 1977).  \footnote{See Donoho (1993) for a formalization of Mallat's heuristic using the unconditional basis property.}

Figure 12 represents the class of Besov spaces schematically in terms of the smoothness index $\alpha$ (= number of derivatives) and the homogeneity index $p$, plotted as $1/p$ as is customary in harmonic analysis. Each point $(\alpha, 1/p)$ corresponds to a class of Besov spaces. The vertical line $p = 2$ represents the Hilbert-Sobolev smoothness spaces traditionally used in statistics, while points to the right are spaces with $p < 2$, hence having some degree of sparsity of representation in the wavelet domain.

![Figure 12](image)

**FIGURE 12.** (a) Schematic diagram of Besov spaces of varying homogeneity $p$ and smoothness. Spaces above the diagonal line $\alpha = 1/p$ consist of functions that are at least continuous. Point at $(1, 1)$ corresponds to the Bump Algebra. (b) Caricature evoking a function in the Bump Algebra - a superposition of Gaussians at widely different spatial scales.

To be more concrete, we consider a single example from the family of Besov spaces, the Bump Algebra, described in Meyer (1990, Section 6.6) \footnote{For a discussion in terms of the space of Bounded Total Variation, see Mallat (1998a).}. Let

$$g_{\mu, \sigma}(t) = \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$$

denote a normalized Gaussian bump with location $\mu$ and scale $\sigma$. The Bump Algebra on $\mathbb{R}$ is the collection of all functions $f$ representable as a convergent
superposition of signed, scaled and located bumps:
\[ f = \sum_{i=1}^{\infty} \alpha_i g_{\mu_i,\sigma_i}, \quad \sum_{i=1}^{\infty} |\alpha_i| < \infty. \]

This might seem a plausible model for (say) signed spectra with peaks of varying location, width and height. Compare Figure 12(b). As Meyer notes, while the simplicity of this description is perhaps deceptive due to lack of uniqueness in the representation, an equivalent and stable description can be given via a smooth wavelet orthonormal\( \{\psi_I\} \). When restricted to \( L_2[0,1] \), we may use the index system \( I = \bigcup_j \ld I_j \) of Section 1.3. A subset \( B_C \) with norm at most \( C \) is defined by those
\[ f = \sum \theta_I \psi_I \]
for which
\[ \sum_{j \geq 0} 2^{j/2} \sum_{k \in \ld I_j} |\theta_{jk}| \leq C. \] (1.14)

The condition (1.14) is a geometrically scale-weighted combination of the \( \ell_1 \) norms of the coefficients at each level \( j \). The full Bump Algebra is simply the union of all \( B_C \). This wavelet representation is the key to statistical results: the appearance of the \( \ell_1 \) sums hints that we may expect to carry over some phenomena from the \( \ell_1 \) statistical game in the monoresolution model of Section 1.2 to the wavelet setting.

Indeed, we now consider a statistical game in the wavelet domain, with the prior constraint family \( \mathcal{P} \) consisting of those priors \( \pi \) such that the \( \pi \)-expectation of the left side of (1.14) is bounded by \( C \). On average, the samples drawn from \( \pi \) should have well behaved norm (in the bump algebra sense).

In view of the \( \ell_1 \) norm analysis in Section 1.2, it is perhaps now not surprising that no linear estimator can achieve optimal rates of convergence. In fact, as is shown in Chapter 13, the minimax risk \( R_e(B_C) \) decreases like \( C^{2/3} \epsilon^{1/3} \), whereas the best rate possible for linear estimators is much slower, namely \( O(\epsilon) \).

The sequence space structure provided by the unconditional basis property also implies that optimal estimators in the statistical game are diagonal: \( \theta_I(y) = \delta_I(y_I) \) depends on \( y_I \) alone. While these optimal estimators cannot be described explicitly, this separation of variables is an important simplification. For example, the least favorable distributions are, at least asymptotically, found to be obtained by making the wavelet coefficients independent and indeed identically distributed within each level. Furthermore, it turns out that thresholding estimators, as described in Section 1.2, but now with thresholds depending on level, have MSE always within a constant multiple (\( \leq 2.2 \)) of the optimal value.

Thresholding in a wavelet basis automatically has the spatial adaptivity that previous algorithmic work sought: the effective window width at a given
time point is proportional to $2^{-j(t_0)}$ where $j(t_0)$ is the finest level wavelet coefficient that survives thresholding among those wavelets whose support contains $t_0$. In regions of strong and narrow signal peaks, $j(t_0)$ will be large, and the implied window width small, while in regions of mostly noise, $j(t_0)$ will be much smaller and the bandwidth correspondingly large.

Sample paths drawn from approximately least favorable distributions for $B_C$ are informative. A sample realization of $f$ can be plotted by substituting a sample draw of coefficients $\theta_{jk}$ into

$$f = \sum \theta_{jk} \psi_{jk}.$$  

By considering only threshold rules, which are nearly minimax as just mentioned, it can be shown that a near least favorable distribution can be constructed from three point distributions $(1 - \epsilon_j)\delta_0 + \frac{1}{2}\epsilon_j(\delta_{\mu_j} + \delta_{-\mu_j})$ that are quite similar to those arising in the monoresolution model at (1.6). Now however, the location $\mu_j$ and size $\epsilon_j/2$ of the non-zero atom and its reflection depends on the level $j$, but within level, the $2^j$ draws are independent, as they are in (1.6). A numerical optimization (Johnstone 1994) allows evaluation of $\mu_j$ and $\epsilon_j$ for a given $\epsilon$ and $B_C$. Figure 13(a) shows part of a representative sample path and Figure 13(b) the corresponding individual wavelet coefficients for this distribution.

Figure 13(c) shows a corresponding part of a sample path drawn from the Gaussian least favorable distribution on wavelet coefficients, Figure 13(d) corresponding to the ellipsoid $\Theta^\alpha(C)$ with $\alpha = 1$ derivatives assumed to be square integrable. Again, the wavelet coefficients are i.i.d. within level, but now are drawn from a Gaussian distribution with variance $\tau_j^2(\alpha, C, \epsilon)$ determined by Pinsker’s solution. The two plots are calibrated to the same indices of smoothness $\alpha = 1$, scale $C = 1$ and noise level $\epsilon = 1/64$.

The qualitative differences between these plots are striking: the Gaussian sample path has a spatially homogeneous irregularity, with the sample wavelet coefficients being ‘dense’, though decreasing in magnitude with increasing scale or “frequency octave”. In contrast, the Bump Algebra sample path has a greater spikiness: the sample wavelet coefficients have increasing sparsity and magnitude with each increasing scale. These differences become even more pronounced if one increases the smoothness $\alpha$ and decreases the homogeneity index $p - \epsilon$, for example the plots for $\alpha = 2, p = 1/2$ in Johnstone (1994).

To summarize: with only the sparsity in mean constraint, and no other restriction on estimators or prior distributions, coordinatewise thresholding and sparse priors emerge as the near optimal strategies for the Bump algebra statistical game. Thresholding has better MSE, indeed faster rates of convergence, than any linear estimate over $B$. These rates of convergence results are reflected visually in the relatively much more noise free reconstruction by wavelet thresholding in Figure 10.

We mention briefly the problem of adaptation. The near optimality of thresholding and sparse priors holds in similar fashion for a large class of
Figure 13. (a): A segment of a sample path on $[0, 1]$ from a prior distribution on wavelet coefficients that is approximately least favorable for a Bump Algebra ball of form (1.14). (b): the corresponding wavelet coefficients - those at level $j$ are i.i.d. draws from a three point distribution $(1 - \epsilon_j)\delta_0 + \frac{1}{2}\epsilon_j(\delta_{\mu_j} + \delta_{-\mu_j})$ as described in the text. The wavelets $\psi_{jk}$ are derived from the $N = 8$ instance of the Daubechies (1992, Ch. 6) “closest to linear phase” filter. (c) sample path from the Gaussian process that is the least favourable distribution for an ellipsoid $\Theta^*(C)$ with $\alpha = 1$ square integrable derivatives. (d): corresponding wavelet coefficients with variance $\tau_j^2$ decreasing with $j$.

<table>
<thead>
<tr>
<th>Prior Constraint</th>
<th>ellipsoid ($\ell_2$)</th>
<th>Bump Algebra ($\ell_1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimax estimator</td>
<td>linear</td>
<td>thresholding (level dep)</td>
</tr>
<tr>
<td>least favorable $\pi$</td>
<td>Gaussian</td>
<td>sparse (three point)</td>
</tr>
<tr>
<td>minimax M.S.E.</td>
<td>$O(\epsilon)$</td>
<td>$\sim C^{2/3}\epsilon^{1/3}$</td>
</tr>
</tbody>
</table>

Table 2. Comparison of structure of optimal strategies in the multiresolution game under traditional and sparsity assumptions.
Besov space constraints described by $(\alpha, 1/p)$ and size parameter $C$. The optimal threshold estimator in each case will depend on $(\alpha, p, C)$ – can one give an estimator with optimal or near-optimal properties without needing to specify $(\alpha, p, C)$? One very simple possibility, already shown in Figure 10 and explored at length in Donoho et al. (1995), is to use hard or soft thresholding at threshold $\sqrt{2\log n}$, where $n$ is the number of observations, or wavelet coefficients. This estimator has a remarkably robust near adaptivity – it nearly achieves, up to logarithmic terms, the minimax rate of convergence simultaneously over a wide range of both functional classes and error measures – not solely mean squared error.

1.5. Concluding Remarks.

Wavelets have enabled the development of theoretical support in statistics for the important notions of sparsity and thresholding.

In contrast, much work in modern curve fitting and regression treats the imposition of smoothness as a guiding principle. Wavelets prompt us to think of smoothness as a particular case of a more general principle, namely sparsity of representation. It is in fact sparsity of representation that determines when good estimation is possible.

Pursuing the sparsity idea for functions of more than one variable leads to systems other than literal wavelets, as discussed by Candès & Donoho (1999).

We may expect to see more use of “dyadic thinking” in areas of statistics and data analysis that have little directly to do with wavelets. This is likely both in the development of methods, and also as a metaphor in making simpler models for theoretical analysis of other more complex procedures.

Many thanks to Marc Raimondo for his help in the preparation of Figure 13. The author gratefully acknowledges financial support from the National Science Foundation (DMS 9505151).
Part 1

SEQUENCE MODELS
CHAPTER 2

The Multivariate Normal Distribution

2.1. Linear Regression and sequence models

The model which will occupy most of our attention can be simply stated:

\[(2.1) \quad y_{\nu} = \alpha_{\nu} \theta + \epsilon_{\nu}, \quad \nu \in I.\]

$p$Here $(y_{\nu})$ represents the observed data: the index set $I$ will typically be \{1\}, \{1, \ldots, n\} or \(\mathbb{N}\). The signal $\theta = (\theta_{\nu}, \nu \in I)$ is unknown and non-negative. The $\epsilon_{\nu}, \nu \in I$ are independent $N(0,1)$ noise variables, and $\epsilon$ is the noise level, generally assumed known.

The major special case of model (2.1) occurs when all $\alpha_{\nu} = 1$—this corresponds to direct estimation, while the setting of general $\alpha_{\nu}$ corresponds to indirect estimation. (Compare [???] below).

As will be discussed in greater detail below, model (2.1) often arises from a model \(y = A\theta + \epsilon z\) after taking coefficients in a suitable orthonormal basis (which diagonalizes \(A\) in the general indirect estimation setting). Typically, then, though not always, the index $\nu$ corresponds in some way to increasing frequency. It may also be a one dimensional encoding of a higher (usually two or three) dimensional index.

Among the issues to be addressed are

(i) we imagine $(\theta_{\nu})$ to be “high dimensional”. In particular, as $\epsilon$ decreases, the number of parameters $n = n(\epsilon)$ may increase. This makes the approach fundamentally nonparametric.

(ii) what are the effects of $(\alpha_{\nu})$, i.e. the consequences of indirect estimation on the ability to recover $\theta$.

(iii) asymptotic behavior as $\epsilon \to 0$. This corresponds to a low-noise (or large sample size) limit.

(iv) optimality questions: can one describe bounds for minimum possible error of estimation and estimators that (more or less) achieve these bounds?

In the direct estimation setting model (2.1) is also written in the multivariate normal mean form $y \sim N_n(\theta, \epsilon^2 I)$ that is the central model for classical parametric statistical theory. We may write $\phi_k(y - \theta) = \Pi \phi_k(y_i - \theta_i)$ for the joint density of $(y_i)$ with respect to Lebesgue measure. The univariate densities $\phi_k(y_i) = (2\pi \epsilon^2)^{-1/2} \exp\{-y_i^2/2\epsilon^2\}$. Also, $\Phi(y) = \int_{-\infty}^{y} \phi(s)ds$ will denote the standard univariate Gaussian cumulative distribution function.
Relating Regression Models to the Sequence Form Consider the basic Gaussian linear model:

$$Y = A\beta + \sigma e$$

where $Y$ is an $N$-dimensional data vector, $\beta$ is an $n \leq N$ dimensional unknown parameter vector and $A$ is an $N \times n$ design matrix. Again $e \sim N_N(0, I)$ and $\sigma$ is the known noise level.

In a number of situations (though by no means all), this model can be usefully converted into sequence form (2.1). The first three examples below concern direct estimation, and the fourth indirect recovery.

(a) If estimation of $\theta = A\beta$ is of interest then one recovers model (3.1) after projecting onto the range of $A$. Such a situation arises, for example, in certain prediction problems. In the “in-sample” setting, one assesses a predictor $\hat{\theta} = A\hat{\beta}$ of a new observation vector $Y^{*} = A\beta + \sigma e^{*}$ via the mean squared error $E\|A\hat{\beta} - Y^{*}\| = E\|A(\hat{\beta} - \beta) - \sigma e^{*}\|^2 = E\|\theta - \theta\|^2 + N\sigma^2$.

(b) If estimation of $\beta$ is of primary interest, now assume that $A$ is orthogonal: $A'A = mI_m$. The columns of $A$ might be orthogonal polynomials or other systems of functions, or orthogonal contrasts in the design of experiments, and so on. Specific examples include weighing designs, Hadamard and Fourier transforms (as in magnetic resonance imaging - More details??) The model can be put in the form (3.1) simply by premultiplying by $m^{-1}A'$: define $y = m^{-1}A'Y, z = m^{-1/2}A'e$, and note especially the noise calibration $\epsilon = \sigma/\sqrt{m}$.

(c) While this formulation appears parametric, it formally also covers the setting of non-parametric regression on a fixed equi-spaced design. Thus, the model

$$Y_i = f(i/n) + \sigma e_i, \quad i = 1, \ldots, n$$

becomes an example of (2.2) if one uses as design matrix an inverse discrete orthogonal wavelet (or Fourier) transform $W^t$ to express $f = \{f(i/n)\} = W^t\theta$. The components of $y$ and $\theta$ are wavelet (or Fourier) coefficients of $Y$ and $f$ respectively.

[Note: it is appropriate to normalize the transform to be distance preserving in the sense that

$$\|\hat{f}\|_{L_2}^2 := (1/n) \sum_{i=1}^n f^2(i/n) = \sum_{i=1}^n \theta_i^2 = \|\theta\|_{L_2}^2.$$]

In matrix terms, this implies that $A'A = nI$ and $\epsilon = \sigma/\sqrt{n}$.]

(d) Change of co-ordinates We use the singular value decomposition of $A = UDV^T$, where $U_{N \times n}$ and $V_{n \times n}$ are orthogonal and $D = \text{diag}(\alpha_i)$ diagonal matrices respectively. Premultiply equation (2.2) by $U^T$, obtaining (2.1) with the definitions

$$y = U^TY, \quad x = U^Te, \quad \theta = V^T\beta, \quad \epsilon = \sigma.$$
Although \( \theta \) are now coefficients in the singular function co-ordinate system, since \( V \) is orthogonal, errors in estimation of \( \beta \) are preserved: if we identify \( \hat{\theta} = V^T \hat{\beta} \), then
\[
\| \hat{\theta} - \theta \|_2 = \| V^T (\hat{\beta} - \beta) \|_2 = \| \hat{\beta} - \beta \|_2.
\]
In this sense, any regression model (2.2) can be put in sequence form (2.1).

**Remark.** If we drop the requirement that the errors be normally distributed, keeping only that \( e \sim (0, I) \), then the same will be true of the transformed errors: \( z \sim (0, I) \). If the matrix \( U \) is in some sense ‘dense’, so that \( z_i = \sum_k u_{ik} \epsilon_k \) has many non-zero terms, then the central limit theorem will push the distribution of the \( z_i \) toward normality.

### 2.2. Penalized Least Squares and thresholding

### 2.3. Priors, posteriors and Bayes estimates

#### 2.4. Stein’s Unbiased Risk formula

#### 2.5. The James Stein estimate

#### 2.6. Mean squared error properties

### 2.7. Differentiation and noise amplification

#### 2.8. Bibliographic Notes

#### 2.9. Problems

1. Suppose that \( \theta \sim N_n(\theta_0, I) \) and that \( y|\theta \sim N_n(\theta, I) \). Let \( p(\theta, y) \) denote the joint density of \((\theta, y)\). Show that
\[
-2 \log p(\theta, y) = \theta^T B \theta - 2 \gamma^T \theta + r(y).
\]
Identify \( B \) and \( \gamma \), and conclude that \( \theta|y \sim N(\theta_y, \Sigma_y) \) and evaluate \( \theta_y \) and \( \Sigma_y \).

2. Establish part (iii) of Proposition, p. 10: If \( \gamma(\mu) \) is symmetric and unimodal, then the posterior median \( \hat{\mu}(x) \) satisfies
\[
0 \leq \hat{\mu}(x) \leq x.
\]
[Hint: show first that if \( x > 0 \),
\[
p(x - v|X = x, \mu \neq 0) \geq p(x + v|X = x, \mu \neq 0)
\]
for all \( v > 0 \).]

3. Let \( F \) be an arbitrary probability distribution function on \( \mathbb{R} \). A median of \( F \) is any point \( a_0 \) for which
\[
F(-\infty, a_0] \geq \frac{1}{2} \quad \text{and} \quad F[a_0, \infty) \geq \frac{1}{2}.
\]
Show (without calculus!) that
\[
a \rightarrow M(a) = \int |a - \theta| dF(\theta)
\]
is minimized at any median $a_0$. 
CHAPTER 3

The Gaussian Sequence model

The basic model throughout is a finite or countably infinite dimensional signal observed in white Gaussian noise:

\[ y_i = \theta_i + \epsilon z_i \quad i \in I \]

where \( I = \{1\}, \{1, \ldots, n\} \) or \( \mathbb{N} \). Here \( z_i, i \in I \) are independent \( N(0,1) \) noise variables, and \( \epsilon \) is the noise level, assumed known. The signal \( \theta = (\theta_i, i \in I) \) is unknown and to be estimated.

Our initial goal is to describe, define and motivate this model, and slight generalizations.

3.1. The Basic Model

**Finite Dimensions.** When \( I \) is finite, model (3.1) is also written in the multivariate normal mean or Gaussian shift form \( y \sim N_n(\theta, \epsilon^2 I) \) that is the central model for classical parametric statistical theory. We may write \( \phi_k(y - \theta) = \prod \phi_k(y_i - \theta_i) \) for the joint density of \( (y_i) \) with respect to Lebesgue measure. The univariate densities \( \phi_k(y_i) = (2\pi \epsilon^2)^{-1/2} \exp \{-y_i^2/2\epsilon^2\} \). Also, \( \Phi(y) = \int_{-\infty}^{y} \phi(s)ds \) denote the standard univariate Gaussian cumulative distribution function.

Although formally a subclass of the parametric models just mentioned, it is useful to highlight regression models:

\[ Y = X\beta + \sigma e \]

where \( Y \) is an \( N \)-dimensional data vector, \( \beta \) is an \( n \leq N \) dimensional unknown parameter vector and \( X \) is an \( N \times n \) design matrix. Again \( e = N_X(0, I) \) and \( \sigma \) is the known noise level.

If estimation of \( \theta = X\beta \) is of interest then one recovers model (3.1) after projecting onto the range of \( X \). Such a situation arises, for example, in certain prediction problems. In the “in-sample” setting, one assesses a predictor \( \hat{\theta} = X\hat{\beta} \) of a new observation vector \( Y^* = X\beta + \sigma e^* \) via the mean squared error \( E\|X\hat{\beta} - Y^*\|^2 = E\|X(\beta - \beta) - \sigma e^*\|^2 = E\|\hat{\theta} - \theta\|^2 + N\sigma^2 \).

If estimation of \( \beta \) is of primary interest, now assume that \( X \) is orthogonal: \( X'X = mI_m \). The columns of \( X \) might be orthogonal polynomials or other systems of functions, or orthogonal contrasts in the design of experiments, and so on. The model can be put in the form (3.1) simply by premultiplying by \( m^{-1}X' \); define \( y = m^{-1}X'Y, z = m^{-1/2}X'e \), and note the noise calibration \( \epsilon = \sigma / \sqrt{m} \).

32
While this formulation appears parametric, it formally also covers the setting of non-parametric regression on a fixed equi-spaced design. Thus, the model

\[ Y_i = f(i/n) + \sigma e_i, \quad i = 1, \ldots, n \]

becomes an example of (3.2) if one uses as design matrix an inverse discrete orthogonal wavelet transform \( W^t \) to express \( f = \{ f(i/n) \} = W^t \theta \). The components of \( y \) and \( \theta \) are wavelet coefficients of \( Y \) and \( f \) respectively.

[Note: it is appropriate to normalize the transform to be distance preserving in the sense that]

\[ \| f \|_{l^2_{2,n}}^2 := (1/n) \sum_{i=1}^{n} f^2(i/n) = \sum_{i=1}^{n} \theta_i^2 = \| \theta \|_{l^2_{2,n}}^2. \]

In matrix terms, this implies that \( X'X = nI \) and \( \epsilon = \sigma/\sqrt{n} \).

**Infinite Dimensions.** Defining Gaussian measures on infinite dimensional spaces is not completely straightforward and we refer to books by Kuo (1975) and Bogachev (1998) for complete accounts. For the sequence model (3.1) with \( I = \mathbb{N} \), the subtleties can usually be safely ignored. For the record, as sample space for model (3.1), we take \( \mathbb{R}^{\infty} \), the space of sequences in the product topology of pointwise convergence, under which it is complete, separable and metrizable. It is endowed with the Borel \( \sigma \)-field, and as dominating measure, we take \( P_0 = P_{0,\epsilon} \), the centered Gaussian Radon measure (see Bogachev (1998, Example 2.3.5)) defined as the product of a countable number of copies of the \( N(0, \epsilon^2) \) measure on \( \mathbb{R} \).

For each \( \theta \in \Theta = \ell_2(\mathbb{N}) \), the measure \( P_0 \) with mean \( \theta \) is absolutely continuous, indeed equivalent, to \( P_0 \), and has density

\[ f_\theta(x) = dP_0/dP_0 = \exp\{ \langle \theta, x \rangle / \epsilon^2 - \| \theta \|^2 / 2\epsilon^2 \}. \]

Note that the random variable \( \langle \theta, x \rangle \) appearing in the density has a \( N(0, \| \theta \|^2) \) distribution under \( \Theta \) and in particular is finite \( P_0 \)-almost surely.

**Remarks.** 1. In fact, it follows from the classical theorem of Kakutani (1948) on product measures that membership of \( \theta \) in \( \ell_2 \) is a necessary and sufficient condition for the distribution of \( y \) to be absolutely continuous with respect to that obtaining when \( \theta = 0 \), and further if \( \theta \notin \ell_2 \), then the two distributions are mutually singular.

2. Bogachev (1998, Theorem 3.4.4) shows that in a certain, admittedly weak sense all infinite dimensional Gaussian measures are isomorphic to the sequence measure \( P_0 \).

Model (3.1) is actually equivalent to the continuous “signal in white Gaussian noise.” Suppose that \( W_t \) is a standard Brownian motion and one observes

\[ Y_t = \int_0^t f(s)ds + \epsilon W_t, \quad 0 \leq t \leq 1, \]
or equivalently
\[ dY_t = f(t)dt + \sigma dW_t, \quad 0 \leq t \leq 1. \]

The goal is to estimate \( f \), assumed to lie in \( L_2[0, 1] \), for example using the integrated squared error loss
\[ \| \hat{f} - f \|_{L_2}^2 = \int_0^1 (\hat{f} - f)^2. \]

If we take the coefficients of model (3.4) in a complete orthonormal basis \( \{ \varphi_i \}_{i=1}^\infty \) for \( L_2[0, 1] \), we recover the basic model (3.1) with \( I = \mathbb{N} \). Indeed, simply define
\[ y_i = \int \varphi_i dY, \quad \theta_i = \int \varphi_i f, \quad z_i = \int \varphi_i dW. \]

The sequence \( \{ z_i \} \) has the required independence and variance structure since, from the basic rules for Wiener integrals,
\[ \text{Cov}(z_i, z_j) = E \int \varphi_i dW \int \varphi_j dW = \int \varphi_i \varphi_j = \delta_{ij}. \]

[The Kronecker delta \( \delta_{ij} = 1 \) if \( i = j \) and 0 otherwise.] The Parseval relation connects the \( L_2 \) and \( \ell_2 \) error measures:
\[ \| \hat{f} - f \|_{L_2}^2 = \int_0^1 (\hat{f} - f)^2 = \sum_i (\hat{\theta} - \theta)^2. \]

One can formally extend the infinitesimal representation (3.5) to a compact set \( D \subset \mathbb{R}^n \) if \( t \to W_t \) is a--parameter Brownian sheet (Hida 1980). If \( \varphi_i \) is an orthonormal basis for \( L_2(D) \), then the operations (3.6) again yield data in the form of model (3.1).

The finite dimensional multivariate normal model is the foundation of parametric statistical theory. For nonparametric statistics, the signal in Gaussian white noise model, or its sequence version expressed in an orthonormal basis, plays an equivalent role. It first emerged in communications theory in work of Kotelnikov (1959). As Ibragimov & Khas’minskii (1981, 1980, for example) have argued, the difficulties thrown up by the “signal+noise” model are essential rather than technical in nature.

Heuristically, the connection between (3.5) and (3.3) arises by forming the partial sum process of the discrete data:
\[ Y_n(t) = n^{-1} \sum_{i=1}^{[nt]} y_i = n^{-1} \sum_{i=1}^{[nt]} f(i/n) + \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} e_i. \]

The signal term is a Riemann sum approximating \( \int_0^t f \), and the error term \( n^{-\frac{1}{2}} \sum_{i=1}^{[nt]} z_i \) converges weakly to standard Brownian motion as \( n \to \infty \). Making the calibration \( \epsilon = \epsilon(n) = \sigma / \sqrt{n} \), and writing \( Y_{\epsilon(n)} \) for the process in (3.4), we see that, formally, the processes \( Y_{\epsilon(n)} \) and \( Y_n \) merge as \( n \to \infty \).
Brown & Low (1996) have given a formal statement of this result using convergence of likelihood ratios and Le Cam’s notion of asymptotic equivalence of experiments. This implies closeness of risks for all decision problems with bounded loss. Here we simply observe that there is convergence of mean average squared errors (for fixed functions \( \hat{f} \) and \( f \in L^2[0, 1] \)):

\[
n^{-1}\|\hat{f} - f\|^2_{L^2} = n^{-1}\sum_{i=1}^{n}[\hat{f}(i/n) - f(i/n)]^2 \rightarrow \int_{0}^{1}[\hat{f} - f]^2.
\]

Brown & Low (1996) also discussed a random design setting in which the \( t_i \) are replaced by i.i.d draws from a fixed design distribution.

3.2. Non-white Gaussian sequence models

**Finite Dimensions.** The general multivariate normal distribution \( X \sim N_n(\mu, \Sigma) \) on \( \mathbb{R}^n \) has symmetric and non-negative definite covariance matrix \( \Sigma \). Thus, \( \Sigma \) is orthogonally similar to a diagonal matrix with non-negative entries: \( \Sigma = U\Lambda U' \), with \( U \) an orthogonal matrix and \( \Lambda = \text{diag}(\lambda_i^2) \). Setting \( y = UX \) yields new, independent coordinates: \( \text{Cov}(y) = U\Sigma U' = \Lambda \).

Writing \( \theta = U\mu \), we have a sequence model with unequal variances:

\[
y_i = \theta_i + \lambda_i z_i,
\]

for \( i = 1, \ldots, n \) and the \( z_i \) independent standard Gaussian variables.

**Infinite Dimensions.** When the index set is countable: \( i \in \mathbb{N} \), we will take (3.10) as our definition of a non-white sequence model, modified slightly to allow for a noise level \( \epsilon \) that may change while the \( \lambda_i \) remain fixed:

\[
y_i = \theta_i + \epsilon \lambda_i z_i, \quad i \in \mathbb{N}.
\]

We continue to take as sample space \( \mathbb{R}^\infty \) with the Borel sigma field for coordinatewise convergence, and as dominating measure \( P_0 \) defined in the previous section. Kakutani’s theorem now says that \( dP_0/dP_0 \) exists if and only if \( \sum \beta_i^2/\lambda_i^2 < \infty \), in which case the density is given in terms of the inner product \( \langle \theta, x \rangle_\Lambda = \sum \theta_i x_i / \lambda_i^2 \) by

\[
\frac{dP_0}{dP_0} = \exp\{\langle \theta, x \rangle_\Lambda / \epsilon^2 - \frac{1}{2}\|\theta\|_\Lambda^2 / \epsilon^2\}.
\]

The Karhunen-Loève transform. Suppose that \{\( Z(t), t \in T \)\} is a zero mean Gaussian random process on an index set \( T \). Assume also that \( Z \) is continuous in quadratic mean, or equivalently that the covariance function (or kernel)

\[
R(s, t) = EZ(s)Z(t)
\]

is jointly continuous in \((s, t) \in T^2\). Then there exist orthonormal functions \( \phi_i \in L^2(T) \) which are eigenfunctions of \( R \), that is

\[
\int R(s, t)\phi_i(t)dt = \lambda_i^2 \phi_i(s), \quad s \in T,
\]

35
and there exist independent $z_i \sim N(0, 1)$ such that the representation

$$Z(t) = \sum_i \lambda_i z_i \phi_i(t)$$

converges in mean-square on $L_2(T)$. If the eigenfunctions $\phi_i$ corresponding to $\lambda_i > 0$ are not complete, then we may add an orthonormal basis for the orthogonal complement of the closure of the range of $R$ in $L_2(T)$ and thereby obtain an orthonormal basis for $L_2(T)$. Since $R$ is symmetric, these $\phi_i$ correspond to $\lambda_i = 0$.

Now suppose that $Z(t)$ is observed with an unknown drift function added:

$$Y(t) = \theta(t) + \varepsilon Z(t), \quad t \in T.$$

If $\theta \in L_2(T)$, then we may take coefficients in the orthonormal set $\{\phi_i\}$:

$$y_i = \langle Y, \phi_i \rangle, \quad \theta_i = \langle \theta, \phi_i \rangle,$$

to obtain exactly the sequence model (3.11). [Of course, co-ordinates corresponding to $\lambda_i = 0$ are observed perfectly, without noise.]

### 3.3. Example: Integrated Wiener process priors.

The $m - 1$-fold integrated Wiener process is defined by

$$Z_m^0(t) = \int_0^1 (t - u)^{m-1} dW(u), \quad t \in [0, 1].$$

The “free” Wiener process (so christened by Shepp (1966)) is derived from this with the aid of i.i.d standard Gaussian variables $\xi_0, \ldots, \xi_{m-1}$ independent of $Z_m^0$:

$$Z_m^\sigma(t) = \sigma \sum_{j=0}^{m-1} \xi_j \frac{t^j}{j!} + Z_m(t).$$

Most interesting is the case $m = 2$, since it corresponds to cubic smoothing splines:

$$Z_2^\sigma(t) = \sigma \xi_0 + \sigma \xi_1 t + \int_0^t (t - u) dW(u).$$ (3.12)

Wahba (1978, 1983, 1990) has advocated the use of $Z_m^\sigma$ as a prior distribution for Bayesian estimation in the context of smoothing splines (actually, she recommends using $\sigma \to \infty$, for reasons that will be apparent.) This prior distribution has some curious features, so we explore its Karhunen-Loève transform now as preparation for later use. The key conclusion is that for each $\sigma \geq 0$, and in the $\sigma \to \infty$ limit, the eigenvalues satisfy

$$\lambda_i \sim 1/(\pi i)^m, \quad \text{as } i \to \infty.$$

We discuss only the cases $m = 1, 2$ here and refer to [REF??] for the general case.
However, it is simpler to discuss the $m = 1$ situation first, with
\[ Z_1^m(t) = \sigma \xi_0 + W(t), \]
and covariance kernel
\[ R_\sigma(s, t) = \text{Cov} \left( Z_1^m(s), Z_1^m(t) \right) = \sigma^2 + s \wedge t. \]
The eigenvalue equation $R_\sigma \phi = \lambda^2 \phi$ becomes
\[ \sigma^2 \int_0^1 \phi(t) dt + \int_0^s t \phi(t) dt + s \int_s^1 \phi(t) dt = \lambda^2 \phi(s). \]
Differentiating with respect to $s$ yields
\[ \int_s^1 \phi(t) dt = \lambda^2 \phi'(s) \]
and differentiating a second time yields the second order ordinary differential equation
\[ -\phi(s) = \lambda^2 \phi''(s) \quad 0 \leq s \leq 1. \]
The homogeneous equation $\lambda^2 \phi'' + \phi = 0$ has two linearly independent solutions given by trigonometric functions
\[ \phi(t) = a \sin(t/\lambda) + b \cos(t/\lambda). \]
The equations (3.13) and (3.14) impose boundary conditions which non-zero eigenfunctions must satisfy:
\[ \phi'(1) = 0, \quad \phi(0) = \phi(0)/\sigma^2. \]
[The first condition is evident from (3.14) while the second follows by combining the two equations: $\lambda^2 \phi(0) = \int \phi = \lambda^2 \phi(0)/\sigma^2$.]

Let us look first at the $\sigma \rightarrow \infty$ limit advocated by Wahlba. In this case the boundary conditions become simply $\phi'(0) = \phi'(1) = 0$. Substituting into (3.16), the first condition implies that $a = 0$ and the second that $\sin(1/\lambda) = 0$. Consequently the eigenvalues and eigenfunctions are given by
\[ \lambda_n = 1/n\pi, \quad \phi_n(s) = \sqrt{2} \cos n\pi s, \quad n = 1, 2, \ldots. \]

Equation (3.15) arises in traditional mathematical physics by separation of variables in the ‘vibrating string’ equation, e.g. Courant & Hilbert (1953, Sec. 5.3). The boundary condition $\phi'(1) = 0$ corresponding to the right end of the string being “free”. In the case of the ordinary Wiener process ($\sigma = 0$), the left hand boundary condition becomes $\phi(0) = 0$, corresponding to the left end of the string being fixed at 0 – recall that $W(0) = 0$ almost surely. The condition for general $\sigma$, $\sigma^2 \phi(0) = \phi'(0)$ corresponds to an ‘elastically attached’ endpoint.

Table 1 shows the eigenvalues $\lambda_n$ and eigenfunctions corresponding to these various natural boundary conditions - all are easily derived from (3.16).

To describe the stochastic process, or “prior distribution” associated with periodic boundary conditions, recall that the Brownian Bridge $\tilde{W}(t) =$
### Table 1. Effect of Boundary Conditions for the vibrating string equation

<table>
<thead>
<tr>
<th>σ = ∞</th>
<th>Boundary Conditions</th>
<th>Eigenvalues</th>
<th>Eigenfunctions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ϕ′(0) = ϕ′(1) = 0</td>
<td>λ_{n-1} = nπ</td>
<td>√2 cos nπt</td>
</tr>
<tr>
<td>σ = 0</td>
<td>ϕ(0) = ϕ(1) = 0</td>
<td>λ_{n-1} = (n + \frac{1}{2})π</td>
<td>√2 sin(n + \frac{1}{2})πt</td>
</tr>
</tbody>
</table>
| 0 < σ < ∞ | \(ϕ′(0) = σ^{-2}ϕ(0),\) \(ϕ′(1) = 0\) | λ_{n-1} ∈ \(nπ, (n + \frac{1}{2})π\) | \(c_n, \sin λ_{n-1} t + \ldots\) c_nσ^{-2}λ_{n-1}^2 cos λ_{n-1}t, 
|       | ϕ(0) = ϕ(1),       | λ_{2n-1} = λ_{2n-1} = 2nπ | √2 sin 2πnt, |
|       | ϕ′′(0) = ϕ′′(1)     |              | √2 cos 2πnt    |

W(t) − tW(1) satisfies \(\ddot{W}(1) = \ddot{W}(0) = 0\) and has \(\text{Cov}(\dot{W}(s), \dot{W}(t)) = s ∧ t − st\). Proceeding as before, define a “free” Brownian Bridge

\[ \ddot{Z}_\sigma(t) = σ_0 + W(t) , \]

and verify that it has covariance kernel

\[ \dddot{R}_\sigma(s, t) = σ^2 + s ∧ t − st . \]

Equations (3.13) and (3.14) change in an obvious way, but the differential equation (3.15) remains the same. The boundary conditions become

\[ \phi(0) = φ(1), \quad \phi′(0) = σ^{-2}\phi(0) + φ′(1), \]

and so the standard periodic boundary conditions and the usual sine and cosine eigenfunctions emerge from the \(σ \to ∞\) limit.

In all cases summarized in Table 1, the eigenfunctions show increasing oscillation with increasing \(n\), as measured by sign crossings, or frequency. This is a general phenomenon for such boundary value problems for second order differential equations (Sturm oscillation theorem — see e.g. Birkhoff & Rota (1969, Sec 10.7)). Note also that in the periodic case, the eigenvalues have multiplicity two — both sines and cosines of the given frequency — but in all cases the asymptotic behavior of the eigenvalues is the same: \(λ_{n-1} \sim nπ\).

The analysis of the integrated Wiener prior (3.12), corresponding to cubic smoothing splines, then proceeds along the same lines, with most details given in Exercise ?? (see also Freedman (1999, Sec. 3)). The eigenvalue equation is a fourth order differential equation:

\[ \dddot{ϕ}(s) = λ^2 \dddot{ϕ}(s) . \]

This equation is associated with the vibrating rod (Courant & Hilbert 1953, Secs IV.10.2 and V.4) — indeed, the roughness penalty \(\int f''^2\) corresponds to the potential energy of deformation of the rod. It is treated analogously to the vibrating string equation. In particular, the (four!) boundary conditions for the \(σ = ∞\) limit become

\[ \ddot{ϕ}(0) = \ddot{ϕ}(0) = 0, \quad \ddot{ϕ}(1) = \ddot{ϕ}(1) = 0, \]

corresponding to “free ends” at both limits.
3.4. Expected Errors of Estimation

Let \( L(a, \theta) \) be a loss function on \( \mathcal{A} \times \ell_2(I) \). Usually \( L(a, \theta) = w(\|a - \theta\|) \) for an appropriate function \( w \) of some norm \( \|\cdot\| \). Indeed, most often, though not exclusively, we use the squared error loss \( \|a - \theta\|^2 \). This choice cases both mathematical arguments and geometric interpretations. Other error measures can be statistically important, and so sometimes we retain the extra generality of the \( L(a, \theta) \) formulation.

The quality of an estimator \( \hat{\theta}(y) \) of \( \theta \) is measured by \( L(\hat{\theta}(y), \theta) \), and most often through its expectation for fixed \( \theta \); the risk function is denoted by

\[
r(\hat{\theta}, \theta) = E_\theta L(\hat{\theta}(y), \theta).
\]

Suppose that \( \theta \) is restricted to lie in a parameter space \( \Theta \subset \ell_2 \) and compare estimators through their worse case risk over \( \Theta \). Thus a particular importance attaches to the best possible worst-case risk, called the minimax risk over \( \Theta \):

\[
R_N(\Theta) = R_N(\Theta, \epsilon) = \inf_{\theta} \sup_{\hat{\theta}} E_\theta L(\hat{\theta}(y), \theta).
\]

The subscript “\( N \)” is a mnemonic for “non-linear” estimators, to emphasise that no restriction is placed on the class of estimators \( \theta \). One is often interested also in the minimax risk when the estimators are restricted to a particular class \( \mathcal{E} \) defined by a property such as linearity. In such cases, we write \( R_\mathcal{E} \) for the \( \mathcal{E} \)-minimax risk, under the assumption that the infimum in (3.17) is taken only over estimators in \( \mathcal{E} \). Note also that we will often drop explicit reference to the noise level \( \epsilon \).

Turning to squared error loss, a classical result of parametric statistics (Ref?) states that

\[
R_N(\mathbb{R}^n, \epsilon) = n \epsilon^2.
\]

Since \( \mathbb{R}^n \subset \ell_2(\mathbb{N}) \) for each \( n \), it is apparent that \( R_N(\ell_2(\mathbb{N}), \epsilon) = \infty \), and in particular for any estimator \( \hat{\theta} \)

\[
\sup_{\theta \in \ell_2(\mathbb{N})} E_\theta \|\hat{\theta} - \theta\|^2 = \infty.
\]

Thus, a fundamental feature of non-parametric estimation is that some a priori restriction on the class of signals \( \theta \) is required in order to make meaningful comparisons of estimators. Fortunately, a large class of such classes is available:

3.1. Lemma. If \( \Theta \) is compact in \( \ell_2 \), then for \( \ell_2 \) error, \( R_N(\Theta, \epsilon) < \infty \).

Proof. Just consider the zero estimator \( \hat{\theta}_0 \equiv 0 \); then \( \theta \to r(\hat{\theta}_0, \theta) = \|\theta\|^2 \) is continuous on the compact \( \Theta \) and so \( R_N(\Theta) \leq \sup_\theta r(\hat{\theta}_0, \theta) < \infty \). \( \Box \)

3.2. Example. Some \( \ell_2 \) compact sets that will appear frequently later include:
• An ellipsoid \( \sum a_k^2 \theta_k^2 \leq C^2 \) is \( \ell_2 \)-compact if and only if \( a_k > 0 \) and \( a_k \to \infty \).

• A hyperrectangle \( \prod [-\tau_k, \tau_k] \) is \( \ell_2 \)-compact if and only if \( \sum \tau_k^2 < \infty \).

In fact, Lemma 3.1 extends to sets of direct product form \( \Theta = \mathbb{R}^r \times \Theta' \), where \( r < \infty \) and \( \Theta' \) is compact. (Exercise??). The argument of the Lemma can also be extended to show that \( R_N(\Theta, \epsilon) < \infty \) if \( L(a, \theta) = w(\|a - \theta\|) \) with \( w \) continuous and \( \Theta \) being \( \| \cdot \| \)-compact. At the same time, of course, compactness is not necessary for finiteness of the minimax risk, as (3.18) shows.

3.3. Example. Ellipsoids and mean-square smoothness. Consider the continuous form of the Gaussian white noise model (3.4). For integer \( m \geq 1 \), let

\[
F = F(m, L) = \{ f \in L_2[0, 1] : \int_0^1 [f^{(m)}(t)]^2 dt \leq L^2 \}.
\]

Historically, considerable interest has focused on the behavior of the minimax estimation risk

\[
R_N(F, \epsilon) = \inf_{f \in F} \sup_{f \in F} \int_0^1 [\hat{f} - f]^2.
\]

in the low noise limit as \( \epsilon \to 0 \). For example, what is the dependence on the parameters describing \( F \); namely \( (m, L) \)? Can one describe minimax estimators, and in turn, how do they depend on \( (m, L, \epsilon) \)?

One convenient way to express this in the sequence model is to use the orthonormal trigonometric basis for \([0, 1]\\):

\[
\varphi_0(t) \equiv 1, \quad \begin{cases}
\varphi_{2k-1}(t) = \sqrt{2} \sin 2\pi kt & k = 1, 2, \ldots \\
\varphi_{2k}(t) = \sqrt{2} \cos 2\pi kt.
\end{cases}
\]

Differentiation is diagonalised in \( 2 \times 2 \) blocks in the Fourier basis:

\[
D^m \begin{bmatrix} \varphi_{2k-1} \\ \varphi_{2k} \end{bmatrix} = \pi^m (2k)^m P^m \begin{bmatrix} \varphi_{2k-1} \\ \varphi_{2k} \end{bmatrix}
\]

where

\[
P = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.
\]

Thus, if \( f = \sum \theta_k \varphi_k \), then \( \|f^{(m)}\|^2 = \pi^{2m} \sum a_k^2 \theta_k^2 \), where

\[
a_0 = 0, \quad a_{2k-1} = a_{2k} = (2k)^m.
\]

Thus \( f \) lies in the functional class \( F \) if and only if the coefficient vector \( \theta \) belongs to the ellipsoid:

\[
\Theta(a, C) = \{ \theta : \sum_0^{\infty} a_k^2 \theta_k^2 \leq C^2 \},
\]

with the calibration \( C^2 = L^2 / \pi^{2m} \). Together with (3.8), this shows that the functional minimax risk problem (3.20) is equivalent to a sequence space problem (3.17) under squared \( \ell_2 \) loss. A complete solution to this problem was given by Pinsker (1980). We shall give Pinsker’s solution in Chapter.
8 as an illustration of tools that we will use for other parameter sets Θ in later chapters.

Remarks. 1. The ellipsoid representation (3.22)-(3.23) of mean-square smoothness extends to non-integer degrees of smoothness. Thus we will sometimes write Θ^θ(C) for the ellipsoid (3.23) when (3.22) holds, or more simply just a_k = k^α. Finiteness of \( \sum k^{2a} \hat{\theta}_k^2 \) can then be taken as a definition of finiteness of the Sobolev seminorm \( \|D^\alpha f\|_2 \) even for non-integer α. Appendix B contains further details and references.

2. (3.22) shows that Θ(a, C) is actually not compact, for the trivial reason that \( a_0 = 0 \). However it does equal \( \mathbb{R} \times \Theta' \) for Θ compact.

3.5. Example: Linear Estimators with Squared Error Loss

Linear estimators are simple and widely used, and are a natural starting point for theoretical study. In practice they may take on various guises: kernel averages, local polynomial fits, spline smoothers, orthogonal series, Wiener filters and so forth. However in the sequence model, all such linear estimators can be written in the form \( \hat{\theta}_C(y) = Cy \) for some matrix C, which when \( I = \mathbb{N} \) has countably many rows and columns. Let us evaluate the mean squared error of \( \hat{\theta}_C \). First, we have a stochastic decomposition of the error

\[ \hat{\theta}_C - \theta = Cz + (C - I)\theta \]

into mean zero stochastic and deterministic bias terms. The stochastic term

\[ E\|Cz\|^2 = E\operatorname{tr}\{C'Czz'\} = c^2 \operatorname{tr} C'C. \]

Consequently, the mean squared error becomes

\[ r(C, \theta) = c^2 \operatorname{tr} C'C + \theta'(C - I)'(C - I)\theta. \]

In particular, for \( r(C, \theta) \) to be finite, C needs to have finite Hilbert-Schmidt norm

\[ \|C\|_{HS}^2 = \operatorname{tr} C'C = \sum_{i,j=1}^\infty c_{ij}^2 < \infty. \]

Thus, C must be a bounded linear operator on \( \ell_2 \) with square summable singular values. In particular, in the infinite sequence case, \( C = I \) must be excluded, and so the bias term is necessarily unbounded over all of \( \ell_2 \):

\[ \sup_{\theta \in \ell_2} r(C, \theta) = \infty, \]

as is expected anyway from the general result (3.19).

3.5.1. Linear Estimators must shrink. Familiar smoothing methods such as the Wiener filter and smoothing splines are linear shrinkers except possibly for a low dimensional subspace on which no shrinkage is done. In fact, all reasonable linear estimators must shrink in all but at most two co-ordinates in some orthonormal basis. To describe this result, we recall that an estimator \( \hat{\theta} \) is inadmissible if there is another estimator \( \hat{\theta}_1 \) which
is better, in the sense that \( r(\hat{\theta}, \theta) \leq r(\hat{\theta}, \theta) \) for all \( \theta \), with strict inequality occurring for some \( \theta \).

3.4. **Theorem.** The linear estimate \( \hat{\theta} \) of \( \theta \) is inadmissible (for squared error) unless \( C \)

(a) is Hilbert-Schmidt: \( \text{tr} C' C < \infty \),
(b) is symmetric: \( C' = C \),
(c) has eigenvalues satisfying \( 0 \leq \lambda_i(C) \leq 1 \), and
(d) has at most two \( \lambda_i(C) = 1 \).

This result is due to Cohen (1966) for the finite sequence model, and to Mandelbaum (1984) for the countable case. Both authors also show that conditions (a)-(d) suffice for the linear estimate to be admissible, thus characterizing the class of admissible linear rules.

**Proof.** Given a matrix \( C \), the first step of the proof is to construct a new matrix \( D \) satisfying (a)-(c) that is as good as \( C \), and better if \( C \) fails to satisfy at least one of these properties. The absolute value \( |I - C| = [(I - C)'(I - C)]^{1/2} \) is positive and self-adjoint, and as coordinates we use\(^1 \) an orthonormal basis \( \{e_i\} \) in which \( |I - C| \) is diagonalised:

\[
|I - C| e_i = (1 - \mu_i) e_i, \quad \mu_i = \max\{\mu_i, 0\},
\]

Defining the modified operator

\[
|I - C|_+ e_i = (1 - \mu_i) e_i, \quad \mu_i = \max\{\mu_i, 0\},
\]

which forces the eigenvalues into \([0,1]\), we will take

\[
D = I - |I - C|_+.
\]

Thus, in the basis \( \{e_i\} \), we have \( D e_i = \mu_i e_i \) so that \( \lambda_i(D) \in [0,1] \).

To show that \( D \) improves the bias term, note that \( |1 - \mu_i| \leq |1 - \mu_i| \), so that

\[
\|(I - D)\theta\|^2 = \| |I - C|_+ \theta \|^2 \leq \| I - C \theta \|^2 = \|(I - C)\theta\|^2,
\]

where the final equality uses the definition of absolute value.

To show that \( D \) also improves the variance term, use the polar decomposition \( I - C = U(I - C) \) where \( U \) is an isometry on the range of \( I - C \), which by (3.26) includes those \( e_i \) for which \( \mu_i \leq 1 \). For such \( i \),

\[
(I - C)e_i = (1 - \mu_i)Ue_i,
\]

and, since \( \|Ue_i\| \leq 1 \),

\[
\langle e_i, (I - C)e_i \rangle = (1 - \mu_i) \langle e_i, Ue_i \rangle \leq 1 - \mu_i = \langle e_i, |I - C|e_i \rangle,
\]

\(^1I + |I - C|\) has a bounded inverse, and the formula

\[
I - |I - C| = [C' + C - C'C][I + |I - C|]^{-1}
\]

shows that \( I - |I - C| \) is Hilbert-Schmidt, and hence compact (and self-adjoint). The existence of the orthonormal basis then follows from the Hilbert-Schmidt theorem.
which implies that
\[ 0 \leq \langle e_i, De_i \rangle \leq \langle e_i, Ce_i \rangle. \]
Finally, since \( \langle e_i, De_i \rangle > 0 \) only when \( \mu_i > 0 \),
\[ \|D\|_{HS}^2 = \sum_i \langle e_i, De_i \rangle^2 = \sum_{i \mu_i > 0} \langle e_i, De_i \rangle^2 \leq \sum_i \langle e_i, Ce_i \rangle^2 \leq \|C\|_{HS}^2. \]
Chasing back the inequalities in the above proof shows that \( D \) improves strictly on \( C \) if any of (a) – (c) fail for \( C \). For example, equality of the variance terms would imply \( \langle Ce_i, e_j \rangle = 0 \) for \( j \neq i \), so that \( Ce_i = \gamma_i e_i \) and so in particular \( C \) must be symmetric.

For the final part of the argument, we use the James-Stein estimator. Suppose that \( C \) satisfies (a) – (c), and decompose \( C = U\Lambda U' \), with \( U \) orthogonal and \( \Lambda = (\text{diag } \lambda_i) \), with \( \lambda_1 = \ldots = \lambda_d = 1 > \lambda_i \) for all \( i > d \geq 3 \). In the new variables \( \eta = U\theta \) and \( x = U'y \), we have
\[ r(C, \theta) = r(\Lambda, \eta) = \sum_i r(\lambda_i, \eta_i) = \sum_i \varepsilon^2 \lambda_i^2 + (1 - \lambda_i)^2 \eta_i^2. \]
Set \( x^d = (x_1, \ldots, x_d) \). For \( d \geq 3 \), the James-Stein positive part estimator,
\[ \hat{\eta}^{JS}(x^d) = \left(1 - \frac{(d - 2)\varepsilon^2}{\|x^d\|^2}\right)x^d, \]
dominates the MSE of the unbiased estimator \( \hat{\eta}(x^d) = x^d \):
\[ r(\hat{\eta}^{JS}, \eta^d) = E \sum_{i=1}^d (\hat{\eta}_{i}^{JS} - \eta_i)^2 < d\varepsilon^2, \quad \eta^d \in \mathbb{R}^d. \]
Hence, the estimator \( \hat{\eta} \) that uses James-Stein in the first \( d \) coordinates and \( \lambda_i x_i \) in the remaining ones dominates \( Ax \):
\[ r(\hat{\eta}, \eta) = r(\hat{\eta}^{JS}, \eta^d) + \sum_{i>d} r(\lambda_i, \eta_i) < r(\Lambda, \eta). \]

[Remark on Cohen’s proof and \( \text{tr} (I-C) = \infty \)? Remark on approach via finite rank approximation?]

### 3.6. Models with Gaussian limits*

Since the earliest days of nonparametric function estimation, striking similarities in large sample results – rates of convergence, distributional structure – have been observed in models as diverse as spectrum estimation, density estimation and nonparametric regression. In recent years, a rigorous expression of this phenomenon has been obtained used LeCam’s notion of asymptotic equivalence of experiments. In each case, a result exists stating that under certain regularity conditions on the unknown function \( f \), in large samples, the model is asymptotically equivalent to the signal in Gaussian white noise model. Informally, this means that conclusions based on estimators, risk functions and asymptotic analysis in the white noise model can
be carried over to corresponding estimators and risks in the other model sequence. The result of Brown & Low (1996) for Gaussian nonparametric regression is of this type, as is a growing list of examples due to Michael Nussbaum and collaborators:

Density Estimation. Suppose that $X_1, \ldots, X_n$ are drawn i.i.d from an unknown density $f$ supported on $[0, 1]$. So long as $f$ has Hölder smoothness greater than $1/2$, the experiment is asymptotically equivalent to

$$dY_t = f^{1/2}(t)dt + \frac{1}{\sqrt{n}} dW_t, \quad 0 \leq t \leq 1.$$  

Nussbaum (1996). The appearance of the root density $f^{1/2}$ is connected with the square root variance stabilizing transformation for Poisson data, which is designed to lead to the constant variance term. Note also that $f^{1/2}$ is square integrable with $L_2$ norm equal to 1!

Here is a heuristic argument, in the spirit of (3.9), that leads to (3.27). Divide the unit interval into $m_n = o(n)$ equal intervals of width $h_n = 1/m_n$. Assume also that $m_n \to \infty$ so that $h_n \to 0$. Write $I_{kn}$ for the $k$th such interval, which at stage $n$ extends from $t_k = k/m_n$ to $t_{k+1}$. First the ‘Poissonization trick’: draw a random number $N_n$ of observations $X_1, \ldots, X_{N_n}$ i.i.d. from $f$, with $N_n \sim \text{Poisson}(n)$. Then, because of the Poisson thinning property, the number of observations falling in the $k$th bin $I_{kn}$ will be Poisson with mean $n \int_{I_{kn}} f = n f(t_k) h_n$. The square root transformation is variance stabilizing for the Poisson family and so

$$y_{kn} := \sqrt{N_n(I_{kn})} \sim N(\sqrt{f(t_k) n} h_n, 1/4) \text{ approximately for large } n.$$  

Thus

$$y_k \approx \sqrt{f(t_k)} \sqrt{n h_n + \frac{1}{4} e_{kn}} \text{ with } e_{kn} \text{ independent and approximately standard Gaussian.}$$

Now form a partial sum process as in (3.9), and premultiply by $\sqrt{h_n/n}$ to obtain

$$Y_n(t) = h_n^{1/2} n^{-1/2} \sum_{1}^{[m_n t]} y_{kn} \approx \sum_{1}^{[m_n t]} \sqrt{f(t_k) h_n} + (1/2) n^{-1/2} m_n^{-1/2} \sum_{1}^{[m_n t]} e_{kn}.$$  

This makes it plausible that the process $Y_n(t)$, based on the density estimation model, merges in large samples with the Gaussian white noise process of (3.27). A rigorous version of the argument is given in Klemela & Nussbaum (1998), along with an explicit recipe for transferring estimates in the white noise model to the density setting.

Spectral density estimation Suppose that $X_1, \ldots, X_n$ is a sample from a stationary Gaussian random process with mean zero and spectral density function $f(\lambda)$ on $[-\pi, \pi]$, related to the covariance function $r(k) = EX_j X_{j+k}$ via $f(\lambda) = (2\pi)^{-1} \int_{-\infty}^{\infty} e^{i \lambda k} r(k)$. Estimation of the spectral density $f$ was the first nonparametric function estimation model to be studied asymptotically – see for example Grenander & Rosenblatt (1957). Let $N_n = o(n^{1/2})$ be a sequence increasing at a rate controlled by the assumed smoothness of $f$, and let $\lambda_k = 2\pi k / N_n$. Then the spectrum estimation experiment is locally
asymptotically equivalent to the homoscedastic regression model (3.3)

\[ Y_k = \log f(\lambda_k) + \sigma_n e_k \quad k = 1, \ldots, N_n, \]

where \( \sigma_n^2 = N_n/n \) and \( e_k \) are i.i.d. \( \mathcal{N}(0, 1) \). See Gohbev & Nussbaum (1998). Again the logarithmic transformation is variance stabilizing for spectrum estimation, since the periodogram ordinates have asymptotically independent exponential distributions.

**Nonparametric Generalized Linear Models.** This is an extension of model (3.3) to errors drawn from an exponential family. Indeed count data with time varying Poisson intensities and dichotomous or categorical valued series with time varying cell probabilities occur naturally in practice (e.g. Kozaczyk (1997), Stoffer (1991)). We suppose that the densities in the family may be written \( P_\theta(dx) = p_\theta(x) \nu(dx) \) with \( p_\theta(x) = e^{\theta U(x) - \psi(\theta)} \). Thus \( \theta \) is the canonical parameter, \( U(x) \) the sufficient statistic, \( \nu(dx) \) the dominating measure on \( \mathbb{R} \) and \( \psi(\theta) = \log \int e^{\theta U(x)} \nu(dx) \) the cumulant generating function. (reference?) All the standard examples – Poisson, Bernoulli, Gaussian mean, Gaussian variance, exponential – are included. The mean value parameterization is given by \( \mu(\theta) = \psi'(\theta) = E_p U(X) \), and \( \psi''(\theta) = \text{Var}_p U(X) \).

Let \( t_i = i/n, i = 1, \ldots, n \) and \( f \) be a sufficiently smooth function, typically with Hölder smoothness greater than 1/2. It is assumed that we have observations \( (t_i, X_i) \) in which \( X_i \) is drawn from \( P_{\theta_i}(dx) \) where \( \theta_i = f(t_i) \). Then this experiment is equivalent to

\[ dY_i = \Gamma(f(t))dt + n^{-1/2}dW_t \quad 0 \leq t \leq 1, \]

where \( \Gamma(\theta) = V \circ \mu(\theta) \) and \( V(\mu) \) is the variance stabilizing transformation for \( \{P_\theta\} \) defined through \( V'(\mu(\theta)) = 1/\sqrt{\psi''(\theta)} \) (Gram & Nussbaum 1998).

Some cautions are in order when interpreting these results. First, there are significant regularity conditions, for example concerning the smoothness of the unknown \( f \). Meaningful error measures for spectral densities may not translate into, say, squared error loss in the Gaussian sequence model. Nevertheless, the asymptotic equivalence results lend further strength to the idea that the Gaussian sequence model is the fundamental setting for nonparametric function estimation, and that theoretical insights won there will have informative analogs in the more concrete practical problems of curve estimation.

### 3.7. Problems

1. *(Compactness criteria.*) Here \( \ell_2 \) denotes square summable sequences with the norm \( \| \theta \|^2 = \sum \theta_k^2 \).

   (a) The ellipsoid \( \Theta = \{ \theta : \sum_{k \geq 1} a_k^2 \theta_k^2 \leq C^2 \} \) is \( \ell_2 \)-compact if and only if \( a_k > 0 \) and \( a_k \to \infty \).

   (b) The hyperrectangle \( \Theta = \prod_{k \geq 1} [-\tau_k, \tau_k] \) is \( \ell_2 \)-compact if and only if \( \sum_{k \geq 1} \tau_k^2 < \infty \).
2. (Equivalence of measures.) Let $P$ and $Q$ be probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$, absolutely continuous with respect to a probability measure $\lambda$. (For example, $\lambda = (P+Q)/2$.) Write $p = dP/d\lambda$ and $q = dQ/d\lambda$. The Hellinger affinity $h(P, Q) = \int \sqrt{pq}d\lambda$ does not depend on the choice of $\lambda$. Let $\{P_n\}$ and $\{Q_n\}$ be two sequences of probability measures on $\mathbb{R}$. Define product measures on sequence space $\mathbb{R}^\infty$, with the product Borel $\sigma$-field, by $P = \prod P_n$ and $Q = \prod Q_n$. Then the celebrated theorem of Kakutani (1948) states that if $P_n \sim Q_n$ for $n = 1, 2, \ldots$ then $P$ and $Q$ are either equivalent or orthogonal. Moreover, $P \sim Q$ if and only if $\prod_{k=1}^\infty h(P_k, Q_k) > 0$. In case $P \sim Q$, $dP/dQ = \prod_{k=1}^\infty dP_k/dQ_k$.

(i) Taking all this as given, show first that

$$h(N(\theta_1, \sigma_1^2), N(\theta_2, \sigma_2^2)) = \left(\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}\right)^{1/2} \exp\left\{ -\frac{(\theta_1 - \theta_2)^2}{4(\sigma_1^2 + \sigma_2^2)} \right\}.$$

(ii) Suppose that $z_i$ are i.i.d $N(0, 1)$ for $i = 1, 2, \ldots$. Show that the measure $P_\theta$ corresponding to $y_i = \theta_i + \lambda_i z_i$ is absolutely continuous with respect to $P_0$ if and only if $\sum \theta_i^2/\lambda_i^2 < \infty$ and write down the likelihood ratio. [WHAT IF $\theta = 0$?]

(iii) In the Gaussian sequence model $y_k = \theta_k + \epsilon_k$, consider priors $\theta_k \sim N(0, \tau_k^2)$, independently with $\tau_k^2 = b k^{-2m}$. Under what conditions on $m$ is the marginal distribution $P_n(dy)$ equivalent to $P_0(dy)$, the distribution conditional on $\theta = 0$?
CHAPTER 4

Interlude: Spline Smoothing

Spline smoothing has become a popular technique in nonparametric regression, and serves as an important example of linear estimation in the Gaussian white noise model. Through the use of a particular orthonormal basis (Demmler-Reinsch) spline smoothing can be understood as a diagonal linear shrinkage method, even for unequally spaced regression designs. With an equally spaced design and Gaussian noise, the use of periodic splines allows a similar but more concrete analysis in the Gaussian sequence model. In particular, it is easy to derive an exact formula for the equivalent kernel in the large $n$, or small noise, limit. This discussion is a first illustration of how the Gaussian sequence model can provide concrete formulas for the “limiting objects” which strengthen understanding of similar finite sample settings. Much more information on spline theory, methods and applications may be found in the books by Wahba (1990), Hastie & Tibshirani (1990) and Green & Silverman (1994).

4.1. Cubic Splines and Periodic Splines

The most usual form of the spline smoothing problem begins with data $(t_i, Y_i)$, $i = 1, \ldots, n$. Suppose, for now, that the knots $t_i$ satisfy $0 = t_0 < t_1 < \ldots < t_n < t_{n+1} = 1$. We seek a function $f : [0,1] \rightarrow \mathbb{R}$ minimizing

$$Q(f) = n^{-1} \sum_{i=1}^{n} |Y_i - f(t_i)|^2 + \lambda \int_{0}^{1} f''^2.$$

The regularization parameter $\lambda$ trades off the first, data fidelity term, with the second, the roughness penalty. As $\lambda$ increases from 0 to $\infty$, the solution will pass from rough (interpolating the data) to smooth (the linear least squares fit). [Remark about $\int(D^m f)^2$].

A cubic spline is a function on $[0,1]$ that is (a) a cubic polynomial on each interval $[t_i, t_{i+1}]$ for $i = 0, 1, \ldots, n$, and (b) is continuous along with its first and second derivatives at the knots $t_i$. It is called a natural cubic spline if, in addition, it is linear on the endpoint intervals containing 0 and 1. Let $S$ denote the linear space of natural cubic splines – by counting coefficients and subtracting the constraints imposed by smoothness at the knots, one sees that the dimension of $S$ equals $n$, the amount of data. Note that the knot locations $t_i$ are held fixed.
A basic property of natural cubic splines (e.g. Green & Silverman (1994, pp. 15-16)) is that for any given data \( Y \in \mathbb{R}^n \), there exists a unique interpolating natural cubic spline \( g \in S \):

\[
g(t_i) = Y_i, \quad i = 1, \ldots, n.
\]

Further, any function \( \tilde{g} \in W_2^2[0,1] \) that interpolates the data (4.1) has roughness at least as large:

\[
\int g''^2 \leq \int \tilde{g}''^2.
\]

As a corollary, a minimizer of \( Q(f) \) may be found within \( S \): given any function \( f \), let \( g_f \) be the natural spline interpolating \( \{f(t_i)\} \). The data fidelity term is clearly the same for \( f \) and \( g_f \), whereas \( f \) improves roughness by (4.2).

These remarks reduce the spline smoothing problem to a finite dimensional question, and indeed it may be solved, for fixed \( \lambda \), using \( O(n) \) algorithms based, for example, on the well known \( B \)-splines of local support.

Associated with a function \( f \) on \([0,1]\) is the vector \( f = (f(t_1), \ldots, f(t_n))' \) of samples. For this section,

\[
\langle f, g \rangle = n^{-1} \sum_{i=1}^{n} f(t_i)g(t_i)
\]

denotes the Euclidean inner product on \( \mathbb{R}^n \), with corresponding norm \( \| \cdot \| \).

Demmler & Reinsch (1975) constructed a basis \{\( \varphi_1, \ldots, \varphi_n \)\} for \( S \) with a useful double orthogonality property:

\[
\langle \varphi_j, \varphi_k \rangle = \delta_{jk},
\]

\[
\int \varphi_j'' \varphi_k'' = w_k \delta_{jk}.
\]

The weights \( w_k \) are non-negative and increasing, indeed \( w_1 = w_2 = 0 \), so that the first two basis functions are linear. For \( k \geq 3 \), Demmler-Reinsch show that \( \varphi_k \) has \( k-1 \) sign changes, so that the basis functions exhibit increasing oscillation with \( k \), and this is reflected in the values \( w_k \) for the roughness penalty. Because of this increasing oscillation with \( k \), we may think of \( k \) as a frequency index, and the Demmler-Reinsch functions as forming a quasi-Fourier basis (which, of course, depends on the knot locations \( \{t_i\} \)).

The spline smoothing problem is diagonalized by the Demmler-Reinsch basis and has an explicit solution. Represent an arbitrary cubic spline \( f \) as \( f(t) = \sum \theta_k \varphi_k(t) \). At the knots, it follows that \( f = \sum \theta_k \varphi_k \), with equality as vectors in \( \mathbb{R}^n \). The corresponding “quasi-Fourier” coefficients of the data are given by \( Y = \sum y_k \varphi_k \). Applying now both parts of (4.4), the quadratic criterion may be rewritten

\[
(4.5) \quad Q(f) = \tilde{Q}(\theta) = \sum_{k=1}^{n} (y_k - \theta_k)^2 + \lambda \sum_{k=1}^{n} w_k \theta_k^2.
\]
The quadratic can be minimized term by term to yield the quasi-Fourier form of the smoothing spline estimate:

\[
\hat{\theta}_{SS,k} = c_{\lambda k} y_k, \quad c_{\lambda k} = \frac{1}{1 + \lambda w_k}.
\]

In the original time domain,

\[
\hat{\mathbf{f}} = \sum_k \hat{\theta}_{SS,k} \varphi_k = \sum_k c_{\lambda k} y_k \varphi_k.
\]

There is no shrinkage on the constant and linear terms: \(a_{\lambda 1} = a_{\lambda 2} = 1\), but for \(k \geq 3\), the shrinkage factor \(a_{\lambda k} < 1\) and decreases with increasing frequency. Large values of smoothing parameter \(\lambda\) lead to greater attenuation of the data, and hence greater smoothing in the estimate.

To represent the solution in terms of the original data, gather the basis functions into an \(n \times n\) orthogonal matrix \(U = [\varphi_1, \ldots, \varphi_n]/\sqrt{n}\). Then \(Y = U^T \hat{\mathbf{f}} \) and \(f = \sqrt{n} U \hat{\mathbf{f}}\), and so

\[
\hat{\mathbf{f}} = \sqrt{n} U \hat{\theta} = U a_{\lambda} U^T \mathbf{y} = A_{\lambda} \mathbf{Y}. \quad a_{\lambda} = \text{diag} (a_{\lambda k}).
\]

Notice that the change of basis matrix \(U\) does not depend on \(\lambda\). Thus, many important aspects of the spline smoothing problem, such as the issue of choosing \(\lambda\) well from data, can be studied in the diagonal sequence form that the quasi-Fourier basis provides.

Software packages, such as \texttt{spline.smooth} in \texttt{S-PLUS}, may use other bases, such as \(B\)-splines, to actually compute the spline estimate. However, because there is a unique solution to the optimization problem, the estimate computed in practice must coincide, up to numerical error, with (4.8).

If the data \((t_i, y_i)\) are assumed to come from model (3.3), then in the Demmler-Reinsch coordinates, \(\mathbf{y} = \theta + \varepsilon \mathbf{z}\) is an instance of the Gaussian sequence model with \(\varepsilon = \sigma / \sqrt{n}\). Efron (1999, 2000) makes an extensive study of the geometric aspects of smoothing parameter choice in this finite model.

**Periodic Splines.** If the points \(t_i = i/n\), for \(i = 0, \ldots, n - 1\) are equally spaced and the unknown function \(f\) is periodic on \([0, 1]\), then a more explicit diagonalization and solution of the spline smoothing problem is available in the Fourier basis itself. Because data is typically real valued, we work with the sine and cosine form of the trigonometric functions:

\[
\varphi_0(t) \equiv 1, \quad \left\{ \begin{array}{ll}
\varphi_{2k-1}(t) = \sqrt{2} \sin 2\pi kt, \\
\varphi_{2k}(t) = \sqrt{2} \cos 2\pi kt.
\end{array} \right. \quad k = 1, 2, \ldots
\]

For convenience in notation, we will consider only \(n = 2m + 1\) odd. Let \(S\) now denote the linear space of trigonometric polynomials of degree \(m\):

\[S = \{ f : f(t) = \sum_{k=0}^{n-1} c_k \varphi_k(t), t \in [0, 1] \} \]

The discrete sines and cosines will be \(\varphi_k = (\varphi_k(t_i))\), and the key point is that the double orthogonality relations (4.4) again hold, with now explicit
weights
\begin{equation}
w_{2k-1} = w_{2k} = (2\pi k)^4.
\end{equation}

Although we now work with trigonometric polynomials rather than cubic splines, much the same argument as before shows that the minimizer of \(Q(f)\) in \(S\) is given by (4.6), but now with explicit shrinkage at frequency \(k\) given by
\[c_{\lambda 2k-1} = c_{\lambda 2k} = [1 + \lambda (2\pi k)^4]^{-1}\]
for \(k \leq m\). Thus the periodic spline problem has many of the qualitative features of general spline smoothing, along with a completely explicit description.

Remark. It is not true that the minimizer of \(Q(f)\) over all functions lies in \(S\), as was the case with cubic splines. The problem lies with aliasing: the fact that when \(0 < r \leq n\) and \(l \in \mathbb{N}\), we have \(\varphi_r = \varphi_{r+2n}\) when restricted to \(t_1, \ldots, t_n\). [MAKE AN EXAMPLE!]

4.2. The Equivalent Kernel for Spline smoothing.

Spline smoothing also has an interpretation in terms of local averaging which is not so apparent from its regularized least-squares formulation. This point of view comes out quite directly using sequence models. With this aim, we jump between the finite sequence model (3.3), namely
\begin{equation}
Y_i = f(i/n) + \sigma e_i, \quad i = 1, \ldots, n
\end{equation}
and the infinite sequence model (3.4)-(3.6), namely
\begin{equation}
Y_t = \int_0^t f(s)ds + \epsilon W_t, \quad t \in [0, 1],
\end{equation}
\[\Leftrightarrow \quad y_k = \theta_k + \varepsilon z_k, \quad k \in \mathbb{N}\]
using the heuristics discussed around (3.9).

In the case of equally spaced data in the finite model, the Priestley-Chao version [INCLUDE A REFERENCE!] of kernel regression sets
\[\hat{f}_h(t) = n^{-1} \sum_{i=1}^{n} K_h(t - t_i)Y_i, \quad K_h(t) = h^{-1}K(h^{-1}t).\]

Here \(K(t)\) is an averaging kernel: \( \int K = 1 \) and is typically symmetric and non-negative. The window width, or bandwidth \(h\), controls the range of \(t_i\) over which the responses \(Y_i\) are averaged together to produce the estimate at \(t\). With the partial sum process notation (3.9), the estimate takes the integral form
\[\hat{f}_h(t) = \int_0^1 K_h(t - s)dY_n(s).\]

Jumping from the finite to the infinite model, we approximate this kernel estimate by its analog shown in the upper right corner of Table 1.
Table 1. The analogy between spline smoothing and regression goes via versions of each method in the infinite sequence model.

Similarly, the periodic spline estimate (4.7) in the finite model has a natural analogue in the infinite case. We define the smoothing spline estimate \( \hat{\theta}_\lambda \) in the infinite sequence model as the minimizer of

\[
\sum_{i=1}^{\infty} (y_k - \theta_k)^2 + \lambda \sum_{i=1}^{\infty} w_k \theta_k^2.
\]

In general, the weights \( w_k \) should be positive and increasing. Just as in the finite case, the estimate \( \hat{\theta}_\lambda \) has diagonal linear form,

\[
\hat{\theta}_\lambda(y) = c_{\lambda k} y_k = (1 + \lambda w_k)^{-1} y_k.
\]

In terms of functions, the spline estimate is given by the series in the lower corner of Table 1.

We can now derive the kernel representation of the infinite sequence spline estimate. Substituting (3.6), \( y_k = \int \varphi_k dY \), we get

\[
\hat{f}_\lambda(s) = \int C(s, t) dY(t), \quad C(s, t) = \sum_{k=0}^{\infty} c_{\lambda k} \varphi_k(s) \varphi_k(t).
\]

Now specialize to the explicit cubic weights for periodic splines in (4.10). Then \( c_{\lambda, 2k-1} = c_{\lambda, 2k} \), and from (4.9) and the addition formula for sines and cosines,

\[
\varphi_{2k-1}(s) \varphi_{2k-1}(t) + \varphi_{2k}(s) \varphi_{2k}(t) = 2 \cos 2\pi k (s - t).
\]

Hence the kernel \( C(s, t) \) has translation form \( K_\lambda(s - t) \), with formula

\[
K_\lambda(s) = 1 + \sum_{k=1}^{\infty} \frac{2 \cos 2\pi k s}{1 + \lambda(2\pi k)^2}.
\]

But we can describe \( K_\lambda \) more explicitly! First, a definition: a function \( f \) on \( \mathbb{R} \) can be made periodic with period 1 by wrapping: \( g(t) = \sum_{j \in \mathbb{Z}} f(t + j) \).

### 4.1. Theorem

The spline estimate with \( \lambda = h^4 \) has the kernel representation

\[
\hat{f}_\lambda(t) = \int_0^1 K_h(t - s) dY(s).
\]
Then $K_h(t)$ is the wrapped version of $L_h(t) = (1/h)L(t/h)$, and the equivalent kernel

$$L(t) = \frac{1}{2} e^{-|t|/\sqrt{2}} \sin\left(\frac{|t|}{\sqrt{2}} + \frac{\pi}{4}\right).$$

The kernel $L$ has exponential decay, and is essentially negligible for $|t| \geq 8$. [FIGURE??]. The wrapped kernel $K_h$ is therefore effectively identical with $L_h$ on $[-\frac{\pi}{2}, \frac{\pi}{2}]$ when $h$ is small: for example $h < 1/16$ will do.

**Proof.** We may write

$$K_h(s) = \sum_{k \in \mathbb{Z}} \frac{e^{-2\pi ik s}}{1 + (2\pi kh)^4}.$$

It follows directly from the Poisson summation formula that $K_h$ is the wrapped version of $L_h$, with Fourier transform given by

$$\hat{L}_h(\xi) = [1 + \xi^4 h^4]^{-1}.$$

The inversion of the transform (e.g. Erdélyi et al. (1954, (19), p.9) ) yields (4.14) - Exercise ?? outlines a direct derivation via contour integration. □

Thus in the infinite sequence model, periodic spline smoothing is identical with a particular kernel estimate. One may therefore interpret finite versions of periodic splines (and by analogy even B-spline estimates for unequally spaced data) as being approximately kernel smoothers. The approximation argument was made rigorous by Silverman (1984) (who also showed that for unequally spaced designs, the bandwidth $h$ varies with the fourth root of the design density.)

### 4.3. The Variance-Bias Lemma

The next calculation occurs so frequently that we do it here once and for all.

**4.2. Lemma (Variance-Bias).** The function $G(h) = vh^{-1} + bh^{2\beta}$, defined for $h \geq 0$ and positive constants $v, b$ and $\beta$, has minimizing value and location

$$G(h_s) = e^{H(r)b^{1-r}} v^-r, \quad h_s = r^{-1} e^{-H(r)(v/b)^{-r}}.$$  

The “rate” $r = 2\beta/(2\beta + 1)$, and $H(r) = -r \log r - (1 - r) \log(1 - r)$ is the binary entropy function.

---

1 Alternately, by successively differentiating (4.15), it is easily seen that

$$h^4 K_h^{(iv)} + K_h = \delta$$

where $\delta$ is the delta function. The solution of $h^4 L_h^{(iv)} + K_h = \delta$ on $\mathbb{R}$ may be found by Fourier transformation, and yields (4.30), and then this is converted into a solution of (4.36) by periodization.
For example, with kernel estimates based on a kernel $K$ of order $\beta$, $h$ can be thought of as a bandwidth and $v$ as a variance factor (such as $n^{-1}$ or $\sigma^2$), while $b$ is a bias factor (for example involving $k(K) \int (D^2 f)^2$).

The proof is straightforward calculus, though the combination of the two terms in $G(h)$ to yield the multiplier $e^{H(r)}$ is instructive: the variance and bias terms contribute in the ratio 1 to $(2\beta)^{-1}$ at the optimum, so that in the typical case $\beta > \frac{1}{2}$, the bias contribution is smaller than the optimum $h_\star$.

### 4.4. Spline Estimates over Sobolev Ellipsoids

So far we have said nothing about the mean squared error performance of the spline estimate, nor anything on the crucial question of how to choose the regularization parameter. These two issues are closely connected, and both depend on the smoothness of the function $f$ being estimated. Our strategy here is to select convenient parameter spaces $\Theta$, to evaluate the worst case MSE of $\hat{\theta}_\lambda$ over $\Theta$, and then to choose the value of $\lambda$ that minimizes this maximum error. This yields information on the rate of convergence of $\hat{\theta}_\lambda$ to $\theta$ as $\epsilon \to 0$; we shall see that such rates of convergence, although crude tools, already yield useful information about estimators.

**Maximum risk over ellipsoids.** A general diagonal linear estimator with components $\hat{\theta}_k = c_k y_k$ has variance-bias decomposition

$$r(\hat{\theta}_c, \theta) = \epsilon^2 \sum_k c_k^2 + \sum_k (1 - c_k)^2 \theta_k^2.$$ 

The worst case risk over $\Theta$ has a corresponding form

$$\bar{r}(\hat{\theta}_c; \epsilon) = \sup_{\theta \in \Theta} r(\hat{\theta}_c, \theta) = \bar{V}(\epsilon) + \bar{B}^2.$$ 

The max variance term $\bar{V}(\epsilon) = \epsilon^2 \sum_k c_k^2$ does not depend on $\Theta$. On the other hand, the max bias term does not depend on the noise level $\epsilon$. It does depend on $\Theta$, but can be easily evaluated on ellipsoids $\Theta(a, C) = \{\theta : \sum a_k^2 \theta_k^2 \leq C^2\}$:

$$\bar{B}^2 = C^2 \sup_k a_k^{-2} (1 - c_k)^2.$$ 

(4.18)

To see this, make new variables $s_k = a_k^2 \theta_k^2 / C^2$ and note that the linear function $\sum d_k s_k$ is maximized over the non-negative simplex $\sum s_k \leq 1$ by $\sup d_k$.

To summarize, over ellipsoids,

$$\bar{r}(\hat{\theta}_c; \epsilon) = \epsilon^2 \sum_k c_k^2 + C^2 \sup_k a_k^{-2} (1 - c_k)^2.$$ 

**Spline estimators for fixed $\lambda$.** We now specialize to shrinkage estimates

$$\hat{\theta}_{\lambda, k} = c_k y_k, \quad c_k = (1 + \lambda k^{2m})^{-1},$$

53
corresponding to roughness penalty \( \int (D^m f)^2 \), and to Sobolev ellipsoids

\[
\Theta_2^2(C) = \{ \theta : \sum k^{2\alpha} \theta_k^2 \leq C^2 \}.
\]

[Clarify that this is an approximation; e.g. in next display. Also discuss min\( k \) versus min\( \alpha \).]

4.3. PROPOSITION. The worst case mean squared error for an \( m \)-th order spline estimate \( \hat{\theta}_\lambda \) over a Sobolev ellipsoid \( \Theta_2^2(C) \) with \( \alpha < 2m \) is

\[
\tilde{\varphi}(\hat{\theta}_\lambda; \epsilon) = v_m \epsilon^2 \lambda^{-1/2m} + b_{\alpha m} C^2 \lambda^{\alpha/m},
\]

where the constants \( v_m \) and \( b_{\alpha m} \) are given below.

PROOF. For the variance term, use an integral approximation

\[
\tilde{V}(\epsilon) = \epsilon^2 \sum (1 + \lambda k^{2m})^{-2} \sim v_m \epsilon^2 \lambda^{-1/2m},
\]

valid for small \( \lambda \). According to, e.g. Gradshteyn & Ryzhik (1980, Formula 3.241.5), the constant

\[
v_m = \int_0^\infty (1 + t^{2m})^{-2} dt = (1 - \delta_m)/(\pi \delta_m),
\]

where \( \delta_m = (2m)^{-1} \) and \( \text{sinc} x = (\sin x)/x \). In the case \( m = 2 \) (cubic splines), \( v_2 = 3\sqrt{2}/16 \).

For the squared bias term, note first that \( 1 - \alpha = [1 + \lambda^{-1} k^{-2m}]^{-1} \), so that (4.18) becomes

\[
\bar{B}^2 = C^2 \{ \inf_k k^{\alpha} + \lambda^{-1} k^{\alpha - 2m} \}^{-2}.
\]

Apply the variance-bias lemma with \( h = k^{2m-\alpha}, v = \lambda^{-1}, b = 1 \) and \( 2\beta = \alpha/(2m - \alpha) \) to obtain \( \bar{B}^2 = b_{\alpha m} C^2 \lambda^{\alpha/m} \), with

\[
b_{\alpha m} = \epsilon^{-2H(\alpha/2m)} = (2m)^{-2} \alpha^{\alpha/m} (2m^{-2} - \alpha/2m). \]

Combining the variance and bias terms yields (4.20). \( \square \)

Optimizing over \( \lambda \). Our interest now turns to the value of \( \lambda \) that minimizes the maximum risk (4.20). This is called the minimax \( \lambda \) for the parameter space \( \Theta \).

4.4. PROPOSITION. Consider an \( m \)-th order spline estimate \( \hat{\theta}_\lambda \) over a Sobolev ellipsoid \( \Theta_2^2(C) \) with \( \alpha < 2m \). Then as \( \epsilon \to 0 \), the minimax value of the regularization parameter

\[
\lambda_s \sim \left( \frac{c_i \epsilon^2}{C^2} \right)^{2m/(2\alpha + 1)}.
\]

Let \( r = 2\alpha/(2\alpha + 1) \). The resulting maximum MSE

\[
\tilde{\varphi}(\hat{\theta}_{\lambda_s}; \epsilon) = c_s C^{2 - 2r} \epsilon^{2r}.
\]

Explicit values for \( c_i = c_s(\alpha, m) \) appear below.
It will be seen in Chapter 8 that the best possible rate of convergence for ellipsoids of functions with $\alpha$ mean-square derivatives is $e^{2r}$. So (4.23) shows that using smoothing splines with roughness penalty $\int (Dmf)^2$ achieves the right rate of convergence so long as $\alpha < 2m$ and $\lambda$ is chosen in accordance with (4.22). Of course, for this result, the choice of $\lambda$ depends on $C, m$ and $\alpha$, but in Section 4.5, data-determined choices of $\lambda$ will be considered.

**Proof.** To minimize (4.20), use the Variance-bias lemma with $h = \lambda^{1/2m}, v = v_m e^2, b = b_m C^2$ and $\beta = \alpha$. This yields the advertised results, with constants

$$c_1 = v_m/(2\sigma b_m), \quad c_2 = e^{H(r) - 2(1-r)H((\alpha/2m)v_m^r)}.$$  

**Remark.** If $\alpha = m$, then $b_m = 1/4$ and $c_1 = 2v_m/m$. This leads to the useful special case

$$(4.24) \quad \bar{r}(\hat{\theta}_\lambda; \epsilon) = e^{H(r)(C^2/4)^{1-r}(v_m e^2)^r}.$$  

In particular, for cubic splines over ellipsoids of twice differentiable functions (in mean square), we get that $\lambda_s = (4v_m^2/C^2)^{1/3}$. For a fixed function $f$, recall that $\int f''^2 = \pi^4 \sum a_k^2 \theta_k^2$. Thus, if $f$ is known (as for example in simulation studies), and a reasonable value of $\lambda$ is desired, one might set $C^2 = \pi^{-1} \int f''^2$ to arrive at the proposal

$$\lambda = \left(\frac{\pi}{2}\right)^4 \left(\frac{6\pi^2 e^2}{\int f''^2}\right)^{1/5}.$$  

4.5. **Exercise.** Consider a slightly different family of shrinkage rules, to appear in Pinsker’s theorem, and also indexed by a positive parameter $k$:

$$\hat{\theta}_{\mu,k}(y) = (1 - k^m/\mu)_+ y_k, \quad k \in \mathbb{N}.$$  

Show that the maximum risk over a Sobolev ellipsoid $\Theta^2(C)$ is approximated by

$$\bar{r}(\hat{\theta}_{\mu}; \epsilon) \sim \bar{e}_m e^{2\mu^{1/m} + C^2 \mu^{-2\min(\alpha/m,1)},}$$

where

$$\bar{e}_m = 2m^2/(m + 1)(2m + 1).$$

If $\alpha = m$, show that the maximum MSE associated with the minimax choice of $\mu$ is given by

$$(4.25) \quad \bar{r}(\hat{\theta}_{\mu}; \epsilon) \sim e^{H(r)} C^{2-2r}(\bar{e}_m e^2)^r.$$  

55
4.5. Integrated Wiener Priors and Spline Smoothing

[Introductory Text.]

A Gaussian prior \( \pi \) on \( \theta \) may be specified by making each coordinate \( \theta_k \) independent and distributed as \( N(0, \tau_k^2) \). The joint distribution of \( (\theta, y) \) then has the pairs \( (\theta_k, y_k) \) independent of one another, and within pairs

\[
\theta_k \sim N(0, \tau_k^2), \quad y_k \mid \theta_k \sim N(\theta_k, \epsilon^2).
\]

Within each pair, marginal distribution of \( y_k \) and the posterior distribution of \( \theta_k \) given \( y_k \) are easily found:

\[
y_k \sim N(0, \epsilon^2 + \tau_k^2), \quad \theta_k \mid y_k \sim N \left( \frac{\tau_k^2 y_k}{\tau_k^2 + \epsilon^2}, \frac{\tau_k^2 \epsilon^2}{\tau_k^2 + \epsilon^2} \right).
\]

The Bayes rule \( \hat{\theta}_x \) is the posterior mean, and shrinks linearly in each coordinate:

\[
\hat{\theta}_x = E[\theta_k \mid y_k] = c_k y_k, \quad c_k = \tau_k^2 / (\tau_k^2 + \epsilon^2).
\]

The Karhunen-Loève analysis of the \((m-1)\)-fold integrated Wiener process in Section 3.2 indicated that, depending on boundary conditions, the eigenvalues decayed exactly or approximately like \( k^{-2m} \). Since the eigenvalues are co-ordinate variances in the Karhunen-Loève basis, we focus in this section on

\[
\tau_k^2 = bk^{-2m}, \quad b, m > 0.
\]

The range of \( m \), which models smoothness, is extended to allow non-integer values. The scale parameter \( b \) controls the magnitude of the oscillations in the sample paths of the prior. The corresponding shrinkage factors \( c_k \) can be written in two ways, if we make the identification \( \lambda = \epsilon^2 / b \)

\[
c_k = \frac{bk^{-2m}}{bk^{-2m} + \epsilon^2} = \frac{1}{1 + \lambda k^{2m}}.
\]

Thus, for the integrated Wiener process prior the Bayes rule is the same as the smoothing spline estimate, so long as we calibrate the spline regularization parameter \( \lambda \) with the prior variance scale factor \( b \) via \( \lambda = \epsilon^2 / b \).

The proofs of the following two preparatory lemmas are left as exercises (See Exercise 5).

4.6. Lemma. Let \( z_k \sim N(0, \sigma_k^2) \) be independent, \( k \in \mathbb{N} \). Then

\[
P(\sum_k z_k^2 < \infty) = 1 \text{ or } 0 \quad \text{according as} \quad \sum_k \sigma_k^2 < \infty \text{ or } = \infty.
\]

4.7. Lemma. The sum

\[
S(\lambda) = \sum_{k=0}^{\infty} k^p(1 + \lambda k^q)^{-r} < \infty
\]
if and only if \( qr > p + 1 \). If so, and if \( p > -1 \), then as \( \lambda \to 0 \),

\[
S(\lambda) \sim \kappa \lambda^{-(p+1)/q},
\]
with \( \kappa = \kappa(p, r; q) = \int_0^\infty v^p(1 + v^q)^{-r} dv \).

Integrated Wiener process priors have some curious properties. For cubic splines \((m = 2)\), for example, the prior often discussed is the once integrated Wiener process. While the roughness penalty \( \int f'^2 \) in the cubic spline formulation penalizes second derivative, it turns out that with probability one, the sample functions of neither prior nor posterior are twice differentiable! On the other hand, the posterior mean \( \hat{\theta}_x \) is smoother, and is twice differentiable almost surely.

To state these results precisely, we again use (4.19) to define the space \( \Theta^2_{\pi} \) of sequences corresponding to functions with \( \alpha \) derivatives in mean square. The notation \( P_{\pi} \) refers to the marginal distribution of the observations \( y \) and \( \pi_y \) to the posterior distribution of \( \theta \) given \( y \) described in (4.26).

4.8. PROPOSITION. (a) Assume that \( m > 1/4 \). The parameter \( \theta \) belongs to \( \Theta^2_{\pi} \) both \( \pi \)-a.s. and \( \pi_y \)-a.s. \((\text{for } P_{\pi \cdot} \text{-a.s. } y)\) if and only if \( \alpha < m - 1/2 \).

(b) The posterior mean \( \hat{\theta}_x(y) \) belongs to \( \Theta^2_{\pi} \) both \( \pi \)-a.s. and \( P \)-a.s. \((\text{for } \theta \in \mathcal{E}_2)\) if and only if \( \alpha < 2m - 1/2 \).

PROOF. The prior support claim is a direct consequence of Lemma 4.6: \( \theta \in \Theta^2_{\pi} \) with \( \pi \)-probability one if and only if \( \sum k^{2\alpha} \tau_k^2 = b \sum k^{2\alpha-2m} < \infty \).

The \( P_{\pi \cdot} \)-marginal behavior of the posterior mean \( \hat{\theta}_x \) follows similarly: the vector \( (c_k y_k) \) lies in \( \Theta^2_{\pi} \) with \( P_{\pi \cdot} \)-probability one if and only if \( \sum k^{2\alpha} c_k^2 (\epsilon^2 + \tau_k^2) < \infty \). Since \( \tau_k^2 \to 0 \), convergence is determined by the \( \epsilon^2 \) term, and so

\[
\sum k^{2\alpha} c_k^2 = \sum k^{2\alpha} (1 + \lambda k^{2m})^{-2} < \infty
\]

if and only if \( 4m > 2\alpha + 1 \).

Now suppose that \( \alpha < 2m - 1/2 \) and confine attention to the \( P_{\pi \cdot} \)-a.s. set \( A \) of \( y \) for which \( \hat{\theta}_x(y) \in \Theta^2_{\pi} \). The posterior distribution of \( \theta \) given \( y \) may be stochastically represented by the sequence \( (\hat{\theta}_{x,k} + v_k \omega_k) \) where the posterior variance \( v_k^2 = \epsilon^2 c_k \) and the \( \omega_k \) are i.i.d. \( N(0, 1) \). Consequently, \( \pi(\theta \in \Theta^2_{\pi} | y) = 1 \) if and only if

\[
\sum k^{2\alpha} (\hat{\theta}_{x,k} + v_k \omega_k)^2 < \infty
\]

with probability one. In the posterior distribution \( \hat{\theta}_x(y) \) is constant and belongs to \( \Theta^2_{\pi} \), since \( y \in A \). So the sum converges if and only if \( \sum k^{2\alpha} v_k^2 = \epsilon^2 \sum k^{2\alpha}(1 + \lambda k^{2m})^{-1} < \infty \), which corresponds to \( \alpha < m - 1/2 \).

It remains to establish when \( \hat{\theta}_x(y) \) belongs to \( \Theta^2_{\pi} \) if \( y \sim P_\theta \). First we establish that \( P_\theta \) is equivalent to \( P_\pi \) when \( m > 1/4 \). Indeed, Kakutani’s theorem [REFERENCE!!] states that the Gaussian product measure with components \( N(0, \epsilon^2) \) is equivalent to that with components \( N(0, \epsilon^2 + \tau_k^2) \) if
and only if
\[
\prod_k \left( \frac{2\epsilon^2 + \sigma_k^2}{2\epsilon^2 + \tau_k^2} \right)^{1/2} > 0.
\]
\[
\prod_k \left( \frac{2\epsilon \sqrt{\epsilon^2 + \tau_k^2}}{2\epsilon^2 + \tau_k^2} \right)^{1/2} > 0.
\]
This condition is equivalent to \(\sum (1 + \delta_k)^{1/4}(1 + \delta_k/2)^{-1/2} < \infty\), for \(\delta_k = \tau_k^2 / \epsilon^2\). [APPENDIX] Taylor expansion of the summands shows that this convergence occurs iff \(\sum \delta_k^2 < \infty\), that is, iff \(m > 1/4\). Finally, again by Kakutani, \(P_g \sim P_0\) for all \(\theta \in \ell_2\) and so almost sure convergence of \(\sum k^{2\alpha} \hat{\theta}_{\alpha, k}(y)\) for \(P_\pi\) is equivalent to a.s. convergence under \(P_g\).

**Rate of convergence of Bayes rules.** Since the integrated Wiener prior Bayes rules can be identified with spline smoothers, we may derive conclusions about their rates of convergence from the earlier calculations with splines. Thus, the maximum risk of the Bayes rule \(\hat{\theta}_\pi\) over \(\Theta_2\) may be read off from (4.20). After substituting the calibration \(\lambda = \epsilon^2 / b\) and absorbing the fixed value of \(b\) into the constants, we get
\[
\bar{r}(\hat{\theta}_\pi; \epsilon) = c_m \epsilon^2 \lambda^{-1/2m} + c_\beta \epsilon^2 \lambda^{\alpha/m},
\]
(4.27)
\[
= c_1 (\epsilon^2)^{1-1/2m} + c_2 (\epsilon^2)^{\alpha/m}.
\]

A number of conclusions are now immediate: first if \(m \leq 1/2\), then the Bayes rule is not even consistent as \(\epsilon \to 0\), since the variance term does not decrease to zero. More importantly, let us focus on the case of twice differentiable functions, \(\alpha = 2\), and in particular on ellipsoids \(\Theta_2(C)\) of functions with \(\int \mu^2 < \pi^2 C^2\). Let us anticipate from Chapter [PRECEIVE REFERENCE!] that the “right” or best possible rate of convergence of estimators over such ellipsoids is \((\epsilon^2)^{4/5}\).

(a) Consider the integrated Wiener process prior with \(m = 2\). From (4.27) it is apparent that the maximum MSE of the Bayes rule \(\bar{r}(\hat{\theta}_\pi; \epsilon) \propto (\epsilon^2)^{3/4}\) which decreases asymptotically more slowly than the correct rate \((\epsilon^2)^{4/5}\).

(b) Suppose instead that \(m\) is chosen to optimize the rate of convergence in (4.27): balancing the variance and bias terms through the equation \(2 - 1/m = 4/m\) entails \(m = 5/2\) and then \(\bar{r}(\hat{\theta}_\pi; \epsilon) \propto (\epsilon^2)^{4/5}\) does attain the right rate. However, Proposition 4.8 (a) shows that the integrated Wiener prior with \(m = 5/2\) also fails to live on twice-differentiable functions!

(c) The situation is even worse than this: Zhao (2000) shows that there is no Gaussian measure \(\pi\) which lives on \(\Theta_2\) for which the Bayes rule \(\hat{\theta}_\pi\) has the right rate of convergence. And of course, similar results hold for the spaces \(\Theta_2\) representing other degrees of smoothness. Zhao (2000) shows that the situation can be rescued by constructing a mixture of Gaussian priors that lives on \(\Theta_2\) such that the Bayes rule does have the right rate of
convergence. However, we take a different approach and allow the prior \( \pi \) to depend on \( \epsilon \).

**Priors depending on \( \epsilon \).** When considering smoothing splines, we saw in Proposition 4.4 that the choice \( \lambda_\epsilon = (e \epsilon^2)^{2m/(2m+1)} \) leads to the right rate of convergence for \( \alpha < 2m \). Using the calibration \( \lambda = \bar{c}^2 / b \), this implies that if the prior scale factor

\[
b = b(\epsilon) = \bar{c} (\epsilon^2)^{1-2m/(2m+1)},
\]

then the Bayes rules \( \hat{\theta}_\pi \) for the sequence of priors \( \pi_\epsilon \) does achieve the right rate of convergence over Sobolev balls \( \Theta^2_\epsilon(C) \).

Inspecting the formula for \( b(\epsilon) \), we see that if \( \alpha > m-1/2 \), then \( b(\epsilon) \to 0 \) with \( \epsilon \), so that the effect of allowing \( b \) to vary is to smooth the prior sample paths by damping their magnitude. Conversely, if \( \alpha < m-1/2 \), in which case the prior does live on \( \Theta^2_\epsilon \), then \( b(\epsilon) \) increases to \( \infty \), with corresponding amplification of the sample paths.

**Remarks.** 1. In Chapter 8, we will see that the least favorable distributions for ellipsoids \( \Theta^2_\epsilon(C) \) form a sequence of Gaussian priors which depend on \( \epsilon \). Incidentally, these priors have \( \tau^2_\epsilon(\epsilon) = 0 \) for \( k \geq k(\epsilon) \), so they automatically live on \( \Theta^2_\epsilon \).

2. Bayesian orthodoxy frowns on prior distributions that depend on sample size. There are perhaps two responses. First, our use of priors is technical rather than subjectivist: we do not claim to have a priori reasons for believing integrated Wiener process priors. The second thesis is that as the sample size \( (n \leftrightarrow \epsilon^{-2}) \) increases, the complexity of models that one entertains naturally increases, and so the prior distribution must in turn adapt to these richer specifications.

3. Another piece of Bayesian anathema is to estimate the prior from the data — “there is no-one less Bayesian than an empirical Bayesian”. To this we now turn!

**Empirical Bayes estimation of \( \lambda \).**

The marginal distribution of \( y_k \) is \( N(0, \bar{c}^2 + \tau^2_\epsilon) \), and in the integrated Wiener prior case, the marginal variances may be written

\[
v_k = \bar{c}^2 + b k^{-2m} = \bar{c}^2 [1 + \lambda^{-1} k^{-2m}], \quad \lambda = \bar{c}^2 / b.
\]

Thus the marginal distribution of the data \( (y_k) \) depends on the parameter \( \lambda \) and we may estimate it using likelihood techniques. The log-likelihood for \( \lambda \) is

\[
l(\lambda) = c - \sum_k \frac{1}{2} \log v_k + \frac{y_k^2}{2v_k}.
\]

Recall the notation for the shrinkage factors \( w_k(\lambda) = [1 + \lambda k^{2m}]^{-1} \). Since \( (\partial v_k / \partial \lambda) / v_k = -\lambda^{-1} w_k(\lambda) \), the likelihood equation \( l'(\lambda) = 0 \) becomes

\[
\bar{c}^2 \lambda^{-1} \sum_k w_k(\lambda) = \sum_k w_k^2(\lambda) k^{2m} y_k^2.
\]

(4.28)
We will give a partially heuristic analysis of the frequentist behavior of $\lambda_{GML}$ as $\epsilon \to 0$. If the true mean vector is $\theta$, let $Q_m(\theta) = \sum k^{2m} \theta_k^2$; which is just a constant times the roughness penalty $\int (D^m f)^2$.

4.9. Claim. If $y = \theta + \epsilon z$ and $\lambda_{GML}$ denotes the generalized maximum likelihood, or empirical Bayes estimate of $\lambda$, then as $\epsilon \to 0$,

\begin{equation}
\hat{\lambda}_{GML}(\epsilon) \sim \left( \frac{v_m \epsilon^2}{Q_m(\theta)} \right)^{2m/(2m+1)}.
\end{equation}

The constant $v_m$ is given in (4.21).

If we interpret (4.29) in terms of the scale parameter $b(\epsilon) = \epsilon^2/\lambda$, we immediately see that the empirical Bayes approach arrives at the same type of sample size dependence on $\epsilon$ as does the maximum-over-ellipsoid strategy, so long as we choose ellipsoids $\Theta^2(C)$ with $\alpha = m$.

Even more remarkably, except for the factor $2\alpha = 2m$, this data-based Empirical Bayes choice of the smoothing parameter $\lambda$ is identical to the choice suggested by optimizing the worst case MSE over an ellipsoid $\Theta_n^2$ with $C^2$ corresponding to $Q(\theta)$.

Details of Claim. We look heuristically at the limiting behavior of solutions to this equation as $\epsilon \to 0$. If we assume, for the moment, that $\lambda = \lambda(\epsilon) \to 0$, then the weights $w_k(\lambda) \to 1$. From the right side, define

$$\hat{Q}_m(\theta; \lambda) = \sum_k w_k^2(\lambda) k^{2m}(y_k^2 - \epsilon^2).$$

Since $E y_k^2 = \epsilon^2 = \theta_k^2$ it follows by the dominated convergence theorem that $E \hat{Q}_m(\theta; \lambda)$ converges to $Q_m(\theta)$. It can be checked easily that $\hat{Q}_m(\theta; \lambda)$ has variance approaching zero with $\epsilon$, and so converges in probability to $Q_m(\theta)$, [and further work shows that the convergence is almost sure].

The left side of (4.28) and that part of the right side not included in $\hat{Q}_m$ are both deterministic, and maybe approximated by integrals using Lemma 4.7. Taking both such terms over to the left side, it follows that the score equation (4.28) is, up to negligible deterministic factors, equivalent to

$$\epsilon^2[\kappa(0,1) - \kappa(2m,2)] \lambda^{-1-1/2m} = \hat{Q}_m(\theta; \lambda).$$

Fortuitous cancellation shows that $\kappa(0,1) - \kappa(2m,2)$ reduces to $v_m$. Bringing in now the convergence of $\hat{Q}_m$, we find that the solution $\lambda = \lambda_{GML}(\epsilon)$ approximately satisfies the equation

$$\epsilon^2 v_m \lambda^{-1-1/2m} \sim Q_m(\theta),$$

or equivalently (4.29).

Remark. According to Wahba, maximum likelihood estimation of $\lambda$ based on a Bayes model for smoothing splines goes back to Wecker & Ansley (1983). It has been further studied, among others, by Wahba (1985) and Efron (1999).
[What happens to the maximum MSE of GML over ellipsoids?]

4.6 Problems

1. (Discrete orthogonality relations). Let $e_k$ denote the vector in $\mathbb{C}^n$ obtained by sampling the $k$-th complex exponential at $t_j = j/n$. Thus $e_k = \{\exp(2\pi ikj/n), j = 0, 1, \ldots, n - 1\}$. For $f, g \in \mathbb{C}^n$, use the usual inner product $\langle f, g \rangle = \sum_{j=0}^{n-1} f_j \bar{g}_j$. Show that for $k, l \in \mathbb{Z}$,

$$\langle e_k, e_l \rangle = \begin{cases} n & \text{if } k - l \in n\mathbb{Z} \\ 0 & \text{otherwise.} \end{cases}$$

Turn now to the real case. For $k \geq 0$, let $c_k = \{\cos(2\pi kj/n), j = 0, 1, \ldots, n - 1\}$ and define $s_k$ analogously using the $k$-th sine frequency. If $n = 2m + 1$ is odd, then take $\{c_0, s_1, \ldots, s_m, c_m\}$ as the basis $B_n$ for $\mathbb{R}^n$. If $n = 2m + 2$ is even, then adjoin $c_{n/2}$ to the previous set to form $B_n$. Show that the following orthogonality relations hold for basis vectors in $B_n$:

$$\langle c_k, s_l \rangle = \frac{n}{2} \delta_{kl}, \quad \langle c_k, c_l \rangle = 0,$$

with the exception of

$$\langle c_0, c_0 \rangle = \langle c_{n/2}, c_{n/2} \rangle = n,$$

where the last equation is only needed if $n$ is even.

Hint. Derive the real relations from the complex by writing $e_k = c_k + is_k$ and using the complex orthogonality relations for pairs $(k, l)$ and $(k, -l)$.

2. (Fourier transform of the equivalent kernel.) The Fourier transform of an integrable function on $\mathbb{R}$ is defined by $\hat{f}(\xi) = \int_{\mathbb{R}} f(x) e^{-i\xi x} dx$. If $f$ is sufficiently nice, it may be recovered from the inversion formula $f(x) = (2\pi)^{-1} \int_{\mathbb{R}} \hat{f}(\xi) e^{i\xi x} d\xi$. The Poisson summation formula (e.g. Mallat, page 28) states that under suitable conditions on $f$, $|(1 + x^2)(|f(x)| + |f'(x)| + |f''(x)|)$ bounded or the same condition on $\hat{f}$ will do, then

$$\sum_{k \in \mathbb{Z}} f(k) = \sum_{k \in \mathbb{Z}} \hat{f}(2\pi k).$$

Use the Poisson summation formula to give an alternate demonstration that the kernel $K_h(s) = 1 + 2 \sum_{k=1}^{\infty} \frac{\cos(2\pi ks)}{1 + (2\pi kh)^2}$ is the wrapped version $\sum_{k} L_h(s + k)$ of $L_h$, with Fourier transform given by

$$\hat{L}_h(\xi) = [1 + \xi^2 h^2]^{-1}.$$

3. (Evaluation of equivalent kernel.) If $\alpha \in \mathbb{C}$ belongs to the upper half plane, show by contour integration that

$$\frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{e^{\gamma x}}{x - \alpha} dx = \begin{cases} e^{i\gamma \alpha} & \text{if } \gamma > 0 \\ 0 & \text{if } \gamma < 0. \end{cases}$$
Use the partial fraction expansion
\[ \prod_{k=1}^{r} \frac{1}{(x - \beta_k)^{-1}} = \sum_{k=1}^{r} \frac{c_k}{x - \beta_k}^{-1}, \quad \frac{1}{\alpha_k} = \prod_{j \neq k} \frac{1}{(\beta_k - \beta_j)}, \]
to compute the equivalent kernel \( L(t) \) given that \( \hat{L}(\xi) = (1 + \xi^4)^{-1} \).

4. *(Wahba’s prior for cubic splines.)* Show that

\[ Z_{\sigma}^2(t) = \sigma \xi_1^2 + \sigma \xi_2^2 + \int_0^1 (t - u)dW(u), \]

the integrated (free) Wiener process, has covariance function

\[ R_\sigma(s, t) = \sigma^2(1 + st) + R_0(s, t), \]

\[ R_0(s, t) = \begin{cases} \frac{1}{2} s^2 t - \frac{1}{6} s^3 & 0 \leq s \leq t \\ \frac{1}{2} s t^2 - \frac{1}{6} t^3 & 0 \leq t \leq s. \end{cases} \]

By differentiating the eigenvalue equation
\[ \int_0^1 R_\sigma(s, t)\varphi(t)dt = \lambda^2 \varphi(s) \]
four times, show that \( \varphi \) satisfies
\[ \varphi(s) = \lambda^2 \varphi^{(4)}(s), \]
with boundary conditions
\[ \varphi''(0) = \sigma^{-2} \varphi'(0), \quad \varphi'''(0) = \sigma^{-2} \varphi(0), \quad \varphi''(1) = \varphi'''(1) = 0. \]

With \( \sigma = 0 \), show that the boundary conditions imply the equation \( \cos \lambda^{-1/2} \cosh \lambda^{-1/2} = -1 \) for the eigenvalues. In the \( \sigma = \infty \) limit, show that the corresponding equation is \( \cos \lambda^{-1/2} \cosh \lambda^{-1/2} = 1 \). In either case, show that the eigenvalues satisfy, for large \( n \)
\[ \lambda_n \sim \frac{1}{(n + \frac{1}{2}) \pi^2} \sim \frac{1}{n^2 \pi^2}. \]

Make plots of the first six eigenfunctions corresponding to the \( \sigma = \infty \) limit.

5. Prove Lemma 4.6. [Hint. To show that \( \sum \sigma_k^2 < \infty \), use characteristic functions.]

6. *(Computational comparison.)* Consider two functions on \([0, 1]\\):
\[ f_1(t) = \sin 4\pi t^2, \quad f_2(t) = (e^{4t} - 1 - t)(1 - t)^2, \]
and consider the model
\[ Y_i = f(i/n) + \sigma z_i, \quad z = 1, \ldots, n, \]
with \( \sigma = 1 \) and \( z_i \sim N(0, 1) \) chosen i.i.d. Let \( \hat{f}_{SS, \lambda} \) and \( \hat{f}_{PER, \lambda} \) denote the solutions to
\[ \min Q(f) = n^{-1} \sum (Y_i - f(i/n))^2 + \lambda \int_0^1 f' \]

62
among cubic splines and trigonometric polynomials respectively. Note that $\hat{f}_{SS,\lambda}$ can be computed in S-PLUS using `smooth.spline()`. For $\hat{f}_{PER,\lambda}$, you’ll need to use the discrete Fourier transform `fft()`, with attention to the real and imaginary parts. For $\lambda$, use the value suggested by the ellipsoid considerations in class:

$$\lambda = (\pi/2)^{1/2}6^{1/5}(n \int f''^2)^{-1/5}.$$  

Run experiments with $R = 100$ replications at $n = 50, 200$ and 1000 to compare the estimates $\hat{f}_{SS,\lambda}$ and $\hat{f}_{PER,\lambda}$ obtained for $f_1$ and $f_2$. Make visual comparisons on selected replications *chosen in advance*, as well as computing averages over replications such as

$$\frac{\text{ave } \| \hat{f}_{SS} - \hat{f}_{PER} \|^2}{\text{ave } \| \hat{f}_{SS} - f \|^2}$$

4.7. *Bibliographic Notes*
CHAPTER 5

Some Linear Inverse Problems

5.1. Some examples
5.2. Singular value decomposition
5.3. Regularization schemes
5.4. Wahba’s generalized spline model
5.5. Prediction error and Estimation error
Part 2

SOME GAUSSIAN DECISION THEORY
CHAPTER 6

Bayes Risks and Minimaxity

This chapter collects some basic tools used to evaluate minimax risks. Bayes estimators play a critical role, through the minimax theorem, which characterizes minimax risk in terms of a least favorable prior distribution.

The exposition is tailored to the Gaussian sequence model, which allows a number of simplifications of the general decision theoretic toolkit. In some places we also rely heavily on the independence of the co-ordinates in the sequence model, in order to decompose minimax problems into simpler subparts.

6.1. Bayes Estimators

It is often useful to consider averages of risk functions with respect to a probability distribution: if \( \pi \) is a probability distribution on \( \ell_2(I) \) : the integrated risk of an estimator \( \hat{\theta} \) is defined by

\[
B(\hat{\theta}, \pi) = \int r(\hat{\theta}, \theta) \pi(d\theta)
\]

\[
= E_{\pi} r(\hat{\theta}, \theta) = E_{\pi} E_{\theta} L(\hat{\theta}(y), \theta).
\]

(6.1)

An estimator \( \hat{\theta}_\pi \) that minimizes \( B(\hat{\theta}, \pi) \) for a fixed prior \( \pi \) is called a Bayes estimator for \( \pi \), and the corresponding minimum value is called the Bayes risk \( B(\pi) \); thus

\[
B(\pi) = \inf_{\hat{\theta}} B(\hat{\theta}, \pi).
\]

(6.2)

Of course \( B(\pi) = B(\pi, \epsilon) \) also depends on the noise level \( \epsilon \), but again this will not always be shown explicitly.

Remark. One of the reasons for using integrated risks is that, unlike the ordinary risk function \( \theta \to r(\hat{\theta}, \theta) \), the mapping \( \pi \to B(\hat{\theta}, \pi) \) is linear. Representation (6.2) then shows that the Bayes risk \( B(\pi) \) is a concave function of \( \pi \).

The decidedly frequentist definition of Bayes estimators fortunately agrees with the subjectivist definition, under mild regularity conditions. Bayes’ theorem states that the joint distribution \( \pi P \) of the pair \( (\theta, y) \) may be decomposed two ways:

\[
\pi P(d\theta, dy) = \pi(d\theta) P(dy|\theta) = P_\pi(dy) \pi(d\theta|y),
\]

66
where $P_\pi(dy)$ is the marginal distribution of $y$ and $\pi(d\theta|y)$ is the posterior distribution of $\theta$ given $y$. The integrated risk of (6.1), which uses the first decomposition, may be written using the second, posterior decomposition as

$$B(\hat{\theta}, \pi) = E_{P_\pi} E_y L(\hat{\theta}(y), \theta).$$

Here, $E_{P_\pi}$ denotes expectation with respect to the marginal distribution $P_\pi(dy)$ and $E_y$ denotes expectation with respect to the posterior $\pi(d\theta|y)$. Thus one sees that $\hat{\theta}_\pi(y)$ is obtained by minimizing the posterior expected loss:

$$\hat{\theta}_\pi(y) = \arg\min_a E_y L(a, \theta).$$

This formula often leads to explicit expressions for the Bayes rules. In particular, if $L(a, \theta) = ||a - \theta||_2^2$, the Bayes estimator is simply given by the mean of the posterior distribution, which we write as

$$\hat{\theta}_\pi(y) = E_\pi(\theta|y).$$

**Uniqueness of the Bayes rule.** The following sufficient condition is proved, for example, in Lehmann (1983, Corollary 4.1.2).

6.1. **Proposition.** Suppose the loss function $L(a, \theta)$ is strictly convex in $a$. The Bayes estimator $\hat{\theta}_\pi$ is unique (a.e. $P_\theta$ for each $\theta$) if $B(\pi) < \infty$, and if a.s. $P_\pi$ implies a.s. $P_\theta$ for each $\theta$.

**Product priors.** Suppose that the coordinates $\theta_i$ of $\theta$ are gathered into groups: $\theta = (\theta_j, j \in J)$ for some finite or infinite set $J$. The $\theta_j$ may just be the individual components of $\theta$, or they may consist of blocks of individual coefficients. For example, in a wavelet decomposition, we re-index the individual coordinates as $\theta_{jk}$ and $\theta_j$ may, for example, represent $(\theta_{jk}, k = 1, \ldots, 2^J)$.

Suppose that the prior $\pi$ makes the groups independent: $\pi(d\theta) = \prod_j \pi_j(d\theta_j)$. Since the individual components in sequence model (3.1) are independent, the likelihood factorizes over the groups $j$: $P(dy|\theta) = \prod_j P(dy_j|\theta_j)$ and so the posterior factorizes also

$$\pi(d\theta|y) = \prod_{j \in J} \pi_j(d\theta_j|y_j).$$

Suppose also that the loss function is additive, so that if $a = (a_j)$,

$$L(a, \theta) = \sum_j L(a_j, \theta_j).$$

Then the minimization of (6.3) for the Bayes rule decomposes into separate components, precisely because the posterior factorizes:

$$\hat{\theta}_\pi(y) = \arg\min_{(a_j)} \sum_j E[L(a_j, \theta_j)|y_j] = (\hat{\theta}_{\pi,j}(y_j)).$$
Here $\hat{\theta}_{\pi_j}(y_j)$ denotes the Bayes rule for estimating $\theta_j$ given observed data $y_j$ and the prior $\pi_j$ on $\theta_j$ alone.

To summarize: given product likelihoods and additive losses, *product priors yield separable Bayes rules*. In such cases, the risk functions are additive

$$ r(\hat{\theta}_\pi, \theta) = \sum_j EL(\hat{\theta}_{\pi_j}(y_j), \theta_j) = \sum_j r(\hat{\theta}_{\pi_j}, \theta_j) $$

and in consequence, so are the Bayes risks

$$ (6.6) \quad B(\pi) = \int r(\hat{\theta}_\pi, \theta)\pi(d\theta) = \sum_j B(\pi_j). $$

*Independence is less favorable.* Here is a trick that often helps in finding least favorable priors. Let $\pi$ be an arbitrary prior, so that the $\theta_j$ are not necessarily independent. Denote by $\pi_j$ the marginal distribution of $\theta_j$. Build a new prior $\bar{\pi}$ by making the $\theta_j$ independent: $\bar{\pi} = \prod_j \pi_j$. This product prior is more difficult, as measured in terms of Bayes risk.

6.2. **Lemma.** $B(\bar{\pi}) \geq B(\pi)$.

**Proof.** Because of the independence structure, the $\bar{\pi}$—posterior distribution of $\theta_j$ given $y$ in fact depends only on $y_j$ – compare (6.5). Hence the $\bar{\pi}$—Bayes rule is separable: $\hat{\theta}_{\bar{\pi}, j}(y) = \hat{\theta}_{\pi, j}(y_j)$. From the additivity of losses and independence of components given $\theta$,

$$ r(\hat{\theta}_{\bar{\pi}}, \theta) = \sum_j \int L_j(\hat{\theta}_{\bar{\pi}, j}(y_j), \theta_j)P_{\theta_j}(dy_j) = \sum_j r(\hat{\theta}_{\bar{\pi}, j}, \theta_j). $$

The $\pi$—average of the rightmost term depends only the marginals $\pi_j$, so

$$ \int r(\hat{\theta}_{\bar{\pi}}, \theta)\pi(d\theta) = \int r(\hat{\theta}_{\bar{\pi}}, \theta)\bar{\pi}(d\theta) = B(\bar{\pi}), $$

The left side is just $B(\hat{\theta}_{\bar{\pi}}, \pi)$, which is at least as large as $B(\pi)$ by definition.

To see intuitively why the product marginal prior $\bar{\pi}$ is harder then $\pi$, consider squared error loss: conditioning on all of $y$ has to be better than conditioning on just $y_j$:

$$ E_{\pi}[E_{\pi}(\theta_j|y) - \theta_j]^2 \leq E_{\pi}[E_{\bar{\pi}}(\theta_j|y_j) - \theta_j|^2]. $$

**$L_2$ error and projection formulas.** The rest of this section uses squared error loss. Brown (1971) made remarkable use of the following simple identity.

6.3. **Proposition.** Suppose that $L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2_2$. For any estimator $\hat{\theta}$ and prior distribution $\pi(d\theta)$,

$$ (6.7) \quad B(\hat{\theta}, \pi) - B(\pi) = \int \|\hat{\theta} - \hat{\theta}_\pi\|^2_2 p. $$
PROOF. An alternative interpretation of the posterior mean arises by thinking of the Bayes risk in terms of the joint distribution $\pi P_\theta$ of $(\theta, y)$:

$$B(\pi) = \inf \{E_{\pi P_\theta}||\theta - \hat{\theta}(y)||^2 : \hat{\theta}(y) \in L_2(P_\pi)\}.$$ 

Now $\hat{\theta}_\pi (y)$ can be viewed as the orthogonal projection of $\theta \in L_2(\pi P_\theta)$ on the closed linear subspace $L_2(P_\pi)$. Consequently, (6.7) is just the Pythagorean identity in $L_2(\pi P_\theta)$:

$$E_{\pi} E_\theta ||\theta - \hat{\theta}||^2 = E_{\pi} E_\theta ||\theta - \hat{\theta}_\pi||^2 + E_{\pi} E_{\hat{\theta}_\pi} ||\hat{\theta}_\pi - \hat{\theta}||^2. \quad \square$$

Consider now the finite dimensional model $y \sim N_\pi(\theta, \epsilon^2 I)$. The posterior mean has representation

$$\hat{\theta}_\pi (y) = \int \theta \phi_c(y - \theta)\pi(d\theta) / \int \phi_c(y - \theta)\pi(d\theta).$$

Since $(\partial/\partial x_i) \phi_c(x) = -(x_i/\epsilon^2) \phi_c(x)$, we may write the numerator integrand as

$$\theta_i \phi_c(y - \theta) = y_i \phi_c(y - \theta) + \epsilon^2 \frac{\partial}{\partial y_i} \phi_c(y - \theta),$$

leading to a representation of the Bayes estimator as a perturbation of the maximum likelihood rule

$$\hat{\theta}_\pi (y) = y + \epsilon^2 \nabla \log p(y),$$

where the marginal density of $y$ is $p(y) = \int \phi_c(y - \theta)\pi(d\theta)$, which is the convolution $\pi * \Phi_c$.

Aside: The celebrated paper of Brown (1971) uses (6.7) and (6.9) to show that statistical admissibility of $\hat{\theta}_\pi$ is equivalent to the recurrence of the diffusion defined by $dX_t = \nabla \log p(X_t)dt + 2dW_t$. In particular the classical and mysterious Stein phenomenon, namely the inadmissibility of the maximum likelihood estimator $\hat{\theta}(y) = y$ in exactly dimensions $n \geq 3$, is identified with the transience of Brownian motion in $\mathbb{R}^n$, $n \geq 3$. See also Srinivasan (1973).


We apply Brown’s identity and some facts about Fisher information to obtain some useful bounds on Bayes risks. For the rest of this section, $n = 1$. If $P$ is a probability measure on $\mathbb{R}$ with absolutely continuous density $p(x)dx$, the Fisher information is defined by

$$I(P) = \int \frac{p'(x)^2}{p(x)}dx.$$ 

The unbiased estimator $\hat{\theta}_0(y) = y$ has variance $\epsilon^2$, and so $B(\hat{\theta}_0, \pi) = E_{\pi} E_\theta (y - \theta)^2 = \epsilon^2$, regardless of the prior $\pi$. Substituting $\hat{\theta}_0$ and formula
(6.9) into (6.7), we have

$$e^2 - B(\pi) = e^4 \int \frac{p'(y)^2}{p(y)^2} p(y) dy.$$ 

Since $p$ is the absolutely continuous density of the marginal distribution $\pi \ast \Phi_\epsilon$, we arrive at

6.4. Corollary. For $y \sim N(\theta, e^2)$ and squared error loss,

$$(6.10) \quad B(\pi) = e^2[1 - e^2 I(\pi \ast \Phi_\epsilon)].$$

Now recall that Fisher information is always bounded below by precision: for any distribution $P$,

$$I(P) \geq 1/\text{Var} P,$$

with equality if and only if $P$ is Gaussian (see appendix). Applying (6.11) to (6.10), we arrive at

6.5. Corollary.

$$B(\pi, \epsilon) \leq \frac{e^2 \text{Var} \pi}{e^2 + \text{Var} \pi},$$

with equality if and only if $\pi$ is Gaussian.

Finally, we give a lower bound for $B(\pi)$ that is sometimes easier to use than (6.10). It is essentially a version of the van Trees inequality (Van Trees 1968) (see appendix).

$$(6.13) \quad B(\pi, \epsilon) \geq e^2/(1 + e^2 I(\pi)).$$

6.2. The Minimax Theorem

The minimax theorem of game and decision theory is a decisive tool in evaluating minimax risks, since it allows them to be calculated (or at least bounded) by finding the maximum Bayes risk over a suitable class of prior distributions. The resulting least favorable distribution and its associated Bayes estimator often give considerable insight into the estimation problem.

First an elementary lower bound for the minimax risk. If a prior probability distribution $\pi$ is supported in $\Theta$, then

$$B(\hat{\theta}, \pi) = \int_{\Theta} r(\hat{\theta}, \theta) \pi(d\theta) \leq \sup_{\theta \in \Theta} r(\hat{\theta}, \theta).$$

Minimizing over $\hat{\theta}$, we have

$$B(\pi) \leq R_N(\Theta).$$

Define the worst-case Bayes risk over a collection $\mathcal{P}$ of probability measures as

$$B(\mathcal{P}) = \sup_{\pi \in \mathcal{P}} B(\pi).$$
Letting \( \text{supp} \mathcal{P} \) denote the union of all \( \text{supp} \pi \) for \( \pi \) in \( \mathcal{P} \), we obtain the lower bound

\[
R_N(\Theta) \geq B(\mathcal{P}) \quad \text{whenever} \quad \text{supp} \mathcal{P} \subset \Theta.
\]

We now turn to the minimax theorem, in a version suited to the Gaussian sequence model. We defer to Appendix A a discussion of its assumptions and proof, and its connections with the classical minimax theorems of game theory.

A function \( f : T \to \mathbb{R} \) on a metric space \( T \) is lower semicontinuous at \( t \) if \( f(t) \leq \liminf_{s \to t} f(s) \). The action \( a \) is typically an infinite sequence \( a = (a_i) \in \mathbb{R}^\infty \). For technical reasons, we want to allow \( a_i = \pm \infty \), and take the action space \( \mathcal{A} = (\mathbb{R})^\infty \).

6.6. Theorem. Consider the estimation problem (3.1) and suppose that for each \( \theta \in \ell_2(I) \) the loss function \( L(a, \theta) \) is convex and lower semicontinuous in \( a \in \mathcal{A} \). Let \( B(\hat{\theta}, \pi) \) denote the integrated risk (6.1). Let \( \mathcal{P} \) be a convex set of probability measures on \( \ell_2(I) \). Then

\[
\inf_{\hat{\theta}} \sup_{\pi \in \mathcal{P}} B(\hat{\theta}, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\hat{\theta}} B(\hat{\theta}, \pi) = B(\mathcal{P})
\]

A maximising \( \pi \) is called a least favorable distribution (with respect to \( \mathcal{P} \)).

Remarks 1. A pair \( (\hat{\theta}_0, \pi_0) \) is called a saddlepoint if for all \( \hat{\theta} \), and all \( \pi \in \mathcal{P} \),

\[
B(\hat{\theta}_0, \pi) \leq B(\hat{\theta}_0, \pi_0) \leq B(\hat{\theta}, \pi_0).
\]

If a saddlepoint exists, then \( \hat{\theta}_0 \) is a Bayes rule for \( \pi_0 \) (from the right side), and \( \pi_0 \) is a least favorable distribution (since the left side implies \( B(\pi_0) \leq B(\pi) \)).

2. Let \( \delta_\theta \) denote a point probability mass concentrated at \( \theta \). Then we may rewrite \( r(\hat{\theta}, \theta) \) as \( B(\hat{\theta}, \delta_\theta) \). If \( \Theta \) is a parameter space and \( \mathcal{P} \) contains all point probability masses \( \delta_\theta, \theta \in \Theta \), then clearly

\[
\sup_{\theta \in \Theta} r(\hat{\theta}, \theta) \leq \sup_{\pi \in \mathcal{P}} B(\hat{\theta}, \pi),
\]

and so minimizing over all estimators \( \hat{\theta} \) and using (6.15) gives an upper bound on minimax risk that we will use frequently:

\[
R_N(\Theta) \leq B(\mathcal{P})
\]

3. Now we may combine the lower and upper bounds (6.14) and (6.16). For example, if \( \mathcal{P} = \mathcal{P}(\Theta) = \{ \pi : \text{supp } \pi \subset \Theta \} \), then

\[
R_N(\Theta) = B(\mathcal{P}(\Theta)).
\]

Remarks. 1. The weakening of continuity to lower semicontinuity seems necessary: even in dimension one with quadratic loss and \( \epsilon = 1 \), one checks that the (absurd) estimator \( \hat{\theta}(y) = e^{y^2/4} / (1 + y) I[y > 0] \) has a risk function discontinuous at 0.
2. It is easy to check that the loss functions $\|a - \theta\|_p^p$ are lower semicontinuous in $a$; if $a_{i}^{(n)} \to a_{i}^{(\infty)}$ for all $i$, then $\|a^{(\infty)} - \theta\|_p^p \leq \liminf_{n} \|a^{(n)} - \theta\|_p^p$.

Notes: $\pi \to B(\pi)$ is usc when $I$ finite = DJ94, p297. BUT, technical difficulty to show when $I = \mathbb{N}$ - no bdd risk functions on $\ell_2(I)$.

6.3. Minimaxity on Products.

Suppose that $\Theta \subset \ell_2(I)$ is a product space $\Theta = \prod_{j \in J} \Theta_j$. The index $j$ may refer to individual coordinates of $\ell_2(I)$, but in some cases each $j$ may represent a cluster of coordinates (for example, in wavelet bases, all coefficients at a fixed scale.) If the loss function is additive and convex, then the minimax risk for $\Theta$ can be built from the minimax risk for each of the subproblems $\Theta_j$.

6.7. Proposition. Suppose that $\Theta = \prod_{j \in J} \Theta_j$ and $L(a, \theta) = \sum L_j(a_j, \theta_j)$. Suppose that $a_j \to L_j(a_j, \theta_j)$ is convex and lower semicontinuous for each $\theta_j$. Then

$$R_N(\prod_j \Theta_j, \epsilon) = \sum_j R_N(\Theta_j, \epsilon).$$

If $\theta_j^*(y_j)$ is separately minimax for each $\Theta_j$, then $\theta^*(y) = (\theta_j^*(y_j))$ is minimax for $\Theta$.

Remarks: 1. There is something to prove here: among estimators $\hat{\theta}$ competing in the left side of (6.18), each coordinate $\hat{\theta}_j(y_j)$ may depend on all components $y_{j'}, j' \in J$. The result says that a minimax estimator need not exhibit such dependencies since $\theta_j^*(y_j)$ depends only on $y_j$.

2. The statement of this result does not involve prior distributions, and yet the simplest proof seems to need priors and the minimax theorem. A direct proof without priors is possible, but is more intricate. (include in Appendix?)

Proof. By the minimax theorem (6.6):

$$R_N(\Theta) = \sup\{B(\pi), \pi \in \mathcal{P}(\Theta)\},$$

where $\mathcal{P}(\Theta)$ denotes the collection of all probability measures supported in $\Theta$. Given any such prior $\pi$, construct a new prior $\tilde{\pi}$ as the product of the marginal distributions $\pi_j$ of $\theta_j$ under $\pi$. Lemma 6.2 shows that $\tilde{\pi}$ is more difficult than $\pi : B(\tilde{\pi}) \geq B(\pi)$. Because of the product structure of $\Theta$, each $\pi_j$ is supported in $\Theta_j$ and $\tilde{\pi}$ still lives on $\Theta$. Thus the maximization can be restricted to priors with independent coordinates. Bayes risk is then additive, by (6.6), so the optimization can be term-by-term:

$$R_N(\Theta) = \sum_j \sup\{B(\pi_j) : \pi_j \in \mathcal{P}(\Theta_j)\} = \sum_j R_N(\Theta_j).$$

The verification that separately minimax $\theta_j^*(y_j)$ combine to yield a minimax $\theta^*(y)$ can now be left to the reader.
6.4. The Bayes Minimax Method*

In this section we outline a general strategy for asymptotic evaluation of minimax risks \( R_N(\Theta) \) that will be useful in several settings.

We start with an upper bound, for fixed \( \epsilon \), based on the minimax theorem. Suppose that \( L(\theta, a) \) is convex in \( a \) for each \( \theta \in \ell_2 \). Let \( \mathcal{M} \) be a convex collection of probability measures on \( \ell_2(I) \) containing \( \Theta \) in the sense that point masses \( \delta_{\theta} \in \mathcal{M} \) for \( \theta \in \Theta \). Then, as we have seen at (6.16) and (6.15),

\[
R_N(\Theta, \epsilon) \leq B(\mathcal{M}, \epsilon) = \sup_{\pi \in \mathcal{M}} B(\pi).
\]

We call the right side the Bayes-minimax risk. Typically \( \mathcal{M} \) is defined by constraints on marginal moments: in general \( \mathcal{M} \) will not be supported on \( \Theta \). The idea is that a judiciously chosen relaxation of the constraints defining \( \Theta \) may make the problem easier to evaluate, and yet still be asymptotically equivalent to \( \Theta \) as \( \epsilon \to 0 \).

The main task, then, is to establish that \( R_N(\Theta, \epsilon) \sim B(\mathcal{M}, \epsilon) \) as \( \epsilon \to 0 \).

(a) Basic Strategy. Suppose that one can find a sequence \( \nu_\epsilon \) supported in \( \Theta \), that is nearly least favorable: \( B(\nu_\epsilon) \sim B(\mathcal{M}, \epsilon) \). Then asymptotic equivalence follows from the chain of inequalities

\[
B(\nu_\epsilon) \leq R_N(\Theta, \epsilon) \leq B(\mathcal{M}, \epsilon) \sim B(\nu_\epsilon).
\]

(b) Asymptotic Concentration. Often it is inconvenient to work directly with priors supported on \( \Theta \). Instead, one may seek a sequence \( \pi_\epsilon \in \mathcal{M} \) that is both asymptotically least favorable, \( B(\pi_\epsilon) \sim B(\mathcal{M}, \epsilon) \) and concentrates on \( \Theta \):

\[
\pi_\epsilon(\Theta) \to 1.
\]

Then construct the conditioned prior \( \nu_\epsilon = \pi_\epsilon(\cdot | \Theta) \). If one additionally shows that

\[
B(\pi_\epsilon) \sim B(\nu_\epsilon),
\]

then asymptotic equivalence follows by replacing the last similarity in (6.20) by \( B(\mathcal{M}, \epsilon) \sim B(\pi_\epsilon) \sim B(\nu_\epsilon) \).

In our examples, there is typically a scale family of minimax problems: \( \Theta(C) = C\Theta(1) \). We assume also that the corresponding prior family scales \( \mathcal{M}(C) = C\mathcal{M}(1) \). Let \( R(C, \epsilon) \) and \( B(C, \epsilon) \) denote the frequentist and Bayes minimax risks over \( \Theta(C) \) and \( \mathcal{M}(C) \) respectively. We exploit the scale structure by taking \( \pi_\epsilon \) as the least favorable prior for \( B(\gamma C, \epsilon) \) for some \( \gamma < 1 \).

Although \( \pi_\epsilon \) will typically not live on \( \Theta(\gamma C) \), it often happens that it is asymptotically concentrated on the larger set \( \Theta(C) \).

We now give some of the technical details needed to carry out this heuristic. Our setting is \( \ell_2 \) loss, but the argument can easily be generalized, at least to other norm based loss functions. Since \( C \) remains fixed, set \( \Theta = \Theta(C) \).

Let \( \pi_\epsilon \) be a prior distribution with \( B(\pi_\epsilon) \geq \gamma B(\gamma C, \epsilon) \) and \( \pi_\epsilon(\Theta) > 0 \). Set \( \nu_\epsilon = \pi_\epsilon(\cdot | \Theta) \), and let \( \hat{\theta}_\nu \) be the Bayes estimator of \( \theta \) for the conditioned prior
\( \nu \). The issue is to relate \( B(\nu) \) to \( B(\pi) \). From the frequentist definition of Bayes risk,

\[
B(\pi) \leq E_{\pi} \left\{ \| \hat{\theta}_\nu - \theta \|^2 | \Theta \right\} + 2E_{\pi} \left\{ \| \hat{\theta}_\nu - \theta \|^2, \Theta^c \right\} \\
\leq B(\nu)\pi_\nu(\Theta) + 2E_{\pi} \left\{ \| \hat{\theta}_\nu \|^2 + \| \theta \|^2, \Theta^c \right\}.
\]

Since \( \nu \) and its corresponding posterior live on the compact set \( \Theta \), we conclude that \( \| \hat{\theta}_\nu \|^2 \leq E(\| \theta \|^2 | y) \leq M \) is bounded. Since also \( B(\nu) \leq R(C, \epsilon) \), on putting everything together, we have

\[
\gamma B(\gamma C, \epsilon) \leq B(\pi) \leq R(C, \epsilon)\pi_\nu(\Theta) + E_{\pi} \left\{ M + \| \theta \|^2, \Theta^c \right\}.
\]

In summary, we now have a lower bound for the minimax risk.

6.8. LEMMA. Suppose that for each \( \gamma < 1 \) one chooses \( \pi_\epsilon \in \mathcal{M}(\gamma C) \) such that

\[
B(\pi_\epsilon) \geq \gamma B(\gamma C, \epsilon),
\]

\[
\pi_\epsilon(\Theta) \to 1,
\]

\[
E_{\pi_\epsilon} \left\{ 1 + \| \theta \|^2, \Theta^c \right\} = o(B(C, \epsilon)) \text{ as } \epsilon \to \infty.
\]

Then for each such \( \gamma \),

\[
R(C, \epsilon) \geq \gamma B(\gamma C, \epsilon)(1 + o(1))
\]

In general, it is easy to establish from simple scaling properties (see for example, Exercise 2) that

\[
\lim_{\gamma \to 1} \lim_{\epsilon \to 0} \inf \frac{B(\gamma C, \epsilon)}{B(C, \epsilon)} = 1.
\]

Combining (6.26) with (6.27), it follows that \( R(C, \epsilon) \sim B(C, \epsilon) \).

Remark. This approach will be applied in the proof of Pinsker’s theorem in Chapter 7. In other settings, slight modifications of the general procedure are needed. For example, \( \Theta \) and \( \mathcal{M} \) may depend on \( \epsilon \) (or on dimension \( n \), as in Chapter 14). We will see that the strategy may still be applied, even though the noise level in each coordinate may not be asymptotically small.

6.5. Appendix: Further details.

Proof of (6.11): We may of course suppose that \( I(P) < \infty \), which entails that the density \( p \) of \( P \) exists and is absolutely continuous, and permits integration by parts in the following chain:

\[
1 = \int p(y)dy = - \int (y - \mu)p'(y)dy \leq \int (y - \mu)^2 p(y)dy \int |p'(y)|^2 / p(y)dy,
\]

with equality if and only if

\[
(p'/p)(y) = (\log p)'(y) = c(y - \mu).
\]

Proof of (6.13): (taken from (Belitser & Levit 1995)). The argument is of the same flavor as the Fisher information bound (6.11). Of course,
by scaling arguments, we may reduce to \( \epsilon = 1 \). Let \( A = \hat{\theta}(y) - \theta \); and
\[
B = (\partial / \partial \theta)[\log \phi(y - \theta)p(\theta)].
\]
Then by Fubini’s theorem,
\[
E_\pi E_\theta AB = \int \int (\hat{\theta}(y) - \theta) \left( \partial / \partial \theta \right)[\log \phi(y - \theta)p(\theta)] \phi(y - \theta)p(\theta)d\theta dy
\]
\[
= \int \int (\hat{\theta}(y) - \theta) \left( \partial / \partial \theta \right)[\phi(y - \theta)p(\theta)] d\theta dy
\]
\[
= \int d\theta \int \phi(y - \theta)p(\theta)d\theta = 1.
\]
Now apply the Cauchy-Schwartz inequality: we have
\[
E_\pi E_\theta A^2 \geq B(\pi) \quad \text{and} \quad E_\pi E_\theta B^2 = 1 + I(\pi),
\]
which establishes (6.13). We note that improved bounds on the Bayes risk are given by Brown & Gajek (1990).

**Exercises**

1. **(Less noise is easier.)** Consider two versions of the sequence model:
\( y = \theta + \epsilon z \), and a lower noise version, \( y' = \theta + \epsilon' z' \), where \( \epsilon' < \epsilon \). Suppose that for each \( \theta \in \ell_2(I) \), the loss function \( L(a, \theta) \) is convex in \( a \).

   (a) Show that for each estimator \( \hat{\theta}(y) \), there is an estimator \( \hat{\theta}'(y') \) such that for all \( \theta \),
\[
r(\hat{\theta}', \theta; \epsilon') \leq r(\hat{\theta}, \theta; \epsilon).
\]

   (b) Conclude that for any parameter space \( \Theta \), and for any set of priors \( \mathcal{P} \),
\[
R_N(\Theta, \epsilon') \leq R_N(\Theta, \epsilon),
\]
\[
B(\mathcal{P}, \epsilon') \leq B(\mathcal{P}, \epsilon).
\]

2. **(Scaling bounds for risks.)** Consider \( y = \theta + \epsilon z \) and squared error loss. Suppose that \( \{ \Theta(C) \} \) is a scale family of parameter spaces in \( \ell_2(I) \), so that \( \Theta(C) = C\Theta(1) \) for \( C > 0 \). Use the abbreviation \( R(C, \epsilon) \) for \( R_N(\Theta(C); \epsilon) \).

   (a) Show that if \( C' \leq C \) and \( \epsilon' \leq \epsilon \), then
\[
R(C, \epsilon) \leq (C/C')^2(\epsilon/\epsilon')^2R(C', \epsilon'),
\]
and that if \( \mathcal{P}(C) = C\mathcal{P}(1) \) is a scale family of priors, that the same result holds for \( B(C, \epsilon) = B(\mathcal{P}(C); \epsilon) \).

   (b) Conclude that
\[
\lim_{\gamma \to 1} \lim_{\epsilon \to 0} \frac{B(\gamma C, \epsilon)}{B(C, \epsilon)} = 1.
\]
CHAPTER 7

Intervals, Hyperrectangles and Quadratic Convexity

A theme of these notes is that conclusions about function estimation can sometimes be built up from very simple, even one dimensional, parametric constituents. We will see this technique at work in this chapter – starting with minimaxity on a bounded interval in a single dimension and progressing through hyperrectangles to more complex quadratically convex sets in \( \ell_2(\mathbb{N}) \).

In this chapter, we confine attention to squared error loss.

7.1. Single Bounded Normal Mean

If \( y \sim N(\theta, \sigma^2) \), and there is no constraint on \( \theta \), then it is a classical result that the minimax mean squared error for estimation of \( \theta \) based on \( y \) equals the variance \( \sigma^2 \). Suppose now that \( \theta \) is known to lie in a bounded interval of length \( 2\tau \), which without any real loss of generality, we may assume to be centered about 0: \( \Theta(\tau) = [-\tau, \tau] \). It is clear that any estimator \( \hat{\theta} \), whether linear or not, can be improved simply by enforcing the interval constraint: if \( \hat{\theta} = [\hat{\theta}]_{[-\tau, \tau]} = \max\{\min\{\hat{\theta}, \tau\}, -\tau\} \), then \( r(\hat{\theta}, \theta) \leq r(\theta, \theta) \). This section asks: how much better is the nonlinear minimax risk

\[
\rho_N(\tau, \epsilon) = \inf_{\theta} \sup_{\hat{\theta} \in [\theta - \tau, \theta + \tau]} E(\hat{\theta} - \theta)^2,
\]

than the corresponding linear minimax risk \( \rho_L(\tau, \epsilon) \) obtained by restricting \( \theta \) to linear estimators of the form \( \hat{\theta}_c(y) = cy \)?

7.1.1. Linear Estimators. For squared error loss, the variance-bias decomposition of mean squared error is a common tool for evaluating risks:

\[
E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2
\]

\[
= E(\hat{\theta} - E\hat{\theta})^2 + (E\hat{\theta} - \theta)^2 = \text{Var} \hat{\theta} + \text{Bias}^2(\hat{\theta}).
\]

If this is applied to a linear estimator \( \hat{\theta}_c(y) = cy \), we obtain

(7.1) \[
E(\hat{\theta}_c - \theta)^2 = c^2\epsilon^2 + (1 - c)^2\theta^2.
\]

If the parameter is known to lie in a bounded interval \( \Theta(\tau) = [-\tau, \tau] \), then the maximum risk occurs at the endpoints:

\[
\sup_{\theta \in [\tau]} E(\hat{\theta}_c - \theta)^2 = c^2\epsilon^2 + (1 - c)^2\tau^2 = r(\hat{\theta}_c, \tau).
\]
The minimax linear estimator is thus found by minimizing the quadratic function \( c \rightarrow r(\hat{\theta}_c, \tau) \). It follows that
\[
(7.2) \quad \rho_L(\tau, \epsilon) = \inf_{\epsilon} r(\hat{\theta}_c, \tau) = \frac{\epsilon^2 \tau^2}{\epsilon^2 + \tau^2}.
\]
with the minimum occurring at \( \epsilon^* = \tau^2/(\epsilon^2 + \tau^2) \in (0, 1] \), and the corresponding minimax linear estimator
\[
(7.3) \quad \hat{\theta}_{LIN}(y) = \frac{\tau^2}{\epsilon^2 + \tau^2} y.
\]
Thus, if the prior constraint information is that \( \tau^2 \ll \epsilon^2 \), then a large amount of linear shrinkage is indicated, while if \( \tau^2 \gg \epsilon^2 \), then essentially the unbiased estimator is used. When \( \tau^2 = \epsilon^2 \), the minimax linear risk is exactly \( \frac{1}{2} \) that if there is no constraint on \( \tau \).

Note the scale invariance relation
\[
(7.4) \quad \rho_L(\tau, \epsilon) = \epsilon^2 \rho_L(\tau/\epsilon, 1).
\]
Writing \( \nu^2 = \tau^2/\epsilon^2 \) for the squared signal-to-noise ratio, we have
\[
(7.5) \quad \rho_L(\nu, 1) = \nu^2/(1 + \nu^2) \sim \begin{cases} \nu^2 & \nu \to 0 \\ 1 & \nu \to \infty. \end{cases}
\]
These results, however simple, are nevertheless a first quantitative indication of the importance of prior information, here quantified through \( \nu \), on possible quality of estimation.

**Linear Estimators are Bayes.** If \( \theta \) has a Gaussian prior distribution \( \theta \sim N(0, \omega^2) \), and if \( y|\theta \sim N(\theta, \epsilon^2) \), then, as also noted in Section 4.5, the posterior distribution is again Gaussian:
\[
L(\theta|y) = \frac{\omega^2}{\epsilon^2 + \omega^2 y} \frac{\epsilon^2 \omega^2}{\epsilon^2 + \omega^2}.
\]
The Bayes estimator, for squared error loss, is linear:
\[
\hat{\theta}_\pi(y) = E_\pi(\theta|y) = \frac{\omega^2}{\epsilon^2 + \omega^2} y.
\]
In particular, setting \( \omega = \tau \), we see that the estimator that is linear minimax for \( \Theta(\tau) = [-\tau, \tau] \) is also Bayes for a prior \( \pi_\tau(d\theta) = N(0, \tau^2) \). The linear minimax risk \( \rho_L(\tau, \epsilon) \) equals the posterior variance. Note that this prior is not concentrated on \( \Theta(\tau) \): only a moment statement is possible: \( E_\pi \theta^2 = \tau^2 \).

**Remark.** If \( \theta(y) = \epsilon y \) is a linear estimator that is Bayes for some prior \( \pi(d\theta) \) under squared error loss, then it can be shown that the prior \( \pi \) is necessarily Gaussian. This property is a special case of a general phenomenon for exponential families: linear estimators are Bayes if and only if the prior comes from the conjugate prior family associated with that exponential family (Diaconis & Ylvisaker 1979).

**Projection Estimators.** Orthogonal projections form an important and simple subclass of linear estimators. In one dimension the situation
is almost trivial - there are only two possibilities: \( \hat{\theta}_0(y) \equiv 0 \) with risk 
\( r(\hat{\theta}_0, \theta) = \theta^2 \) - the pure bias case; or \( \hat{\theta}_1(y) = y \), with risk \( r(\hat{\theta}_1, \theta) = \epsilon^2 \) - the 
case of pure variance. Nevertheless, one can usefully define and evaluate the 
minimax risk over \( \Theta = [-\tau, \tau] \) for projection estimators 
\[
(7.6) \quad \rho_p(\tau, \epsilon) = \inf_{\epsilon \in [0,1]} \sup_{\theta \in [-\tau, \tau]} E(\hat{\theta}_\epsilon - \theta)^2 = \min(\tau^2, \epsilon^2).
\]
Indeed, if the signal to noise ratio \( \tau/\epsilon \) exceeds 1, use \( \hat{\theta}(y) = y \), otherwise use 
\( \hat{\theta} = 0 \).
The inequalities 
\[
\frac{1}{2} \min(\tau^2, \epsilon^2) \leq \frac{\tau^2 \epsilon^2}{\tau^2 + \epsilon^2} \leq \min(\tau^2, \epsilon^2)
\]
imply immediately that 
\[
(7.7) \quad \rho_l(\tau, \epsilon) \in \left[ \frac{1}{2}, 1 \right] \rho_p(\tau, \epsilon)
\]
so that the best projection estimator is always within a factor of 2 of the 
best linear estimator.

7.1.2. Non-linear minimaxity. Of course, rather than confining at-
tention to linear methods, we should be willing to consider arbitrary, non-
linear estimators of \( \theta \). Again assuming that \( \theta \) lies in the interval \( [-\tau, \tau] \), 
define 
\[
(7.8) \quad \rho_N(\tau, \epsilon) = \inf_{\theta} \sup_{\theta \in [-\tau, \tau]} E(\hat{\theta} - \theta)^2.
\]
In contrast to the linear quantity, \( \rho_N(\tau, \epsilon) \) cannot be evaluated explicitly in 
general. However the following properties are easy enough:
\[
(7.9) \quad \rho_N(\tau, \epsilon) \leq \rho_l(\tau, \epsilon), \\
(7.10) \quad \rho_N(\tau, \epsilon) = \epsilon^2 \rho_N(\tau, 1), \\
(7.11) \quad \rho_N(\tau, \epsilon) \text{ is increasing in } \tau, \\
(7.12) \quad \lim_{\tau \to \infty} \rho_N(\tau, \epsilon) = \epsilon^2.
\]
Indeed (7.9) is plain since more estimators are allowed in the nonlinear 
competition, while (7.10) follows by rescaling, and (7.11) is obvious.

The classical result (3.18) says that the minimax risk for \( \theta \) unconstrained
to any interval, \( \rho_N(\infty, \epsilon) = \epsilon^2 \). Thus (7.12) asserts continuity as \( \tau \) increases 
without bound. To prove it, we choose an absolutely continuous distribution 
\( F_1 \) on \([-1,1]\) and rescale it to \( F_\tau \) on \([-\tau, \tau]\); the scaling properties of Fisher 
information ensure that \( I(F_\tau) = I(F_1)/\tau^2 \). The Bayes bound (6.13) together 
with the van Trees bound (6.13) entail 
\[
(7.13) \quad \rho_N(\tau, 1) \geq B(F_\tau, 1) \geq [1 + I(F_\tau)]^{-1} = [1 + I(F_1)/\tau^2]^{-1} \to 1.
\]
and this suffices for (7.12).
Least favorable priors are discrete. The fine structure of minimax rules is in general complicated, although some useful information is available. We begin with a well known criterion for identifying minimax rules.

7.1. Lemma. Given a prior distribution $\pi$, let $M_\pi$ be the set where the Bayes rule attains its maximum risk: $M(\pi) = \{\theta \in \Theta : r(\hat{\theta}_\pi, \theta) = \max_{\theta \in \Theta} r(\hat{\theta}_\pi, \theta)\}$. If $\pi(M_\pi) = 1$, then $\hat{\theta}_\pi$ is the unique minimax rule.

Proof. This is a standard result Lehmann (1983, e.g. Theorem 4.2.1 and Corollary). In particular, from Proposition 6.1, $\hat{\theta}_\pi$ is the unique Bayes rule for $\pi$, and so is unique minimax.

Suppose then that $(\pi_\tau, \hat{\theta}_\tau)$ is a saddlepoint for (7.8), estimation on $[-\tau, \tau]$: $\pi_\tau(d\theta)$ is a least favorable prior and $\hat{\theta}_\tau$ the corresponding minimax estimator. The lemma says that the support of $\pi_\tau$ is contained in the set $M = M(\pi_\tau) \subset [-\tau, \tau]$ of points where the risk function $\theta \rightarrow r(\hat{\theta}_\pi, \theta)$ attains its maximum. The smoothing effect of the Gaussian distribution implies that this risk function is analytic. Furthermore, it cannot be constant on $[-\tau, \tau]$ – the only estimator with constant risk is $\hat{\theta}(y) = y$, and the linear case (7.5) shows that this estimator cannot be minimax for $\tau < \infty$. Hence the set $M$ and thus supp $\pi_\tau$ cannot have any accumulation points in $[-\tau, \tau]$, and so must be a finite set! In general, this finite set and the corresponding minimax estimator can only be determined numerically (see Donoho et al. (1990), Goudin et al. (1994).)

Since the posterior distribution of $\pi_\tau$ must also live on this finite set, and since the mean squared error of $\hat{\theta}_\tau$ must be everywhere less than $\epsilon^2$, one guesses heuristically that the support points of $\pi_\tau$ will be spaced at a distance on the scale of the noise standard deviation $\epsilon$.

For small $\tau$, then, one expects that there will be only a small number of support points, as was shown explicitly by Casella & Strawderman (1981). This observation will be an important baseline for our later study of the least favorable character of sparse signal representations, so we give an outline of the argument. Without loss of generality, set $\epsilon = 1$.

1. From Lemma 7.1, it follows that the symmetric two point prior $\pi_\tau^{(2)} = (1/2)(\delta_\tau + \delta_{-\tau})$ is minimax if $\{-\tau, \tau\} \subset M(\pi_\tau^{(2)})$. For this two point prior, one may calculate (Exercise!) that

$$\pi(\{\tau\}|y) = e^{\tau y}/(e^{\tau y} + e^{-\tau y}),$$

$$\hat{\theta}_\pi(y) = E(\theta|y) = \tau \tanh \tau y,$$

$$B(\pi_\tau) = \tau^2 e^{-\tau^2}/2 \int \phi(y)dy \cosh y\tau$$

(7.14)

2. Heuristically, it is evident that for all $\tau$, $M(\pi_\tau^{(2)}) \subset \{-\tau, 0, \tau\}$. The formal proof uses a sign change argument linked to total positivity of the Gaussian distribution – see Casella & Strawderman (1981).
3. A second sign change argument shows that for $|\tau| < \tau_2$,

$$r(\hat{\theta}^{(2)}_\tau, 0) < r(\hat{\theta}^{(2)}_\tau, \theta).$$

Thus $\hat{\theta}^{(2)}_\tau$ is minimax for $|\tau| < \tau_2$, with numerical work showing that $\tau_2 = 1.05$.

This completes the story for symmetric two point priors. In fact, Casella and Strawderman go on to show that for $\tau_2 \leq |\tau| < \tau_3$, an extra atom of the prior distribution appears at 0, and $\pi_\tau$ has the three-point form

$$\pi^{(3)}_\tau = (1 - \alpha)\delta_0 + (\alpha/2)(\delta_\tau + \delta_{-\tau}).$$

This three point prior appears again in Section ?? below.

As $|\tau|$ increases, prior support points are added successively (numerical calculations may be found in Gourdin et al. (1994). If the least favorable distributions $\pi_\tau$ are rescaled to $[-1, 1]$ by setting $\nu_\tau(A) = \pi_\tau(\tau A)$, then Bickel (1981) derives the weak limit

$$\nu_\tau \Rightarrow \nu_\infty, \quad \nu_\infty(ds) = \cos^2(\pi s/2)ds, \quad |s| \leq 1,$$

and shows that $\rho_N(\tau, 1) = 1 - \pi^2/\tau^2 + O(\tau^{-2})$ as $\tau \to \infty$.

Remark. Levit (1980, 1982, 1985) and Berkhin & Levit (1980) developed a more extensive theory of second order asymptotic minimax estimation of a $d$-dimensional Gaussian mean. Quite generally, they showed that the second order coefficient (here $\pi^2$), could be interpreted as twice the principal eigenvalue of the Laplacian (here $\omega(t) = -2d^2/dt^2$) on the fundamental domain (here $[-1, 1]$), with the asymptotically least favorable distribution having density the square of the principal eigenfunction, here $\omega(t) = \cos(\pi t/2)$. We do not delve further into this beautiful theory since it is essentially parametric in nature: in the nonparametric settings to be considered in these notes, we are still concerned with understanding the first order behaviour of the minimax risk with noise level $\epsilon$ or sample size $n$.

7.1.3. Near minimaxity of linear estimators. In spite of the complex structure of non-linear minimax rules, it is remarkable that they do not, in this univariate setting, offer great improvements over linear estimators.

7.2. Theorem.

(7.15) \[ \mu^* := \sup_{\tau \neq 0} \frac{\rho_L(\tau, \epsilon)}{\rho_N(\tau, \epsilon)} \leq 1.25. \]

The bound $\mu^* < \infty$ is due to Ibragimov & Khas’minskii (1984): the extra work - some numerical - needed to obtain the essentially sharp bound 1.25 is outlined in Donoho et al. (1990) along with references to other work on the same topic.

Proof. We use projection estimators and the identity (7.14) for the two point priors $(1/2)(\delta_\tau + \delta_{-\tau})$ to give a short and instructive proof that $\mu^* \leq 1/B(\pi_1)$, which numerical evaluation of the integral (7.14) shows to be approximately 2.22.
First, it is enough to take $\epsilon = 1$, in view of the scaling invariances (7.4) and (7.10). So consider the ratio $\mu(\tau) = \rho_L(\tau, 1) / \rho_N(\tau, 1)$. For the numerator we have $\rho_L(\tau, 1) \leq \rho_P(\tau, 1) = \min(\tau^2, 1)$ by the simple properties of projection estimators (c.f. (??) and (7.6)). Hence
\[
\mu(\tau) \leq \min(\tau^2, 1) / \rho_N(\tau, 1) =: \mu_P(\tau).
\]
Now use (??) to minorize $\rho_N(\tau, 1)$ by $B(\pi)$, where $\pi$ is the two point prior $(1/2)(\delta_0 + \delta_{-\tau})$ discussed earlier. For $\tau \geq 1$, since $\rho_N(\tau, 1)$ increases with $\tau$,
\[
\mu_P(\tau) = 1 / \rho_N(\tau, 1) \\
\leq 1 / \rho_N(1, 1) \leq 1 / B(\pi).
\]
For $\tau \leq 1$, (7.14) shows that $\tau^2 / B(\pi)$ is increasing in $\tau$, and so
\[
\mu_P(\tau) = \tau^2 / \rho_N(\tau, 1) \leq \tau^2 / B(\pi) \\
\leq 1 / B(\pi).
\]

Remark. The proof also gives sharper information for small and large $\tau$: indeed, the linear minimax risk is then essentially equivalent to the non-linear minimax risk:
\[
\mu(\tau) = \rho_L(\tau, 1) / \rho_N(\tau, 1) \to 1 \quad \text{as } \tau \to 0, \text{ or } \tau \to \infty.
\]
Indeed, for small $\tau$, the right side of (7.18) approaches 1 because of (7.14), while for large $\tau$, the same limit results from (7.17) and the limit (7.12).

Diagram of $\rho_P, \rho_L$ and $\rho_N$

7.2. Hyperrectangles

The set $\Theta \subset \ell_2(I)$ is said to be a hyperrectangle if
\[
\Theta = \Theta(\tau) = \{\theta : |\theta_i| \leq \tau_i \text{ for all } i \in I\}.
\]
For $\Theta(\tau)$ to be compact, it is necessary and sufficient that $\sum \tau_i^2 < \infty$ (Example 3.2.) We suppose that data $y$ from the Gaussian model (3.1) is observed.

7.3. Example. Hyperrectangles and smoothness. If $(\theta_i)$ represent the coefficients of a function $f$ in an appropriate orthonormal basis, then the rate of decay of $\tau_i$ can correspond to smoothness information about $f$.

For periodic functions on $[0, 1]$, the Fourier basis is natural. If $f$ is $C^\alpha$, in the sense of Hölder continuity (see Appendix), then the Fourier coefficients $|\theta_k| \leq M k^{-\alpha}$ for some constant $M$ (e.g. Katzenelson (1968, p. 25)). However, the converse fails, so Fourier hyperrectangles do not exactly capture Hölder smoothness. On the other hand, a periodic function $f$ is analytic if and only if there exist positive constants $M$ and $a$ so that $|\theta_k| \leq Me^{-ak}$ for all $k$ (e.g. Katzenelson (1968, p. 26)). However, analyticity conditions are less used in nonparametric theory than are constraints on a finite number of derivatives. [See Section ?? for some discussion??]
From this perspective, the situation is much better for wavelet bases, since Hölder smoothness is exactly characterized by hyperrectangle conditions. Let $(\theta_{jk})$ for $j \geq 0$ and $k = 1, \ldots, 2^j$ be the coefficients of $f$ in an orthonormal wavelet basis for $L_2[0,1]$ of regularity $m > \alpha$ (for example a CDJV basis discussed in Appendix). Then $f$ is $C^\alpha$ if and only if for some constant $C$, the coefficients $(\theta_{jk})$ leading to the hyperrectangle

\begin{equation}
\Theta^\alpha(C) = \{(\theta_{jk}) : |\theta_{jk}| \leq C2^{-\lfloor (\alpha+1)/2 \rfloor j}, j \in \mathbb{N}, k = 1, \ldots, 2^j\}.
\end{equation}

The subscript $\infty$ indicates that the bounds hold for all $(j,k)$ and emphasizes that Hölder continuity measures uniform smoothness.

### 7.2.1. Linear Estimates

In principle, a linear estimator has the form $\hat{\theta}_C(y) = Cy$, for $C$ a linear operator on $\mathbb{R}^I$, and one might even be tempted to consider affine estimators

\begin{equation}
\hat{\theta}_C(y) = Cy + b \quad b \in \ell_2(I),
\end{equation}

with risk function given from the variance-bias decomposition by

\begin{equation}
E_{\theta} \|Cy + b - \theta\|^2 = \epsilon^2 \text{tr} CC^T + (C\theta - \theta + b)^T (C\theta - \theta + b).
\end{equation}

However because of the symmetry properties of hyperrectangles, we will see that it suffices for minimax purposes to consider diagonal linear estimators

\begin{equation}
\hat{\theta}_c(y) = (c_i y_i)_{i \in I},
\end{equation}

whose mean squared error has the much simpler additive form

\begin{equation}
\tau(\hat{\theta}_c, \theta) = \sum_i c_i^2 \epsilon^2 + (1 - c_i)^2 \theta_i^2.
\end{equation}

#### 7.4. Lemma

On any hyperrectangle, an estimator that is minimax among affine estimators (7.21) must have diagonal linear form (7.23).

For the admissibility result Theorem 3.4, all that was required was that a linear estimator be diagonal in some orthonormal basis. For minimaxity on a hyperrectangle $\Theta(\tau)$, which has product structure in a given basis, the estimator needs to be diagonal in this basis.

**Proof.** The argument uses a simple but generally useful random signs technique. Indeed, given $\tau \in \ell_2(I)$, let $\pi_\tau$ be the prior distribution on $\theta$ obtained by choosing $\theta_i = \pm \tau_i$ independently with probability $1/2$. $\pi$ is concentrated on the vertex set $V(\tau) = \{\theta : |\theta_i| = |\tau_i|, \forall i \in I\}$ of the corresponding hyperrectangle. Since the coordinates $\theta_i$ are now independent mean 0, when we average (7.22) with respect to $\pi_\tau$, the cross terms $E_{\pi_\tau} \theta_i \theta_j$ vanish for $i \neq j$, and it follows that the average risk can be reduced by
dropping non-diagonal terms and ignoring the shift \( b \)

\[
\sup_{V(\tau)} r(\hat{\theta}_{C,b}, \theta) \geq E_{\pi} ||C y + b - \theta||^2 
\]

\[
e^2 \sum_{j,k} c_{j,k}^2 + \sum_j r_j^2 \left[ \sum k (c_{j,k} - \delta_{j,k})^2 \right] + \tilde{b}_j^2 
\]

\[
\geq e^2 \sum_j c_{jj}^2 + \sum_j r_j^2 (c_{jj} - 1)^2. 
\]

The last inequality will be strict if any off-diagonal \( c_{jk} \) is non-zero. The final sum may be further reduced by constraining \( c_{jj} \) to the interval \([0,1]\). Let \( C^0 \) denote the diagonal operator thus obtained:

\[
C^0 = \begin{pmatrix} 
    c_{j,j}^0 & & \\
    & \ddots & \\
    & & c_{n,n}^0 
\end{pmatrix} = \begin{pmatrix} 
    \delta_{j,j} \max(0, \min(c_{jj}, 1)) & & \\
    & \ddots & \\
    & & \delta_{n,n} \max(0, \min(c_{nn}, 1)) 
\end{pmatrix}.
\]

The risk of a diagonal linear estimator is identical at all the vertices of \( V(\tau) \) — compare (7.24) — and so the average risk equals the minimax risk. We have thus established that for all vertex sets \( V(\tau) \),

\[
\sup_{V(\tau)} r(\hat{\theta}_{C_0,0}, \theta) \leq \sup_{V(\tau)} r(\hat{\theta}_{C,b}, \theta). 
\]

Given a hyperrectangle \( \Theta(\tau) \), this conclusion can be applied to the vertex set \( V(\theta) \) constructed from any \( \theta \in \Theta(\tau) \).

After this reduction to diagonal linear estimators, it is straightforward to show that the minimax linear risk for a hyperrectangle is found by adding the component one-dimensional risks.

**7.5. Proposition.**

\[
R_L(\Theta(\tau), \epsilon) = \sum_i \rho_L(\tau_i, \epsilon) = \sum_i e^2 r_i^2 \left/(e^2 + r_i^2 \right) 
\]

**Proof.** The lemma allows us to focus on diagonal estimators. From (7.24),

\[
\sup_{\Theta(\tau)} r(\hat{\theta}_i, \theta) = \sum_i c_i^2 e^2 + (1 - c_i) r_i^2. 
\]

The minimax linear risk \( R_L(\Theta(\tau)) \) is then found by minimizing term by term over \( c_i \) — the relation (7.29) then follows from the univariate version (7.2). \( \square \)

**7.2.2. Non-linear estimators.** Although we cannot expect to evaluate the non-linear minimax risk for a hyperrectangle \( \Theta(\tau) \) exactly, it is possible to exploit the product structure as at (6.18) to express the minimax risk in terms of the one-dimensional component problems. From Proposition 6.7 we obtain

**7.6. Proposition.**

\[
R_N(\Theta(\tau), \epsilon) = \inf_{\theta} \sup_{\theta \in \Theta(\tau)} E \|\hat{\theta} - \theta\|_2^2 = \sum \rho_N(\tau_i, \epsilon). 
\]
The Ibragimov-Hasminskii theorem lifts from intervals to hyperrectangles; simply apply Theorem 7.2 term by term: \( \rho_L(\tau, \epsilon) \leq \mu^* \rho_N(\tau, \epsilon) \). We conclude that, over hyperrectangles, the best non-linear estimate cannot significantly improve on the best linear one.

7.7. COROLLARY.

\[
R_L(\Theta(\tau), \epsilon) \leq \mu^* R_N(\Theta(\tau), \epsilon); \quad \mu^* \leq 1.25.
\]

Comments about behavior on hyperrectangles as \( \epsilon \to 0 \).

7.2.3. Rates of Convergence on Hölder classes. As a first and easy consequence of minimaxity results for hyperrectangles, we derive the optimal rate of convergence for estimation over Hölder classes.

Notation. The expression \( a(\epsilon) = \gamma b(\epsilon) \) means that there exist positive constants \( \gamma_1 < \gamma_2 \) depending only on \( \alpha \) such that for all \( \epsilon \),

\[
\gamma_1 \leq a(\epsilon)/b(\epsilon) \leq \gamma_2.
\]

The constant \( \gamma \) may not be the same at each appearance.

7.8. PROPOSITION. For Hölder balls \( \Theta^\alpha(\epsilon) \) defined at (7.20),

\[
R_N(\Theta^\alpha(\epsilon), \epsilon) = \gamma C^{2/(2\alpha+1)} (\epsilon^2)^{2\alpha/(2\alpha+1)}.
\]

Since the minimax risk tracks the best possible performance among all estimators, this result is a first justification for the statement that, for example, \( n^{-4/5} \) is the “right” rate of convergence for twice differentiable functions. In Chapter 8, we give more precise results for smoothness measured in a mean-square sense.

Corollary 7.7 says that the same rate of convergence applies to the minimax linear risk, and it is this that we really calculate.

PROOF. \( \Theta = \Theta^\alpha(\epsilon) \) is a hyperrectangle with semi-side lengths \( \tau_{jk} = C2^{-(\alpha+1/2)j} \) that do not depend on \( k \). Hence from (7.30),

\[
R_N(\Theta) = \sum_j 2^j \rho_N(C2^{-(\alpha+1/2)j}, \epsilon).
\]

We will use a simple sandwich bound \( \rho_N(\tau, \epsilon) \in [(2\mu^*)^{-1}, 1] \tau^2 \wedge \epsilon^2 \), obtained by combining the Ibragimov-Hasminskii bound (7.15) with the projection bound (7.7). The balance between the variance term \( \epsilon^2 \) at low \( j \) and the bias terms at \( \tau_{jk}^2 \) at high \( j \) comes at \( j_\star \in \mathbb{R} \) given by

\[
C^{2-(\alpha+1/2)j_\star} = \epsilon.
\]

Since \( j_\star \) is in general not an integer, set \( j_0 = \lfloor j_\star \rfloor \), so that

\[
R_N(\Theta) = \gamma \sum_{j \leq j_0} 2^j \epsilon^2 + \gamma \sum_{j > j_0} 2^{-2\alpha j}.
\]
These geometric sums are dominated by their leading terms, multiplied by constants depending only on $\alpha$. Since also $2^{j_0} \in [\frac{1}{2}, 1]$ $2^{j_i}$, and $2^{-2\alpha j_0} \in [2^{-2\alpha}, 1]$ $2^{-2\alpha j_i}$,

$$R_N(\Theta) = \gamma [2^{j_0} \varepsilon^2 + C2^{-2\alpha j_i}],$$

from which (7.32) follows by substituting $j_i$. \hfill \Box

7.3. Orthosymmetry and Hardest rectangular subproblems

Although the minimax structure of hyperrectangles is, as we have just seen, essentially trivial, they are a key tool for obtaining deeper results on minimax risks for more general sets satisfying certain symmetry and convexity properties that we now define.

$\Theta$ is said to be solid and orthosymmetric if $\theta \in \Theta$ and $|\xi_i| \leq |\theta_i|$ for all $i$ implies that $\xi \in \Theta$ also. If a solid, orthosymmetric $\Theta$ contains a point $\tau$, then the same is true for the entire hyperrectangle that it defines: $\Theta(\tau) \subseteq \Theta$.

Examples of solid orthosymmetric sets.

(i) Sets defined by the contours of symmetric increasing functions.

Thus, if $\psi$ is increasing on $\mathbb{R}^+$, then \( \{ \theta : \sum \psi(\theta_i^2) \leq 1 \} \) is solid and orthosymmetric.

(ii) $\ell_p$ bodies: defined by \( \sum_i a_i^p |\theta_i|^p \leq C^p \) for $p > 0$, and

(iii) Besov bodies: defined by \( \sum_j 2^{jsp} (\sum_k |\theta_{jk}|^p)^{q/p} \leq C^q \) for $0 < p, q \leq \infty$.

Nonlinear minimax risk. Since $\Theta$ contains $\Theta(\tau)$ for each $\tau \in \Theta$, it is clear that $R_N(\Theta) \geq R_N(\Theta(\tau))$. Consequently, a simple but often useful lower bound to the non-linear minimax risk is obtained by restricting attention to the hardest rectangular subproblem of $\Theta$:

$$R_N(\Theta) \geq \sup_{\tau \in \Theta} R_N(\Theta(\tau)). \tag{7.33}$$

Linear minimax risk. First, we may again confine attention to diagonal linear estimators: orthosymmetry means that whenever $\tau \in \Theta$, so also is the vertex set $V(\tau)$. Hence any affine linear estimate $\hat{\theta}_{C, b}$, as in (7.21), can be improved upon by $\hat{\theta}_{C^0, b}$, as in (7.28), using Lemma 7.4:

$$\sup_{\theta \in \Theta} r(\hat{\theta}_{C, b}, \theta) = \sup_{\tau \in \Theta} \sup_{\theta \in V(\tau)} r(\hat{\theta}_{C, b}, \theta) \geq \sup_{\tau \in \Theta} r(\hat{\theta}_{C^0, b}, \theta).$$

To summarize: on any solid, orthosymmetric set, a linear minimax estimator must have diagonal form (7.23):

$$R_L(\Theta) = \inf_{\theta \in \Theta} r(\hat{\theta}_c, \theta). \tag{7.34}$$

Quadratic convexity. To relate the linear minimax risk of $\Theta$ to that of the rectangular subproblems $\Theta(\tau)$, we need an extra convexity property. $\Theta$ is said to be quadratically convex if $\Theta^2 = \{ \theta^2 : \theta \in \Theta \}$ is convex. Examples include sets of the form $\{ \theta : \sum a_i \psi(\theta_i^2) \leq 1 \}$ for $\psi$ a convex function. This
makes it clear that quadratic convexity is a stronger property than ordinary (linear) convexity. Particular examples include

(i) $\ell_p$ bodies: for $2 \leq p \leq \infty$, and
(ii) Besov bodies: for $2 \leq p \leq q \leq \infty$.

Just as in (7.33) the linear minimax risk over $\Theta$ is clearly bounded below by that of the hardest rectangular subproblem. However, for quadratically convex $\Theta$, the linear difficulties are actually equal:

7.9. Theorem. (Donoho et al., 1990) If $\Theta$ is compact, solid orthosymmetric and quadratically convex, then

\[(7.35) \quad R_L(\Theta) = \sup_{\tau \in \Theta} R_L(\Theta(\tau)).\]

Proof. First note that (7.35) is in fact a minimax theorem: compare (7.34) with the result of Proposition 7.5 and the definition (7.2):

\[\sup_{\tau \in \Theta} R_L(\Theta(\tau)) = \sup_{\tau \in \Theta} \inf_{(c_1)} \sum c_i^2 \tau_i^2 + (1 - c_i)^2 \tau_i^2 = \inf_{\theta \in \Theta} R_L(\theta, \tau).\]

We simply apply a classical minimax theorem (Corollary A.4) with payoff function

\[f(c, s) = \sum c_i^2 \tau_i^2 + (1 - c_i)^2 s_i,\]

which is convex-concave – indeed, even linear in the second argument – with $c \in \ell_2$ and $s \in \Theta^2 \subset \ell_1$. The latter set is convex and $\ell_1$-compact by the assumption that $\Theta$ is $\ell_2$-compact. Finally, $f(c, s)$ is trivially $\ell_1$-continuous in $s$ for fixed $c$. \hfill $\square$

Example. Let $\Theta_{n, 2}(C)$ denote an $\ell_2$ ball of radius $C$ in $\mathbb{R}^n$: $\{\theta : \sum \theta_i^2 \leq C^2\}$. Theorem 7.9 says that

\[R_L(\Theta_{n, 2}(C), \epsilon) = \sup \left\{ \epsilon^2 \sum \frac{n^2}{\epsilon^2 + \tau_i^2} : \sum \tau_i^2 \leq C^2 \right\},\]

and since $s \to s/(1 + s)$ is concave, it is evident that the maximum is attained at the vector with symmetric components $\tau_i^2 = C^2/n$. Thus,

\[R_L(\Theta_{n, 2}(C), \epsilon) = \frac{C^2}{n \epsilon^2 + C^2},\]

which grows from 0 to the unrestricted minimax risk $n \epsilon^2$ as the signal-to-noise ratio $C^2/n \epsilon^2$ increases from 0 to $\infty$.

While the norm ball in infinite sequence space, $\Theta_2(C) = \{\theta \in \ell_2 : ||\theta||_2 \leq C\}$ is not compact, the preceding argument does yield the lower bound

\[R_L(\Theta_2(C), \epsilon) \geq C^2,\]

which already shows that no linear estimate can be uniformly consistent as $\epsilon \to 0$ over all of $\Theta_2(C)$. Section 8.3 contains an extension of this result.

Combining (7.35), (7.31) and (7.33), we immediately obtain a large class of sets for which the linear minimax estimator is almost as good as the non-linear minimax rule.
7.10. COROLLARY. For such \( \Theta \), \( R_L(\Theta) \leq \mu^t R_N(\Theta) \).

This collection includes \( \ell_p \) bodies for \( p \geq 2 \) and so certainly ellipsoids, solid spheres, etc. and the Besov bodies just discussed.

7.3.1. Non quadratically convex sets. For any set \( \Theta \subset \ell_2(I) \), we recall the notation for the “square” of \( \Theta \), namely \( \Theta_+^2 = \{ (\theta_i)^2 : \theta \in \Theta \} \). The quadratically convex hull of \( \Theta \) is then defined as

\[
QHull(\Theta) = \{ \theta : (\theta_i)^2 \in Hull(\Theta_+^2) \},
\]

where \( Hull(S) \) denotes the closed convex hull of \( S \). Of course, if \( \Theta \) is closed and quadratically convex, then \( QHull(\Theta) = \Theta \). However, for \( \ell_p \)-bodies with \( p < 2 \),

\[
QHull(\Theta_p(a)) = \{ \theta : \sum a_i^2 \theta_i^2 \leq 1 \}
\]

is an ellipsoid. The key property of quadratic convexification is that it preserves the risk of linear estimators.

7.11. THEOREM. Let \( \Theta \) be solid orthosymmetric and compact. Then

\[
R_L(\Theta, \epsilon) = R_L(QHull(\Theta), \epsilon).
\]

PROOF. Since \( \Theta \) is orthosymmetric, Lemma 7.4 shows that linear minimax estimators may be found that are diagonal, with risk functions given by (7.24). Such risk functions are linear in \( s = (\theta_i^2) \) and hence have the same maximum over \( Hull(\Theta_+^2) \) as over \( \Theta_+^2 \). \( \square \)

Remark. Combining Theorems 7.9 and 7.11, we observe that the minimax linear risk of \( \Theta \) is still determined by the hardest rectangular subproblem, but now of the enlarged set \( QHull(\Theta) \). Of course, \( QHull(\Theta) \) may be much larger that \( \Theta \), and so (in contrast to Corollary 7.10) it could certainly happen now that \( R_L(\Theta) \gg R_N(\Theta) \); we will see examples in the later discussion of \( \ell_p \) balls and Besov spaces.

Section on truncation estimators and \( n \)-widths?

7.4. Appendix: Further details.

Direct proof of Theorem 7.9. (from Donoho et al. (1990))

By scale equivariance, we may set \( \epsilon = 1 \). The proof is essentially summarized in identity (7.40) below, whose terms we now define and justify. The minimax linear risk for rectangle \( \Theta(\tau) \), given by (7.29), is in fact a function of \( s = (\tau_i^2) \in \Theta_+^2 \) which we denote by

\[
J(s) := \sum_i s_i/(1 + s_i).
\]

Now \( \ell_2 \)-compactness of \( \Theta \) implies \( \ell_1 \)-compactness of \( \Theta_+^2 \), and so the easily verified \( \ell_1 \)-continuity of \( J \) on \( \Theta_+^2 \) implies the existence of a maximum \( \bar{s} = (\bar{s}_i^2) \in \Theta_+^2 \). For \( \eta \in [0, 1] \) and \( s \in \Theta_+^2 \), define \( s_{\eta i} = (1 - \eta)\bar{s}_i + \eta s_i \); we have \( s_\eta \in \Theta_+^2 \) since the latter is convex by our main hypothesis. Thus
$J(s_\eta) \leq J(\bar{s})$ and a formal calculation, justified by the dominated convergence theorem, gives

\begin{equation}
(d/d\eta) J(s_\eta)\big|_{\eta=0} = \sum_i \frac{s_i - \bar{s}_i}{(1 + \bar{s}_i)^2} \leq 0.
\end{equation}

We denote the linear estimate corresponding to $\bar{s}$ by $\hat{\theta}_c$, where $\hat{\theta}_c = \bar{s}_i/(1 + \bar{s}_i)$. Since $\hat{\theta}_c$ is linear minimax over $\Theta(\bar{s})$, which is the hardest hyperrectangle, it satisfies

\begin{equation}
r(\hat{\theta}_c, \bar{s}) = \sup_{\tau \in \Theta} R_L(\Theta(\tau)) \leq R_L(\Theta).
\end{equation}

To finish the proof, we show that $\hat{\theta}_c$ is actually minimax over all of $\Theta$, which in view of (7.38), would follow if we show that for all $\theta \in \Theta$,

\begin{equation}
r(\hat{\theta}_c, \theta) - r(\hat{\theta}_c, \bar{s}) \leq 0.
\end{equation}

Thus, for any $\theta \in \Theta$, let $s = (s_i(\theta)) = (\theta_i^2)$: using (7.24), the mean squared error of $\hat{\theta}_c$ depends on $\theta$ only through $s$. Hence

\begin{equation*}
r(\hat{\theta}_c, \theta) = \sum \hat{\theta}_c^2 + \frac{s_i}{(1 + \bar{s}_i)^2}.
\end{equation*}

Since the variance term is constant over $\Theta$, and we may calculate the risk difference in (7.39) simply by subtracting biases. This difference is exactly the sum in (7.37), so the proof is completed by concluding that for all $\theta \in \Theta$,

\begin{equation}
r(\hat{\theta}_c, \theta) - r(\hat{\theta}_c, \bar{s}) = (d/d\eta) J(s_\eta)\big|_{\eta=0} \leq 0.
\end{equation}

**Exercises**

1. (Two point priors.) Suppose that $y \sim N(\theta, 1)$, and consider the symmetric two point prior $\pi_\tau^{[2]} = (1/2)(\delta_\tau + \delta_{-\tau})$. Show that for squared error loss,

\begin{align*}
\pi(\{\tau\}\mid y) &= e^{\tau y} / (e^{\tau y} + e^{-\tau y}), \\
\hat{\theta}_c(y) &= E(\theta\mid y) = \tau \tanh \tau y, \\
E[(\theta - \hat{\theta}_c)^2] &= \tau^2 / \cosh^2 \tau y, \\
B(\pi_\tau) &= \tau^2 e^{-\tau^2/2} \int \frac{\phi(y)dy}{\cosh \tau y}.
\end{align*}

2. (Bounded normal mean theory for $L_1$ loss.) Redo the previous question for $L(\theta, a) = |\theta - a|$. In particular, show that

\begin{align*}
\hat{\theta}_c(y) &= \tau \text{ sgn } y, \quad \text{and} \quad B(\pi_\tau) = 2\tau \Phi(\tau),
\end{align*}

where, as usual $\Phi(\tau) = \int_{-\infty}^{\infty} \phi(s)ds$. In addition, show that

\begin{equation*}
\mu^* = \sup_{\tau, \epsilon} \frac{\rho_L(\tau, \epsilon)}{\rho_N(\tau, \epsilon)} \leq \frac{1}{B(\pi_1)} = 1/32 < \infty.
\end{equation*}

Hint: show that $\rho_L(\tau, 1) \leq \rho_P(\tau, 1) = \min(\tau, \sqrt{2/\pi})$.
3. (Continued.) For $L_1$ loss, show that (a) $\rho_N(\tau, \epsilon) = \epsilon \rho_N(\tau/\epsilon, 1)$ is increasing in $\tau$, and (b) $\lim_{\eta \to \infty} \rho_N(\tau, \epsilon) = \epsilon \gamma_0$, where $\gamma_0 = E_0|z| = \sqrt{2/\pi}$.

[Hint: for (b) consider the uniform prior on $[-\tau, \tau]$]

4. (Translation invariance implies diagonal Fourier optimality.) Signals and images often are translation invariant. To make a simplified one-dimensional model, suppose that we observe, in the “time domain”, $x_k = \gamma_k + \sigma \eta_k$ for $k = 1, \ldots, n$. To avoid boundary effects, assume that $x, \gamma$ and $\eta$ are extended to periodic functions of $k \in \mathbb{Z}$, that is $x(k + n) = x(k)$, and so on. Define the shift of $\gamma$ by $(S\gamma)_k = \gamma_{k+1}$. The set $\Gamma$ is called shift-invariant if $\gamma \in \Gamma$ implies $S\gamma \in \Gamma$. Clearly, then, $S^l \gamma \in \Gamma$ for all $l \in \mathbb{Z}$.

(a) Show that $\Gamma = \{ \gamma : \sum_{k=1}^{n} |\gamma_k - \gamma_{k-1}| < C \}$ is an example of a shift-invariant set. Such sets are said to have bounded total variation.

Now rewrite the model in the discrete Fourier domain. Let $\epsilon = e^{2\pi i n}$ and note that the discrete Fourier transform $y = Fx$ can be written

$$y_k = \sum_{l=0}^{n-1} \epsilon^{kl} x_l, \quad k = 0, \ldots, n - 1.$$ Similarly, let $\theta = F\gamma, z = F\eta$ and $\Theta = FT$.

(b) Show that shift-invariance of $\Gamma$ means that $\theta = (\theta_k) \in \Theta$ implies $M^l \theta = (\epsilon^{kl} \theta_k) \in \Theta$ for $l \in \mathbb{Z}$. In particular, we have $FS = M^{-1}F$.

(c) Let $V(\tau) = \{ M^l \tau, l \in \mathbb{Z} \}$ denote the orbit of $\tau$ under the action of $M$. By using a random shift (i.e. $l$ chosen at random from $\{0, \ldots, n - 1\}$), modify the random signs method to show that

$$\sup_{\theta \in V(\tau)} r(\hat{\theta}_{C_0}, \theta) \leq \sup_{\theta \in V(\tau)} r(\hat{\theta}_{C_0}, \theta).$$

Thus, on a translation invariant set $\Gamma$, an estimator that is minimax among affine estimators must have diagonal linear form when expressed in the discrete Fourier basis.
Part 3

LINEAR ESTIMATION
CHAPTER 8

Pinsker’s Theorem and Asymptotics

The theorem of Pinsker (1980) is notable for several reasons. First, it gives an exact evaluation of the linear minimax risk in the Gaussian sequence model for quadratic loss over general ellipsoids in $\ell_2$. Secondly, it shows that in the low noise limit $\epsilon \to 0$, the non-linear minimax risk is actually equivalent to the linear minimax risk: in other words, there exist linear rules that are asymptotically efficient. Pinsker’s argument proceeds by showing that there are Gaussian priors that are nearly least favorable, and that these priors are almost entirely concentrated on the ellipsoid $\Theta$. The Bayes rules for these Gaussian priors are linear, and are essentially the linear minimax rules, which leads to the asymptotic efficiency. Pinsker’s result applies to ellipsoids generally, and thus to all levels of Hilbert-Sobolev smoothness, and so might be considered as a crowning result for linear estimation.

Pinsker’s paper inspired a considerable literature (references later), for now we mention only two recent works which contain, among other developments, shorter proofs of the original result: Belitser & Levit (1995) and Tsybakov (1997).

8.1. Exact evaluation of linear minimax risk.

Suppose that $\Theta$ is an ellipsoid in $\ell_2(N)$:

$$\Theta = \Theta(C) = \{\theta : \sum a_i^2\theta_i^2 \leq C^2\}. \quad (8.1)$$

A pleasant surprise is that there is an explicit solution for the minimax linear estimator over such ellipsoids.

**8.1. Proposition.** Suppose that $y = \theta + \epsilon z$. Assume that $\Theta$ is an ellipsoid $\Theta(C)$ and that $a_i > 0$ and $a_i \to \infty$. Then the minimax linear risk

$$R_L(\Theta, \epsilon) = \epsilon^2 \sum_i (1 - a_i/\mu)_+, \quad (8.2)$$

where $\mu = \mu(\epsilon, C)$ is determined by

$$\epsilon^2 \sum a_i(\mu - a_i)_+ = C^2. \quad (8.3)$$

The linear minimax estimator

$$\hat{\theta}_i^L(y) = c_i y_i = (1 - a_i/\mu)_+ y_i \quad (8.4)$$
is Bayes for a Gaussian prior $\pi_{\epsilon,C}$ having independent components given by
\begin{equation}
\theta_i \sim N(0, \tau_i^2), \quad \tau_i^2 = \epsilon^2/(a_i - 1).
\end{equation}

Proof. The set $\Theta$ is solid, orthosymmetric and quadratically convex. Since $a_i \not\to \infty$ it is also compact. Thus the minimax linear risk is determined by the hardest rectangular subproblem (Theorem 7.9):
\begin{equation}
R_L(\Theta, \epsilon) = \sup_{\tau \in \Theta} R_L(\Theta(\tau), \epsilon) = \sup \left\{ \sum_i \epsilon^2 \tau_i^2/(\epsilon^2 + \tau_i^2) : \sum_i a_i^2 \tau_i^2 \leq C^2 \right\}.
\end{equation}
This maximum may be evaluated by forming the Lagrangian
\[ \mathcal{L} = \sum_i [\epsilon^2 - \epsilon^2/(\tau_i^2 + \epsilon^2)] - \mu \sum_i a_i^2 \tau_i^2.\]
Simple calculus shows that the maximum is attained at $\tau_i^2$ given by (8.5). The positive part constraint arises because $\tau_i^2$ cannot be negative. The Lagrange multiplier parameter $\mu$ is uniquely determined by the equation $\sum a_i^2 \tau_i^2 = C^2$, which on substitution for $\tau_i^2$ yields (8.3). This equation has a unique solution since the left side is a continuous, strictly increasing function of $\mu$. The corresponding maximum is then (8.2) and the linear minimax estimator is given by (8.4). $\square$

[As noted in the proof of Theorem 7.9, identity (8.6) is itself a minimax theorem, indeed Pinsker gave a direct proof.]

An important point is that this minimax estimator downweights higher coordinates, and since $a_i \not\to \infty$, is non-zero in only a finite number of coordinates.

It should also be emphasized that the least favorable sequence $(\tau_i)$ and hence the minimax linear estimator depend on $(\epsilon, a, C)$ - in particular, the difficult configurations change with changing noise-level.

Throughout the chapter, check consistency of $\lambda$ with $\mu^{-2}$ with $\lambda^{2m}$.

8.1.1. Example: Sobolev Ellipsoids. Return to the Hilbert-Sobolev parameter space in the trigonometric basis considered in Section 3.1: we have $a_{2k} = a_{2k-1} = (2k)^{\alpha}$ for $\alpha > 0$. For a fixed $\mu$, the sum in (8.3) extends up to $k_\mu = [\mu^{1/\alpha}]$ and so, defining $S_\mu(k) = \sum_{i=1}^{k_\mu} a_i^2$, the equation becomes
\[ \mu S_1(k_\mu) - S_2(k_\mu) = C^2/\epsilon^2.\]
Using the integral approximation $S_\mu(k) \sim k^{\mu\alpha+1}/(\mu\alpha + 1)$ and so, solving for $\mu$ yields
\begin{equation}
\mu_\epsilon \sim c_\alpha (C/\epsilon)^{2\alpha/(2\alpha+1)},
\end{equation}
where $c_\alpha = [\alpha^{-1}(\alpha + 1)(2\alpha + 1)]^{\alpha/(2\alpha+1)}$. Substitution into (8.2) followed by similar integral approximations then shows that, as $\epsilon \to 0$,
\begin{equation}
R_L(\Theta, \epsilon) \sim P_C C^{2(1-C)} \epsilon^{2\alpha},
\end{equation}
where \( r = \frac{2\alpha}{(2\alpha + 1)} \) and the Pinsker constant

\[ P_r = r'(1 - r)^{-1}(2 - r)^{-r}. \]

**Remarks.** 1. The rate of convergence \( \epsilon^{2r} \) depends on the assumed smoothness \( \alpha \): the greater the smoothness, the closer is the rate to the parametric rate \( \epsilon^{2} \).

2. The dependence on the scale \( C \) of the ellipsoid is also explicit: in fact, it might be written \( C^{2}(\epsilon^{2}/C^{2})^{r} \) to emphasise that the convergence rate \( r \) really applies to the (inverse) signal-to-noise ratio \( \epsilon^{2}/C^{2} \).

3. The shrinkage weights \( w_{k} = (1 - k^{\alpha}/\mu)_{+} \) assign weight close to 1 for low frequencies, and cut off at \( k = \mu^{1/\alpha} \propto (C^{2}/\epsilon^{2})^{1/(2\alpha + 1)} \). Thus, the number of frequencies retained is an algebraic power of \( C/\epsilon \), decreasing as the smoothness \( \alpha \) increases.

**8.1.2. Example: Ellipsoids of analytic functions.** Again consider the trigonometric basis for periodic functions on \([0, 1]\), but now with \( a_{2k} = a_{2k-1} = e^{ik} \), so that \( \Theta(\alpha, C) = \{ \theta : \sum e^{2\alpha k} (\theta_{2k-1}^{2} + \theta_{2k}^{2}) \leq C^{2} \} \). Since the semi-axes decay exponentially with frequency, these ellipsoids contain only infinitely differentiable functions, which are thus much smoother than typical members of the Sobolev classes.

To interpret the exponential decay conditions, it may help to think of the periodic function \( f(t) \) as a Fourier series in complex exponentials \( f(t) = \sum_{-\infty}^{\infty} \theta_{2} e^{2\pi i t} \), where \( \Theta = \sum_{2} \zeta_{2} e^{2\pi i t} \) of the complex variable \( z = re^{2\pi i t} \). Consider then the domain in which the function \( g(z) = \sum_{-\infty}^{\infty} \zeta_{2} e^{2\pi i t} \) remains analytic. On the unit circle \( |z| = 1 \), \( g \) reduces to our periodic function \( f \). Now if \( |\zeta| = O(e^{-\alpha |d|}) \), then \( g \) is analytic in the annulus \( A_{\alpha} = \{ z : e^{-\alpha} < |z| < e^{\alpha} \} \) while a near converse also holds: if \( g \) is analytic in a domain containing \( \overline{A}_{\alpha} \), then \( |\zeta| = O(e^{-\alpha |d|}) \). Thus, the larger the value of \( \alpha \), the greater the domain of analyticity.

We turn to interpretation of the linear minimax solution of Proposition 8.1. For given \( \mu \), the sum in (8.3) cuts off after \( k(\mu) = [\alpha^{-1} \log \mu] \), and so its evaluation involves geometric sums like \( \sum_{k} e^{\alpha j \ell} = e^{\alpha j} e^{\ell \log k} \) for \( \ell = 1 \) and 2 which are, in strong contrast with the Sobolev case, dominated by a single leading term. After solving (8.3) for \( \mu = \mu_{c} \), we find from (8.2)

\[ R_{L}(\Theta, \epsilon) = \epsilon^{2} k_{c} \sum_{1}^{k_{c}} \left[ 1 - \gamma k_{c} e^{-\alpha(k_{c}, k)} \right], \]

where \( k_{c} = k(\mu_{c}) = [\alpha^{-1} \log C/\epsilon] \) and \( \gamma k_{c} \) are unimportant constants confined to the range \([e^{-\alpha}, 1]\). Thus it is apparent that the number of frequencies \( k_{c} \) retained is logarithmic in signal to noise (as opposed to algebraic, in the Sobolev case), and the smoothing weights \( w_{k} = 1 - \gamma k_{c} e^{-\alpha(k_{c}, k)} \) are very close to 1 except for sharp decline near \( k_{c} \). In particular, the minimax linear risk

\[ R_{L}(\Theta, \epsilon) \sim \epsilon^{2} k_{c} \sim (1/2\alpha) \log \epsilon^{-2} \cdot \epsilon^{2} \]
is only logarithmically worse than the parametric rate $\tilde{c}^2$, and the dependence on $\Theta(a, C)$ comes only through the analyticity range $\alpha$ and not via the scale factor $C$.

8.1.3. The minimax estimator compared with smoothing splines.

We return to the Sobolev ellipsoid setting to suggest that information derived from study of the minimax linear estimate and its asymptotic behaviour is quite relevant to the smoothing spline estimates routinely computed in applications by statistical software packages. The following discussion is inspired by Carter et al. (1992).

We have seen in (4.13) of Chapter 4 that the Lagrange multiplier form of smoothing spline problem in the sequence model is

$$\hat{\theta}_{SS} = \arg\min_{\theta} \sum_{i} (y_i - \theta_i)^2 + \lambda \sum_i a_i^2 \theta_i^2,$$

with solution

$$\hat{\theta}_{SS}^* = (1 + \lambda a_k^2)^{-1} y_k.$$

This should be compared with the linear minimax solution

$$\hat{\theta}_{\lambda,k} = (1 - \mu a_k) y_k.$$

If we make the identification $\lambda \leftrightarrow \mu^{-2}$, then the inequality $(1 + x^2)^{-1} \geq (1-x)_+$ valid for positive $x$, shows that the spline estimate shrinks somewhat less in each frequency than the minimax rule.

Pursuing this comparison, we might contrast the worst case mean squared error of the Pinsker and smoothing spline estimates over Sobolev ellipsoids of smooth functions:

$$\overline{r}(\hat{\theta}_{\lambda}; \epsilon) = \sup_{\theta \in \Theta_0^\perp(C)} r(\hat{\theta}_{\lambda}, \theta; \epsilon).$$

To do so, we need to specify the value of the regularization parameter $\lambda$ to be used in each case – a reasonable choice is the optimum, or minimax value:

$$\lambda_* = \arg\min_{\lambda} \overline{r}(\hat{\theta}_{\lambda}; \epsilon).$$

This is exactly the calculation done in Chapter 4 at (4.24) and (4.25) for the spline and minimax families respectively. [Of course, the result for the minimax family must agree with (8.8)]! In both cases, the solutions took the form

$$\lambda_* \sim (c_1 \tilde{c}^2/C^2)^r, \quad \overline{r}(\lambda_*; \epsilon) \sim c_2 e^{H(r)} C^{2(1-r)} 2^r,$$

with $r = 2\alpha/(2\alpha + 1)$, and

$$c_1^{SS} = 2v_{2\alpha}/\alpha, \quad c_2^{SS} = \tilde{v}_{\alpha}/4^{1-r}, \quad v_{\alpha} = (1 - 1/2\alpha)/\text{sinc} (\pi/2\alpha),$$

$$c_1^M = \frac{1}{2} \sigma_{2\alpha}/\alpha, \quad c_2^M = \sigma_{\alpha}, \quad \sigma_{\alpha} = 2\alpha^2/(\alpha + 1)(2\alpha + 1).$$

Thus the methods have the same dependence on noise level $\epsilon$ and scale $C$, with differences appearing only in the coefficients. We may therefore summarize the comparison through the ratio of max MSE's. Remarkably,
the low noise smoothing spline maximal MSE turns out to be only negligibly larger than the minimax linear risk of the Pinsker estimate: for $\Theta = \Theta_2(C)$, as $\epsilon \to 0$,

$$
R_{SS}(\Theta, \epsilon) \sim \left( \frac{v_\alpha}{\sigma_\alpha} \right)^{r \left( \frac{1}{2} \right)} 1^{1-r} = \begin{cases} 
1.083 & \alpha = 2 \\
1.055 & \alpha = 4 \\
\to 1 & \alpha \to \infty
\end{cases}. 
$$

Similarly, we may compare the asymptotic choices of the smoothing parameter:

$$
\frac{\lambda_{SS}}{\lambda_M} \sim \left( \frac{4v_\alpha}{\hat{\sigma}_\alpha} \right)^{r} = \begin{cases} 
4.331 & \alpha = 2 \\
4.219 & \alpha = 4 \\
\to 1 & \alpha \to \infty
\end{cases},
$$

and so $\lambda_{SS}$ is approximately four times $\lambda_M$ and this counteracts the lesser shrinkage of smoothing splines noted earlier.

Furthermore, in the discrete smoothing spline setting of Section 4.1, Carter et al. (1992) present small sample examples in which the efficiency loss of the smoothing spline is even smaller than these asymptotic values. In summary, from the maximum MSE point of view, the minimax linear estimator is not so different from the Reinsch smoothing spline that is routinely computed in statistical software packages.

**8.2. Pinsker’s Theorem**

8.2. **Theorem (Pinsker).** If $\Theta$ is the ellipsoid (8.1) and $a_i \to \infty$ as $i \to \infty$, then

$$
R_N(\Theta, \epsilon) \sim R_L(\Theta, \epsilon) \quad \text{as } \epsilon \to 0,
$$

where $R_L(\Theta, \epsilon)$ is given by (8.2) and $\hat{\theta}_L$ given by (8.4) is asymptotically minimax among all estimators.

**Remarks.** 1. The condition that $a_i \to \infty$ as $i \to \infty$ is equivalent to compactness of $\Theta$ in $\ell_2$. In Section 8.3, it is shown that if $\Theta$ is not compact, then $R_N(\Theta, \epsilon)$ does not even approach 0 as $\epsilon \to 0$.

2. Pinsker’s proof is actually for a considerably more general situation, in which the variance $\sigma_i^2$ of each co-ordinate may be different. This extension is relevant for a number of models with different loss functions, indirectly observed data, or non-white noise. (Add discussion later).

**8.2.1. General comments on the proof.** Formula (8.5) gives an expression for the least favorable vector $\tau_i$, which depends on $\epsilon$ as well as the ellipsoid parameters $(a, C)$. There are essentially three ways in which asymptotic equivalence of linear and non-linear estimates can occur:

(i) $\tau_i/\epsilon \to \infty$. As shown at (7.19), in this setting $\rho_L(\tau_i, \epsilon) \sim \rho_N(\tau_i, \epsilon)$, and one can choose a prior $\nu_{i, \epsilon}$ for $\theta_i$ that is supported on $[-\tau_i, \tau_i]$. 

95
(ii) $\tau_i/\epsilon \to 0$. Again from (7.19), $\rho_L(\tau_i, \epsilon) \sim \rho_N(\tau_i, \epsilon)$, with the zero estimator being asymptotically minimax, and one can simply choose two point priors $\nu_{i,\epsilon} = \frac{1}{2}(\delta_{-\tau_i} + \delta_{\tau_i})$.

(iii) $\zeta^{-1} < \tau_i/\epsilon < \zeta$. In this case one assigns each $\theta_i$ an independent $N(0, \tau_i^{-2})$ distributions, obtaining a product prior $\pi_{\epsilon, C}$ (recall the dependence of $\tau_i$ on $\epsilon, C$). Of course, this prior is not fully supported on the ellipsoid $\Theta(a, C)$. However, if many coordinates are combined, and the variances in these coordinates are not too dissimilar, as is guaranteed by the bounds in $\zeta$, then one can hope that by choosing $\gamma_\epsilon < 1$ appropriately, $\pi_{\epsilon, C}$ will asymptotically concentrate on $\Theta(a, C)$.

Pinsker’s proof handles the three modes simultaneously. For simplicity, we confine attention to two special cases.

(i) **Analytic classes.** For the ellipsoids considered in Section 8.1.2, the large signal-to-noise case (i) suffices, due to the geometric growth of the weights $a_i$. We take this up in Section ???.

(ii) **Equidistributed classes.** Assume that the weights satisfy

\begin{align*}
\text{(8.11a)} & \quad a_k \text{ is non-decreasing and } a_{k+1} = O(a_k), \quad \text{and} \\
\text{(8.11b)} & \quad \text{for some } \gamma > 1, \quad a_k^\gamma = O(\sum_1^k a_i).
\end{align*}

These conditions allow most cases of polynomial and logarithmic growth. For example, if

\[ c_1 k^{\alpha} \leq a_k \leq c_2 k^{\alpha}, \]

then (8.11b) holds with $\gamma = 1 + 1/\alpha$. These bounds immediately cover the Sobolev ellipsoids in the Fourier basis considered in the last section. They also include the ellipsoid weights occurring in the wavelet versions of Hilbert-Sobolev norms:

\[ a_{2^j+k} = 2^{j\alpha}, \quad k = 0, 1, \ldots, 2^j - 1. \]

[Exercise: give an example of a monotone sequence $a_k \leq k^{\alpha}$ for which (8.11b) fails.] The restrictions exclude exponential growth, in which sums are dominated by their leading term. Indeed, since the $a_k$ are increasing, (8.11b) implies $a_k^\gamma \leq c k a_k$, so that $a_k = O(k^{1/(\gamma-1)})$.

**8.2.2. Gaussian priors are least favorable.** To evaluate $R(C, \epsilon) = R_N(\Theta(C), \epsilon)$ asymptotically, we apply the Bayes minimax method of Section 6.4 to the set of priors

\[ \mathcal{M}(C) = \{ \pi(d\theta) : \sum a_i^2 E_\pi \theta_i^2 \leq C^2 \}. \]

Write $B(C, \epsilon)$ for the corresponding Bayes minimax value $B(\mathcal{M}(C), \epsilon)$. Thus, the condition that $\theta \in \Theta(C)$ is replaced with a weaker requirement that the ellipsoid constraint merely hold on average for $\pi$. Clearly $\mathcal{M}(C)$ is
convex and contains point masses $\delta_\theta$ for $\theta \in \Theta(C)$. Hence, by the general considerations of (6.16) and (6.15),

$$R(C, \epsilon) \leq B(C, \epsilon) = \sup_{\pi \in \mathcal{M}(C)} B(\pi).$$

We have observed that the linear minimax rule is Bayes for the Gaussian prior $\pi_{e,C}$ given by (8.5). This prior is actually least favorable for $\mathcal{M}(C)$.

8.3. Lemma. $B(C, \epsilon) = B(\pi_{e,C}) = R_L(\Theta(C), \epsilon)$.

Proof. Given any prior $\pi \in \mathcal{M}$, a more difficult prior for estimation of $\theta$ can be built as the product measure $\tilde{\pi} = \prod_i \pi_i$ of the univariate marginals of $\pi$. Membership in $\mathcal{M}$ depends only on the univariate marginal distributions of $\pi$, so $\tilde{\pi}$ still belongs to $\mathcal{M}$. Lemma 6.2 showed that $B(\pi) \leq B(\tilde{\pi}) = \sum B(\pi_i)$. Hence, the Bayes-minimax risk may be found by coordinatewise maximization:

$$B(\mathcal{M}) = \sup_{\tilde{\pi}} B(\tilde{\pi}) = \sup\{\sum B(\pi_i) : \sum a_i^2 E_{\pi_i} \theta_i^2 \leq C^2\}.$$

Corollary 6.5 maximises univariate Bayes risks under variance constraints:

$$\sup\{B(\pi_i) : E_{\pi_i} \theta_i^2 \leq \tau_i^2\} = \frac{c^2 \tau_i^2}{c^2 + \tau_i^2}. $$

with the least favorable prior $\pi_i$ being Gaussian $N(0, \tau_i^2)$. Thus for squared error loss and ellipsoids, the Bayes minimax problem (8.13) reduces to exactly the linear minimax risk optimization problem (8.6)! Hence, the least favorable prior distribution $\pi_{e,C}$ has independent $N(0, \tau_i^2)$ components with $\tau_i$ given by (8.5).

Remark. Our approach uses the Bayes-minimax inequality and then observes that the Bayes minimax risk equals the linear minimax risk. Of course we could here have used the trivial inequality $R_N(\Theta) \leq R_L(\Theta)$ and then identified the linear minimax risk as arising from an independent components Gaussian prior. However for the non-quadratically convex $\Theta$ to be considered in later chapters, the inequality (8.12) is much tighter than the trivial bound - the asymptotic minimaxity of linear rules being specific to the setting of ellipsoids.

8.2.3. Proof of Pinsker’s Theorem assuming equidistribution. We use the Bayes minimax strategy to show that $B(C, \epsilon) \sim R(C, \epsilon)$ as $\epsilon \to 0$. We need to construct a nearly least favorable prior sequence $\pi_\epsilon$ that concentrates on $\Theta = \Theta(C)$ as $\epsilon \to 0$. We have just seen that the Gaussian prior $\pi_{e,C}$ is least favorable for $\mathcal{M}(C)$. However, it is clearly not supported on $\Theta(C)$. Instead, we exploit the scale family structure: we choose an appropriate sequence $\gamma_\epsilon$ increasing slowly to 1, and pick for $\pi_\epsilon$ the least favorable prior for $\mathcal{M}(\gamma_\epsilon C)$, namely $\pi_{e,\gamma_\epsilon C}$. Lemma 6.8, combined with Exercise 6.2(b) reduce the task to showing that $\pi_\epsilon$ concentrates on $\Theta = \Theta(C)$:

$$P_\epsilon = \pi_\epsilon(\Theta^\epsilon) \to 0,$$

97
and the more technical complement

\[(8.16) \quad \Delta_\epsilon = E_{\pi_\epsilon} \{ 1 + \| \theta \|^2, \theta \} = o(B(C, \epsilon)).\]

First, assumptions (8.11) yield a simple bound on \( \mu_\epsilon \). Indeed, choose the largest integer \( k(\epsilon) \) such that \( 2a_{k(\epsilon)} \leq \mu_\epsilon \). Then (8.11a) provides a constant \( c \) such that

\[(8.17) \quad 2a_{k(\epsilon)} \leq \mu_\epsilon \leq 2a_{k(\epsilon)+1} \leq 2ca_{k(\epsilon)}.
\]

Here, and below, \( c \) denotes a generic constant, not necessarily the same at each appearance. Using now (8.3) and (8.11b),

\[ C^2 / \epsilon^2 \geq \sum a_i (2a_{k(\epsilon)} - a_i) \geq a_{k(\epsilon)} \sum a_i \geq c^2 a_{k(\epsilon)}^{1+\gamma} \]

Hence \( C^2 \geq c^2 \mu_\epsilon^{1+\gamma} = c \epsilon^{1-\gamma}(\epsilon \mu_\epsilon)^{1+\gamma} \), so that for all small \( \epsilon \),

\[(8.18) \quad \epsilon \mu_\epsilon \leq \epsilon^\beta,
\]

with \( \beta = (\gamma - 1)/(\gamma + 1) > 0 \).

With this bound in hand, we turn to showing that \( \pi_\epsilon \) concentrates on \( \Theta \). Since \( E \sum a_i \theta_i^2 = \gamma_\epsilon^2 C^2 \) under \( \pi_\epsilon \), introduce \( \delta_\epsilon = 1 - \gamma_\epsilon^2 \) and write

\[ P_\epsilon = \pi_\epsilon(\Theta^\epsilon) = P\{ \sum a_i^2 (\theta_i^2 - E\theta_i^2) > \delta_\epsilon C^2 \}.
\]

Now use an exponential inequality for mixtures of chi-squared variables, (E.2) of Appendix E, which states that if \( z_i \) are i.i.d. \( N(0, 1) \), and \( t \leq \| a \|_1 \), then

\[ P\{ \sum \alpha_i (z_i^2 - 1) > t \} \leq \exp \{-t^2/(8\| a \|_1 \| a \|_\infty) \}.
\]

This is applied with \( z_i = \theta_i/t_i \), \( a_i = a_i \gamma_\epsilon^2 t_i^2 \) and \( t = \delta_\epsilon C^2 \). Note that \( \| a \|_1 = \sum a_i \gamma_\epsilon^2 t_i^2 = \gamma_\epsilon^2 C^2 \) so that \( t \leq \| a \|_1 \). Using (8.5),

\[ \| a \|_\infty = \max_i c^2 a_i (\mu_\epsilon - a_i) \leq c^2 \mu_\epsilon^2 /4.
\]

Substitution into the exponential bound yields

\[ P_\epsilon \leq \exp \{-\frac{1}{2}(C/\gamma_\epsilon)^2(\delta_\epsilon/\epsilon \mu_\epsilon)^2 \} \leq \exp \{-\alpha^{-\beta} \} \to 0,
\]

if we choose \( \delta_\epsilon = \sqrt{\mu_\epsilon} \) and appeal to (8.18).

We complete the proof by verifying (8.16). First, use Cauchy-Schwartz termwise

\[ E_{\pi_\epsilon}(\| \theta \|^2, \theta) \leq P_\epsilon^{1/2} \sum (E\theta_i^4)^{1/2}.
\]

Now \( E\theta_i^4 = 3\tau_i^4 \) and \( a_i \) is monotone, so the sum is bounded by \( \sqrt{3} a_1^{-2} \sum a_i^2 \tau_i^2 \leq c \). Hence \( \Delta_\epsilon = o(P_\epsilon^{1/2}) \). A simple lower bound on \( B(\pi_\epsilon) \) comes from (8.2):

\[ B(C, \epsilon) \geq c^2 (1 - a_1 / \mu_\epsilon) \geq c^2 /2,
\]

since \( \mu_\epsilon \not\to \infty \) as \( \epsilon \to 0 \). Combining the last three displays yields (8.16).
8.2.4. Pinski’s theorem for the analytic function ellipsoids.

The idea is to use the simpler version of the Bayes minimax strategy (cf. 3(a) of Section 6.4), in which the least favorable priors are already supported on $\Theta$.

To be more specific, define $\nu_i, e$ by giving $\theta_i / \tau_i$ the distribution $F$ with density $\cos^2(\pi s/2) ds$ on $[-1, 1]$. Then setting $\beta = 1/I(F)$, and using (7.13),

$$\frac{B(\nu_i, e)}{\rho_L(\tau_i, e)} \geq \frac{\tau_i^2 / \epsilon^2 + 1}{\tau_i^2 / \epsilon^2 + \beta} \to 1$$

(8.19)

In the analytic function setting of Section 8.1.2, define $\nu_e = \prod_{i=1}^{k(\lambda_i)} \nu_i, e$, and note that $\nu_e$ is supported in $\Theta$. It is then fairly direct to verify, using (8.19), that

$$B(\nu_e) = \sum_{i=1}^{k(\lambda_i)} B(\nu_i, e) \geq \sum_{i=1}^{k(\lambda_i)} \frac{\tau_i^2 + \epsilon^2}{\tau_i^2 + \beta \epsilon^2} \rho_L(\tau_i, e) \sim B(M, \epsilon).$$

Indeed, define a set $N_e = \{ i : \tau_i^2 / \epsilon^2 \geq \beta_i \}$, where $\beta_i = \sqrt{\log \epsilon^{-2}}$, say, and set $\gamma_e = \frac{\beta_i + 1}{\beta_i + \beta} \to 1$. Then, since $|N_e| \leq O(\log \epsilon^{-1})$ in the analytic ellipsoid setting,

$$B(\nu_e) \geq \gamma_e [B(M, \epsilon) - \sum_{N_e \rho_L(\tau_i, e)] \sim B(M, \epsilon).$$
8.3. Interlude: Compactness and Consistency

[This section is somewhat of an interlude, included in part for variety, because of the different methods used.]

We have seen from Pinsker's theorem that if an ellipsoid $\Theta(a)$ is compact, then $R_N(\Theta(a), \epsilon) \to 0$ as $\epsilon \to 0$. In fact, quite generally, compactness of $\Theta$ is both necessary and sufficient for the existence of a uniformly consistent estimator.

8.4. Theorem. In the Gaussian sequence model (3.1), assume that $\Theta$ is bounded in $\ell_2(I)$. Then as $\epsilon \to 0$, $R_N(\Theta, \epsilon) \to 0$ if and only if $\Theta$ is compact.

Of course, if $R_N(\Theta, \epsilon)$ does not converge to 0, then there exists $c > 0$ such that every estimator has maximum risk at least $c$ regardless of how small the noise level might be. This further explains why it is necessary to introduce constraints on the parameter space in order to obtain meaningful results in nonparametric theory.

In particular, there can be no uniformly consistent estimator on $\{\theta \in \ell_2(\mathbb{N}) : \|\theta\|_2 \leq 1\}$, or indeed on any open set in the norm topology.

Because there are no longer any geometric assumptions on $\Theta$, the tools used for the proof change: indeed methods from testing, classification and from information theory now appear. While our proof involves only consistency and so is not at all quantitative, it nevertheless gives a hint of the role that covering numbers and metric entropy play in a much more refined theory (Birgé 1983) that describes how the “massiveness” of $\Theta$ determines the possible rates of convergence of $R_N(\Theta)$.

8.3.1. A lower bound for misclassification error. Any method that chooses between a finite number $m$ of alternative distributions necessarily has an error probability bounded below in terms of $\log m$ and the mutual separation of those distributions.

In detail, let $\{\theta_1, \ldots, \theta_m\}$ be a finite set, and $P_{\theta_1}, \ldots, P_{\theta_m}$ a corresponding set of probability distributions on $(\mathcal{Y}, \mathcal{B}_Y)$. For convenience, assume that the $P_{\theta_i}$ are mutually absolutely continuous, and so have positive densities $p_i$ with respect to some dominating measure $\nu$. Then, the Kullback-Leibler divergence between two probability measures $P$ and $Q$ having densities $p, q$ relative to $\nu$ is

$$K(P, Q) = \int \log \frac{dP}{dQ} dP = \int \log \frac{p}{q} p d\nu.$$  

(8.20)

The following lower bound is a formulation by Birgé (1983, Lemma 2.7) of a lemma due to Ibragimov & Khas'minskii (1981, pages 325–5).

8.5. Lemma. With the above definitions, let $\hat{\theta} : \mathcal{Y} \to \{\theta_1, \ldots, \theta_m\}$ be an arbitrary estimator. Then

$$\text{ave}_{i} P_{\theta_i} \{\hat{\theta} \neq \theta_i\} \geq 1 - \frac{\text{ave}_{i} K(P_{\theta_i}, P_{\theta_\hat{\theta}}) + \log 2}{\log (m - 1)}.$$  

(8.21)
Remark. Both averages in inequality (8.21) can of course be replaced by maxima over $i$ and $(i, j)$ respectively.

Proof. We first recall Fano’s lemma from information theory (e.g. Cover & Thomas (1991, page 39)). Let $\theta$ be a random variable with distribution $P(\theta = \theta_i) = q_i$. The conditional entropy of $\theta$ given $Y$ is defined by

$$H(\theta|Y) = -E \sum_i P(\theta = \theta_i|Y) \log P(\theta = \theta_i|Y),$$

where the expectation is taken over the $Y$ marginal of the joint distribution of $(\theta, Y)$. Let $h(q) = -q \log q - (1 - q) \log(1 - q)$ be the binary entropy function. Write $p_c = P(\theta \neq \theta)$ for the overall error probability when using estimator $\hat{\theta}$. Fano’s lemma provides a lower bound for $p_c$:

$$h(p_c) + p_c \log (m - 1) \geq H(\theta|Y).$$

To apply this, we choose the uniform distribution for $\theta$: $q_i = 1/m$ for all $i$. Hence the marginal density of $Y$ is just $\frac{1}{m} \sum_k p_k$ and the posterior probabilities $P(\theta = \theta_i|Y) = p_i / \sum_j p_j$. Consequently,

$$H(\theta|Y) = -\frac{1}{m} \sum_i \frac{p_i}{\sum_j p_j} \log \frac{p_i}{\sum_j p_j} \sum_k p_k \, dv = \log m - \frac{1}{m} \sum_i \int p_i \log \frac{p_i}{\sum_j p_j} \, dv.$$

Now apply Jensen’s inequality: $\log(\frac{1}{m} \sum_j p_j) \geq \frac{1}{m} \sum_j \log p_j$. Combine this with Fano’s lemma and the bound $h(p_c) \leq \log 2$ to get

$$p_c \log (m - 1) \geq \log m - \frac{1}{m^2} \sum_i \sum_j \int p_i \log \frac{p_i}{p_j} \, dv - \log 2.$$

Divide through by $\log (m - 1)$ and insert definition (8.20) to yield the result.

\[\square\]

8.3.2. Necessity of compactness. For both parts of the proof, we use an equivalent formulation of compactness, valid in complete metric spaces, in terms of total boundedness: $\Theta$ is totally bounded if and only if for every $\delta$, there is a finite set $\{\theta_1, \ldots, \theta_m\}$ such that the open balls $B(\theta_i, \delta)$ of radius $\delta$ centered at $\theta_i$ cover $\Theta$: $\Theta \subset \bigcup_{i=1}^m B(\theta_i, \delta)$.

[Clarify definitions here.]

Since $\Theta$ is bounded, it has a finite diameter $\Delta = \sup\{\|\theta_1 - \theta_2\| : \theta_1, \theta_2 \in \Theta\}$.

If $R_N(\Theta, \epsilon) \to 0$, then given any positive $\eta$, there exists $\epsilon_\eta$ and an estimator $\hat{\theta}_\eta$ such that

$$E_{\theta, \epsilon} \|\hat{\theta}_\eta - \theta\|^2 \leq (\eta/2)^2 \quad \text{for all } \theta \in \Theta.$$

Let $\Theta_\eta$ be a finite and $2\eta$-discernible subset of $\Theta$: each distinct pair $\theta_i, \theta_j$ in $\Theta_\eta$ satisfies $\|\theta_i - \theta_j\| > 2\eta$. From $\hat{\theta}_\eta(y)$ we build an estimator $\hat{\theta}_\eta(y)$ confined
to \( \Theta_{\eta} \) by choosing a closest \( \theta_i \in \Theta_{\eta} \) to \( \hat{\theta}_n(y) \): of course, whenever \( \hat{\theta}_n \neq \theta_i \), it must follow that \( \| \hat{\theta}_n - \theta_i \| \geq \eta \). Consequently, from Markov’s inequality and (8.22), we have for all \( i \)

\[
P_{\theta_i} \{ \| \hat{\theta}_n - \theta_i \| \geq \eta \} \leq \eta^{-1} E \| \hat{\theta}_n - \theta_i \| \leq 1/2.
\]

On the other hand, the misclassification inequality (8.21) provides a lower bound to the error probability: for the noise level \( \epsilon \) Gaussian sequence model, one easily evaluates

\[
K(P_{\theta_i}, P_{\theta_j}) = \| \theta_i - \theta_j \|/2 \epsilon^2 \leq \Delta^2/2 \epsilon^2,
\]

where \( \Delta \) is the diameter of \( \Theta \), and so

\[
\max_i P_{\theta_i} \{ \| \hat{\theta}_n - \theta_i \| \geq \eta \} \geq 1 - \frac{\Delta^2/2 \epsilon^2 + \log 2}{\log(\| \Theta_{\eta} \| - 1)}.
\]

Combining this with (8.23) gives a uniform upper bound for the cardinality of \( \Theta_{\eta} \):

\[
\log(\| \Theta_{\eta} \| - 1) \leq \Delta^2 \epsilon^{-2} + 2 \log 2.
\]

We may therefore speak of a \( 2\eta \)-discernible subset \( \Theta_{\eta} \subset \Theta \) of maximal cardinality, and for such a set, it is easily checked that \( \hat{\Theta} \) is covered by closed balls of radius \( 4\eta \) centered at the points of \( \Theta_{\eta} \). Since \( \eta \) was arbitrary, this establishes that \( \hat{\Theta} \) is totally bounded, and so compact.

### 8.3.3. Sufficiency of Compactness

Given \( \delta > 0 \), we will construct an estimator \( \hat{\theta}_c \) such that \( E_{\theta} \| \hat{\theta}_c - \theta \|^2 \leq 20 \delta^2 \) on \( \Theta \) for all sufficiently small \( \epsilon \). Indeed, compactness of \( \Theta \) supplies a finite set \( \Theta_{\delta} = \{ \theta_1, \ldots, \theta_m \} \) such that \( \hat{\Theta} \subset \bigcup_{i=1}^m B(\theta_i, \delta) \), and then \( \hat{\theta}_c \) is the maximum likelihood estimate on the sieve \( \Theta_{\delta} \). Thus let

\[
L(\theta) = \epsilon^2 \log dP_{\theta,c}/dP_{\theta,c} = \langle y, \theta \rangle - \frac{1}{2} \| \theta \|^2,
\]

\[
\hat{\theta}_c = \arg \max_{\theta \in \Theta_{\delta}} L(\theta).
\]

Since \( \Theta \) has diameter \( \Delta \), we have for any \( \theta \in \Theta \) the simple MSE bound

\[
E_{\theta} \| \hat{\theta}_c - \theta \|^2 \leq (4\delta)^2 + \Delta^2 \sum_{i, \| \theta - \theta_i \| \geq \delta} P_{\theta} \{ \hat{\theta}_c = \theta_i \}.
\]

Let \( \theta \in \Theta \) be fixed, and choose a point in \( \Theta_{\delta} \), renumbered to \( \theta_1 \) if necessary, so that \( \theta \in B(\theta_1, \delta) \). To have \( \hat{\theta}_c = \theta_i \) certainly implies that \( L(\theta_i) \geq L(\theta_1) \), and from (8.24)

\[
L(\theta_i) - L(\theta_1) = \langle y - \frac{\theta_1 + \theta_i}{2}, \theta_i - \theta_1 \rangle.
\]

Substituting \( y = \theta + \epsilon z \), putting \( u = (\theta_i - \theta_1)/\| \theta_i - \theta_1 \| \), and defining the standard Gaussian variate \( Z = \langle z, u \rangle \), we find that \( L(\theta_i) \geq L(\theta_1) \) implies

\[
\epsilon Z \geq \langle \frac{\theta_1 + \theta_i}{2} - \theta, u \rangle \geq \frac{1}{2} \| \theta_i - \theta_1 \| - \delta \geq \delta.
\]

Thus \( P_{\theta} \{ \hat{\theta}_c = \theta_i \} \leq \tilde{\Phi}(\delta/\epsilon) \), and so from (8.25)

\[
E_{\theta} \| \hat{\theta}_c - \theta \|^2 \leq (4\delta)^2 + m \Delta^2 \tilde{\Phi}(\delta/\epsilon) \leq 20 \delta^2,
\]
whenever $\epsilon$ is sufficiently small.
CHAPTER 9

James-Stein Estimation and Adaptive Minimaxity over Ellipsoids

[INTRODUCTION]

9.1. Stein’s Unbiased Risk Formula

Stein (1981) gave a formula for an unbiased estimate of the mean squared error of a nearly arbitrary function of a multivariate Gaussian variate. Although the identity itself involves little more than integration by parts, it has proved powerful and influential. As some important applications of the identity involve functions that are only “almost” differentiable, we begin with some remarks on weak differentiability, referring to standard sources, such as Gilbarg & Trudinger (1983, Chapter 7), for omitted details.

A function \( g : \mathbb{R}^d \to \mathbb{R} \) is said to be weakly differentiable if there exist functions \( h_i : \mathbb{R}^d \to \mathbb{R}, i = 1, \ldots, d \), such that

\[
\int \psi h_i = - \int (D_i \psi) g \quad \text{for all } \psi \in C_0^\infty,
\]

where \( C_0^\infty \) denotes the class of \( C^\infty \) functions on \( \mathbb{R}^d \) of compact support. We write \( h_i = D_i g \).

To verify weak differentiability in particular cases, we note that it can be shown that \( g \) is weakly differentiable if and only if it is equivalent to a function \( \tilde{g} \) that is absolutely continuous on almost all line segments parallel to the co-ordinate axes and whose partial derivatives (which consequently exist almost everywhere) are locally integrable.

9.1. THEOREM. Suppose that \( g : \mathbb{R}^d \to \mathbb{R}^d \) is weakly differentiable.

(a) If \( X \sim N_d(\mu, I) \) and \( E_\mu |X_i g_i(X)| + E|D_i g_i(X)| < \infty \) for \( i = 1, \ldots, d \), then

\[
\begin{align}
(9.1) & \quad E_\mu (X - \mu) \cdot g(X) = E_\mu \nabla \cdot g(X) \\
(9.2) & \quad E_\mu \|X + g(X) - \mu\|^2 = E_\mu \{d + 2 \nabla \cdot g(X) + \|g(X)\|^2\}.
\end{align}
\]

(b) More generally, if \( X \sim N_d(\mu, V) \) and \( E[|X_i g_i(X)| + |D_i g_i(X)|] < \infty \) for all \( i, j \), then

\[
\begin{align}
(9.3) & \quad E_\mu \|X + g(X) - \mu\| = E_\mu \{tr V + 2tr[V Dg(X)] + \|g(X)\|^2\}.
\end{align}
\]

The expression \( U(x) = d + 2 \nabla \cdot g(x) + \|g(x)\|^2 \) in (9.2), and its generalization in (9.3) is called Stein’s unbiased risk estimate (SURE) corresponding
to the estimator \( \hat{\mu}(x) = x + g(x) \). As usual, \( \nabla \cdot g \) denotes the divergence of \( g : \nabla \cdot g(x) = \sum_i D_i g_i(x) \).

**Proof.** First note that by a simple translation of parameter, it suffices to consider \( \mu = 0 \). Next, for scalar \( C^\infty \) functions \( \psi : \mathbb{R}^d \to \mathbb{R} \) of compact support, so that formula (9.1) becomes a simple integration by parts:

\[
\int x_i \psi(x) \phi(x) \, dx = \int \psi(x) [-D_i \phi(x)] \, dx \\
= \int D_i \psi(x) \phi(x) \, dx.
\]

(9.4)

To verify (9.1) for general \( g \) we take limits in (9.4), and exploit a standard convergence criterion: suppose that \( h_i \) and \( g \) belong to \( L_1(\Phi) = L_1(\mathbb{R}^d, \phi(x) \, dx) \). Then \( h_i = D_i g \) if and only if there exists a sequence of \( C^\infty \) functions \( \{ \psi_n \} \) with \( x_i \psi_n(x) \) converging to \( x_i g(x) \) in \( L_1(\Phi) \) such that \( D_i \psi_n \to h_i \) in \( L_1(\Phi) \).

Formula (9.2) follows immediately from (9.1) (since \( E_\mu \|X - \mu\|^2 = d \)) and the extension to (9.3) is left to the reader.

\[\square\]

### 9.2. The James Stein Estimate

For \( X \sim N_d(\mu, I) \), the James-Stein estimator is defined by

\[
\hat{\mu}_{JS}(x) = (1 - \frac{d - 2}{\|x\|^2}) x,
\]

(9.5)

and was used James & Stein (1961) to give a more explicit demonstration of the inadmissibility of the maximum likelihood estimator \( \hat{\mu}_{MLE}(x) = x \) in dimensions \( d \geq 3 \). [The MLE being known to be admissible for \( d = 1, 2 \)] Later, Stein (1981) showed that the inadmissibility may be verified immediately from his unbiased risk formula (9.2). Indeed, \( g(x) = -(d - 2)\|x\|^{-2} x \) is weakly differentiable, and

\[
D_i g(x) = -(d - 2)\left( \frac{1}{\|x\|^2} - \frac{2x_i^2}{\|x\|^4} \right)
\]

so that \( \nabla \cdot g(x) = -(d - 2)^2 \|x\|^{-2} \) and so the unbiased risk estimator

\[
U(x) = d - (d - 2)^2 \|x\|^{-2}.
\]

Consequently

\[
r(\hat{\mu}_{JS}, \mu) = d - (d - 2)^2 E_\mu \|X\|^{-2},
\]

(9.6)

which is everywhere smaller than \( r(\hat{\mu}_{MLE}, \mu) = E_\mu \|x - \mu\|^2 \equiv d \) so long as \( d \geq 3 \).

**Remarks.** 1. Where does the factor \( d - 2 \) come from? Here is a partial explanation: the estimator \( \hat{\mu}(x) = (1 - \beta/\|x\|^2) x \) has unbiased risk estimate \( U_{\beta}(x) = d - \{2\beta(d - 2) - \beta^2 \}/\|x\|^2 \), and this quantity is minimized by the choice \( \beta = d - 2 \).
2. We recall that the positive part James-Stein estimator

\[ \hat{\mu}^{JS+}(x) = (1 - \frac{d-2}{\|x\|^2})_+ x \]

has necessarily even better MSE than \( \hat{\mu}^{JS} \).

The unbiased risk estimate leads to an informative bound on the mean squared error of the James-Stein rule.

9.2. Proposition. If \( X \sim N_d(\mu, I) \), then the James-Stein rule satisfies

\[ E_\mu \| \hat{\mu}^{JS} - \mu \|^2 \leq 2 + \frac{1}{(d-2) + \|\mu\|^2}. \]

Proof. For general \( \mu \), the sum of squares \( \|X\|^2 \) follows a non-central chi-squared distribution with non-centrality parameter \( \|\mu\|^2 \). The non-central distribution may be realized as a mixture of central chi-squared distributions \( \chi^2_{d+2N} \), where \( N \) is a Poisson variate with mean \( \|\mu\|^2/2 \). (cf. e.g. Johnson & Kotz (1970, p. 132)). Recall also the formula

\[ E[\chi^2_{d+2N}] = 1/(d-2). \]

Hence, by conditioning first on \( N \), and then using (9.9) and Jensen’s inequality,

\[ E1/\chi^2_{d+2N} = E1/(d-2 + 2N) \geq 1/(d-2 + \|\mu\|^2). \]

Substituting into the unbiased risk formula (9.6), we obtain

\[ r(\hat{\mu}^{JS}, \mu) \leq 2 + (d-2) - \frac{(d-2)^2}{d-2 + \|\mu\|^2}, \]

which yields the desired result after rearrangement. \( \square \)

Remark. A cruder use of Jensen’s inequality directly on the inverse sum of squares: \( E_\mu \|x\|^{-2} \geq 1/((d+\|\mu\|^2)) \) yields only the upper bound \( 4 + r(\hat{\mu}^{JS}, \mu) \).

9.2.1. Ideal Linear Shrinkage and an Oracle Inequality. Still with the model \( X \sim N_d(\mu, I) \), we consider a family of diagonal shrinkage linear estimators \( \hat{\mu}^c = (1 - c)x \). As at (7.1), we have

\[ E\|\hat{\mu}^c - \mu\|^2 = (1 - c)^2 d + c^2 \|\mu\|^2. \]

In an idealized situation in which \( \|\mu\| \) is known, the ideal choice of estimator would be

\[ \hat{\mu}^{IS}(x) = (1 - c^{IS})x, \quad c^{IS}(\mu) = \frac{d}{d + \|\mu\|^2}, \]

where the superscript \( IS \) denotes “ideal shrinkage”, and the estimator has mean squared error

\[ r(\hat{\mu}^{IS}, \mu) = d\|\mu\|^2/(d + \|\mu\|^2). \]

Combining these observations with Proposition 9.2 yields
9.3. COROLLARY. For all $\mu \in \mathbb{R}^d$,
\begin{equation}
 r(\hat{\mu}^{JS}, \mu) \leq 2 + \inf_c r(\hat{\mu}_c, \mu) = 2 + r(\hat{\mu}^{JS}, \mu).
\end{equation}

This is an example of an oracle inequality: the risk of a bona fide estimator $\hat{\mu}^{JS}$ is bounded by the risk of the ideal estimator (unusable in practice, of course) plus an additive constant. In high dimensions, the constant 2 is small in comparison with the risk of the MLE, everywhere equal to $d$. On the other hand the bound (9.13) is sharp: at $\mu = 0$, the unbiased risk equality (9.6) shows that $r(\hat{\mu}^{JS}, 0) = 2$, while the ideal risk is zero.

We now give an interpretation of the James-Stein estimator $\hat{\mu}^{JS}$ as an adaptive (quasi-) linear estimator, based on the observation that $E\mu \|X\|^2 = d + \|\mu\|^2$. This suggests that one might attempt to estimate the ideal shrinkage constant $c^{JS}$ in (9.10) by $\beta/\|x\|^2$ for an appropriate choice of $\beta$. To fix the value of $\beta$, note that when $\mu = 0$, $\|X\|^2 \sim \chi^2_0$. The formula (9.9) shows that we get an unbiased estimate of $c^{JS}$ when $\mu = 0$ by setting $\beta = d - 2$. This is precisely the James-Stein estimator.

Remark. Another approach is to compute the unbiased estimate $U$, of the risk of an estimate $\hat{\mu} = (1 - c)x$. We have $g(x) = -cx$ and $U(c, x) = d - 2dc + c^2\|x\|^2$. Minimising over $c$ gives $c_{\text{min}} = d/\|x\|^2$, which is the James-Stein estimator except for the constant multiplier.

9.3. Adaptive minimaxity

9.3.1. The problem of adaptive estimation. Consider the family of ellipsoids corresponding to smoothness constraints $\int (D^p f)^2 \leq L^2$ on periodic functions in $L_2[0, 1]$ when represented in the Fourier basis:
\begin{equation}
\Theta^\alpha(C) = \{ \theta \in \ell_2 : \sum_{l \geq 0} l^{2\alpha} (\hat{\theta}_{2l-1}^2 + \hat{\theta}_{2l}^2) \leq C^2 \} \quad \alpha, C > 0.
\end{equation}

As we have seen, Pinsker’s theorem delivers a linear estimator $\hat{\theta}(\alpha, C, \epsilon)$, given by (77), which is minimax linear for all $\epsilon > 0$, and asymptotically minimax among all estimators as $\epsilon \to 0$.

As a practical matter, the constants $(\alpha, C)$ are generally unknown, and even if one believed a certain value $(\alpha_0, C_0)$ to be appropriate, there is an issue of robustness of MSE performance of $\hat{\theta}(\alpha_0, C_0, \epsilon)$ to misspecification of $(\alpha, C)$. One possible way around this problem is to construct an estimator family $\hat{\theta}_\epsilon$, whose definition does not depend on $(\alpha, C)$, such that if $\theta$ is in fact restricted to some $\Theta^\alpha(C)$, then $\hat{\theta}_\epsilon$ has MSE appropriate to that space:
\begin{equation}
\sup_{\theta \in \Theta^\alpha(C)} r_\epsilon(\hat{\theta}_\epsilon, \theta) \leq c_\epsilon(\Theta) R_N(\Theta^\alpha(C), \epsilon) \quad \text{as } \epsilon \to 0.
\end{equation}

where $c_\epsilon(\Theta)$ is a bounded sequence. Write $\mathcal{T}_2$ for the collection of all ellipsoids $\{ \Theta^\alpha(C) : \alpha, C > 0 \}$. One then calls $\hat{\theta}_\epsilon$ rate-adaptive: it “learns” the right rate of convergence for all $\Theta \in \mathcal{T}_2$. 

107
If $\theta^j$ has the stronger property that $c_\epsilon(\Theta) \to 1$ for each $\Theta \in \mathcal{T}_2$, then it is called *adaptively asymptotically minimax*: it gets the constant right as well. An adaptive minimax estimator sequence for Sobolev ellipsoids $\mathcal{T}_2$ was constructed by Efroimovich & Pinsker (1984), and this section presents their blockwise estimator approach, fortified with use of the James–Stein method. We will see that good non-asymptotic bounds are also possible, and that the dyadic-blocks James–Stein estimator is a plausible estimator for practical use in appropriate settings.

**9.3.2. Blockwise Estimators.** Consider the Gaussian sequence model, at first with abstract countable index set $\mathcal{I}$: $y_I = \theta_I + \epsilon z_I$. Suppose that $\mathcal{I}$ is partitioned into an ordered sequence of blocks $B_j$ of finite cardinality $n_j$. We write $y_j$ for the vector of coefficients $\{y_I, I \in B_j\}$, and similarly for $\theta_j, z_j$ etc.

In the case that $\mathcal{I} = \mathbb{N}$, suppose that the blocks are defined by an increasing sequence $l_j$:

\[(9.16) \quad B_j = \{l_j + 1, l_j + 2, \ldots, l_{j+1}\}, \quad n_j = l_{j+1} - l_j.\]

**Example. Dyadic blocks.** Here $n_j = 2^j$. In the dyadic sequence model, we have $B_j = \{(j, k), k = 1, \ldots, 2^j\}$. In the linear sequence model, set $l_j = 2^j$ and $B_j = \{2^j + 1, 2^j + 2, \ldots, 2^{j+1}\}$. Again in the linear sequence model, define *dyadic Sobolev ellipsoids* using weights $a_l = 2^{j\alpha}$ if $l \in B_j$:

\[(9.17) \quad \Theta^2_j(C) = \{\theta : \sum_j 2^{2j\alpha} \sum_{l \in B_j} \theta^2_l \leq C^2\}.\]

Let $\mathcal{T}^{D, 2}$ denote the class of dyadic Sobolev ellipsoids $\{\Theta^2_j(C), \alpha, C > 0\}$. When functions are represented using wavelet bases, the dyadic ellipsoids (9.17) are the appropriate way to measure Hilbert–Sobolev smoothness (cf. Appendix ??). In any event, the two approaches are norm-equivalent: write $\|\theta\|_{\ell_{2}, \alpha}$ for the squared seminorm appearing in (9.14) and $\|\theta\|_{\ell_{\alpha}, \alpha}$ for that appearing in (9.17). It is easily seen that for all $\theta \in \ell_2$:

\[(9.18) \quad \|\theta\|_{\ell_{2}, \alpha} \leq \|\theta\|_{\ell_{\alpha}, \alpha} \leq 2^{\alpha} \|\theta\|_{\ell_{2}, \alpha}.\]

**Block diagonal linear estimators.** This term refers to diagonal linear estimators in which the shrinkage factor is constant within blocks: for all blocks $j$:

\[\hat{\theta}_{j,c_j}(y) = c_j y_j \quad c_j \in \mathbb{R}.\]

The corresponding minimax risk over block linear estimators is then

\[R_{BL}(\Theta, \epsilon) = \inf_{\theta} \sup_{(c_j)} \sum_j r_\epsilon(\hat{\theta}_{j,c_j}, \theta_j).\]

The minimax theorem for diagonal linear estimators, Theorem 7.9 and its proof, have an immediate analog in the block case.
9.4. Proposition. If $\Theta$ is compact, solid-orthosymmetric and quadratically convex, then

$$R_{BL}(\Theta, \varepsilon) = \sup_{\Theta} \inf_{\varepsilon_j} \sum_j r_{\varepsilon} \langle \hat{\theta}_j, \varepsilon_j, \hat{\theta}_j \rangle. \quad (9.19)$$

To rewrite (9.19), we adopt the ideal linear shrinkage interpretation from Section 9.2.1. Indeed the sum in (9.19) is minimized term by term, by the blockwise ideal shrinkage estimators given by (9.11) and with corresponding ideal MSE’s given by (9.12). Thus

$$R_{BL}(\Theta, \varepsilon) = \sup_{\Theta} \sum_j r_{\varepsilon} \langle \hat{\theta}_j^S, \hat{\theta}_j \rangle = \sup_{\Theta} \sum_j \frac{n_j \varepsilon^2 \|\theta_j\|^2}{n_j \varepsilon^2 + \|\theta_j\|^2}. \quad (9.20)$$

**Block Linear versus Linear.** Clearly $R_L(\Theta, \varepsilon) \leq R_{BL}(\Theta, \varepsilon)$. In two cases, more can be said:

(i) Call $\Theta$ block symmetric if $\Theta$ is invariant to permutations of the indices $I$ within blocks. A variant of the argument in Section 7.2.1 employing random block-preserving permutations (instead of random signs) shows that if $\Theta$ is solid, ortho- and block-symmetric, then

$$R_L(\Theta, \varepsilon) = R_{BL}(\Theta, \varepsilon) \quad \text{for all } \varepsilon > 0. \quad (9.21)$$

The dyadic Sobolev ellipsoids $\Theta^d_j(C)$ are block symmetric and so are an example for (9.21).

(ii) For general ellipsoids $\Theta(a, C)$ as in (8.1), and a block scheme (9.16), measure the oscillation of the weights $a_l$ within blocks by

$$\text{osc}(B_j) = \max_{l,l' \in B_j} \frac{a_l}{a_{l'}}$$

It follows (exercise) from the linear minimax risk formula (?) that if $a_k \to \infty$ and $\text{osc}(B_j) \to 1$, then

$$R_L(\Theta, \varepsilon) \sim R_{BL}(\Theta, \varepsilon) \quad \text{as } \varepsilon \to 0. \quad (9.22)$$

In the Fourier ellipsoid case, (9.22) applies to all $\Theta(\alpha, C)$ if one uses blocks $B_j$ defined by either $l_j = (j + 1)^\beta$ for $\beta > 0$, or $l_j = e^{\sqrt{7} - j}$ in either case

$$\text{osc}(B_j) = (l_j+1/l_j)^\alpha \to 1.$$ The block sizes are necessarily subgeometric in growth: for dyadic blocks, $l_j = 2^j$, the condition fails: $\text{osc}(B_j) \to 2^\alpha$.

9.3.3. Blockwise James Stein Estimation. We construct an estimator which on each block $B_j$ applies the positive part James-Stein estimator (9.7):

$$\hat{\theta}_j^S(y_j) = \left(1 - \frac{(n_j - 2)\varepsilon^2}{\|y_j\|^2}\right) y_j. \quad (9.23)$$

A key benefit of the James-Stein estimate is the good bounds for its MSE: the consequences (9.13) and (9.12) of Proposition 9.2 show that when $n_j \geq 3$,

$$r_e(\hat{\theta}_j^S, \theta_j) \leq 2\varepsilon^2 + \frac{n_j \varepsilon^2 \|\theta_j\|^2}{n_j \varepsilon^2 + \|\theta_j\|^2} = 2\varepsilon^2 + r_e(\hat{\theta}_j^S, \theta_j). \quad (9.24)$$
The full blockwise estimator, $\hat{\theta}^{BJS}$, is then defined by

\begin{equation}
\hat{\theta}^{BJS}_j(y) = \begin{cases} y_j & j < L \\
\hat{\theta}^{JS}_j(y_j) & L \leq j < J_e \\
0 & j \geq J_e \end{cases}
\end{equation}

(9.25)

For the ‘earliest’ blocks, specified by $L$, no shrinkage is performed. This may be because the blocks are of small size ($n_j \leq 2$), or are known to contain very strong signal, as is often the case if the blocks represent the lowest frequency components.

No blocks are estimated after $J_e$. Usually $J_e$ is chosen so that $l_e = l_J = \epsilon^{-2}$, which equals the sample size $n$ in the usual calibration. This restriction corresponds to not attempting to estimate, even by shrinkage, more coefficients than there is data.

It is now straightforward to combine earlier results to obtain risk bounds for $\hat{\theta}^{BJS}$ that will also show in many cases that it is asymptotically minimax. Since $r_e(\hat{\theta}^{BJS}, \theta) = \sum_j r_e(\hat{\theta}^{JS}_j, \theta_j)$, we obtain from (9.24) and (9.25) the pointwise (i.e. fixed $\theta$) bound

\begin{equation}
r_e(\hat{\theta}^{BJS}, \theta) \leq (l_L + 2J_e - 2L)\epsilon^2 + \sum_{j=L}^{J_e-1} r_e(\hat{\theta}^{JS}_j, \theta_j) + \sum_{l \geq l_e} \theta_l^2.
\end{equation}

Maximizing over any parameter space $\Theta$ for which the consequence (9.20) of Proposition 9.4 applies, we find

\begin{equation}
\sup_{\Theta} r_e(\hat{\theta}^{BJS}, \theta) \leq (l_L + 2J_e)\epsilon^2 + R_{BL}(\Theta, \epsilon) + \Delta_e(\Theta).
\end{equation}

(9.27)

In the case of ellipsoids $\Theta(a, C)$ with $a_l \not\to \infty$, the squared maximal tail bias

\begin{equation}
\Delta_e(\Theta) = \sup\{\sum_{l \geq l_e} \theta_l^2 : \sum_{l \geq l_e} a_l^2 \theta_l^2 \leq C^2\} = C^2 a_l^{-2}.
\end{equation}

(9.28)

9.5. Example. Dyadic Blocks. In the case of dyadic blocks, one might therefore reasonably take $L = 2$ and $J_e = \log_2 \epsilon^{-2}$. The success of octave based thinking in harmonic analysis and wavelet methodology gives a distinguished status to the use of dyadic blocks vis-a-vis other choices, and we will focus most attention on this version. Note also that there are no parameters left unspecified, except the noise level $\epsilon$, here assumed known.

(To be discussed in estimation of $\epsilon$)

9.6. Theorem. The dyadic blocks James-Stein estimator is adaptive minimax over the full class of dyadic Sobolev ellipsoids: for each $\Theta \in \mathcal{T}_{D,2}$:

\begin{equation}
\sup_{\Theta \in \Theta} r_e(\hat{\theta}^{BJS}, \theta) \sim R_N(\Theta, \epsilon) \quad \epsilon \to 0.
\end{equation}

Proof. This is now a simple consequence of (9.27). Let $\Theta = \Theta_{D}(C) \in \mathcal{T}_{D,2}$. From block symmetry, (9.22) and (8.8),

\begin{equation}
R_{BL}(\Theta, \epsilon) = R_L(\Theta, \epsilon) \asymp \epsilon^r \quad r = 2\alpha/(2\alpha + 1).
\end{equation}

110
Since $L = 2$ and $J_e = \log_2 \epsilon^{-2}$, we have $l_e = 4$ and so the first right hand term in (9.27) is $o(\epsilon^{2\nu})$. For the tail bias term, we note that $l_e = \epsilon^{-2}$ and so, from (9.28), $\Delta_\epsilon(\Theta) \leq C(\epsilon^2)^{2\alpha} = o(\epsilon^{2\nu})$. Thus, as $\epsilon \to 0$,
\[ \sup_{\Theta} \epsilon_\Theta(B_{\text{JS}}, \theta) \leq R_{\text{NL}}(\Theta, \epsilon)(1 + o(1)) \sim R_N(\Theta, \epsilon), \]
Since $R_N(\Theta, \epsilon)$ is by definition a lower bound for the left side, the result is proved. 

The proof just given clearly extends to other blocking schemes $\{B_j\}$ and other classes of ellipsoids. Rather than formulate a very general result, we make a few remarks on particular cases.

For traditional Sobolev ellipsoids, dyadic blocks are too large, since with $a_l \sim 1^{1/3}, \text{osc} (B_j) \to 2^\alpha$, and so one has only rate adaptivity: $R_{\text{BL}}(\Theta, \epsilon) \leq 2^{2\alpha} R_N(\Theta, \epsilon)(1 + o(1))$. However exact adaptation can be achieved with smaller block sizes, for which $\text{osc} B_j \to 1$. Thus $l_j = e^\nu T$ works, for example. However, the sequence $l_j = (j + 1)^{\beta}$ is less satisfactory, since $l_e = \epsilon^{-2}$ implies that $J_e = \epsilon^{-2\beta}$ and so $\epsilon^2 J_e$ is not $o(\epsilon^{2\nu})$ in the smoother cases, when $2\alpha + 1 \geq \beta$.

In fact, this last problem arises from the bound $2\epsilon^2$ in (9.24), and could be reduced by using a modified estimator $\hat{\theta}_j = (1 - \gamma \epsilon^2 / \|u_j\|^2) u_j$ with $\gamma \in (n_j - 2, 2n_j]$. This reduces the error at zero to essentially a large deviation probability (see e.g. Brown et al. (1997), who use $\frac{1}{2} n_j$). However, in overall practical and MSE performance, the choice $n_j - 2$ has been preferred, and we have chosen to establish theoretical results for an estimator closer to that which one might use in practice.

9.3.4. WaveJS as a representative adaptive linear rule. Consider the continuous Gaussian white noise model (3.4) or equivalently its sequence space counterpart (3.1) in the Fourier basis. Many standard linear estimators can be represented in this basis in the form

(9.29) \[ \hat{\theta}_k = \hat{h}(hk) v_k. \]

As examples, we cite

1. Weighted Fourier series. The function $\hat{h}$ decreases with increasing frequency, corresponding to a downweighting of signals at higher frequencies. The parameter $h$ controls the actual location of the “cutoff” frequency band.

2. Kernel estimators. In the time domain, the estimator has the form $\hat{\theta}(t) = \int h^{-1} K(h^{-1}(t - s)) dY(s)$, for a suitable kernel function $K(\cdot)$, typically symmetric about zero. The parameter $h$ is the bandwidth of the kernel. The representation (9.29) follows after taking Fourier coefficients.

3. Smoothing splines. The estimator $\hat{\theta}_k$ minimizes

\[ \sum (v_k - \hat{\theta}_k)^2 + \lambda^{2r} \sum k^{2r} \hat{\theta}_k^2, \]

where the penalty term viewed in the time domain takes the form of a derivative penalty $\int (D^r \theta)^2$ for some integer $r$. In this case, $\hat{\theta}_k$ again has the representation (9.29) with $\hat{h}(\lambda k) = [1 + (\lambda k)^{2r}]^{-1}$. 

111
(comment on local polynomial fitting.)

In addition, many methods of choosing \( h \) from the data \( y \) have been shown to be asymptotically equivalent to first order (see e.g. Härdle et al. (1988)) - these include cross validation, Generalized cross validation, Rice’s method based on unbiased estimates of risk, final prediction error, Akaike information criterion.

To see the connection with estimator (9.25), consider a Taylor expansion of \( \tilde{h}(s) \) about \( s = 0 \). If the time domain kernel \( K(t) \) corresponding to \( \tilde{h} \) is even about 0, then the odd order terms vanish and \( \tilde{h}(s) = 1 + k_2 s^2/2 + k_4 s^4/4! + \ldots \), so that for \( h \) small and a positive integer \( r \), \( \tilde{h}(hk) = 1 - c_r k^{2r} \).

Now consider grouping the indices \( k \) into blocks \( B_j \) - for example, dyadic blocks \( B_j = \{ k : 2^j < k \leq 2^{j+1} \} \). Then the weights corresponding to two indices \( k, \tilde{k} \) in the same block are essentially equivalent: \( k^{2r} / \tilde{k}^{2r} \in [c_r^{-1}, c_r] \) so that we may approximately write

\[
\hat{\theta}_k = (1 - c_j) y_k, \quad k \in B_j
\]

Here \( c_j \) depends on \( h \), but this is not shown explicitly, since we are about to determine \( c_j \) from the data \( y \) anyway.

For example, we might use Stein’s unbiased risk criterion to estimate \( c_j \) as follows. Indeed, \( \theta_j(y) = y_j - c_j y_j \), so that in the formulation of Section 9.2, \( g(y) = -c_j y_j \), with noise level \( \epsilon^2 \). From formula (9.3) we find

\[
U_c(y) = n_j \epsilon^2 - 2n_j \epsilon^2 c_j + c_j^2 |y_j|^2.
\]

This is in fact the same as the \( C_L \) formula of Mallows (1973). The value of \( c_j \) that minimizes (9.31) is \( c_j = n_j \epsilon^2 / \|y_j\|^2 \), which differs from the James-Stein estimate (9.23) only in the use of \( n_j \) rather than \( n_j - 2 \).

Thus, many standard linear methods are closely related to the diagonal linear shrinkage estimator (9.30). Denote by \( \hat{\theta}^{LPS}(y) = (\hat{\theta}^{LPS}_j(y_j)) \) the estimator obtained by choosing for \( c_j \) the James-Stein shrinkage factor in (9.25). Dyadic blocking in the frequency domain is a key feature of Littlewood-Paley theory in harmonic analysis, hence the letters LP.

Figure adfig04.ps shows a reconstruction of the four basic functions using \( \hat{\theta}^{LPS}(y) \). Figure adfig13.ps shows a corresponding reconstruction using the estimator \( \hat{\theta}^{WJS}(y)(t) = \sum_j \hat{\theta}^{WJS}_j \psi_j(t) \) in a wavelet basis. The qualitative similarity of the two reconstructions may be explained as follows. The estimator (9.30) is invariant with respect to orthogonal changes of basis for the vector \( y_j = (y_k : k \in B_j) \). To the extent that the frequency content of the wavelets spanning the wavelet multiresolution space \( W_j \) is concentrated on a single frequency octave (only true approximately), it represents an orthogonal change of basis from the sinusoids belonging to that octave.

asfig05.ps and asfig06.ps show reconstructions obtained using smoothing splines and truncated Fourier series respectively. In each case the regularization parameter/truncation parameter is chosen from the data to minimize an unbiased estimate of mean-squared error. The qualitative similarity
of the reconstructions with those of $\hat{\theta}^{LPJS}$ and $\hat{\theta}^{WJS}$ is not surprising in view of the arguments given above.

Note that none of these methods are effective at simultaneously removing high frequency noise and maintaining the sharpness of jumps and peaks.

**Issues for figures here:** a) Cut down number?, b) introduce four functions and pictures of noise level?, c) ensure that same noise is used in all plots.

**Figure 1.** adfig04.eps from AdaptDemo
**Figure 2. adfig13.eps from AdaptDemo**

**Figure 3. asfig05.eps from AsympDemo**
Figure 4. asfig06.eps from AsympDemo
9.4. Interlude: Superefficiency

This section tries to illuminate one of the motivations that underlies the use of worst-case and minimax analyses both in these notes and in the literature of non-parametrics: a desire for a robust alternative to “fixed θ” asymptotics.

Fixed θ asymptotics refers to a mode of analysis in which the unknown function θ is kept fixed, and the behavior of an estimator sequence \( \hat{\theta}_n \) is analysed, for example through mean squared error, as \( \epsilon \to 0 \). Asymptotic approximations might then be used to optimize parameters of the estimator — such as bandwidths or regularization parameters — or to assert optimality properties.

This mode of analysis has generally been very effective in large sample analysis of finite dimensional models. While problems such as superefficiency surfaced in trying to make the approach theoretically watertight, the phenomena encountered are not serious enough to affect the broad validity of the practical implications widely drawn, for example, from Fisher’s asymptotic theory of maximum likelihood. Thus, the set of superefficiency for any estimator sequence has measure zero, and estimators superefficient at one point typically have sufficiently poor behavior at neighboring points as to render them practically irrelevant.

In nonparametric problems with infinite dimensional parameter spaces, however, fixed θ asymptotics is a more fragile tool. When used with care, it still of course yields much useful information. However, if optimization is pushed too far, it can suggest conclusions valid only for implausibly large sample sizes, and seriously misleading for actual practice. In nonparametrics, superefficiency is a much more pervasive phenomenon: even very practical estimators can exhibit superefficiency at every parameter point, and poor behaviour in a neighbourhood of any fixed parameter point is a necessary property of every estimator sequence.

After reviewing Hodges’ classical example of parametric superefficiency, the bulk of this section is devoted to illustrations of these points - some concluding contrasting remarks are made about worst-case and minimax analysis.

9.4.1. Parametric Estimation: the Hodges example. Suppose that \( y \sim N(\theta, \epsilon^2) \) is a single scalar observation with \( \epsilon \) small. A rather special case of Fisherian parametric asymptotics asserts that if \( \hat{\theta}_n \) is an asymptotically normal and unbiased estimator sequence, \( \epsilon^{-1}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, v(\theta)) \) when \( \theta \) is true, then necessarily \( v(\theta) \geq 1 \). A consequence for mean squared error would then be that

\[
\liminf_{\epsilon \to 0} \epsilon^{-2} E_{\theta}(\hat{\theta}_n - \theta)^2 = \liminf_{\epsilon \to 0} r_\epsilon(\hat{\theta}_n, \theta)/R_N(\Theta, \epsilon) \geq 1.
\]
For this subsection, $\Theta = \mathbb{R}$, Hodges' counterexample modifies the MLE

$$\hat{\theta}(y) = \begin{cases} 0 & |y| < \sqrt{\epsilon} \\ y & \text{otherwise.} \end{cases}$$

Since $\sqrt{\epsilon} = \frac{1}{\sqrt{\epsilon}}$, $\epsilon$ is many standard deviations in size, it is clear that if $\theta = 0$, this estimator has MSE equal to $2\epsilon^2 \int_{-\epsilon}^{\epsilon} y^2 \phi(y) dy \approx 2\epsilon^2$. On the other hand, if $\theta \neq 0$ and $\epsilon$ is small, and noting the rapid decay of the tails of the Gaussian distribution, then the interval $[-\sqrt{\epsilon}, \sqrt{\epsilon}]$ is essentially irrelevant to estimation of $\theta$, and so

$$e^{-2}E_\theta(\hat{\theta} - \theta)^2 \to \begin{cases} 0 & \text{if } \theta = 0, \\ 1 & \text{otherwise}, \end{cases}$$

in clear violation of the Fisherian program. A fuller introduction to this and related superefficiency issues appears in Lehmann (1983, Section 6.1), along with references to the primary research of LeCam, Bahadur, Huber and Hajek. Here we simply note two phenomena which are also characteristic of more general parametric settings:

(i) points of superefficiency are rare: in Hodges' example, only at $\theta = 0$. More generally, the set of fixed parameter points $\theta$ for which

$$\lim_{\epsilon \to 0} \sup \frac{R_N(\Theta, \epsilon)}{\tau_\epsilon(\hat{\theta}, \theta)} < 1$$

has measure zero.

(ii) Superefficiency entails poor performance at nearby points. For Hodges' example, consider $\theta_c = \sqrt{\epsilon}/2$. Since the threshold zone extends $1/(2\sqrt{\epsilon})$ standard deviations to the right of $\theta_c$, it is clear that $\hat{\theta}_c$ makes a squared error of $(\sqrt{\epsilon}/2)^2$ with overwhelming probability, and so

$$e^{-2}r(\hat{\theta}_c, \frac{\sqrt{\epsilon}}{2}) = \frac{1}{\epsilon}(\frac{\sqrt{\epsilon}}{2})^2 \to \infty,$$

so that

$$\sup_{|\theta| \leq \sqrt{\epsilon}} \frac{r(\hat{\theta}_c, \theta)}{R_N(\Theta, \epsilon)} \to \infty.$$  

LeCam, Huber and Hajek showed that more generally, superefficiency at $\theta_0$ forces poor properties in a neighborhood of $\theta_0$. Since broadly efficient estimators such as maximum likelihood are typically available with good risk properties, this phenomenon makes superefficiency of little relevance to the use of reasonable estimators in the parametric setting.

Remark. The estimator of Hodges' counterexample is an example of hard thresholding, to be discussed in some detail in Chapters 10 and 11 and later in connection with wavelet shrinkage in non-parametric estimation. It is indeed curious that the points and phenomena of superefficiency that are
ultimately unimportant in the practical implications of the one-dimensional theory go over into essential features associated with sparsity of high-dimensional signals in the nonparametric setting. The “poor performance at nearby points” re-emerges in the non-parametric theory as the least favorable situations that determine the best upper bounds in oracle inequalities.

9.4.2. Nonparametrics: Superefficiency everywhere. We return to the nonparametric setting, always in the Gaussian sequence model. Previous sections argued that the dyadic blocks James-Stein estimate (cf. (9.25) and Example 9.5) is a theoretically and practically promising method. Nevertheless, every fixed $\theta$ is a point of superefficiency in the sense of (9.32):

9.7. Proposition. (Brown et al. 1997) Let $\Theta = \Theta^\prime(C)$ be a Sobolev ellipsoid (9.14). Then for every $\theta \in \Theta$,

$$r_\epsilon(\hat{\theta}_{\text{BJS}}, \theta) R_N(\Theta, \epsilon) \to 0.$$  

(9.34)

Thus, if $\Theta$ corresponds to functions with second derivative ($m = 2$) having $L_2$ norm bounded by 1, say, then for any fixed such function, the blockwise James-Stein estimator has rate of convergence faster than $\epsilon^{8/5} \to n^{-4/5}$. Brown et al. (1997) also show that convergence cannot, in general, be very much faster – at best of logarithmic order in $\epsilon^{-1}$ – but the fact remains that the fixed $\theta$ rate is always slightly different from that of a natural minimax benchmark. Of course, in parametric problems, the rate of convergence is the same at almost all points.

Proof. Fix $\Theta = \Theta^\prime(C)$ and recall from (8.8) that $R_N(\Theta, \epsilon) \approx \epsilon^{2r}$ as $\epsilon \to 0$, with $r = 2m/(2m + 1)$. A “fixed $\theta$” bound for the risk of $\hat{\theta}_{\text{BJS}}$ follows from (9.26) : indeed, since $L = 2$ and $ab/(a + b) \leq \min(a, b)$, we may write

$$r_\epsilon(\hat{\theta}_{\text{BJS}}, \theta) \leq 2J_\epsilon \epsilon^2 + \sum_j \min(n_j \epsilon^2, \| \theta_j \|^2) + \sum_{j > 0} \theta_j^2.$$  

The proof of Theorem 9.6 showed that the first and third terms were $o(\epsilon^{2r})$, uniformly over $\theta \in \Theta$. Consider, therefore, the second term, which we write as $R_1(\theta, \epsilon)$. For any $j$, use the variance component below $j$, and the bias term thereafter:

$$R_1(\theta, \epsilon) \leq 2J_\epsilon \epsilon^2 + 2^{-2m} \sum_{j \geq j_\epsilon} 2^{m_j} \| \theta_j \|^2.$$  

To show that $R_1(\theta, \epsilon) = o(\epsilon^{2r})$, first fix a $\delta > 0$ and then choose $j_\epsilon$ so that $2^{j_\epsilon} \epsilon^2 = \delta \epsilon^{2r}$. Of course, $j_\epsilon$ should be an integer, but there is no harm in ignoring this point. It follows that $2^{-2m_j} = \delta^{-2m\epsilon^{2r}}$, and so

$$\epsilon^{-2r} R_1(\theta, \epsilon) \leq \delta + \delta^{-2m} \sum_{j \geq j_\epsilon} 2^{m_j} \| \theta_j \|^2 = \delta + o(1),$$  

since the tail sum vanishes as $\epsilon \to 0$, for $\theta \in \Theta^\prime(C)$. Since $\delta > 0$ is arbitrary, this shows that $R_1(\theta, \epsilon) = o(\epsilon^{2r})$ and establishes (9.34). \qed

118
9.4.3. Nearby instability is unavoidable. The next result shows that for every consistent estimator sequence, and every parameter point \( \theta \in \ell_2 \), there exists a shrinking \( \ell_2 \) neighborhood of \( \theta \) over which the worst case risk of the estimator sequence is arbitrarily worse than it is at \( \theta \) itself. Compare (9.33). In parametric settings, such as the Hodges example, this phenomenon occurs only for unattractive, superefficient estimators, but in nonparametric estimation the property is ubiquitous. Here, neighborhood refers to balls in \( \ell_2 \) norm: \( B(\theta_0, \eta) = \{ \theta : \| \theta - \theta_0 \|_2 < \eta \} \). The root of the problem is that such neighborhoods do not have compact closure in \( \ell_2 \), but the fact remains that fixed \( \theta \) asymptotics does not give any hint of the perils that lie arbitrarily close nearby.

9.8. Proposition. Suppose that \( \hat{\theta}_e \) is any estimator sequence such that \( r_e(\hat{\theta}_e, \theta) \to 0 \). Then there exists \( \eta_e \to 0 \) such that as \( \epsilon \to 0 \),

\[
(9.35) \quad \sup_{\hat{\theta} \in B(\theta_0, \eta_e)} \frac{r_e(\hat{\theta}_e, \theta)}{r_e(\hat{\theta}_e, \theta_0)} \to \infty
\]

Remark. The result remains true if the neighborhood \( B(\theta_0, \eta_e) \) is replaced by its intersection with any dense set: for example, the class of infinitely differentiable functions (represented, for example, in a wavelet or Fourier basis).

Proof. Let \( \gamma^2 = r_e(\hat{\theta}_e, \theta_0) : \) we show that \( \eta_e = \sqrt{\gamma} \) will suffice for the argument. The proof is a simple consequence of the fact that \( B(1) \) is not compact (compare Theorem 8.4 or the example following Theorem 7.9), so that \( R_N(B(1), \epsilon) \geq c_0 > 0 \) even as \( \epsilon \to 0 \). All that is necessary is to rescale the estimation problem by defining \( \tilde{\theta} = \eta_e^{-1}(\theta - \theta_0) \), \( \tilde{y} = \eta_e^{-1}(y - \theta_0) \), \( \tilde{\epsilon} = \eta_e^{-1} \epsilon \), and so on. Then \( \tilde{y} = \tilde{\theta} + \tilde{\epsilon} \) is an instance of the original Gaussian sequence model, and \( B(\theta_0, \eta_e) \) corresponds to the unit ball \( B(1) = \{ \| \tilde{\theta} \|_2 \leq 1 \} \). Rescaling the estimator also via \( \hat{\theta}_e(\tilde{y}) = \eta_e^{-1}(\hat{\theta}_e(y) - \theta_0) \),

\[
\gamma^{-2} E \| \hat{\theta}_e - \theta \|^2 = \eta_e^{-2} E \| \hat{\theta}_e(\tilde{y}) - \tilde{\theta} \|^2,
\]

and so, writing \( S_e \) for the left side of (9.35), we obtain

\[
S_e \geq \gamma^{-1} R_N(B(1), \epsilon) \geq c_0 \gamma^{-1} \to \infty.
\]

\[ \square \]

9.4.4. Ultra-asymptotic bandwidth selection. Here is a “fixed-f” argument that is frequently encountered in asymptotics. In the setting of density estimation, suppose that \( X_1, \ldots, X_n \) are i.i.d. samples from an unknown smooth probability density \( f(x) \) on \( \mathbb{R} \), and that a kernel estimator

\[
(9.36) \quad f_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^n K\left( \frac{x - x_i}{h} \right)
\]

is constructed using a \( q \)-th order kernel: \( \int K = 1 \), \( \int t^k K(t) dt = 0 \) for \( k = 1, \ldots, q - 1 \), and \( = c_0 \neq 0 \) for \( k = q \). Then a variance-bias asymptotic
expression for the mean integrated squared error \( r_n(h) = E\langle \hat{f}_{n,h} - f \rangle^2 \) is given by

\[
r_n(h) \sim \frac{c_0(K)}{nh} + c_1(K)h^{2q} \int (D^q f)^2
\]

as \( n \to \infty \). Then \( r_n(h) \) is minimized at a bandwidth \( h = h_n(f) \), and the minimum value \( r_n(h_n(f)) \) converges to zero at rate \( n^{-2q/(2q+1)} \). Since \( h_n(f) \) still depends on the unknown function \( f \), the “plug-in” approach inserts a preliminary estimator \( \hat{f}_n \) of \( f \), and uses \( h_n(\hat{f}_n) \) in (9.36). This approach goes back at least to Woodroofe (1970), for further references and discussion see Brown et al. (1997). (and Scott or Jones/Wand books?)

**Analog in the white noise model.** Here is a version of this argument in the sequence model (3.1), which allows exact calculation of the small sample consequences of this asymptotic bandwidth selection argument. (Ref to Marron?)

We use the Fourier basis with \( Z \) as index, so that positive integers \( l \) label cosine terms of frequency \( l \) and negative \( l \) label the sine terms. As in Section 9.3.4, we represent a kernel estimator in the Fourier domain by diagonal shrinkage

\[
\hat{h}_{l,t} = \kappa(hl)y_l,
\]

where \( \kappa(s) = \int e^{-ist}K(t)dt \) is the Fourier transform of kernel \( K \). The \( q \)-th order moment condition becomes a statement about derivatives at zero: \( \kappa(0) = 1, \kappa^{(k)}(0) = 0 \) for \( k = 1, \ldots, q-1 \) and equals \( c_q \neq 0 \) for \( k = q \). To simplify calculations, we use a specific choice of \( q \)-th order kernel:

\[
\kappa(s) = (1 - |s|^q)^+.
\]

For this kernel, the mean squared error of (9.38) can be written explicitly as

\[
r_n(\hat{h}, \theta) = \sum_{|l| \leq |h|^{-1}} c^2(1 - |hl|^q)^2 + |hl|^{2q}\theta^2 + \sum_{|l| > |h|^{-1}} \theta^2.
\]

By using integral approximations to sums, an asymptotic approximation to (9.40) is given by

\[
r_{n,e}(\hat{h}, \theta) = a_q \epsilon^2 h^{-1} + b_q(\theta) h^{2q},
\]

which is exactly analogous to (9.37). Here \( a_q = 4q^2 (2q+1)^{-1}(q+1)^{-1} \), and \( b_q(\theta) = \sum l^{2q}\theta^2 \), is proportional to \( \int (D^q f)^2 \) when expressed in terms of \( f \). In order that \( b_q(\theta) < \infty \) for all \( q \), we assume that \( f \) is infinitely differentiable. The bandwidth that is asymptotically MSE-optimal is found by minimizing \( h \to r_{n,e}(\hat{h}, \theta) \) - using the Variance-Bias Lemma 4.2 gives

\[
h_e = h_e(\theta) = \left[ \frac{a_q}{2qb_q(\theta)} \right]^{1/2q+1},
\]
with the corresponding optimal MSE being

\[(9.42) \quad r_c(\hat{\theta}_h(\theta), \theta) \sim c_q \left(2q \beta_0(\theta)\right)^{-1} \left(a_q c^2\right)^{2q+1},\]

with \(c_q = 1 + (2q)^{-1}\). Thus the rate of convergence is \(2q/(2q + 1)\), and hence reflects only the order of the kernel used and nothing of the properties of \(f\). Although this already is suspicious, it would seem, so long as \(f\) is smooth, that the rate of convergence can be increased arbitrarily closely to 1, simply by using a kernel of sufficiently high order \(q\).

Of course, this is an over literal use of fixed \(\theta\) asymptotics – a hint of the problem is already suggested by the constant term in (9.42), which depends on \(\beta_0(\theta)\) and could grow rapidly with \(q\). However, we may go further and do exact MSE calculations with formula (9.40) using kernel (9.39). As specific test configurations, we use the functions

\[(9.43) \quad f(x) = \sum_{l \geq 0} \theta_l \cos 2\pi lx + \sum_{l < 0} \theta_l \sin 2\pi lx\]

with

\[(9.44) \quad \theta_l = c(l_1, l_2) \begin{cases} |l|^{-3} & \text{even, } l \in [l_1, l_2] \\ |l|^{-3} & \text{odd, } -l \in [l_1, l_2] \\ 0 & \text{otherwise,} \end{cases}\]

and with \(c(l_1, l_2)\) chosen so that a Sobolev 2nd derivative smoothness condition holds: \(\sum l^4 \theta_l^2 = C^2\). Two specific choices are

(I) \(l_1 = 4, \ l_2 = 20, \ C = 60,\)

(II) \(l_1 = 4, \ l_2 = 400, \ C = 60,\)

which differ only in the number of high frequency terms retained. Note that \(\theta^I\) and \(\theta^{II}\) both correspond to \(C^\infty\) functions, and Figure 5 shows that they are visually similar, and would be difficult to distinguish without very large amounts of data in the regression model (3.3).

![Figure 5](image-url)  

**Figure 5.** Two \(C^\infty\) functions, defined at (9.43) - (9.44). Solid line is \(\theta^I\), containing frequencies only through \(l = 20\), dashed line is \(\theta^{II}\), with frequencies up to \(l = 400\).
Figure 6. MSE of ideal bandwidth choice for $\theta^{II}$: $r_c(\hat{h}_c(\theta^{II}), \theta^{II})$ resulting from $q$-th order optimal bandwidth (9.41) for $q = 2, 4, 8$ with exact risks calculated using (9.40). Also shown is the upper bound (9.26) for the risk of the dyadic blocks James Stein estimator (9.25).

Figure 7. Corresponding plot of MSEs and James-Stein bound for ideal bandwidth choice for $\theta^{I}$.

Figure 8. Ratio of James Stein MSE bound to actual MSE for kernels of order $q = 2, 4, 8$ at $\theta = \theta^I$ (dotted) and $\theta^{II}$ (solid) over a wide range of sample sizes $n = \epsilon^{-2}$. 

122
Figure 7 shows the MSE $r_e(\hat{\theta}_n(\theta^I), \theta^I)$ occasioned by using the $q$-th order optimal bandwidth (9.41) for $q = 2, 4, 8$ with exact risks calculated using (9.40). Clearly the 8th order kernel is always several times worse than the 2nd order kernel for $n = e^{-2}$ less than $10^6$. The 4th order kernel will dominate $q = 2$ for $n$ somewhat larger than $10^6$, but $q = 8$ will dominate only at absurdly large sample sizes.

Figure ?? shows that the situation is not so bad in the case of curve I: because the higher frequencies are absent, the variance term in (9.40) is not so inflated in the $q = 8$ case. However, with moderate noise levels $\epsilon$, a test would not be able to discriminate between $\theta^I$ and $\theta^I$. This is an instance of the nearby instability of MSE as discussed in Section 9.4.3.

At least for $q > 2$, this is a case of over-use of fixed $\theta$ asymptotic approximation – both $h_v(\theta)$ and $r_v(\theta_h, \theta)$ are very sensitive to aspects of the function that are difficult or impossible to estimate at small sample sizes. Of course, the density estimation literature also cautions against the use of higher order ($q > 2$) kernels due to these poor finite sample properties. We did not even attempt to consider the behavior of “plug-in” methods that attempt to estimate $h_v(\theta)$: variability in the data based estimates of $h_v(\theta)$ would of course also contribute to the overall mean squared error. Loader (1999) provides a somewhat critical review of “plug-in” methods in the case $q = 2$.

However, our main point remains that fixed $\theta$ expansions such as (9.37) and (9.42) are potentially unstable tools. While the choice $q = 8$ may seem extreme in the setting of traditional density estimation, it is actually standard to use wavelets with higher order vanishing moments – for example, the Daubechies Symmlet 8 discussed in Daubechies (1992, p. 198-199) or Mallat (1999b, p. 252). Analogs of (9.37) and (9.42) for wavelet based density estimates appear in Hall & Patil (1993), though of course these authors do not use the expansions for bandwidth selection.

**Remarks.** 1. **Block James Stein estimation.** Figures 7 and ?? also show the upper bounds (9.26) for the MSE of the dyadic blocks James-Stein estimator, and it can be seen that its MSE performance is generally satisfactory, and close to the $q = 2$ kernel over small sample sizes. Figure 8 compares the ratio $r_e(\hat{\theta}^{BS}, \theta)/r_e(\hat{\theta}, \theta)$ of the Block JS mean squared error to the $q$-th order kernel MSE over a much larger range of $n = e^{-2}$. The James Stein MSE bound is never much worse than the MSE of the $q$-th order optimal bandwidth, and in many cases is much better.

2. **Smoothness assumptions.** Since $\theta^I$ and $\theta^{II}$ have finite Fourier expansions, they are certainly $C^\infty$, but for all practical purposes, they behave more like functions with about two square summable derivatives. Since $\theta^{II}$ for example, is technically $C^\infty$, one expects from the adaptivity Theorem 9.6 that Block JS should eventually improve on the $q = 4$ and $q = 8$ kernels, and this indeed occurs in Figure 8 on the right side of the plot. However, the
huge sample sizes suggest how impractical this “theoretical” high smooth-ness actually is. Such considerations point toward the need for quantitative measures of smoothness (Sobolev, Besov norms) rather than qualitative hypotheses such as the mere existence of derivatives.

3. Speed limits. There is a uniform version of (9.42) that says that over ellipsoids of functions with $\alpha$ mean-square derivatives, the uniform rate of convergence using the $q$-th order kernel is at best $(\epsilon^3)^{2q/(2q+1)}$, no matter how large $\alpha$ is. By contrast, the adaptivity results of Theorem 9.6 (and its extensions) for the block James-Stein estimate show that it suffers no such speed limit, and so might effectively be regarded as acting like an infinite order kernel. (Exercise 1 below has further details.)

Concluding discussion. Worst case analysis is, in a certain sense, the antithesis of fixed $\theta$ analysis. The least favorable configuration — whether parameter point $\theta_0$ or prior distribution $\pi_0$ — will generally change with noise level $\epsilon$. This is perfectly natural, since the such configurations are representative of the “limit of resolution” attainable, which improves as the noise diminishes.

The specification of the space $\Theta$ to be maximized over is certainly critical, and greatly affects the least favorable configurations found. This at least has the virtue of making the consequences of assumptions (far more potent in nonparametrics, even if hidden) clearer. It might be desirable to have some compromise inbetween the very local nature of fixed $\theta$ asymptotics, and the very global aspect of minimax analysis (perhaps in the spirit of the local asymptotically minimax approach used in parametric asymptotics.) Nevertheless, if one can construct estimators that deal successfully with many least favorable configurations from the global minimax framework – as in the blockwise James-Stein constructions – then one can have some degree of confidence in such estimators for practical use in settings not too distant from the assumptions.

9.5. Discussion

Discuss importance of changes in least favorable prior and estimator with sample size $\epsilon$ -- in connection with superefficiency.

Also comment on the value in doing the exact Pinsker result, even given the Ibragimov-Hasminskii constant.

Visualizing least favorable distributions. Pinsker’s theorem gives an explicit construction of the asymptotically least favorable distribution associated with the ellipsoid $\Theta = \{\theta : \sum a_i^2 \theta_i^2 \leq C^2 \}$: simply take independent variables $\theta_i \sim N(0, \tau_i^2)$, with $\tau_i$ given by (??). Recalling that the $\theta_i$ can be thought of as coefficients of the unknown function in an orthonormal basis $\{\varphi_i\}$ of $L^2[0,1]$, it is then instructive to plot sample paths from the random function

$$X(t) = \sum \theta_i \varphi_i(t).$$
Figure ??? shows two such sample paths, corresponding to smoothness \( m = 1 \) and \( m = 2 \) respectively (and with \( \epsilon = 2^{-6} \) and \( 2^{-6.5} \) respectively). In fact, the pictures were generated using a wavelet basis, and coefficient sequence \( \bar{a}_i = 2^{n[i \epsilon_2]} \), but since \( \bar{a}_i/a_i \in [2^{-m}, 1] \) relative to the trigonometric basis weight sequence (??), this has little influence on our qualitative conclusions – see Johnstone (1994) for details.

Notice the spatial homogeneity of the sample paths – even though the smoothness of the paths is, as expected, very different in the two cases, the degree of oscillation within each figure is essentially constant as one moves from left to right in the domain of the function.

**Challenges to the ellipsoid model.**

Of course, not all signals of scientific interest will necessarily have this spatial homogeneity:

*Ex: NMR spectrum of tryptophan in heavy water from DJHS*

*Ex: plethysmograph signal from Nason & Silverman*

In each case, there are regions of great “activity” or “oscillation” in the signal, and other regions of relative smoothness.

Thus, by comparing sample paths from the Gaussian priors with the data examples, one naturally suspects that the ellipsoid model is not relevant in these cases, and to ask whether linear estimators are likely to perform near optimally (and in fact, they don’t).

Another implicit challenge to the ellipsoid model and the fixed bandwidth smoothers implied by ?? and ?? begins to appear in the methodological and applied statistical literature at about the same time as Pinsker (1980). Cleveland (1979) investigates local smoothing, and Friedman & Stuetze (1981), in describing the univariate smoother they constructed for projection pursuit regression say explicitly “the actual bandwidth used for local averaging at a particular value of (the predictor) can be larger or smaller than the average bandwidth. Larger bandwidths are used in regions of high local variability of the response.”

**Commentary on the minimax approach.** One may think of minimax decision theory as a strategy for evaluating the consequences of assumptions - the sampling model, loss function, and particularly the structure of the postulated parameter space \( \Theta \). The results of a minimax solution consist, of course, of the minimax value, the minimax strategy, the least favorable prior, and also, information gained in the course of the analysis.

A particular feature of least favorable distributions is that they indicate the “typical enemy” corresponding to the parameter space \( \Theta \) chosen. The least favorable prior avoids both the arbitrariness of an effort to choose a single “representative” function, and yet focuses attention on elements of \( \Theta \) that are “relevant” to the minimax problem. [Of course, whether the “typical enemies” thus exhibited are scientifically relevant depends on the particular application.]
The minimax strategy is can be successful if the structure of $\Theta$ is intellecutally and/or scientifically significant, and if it is possible to get close enough to a solution of the resulting minimax problem that some significant and interpretable structure emerges.

Pinsker's theorem is an outstanding success for the approach, since it yields an (asymptotically) sharp solution, along with the important structure of linear estimators, independent Gaussian least favorable priors, decay of shrinkage weights with frequency to a finite cutoff, and so on.

The clarity of the solution, paradoxically, also reveals some limitations of the result, or rather, of the formulation. The juxtaposition of the Pinsker priors and particular datasets suggests that for some scientific problems, one needs richer models of parameter spaces than ellipsoids (and their quadratically convex relatives.) This is one motivation for the introduction of Besov and Triebel bodies in Chapter ??? below.

Just a test of todo in jamesstein.tex

Exercises

1. (Speed limits for $q$-th order kernels.)

We have argued that in the Gaussian sequence model in the Fourier basis, it is reasonable to think of a kernel estimate with bandwidth $h$ as represented by $\hat{\theta}_{h,t} = \kappa(hd)g_t$.

(a) Explain why it is reasonable to express the statement “$K$ is a $q$-th order kernel,” $q \in \mathbb{N}$, by the assumption $\kappa(s) = 1 - c_q s^q + o(s^q)$ as $s \to 0$ for some $c_q \neq 0$.

(b) Let $\Theta_\alpha(C) = \{\theta : \sum a_\alpha^2 \theta^2 \leq C^2\}$ with $a_{\alpha-1} = a_\alpha = (2l)^\alpha$ be, as usual, an ellipsoid of $\alpha$-mean square differentiable functions. If $K$ is a $q$-th order kernel in the sense of part (a), show that for each $\alpha > q$,

$$\inf_{h > 0} \sup_{\Theta \in \Theta_\alpha(C)} r_s(\hat{\theta}_h, \theta) \geq c(\alpha,q,C)(\mathcal{L}^{2q/(2q+1)}).$$

[Thus, for a second order kernel, the (uniform) rate of convergence is $n^{-1/5}$, even if we consider ellipsoids of functions with 10 or 10^6 derivatives. Since the (dyadic) block James Stein estimate has rate $n^{-2\alpha/(2\alpha+1)}$ over each $\Theta_\alpha(C)$, we might say that it corresponds to an infinite order kernel.]
Part 4

Interlude:
CHAPTER 10

A Primer on Estimation by Wavelet Shrinkage

[INTRODUCTION]

Our goal in this chapter is to give an account of some of the main issues and ideas behind wavelet thresholding as applied to equally spaced data. The purpose is both to give the flavor of how wavelet shrinkage can be used in practice, as well as provide the setting and motivation for theoretical developments in subsequent chapters.

10.1. The Cascade algorithm for the Discrete Wavelet Transform

[PREPARATORY TEXT!]

[Advantages of wavelet bases.]

A further key feature of wavelet bases is the availability of fast $O(N)$ algorithms for computing both the wavelet transform of discrete data and its inverse. This “cascade” algorithm is often derived, as we do below, by studying the structure of a multiresolution analysis of functions of a continuous real variable. In practice, it is used on finite data sequences, and the scaling function $\varphi$ and wavelet $\psi$ of the MRA are not used at all. This is fortunate, because the latter are typically only defined by limiting processes and so are hard to compute (compare ?? and ??). Thus there is a (most helpful) gap between the motivating mathematics and the actual data manipulations. Since our goal later is to give a theoretical account of the statistical properties of these data manipulations, our presentation here will try to be explicit about the manner in which discrete orthogonal wavelet coefficients in fact approximate their multiresolution relatives.

We use the results, notations and conventions developed in Appendix B.

Suppose that $\{h[k], k \in \mathbb{Z}\}$ is a filter generating a multiresolution analysis with scaling function $\varphi \in L_2(\mathbb{R})$, and that the corresponding filter

$$g[k] = (-1)^{1-k} h[1 - k]$$

generates a wavelet $\psi \in L_2(\mathbb{R})$ such that the scales and translates $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ for $j, k \in \mathbb{Z}$ form an orthonormal basis of square integrable functions $L_2(\mathbb{R})$. Typically we are interested in filters $h$ and $g$ which have only finitely many non-zero values, or ‘taps’, but this restriction is not needed for much of the development to follow.

Exploiting the Two Scale equations. Since $\varphi$ and $\psi$ are derived from a multiresolution analysis, they satisfy two-scale equations which we recall
here (cf. (B.4) and (B.8)): 
\[ \frac{1}{\sqrt{2}} \varphi(t/2) = \sum_{l} h[l] \varphi(t - l), \quad \frac{1}{\sqrt{2}} \psi(t/2) = \sum_{l} g[l] \varphi(t - l). \]

First we express \( \varphi_{j-1,k} \) and \( \psi_{j-1,k} \) in terms of \( \varphi_{jk} \) (which can be done because \( V_{j-1} \) and \( W_{j-1} \) are contained in \( V_{j} \)). Rescale by setting \( t = 2^j x - 2k \) and multiply both equations by \( 2^{j/2} \). Recalling the notation \( \varphi_{jk}(x) = 2^{j/2} \varphi(2^j x - k) \), we have 
\[ \varphi_{j-1,k}(x) = \sum_{l} h[l] \varphi_{j,k+l}(x) = \sum_{l} h[l - 2k] \varphi_{jl}(x). \] 
The corresponding relation for the coarse scale wavelet reads 
\[ \psi_{j-1,k}(x) = \sum_{l} g[l - 2k] \varphi_{jl}(x). \]

Since the collection \( \{ \varphi_{jl}, l \in \mathbb{Z} \} \) is an orthobasis for \( V_{j} \), we obtain 
\[ h[l - 2k] = \langle \varphi_{j-1,k}, \varphi_{jl} \rangle, \quad g[l - 2k] = \langle \psi_{j-1,k}, \varphi_{jl} \rangle \]

Conversely, we may express \( \varphi_{jl} \) in terms of \( \varphi_{j-1,k} \) and \( \psi_{j-1,k} \) since \( V_{j} = V_{j-1} \oplus W_{j-1} \). Thus 
\[ \varphi_{jl} = \sum_{k} \langle \phi_{jl}, \varphi_{j-1,k} \rangle \varphi_{j-1,k} + \langle \phi_{jl}, \psi_{j-1,k} \rangle \psi_{j-1,k} \]
\[ = \sum_{k} h[l - 2k] \varphi_{j-1,k} + g[l - 2k] \psi_{j-1,k}. \]

**Notation for operators on sequences.** In this section, we adopt a notation \( a[k] \) for components of sequences indexed by \( \mathbb{Z} \), or \( \mathbb{Z}_{\omega} \). This convention, often used in engineering texts, facilitates the description of filtering algorithms applied to discrete sequences. In addition, we list some notation for various simple filtering operations:

**Convolution:** 
\[ a \ast b[k] = \sum_{l} a[l] b[k - l], \]

**Reversal:** 
\[ R a[k] = a[-k], \]

**Downsampling:** 
\[ D a[k] = a[2k], \]

**Zero padding:** 
\[ Z a[2k] = a[k], \quad Z a[2k + 1] = 0. \]

For the Fourier transform of a zero padded sequence, we have 
\[ \tilde{Z} a(\omega) = \sum_{l} e^{-i \omega l} \tilde{Z} a[l] = \sum_{l} e^{-2i \omega k \omega} a[k] = \hat{a}(2\omega). \]

**The Building Blocks.** Suppose that \( f \in L_2(\mathbb{R}) \), and set 
\[ a_j[k] = \langle f, \varphi_{jk} \rangle, \quad d_j[k] = \langle f, \psi_{jk} \rangle. \]
Here \( a \) is mnemonic for the “approximation” and \( d \) for “detail” coefficients at level \( j \) in continuous time.

**Analysis Step:** By taking inner products with \( f \) in (10.1) and (10.2), we obtain

\[
\begin{align*}
a_{j-1}[k] &= \sum_l h[l - 2k]a_j[l] = Rh * a_j[2k] \\
d_{j-1}[k] &= \sum_l g[l - 2k]a_j[l] = Rg * a_j[2k]
\end{align*}
\]

In other words, “filter with \( Rh \) (resp. \( Rg \)), and then downsample”. The analysis step is sometimes referred to as the “decomposition” or “fine-to-coarse” step.

**Synthesis Step:** By taking inner products with \( f \) in (10.3), we obtain

\[
\begin{align*}
a_j[l] &= \sum_k h[l - 2k]a_{j-1}[k] + g[l - 2k]d_{j-1}[k] \\
&= h * Z a_{j-1}[l] + g * Z d_{j-1}[l].
\end{align*}
\]

In other words, “zero-pad, then filter with \( h \) (resp. \( g \)) and add”. The synthesis step is correspondingly described as the “reconstruction” or “coarse-to-fine” step.

The analysis map \( A_j : a_j \rightarrow (a_{j-1}, d_{j-1}) \) and its inverse, the synthesis map \( S_j : (a_{j-1}, d_{j-1}) \rightarrow a_j \) are orthogonal operators on \( \ell_2(\mathbb{Z}) \) since they express the change of basis between \( V_j = \text{span}\{\varphi_{jk}, k \in \mathbb{Z}\} \) and \( V_{j-1} \oplus W_{j-1} = \text{span}\{\varphi_{j-1,k}, \psi_{j-1,k}, k \in \mathbb{Z}\} \). Note also that if \( h \) has \( B \) non-zero coefficients, or “taps”, then each convolution requires only \( B \) multiplies and additions.

**The Cascade Algorithm**

[DIOGONAL ARROWS PICTURE HERE]

\[
(10.7) \quad a_j \longleftrightarrow \{d_{j-1}, d_{j-2}, \ldots, d_L, a_L\}
\]

The forward direction is the analysis operator, given by the orthogonal discrete wavelet transform \( W \). The reverse direction is the synthesis operator, given by its inverse, \( W^t \).

**10.1.1. Finite data sequences.** So far we have worked with infinite sequences \( a_j \) and \( d_j \in \ell_2(\mathbb{Z}) \). Consider for a moment the action of the transform and its inverse on a *finite* data sequence \( a_j \) of length \( N = 2^J \). It is now necessary to say how the boundaries of the data are treated. The transform \( W \) remains orthogonal so long as \( h \) is a filter generating an orthonormal wavelet basis, and either

(i) boundaries are treated periodically, or

(ii) we use boundary filters (such as those of Cohen et al. (1993a)) that preserve orthogonality.
In either case, the detail vectors $d_j$ in (10.7) are of length $2^j$, and the final approximation vector $a_L$ is of length $2^L$. The orthogonal transform is then “non-redundant”, as it takes $N = 2^j$ coefficients $a_j$ into $2^{j-1} + 2^{j-2} + \ldots + 2^L = N$ coefficients in the transform domain. If $h$ has $B$ non-zero coefficients, then the computational complexity of both $W$ and $W^t$ is of order $2B(2^{j-1} + 2^{j-2} + \ldots + 2^L) \leq 2BN = O(N)$.

$W$ maps a vector of data $y = (y_l, l = 1, \ldots, N)$ of length $N = 2^j$ into $N$ wavelet coefficients $w = Wy$. Identifying $y$ with $a_j$, we may identify $w$ with $\{d_{j-1}, d_{j-2}, \ldots, d_L, a_L\}$. More specifically, we index $w = (w_I)$ with $I = (j, k)$ and

\[
\begin{align*}
    w_{jk} &= d_{jk} & j = L, \ldots, J - 1 & \quad k = 1, \ldots, 2^j \\
    w_{L-1,k} &= a_{Lk} & k = 1, \ldots, 2^L.
\end{align*}
\]

With this notation, we may write $y = W^t w$ in the form

\[
y = \sum w_I \psi_I
\]

with $\psi_I$ denoting the columns of the inverse discrete wavelet transform matrix $W^t$.

Our goal now is to describe more explicitly how the vectors $\psi_I$ are related to the $L_2(\mathbb{R})$ wavelets $\psi_{jk}(x) = 2^{j/2} \phi(2^j - k)$. For simplicity, we ignore boundary effects and return to the setting of $\ell_2(\mathbb{Z})$.

10.1.2. Discrete and Continuous Wavelets. The discrete filtering operations of the cascade algorithm make no explicit use of the wavelet $\phi$ and scaling function $\varphi$. Yet they are derived from the multiresolution analysis generated by $(\varphi, \psi)$, and it is our goal in this subsection to show more explicitly how the orthonormal columns of the discrete wavelet transform are approximations to the orthonormal functions $\varphi_{jk}$ and $\psi_{jk}$.

Approximating $\varphi$ and $\psi$ from the filter cascade. So far, the cascade algorithm has been described implicitly, by iteration. We now seek a more explicit representation. Let $h^{(r)} = h \star Z \ast \cdots \ast Z^{r-1} h$ and $g^{(r)} = h^{(r-1)} \star Z^{r-1} g$.

10.1. Lemma.

\[
\begin{align*}
    a_{j-1}[k] &= \sum_{n} h^{(r)}[n - 2^r k] a_j[n] = R h^{(r)} \ast a_j[2^r k]. \\
    d_{j-1}[k] &= \sum_{n} g^{(r)}[n - 2^r k] a_j[n] = R g^{(r)} \ast a_j[2^r k].
\end{align*}
\]

This formula says that the $2^r$-fold downsampling can be done at the end of the calculation if appropriate infilling of zeros is done at each stage. While not necessarily sensible in computation, this is helpful in deriving a formula.
Figure 1. The function $\varphi(r)$ for the Daubechies D4 filter with $r = 4$.

Proof. [Also include proof for $d_{j-r}[k]$??] The argument, of course, uses induction. The case $r = 1$ is the analysis step (10.5). For general $r$, (10.5) gives

$$a_{j-r}[k] = Rh * a_{j-r+1}[2k],$$

and using the induction hypothesis for $r - 1$, we obtain

$$a_{j-r}[k] = \sum_l h[l - 2k] \sum_n h^{(r-1)}[n - 2^{r-1}l] a_j[n]$$

$$= \sum_n a_j[n] \sum_l h^{(r-1)}[n - 2^{r-1}l] h[l - 2k],$$

Now $h[l - 2k] = Z^{-1} h[2^{r-1}l - 2^r k]$ and since $Z^{-1} h[m] = 0$ unless $m = 2^{r-1}l$,

$$a_{j-r}[k] = \sum_n a_j[n] \sum_m h^{(r-1)}[n - m] Z^{-1} h[m - 2^r k].$$

The inner sum is $h^{(r-1)} * Z^{-1} h[n - 2^r k] = h^{(r)}[n - 2^r k]$, so we are done. □

Relating $h^{(r)}$ to $\varphi$. The Fourier transform of the cascadecom filter $h^{(r)}$ is $\hat{h}^{(r)}(\omega) = \prod_{j=0}^{r-1} \hat{h}(2^j \omega)$, as follows from (10.4). Substituting $\omega = 2^{-r} \xi$, we are led to define

$$(10.8) \quad \hat{\varphi}(r)(\xi) = 2^{-r/2} \hat{h}^{(r)}(2^{-r} \xi) = \prod_{j=1}^{r} \frac{\hat{h}(2^{-j} \xi)}{\sqrt{2}},$$

since $\hat{\varphi}(r)(\xi) \rightarrow \hat{\varphi}(\xi)$ as $r \rightarrow \infty$. Since $\hat{\varphi}(r)(\xi)$ has period $2^{r+1} \pi$, we think of $\phi(r)$ as a function on $2^{-r} \mathbb{Z}$, given by

$$(10.9) \quad 2^{-r} \varphi^{(r)}(2^{-r} x) = 2^{-r/2} \hat{h}^{(r)}[n].$$

See Figure 1 for an example. Since $\varphi(r)$ is a function on $2^{-r} \mathbb{Z}$, we may then define scaled and translated versions $\varphi^{(r)}_{jk}$ on $2^{-(r+j)} \mathbb{Z}$ by

$$\varphi^{(r)}_{jk}(x) = 2^{j/2} \varphi^{(r)}(2^{j} x - k).$$

Remark. Identification (10.9) may be understood via the theory of distributions. Let $\delta_t$ denote the Dirac delta function located at $t$ : if $\chi$ is a smooth test function, then $\delta_t(\chi) = \chi(t)$. A function $g^{(r)} : 2^{-r} \mathbb{Z} \rightarrow \mathbb{R}$ corresponds to the distribution $\sum_k 2^{-r} g^{(r)}(2^{-r} k) \delta_{2^{-r} k}$. Indeed, think of $g^{(r)}$ as interpolated in some manner from $2^{-r} \mathbb{Z}$ to $\mathbb{R}$, then $g^{(r)}(\chi) \approx \int g^{(r)} \chi$ by Riemann sum approximation. The Fourier transform of the delta function located at $t$ is $\mathcal{F}(\delta_t)(\xi) = e^{-it \xi}$. With the identification

$$2^{-r} g^{(r)}(2^{-r} k) = g[k],$$

the Fourier transform of the distribution $g^{(r)}$ becomes

$$\hat{g}^{(r)}(\xi) = \sum g[k] e^{-ik2^{-r} \xi} = \hat{g}(2^{-r} \xi).$$

132
Continuous world \hspace{2cm} Discrete World

\[ a_J[k] = \langle \varphi_{jk}, f \rangle \hspace{2cm} a_J[k] = N^{-1/2} f(nN^{-1}) \]

\[ \downarrow \hspace{2cm} \downarrow \]

\[ a_J[k] = \langle \varphi_{jk}, f \rangle \hspace{2cm} a_J[k] = \langle \varphi_{J-j}^{(J-j)}, f \rangle_N \]

Table 1. Schematic comparing the orthogonal wavelet transform of functions \( f \in L_2(\mathbb{R}) \) with the discrete orthogonal wavelet transform of square summable sequences formed by sampling such functions on a lattice with spacing \( N^{-1} \).

The vertical arrows represent the outcome of \( r = J - j \) iterations of the cascade algorithm in each case.

Interpreting (10.8) in this light yields the identification (10.9).

Suppose that \( f \) is a square integrable function on \( 2^{-J} \mathbb{Z} = N^{-1} \mathbb{Z} \). The columns of the discrete wavelet transform will be orthogonal with respect to the inner product

\[ \langle f, g \rangle_N = N^{-1} \sum_{n \in \mathbb{Z}} f(N^{-1} n) g(N^{-1} n) . \tag{10.10} \]

Now start the cascade algorithm at level \( J \) with a sampled version of \( f \):

\[ a_J[n] = N^{-1/2} f(N^{-1} n), \quad n \in \mathbb{Z} \]

10.2. **Proposition.** If \( a_J[n] = N^{-1/2} f(N^{-1} n) \), and \( N = 2^J \), then for \( j \leq J \),

\[ a_J[k] = \langle \varphi_{jk}^{(J-j)}, f \rangle_N, \quad d_J[k] = \langle \psi_{jk}^{(J-j)}, f \rangle_N, \quad k \in \mathbb{Z}. \tag{10.11} \]

Formulas (10.11) are an explicit representation of our earlier description that the sequences \( \{a_J[k], k \in \mathbb{Z}\} \) and \( \{d_J[k], k \in \mathbb{Z}\} \) are found from \( \{a_J[k], k \in \mathbb{Z}\} \) by repeated filtering and downsampling. Formulas (10.11) suggest, without complete proof, that the iteration of this process is stable, in the sense that as \( J - j \) increases (the number of levels of cascade between the data level \( J \) and the coefficient level \( j \)), the coefficients look progressively more like the continuous-time coefficients \( \langle \varphi_{jk}, f \rangle \).

**Proof.** Let \( r = J - j \), so that \( a_j = a_{J-r} \) and, using Lemma 10.1, \( a_J[k] = \sum_n h^{(r)}[n - 2^r k] a_J[n] \). From (10.9),

\[ h^{(r)}[n - 2^r k] = 2^{-r/2} \varphi^{(r)}(2^{-r} n - k) = N^{-1/2} \varphi^{(r)}_{jk}(N^{-1} n), \]

which implies that \( a_J[k] = N^{-1} \sum_n \varphi^{(r)}_{jk}(N^{-1} n) f(N^{-1} n) = \langle \varphi^{(J-j)}_{jk}, f \rangle_N \).

The argument for \( d_J[k] \) is exactly analogous. \( \square \)
10.2. Wavelet shrinkage estimation

Basic model. Observations are taken at equally spaced points \( t_i = i/n, \)
\( i = 1, \ldots, n = 2^J, \) and are assumed to satisfy
\[
y_i = f(t_i) + \sigma z_i, \quad z_i \overset{i.i.d.}{\sim} N(0,1).
\]
It is assumed, for now, that \( \sigma \) is known. The goal is to estimate \( f \), at least
at the observation points \( t_i \).

The assumption that the observation points are equally spaced is quite
important (but see Section ??? below), whereas the specific form of the error
model and knowledge of \( \sigma \) are less crucial.

[Remarks about wavelet shrinkage as an exploratory method?]

Basic strategy. The outline is simply described. First, the transform step,
which uses an orthogonal wavelet transform \( W \). Second, a processing step in
the wavelet domain, and finally an inverse transform, which is accomplished
by \( W^t \), since \( W \) is orthogonal.

\[
\begin{array}{c}
(y_i) \\
\downarrow W \\
(\hat{w}_j) \\
\end{array}
\]

\[
(f(t_i)) \quad \iff \quad (\hat{w}_j)
\]

Transform step. Being an orthogonal transform, \( W \) is non-redundant,
and given \( n = 2^J \) data \( (y_i) \) in the “time” domain, produces \( n \) transform
coefficients in the wavelet domain, by use of the cascade algorithm derived
from a filter \( h \), as described at ????. We assume that either (i) the filters
are applied to the periodized data, or, more commonly in practice, that (ii)
boundary adjustments to the filters are made to preserve the orthogonality
and vanishing moments properties of the transform, such as those of Cohen
et al. [CDVJ].

We use \( f \) as an abstract index to denote the pair \((j, k)\), and \( w_j \) to denote
the collection \( \{w_{jk}, k = 1, \ldots, 2^j\} \) of coefficients at level \( j \).

The algorithm requires specification of a finest scale \( J = \log_2 n \) and a
coarse scale \( L \), a number between 0 and \( J - 1 \). Referring to the diagram
??? of the cascade, we see that \( w_j \) corresponds to the detail coefficients \( d_j \)
for \( j = J - 1, J - 2, \ldots, L \) and we adopt the slightly anomalous convention
that \( w_{L-1} = a_L \) represents the \( 2^L \)-coarse scale approximation coefficients.

If the filter \( h \) has \( B \) non-zero coefficients, then the transform requires
\( O(BN) \) steps.

The choice of filter \( h \) depends on a number of factors that influence the
properties of the resulting wavelet, such as support length, symmetry, and
number of vanishing moments (both for the wavelet and the scaling function).
The tradeoffs between these criteria are discussed by Mallat (1999,
Chapter 7). Common choices in WaveLab include (boundary adjusted) versions of \( D4 \) or the symmlet \( S8 \).
Figure 2. Graphs of soft and hard thresholding functions

Processing Step. Generally the estimated coefficients \( \hat{w} = \eta(w) \) are found by the following recipe

\[
\hat{w}_I = \begin{cases} 
\eta(w_I; t) & I \in \mathcal{D} \\
w_I & I \in \mathcal{A} 
\end{cases}
\]

Here \( \eta(w_I; t) \) is a scalar function of the observed coefficient \( w_I \), usually non-linear and depending on a parameter \( t \). Often, this parameter \( t \) is estimated, usually from all or some of the data at the same level as \( I \), yielding the modified expression \( \eta(w_I; t(w_j)) \), where \( I \in \mathcal{I}_j \). In some cases, the function \( \eta \) itself may depend on the coefficient index \( I \) or level \( j \). Common examples include (compare Figure 2)

a) hard thresholding

\[ \eta_H(w_I; t) = w_I I\{|w_I| \geq t\}, \]

b) soft thresholding

\[ \eta_S(w_I; t) = \begin{cases} 
    w_I - t & w_I > t \\
    0 & |w_I| \leq t \\
    w_I + t & w_I < -t.
\end{cases} \]

Other choices for \( \eta \), and methods for estimating \( t \) from data will be discussed later ???. For now, we simply remark that James-Stein shrinkage, though not a coordinate-wise thresholding method, also falls naturally into this framework:

\[ \eta_{JS}(w_I; s(w_j)) = s(w_j)w_I, \]

\[ s(w_j) = (1 - (n_j - 2)\sigma^2/|w_j|^2)^+. \]

The estimator. Writing \( \hat{f} \) for \( (\hat{f}(t_i)) \), we may summarize the estimation process as

\[ \hat{f} = W'\eta(Wy). \]

This representation makes the important point that the scaling and wavelet functions \( \varphi \) and \( \psi \) are not required or used in the calculation. So long as the filter \( h \) is of finite length, and the wavelet coefficient processing \( w \rightarrow \hat{w} \) is \( O(N) \), then so is the whole calculation.

Nevertheless, the iteration that occurs within the cascade algorithm generates approximations to the wavelet, as we have seen at ???. Thus, we may write the estimator more explicitly as

\[
\hat{f}(t_i) = \sum_I \eta_I(w)\psi_I(t_i), \\
= \sum_{I \in \mathcal{A}} w_I \varphi_I(t_i) + \sum_{I \in \mathcal{D}} \eta(w_I)\psi_I(t_i),
\]

(10.14)
Thus, $\psi_I = \psi_{jk}^{(J-j)}$ here is not the continuous time wavelet $\psi_{jk} = 2^{j/2}\psi(2^j \cdot -k)$, but rather the $(J-j)^{th}$ iterate of the cascade, defined at $??$, after being scaled and located to match $\psi_{jk}$.

The $(I,l)^{th}$ entry in the discrete wavelet transform matrix $W$ is given by $\psi_{jk}^{(J-j)}(N^{-l}l)$ and in terms of the columns $\psi_I$ of $W$, we have

$$y_l = \sum_I w_I^l \psi_I(N^{-l}l).$$

Examples: NMR

4 functions: Hard and Soft Thresholding.

Compare linear

Also need a figure with HARD thresholding at $\sqrt{2 \log n}$ Look for emfig5.eps on rgmiller

The thresholding estimates are

- simple, based on co-ordinatewise operations
- non-linear, and yeat
- fast to compute ($O(n)$ time).

The appearance of the estimates constructed with the $\sqrt{2 \log n}$ thresholds is

- noise free, with
- no peak broadening, and thus showing
- spatial adaptivity,

in the sense that more averaging is done in regions of low variability. Comparison with Figures $??$ shows that linear methods fail to exhibit these properties.

The hidden sparsity heuristic. A rough explanation for the success of thresholding goes as follows. The model (10.12) is converted by the orthogonal wavelet transform into

$$w_I = \theta_I + \epsilon \tilde{z}_I, \quad \epsilon = \sigma/\sqrt{n}, \quad \tilde{z}_I \sim \mathcal{N}(0,1).$$

Since the noise is white (i.e. independent with constant variance) in the time domain, and the wavelet transform is orthogonal, the same property holds for the noise variables $\tilde{z}_I$ in the wavelet domain - they each contribute noise at level $\tilde{c}^2$. On the other hand, in our sample signals, and more generally, it is often the case that the signal in the wavelet domain is sparse, i.e. its energy is largely concentrated in a few components. With concentrated signal and dispersed noise, a threshold strategy is both natural and effective, as we have seen in examples, and will see from a theoretical perspective in Chapter $??$. The sparsity of the wavelet representation may be said to be hidden, since it is not immediately apparent from the form of the signal in the time domain.

Estimation of $\sigma$. Assume that the signal is sparsely represented, and so most, if not all, data coefficients at the finest level are essentially pure noise. Since there are many $(2^l-1)$ such coefficients, one can estimate $\sigma^2$ well using a robust estimator

$$\hat{\sigma}^2 = \text{MAD}\{w_{J-1,k}, k \in I_{J-1}\}/0.6745,$$
which is not affected by the few coefficients which may contain large signal. Here $MAD$ denotes the median absolute deviation (from zero). The factor $0.6745$ is the population $MAD$ of the standard normal distribution, and is used to calibrate the estimate.

**Soft vs. Hard thresholding**

**Correlated data.**

**Wavelet shrinkage as a spatially adaptive kernel method.** Using the discrete inner product notation of (10.10), the empirical wavelet coefficients become

$$w_I = \langle y / \sqrt{N}, \psi_I \rangle_N = \sum_{l'} y_{l'} \psi_I(N^{-1}l')N^{-1},$$

where $\psi_I$ is an abbreviation for $\varphi^{(j-i)}_{jk}$ when $I \in \mathcal{D}$ and for $\varphi^{(j-L)}_{Ik}$ when $I \in \mathcal{A}$. Writing $\eta_I(w)$ in the form $c_I(y)w_I$, we may recast the expression (10.14) as

$$\hat{f}(t) = \sum_I c_I(y) \sum_{l'} y_{l'} \psi_I(N^{-1}l') \psi_I(N^{-1}l) N^{-1} = N^{-1} \sum_{l'} \hat{K}(t, t_{l'}) y_{l'},$$

where

$$\hat{K}(t, t_m) = \sum_I c_I(y) \psi_I(t) \psi_I(t_m).$$

The hat in this kernel emphasises that it depends on the data through the coefficients $c_I(y)$. The individual component kernels $K_I(t, s) = \psi_I(t) \psi_I(s)$ have bandwidth $2^{-j}B$ where $B$ is the support length of the filter $h$. Hence, one may say that the bandwidth of $\hat{K}$ at $t_i$ is of order $2^{-j(t_i)}$, where

$$j(t_i) = \max\{j : c_I(y) \psi_I(t_i) \neq 0, \text{ some } I \in \mathcal{I}_j\}.$$ 

In other words, $t_i$ must lie within the support of a level $j$ wavelet for which the corresponding data coefficient is not thresholded to zero. Alternatively, if a fine scale coefficient estimate $\hat{w}_{jk} \neq 0$, then there is a narrow effective bandwidth near $2^{-j}k$. [INCLUDE FIGURE] By separating the terms in (10.15) corresponding to the approximation set $\mathcal{A}$ and the detail set $\mathcal{D}$, we may decompose

$$\hat{K} = K_A + \hat{K}_D$$

where the approximation kernel $K_A(t_i, t_m) = \sum_k \varphi_I(t_i) \varphi_I(t_m)$ does not depend on the observed data $y$.

**Exercise.** With $W$ the $N \times N$ discrete wavelet transform matrix, let $C = \text{diag}(c_I)$ be a diagonal matrix with entries $c_I$ defined as above and let $\delta_t \in \mathbb{R}^N$ have zero entries except for a 1 in the $l$-th place. Show that the adaptive kernel at $t_i$, namely the vector $\tilde{K}_t = \{\tilde{K}(t_i, t_m)\}_{m=1}^N$, may be calculated using the wavelet transform via $\tilde{K}_t = W^T C W \delta_t$.

This is another test of todo
Translation invariant versions. The discrete wavelet transform (DWT) is not shift invariant: the transform of a shifted signal is not the same as a shift of the transformed original. This arises because of the dyadic downsampling between levels that makes the DWT non-redundant. For example, the Haar transform of a step function with jump at 1/2 has only one non-zero coefficient, whereas if the step is shifted to say, 1/3, then there are \(\log_2 N\) non-zero coefficients.

This is yet another test of todo

The transform, and the resulting threshold estimates, can be made invariant to shifts by multiples of \(N^{-1}\) by the simple device of averaging. Let \(S\) denote the operation of circular shifting by \(N^{-1}\): \(Sf(k/N) = f((k+1)/N)\), except for the endpoint which is wrapped around: \(Sf(1) = f(1/N)\). Define

\[
\hat{f}' = Aw_{1\leq k \leq N} (S^{-k} \circ \hat{f} \circ S^k).
\]

The translation invariant (TI) estimator averages over all \(N\) shifts, and so appears to involve at least \(O(N^2)\) calculation. However, the proposers of this method, Coifman & Donoho (1995), describe how the algorithm can in fact be implemented in \(O(N \log N)\) operations.

It can be seen from Figure 8 that the extra averaging implicit in \(\hat{f}'\) reduces artefacts considerably. Experience in practice has generally been that translation invariant averaging improves the performance of virtually every method of thresholding, and its use is encouraged in situations where the \(\log N\) computational penalty is not serious.

[COMMENT ON THEORY?]

10.3. Choice of threshold.

We give only a partial discussion of this large topic here, and choose only among methods that have some theoretical support.

The key features of a threshold method are firstly, the existence of a threshold zone \([-t, t]\) in which all observed data is set to zero. This allows the estimator to exploit sparse signal representations by ensuring that the mean squared error is very small in the majority of co-ordinates in which the true signal is negligible.

Secondly the tail behavior of the estimate as \(|x| \to \infty\) is also significant. More specifically, the growth of \(x - \eta(x)\) – approaching zero or a constant or diverging – influences the bias properties of the estimate, particularly for large signal components.

Previous experience and subjective experimentation can play important roles in threshold selection. ‘Automatic’ thresholding methods can be broadly divided into fixed versus data-dependent.

1. ‘Universal’ threshold \(\lambda_n = \sqrt{2 \log n}\). This is a fixed threshold method, and can be used with either soft or hard thresholding. If \(Z_1, \ldots, Z_n\) are i.i.d. \(N(0,1)\) variates, then it can be shown that

\[
P_n = P\{\max_{1\leq i \leq n} |Z_i| > \sqrt{2 \log n}\} \leq 1/\sqrt{4\pi \log n} \to 0.
\]
(Thus, for a wide range of values of \( n \), including \( 64 = 2^6 \leq n \leq 2^{10} \), the expected number of \( |Z| \) that exceed the threshold will be between 0.075 and 0.125, so only in about a tenth of realizations will any pure noise variables exceed the threshold.

Since the wavelet transform is orthogonal,

\[
P(\hat{f}_n = 0 | f = 0) = 1 - P_n \rightarrow 1.
\]

Thus, with high probability, no “spurious structure” is declared, and in this sense, the universal threshold leads to a “noise free” reconstruction. [Note however that this does not mean that \( \hat{f} = f \) with high probability when \( f \neq 0 \), since \( \hat{f} \) is not linear in \( y \).]

The price for this admirably conservative performance is that the method chooses large thresholds, which can lead to noticeable bias at certain signal strengths. This shows up in the theory as extra logarithmic terms in the rate of convergence of this estimator. **REF to Ch??**

2. **False discovery rate (FDR) thresholding.** This is a data dependent method for hard thresholding that is typically applied levelwise in the wavelet transform. Suppose that \( y_i \sim N(\theta_i, \sigma^2) \) are independent, and form the order statistics of the magnitudes:

\[
|y|_{(1)} \geq |y|_{(2)} \geq \ldots \geq |y|_{(n)}.
\]

Fix the false discovery rate parameter \( q \in (0, 1/2] \). Form quantiles \( t_k = \sigma \tilde{z}(q/2 \cdot k/n) \). Let \( \tilde{k} = \max\{k : |y|_{(k)} \geq t_k\} \), and set \( \tilde{t}_F = \tilde{k}_F \) and use this as the hard threshold

\[
\hat{\theta}_k(y) = y_k I\{|y|_{(k)} \geq \tilde{t}_F\}.
\]

[SHOW FIGURE 3 from FDR paper]. The boundary sequence \( (t_k) \) may be thought of as a sequence of thresholds for \( t\)-statistics in model selection: the more variables (i.e. coefficients in our setting) enter, the easier it is for still more to be accepted (i.e. pass the threshold unscathed.)

As is shown in Abramovich et al. (2000), the FDR estimator has excellent mean squared error performance in sparse multinormal mean situations - for example being asymptotically adaptive minimax over \( \ell_p \) balls. In addition (unpublished), it achieves the “right” rates of convergence over Besov function classes - thus removing the logarithmic terms present when the \( \sqrt{2 \log n} \) threshold is used.

However, the choice of \( q \) is an issue requiring further study - the smaller the value of \( q \), the larger the thresholds, and the more conservative the threshold behavior becomes.

3. **Stein’s unbiased risk estimate (SURE) thresholding.** This is a data dependent method for use with soft thresholding, again typically level by level. It has the special feature of allowing for certain kinds of correlation in the noise. Thus, assume that \( y \sim N_n(\theta, V) \), and assume that the diagonal elements \( \sigma_{kk} \) of the covariance matrix are constant and equal
to $\sigma^2$. This situation arises, for example, if in the wavelet domain, $k \rightarrow y_{jk}$ is a stationary process.

Now apply Stein’s unbiased risk estimate, REF??, to $\hat{\theta}(x) = x + g(x)$, with

$$g_k(x) = \begin{cases} -t & x_k > t \\ -x_k & |x_k| \leq t \\ t & x_k < -t. \end{cases}$$

In this case, $E_\theta \|\hat{\theta} - \theta\|^2 = E_\theta \hat{U}(t)$, where

$$\hat{U}(t) = \sigma^2 n + \sum_k y_k^2 \wedge t^2 - 2 \sum_k I\{|y_k| \leq t\}.$$

Now set

$$\hat{t}_{SURE} = \arg\min_{0 \leq t \leq \sigma \sqrt{2 \log n}} \hat{U}(t).$$

The criterion $\hat{U}(t)$ does not depend on details of the correlation $(\sigma_{jk}, j \neq k)$ and so can be used in correlated data settings when the correlation structure is unknown, without the need of estimating it.

The minimization can be carried out in $O(n \log n)$ time.

The SURE estimate also removes logarithmic terms in the rates of convergence of wavelet shrinkage estimates over Besov classes (though a ‘pretest’ is needed in certain cases to complete the proofs).

4. **Empirical Bayes.** This data dependent method for levelwise thresholding provides a family of variants on soft and hard thresholding. Again assume an independent normal means model, $y_i = \theta_i + \sigma z_i$, with $z_i$ i.i.d standard normal. Allow $\theta_i$ to independently be drawn from a mixture prior distribution

$$\theta_i \sim (1 - w)\delta_0 + w\gamma_\alpha.$$

Here $w$ is the probability that $\theta_i$ is non-zero, and $\gamma_\alpha(d\theta)$ is a family of distributions with scale parameter $\alpha > 0$, for example the double exponential

$$\gamma_\alpha(d\theta) = (\alpha/2)e^{-|\theta|/\alpha}d\theta.$$

Using $L_1$ loss $\|\hat{\theta} - \theta\|_1 = \sum_i |\hat{\theta}_i - \theta_i|$, the Bayes rule for this prior is the median $\hat{\theta}_{EB}(y)$ of the posterior distribution of $\theta$ given $y$:

$$\hat{\theta}_{EB}(y) = \eta(y; w, \alpha).$$

The reason for using the posterior median is that it turns out that $\eta$ has threshold structure:

$$\eta(y; w, \alpha) = 0 \quad \text{if} \quad |y| \leq \sigma t(w, \alpha),$$

while for large $|y|$, it turns out that $|y - \eta(y)| \sim \sigma a$.

The hyperparameters $(w, \alpha)$ can be estimated by maximizing the marginal likelihood of $(w, \alpha)$ given data $(y_i)$. Theory shows that the method achieves the optimal rates of convergence, while simulations suggest that
the method adapts gracefully to differing levels of sparsity at different resolution levels in the wavelet transform.

5. Block Thresholding?

10.4. Concluding remarks

A fourth and final test

Unequally spaced data?

More remarks on $L_1$ loss leading to posterior median.

Include Eisenberg example?

Topics not covered here: Extensions to other data formats: time series spectral density estimation, count data and Poisson estimation.
Figure 3. From toons0551.m

Figure 4. From toons0552.m
FIGURE 5. From toons0553.m

FIGURE 6. From toons0554.m
Figure 7. Produces figure showing spatially adaptive kernel features of hard thresholding as applied to the RaphaelNMR signal.

Figure 8. Figure to show SoftHaarTI-Blocks and HardHaarTI-Blocks with reference to CoDo paper for other test functions.
Part 5

MONORESOLUTION MODELS
CHAPTER 11

Thresholding and Oracle inequalities

Thresholding is very common, even if much of the time it is conducted informally, or perhaps most often, unconsciously. Most empirical data analyses involve, at the exploration stage, some sort of search for large regression coefficients, correlations or variances, with only those that appear “large”, or “interesting” being retained for reporting purposes, or in order to guide further analysis.

For all its ubiquity, thresholding has received much less theoretical attention than linear estimation methods, such as those we have considered until now. This is perhaps due, in part, to the non-linearity that is inherent to thresholding: a scaled up version of the data does not always yield a proportionately scaled-up version of the estimate, since the very act of scaling up the data may put it over the retention threshold.

Consequently, the bias-variance decomposition cannot be used as directly for threshold estimators as for linear ones: one needs other features of the distribution of the data beyond first and second moments. The main concern of this chapter will therefore be to develop tools for analysing and understanding the mean squared error of soft and hard thresholding and its dependence on both the unknown mean and the threshold level.

11.1. A crude MSE bound for hard thresholding.

Consider a single observation \( y \sim N(\theta, \epsilon^2) \). The thresholding estimator may be written as \( \hat{\theta}(y) = yI_E \) where \( E \) is the event \( \{|y| > \lambda \epsilon\} \) on which \( y \) exceeds the threshold and is retained.

Denote the mean squared error of \( \hat{\theta} \) by \( r_H(\lambda, \theta) = E_{\theta}[|yI_E - \theta|^2] \). We construct two bounds for the mean squared error, according as the signal \( \theta \) is smaller than the noise \( \epsilon \) or not. It will be seen that this has the character of a bias or variance decomposition – since such a thing is of course not really possible, we are forced to accept extra terms, either additive or multiplicative, in the analogs of bias and variance.

11.1. PROPOSITION. If \( y \sim N(\theta, \epsilon^2) \), there exists a constant \( M \) such that if \( \lambda > 4 \)

\[
r_H(\lambda, \theta) \leq \begin{cases} 
M[\theta^2 + \lambda \phi(\lambda - 1) \epsilon^2] & \text{if } |\theta| \leq \epsilon \\
M \lambda^2 \epsilon^2 & \text{if } |\theta| > \epsilon.
\end{cases}
\]
PROOF. Consider first the small signal case $|\theta| < \epsilon$. Arguing cruelly,

$$E_0[yI_E - \theta]^2 \leq 2E_0y^2I_E + 2\theta^2$$

The first term is largest when $|\theta| = \epsilon$. In this case, if we set $x = y/\epsilon \sim N(1, 1)$ and suppose that $\lambda > 4$, then

$$E_0y^2I_E \leq \epsilon^2 \cdot 2 \int_{\lambda}^{\infty} x^2 \varphi(x - 1)dx \leq 4\lambda \varphi(\lambda - 1)\epsilon^2;$$

where we used the fact that for $x \geq 3$, $(x + 1)^2\varphi(x) \leq 2(x^2 - 1)\varphi(x) = 2(d/dx)[-x\varphi(x)]$.

In the large signal case, $|\theta| > \epsilon$, we use the relation $y = \theta + \epsilon z$ to analyse by cases, obtaining

$$yI_E - \theta = \begin{cases} 
\epsilon z & \text{if } |y| > \lambda \epsilon, \\
\epsilon z - y & \text{if } |y| \leq \lambda \epsilon,
\end{cases}$$

so that in either case

$$(yI_E - \theta)^2 \leq 2\epsilon^2(\epsilon^2 + \lambda^2).$$

Taking expectations gives the result. We have de-emphasised the explicit constants (which will be improved later anyway in Lemma 11.5) to emphasise the structure of the bound, which is the most important point here. □

From the proof, one sees that when the signal is small, the threshold produces zero most of the time and the MSE is essentially the resulting bias plus a term for ‘rare’ errors which push the data beyond the threshold. When the signal is large, the data is left alone, and hence has variance of order $\epsilon$, except that errors of order $\lambda \epsilon$ are produced about half the time when $\theta = \lambda \epsilon$!

EXAMPLE 14.3 continued. Let us apply inequality (11.1) to $\Theta_{n,1} = \{\theta : \sum |\theta_i| \leq 1\}$. Let $B_n$ be the set of “big” coordinates $|\theta_i| \geq \epsilon = n^{-1/2}$, and $S_n = B_n^c$. Clearly, when $\theta \in \Theta_{n,1}$, the number of big coordinates is relatively limited: $|B_n| \leq n^{1/2}$. For the ‘small’ coordinates, $\theta_i^2 \leq n^{-1/2}|\theta_i|$, so $\sum S_n \theta_i^2 \leq n^{-1/2}$. Now using (11.1)

$$\sum r_H(\lambda, \theta_i) \leq M \sum _{B_n} \lambda^2 \epsilon^2 + M \sum _{S_n} [\theta_i^2 + \lambda \phi(\lambda - 1)\epsilon^2] \leq M \lambda^2 n^{-1/2} + M[n^{-1/2} + \lambda \phi(\lambda - 1)].$$

Choosing, for now, $\lambda = 1 + \sqrt{\log n}$, so that $\phi(\lambda - 1) = \phi(0)n^{-1/2}$, we finally arrive at

$$E\|\hat{\theta}_\lambda - \theta\|^2 \leq M \log n / \sqrt{n}.$$

While this argument does not give exactly the right rate of convergence, which is $(\log n / n)^{1/2}$, let alone the correct constant, it already shows clearly that thresholding is much superior to linear estimation on the $\ell_1$ ball.
11.2. Properties of Thresholding Estimators

In this section we consider two types of thresholding estimators \( \hat{\delta}(x) \) in the simplest univariate case: \( x \sim N(\mu, 1) \). Note the special notation that we henceforth use for the standard noise level \( \epsilon = 1 \).

**Hard Thresholding.**

\[
\hat{\delta}_H(x, \lambda) = \begin{cases} 
  x & |x| > \lambda \\
  0 & |x| \leq \lambda 
\end{cases}
\]

(11.3)

**Soft Thresholding.**

\[
\hat{\delta}_S(x, \lambda) = \begin{cases} 
  x - \lambda & x > \lambda \\
  0 & |x| \leq \lambda \\
  x + \lambda & x < -\lambda 
\end{cases}
\]

(11.4)

*Similarities.* These two estimators are both non-linear, and in particular have in common the notion of a threshold region \( |x| \leq \lambda \) in which no signal is estimated. Of course, hard thresholding is discontinuous, while soft thresholding is constructed to be continuous, which explains the names. The threshold parameter in principle can vary over the entire range \([0, \infty)\), so the family includes the special linear estimators \( \hat{\delta}(x, 0) = x \) and \( \hat{\delta}(x, \infty) = 0 \) that “keep” and “kill” the data respectively. In general, however, we will be interested in thresholds in the range between about 1.5 and a value proportional to the square root of log-sample-size. We now make some comments specific to each class.

*Differences.* Hard thresholding preserves the data outside the threshold zone - this can be important in certain applications - for example in denoising where it is desired to preserve as much as possible the heights of true peaks in estimated spectra. The mathematical consequence of the discontinuity is that the risk properties of hard thresholding are a little more awkward - for example the mean squared error is not monotonic increasing in \( \mu \geq 0 \). Hard thresholding also has the interesting property that it arises as the solution of a penalized least squares problem

\[
\hat{\delta}_H(x, \lambda) = \arg \min_{\mu} (x - \mu)^2 + \lambda^2 I[\mu \neq 0].
\]

Indeed, when \( \mu \neq 0 \) the criterion has minimum value \( \lambda^2 \) when \( \mu = x \) and when \( \mu \) vanishes, the criterion equals \( x^2 \). Hard thresholding amounts to choosing the better of these two values.

Soft thresholding, on the other hand, shrinks the data towards 0 outside the threshold zone. The mean squared error function is now monotone in \( \mu \geq 0 \), and we will see later that the shrinkage aspect leads to significant smoothing properties in function estimation (REF) - in practice, however, neither soft nor hard thresholding is universally preferable - the particular features of the application play an important role. The estimator that we call soft thresholding has appeared frequently in the statistics literature, for
example Efron & Morris (1971), who term it a “limited-translation” rule. Soft thresholding also arises from a penalized least squares problem

\[ \hat{\delta}_S(x, \lambda) = \arg \min_\mu (x - \mu)^2 + 2\lambda |\mu|, \]

as may be verified directly.

Notice that in the case of \( n \)-dimensional data, the same calculation can be conducted co-ordinatewise:

\[ \hat{\delta}_S(x, \lambda) = \arg \min_{\mu \in \mathbb{R}^n} \sum_i (x_i - \mu_i)^2 + 2\lambda \sum_i |\mu_i|, \]

Since the penalty term is an \( \ell_1 \) norm of \( \mu \), soft thresholding is sometimes also called the \( \ell_1 \) rule. In the same vein, the corresponding penalty for hard thresholding in \( \mathbb{R}^n \) is \( \sum_i I\{\mu_i \neq 0\} \), which with slight abuse of notation might be called an \( \ell_0 \) penalty.

Compromises. Many compromises between soft and hard thresholding are possible that appear in principle to offer many of the advantages of both methods: a threshold region for small \( x \) and exact or near fidelity to the data when \( x \) is large.

1) soft-hard thresholding (Gao & Bruce 1997): This is a compromise between soft and hard thresholding defined by

\[ \hat{\delta}_{\lambda_1, \lambda_2}(x) = \begin{cases} 0 & \text{if } |x| \leq \lambda_1 \\ \text{sgn}(x) \frac{\lambda_1 |x| - \lambda_1}{\lambda_1 - \lambda_2} & \text{if } \lambda_1 < |x| \leq \lambda_2 \\ x & \text{if } |x| > \lambda_2. \end{cases} \]

2) \( \hat{\delta}(x) = (x - \lambda^2 / x)_+ \) suggested by Gao (1998) based the “garrote” of Breiman (1995).

3) \( \hat{\delta}(x) \) constructed as the posterior median for a prior distribution that mixes a point mass at zero with a Gaussian of specified variance. (Abramovich et al. 1998).

While these and other proposals can offer useful advantages in practice, for these notes we concentrate on soft and hard thresholding, because of their simplicity and the fact that they encompass the main theoretical phenomena.

11.2.1. Soft thresholding. The explicit risk function \( r_S(\mu, \lambda) = E[\hat{\delta}_S(x, \lambda) - \mu]^2 \) can be calculated by considering the various zones separately – explicit formulas are in the appendix. Here we focus on qualitative properties and bounds. Write \( \Phi(A) = \int_A \phi(z)dz \) for the standard Gaussian measure of an interval \( A \) and let \( I_\lambda = [-\lambda, \lambda] \). The risk function of soft thresholding is increasing, with derivative bounded by \( 2\mu \):

11.2. Lemma. For \( \mu \geq 0 \),

\[ D_\mu r_S(\lambda, \mu) = 2\mu \Phi(I_\lambda - \mu) \geq 0. \]

Proof. The formula may be established by direct calculation. Write

\[ \hat{\delta}_S(x, \lambda) = (x - \lambda)_+ + (x + \lambda)_-, \]
where \( x_\rightarrow = \min(x, 0) \). Hence, look first at the estimator \( \hat{\delta}(x) = (x - \lambda)_+ \) which, after changing variables to \( z = x - \mu \), has risk function

\[
E_\mu(\delta - \mu)^2 = \mu^2 \int_{-\infty}^{\mu} \phi(z)dz + \int_{\mu}^{\infty} (z - \lambda)^2 \phi(z)dz.
\]

Differentiating the integrals with respect to \( \mu \) and observing the cancellation of terms, one obtains \( 2\mu \Phi(-\infty, \lambda - \mu) \). The advertised result (11.5) for \( \hat{\delta}_S \) follows in exactly the same way, since the two limits of integration \( \lambda - \mu \) and \( -\lambda - \mu \) act completely separately. \( \square \)

The limiting values for \( \mu \) small and large are easily found:

\[
(11.6) \quad r_S(\lambda, \infty) = 1 + \lambda^2,
\]

which shows the effect of the bias due to the shrinkage by \( \lambda \), and

\[
(11.7) \quad r_S(\lambda, 0) = 2 \int_0^\infty (z - \lambda)^2 \phi(z)dz \begin{cases} \leq \frac{\lambda^2}{2} & (\text{all } \lambda) \leq 4\lambda^{-3} \phi(\lambda) & (\text{all } \lambda) \end{cases}
\]

the risk at \( \mu = 0 \) is small because errors are only made when the observation falls outside the threshold zone.

We summarize and extend some of these conclusions about the risk properties:

11.3. Lemma. Let \( \overline{r}(\lambda, \mu) = \min\{r_S(\lambda, 0) + \mu^2, 1 + \lambda^2\} \). For all \( \lambda > 0 \) and \( \mu \in \mathbb{R} \),

\[
(11.8) \quad \frac{1}{2}r(\lambda, \mu) \leq r_S(\lambda, \mu) \leq \overline{r}(\lambda, \mu).
\]

Note that the risk bound \( \overline{r}(\lambda, \mu) \) has the same qualitative flavor as the crude bound (11.1) derived earlier for hard thresholding, only now the constants are correct. In fact, the bound is sharp when \( \mu \) is close to 0 or \( \infty \).

We may interpret \( r_S(\lambda, 0) + \mu^2 \) as a “bias” term, adjusted for risk at zero, and \( 1 + \lambda^2 \) as a “variance” term, reflecting the risk for large \( \mu \).

Proof. Symmetry of the risk function means that we may assume without loss that \( \mu \geq 0 \). Write \( r_\mu(\lambda, s) = (\partial/\partial \mu)r_S(\lambda, \mu)|_{\mu=s} \). By (11.5), the partial derivative \( r_\mu \leq 2 \), and so

\[
(11.9) \quad r_S(\lambda, \mu) - r_S(\lambda, 0) = \int_0^\mu r_\mu(\lambda, s)ds \leq \mu^2.
\]

The upper bound follows from this and (11.6). For the lower bound, write \( x = \mu + z \), and use the simple decomposition

\[
(11.10) \quad E_\mu(|\hat{\delta}_S(x, \lambda) - \mu|^2) \geq E[(z - \lambda)^2, z + \mu > \lambda] + \mu^2P(z + \mu < \lambda).
\]

If \( \mu \leq \lambda \), the right side is evidently bounded below by

\[
E[(z - \lambda)^2, z > \lambda] + \mu^2/2 = (r_S(\lambda, 0) + \mu^2)/2,
\]
using (11.7). If \(\mu \geq \lambda\), then from monotonicity of the risk function, \(r_S(\lambda, \mu) \geq r_S(\lambda, \lambda)\), and applying (11.10) at \(\mu = \lambda\),
\[
  r_S(\lambda, \mu) \geq E[(z - \lambda)^2, z > 0] + \lambda^2/2 = \lambda^2 - 2\lambda\phi(0) + 1/2 \geq (\lambda^2 + 1)/2
\]
with the last inequality valid if and only if \(\lambda \geq \sqrt{8/\pi}\). In this case, the right sides of the last two display both exceed \(\bar{r}(\lambda, \mu)\) and we are done. The proof of the lower bound for \(\lambda < \sqrt{8/\pi}\) is deferred to the Appendix. \(\square\)

fig:risks. Make a figure of risk functions of soft and hard thresholding - p4 of 11/28/00 notes.

We now derive a version of (11.8) that is well suited to showing the relation between sparsity and quality of estimation. Using elementary properties of minima, one may write
\[
  \bar{r}(\lambda, \mu) \leq r_S(\lambda, 0) + (1 + \lambda^2) \min(\mu^2, 1).
\]
In conjunction with the bound \(r_S(\lambda, 0) \leq e^{-\lambda^2/2}\) (cf. (11.7)), we arrive at

11.4. COROLLARY. Suppose \(\delta > 1\) and let \(\lambda_\delta = \sqrt{2 \log \delta^{-1}}\). Then
\[
  r_S(\lambda_\delta, \mu) \leq \delta + (1 + 2 \log \delta^{-1})(\mu^2 \wedge 1).
\]

11.2.2. Hard thresholding. The risk function is easily written in the form
\[
  r_H(\lambda, \mu) = \mu^2 \Phi(I_\lambda - \mu) + \int_{|z| + \mu > \lambda} z^2 \phi(z)dz.
\]
The extreme values for small and large \(\mu\) are:
\[
  r_H(\lambda, \infty) = 1
\]
\[
  r_H(\lambda, 0) = 2 \int_\lambda^\infty z^2 \phi(z)dz \sim 2\lambda \phi(\lambda).
\]
Note that the value at \(\infty\) reflects only variance and no bias, while the value at zero is small, though larger than that for soft thresholding due to the discontinuity at \(\lambda\). However (11.12) also shows that there is a large risk near \(\mu = \lambda\): for large \(\lambda\):
\[
  r_H(\lambda, \lambda) \sim \lambda^2/2.
\]
An analogue of the upper bound of Lemma 11.3 is available for hard thresholding. In this case, define
\[
  \bar{r}(\lambda, \mu) = \begin{cases} r(\lambda, 0) + \mu^2 & 0 \leq \mu \leq \lambda \\ 1 + \mu^2 \Phi(\mu - \lambda) & \mu \geq \lambda, \end{cases}
\]
and extend \(\bar{r}\) to negative \(\mu\) by making it an even function.
11.5. Lemma. (a) For $\lambda > 0$ and $\mu \in \mathbb{R}$, 
\begin{equation}
\frac{1}{\bar{r}}(\lambda, \mu) \leq r_H(\lambda, \mu) \leq (1.2)\bar{r}(\lambda, \mu).
\end{equation}

(b) The large $\mu$ component of $\bar{r}$ has the bound 
\[ \sup_{\mu \geq \lambda} \mu^2 \Phi(\mu - \lambda) \leq \begin{cases} 
\lambda^2/2 & \text{if } \lambda \geq \sqrt{2\pi}, \\
\lambda^2 & \text{if } \lambda \geq 1.
\end{cases} \]

Proof. Again we assume with loss that $\mu \geq 0$. The upper bound for 
$\mu \geq \lambda$ is a direct consequence of (11.12) – the factor 1.2 is not needed here. 
For $0 \leq \mu \leq \lambda$, the approach is as used for (11.9), but for the details of the 
bound $0 \leq (\partial \phi/\partial \mu)r_H(\lambda, \mu) \leq 2\mu$, we refer to Donoho & Johnstone (1994a, 
Lemma 1).

The lower bound is actually easier - it is a direct consequence of an 
inequality analogous to (11.10):
\[ E_{\mu}[\delta_H(x, \lambda) - \mu]^2 \geq E[z^2, z + \mu > \lambda] + \mu^2 P(z + \mu < \lambda). \]

For part (b), set $\alpha = \mu - \lambda \geq 0$ and define $g(\alpha) = (\lambda + \alpha)^2 \Phi(\alpha)$. We have 
\[ g'(\alpha) = (\lambda + \alpha)\phi(\alpha)h(\alpha), \quad h(\alpha) = 2(\Phi(\alpha)/\phi(\alpha)) - \lambda - \alpha, \]
and $h(0) = \sqrt{2\pi} - \lambda \leq 0$ if $\lambda \geq \sqrt{2\pi}$. Differentiation and the bound $\Phi(\alpha) \leq 
\phi(\alpha)/\alpha$ show that $h$ is decreasing and hence negative on $[0, \infty)$, so that 
$g(\alpha) \leq g(0) = \lambda^2/2$. In the case where we only assume that $\lambda \geq 1$, we have 
$g(\alpha) \leq \lambda^2(1 + \alpha)^2 \Phi(\alpha) \leq \lambda^2$, as may be checked numerically, or by 
calculator. \hfill \square

(Figure fig:risks comparing soft and hard thresholding risks.

Remark. In both cases, we have seen that the maximum risk of soft 
and hard thresholding is $O(\lambda^2)$. This is a necessary consequence of having 
a threshold region $[-\lambda, \lambda]$: if $\delta(x)$ is any estimator vanishing for $|x| \leq \lambda$, 
then simply by considering the error made by estimating the error made by estimating 0 when $\mu = \lambda$, we find that
\begin{equation}
E_{\lambda}(\delta(x) - \lambda)^2 \geq \lambda^2 P_{\lambda} \{|x| \leq \lambda\} \approx \lambda^2/2 \quad \text{for large } \lambda.
\end{equation}

Problem. Find bounds sharper than (11.14) for the smallest maximum risk 
attainable by an estimator having threshold zone $[\lambda, \lambda]$. Since such estimators form 
a convex set, this should be accessible via the minimax theorem and the formulas 
for posterior mean given in Chapter ???. The problem is a little reminiscent of 
bounding minimax risk subject to specified risk properties at a point (compare 
Bickel (1983).)
### 11.3. Oracle Inequalities

Suppose that $y_i = \theta_i + \epsilon z_i$, $i = 1, \ldots, n$, with $z_i$ being i.i.d. $N(0,1)$. Given a fixed value of $\theta$, an ideal linear estimator $\theta_{c,i}^* = c_i^\top y_i$ would achieve the best possible mean squared error among linear estimators for the given $\theta$:

$$\min_{c_i} r(\theta_{c,i}^*, \theta) = \frac{\theta_i^2 \epsilon^2}{\theta_i^2 + \epsilon^2} \in \left[ \frac{1}{2}, 1 \right] \cdot \theta_i^2 \wedge \epsilon^2.$$  

Because of the final bound, we might even restrict attention to the ideal projection, which chooses $c_i$ from 0 or 1 to attain

$$\min_{c_i \in \{0,1\}} r(\theta_{c,i}^*, \theta) = \theta_i^2 \wedge \epsilon^2.$$  

Thus the optimal projection choice $c_i(\theta)$ equals 1 if $\theta_i^2 \geq \epsilon^2$ and 0 otherwise, so that

$$\theta_{c}^*(y) = \begin{cases} y_i & \text{if } \theta_i^2 \geq \epsilon^2 \\ 0 & \text{if } \theta_i^2 \leq \epsilon^2. \end{cases}$$

Of course, such estimators depend on $\theta$, and so are unavailable to the statistician without access to an oracle who divulges the ideal shrinkage factor. Thus, with the aid of a “projection oracle”, the best mean squared error attainable is the ideal risk:

$$\mathcal{R}(\theta, \epsilon^2) = \sum_i \theta_i^2 \wedge \epsilon^2.$$  

In later sections (REF), we will discuss further the significance of the ideal risk, and especially its interpretation in terms of sparsity.

We now show that it is possible to mimic the ideal risk with threshold estimators, at least up to a precise logarithmic factor. Let $\hat{\theta}^S_{\lambda_n}$ denote soft thresholding at $\lambda_n = \sqrt{2 \log n}$:

$$\hat{\theta}^S_{\lambda_n,i} = \eta_S(y_i, \epsilon \sqrt{2 \log n}),$$  

and let $\hat{\theta}^H_{\lambda_n}$ denote hard thresholding at the same level.

**Remarks.** 1. Here is one reason for the specific choice $\lambda_n = \sqrt{2 \log n}$ (other choices will be discussed later.) We show that this threshold level is conservative, in the sense that

$$P\{\hat{\theta} = 0 | \theta = 0\} \rightarrow 1$$

as $n \rightarrow \infty$, so that with high probability, $\hat{\theta}$ does not assert the presence of “spurious structure”. To verify this, note that if each $y_i$ is distributed independently as $N(0, \epsilon^2)$, then the probability that no observation exceeds the threshold $\lambda_n$ equals the extreme value probability

$$P\{\max_{i=1,\ldots,n} |Z_i| \geq \sqrt{2 \log n}\} = 1 - [1 - 2\Phi(\sqrt{2 \log n})]^n$$

$$\leq 1 - \exp\{-\frac{2}{\sqrt{\log n}}\} \leq \frac{1}{\sqrt{\log n}},$$

153
where we have used bounds \( \hat{\Phi}(y) \leq \phi(y)/y \), then \( \log(1 - x) \geq -2x \) (for small positive \( x \)) and then \( e^{-y} \geq 1 - y \). While the final bound may not be sharpest possible, it does at least indicate that the rate of approach of the exceedence probability to zero is not very fast.

**Alan Miller’s variable selection scheme.** A method of Miller (1984, 1990) offers an interesting perspective on \( \sqrt{2 \log n} \) thresholding. Consider a traditional linear regression model

\[
y = X\beta + \sigma^2z,
\]

where \( y \) has \( N \) components and \( X \) has \( n < N \) columns \([x_1 \cdots x_n]\) and the noise \( z \sim N_N(0, I) \). For convenience only, assume that the columns are centered and scaled: \( x_i^T 1 = 0 \) and \( |x_i|^2 = 1 \). Now create “fake” regression variables \( x_i^* \), each as an independent random permutation of the entries in the corresponding column \( x_i \). Assemble \( X \) and \( X^* = [x_1^* \cdots x_n^*] \) into a larger design matrix \( \tilde{X} = [X \; X^*] \) with coefficients \( \hat{\beta} = [\beta^T \; \beta^*] \) and fit the enlarged regression model \( y = X\beta \) by a forward stepwise method. Let the method stop just before the first ‘fake’ variable \( x_i^* \) enters the model. Since the new variables \( x_i^* \) are approximately orthonormal among themselves and approximately orthogonal to each \( x_i \), the estimated coefficients \( \hat{\beta}_i \) are essentially i.i.d. \( \mathcal{N}(0,1) \), and so the stopping criterion amounts to “enter variables above the threshold given by \( \max_{i=1,\ldots,n} |\hat{\beta}_i| = \sqrt{2 \log n} \).”

(Comments on choice \( \sqrt{2 \log n} \): extreme value distribution and concentration inequality)

The classical extreme value theory result for the maximum of \( n \) i.i.d. \( N(0,1) \) variables \( Z_i \) states that

\[
l_n^{-1} \left( \max_{i=1,\ldots,n} |Z_i| - a_n \right) \leq W, \quad P(W \leq t) = \exp \{-e^{-t}\},
\]

where \( a_n = \sqrt{2 \log n} \) and \( b_n = (\log \log n + 4\pi)/(2\sqrt{2 \log n}) \).

A non-asymptotic bound follows from the Tsirelson-Sudakov-Ibragimov bound [ref, check] for a Lipschitz(1) function \( f : \mathbb{R}^n \to \mathbb{R} \) of a standard Gaussian \( n \)-vector \( Z \sim \mathcal{N}_n(0, I) \):

\[
P \{ |f(Z) - Ef(Z)| \geq t \} \leq 2e^{-t^2/2}.
\]

When applied to \( f(z) = \max|z_i| \), this says that the tails of \( \max|Z_i| \) are sub-Gaussian, while the extreme value result in fact says that the limiting distribution has negligible variability around \( a_n \).

11.6. **Proposition.**

\[
(11.15) \quad E \|\hat{\theta}^S_{\Lambda_n} - \theta\|_2^2 \leq (1 + 2\log n)[\epsilon^2 + \sum_{i=1}^n \theta_i^2 \vee \epsilon^2].
\]

A similar result holds for \( \hat{\theta}^H_{\Lambda_n} \), with the multiplier \( (1 + 2\log n) \) replaced by \( \Lambda_n \leq (1.2)((1 + 2\log n) \) with \( \Lambda_n \sim 2\log n \) as \( n \to \infty \).
The factor $2 \log n$ is optimal without further restrictions on $\theta$:

$$
\inf_{\theta} \sup_{\theta \in \mathbb{R}^n} \frac{E[\|\hat{\theta} - \theta\|^2]}{\epsilon^2 + \sum_{i=1}^{n} \theta_i^2} \geq (2 \log n)(1 + o(1)), \quad n \to \infty.
$$

Results of this type help to render the idea of ideal risk statistically meaningful: a genuine estimator, depending on only on available data, and not upon access to an oracle, can achieve the ideal risk $R(\theta, \epsilon)$ up to the (usually trivial) additive factor $\epsilon^2$ and the multiplicative factor $2 \log n + 1$. This logarithmic penalty can certainly be improved if extra constraints upon $\theta$ are added: for example that $\theta$ belong to some $\ell_p$ ball, weak or strong (Reference!!). However, the lower bound (11.16) shows that the $2 \log n$ factor is optimal for unrestricted $\theta$, at least asymptotically.

Note that the upper bounds are non-asymptotic, holding for all $\theta \in \mathbb{R}^n$ and $n \geq 1$.

The upper bound extends trivially to correlated, heteroscedastic data, since it thresholding depends only on the univariate marginal distributions of the data. The only change is to replace $\epsilon^2$ by $\epsilon_i^2$, the variance of the $i$th coordinate, in the ideal risk, and to modify the additive factor to $\epsilon_i^2 \leq n_i^2$. There is also a version of the lower bound under some conditions on the correlation structure: for details see Johnstone & Silverman (1997).

**Proof.** Upper bound. For soft thresholding, letting $\delta = 1/n$ in Corollary 11.4 and then summing over coordinates yields

$$
E \sum_i [\mu_{\lambda, i} - \mu_i]^2 \leq 1 + (1 + 2 \log n) \sum_i \mu_i^2 \land 1.
$$

The result (11.15), and indeed a slightly stronger version, follows by rescaling to noise level $\epsilon$.

For hard thresholding, essentially the same argument provides the bound

$$
\Lambda_n \leq (1.2)(1 + 2 \log n),
$$

starting from the inequalities (see Donoho & Johnstone (1994a)):

$$
\tau_H(\lambda, \mu) \leq \begin{cases} 
\tau_H(\lambda, 0) + 1.2\mu^2 & \mu \leq \lambda \\
1 + \lambda^2 & \mu > \lambda.
\end{cases}
$$

**Lower bound.** THIS NEEDS TO BE REDONE. The proof is deferred till the next section, since it is based on a univariate result. An alternate $n$-variate argument is in Donoho & Johnstone (1994a).

**Comparison of James Stein and Thresholding.** It is instructive to compare the bounds available for the mean squared error of James-Stein estimation and thresholding. For the James-Stein (positive part) estimator, we
have from ??? that

\[ E \left\| \tilde{\theta}_{J\tilde{S}} - \theta \right\|^2 \leq 2c^2 + \frac{(n - 2)c^2 \left\| \theta \right\|^2}{(n - 2)c^2 + \left\| \theta \right\|^2}, \]

\[ \in 2c^2 + \left| \frac{1}{n} - \frac{1}{m} \right| \cdot \min(n, \theta^2, \sum_i \theta_i^2). \]

On the other hand, when rescaled to noise level \( \epsilon \), (11.17) becomes

\[ E \left\| \tilde{\theta}_{\lambda \epsilon} - \theta \right\|^2 \leq c^2 + (2 \log n + 1) \sum_i \min(c^2, \theta_i^2). \]

Clearly, thresholding dominates if

\[ \sum_i \min(c^2, \theta_i^2) \ll \min(n, \theta^2, \sum_i \theta_i^2). \]

This will happen if there is considerable sparsity among the \( \theta_i \): if some \( |\theta_i| \) are much larger than \( \epsilon \) while others, indeed most, are smaller. On the other hand, James-Stein shrinkage wins (by a factor of \( 2 \log n \), not more) if all \( |\theta_i| \) are nearly equal.

11.4. Optimality of risk bounds.

One may reasonably ask how tight are the risk bounds established in the previous sections. Consider, for example, (11.11), which implies

\[ r_S(\lambda, \mu) \leq (1 + 2 \log \delta^{-1})(\delta + \mu^2 \wedge 1). \]

We consider \( 3(\mu) = \delta + \mu^2 \wedge 1 \) as a simple and convenient bound for the risk of non-linear thresholding. We study the tightness of this bound by considering the ratio \( \sup_{\mu} r(\hat{\mu}, \mu)/3(\mu) \) for any estimator \( \hat{\mu} \). [Note that we cannot put \( \delta = 0 \) since \( r(\hat{\mu}, \mu) \leq C(\mu^2 \wedge 1) \) at \( \mu = 0 \) would force \( E\hat{\mu}^2 = 0 \) and hence that \( \hat{\mu} \equiv 0 \).]

11.7. PROPOSITION. As \( \delta \to 0 \),

\[ (11.18) \quad \tilde{R} = \inf_{\hat{\mu}} \sup_{\mu} \frac{r(\hat{\mu}, \mu)}{\delta + \mu^2 \wedge 1} \geq (2 \log \delta^{-1})(1 + o(1)). \]

**PROOF.** Consider the asymmetric loss function

\[ \tilde{L}(a, \mu) = l(\mu)(a - \mu)^2, \]

where in our case, \( l(\mu) = [\delta + \mu^2 \wedge 1]^{-1} \). The minimax risk \( \tilde{R} \) is bounded below by the Bayes risk \( \tilde{B}(\pi) \) of any prior \( \pi \) with respect to loss function \( \tilde{L} \). The Bayes rule is found by minimising posterior expected loss:

\[ (11.19) \quad \tilde{\mu}_\pi(x) = \frac{E\left[ \mu l(\mu) | x \right]}{E[l(\mu) | x]} \]
Consider two point priors $\pi = \eta \mu_0 + \bar{\eta} \mu_{00}$, where $\bar{\eta} = 1 - \eta$, and $\mu_0 \geq 1$ is to be specified. For such priors, the posterior distribution is also concentrated on 0 and $\mu_0$, with

$$\eta(x) = P(\{0\} | x) = \frac{1}{1 + \gamma \Lambda(x)},$$

where $\gamma = \gamma_0 = \bar{\eta} / \eta$ is the prior odds of $\mu_0$ and the likelihood ratio

$$\Lambda(x) = \phi(x - \mu_0) / \phi(x) = \exp(\mu_0 x - \mu_0^2 / 2).$$

Writing also $\bar{\eta}(x) = 1 - \eta(x)$, we find that (11.19) becomes

$$\tilde{\mu}_\pi(x) = \frac{\mu_0 \cdot l(\mu_0) \bar{\eta}(x)}{l(0) \eta(x) + l(\mu_0) \bar{\eta}(x)}$$

A lower bound for the Bayes risk $\tilde{B}(\pi)$ follows by ignoring the contribution from one of the prior support points:

$$\tilde{R} \geq \tilde{B}(\pi) \geq \bar{\eta} l(\mu_0) E_{\mu_0} [\tilde{\mu}_\pi(X) - \mu_0]^2.$$

In our case, $l(0) = \delta^{-1} \gg l(\mu_0) = (1 + \delta)^{-1}$. Suppose that we can arrange that

$$\mu_0^2 \sim 2 \log \delta^{-1}, \quad \bar{\eta}_0 = \pi \{\mu_0\} \rightarrow 1$$

and that with high probability under $X \sim N(\mu_0, 1)$

$$\delta \ll \eta(x) \ll 1.$$

Then from (11.22), $\tilde{\mu}_\pi(x) = \mu_0 \bar{\eta}(x) / [\delta^{-1} \eta(x) + 1] \approx 0$, and from (11.23) and (11.24), we have approximately

$$\tilde{R} \geq \frac{\bar{\eta}}{1 + \delta} \mu_0^2 \sim 2 \log \delta^{-1} \quad \text{as} \quad \delta \rightarrow 0.$$  

To accomplish (11.24) and (11.25), choose $\eta_0 = \pi(0) = 1 / \log \delta^{-1}$, and then $a_\delta = \log \log \delta^{-1}$ and finally $\mu_0$ to satisfy

$$\eta(\mu_0 + a_\delta) = \delta.$$

The heuristic idea is that since $a_\delta \rightarrow \infty$, $\eta(\mu_0 + y) \gg \delta$ for any fixed $y$ due to the fast decay of Gaussian tails. More explicitly, some algebra with (11.20) and (11.21) shows that $\mu_0^2 / 2 + a_\delta \mu_0 = \log \delta^{-1} - \log \log \delta^{-1} + o(1)$ and hence that $\mu_0^2 \sim 2 \log \delta^{-1}$ as needed for (11.24). Thus, for fixed $y$,

$$\eta(y + \mu_0) = [1 + \gamma_0 \exp(\mu_0 y + \mu_0^2 / 2)]^{-1} \rightarrow 0,$$

and since $\Lambda(y + \mu_0) / \Lambda(a_\delta + \mu_0) = \exp(\mu_0 (y - a_\delta)) \rightarrow 0$, we have also from (11.20) that

$$\frac{\delta}{\eta(y + \mu_0)} = \frac{\eta(a_\delta + \mu_0)}{\eta(y + \mu_0)} \rightarrow 0.$$

Together (11.27) and (11.28) establish (11.25). Details for (11.26) are in the Appendix.
Verification of (11.26). We note first from (11.22) that
\[ \mu_\delta - \hat{\mu}_\pi(x) = \mu_\delta \left[ 1 + \frac{I(\mu_\delta \eta(x))}{I(0) \eta(x)} \right]^{-1}. \]
Putting this together with (11.23) and setting \( y = x - \mu_\delta \) gives
\[ \tilde{R} \geq \frac{\tilde{g}}{1 + \delta^2} \int \left[ 1 + \frac{\delta \tilde{g}(y + \mu_\delta)}{1 + \delta \eta(y + \mu_\delta)} \right]^{-2} \phi(y)dy. \]
It follows from (11.27) and (11.28) that the integrand in parentheses converges to 1 for each fixed \( y \). The bound (11.26) is then a consequence of the dominated convergence theorem. \( \square \)

11.4.1. Extension to \( n \) dimensions. (Extension to give proof of lower bound of oracle ineq. theorem - REMOVE!)

11.5. Appendix: Further details

2°. The mean squared error of a thresholding rule \( \hat{\delta}(x, \lambda) \) (either hard or soft) may be written
\[ r(\lambda, \mu) = E_\mu[\hat{\delta}(x, \lambda) - \mu]^2 = \int (\hat{\delta}(x, \lambda) - \mu)^2 \phi(x - \mu)dx. \]
Breaking the range of integration into regions \((-\infty, -\lambda), [-\lambda, \lambda], (\lambda, \infty)\) to match the thresholding structure, one obtains the following basic mean squared error formulas:
\[ r_S(\lambda, \mu) = 1 + \lambda^2 + (\mu^2 - \lambda^2 - 1)[\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)] - (\lambda - \mu)\phi(\lambda + \mu) - (\lambda + \mu)\phi(\lambda - \mu), \]
\[ r_H(\lambda, \mu) = \mu^2[\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)] + \tilde{\Phi}(\lambda - \mu) + \tilde{\Phi}(\lambda + \mu) + (\lambda - \mu)\phi(\lambda + \mu) + (\lambda + \mu)\phi(\lambda - \mu) \]
where \( \phi \) and \( \Phi \) denote the standard Gaussian density and cumulative distribution functions respectively, and \( \tilde{\Phi}(x) = 1 - \Phi(x) \).

3°. Proof of lower bound in Lemma 11.3 for \( 0 \leq \lambda \leq 2 \). Let \( \mu_\lambda \) be the solution in \( \mu \) of \( r(\lambda, 0) + \mu^2 = 1 + \lambda^2 \). Since \( r(\lambda, 0) \leq e^{-\lambda^2/2} < 1 \), (compare (11.7) ), it is clear that \( \mu_\lambda > \lambda \). For \( \mu \leq \mu_\lambda \) we may write, using (11.5),
\[ R(\lambda, \mu) = \frac{r(\lambda, \mu)}{r(\lambda, \mu)} = \frac{r(\lambda, 0) + \int_0^\mu 2s \Phi(I_\lambda - s)ds}{r(\lambda, 0) + \mu^2}. \]
We first verify that \( R(\lambda, \mu) \) is decreasing in \( \mu \leq \mu_\lambda \). Indeed \( \mu \to [c + f_1(\mu)]/[c + f_2(\mu)] \) is decreasing if both \( f_1(\mu) \leq f_2(\mu) \) and \( (f_1/f_2)(\mu) \) is decreasing. The former condition is evident, while the latter follows by the rescaling \( v = s/\mu ; \) for then \((f_1/f_2)(\mu) = 2 \int_0^1 \Phi(I_\lambda - \mu v)dv.\)
For $\mu \geq \mu_\lambda$, we also have $R(\lambda, \mu) \geq R(\lambda, \mu_\lambda)$ since $r(\lambda, \mu) \geq r(\lambda, \mu_\lambda)$ while $\bar{r}(\lambda, \mu) \equiv 1 + \lambda^2$. Consequently, for all $\mu$

\[
R(\lambda, \mu) \geq r(\lambda, \mu_\lambda)/[r(\lambda, 0) + \mu^2],
\]

and numerical evaluation for $0 \leq \lambda \leq 2$ shows the right side to be bounded below by $.516$, with the minimum occurring for $\lambda \in [.73, .74]$.

**Exercises**

1. *Mill’s ratio and Gaussian tails.* The function $R(\lambda) = \tilde{\Phi}(\lambda)/\phi(\lambda)$ is sometimes called Mill’s ratio. Show that the modified form

\[
M(\lambda) = \frac{\lambda \tilde{\Phi}(\lambda)}{\phi(\lambda)} = \int_0^\infty e^{-v^2/(2\lambda^2)}dv,
\]

and hence that $M(\lambda)$ is increasing from 0 at $\lambda = 0$ up to 1 at $\lambda = \infty$.

Define the $l$-th approximation to the Gaussian tail integral by

\[
\tilde{\Phi}_l(\lambda) = \lambda^{-l} \phi(\lambda) \sum_{k=0}^l \frac{(-1)^k \Gamma(2k + 1)}{k! 2^k \lambda^{2k}}.
\]

Show that for each $k \geq 0$ and all $\lambda > 0$ that

\[
\tilde{\Phi}_{2k+1}(\lambda) \leq \tilde{\Phi}(\lambda) \leq \tilde{\Phi}_{2k}(\lambda).
\]

[Hint: induction shows that $(-1)^{l-1}[e^{-x} - \sum_{k=0}^l (-1)^k x^k/k!] \geq 0$ for $x \geq 0$.]

As consequences, we obtain, for example, the bounds

\[
\lambda^{-1}\phi(\lambda)(1 - \lambda^{-2}) \leq \tilde{\Phi}(\lambda) \leq \lambda^{-1}\phi(\lambda),
\]

and the expansion, for large $\lambda$,

\[
\tilde{\Phi}(\lambda) \sim \lambda^{-1}\phi(\lambda)[1 - \lambda^{-2} + 3\lambda^{-4} - 15\lambda^{-6} + O(\lambda^{-8})].
\]

2. *Risk of soft thresholding at 0.* Let $z \sim N(0, 1)$, and $r(\lambda, 0) = E\tilde{\nu}^2(z)$ denote the mean squared error of soft thresholding at $\lambda = 0$.

(a) Use (11.32) to show that $r(\lambda, 0) \sim 4\lambda^{-3}\phi(\lambda)$ as $\lambda \rightarrow \infty$.

(b) Let $\delta(\lambda) = e^{-\lambda^2/2} - r(\lambda, 0)$. Use (11.31) to show that $\delta(\lambda) > 0$ for $\lambda \geq \lambda_0 = 2\phi(0)$.

(c) Show that $\delta(\lambda)$ is concave for $\lambda \in [0, 1]$, and conclude that $r(\lambda, 0) \leq e^{-\lambda^2/2}$ for all $\lambda \geq 0$.

3. Derive the following inequalities for hard thresholding, which are sharper than direct application of the bounds in (11.13):

\[
\begin{align*}
 r_H(\lambda, \lambda) &\geq (\lambda^2 + 1)/2, \\
r_H(\lambda, 0) &\geq (2\lambda \vee \sqrt{2\pi})\phi(\lambda).
\end{align*}
\]

(Birgé & Massart 2001)
4. Refer to Alan Miller’s variable selection scheme, and assume as there that the columns are centered and scaled: \( \langle x_i, 1 \rangle = 0 \) and \( \langle x_i, x_i \rangle = 1 \). Show that the permuted columns are approximately orthogonal to each other and to the original columns. More precisely, show that

(i) if \( j \neq k \), then \( \langle x_j^*, x_k^* \rangle \) has mean 0 and standard deviation \( 1/\sqrt{N-1} \), and

(ii) for any pair \( (j, k) \), similarly \( \langle x_j^*, x_k \rangle \) has mean 0 and standard deviation \( 1/\sqrt{N-1} \).
CHAPTER 12

The optimal recovery approach to thresholding.

[NEEDS AN INTRODUCTION!]

We have seen that the fact that the maximum of $n$ independent standard normal
variates is usually bounded by $\sqrt{2 \log n}$ leads to some attractive
properties for threshold estimators which use this relatively high threshold.

In this chapter (and its multiresolution cousin, REF), we will see how
some quite general conclusions about $\sqrt{2 \log n}$ thresholding may be drawn
by analyzing a related optimal recovery problem with deterministic noise.

The plan is to consider a whole class of parameter spaces $\Theta$ and loss
functions $\|\hat{\theta} - \theta\|$. The point is that we establish near optimality properties
for a single estimator over many settings, rather than an exact optimality
result for a single setting which may be dangerously misleading if that setting
is not, in fact, the appropriate one.

12.1. An Optimal Recovery Model

Consider the following deterministic version of the sequence model. Data
$x = (x_I)$ is observed that satisfies

\[ x_I = \theta_I + \delta u_I \quad |u_I| \leq 1 \quad I \in I \]

It is desired to recover the unknown vector $\theta$, but it is assumed that the
deterministic noise $u$ might be chosen maliciously by an opponent, subject
only to the uniform size bound. The noise level $\delta$ is assumed known. The
worst case error suffered by an estimator $\hat{\theta}$ is then

\[ e(\hat{\theta},\theta;\delta) = \sup_{|u_I| \leq 1} \|\hat{\theta}(x) - \theta\|. \]

We will see that a number of conclusions for the statistical (Gaussian)
sequence model can be drawn, after appropriate calibration, from the deter-
ministic model (12.1).

Assumptions on loss function and parameter space. Throughout this
chapter we will assume:

(i) $\Theta \subset \ell_2(I)$ is solid, orthosymmetric and convex$^1$.

(ii) The error norm $\| \cdot \|$ is also solid and orthosymmetric, in the sense that

\[ |\xi| \leq |\theta_I| \quad \forall I \quad \Rightarrow \quad \|\xi\| \leq \|\theta\|. \]

Remark. The literature on optimal recovery goes back to Golomb &
Weinberger (1959) and a 1965 Moscow dissertation of Smolyak. See also
Micchelli (1975), Micchelli & Rivlin (1977) and Donoho (1994), who makes the connection with statistical estimation. These latter references are concerned with estimation of a linear functional, while here we are concerned with the whole object $\theta$.

**12.1.1. The Uniform Shrinkage Property of Soft Thresholding.**

Soft thresholding at threshold $\lambda$ can be used in the optimal recovery setting:

$$\hat{\theta}_{\lambda I}(x_I) = \text{sgn}(x_I)(|x_I| - \lambda)_+.$$  

The shrinkage aspect of soft thresholding has the simple but important consequence that the estimate remains confined to the parameter space:

**12.1. Lemma.** If $\Theta$ is solid orthosymmetric and $\lambda \geq \delta$, then $\theta \in \Theta$ implies $\hat{\theta}_\lambda \in \Theta$.

**Proof.** Since soft thresholding shrinks each data coordinate $x_I$ towards 0 (but not past 0!) by an amount $\lambda$ that is greater than the largest possible noise value $\delta$ that could be used to expand $\theta_I$ in generating $x_I$, it is clear that $|\hat{\theta}_{\lambda I}| \leq |\hat{\theta}_I|$. Since $\Theta$ is solid orthosymmetric, this implies $\hat{\theta}_\lambda \in \Theta$. □

include figure.

**12.1.2. Minimax Error.** The minimax error of recovery in the deterministic model is

$$E(\Theta, \delta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} e(\hat{\theta}, \theta),$$

where $e(\hat{\theta}, \theta)$ is given by (12.2). Good bounds on this minimax error can be found in terms of a *modulus of continuity* defined by

$$\Omega(\delta) = \Omega(\delta; \Theta; \| \cdot \|) = \sup_{\theta \in \Theta} \{ \| \theta \| : \| \theta \|_\infty \leq \delta \}.$$  

We may obtain an *upper bound* for the minimax error $E(\Theta, \delta)$ by using soft thresholding with threshold $\delta$. The fact that $\hat{\theta}_\delta$ shrinks the data towards 0 by $\delta$ means that $|\hat{\theta}_{\delta I} - x_I| \leq \delta$ for all $I$. Since the model (12.1) says that $|x_I - \theta_I| \leq \delta$ for all $I$, we obtain

$$\| \hat{\theta}_\delta - \theta \|_\infty \leq \| \hat{\theta}_\delta - \theta \|_\infty + \| \theta - \theta \|_\infty \leq 2\delta.$$  

Both $\theta$ and $\hat{\theta}_\delta \in \Theta$ (by the uniform shrinkage property), and since $\Theta$ is convex and symmetric about 0, this implies that $(\hat{\theta}_\delta - \theta)/2 \in \Theta$. It follows that

$$e(\hat{\theta}_\delta, \theta) \leq 2\Omega(\delta).$$

Turning now to a *lower bound*, suppose that $\hat{\theta} \in \Theta$ attains the value $\Omega(\delta)$ defining the modulus. The data sequence $x = 0$ is potentially observable under (12.1) if either $\theta = 0$ or $\theta = \hat{\theta}$, and so for any estimator $\theta$,

$$\sup_{\theta \in \Theta} e(\hat{\theta}, \theta) \geq \sup_{\theta \in \Theta} \| \hat{\theta}(0) - \theta \| \geq \Omega(\delta)/2.$$
12.2. Theorem. Suppose that $\Theta$ is solid, orthosymmetric and convex, and that the error norm $\| \cdot \|$ is solid and orthosymmetric. Then

$$(1/2)\Omega(\delta) \leq E(\Theta, \delta) \leq 2\Omega(\delta).$$

In addition, soft thresholding $\hat{\theta}_\delta$ is near minimax simultaneously for all such parameter spaces and error norms.

12.2. Evaluation of Moduli of Continuity

12.2.1. Modulus for $\ell_p$ balls. In the definition of the modulus $\Omega(\delta)$, we take $\Theta = \Theta_{n,p}(C)$ as defined at (14.22) and $\| \cdot \|$ equal to the norm of $\ell_{2,n}$. Thus, we have,

$$\Omega^2(\delta) = \Omega^2(\delta; \Theta_{n,p}(C), \| \cdot \|_2) = \sup_{\theta \in \Theta_{n,p}(C)} \sum_{i=1}^n \theta_i^2 \wedge \delta^2.$$

[Note that we often drop the explicit reference to $C$ in $\Omega(\delta).$] In the case $p < 2$ this was analysed carefully in Chapter 14 [INCLUDE PROOF HERE??], and a similar approach yields the results for $p \geq 2$. We summarize the conclusions here and in Table 1 below. A key notion is that of the least favorable configurations $\theta_{n_0, \delta_0}$ for the modulus: These may be essentially completely described by permutations of vectors with $n_0$ non-zero entries

$$(12.4) \quad \theta_{n_0, \delta_0} = (\delta_0, \ldots, \delta_0, 0, \ldots, 0) \quad \delta_0 \leq \delta, \ 1 \leq n_0 \leq n.$$  

and in terms of $\theta_{n_0, \delta_0}$, the value of the modulus is essentially given by

$$\Omega^2(\delta) = n_0 \delta_0^2.$$  

Thus $n_0$ (or the ratio $n_0/n$) measures the sparsity of the least favorable configuration. When $p > 2$, the least favorable configurations are always dense, since the contours of the $\ell_2$ loss touch those of the $\ell_p$ norm along the direction $(1, \ldots, 1)$. On the other hand, when $p < 2$, the maximum value of $\ell_2$ error over the intersection of the $\ell_p$ ball and $\delta-$cube is always attained on the boundary of the cube, which leads to sparser configurations when $C < \delta^{n_0}/p$.

Note: In the case $p < 2$, $\delta \leq C \leq \delta n^{1/p}$ (and only in this case), the description above and in the table needs to be slightly modified to incorporate an $n_0 + 1-$th atom of size $\eta \in [0, \delta_0]$ chosen so that the constraint $n_0 \delta_0^p + \eta^p = C^p$ holds. In this case $\Omega^2$ (given exactly in (14.56)) is smaller than the value $C^p \delta^{2-p}$ given in the table but by a factor of at most $n_0/(n_0 + 1)$. Likewise, $\Omega^2$ exceeds $n_0 \delta_0^2$ but again by a factor of at most $1 + 1/n_0$.

[Include figure of growing cubes, as in SU lecture notes?]
Table 1. Value of modulus $\Omega^2(\delta, C)$ and the number $n_0$ and size $\delta_0$ of spikes in the least favorable configuration.

<table>
<thead>
<tr>
<th>$p &lt; 2$</th>
<th>$\Omega^2$</th>
<th>$\delta_0$</th>
<th>$n_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta \leq Cn^{-1/p}$</td>
<td>$n\delta^2$</td>
<td>$\delta$</td>
<td>$n$</td>
</tr>
<tr>
<td>$Cn^{-1/p} \leq \delta \leq C$</td>
<td>$C^p\delta^{2-p}$</td>
<td>$\delta$</td>
<td>$[C/\delta]^p$</td>
</tr>
<tr>
<td>$\delta \geq C$</td>
<td>$C^2$</td>
<td>$C$</td>
<td>$1$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$p \geq 2$</th>
<th>$\Omega^2$</th>
<th>$\delta_0$</th>
<th>$n_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta \leq Cn^{-1/p}$</td>
<td>$n\delta^2$</td>
<td>$\delta$</td>
<td>$n$</td>
</tr>
<tr>
<td>$\delta \geq Cn^{-1/p}$</td>
<td>$n^{1-2/p}C^2$</td>
<td>$Cn^{-1/p}$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

12.3. The Statistical Sequence Model

In the deterministic model of optimal recovery, Theorem 12.2 is a strong statement of the near optimality of soft thresholding over a range of parameter spaces and error norms, phrased in terms of the modulus of continuity $\Omega(\delta)$.

Consider now a monoresolution Gaussian error model

\[(12.5) \quad y_i = \theta_i + \epsilon z_i, \quad z_i \overset{i.i.d.}{\sim} N(0, 1), \quad i = 1, \ldots, n.\]

(multiresolution models are deferred to Chapter ??). The connection with the optimal recovery model (with $I = \{1, \ldots, n\}$) is made by considering the event

\[ A_n = \{ \sup_{I \in \mathcal{I}} |z_I| \leq \sqrt{2\log n} \}, \]

which because of the properties of maxima of i.i.d. Gaussians has probability approaching one:

\[ P(A_n) = \pi_n \geq 1 - 1/\sqrt{\pi \log n} \to 1 \quad \text{as} \quad n \to \infty. \]

The key idea is to apply results from the optimal recovery model with deterministic noise level $\delta_n = \epsilon \sqrt{2\log n}$ on the set $A_n$. Thus, in the statistical model we consider the soft thresholding estimator $\hat{\theta}_n$ of (16.2) at level $\epsilon \sqrt{2\log n}$. We therefore obtain immediately

12.3. Proposition. If $(\Theta, \| \cdot \|)$ is solid, orthosymmetric and $\Theta$ is convex, then

\[ \sup_{\theta \in \Theta} P(\|\hat{\theta}_n - \theta\| \leq 2\Omega(\epsilon \sqrt{2\log n})) \geq \pi_n \to 1. \]

12.3.1. Lower Bounds. In the statistical problem, one does not have an overtly malicious opponent choosing the noise, which suggests that statistical estimation might not be as hard as optimal recovery. However, a statistical lower bound argument, based on hypercubes, will show that in
fact this is not true, and that in many cases, the modulus yields (up to logarithmic factors), a description of the difficulty of the statistical problem as well.

From now on, we restrict to parameter spaces which are \( \ell_p \) balls: \( \Theta_{n,p}(C) = \{ \theta \in \mathbb{R}^n : \sum_i^n |\theta_i|^p \leq C^n \}. \) At one point, it will be convenient to consider a set in which vectors have at most \( k \) non-zero components:

\[
\Theta_{n,0}(C, k) = \{ \theta \in \mathbb{R}^n : \# \{ \theta_i \neq 0 \} \leq k, |\theta| \leq C \}.
\]
The subscript 0 emphasizes that in some sense, this is the \( p = 0 \) limit of \( \ell_p \) balls: for example \( \Theta_{n,0}(C, 1) \subset \Theta_{n,p}(C) \) for all \( p > 0 \).

In stating lower bounds for the statistical model over \( \ell_p \) balls, we recall the structure of extremal configurations for the modulus \( \Omega(\varepsilon) = \Omega(\varepsilon; \Theta_{n,p}(C), \| \cdot \|_2) \). Indeed, let \( n_0 = n_0(p, C, \varepsilon, n) \) be the number of non-zero components in the extremal vectors \( \theta_{n_0,\delta_0} \) of (12.4). We develop two bounds, for the dense (\( n_0 = 1 \)) and sparse (\( n_0 = 1 \)) cases respectively.

12.4. Proposition. Assume data is taken from model (12.5).

(i) (Dense case). Let \( \pi_0 = \Phi(-1)/2 \).

\[
\inf_{\theta} \sup_{\Theta_{n,p}(C)} P\{ \| \hat{\theta} - \theta \| \geq \sqrt{n} \Omega(\varepsilon) \} \geq 1 - e^{-2^{n_0} \pi_0^2}.
\]

(ii) (Sparse case). Fix \( \eta > 0 \) small, \( \delta_{0n} = \varepsilon \sqrt{(2 - \eta) \log n} \). Then, as \( n \to \infty \),

\[
\inf_{\theta} \sup_{\Theta_{n,0}(\delta_{0n}, 1)} P\{ \| \hat{\theta} - \theta \| \geq \varepsilon \delta_{0n} \} \to 1.
\]

Remarks. 1. In the sparse case, one can rewrite the lower bound in terms of the modulus \( \Omega \) by setting \( \lambda_n = \sqrt{2 \log n} \) and \( c_n = (1 - \eta/2)^{1/2} \). Then, when \( p < 2 \), \( \delta_{0n}/2 = (c_n/2)\Omega_{n,p}(\lambda_n, \varepsilon, \delta_{0n}) \).

2. Thus, in the statistical model, an upper bound for estimation over \( \Theta_{n,p}(C) \) is given, on a set of high probability, by \( \Omega(c_n \sqrt{2 \log n}) \), whereas the lower bound in the dense case (12.6) is of order \( \Omega(\varepsilon_n) \). Thus there is a logarithmic gap between the two bounds. However, the near optimality of \( \sqrt{2 \log n} \) soft thresholding holds quite generally: the method of proof works for all \( \ell_q \) losses, not just the \( \ell_2 \) case given explicitly here, and over all \( \ell_p \) balls \( \Theta_{n,p}(C) \).

3. In the sparse case, \( p < 2 \), remark 1) shows that in fact the logarithmic term appears in the lower bound also, so that there are cases in which the optimal recovery method yields exact rate results in the statistical model.

4. It would be possible to give an explicit lower bound for (12.7) depending on \( n \) and \( \eta \), for example by working further from the proof given below.

Proof. Dense Case. The argument uses a version of the hypercube method. [cf. Assouad?] Let \( (n_0, \delta_0) \) be parameters of the worst case configuration for \( \Omega(\varepsilon) \), with \( \delta_0 \leq \varepsilon \). Let \( \pi \) be the distribution on \( \theta \) which makes
\( \theta_i \) independently equal to \( \pm \delta_0 \) with probability \( \frac{1}{2} \) for \( i = 1, \ldots, n_0 \), and all other co-ordinates 0. Since \( \text{supp} \pi \subset \Theta \), we have for any \( (\theta, y) \)-measurable event \( A \),

\[
\sup_{\theta \in \Theta} P_\theta(A) \geq P_\pi(A).
\]

Suppose now that \( \hat{\theta}(y) \) is an arbitrary estimator and let \( N(\hat{\theta}(y), \theta) = \sum_i I[\hat{\theta}_i(y) \theta_i < 0] \) be the number of sign errors made by \( \hat{\theta} \). Under \( P_\pi \),

\[
\|\hat{\theta} - \theta\|^2 \geq \delta_0^2 N(\hat{\theta}(y), \theta).
\]

Consider the posterior distribution of \( \theta \): because of coordinate-wise independence

\[
P_\pi(\theta_i = \delta_0 | y_i) > P_\pi(\theta_i = -\delta_0 | y_i) \iff y_i > 0.
\]

Consequently, the Bayes rule for the loss function which counts sign errors is 
\( \hat{\theta}_{\pi,i}(y) = \delta_0 \text{sgn}(y_i) \), and by its construction

\[
N(\hat{\theta}_{\pi}(y), \theta) \leq N(\hat{\theta}(y), \theta)
\]

for any other estimator \( \hat{\theta}(y) \). [Here \( \leq \) denotes stochastic inequality in the joint distribution of \( (\theta, y) \)]. Consequently, by combining (12.8), (12.9) and (12.10), we obtain

\[
S(c) = \inf_{\theta} \sup_{\delta} \mathbb{P}_{\theta} (\|\hat{\theta} - \theta\|^2 \geq c \delta_0^2) \geq P_\pi \{ N(\hat{\theta}_{\pi}(y), \theta) \geq c \}.
\]

Under \( P_\pi \), the Bayes rule error rate \( N(\hat{\theta}_{\pi}, \theta) \sim \text{Bin}(n_0 \pi_1) \) where \( \pi_1 = P_\pi \{ y_i \theta_1 < 0 \} = P \{ \delta_0 + \varepsilon \zeta < 0 \} = \Phi(-\delta_0 / \varepsilon) \geq 2 \pi_0 \). We recall the Cramér-Chernoff large deviations principle\(^5\): if \( \pi_1 > \pi_0 \), then

\[
P \{ \text{Bin}(n_0 \pi_1) < n_0 \pi_0 \} \leq e^{-n_0 D(\pi_0, \pi_1)}
\]

where \( D(\pi_0, \pi_1) = K(\text{Be}(\pi_0), \text{Be}(\pi_1)) = \pi_0 \log(\pi_0 / \pi_1) + \pi_1 \log(\pi_1 / \pi_0) \). Here \( D(\pi_0, \pi_1) \) denotes the Kullback-Leibler divergence between two Bernoulli distributions, and \( \pi_i = 1 - \pi_i \). Noting also\(^6\) that \( D(\pi_0, \pi_1) \geq 2(\pi_1 - \pi_0)^2 \), we conclude that

\[
1 - S(n_0 \pi_0) \leq e^{-n_0 D(\pi_0, \pi_1)} \leq e^{-2n_0 \pi_0^2}
\]

and since \( \Omega(\varepsilon) \geq n_0 \delta_0^2 \), this establishes (12.6).

\textbf{Proof in Sparse Case.} [Needle in haystack reference?] Consider a prior \( \pi \) on \( \Theta_{n,0}(\delta_0, 1) \) in which \( \theta = \delta_0 e_i \) and \( I \) is uniformly distributed on \( \{1, \ldots, n\} \). The posterior distribution of \( I \) is

\[
P(I = i | y) = \frac{\phi(y - \delta_0 e_i)}{\sum_{k=1}^n \phi(y - \delta_0 e_k)} = \frac{e^{\delta_0 y_i}}{\sum_k e^{\delta_0 y_k}}.
\]
The maximum a posteriori estimator \( \hat{\theta}_n = \delta_0 \bar{e}_I \) where \( \bar{I} = \text{argmax}_i P(I = i|y) = \text{argmax}_i y_i \). As at (12.8),

\[
\sup_{\theta \in \Theta} P\{\|\hat{\theta} - \theta\| \geq \delta_0 / 2\} \geq P_\pi\{\|\hat{\theta} - \theta\| \geq \delta_0 / 2\} = P_\pi\{\bar{I} \neq I\} = P_{\delta_0 \bar{e}_n}\{y_n \neq \max_{i=1, \ldots, n} y_k\}
\]

\[
= P\{e^{-1} \delta \bar{e}_n + z_n < \max_{i=1, \ldots, n-1} z_i\} \to 1,
\]

since \( \max_{i=1, \ldots, n-1} z_i = \sqrt{2 \log(n - 1)} + o_p(1) \). [More detailed bounds?]

\[
\square
\]

12.4. Further Details

1. Note. The theory can be developed without the convexity assumption in (i) and weakening the error norm in (ii) to a quasi-norm: i.e. there exists a constant \( \rho \) such that \( \|\theta + \xi\| \leq \|\theta\| + \|\xi\| \rho \). (see Donoho et al. (1995) and also DeVore & Lorentz (1993)[Chapter 2, Theorem 1.1]).

2. Remark. The assumption that \( \Theta \) is convex can be removed by using the original definition of the modulus employed by Donoho et al. (1995):

\[
\Omega(\delta) = \sup_{(\theta_0, \theta_1) \in \Theta \times \Theta} \{\|\theta_0 - \theta_1\| : \|\theta_0 - \theta_1\|_{\infty} \leq \delta\}.
\]

This definition makes the reason for the name “modulus of continuity” clearer: we measure the continuity of the norm \( \cdot \) with respect to the uniform norm \( \cdot \| \cdot \|_{\infty} \).

3. If the supremum in (12.3) is not attained, the argument above can be repeated for an approximating sequence.

4. Placeholder for proof of Gaussian bound.

5. Here is a proof of the Cramér-Chernoff result, using a standard change of measure argument. Let \( P_\pi \) denote a binomial distribution \( \text{Bin}(n_0, \pi) \), and let \( B \) denote the corresponding random variable. The likelihood ratio

\[
\frac{dP_{\pi_0}}{dP_{\pi_1}} = (\pi_0 / \pi_1)^B (\pi_1 / \pi_0)^{n_0 - B}.
\]

Defining \( \lambda = \log \pi_0 / \pi_1 \) and \( \bar{\lambda} = \log \pi_0 / \bar{\pi}_1 \), rewrite the log likelihood ratio as

\[
L = \log \frac{dP_{\pi_0}}{dP_{\pi_1}} = (\lambda - \bar{\lambda})B + n_0 \bar{\lambda}.
\]

Since \( \pi_0 < \pi_1 \) implies \( \lambda < \bar{\lambda} \), it follows that \( \{B \leq E_{\pi_0} B\} = \{L \geq E_{\pi_0} L\} \), while

\[
E_{\pi_0} L = n_0 D(\pi_0, \pi_1) = n_0 (\pi_0 \lambda + \bar{\pi}_0 \bar{\lambda})
\]

Consequently, using Markov’s inequality along with \( E_{\pi_1} e^L = 1 \), we have

\[
P_{\pi_1}\{B \leq n \pi_0\} = P_{\pi_1}\{e^L \geq e^{n_0 D}\} \leq e^{-n_0 D} E_{\pi_1} e^L = e^{-n_0 D}.
\]

167
6.

\[ D(\pi_0, \pi_1) = \pi_0 \log \frac{\pi_0}{\pi_1} + (1 - \pi_0) \log \frac{1 - \pi_0}{1 - \pi_1} \]

\[ = \pi_0 \int_{\pi_0}^{\pi_1} \frac{-du}{u} + (1 - \pi_0) \int_{\pi_0}^{\pi_1} \frac{du}{1 - u} \]

\[ = \int_{\pi_0}^{\pi_1} \frac{u - \pi_0}{u(1 - u)} \, du \geq 4 \int_{\pi_0}^{\pi_1} (u - \pi_0) \, du = 2(\pi_1 - \pi_0)^2. \]
CHAPTER 13

Model Selection, Penalization and Oracle Inequalities

Our ultimate goal is to obtain sharper bounds on rates of convergence - in fact exactly optimal rates, rather than with spurious log terms. However, this is a situation where the tools introduced are perhaps of independent interest. These include model selection via penalized least squares, where the penalty function is not $\ell_2$ or even $\ell_1$ but instead a function of the number of terms in the model. We will call such things complexity penalties.

Many of the arguments work for general (i.e. non-orthogonal) linear models. While we will not ultimately use this extra generality in this book, there are important applications (E.A.D.) and the model is of such importance that it seems reasonable to present part of the theory in this setting.

While it is natural to start with penalties proportional to the number of terms in the model, it will turn out that for our later results on exact rates, it will be necessary to consider a larger class of \(2k \log(p/k)\) penalties, in which, roughly speaking, the penalty to enter the \(k^{th}\) variable is a function that decreases with \(k\) approximately like \(2\log(p/k)\).

We will be looking essentially at "all subsets" versions of the model selection problem. If there are \(p\) variables, then there are \(\binom{p}{k}\) distinct submodels with \(k\) variables, and this grows very quickly with \(k\). In order to control the resulting model explosion, good exponential probability inequalities for the tails of chi-square distributions are needed. We will derive these as a consequence of a powerful concentration inequality for Gaussian measures in \(\mathbb{R}^n\). We give a separate exposition of this result, as it is finding increasing application in statistics.

13.1. A Gaussian concentration inequality

The Lipschitz norm of a function \(f : \mathbb{R}^n \to \mathbb{R}\) is
\[
\|f\|_{\text{Lip}} = \sup\{ |f(x) - f(y)|/\|x - y\| \}
\]
where \(\|x\|\) is the usual Euclidean norm on \(\mathbb{R}^n\).

13.1. Proposition. If \(Z \sim N_n(0, I)\), and \(f : \mathbb{R}^n \to \mathbb{R}\) is Lipschitz, then
\[
P\{f(Z) \geq Ef(Z) + t\} \leq e^{-t^2/(2\|\|f\|_{\text{Lip}}^2)}.
\]
Thus the tails of the distribution of a Lipschitz function of a Gaussian vector are sub Gaussian. Some statistically relevant examples of Lipschitz functions include

(i) order statistics. If \( z^{(1)} \geq z^{(2)} \geq \cdots \geq z^{(n)} \) denote the order statistics of a data vector \( z \), then \( f(z) = z^{(k)} \) has Lipschitz constant \( \|f\|_{\text{lip}} = 1 \).

(ii) ordered eigenvalues of symmetric matrices. Let \( A \) be an \( n \times n \) symmetric matrix with eigenvalues \( \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A) \). If \( E \) is also symmetric, then (e.g. (Gohub & Van Loan 1996, p. 56 and 396))

\[
|\lambda_k(A + E) - \lambda_k(A)| \leq \|E\|_F,
\]

where \( \|E\|_F^2 = \sum_{i,j} e_{i,j}^2 \) denotes the square of the Frobenius norm, which is the Euclidean norm on \( n \times n \) matrices,

(iii) orthogonal projections. If \( S \) is a linear subspace of \( \mathbb{R}^n \), then \( f(z) = \|P_S z\| \) has Lipschitz constant 1. This is the example we use below. If \( \dim S = k \), then \( \|P_S z\|^2 = \frac{D}{\lambda_{(k)}^2} \) and so

\[
E\|P_S z\| \leq \{E\|P_S z\|^2\}^{1/2} = \sqrt{k}
\]

and so the inequality implies

\[
P\{ \|P_S z\| \geq \sqrt{k} + t \} \leq e^{-t^2/2}.
\]

Note that the dimension \( k \) plays a very weak role in the inequality, which is sometimes said to be “infinite-dimensional”. The phrase “concentration of measure” refers at least in part to the fact that the distribution of a Lipschitz(1) function of \( n \) variables is concentrated about its mean, in the sense that the tails are no heavier than those of a univariate standard Gaussian, regardless of the value of \( n! \).

### 13.2. All subsets regression and complexity penalized least squares

We begin with the usual form of the general linear model with Gaussian errors:

\[
y = X \beta + \epsilon z = \mu + \epsilon z, \quad z \sim N_n(0, I).
\]

There are \( n \) observations \( y \) and \( p \) unknown parameters \( \beta \), connected by an \( n \times p \) design matrix \( X \) with columns

\[X = [x_1, \ldots, x_p].\]

There is no restriction on \( p \); indeed, we particularly wish to allow for situations in which \( p \gg n \). We will assume that the noise level \( \epsilon \) is known.

**Example:** Overcomplete dictionaries. Here is a brief indication of why one might wish to take \( p \gg n \). Consider estimation of \( f \) in the continuous Gaussian

\footnote{In fact, sharper bounds for the tail of \( \chi^2 \) random variables are available (Laurent & Massart 1998), Johnstone (2001), Birgé & Massart (2001), [CHECK1]), but this bound will suffice for our purposes and serves as an illustration of the power of the general inequality (13.1).}
white noise model (REF) \( dY(t) = f(t)\, dt + \varepsilon dW(t) \), and suppose that the observed data are inner products of \( Y \) with \( n \) orthonormal functions \( \psi_1, \ldots, \psi_n \). Thus
\[
y_i = \langle f, \psi_i \rangle + \varepsilon z_i, \quad i = 1, \ldots, n.
\]

Now consider the possibility of approximating \( f \) by elements from a dictionary \( \mathcal{D} = \{ \varphi_1, \varphi_2, \ldots, \varphi_p \} \). The hope is that by making \( \mathcal{D} \) sufficiently rich, one might be able to represent \( f \) well by a linear combination of a very few elements of \( \mathcal{D} \). This idea has been advanced by a number of authors ADD REFERENCES. As a simple illustration, the \( \psi_i \) might be sinusoids at the first \( n \) frequencies, while the dictionary elements might allow a much finer sampling of frequencies
\[
\varphi_k(t) = \sin(2\pi kt/p), \quad k = 1, \ldots, p = n^\beta \gg n.
\]
with \( p = n^\beta \) for some \( \beta > 1 \). If there is a single dominant frequency in the data, it is possible that it will be essentially captured by an element of the dictionary even if it does not complete an integer number of cycles in the sampling interval.

If we suppose that \( f \) has the form \( f = \sum_{j=1}^p \beta_j \phi_j \), then these observation equation become an instance of the general linear model (13.2) with
\[
X_{ij} = \langle \psi_i, \phi_j \rangle.
\]
Again, the hope is that one can find an estimate \( \hat{\beta} \) for which only a small number of components \( \beta_j \neq 0 \).

All subsets regression. To each subset \( J \subset \{1, \ldots, p\} \) of cardinality \( n_J = |J| \) corresponds a regression model which fits only the variables \( x_j \) for \( j \in J \). The possible fitted vectors \( \mu \) that could arise from these variables lie in the model space
\[
S_J = \text{span}\{x_j : j \in J\}.
\]
The dimension of \( S_J \) is at most \( n_J \), and could be less in the case of collinearity.

Let \( P_J \) denote orthogonal projection onto \( S_J \): the least squares estimator \( \hat{\mu}_J \) of \( \mu \) is given by \( \hat{\mu}_J = P_J y \). The issue in all subsets regression consists in deciding how to select a subset \( J \) on the basis of data \( y \): the resulting estimate of \( \mu \) is then \( \hat{\mu} = P_J y \).

Mean squared error properties can be used to motivate all subsets regression. We will use a predictive risk\(^2\) criterion to judge an estimator \( \hat{\beta} \) through the fit \( \hat{\beta} = X \hat{\beta} \) that it generates:
\[
E \|X \hat{\beta} - X \beta\|^2 = E \|\hat{\mu} - \mu\|^2.
\]

The mean of a projection estimator \( \hat{\mu}_J \) is just the projection of \( \mu \), namely \( E \hat{\mu}_J = P_J \mu \), while its variance is \( \varepsilon^2 \text{tr} P_J = \varepsilon^2 \dim S_J \). From the variance-bias
\[\text{Why the name “predictive risk”? Imagine that new data will be taken from the same design as used to generate the original observations \( y \) and estimator \( \hat{\beta} \): \( y^* = X \beta + \varepsilon z^* \). A natural prediction of \( y^* \) is \( X \hat{\beta} \), and its mean squared error, averaging over the distributions of both \( z \) and \( z^* \), is}
\[
E \|y^* - X \hat{\beta}\|^2 = E \|X \beta - X \hat{\beta}\|^2 + n\varepsilon^2,
\]
so that the mean squared error of prediction equals \( E \|\hat{\mu} - \mu\|^2 \), up to an additive factor that doesn’t depend on the model chosen.\]

171
decomposition of MSE,
\[ E\|\hat{\mu}_J - \mu\|^2 = \|P_J \mu - \mu\|^2 + \epsilon^2 \dim S_J. \]

A saturated model arises from any subset with \( \dim S_J = n \), so that \( \hat{\mu}_J = y \) "interpolates the data". In this case the MSE is just the unrestricted minimax risk for \( \mathbb{R}^n \):
\[ E\|\hat{\mu} - \mu\|^2 = n\epsilon^2. \]

Comparing the last two displays, we see that if \( \mu \) lies close to a low rank subspace \( \mu = \sum_{j \in J} \beta_j x_j \) for \(|J|\) small—then \( \hat{\mu}_J \) offers substantial risk savings over a saturated model. Thus, it seems that one would wish to expand the dictionary \( D \) as much as possible to increase the possibilities for sparse representation. Against this must be set the dangers inherent in fitting over-parametrized models — principally overfitting of the data. Penalized least squares estimators are designed specifically to address this tradeoff.

This discussion leads to a natural generalization of the notion of ideal risk introduced in Chapter ???. For each mean vector \( \mu \), there will be an optimal model subset \( J = J(\mu) \) which attains the ideal risk
\[ \mathcal{R}(\mu, \epsilon) = \min \|\mu - P_J \mu\|^2 + \epsilon^2 \dim S_J. \]

Of course, this choice \( J(\mu) \) is not available to the statistician, since \( \mu \) is unknown. The challenge, taken up below, is to see to what extent penalized least squares estimators can "mimick" ideal risk, in a fashion analogous to the mimicking achieved by threshold estimators in the orthogonal setting.

**Complexity penalized least squares** The residual sum of squares (RSS) after fitting model \( J \) is
\[ \|y - \hat{\mu}_J\|^2 = \|y - P_J y\|^2, \]
and clearly decreases as the model \( J \) increases. To discourage simply using a saturated model, we introduce a penalty on the size of the model, \( \text{pen}(n, J) \), and define a complexity criterion
\[ (13.3) \quad C(J, y) = \|y - \hat{\mu}_J\|^2 + \epsilon^2 \text{pen}(n, J). \]

The complexity penalized RSS estimate \( \hat{\mu}_\text{pen} \) is then given by orthogonal projection onto the subset that minimizes the penalized criterion:
\[ (13.4) \quad \hat{J}_\text{pen} = \arg\min_J C(J, y) \]
\[ \hat{\mu}_\text{pen} = P_{\hat{J}_\text{pen}} y. \]

The simplest penalty function is simply proportional to the number of variables in the model:
\[ (13.5) \quad \text{pen}_0(k) = \lambda_p^2 k, \]
where we will take \( \lambda_p^2 \) to be roughly of order 2 log \( p \). [The well known AIC criterion would set \( \lambda_p^2 = 2 \): this is effective for selection among a nested
sequence of models, but is known to overfit in all-subsets settings. [mention BIC $\lambda_p^2 = \log p$?]

For this particular case, we describe the kind of oracle inequality to be proved in this chapter. First, note that for $\text{pen}_0(k)$, minimal complexity and ideal risk are related:

$$\min_j C(J, \mu) = \min_j \left[ \|\mu - P_j \mu\|^2 + \epsilon^2 \text{pen}_0(n, J) \right]$$

$$\leq \lambda_p^2 \min_j \left[ \|\mu - P_j \mu\|^2 + \epsilon^2 n_j \right] = \lambda_p^2 \mathcal{R}(\mu, \epsilon).$$

Let $\lambda_p = \zeta (1 + \sqrt{2 \log p})$ for $\zeta > 1$ and $A(\zeta) = (1 - \zeta^{-1})^{-1}$. Then for penalty function (13.5) and arbitrary $\mu$,

$$E[\|\hat{\mu}_{\text{pen}} - \mu\|^2] \leq A(\zeta) \lambda_p^2 C \epsilon^2 + \mathcal{R}(\mu, \epsilon).$$

Thus, the complexity penalized RSS estimator, for non-orthogonal and possibly over-complete dictionaries, comes within a factor of order $2 \log p$ of the ideal risk.

Remark. Another possibility is to use penalty functions monotone in the rank of the model, $\text{pen}(\dim J)$. However, when $k \to \text{pen}(k)$ is strictly monotone, this will yield the same models as minimizing (13.3), since a collinear model will always be rejected in favor of a sub-model with the same span.

### 13.3. Oracle inequalities for $2k \log(p/k)$ penalties

Class $\mathcal{P}$. Assume that $k \to \text{pen}(k)$ is strictly increasing, and has the specific form

$$\text{pen}(k) = \zeta^2 k (1 + \sqrt{2L_{p,k}})^2 \quad (\zeta > 1)$$

with $k \to L_{p,k}$ non-increasing and

$$L_{p,k} \geq \log(p/k) + \gamma_k$$

for some sequence $\gamma_k$ not depending on $p$ and satisfying $\gamma_k \geq \gamma > 1$ for all large $k$.

The class $\mathcal{P}$ is broad enough to include penalties proportional to model size, c.f. (13.5), with $L_{p,k} \equiv \lambda_p \geq \log p$. However, the possibility that $k \to L_{p,k}$ be strictly decreasing is critical for the later application of the oracle inequalities to derive exact rates of convergence. Since the dominant terms in the two preceding displays are $2k \log(p/k)$, we will loosely refer to (P) as a class of “$2k \log(p/k)$ penalties”.

### 13.2. Theorem. Let $\hat{\mu}$ be a penalized least squares estimate of the form (13.3)- (13.4) for a penalty of class $\mathcal{P}$ defined above. Then, for all $\mu$,

$$E[\|\hat{\mu} - \mu\|^2] \leq (1 - \zeta^{-1})^{-1} \left[ \zeta C, L_{p,k} \epsilon^2 + \min J \hat{C}(J, \mu) \right].$$

173
We will see that $\min_{J} C(J, \mu)$ is an extension of the notion of ideal risk, and describes the “intrinsic accuracy” of approximation of $\mu$ by models $\{S_j : j \subset \{1, \ldots, p\}\}$.

Our key examples will have the form $L_{p,k} \approx \log(p/k) + \gamma$, for $\gamma > 1$. For example, in the orthogonal case, the False Discovery Rate thresholds satisfy

$$t_{n,k} \sim \sqrt{2\log(2n/kq)}$$

corresponding roughly to $L_{p,k} = \log(n/k) + \gamma$ with $\gamma = \log(2/q) > 1$ for $q < 2/e$.

“Threshold” interpretation. In the orthogonal setting ($n = p$), suppose that

$$\text{pen}(k) = \sum_{j=1}^{k} t_{n,j}^2,$$

so that $\hat{\mu}_{\text{pen}}$ becomes thresholding:

$$\hat{\mu}_{\text{pen},j} = \begin{cases} y_j & |y_j| \geq t_{n,k} \\ 0 & \text{otherwise} \end{cases}$$

where

$$\hat{k} = \arg\min_{0 \leq k \leq n} \sum_{j=1}^{k} y_{(j)}^2 + \epsilon^2 \sum_{j=1}^{k} t_{p,j}^2.$$ 

Typically $\sum_{j=1}^{k} t_{n,j}^2 \sim k t_{n,k}^2$, at least for $k = o(n)$, and so the thresholds covered by the Theorem have the form

$$t_{n,k} \approx \zeta(1 + \sqrt{2L_{p,k}}) \geq \zeta(1 + \sqrt{2\log(p/k) + 2\gamma n}).$$

The fact that $\zeta > 1$ (rather than $= 1$) and the extra “1” are needed for the proof, but do make the thresholds somewhat larger than is desirable in practice.

The idea to use penalties of the general form $2\epsilon^2 k \log(n/k)$ arose among several authors more or less simultaneously:

- Foster & Stine (1997) $\text{pen}(k) = \epsilon^2 \sum_{j=1}^{k} 2\log(n/j)$ via information theory.
- George & Foster (2000) Empirical Bayes approach. $[\mu_1 \ldots \mu_p, \epsilon_d, (1 - w)\delta_0 + wN(0, C)]$ followed by estimation of $(w, C)$. They argue that this approach penalizes the $k^{th}$ variable by about $2\epsilon^2 \log(((n + 1)/k) - 1)$.
- The covariance inflation criterion of Tibshirani & Knight (1999) in the orthogonal case leads to $\text{pen}(k) = 2\epsilon^2 \sum_{j=1}^{k} 2\log(n/j)$.
- FDR - discussed above (?).
- Birgé & Massart (2001) contains a systematic study of complexity penalized model selection from the specific viewpoint of obtaining non-asymptotic bounds, using a penalty class similar to, but more general than that used here.
13.4. Proof of Oracle Inequality

Penalized complexity and complexity functionals The definition of the complexity penalized RSS estimator uses the minimum over submodels $J$ of

$$C(J, y) = \|y - P_J y\|^2 + \text{pen}(n_J).$$

It will be helpful to introduce a related complexity functional $K(\mu, y)$, defined for all $\mu \in \mathbb{R}^n$. First, a definition. Given $\mu$, the minimal dimension of a model containing $\mu$ is

$$N(\mu) = \text{inf}\{n_J : \mu \in S_J\}.$$

Now define

$$K(\mu, y) = \|y - \mu\|^2 + \text{pen}(N(\mu)).$$

The criteria $C$ and $K$ yield the same estimators:

**13.3. Lemma.** Suppose that $\text{pen}(k)$ is strictly increasing in $k$. Given data $y$, let $\hat{J}$ be a minimizer of $C(J, y)$ and set $\hat{\mu} = P_{\hat{J}} y$. Then $\hat{\mu}$ minimizes $K(\mu, y)$ and

$$\inf_{\mu} K(\mu, y) = \min_{J} C(J, y).$$

**Proof.** We chase definitions resolutely. First, fix $\mu$ and let $J(\mu) = \text{argmin}\{n_J : \mu \in S_J\}$. Consequently $\mu \in S_{J(\mu)}$ and hence both

$$\|y - \mu\|^2 \geq \|y - P_{J(\mu)} y\|^2$$

and $N(\mu) = n_{J(\mu)}$, so that for every $\mu$,

$$K(\mu, y) \geq C(J(\mu), y) \geq \min_{J} C(J, y).$$

Now turn to $\hat{J} = \text{argmin} C(J, y)$ and $\hat{\mu} = P_{\hat{J}} y$. Let us show that $N(\hat{\mu}) = n_{\hat{J}}$. Indeed, suppose to the contrary that there were a subset $J$ with $n_J < n_{\hat{J}}$ and $\hat{\mu} \in S_J$. Then we would have both $\text{pen}(n_J) < \text{pen}(n_{\hat{J}})$ and

$$\|y - P_{\hat{J}} y\|^2 \leq \|y - \hat{\mu}\|^2 = \|y - P_{\hat{J}} y\|^2.$$

But this means that $C(J, y) < C(J, y)$, in contradiction to the definition of $\hat{J}$. Hence $\text{pen}(N(\hat{\mu})) = \text{pen}(n_{\hat{J}})$ and so

$$K(\hat{\mu}, y) = \|y - P_{\hat{J}} y\|^2 + \text{pen}(n_{\hat{J}}) = C(\hat{J}, y) = \min_{J} C(J, y).$$

Combining the first and third displays, we obtain the result. \qed

It is now clear that Theorem 13.2 is a consequence of

**13.4. Theorem.** Let $K(\mu, y) = \|y - \mu\|^2 + \epsilon^2 \text{pen}(N(\mu))$ and assume that $\text{pen}(k)$ satisfies assumptions $\mathcal{P}$. If $\hat{\mu}_P = \text{argmin}_\mu K(\hat{\mu}, y)$ and $K_0(\mu) = \inf_\mu K(\hat{\mu}, \mu)$, then

$$EK(\hat{\mu}_P, \mu) \leq (1 - \zeta^{-1})^{-1} [\zeta C_L \epsilon^2 + K_0(\mu)].$$

175
Proof. 1°. Reduction to $\epsilon = 1$. Showing the $\epsilon$-dependence explicitly for now in $K_\epsilon(\mu, y)$, we see immediately that

$$K_\epsilon(\mu, y) = \epsilon^2 K_1(\mu/\epsilon, y/\epsilon),$$

and so it suffices to establish the inequality when $\epsilon = 1$.

2°. A basic inequality. One reason to introduce the functional $K(\mu, y)$ is a useful inequality which we now derive. Indeed, use of $y = \mu + \varepsilon z$ and the expansion $\|y - \tilde{\mu}\|^2 = \|\varepsilon z\|^2 + 2\langle \mu - \tilde{\mu}, \varepsilon z \rangle + \|\mu - \tilde{\mu}\|^2$ lead to the identity

$$K(\tilde{\mu}, y) = \|\varepsilon z\|^2 + 2\langle \mu - \tilde{\mu}, \varepsilon z \rangle + K(\tilde{\mu}, \mu).$$

Use this identity for both $K(\tilde{\mu}, y)$ and $K(\hat{\mu}, y)$. Since $K(\tilde{\mu}, y) \leq K(\hat{\mu}, y)$ for all $\tilde{\mu}$ by definition, we obtain by subtraction the basic inequality

$$K(\hat{\mu}, \mu) \leq K(\tilde{\mu}, \mu) + 2\langle \mu - \tilde{\mu}, \varepsilon z \rangle.$$  

The left side exceeds the quadratic loss $\|\mu - \tilde{\mu}\|^2$, while on the right side, $\tilde{\mu}$ can be chosen to minimize the theoretical complexity $K_0(\mu) = \inf_{\tilde{\mu}} K(\tilde{\mu}, \mu)$. If a minimizing value is $\mu_0$, the chief task in obtaining an upper bound for $E\|\mu - \tilde{\mu}\|^2$ in terms of $K_0(\mu)$ becomes that of bounding the error term $\langle \mu - \mu_0, \varepsilon z \rangle$.

We return to the basic inequality (13.6). Inserting $\mu_0$ for $\tilde{\mu}$ and noting that $E\langle \mu - \mu_0, z \rangle = 0$, we have

$$E K(\hat{\mu}, \mu) \leq K(\mu_0, \mu) + 2E\langle \mu - \mu_0, z \rangle.$$  

Our goal will be to derive an “a priori” bound for $E\langle \mu - \tilde{\mu}, z \rangle$ in terms of $\zeta^{-2} E K(\hat{\mu}, \mu)$ and an error term. This will require control on deviations of $\|P_Sz\|^2$ as $S$ ranges over the possible subset models.

3°. Rejection of individual models. Given a model $S$, we will say that it is rejected if

$$\|P_Sz\|^2 \geq \zeta^{-2} \text{pen}(\text{dim} S).$$

In a ‘null hypothesis’ situation, this will constitute a type I error, and so the penalty should be large enough so as to make this type I error probability small, simultaneously over all sufficiently large models $S$.

With the choice pen$(k) = \zeta^2 k(1 + \sqrt{2L_{p,k}})^2$, we find that when $\text{dim} S = k$, (13.8) is equivalent to

$$\|P_Sz\| \geq k + \sqrt{2kL_{p,k}}.$$  

Now $E\|P_Sz\|^2 = E\chi^2(k) = k$, and so $E\|P_Sz\| \leq \sqrt{k}$. Consequently, from the concentration inequality (13.1),

$$P\{\text{reject model } S\} \leq e^{-kL_{p,k}}.$$  

4°. A Priori Bound. Fix $\mu$, and for each model $J$, let $S_{J,\mu} = \text{span} \{\mu, S_J\}$. Denote the full augmented collection of models by

$$\mathcal{M}_\mu = \{S_J, S_{J,\mu} : J \subset 1:p\}.$$
We need to track the size of the largest rejection model. So, let
\[
\hat{N} = \begin{cases} 
\min\{N: \text{ all models } S \in \mathcal{M}_N \text{ with } \dim S \geq n \text{ are "accepted"} \} \\
\infty \quad \text{if } \|z\|^2 \geq \zeta^{-2} \text{pen}(n),
\end{cases}
\]
since in the latter case, even the saturated model is not accepted. Correspondingly, set
\[
\hat{P}(z, \mu) = \begin{cases} 
\text{pen}(\hat{N}) \quad \text{if } \hat{N} \leq n \\
\|\zeta z\|^2 \quad \text{if } \hat{N} = \infty.
\end{cases}
\]
We will establish that
\[
(13.10) \quad 2\zeta(z, \mu' - \mu) \leq \hat{P}(z, \mu) + K(\mu', \mu).
\]
Assuming for now, both (13.10) and that $E\hat{P} < \infty$, we have, on substitution into (13.7),
\[
EK(\hat{\mu}, \mu) \leq \mathbb{K}(\mu) + \zeta^{-1} E\hat{P} + \zeta^{-1} \mathbb{K}(\hat{\mu}, \mu).
\]
Moving the unknown to the left side and then ignoring the penalty term in $K(\hat{\mu}, \mu) = \|\hat{\mu} - \mu\|^2 + \text{pen}(\hat{\mu})$, we get
\[
E\|\hat{\mu} - \mu\|^2 \leq (1 - \zeta^{-1})^{-1} \left[\zeta^{-1} E\hat{P} + \mathbb{K}(\mu)\right].
\]

5°. Proof of (13.10). As before, let $J(\mu') = \arg\min\{n_J : \mu' \in S_j\}$ be the minimal model containing $\mu'$, and set
\[
S' = S_{J(\mu'), \mu} = \text{span}\{\mu, S_{J(\mu')}\}.
\]
Clearly $\mu' - \mu \in S'$ and so
\[
2\zeta(z, \mu' - \mu) \leq 2\zeta \|P_{S'} z\| \|\mu' - \mu\|
\leq \|\zeta P_{S'} z\|^2 + \|\mu' - \mu\|^2.
\]
If $\hat{N} = \infty$, the right side is bounded by $\|\zeta z\|^2 + K(\mu', \mu)$, so the Claim is straightforward. So we now suppose that $\hat{N} \leq n$, and propose to show that
\[
\|\zeta P_{S'} z\|^2 \leq \text{pen}(\hat{N}) + \text{pen}(N(\mu')),
\]
which suffices for (13.10).

Two cases arise. If $\dim S' \leq \hat{N}$, choose $S'' \supset S'$ with $\dim S'' = \hat{N}$. By definition of $\hat{N}$
\[
\|\zeta P_{S''} z\|^2 \leq \|\zeta P_{S'} z\|^2 \leq \text{pen}(\hat{N}).
\]
In the second case, $\dim S' > \hat{N}$. Since $N(\mu') \leq \dim S' \leq N(\mu') + 1$, we know also that $N(\mu') \geq \hat{N}$. We remark in addition that for any $r$, $\text{pen}(r+1) \leq \text{pen}(r) + \text{pen}(r)/r$ since $\text{pen}(r)/r$ is decreasing. Since model
\( S' \) is accepted, we therefore find
\[
\|z_{P_{S'}z}\|^2 \leq \text{pen}(\dim S') \\
\leq \text{pen}(N(\mu')) + \text{pen}(N(\mu'))/N(\mu') \\
\leq \text{pen}(N(\mu')) + \text{pen}(\hat{N})/\hat{N} \\
\leq \text{pen}(N(\mu')) + \text{pen}(\hat{N}).
\]

6°. Bounding \( E\hat{P} \). By definition, we have
\[
E\hat{P}(z,\mu) = E\{\text{pen}(\hat{N}), \hat{N} \leq n\} + E\{\|z\|^2, \|z\|^2 \geq \text{pen}(n)\},
\]
and, using monotonicity of \( k \rightarrow L_{p,k} \) and setting \( \lambda_{p,1} = \zeta(1 + \sqrt{2L_{p,1}}) \),
\[
E\{\text{pen}(\hat{N}), \hat{N} \leq n\} \leq \lambda_{p,1}^2 \{P(\hat{N} = 1) + \sum_{k=1}^{n-1} (k+1)P(\hat{N} = k+1)\}.
\]
If \( \hat{N} = k+1 \), there is some \( k \)-dimensional model in \( \mathcal{M}_\mu \) that is rejected. For each individual model, the rejection probability is bounded by (13.9). We turn to bounding the number of such models.

Let \( S_k \) be the set of models \( S_J \) with dimension \( k \)—clearly \( |\{S_k\}| \leq \binom{p}{k} \). A model in \( \mathcal{M}_\mu \) of dimension \( k \) is either from \( S_k \) or is of the form \( S_J \oplus \mu \), for some \( S_J \in S_{k-1} \) which does not already contain \( \mu \). Hence
\[
|\{S \in \mathcal{M}_\mu : \dim S = k\}| \leq |S_k| + |S_{k-1}|
\leq \binom{p}{k} + \binom{p}{k-1} \leq \frac{2^{\frac{k}{k!}}}{k!},
\]
since \( k < n \leq p \). Putting this together with (13.9) gives
\[
P(\hat{N} = k+1) \leq \frac{2^{\frac{k}{k!}}}{k!} \exp\{\frac{1}{2}\sqrt{2k}e^{-k}\}
\leq \frac{1}{2}\sqrt{2\pi k} \exp\{-kL_{p,k} - \log(p/k) - 1\},
\]
where we have used Stirling’s inequality: \( k! \geq \sqrt{2\pi k}e^{-k}k^{k+1/2} \).

Now use the assumption \( L_{p,k} \geq \log(p/k) + \gamma \) for some \( \gamma > 1 \). This yields
\[
E[\hat{N}, \hat{N} \leq n] \leq 1 + \sum_{k=1}^{n-1} \frac{2(k+1)}{\sqrt{2\pi k}} e^{-k(\gamma-1)} \leq c(\gamma).
\]

We must finally deal with the second term in (13.11). The Cauchy-Schwarz inequality shows it to be bounded by
\[
\{E\|z\|^4\}^{1/2} P\{\|z\| > \sqrt{n} + \sqrt{2nL_{p,n}}\}^{1/2}.
\]
Since \( \|z\|^2 \sim \chi_n^2 \), we have \( E\|z\|^4 = 2n + n^2 \leq (n+1)^2 \). Applying (13.1) to the second term, we obtain as upper bound
\[
\zeta^2(n+1)e^{-\frac{1}{2}nL_{p,n}} \leq \zeta^2(n+1)e^{-n/2},
\]

178
using again the assumption on $L_{p,k}$.

Finally assembling the bounds, we have, so long as $L_{p,1} \geq 1$,

$$EP(z, \mu) \leq \zeta^2(1 + \sqrt{2L_{p,1}})^2\epsilon(\gamma) + \zeta^2(n + 1)e^{-n/2}$$

$$\leq \zeta^2(1 + \sqrt{2L_{p,1}})^2\epsilon(\gamma) \leq \zeta^2\epsilon'(\gamma)L_{p,1}.$$ 

\[\square\]

13.5. Aside: Stepwise methods vs. complexity penalization.

Stepwise model selection methods have long been used as heuristic tools for model selection. In this aside, we explain a connection between such methods and a class of penalties for penalized least squares.

The basic idea with stepwise methods is to use a test statistic—in application, often an $F$-test—and a threshold to decide whether to add or delete a variable from the current fitted model. Let $\hat{J}_k$ denote the best submodel of size $k$:

$$\hat{J}_k = \arg\max_k \{\|P_j y\|^2 : n_j = k\},$$

and denote the resulting best $k$-variable estimator by $Q_k y = P_{\hat{J}_k} y$. The mapping $y \to Q_k(y)$ is non-linear since the optimal set $\hat{J}_k(y)$ will in general vary with $y$.

In the forward stepwise approach, the model size is progressively increased until a threshold criterion suggests that no further benefit will accrue by continuing. Thus, define

$$k_G = \text{first } k \text{ s.t. } \|Q_{k+1}y\|^2 - \|Q_k y\|^2 \leq \epsilon^2 t_{p,k+1}^2.$$ 

Note that we allow the threshold to depend on $k$; in practice it is often constant, but we wish to allow $k \to t_{p,k}^2$ to be decreasing.

In contrast, the backward stepwise approach starts with a saturated model and gradually decreases model size until there appears to be no further advantage in going on. So, define

$$k_F = \text{last } k \text{ s.t. } \|Q_k y\|^2 - \|Q_{k+1} y\|^2 \geq \epsilon^2 t_{p,k}^2.$$ 

Remarks. 1. In the orthogonal case, $y_i = \mu_i + \epsilon_i, i = 1, \ldots, n$ with order statistics $|y|_{(1)} \geq |y|_{(2)} \geq \ldots \geq |y|_{(n)}$, we find that

$$\|Q_k y\|^2 = \sum_{j=1}^k |y|_{(j)}^2,$$

so that

$$k_F = \max\{k : |y|_{(k)} \geq \epsilon t_{p,k}\},$$

and that $k_F$ agrees with the FDR definition with $t_{p,k} = z(qk/2n)$. [Fuller reference??] In this case, it is critical to the method that the thresholds $k \to t_{p,k}$ be (slowly) decreasing.
2. In practice, for reasons of computational simplicity, the forward and backward stepwise algorithms are often “greedy”, i.e., they look for the best variable to add (or delete) without optimizing over all sets of size $k$.

The stepwise schemes are related to a penalized least squares estimator. Let

$$S(k) = \| y - Q_k y \|^2 + \sum_{j=1}^{\ell_k^2} t_{p,j}^2, \tag{13.15}$$

$$\hat{k}_2 = \arg \min_{0 \leq k \leq n} S(k).$$

Thus the associated penalty function is $\text{pen}(k) = \sum_{j=1}^{\ell_k^2} t_{p,j}^2$ and the corresponding estimator is given by (13.3) and (13.4). [Remark that $\text{pen}(k)$ satisfies assumptions (P) if $j \to t_{p,j}$ is decreasing with $\gamma t_{p,k} \geq \log(p/k) + \gamma_k$ etc. ??]

The optimal model size for pen$(k)$ is bracketed between the stepwise quantities.

13.5. Proposition. Let $\hat{k}_G, \hat{k}_F$ be the forward and backward stepwise variable numbers defined at (13.12) and (13.13) respectively, and let $\hat{k}_2$ be the global optimum model size for pen$(k)$ defined at (14.9). Then

$$\hat{k}_G \leq \hat{k}_2 \leq \hat{k}_F.$$

Proof. Since $\| y - Q_k y \|^2 = \| y \|^2 - \| Q_k y \|^2$,

$$S(k+1) - S(k) = \| Q_{k+1} y \|^2 - \| Q_k y \|^2 + \| Q_{k+1} y \|^2 - \| Q_k y \|^2 \geq 2 t_{p,k+1}^2.$$

Thus

$$S(k+1) \begin{cases} < & S(k) \text{ according as } \| Q_{k+1} y \|^2 - \| Q_k y \|^2 \begin{cases} > & 2 t_{p,k+1}^2. \end{cases} \end{cases}$$

Thus, if it were the case that $\hat{k}_2 > \hat{k}_F$, then necessarily $S(\hat{k}_2) > S(\hat{k}_2 - 1)$, which would contradict the definition of $\hat{k}_2$ as a global minimum of $S(k)$. Likewise, $\hat{k}_2 < \hat{k}_G$ is not possible, since it would imply that $S(\hat{k}_2 + 1) < S(\hat{k}_2)$. \hfill \Box

[Include diagram, notes 11/26/02 p.8 showing local and global minima.]

[Remark?: Experience with FDR in orthogonal case shows that often $\hat{k}_G = \hat{k}_2 = \hat{k}_F$, and that in sparse cases can bound $\hat{k}_F - \hat{k}_G$.]

13.6. A version for orthogonal regression

We now specialize to the $n$-dimensional white Gaussian sequence model:

$$y_i = \mu_i + e_i, \quad i = 1, \ldots, n, \quad z_i \sim i.i.d. N(0,1). \tag{13.16}$$

3Of course, this is the canonical form of the more general orthogonal regression setting $\mathbf{Y} = \mathbf{X}\beta + e\mathbf{Z}$.
The columns of the design matrix implicit in (13.16) are the unit co-
ordinate vectors $e_i$, consisting of zeros except for a 1 in the $i^{th}$ position. The
least squares estimator corresponding to a subset $J \subset \{1, \ldots, n\}$ is simply
given by co-ordinate projection:

$$
\hat{\mu}_{J,k}(y) = \begin{cases} 
y_k & k \in J \\
0 & k \notin J.
\end{cases}
$$

In the orthogonal setting

$$
N(\mu) = \# \{i : \mu_i \neq 0\}
$$

simply counts the number of non-zero co-ordinates. To minimize the em-
pirical complexity

$$
K(\mu, y) = \|y - \mu\|^2 + \epsilon^2 \text{pen}(N(\mu)),
$$

we observe that the best fitting model of dimension $k$ just picks off the
largest $k$ observations. Thus, if $y_{(1)} \geq y_{(2)} \geq \ldots \geq y_{(n)}$ denote the ordered
squared observations,

$$
\inf_{\mu} K(\mu, y) = \min_k \inf_{N(\mu)=k} \|y - \mu\|^2 + \epsilon^2 \text{pen}(k)
$$

$$
= \min_k \sum_{j>k} y_{(j)}^2 + \epsilon^2 \text{pen}(k).
$$

If $\hat{k}$ denotes the (data dependent!) minimizing value of $k$, then the com-
plexity penalized least squares estimate $\hat{\mu}_P$ is given by hard thresholding at
threshold $\hat{t} = \|y\|_{\hat{k}}$:

$$
\hat{\mu}_{P,\hat{k}}(y) = \begin{cases} 
y_j & |y_j| \geq \hat{t} \\
0 & \text{otherwise}.
\end{cases}
$$

To state the oracle inequality established in the last section (7), recall
the definition of theoretical complexity:

$$
K_*(\mu) = \inf_{\mu} \|\mu - \hat{\mu}\| + \epsilon^2 \text{pen}(N(\hat{\mu})),
$$

and the assumptions (P) on the penalty function, namely that

$$
\text{pen}(k) = \zeta^2 k(1 + \sqrt{2L_{n,k}})^2, \quad \zeta > 1,
$$

and (i) $k \to \text{pen}(k)$ is strictly increasing, (ii) $k \to L_{n,k}$ is nonincreasing, and
(iii) $L_{n,k} \geq \log(n/k) + \gamma_k$, with $\gamma_k \geq \gamma > 1$ for $k \geq k_0$.

Then the oracle inequality becomes

$$
E\|\hat{\mu}_P - \mu\|^2 \leq c_1 L_{n,1} \epsilon^2 + c_2 K_*(\mu),
$$

with $c_1 = c_1(\zeta, \gamma)$ and $c_2 = c_2(\zeta)$.

A simple bound for theoretical complexity

with $N$ dimensional response and $n$ dimensional parameter vector $\beta$ linked by an ortho-
gonal design matrix $X$ satisfying $X^TX = I_n$, and with the noise $Z \sim N(0, I)$. This reduces
to (13.16) after premultiplying by $X^T$ and setting $y = X^TY$, $\mu = \beta$ and $z = X^TZ$. 

181
13.6. **Lemma.** If \( \text{pen}(k) = k \lambda_k^2 \) with \( k \to \lambda_k \) non-increasing, then

\[
(13.18) \quad K_c(\mu) \leq \sum_{k=1}^{n} \mu_{(k)}^2 \land \lambda_k^2 \epsilon^2.
\]

This and the following lemma will be proved below. For now, look at the quantity

\[
\mathcal{R}_\lambda(\mu, \epsilon) = \sum_{k=1}^{n} \mu_{(k)}^2 \land \lambda_k^2 \epsilon^2.
\]

If \( \lambda_k \equiv \lambda \), this reduces to the ideal risk \( \mathcal{R}(\mu, \epsilon) \) studied at length earlier — and in fact, equality holds in (13.18) in this case. However, we are now more interested in cases in which \( \lambda_k \) is strictly decreasing, for example like \( k \to \zeta(1 + \sqrt{2 \log(n \beta/k)}) \). Although the rate of decrease is slow, we will see that it suffices to remove spurious logarithmic terms from rates of convergence.

We remark also that in typical cases, the inequality in (13.18) is sharp at the level of rates:

13.7. **Lemma.** Suppose that \( \lambda_k = \ell(k/n) \) for a function \( \ell(x) \), positive and decreasing in \( x \in [0, 1] \), that satisfies

\[
\lim_{x \to 0} x \ell(x) = 0, \quad \sup_{0 \leq x \leq 1} x \ell' (x) \leq c_1.
\]

Then

\[
\sum_{k=1}^{n} \mu_{(k)}^2 \land \lambda_k^2 \epsilon^2 \leq c_2 K_c(\mu), \quad c_2 \leq 1 + c_1 / \ell(1).
\]

This leads immediately to an important bound.

13.8. **Corollary.** If \( \hat{\mu}_P = \arg\min_{\mu} ||y - \mu||^2 + \epsilon^2 \text{pen}(N(\mu)) \), for \( \text{pen}(k) = k \lambda_k^2 \) and \( \lambda_k = \zeta(1 + \sqrt{2 \log(n \beta/k)}) \) satisfying assumptions (P), then

\[
E||\hat{\mu}_P - \mu||^2 \leq c_1 L_n, \epsilon^2 + c_2 \sum_{k} \mu_{(k)}^2 \land \lambda_k^2 \epsilon^2.
\]

[Include proofs, pp 5 & 6.]

13.9. **Lemma.** (a) Suppose that \( \{s_k\}_1^n \) and \( \{\gamma_k\}_1^n \) are positive, non-decreasing sequences. Then

\[
\min_{0 \leq k \leq n} [ks_k + \sum_{k+1}^{n} \gamma_n] \leq \sum_{k=1}^{n} s_k \land \gamma_k.
\]

(b) Conversely, suppose that \( s_k = \sigma(k/n) \) with \( \sigma(u) \) a positive, decreasing function on \([0, 1]\) that satisfies

\[
\lim_{u \to 0} u \sigma(u) = 0, \quad \sup_{0 \leq u \leq 1} |u \sigma'(u)| \leq c_1.
\]
Then, with $c_2 \leq 1 + (c_1 / \sigma(1))$,
\[
\sum_{k=1}^{n} s_k \wedge \gamma_k \leq c_2 \min_{0 \leq k \leq n} \left[ k s_k + \sum_{k+1}^{n} \gamma_n \right].
\]

Proof. Let $\Gamma_k = \sum_{i=k+1}^{n} \gamma_i$, and let $\kappa = \max\{ k \geq 1 : s_k \leq \gamma_k \}$ if such an index exists, otherwise set $\kappa = 0$. Using the monotonicity of both sequences and the definition of $\kappa$, we have

\begin{equation}
\sum_{i=1}^{n} s_i \wedge \gamma_i = \sum_{i=1}^{\kappa} s_i \wedge \gamma_i + \Gamma_\kappa \geq \kappa(s_\kappa \wedge \gamma_\kappa) + \Gamma_\kappa \geq \kappa s_\kappa + \Gamma_\kappa \geq \inf_{k \geq 0} k s_k + \Gamma_k.
\end{equation}

(b) For each $k$, including $k = 0$ and $n$, we have $\sum_{i=1}^{n} s_i \wedge \gamma_i \leq \sum_{i=1}^{k} s_i + \sum_{i=k+1}^{n} \gamma_i$, so that

\[
\sum_{i=1}^{n} s_i \wedge \gamma_i \leq \min_{0 \leq k \leq n} \sum_{i=1}^{k} s_i + \Gamma_k.
\]

The result will follow if we show that $\sum_{i=1}^{k} s_i \leq c_2 k s_k$ for $k = 0, \ldots, n$. But

\[
\sum_{i=1}^{k} s_i = \sum_{i=1}^{k} \sigma(i/n) \leq \int_{0}^{k/n} \sigma(u)du.
\]

Hence $(k s_k)^{-1} \leq \sup_{0 \leq u \leq [\alpha \sigma(x)]^{-1}} \int_{0}^{x} \sigma(u)du$. The result now follows from

Remark. If $\sigma(u)$ is a positive, decreasing function on $[0, 1]$ that satisfies

\[
\lim_{u \to 0} u \sigma(u) = 0, \quad \sup_{0 \leq u \leq 1} |u \sigma'(u)| \leq c_1.
\]

Then, with $c_2 \leq 1 + (c_1 / \sigma(1))$ and $v \in [0, 1],$

\[
\int_{0}^{v} \sigma(u)du \leq c_2 v \sigma(v).
\]

Indeed, by partial integration,

\[
\int_{0}^{v} \sigma(u)du = v \sigma(v) + \int_{0}^{v} u |\sigma'(u)|du \leq v |\sigma(v) + c_1| \leq v \sigma(v)[1 + c_1 / \sigma(1)].
\]

\]

13.7. False discovery rate estimation

The notion of False Discovery Rate (FDR) originated in the area of multiple inference, which is concerned with the simultaneous testing of a possibly large number of null hypotheses. When cast as a prescription for estimation, the FDR point of view leads to an estimator closely connected with the $2k \log(n/k)$ penalty class. We describe this connection, first by reviewing the simultaneous testing setting for FDR.
Consider the orthogonal regression model \( y = \mu + \varepsilon z \) in \( \mathbb{R}^n \), and consider the \( n \) separate null hypotheses \( H_i : \mu_i = 0 \). Suppose that \( n \) independent test statistics are given, which we here take to be simply the components \( y_i \) of the data. Traditional \textit{familywise error rate} (FWER) control methods seek to bound the chance of even one type I error among the \( n \) tests. The goal then is to guarantee that

\[
P\{\text{reject } \geq 1 \, H_i \mid \text{all } H_i \text{ true} \} \leq q,
\]

for a specified value \( q \). The standard way to achieve this is to use the Bonferroni approach: assign error probability \( q/n \) to each test and set the rejection regions as \( \{|y_i| > t\} \) where \( t \) is chosen so that under the null distribution, \( P\{|z_i| > t\} = \frac{q}{n} \). As usual, let \( z(\eta) \) denote the upper \( 100(1 - \eta)\% \) quantile of the standard Gaussian distribution. Thus \( t = t_B \), where

\[
t_B = z(q/2n) \sim \sqrt{2 \log(2n/q)} \sim \sqrt{2 \log n}
\]

for \( n \) large. This shows the fundamental limitation of the FWER control approach—since the thresholds are chosen quite high, the overall procedure has insufficient power to detect false \( H_i \).

Here is an alternate sequential procedure, originally due to Seeger (1968) and Simes (1986), to decide which hypotheses to reject. Form the 2-sided \( P \)-values \( P_i = 2\Phi(|y_i|/\varepsilon) \), order them: \( P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(n)} \), and look for the last crossing time of a linear boundary with slope \( q/n \):

\[ k_F = \max \{ i : P_{(i)} \leq q/n \}. \] (13.20)

INCLUDE DIAGRAM! Now reject all null hypotheses \( H_{(i)} \) corresponding to \( i = 1, \ldots, k_F \). By contrast, the Bonferroni method rejects some hypotheses if and only if \( P_{(1)} \leq q/n \). If \( k_B = \#\{ i : P_{(i)} \leq q/n \} \) is the number of Bonferroni-rejected hypotheses, then necessarily \( k_F \geq k_B \), so that FDR always generates at least as many “discoveries” as Bonferroni, and typically many more.

A major contribution of Benjamini & Hochberg (1995) was to provide a definition of false discovery rate that the \( k_F \) parameter satisfies. Let \( \mathcal{N} = \{ k : \mu_k = 0 \} \) be the set of “true” null hypotheses, and \( \mathcal{D} = \{ k : H_k \text{ is rejected} \} \), the set of “discoveries” made from the data. Then set

\[
FDR = |\mathcal{D} \cap \mathcal{N}|/|\mathcal{D}|,
\]

the fraction of discoveries that are false, i.e., come from true null hypotheses.

The key result of Benjamini & Hochberg (1995) is that the sequential procedure producing \( k_F \) guarantees that

\[
E_{\mu} \{ FDR \} \leq q, \quad \text{for all } \mu \in \mathbb{R}^n.
\]

Thus, the expected fraction of spurious results is controlled, \textit{regardless of} the configuration of the true means. [In fact, Benjamini & Hochberg (1995) prove a somewhat sharper result, with the upper bound being \( n_0 q/n \), where \( n_0 = \#\{ i : \mu_i = 0 \} \).]
**FDR Estimation.** As proposed by Abramovich & Benjamini (1995), the definition (13.20) can be converted into a prescription for estimation in the sequence model via the switching relation

\begin{equation}
P_{(k)} = 2 \Phi(|y|_{(k)}/\epsilon) \leq kq/n \iff |y|_{(k)}/\epsilon > z(kq/2n).
\end{equation}

(Here, \(|y|_{(1)} \geq \ldots \geq |y|_{(n)}\) are the order statistics of the data y). This suggests that we define a boundary sequence \(k \to t_{n,k}\) via the expression

\[ t_{n,k}^2 = z(kq/2n). \]

This sequence is decreasing and satisfies

\begin{equation}
t_{n,k}^2 \sim 2 \log(2n/kq)
\end{equation}
as \(k/n \to 0\). From (13.20) and (13.21), we see that the FDR index \(\hat{k}_F\) is the last crossing time

\begin{equation}
\hat{k}_F = \max\{k : |y|_{(k)} \geq \epsilon t_{n,k}\}.
\end{equation}
The corresponding estimator is just hard thresholding with a data determined threshold \(t_F = t_{\hat{k}_F}^*:\)

\[ \hat{\mu}_{F,k} = \eta_H(y_k; t_F^*) = \begin{cases} y_k & |y_k| \geq t_F^* \\ 0 & \text{otherwise.} \end{cases} \]

Now the connection with penalized estimation is apparent: (13.23) is just the backward stepwise model selection method discussed in Section 13.5 for the threshold sequence \(t_{n,k}^2 = z(kq/2n)\). The associated penalized least squares criterion is given by (14.9), and the relation (13.22) indicates that the penalty may belong to the \(2k \log(n/k)\) class. **FIX UP!**
CHAPTER 14

[Unrevised] Minimax estimation on $\ell_p$ balls

Suppose that we observe $n$-dimensional data

\begin{equation}
    y_i = \mu_i + \varepsilon_i \quad i = 1, \ldots, n
\end{equation}

where $\mu$ is constrained to lie in a ball of radius $C$ defined by the $\ell_p$ norm:

\begin{equation}
    \Theta = \Theta_{n,p}(C) = \{ \mu \in \mathbb{R}^n : \sum_{i=1}^n |\mu_i|^p \leq C^p \}.
\end{equation}

We seek to estimate $\mu$ using squared error loss $\| \hat{\mu} - \mu \|^2 = \sum_i (\hat{\mu}_i - \mu_i)^2$, and in particular to evaluate the nonlinear minimax risk

$$R_N(\Theta) = \inf_{\hat{\mu}} \sup_{\mu \in \Theta} E \| \hat{\mu} - \mu \|_2^2,$$

In this section we will study non-asymptotic upper and lower bounds for the minimax risk – and will see that these lead to the optimal rates of convergence for these classes of parameter spaces.

The optimal constants can be found from an asymptotic analysis as $\epsilon \to 0$ (see Section ?? below).

The non-asymptotic bounds will have a number of consequences. We will see a sharp transition between the sparse case $p < 2$, in which non-linear methods clearly outperform linear ones, and the more regular setting of $p \geq 2$.

The upper bounds will illustrate the use of the $2k\log(n/k)$ type oracle inequalities established in the preceding chapter. They will also be used in Chapter ??? to derive exactly optimal rates of convergence over Besov spaces for certain wavelet shrinkage estimators.

The lower bounds exemplify the use of minimax risk tools based on hyperrectangles.

A note on parametrizations: We sometimes use $\eta^p = n^{-1} C^p$ below, in which case the membership condition $\mu \in \Theta_{n,p}(C)$ becomes

$$n^{-1} \sum_{i=1}^n |\mu_i|^p \leq \eta^p.$$

A control function. The non-asymptotic bounds will be expressed in terms of a control function $r_{n,p}(C) = \tilde{r}_{n,p}(\eta)$ defined separately for $p \geq 2$
and \( p < 2 \). For \( p < 2 \), set

\[
(14.3) \quad r_{n,p}(C) = \begin{cases} 
C^2 & \text{if } C \leq \sqrt{1 + \log n}, \\
\frac{C^p [1 + \log(n/C^p)]^{1-p/2}}{n} & \text{if } \sqrt{1 + \log n} \leq C \leq n^{1/p}, \\
\sqrt{n} & \text{if } C \geq n^{1/p}.
\end{cases}
\]

For \( p \geq 2 \), it is more convenient to use the \( \eta \) parameter:

\[
(14.4) \quad \tilde{r}_{n,p}(\eta) = \begin{cases} 
\eta n^2 & \text{if } \eta \leq 1, \\
n & \text{if } \eta \geq 1.
\end{cases}
\]

To show that the bounds can be attained, we use a penalized least squares estimator \( \hat{\mu}_P \) with a specific choice of penalty:

\[
\text{pen}(k) = k\lambda_k^2, \quad \lambda_k = \zeta(1 + \sqrt{2\log n\beta/k}), \quad \beta > 1.
\]

14.1. Theorem. For \( n \geq 1, 0 < p \leq \infty, 0 < C < \infty \), there exist constants \( c_i(\zeta, \beta) \) so that

\[
(14.5) \quad c_1 r_{n,p}(C) \leq R_N(\Theta_{n,p}(C)) \\
(14.6) \quad \leq \sup_{\Theta_{n,p}(C)} E\|\hat{\mu}_P - \mu\|^2 \leq c_2 |\log n + r_{n,p}(C)|.
\]

Remarks. 1. It is straightforward to extend to the case of arbitrary noise level \( \epsilon \). We obtain, for \( \Theta = \Theta_{n,p}(C) \),

\[
c_1 c^2 r_{n,p}(C/\epsilon) \leq R_N(\Theta, \epsilon) \leq \sup_{\Theta} E\|\hat{\mu}_P - \mu\|^2 \leq c_2 c^2 |\log n + r_{n,p}(C/\epsilon)|.
\]

2. Note that a single estimator \( \hat{\mu}_P \), defined without reference to either \( p \) or \( C \), achieves the upper bound. We may thus speak of \( \hat{\mu}_P \) as being adaptively optimal at the level of rates of convergence.

3. The formulation and methods used here owe much to the account of Birgé & Massart (2001). Indeed, their bounds are stronger in the sense that the \( \log n \) term on the right of (14.5) can be replaced by a constant. [On the other hand, the constants \( c_1, c_2 \) appearing in (14.5) have the advantage of not depending on \( p \).]

**Upper Bounds** For this choice of penalty, formula (13.17) yields – after noting that \( L_{n,1} = \log \beta n \), with \( c_1 = c_1(\zeta, \beta), c_2 = c_2(\zeta) \),

\[
E\|\hat{\mu}_P - \mu\|^2 \leq c_1 |\log n| + c_2 K_0(\mu).
\]

For the upper bound, we need then to show that

\[
(14.7) \quad s_{n,p}(C) := \sup\{K_0(\mu) : \mu \in \Theta_{n,p}(C)\} \leq c_3 r_{n,p}(C).
\]

As argued earlier,

\[
K_0(\mu) = \inf_{0 \leq k \leq n} \inf_{N(\hat{\mu}) = k} \|\mu - \hat{\mu}\|^2 + k\lambda_k^2
\]

\[
= \inf_{0 \leq k \leq n} \sum_{j > k} \mu_{(j)}^2 + k\lambda_k^2.
\]

187
Now $\lambda_n = (1 + \sqrt{2\log n}) = \sqrt{c_2}$, say, and so by choosing $k = n$ in (14.8), we obtain the global bound

$$S_{n,p}(C) \leq c_2 n. \tag{14.9}$$

For $p \geq 2$. Simply by choosing $k = 0$ in (14.8), we also have

$$K_0(\mu) \leq n \cdot n^{-1} \sum_{j=0}^{n-1} \mu_j^2 \leq n(n^{-1} \sum_{j=0}^{n-1} |\mu_j|^p)^{2/p} \leq m n^2. \tag{14.10}$$

Combining (14.9) and (14.10) suffices to establish (14.7) in the $p \geq 2$ case.

For $p < 2$. We first note that $\sum |\mu_j|^p \leq C^p$ implies that $|\mu|_{(j)} \leq C_j^{-1/p}$, and hence that

$$\sum_{j \geq k} \mu_{(j)}^2 \leq C^{2-p}(k + 1)^{1-2/p} \sum_{j \geq k} |\mu_j|^p \leq C^2(1 + \log n)^{1-2/p}. \tag{14.11}$$

Noting that the sum vanishes in the case $k = n$, we obtain from (14.8):

$$K_0(\mu) \leq \min\{n \lambda_n^2, \ \inf_{0 \leq k \leq n} C^2(k + 1)^{1-2/p} + k n^2 \} \tag{14.12}$$

$$\leq c_3 \inf_{1 \leq k \leq n} \{C^2 k^{1-2/p} + k(1 + \log n/k)\}.$$ 

Let us immediately note two cases of the bound (14.7): for $C \leq \sqrt{1 + \log n}$, take $k = 0$ in (14.11) to get $K_0(\mu) \leq C^2$, while for $C \geq n^{1/p}$, we again use bound (14.9).

**Lower Bounds.** *Case $p \geq 2$.* These are a straightforward consequence of the $\ell_p$ ball modulus lower bounds in the optimal recovery discussion of Chapter 12. Since for any estimator $\hat{\theta}$ and any constant $\gamma$,

$$E \|\hat{\theta} - \theta\|^2 \geq \gamma^2 P\{\|\hat{\theta} - \theta\| \geq \gamma\},$$

we obtain from ... (with $\epsilon = 1$) that

$$R_N(\Theta_{n,p}(C)) \geq (\pi_0/2) W^2(1, C)[1 - e^{-2\pi n^2}] \geq c_0 W^2(1, C) = c_0 n \min(1, C^2 n^{-2/p}) = c_0 n \min(1, \eta^2),$$

where we have used the evaluation of $W^2(\epsilon, C)$ given at ...

*Case $p < 2$.* It will be useful to consider sets with at most $k$ non-zero co-ordinates, each bounded by $\tau$:

$$\Theta_{n,0}(\tau; k) = \{\mu \in \mathbb{R}^n : N(\mu) \leq k, \ |\mu_i| \leq \tau \text{ for all } i\}.$$ 

$\Theta_{n,0}(\tau; k)$ is a union of $n^k$ hyperrectangles of dimension $k$ and side length $\tau$. In particular, these sets include the sparse sets

$$\Theta(\tau, n) := \Theta_{n,0}(\tau; 1)$$

(i.e. with at most one non-zero entry), which we studied in Chapter 12.
A simple consequence of the sparse lower bound lemma proved there (Ref.?) is a squared error minimax risk bound
\begin{equation}
R_N(\Theta(\tau; n)) \geq c_0 \tau^2 \wedge (1 + \log n).
\end{equation}

For any fixed number \(k\), we can obtain a subset of \(\Theta_{n,0}(\tau; k)\) by decreeing that each block of \([n/k]\) co-ordinates contain at most one non-zero entry. Thus
\[
\Theta_{n,0}(\tau; k) \supset \Pi^k \Theta(\tau; [n/k]).
\]

Since minimax risk is additive on product spaces, (14.13) gives a lower bound
\begin{equation}
R_N(\Theta_{n,0}(\tau; k)) \geq c_0 k^2 \wedge (1 + \log[n/k])
\end{equation}
\begin{equation}
\geq c_1 \{k \tau^2 \wedge (k + \log n/k)\}. 
\end{equation}

We return to \(\ell^p\)-balls and note now that
\[
\Theta_{n,0}(\tau; k) \supset \Theta_{n,p}(C) \iff k \tau^p \leq C^p.
\]

We have a bound (14.14) for each dimension \(k\) of hyperrectangles, and so we arrive at
\begin{equation}
R_N(\Theta_{n,p}(C)) \geq c_1 \sup_{1 \leq k \leq n} \{C^2 k^{1-2/p} \wedge k(1 + \log n/k)\},
\end{equation}
a bound which contains terms very similar to those in the upper bound (14.12).

Again we draw two quick conclusions: for \(C \leq \sqrt{1 + \log n}\), the choice \(k = 1\) yield the bound \(C^2\), while for \(C \geq n^{1/p}\), the choice \(k = n\) gives the lower bound \(n\).

Let us summarize what remains to be shown. Define two functions
\[
g_1(x) = C^2 x^{1-2/p}, \quad g_2(x) = x + x\log n/x.
\]

Then, for \(\sqrt{1 + \log n} \leq C \leq n^{1/p}\), we seek absolute constants \(c_3\) and \(c_4\) so that
\begin{equation}
c_3 r_{n,p}(C) \leq \sup_{1 \leq k \leq n} g_1(k) \wedge g_2(k)
\end{equation}
\begin{equation}
\leq \inf_{1 \leq k \leq n} g_1(k) + g_2(k) \leq c_4 r_{n,p}(C),
\end{equation}

where
\[
r_{n,p}(C) = C^p \left[1 + \log(n/C^p)\right]^{1-p/2}.
\]

Since \(g_1\) is decreasing and \(g_2\) is increasing for \(0 \leq x \leq n\), it is natural to look for the point \(x_* = x_*(C)\) at which \(g_1(x_*) = g_2(x_*)\), i.e. the point defined by
\begin{equation}
x_* = C^p \left[1 + \log(n/x_*)\right]^{-p/2},
\end{equation}
at which
\begin{equation}
g_1(x_*) = C^p \left[1 + \log(n/x_*)\right]^{1-p/2}.
\end{equation}

It is clear that \(C \to x_*(C)\) is strictly increasing, and that
\[ x_* = 1 \quad \iff \quad C = \sqrt{1 + \log n}, \]
\[ x_* = n \quad \iff \quad C = n^{1/p}. \]

Hence \( 1 \leq x_* \leq n \) if and only if \( \sqrt{1 + \log n} \leq C \leq n^{1/p} \), which explains the choice of transition points for \( C \) in the definition of \( r_{n,p}(C) \).

We now relate \( g_1(x_*(C)) \) to \( r_{n,p}(C) \). One direction is easy: putting \( x_* \leq n \) into (14.17) shows that \( x_* \leq C^p \), and hence from (14.18) that \( g_1(x_*) \geq r_{n,p}(C) \). For the other direction, make the abbreviations

\[ y = 1 + \log(n/x_*), \quad \text{and} \quad b = 1 + \log(n/C^p). \]

Now equation (14.17) shows that \( y \leq b + \log y \). But \( \log y \leq y/2 \) (since \( y \geq 1 \) whenever \( x_* \leq n \)), and so \( y \leq 2b \). Plugging this into (14.18), we conclude that

\[ r_{n,p}(C) \leq g_1(x_*(C)) \leq 2r_{n,p}(C). \]

**A detail.** We are not quite done since the extrema in the bounds (14.16) should be computed over integers \( k, 1 \leq k \leq n \). The following remark is convenient: for \( 1 \leq x \leq n \), the function \( h(x) = x + x \log(n/x) \) satisfies

\[ \frac{1}{2} h([x]) \leq h(x) \leq 2h([x]). \]

Indeed, \( h \) is concave and \( h(0) = 0 \), and so for \( x \) positive, \( h(x) \leq 2h(x/2) \). Since \( h \) is increasing for \( x \leq n \), it follows that if \( y \geq x/2 \), then \( h(x) \leq 2h(y) \). Since \( x \geq 1 \) implies both \( [x] \geq x/2 \) and \( x \geq [x]/2 \), the bounds (14.20) follow.

For the upper bound in (14.16), take \( k = [x_*] \): since \( g_1 \) is decreasing, and using (14.20), we find

\[ \inf_{1 \leq k \leq n} g_1 + g_2 \leq (g_1 + g_2)([x_*]) \leq g_1(x_*) + 2g_2(x_*) = 3g_1(x_*). \]

For the lower bound, take \( k = [x_*] \), and again from (14.20),

\[ \sup_{1 \leq k \leq n} g_1 \wedge g_2 \geq (g_1 \wedge g_2)([x_*]) = g_2([x_*]) \geq \frac{1}{2} g_2(x_*) = \frac{1}{2} g_1(x_*). \]

We use (14.19) to complete the proof of (14.16).
Suppose we observe $n$-dimensional data

$$y_i = \theta_i + \varepsilon_i \quad i = 1, \ldots, n$$

where $\theta$ is constrained to lie in a ball of radius $C$ defined by the $\ell_p$ norm:\n
$$\Theta = \Theta_{n,p}(C) = \{ \theta \in \mathbb{R}^n : \sum_{i=1}^{n} |\theta_i|^p \leq C^p \}.$$\n
We seek to estimate $\theta$ using squared error loss $\| \hat{\theta} - \theta \|^2 = \sum_i (\hat{\theta}_i - \theta_i)^2$, and in particular to evaluate the nonlinear minimax risk

$$R_N(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} E\| \hat{\theta} - \theta \|^2_2,$$

and make comparisons with the corresponding linear minimax risk $R_L(\Theta)$.

Although this model is finite dimensional, it is non-parametric in character since the dimension of the unknown parameter equals that of the data (and we will consider asymptotics as $n \to \infty$). The $\ell_p$-constrained parameter space $\Theta$ is permutation symmetric and certainly solid, orthosymmetric and compact. It is thus relatively simple to study and yet yields a very sharp distinction between linear and non-linear estimators when $p < 2$. The setting also illustrates the Bayes minimax method discussed in Chapter 7.

Remark. A slightly artificial motivation for the $\ell_p$ balls model comes from the continuous Gaussian white noise model $dY_t = f(t)dt + n^{-1/2}dW_t$, $t \in [0,1]$ in which $f$ has the form $f = \sum \theta_k \phi_{n,k}$, where $\phi_{n,k}(t) = n^{1/2} \phi(nt-k)$. If $\phi$ is the indicator of the unit interval $[0,1]$, then $\| f - \hat{f} \|_2^2 = \sum (\hat{\theta}_i - \theta_i)^2$ and since

$$\int |f|^p = n^{p/2-1} \sum_{i=1}^{n} |\theta_k|^p,$$

an $L_p$ norm constraint on $f$ corresponds to an $\ell_p$ constraint on $\theta$. This connection becomes much more natural and useful in the context of sequence space characterizations of Besov and Triebel classes of functions to be discussed later (ref).

Remark. The discussion here will be confined to squared error loss, but the main results and phenomena remain valid for $\ell_q$-loss, $0 < q < \infty$, with the non-linearity phenomena appearing in case $q < p$. Details are given in Donoho & Johnstone (1994b).

More Outline and Intro

14.1. Linear Estimators.

The $\ell_p$ balls $\Theta_{n,p}(C)$ are solid and orthosymmetric and compact for all $0 < p \leq \infty$. However it is also easy to see that they are quadratically convex only if $p \geq 2$, and that

for $p < 2$, \quad $QHull[\Theta_{n,p}(C)] = \Theta_{n,2}(C)$.\n
191
Theorem 7.11 says that the linear minimax risk is determined by the quadratic hull, and so we may suppose that $p \geq 2$. Theorem 7.9 says that the linear minimax risk is found by looking for the hardest rectangular subproblem:

$$R_L(\Theta) = \sup \{ \sum_i e^2 \tau_i^2 / (e^2 + \tau_i^2) : \sum_i \tau_i^p \leq C^p \}.$$ 

Introducing new variables $u_i = \tau_i^p / C^p$, this optimization is rephrased as that of maximizing

$$f(u) = e^2 \sum_i \ell( C e^{-1} u_i^{1/p} )$$

over the simplex $\sum_i u_i \leq 1$ in the non-negative orthant of $\mathbb{R}^n$. Since $f$ is symmetric and increasing in the co-ordinates $u_i$, and concave when $p \geq 2$, it follows that the maximum is attained at the centroid $u_i = n^{-1}(1, \ldots, 1)$. Introducing the normalized radius $\bar{\eta} = n^{-1/p} \sqrt[2]{C/e}$, we may write the corresponding minimax risk as $n e^2 \bar{\eta}^2$. From (7.3), the corresponding linear minimax estimate is $\hat{\theta}_L = \ell(\bar{\eta}_n) y$. To summarize:

14.2. Proposition. $\Theta_{n,p}(C)$ is quadratically convex iff $p \geq 2$, and for $p < 2$, $QHull(\Theta_{n,p}(C)) = \Theta_{n,2}(C)$. Let $\bar{p} = p \lor 2$ and $\bar{\eta} = n^{-1/\bar{p}} C/e$. The minimax linear risk for squared error loss is

$$R_L(\Theta_{n,p}(C), e) = n e^2 \bar{\eta}^2 / (1 + \bar{\eta}^2),$$

with minimax linear estimator $\hat{\theta}_L$ given coordinatewise by

$$\hat{\theta}_{L,i}(y) = [\bar{\eta}^2 / (1 + \bar{\eta}^2)] y_i.$$ 

Remarks. 1. For large $C$, and hence large $\bar{\eta}$, the minimax linear risk approaches the unconstrained minimax risk for $\mathbb{R}^n$, namely $n e^2$.

14.3. Example. The calibration $e = 1/\sqrt{n}$ arises frequently in studying sequence model versions of nonparametric problems (compare (3.9) in Chapter ??). Consider the $\ell_1$ ball of radius $C = 1 : \Theta_{n,1} = \{ \theta : \sum_i |\theta_i| \leq 1 \}$. We see that $\bar{\eta} = n^{-1/2} \cdot n^{1/2} = 1$ and that

$$R_L(\Theta_{n,1}) = 1/2, \quad \hat{\theta}_L(y) = y/2.$$ 

Of course $\Theta_{n,1}$ has the same linear minimax risk as the solid sphere $\Theta_{n,2}$ which is much larger, for example in terms of volume. Indeed a heuristic argument suggests that we should be able to improve on $\hat{\theta}_L$ by using a threshold estimate, which also operates on each coordinate separately, and defined by

$$\hat{\theta}_{\lambda,i}(y) = \begin{cases} y_i & \text{if } |y_i| > \lambda e_n \\ 0 & \text{otherwise} \end{cases}.$$ 

Suppose, for now, that $\lambda$ is chosen at a fixed number of standard deviations of $y_i$ — say $\lambda = 5$. Since on $\Theta_{n,1}$, the coordinates $\theta_i$ are on average $O(n^{-1})$, very few of the data values $y_i$ will exceed the threshold $\lambda$, and so in all of these coordinates, the error will be $O(1/n)$. In those few coordinates where
$|y_i| > \lambda \epsilon_n$, the error will be $O(n^{-1/2})$, but overall, the MSE will be much smaller than that of the linear minimax estimator $y/2$, which is forced to make an error of $O(n^{-1/2})$ in essentially all coordinates.

### 14.2. Univariate Bayes Minimax Problem

We now consider a univariate Bayes minimax problem which may be viewed as a generalization of the bounded normal mean problem of Section 7.1. Suppose that $y \sim N(\theta, \sigma^2)$, and that $\theta$ is distributed according to a prior $\pi(d\theta)$ on $\mathbb{R}$. The integrated risk of an estimator $\hat{\theta}(y)$ is again $R(\theta, \tau) = \int \int f(y|x)P(\theta|x) \pi(\theta) \, dx \, dy$. It is now assumed that $\pi$ belongs to a class satisfying the $p-$th moment constraint

$$\mathcal{M}_p(\tau) = \{ \pi(d\theta) : \int |\theta|^p \pi(d\theta) \leq \tau^p \},$$

which is convex and weakly compact for all $p \leq \infty$ and $\tau < \infty$. Such moment constraints are a “population” version of the “empirical” constraints on $(\hat{\theta}_1, \ldots, \hat{\theta}_n)$ defining an $\ell_p$-ball – compare (14.22). We study the Bayes minimax risk

$$\beta_p(\tau, \epsilon) = \inf_{\hat{\theta}} \sup_{\pi \in \mathcal{M}_p(\tau)} B(\hat{\theta}, \pi) = \sup_{\pi \in \mathcal{M}_p(\tau)} B(\pi),$$

where the second equality uses the minimax theorem (6.15) and (??) of Chapter ??.

Of course, $\mathcal{M}_\infty(\tau)$ equals the set of priors supported on the bounded interval $[-\tau, \tau]$, and so $\beta_\infty(\tau, \epsilon) = \rho_N(\tau, \epsilon)$. We saw in Section 7.1.2 that the least favorable priors are discrete and only available numerically. In the case of $\mathcal{M}_2(\tau)$, Gaussian priors $\pi_\tau = N(0, \tau^2)$ are least favorable and

$$\beta_2(\tau, \epsilon) = \frac{\tau^2 \epsilon^2}{\tau^2 + \epsilon^2} = \rho_\pi(\tau, \epsilon).$$

(Corollary 6.5). From now on, however, we will be especially interested in $p < 2$, and in general we will not have such explicit information about the value of $\beta_p(\tau, \epsilon)$, least favorable priors or corresponding estimators. We will therefore be interested in approximations, either by linear rules when $p \geq 2$, or more importantly, by threshold estimators for all $p > 0$.

**Remark on Notation.** We use the lower case letters $\beta$ and $\rho$ for Bayes and frequentist minimax risk in univariate problems, and the upper case letters $B$ and $R$ for the corresponding multivariate minimax risks.

#### 14.4. Proposition

**For the Bayes minimax risk defined at (14.23)**

1. $\beta_p(\tau, \epsilon)$ is decreasing in $p$,
2. $\beta_p(\tau, \epsilon) = \epsilon^2 \beta_p(\tau/\epsilon, 1)$,
3. $\beta_p(\tau, \epsilon)$ is increasing and concave in $\tau^p$, and increasing in $\epsilon$,
4. $\beta_p(\tau, \epsilon) \leq a^2 \beta_p(\tau, \epsilon)$ for all $a \geq 1$,
5. $\beta_p(\tau, \epsilon)$ is continuous in $\tau$. 

193
PROOF (AND DISCUSSION). (1) is straightforward (and recall that Theorem 7.2 showed that \( \beta_2(\tau, \epsilon)/\beta_\infty(\tau, \epsilon) \leq 1.25 \)). (2) follows by change of scale. Assuming (3) for now, (4) follows from (2) and (3), while (5) follows from monotonicity and (4).

We turn to (3). Let \( t = \tau^p \); the function \( \tilde{\beta}(t) = \sup\{ B(\pi) : \int [\theta^p d\tau = t] \} \) is concave in \( t \) because \( \pi \rightarrow B(\pi) \) is concave and the constraint on \( \pi \) is linear. While monotonicity of \( \beta_p(\tau, \epsilon) \) in \( \tau \) is trivial, a separate argument is required to show that \( \tilde{\epsilon} < \epsilon \) entails \( \beta_p(\tau, \tilde{\epsilon}) \leq \beta_p(\tau, \epsilon) \). Indeed, let \( \tilde{y} \sim N(\theta, \tilde{\epsilon}^2) \) and \( w \sim N(0, \epsilon^2 - \tilde{\epsilon}^2) \). Thus \( y = \tilde{y} + w \) in distribution. Given an estimator \( \hat{\theta}(\tilde{y}) \), define a corresponding estimator in the \( \tilde{y} \)-problem by

\[
\hat{\theta}(\tilde{y}) = E[\hat{\theta}(\tilde{y} + w)|\tilde{y}],
\]

- this is a legitimate estimator because \( \tilde{y} \) is sufficient for \( \theta \). By Jensen’s inequality, \( \hat{\theta} \) has smaller mean squared error than \( \tilde{\theta} \), so \( B(\hat{\theta}, \pi; \tilde{\epsilon}) \leq B(\tilde{\theta}, \pi; \epsilon) \), and hence \( B(\pi, \tilde{\epsilon}) \leq B(\pi, \epsilon) \). Minimizing over \( \pi \) in \( M_p(\tau) \) gives \( \beta_p(\tau, \tilde{\epsilon}) \leq \beta_p(\tau, \epsilon) \).

The scaling property (2) means that it suffices to study the unit noise situation. As in the previous subsection, we use a special notation for this case: \( x \sim N(\mu, 1) \), and write \( \beta_p(\eta) \) for \( \beta_p(\eta, 1) \) where \( \eta = \tau/\epsilon \) denotes the signal to noise ratio.

14.2.1. \( p \geq 2 \) versus \( p < 2 \) in low signal-to-noise. We are interested in comparing threshold rules to the Bayes minimax value over all values of the parameter \( \eta \), but with a particular focus on \( \eta \rightarrow 0 \), so we begin with this.

Let us explore the differences between \( p \geq 2 \) and \( p < 2 \). Consider the trivial estimator \( \hat{\delta}_0 \equiv 0 \). For any prior \( \pi(d\mu) \in M_p(\eta) \), when \( p \geq 2 \),

\[
B(\hat{\delta}_0, \pi) = E_{\pi} \mu^2 \leq (E_{\pi} \mu^p)^{2/p} \leq \eta^2.
\]

Consequently \( B(\pi) \leq \eta^2 \), and since this is true for any \( \pi \in M_p(\eta) \), we obtain for \( p \geq 2 \)

\[
\beta_p(\eta) \leq \eta^2.
\]

However, when \( p < 2 \), the moment inequality (14.24) fails completely. To see the consequences, consider the simple two point priors

\[
\pi = (1 - \alpha)\delta_0 + \alpha \delta_\mu
\]

which will lie in \( M_p(\eta) \) if

\[
\alpha \mu^p = \eta^p.
\]

Let \( \pi_\alpha = \pi_{\alpha, \mu_\alpha} \) denote the prior defined by both these conditions: of course \( \mu_\alpha = \eta \alpha^{-1/p} \) increases without bound as \( \alpha \searrow 0 \). In this case the zero estimator has unbounded integrated risk: as \( \alpha \searrow 0 \)

\[
B(\hat{\delta}_0, \pi_\alpha) = E_{\mu_\alpha} \mu^2 = \alpha \mu_\alpha^2 = \eta^p \mu_\alpha^{2-p} \nearrow \infty.
\]
Formula (14.26) describes an example of a sparse prior: in i.i.d. samples 
\( \mu_i \) from \( \pi_\alpha \), most values \( \mu_i \) are zero and a small fraction take the relatively 
large value \( \mu_\alpha = \eta \alpha^{-1/p} \). Our discussion so far suggests that these sparse 
riors will play an important role when \( p < 2 \). The next subsections will be 
denoted to fleshing out these ideas while establishing the following important 
result.

14.5. Theorem. As \( \eta \to 0 \),

\[
\beta_p(\eta) \sim \begin{cases} 
\eta^2 & 2 \leq p \leq \infty \\
\eta^p(2 \log \eta^{-p})^{1-p/2} & 0 < p < 2 
\end{cases}
\]

If \( p \geq 2 \), then \( \hat{\delta}_0 \) is asymptotically minimax and \( \pi = (\delta_{-\eta} + \delta_{\eta})/2 \) is 
asymptotically least favorable.

If \( p < 2 \), then \( \hat{\delta}_\lambda \) with threshold \( \lambda = \sqrt{2 \log \eta^{-p}} \) is asymptotically mini-
max and \( \pi_\alpha \) is asymptotically least favorable, with \( \alpha = \alpha(\eta) \) specified below.

Remark. The threshold estimator \( \hat{\delta} \sqrt{2 \log \eta^{-p}} \) is also asymptotically mini-
max when \( p \geq 2 \).

14.2.2. Integrated risk of threshold estimators. Emboldened by 
the heuristic analysis of hard thresholding over \( \ell_1 \) balls, let us now consider 
the worst case behavior of soft threshold estimators over \( M_p(\eta) \):

14.6. Proposition. Let a threshold \( \lambda \) and moment space \( M_p(\eta) \) be 
given. Then

\[
sup_{\pi \in M_p(\eta)} \int r_S(\lambda, \mu)\pi(d\mu) \leq \begin{cases} 
r_S(\lambda, 0) + \eta^p(1 + \lambda^2)^{1-p/2} & p < 2 \\
r_S(\lambda, 0) + \eta^2 & p \geq 2
\end{cases}
\]

Thus, in contrast to (14.28), threshold estimators have well controlled 
maximum risk properties over \( M_p(\eta) \) – we will see below that (14.30) is 
actually the optimal bound as \( \eta \to 0 \).

Proof. Since the risk function \( \mu \to r_S(\lambda, \mu) \) of soft thresholding is an 
even function of \( \mu \in \mathbb{R} \) it suffices to consider priors \( \pi \) supported in \([0, \infty)\). 
The result follows directly from the bound (11.8) for the frequentist risk of 
soft thresholding, which implies

\[
r_S(\lambda, \mu) \leq r_S(\lambda, 0) + (1 + \lambda^2) \land \mu^2.
\]

Consider first the case \( p \geq 2 \). Introducing the notation \( B_S(\lambda, \pi) = \int r_S(\lambda, \mu)\pi(d\mu) \), 
we have, using (14.24)

\[
B_S(\lambda, \pi) \leq r_S(\lambda, 0) + \int \mu^2\pi(d\mu) \leq r_S(\lambda, 0) + \eta^2.
\]

Suppose now that \( p < 2 \). For sufficiently large \( c \), \( (1 + \lambda^2) \land \mu^2 \leq c\mu^p \) for 
all positive \( \mu \) – indeed, a graph shows that the smallest value of \( c \) is obtained.
by solving $c\mu^2 = \mu_\lambda^2$, where $\mu_\lambda = \sqrt{1 + \lambda^2}$. Hence $c = (1 + \lambda^2)^{1-p/2}$, and so

$$B_S(\lambda, \pi) \leq r_S(\lambda, 0) + c \int \mu^p \pi(d\mu),$$

and hence

$$\sup_{\mathcal{M}_p(\eta)} B_S(\lambda, \pi) \leq r_S(\lambda, 0) + (1 + \lambda^2)^{1-p/2} \eta^p.$$  

Set the threshold in (14.30) at $\lambda = \sqrt{2 \log \eta^{-p}}$, so that $\phi(\lambda) = \phi(0) \eta^p$ and hence $r(\lambda, 0) \sim 4\lambda^{-3} \phi(\lambda) = o(\eta^p)$ makes a negligible contribution. Using

$$\beta_p(\eta) \leq \inf_{\lambda} \sup_{\pi} B_S(\lambda, \pi)$$

we obtain as a corollary the upper bound part of Theorem 14.5 using threshold estimators for all values of $p$.

14.7. COROLLARY.  

$$\beta_p(\eta) \leq \eta^p (2\log \eta^{-p})^{1-p/2}(1 + o(1)) \quad \text{as} \ \eta \to 0.$$  

Remark. There is an alternative approach to bounding $\sup_{\mathcal{M}_p(\eta)} B_S(\lambda, \pi)$ which looks for the maximum of the linear function $\pi \to B_S(\lambda, \pi)$ among the extreme points of the convex $\mathcal{M}_p(\eta)$ and shows that the maximum is actually of the two point form (14.26). This approach yields

14.8. PROPOSITION. Let a threshold $\lambda$ and moment space $\mathcal{M}_p(\eta)$ be given. Then

$$\sup \{B(\hat{\delta}_\lambda, \pi) : \pi \in \mathcal{M}_p(\eta)\} \leq r(\lambda, 0) + \eta^p \mu_\lambda^{2-p}$$

where $\mu_\lambda$ is the unique solution of

$$r(\lambda, \mu_\lambda) - r(\lambda, 0) = (\mu_\lambda/p) r_\mu(\lambda, \mu_\lambda).$$

The least favorable prior for $\hat{\delta}_\lambda$ over $\mathcal{M}_p(\eta)$ is of the two point prior form (14.26) with $\alpha$ determined from $\eta$ and $\alpha = \mu_\lambda$ by (14.27). As $\lambda \to \infty$, we have

$$\mu_\lambda \sim \lambda + \hat{\phi}^{-1}(p/2).$$

The proof, being longer than that of Proposition 14.6 is given in the appendix.

14.2.3. Digression on the choice of threshold. A first comment derives from the $n$-dimensional setting $y_i = \theta_i + z_i$, $i = 1, \ldots, n$ with noise $\epsilon = 1$ for convenience. A key property of i.i.d. samples from the Gaussian distribution is that $\max_{1 \leq i \leq n} |z_i|$ is concentrated about $\sqrt{2 \log n}$. Thus it is difficult to distinguish a single sample with true signal size smaller than this level from the most extreme of $n$ samples of noise. This suggests a threshold set at around $\lambda = \sqrt{2 \log n}$ at least in very sparse cases. This choice will be discussed in greater detail in later chapters.
Here is an heuristic argument guiding the choice of \( \lambda \) in the less sparse minimax cases that we have been considering. One can view the choice of threshold as a decision between presence and absence of signal, and this decision is hardest when the signal strength is exactly at the threshold. So we switch temporarily to a Bayesian testing situation in which one tests a null hypothesis \( H_0 : \mu = 0 \) against the alternative \( H_1 : \mu = \lambda \). We decree that the loss associated with false decisions will be \( \lambda^2 \), in order to make the testing problem most comparable with the estimation one. Assign prior probabilities \( \pi_0(H_0) = 1 - \alpha \) and \( \pi_0(H_1) = \alpha \) and let \( P_{x_0} \) be the joint distribution of \( \mu \) and \( x \). Of course, we suppose that \( \alpha \) and \( \lambda \) are connected by the moment constraint \( \alpha \lambda^p = \eta \). The two error probabilities of the test which rejects \( H_0 \) if and only if \( y > \lambda \) are given by

\[
e_0 = P_{x_0}(H_0; y > \lambda) = (1 - \alpha) \Phi(\lambda), \quad e_1 = P_{x_0}(H_1; y \leq \lambda) = \alpha/2.
\]

Let us choose \( \lambda \) to minimise the total expected loss under the prior \( \pi_0 \), namely \( b(\lambda) = \lambda^2[(1 - \alpha) \Phi(\lambda) + \alpha/2] \). Using the relation \( \Phi(\lambda) \sim \phi(\lambda)/\lambda \) and choosing \( \alpha \) small and satisfying \( \alpha = (\eta/\lambda)^p \), we find by differentiating \( b(\lambda) \) that \( \lambda \) approximately satisfies the equation

\[
\lambda^2 \phi(\lambda) = (1 - p/2) \eta^p \lambda^{1-p},
\]

which in turn implies that \( \lambda^2 \approx 2 \log \eta^{-p} \), the value that appears in Theorem 14.5.

Remark. Consider the special choice \( \epsilon = n^{-1/2} \). Then \( \eta_n^p = n^{-1}(C/\epsilon)^p = n^{-1 + p/2} C \) and so \( \lambda_n^2 = 2 \log \eta_n^{-p} = (2 - p) \log n - 2p \log C \). Hence larger signal strength, represented both in index \( p \) and in radius \( C \), translates into a smaller choice of minimax threshold. Note that in a very small signal setting, \( \eta_n = 1/n \), we recover the choice \( \lambda_n = \sqrt{2 \log n} \) discussed at the start of this subsection.

14.2.4. Lower bounds for Bayes risk of sparse two point priors. For \( p \geq 2 \), the upper bound (14.25) using the zero estimator is shown to be sharp for small \( \eta \) by the lower bound (7.14) using symmetric two point priors.

For \( p < 2 \), it is easy to compute the Bayes risk of the sparse two point priors (14.26): they provide lower bounds to \( \beta_p(\eta) \) and information about the nature of least favorable priors for small \( \eta \). 

Posterior Distribution. We begin with the Bayes estimator \( \hat{\delta}_n(x) \) corresponding to prior (14.26), which is just the posterior mean:

\[
(14.34) \quad \hat{\delta}_n(x) = \mu P(\{\mu\}|x) = \frac{\mu}{1 + m(x)},
\]

where \( m(x) \) is the Bayes factor (7) for 0 versus \( \mu \):

\[
(14.35) \quad m(x) = \frac{(1 - \alpha) \phi(x)}{\alpha \phi(x - \mu)} = \frac{P(\{0\}|x)}{P(\{\mu\}|x)}
\]

197
The Bayes factor thus gives the ratio of posterior probabilities of the support points 0 and $\mu$ of prior $\pi$. Obviously $m(x)$ decreases as $x$ increases: indeed

$$m(x)/m(y) = e^{-\mu(x-y)},$$

The value of $x$ at which $m(x) = 1$ is the point of posterior indifference between 0 and $\mu$. We will be interested in the apparently peculiar situation where this posterior indifference point lies to the right of $\mu$! Thus, given $\alpha > 0$ (small) and $a > 0$ (large), choose $\mu$ in prior (14.26) to satisfy $m(\mu + a) = 1$, that is

$$\alpha \phi(a + \mu) = \alpha \phi(a).$$

[In effect, the prior probability on 0 is so large that even if $x > \mu$ (but is less than $\mu + a$), the posterior distribution places more weight on 0 than $\mu$. (PICTURE!)] We may reexpress the Bayes rule (14.34) using the condition $m(\mu + a) = 1$ as

$$\hat{\delta}_n(\mu + z) = \mu/[1 + e^{-\mu(z-a)}].$$

Bayes risk of $\pi_{\alpha,\mu}$. Expressed in terms of $z \sim N(0,1)$, we have

$$B(\pi_{\alpha,\mu}) = (1 - \alpha) E^2[\hat{\delta}_n(z) + \alpha E[\hat{\delta}_n(\mu + z) - \mu]^2].$$

(14.37) shows that $\hat{\delta}_n(x) \approx 0$ for all $x \leq \mu + a$ - if $\mu$ is at all large, the primary contribution will come from the second term. In any event,

$$B(\pi_{\alpha,\mu}) \geq \alpha \mu^2 \int_{-\infty}^{\infty} \frac{\phi(z)dz}{1 + e^{\mu(z-a)}} \sim \alpha \mu^2 \Phi(a),$$

as $\mu = \mu_\alpha \sim \infty$ (the integral increases to $\Phi(a)$ by the monotone convergence theorem.)

Lower bound for $\beta_p(\eta)$. We assume that $a$ is fixed and large, and determine the parameters $(\alpha, \mu)$ by the moment condition (14.27) and the likelihood ratio condition (14.36). These may be rewritten as

$$\alpha = \eta^p/\mu^p; \quad \mu^2 + 2a \mu = 2 \log(1 - \alpha)/\alpha.$$ 

It follows that for small $\eta$, solutions $(\alpha_\eta, \mu_\eta)$ exist and $\alpha_\eta \to 0$ and that $\mu_\eta \sim \infty$, and so from (14.39),

$$\mu_\eta \sim \sqrt{2 \log \alpha_\eta^{-1}} \sim \sqrt{2 \log \eta^{-p}}.$$ 

Writing $\pi_\eta$ for $\pi_{\alpha_\eta,\mu_\eta}$, we get from (14.38) and these considerations that

$$\beta_p(\eta) \geq B(\pi_\eta) \sim \alpha \mu^p \cdot \mu^{2-p} \Phi(a) \sim \eta^p(2 \log \eta^{-p})^{1-p/2} \Phi(a)(1 + o(1))$$

Since $a$ can be taken arbitrarily large, when taken together with Corollary 14.7 this completes the proof of Theorem 14.5.
14.2.5. Near Minimaxity of univariate thresholding. Move this section later??

We consider the optimal performance of the best threshold rule over the moment space $\mathcal{M}_p(\tau)$ with the goal of comparing it to the minimax Bayes estimator, which although optimal, is not available explicitly. Define therefore

$$
\beta_{S,p}(\tau, \epsilon) = \inf_{\Lambda} \sup_{\pi \in \mathcal{M}_p(\tau)} B(\hat{\lambda}, \pi),
$$

where $\hat{\lambda}$ refers to a soft threshold estimator (11.4) with threshold $\lambda$. Throughout this section, we work with soft thresholding (sometimes emphasised by the subscript “S”), though analogous results are possible for hard thresholding (see Donoho & Johnstone (1994b). A goal of this subsection is to establish an analogue of Theorem 7.2, which in the case of a bounded normal mean, bounds the worst case risk of linear estimators relative to all non-linear ones. Over the more general moment spaces $\mathcal{M}_p(\tau)$, the preceding sections show that we have to replace linear by threshold estimators.

To emphasize that the choice of estimator in (14.40) is restricted to thresholds, we sometimes write $B_S(\lambda, \pi)$ for $B(\hat{\lambda}, \pi)$. Just as in the full non-linear case, it is useful to think in terms of least favorable distributions for thresholding. The payoff function $B_S(\lambda, \pi)$ is not convex in $\lambda$ as is shown by consideration of, for example, the risk function $\lambda \rightarrow r_S(\lambda, 0)$ corresponding to $\pi = \delta_0$. On the other hand, $B_S(\lambda, \pi)$ is still linear in $\pi$, and this makes it possible to establish the following minimax theorem directly (Appendix, to come).


$$
\inf_{\Lambda} \sup_{\mathcal{M}_p(\tau)} B_S(\lambda, \pi) = \sup_{\mathcal{M}_p(\tau)} \inf_{\Lambda} B_S(\lambda, \pi) = \sup_{\mathcal{M}_p(\tau)} B_S(\pi).
$$

14.10. Proposition. The minimax Bayes threshold risk $\beta_{S,p}(\tau, \epsilon)$ also satisfies the properties (1) - (5) of $\beta_p(\tau, \epsilon)$ enumerated in Proposition 14.4.

Given the minimax result Proposition 14.9, the proof of this result is entirely analogous to that of Proposition 14.4 except in the case of monotonicity of $\beta_{S,p}(\tau, \epsilon)$ in $\epsilon$ (Proof to follow.)

14.11. Proposition. (i) For $0 < p \leq \infty$,

$$
\sup_{\tau, \epsilon} \frac{\beta_{S,p}(\tau, \epsilon)}{\beta_p(\tau, \epsilon)} = \Lambda(p) < \infty.
$$

(ii) For $p \geq 2$, $\Lambda(p) \leq 2.22$.

Numerical work indicates that $\Lambda(1) = 1.6$, so that one may expect that even for $p < 2$, the inefficiency of the best threshold estimator is quite moderate. In addition, the proof below shows that the ratio

$$
R_S(\eta) = \frac{\beta_{S,p}(\eta, 1)}{\beta_p(\eta, 1)}
$$

190
approaches 1 as $\eta = \tau/\epsilon$ approaches both 0 and $\infty$.

**Proof.** Of course, by the scaling property Proposition 14.4(2), it suffices to work with noise level $\epsilon = 1$. The continuity property (5), applied to both numerator and denominator in $R_S(\eta)$ shows that the latter is a continuous function of $\eta \in (0, \infty)$. Thus it suffices show show that the limiting values of $R_S(\eta)$ are bounded (in fact equal to 1) for $\eta$ small and large.

For small $\eta$, the argument leading to Corollary 3.9 showed that the choice
\[ \lambda = \sqrt{2\log \eta^{-\beta}} \text{ implies } \sup_{\lambda \leq \eta} R_S(\lambda, \pi) \leq \eta^p(2 \log \eta^{-\beta})^{1/2}(1 + o(1)). \]
The two point prior argument of the previous section established that $\beta_p(\eta)$ was bounded below by the same quantity, so $R_S(\eta) \to 1$ as $\eta \to 0$.

For large $\eta$, use the trivial bound $\beta_{S, p}(\eta, 1) \leq 1$, along with the property (1) that $\beta_p(\eta)$ is decreasing in $\eta$ to write
\[ R_S(\eta) \leq 1/\beta_\infty(\eta) = 1/\rho_N(\eta, 1) \lesssim 1 \]
as $\eta \to \infty$. This completes the proof of (i).

For part (ii), we again use properties of the bounded normal mean minimax risk $\rho_N(\eta, 1)$ discussed in Chapter 7. Indeed, together with (14.43), we conclude for any $p$ and for $\eta \geq 1$, that $R_S(\eta) \leq 1/\rho_N(1, 1) = 2.22$. For $\eta \leq 1$ and now using $p \geq 2$, we use $\beta_{S, p}(\eta) \leq \eta^2$ (compare (14.24)) to write
\[ R_S(\eta) \leq \eta^2/\beta_\infty(\eta) = \eta^2/\rho_N(\eta, 1) \leq 1/\rho_N(1, 1) = 2.22, \]
where the last inequality uses (7.18). \qed

Comment on numerics?

14.3. Minimax Bayes Risk for $n$-dimensional data.

We are at last able to return to the estimation of a $n$-dimensional parameter constrained to an $\ell_p$ ball and observed in white Gaussian noise of scale $\epsilon_n$ - compare model (14.21) and (14.22).

Our object is to study the asymptotic behavior of the minimax risk $R_N$ as $n$, the number of parameters, increases. We regard the noise level $\epsilon = \epsilon_n$ and ball radius $C = C_n$ as known functions of $n$. This framework accommodates a common feature of statistical practice: as the amount of data increases (here thought of as a decreasing noise level $\epsilon$ per parameter), so too does the number of parameters that one may contemplate estimating.

If there were no prior constraints, $\Theta = \mathbb{R}^n$, then the unmodified raw data would give a minimax estimator $\hat{\theta}(y) = y$, with unconstrained minimax risk equalling $n\epsilon^2$. Asymptotically, we will see that $R_N$ depends on the size of $\Theta_{n, p}(C)$ through $n\epsilon^2$ times the dimension normalized radius $\eta_n = n^{-1/p}(C/\epsilon)$. This may be interpreted as the maximum scalar multiple in standard deviation units of the vector $(1, \ldots, n)$ that is contained within $\Theta_{n, p}(C)$. Alternatively, it is the average signal to noise ratio measured in $\ell_p$-norm: $n^{-1} \sum \delta_i/|\delta_i|^{1/p} \leq n^{-1/p}(C/\epsilon)$.

The asymptotics of $R_N$ will be evaluated by the Bayes minimax approach of Section 6.4. This approach allows reduction to the basic one dimensional
Bayes minimax problem studied in the previous section. More specifically, we choose a collection of priors \( \mathcal{M}_n \) depending on \( n \) using the condition that the vector \( \tau \) of marginal moments \( \tau_i(\pi) = (E_r|\theta_i|^p)^{1/p} \) belong to \( \Theta_n \):

\[
\mathcal{M}_n = \{ \pi(d\theta) : E_n \sum_{i=1}^{n} |\theta_i|^p \leq C_n^p \}.
\]

The set \( \mathcal{M}_n \) contains all point masses \( \delta_\theta \in \Theta_n \), is convex and is easily shown to be weakly compact. Thus we have the upper bound

\[
R_N(\Theta_n) \leq B(\mathcal{M}_n) = \sup \{ B(\pi), \pi \in \mathcal{M}_n \} = B_{n,p}(C_n, \epsilon_n),
\]

where the equality uses the minimax theorem 6.6. For any \( \pi \in \mathcal{M}_n \), we construct a prior \( \tilde{\pi} \) that is more difficult for Bayes risk (in the sense that \( B(\tilde{\pi}) \geq B(\pi) \)) as the product of the univariate marginal distributions \( \pi_i \) of \( \pi \). Since membership of \( \mathcal{M}_n \) is determined only by these marginal distributions, \( \tilde{\pi} \in \mathcal{M}_n \), and exactly as in the Pinsker case (Section 8.2), the posterior variance of \( E(\theta_i|y_i) \) must be at least as large as that of \( E(\theta_i|y) \).

For a product prior \( \tilde{\pi} \), the Bayes risk becomes the sum of the component univariate Bayes risks: \( B(\tilde{\pi}) = \sum B(\pi_i) \), so that it is possible to maximise the univariate terms separately subject to the overall moment constraint. Hence, we may now make use of the univariate Bayes minimax quantities \( \beta_p(\tau, \epsilon) \) by writing

\[
B_{n,p}(C_n, \epsilon_n) = \sup \{ \sum_{i=1}^{n} \beta_p(\tau_i, \epsilon_n) : \sum_{i=1}^{n} \tau_i^0 \leq C_n^0 \}.
\]

Since \( \beta_p(\tau, \epsilon) \) is a concave function of \( \tau^0 \) (Proposition 14.4, (3)), the maximum in this expression occurs when all \( \tau_i^0 = n^{-1}C_n^0 \). Using now also the scale invariance (Proposition 14.4, (2)), we arrive at

\[
B(\mathcal{M}_n) = B_{n,p}(C_n, \epsilon_n) = n^{-1/p}C_n, \epsilon_n = n \epsilon_n^2 \beta_p(\eta_n).
\]

where \( \eta_n = n^{-1/p}(C_n/\epsilon_n) \) denotes the normalized radius of the \( \ell_p \)-ball introduced earlier.

**Example 14.3 continued.** Let us return to our original example in which \( p = 1 \), the noise \( \epsilon_n = 1/\sqrt{n} \), and the radius \( C_n = 1 \). Thus \( \eta_n = n^{-1} \cdot \sqrt{n} = n^{-1/2} \). It follows that

\[
R_N(\Theta_{1,n}) \leq B_{n,1}(C_n, \epsilon_n) = n \cdot (1/n) \cdot \beta_p(1/\sqrt{n}) \sim (\log n/n)^{1/2},
\]

and the next theorem will show that this rate and constant are optimal. Recall, for comparison, that \( R_k(\Theta_{1,1}, \epsilon_n) = 1/2 \).

The main result of this chapter asserts that the Bayes minimax risk is asymptotically equivalent to the non linear minimax risk \( R_N(\Theta) \). We can thus incorporate the small \( \eta \) results from the previous section. In particular, except in highly sparse conditions, the least favorable distribution for \( R_N(\Theta) \) is essentially found by drawing \( n \) i.i.d rescaled observations from the least favorable distribution \( \pi_p(\eta_n) \) for \( \mathcal{M}_p(\eta_n) \).
14.12. Theorem. Suppose that \( 0 < p \leq \infty \) and \( \eta_n = n^{-1/p}(C/\epsilon) \to \eta \in [0, \infty] \). If either (i) \( p \geq 2 \), or

\[
0 < p < 2 \quad \text{and} \quad (\epsilon/C)^2 \log n(\epsilon/C)^p \to 0,
\]

then

\[
R_N(\Theta_n, \epsilon) \sim n^{2} \beta_{p}(\eta_n, 1) \quad \text{as } n \to \infty.
\]

**WHAT FOLLOWS SEEMS REPETITIVE - INCLUDE IN TEXT?** In specific cases, there is more information on the minimax risk \( R_N \) and asymptotically least favorable estimators \( \hat{\theta}_N \):

(i) \( \eta_n \to \infty \) \( R_N \sim n^{2} \quad \hat{\theta}_N(y) = y \),

(ii) \( \eta_n \to \eta \in (0, \infty) \) \( R_N \sim n^{2} \beta_{p}(\eta) \quad \hat{\theta}_N(y) = d \beta_{\pi_n}(\eta_n) \left( y_k / \epsilon \right) \),

(iii, a) \( \eta_n \to 0, p \geq 2 \) \( R_N \sim n^{2} \eta_n \quad \hat{\theta}_N(y) = 0 \),

(iii, b) \( \eta_n \to 0, p \leq 2 \) \( R_N \sim n^{2} \eta_n^{-1}(2 \log \eta_n^{-p})^{1-p/2} \quad \hat{\theta}_N(y) = \text{sgn} \ y_i \left( \left| y_k \right| - \lambda_n \epsilon \right) \),

\[
\lambda_n^2 = 2 \log \eta_n^{-p}.
\]

When \( \eta_n \to 0 \), for \( p \geq 2 \), the asymptotically least favorable prior is \( \pi_{1n} = \left( \delta_{\eta_n} + \delta_{-\eta_n} \right) / 2 \). For \( p \leq 2 \), the asymptotically least favorable prior is \( \pi_{1n} = (1 - \alpha) \delta_0 + \alpha \delta_\mu \), where \( \alpha = \alpha_n \) and \( \mu = \mu_n \sim \sqrt{2 \log \epsilon} \) are determined from

\[
\alpha \mu^p = \eta^p \quad \phi(a_n + \mu) = \alpha \phi(a_n),
\]

where \( a_n = a(\eta_n) \to \infty \), but \( a_n^2 = o(\log \eta_n^{-p}) \).

**Proof.** The approximation (14.47) follows from (14.45) once we establish the asymptotic equivalence of frequentist and Bayes minimax risks. The particular expressions for \( R_N \) and the structure of the asymptotically least favorable priors and estimators follow from the results of the previous subsections on the univariate quantity \( \beta_{p}(\eta, 1) \).

Asymptotic equivalence of \( R_N \) and \( B \). To show that the Bayes minimax bound is asymptotically sharp, we construct a series of asymptotically least favorable priors \( \pi_n \) that essentially concentrate on \( \Theta_n \). More precisely, following the recipe of Chapter 6.4, for each \( \gamma < 1 \) we construct priors \( \pi_n \) satisfying

\[
B(\pi_n) \geq \gamma B_{n, p}(\gamma C_n, \epsilon_n)
\]

\[
\pi_n(\Theta_n) \to 1, \text{ and}
\]

\[
E_{\pi_n} \{ \| \hat{\theta}_n' \|^2 + \| \theta' \|^2, \Theta_n \} = o(B_{n, p}(C_n, \epsilon_n))
\]

where \( \hat{\theta}_n(y) = E_{\pi_n}(\theta' \in \Theta_n, y) \).

We will always define \( \pi_n \) by i.i.d rescaled draws from a univariate distribution \( \pi(d\mu) \) on \( \mathbb{R} \) (in some cases \( \pi_1 = \pi_{1n} \) depends on \( n \)): thus \( \pi_n(d\theta) = \).
\[ \pi_n^*(d\theta/\epsilon_n). \] In this case, using (14.45), condition (14.48) can be reexpressed as

(14.51) \[ B(\pi_n) \geq \gamma \beta_\rho(\gamma \eta_n), \]

and property (14.49) may be rewritten as

\[ \pi_n(\Theta_n) = P_{\pi_n} \{ n^{-1} \sum |\mu_i|^p \leq \eta_n^p \}. \]

Suppose first that \( \eta_n \to \eta \in (0, \infty) \). Given \( \gamma \), there exists \( M < \infty \) and a prior \( \pi \) in \( \mathcal{M}_\rho(\gamma \eta) \) supported on \([-M, M]\) whose Bayes risk satisfies (14.51). Noting \( E_\pi[\mu]^p \leq \gamma^p \eta^p \) and that \( |\mu_i| \leq M \), property (14.49) follows from an exponential inequality applied to the i.i.d. draws from \( \pi \). Finally, again since \( |\mu_i| \leq M \), the left side of (14.50) is bounded by a fixed multiple of \( 2\gamma^2 \pi_n(\Theta_n) \), while \( B_{n,p}(C_n, \epsilon_n) \) is of exact order \( n_\eta^2 \), and so (14.50) follows from (14.49).

Now suppose that \( \eta_n \to 0 \). The case \( p \geq 2 \) is straightforward: we know from the univariate case that the symmetric two point priors \( \pi_n = \pi_{\eta_n} = (\delta_{\eta_n} + \delta_{-\eta_n})/2 \) are asymptotically least favorable, so \( \pi_n \) satisfies (14.51) for large \( n \). The corresponding measure \( \pi_n \) is already supported on \( \Theta_n \), so the remaining conditions are vacuous here.

The case \( p < 2 \) is more interesting. Given \( \gamma < 1 \), let \( \pi_n = \pi_{\alpha_n, \mu_n} \) be the symmetric three point prior of Theorem 14.5 corresponding to \( \gamma \eta_n \) with this substitution, \( \alpha_n \) and \( \mu_n \) are determined by (14.39).

Let \( N_n \) be a Binomial\( (n, \alpha_n) \) random variable that counts the number of non-zero components of \( \mu_n \) drawn from \( \pi_n^* \): Since \( \sum |\mu_i|^p = N_n \mu_n^p \), it follows that

(14.52) \[ \{ \theta \in \Theta \} = \{ N_n \leq n \eta_n^p / \mu_n^p \}. \]

Using (14.39),

(14.53) \[ EN_n = n \alpha_n = n \gamma^p \eta_n^p / \mu_n^p \]

Note that \( n \eta_n^p / \mu_n^p = (C_n / \epsilon_n \mu_n)^p \to \infty \) by our assumption that \( (\epsilon_n \mu_n / C_n)^2 \sim 2(\epsilon_n / C_n)^2 \log n(\epsilon_n / C_n)^p \to 0 \). This establishes (14.49).

The proof of (14.50) also follows from the fact that \( n \alpha_n \to \infty \), but is postponed to the appendix. \( \square \)

The role of the assumption that \( (\epsilon / C)^2 \log n(\epsilon / C)^p \to 0 \) is to ensure that \( EN_n \to \infty \). In other words, that \( \Theta_n \) has large enough radius that the least favorable distribution in the Bayes minimax problem generates an infinite number of sparse spikes. Without this condition, asymptotic equivalence of Bayes and frequentist minimax risks can fail. For an example, return to the case \( p = 1, \epsilon = n^{-1/2} \), but now with small radius \( C_n = n^{-1/2} \). We have \( \eta_n = n^{-1} \) and hence \( B(C_n, \epsilon_n) \sim n^{-1} \sqrt{2 \log n} \). However, the linear minimax risk is smaller: \( R_l \sim n^{-1} \eta_n^2 \sim n^{-1} \), and of course the non-linear minimax risk \( R_N \) is smaller still.

In this case \( EN_n = n \alpha_n = n \eta_n / \mu_n = 1 / \mu_n \to 0 \), since \( \mu_n \sim \sqrt{2 \log n} \).
The proof of equivalence thus demonstrates the existence of three different regimes for the least favorable distribution.

(i) **Dense:** \( \eta_n \to \eta > 0 \). The least favorable distribution \( \pi_p(\eta) \in \mathcal{M}_p(\eta) \) has high probability of yielding non-zero values \( \mu_i \).

(ii) **Sparse:** \( \eta_n \to 0, EN_n \to \infty \). The least favorable sequence of two point distributions have an atom at 0 with probability increasing to 1, but still produce on average \( EN_n = n\alpha_n \to \infty \) non-zero "spikes" at \( \mu_n = \sqrt{2\log \eta_n^{-p}} \) as \( n \to \infty \).

(iii) **Highly sparse:** \( \eta_n \to 0, \limsup_n EN_n < \infty \). In this case the signal to noise ratio \( n/\epsilon_n \) is so small that only a finite number of non-zero spikes appear. The practical importance of this case has been highlighted by Mallat in a satellite image deconvolution/denoising application. Hence we devote the next section to its analysis.

### 14.4. Minimax Risk in the Highly Sparse Case

The moment constrained Bayesian approach to evaluating minimax risk \( R_N(\Theta_n) \) fails in the highly sparse case because i.i.d. samples from the least favorable prior for the moment space \( \mathcal{M}_p(\eta_n) \) do not concentrate on \( \Theta_n \) even asymptotically. In turn, this is a consequence of the small size of \( \Theta_n \), which entails that the expected number of non-zero ‘spikes’ is finite.

Thus in the highly sparse case, to obtain sharp lower bounds, we are forced to work with priors that concentrate entirely on \( \Theta_n \). We therefore abandon independence priors, and use instead exchangeable priors with a fixed number of non-zero components.

The next theorem fills the gap left by the Minimax Bayes theorem 14.12. Let \( \{x\} \) denote the fractional part of \( x \).

**14.13. Theorem.** (??) Suppose that \( 0 < p < 2 \) Let \( \lambda_n = \sqrt{2\log n} \) and suppose that \( \gamma_n = C_n/(\epsilon_n \lambda_n) \to \gamma \in [0, \infty) \). Then

\[
R_N(\Theta_n,p(C_n),\epsilon_n) \sim \begin{cases} 
C_n^2 & \text{if } \gamma \leq 1, \\
C_n^p(\epsilon_n \lambda_n)^{2-p} \{1 - P(\gamma_n^p)/\gamma_n^p\} & \text{if } \gamma \geq 1 
\end{cases}
\]

where

\[
P(r) = \{r\} - \{r\}^{2/p}
\]

is a 1-periodic function vanishing on integers. In all these cases, soft thresholding at \( \lambda_n = \sqrt{2\log n} \) is asymptotically minimax.

**Remarks.**

1. The ?? in the statement refers to the fact that the lower bound has been checked only for \( \gamma \leq 1 \) and \( \gamma \in \mathbb{N} \). We conjecture however that the result is correct as stated.

2. The condition (14.46) of Theorem 14.12 is easily shown to be equivalent to \( \gamma_n = C_n/\epsilon_n \sqrt{2\log n} \to \infty \).
(3) In this setting, $2\log \eta_n^{-\gamma} = 2\log n - 2p \log(C_n/\epsilon_n) \sim 2\log n = \lambda_n^2$
and so the Minimax Bayes expression in the $p < 2$ sparse case, $n\epsilon_n^2\eta_n(2\log \eta_n^{-\gamma})^{1-\gamma/2} \sim C_n^p(\epsilon_n\lambda_n)^{2-\gamma}$, agrees with the present result
when $\gamma \to \infty$, as expected. For finite $\gamma$, however, it is strictly larger except when $\gamma$ is an integer.

(4) In particular, note that when $\gamma < 1$ the bound $\epsilon_n^2\eta_n^{2-\gamma} \lambda_n^{2-\gamma}$ predicted by the Minimax Bayes method is too large and hence incorrect as the limiting value of $R_N$.

(5) The parameter $\gamma_n$ measures the number of 'spikes' of height about $\epsilon_n \sqrt{2\log n}$ that appear in the least favorable configuration for $\Theta_n, (\epsilon_n)$.

(6) The $\gamma \geq 1$ formula actually applies for $\gamma < 1$ also, but the form given is clearer.

**Upper Bound.** By scaling, it suffices to carry out the proof in the unit noise case $\epsilon_n = 1$. We use the risk bound (11.8) for soft thresholding with $
abla \Theta = \lambda_n = \sqrt{2\log n}$ and maximize over $\Theta$:

$$R_N(\Theta) \leq \sup_{\Theta} \sum_i r_S(\lambda_n, \theta_i) \leq n r_S(\lambda_n, 0) + \sup_{\Theta} \sum_i \theta_i^2 \wedge \lambda_n^2 + 1.$$

Using (11.7), $nr_S(\lambda_n, 0) \leq c_1 n \phi(\lambda_n)/\lambda_n^3 \leq c_2/(\log n)^{3/2} \to 0$. The upper bound now follows from the next lemma, on setting $\lambda^2 = \lambda_n^2 + 1 \sim \lambda_n^2$ as $n \to \infty$.

14.14. **Lemma.**

(14.56) \[
\sup_{\|\theta\|_p \leq C} \sum_{i=1}^n \theta_i^2 \wedge \lambda^2 = \begin{cases} C^2 & \text{if } C \leq \lambda \\
\lambda^2 & \text{if } C > n^{1/p} \lambda \\
\lambda^2 & \text{if } C \geq n^{1/p} \lambda \end{cases} \]

where $\gamma = C/\lambda$ and $P(\cdot)$ is the 1-periodic function defined at (14.55).

**Proof.** If $C \leq \lambda$, then the $\ell_p$ ball is entirely contained in the $\ell_\infty$ cube of side $\lambda$, and the maximum of $\sum \theta_i^2$ over the $\ell_p$ ball is attained at the spike $\theta^* = C(1, 0, \ldots, 0)$ or permutations. This yields the first bound in (14.56). At the other extreme, if $C \geq n^{1/p} \lambda$, then the $\ell_\infty$ cube is contained entirely within the $\ell_p$ ball and the maximum of $\sum \theta_i^2$ is attained at the dense configuration $\theta^* = \lambda(1, \ldots, 1)$.

If $\lambda < C < n^{1/p} \lambda$, the worst case vectors are subject to the $\ell_\infty$ constraint and are then permutations of the vector $\theta^* = (\lambda, \ldots, \lambda, 0, \ldots, 0)$ with $n_0$ components of size $\lambda$ and the remainder $\mu$ being determined by the $\ell_p$ condition:

$$n_0 \lambda^p + \mu^p = C^p.$$

To verify that this is indeed the worst case configuration, change variables to $u_i = \theta_i^p$: the problem is then to maximize the convex function $u \to \sum u_i^{2/p}$ subject to the convex constraints $\|u\|_1 \leq C$ and $\|u\|_\infty \leq \lambda$. This forces an extremal solution to occur on the boundary of the constraint set and to
have the form described.] Thus \( n_0 = \left[ C^p \right. / \left. \lambda^p \right] \) and \( \mu^p = \lambda^p \left[ C^p \right. / \left. \lambda^p \right] \). Setting \( \gamma^p = C^p / \lambda^p \), we obtain

\[
\sum \theta_i^2 + \lambda^2 = n_0 \lambda^2 + \mu^2
= \lambda^2 \gamma^p + \lambda^2 \left[ 2 \gamma^p \right. / \left. p \right]
= \lambda^2 \gamma^p C^p - \lambda^2 \left[ \gamma^p \right. \right. / \left. \gamma^p \right] \cdot \gamma^p,
\]

which completes the proof. \( \square \)

**Lower Bound.** We use the consequence (6.17) of the minimax theorem: \( R_{\lambda}(\Theta) = \sup \{ B(\pi) : \sup \pi \subset \Theta \} \). The priors will be exchangeable on permutations of the least favorable vectors described in the proof of Lemma 14.56.

We provide details for the case \( \gamma < 1 \) for which there is a single non-zero component: a “needle in a haystack”. Thus, let \( I \) be uniform on \( \{ 1, \ldots, n \} \) and define \( \pi_n \) by setting \( \theta = \delta_n e_I \), where to ensure that \( \pi_n \) concentrates on \( \Theta_n \) we require that \( \delta_n \leq C_n \). By symmetry, and then decomposing on the event \( \{ I = 1 \} \), we have

\[
B(\pi_n) = n E_\pi [\theta(\pi_n) - \theta] \cdot 2
= (n - 1) E_\pi [\theta(\pi_n) - \theta] / I \neq 1 + E_\delta_n e_I (\hat{\theta}(\pi_n) - \delta_n)^2,
\]

where

\[
\hat{\theta}(\pi_n) = E(\theta|y) = \delta_n P(I = 1|y) = \delta_n p_{1n}(y).
\]

We use the earlier heuristic that the primary contribution to the Bayes risk comes from the error made by \( \hat{\theta}_n \) when \( \theta_1 \) in fact takes the rare value \( \delta_n \): this yields

\[
B(\pi_n) \geq \delta_n^2 E_\delta_n e_I |p_{1n}(y) - 1|^2
\]

and our strategy will be to show that \( p_{1n}(y) \to 0 \) for all \( y \) and use the dominated convergence theorem to conclude that \( B(\pi_n) \geq \delta_n^2 (1 + o(1)) \). From Bayes formula, we have

\[
p_{1n}(y) = P(I = 1|y) = \frac{P(I = 1) \phi(y - \delta_n e_1)}{\sum_j P(I = j) \phi(y - \delta_n e_j)} = \frac{e^{\delta_n y_1}}{\sum_j e^{\delta_n y_j}}.
\]

We now write this ratio in terms of i.i.d. standard Gaussian variables \( z_i \): of course \( y_1 = \delta_n + z_1 \) and \( y_j = z_j \) for \( j \geq 2 \). We obtain

\[
p_{1n}(y) = e^{\delta_n z_1} + (n - 1) e^{-\delta_n^2 / 2 W_{n-1}} - W_{n-1},
\]

where \( W_{n-1} = e^{-\delta_n^2 / 2 (n - 1)} - 1 \sum_2^n e^{\delta_n z_i} \to 1 \) a.s. as \( n \to \infty \) (since \( E e^{\delta z} = e^{\delta^2 / 2} \)). If \( \delta_n \) is slightly smaller than \( \sqrt{2 \log n} \), for example \( \delta_n \leq \sqrt{2 \log n} \), then \( n e^{-\delta_n^2 / 2} / e^{\delta_n z_1} \to \infty \) for any fixed value of \( z_1 \) and so \( p_{1n}(y) \to 0 \).

**Remarks on weak \( \ell_p \) balls**
Appendix: Further details

2°. The mean squared error of a thresholding rule \( \hat{\omega}(x, \lambda) \) (either hard or soft) may be written
\[
E_\omega(\hat{\omega}(x, \lambda) - \mu)^2 = \int_{-\infty}^{\infty} (\hat{\omega}(x, \lambda) - \mu)^2 \phi(x - \mu) dx.
\]

Breaking the range of integration into regions \((-\infty, -\lambda), [-\lambda, \lambda], \text{ and } (\lambda, \infty)\) to match the thresholding structure, one obtains the following basic mean squared error formulas:
\[
\begin{align*}
(14.57) & \quad r_S(\lambda, \mu) = 1 + \lambda^2 + (\mu^2 - \lambda^2 - 1)[\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)] \\
& \quad - (\lambda - \mu)\phi(\lambda + \mu) - (\lambda + \mu)\phi(\lambda - \mu), \\
(14.58) & \quad r_H(\lambda, \mu) = \mu^2[\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)] + \Phi(\lambda - \mu) + \Phi(\lambda + \mu) \\
& \quad + (\lambda - \mu)\phi(\lambda - \mu) + (\lambda + \mu)\phi(\lambda + \mu)
\end{align*}
\]
where \( \phi \) and \( \Phi \) denote the standard Gaussian density and cumulative distribution functions respectively, and \( \Phi(x) = 1 - \Phi(x) \).

3°. Proof of Proposition 14.8 Since the risk function \( \mu \to r_S(\lambda, \mu) \) of soft thresholding is an even function of \( \mu \in \mathbb{R} \) and is increasing in \( \mu \geq 0 \), the maximum of \( B(\hat{\omega}_\lambda, \pi) = \int r(\lambda, \mu)\pi(d\mu) \) over \( \mathcal{M}_\mu(\eta) \) in fact occurs on the convex subset of distributions \( \pi \) that are supported in \([0, \infty)\) and achieve the moment bound: \( \int \mu^p\pi(d\mu) = \eta^p \). Since \( \pi \to B(\hat{\omega}_\lambda, \pi) \) is linear, we can confine attention to the extreme points of this subset, which are exactly those two point priors \( \pi_{\mu_0, \mu_1} = (1 - \alpha)\delta_{\mu_0} + \alpha\delta_{\mu_1} \) for which the moment condition becomes \( (1 - \alpha)\mu_{0}^p + \alpha\mu_{1}^p = \eta^p \). We change variables to \( \kappa = \mu^p \) so that the moment condition itself assumes a linear form. Writing \( B(\lambda, \eta) \) for the left side of (14.31), the optimization then becomes
\[
B(\lambda, \eta) = \sup \{ (1 - \alpha)r(\lambda, \kappa_{0}^{1/p}) + \alpha r(\lambda, \kappa_{1}^{1/p}) : (1 - \alpha)\kappa_{0} + \alpha\kappa_{1} = \eta \}.
\]

14.15. Lemma. For \( 0 < p \leq 2 \), the function \( \kappa \to r(\lambda, \kappa^{1/p}) \) is convex-concave on \((0, \infty)\).

By convex-concave, we mean that there is a constant \( \kappa_c > 0 \) such that the function is convex on \((0, \kappa_c)\) and concave on \([\kappa_c, \infty)\). We may conclude from Lemma 14.15 (whose proof is outlined below) that it suffices to take \( \kappa_0 = 0 \) in (14.59) and that there exists a unique \( \mu_\lambda = \kappa_\lambda^{1/p} \) attaining the supremum, which in the original variables now becomes
\[
B(\lambda, \eta) = \sup_{\mu_\lambda \geq \eta} r(\lambda, 0) + (\eta/\mu)^p[r(\lambda, \mu) - r(\lambda, 0)].
\]
Differentiation with respect to \( \mu \) shows that the equation for this maximum is given by (14.32). Inequality (??) showed that \( r(\lambda, \mu) - r(\lambda, 0) \leq \mu^2 \); applying this at the particular point \( \mu = \mu_\lambda \) we arrive at (14.31).
To obtain more explicit information about \( \mu_\lambda \), use formula (11.5) for the derivative of the risk of soft thresholding to write
\[
    r_\mu(\lambda, \mu) = 2\mu \Phi_\mu(I_\lambda),
\]
where we have put \( \Phi_\mu \) for the measure corresponding to the \( N(\mu, 1) \) distribution. We see that (14.32) is equivalent to solving for \( \mu = \mu_\lambda \) in
\[
    R(\mu) := \frac{\int_0^\mu s \Phi_\mu(I_\lambda)ds}{\Phi_\mu(I_\lambda) \int_0^\mu s ds} = \frac{2}{p}
\]
Set \( \mu = \lambda + u \); since \( \Phi_{\lambda+u}(I_\lambda) = \Phi(-u) - \Phi(-2\lambda - u) \), as \( \lambda \to \infty \), we have, uniformly in \( u \) in compact intervals, that \( R(\lambda + u) \to 1/\Phi(u) \). Solving \( \phi(u_p) = p/2 \) gives \( u_p = \Phi^{-1}(p/2) \) and \( \mu_\lambda \sim \lambda + u_p \) as \( \lambda \to \infty \).

4°. Comments on the proof of Lemma 14.15. 1. General \( p \in (0, 2] \). Some calculus shows that
\[
    p^2 \kappa^2 D^2_{\mu P}(\lambda, \kappa^{1/p}) = \mu^2 \Phi_\mu(I_\lambda) \{(2 - p)\mu^{-1} + D_\mu \log \Phi_\mu(I_\lambda)\}.
\]
The lemma is established by observing that the quantity in parentheses is strictly decreasing, as follows from the assertion
\[
    \mu \to \log \int_{-\lambda}^\lambda \phi(x - \mu)dx \text{ is concave on } (0, \infty).
\]
For this I have only a proof by tedious calculus, which is omitted: it seems that a short proof should exist.

2. There is also a simple sign change argument using Stein’s unbiased estimate of risk when \( p = 1 \).

5°. Proof of (14.50). First we note that (14.52) and (14.53) imply that on \( \Theta, N_n \leq E N_n/\gamma^p \), and so
\[
    \|\hat{\theta}_n\|^2 \leq E\{\|\theta\|^2 | \theta \in \Theta_n, y\} = c_n^2 \mu_n^2 E\{N_n | \theta \in \Theta, y\} \leq c_n^2 \mu_n^2 EN_n/\gamma^p.
\]
Thus
\[
    E_{\mu_n}\{\|\hat{\theta}_n\|^2 + \|\theta\|^2, \Theta_n^c\} \leq \gamma^{-p} c_n^2 \mu_n^2 EN_n + N_n, \Theta_n^c, \}
\]
whereas
\[
    B_{\alpha, p}(C_n, \epsilon_n) = n c_n^2 \alpha_n \mu_n^2 \Phi(a)(1 + o(1)) = c_n^2 \mu_n^2 \Phi(a) EN_n(1 + o(1)).
\]
Now (14.50) follows because \( P(\Theta_n^c) \to 0 \) and because \( E|N_n - EN_n|/EN_n \leq \sqrt{\text{Var}N_n}/EN_n \leq (n \epsilon_n)^{-1/2} \to 0 \) (consequence of \( n \epsilon_n \to \infty \).)

20
Part 6

MULTIRESOLUTION MODELS
CHAPTER 15

Sparsity, Weak $\ell_p$ and Thresholding

In this chapter, we explore various measures for quantifying sparsity and the connections among them. In the process, we will see hints of the links that these measures suggest with approximation theory and compression.

We first contrast traditional linear approximation with a version of non-linear approximation that greedily picks off the largest coefficients in turn. This is clearly related to counting the number $N(\theta, \epsilon)$ of coefficients of $\theta$ that exceed $\epsilon$ in size. We then take a more explicitly statistical point of view and relate the size of ideal risk to the non-linear approximation error. It turns out that ideal risk captures the idea of sparsity: ideal risk is small exactly when both $N(\theta, \epsilon)$ and the tail sum of small coefficients are small.

Thirdly, we look at the decay of individual ordered coefficients: this is expressed in terms of a weak-$\ell_p$ condition, which is related to the more familiar $\ell_p$ norms.

The intuitively natural connections between these viewpoints can be formalized as an equivalence of (quasi-)norms – Proposition 15.1.

The weak-$\ell_p$ spaces are a convenient tool in studying estimation over Besov bodies $\Theta^\alpha_{p,q}(C)$: Proposition 15.2 shows that $\Theta^\alpha_{p,q}(C)$ embeds in $w\ell_p(\alpha)$ with $p(\alpha) = 2/(2\alpha + 1)$, and that this is the smallest $w\ell_p$ space for which the embedding holds.

Finally, we use this weak $\ell_p$ embedding to give a fairly direct proof of the adaptive near minimaxity of $\sqrt{2 \log n}$ thresholding over almost the full array of Besov bodies $\Theta^\alpha_{p,q}(C)$. The key ingredients are the ideal risk oracle inequality for thresholding, and the equivalences between ideal risk decay and weak $\ell_p$ norms.

15.1. Approximation, Ideal Risk and Weak $\ell_p$

15.1.1. Non-linear approximation. Let $\{\psi_i, i \in \mathbb{N}\}$ be an orthonormal basis for $L_2[0,1]$, and consider approximating $f \in L_2[0,1]$ by a linear combination of basis functions from a subset $K \subset \mathbb{N}$:

$$P_K f = \sum_{i \in K} \theta_i \psi_i.$$

The coefficients $\theta_i = \langle f, \psi_i \rangle$, and we will not distinguish between $f$ and the corresponding coefficient sequence $\theta = \theta[f]$. Again, using the orthonormal
basis property, we have
\[ \|f - P_K f\|_2^2 = \sum_{i \in K} \theta_i^2. \]

The operator \( P_K \) is simply orthogonal projection onto the subspace spanned by \( \{\psi_i, i \in K\} \), and yields the best \( L_2 \) approximation of \( f \) from this subspace. In particular, \( P_K \) is linear, and we speak of best linear approximation.

Now consider the best choice of subset \( K \) of size \( m \): we have
\[ c_k^2(f) = \inf \{ \|f - P_K f\|_2^2 : \#(K) \leq k \} \]
\[ = \inf \{ \sum_{i \in K} \theta_i^2 : \#(K) \leq k \}. \]

Let \( |\theta|_{(1)} \geq |\theta|_{(2)} \geq \ldots \) denote the amplitudes of \( \theta \) in decreasing order. Then
\[ c_k^2(f) = c_k^2(\theta) = \sum_{l > k} |\theta(l)|^2, \]
and we call \( c_k(\theta) \) the compression numbers associated with \( \theta = \theta[f] \).

Let \( K_k(\theta) \) be the set of indices corresponding to the \( k \) largest magnitudes. Since \( K_k(f) \) depends strongly on \( f \), the best approximation operator \( Q_k f = P_{K_k(\theta)} f \) is non-linear: \( Q_k(f + g) \neq Q_k f + Q_k g \).

Thus the rate of decay of \( c_k(\theta) \) with \( k \) measures the rate of non-linear approximation of \( f \) using the best choice of \( k \) functions from the basis. To quantify this, define a sequence (quasi)norm
\[ |\theta|_{c,\alpha}^2 = \sup_{k \geq 0} k^{2\alpha} \sum_{l > k} |\theta(l)|^2, \]
with the convention that \( k^{2\alpha} = 1 \) when \( k = 0 \). In other words, \( |\theta|_{c,\alpha} = C \) means that \((\sum_{l > k} |\theta(l)|^2)^{1/2} \leq C k^{-\alpha} \) for all \( k \) and that \( C \) is the smallest constant with this property.

15.1.2. Ideal Risk. Return to estimation in a Gaussian white sequence model
\[ y_i = \theta_i + \varepsilon_i, \quad i \in I, \]
thought of, as usual, as the coefficients of the continuous Gaussian white noise model (3.4) in the orthonormal basis \( \{\psi_i\} \).

We now take an ideal subset regression point of view, seeking that subset \( K \) of variables which minimizes MSE for estimating \( f \). Using the variance-bias decomposition for projection estimation, we have
\[ \mathcal{R}(f, \varepsilon) := \inf_{K \subset I} \mathbb{E} \| P_K y - f \|^2 \]
\[ = \inf_k k c_k^2 + \inf_{K: \#(K) = k} \| P_K f - f \|^2 \]
\[ = \inf_k k c_k^2 + c_k^2(\theta). \]
The second and third forms show the connection between ideal estimation and non-linear approximation, and hint at the manner in which approximation theoretic results have a direct implication for statistical estimation.

Denote the objective function appearing in (15.3) by $S_k$: clearly the differences $S_k - S_{k-1} = e^2 - |\theta_k|^2$ are increasing and the unique minimum of $S_k$ occurs as $k \to |\theta_k|^2$ crosses the threshold $e^2$, at the index $k$ given by

$$N(e) = N(\theta, e) = \# \{i : |\theta_i| \geq e\},$$

Compare Figure ?? [in approximation theory, this is called the distribution function of $|\theta|$, a usage related to, but not identical with the standard statistical term.]

It is thus apparent that, in an orthonormal basis, the ideal subset estimation risk coincides with our earlier notion of ideal risk (Section 11.3):

$$R(f, e) = R(\theta, e) = \sum \theta_i^2 \wedge e^2$$

The ideal risk measures the intrinsic difficulty of estimation in the basis $\{\psi_i\}$. Of course, it is attainable only with the aid of an oracle who knows $\{i : |\theta_i| > e\}$.

In addition, (15.3) and (15.4) yield the decomposition

$$R(\theta, e) = N(e, \theta) e^2 + c_{N(\epsilon)}^2(\theta).$$

Thus, the ideal risk is small precisely when both $N(e)$ and $c_{N(\epsilon)}$ are. This has the following interpretation: suppose that $N(e, \theta) = k$ and let $K_k(\theta)$ be the best approximating set of size $k$. Then the ideal risk consists of a variance term $ke^2$ corresponding to estimation of the $k$ coefficients in $K_k(\theta)$ and a bias term $c_k^2(\theta)$ which comes from not estimating all other coefficients. Because the oracle specifies $K_k(\theta) = \{i : |\theta_i| > e\}$, the bias term is as small as it can be for any projection estimator estimating only $k$ coefficients.

The rate of decay of $R(\theta, e)$ with $e$ measures the rate of estimation of $\theta$ (or $f(\theta)$) using the ideal projection estimator for the given basis. Again to quantify this, we define a second sequence norm

$$|\theta|^2_{IR} = \sup_{\epsilon > 0} \epsilon^{-2r} \sum \theta_i^2 \wedge \epsilon^2.$$

In other words, $|\theta|_{IR} = B$ means that $R(\theta, e) \leq B^2 \epsilon^{-2r}$ for all $\epsilon > 0$, and that $B$ is the smallest constant for which this is true.

Identity (15.5) says that good estimation is possible precisely when $\theta$ compresses well in basis $\{\psi_i\}$, in the sense that both the number of large coefficients $N(e)$ and the compression number $c_{N(\epsilon)}^2$ are small. Proposition ?? below uses (15.5) to show that the compression number and ideal risk sequence quasinorms are equivalent.

15.1.3. Weak $\ell_p$ and Coefficient decay. A further natural measure of the “compressibility” of $\theta$ is the rate at which the individual magnitudes
$|\theta_j|$ decay. More formally, we say that $\theta = (\theta_i, i \in I) \in w\ell_{1,p}$, for $I = \{1, \ldots, n\}$ or $\mathbb{N}$ if the decreasing rearrangement $|\theta|_{(1)} \geq |\theta|_{(2)} \geq \ldots$ satisfies
\[
|\theta|_{(l)} \leq Cl^{-1/p} \quad l \in I,
\]
and we set $\|\theta\|_{w\ell_p}$ equal to the smallest such $C$. We write $w\ell_p(C)$ for the norm ball of radius $C$.

Smaller values of $p$ correspond to faster decay for the components of $\theta$. We will be especially interested in cases where $p < 1$, since these correspond to the greatest sparsity. **Comment on quasi-norms**

We note some relations satisfied by $w\ell_p(C)$.

1. $\ell_p(C) \subset w\ell_p(C)$. This follows from
\[
[k^{1/p}]|\theta|_{(k)}| \leq k : (1/k) \sum_{i=1}^{k} |\theta|_{(i)}^{p} \leq \|\theta\|_{\ell_p}^{p}.
\]

2. $w\ell_p \subset \ell_{p'}$ for all $p' > p$, since if $\theta \in w\ell_p$, then
\[
\sum_{i=1}^{\infty} |\theta|_{(k)}^{p'} \leq C^{p'} \sum_{i=1}^{\infty} k^{-i/p} = C^{p'} \zeta(p'/p).
\]

3. A plot of $N(\theta, \epsilon)$ versus $\epsilon$ shows that the maximum of $\epsilon \to \epsilon N(\theta, \epsilon)$ may be found among the values $\epsilon = |\theta|_{(k)}$. Hence we obtain
\[
(15.6) \quad \|\theta\|_{w\ell_p}^{p} = \sup_{\epsilon > 0} \epsilon^{p} N(\theta, \epsilon).
\]
This representation makes it easy to establish the quasinorm property. Indeed, since
\[
N(\theta + \theta', \epsilon) \leq N(\theta, \epsilon/2) + N(\theta', \epsilon/2),
\]
we have
\[
\|\theta + \theta'\|_{w\ell_p}^{p} \leq 2^{p}(\|\theta\|_{w\ell_p}^{p} + \|\theta'\|_{w\ell_p}^{p}),
\]
which in turn implies $\|\theta + \theta'\| \leq c_p(\|\theta\| + \|\theta'\|)$.

### 15.2. A Risk Lower Bound via Embedding of hypercubes.

We have just seen that $N(\theta, \epsilon)$, the number of coefficients with modulus larger than $\epsilon$, is a useful measure of sparsity. In combination with earlier minimax estimation results for hyperrectangles, it also leads to a simple, but important lower bound for minimax risk for solid, orthosymmetric $\Theta$ under squared error loss.

Suppose $\Theta$ is solid and orthosymmetric. For $\theta \in \Theta$ and $\epsilon > 0$, the very definition of $N(\theta, \epsilon)$ shows that $\Theta$ contains a hypercube $\Theta(\epsilon)$ with center $0$, side length $2\epsilon$ and dimension $N(\theta, \epsilon)$. Let
\[
(15.7) \quad N(\Theta, \epsilon) := \sup_{\theta \in \Theta} N(\theta, \epsilon)
\]
denote the maximal dimension of a hypercube embedded in Θ of side 2ε
centered at 0. This leads, using Proposition 7.6 and then the scale invariance
(7.10), to a lower bound on the minimax risk:
\[ R_N(Θ, ε) ≥ R_N(Θ(ε), ε) = N(Θ, ε)ρ_N(ε, ε) \]
\[ = ρ_N(1, 1)N(Θ, ε)ε^2. \]

15.3. Quasi-norm equivalences

In preceding subsections, we have defined three quantitative measures
of the sparseness of a coefficient vector θ.

(a) \( |θ|_{c, α} \) as a measure of the rate α of non-linear \( ℓ_2 \) approximation of
θ using a given number of coefficients,

(b) \( |θ|_{IR, r} \) as a measure of the rate r of mean squared error decrease in
ideal statistical estimation of θ in the presence of noise of scale \( ε \), and

(c) \( |θ|_{w, p} \) as a measure of the rate 1/p of decay of the individual coeffi-
cients \( |θ|_{(l)} \).

We now show that these measures are actually equivalent, so long as one
makes the calibrations
\[ r = 2α/(2α + 1), \quad p = 2/(2α + 1) \]
and note the trivial consequence
\[ p = 2(1 - r). \]

15.1. Proposition. Let \( α > 0 \), and suppose that \( r \) and \( p \) are given by
(15.10). Then
\[ 3^{-1/p}|θ|_{w, p} ≤ |θ|_{c, α} ≤ |θ|_{IR, r}^{2/p} ≤ (\frac{2}{2^p})^{1/p}|θ|_{w, p}. \]

Proof. We establish the inequalities proceeding from right to left in
(15.11). By rearrangement we may suppose without loss of generality that
θ is positive and decreasing, so that \( θ_k = |θ|_{(k)} \).

1°. Suppose first that \( C = |θ|_{w, p} \), so that \( θ_k ≤ CK^{-1/p} \). Hence
\[ \sum \theta_k^2 \wedge t^2 ≤ \sum_1 C^2 k^{-2/p} \wedge t^2 ≤ \int_0^∞ (Cu^{-1/p})^2 \wedge t^2 du \]
\[ = u_1 t^2 + \frac{p}{2 − p} C^2 u_1^{-1+2/p} = (1 + \frac{p}{2 − p})C^{-1/p} t^{2p}. \]

Here \( u_1 = C^{-1/p} \) is the point of balance in the pairwise minimum. Hence
\( |θ|_{IR} = \sup t > 0 t^{-2r} \sum \theta_k^2 \wedge t^2 ≤ 2^{−p}|θ|_{w, p}. \)

(Remark on \( p/(2−p) \) as difference between weak and strong \( ℓ_p \)
norm minimax risks. Also FDR connections?)

2°. Now let \( C = |θ|_{IR, r} \), so that for all positive \( t \), \( t^{-2r} \sum \theta_k^2 \wedge t^2 ≤ C^2 \).
In particular, when \( t = θ_k \), we obtain, for all \( k ≥ 1 \),
\[ \theta_k^{-2r}[kθ_k^2 + c_1^2(θ)] ≤ C^2. \]
Hence $\theta^2_k \leq k^{-1} C^2$ and so
\[
c_k^2(\theta) \leq \theta^2_k C^2 \leq k^{-2r/p}(C^2)^{1+2r/p}.
\]
Since $2r/p = 2\alpha$, we conclude for every $k \geq 1$, that $k^{2\alpha} c_k^2(\theta) \leq C^{2(1+2\alpha)} = |\theta|_{1R}^{4/p}$. It remains to consider the exceptional case $k = 0$ : putting $t = \theta_1$ in the definition of $|\theta|_{1R,r}$, we find $c_0^2(\theta) \leq C^2 \theta_1^{2r}$ and also that $\theta^2_0 \leq C^2 \theta_1^{2r}$. Hence $\theta^2_0 \leq C^2$ and so $c_0^2(\theta) \leq C^{4/p}$, which completes the verification.

3. Let $C = |\theta|_{c,\alpha}$, so that $c_k^2(\theta) \leq C^2 k^{-2\alpha}$ for $k \geq 1$ and $c_0^2(\theta) \leq C^2$. This implies that $\theta^2_k \leq C^2$ and for $k \geq 2$ and $1 \leq r < k$ that
\[
\theta^2_k \leq (1/r) \sum_{k-r+1}^k \theta^2_r \leq C^2 / r(k-r)^{2\alpha}.
\]
Consequently, for all $k \geq 1$,
\[
|\theta|^2_{w\ell_p} = \sup_k k^{2r/p} \theta^2_k \leq c_\alpha C^2,
\]
where
\[
c_\alpha \leq \sup_{k \geq 2} \inf_{1 \leq r < k} \frac{k^{1+2\alpha}}{r(k-r)^{2\alpha}} \leq 3^{1+2\alpha} = 3^{2/p}.
\]

\[\square\]

15.4. Besov Bodies and weak $\ell_p$

We have seen that the weak $\ell_p$ quasi-norm measures the sparsity of a coefficient sequence $\theta$, with smaller $p$ corresponding to greater sparsity. If a parameter set $\Theta$ is contained within $w\ell_p$, then all elements $\theta \in \Theta$ satisfy the same decay estimate. Define the critical exponent of a set $\Theta$:
\[
p^*(\Theta) = \inf\{p : \Theta \subset w\ell_p\}
\]
The next result (Donoho (1993)) says essentially that $p^*(\Theta^p_q) = 2/(1+2\alpha)$. As a matter of notation, we note that $c^\alpha_{p,q}$ will denote a constant depending only on $\alpha$ and $p$, and not necessarily the same at each appearance.

15.2. Proposition. Suppose that $\alpha > 1/p - 1/2$ and that $p^* \in (0,2)$ is given. (i) If $\alpha \geq 1/p^* - 1/2$, then
\[
\Theta^\alpha_{p,q}(C) \subset w\ell_{p^*},
\]
(ii) in particular, $\Theta^\alpha_{p,q} \subset w\ell_{(\alpha+1/2)^{-1}}(c^\alpha_{p,q} \cdot C)$ but we also have
\[
\Theta^\alpha_{p,q} \nsubseteq w\ell_s \quad \text{if} \quad s < (\alpha+1/2)^{-1/2}.
\]

Proof. First, since $\Theta^\alpha_{p,q}(C) \subset \Theta^p_{p,\infty}(C)$, it suffices to establish part (i) for the latter, which consists of the product of $\ell_p$ balls $\|\theta_j\|_p \leq C^2^{-ja}$ for all $j \geq 0$ and $a = \alpha + 1/2 - 1/p > 0$ by assumption. For an $\ell_p$ ball it is clear that
\[
N(\ell_{n,p}(C),\epsilon) = \min\{n, C^p \epsilon^{-p}\}.
\]

215
Hence, for $\theta \in \Theta_{p,\infty}^\alpha(C)$,
\[ N(\theta, \epsilon) \leq 1 + \sum_j \min\{2^j, (C\epsilon^{-1}2^{-\alpha j})^p\}. \]

The terms in the sum have geometric growth and decay away from the maximum $j_0$ defined by equality between the two terms: thus $2^{j_0(\alpha + 1/2)} = C/\epsilon$. Hence, for all $\theta \in \Theta_{p,\infty}^\alpha(C)$,
\begin{equation}
\epsilon^s N(\theta, \epsilon) \leq c_{p,\alpha} \epsilon^s 2^{j_0(\epsilon)} = c_{p,\alpha} C^p(\alpha) \epsilon^{p(\alpha)} \leq \epsilon^s
\end{equation}
for all $\epsilon \leq 1$ so long as $p^s \geq p(\alpha) = (\alpha + 1/2)^{-1}$. Part (i) now follows from the characterization (15.6).

For the second part, consider the Besov shells $\Theta_{p,q}^{(j_0)} = \{\theta \in \Theta_{p,q}^\alpha(C) : \theta_{jk} = 0 \text{ unless } j = j_0\} \equiv \ell_{2^{-j_0}}(C2^{-j_0})$. Consider the shell corresponding to level $j_0$ determined above: since this shell belongs to $\Theta = \Theta_{p,q}^\alpha(C)$ for all $q$, we have
\begin{equation}
\epsilon^s N(\Theta, \epsilon) \geq \epsilon^s 2^{j_0(\epsilon)},
\end{equation}
and hence from (15.12), this is unbounded in $\epsilon$ if $s < (\alpha + 1/2)^{-1}$. \hfill \Box

Remarks.
1. $(1/p, \alpha)$ Diagram showing spaces embedding in $w\ell_{p^s}$.
2. Note that in the case $\alpha = 1/p - 1/2$, we have $a = 0$, and so
\[ \Theta_{p,p}^\alpha(C) = \{\theta : \sum_j \sum_k |\theta_{jk}|^p \leq C^p\} = \ell_p(C). \]
Note that there is no compactness here!
3. What happens to the embedding results at $\alpha^s = 1/p^s - 1/2$? For $q \leq p^s$ we have
\[ \Theta_{p,q}^{p^s}(C) \subset \Theta_{p,q}^{p^s}(C) = \ell_{p^s}(C) \subset w\ell_{p^s}(C). \]
It can also be seen that $\ell_{p^s}(C) \subset \Theta_{p^s,\infty}^{p^s}(C)$.
4. However, there is no containment relation between $w\ell_{p^s}(C)$ and $\Theta_{p^s,\infty}^{p^s}(C)$:
   (i) The vector $\theta$ defined by $\theta_{jk} = C\delta_{jk} \in \Theta_{p^s,\infty}^{p^s}(C)$ but is not in $w\ell_{p^s}(C)$
       for any $C$.
   (ii) The vectors $\theta_{jk}$ defined by $\theta_{jk}^{s} = \delta_{jk} C^{-1/p^s}$ for $k = 1, \ldots 2^j$
       are each in $w\ell_{p^s}(C)$, but $\|\theta_{jk}^{s}\|_{p^s,\infty} \sim C_{jk}^{1/p^s}$.

15.5. Adaptive minimaxity for $\sqrt{2\log n}$ thresholding

We combine the preceding results with properties of thresholding established in Chapter 11 to derive adaptive near minimaxity results for $\sqrt{2\log n}$ thresholding over Besov bodies $\Theta_{p,q}^\alpha(C)$. Consider the dyadic sequence model
\begin{equation}
y_{jk} = \theta_{jk} + \epsilon_{jk},
\end{equation}
and apply soft thresholding to the first \( n = \epsilon^{-2} = 2^J \) coefficients, using threshold \( \lambda_\epsilon = \sqrt{2 \log \epsilon^{-2}} = \sqrt{2 \log n} \):

\[
\hat{\theta}^{(U)}_{jk} = \begin{cases} 
\eta_S(y_{jk}, \lambda_\epsilon \epsilon) & j < J \\
0 & j \geq J.
\end{cases}
\]

(15.15)

Remark. Although not strictly necessary for the discussion that follows, we have in mind the situation of fixed equi-spaced regression: \( y_i = f(i/n) + \sigma e_i \) — compare (3.3). After a discrete orthogonal wavelet transform, we would arrive at (15.14), restricted to \( j < J = \log_2 n \), and with calibration \( \epsilon = \sigma n^{-1/2} \). The restriction of thresholding in (15.15) to levels \( j < J \) corresponds to what we might do with real data: namely threshold the \( n \) empirical discrete orthogonal wavelet transform coefficients.

Let \( \theta^{(n)} \) and \( \hat{\theta}^{(n)} \) denote the first \( n \) coordinates — i.e. \( (j, k) \) with \( j < J \) — of \( \theta \) and \( \hat{\theta} \) respectively. To compute a bound on the risk (mean squared error) \( \hat{\theta} \), we apply the soft thresholding risk bound (11.15) of Proposition 11.6 to \( \theta^{(n)} \). Since \( \theta_{jk} \equiv 0 \) except in these first \( n \) coordinates, what remains is a “tail bias” term:

\[
r(\hat{\theta}^{(U)}, \theta) = E_\theta \| \hat{\theta}^{(n)} - \theta^{(n)} \|^2 + \| \theta^{(n)} - \theta \|^2 \leq (2 \log \epsilon^{-2} + 1)|\epsilon^2 + \mathcal{R}(\theta^{(n)}, \epsilon)| + \sum_{j \geq J} \| \theta_j \|^2.
\]

(15.16)

Bound (15.16) is a pointwise estimate — valid for each coefficient vector \( \theta \). We now investigate its consequences for the worst case MSE of thresholding over Besov bodies \( \Theta = \Theta_{p,q}^\alpha(C) \). (Recall definition?) Given \( \alpha \), we set, as before,

\[
r = \frac{2\alpha}{2\alpha + 1}, \quad p(\alpha) = \frac{2}{2\alpha + 1} = 2(1 - r).
\]

Then, using the definition of the ideal risk seminorm, followed by Proposition 11.6, we have for any \( \theta \in \Theta_{p,q}^\alpha(C) \):

\[
\mathcal{R}(\theta^{(n)}, \epsilon) \leq |\theta|_{T_{p,r}} \epsilon^{2r} \leq \frac{2}{2-p} |\theta|_{w^{p(\alpha)}_{p,q}} \epsilon^{2r} \leq c_{\alpha p} C^{2(1-r)} \epsilon^{2r},
\]

(15.17)

where the final inequality uses the Besov space embedding result of Proposition 15.2: \( |\theta|_{w^{p(\alpha)}_{p,q}} \leq c_{\alpha p} \| \theta \|_{p,q} \leq c_{\alpha p} C \).

Tail bias. We digress briefly to show that the tail bias term in (15.16)

\[
\Delta_j^T(\Theta) := \sup \{ \| \theta \|^2 : \theta_{jk} = 0 \text{ if } j < J \}
\]

is negligible relative to the ideal risk bound just established. First, note the simple

\[\text{Lemma.} \quad \text{For } \theta \in \mathbb{R}^n, \quad \sup \{ \| \theta \|_2 : \| \theta \|_p \leq C \} = C n^{1/2 - 1/p} \epsilon.\]
Proof. This follows from a picture: when \( p < 2 \), the vectors having largest \( \ell_2 \) norm in an \( \ell_p \) ball are sparse, being signed permutations of the "spike" \( C(1,0,\ldots,0) \). When \( p \geq 2 \), the extremal vectors are dense, being sign flips of \( Cn^{-1/p}(1,\ldots,1) \).

Now combine across levels to obtain a tail bias bound.

15.4. Proposition. Let \( \alpha' = \alpha \) if \( p \geq 2 \) and \( \alpha = \alpha + 1/2 - 1/p \) if \( p < 2 \). Then for \( \Theta = \Theta_{p,q}^\alpha(C) \), the tail bias

\[
\Delta^2_j \leq c_{op}C^2(e^2)^{2\alpha'},
\]

(15.18)

\[
= o(e^2),
\]

(15.19)

(in the case \( p < 2 \), we assume also that \( \alpha \geq 1/p \) in (15.19).)

Proof. Since \( \Theta_{p,q}^\alpha(C) \subset \Theta_{p,\infty}^\alpha(C) \), we apply Lemma ?? along with the bounds \( \|\theta_j\|_p \leq C2^{-\alpha j} \) to conclude that

\[
\|\theta_j\|_2 \leq C2^{-\alpha j}.
\]

Since \( \Delta_j^2(\Theta) = \sum_{j \geq j} \|\theta_j\|_2^2 \), we obtain (15.18) by summing the geometric series and inserting \( 2^j = e^2 \). Bound (15.19) is trivial when \( p \geq 2 \), since \( 2\alpha > r \). When \( p < 2 \), the condition \( \alpha \geq 1/p \) implies \( 2\alpha = 2 \alpha' = 1 > r \). □

Remark. The condition \( \alpha \geq 1/p \) in the \( p < 2 \) case could be weakened to \( \alpha > 1/p - 1/2 \) by choosing to threshold, say \( (\log_2 e^2)^2 \) levels rather than \( \log_2 e^2 \). However, we retain the latter choice in order to stay closer to what one does with data in practice. The condition \( \alpha \geq 1/p \) implies, by embedding results in Appendix ??, that the functions \( f[\theta] \) are continuous, which seems a reasonable condition in order to speak sensibly of point evaluation in model (3.3).

Lower Bounds. We saw in the proof of Proposition 15.2 that \( \Theta_{p,q}^\alpha(C) \) contains \( \epsilon \) hypercubes of dimension

\[
N(\Theta, \epsilon) \geq c_0(C/\epsilon)^{p(\alpha)}.
\]

Hence the general hypercube lower bound (15.9) implies that

\[
R_N(\Theta, \epsilon) \geq c_1(C/\epsilon)^{p(\alpha)} \epsilon^2 = c_1C^{2(1-r)} \epsilon^{2r},
\]

(15.20)

Let us summarize the conclusions. Combining (15.16), (15.17), (15.19) and (15.20), we arrive at

15.5. Theorem. Assume that \( \alpha > (1/p - 1/2)_+ \), \( 0 < p,q \leq \infty \), \( 0 < C < \infty \). If \( p < 2 \), then assume also that \( \alpha \geq 1/p \). Let \( \tilde{\theta}^U \) denote soft thresholding at \( \epsilon \sqrt{2 \log n} \), defined at (15.15) Then for any Besov body \( \Theta = \Theta_{p,q}^\alpha(C) \) and all \( \epsilon < 1 \),

\[
\sup_{\Theta} r(\tilde{\theta}^U,\theta) \leq c_{op}(2 \log \epsilon^{-2})C^{2(1-r)} \epsilon^{2r}
\]

\[
\leq c_{op}(2 \log \epsilon^{-2})R_N(\Theta, \epsilon).
\]

218
A key aspect of this Theorem is that thresholding “learns” the rate of convergence appropriate to the parameter space $\Theta$. The definition (15.15) of $\hat{\theta}^\nu$ does not depend at all on the parameters of $\Theta^a_h(C)$, and yet, when restricted to such a set, the MSE attains the rate of convergence appropriate to that set, subject only to extra logarithmic terms.

(Comparison with linear....)
In previous chapters, we have taken the minimax principle literally, and have described minimax risks and estimators that correspond to a particular choice of parameter space $\Theta$ and loss function. While such problems have the virtue of being clearly defined, we have often seen that the resulting estimators are not available explicitly, and in any case, depend strongly on aspects of $\Theta$ that may not be known in practice.

In this chapter, we change the point of view to that of describing a single explicit estimator and investigating its properties across a variety of contexts. It will be seen that the understanding gained in previous chapters provides an important basis for the development in this chapter.

The setting is the projected white noise model (19.8) with $n = 2^J$ observations, expressed in the wavelet domain as

$$y_I = \theta_I + \varepsilon z_I, \quad I \in I^J,$$

with $\varepsilon$ known and $z_I \sim N(0, 1)$. We consider a soft threshold estimate, with threshold set at $\delta_n = \varepsilon \sqrt{2 \log n}$:

$$\hat{\delta}_{\delta_n, f}(y) = \begin{cases} \eta_S(y_I; \delta_n) & I \in I^J \\ 0 & \text{otherwise.} \end{cases}$$

This estimator will be seen to have several useful features:

First, it can be easily used in practice. If data is observed that can be reasonably approximated by the finite sample regression model

$$\hat{y}_i = f(i/n) + \sigma \tilde{z}_i, \quad i = 1, \ldots n,$$

then the estimator (16.2) can be applied to the discrete wavelet transform of $\hat{y}_i$. [Compare diagram (19.15) in the previous chapter.] Indeed, the algorithm runs in $O(n)$ time (if the wavelet filters have finite support) and is widely distributed in software. The results of the previous chapter show that properties of (16.2) proved in model (16.1) are at least asymptotically equally valid when the same estimator is applied to the wavelet transform of discrete data (16.3). Picture of NMR reconstruction.

Secondly, estimator (16.2) possesses a number of near optimality properties simultaneously over a large class of function spaces capturing varying degrees and homogeneity of smoothness.
(a) The function estimates \( f[\hat{\theta}] \) corresponding to (16.2) are in a strong sense “as smooth as” \( f \), so that one has, with high probability, a guarantee of not “discovering” non-existent features. (Theorem 16.4)

(b) The local nature of wavelet bases ensures that \( f[\hat{\theta}] \) has strong properties of spatial adaptation to local variations in the function \( f \). (See Chapter ???)

(c) For a large class of parameter spaces \( \Theta \) and global error measures \( \| \cdot \| \) derived from Besov and Triebel norms, the estimator (16.2) is simultaneously near minimax (Corollary ?? and Theorems ??).

(d) For local estimation of a function at a fixed point \( t_0 \), (16.2) again is simultaneously near minimax over a range of smoothness classes - indeed no better rate adaptivity property is possible for any estimator. (Theorem 16.9)

Finally, the proofs of the properties (a) and (c) exploit a useful connection with a deterministic problem of optimal recovery, and highlight the key role played by the concept of shrinkage in unconditional bases, of which wavelet bases are a prime example.

16.1. Unconditional Bases

An unconditional basis \( \{\psi_t\} \) for a Banach space \( B \) can be defined by two properties: (i) (Schauder basis) \( \forall v \in B, \exists \) unique sequence \( \{\theta_t\} \subset \mathbb{C} \) such that \( v = \sum_{t}^{\infty} \theta_t \psi_t \), and (ii) (Multipliers) \( \exists \) \( C \) s.t. \( \forall N \) and sequences \( \{m_t\} \subset \mathbb{C} \) with \( |m_t| \leq 1 \)

\[
\| \sum_{t}^{N} m_t \theta_t \psi_t \| \leq C \sum_{t}^{N} \| \theta_t \psi_t \|.
\]

Several equivalent forms and interpretations of the definition are given by Meyer (1990, I, Ch. VI). Here we note only that (16.4) says that shrinkage of coefficients can not grossly inflate the norm in unconditional bases. This suggests that traditional statistical shrinkage operations - usually introduced for smoothing or stabilization purposes - are best performed in unconditional bases.

A key consequence of the sequence norm characterisation results described in Section B.3.2 is that wavelets form unconditional bases for the Besov and Triebel scales of function spaces. Indeed, when viewed in terms of the sequence norms

\[
\|f\|_{B^{s}_{p,q}} \approx \sum_{j} 2^{j} \sum_{k} |\alpha_{jk}|^{q/p},
\]

the multiplier property is trivially satisfied, since \( \|f\| \) depends on \( \alpha_{jk} \) only through \( |\alpha_{jk}| \). Donoho (1993, 1996) has shown that unconditional bases are in a certain sense optimally suited for compression and statistical estimation.
16.1. Example. Suppose that the orthonormal wavelet $\psi$ is $C^R$ and has $D$ vanishing moments. Consider a scale of functional spaces

$$
C(R, D) = \{ B_{p,q}^\alpha [0,1], F_{p,q}^\alpha [0,1] : 1/p < \alpha < \min(R, D) \}.
$$

These spaces are (i) all embedded in $C[0,1]$ (since $\alpha > 1/p$), and (ii) the wavelet system $\{\psi_{jk}\}$ forms an unconditional basis for each of the spaces in the scale (since $\alpha < \min(R, D)$).

16.2. Example. Preservation of Smoothness. Suppose now that $\{\psi_I\}$ is an unconditional basis for a function space $\mathcal{F}$ with norm $\| \cdot \|_\mathcal{F}$. Data from model (12.1) can be used to construct an estimator of $f = \sum \theta_I \psi_I$ by setting

$$
\hat{f} = \sum \hat{\theta}_I \psi_I.
$$

The uniform shrinkage property combined with the multiplier property (16.4) implies that whatever be the noise $u$,

$$
\| \hat{f} \|_\mathcal{F} \leq C \| f \|_\mathcal{F}.
$$

This means that one can assert that $\hat{f}$ is as smooth as $f$. In particular, if $f$ is identically 0, then so is $\hat{f}$! Furthermore, for a $C^R$ wavelet $\psi$ with $D$ vanishing moments, this property holds simultaneously for all spaces $\mathcal{F}$ in the scale $C(R, D)$ of (16.5).

Figure of soft thresholding on the 4 functions.

16.2. Optimal Recovery Model - Finite Data

For much the same reasons as in Section 19.2, we will also need to consider a projected data version of the optimal recovery model in which

$$
x_I = \theta_I + \delta u_I \quad I \in \mathcal{I}_n, \quad |\mathcal{I}_n| = n.
$$

Again, one still attempts to recover the entire object $\theta$, and the corresponding minimax recovery error is

$$
E(\Theta, \delta; n) = \inf_{\hat{\theta}(x^{(n)})} \sup_{\theta} e(\hat{\theta}(x^{(n)}), \theta).
$$

Projection onto the $n-$data model is defined by

$$
(P_\delta \theta)_I = \begin{cases} \theta_I & I \in \mathcal{I}_n \\ 0 & \text{otherwise} \end{cases}
$$

Even when the noise level $\delta = 0$, there is still an error of recovery due to the attempt to infer the full vector $\theta$ from only $n$ components. Hence we make the

Definition. The tail $n-$width of $\Theta$ in norm $\| \cdot \|$ is

$$
\Delta(n, \Theta, \| \cdot \|) = \sup_{\theta \in \Theta} \{ \| \theta \| : P_\delta \theta = 0 \} = E(\Theta, 0; n).
$$

It is then straightforward to establish the following finite data analog of Theorem 12.2.

222
16.3. PROPOSITION. Suppose that $\Theta$ is solid, orthosymmetric and convex, and that the error norm $\| \cdot \|$ is solid and orthosymmetric. Then
\[
\max \left\{ \frac{\Omega(\delta)}{2}, \Delta(n) \right\} \leq E(\Theta, \delta; n) \leq 2\Omega(\delta) + \Delta(n).
\]
In addition, soft thresholding $\hat{\theta}_n$ is near minimax simultaneously for all such parameter spaces and error norms.

16.3. Statistical Model

We return to the projected Gaussian white noise model (16.1). The connection with the optimal recovery model is made by considering the event
\[
A_n = \{ \sup_{I \in \mathcal{I}} |z_I| \leq \sqrt{2 \log n} \},
\]
which because of the properties$^1$ of maxima of i.i.d. Gaussians has probability approaching one:
\[
P(A_n) = \pi_n \geq 1 - \frac{1}{\sqrt{n \log n}} > 1 \quad \text{as } n \to \infty.
\]
The key idea is to apply results from the optimal recovery model with deterministic noise level $\delta_n = \epsilon_n \sqrt{2 \log n}$ on the set $A_n$. Thus, in the statistical model we consider the soft thresholding estimator (16.2) at level $\sqrt{2 \log n/n}$.

16.3.1. Preservation of Smoothness. As a first illustration, consider the smoothness preservation property of Example 16.2. On the event $A_n$, the uniform shrinkage property Lemma 12.1 implies that $\hat{\theta}_n \in \Theta$ whenever $\theta \in \Theta$. Hence, for function spaces in the scale $C(R, D)$, we have on $A_n$ that
\[
\| \hat{f}_n \|_{\mathcal{F}} \leq C(\mathcal{F}) \| f \|_{\mathcal{F}}.
\]

16.4. THEOREM. For each function space $\mathcal{F} \in C(R, D)$ there exists a constant $C(\mathcal{F})$ such that
\[
P\{\| \hat{f}_n \|_{\mathcal{F}} \leq C(\mathcal{F}) \| f \|_{\mathcal{F}} \forall \mathcal{F} \in \mathcal{C}\} \geq \pi_n \to 1.
\]

Thus, one can assert that with high probability, the estimator $\hat{f}_n$ is as smooth as the “truth” $f$ simultaneously over many smoothness classes. In particular, if $f \equiv 0$, then $\hat{f}_n \equiv 0$ with probability at least $\pi_n$ so that one can assert that $\hat{f}_n$ does not find “spurious structure”.

16.3.2. Global Estimation Bounds. In a similar manner, we can immediately convert the upper-bound part of Proposition 16.3 to a statement in the projected Gaussian model with $\delta_n = \epsilon_n \sqrt{2 \log n}$: for the soft threshold estimator $\hat{\theta}_n$, we have for all solid, orthosymmetric (and convex) $\Theta$ that
\[
\sup_{\Theta} P\{\| \hat{\theta}_n - \theta \| \leq 2\Omega(\delta) + \Delta(n)\} \geq \pi_n \to 1.
\]
Thus the statistical model is not harder than the optimal recovery model, up to factors involving $\sqrt{\log n}$. We may say, using the language of Stone
(1980) (CHECK), that \(2\Omega(\delta) + \Delta(n)\) is an achievable rate of convergence for all qualifying \((\Theta, \| \cdot \|)\).

Now specialize to the case of parameter space \(\Theta\) and error norm \(\| \cdot \|\) taken from the Besov scale.

We first summarize the results of calculation of the Besov modulus and bounds for the tail bias, the details being deferred to the next section.

16.5. THEOREM. Let \(\Theta = \Theta_{p,q}^r(C)\) and \(\| \cdot \| = \| \cdot \|_{b_{p',q'}}\). Assume that
\[
\alpha = \alpha - \alpha' - (1/p - 1/p') > 0.
\]
(a) Then
\[
\Omega(\delta) \geq C^{1-r} \delta^r \quad \text{as } \delta \to 0.
\]
where
\[
r = \begin{cases} 
\frac{2(\alpha - \alpha')}{(\alpha + 1/2)p} & (\alpha + 1/2)p > (\alpha' + 1/2)p' \quad \text{"regular"} \\
\frac{2(\alpha - \alpha')}{(\alpha + 1/2)p - (\alpha' + 1/2)p'} & (\alpha + 1/2)p < (\alpha' + 1/2)p' \quad \text{"logarithmic"}
\end{cases}
\]
(b) If, in addition, \(\alpha > 1/p\), then the tail bias
\[
\Delta(n) \leq c_2 C n^{-\delta} = o(\Omega(n^{-1/2} \sqrt{\log n})).
\]
with \(c_2 = (1 - 2^{-\alpha/d})^{-1/d} \).

Remark. In the “critical case” \((\alpha + 1/2)p = (\alpha' + 1/2)p'\), behavior is more complicated and discussed in Donoho et al. (1997).

Make the normalization \(\epsilon = n^{-1/2}\). Using the bounds derived for the Besov modulus in Theorem 16.5 and for the tail bias in Proposition ?? we obtain

16.6. THEOREM. Let \(\Theta = \Theta_{p,q}^r(C)\) and \(\| \cdot \| = \| \cdot \|_{b_{p',q'}}\). Assume that
\[
\alpha = \alpha - \alpha' - (1/p - 1/p') > 0 \quad \text{and that } \alpha > 1/p.
\]
Then
\[
\sup_{\theta \in \Theta(n)} P\{|\hat{\theta}_{\delta_n} - \theta| \leq c\Omega(n^{-1/2} \sqrt{\log n})\} \geq \pi_n \to 1.
\]
There exists a constant \(c = c(p)\) such that
\[
\inf_{\theta} \sup_{\hat{\theta}} P\{|\hat{\theta} - \theta| \geq c\Omega(n^{-1/2})\} \to 1.
\]
In the logarithmic case, the lower bound can be strengthened to \(\Omega(n^{-1/2} \sqrt{\log n})\).

Thus, soft thresholding at \(\delta_n = \epsilon_n \sqrt{\log n}\) is simultaneously nearly minimax (up to a logarithmic term) over all parameter spaces and loss functions in the (seven parameter) scale \(C(R, D)\), and indeed attains the optimal rate of convergence in the logarithmic case.

To appreciate the significance of adaptive estimation results such as this, note that an estimator that is exactly optimal for one pair \((\Theta, \| \cdot \|)\) may well have very poor properties for other pairs; one need only imagine taking a linear estimator (e.g. from Pinsker’s theorem) that would be optimal for an ellipsoid \(\Theta_{2,2}\) and using it on another space \(\Theta_{p,q}\) with \(p < 2\) in which
linear estimators are known (e.g. Chapter ???) to have suboptimal rates of convergence.

16.4. Besov Modulus and Tail Bias

[REVISE:] In this section we evaluate the asymptotic order of the modulus of continuity $\Omega(\delta)$ when both parameter space $\Theta_{p,q}^r$ and error measure $\| \cdot \|_{b_{p',q'}}$ are taken from the Besov scale. The approach is to reduce the optimisation defining the modulus to a hardest resolution level $j$, where one is effectively dealing with scaled versions $\ell_p$ norms in both the error measure and in the reduced parameter space. It is therefore useful to begin with the evaluation of the modulus over such $\ell_p$ balls and norms.

In this exposition, we consider only quadratic error norms with $p' = 2$, while keeping $p$ unrestricted. This makes some of the formulas easier to read, while retaining the important features of the general $p'$ case, which is treated (along with the Triebel and “critical” cases) by the same methods in Donoho et al. (1995, 1997).

The approximate calculation of the modulus $\Omega(\delta)$ for $\Theta$ and $\| \cdot \|$ taken from the Besov scale can be achieved by a reduction to the $\ell_p$ modulus just discussed. This leads to the [END-REVISE]

PROOF. First define the Besov shells

$$\Theta^{(j)} = \{ \theta \in \Theta : \theta_I = 0, I \notin I_j \}.$$ 

If $\theta^{(j)}$ is derived from $\theta$ by setting to zero all components $\theta_I$ with $I \notin I_j$, then

$$\| \theta^{(j)} \|_{b_{p,q}} = 2^{aj} \| \theta_j \|_p.$$ 

This shows that $\Theta^{(j)}$ is isomorphic to a scaled $\ell_p$-ball:

$$\Theta^{(j)} \cong \Theta_{2^j,p}(C2^{-aj}).$$

The modulus of continuity, when restricted to the $j$th shell, reduces in turn to a scaled form of the $\ell_p$-modulus:

$$\Omega_j(\delta) := \Omega(\delta; \Theta^{(j)}, \| \cdot \|) = 2^{aj} j W_{2^j,p}(\delta, C2^{-aj}).$$

It is easy to verify that nothing essential (at the level of rates of convergence) is lost by considering the shell moduli: In the convex case,

$$\| \Omega_j(\delta) \|_{\ell_p} \leq \Omega(\delta) \leq \| \Omega_j(\delta) \|_{\ell_p}.$$ 

Initially, we give the rest of the argument only in the case $p < 2$. The approach is simply to use Table 1 to compute the $\ell_p$-modulus appearing in (16.11) as a function of $j$, determining which case in the table is appropriate for each value of $j$. This leads to the following summary table:
\[
\begin{array}{c|c|c|c}
 & p < p' = 2 & \Omega_j^0 & n_0 \\
\hline
j < j_0 = (\alpha + 1/2)^{-1} \log_2 C/\delta & \delta^2 2^{(1+2\alpha)j} & 2^{\alpha} \delta & 2^j \\
\hline
j_0 < j < j_1 & C^n \delta^{3-p-2j(\alpha-2\delta')} & 2^{\alpha} \delta & 2^{2p(j_1-j)} \\
\hline
j > j_1 = a^{-1} \log_2 C/\delta & C^{2-2\delta}(a-\delta') & C^{2-2\delta}(a-\delta') & 1 \\
\end{array}
\]

Thus, for \( j < j_0 \), the function \( j \rightarrow \Omega_j(\delta) \) is geometrically increasing, and for \( j > j_1 \) it is geometrically decreasing. In the non-critical case, \( \Omega_j \) is not constant on the interval \((j_0, j_1)\) and has geometric growth or decay: the location of the overall maximum \( j_0 \) is at \( j_0 \) if \( pa > 2\delta' \) (the “regular case”), and at \( j_1 \) if \( pa < 2\delta' \) (the “logarithmic case”). At these critical points \( j_0, j_1 \), one has \( \Omega_j = C^{2(1-r)}\delta^{2r} \), with

\[
r_0 = 2(\alpha - \alpha')/(2\alpha + 1), \quad r_1 = 1 - \alpha'/\alpha.
\]

In the regular case, the least favorable configuration is dense, having \( 2^j \) spikes at level \( j_0 \), while in the logarithmic case, the favorable configuration is sparse, with a single spike at the (higher) level \( j_1 \).

In particular, in the non-critical case, we have geometric decay of \( j \rightarrow \Omega_j \) away from the maximum \( j_0 \) (\( = j_0 \) or \( j_1 \)): there exists \( \eta > 0 \) such that

\[
\Omega_j(\delta) \leq \delta^\eta C^{1-r} 2^{-n j - j',-1}.
\]

The evaluation (16.7) follows from this and (16.12). \( \square \)

**Evaluation of Besov tail widths**

**Proof.** In keeping with the rest of this chapter, we consider only quadratic type error measures, namely \( p' = 2 \). First, a remark on the form taken by the modulus \( W_{n,p}(\delta, C) \) for \( \ell_p \) balls when the constraint \( \delta \) is absent, or equivalently, when \( \delta = \infty \). Inspection of Table 1 shows that

\[
W_{n,p}(\infty, C) = n^{1/2-1/p} + C.
\]

Now, by analogy with the Besov modulus argument, consider the tail width restricted to the \( j \)th Besov shell.

\[
\Delta_j = \sup \{ \| \theta_j \|_{l_2^{p',\alpha}} : \| \theta_j \|_{l_2^{p',\alpha}} \leq C \}.
\]

The full tail width is related to these shell widths by

\[
\Delta_{j+1} \leq \Delta(2^j, \Theta) \leq \| (\Delta_j)_{j>0} \|_{l_2^{p',\alpha}}.
\]

Now using (16.10) and (16.14),

\[
\Delta_j = 2^{j \alpha} \sup \{ \| \theta_j \|_2 : \| \theta_j \|_p \leq C^{2^{-\alpha}j} \}
\]

\[
= 2^{j \alpha} W_{2, p}(\infty, C^{2^{-\alpha}j})
\]

\[
= 2^{j \alpha} 2^{j(1/2-1/p)+C^{2^{-\alpha}j}} = C^{2^{-\alpha}j}.
\]

Consequently,

\[
\Delta_j(2^j, \Theta) \leq C^{\alpha} 2^{-j \alpha} \sum_{j=1}^{\infty} 2^{-j \alpha j}.
\]

226
which yields (16.8) on setting \( n = 2^j \).

Note also that so long as \( \tilde{\alpha} > 0 \) and \( \alpha > 1/p \), it can be verified (proof?) that the tail bias is asymptotically negligible relative to the Besov modulus \( \Omega(n^{-1/2}) \) corresponding to the noise level \( \epsilon_n = n^{-1/2} \).

Include graph of \( j \to \Omega_j^2 \).

### 16.5. Lower Bounds in the Statistical model

As with the evaluation of the Besov modulus, our approach to statistical lower bounds will be to use the hardest Besov subshell, and there to use a lower bound on estimation in \( \ell_p \) balls. The lower bounds will be developed using observations from a model

\[
v_i = \xi_i + \varepsilon_i, \quad i = 1, \ldots, m
\]

with \( z_i \) i.i.d. \( N(0, 1) \) and \( \xi_i \) sampled at random from least favorable configurations found in the optimal recovery analysis. We use a sample size index \( m \) (which will later be \( 2^j \) for an appropriate level \( j \)) to distinguish from the total effective sample size \( n = \epsilon^{-1/2} \). In the non-critical case, there are two situations: in which the least favorable configurations are respectively dense and sparse.

**Dense Case.** Return to the \( \ell_p \) modulus \( W_{m,p}^2(\delta, C) \) considered in Section 12.2.1 and let \( (n_0, \delta_0) \) be the number and size of the spikes corresponding to the least favorable configuration for \( (\delta, C) \).

#### 16.7. Proposition. In the setting of model (16.15), let \( \pi_0 = \Phi(-1)/2 \approx .08 \)

\[
\inf_{\xi} \sup_{\Theta_{m,p}(C)} P\{ \| \hat{\xi} - \xi \|^2 \geq (\pi_0/2)W^2(\delta, C) \} \geq 1 - (n_0 + 1)^2 e^{-2n_0\delta^2}.
\]

Note that this result (proof to be added) applies uniformly over \( (m, \delta, C) \): we will use it with \( m = 2^j \), and \( \delta = \epsilon_n = n^{-1/2} \). To apply this result in the Besov estimation context, we recall two consequences of the definition ?? of the Besov shells \( \Theta^{(j)} \). Suppose that \( \hat{\theta} \) is an arbitrary estimator. Then

\[
\| \hat{\theta}^{(j)} - \theta^{(j)} \|_{p, m} = 2^{d/j} \| \hat{\theta}_j - \theta_j \|_p \quad \text{and}
\]

\[
\theta^{(j)} \in \Theta \iff \| \theta_j \| \leq C2^{-aj}.
\]

As a result,

\[
\sup_{\theta \in \Theta} P\{ \| \hat{\theta} - \theta \| > c_1 \} \geq \sup_{\theta_j \in \Theta_{m, p}(C2^{-aj})} P\{ \| \hat{\theta}_j - \theta_j \|_p \geq c_1 \}
\]

\[
\geq 1 - (n_0 + 1)^2 e^{-2n_0\delta^2}
\]

so long as \( c_1 = (\pi_0/2)2^{d/j}W_{2^j, p}(n^{-1/2}, C2^{-aj}) = (\pi_0/2)\Omega_j(n^{-1/2}) \). In the regular case, \( n_0 = 2^{3j} = (C/\delta)^{2/\delta^{2+1}} = (C\sqrt{n})^{2/\delta^{2+1}} \to \infty \) and so the probability above approaches 1 and we obtain the lower bound (16.9).
Sparse Case. Again suppose the model (16.15), but now suppose that we are in a sparse “needle in a haystack” situation: all the $\xi_j$ are zero, with the exception of at most one, of size $\delta_0$, whose location among the $m$ observations is unknown. Let $\Theta_{m,0}(\delta_0,1)$ denote the collection of all such sequences. The following lemma says that such solitary spikes are asymptotically undetectable if they are of smaller order than the maximum of $m$ observations on pure noise.

16.8. Proposition. With $\eta \in (0,2)$, let $\delta_0 \leq \sqrt{(2-\eta)\log m} \cdot \epsilon$ for all $m$. Then

$$\inf_{\hat{\xi}} \sup_{\Theta_{m,0}(\delta_0,1)} P\{ \| \hat{\xi} - \xi \|^2 \geq (\delta_0/3)^2 \} \to 1.$$ 

Again, we apply this result (proof to be added) to estimation over Besov bodies in the dyadic sequence model. At level $j$, the sample size $m$ in (16.15) corresponds to $2^j$ and indistinguishable spikes have maximum size up to $\epsilon \sqrt{(2-\eta)\log 2^j}$. This size increases with level $j$ - to maximize it, choose the largest $j$ such that the spike still belongs to $\Theta(j)$; in other words, choose the largest $j_1$ such that

$$\epsilon \sqrt{(2-\eta)\log 2^j} \leq C2^{-a j_1}.$$ 

This condition implies that $2^j \eta \sqrt{j_1} \leq C/\epsilon$ or equivalently that $j_1 \geq a^{-1} \log C/\epsilon$, and hence that $2^{-j_1} \simeq (\epsilon/C)\sqrt{\log C/\epsilon}^{1/a}$. For an arbitrary estimator $\hat{\theta}$, Proposition 16.8 implies that with probability increasing to 1,

$$\| \hat{\theta} - \theta \|_\nu \geq c_2 2^j \cdot C2^{-j_1} \simeq \epsilon C\sqrt{\log C/\epsilon}^{1/a} \simeq \Omega(\sqrt{\log n/n}),$$

after substituting $\epsilon = n^{-1/2}$. This yields the logarithmic case of Theorem 77.

16.6. Estimation at a point.

In this section, we change point of view and consider the estimation of the value $f(t_0)$ of a function at a point $t_0 \in (0,1)$ on the basis of observations from the projected Gaussian white noise model (16.1) with $n = 2^j$ observations. We again consider the wavelet threshold estimator with threshold $\delta_n = \epsilon_n \sqrt{2 \log n}$:

$$\hat{f}_n(t_0) = \sum_{j \in I_{\delta_n}} \hat{\theta}_{\delta_n,t} \psi_j(t_0).$$

(16.16)

Suppose that $f$ satisfies a Hölder smoothness condition $f \in \Lambda^\alpha(C)[0,1]$ for $\alpha > 0$. If $\alpha$ is not an integer,

$$\Lambda^\alpha(C) = \{ f : |f^{(m)}(s) - f^{(m)}(t)| \leq C|s-t|^{\delta} \}$$

with $m = [\alpha]$ and $\delta = \alpha - m$. (If $\alpha$ is an integer, we may even extend to the Zygmund definition (Meyer (1990)?)). connect to other defn.
16.9. **Theorem.** Suppose that the wavelet $\psi$ has compact support and $\min(R, D) > 1$. For each space $\Lambda^\alpha(C)$ with $0 < \alpha < \min(R, D)$, and $r = 2\alpha/2\alpha + 1$

\begin{equation}
(16.17) \sup_{f \in \Lambda^\alpha(C)} E[\hat{f}_n(t_0) - f(t_0)]^2 \leq c_{\psi, \alpha} C^{2(1-r)}(\frac{\log n}{n})^r (1 + o(1)).
\end{equation}

Remarks. 1. If we knew $\alpha$ and $C$, then we could construct a linear minimax estimator $\hat{f}_n = \sum \alpha y_j$ where the $(\alpha_j)$ are the solution of a quadratic programming problem depending on $C, \alpha, n$ (Ibragimov & Khas'minskii (1982), Donoho & Liu (1991), Donoho (1994)). This estimator has worst case risk over $\Lambda^\alpha(C)$ asymptotic to $c_\alpha C^{2(1-r)} n^{-r}$. However, as in section 12.3.1, if the Hölder class is incorrectly specified, then this linear estimator will have a suboptimal rate of convergence over the true Hölder class.

In contrast, the wavelet threshold estimator (16.16) does not depend on the parameters $(C, \alpha)$, and yet achieves nearly the optimal rate of convergence — up to a factor $\log^2 n$ — over all the Hölder classes.

Lepskii (1991) and Brown & Low (1992) have shown that this rate penalty $\log^2 n$ is in fact optimal: even if the correct Hölder class is one of two, specified by pairs $(\alpha_0, C_0)$ and $(\alpha_1, C_1$ with $\alpha_0 < \alpha_1$, then

\begin{equation}
\inf_{\hat{f}_n} \max_{L \in 0, 1} (C, \alpha, n) \sup_{L \in 0} E[\hat{f}_n(t_0) - f(t_0)]^2 \geq C_2 \log^{\alpha_0} n.
\end{equation}

2. It would certainly be possible to give a version of Theorem 16.9 using optimal recovery ideas, but we have chosen here to give a mean squared error version that uses the risk inequality (11.8) for soft thresholding.

Proof. We begin with some preliminary remarks. First, membership of $f$ in $\Lambda^\alpha(C)$ entails (Lemma B.7) that the corresponding wavelet coefficients $\theta = \theta[f]$ satisfy

\begin{equation}
|\theta| \leq c C^{-\alpha/2} \varepsilon.
\end{equation}

Next, we recall from (11.8) the mean squared error bound for a soft threshold estimator with threshold $\lambda$, here given for noise level $\varepsilon$:

\begin{equation}
(16.19) \quad r_s(\lambda, \theta; \varepsilon) \leq \varepsilon^2 r(\lambda, 0) + \varepsilon^2 + \lambda^2 \varepsilon^2 \quad \lambda^2 = 1 + \varepsilon^2.
\end{equation}

Since $\lambda = \sqrt{2\log n}$, we have from (11.7) that $r(\lambda, 0) \leq c n^{-1}$. The following identity says that the root mean squared (RMS) error of a sum is bounded by the sum of the RMS errors: from the Cauchy-Schwarz inequality,

\begin{equation}
(16.20) \quad E\left(\sum_{i} a_i^2\right)^2 \leq \sum_{i,j} \sqrt{Ea_i^2 E\bar{a}_j^2} = \left(\sum_{i} \sqrt{Ea_i^2}\right)^2.
\end{equation}

Now, a useful bound for sums with geometric decay from a single dominant term:

\begin{equation}
(16.21) \quad \sum_{j} 2^j (C^{2-\omega} \wedge \delta) \leq c(\gamma, \omega) C^{\gamma/\omega} \delta^{1-\gamma/\omega} \quad \gamma < \omega.
\end{equation}
[First, verify the case $C = 1$, and then apply that result with $\delta' = \delta/C$ to recover (16.21).] Finally, constants $c = c(\psi, \alpha)$ may change from line to line, but depend only on $\psi, \alpha$), and could be made more explicit.

Now write the estimation error as

$$
\hat{f}_n(0) - f(0) = \sum_{I \in \mathcal{I}_{(n)}} (\hat{\theta}_I - \theta_I)\psi_I(0) + \sum_{I \notin \mathcal{I}_{(n)}} \theta_I\psi_I(0).
$$

The second term represents “tail bias” and is estimated in a straightforward way from (16.18) and the compact support property of $\psi$:

$$
|\text{tail}| \leq c \sum_{j \geq J} C2^{-(\alpha+1/2)j} \cdot 2^{j/2} \leq cC2^{-\alpha J} = O(n^{-\alpha}) = o(n^{-r/2}).
$$

Set now $a_I = (\hat{\theta}_I - \theta_I)\psi_I(0)$ and use (16.19) in conjunction with $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$:

$$
\sqrt{Ea_I^2} \leq |\psi_I(0)| [c\sqrt{r(\lambda, 0)} + |\theta_I| \wedge \lambda]\]
\leq c2^{j/2} [c/n + C2^{-\omega j} \wedge \delta]
$$

where $\delta = \sqrt{2 \log n / n}$. Using now the compact support of the wavelets $\psi_I$, and then (16.21)

$$
\sum_{I \in \mathcal{I}_{(n)}} \sqrt{Ea_I^2} \leq c2^{j/2}/n + c \sum_{j < J} 2^{j/2}(C2^{-\omega j} \wedge \delta)
$$

$$
\leq c/\sqrt{n} + c_{\alpha, \psi} C^{1-r} \delta^r
$$

$$
r = 1 - 1/2\omega = 2\alpha/2\alpha + 1
$$

$$
\sim c_{\alpha, \psi} C^{1-r} (\log n)^{r/2}.
$$

Combining (16.22) and (16.20) yields the result (16.17). □

**Remark.** It is evident both intuitively from (16.18) that the full global constraint of Hölder regularity on $[0, 1]$ is not needed: a notion of local Hölder smoothness near $t_0$ is all that is used [Ref to Jaffard-Meyer or Jaffard for more details?] Indeed (16.18) is only needed for indices $I$ with $\psi_I(0) \neq 0$.

**Further Remarks:** $\epsilon$ unknown

ALSO: more on discretization,

remarks on estimation of derivatives

remains on transition to function spaces

historical note on Nemirovskii(1985)
CHAPTER 17

*Adaptive indirect estimation on Besov spaces*

The claim made for wavelet shrinkage has been that it takes advantage of the local nature of wavelet basis functions to achieve spatial adaptation to inhomogeneous smoothness.

We have modelled inhomogeneous smoothness theoretically using Besov spaces $B^s_{p,q}$ for $p < 2$, and their sequence space norm balls $\Theta^s_{p,q}(C)$.

In studying $\sqrt{2\log n}$ thresholding, we showed that it was adaptively optimal up to a logarithmic factor of order $2\log \epsilon^{-1}$. That is, we showed that $\hat{\theta}^U$ satisfies (for $\alpha > (1/p - 1/2)_+, \alpha \geq 1/p, 0 < p, q \leq \infty, C > 0$)

$$
\sup_{\Theta^s_{p,q}(C)} E[\|\hat{\theta}^U - \theta\|^2] \leq c(\log \epsilon^{-1})R_N(\Theta^s_{p,q}(C), \epsilon)(1 + o(1))
$$
as $\epsilon \to 0$.

While this already a quite strong adaptivity statement, the extra $\log \epsilon^{-1}$ is undesirable, and indeed reflects a practically important phenomenon: $\sqrt{2\log n}$ thresholds can be too high in some settings (e.g. ion channel example), and lower choices of threshold can yield much improved reconstructions and MSE performance.

In this chapter, we apply the $2k\log n/k$ oracle inequality of Chapter 13 and its $\ell_p$ ball consequences (Chapter 14 to show that appropriate penalized least squares estimates (i.e. data dependent thresholding) adapt exactly to the correct rates of convergence over essentially all reasonable Besov bodies. Thus, we show that for an explicit $\hat{\theta}^P$,

$$
\sup_{\Theta^s_{p,q}(C)} E[\|\hat{\theta}^P - \theta\|^2] \leq c(\alpha, p, q)R_N(\Theta^s_{p,q}(C), \epsilon)(1 + o(1)).
$$

We work with the projected sequence model

$$
y_{jk} = \theta_{jk} + \epsilon z_{jk} \quad k = 1, \ldots, 2^j; \quad 0 \leq j < J = \log \epsilon^{-2},
$$

and $z_{jk} \sim N(0, 1)$ independently.

**Estimator:** Use a penalized least squares estimator at each level $j$:

$$
\hat{\theta}_P(y_j) = \arg\min_{\theta_j} \|y_j - \theta_j\|^2 + \epsilon^2 \text{pen}_j(N(\theta_j))
$$

where the $j^{th}$ penalty function is given by

$$
\text{pen}_j(k) = k\lambda^2_{j,k},
$$

$$
\lambda_{j,k} = \zeta\left(1 + \sqrt{2\log(2^j/\beta/k)}\right),
$$

231
Finally, put these levelwise estimates together to get $\hat{\theta}_j^P$:

$$
\hat{\theta}_j^P(y) = \begin{cases} 
\hat{\theta}_P(y_j) & j < J = \log_2 e^{-2} \\
0 & j \geq J.
\end{cases}
$$

**MSE decomposition.**

(17.1) 
$$E\|\hat{\theta}_P - \theta\|^2 = \sum_{j < J} E\|\hat{\theta}_P(y_j) - \theta_j\|^2 + \Delta_j(\theta),$$

where $\Delta_j(\theta) = \sum_{j \geq J} \|\theta_j\|^2$ is the “tail bias” due to not estimating beyond level $j$.

Our earlier oracle inequalities [REFS!] lead to explicit bounds for MSE at each level $j$:

(17.2) 
$$E\|\hat{\theta}_P(y_j) - \theta_j\|^2 \leq c_1 j e^2 + c_2 K_{e,j}(\theta_j)$$

$$\leq c_1 j e^2 + c_3 \sum_{j=1}^{J} \theta_{j(k)}^2 \wedge \lambda_{jk}^2.$$

Here $c_1$ depends on both $\zeta$ and $\beta$, while $c_2$ and $c_3$ depend on $\zeta$ alone. The $j$th theoretical complexity

$$K_{e,j}(\theta_j) = \inf_{\theta} \|\theta_j - \hat{\theta}_j\|^2 + c^2 \text{pen}_j(N(\hat{\theta})).$$

17.1. **Theorem.** Let $\hat{\theta}_P$ be the wavelet penalized least squares estimate described above. For $\alpha > (1/p - 1/2)_+$ along with $0 < p, q \leq \infty$ and $C > 0$,

$$\sup_{\theta \in \mathcal{B}} E\|\hat{\theta}_P - \theta\|^2 \leq c_3 C^{2(1-r)} e^{2r} + c_2 C^2 (e^2)^{2\alpha'} + c_1 e^2 \log^2 e^{-2},$$

where $r = 2\alpha/(2\alpha + 1)$ while $\alpha' = \alpha$ if $p \geq 2$ and $a = \alpha + 1/2 - 1/p$ if $p < 2$.

**Remarks.** 1. The dependence of the constants on the parameters defining the estimator and Besov space is given by $c_1 = c_1(\zeta, \beta, c_2 = c_2(\alpha, p)$ and $c_3 = c_3(\zeta, \beta, \alpha, p)$.

2. Let us examine when the $C^{2(1-r)} e^{2r}$ term dominates as $\epsilon \to 0$. Since $r < 1$, the $e^2 \log^2 e^{-2}$ term is always negligible. If $p \geq 2$, then $2\alpha' = 2\alpha > r$ and so the tail bias term is also of smaller order. If $p < 2$, a convenient extra assumption is that $\alpha \geq 1/p$, for then $\alpha' = a \geq 1/2 > r/2$. Note that the condition $\alpha \geq 1/p$ is necessary for the Besov space $B^\alpha_{p,q}$ to embed in spaces of continuous functions.

3. One may ask more explicitly for what values of $\epsilon$ the tail bias $C^{2(\epsilon^2)^{2\alpha'}} < C^{2(1-r)} e^{2r}$. Simple algebra shows that this occurs when

$$\epsilon < C^{-r/(2\alpha'-r)},$$

showing the key role of the radius $C$. 

232
PROOF. We may suppose that the parameter space $\Theta_{p,q}$ is fixed, and we abbreviate it by $\Theta$. Combining (17.1) and (17.2), we have

$$\sup_{\theta \in \Theta} E \| \hat{\theta}^p - \theta \|^2 \leq \sum_{j < J} [c_1 \epsilon^2 + c_2 S_j(\epsilon)] + \sup_{\theta \in \Theta} \Delta J(\theta),$$

where

$$S_j(\epsilon) = \sup_{\theta \in \Theta} K_{\epsilon,j}(\theta_j).$$

Note that $\sum_{j < J} j \leq \frac{1}{2} J^2 = \frac{1}{2} \log^2 \epsilon^{-2}$, and that we have earlier [REF1] obtained a bound for the maximal tail bias:

$$\sup_{\theta \in \Theta} \Delta J(\theta) \leq c C^2 \epsilon^{-2 \log \epsilon \alpha' \Delta \epsilon^{-1}}.$$

The main remaining step is to show that $S_j(\epsilon)$ decays geometrically as $j$ departs from a critical level defined by

$$j_* = (\alpha + 1/2)^{-1} \log_2 (C/\epsilon).$$

Indeed, we shall show the existence of $c = c(\alpha, p, \epsilon, \beta)$ and $\delta = \delta(\alpha, p)$ so that

(17.3) $$S_j(\epsilon) \leq c C^2(1-\epsilon) \rho(j - j_*) 2^{-\delta j - j_*},$$

where

$$\rho(\ell) = \begin{cases} 1 & p \geq 2 \\ 1 + \ell & p < 2 \end{cases}$$

If this is taken for granted, the theorem follows immediately because of convergence of the sum

$$\sum_j \rho(j - j_*) 2^{-\delta j - j_*}.$$

We turn to the proof of (17.3). This depends on the bounds developed earlier [REF1] for empirical complexity over $\ell_p-$balls, which recall here for convenience in the case of noise level $\epsilon = 1$. In terms of the parameter $\eta^p = n^{-1} C^p$, it was shown that

$$S_{n,p}(C) = \sup \{ K_0(\mu) : \mu \in \Theta_{n,p}(C) \} \leq c_3 \bar{\tau}_{n,p}(\eta).$$

When $p \geq 2$,

(17.4) $$\bar{\tau}_{n,p}(\eta) \leq n (\eta^2 \land 1),$$

while for $p < 2$,

(17.5) $$\bar{\tau}_{n,p}(\eta) \leq \begin{cases} n \eta^p (1 + \log \eta^{-p})^{1-p} /2 & \eta \leq 1 \\ n & \eta \geq 1 \end{cases}$$

[Note that the first bound in (14.3), valid for $C \leq \sqrt{1 + \log n}$, is, up to a constant multiple of 2, less than the second bound over this range, and that is all that will be needed below in the applications of (17.5).]
Recall that the Besov shell $\Theta^{(j)} = \{ \theta \in \Theta : \theta_{j+k} = 0 \} \text{ unless } j' = j \}$ and that it is equivalent to an $\ell_p$–ball

$$\Theta^{(j)} \equiv \{ \theta_j : \|\theta_j\|_p \leq C2^{-a j} \}.$$

Bounds (17.4) and (17.5) are applied to $\mu = (\mu_k)$, with $\mu_k = \theta_{jk}/\epsilon$, which belongs to $\ell_p$ with $\|\mu\|_p = \|\theta_j\|_p = \|\theta_j\|_2$ with

$$\eta_j = 2^{-j/p/(\alpha + 1/2)} = (C/\epsilon)2^{-((\alpha+1)/2)j} \quad = 2^{-((\alpha+1)/2)(j - j_*)}$$

after substituting the definition of $j_*$. Thus $\eta_j \geq 1$ if and only if $j \leq j_*$. Recall that $S_j(\epsilon) = \sup_{\Theta} K_{\epsilon,j}(\theta_j)$. Starting with $j \leq j_*$, (17.4) and (17.5) yield

$$S_j \leq c_12^j \epsilon^2 = c_1 \epsilon^2 2(\alpha - j_*)$$

where the definition of $j_*$ implies

$$2^{j_*} \epsilon^2 = (C/\epsilon)^{2(\alpha + 1)} \epsilon^2 = C^{2(1-r)} \epsilon^2.$$

This establishes the small-$j$ side of (17.3) with $\delta = 1$.

For $j \geq j_*$, we have $\eta_j \leq 1$. For $p \geq 2$, (17.4) gives

$$S_j \leq 2^j \epsilon^2 \eta_j^p = 2^j \epsilon^2 2^{-2\alpha(j - j_*)},$$

which (17.7) shows is of the form needed for (17.3).

Finally, for $p < 2$, we note that $\log \eta_j = p(\alpha + 1/2)(j - j_*) \log 2$, and so

$$S_j \leq c_22^j \epsilon^2 2^{-((\alpha+1)/2)p(j - j_*)}[1 + c_3(j - j_*)],$$

since

$$2^j (\alpha + 1/2)p(j - j_*) = 2^j - \alpha \epsilon 2^{j - j_*},$$

we again obtain (17.3), with $\delta = 2\alpha \land 1$ for $p \geq 2$ and $\delta = \alpha \land 1$ for $p < 2$. \qed

**Remarks:** (i) designed to enable a convergence argument; (ii) not directed immediately at threshold estimation, unlike E-Bayes (iii) however, *does* illustrate phenomenon of exact rates. Also, (a) show pictures of ion channel data! (b) review abbreviated notes

17.1. *Wavelet vaguelette decomposition*

17.2. *Near independence*

17.3. *Use of oracle inequalities*
Minimax estimation on Besov spaces

18.1. Introduction...

[Notes for introduction:]

We have now developed the statistical tools for a systematic treatment of minimax function estimation over Besov and Triebel spaces. Exact asymptotic evaluations of minimax risks in global $L_2$ norms are obtained, and the results show clearly the utility of non-linear estimates in the presence of inhomogeneous smoothness.

In fact, it is possible to present the main results in a dyadic sequence model, without any direct mention of function spaces, wavelets and so forth. In order to justify the dyadic sequence model, we of course have to describe the connections to nonparametric function estimation, wavelets and function spaces, but this can be done separately.

This work began in 1990 after Donoho heard Kerkyacharian and Picard speak at St. Flour on the use of wavelets in linear estimation of densities. The results in this chapter are taken from Donoho & Johnstone (1998) and Donoho & Johnstone (1999), which have not yet appeared, but the latest versions can be downloaded from http://wwwmaths.anu.edu.au/research.reports/98srr.html

18.2. The Dyadic Sequence Model

We consider basic the Gaussian sequence model (3.1) with countable index set. However, we write it using a dyadic indexing regime

\begin{equation}
 y_I = \theta_I + \varepsilon_I
\end{equation}

where $I$ denotes the pair $(j, k)$, supposed to lie in the set $\mathcal{I} = \cup_{j \geq -1} \mathcal{I}_j$, where for $j \geq 0$, $\mathcal{I}_j = \{(j, k) : k = 1, \ldots, 2^j\}$ and the exceptional $\mathcal{I}_{-1} = \{(-1, 0)\}$.

Remark. The indexing scheme may be understood in terms of an orthonormal wavelet basis for $L_2[0, 1]$ - we use the Haar basis for illustration here. For $j \geq 0$, the $(j, k)$th basis function is supported on the interval $I_{jk} = [(k - 1)2^{-j}, k2^{-j}]$. The exceptional index $(-1, 0)$ is attached to the Haar scaling function $\phi(x) = I\{0 \leq x \leq 1\}$ needed to complete the orthogonal system.
Parameter Space. It is assumed that \( \theta \) belongs to a Besov body \( \Theta = \Theta^0_{p,q}(C) \), defined by

\[
(18.2) \quad \Theta^0_{p,q}(C) = \{ \theta = (\theta_I) : \sum_j 2^{j\alpha} \| \theta_j \|_p^q \leq C^q \} \quad \alpha = \alpha + 1/2 - 1/p.
\]

Here \( \theta_I \) denotes \( \{ \theta_I : I \in I_j \} \), and as usual \( \| \theta_j \|_p = \sum_{k=1}^{2^j} |\theta_{jk}|^p \), except in the case \( p = \infty \), where \( \| \theta \|_\infty = \sup_k |\theta_{jk}| \).

We note that \( \Theta \) is solid and orthosymmetric, and compact when \( \alpha > 0 \), i.e. when \( \alpha > 1/p - 1/2 \).

As usual, we will be interested in global \( \ell_2 \) estimation: that is we evaluate estimators with the loss function \( \| \hat{\theta} - \theta \|_2^2 = \sum (\hat{\theta}_I - \theta_I)^2 \) and the minimax risk

\[
R_N(\Theta, \varepsilon) = \inf_{\hat{\theta}} \sup_{\theta} E_\theta \| \hat{\theta} - \theta \|_2^2.
\]

In principle, a similar development could be carried out for the \( \ell_p \) loss \( \| \hat{\theta} - \theta \|_p = \sum |\hat{\theta}_I - \theta_I|^p \), or weighted losses of the form \( \sum_j 2^{j\alpha} \sum_k |\hat{\theta}_{jk} - \theta_{jk}|^p \).

The theory will be developed for this sequence model below. Of course, the reason for studying this model, and specifically for the choice of the Besov body parameter spaces, lies in the sequence space characterization of the Besov function spaces in wavelet bases. More later.

18.3. Bayes minimax problem

We relax the ‘hard’ constraint that \( \| \theta \|_{\Theta^0_{p,q}} \leq C \) by a constraint ‘in mean’ with respect to a prior \( \pi \). Given \( \pi \), define the \( s \)-th moment sequence \( \tau_s = \tau_s(\pi) \) by

\[
\tau_s(\pi) = \left( E_\pi |\theta_I|^s \right)^{1/s}.
\]

If \( s = \infty \), put \( \tau_\infty(\pi) = \text{ess sup} |\theta_I| \). Define a class of priors\(^2\) \( \mathcal{M}_s \) by

\[
\mathcal{M}_s = \mathcal{M}_s(\Theta^0_{p,q}(C)) = \{ \pi : \tau_s(\pi) \in \Theta^0_{p,q}(C) \}.
\]

Clearly \( \Theta \subset \mathcal{M}_s \) for all \( s \leq \infty \), and the moment spaces \( \mathcal{M}_s \) are decreasing as \( s \) increases: indeed \( \mathcal{M}_\infty \subset \mathcal{P}(\Theta) \), the collection of priors supported on \( \Theta \). However it turns out\(^3\) that \( \mathcal{M}_s \) is convex only if \( s \leq p \wedge q \), and so we will adopt the specific choice \( \mathcal{M} = \mathcal{M}_{p,q} \) in this chapter, except for the lower bound in Section 18.6, which uses \( \mathcal{M}_\infty \).

As in earlier chapters, define the integrated risk \( B(\hat{\theta}, \pi) = E_\pi E_{\theta} \| \hat{\theta} - \theta \|_2^2 \) and the Bayes minimax risk

\[
B(\mathcal{M}, \varepsilon) = \inf_{\theta} \sup_{\pi \in \mathcal{M}} B(\hat{\theta}, \pi).
\]

\(^1\)w

\(^2\)See the Notes section at the end of the chapter for notes corresponding to footnote marks

\(^3\)
Since $\Theta \subset \mathcal{M}$, $R_N(\Theta, \epsilon) \leq B(\mathcal{M}, \epsilon)$. To emphasize the dependence on $C$ and $\epsilon$, we sometimes write $B(C, \epsilon)$ for $B(\mathcal{M}, \epsilon)$.

A first key property of the Bayes minimax problem is that minimax estimators are separable into functions of each individual coordinate:

18.1. Theorem. Suppose that $0 < p, q \leq \infty$ and $\alpha + 1/2 > 1/(2^p \wedge q)$. A minimax estimator for $B(\mathcal{M}, \epsilon)$ has the form

\begin{equation}
\hat{\theta}_I^*(y) = \hat{\theta}_I^*(y_I), \quad I \in \mathcal{I},
\end{equation}

where $\hat{\theta}_I^*(y)$ is a scalar non-linear function of the scalar $y$. In fact there is a two parameter family of functions from which the minimax estimator is built. Let $\hat{\delta}(x; \eta, p)$ be the Bayes minimax estimator for the univariate Bayes minimax problem $\beta_p(\eta, 1)$ defined at (14.23). Then

\begin{equation}
\hat{\delta}_I^*(y_I) = \varepsilon \hat{\delta}(y_I / \epsilon; \eta_I, p \wedge q),
\end{equation}

where $\eta_I = \eta_I(\alpha, p, q, C\epsilon^{-1})$.

For $p \neq 2$, neither $\eta_I$ nor the explicit form of $\hat{\delta}(\cdot; \eta, p)$ is available. However renormalization arguments will at least give information on the asymptotic dependence of $\eta_I$ on $C$ and $\epsilon$, and we will also see that useful approximations of $\hat{\delta}(\cdot; \eta, p)$ by threshold rules are possible.

Second, the exact asymptotic structure of the Bayes minimax risk can be determined.

18.2. Theorem. Suppose that $0 < p, q \leq \infty$ and $\alpha + 1/2 > 1/(2^p \wedge q)$. Then $B(C, \epsilon) < \infty$ and

\begin{equation}
B(C, \epsilon) \sim P(C / \epsilon) \cdot C^{2(1-r)} \epsilon^{2r}, \quad \epsilon \to 0,
\end{equation}

where $r = 2\alpha / (2\alpha + 1)$ and $P(\cdot) = P(\cdot; \alpha + 1/2, p \wedge q, q)$ is a continuous, positive periodic function of $\log_2(C / \epsilon)$.

This periodic function might be viewed as reflecting the arbitrary choice of the location of frequency octaves that is implicit in discrete dyadic wavelet bases.

Third, we provide lower bounds for $R_N(\Theta, \epsilon)$, and, when $q \geq p$, establish asymptotic equivalence of frequentist and minimax Bayes risk.

18.3. Theorem. For $\alpha > 0$, $a, p, q > 0$,

\begin{equation}
R_N(\Theta, \epsilon) \geq \gamma(C / \epsilon) C^{2(1-r)} \epsilon^{2r} - \epsilon^2, \quad \epsilon > 0,
\end{equation}

where $r$ is as above, and $P(\cdot) = P(\cdot; \alpha + 1/2, \infty, q)$ is a continuous, positive, periodic function of $\log_2(C / \epsilon)$. If $q \geq p$, then

\begin{equation}
R_N(\Theta, \epsilon) = B(C, \epsilon)(1 + o(1)), \quad \epsilon \to 0.
\end{equation}
Combining Theorems 18.2-18.3, we have in the case \( p \leq q \) that the estimator \( \hat{\theta}^* \) is asymptotically minimax for \( R \) as \( \epsilon \to 0 \). In short: a separable nonlinear rule is asymptotically minimax. In the case \( p > q \), the Bayes-Minimax estimator is within a constant factor of minimax.

We have given statements of the results for the Besov case in full generality, and will give an indication of the structure of the proofs. When technical details obtrude, we retreat to the particular case \( p = 1, q = \infty \), for which a more or less complete description is given. For fuller details, and the Triebel case, we refer to Donoho & Johnstone (1998).

### 18.4. Separable rules

We begin the proof of Theorem 18.1 by noting three important properties of \( \mathcal{M} \), which in the case \( p = 1, q = \infty \) becomes

\[
\mathcal{M} = \mathcal{M}_{1,\infty}(C) = \{ \pi : \sum_k E_\pi |\theta_k| \leq C2^{-aj} \quad \forall j \}.
\]

(i) **Compactness.** Via Prohorov’s theorem (REF), one shows that \( \mathcal{M} \) is uniformly tight: in other words that for all positive \( \epsilon \), there exists a compact \( K_\epsilon \) such that for all \( \pi \in \mathcal{M} \), one has \( \pi(K_\epsilon) < \epsilon \). This can be accomplished using sets of the form \( K(M) = \{ \theta : \|\theta_j\|_1 \leq MC2^{-nj} \text{ for all } j \} \) for small positive \( \eta \) and large \( M \).

(ii) **Convexity.** This is trivial when \( p = 1, q = \infty \), since the moment condition is linear in \( \pi \).

(iii) **Closure under product formation.** Given a prior \( \pi \in \mathcal{M} \), form the univariate marginals \( \pi_{jk} \) and then levelwise averages \( \bar{\pi}_j = \text{ave}_k(\pi_{jk}) \). Form a new prior \( \bar{\pi} \) by making \( \theta_{jk} \) independent, with \( \theta_{jk} \sim \bar{\pi}_j \). The closure property requires that \( \bar{\pi} \in \mathcal{M} \), but this is again trivial when \( p = 1, q = \infty \) : because by construction

\[
\sum_k E_{\bar{\pi}} |\theta_k| = \sum_k E_{\pi} |\theta_k|.
\]

**Remark.** These properties are less trivial for general \( (p, q) \). In particular, the convexity and product closure properties follow from the use of the \( p \wedge q \)-th moments, rather than the \( p \)-th moments one might prefer to use. The assumption \( \alpha + 1/2 > 1/(2 \wedge p \wedge q) \) is then used to establish compactness.

The compactness and convexity properties allow use of the minimax theorem Theorem 6.6 to write that \( B(\mathcal{M}) = \sup_{\mathcal{M}} B(\pi) \), so that we may look for a least favorable prior. Given any prior \( \pi \in \mathcal{M} \), form the levelwise product prior \( \bar{\pi} \) as described earlier: the closure property shows that \( \bar{\pi} \) is still in \( \mathcal{M} \). As we showed in earlier chapters (e.g., (??)), the prior \( \bar{\pi} \) is more difficult for Bayes estimation, so \( B(\bar{\pi}) \geq B(\pi) \). Thus it suffices to maximise over priors \( \bar{\pi} \in \mathcal{M} \).
The independence structure of $\pi$ means that the Bayes estimator $\hat{\theta}_\pi$ is separable - since prior and likelihood factorize, so does the posterior, and so

$$\hat{\theta}_{\pi,I} = E_{\hat{\pi}_I}(\theta_I|y_I).$$

In addition, the Bayes risk is additive: $B(\pi) = \sum I B(\pi_I)$. Recalling now from (18.3) that $M$ is defined by moment constraints, we have

$$B(M) = \sup \{ \sum I B(\pi_I) : (\pi_I(\pi_I)) \in \Theta \},$$

where in addition, because of the symmetry induced by averaging over coordinates $k$ within a fixed level, we may restrict to sequences $\pi_I = \tau_{ik} \equiv \tau_I$ that depend on the level $j$ only.

Since $\Theta$ is solid, we may rewrite

$$B(M) = \sup_{\tau \in \Theta} \{ \sum I B(\pi_I) : \tau(\pi_I) \leq \tau_I \forall I \}.$$

The inner supremum can be performed co-ordinatewise, and so we may exploit the univariate minimax Bayes problem studied in Section 14.2: $\beta_p(\tau, \epsilon) = \sup \{ B(\pi, \epsilon) : E_x[\Theta^p \leq \tau^p] \}. \quad \text{Introducing the notation } B(C, \epsilon) \text{ for } B(M) \text{ in order to emphasize the dependence on } (C, \epsilon), \text{ we have}

$$B(C, \epsilon) = \sup \{ \sum I \beta_{p/q}(\tau_I, \epsilon) : \tau \in \Theta(C) \}.$$
18.5. Exact Bayes minimax asymptotics.

To start the proof of Theorem 18.2, we rewrite (18.7) as
\[ B(C, \epsilon) = \sup \left\{ \sum_{j=0}^{\infty} 2^j \beta(t_j, \epsilon) \mid \sum_{j=0}^{\infty} (2^j \alpha^{-1/p})^q \leq C^q \right\}, \]
with the usual reformulation if \( p = \infty \) or \( q = \infty \). Here \( \beta = \beta_{\alpha/\epsilon} \). At first glance, solution of this problem would appear to be beyond reach, owing to the fact that we have no closed form expression for \( \beta_{\alpha/\epsilon} \) when \( p \neq 2 \). However, a certain "renormalizability" of the problem provides a tool to get qualitative insights.

First, define a modified optimization problem on the space of bilateral sequences \( (t_j)_{j \in \mathbb{Z}} \) : set \( \omega = a + 1/p = \alpha + 1/2 \), and
\[ Q(C, \epsilon) = \sup \left\{ \sum_{j=-\infty}^{\infty} 2^j \beta(t_j, \epsilon) \mid \sum_{j=-\infty}^{\infty} (2^j \omega^{-1/p})^q \leq C^q \right\}. \]
If a unilateral sequence \( (t_j)_{j \geq 0} \) is feasible for \( B(C, \epsilon) \), then defining \( t_j = 0 \) for negative \( j \) creates a bilateral sequence feasible for \( Q(C, \epsilon) \). On the other hand, since \( \beta(t, \epsilon) \leq \epsilon^2 \), the component of \( Q(C, \epsilon) \) due to negative indices is a geometric series with sum \( \epsilon^2 \) and so \( Q \leq B + \epsilon^2 \). Hence
\[ (18.10) \quad Q(C, \epsilon) - \epsilon^2 \leq B(C, \epsilon) \leq Q(C, \epsilon) \quad \forall \epsilon > 0, \ C > 0. \]
Since a discrepancy of order \( \epsilon^2 \) is negligible in non-parametric problems as \( \epsilon \to 0 \), we may safely proceed to study \( Q(C, \epsilon) \).

**Special case \( p = 1, q = \infty \):** When \( q = \infty \), the moment constraint becomes \( t_j \leq C2^{-\omega j} \) for all \( j \), and so the monotonicity of \( \beta_{1}(t, \epsilon) \), along with the scaling relation Proposition 3.6(2) allows the more explicit evaluation
\[ (18.11) \quad Q(C, \epsilon) = \epsilon^2 \sum_{j=-\infty}^{\infty} 2^j \beta_{1}(C/\epsilon 2^{-\omega j}, 1). \]
Let us check convergence of the sum: the only difficulty comes from large positive \( j \) : for this we use the approximation derived in Theorem 14.5:
\[ \beta_{1}(\eta) \sim \eta (2 \log \eta^{-1})^{1/2} \quad \text{as} \quad \eta \to 0 . \]
It follows that
\[ 2^j \beta_{1}(C e^{-1} 2^{-\omega j}) \approx (C/\epsilon)2^{-\omega j} \sqrt{j} \sim \sqrt{j} 2^{-\omega (\alpha - 1/2) j} \]
which converges if and only if \( \alpha > 1/2 \). In the general case, such convergence considerations yield the condition \( \alpha + 1/2 > 1/(2 \wedge p \wedge q) \) in Theorem 18.2.

As \( \epsilon \) decreases, the location \( j \) of the maximum summand increases. To account for this effect and to stabilize the sum, write \( C/\epsilon = 2^{v \epsilon} \) and change variables to \( v = j - \zeta \). We obtain
\[ Q(C, \epsilon) = \epsilon^2 2^{\zeta} \sum_{v \in \mathbb{Z} - \zeta} 2^v \beta_{1}(2^{-v \epsilon}) = \epsilon^2 2^{\zeta} P_{1}(\{\zeta\}), \]
where the last inequality reflects the fact that the sum depends only on the fractional part \( \{\zeta\} \) of \( \zeta \), and hence is 1-periodic.
Since \(1/\omega = 2/(2\alpha + 1) = 2(1 - r)\) and \(2^\zeta = (C/\epsilon)^{1/\omega}\), we arrive at
\[
Q(C, \epsilon) = P_1(\{\omega^{-1} \log_2(C\epsilon^{-1})\}) \cdot 2^{2(1-r)\epsilon^2}. 
\]

**Remark.** How does the location of the maximum \(j\) in \(Q(C, \epsilon)\) depend on \(\epsilon\)? Suppose that \(v_s\) is the location of the maximum of the function \(v \rightarrow 2^j \beta(2^{-\omega j})\) appearing in the unit-noise, renormalized problem \(P(\{\zeta\}) = 2^{-|\zeta|}Q(2^{-|\zeta|}, 1)\). Then the maximum in \(Q(C, \epsilon)\) occurs at \(u_s = v_s + \zeta = v_s + \omega^{-1} \log_2(C/\epsilon)\). Using the calibration \(\epsilon = n^{-1/2}\) and \(\omega = 1/(2\alpha + 1)\), we can interpret this in terms of equivalent sample sizes as
\[
u_s = c_s + \frac{\log_2 n}{1 + 2\alpha}, \quad c_s = v_s + \frac{\log_2 C}{\alpha + 1/2}.
\]
The “most difficult” resolution level for estimation is therefore at about \((\log_2 n)/(1 + 2\alpha)\). This is strictly smaller than \(\log_2 n\) for \(\alpha > 0\), meaning that so long as the sum (18.11) converges, the primary contributions to the risk \(B(C, \epsilon)\) come from levels below the finest (at \(\log_2 n\) corresponding to a sample of size \(n\)).

**General case.** When \(q < \infty\), the levels \(j\) do not decouple in the fashion that led to (18.11). We may obtain similar asymptotic behavior by using homogeneity properties of the \(Q(C, \epsilon)\) problem with respect to scaling and level shifts. Indeed
\[
Q(C, \epsilon) = \sup\{J_{\beta, \epsilon}(t) \text{ subject to } J_{q, \omega}(t) \leq C\},
\]
where
\[
J_{\beta, \epsilon}(t) = \sum_{-\infty}^{\infty} 2^j \beta(t_j, \epsilon), \quad J_{q, \omega}(t) = [\sum_{-\infty}^{\infty} (2^{\omega j} t_j)^q]^{1/q}.
\]

To emphasize the dependence on \(C\) and \(\epsilon\), write \(\tau(C, \epsilon)\) for the bilateral sequence that maximizes \(Q(C, \epsilon)\). Introduce transformations \((U_{c,h} t)_j = c t_{j-h}\): simple changes of variables and the scaling property for \(\beta\) show that
\[
J_{\beta, \epsilon}(U_{c,h} t) = 2^{2h} J_{\beta, 1}(t), \quad J_{q, \omega}(U_{c,h} t) = 2^{2h} J_{q, \omega}(t).
\]
In other words, \(J_{q, \omega}(t) \leq C e^{-1} 2^{-\omega h}\) if and only if \(J_{q, \omega}(U_{c,h} t) \leq C\). Hence \(\tau\) maximizes \(Q(\epsilon/12^{-\omega h}, 1)\) if and only if \(U_{c,h} \tau\) maximizes \(Q(C, \epsilon)\). So far \(h \in \mathbb{Z}\) has been a free variable: now we choose it so that \(C/(\epsilon 2^{-\omega h}) = 2^{-\eta}\) for \(\eta \in [0, 1)\) - thus \(h\) and \(\eta\) are respectively the integer and fractional parts of \(\omega^{-1} \log_2(C/\epsilon)\). We have \(\tau(C, \epsilon) = U_{c,h} \tau(2^{\omega \eta}, 1)\), or in other words
\[
(18.13) \quad \tau(C, \epsilon)_j = \epsilon \tau(2^{2\eta}, 1)_{j-h},
\]
and
\[
Q(C, \epsilon) = \epsilon^{2h} Q(2^{2\eta}, 1).
\]
Since \(2^{h+\eta} = (C/\epsilon)^{1/\omega}\), we again recover (18.12), except that the 1-periodic function \(P_1\) is now given by \(P_1(\eta) = 2^{-\eta} Q(2^{2\eta}, 1)\) and the function \(P\) of Theorem 18.2 is given by \(P(C/\epsilon) = P(\omega^{-1} \log_2 C/\epsilon)\).
By combining (18.13) with (18.9) we obtain a representation for the Bayes minimax estimator in terms of the two dimensional family \( \tilde{\beta}(\cdot, \eta, p) \) that incorporates the renormalization structure:

\[
\tilde{\beta}_I(y; C, \epsilon) = c\tilde{\beta}(y_I / \epsilon; \eta_{j-h(2^N)}, p \wedge q).
\]

Example: \( p = q = 2 \). In this case \( \beta_2(\eta) = \eta^2/(1 + \eta^2) \) and the moment constraints are ellipsoidal, so Pinsker’s method allows a more explicit solution to be given. In addition, \( \tilde{\beta}(x; \eta, 2) = ux = [\eta^2/(1 + \eta^2)]x \). Consider \( \epsilon \) of the form \( C/\epsilon = 2^{-h} \) for \( h \) integer. The vector \( \eta = (\eta_j(1)) \) that solves the unit noise problem

\[
\sup \left\{ \sum_{-\infty}^{\infty} 2^j \beta_2(\eta_j, 1) : \sum_{-\infty}^{\infty} 2^{2^j} \eta_j^2 \leq 1 \right\}
\]

is found by a calculation completely analogous to (8.6), after accounting for the weight factor \( 2^j \) in the objective. One finds

\[
\eta_j^2 = (\lambda 2^{-\alpha j} - 1)_+, \quad w_j = (1 - 2^{\alpha j}/\lambda)_+,
\]

where \( \lambda \) is chosen to satisfy the constraint \( \sum 2^{\alpha j}(2^{\alpha}/2 \lambda - 2^{\alpha j})_+ = 1 \). This can be evaluated by writing \( \lambda = 2^\rho \psi : \) the integer part of \( \rho \) can be shown to be given by the integer part of \( (2\alpha + 1)^{-1} \log_2 R(\alpha) \) where \( R(\alpha) \) is a rational function of \( 2^\alpha \) given by ?????.

It can be shown that \( r \geq 0 \) for all \( \alpha \), and that \( r = 0 \) if \( \alpha > 1 \); in this case only non-positive terms in the sum (18.15) contribute. On the other hand, \( r \) increases to \( \infty \) as \( \alpha \) decreases to 0. Finally, from (18.14) and (18.16), we have an explicit formula for the Bayes minimax estimator

\[
\tilde{\beta}_I(y; C, \epsilon) = (1 - 2^{\alpha(j-h-\rho)})_+ y_I.
\]

18.6. Lower Bounds and Asymptotic Efficiency when \( p \leq q \).

Lower Bound A lower bound valid for all \( p \) and \( q \) can be obtained from the minimax theorem (6.17) – which expresses \( R_N(\Theta) = \sup \{ B(\pi) : \sup \pi \subseteq \Theta \} \). Returning to (18.3), since every prior in \( \mathcal{M}_\infty \) is supported in \( \Theta \), we have

\[
R_N(\Theta) \geq \sup \{ B(\pi), \pi \in \mathcal{M}_\infty \}.
\]

The separability arguments of Section 18.4 may now be applied to yield

\[
R_N(\Theta) \geq \sup \{ \sum \beta_\infty(\tau_I, \epsilon) : \tau \in \Theta \},
\]

which is exactly of the form considered in the asymptotic renormalization analyses of the previous section, only with the replacement of \( \beta_{\rho \mid q} \) by \( \beta_\infty \). In combination with (18.10), this yields the lower bound (18.5) of Theorem 18.3.

Equivalence for \( q \geq p \). We again use the approach outlined in Chapter 6.4, which involves constructing near least favorable priors \( \pi_\epsilon \) that asymptotically concentrate on \( \Theta \) as \( \epsilon \rightarrow 0 \). More specifically, in line with the
strategy (6.23) - (6.25), for each \( \gamma < 1 \), we construct \( \pi_\epsilon \in \mathcal{M}_{p,q} \) such that \( B(\pi_\epsilon) > \gamma B(\epsilon, \gamma C) \) and verify that \( \pi_\epsilon(\Theta) \to 1 \). We will again focus on the case \( p = 1, q = \infty \), and omit the technical step (6.25). Finally, to indicate why the assumption \( p \leq q \) is needed, we show how the corresponding construction fails when \( p = \infty, q = 1 \).

We also focus on a subsequence of noise levels \( \epsilon_h \) defined by \( C/\epsilon_h = 2^{\omega h} \). Substituting this and the shrunken radius constraint \( \gamma C \) into (18.11) yields

\[
Q(\gamma C, \epsilon_h) = \epsilon_h^{2\beta_h} \sum_j 2^j \beta_h(\gamma 2^{-\omega j}).
\]

For each level \( j \), choose a nearly least favorable prior subject to the corresponding first moment constraint: given \( \gamma < 1 \) it is clear that one can choose a positive integer \( J \) and univariate priors \( \pi_j \) for \(-J \leq j \leq J \) such that

\[
E_{\pi_j} |\mu| \leq \gamma 2^{-\omega j} \quad \text{for} \quad |j| \leq J, \quad \text{and} \quad \sum_{-J}^J 2^j B(\pi_j) > \gamma \sum_{-\infty}^\infty 2^j \beta_h(\gamma 2^{-\omega j}).
\]

The approximately least favorable prior \( \pi_{\epsilon_h} \) on \( \Theta \) is built as follows. Let \( \{\mu_{j,k}, k \in \mathbb{N}\} \) be an i.i.d. sequence drawn from \( \pi_j \). For \( |j| \leq J \), set

\[
\theta_{h+j,k} = \epsilon_h \mu_{j,k} \quad k = 1, \ldots, 2^{h+j}
\]

Thus \( \pi_{\epsilon_h} \) concentrates on the \( 2J + 1 \) levels centered at \( h = \omega^{-1} \log_2 C/\epsilon \). Hence, as \( \epsilon \to 0 \), the near least favorable priors charge (a fixed number of) ever higher frequency bands.

After chasing definitions and scalings, one verifies from (18.18) that \( B(\pi_{\epsilon_h}) \geq \gamma B(\epsilon_h, \gamma C) \). Recalling the definition of \( \epsilon_h \) and that \( a = \alpha - 1/2 = \omega - 1 \),

\[
\theta \in \Theta \iff \sum_k |\theta_{h+j,k}| \leq C 2^{-\alpha(h+j)} \quad \text{for} \quad |j| \leq J,
\]

\[
\iff 2^{-(h+j)} \sum_k |\mu_{j,k}| \leq 2^{-\omega j} \quad \text{for} \quad |j| \leq J.
\]

Now apply the strong law of large numbers to each of the fixed number of levels: from the moment condition (18.17) it follows that \( \pi_{\epsilon_h}(\Theta) \to 1 \) as \( h \to \infty \) and \( \epsilon_h \to 0 \).

**Remark.** Again consider priors of the form (18.19). What goes wrong when \( p > q \)? If \( p = \infty, q = 1 \), then the event \( \theta \in \Theta \) is now equivalent to

\[
\sum_j 2^{\omega j} \sup_k |\theta_{j,k}| \leq C,
\]

while since \( \tau_{p,q}(\pi) = E_{\pi} |\theta_{j,k}| \), the condition \( \tau(\pi) \in \Theta(\gamma C) \) amounts to

\[
\sum_j 2^{\omega j} \sup_k E|\theta_{j,k}| \leq \gamma C.
\]
However, on passing to the variables $\mu_{jk}$ at levels $h + j$ for $|j| \leq J$, one sees that
\[
\theta \in \Theta \iff \sum_{-J}^{J} 2^{2j} \sup_{1 \leq k \leq 2^{h+j}} |\mu_{jk}| \leq C
\]
and this does not follow from a condition on first moments: $\sum_{-J}^{J} 2^{2j} E|\mu_{j1}| \leq \gamma C$, since the tails of the least favorable distribution of $\mu_{jk}$ will be large.

18.7. Linear Estimates

Using results from Chapter 7, it is relatively straightforward to show that over Besov bodies with $p < 2$, linear estimates are suboptimal, even at the level of rates of convergence.

First, we recall that the Besov bodies $\Theta = \Theta^p_{p,q}(C)$ are solid and orthosymmetric, so that by Theorem 7.11 the linear minimax risk is determined by the quadratic hull of $\Theta$. It is straightforward to check that
\[
QHull(\Theta^p_{p,q}) = \Theta^q_{p',q'} \quad p' = p \lor 2, q' = q \lor 2, \alpha' = \alpha - 1/p + 1/p'.
\]
In particular, $\Theta^p_{p,q}$ is quadratically convex only if both $p$ and $q$ are at least 2. The Ibragimov-Hasminskii theorem shows that the linear minimax risk of a quadratically convex solid orthosymmetric set is between 1 and $5/4$ times the non-linear minimax risk. Hence
\[
R_L(\Theta^p_{p,q}(C), \epsilon) \asymp R_N(\Theta^q_{p',q'}, \epsilon)
\]
(18.20)
\[
\asymp C^{2(1-r')} C^{2r'} \quad r' = 2\alpha'/2\alpha' + 1.
\]

In particular, when $p < 2$, we have $\alpha' = \alpha - (1/p - 1/2)$, so that the linear rate $r'$ is strictly smaller than the minimax rate $r$. For example, on the Besov body $\Theta^1_{1,1}$ corresponding to the Bump Algebra, one finds that $\alpha' = 1/2$ and so the linear minimax rate is $O(\epsilon)$, whereas the non-linear rate is much faster, at $O(\epsilon^{4/3})$.

Let us conclude this section with some remarks about the structure of minimax linear estimators. Since the spaces $\Theta = \Theta^p_{p,q}(C)$ are symmetric with respect to permutation of co-ordinates within resolution levels, it is intuitively clear that a minimax linear estimator will have the form $\theta = \hat{\theta}_{j,c_j}$, where for each $j$, $c_j \in [0,1]$ is a scalar and
\[
\hat{\theta}_{j,c_j} = c_j y_j,
\]
(18.21)
and hence that
\[
R_L(\Theta, \epsilon) = \inf_{\{c_j\}} \sup_{\theta} \sum_j E\|\hat{\theta}_{j,c_j} - \theta_j\|^2.
\]
(18.22)

A formal verification again uses the observation that $R_L(\Theta) = R_L(\tilde{\Theta})$ where $\tilde{\Theta} = QHull(\Theta) = \Theta^{q}_{p',q'}$ as described earlier. Given $r \in \tilde{\Theta}$, construct $\tau$ by setting $\tau^2_{jk} \equiv \text{ave}_{k} \tilde{\tau}^2_{jk}$; since $p' \geq 2$, one verifies that $\tau \in \tilde{\Theta}$ also. Formula (7.29) shows that $R(\Theta(\tau))$ is a concave function of $(\tau^2_{jk})$, and hence that
\( R(\Theta(\bar{r})) \geq R(\Theta(\tau)) \). Consequently, the hardest rectangular subproblem lies among those hyperrectangles that are symmetric within levels \( j \). Since the minimax linear estimator for rectangle \( \bar{r} \) has the form \( \hat{\theta}_{l(\bar{r}), l} = \frac{\bar{r}_j^2}{(\bar{r}_j^2 + \epsilon^2)}y \), it follows that the minimax linear estimator for \( \Theta \) has the form (18.21), which establishes (18.22).

### 18.8. Near Minimaxity of Threshold Estimators

Summary to be expanded later.

Although described in terms of a two parameter family of co-ordinatewise Bayes estimators, the asymptotic minimax estimators derived at (18.14) are still not available in fully explicit form. In this section, we show that nearly minimax estimators exist within the family of soft threshold estimators.

The approach is a multilevel version of that taken for \( \ell_p \) balls in Chapter 14. Given the moment constrained family of probability measures \( \mathcal{M} = \mathcal{M}_{p,q}(\Theta) \), define a minimax Bayes risk among threshold estimators

\[
B_S(\mathcal{M}, \epsilon) = \inf_{(\lambda)} \sup_{\pi \in \mathcal{M}} B(\hat{\theta}_{\lambda}, \pi) = \sup_{\mathcal{M}} B_S(\pi),
\]

where the second inequality requires a special purpose minimax theorem in the spirit of Theorem ???.

Proposition 14.11 showed that the univariate Bayes minimax risk for soft thresholding subject to a \( p \)-th moment constraint \( \tau \) satisfies the inequality \( \beta_{S,p}(\tau, \epsilon) \leq \Lambda(p)\beta_p(\tau, \epsilon) \) for all \( 0 < \tau, \epsilon < \infty \).

The result lifts to sequence space:

**18.4. Theorem.** \( B_S(\mathcal{M}, \epsilon) \leq \Lambda(p \wedge q)B(\mathcal{M}, \epsilon) \).

When \( p \leq q \), the asymptotic equivalence of Bayes and frequentist minimax risks allows a corresponding statement for thresholding over \( \Theta_{p,q}^\infty \).

Define

\[
(18.23) \quad R_S(\Theta, \epsilon) = \inf_{(\lambda)} \sup_{\theta \in \Theta} r(\hat{\theta}_{\lambda}, \theta)
\]

**18.5. Corollary.** If \( p \leq q \), then \( R_S(\Theta, \epsilon) \leq \Lambda(p)R_N(\Theta, \epsilon)(1 + o(1)) \) as \( \epsilon \to 0 \).

Since \( \Lambda(p) \leq 2.22 \) for \( p \geq 2 \), and \( \Lambda(1) \approx 1.6 \), these results provide some assurance that threshold estimators achieve nearly optimal minimax performance. The particular choice of threshold still depends on the parameters \( (\alpha, p, q, C) \), however. Special choices of threshold not depending on a prior specifications of these parameters will be discussed in later chapters.

Later: least favorable priors over \( \mathcal{M} \) for soft thresholding allow visualization of sample paths.

Similar results may be established for hard thresholding.
18.9. Notes

Remark. Here and in preceding chapters we have introduced various spaces of moment-constrained probability measures. These are all instances of a single method, as is shown by the following slightly cumbersome notation. If \( \pi \) is a probability measure on \( \ell_2(\mathcal{I}) \), let \( \tau_p(\pi) \) denote the sequence of marginal \( p \)th moments

\[
\tau_p(\pi)_I = (E[I \theta_I|^p])^{1/p}, \quad I \in \mathcal{I}, \quad p \in (0, \infty].
\]

If \( \Theta \) is a parameter space contained in \( \ell_2(\mathcal{I}) \), then set

\[
\mathcal{M}_p(\Theta) = \{ \pi \in \mathcal{P}(\ell_2(\mathcal{I})): \tau_p(\pi) \in \Theta \}.
\]

In the following examples, the left side gives the notation used in the text, and the right side the notation according to the convention just introduced.

(i) Intervals \( \Theta = [-\tau, \tau] \subset \mathbb{R} \):

\[
\mathcal{M}_p(\tau) = \mathcal{M}_p([-\tau, \tau]).
\]

(ii) \( \ell_p \) balls:

\[
\mathcal{M}_n = \mathcal{M}_p(\Theta_{n,p}(C)),
\]

(iii) Ellipsoids in Pinsker’s Theorem:

\[
\mathcal{M}(C) = \mathcal{M}_2(\Theta(C)),
\]

(iv) Besov bodies:

\[
\mathcal{M}^0_{p,q}(C) = \mathcal{M}_p(\Theta_{p,q}(C)).
\]
CHAPTER 19

[Unrev'd] Pesky Epilogue: Continuous v. Sampled Data

Our theory has been developed so far exclusively in the Gaussian sequence model (3.1). In this chapter, we indicate some implications of the theory for models that are more explicitly associated with function estimation. We first consider the continuous white noise model

\begin{equation}
    dY_t = f dt + dW_t \quad t \in [0, 1],
\end{equation}

which we will see is in fact an equivalent representation of (3.1).

Less trivial, but closer to many applications is the sampled data model in which one observes

\begin{equation}
    \hat{y}_i = n^{-1/2} f(i/n) + e_i, \quad i = 1, \ldots n,
\end{equation}

and it is desired to estimate the function \( f \in L_2[0,1] \).

For many purposes, the models (19.1) and (19.2) are very similar, and methods and results developed in one should apply equally well in the other. A general equivalence result of Brown & Low (1996) implies that for bounded loss function \( \ell(\cdot) \) and for collections \( \mathcal{F} \) which are bounded subsets of Hölder classes \( C^\alpha, \alpha > 1/2 \), we have as \( \epsilon \to 0 \),

\begin{equation}
    \inf_{f} \sup_{f \in \mathcal{F}} E\ell \left( \| \hat{f}(Y) - f \|_{L_2[0,1]}^2 \right) \sim \inf_{f} \sup_{f \in \mathcal{F}} E\ell \left( \| \hat{f}(\hat{y}) - f \|_{L_2[0,1]}^2 \right)
\end{equation}

the expectation on the left-hand side being with respect to white noise observations \( Y \) in (19.1) and on the right hand-side being with respect to \( \hat{y} \) in (19.2). However, the general equivalence result fails for \( \alpha \leq 1/2 \) and we wish to establish results for the global estimation problem for the unbounded loss function \( \| \hat{f} - f \|_2 \) that are valid also for Besov and Triebel classes satisfying \( \alpha > 1/p \), where \( p \) might be arbitrarily large.

In addition our development will address directly the common and valid complaint that theory is often developed for “theoretical” wavelet coefficients in model (19.1) while computer algorithms work with empirical wavelet coefficients derived from the sampled data model (19.2). We compare explicitly the sampling operators corresponding to pointwise evaluation and integration against a localized scaling function. The results obtained will be useful in later chapters, even beyond the current setting of exact asymptotic minimaxity.

247
The approach taken in this chapter is based on Donoho & Johnstone (1999). That paper studies the sampled data problem using a suitable order of the interpolating wavelet transform developed by Donoho (1992). Here we develop the theory requiring instead a suitable number of vanishing moments on the scaling function, which allows a more direct treatment of the transition from sampled data to the white noise model. The ideas underlying the proofs are very similar however, so it is relatively safe to refer the reader to Donoho & Johnstone (1999) for details omitted here.

19.1. The Continuous White Noise Model

We describe this in a fashion that makes it entirely equivalent to the sequence model. Let \(\{\psi_I, I \in I\} \) be an orthonormal wavelet basis adapted to \(L_2[0, 1]\) (compare Section B.2.4).

Just as we did with the trigonometric basis in (3.6), we compute the representation of the continuous white noise model in the wavelet basis. Let \(\langle \cdot, \cdot \rangle \) denote the inner product in \(L_2[0, 1]\) and define

\[ y_I = \langle \psi_I, dY \rangle, \quad \theta_I = \langle \psi_I, f \rangle, \quad z_I = \langle \psi_I, dW \rangle \quad I \in I. \]

From (19.1) we obtain \(y_I = \theta_I + \varepsilon_I, \quad I \in I\), and just as in (3.7), the sequence \(z_I\) is i.i.d standard Gaussian. This is precisely the sequence model (18.1).

Every function \(f \in L_2[0, 1]\) has the expansion \(f = \sum \theta_I \psi_I\), and the Parseval relation \(\int f \, d^2 = \sum \theta_I^2\) shows that the mapping from \(f\) to \(\theta\) is an isometry, which we sometimes write \(\theta[f]\). Thus \(\theta[f]_I = \langle f, \psi_I \rangle\) for \(I \in I\). For the inverse mapping, we write \(f[\theta]\) for the function defined by \(f[\theta](t) = \sum I \theta_I \psi_I(t)\).

In the continuous white noise model, we estimate the function \(f\) using mean integrated squared error \(\int (\hat{f} - f)^2\), and of course

\[ \| \hat{f} - f \|_2^2 = \sum_I (\theta_I - \hat{\theta}_I)^2 = \| \theta - \hat{\theta} \|_2^2. \]  

(19.4)

An important property of wavelet bases, reviewed in Section B.3.2, is that the sequence of suitable wavelet coefficients can be used to define norms equivalent to those of the classical Besov and Triebel function spaces.

Since we consider here minimaxity with respect to a fixed function space \(F\), we assume that the choice of \((\alpha, p, q)\) has been fixed (and for convenience within the Besov scale \(B_{p,q}^\alpha[0, 1]\), though corresponding results hold for Triebel spaces.) We choose a wavelet basis compatible with this space.

We therefore assume that the wavelet basis is built from a wavelet \(\psi\) satisfying

\[ (R_\alpha) \quad \psi \text{ has compact support, } R \text{ continuous derivatives and } D \text{ vanishing moments with } \min(R, D) > \alpha. \]

Then the wavelet basis will be an unconditional basis of the corresponding space of interest. (Compare Section ??).
Recalling the definition (18.2) of the sequence norms \( \| \cdot \|_{p,q} \) and Besov bodies \( \Theta^\alpha_{p,q} \), we use the norm equivalence to define function spaces

\[
\mathcal{F} = \mathcal{F}^\alpha_{p,q}(C) = \{ f : \theta[f] \in \Theta^\alpha_{p,q}(C) \}.
\]

Our choice of definitions has made the continuous white noise estimation problem exactly equivalent to the sequence model. Using the natural definition of minimax risks, we therefore have the identity

\[
R_\varepsilon(\mathcal{F}, \epsilon) = \inf_{f \in \mathcal{F}} \sup_{\theta \in \Theta} E_\theta \| \hat{\theta} - \theta \|^2 = R_\varepsilon(\Theta, \epsilon).
\]

Here \( \mathcal{E} \) might denote the class of all estimators. We will also be particularly interested in certain classes of coordinatewise estimators applied to the wavelet coefficients. In the sequence model, this means that the estimator has the form \( \hat{\theta}_I(y) = \hat{\delta}_I \), where \( \hat{\delta} \) belongs to one of the four families in the following table.

<table>
<thead>
<tr>
<th>Family</th>
<th>Description</th>
<th>Form of ( \hat{\delta}_I(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{E}_L )</td>
<td>Diagonal linear procedures in the wavelet domain</td>
<td>( \hat{\delta}_I^L(y) = \alpha \cdot y )</td>
</tr>
<tr>
<td>( \mathcal{E}_S )</td>
<td>Soft thresholding of wavelet coefficients</td>
<td>( \hat{\delta}_I^S(y) = (</td>
</tr>
<tr>
<td>( \mathcal{E}_H )</td>
<td>Hard thresholding of wavelet coefficients</td>
<td>( \hat{\delta}<em>I^H(y) = y1</em>{(</td>
</tr>
<tr>
<td>( \mathcal{E}_N )</td>
<td>Scalar nonlinearities of wavelet coefficients</td>
<td>Arbitrary ( \hat{\delta}_I^N(y) )</td>
</tr>
</tbody>
</table>

The corresponding estimators in classes \( \mathcal{E} \) in (19.7) in the continuous white noise model are defined by \( \hat{f} = f[\hat{\theta}] = \sum_I \hat{\delta}_I(\langle \psi_I, dX \rangle)\psi_I \), where \( \theta \in \mathcal{E}_S, \mathcal{E}_L \) and so on.

### 19.2. The Projected White Noise Model

Finite dimensional submodels of (19.1) are of interest for a number of reasons. Firstly, when the noise level \( \epsilon \) is of order \( n^{-1/2} \), a model with \( n \) observed coefficients is a closer relative of the regression model (19.2). Secondly, for a given parameter space \( \Theta \), finite dimensional submodels can be found with dimension \( m(\epsilon) \) depending on \( \epsilon \) that are asymptotically as difficult as the full model. This proves to be a useful technical tool, for example in proving results for the sampling model.

Let \( \phi \) be the scaling function corresponding to the orthonormal wavelet \( \psi \) used in the previous section. We consider only projections on to the increasing sequence of multiresolution spaces \( V_j = \text{span} \{ \phi_{j,i}, i = 1, \ldots 2^j \} \). Given \( \epsilon \), fix a level \( J = J(\epsilon) \), set \( m = m_\epsilon = 2^J(\epsilon) \) and define

\[
y_i = \langle \phi_{J,i}, dY \rangle, \quad z_i = \langle \phi_{J,i}, dW \rangle, \quad i = 1, \ldots m.
\]
The projected white noise model refers to observations

\begin{equation}(19.8)\quad y_i = \langle f, \phi_{j_i} \rangle + \varepsilon_i, \quad i = 1, \ldots, m.\end{equation}

Write \(y^{(m)}\) for the projected data \(y_1, \ldots, y_m\). When \(\varepsilon = n^{-1/2}\), the choice \(J = \log_2 n\) yields an \(n\)-dimensional model which is an approximation to (19.2), in a sense to be explored below. Since \(V_J = \bigoplus_{j<J} W_j\), it is equivalent to the \(2^J\) dimensional submodel of the sequence model given by

\begin{equation}(19.9)\quad y_I = \theta_I + \varepsilon_I, \quad I \in \mathcal{I}^J,\end{equation}

where we define \(\mathcal{I}^J = \cup_{j<J} \mathcal{I}_j\).

Estimation of the unknown coefficients \(\langle f, \phi_{j_i} \rangle\) is done in the wavelet basis. Recall that \(\phi_{j_i}\) is an orthobasis for \(V_J\) and that \(\{\psi_I, I \in \mathcal{I}^J\}\) is an orthobasis for the wavelet spaces \(\{W_j, j < J\}\). The orthogonal change of basis transformation \(W\) on \(\mathbb{R}^{2^J}\) that maps \(\langle f, \phi_{j_i} \rangle\) to \(\langle f, \psi_I \rangle = \theta_I\) is called the discrete wavelet transform \(W\). Its matrix elements \(W_{ij}\) are just the inner products \(\langle \psi_I, \phi_{j_i} \rangle\).

The estimation procedure could then be summarized by the diagram

\begin{equation}(19.10)\quad \begin{array}{c}
(y_i) \\
\downarrow
\end{array} \xrightarrow{W} \begin{array}{c}
(y_I)
\end{array} \quad \begin{array}{c}
(\hat{f}_{m,i})
\downarrow
\end{array} \xleftarrow{W^T} \begin{array}{c}
(\hat{\theta}_I(y_I))
\end{array}\end{equation}

There are thus three steps:

(i) transform into the wavelet coefficient domain,

(ii) apply a co-ordinatewise estimator \(\hat{\theta}_I\) (which might be a minimax rule, thresholding, linear shrinkage ...) to each coefficient \(y_I\),

(iii) transform back to the observation domain, which may be accomplished using \(W^T\) since \(W\) is orthogonal.

Consider now the minimax risk of estimation of \(f \in \mathcal{F}\) using data from the projected model (19.8). We look for a condition on the dimension \(m = 2^J\) so that the minimax risk in the projected model is asymptotically equivalent to (i.e. not easier than) the full model. Because of the Parseval relation (19.4), we may work in the sequence model/wavelet coefficient domain. Define the projected minimax risk by

\[ R_N(\Theta, \varepsilon; J) = \inf_{\hat{\theta}} \sup_{\Theta} E \|\hat{\theta}(y^{(m)}) - \Theta\|^2, \quad m = 2^J. \]

Note that we still attempt to estimate the entire sequence \(\Theta\).

Suppose, as would be natural in the projected model, that \(\hat{\theta}\) is an estimator which has non-zero co-ordinates only in \(\mathcal{I}^J\). Set \(\|\Theta\|_{2,m}^2 = \sum_{I \in \mathcal{I}^J} \theta_I^2\) and \(\|\Theta\|_{2,m}^2 = \sum_{I \in \mathcal{I}^J} \theta_I^2\). The following decomposition emphasises the “tail bias” term that results from estimating only up to level \(J:\)

\[ \|\hat{\theta} - \Theta\|^2 = \|\hat{\theta} - \hat{\Theta}\|_{2,m}^2 + \|\Theta\|_{2,m}^2. \]
Of course, in terms of the equivalent $f = f[\theta]$, and with $P_m$ denoting
the orthogonal projection of $L_2[0,1]$ onto $V_J$, the tail bias $\|\theta\|_{2,m^+}^2 = \|f - P_m f\|^2$.

A bound on the tail bias is implied by the smoothness of parameter
space $\Theta^\alpha_{p,q}$.

19.1. Lemma. Let $\alpha' = \alpha - (1/p - 1/2)_+ > 0$. Then for a constant
$K = K(\alpha')$,

$$\sup_{\theta \in \Theta^\alpha_{p,q}(C)} \|f - P_m f\|^2 = \sup_{\theta \in \Theta^\alpha_{p,q}(C)} \|\theta\|_{2,m^+}^2 \leq KC^{2-2\alpha'}.\quad (19.11)$$

Proof. This is a consequence of a basic inequality for $\ell_p$ norms in $\mathbb{R}^m$:

applied to $v = \theta_j, \in \mathbb{R}^2$ for $j \geq J$. Since $\Theta^\alpha_{p,q}(C) \subset \Theta^\alpha_{p,\infty}(C)$ and the latter
entails $\|\theta_j\|_p \leq 2^{-\alpha'j}C$, we obtain, in combination with (19.11) that

$$\|\theta_j\|_2^2 \leq 2^{(1-1/2)\alpha'} \|\theta_j\|_p^2 \leq 2^{-\alpha'}C.$$ 

Since $\|\theta\|_{2,m^+}^2 = \sum_{j > J} \|\theta_j\|_2^2$, summing over $j$ yields the lemma. \hfill $\Box$

If $J(\epsilon)$ is chosen to be a large enough multiple of $\log_2 \epsilon^{-2}$, then the tail
bias term becomes negligible relative to the order $\epsilon^{2\alpha/(2\alpha + 1)}$ of the minimax
risk. This is the key ingredient in

19.2. Proposition. Let $J(\epsilon) = \gamma \log_2 \epsilon^{-2}$. If $\gamma > (1/(2\alpha + 1))(\alpha/\alpha')$, then

$$R_N(\Theta, \epsilon, J(\epsilon)) \sim R_N(\Theta, \epsilon) \quad \epsilon \to 0.$$ 

Remark. The ratio $(1/(2\alpha + 1))(\alpha/\alpha')$ is certainly less than 1 whenever
(i) $p \geq 2$ and $\alpha > 0$, or (ii) $p < 2$ and $\alpha \geq 1/p$. In these cases, if $\epsilon = n^{-1/2}$,
then the contributions from levels above $\log_2 \epsilon$ are always negligible, and

$$R_N(\Theta, n^{-1/2}; \log_2 \epsilon) \sim R_N(\Theta, n^{-1/2}).$$

Thus, when $\epsilon = n^{-1/2}$, the projected model with $n$ data values is equivalent
to the full sequence model over all the indicated smoothness spaces.

Proposition 19.2 shows more: given $\Theta^\alpha_{p,q}$, it is possible to find $\eta > 0$ such that

$$\gamma = \frac{1}{2\alpha + 1} \frac{\alpha}{\alpha'} + \eta < 1,$$

and so the choices

$$m_n = 2^{4n} \quad J_n = \gamma \log_2 n = \gamma J_n$$

lead to a projected model sequence indexed by dyadic powers of $n$, using
less than $\log_2 \epsilon$ levels, but of full asymptotic difficulty. This will be used in
Section 19.5.
19.3. The Sampled Data Model: A Wavelet Crime?

The simplest non-parametric regression model posits an unknown function observed in homoscedastic Gaussian noise at equally spaced points $t_i = i/n$:

$$\hat{y}_i = n^{-1/2} f(t_i) + \epsilon \tilde{z}_i \quad i = 1, \ldots, n. \quad (19.14)$$

We assume that the $\tilde{z}_i$ are i.i.d standard Gaussian variables and that the noise level $\epsilon$ is known. For convenience, we suppose throughout that $n = 2^J$ for some integer $J$.

A version of the diagram (19.10) could be applied to the observed data $(\hat{y}_i)$ to produce an estimate $\hat{f}(t_i)$:

$$\begin{align*}
\begin{array}{c}
(\hat{y}_i) \\
\downarrow \\
(\hat{f}(t_i))
\end{array}
\xrightarrow{W} 
\begin{array}{c}
(\hat{y}_i) \\
\downarrow \\
(\hat{f}(t_i))
\end{array}
\xleftarrow{W^T} 
\begin{array}{c}
(\hat{y}_i) \\
\downarrow \\
(\hat{f}(t_i))
\end{array}
\end{align*} \quad (19.15)$$

Indeed, this general scheme is widely implemented in wavelet shrinkage and denoising packages. In addition to the structural advantages of wavelet bases, a key practical point is that the discrete wavelet transform $W$ and its transpose can be applied to data of size $n$ in $O(n)$ time, thanks to the cascade structure of the algorithm (see Section ???).

This leads to a possibly troubling dichotomy. Much of the theory developed to study wavelet methods is carried out using functions of a continuous variable, and uses the multiresolution analysis and smoothness classes of functions on $\mathbb{R}$ or $[0, 1]$. Almost inevitably, most actual data processing is carried out on discrete, sampled data, which in simple cases might be modeled by (19.14).

There is therefore a clear need to make a connection between the continuous and sampled models, and to show that, under appropriate conditions, that conclusions in one model are valid for the other and vice versa. For this purpose, we introduce a minimax risk for estimation of $f$ based on the sampled data $\hat{y}$ from (19.14). Set

$$R(\mathcal{F}, n) = \inf_{f(\hat{y}) \in \mathcal{F}} \sup_{\hat{f}(\hat{y})} E \| \hat{f}(\hat{y}) - f \|^2_2$$

Note that the error estimation is evaluated in the norm of $L_2[0, 1]$. One might also be interested in the error measured in the discrete norm $n^{-1} \| f \|^2_{2,n} = (1/n) \sum f^2(t_i)$. To distinguish the two, we will write $R(\mathcal{F}, n; L_2)$ for the former and $R(\mathcal{F}, n; \ell_{2,n})$ for the latter whenever necessary.

Use the scaling functions $\phi_{j+i}$ to form an interpolation of the sampled data $(\hat{y}_i)$:

$$\hat{y}(t) = \sum_{i=1}^{n} \hat{y}_i \phi_{j+i}(t).$$
The expectation of $\hat{y}(t)$ is given by
\[ P_n f(t) = \sum n^{-1/2} f(t_i) \phi_i(t). \]
We will need to compare $P_n f$ with $P_n f$, the projection of $L_2[0, 1]$ onto $V_j$, which has corresponding representation
\[ P_n f(t) = \sum \langle f, \phi_i \rangle \phi_i(t). \]
Since $E \hat{y} = P_n f$, it is natural to evaluate the error of an estimator $\hat{f}(\hat{y})$ based on sampled data using the decomposition
\[ \| \hat{f}(\hat{y}) - f \|_2 \leq \| \hat{f}(\hat{y}) - \hat{P}_n f \|_2 + \| \hat{P}_n f - f \|_2. \]
Now, if $f \in V_j$, so that $P_n f = f$, then from orthonormality of $\{ \phi_i \}$,
\begin{equation}
\| \hat{P}_n f - f \|_2^2 = (1/n) \sum_i [f(t_i) - \sqrt{n} \langle f, \phi_i \rangle]^2. \tag{19.16}
\end{equation}

Equality (19.16) shows that a fundamental role is played by the quality of approximation of the pointwise sampling operator $S_\delta f$ by the inner product sampling operator $S_\phi f$, where [Define these terms first!]
\[ (S_\delta f)_i = f(t_i), \quad (S_\phi f)_i = \sqrt{n} \langle f, \phi_i \rangle. \]

As reviewed in Section B.2.3, the quality of approximation is closely connected with the number of vanishing moments of the scaling function $\phi$. Lemma B.9 there shows that if $f$ is $C^\alpha$ and $\phi$ has at least $r = \lfloor \alpha \rfloor$ vanishing moments, then
\[ \| f(t_i) - 2^{-j/2} f(t_k) \| \leq M 2^{-j(\alpha+1/2)}, \]
with the constant $M$ depending on $\phi$ and the $C^\alpha$ norm of $f$.

To state a more general result we will require that $\phi$ also satisfy the regularity condition (19.5): thus $\phi$ has more than $\alpha$ vanishing moments and continuous derivatives. We note that these assumptions are met by the Coiflet family of compact support wavelets (REF) – indeed they were constructed to satisfy precisely these properties. We may now formulate a result in which the $L_2$ error of approximation of $f$ by the interpolant $\hat{P}_n f$ is negligible relative to the minimax estimation error.

19.3. **Proposition.** Suppose that $\alpha > 1/p, 1 \leq p, q \leq \infty$ and that $(\phi, \psi)$ satisfy (19.5). Then
\begin{equation}
\sup_{\mathcal{F}} \| \hat{P}_n f - f \|_2^2 \leq o(n^{-r}). \tag{19.17}
\end{equation}

**Remark.** Proposition 19.3 has a re-expression in terms of wavelet coefficients. Let $\hat{\theta}$ denote the wavelet coefficients of $\hat{P}_n f$: under the hypotheses of Proposition 19.3, we have
\begin{equation}
\sup_{\mathcal{F}} \| \hat{\theta} - \theta \|_2^2 = o(n^{-r}). \tag{19.18}
\end{equation}
Proof. We consider here only the case $F$ corresponding to $C^\alpha$, i.e. $B^\alpha_{\infty,\infty}$. Some indications of the argument in the general case are in the Appendix. In this case, (19.16) combined with Proposition 19.2 shows that
\[
\|\hat{P}_n f - P_n f\|^2 \leq M^2 n^{-2\alpha} = o(n^{-2\alpha/(2\alpha+1)}).
\]
In addition, Lemma 19.1 shows that $\|P_n f - f\|^2 = o(n^{-r})$. □

Remark: $1 \leq p, q$ really needed?

Remark: Comments on $\alpha = 1 = p = q$

Remark: On continuity of point evaluation

19.4. Sampling is not easier

It is perhaps intuitively clear that sampled data does not provide as much information as the continuous white noise model, but a formal argument is still necessary. Thus, in this section, we outline a proof of a lower bound to minimax risk in the sampling problem.

19.4. Theorem. Suppose that $\alpha > 1/p, 1 \leq p, q \leq \infty$ and that $(\phi, \psi)$ satisfy (19.5). Let $\epsilon_n = n^{-1/2}$.

(19.19)
\[
\hat{R}(F, n) \geq R(F, \epsilon_n)(1 + o(1)).
\]

The approach is to show that a prior distribution that is difficult in the continuous model sequence problem induces a difficult prior distribution in the sampled data setting. Proposition 19.2 shows that the continuous problem, in sequence space form, can be projected to a level $J_{0n} = \gamma \log_2 \epsilon_n^{-2}$ given by (19.13) without loss of difficulty. It is convenient to first show that
\[
\hat{R}(F, n; \ell_{2,n}) \geq R(\Theta, \epsilon_n; J_{0n}),
\]
where the discrete $\ell_{2,n}$ loss is used for the sampled data problem. Indeed, consider two multivariate Gaussian estimation problems:

(I) \quad $y_i = \theta_I + \epsilon z_i, \quad I = 1, \ldots, m, \quad L(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_{2,m}^2.$

(II) \quad $\hat{y}_i = \hat{\xi}_i + \epsilon \hat{\epsilon}_i, \quad i = 1, \ldots, n, \quad L(\hat{\xi}, \hat{\xi}) = \|\hat{\xi} - \xi\|_{2,n}^2.$

We suppose that the first problem is of smaller dimension $m < n$. Let $\pi$ be a prior distribution on $\mathbb{R}^m$, with support contained in a set $\Theta$, and having associated Bayes risk $B(\pi) = \mathbb{E}_\pi E_\theta \|\hat{\theta}_\pi - \theta\|_{2,m}^2$. Let $T : \mathbb{R}^m \to \mathbb{R}^n$ be a linear mapping. If $T$ were a partial isometry (i.e. $T^*T = Id$ on $\mathbb{R}^m$) then the induced distribution $\hat{\pi} = \pi \circ T^{-1}$ on $\mathbb{R}^n$ would have identical Bayes risk in problem (II): $\hat{B}(\hat{\pi}) = \mathbb{E}_\hat{\pi} E_\hat{\xi} \|\hat{\xi} - \hat{\xi}\|_{2,n}^2 = B(\pi)$. If $T$ is close to a partial isometry $U : \mathbb{R}^m \to \mathbb{R}^n$, then there is a lower bound for $\hat{B}(\hat{\pi})$ that is close to $B(\pi)$.

19.5. Proposition. With the definitions above, set

(19.20) \quad $\delta := \sup_{\Theta} \|T - U\|_{2,n}, \quad \lambda = \|T^*T\|_2.$

254
Then for \( \hat{\pi} = \pi \circ T^{-1} \),
\[
\sqrt{B(\hat{\pi})} \geq \frac{\sqrt{B(\pi)} - \delta}{1 \lor \lambda}.
\]

**Proof to follow later.**

In our application, problem (I) is of course the projected sequence model and problem (II) is the sampled data model with \( \xi_i = f(t_i)/\sqrt{n} \). The partial isometry \( U \) is just the inverse discrete wavelet transform restricted to indices in \( I^d \), while \( T \) samples the function \( f[\theta] \) associated with wavelet coefficients \( \theta \). Thus
\[
T\theta = n^{-1/2} S_\delta f[\theta], \quad \text{i.e.} \quad (T\theta)_i = (n^{-1/2} \sum_I \theta_I \psi_I(t_i))
\]
\[
U\theta = n^{-1/2} S_\phi f[\theta], \quad \text{i.e.} \quad (U\theta)_i = (\sum_I \theta_I \langle \phi_I, \psi_I \rangle)
\]

Thus
\[
\|T - I\|_{2,n}^2 = \|\theta - \theta\|_{2,m}^2
\]
and from (19.18), we conclude that
\[
\delta = \delta_n = o(n^{-r}).
\]

*Remark.* The condition (19.17) in fact shows that the quality of estimation in continuous and discrete norms is in fact equivalent:
\[
\hat{R}(F, n; L_2) \sim \hat{R}(F, n; \ell_{2,n}).
\]
(and similarly for \( R \)). Indeed it suffices to give a recipe for going from a good estimator for one loss function to an essentially equally good estimator for the other loss function. Before doing so, we recall that if \( f, g \in L_2[0,1] \) then \( P_n f \) has coefficients \( n^{-1/2} S_\delta f \) in the basis \( \phi_I \), while \( P_n f \) has coefficients \( n^{-1/2} S_\phi f \). This leads to Parseval identities like
\[
\langle P_n f, \tilde{P}_n g \rangle_{2,n} = n^{-1} \langle S_\delta f, S_\phi g \rangle_{2,n}
\]
with obvious analogs for \( \|P_n f\|_{2}^2 \) and \( \|\tilde{P}_n g\|_{2,n}^2 \).

First suppose that \( \tilde{f} = (f(t_i)) \) is a good estimator for \( \ell_{2,n} \) loss. Construct the interpolation \( \hat{f}(t) = (U\tilde{f})(t) = n^{-1/2} \sum_i \tilde{f}(t_i) \phi_I(t) \). From the decomposition
\[
\hat{f} - f = \hat{f} - \tilde{P}_n f + \tilde{P}_n f - f
\]
and the identity \( \|\hat{f} - \tilde{P}_n f\|_{2,n} = n^{-1/2} \|\hat{f} - S_\delta f\|_{2,n} \), we obtain from Proposition 19.3
\[
\|\hat{f} - f\|_2 \leq n^{-1/2}\|\hat{f} - \tilde{f}\|_{2,n} + o(n^{-r/2})
\]
so that \( \hat{f} \) has essentially as good performance for \( L_2 \) loss as does \( \tilde{f} \) for loss \( \ell_{2,n} \).

Now suppose on the other hand that \( \tilde{f}(t) \) is a good estimator for \( L_2 \) loss. Construct a discrete estimator using scaling function coefficients \( \tilde{f}(t_i) = \)
\((S_n f)\). From the identity \(n^{-1/2} \| \hat{f} - S_n f \|_2, n = \| P_n \hat{f} - \hat{P}_n f \|_2\) and the decomposition

\[
P_n \hat{f} - \hat{P}_n f = P_n (\hat{f} - f) + P_n f - f + f - \hat{P}_n f
\]

we obtain from Lemma 19.1, Proposition 19.3 and the fact that \(P_n\) is a projection that

\[
n^{-1/2} \| \hat{f} - S_n f \|_2 \leq \| \hat{f} - f \|_2 + o(n^{-r/2}).
\]

comment on need for vanishing moments on \(\phi\)

19.5. Sampling is not harder

In this section, our goal is to show that, at least when using scaling functions and wavelets with adequate smoothness and vanishing moments, the standard algorithmic practice of using the cascade algorithm on discrete data (diagram (19.15)) does not significantly inflate minimax risk relative to its use on genuine wavelet coefficients (diagram (19.10)). Specifically, we prove

19.6. THEOREM. Suppose that \(\alpha > 1/p, 1 \leq p, q \leq \infty\) and that \((\phi, \psi)\) satisfy (19.5). If \(E\) is any one of the four coordinatewise estimator classes of Table ??, then

\[
R_E(\mathcal{F}(C), n) \leq R_E(\mathcal{F}(C), \epsilon_n)(1 + o(1)) \quad \text{as } n \to \infty.
\]

To emphasize the dependence on the parameters \(C\) and \(\epsilon_n\), we will write \(R_E(C, \epsilon)\) for \(R_E(\mathcal{F}(C), \epsilon) = R_E(\Theta(C), \epsilon)\).

Outline of proof. We begin with the discrete wavelet transform by \(W\) of the sampling model (19.14) and obtain

\[
\hat{y}_I = \hat{\theta}_I + \hat{e}_I, \quad I \in \mathcal{I}^J.
\]

Since \(W\) is orthogonal, the transfoemed noise \(\hat{e}_I\) is still i.i.d. \(N(0,1)\). As noted earlier (below (19.9)), the sampled signal has wavelet coefficients

\[
\hat{\theta}_I = \sum_i \langle \psi_I, \phi_{Ij} \rangle f(t_i)/\sqrt{n} = \langle \psi_I, \hat{P}_n f \rangle.
\]

CONNECT with ch:primer!] In other words, \(\hat{\theta} = \hat{\theta}[\hat{P}_n f]\), and so we have the identity

\[
\| \hat{\theta} - \theta \|_2 = \| \hat{P}_n f - f \|_2.
\]

Our strategy will be to take an good estimator \(\hat{\theta}(\hat{y})\) and apply it with the sample data wavelet coefficients \(\hat{y} = (\hat{y}_I)\) used in place of \(y\). We will restrict attention to estimators vanishing for levels \(j \geq J_0\), where \(J_0\) is defined as at (19.13) and represents a projected version of the sequence model which is asymptotically as difficult as the full model. We also write \(m = m_n = 2^{J_0}\).

In view of (19.22), it is natural to decompose the error of estimation of \(\theta\) in terms of \(\hat{\theta}\):

\[
\| \hat{\theta}(\hat{y}) - \theta \|_2 \leq \| \hat{\theta}(\hat{y}) - \hat{\theta} \|_{2,m} + \| \hat{\theta} - \theta \|_{2,m} + \| \theta \|_{2,m^\perp}.
\]

256
Over smoothness classes $\Theta^\alpha_{p,q}$, the second term on the right side is controlled by (19.23) and (19.17). The third term is controlled by Lemma 19.1. Turning to the first term, the key remaining issue is to show that if $\theta$ has bounded Besov norm, then the Besov norm of the interpolant coefficients $\tilde{\theta}$ below level $J_{on}$ is not much larger. To emphasize this, we write $P_m \tilde{\theta}$ for the vector whose $(j,k)$-th coefficient is $\tilde{\theta}_{j,k}$ if $j < J_{on}$ and 0 otherwise.

19.7. Theorem. Suppose that $\alpha > 1/p, 1 \leq p, q \leq \infty$ and that $(\phi, \psi)$ satisfy (19.5). There exist constants $\Delta_n = \Delta_n(\phi, \psi, \alpha, p, q)$ such that

$$\|P_m \tilde{\theta}\|_{L^p_{\alpha,q}} \leq (1 + \Delta_n) \|\theta\|_{L^p_{\alpha,q}}.$$  

(19.25)

Some remarks on the proof are in the Appendix.

Hence, if we set $C_n = (1 + \Delta_n)C$, then $\theta \in \Theta(C)$ implies that $P_m \tilde{\theta} \in \Theta(C_n)$. Suppose now that $\hat{\theta}^*$ is asymptotically $\mathcal{E}$-minimax over $\Theta(C_n)$ - note that we have chosen $J_{on}$ explicitly so that this can be achieved with an estimator that vanishes for $j \geq J_{on}$. Thus, since we only attempt to estimate the first $m$ components of $\theta$,

$$\sup_{\theta \in \Theta(C)} E \{\|\hat{\theta}^*(\tilde{y}) - \tilde{\theta}\|^2_{2,m} \leq 0 \} \leq \sup_{\theta \in \Theta(C_n)} E \{\|\hat{\theta}^*(\tilde{y}) - \theta\|^2_{2,m} \}

\leq R\mathcal{E}(C_n, \epsilon_n)(1 + o(1)).$$

19.8. Lemma. If $C_1 \geq C_0$, then for any of the four estimator classes $\mathcal{E}$

$$R\mathcal{E}(C_1, \epsilon) \leq (C_1/C_0)^2 R\mathcal{E}(C_0, \epsilon).$$

(19.26)

For the proof, see Donoho & Johnstone (1999). If $\hat{f}^*(\tilde{y})$ is the estimator that corresponds to $\hat{\theta}^*$, then by combining (19.24) and (19.26), we obtain

$$\sup_{\hat{f} \in \mathcal{F}(C)} E \{\|\hat{f}^*(\tilde{y}) - f\|^2 = \sup_{\theta \in \Theta(C)} E \|\hat{\theta}^*(\tilde{y}) - \theta\|^2 

\leq (C_n/C)^2 R\mathcal{E}(C, \epsilon_n)(1 + o(1))

= R\mathcal{E}(C, \epsilon_n)(1 + o(1)).$$

Formulate statements for specific estimators?

19.6. Further Details

Proof of Proposition 19.2. Since the class of estimators allowed in the projected problem is a subset of those in the full problem, $R_N(\Theta, \epsilon, J) \geq R_N(\Theta, \epsilon)$, and it suffices to show the reverse inequality. Suppose that $\hat{\theta}_e$ is the Bayes estimator for the least favorable prior $\pi = \pi_e$ for $R_N(\Theta, \epsilon)$. Define $\pi_m = \pi_{m,e}$ as the marginal distribution of the first $m$ coordinates of $\theta$ under $\pi_e$; clearly the corresponding Bayes estimate $\hat{\theta}_{m} = E_{\pi_m}(\theta|y)$ depends only on $y^{(m)}$ and so is feasible in the projected problem. Since both $\pi$ and $\pi_m$ are supported on $\Theta$, $\hat{\theta}_e - \hat{\theta}_{m} = E_{\pi}(\theta - P_m \theta|y)$ and so by Jensen’s inequality and Lemma 5.1???

$$\|\hat{\theta}_e - \hat{\theta}_{m}\|_2 \leq E(\|\theta - P_m \theta\|_2|y) \leq K C^2 2^{-J_{on}} = o(m^{-r})$$

257
by the choice of $J(\epsilon)$ specified in the hypotheses. Thus, for all $\theta \in \Theta$,
\[
E\|\hat{\theta}_{\pi_n} - \theta\|^2 \leq \left(\{E_\theta \|\hat{\theta}_\pi - \theta\|^2\}^{1/2} + o(m^{-r/2})\right)^2 \\
= R_N(\Theta, \epsilon)(1 + o(1))
\]
since $\hat{\theta}_\pi$ is minimax for $\Theta$. \hfill \Box

**Remarks on proof of Proposition 19.3.** Decompose $f = \sum e_j$, where $e_j$ is the projection of $f$ on the detail space $W_j$.

1°. For $j \leq J$, consider the restriction of $\hat{P}_n$ to the approximation spaces $V_j$ and obtain $\| (I - \hat{P}_n)|_{V_j} \|_2 \leq c2^{-R(J-j)}$. Of course, the regularity and vanishing moments properties are crucial here.

2°. For $j \geq J$, $\sup_{\theta} \|\hat{P}_n e_j\|_2 \leq \alpha Cn^{-1/2}\beta^{-j(\alpha^{-1}/p)}$. This part does not use vanishing moments - the key property is to use the compact support of $\psi_{jk}$ at high levels $j$.

3°. Combine steps 1° and 2°, using the constraints on $e_j$ implied by $\theta \in \Theta$. \hfill \Box

**Remarks on proof of Theorem 19.7.** Write $P_n.\hat{\theta} = T^n.\theta$ where $T^n : \ell_2(\mathcal{I}) \rightarrow \ell_2(\mathcal{I}^0)$ has matrix elements $T_{ij} = \langle \psi_i, \hat{P}_n \psi_j \rangle$. Define corresponding operators $\Delta^n : \ell_2(\mathcal{I}) \rightarrow \ell_2(\mathcal{I}^0)$ by $U_{ij} = \delta_{ij}$ and $\Delta^n = T^n - U^n$. We emphasize that $j_0 = \log_2 m < j_1 = \log_2 n$. Theorem 19.7 follows by showing
\[
\|\Delta^n\|_{p, q} \leq \Delta_n(\alpha, p, q) \rightarrow 0.
\]
Define $\hat{\Delta}_n^I = 2^{2(i-j)}(T^n - U^n)|_{IJ}$ for $I \in \mathcal{I}^0$, $J \in \mathcal{I}$. By results on interpolation of operators, one obtains
\[
\|\Delta^n\|_{p, q} \leq C(\log n) \max(\\|\Delta^n\|_{1, \infty}, \\|\Delta^n\|_{\infty, \infty})
\]
where $\| \cdot \|_{p, \infty}$ are operator norms on $\ell_\infty(\ell_p)$, which for $p = 1, \infty$ can be expressed in terms of various row and column sums of $\Delta^n$. A (somewhat tedious) development and combination of bounds for sums of the form
\[
\sum_{J \in \mathcal{I}^0} |\Delta^n_{IJ}|, \quad J \in \mathcal{I}, \quad \text{and} \quad \sum_{I \in \mathcal{I}^0} |\Delta^n_{IJ}|, \quad I \in \mathcal{I}^0
\]
proceeds using the ideas from steps 1° and 2° of the proof of Proposition 19.3 and completes the proof. \hfill \Box

258
APPENDIX A

Appendix: The Minimax Theorem

The aim of this appendix is to give some justification for the minimax Theorem 6.6, restated below as Theorem A.5. Such statistical minimax theorems are a staple of statistical decision theory as initiated by Abraham Wald, who built upon the foundation of the two person zero-sum game theory of von Neumann & Morgenstern (1944). It is, however, difficult to find in the published literature a statement of a statistical minimax theorem which is readily seen to cover the situation of our nonparametric result Theorem A.5. In addition, published versions (e.g. Le Cam (1986, Theorem ??)) often do not pause to indicate the connections with game theoretic origins.

This appendix gives a brief account of von Neumann’s theorem and one of its infinite-dimensional extensions (Kneser 1952) which aptly indicates what compactness and continuity conditions are needed. Following Brown (1978), we then attempt an account of how statistical minimax theorems are derived, orienting the discussion towards the Gaussian sequence model. While the story does not in fact use much of the special structure of the sequence model, the Gaussian assumption is used at one point to assure the separability of $L_1$.

In later sections, a number of concepts and results from point set topology and functional analysis are needed, which for reasons of space we do not fully recall here. They may of course be found in standard texts such as Dugundji (1966) and Rudin (1973).

A.1. Finite two person zero sum games.

A finite two person, zero sum game can be described by an $m \times n$ payoff matrix $A = \{A(i, j)\}$, with the interpretation that if player I uses strategy $i \in \{1, \ldots, m\}$ and player II chooses strategy $j \in \{1, \ldots, n\}$, then player II receives a payoff $A(i, j)$ from player I.

If player I declares his strategy, $i$ say, first, then naturally player II will choose the maximum payoff available in that row, namely $\max_j A(i, j)$. Expecting this, player I will therefore choose $i$ to achieve $\min_i \max_j A(i, j)$. On the other hand, if player II declares his strategy $j$ first, player I will certainly pay only $\min_i A(i, j)$, so that II will receive at most $\max_j \min_i A(i, j)$. Intuitively, II is better off if I has to declare first: indeed one may easily
verify that
\[
\max_i \min_j A(i, j) \leq \min_i \max_j A(i, j).
\]

When equality holds in (A.1), the game is said to have a value. This
occurs, for example, if the game has a saddlepoint \((i_0, j_0)\), defined by the
property
\[
A(i_0, j) \leq A(i_0, j_0) \leq A(i, j_0) \quad \text{for all } i, j.
\]
However, saddlepoints do not exist in general, as is demonstrated already by
the matrix \(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\). The situation is rescued by allowing mixed or randomized
strategies, which are probability distributions \(x = (x(i))_n^i\) and \(y = (y(j))_n^j\)
on the space of nonrandomized rules for each player. If the players use the
mixed strategies \(x\) and \(y\), then the expected payoff from \(I\) to \(II\) is given by
\[
f(x, y) = x^T Ay = \sum_{i,j} x(i)A(i, j)y(j).
\]
Write \(S_m\) for the simplex of probability vectors \(\{x \in \mathbb{R}^n : x_i \geq 0, \sum x_i = 1\}\).
The classical minimax theorem of von Neumann states that for an arbitrary
\(m \times n\) matrix \(A\) in (A.2),
\[
\min_{x \in S_m} \max_{y \in S_n} f(x, y) = \max_{y \in S_n} \min_{x \in S_m} f(x, y).
\]
We establish below a more general result that implies (A.3).

### A.2. Bilinear semicontinuous payoffs

In (A.2) - (A.3), we observe that \(f\) is a bilinear function defined on compact, convex sets in Euclidean space. There have been numerous generalizations of this result, either relaxing bilinearity in the direction of convexity-
convexity type assumptions on \(f\), or in allowing more general convex spaces of
strategies, or in relaxing the continuity assumptions on \(f\). Frequently cited
papers include those of Fan (1953) and Sion (1958), and a recent survey is

We give here a result for bilinear functions on general convex sets due to
Kneser (1952) that has a particularly elegant and simple proof. In addition,
Kuhn (1953), Peck & Duhmace (1957) observed that the method extends
directly to convex-concave \(f\). First recall that a function \(f : X \to \mathbb{R}\) on a
topological space \(X\) is lower semicontinuous (lsc) iff \(\{x : f(x) > t\}\) is open
for all \(t\), or equivalently if \(\{x : f(x) \leq t\}\) is closed for all \(t\). [If \(X\) is lsc
countable, then these conditions may be rewritten in terms of sequences as
\(f(x) \leq \limf(x_n)\) whenever \(x_n \to x\).] If \(X\) is also compact, then an lsc
function \(f\) attains its infimum: \(\inf_{x \in X} f = f(x_0)\) for some \(x_0 \in X\).

A.1. THEOREM (Kneser, Kuhn). Let \(K, L\) be convex subsets of real vector
spaces and \(f : K \times L \to \mathbb{R}\) be convex in \(x\) for each \(y \in L\), and concave in \(y\)
for each \( x \in K \). Suppose also that \( K \) is compact and that \( x \to f(x,y) \) is lsc for all \( y \in L \). Then

\[
\inf_{x \in K} \sup_{y \in L} f(x,y) = \sup_{y \in L} \inf_{x \in K} f(x,y).
\]

(A.4)

A notable aspect of this extension of the von Neumann theorem is that there are no compactness conditions on \( L \), nor continuity conditions on \( y \to f(x,y) \): the topological conditions are confined to the \( x \)-slot.

Here is an example where \( f \) is not continuous, and only the semi-continuity condition of the theorem holds. Let \( \mathbb{R}^\infty \) denote the space of sequences: a countable product of \( \mathbb{R} \) with the product topology: \( x^{(n)} \to x \) if for each coordinate \( i \), \( x_i^{(n)} \to x_i \). Then the infinite simplex \( K = \{ x \in \mathbb{R}^\infty : x_i \geq 0, \sum_i x_i \leq 1 \} \) is compact. Consider a simple extension of the payoff function (A.2), \( f(x,y) = \sum x_i y_i \) for \( y \in L = \{ y : 0 \leq y_i \leq C \) for all \( i \). Then the function \( x \to f(x,1) \) is not continuous: the sequence \( x^{(n)} = (1/n, \ldots, 1/n, 0, 0, \ldots) \) converges to 0 but \( f(x^{(n)}, 1) \equiv 1 \). However, \( f(x,y) \) is lsc in \( x \), as is easily verified.

Kneser’s proof nicely brings out the role of compactness and semi-continuity, so we present it here through a couple of lemmas.

A.2. Lemma. Let \( f_1, \ldots, f_n \) be convex, lsc real functions on a compact convex set \( K \). Suppose for each \( x \in K \) that \( \max_i f_i(x) > 0 \). Then there exists a convex combination that is positive on \( K \): for some \( \sigma \in S_n \),

\[
\sum_{i=1}^{n} \sigma_i f_i(x) > 0 \quad \text{for all } x \in K.
\]

Remark. This lemma implies the standard separating hyperplane theorem in \( \mathbb{R}^m \): if \( K \) is compact, convex with \( 0 \notin K \), then there exists a hyperplane separating 0 from \( K \). Indeed, simply let \( n = 2m \) and \( f_i(x) = x_i \) and \( f_{m+i}(x) = -x_i \).

Proof. Once the case \( n = 2 \) is established (\( n = 1 \) is vacuous), an induction argument can be used. So, with a slight change of notation, assume for all \( x \) that \( \max \{ f(x), g(x) \} > 0 \). By lower semi-continuity, the sets \( M = \{ x : f(x) \leq 0 \} \) and \( N = \{ x : g(x) \leq 0 \} \) are closed, and hence compact. The positivity condition implies that \( M \) and \( N \) are disjoint, and we may assume they are both non-empty, since otherwise the conclusion is trivial. Again by the positivity, on \( M \), both \( g > 0 \) and the ratio \(-f/g\) is defined and use. Arguing similarly on \( N \), we obtain

\[
\max_{M} -\frac{f}{g} = \frac{-f}{g}(p) = \alpha \geq 0, \quad \max_{N} -\frac{g}{f} = \frac{-g}{f}(q) = \beta \geq 0.
\]

(A.5)

Since \( f(p) \leq 0 \) and \( f(q) > 0 \), there exists \( \eta > 0 \) such that \( \eta f(p) + \eta f(q) = 0 \) [we have set \( \eta = 1 - \eta \)] Thus, writing \( p_\eta = \eta p + \eta q \), convexity of \( f \) implies \( f(p_\eta) \leq 0 \). By convexity of \( g \) and the positivity condition, \( \eta g(p) + \eta g(q) \geq \eta g(p) \leq 0 \).
\( g(p_\eta) > 0 \). Combining these conclusions with (A.5),
\[
\eta g(p) > -\bar{\eta}g(q) = \bar{\eta} \beta f(q) = -\eta \beta f(p) = \eta \alpha \beta g(p),
\]
which implies that \( \alpha \beta < 1 \).

Thus, we may increase \( \alpha \) to \( \gamma \) and \( \beta \) to \( \delta \) in such a way that \( \gamma \delta = 1 \). Equalities (A.5) then become strict inequalities:
\[
\text{On } M, \quad \frac{f + \gamma g}{1 + \gamma} > 0, \quad \text{On } N, \quad \frac{\delta f + g}{1 + \delta} > 0.
\]
Since \( \gamma \delta = 1 \), define \( \sigma = 1/(1 + \gamma) = \delta/(1 + \delta) \). Thus on \( M \cup N \), we get \( \sigma f + \bar{\sigma} g > 0 \), and on the rest of \( K \) this holds trivially, so the proof is done. \( \square \)

A.3. Lemma. Either (I) for some \( x \), \( \sup_y f(x,y) \leq 0 \), or (II) for some \( y \), \( \min_x f(x,y) > 0 \).

Proof. If (I) is false, then for every \( x \), there exists some value of \( y \), which we call \( p(x) \), such that \( f(x, p(x)) > 0 \). Lower semicontinuity implies that each of the sets \( A_y = \{ x : f(x, y) > 0 \} \) are open, and we have just shown that \( x \in A_{p(x)} \). Hence \( K \) is covered by \( \{ A_{p(x)} \} \), so extract a finite subcover indexed by \( y_i = p(x_i) \) for some \( x_1, \ldots, x_n \). This means exactly that for each \( x \), \( \max_i f(x, y_i) > 0 \). The previous lemma then gives a probability vector \( \sigma \in S_n \), such that for each \( x \), \( \sum_i \sigma_i f(x, y_i) > 0 \). By concavity, at \( y^* = \sum \sigma_i y_i \), we have \( f(x, y^*) > 0 \) for each \( x \). Again using compactness and lsc, \( \min_{x \in K} f(x, y^*) > 0 \), which implies alternative II. \( \square \)

Proof of Theorem A.1. That the right side of (A.7) is less than or equal to the left side is elementary, just as in (A.1). Let us suppose, then, that the inequality is strict, so that for some \( c \),
\[
\sup_y \inf_x f \leq c < \inf_x \sup_y f.
\]
Replacing \( f \) by \( f - c \) does not harm any of the hypotheses, so we may assume that \( c = 0 \). The left inequality in (A.6) implies that Alternative II in the previous lemma fails, so Alternative I holds, and so \( \inf_x \sup_y f \leq 0 \), in contradiction with the right hand inequality of (A.6)! Hence there must be equality in (A.6). \( \square \)

The following corollary is a trivial restatement of Theorem A.1 for the case when compactness and semicontinuity is known for the variable which is being maximised.

A.4. Corollary. Let \( K, L \) be convex subsets of real vector spaces and \( f : K \times L \to \mathbb{R} \) be convex in \( x \) for each \( y \in L \), and concave in \( y \) for each \( x \in K \). Suppose also that \( L \) is compact and that \( y \to f(x, y) \) is upper semicontinuous for each \( x \in K \). Then
\[
\inf_{x \in K} \sup_{y \in L} f(x, y) = \sup_{y \in L} \inf_{x \in K} f(x, y).
\]

Proof. Apply Theorem A.1 to \( \bar{f}(y, x) = -f(x, y) \). \( \square \)
A.3. A statistical minimax theorem

First, we state the Gaussian sequence model in a little more detail. The sample space $\mathcal{X} = \mathbb{R}^\infty$, the space of sequences in the product topology of pointwise convergence, under which it is complete, separable and metrizable. It is endowed with the Borel $\sigma$-field, and as dominating measure, we take $P_0$, the centered Gaussian Radon measure (see Bogachev (1998, Example 2.3.5)) defined as the product of a countable number of copies of the standard $N(0,1)$ measure on $\mathbb{R}$. For each $\theta \in \Theta = \ell_2(\mathbb{N}, \lambda)$, the measure $P_\theta$ with mean $\theta$ is absolutely continuous (indeed equivalent) to $P_0$, and has density $f_\theta(x) = dP_\theta/dP_0 = \exp\{\langle \theta, x \rangle - \|\theta\|^2_\lambda/2\}$. Because $P_0$ is Gaussian, the space $L_2(\mathcal{X}, P_0)$ of square integrable functions is separable (Bogachev 1998, Corollary 3.2.8), and hence so also is $L_1 = L_1(\mathcal{X}, P_0)$.

Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ denote the two point compactification of $\mathbb{R}$. As action space we take the countable product $\mathcal{A} = (\mathbb{R})^\infty$ which with the product topology is compact, $\mathcal{C}$ countable and Hausdorff, and again equip it with the Borel $\sigma$-field.

We consider loss functions $L(a, \theta)$ that are non-negative, and perhaps extended-real valued: $L : \mathcal{A} \times \Theta \rightarrow [0, \infty]$. A.5. Theorem. For the above Gaussian sequence model, we assume (i) that for each $\theta$, the map $a \rightarrow L(a, \theta)$ is convex and lsc for the product topology on $\mathcal{A}$, and (ii) that $\mathcal{P}$ is a convex set of prior probability measures on $\ell_2(\mathbb{N}, \lambda)$. Then

(A.8) \[ \inf_{\tilde{\theta}} \sup_{\pi \in \mathcal{P}} B(\tilde{\theta}, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\tilde{\theta}} B(\tilde{\theta}, \pi). \]

Remark. Our applications of this theorem will typically be to loss functions of the form $L(a, \theta) = w(\|a - \theta\|_p)$, with $w(\cdot)$ a continuous, increasing function. It is easy to verify that such loss functions are lsc in $a$ in the topology of pointwise convergence. Indeed, if $a_i^{(n)} \rightarrow a_i^{(\infty)}$ for each $i$, then for each fixed $m$, one has

\[ \sum_{i=1}^m |a_i^{(\infty)} - \theta_i|^p = \lim_n \sum_{i=1}^m |a_i^{(n)} - \theta_i|^p \leq \liminf_n \|a^{(n)} - \theta\|^p. \]

Theorem A.5 and other statistical minimax theorems, while closely related to Theorem A.1, as will be seen below, do not seem to follow directly from it, using instead separating hyperplane results (compare Lemma A.2).

A general framework for statistical decision theory, including minimax and complete class results, has been developed by its chief exponents, including A. Wald, L. LeCam, C. Stein, and L. Brown, in published and unpublished works. A selection of references includes Wald (1950), LeCam (1955), Le Cam (1986), Diaconis & Stein (1983), Brown (1977, 1978).

The theory is general enough to handle abstract sample spaces and unbounded loss functions, but it is difficult to find a statement that immediately covers our Theorem A.5. We therefore give a summary description
of the steps in the argument for Theorem A.5, using freely the version of
the Wald-LeCam-Brown approach set out in Brown (1978). The theory of
Brown (1978) was developed specifically to handle both parametric and non-
parametric settings, but few nonparametric examples were then discussed
explicitly. Proofs of results given there will be omitted, but we hope that this
outline nevertheless has some pedagogic value in stepping through the gen-
eral method in the concrete setting of the nonparametric Gaussian sequence
model.

A.4. Randomized decision rules.

The payoff function $B(\hat{\theta}, \pi)$ appearing in Theorem A.5 is linear in $\pi$,
but not in $\hat{\theta}$. Just as in the two-person game case, the standard method in
statistical decision theory for obtaining linearity is to introduce randomized
decision rules. These are Markov kernels $\delta(da|x)$ with two properties: (i)
for each $x \in \mathcal{X}$, $\delta(\cdot|x)$ is a probability measure on $\mathcal{A}$ which describes
the distribution of the random action $\alpha$ given that $x$ is observed, and (ii), for
each measurable $A$, the map $x \rightarrow \delta(A|x)$ is measurable. The risk function
of a randomized rule $\delta$ is

$$r(\delta, \theta) = \int \int L(a, \theta) \delta(da|x)P_\theta(dx),$$

and the payoff function we consider is the integrated risk against a proba-
bility measure $\pi$:

$$B(\delta, \pi) = \int r(\delta; \theta)\pi(d\theta).$$

A major reason for introducing $B(\delta, \pi)$ is that it is bilinear in $\delta$ and $\pi$.
Further, writing $\mathcal{D}$ for the class of all randomized decision rules, we note
that both it and $\mathcal{P}$ are convex. To establish a minimax statement

$$\inf_{\delta \in \mathcal{D}} \sup_{\pi \in \mathcal{P}} B(\delta, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\delta \in \mathcal{D}} B(\delta, \pi),$$

Kneser’s theorem suggests that we need a topology on decision rules $\delta$ with
two key properties:

1. (P1): $\mathcal{D}$ is compact, and
2. (P2): the risk functions $\delta \rightarrow B(\delta, \pi)$ are lower semicontinuous.

Before describing how this is done, we explain how (A.8) follows from
(A.10) using the convexity assumption on the loss function. Indeed, given
a randomized rule $\delta$, the standard method is to construct a non-randomized
rule by averaging: $\hat{\theta}_\delta(x) = \int a\delta(da|x)$. Convexity of $\alpha \rightarrow L(a, \theta)$ and
Jensen’s inequality then imply that

$$L(\hat{\theta}_\delta(x), \theta) \leq \int L(a, \theta) \delta(da|x).$$

Averaging over $X \sim P_\theta$, and recalling (A.9) shows that $\hat{\theta}_\delta$ is at least as good
as $\delta$:

$$r(\hat{\theta}_\delta, \theta) \leq r(\delta, \theta) \quad \text{for all } \theta \in \Theta.$$
Consequently, with convex loss functions, there is no reason ever to use a randomized decision rule, since there is always a better non-randomized one. In particular, integrating with respect to an arbitrary $\pi$ yields
\[ \sup_{\pi} B(\hat{\theta}, \pi) \leq \sup_{\pi} B(\delta, \pi). \]

We then recover (A.8) from (A.10) via a simple chain of inequalities:
\[
\begin{align*}
\inf_{\pi} \sup_{\delta} B(\hat{\theta}, \pi) &\leq \inf_{\pi} \sup_{\delta} B(\hat{\theta}_\delta, \pi) \leq \inf_{\pi} \sup_{\delta} B(\delta, \pi) \\
&= \sup_{\pi} \inf_{\delta} B(\delta, \pi) \leq \sup_{\pi} \inf_{\delta} B(\hat{\theta}, \pi),
\end{align*}
\]
and since the first and last terms are the same, all terms are equal.

### A.5. A compact topology for $\mathcal{D}$

We return to establishing properties [P1] and [P2]. The approach is to identify decision rules $\delta$ with bilinear, bicontinuous functionals, and then use the Alaoglu theorem on weak compactness to induce a topology on $\mathcal{D}$.

For this section, we write $L_0(a)$ for the loss function to emphasise the dependence on $a$. The risk function of a rule $\delta$ may then be written
\[
r(\delta, \theta) = \int \int L_0(a) f_0(x) \delta(da|x) P_0(dx) = b_\delta(f_0, L_0),
\]
Here the probability density $f_0$ is regarded as a non-negative function in the Banach space $L_1 = L_1(\mathbb{R}^\infty, B(\mathbb{R}^\infty), P_0)$ which is separable as noted earlier.

Since $\mathcal{A} = (\mathbb{R})^\infty$ is compact, metrizable and second countable, the Banach space $C = C(\mathcal{A})$ of continuous functions on $\mathcal{A}$, equipped with the uniform norm, is also separable. The functional
\[
b_\delta(g, c) = \int \int g(x) c(a) \delta(da|x) P_0(dx)
\]
belongs to the Banach space $B$ of bilinear, bicontinuous functionals on $L_1 \times C$ with the operator norm $\|b\|_B = \sup\{|b(g, c)| : \|g\|_{L_1} = \|c\|_C = 1\}$. Under assumptions satisfied here, Brown (1978) shows that the mapping $\iota : \delta \rightarrow b_\delta$ is a bijection of $\mathcal{D}$ onto
\[
B_1^+ = \{b \in B : b \geq 0 \text{ and } b(g, 1) = \|g\|_{L_1} \forall g \geq 0\}
\subset \{b : \|b\|_B \leq 1\},
\]
and the latter set, by Alaoglu’s theorem, is compact in the weak topology, which by separability of $L_1$ and $C$ is also metrizable on such norm bounded sets. Thus, $B_1^+$, being a closed subset, is also compact. The map $\iota$ is then used to induce a compact metrizable topology on $\mathcal{D} = \iota^{-1}(B_1^+)$ in which convergence may be described by sequences: thus $\delta_k \rightarrow \delta$ means that
\[
(A.11) \quad b_{\delta_k}(g, c) \rightarrow b_\delta(g, c) \quad \forall (g, c) \in L_1 \times C.
\]
This topology also satisfies our second requirement: that the maps $\delta \rightarrow B(\delta, \pi)$ be lsc. Indeed, since $\mathcal{A}$ is second countable, the lsc loss functions can
be approximated by an increasing sequence of continuous functions \( c_i \in C \): 
\[ L_\theta(a) = \lim_i c_i(a). \]
This implies that 
\[ r(\delta, \theta) = \sup_c b_\theta(f_\theta, c) : c \leq L_\theta \].

The definition (A.11) says that the maps \( \delta \to b_\theta(f_\theta, c) \) are each continuous, and so \( \delta \to r(\delta, \theta) \) appears as the upper envelope of a family of continuous functions, and is hence lsc. Finally Fatou’s lemma implies that 
\[ \delta \to B(\delta, \pi) = \int r(\delta, \theta) \pi \, (d\theta) \] is lsc.

### A.6. A separation theorem

We have now established \( B(\delta, \pi) \) as a bilinear function on \( \mathcal{D} \times \mathcal{P} \) which for each \( \pi \) fixed is lsc on the compact \( \mathcal{D} \). What prevents us from applying Kneser’s minimax theorem directly is that \( B(\delta, \pi) \) can be infinite. The strategy used by Brown (1978) for handling this difficulty is to prove a separation theorem for extended real valued functions, and derive from this the minimax result.

Slightly modified for our context, this approach works as follows. Let 
\[ T = T(\mathcal{P}, [0, \infty]) \]
denote the collection of all functions \( b : \mathcal{P} \to [0, \infty] \) - with the product topology, this space is compact by Tychoff’s theorem. Now define an upper envelope of the risk functions by setting \( \Gamma = \rho(\mathcal{D}) \) and then defining
\[ \bar{\Gamma} = \{ b \in T : \text{there exists } b' \in \Gamma \text{ with } b' \leq b \}. \]
Brown uses the \( \mathcal{D} \) topology constructed above, along with the compactness and lower semicontinuity properties [P1] and [P2] to show that \( \bar{\Gamma} \) is closed and hence compact in \( T \).

Using the separating hyperplane theorem for Euclidean spaces - a consequence of Lemma A.2 - Brown shows

**A.6. THEOREM.** Suppose that \( \bar{\Gamma} \) is convex and closed in \( T \) and that \( b_0 \in T \backslash \bar{\Gamma} \). Then there exists \( c > 0 \), a finite set \( \{ \pi_i \}_m \subset \mathcal{P} \) and a probability vector \( (\xi_i)_m^{\pi} \) such that the convex combination \( \pi_\xi = \sum \xi_i \pi_i \in \mathcal{P} \) satisfies
\[ b_0(\pi_\xi) < c < l(\pi_\xi) \quad \text{for all } b \in \bar{\Gamma}. \]

It is now easy to derive the minimax conclusion (A.10). Indeed, write 
\[ V = \inf_\pi \sup_b B(\delta, \pi). \] If \( V < \infty \), let \( \epsilon > 0 \) and choose \( b_0 \equiv V - \epsilon \) - clearly \( b_0 \notin \bar{\Gamma} \). Convexity of \( \mathcal{D} \) entails convexity of \( \bar{\Gamma} \), which is also closed in \( T \) as we saw earlier. Hence, the separation theorem produces \( \pi_\xi \in \mathcal{P} \) such that 
\[ V - \epsilon = b_0(\pi_\xi) < \inf_\delta B(\delta, \pi_\xi). \]
In other words, \( \sup_\delta \inf_\pi B(\delta, \pi) > V - \epsilon \) for each \( \epsilon > 0 \), and hence it must equal \( V \). If \( V = \infty \), a similar argument using \( b_0 \equiv m \) for each finite \( m \) also yields (A.10).

**Exercise.** Complete the induction step for the proof of Lemma A.2.
APPENDIX B

Background material on Wavelets and Function Spaces

B.1. Wavelets.

This is not an ab initio exposition of wavelet ideas and theorems. There are now many authoritative books to which we encourage readers to turn: a partial list includes Meyer (1990), Daubechies (1992), Mallat (1998b), Hernández & Weiss (1996), Härdle et al. (1998), Cohen & Ryan (1995) and Chui (1992). Rather we present, without proofs, a summary of definitions, concepts and results relevant to our statistical theory and algorithms. In this way, we also establish the particular notation that we use, since there are significantly different conventions in the literature.

B.2. Multiresolution Analysis

Many expositions rightly begin with the continuous wavelet transform, and then discuss frames in detail before specialising to orthogonal wavelet bases. However, as the statistical theory mostly uses orthonormal bases, we jump directly to the definition of multiresolution analysis due to Mallat and Meyer [GIVE REFS] here in a unidimensional form given by Hernández & Weiss (1996):

Definition. A multiresolution analysis (MRA) of $L_2(\mathbb{R})$ is given by a sequence of closed subspaces \( \{ V_j, j \in \mathbb{Z} \} \) satisfying the following conditions:

(i) \( V_j \subseteq V_{j+1} \),
(ii) \( f(x) \in V_j \) if and only if \( f(2^j x) \in V_j \), \( \forall j \in \mathbb{Z} \),
(iii) \( \bigcap_{j \in \mathbb{Z}} V_j = \{ 0 \} \), \( \bigcup_{j \in \mathbb{Z}} V_j = L_2(\mathbb{R}) \).
(iv) \( \exists \varphi \in V_0 \) such that \( \{ \varphi(x - k) : k \in \mathbb{Z} \} \) is an orthonormal basis (o.n.b) for \( V_0 \).

The function \( \varphi \) in (4) is called the scaling function of the given MRA. Set \( \varphi_{jk} = 2^{j/2} \varphi(2^j x - k) \), and let \( t_{jk} = 2^{-j} k \). If \( \varphi \) is supported on \([-S, S]\) and bounded by \( B \), then \( \varphi_{jk} \) is supported on \( t_{jk} + 2^{-j} [-S, S] \) and is bounded by \( 2^j B \). Property (2) implies that \( \{ \varphi_{jk}, k \in \mathbb{Z} \} \) is an orthonormal basis for \( V_j \). The orthogonal projection from \( L_2(\mathbb{R}) \rightarrow V_j \) is then

\[
P_j f = \sum_k \langle f, \varphi_{jk} \rangle \varphi_{jk}.
\]
The spaces $V_j$ form an increasing sequence of approximations to $L_2(\mathbb{R})$: indeed property (3) implies that $P_j f \to f$ in $L_2(\mathbb{R})$ as $j \to \infty$.

Example. Set $I_{j,k} = [2^{-j} k, 2^{-j} (k + 1)]$. The “Haar multiresolution analysis” is defined by

$$V_j = \{ f \in L_2(\mathbb{R}) : f|_{I_{j,k}} = c_{j,k} \}, \quad \varphi = I_{[0,1]}.$$

Remarks. 1. Warning: many authors use the opposite convention $V_{j+1} \subset V_j!$

2. A family $\{c_k\}_{k \in \mathbb{N}}$ is a Riesz basis for a Hilbert space $H$ if (i) for all $h \in H$, there is a unique representation $h = \sum a_k c_k$, and (ii) there exist positive absolute constants $C_1, C_2$ such that for all $h \in H$, $C_1 \|h\|^2 \leq \sum_k |a_k|^2 \leq C_2 \|h\|^2$.

It is more common to replace condition (iv) by the weaker condition

(iv') $\exists g \in V_0$ such that $\{g(x - k) : k \in \mathbb{Z}\}$ is a Riesz basis for $V_0$.

That (iv') is equivalent to (iv) follows from the “orthonormalization trick” discussed below.

3. Conditions (i) - (iv) are not mutually independent - see Hernández & Weiss (1996).

A key point is the existence of approximations by $V_j$ with better smoothness and approximation properties than those of the Haar multiresolution. A function $f$ on $\mathbb{R}$ is rapidly decreasing if for all $m \in \mathbb{N}$, then $|f(x)| \leq C_m (1 + |x|)^{-m}$. Following Meyer (1990, p22), we say that a multiresolution analysis $\{V_j\}$ of $L_2(\mathbb{R})$ is $r$-regular if $D^r g(x)$ is rapidly decreasing for $0 \leq k \leq r \in \mathbb{N}$.

Example. Box spline MRA. Given $r \in \mathbb{N}$, set $\chi = I_{[0,1]}$ and

$$V_j = \{ f \in L_2 \cap C^{r-1} \text{ and } f|_{I_{j,k}} \text{ is a polynomial of degree } r \}.$$

$$g = g_r = \chi \ast \cdots \ast \chi = \chi^{k(r+1)}.$$

If $r = 0$, this reduces to the piecewise constant MRA given earlier. If $r = 1$, we get piecewise linear functions and if $r = 3$, cubic splines.

A key role in constructions and interpretations is played by the frequency domain and the Fourier transform, defined first for $f \in L_1(\mathbb{R})$ by

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} e^{-i\xi x} f(x) dx.$$

and then extended to $L_2$. The Plancherel identity leads to a frequency domain characterization of the orthonormality and Riesz basis conditions (iv) and (iv'):

B.1. Lemma. Suppose $\varphi \in L_2$. The set $\{\varphi(x - k), k \in \mathbb{Z}\}$ is (i) orthonormal iff

$$(B.1) \quad \sum_k |\hat{\varphi}(\xi + 2k\pi)|^2 = 1 \quad a.e.,$$
and (ii) a Riesz basis iff there exist positive constants $C_1, C_2$ such that

\begin{equation}
C_1 \leq \sum_k |\hat{\varphi}(\xi + 2k\pi)|^2 \leq C_2 \quad \text{a.e.}
\end{equation}

We give the easy proof of (B.1) since it gives a hint of the role of frequency domain methods. The Fourier transform of $\xi \to \varphi(x - n)$ is $e^{-inx}\hat{\varphi}(\xi)$. Thus, orthonormality combined with Parseval’s relation gives

$$\delta_{0n} = \int_{-\infty}^{\infty} \varphi(x)\overline{\varphi(x-n)}dx = \int_{-\infty}^{\infty} e^{inx}|\hat{\varphi}(\xi)|^2d\xi.$$  

Now partition $\mathbb{R}$ into segments of length $2\pi$, add the integrals, and exploit periodicity of $e^{inx}$ to rewrite the right hand side as

$$\int_0^{2\pi} e^{inx}\sum_k |\varphi(\xi + 2k\pi)|^2d\xi = \delta_{0n}.$$  

The function in (B.1) has as Fourier coefficients the delta sequence $\delta_{0n}$ and so equals $1$ a.e.

We describe two common constructions of a scaling function $\varphi$ (and then the corresponding wavelet $\psi$):

(a) beginning from a Riesz basis and the function $g$ in condition (iv’),

and

(b) starting from discrete (especially finite) filters (see Section B.2.1).

The first method uses the “orthonormalization trick”; which creates (B.1) by fiat:

B.2. THEOREM. Suppose that $\{V_j\}$ is an MRA, and that $\{g(x-k), k \in \mathbb{Z}\}$ is a Riesz basis for $V_0$. Define

\begin{equation}
\hat{\varphi}(\xi) = \hat{g}(\xi)/\left(\sum_k |g(\xi + 2k\pi)|^2\right)^{1/2}
\end{equation}

Then $\varphi$ is a scaling function for the MRA, i.e. $\forall j \in \mathbb{Z}$, $\{\varphi_{jk} : k \in \mathbb{Z}\}$ is an o.n.b. for $V_j$.

Example. (Box splines continued.) Without any loss of generality, we may shift $g_r = \chi^{(r+1)}$ by an integer so that the center of the support is at 0 if $r$ is odd, and at 1/2 if $r$ is even. Then it can be shown (Meyer 1990, p61), (Mallat 1999, Sec 7.1) that

$$\hat{g}_r(\xi) = \left(\frac{\sin \xi/2}{\xi/2}\right)^{r+1} e^{-i\xi/2} \quad \epsilon = \begin{cases} 1 & r \text{ even} \\ 0 & r \text{ odd} \end{cases},$$

$$\sum_k |\hat{g}_r(\xi + 2k\pi)|^2 = P_{2r}(\cos \xi/2),$$

where $P_{2r}$ is a polynomial of degree $2r$. For example, in the piecewise linear case $r = 1$, $P_2(\nu) = (1/3)(1 + 2\nu^2)$. Using (B.2), this establishes the Riesz basis condition (iv’) for this MRA. Thus (B.3) gives an explicit Fourier domain expression for $\varphi$ which is amenable to numerical calculation. Mallat
(1999, pp. 226-228) gives corresponding formulas and pictures for cubic splines.

**B.2.1. Filters.** Given an MRA with scaling function \( \varphi \), since \( V_{-1} \subset V_0 \), one may express \( \varphi_{-1,0} \) in terms of \( \varphi_{0,k} \) using the two scale equation

\[
\frac{1}{\sqrt{2}} \varphi\left( \frac{x}{2} \right) = \sum_k h_k \varphi(x - k).
\]

Now take Fourier transforms of both sides: since \( \overline{\varphi_{0,k}}(\xi) = e^{-ik\xi} \varphi(\xi) \), the two scale equation has the reexpression

\[
\hat{\varphi}(2\xi) = 2^{-1/2} \hat{h}(\xi) \hat{\varphi}(\xi),
\]

where

\[
\hat{h}(\xi) = \sum_k h_k e^{-ik\xi}.
\]

The sequence \( \{h_k\} \) is called the discrete filter associated with \( \varphi \), and \( \hat{h}(\xi) \) is its transfer function.

The MRA conditions imply important structural constraints on \( \hat{h}(\xi) \): using (B.1) and (B.5) it can be shown that

**B.3. Proposition.** If \( \varphi \) is an integrable scaling function for an MRA, then

\[
\begin{align*}
\text{(CMF)} & \quad |\hat{h}(\xi)|^2 + |\hat{h}(\xi + \pi)|^2 = 2 \quad \forall \xi \in \mathbb{R} \\
\text{(NORM)} & \quad \hat{h}(0) = \sqrt{2}.
\end{align*}
\]

(B.6) is called the conjugate mirror filter (CMF) condition, while (B.7) is a normalization requirement.

Conditions (B.7) and (B.6) respectively imply constraints on the discrete filters:

\[
\sum h_k = \sqrt{2}, \quad \sum h_k^2 = 1.
\]

The conditions (B.6) and (B.7) are the starting point for a unified construction of many of the important wavelet families (Daubechies variants, Meyer, . . . ) that starts with the filter \( \{h_k\} \), or equivalently \( \hat{h}(\xi) \). Here is a key result in this construction.

**B.4. Theorem** (Meyer, Mallat). If \( \hat{h}(\xi) \) is \( 2\pi \)-periodic, \( C^1 \) near \( \xi = 0 \) and

(a) satisfies (B.6) and (B.7), and (b) \( \inf_{[-\pi,\pi]} |\hat{h}(\xi)| > 0 \), then

\[
\hat{\varphi}(\xi) = \prod_{l=1}^{\infty} \frac{\hat{h}(2^{-l}\xi)}{\sqrt{2}}
\]

is the Fourier transform of a scaling function \( \varphi \in L_2 \) that generates an MRA.
Notes: 1. That \( \hat{\phi} \) is generated by an infinite product might be guessed by iteration of the two scale relation (B.5): the work lies in establishing that all MRA properties hold.

2. Condition (b) can be weakened to a necessary and sufficient condition due to Cohen (1990) (see also Cohen & Ryan (1995)).

B.2.2. From Scaling functions to wavelets. Define the detail subspace \( W_j \subset L_2 \) as the orthogonal complement of \( V_j \) in \( V_{j+1} \): \( V_{j+1} = V_j \oplus W_j \). A candidate wavelet \( \psi \in W_j \subset V_0 \) satisfies the two scale equation

\[
\frac{1}{\sqrt{2}} \hat{\phi}(\frac{x}{2}) = \sum_k g_k \varphi(x - k).
\]

Again, taking the Fourier transform of both sides and defining \( \hat{g}(\xi) = \sum g_k e^{-ik\xi} \),

\[
\hat{\psi}(2\xi) = 2^{-1/2} \hat{g}(\xi) \hat{\phi}(\xi).
\]

Define \( \psi_{jk}(x) = 2^{j/2} \psi(2^j x - k) \). The next lemma gives the conditions on \( g \) in order that \( \psi \) be an orthonormal wavelet: it is an analog of Proposition B.3.

B.5. Lemma. \( \{\psi_{jk}, k \in \mathbb{Z}\} \) is an orthonormal basis for \( W_j \), the orthogonal complement of \( V_j \) in \( V_{j+1} \) if and only if, for all \( \xi \in \mathbb{R} \),

\[
|\hat{g}(\xi)|^2 + |\hat{g}(\xi + \pi)|^2 = 2
\]

\[
\hat{g}(\xi) \hat{h}^*(\xi) + \hat{g}(\xi + \pi) \hat{h}^*(\xi + \pi) = 0.
\]

Here \( \hat{h}^* \) denotes the complex conjugate of \( \hat{h} \).

One way to satisfy (B.10) and (B.11) is to set

\[
\hat{g}(\xi) = e^{-i\xi} \hat{h}^*(\xi + \pi)
\]

Interpreting this in terms of filter coefficients, one obtains the “mirror” relation

\[
g_k = (-1)^{1-k} h_{1-k}.
\]

To better appreciate the time domain consequences of the choice (B.12), note that if \( s(\xi) \) has coefficients \( s_k \), then conjugation corresponds to time reversal: \( s^*(\xi) \leftrightarrow s_{-k} \), while modulation corresponds to time shift: \( e^{i\xi s(\xi)} \leftrightarrow s_{k+1} \), and frequency shift goes over to time modulation: \( s(\xi + \pi) \leftrightarrow (-1)^k s_k \).

Together, (B.9) and (B.12) provide a frequency domain recipe for constructing a candidate wavelet from \( \varphi \):

\[
\hat{\psi}(2\xi) = e^{-i\xi} \hat{h}^*(\xi + \pi) \hat{\phi}(\xi).
\]

Of course, there is still work to do to show that this does the job:

B.6. Theorem. If \( g \) is defined by (B.12), and \( \psi \) by (B.9), then \( \{\psi_{jk}, (j,k) \in \mathbb{Z}^2\} \) is an orthonormal basis for \( L_2(\mathbb{R}) \).
Thus we have decompositions

\[ L_2(\mathbb{R}) = \oplus_{j \in \mathbb{Z}} W_j = V_J \oplus \bigoplus_{j \geq J} W_j, \]

with corresponding expansions

\[ f = \sum_{j,k} \langle f, \psi_{jk} \rangle \psi_{jk} = \sum_{k} \langle f, \varphi_{jk} \rangle \varphi_{jk} + \sum_{j \geq J} \sum_{k} \langle f, \psi_{jk} \rangle \psi_{jk}. \]

The first is called a *homogeneous* expansion, while the second is inhomogeneous since it combines only the detail spaces at scales finer than \( J \).

*Example.* Box splines again. Given \( \hat{\varphi} \), one constructs \( \hat{h} \) from (B.5), \( \hat{g} \) from (B.12) and \( \hat{\psi} \) from (B.9). This leads to the *Battle-Lemarié spline wavelets* (see also Chui (1992)). The case \( r = 0 \) yields the Haar wavelet:

\[ \psi(x) = I_{[1/2,1]}(x) - I_{[0,1/2]}(x) \]

- verifying this via this construction is possibly a useful exercise in chasing definitions. However, the point of the construction is to yield wavelets with increasing regularity properties as \( r \) increases.

Figure ?? shows a schematic of the qualitative frequency domain properties of the squared modulus of \( \varphi, \hat{\varphi}, \hat{g} \) and finally \( \hat{\psi} \). It can be seen that the space \( V_0 \) generated by translates of \( \varphi \) corresponds roughly to frequencies around \( \pm [0, \pi] \), while the space \( W_j \) contains frequencies around \( \pm [2^j \pi, 2^{j+1} \pi] \). More precisely, it can be shown (Hernández & Weiss 1996, p.332, and p.61) that \( \varphi \) and the dilations of \( \psi \) form a partition of frequency space in the sense that

\[ |\hat{\psi}(\xi)|^2 + \sum_{j=0}^{\infty} |\hat{\psi}(2^{-j}\xi)|^2 = 1 \quad \text{a.e.} \]  

(*B.15*)

*Example.* The class of *Meyer wavelets* (Meyer 1986) is built from a filter \( \hat{h}(\xi) \) on \( [-\pi, \pi] \) satisfying

\[ \hat{h}(\xi) = \begin{cases} 1 & |\xi| \leq \pi/3 \\ 0 & |\xi| \geq 2\pi/3, \end{cases} \]

the CMF condition (B.6), and that is also required to be \( C^n \) at the join points \( \pm \pi/3 \) and \( \pm 2\pi/3 \). In fact \( C^\infty \) functions exist with these properties, but for numerical implementation one is content with finite values of \( n \), for which computable descriptions are available: for example \( n = 3 \) in the case given by Daubechies (1992, p.137-8).

The scaling function \( \hat{\varphi}(\xi) = \prod_1^\infty 2^{-1/2} \hat{h}(2^{-j}\omega) \) then has support in \( [-4\pi/3, 4\pi/3] \), and the corresponding wavelet (defined from (B.9) and (B.12)) has support in the interval \( \pm [2\pi/3, 8\pi/3] \). Since \( \hat{\varphi} \) and \( \hat{\psi} \) have compact support, both \( \varphi(x) \) and \( \psi(x) \) are \( C^\infty \) \(-\) unlike, say, Daubechies wavelets. However, they cannot have exponential decay in the time domain (which is impossible for \( C^\infty \) orthogonal wavelets, according to Daubechies (1992, Corollary 5.5.3)) \(-\) at least they are \( O(|x|^{-n-1}) \) if \( \hat{h} \) is \( C^n \). Finally, since \( \hat{\psi} \)
vanishes in a neighborhood of the origin, all its derivatives are zero at 0 and so \( \psi \) has an infinite number of vanishing moments.

**Remark.** The two scale relations (B.4) and (B.8) yield information on the support of the scaling function \( \varphi \) and wavelet \( \psi \). For a filter \( h = (h_k, k \in \mathbb{Z}) \), its support is the smallest closed interval containing the non-zero values of \( h_k \). For example, Mallet (1999, Chapter 7) shows that

(i) \( \text{supp } \varphi = \text{supp } h \) if one of the two is compact, and

(ii) if \( \text{supp } \varphi = [N_1, N_2] \), then \( \text{supp } \psi = \left[ \frac{N_1 - N_2 + 1}{2}, \frac{N_2 - N_1 + 1}{2} \right] \).

**B.2.3. The significance of vanishing moments.** A key feature of wavelet analysis is that the wavelet coefficients of a function are large only at low frequencies and in the vicinity of singularities. More precisely, the rate of decay of wavelet coefficients of a smooth function is governed by the number of vanishing moments of the wavelet \( \psi \).

The wavelet \( \psi \) is said to have \( r \) vanishing moments if

\[
\int x^k \psi(x) dx = 0 \quad k = 0, 1, \ldots, r - 1.
\]

Thus \( \psi \) is orthogonal to all polynomials of degree \( r - 1 \).

**Hölder regularity.** A function \( f \) is said to be \( C^\alpha \) \((\alpha > 0)\) on the interval \( I \subset \mathbb{R} \) if there exists a constant \( C \) and for every \( x \in I \), there exists a polynomial \( p_x(y) \) of degree \( \lfloor \alpha \rfloor \) such that

\[
|f(x + y) - p_x(y)| \leq C|y|^\alpha, \quad x + y \in I.
\]

In particular, if \( f \) is \( C^\alpha \) on \( I \), then it is \( m \) times continuously differentiable on \( I \) for integers \( m < \alpha \).

**B.7. Lemma.** If \( f \) is \( C^\alpha \) on \( \mathbb{R} \) and \( \psi \) has at least \( r = \lfloor \alpha \rfloor + 1 \) vanishing moments,

\[
|\langle f, \psi_{jk} \rangle| \leq c_\psi C 2^{-j(\alpha + 1/2)},
\]

where \( C = C(f) \) is the constant appearing in (B.16).

**Proof.** Let \( p(y) \) be the approximating polynomial of degree \( \lfloor \alpha \rfloor = r - 1 \) at \( x_k = k2^{-j} \). Using a change of variable and the vanishing moments property,

\[
\int f(x)2^{j/2}\psi(2^jx - k)dx = 2^{-j/2}\int [f(x_k + 2^{-j}v) - p(2^{-j}v)]\psi(v)dv.
\]

Hence, using the Hölder bound

\[
|\langle f, \psi_{jk} \rangle| \leq 2^{-j/2}C 2^{-j\alpha} \int |v|^\alpha |\psi(v)|dv.
\]

Setting \( c_\psi \) equal to the latter integral yields the result. \( \square \)

The condition that \( \psi \) have \( r \) vanishing moments has equivalent formulations in terms of the Fourier transform of \( \psi \) and the filter \( h \).
Table 1. Desirable properties of orthonormal wavelet family, together with corresponding conditions on the filter $h$

1. Orthonormal wavelet $\psi$ $\leftrightarrow$ CMF (B.6) and NORM (B.7)
2. $p$ vanishing moments $\leftrightarrow$ VM$_p$ (B.17)
3. (small) compact support $\leftrightarrow$ $N$ small
4. (high) regularity of $\psi$

B.8. Lemma. Let $\psi$ be an orthonormal wavelet. If $\hat{\psi}$ is $C^p$ at $\xi = 0$, then the following are equivalent:

$$(i) \int t^j \psi = 0, \quad j = 0, \ldots, p - 1.$$

$$(ii) D^j \hat{\psi}(0) = 0, \quad j = 0, \ldots, p - 1.$$ (B.17)

$$(iii) D^j \hat{h}(\pi) = 0 \quad j = 0, \ldots, p - 1. \quad (VM_p)$$

See for example Mallat (1999)[Theorem 7.4] or Härdele et al. (1998)[Theorem 8.3].

Meyer (1990)[p.38] shows that a wavelet deriving from an $r$-regular multiresolution analysis necessarily has $r$ vanishing moments.

Here is a brief sketch, with a probabilistic twist, of some of the steps in Daubechies’ construction of orthonormal wavelets of compact support. Of course, there is no substitute for reading the original accounts (see Daubechies (1988), Daubechies (1992, Ch. 6), and for example the descriptions by Mallat (1999, Ch. 7) and Meyer (1990, Vol I, Ch. 3)). Table 1 sets out some desiderata for a wavelet basis. The last three requirements are in a sense contradictory: it turns out that higher regularity of $\psi$ can only be achieved with longer filters. One advantage of Daubechies’ family of wavelets $\psi_N$, indexed by support size $N$, is to make this tradeoff directly apparent: the smoothness of $\psi$ increases with $N$ at approximate rate 0.2 $N$.

[CHECK]

The approach is to build a filter $h = \{h_k\}_{0}^{N-1}$ with $h_k \in \mathbb{R}$ and transfer function $\hat{h}(\xi) = \sum_{k=0}^{N-1} h_k e^{-i k \xi}$ satisfying these conditions and then derive the conjugate filter $g$ and the wavelet $\psi$ from Theorem B.6. The vanishing moment condition of order $p$ (VM$_p$) implies that $\hat{h}(\xi)$ may be written

$$\hat{h}(\xi) = \left(1 + \frac{e^{-i \xi}}{2}\right)^p r(\xi), \quad r(\xi) = \sum_{k=0}^{m} r_k e^{-i k \xi},$$

with $N = p + m + 1$ and $r_k \in \mathbb{R}$. Passing to squared moduli, one may write

$$|\hat{h}(\xi)|^2 = 2(\cos^2 \frac{\xi}{2})^p P(\sin^2 \frac{\xi}{2})$$

for some real polynomial $P$ of degree $m$. The conjugate mirror filter condition (B.6) then forces, on putting $y = \sin^2 \xi/2$,

(B.18) \[ (1 - y)^m P(y) + y^m P(1 - y) = 1 \quad 0 \leq y \leq 1. \]
To have the support length $N$ as small as possible, one seeks solutions of (B.18) of minimal degree $m$. One solution can be described probabilistically in terms of repeated independent tosses of a coin with $Pr(\text{Heads}) = y$. Either $p$ tails occur before $p$ heads or vice versa, so

$$P(y) = Pr\{p \text{ T's occur before } p \text{ H's}\}/(1 - y)^p$$

$$= \sum_{k=0}^{p-1} \binom{p + k - 1}{k} y^k$$

certainly solves (B.18). Further, it is the unique solution of degree $p - 1$ or less.

To return from the squared modulus scale, appeal to the F. Riesz lemma: if $s(\xi) = \sum_{m} s_k e^{-ik\xi} \geq 0$, then there exists $r(\xi) = \sum_{n} r_k e^{-ink\xi}$ such that $s(\xi) = |r(\xi)|^2$, and if $\{s_k\}$ are real, then the $\{r_k\}$ can be chosen to be real also.

The lemma is applied to $s(\xi) = P(\sin^2 \frac{\xi}{2}) \geq 0$, and so one arrives at orthonormal wavelets with support length $N = 2p$ for $p = 1, 2, \ldots$. The uniqueness argument shows that $N < 2p$ is not possible. The choice $N = 2$ yields Haar wavelets and $N = 4$ gives the celebrated D4 wavelet of Daubechies. For $N \geq 6$ there are non-unique choices of solution to the construction of the “square root” $r(\xi)$ (a process called spectral factorization), and Daubechies (1992, Ch. 6) describes some families of solutions (for example, directed towards least asymmetry) along with explicit listings of coefficients.

**Vanishing moments for the scaling function.** The approximation of point values $f(t_k)$ of a function by scaling function coefficients $\langle f, 2^{j/2}\varphi_{jk}\rangle$ is similarly dependent on the smoothness of $f$ and the number of vanishing moments of $\varphi$. Bearing in mind that the scaling function itself has $\int \varphi = 1$, we say that $\varphi$ has $r$ vanishing moments if

$$\int x^k \varphi(x) dx = 0 \quad k = 1, \ldots r - 1.$$ 

**B.9. Lemma.** If $f$ is $C^\alpha$ on $\mathbb{R}$ and $\varphi$ has at least $r = [\alpha] + 1$ vanishing moments,

$$|\langle f, \varphi_{jk}\rangle - 2^{-j/2} f(k2^{-j})| \leq c_\varphi C 2^{-j(\alpha + 1/2)}.$$

**Proof.** Modify the proof of Lemma B.7 by writing the approximating polynomial at $x_k = k2^{-j}$ in the form $p(y) = f(x_k) + p_1(y)$ where $p_1$ is also of degree $r - 1$, but with no constant term, so that $\int p_1 \varphi = 0$. Then

$$\int \varphi_{jk} - 2^{-j/2} f(x_k) = 2^{-j/2} \int [f(x_k + 2^{-j}v) - f(x_k) - p_1(2^{-j}v)] \varphi(v) dv$$

and so

$$|\langle f, \varphi_{jk}\rangle - 2^{-j/2} f(x_k)| \leq 2^{-j/2} C 2^{-j\alpha} c_\varphi,$$

where again $c_\varphi = \int |v|^\alpha |\varphi(v)| dv$. \qed

275
B.2.4. Life on the interval. In statistical applications, one is often interested in an unknown function \( f \) defined on an interval, say \( I = [0,1] \) after rescaling. Brutal extension of \( f \) to \( \mathbb{R} \) by setting it to 0 outside \( I \), or even more sophisticated reflection/folding extensions, introduce a discontinuity in \( f \) or its derivatives at the edges of \( I \).

If one works with wavelets of compact support (of length \( S \), say), these discontinuities affect only a fixed number \( 2S \) of coefficients at each level \( j \) and so will often not affect the asymptotic behavior of global measures of estimation error on \( I \). Nevertheless, both in theory and in practice, it is desirable to avoid such artificially created discontinuities. We refer here to two approaches that have been taken in the literature. [A third, “folding” across boundaries, is discussed in Mallat (1999, Sec. 7.5.2.).]

(i) Periodization. One restricts attention to periodic functions on \( I \). Meyer (1990, Vol 1, Chapter III.11) shows that one can build an orthonormal basis for \( L_{2,\text{per}}(I) \) by periodization. Suppose that \( \varphi \) and \( \psi \) are nice orthonormal scaling and wavelet functions for \( L_2(\mathbb{R}) \) and define

\[
\varphi_{j,k}^{\text{per}}(x) = \sum_{\ell \in \mathbb{Z}} \varphi_{j,k}(x + \ell), \quad \psi_{j,k}^{\text{per}}(x) = \sum_{\ell \in \mathbb{Z}} \psi_{j,k}(x + \ell).
\]

If \( \varphi, \psi \) have compact support, then for \( j \) larger than some \( j_1 \), these sums reduce to a single term for each \( x \in I \). Define \( V_{j}^{\text{per}} = \text{span} \{ \varphi_{j,k}^{\text{per}}, k \in \mathbb{Z} \} \), and \( W_{j}^{\text{per}} = \text{span} \{ \psi_{j,k}^{\text{per}}, k \in \mathbb{Z} \} \) : this yields an orthogonal decomposition

\[
L_{2,\text{per}}(I) = V_{j}^{\text{per}} \oplus \bigoplus_{j \geq 0} W_{j}^{\text{per}},
\]

with \( \dim V_{j}^{\text{per}} = \dim W_{j}^{\text{per}} = 2^j \) for \( j \geq 0 \). Meyer (1990, Vol 1, Chapter III.11) has a detailed comparison of Fourier series and wavelets on \([0,1]\), including remarkable properties such as uniform convergence of the wavelet approximations of any continuous function on \([0,1]\).

(ii) Orthonormalization on \([0,1]\) [see Meyer (1992?a), Cohen et al. (1993b,a).] For non-periodic functions on \([0,1]\), one must take a different approach. Suppose \( \psi \) is an orthogonal wavelet of compact support, which for convenience we suppose to have been shifted to the interval \([-k_0, k_0+1]\). An orthonormal basis \( \{ \psi_{j,k}^{\text{int}} \} \) for \( L_2[0,1] \) is defined in terms of the orthonormal basis \( \{ \psi_{j,k} \} \) for \( L_2(\mathbb{R}) \) as follows. For \( k_0 \leq k < 2^j - k_0 \), the support of \( \psi_{j,k} \) is contained in \( I \), so set \( \psi_{j,k}^{\text{int}} = \psi_{j,k} \). In the boundary cases, \( \psi_{j,k}^{\text{int}} \) are obtained from linear combinations of \( \psi_{j,k} \) by orthonormalization on the interval \([0,1]\). In particular, there exist \( k_0 \times k_0 \) matrices \( E^j = (e_{k,k'}^{j}) \), \( E' = (e_{k,k'}^{j}) \), not depending on \( j \), such that for \( 0 \leq k < k_0 \):

\[
\psi_{j,k}^{\text{int}}(t) = \sum_{k'=0}^{k_0-1} e_{k,k'}^{j} \psi_{j,k'}(t), \quad \psi_{j,2^j-k-1}^{\text{int}}(t) = \sum_{k'=0}^{k_0-1} e_{k,k'}^{j} \psi_{j,2^j-k'-1}(t).
\]
In particular, the 'boundary' wavelets have the same smoothness properties as the interior wavelets.

Vanishing moments for wavelets on $[0,1]$. Let $\mathcal{P}_p$ denote the space of polynomials of degree strictly less than $p$. The vanishing moments theorem (e.g. Mallat (1999, Theorem 7.4)) states that if $\varphi$ and $\psi$ have sufficiently rapid decay, then $\psi$ has $p$ vanishing moments if and only if the Strang-Fix condition is satisfied:

$$\theta_l(t) = \sum_{k=-\infty}^{\infty} k^l \varphi(t - k) \in \mathcal{P}_p \quad l = 0, 1, \ldots, p - 1.$$  \hfill (B.20)

The condition (B.20) says that $\mathcal{P}_p \subset V_j$ and further (see Cohen et al. (1993a)) that for $j \geq J_*, \mathcal{P}_p \subset V_j[0,1] = \cdots$ the multiresolution spaces corresponding to the CDJF construction. Consequently $\mathcal{P}_p \perp W_j[0,1]$ and so for $j \geq J_*$, $k = 1, \ldots, 2^j$, we have

$$\int t^j \psi_{jk}^i(t) dt = 0, \quad l = 0, 1, \ldots, p - 1.$$

[Fast algorithms??]

B.3. Function Spaces.


They scales are of interest in statistical theory because they capture a broad range of smoothness and spatial adaptation properties.

In this section we summarise some definitions and properties, with a particular view towards the sequence space characterization. These are taken primarily from Frazier et al. (1991) and Meyer (1990, Volume 1), which should be consulted for further details.

If $\alpha$ is a positive integer, the (Hilbert-)Sobolev measure of smoothness of the $\alpha^{th}$ derivative $D^\alpha f$ is an $L_2$ integral, which by Parseval’s relation has a simple expression also in the Fourier domain:

$$\int (D^\alpha f)^2 = \int |\xi|^{2\alpha} |\widehat{f}(\xi)|^2 d\xi.$$ \hfill (B.21)

As motivation for the Littlewood-Paley definition of the Besov and Triebel spaces, we rewrite the $L_2$-Sobolev measure using a dyadic decomposition of frequency. Suppose that the wavelet $\psi$ satisfies a homogeneous version of (B.15): $\sum_{j \in \mathbb{Z}} |\psi(2^{-j} \xi)|^2 = 1$ a.e. – this holds for a wide class of wavelets, including the Meyer wavelet (Hernández & Weiss 1996).
Let \( \varphi_t(x) = e^{-\alpha x} \varphi(x/e) \). Given a function \( f \), define filtered versions \( f_j \) via either of the (equivalent) forms
\[
f_j(x) = \psi_{2^{-j}} \ast f, \quad \hat{f}_j(\xi) = \hat{\psi}(2^{-j}\xi)\hat{f}(\xi).
\]
Suppose the the support of \( \hat{\psi}(\xi) \) is contained in \( \pm[\pi/2, 2\pi] \); then convolution with \( \psi_{2^{-j}}(t) = 2^j \psi(2^j t) \) corresponds to a bandpass filter that preserves frequencies near \( \pm 2^j \pi \).

Using (B.21) along with the dyadic decomposition and the support constraint on \( \psi \),
\[
\int (D^\alpha f)^2 = \sum_j \int |\xi|^{2\alpha} |\varphi(2^{-j}\xi)\hat{f}(\xi)|^2 d\xi
\leq \sum_j 2^{2\alpha j} \int |\hat{f}_j(\xi)|^2 d\xi
= \sum_j 2^{2\alpha j} \|f_j\|_2^2,
\]  
(B.22)
where \( \leq \) denotes equivalence up to multiplicative constants depending only on \( \alpha \). The \( L_2 \)-Sobolev norm is thus expressed using \( f_j(x) \) in terms of a bivariate decomposition over space \( x \) and frequency octave \( j \).

The function space norms are defined by integrating over space in \( L_p \) and weighted frequency in \( \ell_q \). Let \( \varphi \) be a “window” function (no longer specifically a wavelet as above) belonging to the (Comments on \( L_p \) norms and inhomogeneity) Schwartz space \( S \) of rapidly decreasing functions. It is assumed that \( \varphi \) satisfies
\[
\text{supp } \varphi \subset \{1/2 < |\xi| < 2\}, \quad |\hat{\varphi}| > c \text{ on } \{3/5 < |\xi| < 5/3\}.
\]
Again, given a function \( f \), define filtered versions \( f_j \) such that \( \hat{f}_j \) is concentrated on \( \{2^{j-1} < |\xi| < 2^{j+1}\} \), via \( f_j(x) = \varphi_{2^{-j}} \ast f(x) \), or equivalently, \( \hat{f}_j(\xi) = \hat{\varphi}(2^{-j}\xi)\hat{f}(\xi) \).

Definitions. For \( \alpha \in \mathbb{R}, 0 < p, q < \infty, f \in S' \); the Triebel-Lizorkin seminorm is defined by
\[
\|f\|_{F_{p,q}^{\alpha}} = \left\{ \sum_j (2^{\alpha j} |f_j|^q)^{1/q} \right\}^{1/q} \|f\|_{L_p},
\]
while for the same \( (\alpha,p,q) \), and also \( p = \infty \); the Besov seminorm is given by
\[
\|f\|_{B_{p,q}^{\alpha}} = \left\{ \sum_j (2^{\alpha j} |f_j|^q)^{1/q} \right\}^{1/q}.
\]
Note that the Triebel norm integrates first over scale and then location, whereas the Besov norm reverses the order: location and then scale. In either event, (C.4) shows that \( \int (D^\alpha f)^2 = \|f\|_{B_{p,2}^{\alpha}}^2 \). In each case, there is a geometric weighting of octaves \( j \) according to the smoothness index \( \alpha \). It can be shown that the spaces do not depend on the particular choice of \( \varphi \).
The usual modification to the definition of the norms is made if \( p = \infty \) or \( q = \infty \): for example
\[
\| f \|_{\tilde{B}^0_{\infty, \infty}} = \sup_j 2^j \| f_j \|_{\infty}.
\]

Remarks and properties. (for further details, see Frazier et al. (1991))

1° These are the homogeneous definitions: if \( t = 2^{-j} \), then
\[
\| f_t \|_{\tilde{F}^\alpha_{p, q}} = 2^{\alpha(1/p - 1)} \| f \|_{\tilde{F}^\alpha_{p, q}}.
\]

2° The homogeneous norms are only defined up to polynomials. That is, if \( f \in S' \) (the space of tempered distributions) has zero norm, then \( f_j = 0 \) for all \( j \) which implies that \( \hat{f}(\xi) = 0 \) for all \( \xi \neq 0 \), which implies that \( f \) is polynomial. The need to work modulo polynomials is avoided in the inhomogeneous case, defined below.

3° \( \| \cdot \|_{F, \tilde{B}} \) are norms for \( 1 \leq p, q \leq \infty \); otherwise they are still quasi-norms: the triangle inequality is weakened to the existence of a constant of \( c \geq 1 \) such that
\[
\| f + g \| \leq c \| f \| + c \| g \|.
\]

4° Both Besov and Triebel norms can be thought of as examples of vector norms: given \( g = \{ g_j(x) \}_{j \in \mathbb{Z}} \); define
\[
\| g \|_{L_p(t_\alpha)} = \left\{ \sum_j |g_j|^q \right\}^{1/q} \quad \| g \|_{\ell_q(L_p)} = \left\{ \sum_j |g_j|^q \right\}^{1/q}.
\]
Indeed, if \( g_j = 2^j \varphi_{2^{-j}} * f \); then clearly
\[
\| f \|_{\tilde{F}^\alpha_{p, q}} = \| g \|_{L_p(t_\alpha)}, \quad \| f \|_{\tilde{B}^\alpha_{p, q}} = \| g \|_{\ell_q(L_p)}.
\]

5° The parameter \( \alpha \) measures smoothness: if \( \sim \) denotes equivalence of norms,
\[
\| f \|_{\tilde{F}^\alpha_{p, q}} \sim \| f' \|_{\tilde{F}^{\alpha'}_{p', q'}},
\]

6° If \( p = q \), then \( F^\alpha_{p, p} = B^\alpha_{p, p} \) and
\[
\| f \|_{\tilde{F}^\alpha_{p, p}} = \int \sum_j 2^{\alpha p} |f_j|^p.
\]

The notation \( B_1 \subset B_2 \) means that \( B_1 \) is embedded in \( B_2 \): for some constant \( C \) one has \( \| \cdot \|_{B_2} \leq C \| \cdot \|_{B_1} \). The embedding relations between Besov and Triebel norms are then given by:

- if \( p < q \leq \infty \), \( \tilde{F} \subset \tilde{B} \);
- if \( p > q \), \( B \subset \tilde{F} \).
Also, if $q \leq q'$ then $B^\alpha_{p,q} \subset B^\alpha_{p,q'}$.

Inhomogeneous versions of the norms are defined by bringing in a “low frequency” function $\varphi$ with the properties that $\text{Supp } \hat{\varphi} \subset [-2,2]$, and $\hat{\varphi} > c$ on $[-5/3,5/3]$. Then define

$$
\| f \|_{F^\alpha_{p,q}} = \| \varphi \ast f \|_{L_p} + \| (\sum_{j \geq 1} (2^\alpha \| f_j \|_p)^q \|_\frac{1}{q} \|_{L_p},
$$

$$
\| f \|_{B^\alpha_{p,q}} = \| \varphi \ast f \|_{L_p} + (\sum_{j \geq 1} (2^\alpha \| f_j \|_p)^q)^{\frac{1}{q}}.
$$

Now $\| f \| = 0 \Rightarrow f = 0$ a.e. (and not just modulo polynomials). Again, finiteness of the norms does not depend on the specific choice of the functions $\varphi$. If one adds the additional requirement that $\varphi, \varphi$ satisfy

$$
\hat{\varphi}(\xi) + \sum_{j \geq 1} \hat{\varphi}(2^{-j} \xi) = 1,
$$

then for $\alpha > 0, 1 \leq p < \infty, 0 < q \leq \infty$, it can be shown that

$$
\| f \|_{B^\alpha_{p,q}} \sim \| f \|_{L_p} + \| f \|_{B^\alpha_{p,q}}
$$

with the same relation holding for Triebel norms as well. The latter relation is both simpler than the general definition, and clarifies the relation between homogeneous and inhomogeneous definitions.

**Identifications.** Many of the traditional function spaces of analysis (and non-parametric statistics) can be identified as members of either or both of the Besov and Triebel scales. Some examples from Frazier et al. (1991) are summarized in tabular form below.

<table>
<thead>
<tr>
<th>Space</th>
<th>Identification</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hölder</td>
<td>$C^\alpha \sim B^\alpha_{\infty,\infty} (= F^\alpha_{\infty,\infty})$</td>
<td>$\alpha &gt; 0$</td>
</tr>
<tr>
<td>Hilbert-Sobolev</td>
<td>$W^\alpha_2 \sim B^\alpha_{2,2}$</td>
<td>$\alpha &gt; 0, 1 &lt; p &lt; \infty$</td>
</tr>
<tr>
<td>Sobolev</td>
<td>$W^\alpha_p \sim F^\alpha_{p,2}$</td>
<td>$\alpha &gt; 0, 1 &lt; p &lt; \infty$</td>
</tr>
<tr>
<td>$L_p$</td>
<td>$L_p \sim F^0_{p,2}$</td>
<td>$1 &lt; p &lt; \infty$</td>
</tr>
<tr>
<td>$H_p$</td>
<td>$H_p \sim F^0_{p,2}$</td>
<td>$0 &lt; p \leq 1$</td>
</tr>
<tr>
<td>BMO</td>
<td>$BMO \sim F^0_{\infty,2}$</td>
<td></td>
</tr>
</tbody>
</table>

**Remark.** A little care is needed with the description of the Hölder spaces. If $0 < \alpha < 1$, the inhomogeneous Hölder norm is defined as usual by

$$
\| f \|_{C^\alpha} = \| f \|_\infty + \sup_{x,y} \frac{|f(x) - f(y)|}{|x - y|^\alpha}.
$$

However, when $\alpha = 1$, for compatibility with the general definitions (and for its better properties in mathematical applications) the Zygmund extension of the traditional Lipschitz definition is used:

$$
\| f \|_{C^1} = \| f \|_\infty + \sup_{x,y} \frac{|f(x+y) - 2f(x) + f(x-y)|}{|y|}.
$$
(for example, the second term is finite for \( f(x) = |x| \log |x| \), even though it has unbounded derivative.) Finally, if \( \alpha \in (m, m + 1) \) for \( m \in \mathbb{N} \), then \( f \) is said to belong to \( C^\alpha \) if \( D^\alpha f \in C^{\alpha-m} \).

**Inhomogenous smoothness and Imbedding theorems.** The Hölder spaces \( (p = q = \infty) \) measure smoothness in a uniform norm: equally good behaviour is demanded at every point. The use of \( L_p \) norms, \( p < \infty \) to measure smoothness allows non-uniformity with respect to position - to an increasingly large degree as \( p \) decreases. A trivial example is given by setting \( f = M I_A \), a scale multiple of the indicator function of a set \( A \subset [0,1] \). The Lebesgue measure \( |A| \) of \( A \) describes the size of the set where \( f \) is not zero, and as such, is a measure of inhomogeneity. Of course \( \|f\|_p = |f||A|^{1/p} \to M \) as \( p \to \infty \) as soon as \( |A| > 0 \), regardless of its size. However, for fixed \( p \), and for \( \|f\|_p = M/2 \), one needs \( |A| = 2^{-p} \) and the latter approaches 1 as \( p \to 0 \). Thus, smaller values of \( p \) are much more forgiving of inhomogeneity.

Thus, a function with a small number of isolated discontinuities (or more generally singularities), is nevertheless smooth “on average.” If non-parametric estimation is being assessed via a global norm, then one should expect the rate of convergence of good estimators to be related reflect the average rather than worst case smoothness.

The relation between smoothness measured in different \( L_p \) norms as \( p \) varies is expressed by embedding theorems (see e.g. Peetre (1975, p. 63))

**B.10. Proposition.** If \( \alpha < \alpha' \) and \( p > p' \) are related by \( \alpha = 1/p \), then

\[
B_{p,q}^{\alpha'} \subset B_{p,q}^{\alpha}
\]

In fact, the proof becomes trivial using the sequence space characterization given in Section B.3.2 below.

The situation can be summarized in Figure ??, which represents smoothness \( \alpha \) in the vertical direction, and \( 1/p \) in the horizontal, for a fixed value of \( q \). Thus the \( y \)-axis corresponds to uniform smoothness, and increasing spatial inhomogeneity to \( 1/p \). The imbeddings proceed down the lines of unit slope: for example, inhomogeneous smoothness \((\alpha', 1/p')\) with \( \alpha' > 1/p' \) implies uniform smoothness of lower degree \( \alpha = \alpha' - 1/p' \).

The line \( \alpha = 1/p \) represents the boundary of continuity. If \( \alpha > 1/p \), then functions in \( B_{p,q}^{\alpha} \) are continuous by the embedding theorem just cited. However in general, the spaces with \( \alpha = 1/p \) may contain discontinuous functions – one example is given by the containment \( B_{1,1}^{1} \subset TV \subset B_{1,\infty}^{1} \).

Finally, for \( B_{p,q}^{\alpha}([0,1]) \), the line \( \alpha = 1/p - 1/2 \) represents the boundary of \( L_2 \) compactness - if \( \alpha > 1/p - 1/2 \), then \( B_{p,q}^{\alpha} \) norm balls are compact in \( L_2 \); this observation is basic to estimation in the \( L_2 \) norm.

Thus, one sees that in passing from the Hilbert space decomposition of scale and location (C.4) to the Besov and Triebel definitions, it is the substitution of integration in \( L_p(dx) \) for \( L_2(dx) \) that is most important. The use of \( l_q \) to combine across octaves is usually much less critical – indeed,
since, for example, $B^\alpha_{p,0} \subseteq B^\alpha_{p,\infty}$, it often suffices to work in $B^\alpha_{p,\infty}$ where, for example, $\|f\|_{B^\alpha_{p,\infty}} \leq C$ is equivalent to $\|f\|_{L_p} \leq C2^{-\alpha}$ for each $j$.

**B.3.2. Sequence spaces and smoothness.** Our statistical work is greatly facilitated by the fact that these function space norms are equivalent to various graded norms on sequences of wavelet coefficients. To describe this we first define some sequence norms, beginning with inhomogeneous forms. We write

$$\theta = \{\alpha_{jk}, j \geq j_0, k \in \mathbb{Z}\} \cup \{\beta_{jk}, k \in \mathbb{Z}\}.$$ 

To use $\theta$ as an abbreviation for a doubly indexed array of coefficients $\alpha$ (which might be orthonormal wavelet coefficients and a “coarse level” wave sequence $\beta$).

The Besov sequence norms are defined as follows. Suppose $\alpha \in \mathbb{R}$ and $0 < p, q \leq \infty$ and that we write $s = \alpha + 1/2 - 1/p$. Then

$$\|\theta\|_{\ell^p_{s,q}} = \|\beta_{jk}\|_{\ell^p} + \left\{ \sum_{j \geq j_0} 2^{jq} \left( \sum_k |\alpha_{jk}|^q \right)^{q/p} \right\}^{1/q}.$$ 

For Triebel sequence norms, we take $(\alpha,p,q)$ as above, but exclude $p = \infty$.

Set $\chi_{jk}(x) = I\{k \leq 2^j x < k + 1\}$, and now $s = \alpha + 1/2$ and define

$$\|\theta\|_{\ell^p_{s,q}} = \|\beta_{jk}\|_{\ell^p} + \left\{ \sum_{j \geq j_0} \left( 2^{jq} \sum_k |\alpha_{jk}|^{q/2} \chi_{jk} \right)^{1/q} \right\} \|L_p\|.$$

To define homogeneous forms of these norms and corresponding spaces $j^\alpha_{p,q}, \ell^p_{s,q}$, take sums over $j \in \mathbb{Z}$, and omit the sequence $\beta$. Again note that $j^\alpha_{0,p} = \ell^0_{s,p}$ and $\ell^0_{s,0} = \ell^0_2$.

The basic smoothness characterisation results say that under conditions on orthonormal scaling and wavelet functions $(\varphi, \psi)$, the representation

$$f = \sum_k \beta_{j_0,k} \varphi_{j_0,k} + \sum_{j \geq j_0} \sum_k \alpha_{jk} \psi_{jk},$$ 

where $\beta_{j_0,k} = \langle f, \varphi_{j_0,k} \rangle$, $\alpha_{jk} = \langle f, \psi_{jk} \rangle$, establishes equivalence of norms (with $\theta = ((\beta_{jk}) \cup (\alpha_{jk}))$):

(B.23) $C_1 \|\theta\|_{\ell^p_{s,q}} \leq \|f\|_{\ell^p_{s,q}} \leq C_2 \|\theta\|_{\ell^p_{s,q}}$

(B.24) $C_1 \|\theta\|_{\ell^p_{s,q}} \leq \|f\|_{\ell^p_{s,q}} \leq C_2 \|\theta\|_{\ell^p_{s,q}}$

Here $C_1 = C_1(\varphi, \psi, \alpha, p, q, j_0)$ (but: does not depend on $f$).

Using the Meyer wavelet, Lemarié & Meyer (1986) established, among other things, the equivalence for $\alpha \in \mathbb{R}$ and $1 \leq p, q \leq \infty$ for homogeneous Besov norms. This result is extended to $0 < p, q \leq \infty$ and the Triebel scale by Frazier et al. (1991, Theorem 7.20) After a discussion of numerous particular spaces, the inhomogenous Besov case is written out in Meyer (1990, Volume 1, Chapter VI.10).
If $(\varphi, \psi)$ have lower regularity (e.g. the Daubechies families of wavelets), then these characterisations hold for restricted ranges of $(\alpha, p, q)$. For example 
Meyer (1990, I, ch. VI.10), if $\varphi$ generates an $r-$regular MRA, then 
((B.23)) holds for $p, q \geq 1, |\alpha| < r$.

**B.3.3. Besov spaces via moduli of smoothness on $I \subset \mathbb{R}^n$.** A historically precedent definition of Besov spaces in terms of finite differences makes the connection with traditional notions of smoothness more explicit than the Littlewood-Paley definition. First some notations for finite differences:

$$
\Delta_h f(x) = f(x) - f(x + h) = (I - S_h) f(x), \text{ and}
$$

$$
\Delta_h^m f(x) = (I - S_h)^m f(x) = \sum_{k=0}^{m} \binom{m}{k} (-1)^k f(x + kh).
$$

Now let $I_h = \{ x \in I : x + h \in I \}$ and note that one definition of a *modulus of continuity*, can be given by

$$
\omega_1(f, t)_\infty = \sup_{h \leq t} \sup_{x \in I_h} |f(x) - f(x + h)|.
$$

In general, we have an $m$th order *modulus of smoothness*

$$
\omega_m(f, t)_p = \sup_{h \leq t} \| \Delta_h^m f \|_p
$$

(on $I$, $\| \Delta_h^m f \|_p$ means $\int_{I_{\Delta_h^m}} |\Delta_h^m f(x)|^p dx$.)

**Definition:** For $\alpha > 0$, $0 < p, q \leq \infty$, fix integer $m > \alpha$ and set

$$
\| f \|_{B^{\alpha, q}_{p, \infty}(I)} = \| f \|_p + \left( \int_0^\infty \left( \frac{\omega_m(f, t)_p}{t^\alpha} \right)^q \frac{dt}{t} \right)^{1/q}
$$

This modulus based definition is equivalent to the earlier Fourier form if $\alpha > n(p^{-1} - 1) + 0 < p, q \leq \infty$, (e.g. Triebel (1983, p. 110), [For $\alpha > 0, 1 \leq p, q \leq \infty$, see also Bergh & Löfström (1976, Th. 6.2.5)].

As an example, we may note the case $p = q = \infty, \alpha < 1$ (so choose $m = 1$):

$$
\| f \|_{B^{\alpha, \infty}_{\infty, \infty}(I)} \approx \| f \|_{\infty} + \sup_{x, y} \frac{|f(x) - f(y)|}{|x - y|^\alpha} \approx \| f \|_{C^\alpha}.
$$

**Remarks.** A frequent tool in working with integral norms such as (B.25) is dyadic discretisation: If $g : \mathbb{R}^+ \to \mathbb{R}$, and if $\alpha > 0$; then

$$
2^{-\alpha} \sum_j 2^{j\alpha} g(2^{-j}) \leq \int_0^\infty \frac{g(t)}{t^\alpha} \frac{dt}{t} \leq 2^{\alpha} \sum_j 2^{j\alpha} g(2^{-j}).
$$

Thus, an equivalent form of (B.25) is given by

$$
\| f \|_{B^{\alpha, q}_{p, \infty}(I)} \approx \| f \|_p + \left( \sum_j 2^{j\alpha} \omega_m(f, 2^{-j})_p^{q \frac{1}{q}} \right)^{1/q}.
$$

Note also that because of the presence of $\| f \|_p$, it is enough to consider the integral in (B.25) over $t \in [0, 1]$ or the sum in (B.26) over $j \geq 0$. 283
Notes:

1. Footnotes such as this will be used to explain certain simple details. Thus, if \( r(\xi) = \sum_{k=0}^{m} r_k e^{-ik\xi} \), with \( r_k \in \mathbb{R} \), then \( |r(\xi)|^2 = r(\xi)^* r(\xi) = r(\xi)r(-\xi) = \sum_{k=-m}^{m} s_k e^{-ik\xi} \) is both real and even, so \( s_{-k} = s_k \) and hence it is a polynomial of degree \( m \) in \( \cos \xi = 1 - 2 \sin^2(\xi/2) \). In addition, \(|(1 + e^{-i\xi})/2|^2 = \cos^2 \xi/2\).

2. If \( P_1, P_2 \) are degree \( p - 1 \) solutions of (B.18), then \( Q = P_1 - P_2 \) satisfies \( (1 - y)^p Q(y) + y^p Q(1 - y) \equiv 0 \), which implies that the degree \( p - 1 \) polynomial \( Q \) has \( Q^{(j)}(0) = 0 \) for \( 0 \leq j < p \) and so \( Q \equiv 0 \).
APPENDIX C

Wavelets, Smoothness and Function Classes

C.0.4. Some Heuristics. Some of the traditional measures of smoothness are based on using $L_p$ norms to measure the size of derivatives of the function: through the seminorms $\int |D^k f|^p$, where $1 \leq p \leq \infty$. When $p = \infty$, the integral is replaced by a supremum $\sup_x |D^k f(x)|$.

These seminorms vanish on polynomials of degree less than $k$, and so it is customary to add the $L_p$ norm of the function in order to obtain an actual norm. Thus the ($p^{th}$ power) of the Sobolev norm is defined by

$$\| f \|_{W^k_p} = \int |f|^p + \int |D^k f|^p.$$  

The Sobolev space $W^k_p$ of functions with $k$ derivatives existing and integrable in $L_p$, is then the (Banach) space of functions for which the norm is finite. Again, in the case $p = \infty$, the norm is modified to yield the Hölder norms

$$\| f \|_{C^k} = \| f \|_{\infty} + \| D^k f \|_{\infty}.$$  

![Figure 1](image)

Figure 1 contains some examples to illustrate how smaller $p$ corresponds to a more averaged and less worst-case measure of smoothness. For the function in the first panel, one calculates that

$$\| f' \|_1 = 2, \quad \| f' \|_2 = \sqrt{1/a + 1/b}, \quad \| f' \|_{\infty} = 1/a.$$  

In the $1$–norm the peaks have equal weight, while in the $2$–norm the narrower peak dominates, and finally in the $\infty$–norm, the wider peak has no influence at all.

285
In the second panel, we compare the norms of a function with $M$ peaks each of width $1/N$:

$$\|f'\|_1 = M, \quad \|f'\|_2 = \sqrt{MN}, \quad \|f'\|_\infty = N.$$  

The $1$-norm is proportional to the number of peaks, while the $\infty$-norm measures the slope of the narrowest peak (and so is unaffected by the number of spikes), while the $2$-norm is a compromise between the two.

**C.0.5. Wavelet Coefficients: some examples.** We are going to use wavelet coefficients to describe the local and global smoothness of functions. We already did this in the case of uniform (Hölder) smoothness—now this is extended to smoothness measures with general $p$.

We begin with some basic bounds on wavelet coefficients. The support of a function $\text{supp } f = \{x : f(x) \neq 0\}$ will be abbreviated as $S(f)$.

![Figure 2](image)

By choosing whether to take $f$ or $\psi_I$ out of the integral, one obtains two bounds for the wavelet coefficient. While both bounds are valid in general, the first is better for very coarse scales, panel (a) in Figure 2, while the second is preferred on very fine scales such as shown in panel (b).

$$\int f \psi_I \leq \int_{S(f) \cap S(\psi_I)} |f| |\psi_I| \leq \begin{cases} 2^{j/2} \|\psi\|_\infty \|f\|_1 & \text{best if } S(f) \subset S(\psi_I) \\ 2^{-j/2} \|\psi\|_1 \|f\|_{\infty, S(\psi_I)} & \text{best if } S(\psi_I) \subset S(f) \end{cases}$$

As suggested by these inequalities, it will be useful to keep track of the intersection of the support of the wavelet with the support of the function. We will say that $S(g)$ hits $A$ if $S(g) \cap A \neq \emptyset$, and otherwise that $S(g)$ misses $A$. In particular, we count the number of wavelets at level $j$ that hit $A$:

$$N_j[A] = \# \{ k : S(\psi_{jk}) \text{ hits } A \}.$$  

If $\psi$ has compact support, $S(\psi) = [-S, S]$, then $S(\psi_{jk}) = k2^{-j} + 2^{-j}[-S, S]$. For every $j$, $N_j[\{x_0\}] \approx 2S$. 

296
**Example 1.** Consider the hat function \( f(x) = (1 - |x|/\epsilon)_+ \), and for convenience set \( \epsilon = 2^{-r} \). First note that

\[
\| f \|_1 = \epsilon = 2^{-r}, \quad \| f \|_2 = \sqrt{\epsilon}.
\]

Consider first the coarse scales, when \( j < r \). We have \( N_j[S(f)] \approx 2^j \), and from inequality (a),

\[
|\theta_{jk}| = \int f \psi_{jk} \leq C 2^{j/2} \int |f| = C 2^{j/2 - r}.
\]

Turning to the fine scales, when \( j \geq r \), we have from (b)

\[
|\theta_{jk}| \leq C 2^{-j/2}.
\]

In this case, many wavelets hit the support of \( f \), and \( N_j[S(f)] \approx 2^j \cdot 2^{j-r} \). However, if \( \psi \) has at least two vanishing moments: \( \int \psi = \int x \psi = 0 \), this doesn’t matter so much; so long as \( S(\psi_{jk}) \) misses \( \{0, \pm \epsilon\} \), then \( \theta_{jk} = 0 \).

Let \( \|\theta_{j, k}\|_p = \left( \sum_k |\theta_{jk}|^p \right)^{1/p} \) denote the \( \ell_p \) norm of the coefficients at level \( j \).

Since only \( O(S) \) coefficients are non-zero at any level, our earlier calculations imply that

\[
\|\theta_{j, k}\|_p \leq \begin{cases} c 2^{j/2 - r} & j < r \\ c 2^{-j/2} & j \geq r \end{cases}
\]

Besov sequence spaces measure the rate of decay of the levelwise norms \( \|\theta_{j, k}\|_p \). For example, with \( a = 1/2 \),

\[
2^{-j} \|\theta_{j, k}\|_p \leq \begin{cases} c 2^{-(r-j)} & j < r \\ c & j \geq r \end{cases}
\]

**Example 2.** Consider now \( f(x) = |x|^\beta I\{|x| \leq 1\} \). [Include a graph??] Here \( \beta \) measures the severity of the singularity at 0. Again, suppose that \( \psi \) has compact support in \([-S, S]\). Note that \( f \) is (infinitely) differentiable away from \( \{0, \pm 1\} \). We consider two cases:

a) near 0. Making the change of variables \( y = 2^j x - k \), we calculate

\[
\theta_{jk} = \int_{-1}^1 |x|^\beta \psi_{jk}(x) dx = 2^{-j(\beta + 1/2)} \int |y + k|^\beta \psi(y) dy \approx 2^{-j(\beta + 1/2)},
\]

at least if \( |k| \leq S \), while for \( |k| > S \), \( S(\psi_{jk}) \) misses 0, and the next calculation is more appropriate.

b) away from 0. Here we use the smoothness of \( f \). If \( \psi \) has \( \alpha \) vanishing moments, so that \( \int x^j \psi(x) dx = 0 \) for \( 0 \leq j \leq \alpha \), then

\[
\int f \psi_I \leq C 2^{-j(\alpha + 1/2)} \| D^\alpha f \|_{\infty, S(\psi_I)}.
\]

Clearly, for \( x \neq 0 \) in the interior of \([-1, 1]\), we have \( |D^\alpha f(x)| = c_{\alpha, \beta} |x|^{\beta - \alpha} \). If \( k - S > 0 \), then \( S(\psi_{jk}) \) misses 0 and

\[
|\theta_{j, k+S}| \leq C 2^{-j(\alpha + 1/2)} |k-2^{-j}|^{\beta - \alpha} = C 2^{-j(\beta + 1/2)} |k|^{1-(\alpha - \beta)}.
\]

287
To summarize, then, we have in both cases
\[ |\theta_{jk}| \leq c 2^{-j(\beta+1/2)} |k|^{-\alpha-\beta}. \]
Thus, if \( \alpha > \beta + 1 \), then \( \|\theta_j\|_p \leq c 2^{-j(\beta+1/2)} \). With \( a = \alpha + 1/2 - 1/p \), we have
\[ 2^j a \|\theta_j\|_p \leq c 2^{-j(\beta - \alpha + 1/p)}, \]
and the exponent is negative if and only if \( \alpha \leq \beta + 1/p \). One might say that the smoothness of \( f \) in \( L_p \) is \( \alpha = \beta + 1/p \), showing that the smoothness index increases as \( p \) decreases away from the uniform norm toward more averaged smoothness measures.

C.0.6. Besov sequence norms. Suppose that \( f \) has wavelet expansion
\[
 f(x) = \sum_k \beta_k \phi_k(x) + \sum_{j \geq 0} \sum_k \alpha_{jk} \psi_{jk}(x).
\]
The most general definition of a Besov sequence norm on wavelet coefficients involves three indices \((\alpha, p, q)\): if we use \( \theta \) as an abbreviation for \( \{\beta_k\} \cup \{\alpha_{jk}, j \geq 0, k \in \mathbb{Z}\} \), then
\[
\|\theta\|_{B^{\alpha}_{p,q}} = \left( \sum |\beta_k|^p \right)^{1/p} + \left\{ \sum_{j \geq 0} 2^{ajq} \left( \sum_k |\alpha_{jk}|^q \right)^{q/p} \right\}^{1/q},
\]
with \( a = \alpha + 1/2 - 1/p \). The \( q = \infty \) case is given by
\[
\|\theta\|_{B^{\alpha}_{p,\infty}} = \left( \sum |\beta_k|^p \right)^{1/p} + \sup_{j \geq 0} 2^{aj} \left( \sum_k |\alpha_{jk}|^p \right)^{1/p}.
\]
With the abbreviations \( \|\beta\|_p = (\sum |\beta_k|^p)^{1/p} \) and \( \|\alpha_j\|_p = (\sum_k |\alpha_{jk}|^p)^{1/p} \), these definitions become
\[
\|\theta\|_{B^{\alpha}_{p,\infty}} = \|\beta\|_p + \left\{ \sum_{j \geq 0} 2^{ajq} \|\alpha_j\|_q \right\}^{1/q},
\]
\[
\|\theta\|_{B^{\alpha}_{p,\infty}} = \|\beta\|_p + \sup_{j \geq 0} 2^{aj} \|\alpha_j\|_p, \quad a = \alpha + 1/2 - 1/p.
\]

C.0.7. Connection with Function Classes. In what follows, we explore the connection with function classes in detail in three cases:

(i) Hölder smoothness \((p = \infty)\). We already saw that \( \|\theta\|_{B^{\infty}_{p,\infty}} \leq \|f\|_{C^\alpha} \) for \( \alpha > 0 \). [CROSS-REF] (with the Zygmund class interpretation of \( C^\alpha \) when \( \alpha \in \mathbb{N} \)) \( \|\theta\|_{B^{\infty}_{p,\infty}} = \sup\{2^{(\alpha+1/2)j} |\alpha_{jk}|, |\beta_k|\} \).

(ii) Mean-square smoothness \((p = 2)\). We will see that
\[
\|\theta\|^2_{B^{2}_{2,2}} \geq \int |f|^2 + |D^\alpha f|^2.
\]
(iii) Total variation ($p = 1$). We will show that

$$C_1 \|\theta\|_{B^1_{1,\infty}} \leq \int |f| + |Df| \leq C_2 \|\theta\|_{B^1_{1,\infty}},$$

(i.e. $\alpha = 1$ here).
C.0.8. **Total Variation and Wavelet Coefficients.** When $I = [a, b]$, this norm is defined by
\[
\|f\|_{TV(I)} = \sup \left\{ \sum_{i=1}^{n} |f(t_i) - f(t_{i-1})| : a = t_0 < t_1 < \cdots < t_n = b, n \in \mathbb{N} \right\}.
\]
It represents a scientifically interesting enlargement of $W_1^1$, since when $f \in W_1^1$, we may write
\[
\|f\|_{TV} = \int |Df|.
\] (C.2)

Suppose that $f$ has a wavelet expansion given by (C.1), and that we set $C = \max(\|\phi\|_{TV}, \|\psi\|_{TV})$. The triangle inequality yields immediately the bound
\[
\|f\|_{TV} \leq C \sum_{k} |\beta_k| + C \sum_{j} 2^{j/2} \sum_{k} |\alpha_{jk}|.
\]
and similarly
\[
\|f\|_1 \leq C \sum_{k} |\beta_k| + C \sum_{j} 2^{-j/2} \sum_{k} |\alpha_{jk}|.
\]

Conversely, what can be said about the coefficients $(\beta_k), (\alpha_{jk})$ if it is assumed that $f \in TV$? We need a preliminary remark.

C.1. **Lemma.** If $f \in TV$ and $\int \psi = 0$, then
\[
\int f \psi \leq \frac{1}{2} \|\psi\| \|f\|_{TV(I)}
\] (C.3)

**Proof.** First assume that $f$ is smooth. Since $\int \psi = 0$, we can define the primitive $\psi^{(1)}(x) = -\int_{-\infty}^{x} \psi(s)ds$ and note that $\int f \psi = -\int \psi^{(1)} f'$ implies that $|\int f \psi| \leq \|\psi^{(1)}\| \|f\|_1$. For each $x$, $2|x\psi^{(1)}(x)| = |\int_{-\infty}^{x} \psi| + |\int_{x}^{\infty} \psi| \leq \int |\psi|$, so that $\|\psi^{(1)}\|_{\infty} \leq \frac{1}{2} \|\psi\|_1$ and the result follows from (C.2). Now approximate a general $f \in TV$ by a suitable sequence of smooth functions such that both $\int f_n \psi \to \int f \psi$ and $\|f_n\|_{TV} \to \|f\|_{TV}$. \qed

We now seek an overall bound on the wavelet coefficients of $f \in TV$. By a change of variable,
\[
\alpha_{jk}(f) = \int f(x) 2^{j/2} \psi(2^j x - k)dx = 2^{-j/2} \int f(2^{-j}v) \psi(v - k)dv.
\]
Suppose that $\psi$ has support contained in the interval $[-S, S]$ and use (C.3) to obtain
\[
|\alpha_{jk}| \leq C_{\psi} 2^{-j/2} \|f\|_{TV[2^{-j}(k-S),2^{-j}(k+S)]}.
\]
Adding over $k$ yields
\[
2^{j/2} \sum_{k} |\alpha_{jk}| \leq C_{\psi}' \|f\|_{TV(\mathbb{R})}
\]
for all \( j \). An easier argument shows that, setting \( C_\phi = \sup_x \sum |\phi(x - k)| \), we also have
\[
\sum |\beta_k| \leq \int |f(x)| \sum |\phi(x - k)| dx \leq C_\phi \| f \|_1.
\]

In summary, we have shown that \( TV \) is sandwiched between these two spaces: \( b_1^{1,1} \subset TV \subset b_1^{1,\infty} \). More specifically, we have the bounds
\[
C_1 \| f \|_{b_1^{1,\infty}} \leq TV(f) + \| f \|_1 \leq C_2 \| f \|_{b_1^{1,1}}.
\]
C.0.9. **Traditional Hilbert-Sobolev spaces.** Traditional Hilbert-Sobolev space $H_r^r(\mathbb{R})$ for $r \in \mathbb{N}$ is defined as consisting of those $f \in L_2(\mathbb{R})$ for which

$$\int (D^r f)^2(x) dx < \infty$$

with $\|f\|_{H_r}^2 = \int f^2 + \int (D^r f)^2$.

Suppose that $f \in L_2(\mathbb{R})$ has wavelet expansion

$$f(x) = \sum_k \beta_k \phi_k(x) + \sum_{j \geq 0} \sum_k \alpha_{jk} \psi_{jk}(x).$$

We want to prove that, under appropriate conditions on $(\phi, \psi)$,

(C.4) \quad \|f\|^2_{B_{2,2}^r} = \sum_k \beta_k^2 + \sum_{j \geq 0} 2^{2j} \sum_k \alpha_{jk}^2$

defines an equivalent norm, in the sense that there exist constants $C_1, C_2$, depending on $r$ and the wavelet/scaling functions, but not on $f$, such that

$$C_1 \|f\|_{H_r} \leq \|f\|_{B_{2,2}^r} \leq C_2 \|f\|_{H_r}.$$  

This would justify the use of *dyadic ellipsoids* of the form

$$\{ f : \sum_k \beta_k^2 + \sum_{j \geq 0} 2^{2j} \alpha_{jk}^2 \leq C^2 \}$$

to model quantitative bounds on smoothness as measured by mean-squared integrability of derivatives. Compare Section 9.3.2.

The Besov-type norm (C.4) makes sense for any real value of $r$, so we recall the traditional extension of the Hilbert-Sobolev spaces $H^\alpha(\mathbb{R})$ for $\alpha \in \mathbb{R}$, not necessarily integer. This uses the Fourier domain

$$H^\alpha(\mathbb{R}) = \{ f : \|f\|_\alpha^2 = \int (1 + \xi^2)^\alpha |\hat{f}(\xi)|^2 d\xi < \infty \}.$$  

Since $D^r f(\xi) = -i\xi^r \hat{f}(\xi)$ and $1 + \xi^2 \leq (1 + \xi^2)^\alpha \leq c_\alpha (1 + \xi^2)$, it is clear that this is equivalent to the original definition when $\alpha \in \mathbb{N}$.

Assume that $(\phi, \psi)$ define an $r-$regular multiresolution analysis (MRA) (c.f. (Meyer 1990, p. 21)). This is the usual definition of an MRA (cf. Section B.2) augmented with the property that $\phi$ and its first $r$ derivatives have rapid decay. [Recall that a function $g$ has rapid decay if $|g(x)| \leq C_m(1 + |x|)^{-m}$ for all $m \in \mathbb{N}$.]

It can be shown from the two scale equation satisfied by $\psi$ that if $\phi$ has rapid decay along with its derivatives, then so does $\psi$ up to the same number of derivatives.

The MRA gives rise to approximation spaces $V_j$ and associated detail spaces $W_j = V_{j+1} \ominus V_j$ along with the orthogonal projections

$$P_j : L_2(\mathbb{R}) \to V_j; \quad Q_j : L_2(\mathbb{R}) \to W_j.$$
For \( f \in L_2(\mathbb{R}) \), let \( f_j = Q_j f \) for \( j \geq 0 \). Then we define, for \( \alpha \in \mathbb{R} \), the Besov-Sobolev norm

\[
\| f \|_{B_{\alpha}}^2 = \| P_0 f \|^2 + \sum_{j \geq 0} 2^{2j\alpha} \| f_j \|^2.
\]

Note that \( f_j = Q_j f = \sum_k \langle f, \psi_{jk} \rangle \psi_{jk} \) so that \( \| f_j \|^2 = \sum_k \langle f, \psi_{jk} \rangle^2 \) and so this definition agrees with that given at (C.4). The notation \( \| f \|_{B_{\alpha}} \) will, for this section, be used to abbreviate \( \| f \|_{B_{2\alpha}} \).

C.2. THEOREM. Suppose \((\phi, \psi)\) derive from an \( r \)-regular MRA, and that \( \int x^j \psi(x) dx = 0, 0 \leq j \leq r \). Then the norms \( \| f \|_{B_{\alpha}} \) and \( \| f \|_{\alpha} \) are equivalent for \( |\alpha| < r \).

Remark. The vanishing moments condition is actually superfluous: while the condition itself is not implied by the \( r \)-regular MRA assumption, Meyer (1992b, Theorem 2.4) shows that \( f_0 x^j = x^j \) for \( 0 \leq j \leq r \) and that this in fact suffices for the conclusion of the theorem.

A key tool in the proof is an inequality of 'Bernstein type' that describes bounds on \( L_2 \) norms of derivatives. The proof is deferred till after that of the theorem.

C.3. LEMMA. If \((\phi, \psi)\) comes from an \( r \)-regular MRA with \( \int x^j \psi(x) dx = 0, 0 \leq j \leq r \), then for \(-r \leq \alpha \leq r\) and all \( f \in W_j \) and \( j \geq 0 \),

\[
\| f \|_{\alpha} \leq C_\alpha 2^{j\alpha} \| f \|.
\]

PROOF OF THEOREM C.2. We begin by showing that \( \| f \|_{\alpha} \leq C \| f \|_{B_{\alpha}} \) for \( |\alpha| < r \). Write \( f = P_0 f + \sum_{j \geq 0} f_j \). From the Bernstein inequality

\[
\| P_0 f \|_{\alpha} \leq \| P_0 f \|_r \leq C \| P_0 f \|.
\]

For the sum, we need a trick to bound the cross terms \( \langle f_j, f_k \rangle \). Choose \( \delta \) so that \( |\alpha| + \delta \leq r \); then

\[
\langle u, v \rangle_{\alpha} = \int \left( 1 + \xi^2 \right)^{\frac{\alpha-\delta}{2}} \left( 1 + \xi^2 \right)^{\frac{\alpha+\delta}{2}} u(\xi) \bar{v}(\xi) d\xi \leq \| u \|_{\alpha-\delta} \| v \|_{\alpha+\delta}.
\]

Consequently

\[
\left\| \sum_{j} f_j \right\|_{\alpha}^2 \leq 2 \sum_{j \geq k} \| f_j \|_{\alpha-\delta} \| f_k \|_{\alpha+\delta} + \sum_{j} \| f_j \|_{\alpha}^2.
\]

Now use the Bernstein inequality of the Lemma C.3, \( \| f_j \|_{\alpha} \leq C 2^{j\alpha} \| f \|_j \) and for abbreviation let \( \gamma_j = 2^{j\alpha} \| f_j \|_j \). We obtain a bound of \( C^2 \) times

\[
2 \sum_{j \geq k} 2^{-\delta(j-k)} \gamma_j \gamma_k + \sum_{j} \gamma_j^2 = [\gamma, T\gamma],
\]

[DEFINE \([,]\)] where \( T_{jk} = (2^{-\delta(j-k)}) \) is a matrix of convolution with the summable sequence \( 2^{-\delta} \), so that

\[
\left\| \sum_{j} f_j \right\|_{\alpha}^2 \leq C^2 \sum_{j} \gamma_j^2 \leq C^2 \sum_{j \geq 0} 2^{2j\alpha} \| f_j \|_j^2 \leq C^2 \| f \|_{B_{2\alpha}}^2.
\]

293
The converse inequality \( \|f\|_{B,\alpha} \leq C \|f\|_{\alpha} \) will follow a simple duality argument. First, we modify the notation to put \( P_0 f = f_{-1} \) and introduce weights \( w_j = 2^j \) for \( j \geq 0 \) and \( w_{-1} = 1 \). The inner products in \( L_2(\mathbb{R}) \) and \( H^\alpha(\mathbb{R}) \) become
\[
(f, g) = \sum_{j \geq -1} (f_j, g_j), \quad \text{and} \quad \langle f, g \rangle_{B,\alpha} = \sum_{j \geq -1} w_j^\alpha (f_j, g_j).
\]
By simple transformations, we obtain
\[
\|f\|_{B,\alpha} = \sup \left\{ \langle f, g \rangle_{B,\alpha} \mid \|g\|_{B,\alpha} \leq 1 \right\}
= \sup \left\{ \sum_j w_j^\alpha (f_j, g_j) \mid \sum_j w_j^\alpha \|g_j\|^2 \leq 1 \right\}
= \sup \left\{ \sum_j (f_j, h_j) \mid \sum_j w_j^\alpha \|h_j\|^2 \leq 1 \right\}
= \sup \left\{ (f, h) \mid \|h\|_{B,-\alpha} \leq 1 \right\}.
\]
Now use (C.5) and then the result of the first half of the proof (with \(-\alpha\!\)): \( (f, h) \leq \|f\|_{\alpha} \|h\|_{-\alpha} \leq C \|f\|_{\alpha} \|h\|_{B,-\alpha} \). We therefore immediately conclude that \( \|f\|_{B,\alpha} \leq C \|f\|_{\alpha} \). \quad \Box

We prepare for the proof of Lemma C.3 with a sequence of lemmas. First the special case of \( s \) a non-negative integer.

C.4. Lemma. If \( 0 \leq s \leq r \) is an integer, then for \( f \in W_j \) or \( V_j \),
\]
\begin{equation}
(C.6) \quad \|D^s f\|_2 \leq C 2^j s \|f\|_2.
\end{equation}

Remark. The classical Bernstein inequality says that if \( \text{supp} \hat{f}(\xi) \subset [-R, R]^j \), then \( \|D^s f\|_p \leq 2^R \|f\|_p \).

Proof. We establish the results for the \( V_j \) spaces: exactly the same method works for the \( W_j \). The scale invariance property of the MRA says that \( f(x) \in V_j \) if \( g(x) = f(2^{-j} x) \in V_0 \). Since \( D^s g(x) = 2^{-js} D^s f(2^{-j} x) \), it follows that it is enough to establish (C.6) for \( f \in V_0 \). For such \( f \) we have
\[
D^s f(x) = \sum_k c_k D^s \phi(x - k),
\]
so that, after using the Cauchy-Schwartz inequality
\[
|D^s f(x)|^2 \leq \sum_k c_k^2 |D^s \phi(x - k)| \sum_k D^s \phi(x - k)|.
\]
Finiteness of \( c_\phi = \sup_x \sum_k |D^s \phi(x - k)| \) follows from the rapid decay of \( D^s \phi \). Integrating over \( x \) yields
\[
\|D^s f\|^2 \leq c_\phi \sum_k c_k^2 \|D^s \phi\|_1 \leq C \|f\|^2,
\]
since rapid decay of \( D^s \phi \) implies \( \|D^s \phi\|_1 < \infty \) and because \( \|f\|^2 = \sum c_k^2 \). \quad \Box
C. 5. Lemma. Suppose that the kernel $D(x,y)$ satisfies
\[
\int |D(x,y)|dy \leq M \quad \text{for all } x \text{ and } \int |D(x,y)|dx \leq M \quad \text{for all } y.
\]
Then the operator $D$ defined by
\[
Df(x) = \int D(x,y)f(y)dy
\]
is bounded on $L_2(\mathbb{R})$ with $\|D\|_2 \leq M$.

Proof. From the Cauchy-Schwartz inequality,
\[
|Df(x)|^2 \leq \int |D(x,y)|f^2(y)dy \int |D(x,y)|dy \leq M \int |D(x,y)|f^2(y)dy.
\]
Integrate over $x$ to find that
\[
\|Df\|_2^2 \leq M \int \int |D(x,y)|dx f^2(y)dy \leq M^2 \|f\|_2^2. \quad \square
\]

C. 6. Corollary. If $\psi$ and $\chi$ satisfy a decay condition $|\psi(x)| \leq C(1 + |x|)^{-\beta}$ for some $C < \infty$ and $\beta > 1$, then the kernel
\[
D(x,y) = \sum_k \psi(x-k)\chi(y-k)
\]
defines an $L_2$-bounded operator.

Proof. It is enough to verify conditions (C.7), and indeed, just the first of these, as the second follows by symmetry. We have
\[
\int |D(x,y)|dy \leq \sum_k |\psi(x-k)| \int |\chi(y-k)|dy \leq C_\psi \|\chi\|_1,
\]
where $C_\psi = \sup_x \sum_k |\psi(x-k)| \leq C \sum_{k=1}^{\infty} k^{-\beta}. \quad \square
\]

Now suppose in addition that the wavelet $\psi$ has $r$ vanishing moments. We deduce some regularity properties of the detail projection $Q_0$.

C. 7. Lemma. If $\psi$ has rapid decay and $\int x^j \psi(x)dx = 0$, $0 \leq j \leq r$ then the primitives $\psi^{(j)}$, $0 \leq j \leq r$ have rapid decay and satisfy $\int \psi^{(j)}(x)dx = 0$.

Proof. Consider first $\psi^{(1)}(x) := \int_{-\infty}^{x} \psi(y)dy = -\int_{x}^{\infty} \psi(y)dy$. For $x \geq 1$,
\[
x^m \psi^{(1)}(x) \leq \int_{x}^{\infty} y^{m+2} \psi(y)y^{-2}dy \leq cC_{m+2},
\]
so $\psi^{(1)}$ has rapid decay, and consequently partial integration shows for $0 \leq j \leq r - 1$ that
\[
\int x^j \psi^{(1)}(x)dx = -(j+1)^{-1} \int x^{j+1} \psi(x)dx = 0.
\]
The proof is now completed by induction. □

C.8. Corollary. If \((\phi, \psi)\) is \(r\)-regular with \(r\) vanishing moments for \(\psi\), then

\[
Q_0 = D^r H
\]

where the kernel \(H(x, y) = \sum_k \psi^{(r)}(x - k)\psi(y - k)\) is bounded on \(L_2\).

Proof. We have \(Q_0 f(x) = \sum_k \inner{f, \psi^{(r)}(x)} \psi^{(r)}(x) = \int \sum_k \psi(x - k)\psi(y - k)f(y)dy\). We write \(\psi(x - k) = D^r \psi^{(r)}(x - k)\), which yields (C.8) and the formula for \(H(x, y)\). Since both \(\psi\) and \(\psi^{(r)}\) have rapid decay, the latter from the preceding lemma, \(L_2\)-boundedness of \(H(x, y)\) follows from Corollary C.6. □

Finally, we recall the notion of Fourier multiplier. A function \(m(\xi)\) of the frequency variable \(\xi\) can be used to define a linear operator via the expression \(M f(\xi) = m(\xi) \hat{f}(\xi)\), or, more explicitly,

\[
M f(x) = (2\pi)^{-1} \int e^{ix\xi} m(\xi) \hat{f}(\xi) d\xi.
\]

Crucially, if \(|m(\xi)| \leq C_m\), then the operator \(M\) is bounded on \(L_2(\mathbb{R})\) with \(\|M\| \leq C_m\).

Proof of Lemma 3. The inequality we seek may be written more explicitly as, for \(f \in W_j\):

\[
\int (1 + \xi^2)^\sigma |\hat{f}(\xi)|^2 d\xi \leq C_\sigma 2^{2j\sigma} \int |\hat{f}(\xi)|^2 d\xi.
\]

By the rescaling device used in the proof of the Bernstein lemma C.4, this is equivalent to an inequality for \(f \in W_0\),

\[
\int (\delta_j + \xi^2)^\sigma |\hat{f}(\xi)|^2 d\xi \leq C_\sigma \int |\hat{f}(\xi)|^2 d\xi,
\]

with \(\delta_j = 2^{-2j}\). Consider first \(\sigma \geq 0\). In this case \((\delta_j + \xi^2)^\sigma \leq C_\sigma (1 + \xi^{2\sigma})\) and so the left side is bounded by \(C_\sigma^2 \|f\|^2 + \|D^r f\|^2\), which reduces our bound to the previous Bernstein result.

For \(\sigma = -t\) with \(0 < t \leq r\), introduce a smooth frequency localization function \(u(\xi)\) of support contained in \([-1, 1]\) and equal to 1 on \([-\frac{1}{2}, \frac{1}{2}]\). We write the left side of (C.9) as

\[
\int (\delta_j + \xi^2)^{-t} u(\xi) |\hat{f}(\xi)|^2 d\xi + \int (\delta_j + \xi^2)^{-t} [1 - u(\xi)] |\hat{f}(\xi)|^2 d\xi.
\]

Since \(\xi \rightarrow (\delta_j + \xi^2)^{-t}[1 - u(\xi)]\) is bounded (e.g. by \(2^{2t}\), the second integral is bounded by \(2^{2t} \|f\|^2\).
For the first integral, note that for \( f \in W_0 \), we can write \( f = Q_0 f = D^r H f \) using Corollary C.8. Thus \( |\tilde{f}(\xi)|^2 = |\xi|^{2r} |\tilde{Hf}(\xi)|^2 \). The modulus squared of the multiplier of \( \tilde{Hf}(\xi) \) satisfies

\[
(\delta_j + \xi^2)^{-l} \xi^{2r} u(\xi) \leq \xi^{2(r-l)} u(\xi) \leq 1,
\]

and, combined with \( L_2 \)-boundedness of \( H \), this shows that the first integral in (C.10) is also bounded by \( C \| f \|^2 \), which finishes the proof. \( \square \)
APPENDIX D

Besov spaces and wavelet coefficients

Let \((\phi, \psi)\) be an orthonormal scaling and wavelet function pair, complemented by boundary scaling functions and wavelets to yield an orthonormal basis for \(L_2[0, 1]\):

\[
f = \sum_k \beta_k \phi_{L,k} + \sum_{j \geq L} \sum_k \theta_{j,k} \psi_{j,k}.
\]

[More detail here!] We have made frequent use of what we have called Besov norms on the coefficients \(\beta = (\beta_k)\) and \(\theta = (\theta_j) = (\theta_{j,k})\). To be specific, for \(\alpha > 1/p - 1/2\) and \(p, q \in [1, \infty]\), define

\[
\|f\|_{\nu, \alpha}^p = \|\beta\|_p + \|\theta\|_{\nu, \alpha}^q,
\]

where, setting \(a = \alpha + 1/2 - 1/p\)

\[
\|\theta\|_{\nu, \alpha}^q = \sum_{j \geq L} \|\theta_j\|_p^q.
\]

In this chapter we justify the term 'Besov norm' by showing that these sequence norms are equivalent to standard definitions of Besov norms on functions on \(L_p(I)\).

These equivalences were first described by Lemarié & Meyer (1986) and developed in detail in Meyer (1992b, Chapters 6 - 8). for \(I = \mathbb{R}\). The Calderón-Zygmund operator methods make extensive use of the Fourier transform and the translation invariance of \(\mathbb{R}\).

The exposition here focuses on a bounded interval, for convenience \([0, 1]\), since this is needed for the white noise models of nonparametric regression. On bounded intervals, Fourier tools are less convenient, and our approach is an approximation theoretic one, inspired by Cohen et al. (2000) and DeVore & Lorentz (1993).

Outline of approach in subsequent sections. One classical definition of the Besov function norm uses a modulus of smoothness based on averaged finite differences. This is described in Section D.0.10. The module of smoothness turns out to be equivalent to the \(K\)-functional

\[
K(f, t) = \inf \{ \|f - g\|_p + t\|f^{(r)}\|_p : g \in W_p^r(I) \}
\]

which leads to the view of Besov spaces as being interpolation spaces, i.e. intermediate between \(L_p(I)\) and \(W_p(I)\).
The connection between multiresolution analyses \( \{ V_j \} \) and Besov spaces arises by comparing the \( K \)-functional at scale \( 2^{-r} k \), namely \( K(f, 2^{-r} k) \), with the approximation error
\[
e_k(f) = \| f - P_k f \|_p
\]
due to projection onto \( V_k \).

This comparison is a consequence of two key inequalities. The ‘direct’ or ‘Jackson’ inequality bounds the approximation error in terms of the \( r \)-th derivative
\[
\| f - P_k f \|_p \leq C 2^{-r} k \| f^{(r)} \|_p.
\]
Its proof uses bounds on kernel approximation, along with the key property that each \( V_j \) contains \( \mathcal{P}_{j-1} \). The ‘inverse’ or ‘Bernstein’ inequality bounds the derivatives of functions \( g \in V_k \):
\[
\| g^{(r)} \|_p \leq C 2^r k \| g \|_p.
\]

From this point, it is relatively straightforward to relate the approximation errors \( e_k(f) \) with the wavelet coefficient norms (D.2).

**D.0.10. Moduli of smoothness and Besov spaces.** This section sets out one of the classical definitions of Besov spaces, based on moduli of smoothness, and drawing on DeVore & Lorentz (1993), which contains a wealth of extra material. For more on the extensive literature on Besov spaces and the many equivalent definitions, see Peetre (1975), Triebel (1983, 1992). An expository account, limited to \( \mathbb{R} \) and \( 0 < \alpha < 1 \) is Wojtaszczyk (1997).

The definition does not explicitly use derivatives; instead it is built up from averages, in the \( L_p \) sense, of approximate derivatives given by finite differences. For \( L_p \) norms restricted to an interval \( A \), write
\[
\| f \|_p(A) = \left( \int_A |f(x)|^p dx \right)^{1/p},
\]
and, as usual, \( \| f \|_\infty(A) = \sup_{x \in A} |f(x)| \).

The first difference of a function will be written
\[
\Delta_h(f, x) = f(x + h) - f(x) = (T_h - I) f(x),
\]
where \( T_h f(x) = f(x + h) \) denotes translation by \( h \). Higher order differences, for \( r \in \mathbb{N} \), are given by
\[
\Delta^r_h(f, x) = (T_h - I)^r f(x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x + kh).
\]

To describe sets over which averages of differences can be computed, we need the (one sided) erosion of \( A \): set \( A_h = \{ x \in A : x + h \in A \} \). The main example: if \( A = [a, b] \), then \( A_h = [a, b - h] \).

The \( r \)-th integral modulus of smoothness of \( f \in L_p(A) \) is then
\[
\omega_r(f, t)_p = \sup_{0 \leq h \leq t} \| \Delta^r_h(f, \cdot) \|_p(A_h).
\]
For $p < \infty$, this is a measure of smoothness averaged over $A$; the supremum ensures monotonicity in $t$.

The differences $\Delta_h^r(f, x)$ are linear in $f$, and so for $p \geq 1$, there is a triangle inequality

(D.3) \[ \omega_r(f + g, t)_p \leq \omega_r(f, t)_p + \omega_r(g, t)_p. \]

Again by linearity, $\|\Delta_h^r(f, \cdot)\|_p \leq 2^r \|f\|_p$ and so also

(D.4) \[ \omega_r(f, t)_p \leq 2^r \|f\|_p. \]

For $n \in \mathbb{N}$ and $1 \leq p \leq \infty$ \[ \longrightarrow \text{exercise!}] ,

(D.5) \[ \omega_r(f, nt)_p \leq n^r \omega_r(f, t)_p. \]

[Exercise: Show by induction on $r$ that

\[ \Delta^r_{nh}(f, x) = \sum_{k_1=0}^{n-1} \cdots \sum_{k_r=0}^{n-1} \Delta^r_h(f, x + k_1 h + \cdots + k_r h). \]

From this, deduce (D.5).]

When derivatives exist, the finite difference can be expressed as a kernel smooth of bandwidth $h$ of these derivatives.

D.1. Lemma. \textit{Let} $\chi$ \textit{be the indicator of the unit interval} $[0, 1]$, \textit{and} $\chi^*$ \textit{be its $r$th convolution power}. \textit{Then, for} $f \in W^r_p$,

(D.6) \[ \Delta^r_h(f, x) = h^r \int f^{(r)}(x + hu)\chi^*(u)du. \]

\textbf{Proof.} Via induction, with the case $r = 1$ being just

\[ f(x + h) - f(x) = h \int_0^1 f'(x + hu)du. \]

For general $r \in \mathbb{N}$, apply this last to $\Delta^{-1}_h f$ and note that differentiation commutes with $\Delta^{-1}_h$ to find

\[ \Delta^r_h(f, x) = h \int \Delta^{-1}_h(f', x + hu)\chi(u)du. \]

Now apply the induction hypothesis. \qed

A simple consequence of (D.6) is the bound

(D.7) \[ \omega_r(f, t)_p \leq \|f\|_{W^r_p(I)}, \]

valid for all $t \geq 0$. Indeed, rewrite the right side of (D.6) as $h^r \int K(x, v)f^{(r)}(v)dv$, using the kernel

\[ K(x, v) = h^{-1}\chi^*(h^{-1}(v - x)) \]
for \( x \in I_k \) and \( v = x + h u \in I \). Now apply Young’s inequality (D.33), and note that both \( M_1 \) and \( M_2 \leq 1 \) since \( \chi^*(\cdot) \) is a probability density. Hence

\[
\| \Delta_h^v(f, \cdot) \|_{p(H)} \leq h^q |f|_{W^r_h(I)},
\]

and the result follows from the definition of \( \omega_r \).

**Material from Woytaszyck?**

**Besov spaces.** Let \( \alpha > 0 \) and \( r = |\alpha| \). Let \( A = \mathbb{R} \), or an interval \([a, b]\). The Besov space \( B^\alpha_{p,q}(A) \) is the collection of \( f \in L_p(A) \) for which the seminorm

\[
|f|_{B^\alpha_{p,q}} = \begin{cases} 
\left( \int_0^\infty \left( \frac{\omega_r(f, t)}{t^{\alpha}} \right)^q \frac{dt}{t} \right)^{1/q} & 0 < q < \infty \\
\sup_{t > 0} t^{-\alpha} \omega_r(f, t) & q = \infty
\end{cases}
\]

is finite. The seminorm vanishes if \( f \) is a polynomial of degree less than \( r \). As norm on \( B^\alpha_{p,q}(A) \), we take

\[
\| f \|_{B^\alpha_{p,q}} = \| f \|_p + |f|_{B^\alpha_{p,q}}.
\]
**D.0.11. Besov spaces as interpolation spaces.** This section shows that Besov spaces are *intermediate* spaces between $L_p(I)$ and $W_p^r(I)$. First we need the notion of *K-functional*, reminiscent of a smoothness penalized approximation:

$$K(f, t) = K(f, t; L_p, W_p^r) = \inf \{ \| f - g \|_p + t\| D^r g \|_p : g \in W_p^r \}.$$  

The main fact about $K(f, t)$ for us is that it is equivalent to the $r$th modulus of smoothness $\omega_r(f, t)_p$—see Theorem D.3 below.

First some elementary remarks about $K(f, t)$. Since smooth functions are dense in $L_p$, it is clear that $K(f, 0) = 0$. But $K(f, t)$ vanishes for all $t > 0$ if and only if $f$ is a polynomial of degree at most $r - 1$.

Since $K$ is the pointwise infimum of a collection of increasing linear functions, it is itself increasing and concave. Further, for any $f$

$$K(f, t) \geq \min(t, 1)K(f, 1),$$

while if $f \in W_p^r$ then by choosing $g$ equal to $f$ or 0 as $t \leq 1$ or $t > 1$,

$$K(f, t) \leq \min(t, 1)\| f \|_{W_p^r}.$$  

A sort of converse to (D.10) will be useful. Let $f \in L_p$ and $g \in W_p^r$ be given. Pick $c \in I$ and let $T_{-1}g$ be the Taylor polynomial approximation of degree $r - 1$ to $g$ at $c$. From the definition of $K$ and Lemma D.14,

$$K(f, t) \leq \| f - T_{r-1}g \|_p \leq \| f - g \|_p + \| g - T_{r-1}g \|_p \leq \| f - g \|_p + C\| D^r g \|_p,$$

where $C = C(I, r)$. Hence, for all $t \geq a$,

$$K(f, t) \leq \max(Ca^{-1}, 1)K(f, a).$$

The $K$–function $K(f, t)$ trades off between $L_p$ and $W_p^r$ at scale $t$. Information across scales can be combined via various weighting functions by defining, for $0 < \theta < 1$,

$$\rho(f)_{\theta, a} = \left( \int_0^{\infty} \left[ \frac{K(f, t)}{t^{\theta}} \right]^{\frac{1}{\theta}} dt \right)^{1/q} \quad 0 < q < \infty$$

and

$$\rho(f)_{\theta, \infty} = \sup_{0 \leq t \leq \infty} \frac{K(f, t)}{t^{\theta}}.$$

Replacing $K(f, t)$ by $\min(1, t)$ in the integral (D.13) leads to the sum of two integrals $\int_0^1 t^{1-\theta} dt$ and $\int_1^{\infty} t^{-\theta} dt$, which both converge if and only if $0 < \theta < 1$. Hence property (D.10) shows that in order for $\rho(f)_{\theta, a}$ to finite for any $f$ other than polynomials, it is necessary that $0 < \theta < 1$.

On the other hand, property (D.11) shows that

$$\rho(f)_{\theta, a} \leq \omega_q\| f \|_{W_p^r}.$$
We therefore define intermediate, or interpolation spaces

\[ X_{\theta,q} = (L_p, W^r_p)_{\theta,q} = \{ f \in L_p : \rho(f)_{\theta,q} < \infty \} \]

for \( 0 < q \leq \infty \) and \( 0 < \theta < 1 \), and set \( \| f \|_{X_{\theta,q}} = \| f \|_p + \rho(f)_{\theta,q} \).

From the definition and (D.14),

\[ W^r_p \subset (L_p, W^r_p)_{\theta,q} \subset L_p. \]

The parameters \((\theta, q)\) yield a lexicographic ordering: (Exercise??)

\[ X_{\theta_1,q_1} \subset X_{\theta_2,q_2}, \quad \text{if } \theta_1 > \theta_2, \text{ or if } \theta_1 = \theta_2 \text{ and } q_1 \leq q_2. \]

The main reason for introducing interpolation spaces here is that they are in fact Besov spaces.

D.2. THEOREM. For \( r \in \mathbb{N} \), and \( 1 \leq p \leq \infty, \ 0 < q \leq \infty, 0 < \alpha < r, \)

\[ (L_p, W^r_p)_{\alpha/r, q} = B^\alpha_{p,q}. \]

This follows from the definitions and the next key theorem, which shows that the \( K \)-function is equivalent to the integral modulus of continuity.

D.3. THEOREM (Johnen, ref). Let \( A = \mathbb{R}, \mathbb{R}_+, \Theta \) or \([0,1] \). For \( 1 \leq p \leq \infty \), and \( r \in \mathbb{N} \), there exist \( C_1, C_2 > 0 \) depending only on \( r \), such that for all \( f \in L_p \),

\[ (D.15) \quad C_1 \omega_r \omega_r(f, t)_p \leq K(f, t'; L_p, W^r_p) \leq C_2 \omega_r \omega_r(f, t)_p, \quad t > 0. \]

PROOF. We work on the left inequality first: from the triangle inequality (D.3) followed by (D.4) and derivative bound (D.7), we have for arbitrary \( g \),

\[ \omega_r(f, t)_p \leq \omega_r(f - g, t)_p + \omega_r(g, t)_p \leq 2 \| f - g \|_p + t^r \| g \|_{W^r_p}. \]

Minimizing over \( g \), the right side is bounded by \( 2^r K(f, t') \).

For the right inequality, we only give full details for \( A = \mathbb{R} \).

\[ (D.16) \quad K(f, t) \leq \| f - g \|_p + t^r \| g \|_{W^r_p}, \]

and we will take in particular

\[ (D.17) \quad g(x) = f(x) + (-1)^{r+1} \int \Delta_{t^r}(f, x) \ell^{*_r}(u) du. \]

By the Minkowski integral inequality (E.13),

\[ \| g - f \|_p \leq \int \| \Delta_{t^r}(f, \cdot) \|_p \ell^{*_r}(u) du \leq \omega_r(f, rt)_p \leq r^r \omega_r(f, t)_p, \quad (D.18) \]

where the second inequality follows because \( \ell^{*_r} \) is a probability density supported on \([0, r]\), and the third uses (D.5).
We turn to estimating \( \|g^{(r)}\|_p \). Use the series expansion of \( \Delta_{kt}^{(r)}(f, x) \) and differentiate (D.17) to get

\[
g^{(r)}(x) = \sum_{k=1}^{r} \binom{r}{k} (-1)^{k+1} \int f^{(r)}(x + ktu) \chi_r(u) du
\]

\[
= \sum_{k=1}^{r} \binom{r}{k} (-1)^{k+1} (kt)^{-r} \Delta_{kt}^{(r)}(f, x),
\]

where the second equality uses identity (D.6). Again using (D.5), we find

\[
\|g^{(r)}\|_p \leq \sum_{k=1}^{r} \binom{r}{k} k^{-r} \omega_r(f, kt) \leq 2^r \omega_r(f, t)_p.
\]

Putting this last inequality and (D.18) into (D.16) yields the right hand bound.

If \( A = [0, 1] \), then \( g \) is defined in (D.17) for \( x \in I_1 = [0, 3/4] \) if \( t \leq 1/4r^2 \). By symmetry, one can make an analogous definition and argument for \( I_2 = [1/4, 1] \). One patches together the two subinterval results, and takes care separately of \( t > 1/4r^2 \). For details see ( DeVore & Lorentz 1993, p. 176, 178).

For work with wavelet coefficients, we need a discretized version of these measures.

D.4. Lemma. Let \( L \in \mathbb{N} \) be fixed. With constants of proportionality depending on \( I, r, \theta, q \) and \( L \) but not on \( f \),

\[
\rho(f) \omega_q \geq \sum_{j=L}^{\infty} [2^{\theta j} K(f, 2^{-r j})]^q.
\]

Proof. Since \( K(f, t) \) is concave in \( t \) with \( K(f, 0) = 0 \), we have \( \epsilon K(f, t) \leq K(f, \epsilon t) \), and since it is increasing in \( t \), we have for \( 2^{-r(j+1)} \leq t \leq 2^{-r j} \),

\[
2^{-r} K(f, 2^{-r j}) \leq K(f, 2^{-r(j+1)}) \leq K(f, t) \leq K(f, 2^{-r j}).
\]

From this it is immediate that, with \( a = 2^{-r L} \), the sum \( S_L(f) \) in (D.19) satisfies

\[
S_L(f) \approx \int_0^a \left[ \frac{K(f, t)}{t^\theta} \right] \frac{q dt}{t}
\]

with constants of proportionality depending only on \( (\theta, q, r) \). From (D.12),

\[
\int_0^\infty \left[ \frac{K(f, t)}{t^\theta} \right] \frac{q dt}{t} \leq C[K(f, a)] a^{-\theta/q}
\]

where \( C \) depends on \( (I, L, r, \theta, q) \). With \( a = 2^{-r L} \), this last term can be absorbed in the sum \( S_L(f) \), completing the proof. \( \square \)
D.0.12. Approximation Properties of Kernels and MRA’s. We first look at the approximation power of a family of kernels $K_h(x, y)$. Let $I \subset \mathbb{R}$ be an interval - typically $I = [0, 1]$ or $\mathbb{R}$ itself. Define

$$K_h f(x) = \int_I K_h(x, y)f(y)dy$$

for $x \in I$.

Suppose that $K_h(x, y)$ satisfies

1. $K_h \pi = \pi$ for $\pi \in \mathcal{P}_{r-1}$,
2. $K_h(x, y) = 0$ if $|y - x| > Lh$,
3. $|K_h(x, y)| \leq Mh^{-1}$.

The key requirement is that $K_h$ preserve polynomials of degree at most $r - 1$. Assumption (ii) could be weakened to require sufficient decay of $K_h$ as $|x - y|$ grows.

If $K_h(x, y) = h^{-1}K(h^{-1}(x - y))$ is a scaled translation invariant kernel on $\mathbb{R}$, then condition (i) is equivalent to the vanishing moment property

$$\int e^{ik(x-y)}K(y)dy = \delta_{k0}$$

for $k = 0, 1, \ldots, r - 1$. If $K(y)$ is bounded and has compact support, then properties (ii) and (iii) are immediate.

In the proof to follow, $\|f\|_p = (\int_I |f|^p)^{1/p}$ is the $L_p$ norm on $I$.

D.5. Proposition. Suppose that the kernel $K_h(x, y)$ satisfies (i)-(iii) on an interval $I \subset \mathbb{R}$. For $p \geq 1$, there exists a constant $C = C(L, M, r, p)$ such that for $f \in W^r_p(I)$,

$$\|f - K_h f\|_p \leq C h^r \|D^r f\|_p, \quad h > 0.$$

Proof. A function $f \in W^r_p(I)$ has continuous derivatives of order $k = 0, 1, \ldots, r - 1$. If $x \in I$, we may therefore use the Taylor approximation to $f$ at $x$ by a polynomial $\pi_x$ of degree $r - 1$, so that $f(y) = \pi_x(y) + R_x(y)$ with the integral form of the remainder term

$$R_x(y) = \int_x^y (D^r f)(u)\frac{(y-u)^{r-1}}{(r-1)!}du.$$

Since $K_h$ leaves such polynomials invariant, $K_h f = \pi_x + K_h R_x$, and since $\pi_x(x) = f(x)$, we have (with $c_r = 1/(r-1)!$),

$$(K_h f)(x) - f(x) = \int_I K_h(x, y)R_x(y)dy$$

$$= c \int\int K_h(x, y) \int_x^y (y-u)^{r-1} f^{(r)}(u)du dy$$

$$= \int_I K_h(x, u)f^{(r)}(u)du,$$

where $\tilde{K}_h(x, u)$ is a new kernel on $I \times I$, about which we need only know a bound, easily derived from the above, along with conditions (ii) and (iii):

$$|\tilde{K}_h(x, u)| \leq \begin{cases} cMh^{-1}(Lh)^r & \text{if } |x - u| \leq Lh \\ 0 & \text{otherwise} \end{cases}.$$
Since \( \int |\tilde{K}_0(x,u)|du \leq 2cL^{r+1}Mh^r \), with a similar bound for the corresponding integral over \( x \in I \), our result follows from Young’s inequality (D.33) with \( M_1 = M_2 = 2cL^{r+1}Mh^r \). \( \square \)

**Remark.** Polynomials on \([a, b]\). For \( I = [a, b] \) a compact interval, let \( \{p_k, k = 0, \ldots, r - 1\} \) be orthonormal polynomials of exact degree \( k \). The operation of orthogonal projection onto polynomials of degree at most \( r - 1 \) on the interval \( I \) is given by \( \Pi_{r-1}f \), where \( \Pi_{r-1} \) has kernel

\[
\Pi_{r-1}(x, y) = \sum_{k=0}^{r-1} p_k(x)p_k(y) \quad x, y \in I.
\]

Proposition D.5 applies to this kernel also: assumptions (i) - (iii) are satisfied if we set \( h = 1, L = b - a \) and for \( M \) we derive a bound depending only on \( r \) and \( b - a \). [An explicit value for \( M \) follows by expressing \( p_k \) in terms of the Legendre polynomials \( P_n(x) \) on \([-1,1]\) which satisfy \( |P_n(x)| \leq 1 \) there.]

Consequently, for \( f \in W^r_p(I) \),

\[
\|f - \Pi_{r-1}f\|_p \leq C\|f^{(r)}\|_p
\]

with \( C = C(r, b - a) \).

To apply this to wavelet systems, we make the following assumption:

**(VMe)** Suppose that \((\psi, \psi)\) is an orthonormal scaling function and wavelet pair, with support \([-L + 1, L]\). Suppose that \( \psi \) has \( r \) vanishing moments \( \int t^k\psi(t)dt = 0 \), for \( k = 0, 1, \ldots, r - 1 \).

Let \( \{V_j\} \) be the associated MRA, either for \( L_2(\mathbb{R}) \) or for \( L_2[0,1] \) via the (extended) CDV construction. Let \( \{\phi_{jk}\} \) be the associated orthonormal basis for \( V_j \). The orthogonal projection operators \( P_j : L_2(I) \to V_j \) have associated kernels

\[
E_j(x, y) = \sum_k \phi_{jk}(x)\phi_{jk}(y).
\]

Indeed

\[
P_jf(x) = \sum_k \langle f, \phi_{jk} \rangle \phi_{jk}(x) = \int \sum_k \phi_{jk}(x)\phi_{jk}(y)f(y)dy
= \int E_j(x, y)f(y)dy.
\]

We claim that assumptions (i)-(iii) hold for the kernel \( E_j \) with \( h \) taken as \( 2^{-j} \). First, that \( E_j \) preserves polynomials up to degree \( r - 1 \) follows from the vanishing moments condition. For \( L_2(\mathbb{R}) \), this is (Mallat 1999, Theorem 7.4) [Modify App. B. to include this!]. For \( L_2[0,1] \), this is a key property of the extended CDV construction.

That \( |E_j(x, y)| \leq M2^j \) follows from the scaling properties of \( \phi_{jk} \) along with their compact support. Finally, if \( \text{supp } \phi \subset [-L + 1, L] \), then \( \phi(x -
\( k \phi(y - k) = 0 \) unless \(|x - y| \leq 2L - 1\), which suffices, along with scaling, for property (ii).

**D.6. Corollary.** Suppose that \( \{V_j\} \) is an MRA for \( L_2(I) \), with \( I = \mathbb{R} \) or \([0,1]\), derived from a pair \((\phi, \psi)\) satisfying assumption \((V)M_r\). Let \( P_j \) be the associated orthogonal projection onto \( V_j \). Then there exists a constant \( C = C(L,M,p,r) \) such that for all \( f \in W_p^r(I) \),

\[
\|f - P_j f\|_p \leq C2^{-rj} |f|_{W_p^r}.
\]

**Remark.** For \(1 \leq p \leq \infty\), Young’s inequality (D.33) shows that \( P_j \) is a bounded operator on \( L_p(I) \).

(a). If \( \phi \) is an orthonormal scaling function and \(|\text{supp}\phi| \leq B\), set \( a_1 = \|\phi\|_1 \) and \( a_\infty = B\|\phi\|_\infty \).

(b). If \( \{\gamma_j\} = \{\phi_j\} \) correspond to a CDJV boundary MRA for \([0,1]\) derived from a scaling function \( \phi \) with \( \text{supp}\phi \subset [-L + 1,L] \), then set

\[
\begin{align*}
 a_1 &= \max \{\|\phi\|_1, \|\phi_k^0\|_1, \|\phi_k^1\|_1, k \in \mathcal{K}\}, \\
 a_\infty &= 2L \max \{\|\phi\|_\infty, \|\phi_k^0\|_\infty, \|\phi_k^1\|_\infty, k \in \mathcal{K}\}.
\end{align*}
\]

In either case, we find that

\[
\int |E_j(x,y)| dy \leq a_1 2^{-j/2} \sum_k |\phi_j^k(x)| \leq a_1 a_\infty,
\]

and similarly \( \int |E_j(x,y)| dx \leq a_1 a_\infty \). Consequently, Young’s inequality implies that

\[
(P.22) \quad \|P_j\|_p \leq a_1 a_\infty.
\]

A similar argument applied to \( E_{L-1}(x,y) = \Pi_{r-1}(x,y) \) given by (D.20) shows that

\[
(P.23) \quad \|P_{L-1}\|_p \leq C.
\]

D.7. Lemma. Let \( \{\gamma_{jk}, k \in K\} \) be an orthonormal sequence of functions satisfying

\[
\begin{align*}
(i) & \quad \sum_{k} |\gamma_{jk}(x)| \leq a_{\infty}2^{j/2}, \quad \text{and} \\
(ii) & \quad \max_{k} \int |\gamma_{jk}| \leq a_{1}2^{-j/2}.
\end{align*}
\]

Then for all \( 1 \leq p \leq \infty \), there exist constants \( C_1 = a_{\infty}^{-1}(a_{1}/a_{\infty})^{1/p} \) and \( C_2 = a_{\infty}(a_{1}/a_{\infty})^{1/p} \) such that for any sequence \( \lambda = (\lambda_k, k \in K) \),

\[
C_12^{j(1/2-1/p)}\|\lambda\|_p \leq \left\| \sum_{k} \lambda_k \gamma_{jk} \right\|_p \leq C_22^{j(1/2-1/p)}\|\lambda\|_p.
\]

Remarks. 1. If \( \phi \) is an orthonormal scaling function and \( \gamma_{jk}(x) = 2^{j/2}\phi(2^{j}x - k) \) for \( k \in \mathbb{Z} \), and \( |\text{supp } \phi| \leq B \), then (i) and (ii) are trivially satisfied with \( a_{\infty} = B\|\phi\|_{\infty} \) and \( a_{1} = \|\phi\|_{1} \).

2. If \( \{\gamma_{jk}\} = \{\phi_{jk}\} \) correspond to a CDJV boundary MRA for \([0, 1]\) derived from a scaling function \( \phi \) with \( |\text{supp } \phi| \leq [L + 1, L] \), then (i) and (ii) hold with

\[
a_{\infty} = 2L \max\{\|\phi\|_{\infty}, \|\phi_{k}^{0}\|_{\infty}, \|\phi_{k}^{1}\|_{\infty}, k \in K\},
\]

\[
a_{1} = \max\{\|\phi\|_{1}, \|\phi_{k}^{0}\|_{1}, \|\phi_{k}^{1}\|_{1}, k \in K\}.
\]

3. The right hand inequality in (D.24) does not require the assumption of orthonormality for \( \{\gamma_{jk}\} \).

Proof. This is really just the extended Young inequality, Theorem D.12. Identify \( \mu(dx) \) with Lebesgue measure on \( \mathbb{R} \) and \( \nu(dy) \) with counting measure on \( k \in K \). Then match \( K(x, y) \) with \( \gamma_{jk}(x) \) and \( f(y) \) with \( \lambda_{k} \). Conditions (i) and (ii) imply that \( M_1 = a_{1}2^{-j/2} \) and \( M_2 = a_{\infty}2^{j/2} \) suffice for the conditions of the theorem. The right hand inequality above now follows from (D.33). Note that orthonormality of \( \{\gamma_{jk}\} \) is not used.

For the left hand inequality, assume that \( g(x) = \sum_{k} \lambda_{k} \gamma_{jk} \). Since the \( \{\gamma_{jk}\} \) are orthonormal,

\[
(K^{*}g)_{k} = \int \gamma_{jk}(x)g(x)dx = \lambda_{k}
\]

and now the result follows from (D.34). \( \square \)

Now to the variant of the Bernstein inequality that we need.

D.8. Lemma. Conditions on MRA For \( g \in V_{j} \) and \( 1 \leq p \leq \infty \),

\[
\|D^{r}g\|_{p} \leq C\|g\|_{n^{p}},
\]

308
PROOF. Since \( g \in V_j \), it has an expansion \( g = \sum \lambda_k \phi_{jk} \), and so
\[
D^r g = \sum \lambda_k D^r \phi_{jk} = 2^{jr} \sum \lambda_k \gamma_{jk},
\]
where the functions \( \gamma_{jk} \) are formed from the finite set \( \{ D^r \phi, D^r \phi_0^k, D^r \phi_k \} \) by exactly the same set of linear operations as used to form \( \phi_{jk} \) from the set \( \{ \phi, \phi_0^k, \phi_k \} \).

Since the \( \{ \phi_{jk} \} \) system satisfy the conditions (i) and (ii) of Lemma D.7, the same is true of the \( \{ \gamma_{jk} \} \) system. From the right side of that Lemma,
\[
\| D^r g \|_p = 2^{jr} \sum \lambda_k \gamma_{jk} \leq c_2 2^{jr} 2^{(1/2 - 1/p)} \| \lambda \|_p.
\]
Now apply the left side of the same Lemma to the (orthogonal) \( \{ \phi_{jk} \} \) system to get
\[
\| D^r g \|_p \leq c_2 c_1 2^{jr} \| \sum \lambda_k \phi_{jk} \|_p = c_2 c_1 2^{jr} \| g \|_p,
\]
as required. \( \square \)

D.0.14. Approximation Spaces and Besov Spaces. This section relates the approximation properties of a multiresolution analysis to the behaviour of the \( K \)-functional near 0. Specifically, let the approximation error of a function \( f \in W^r_p(I) \) by its orthogonal projection \( P_k f \) onto the space \( V_k \) be given by
\[
e_k(f) = \| f - P_k f \|_p.
\]
We will show that the rate of decay of \( e_k(f) \) is comparable to that of \( K(f, 2^{-r}k) \). To the spaces \( V_k, k \geq l \) of the multiresolution analysis, we adjoin \( V_{L-1} \) consisting of polynomials of degree at most \( r - 1 \).

D.9. Theorem. Let \( r \in \mathbb{N} \) be given. For \( 1 \leq p \leq \infty, 0 < q < \infty \) and \( 0 < \alpha < r \). With constants depending on ???, but not on \( f \), we have
\[
\sum_{L=1}^\infty [2^{ak} e_k(f)]^q = \sum_{L=1}^\infty [2^{ak} K(f, 2^{-r}k)]^q.
\]

PROOF. 1°. We first show the main inequality: for \( k \geq L-1 \)
\[
C_1 e_k(f) \leq K(f, 2^{-kr}) \leq C_2 \sum_{j=L-1}^k 2^{-(k-j)r} e_j(f).
\]
For the left hand inequality, let \( f \in L_p \) and \( g \in W^r_p \) be fixed. Write \( f - P_k f \) as the sum of \( (I - P_k)(f - g) \) and \( g - P_k g \), so that
\[
e_k(f) \leq \|(I - P_k)(f - g)\|_p + e_k(g).
\]
It follows from (D.22) that \( \|I - P_k\|_p \leq 1 + a_1 a_\infty \). Together with Corollary D.6 and (D.21), this yields \( e_k(f) \leq C\|f - g\|_p + 2^{-rk} \|g\|_{W^r_p} \).

Minimizing now over \( g \) yields the left side of (D.26).

309
For the right inequality, set $\psi_j = P_j f - P_{j-1} f$ for $j \geq L$. Clearly $\|\psi_j\|_p \leq e_j(f) + e_{j-1}(f)$ and

$$P_k f = \sum_{j=L}^{k} \psi_j + P_{L-1} f.$$ 

Now $P_{L-1} f$ is a polynomial of degree at most $r - 1$, so $|P_{L-1} f|_{W^r_*} = 0$. For the other terms, apply the Bernstein inequality Lemma D.8 to obtain

$$|P_k f|_{W^r_*} \leq \sum_{j=L}^{k} |\psi_j|_{W^r_*} \leq c_2 \sum_{j=L}^{k} 2^{rj} \|\psi_j\|_p \leq c_2 \sum_{j=L}^{k} 2^{rj} [e_{j-1}(f) + e_j(f)].$$

Finally, put this into the $K$–function definition:

$$K(f, 2^{-kr}) \leq \|f - P_k f\|_p + 2^{-kr} |P_k f|_{W^r_*}$$

$$\leq (1 + 2^{r-1} c_2) \sum_{j=L-1}^{k} 2^{-(k-j)r} e_j(f).$$

2°. The left to right inequality in (D.25) is immediate from (D.26). For the other inequality, let $b_k = 2^{rk} f_k(f)$ for $k \geq L - 1$ and 0 otherwise, and $a_k = 2^{rk} K(f, 2^{-rk})$. Then bound (D.26) says that for $k \geq L - 1$,

$$a_k \leq c_2 \sum_{j=L-1}^{k} 2^{-(k-j)(r-\alpha)} b_j = \sum_{j=L-1}^{\infty} a_{k-j} b_j$$

for $a_k = C_2 2^{-k(r-\alpha)} I\{k \geq 0\}$. Young’s inequality (D.33) applied to $\mu = \nu =$counting measure on $\{L - 1, L, \ldots\}$ implies that $\|c\|_q \leq c_{r\alpha} C_2 \|b\|_q$ and completes the proof. □

**D.0.15. Wavelet coefficients, finally.** The last step in this chain is now quite easy, namely to relate seminorms on wavelet coefficients to those on approximation errors. Let $Q_j$ be orthogonal projection onto the details space $W_j$, thus $Q_j = P_{j+1} - P_j$. Suppose that for fixed $j$, $\{\psi_{jk}\}$ is the orthonormal basis for $W_j$ so that

$$Q_j f = \sum_k \theta_{jk} \psi_{jk}, \quad \theta_{jk} = \langle f, \psi_{jk} \rangle.$$ 

Under conditions on the multiresolution analysis, conditions (i) and (ii) of Lemma D.7 will hold for $\gamma_{jk} = \psi_{jk}$ as specified there in Remark 2. Consequently, Lemma D.7 implies

$$\|Q_j f\|_p \geq 2^j (1/2^{1/p}) \|\theta_j\|_p,$$

where $\|\theta_j\|_p$ denotes the $\ell_p$–norm of $(\theta_{jk})$.

The same argument applies also to the projection onto $V_L$, given by $P_L f = \sum_k \beta_k \phi_{Lk}$ to show that, with $\beta = (\beta_k)$,

$$\|P_L f\|_p \geq 2^L (1/2^{1/p}) \|\beta\|_p.$$
Setting \( a = \alpha + 1/2 - 1/p \), the equivalence (D.27) implies that
\[
\sum_{j \geq L} [2^{\alpha j} \| \theta_j \|_p]^q = \sum_{j \geq L} [2^{\alpha j} \| Q_j f \|_p]^q.
\]
Hölder’s inequality yields a second consequence of (D.27) for later use:
\[
\sum_{j \geq L} \| Q_j f \|_p \geq \sum_{j \geq L} 2^{-\alpha j} \cdot 2^{\alpha j} \| \theta_j \|_p \leq c_\alpha \| \theta \|_{b^{\alpha \cdot \cdot}}.
\]
We now argue that
\[
\sum_{j \geq L} [2^{\alpha j} \| Q_j f \|_p]^q \geq \sum_{j \geq L} [2^{\alpha j} e_j(f)]^q.
\]
For convenience, let \( \delta_k = \| Q_k f \|_p \) and \( e_k = e_k(f) = \| f - P_k f \|_p \). Clearly \( \delta_k \leq e_k + e_{k+1} \), which suffices for one of the inequalities. On the other hand, \( f - P_j f = \sum_{k \geq j} Q_k f \), and so \( e_j \leq \sum_{k \geq j} \delta_k \), or equivalently
\[
2^{\alpha j} e_j \leq \sum_{k \geq j} 2^{-\alpha(k-j)} 2^{\alpha k} \delta_k.
\]
The other inequality now follows from Young’s inequality (D.33) applied to \( \mu = \nu = \text{counting measure on } \{L, L + 1, \ldots\} \) with kernel \( K(j, k) = 2^{-\alpha(k-j)} I\{k \geq j\} \).
D.0.16. Summary.

D.10. Theorem. Let $A = [0, 1]$ and $1 \leq p, q \leq \infty$ and $\alpha > \max\{1/p - 1/2, 0\}$. Suppose that the Besov function space norm $\|f\|_{B^\alpha_{p,q}}$ is defined by (D.9), and the Besov sequence norm $\|f\|_{\ell^p_{\ell^q}}$ by (D.1). Then the two norms are equivalent: there exist constants $C_1, C_2$ depending on $\alpha, p, q$ and the functions $(\phi, \psi)$, but not on $f$ so that

$$C_1 \|f\|_{\ell^p_{\ell^q}} \leq \|f\|_{B^\alpha_{p,q}} \leq C_2 \|f\|_{\ell^p_{\ell^q}}.$$  

Proof. This assembles the steps carried out in earlier sections. We combine the definition of the Besov seminorm (D.8), the equivalence of modulus and $K$–functional (D.15), the dyadic discretization (D.19) and the $(\theta, q)$–equivalence of $K$–functional and MRA-approximation errors (D.25) to find

$$\|f\|_{B^\alpha_{p,q}}^q = \int_0^\infty \left[ \frac{\omega_r(f, s)_p}{s^\alpha} \right]^q ds \leq \int_0^\infty \left[ \frac{K(f, t)}{t^\beta} \right]^q dt \leq \sum_{j \geq L - 1} [2^{rj}K(f, 2^{-rj})]^q \leq \sum_{j \geq L - 1} [2^{rj}e_j(f)]^q.$$  

Note that the sums here begin at $L - 1$.

On the other hand, the previous section showed that for sums beginning at $L$, we may pass from the MRA approximation errors to the Besov seminorm on wavelet coefficients:

$$\sum_{j \geq L} [2^{rj}e_j(f)]^q \leq \sum_{j \geq L} [2^{rj}\|Q_j f\|_p]^q \leq \sum_{j \geq L} [2^{rj}\|\theta_j\|_p]^q = \|\theta\|_q^q.$$  

Although the ranges of summation differ, this is taken care of by inclusion of $L_p$ norm of $f$.

In one direction this is trivial since the sum from $L$ is no larger than the sum from $L - 1$. So, moving up the preceding chain and using also (D.28) with (D.23), we get

$$\|f\|_b \leq \|\beta\|_p + |\theta|_b \leq C \|P_\ell f\|_p + C |f|_B \leq C (\|f\|_p + |f|_B) = C \|f\|_B.$$  

In the other direction, we connect the two chains by writing $|f|_B \leq C[e_{L-1}(f) + |f|_B]$ and observing from (D.23) that

$$e_{L-1}(f) = \|(I - P_{L-1}) f\|_p \leq \|I - P_{L-1}\|_p \|f\|_p \leq C \|f\|_p.$$  

Hence, again using (D.28),

$$\|f\|_B \leq (1 + c_q C) \|f\|_p + c_q |f|_B \leq C (\|\beta\|_p + |\theta|_b) = C \|f\|_b.$$

\[\square\]
D.0.17. Strang-Fix conditions.

D.11. Proposition. Suppose that $\phi$ is continuous with compact support, with $c_0 = \sum \phi(n) \neq 0$. Let $\Phi = \text{span}\{\phi(x-n), n \in \mathbb{Z}\}$. The following are equivalent:

(i) $\tilde{\phi}^{(k)}(2\pi l) = 0$, $l \in \mathbb{Z}\setminus\{0\}$, $k = 0, \ldots, p - 1$.
(ii) $q_k(x) = \sum_n n^k \phi(x-n)$ is a polynomial of exact degree $k$, $k = 0, \ldots, p - 1$.

(iii) $P_{p-1} \subset \Phi$.

Proof. We assume without loss of generality that $c_0 = \sum \phi(n) = 1$. Let $c_r = \sum_n n^r \phi(n)$. We will show that (i) $\iff$ (i’) $\iff$ (ii’) and that (ii) $\implies$ (ii’), where

(i’) For $k = 0, \ldots, p - 1$,

$$r_k(x) = \sum (x-n)^k \phi(x-n) \equiv c_k.$$

(ii’) For $k = 0, \ldots, p - 1$,

$$q_k(x) = \sum n^k \phi(x-n) = \sum_{j=0}^k \binom{k}{j} (-1)^j c_j x^{k-j}.$$

To show that (i) $\iff$ (ii), note that $r_k$ has Fourier coefficients

$$\int_0^1 e^{-2\pi ilx} \sum_n (x-n)^k \phi(x-n) dx = \int_{-\infty}^\infty e^{-2\pi ilx} x^k \phi(x) dx = \phi^{(k)}(2\pi l).$$

Condition (i’) thus says that all non-zero Fourier coefficients of $r_k$ vanish, which is equivalent to $r_k$ being identically constant and thus equal to $r_k(0) = c_k$.

(i’) $\iff$ (ii’). Write $n^k = [(x-n) + x]^k$, then expand the binomial and use (D.31). The converse follows by expanding $(x-n)^k$, using (D.32) and rearranging:

$$\sum (n-x)^k \phi(x-n) = \sum_{j=0}^k \binom{k}{j} x^{k-j} (-1)^j q_j(x)$$

$$= \sum_{j=0}^k \binom{k}{j} x^{k-j} (-1)^j \sum_{i=0}^j \binom{j}{i} (-1)^i c_i x^{j-i}$$

$$= \sum_{i=0}^k \binom{k}{i} c_i x^{k-i} (1-1)^{k-i} = c_k.$$

(ii) $\implies$ (iii). This is immediate since the $q_k$ are linearly independent.
(iii) ⇒ (ii'). We have \( x^k = \sum_m \alpha_{km} \phi(x - m) \), and so substitution into the definition of \( q_k \) gives

\[
q_k(x) = \sum_n \sum_m \alpha_{km} \phi(n - m) \phi(x - m) = \sum_r \sum_m \alpha_{km} \phi(x - r - m)
\]

\[= \sum_r (x - k)^r \phi(r) = \sum_{j=0}^{k} \binom{k}{j} (-1)^j c_j x^{k-j}.
\]

\[\Box\]
First a straightforward extension of Young’s inequality for convolutions.

D.12. **Theorem.** Let \((X, \mathcal{B}_X, \mu)\) and \((Y, \mathcal{B}_Y, \nu)\) be \(\sigma\)-finite measure spaces, and let \(K(x, y)\) be a jointly measurable function. Suppose that

\[
\begin{align*}
& (i) \quad \int |K(x, y)| \mu(dx) \leq M_1 \quad \text{a.e. } (\nu), \quad \text{and} \\
& (ii) \quad \int |K(x, y)| \nu(dy) \leq M_2 \quad \text{a.e. } (\mu).
\end{align*}
\]

For \(1 \leq p \leq \infty\), the operator 

\[
(Kf)(x) = \int K(x, y) f(y) \nu(dy)
\]

maps \(L_p(Y) \to L_p(X)\) with

\[
\|Kf\|_p \leq M_1^{1/p} M_2^{1-1/p} \|f\|_p.
\]

**Proof.** For \(p = \infty\) the result is immediate. For \(1 < p < \infty\), let \(q\) be the conjugate exponent \(1/q = 1 - 1/p\). Expand \(|K(x, y)|\) as \(|K(x, y)|^{1/q} |K(x, y)|^{1/p}\) and use Hölder’s inequality:

\[
|Kf(x)| \leq \left[ \int |K(x, y)|^{1/q} \nu(dy) \right]^{1/q} \left[ \int |K(x, y)|^{1/p} f(y) \nu(dy) \right]^{1/p},
\]

so that, using (ii),

\[
|Kf(x)|^p \leq M_2^{p/q} \int |K(x, y)|^{1/p} f(y)^p \nu(dy).
\]

Now integrate over \(x\), use Fubini’s theorem and bound (i) to obtain (D.33). The proof for \(p = 1\) is similar and easier. \(\square\)

**Remark.** Similarly, the adjoint operator \((K^* g)(y) = f(x)K(x, y) \mu(dx)\) maps \(L_p(X) \to L_p(Y)\) with

\[
\|K^* g\|_p \leq M_1^{1-1/p} M_2^{1/p} \|g\|_p.
\]

Two traditional forms of Young’s inequality are immediate consequences.

D.13. **Corollary.** Suppose that \(1 \leq p \leq \infty\).

\(i\) If \(Kf(x) = \int_0^\infty K(x - y) f(y) dy\), then

\[
\|Kf\|_p \leq \|K\|_1 \|f\|_p.
\]

\(ii\) If \(c_k = \sum_{j \in \mathbb{Z}} a_k \beta_j\), then

\[
\|c\|_p \leq \|a\|_1 \|\beta\|_p.
\]

D.14. **Lemma.** If \(f \in W^r_p(I)\) for \(1 \leq p \leq \infty\) and \(T_{r-1}\) is the Taylor polynomial approximation to \(f\) at \(c \in I\), then

\[
\|f - T_{r-1}\|_p(I) \leq \frac{|I|^r}{(r-1)!} \|f^{(r)}\|_p(I).
\]

315
See, for example, DeVore & Lorentz (1993, p. 37, Proposition 5.5)

Now for some $L_p$ error bounds on two common forms of polynomial approximation. The Taylor polynomial of degree $r - 1$ associated with $f \in \mathcal{W}_p(I)$ and a point $c \in I$ is

$$T_{r-1} f(x) = \sum_{k=0}^{r-1} \frac{(x-c)^k}{k!} f^{(k)}(c).$$

Integration by parts shows that

$$f(x) - T_{r-1} f(x) = \int_c^x \frac{(x-y)^{r-1}}{(r-1)!} f^{(r)}(y) dy.$$  

Orthogonal polynomials with respect to Lebesgue measure on $I = [a, b]$ are given by scaled versions of the Legendre polynomials $P_n(x)$, which are classically defined relative to the standard interval $[-1, 1]$, e.g. Szegö (1967). If $u = ax + \beta$ maps $[a, b]$ to $[-1, 1]$, then $p_k(x) = [2a^{-1} (2k+1)]^{1/2} P_k(ax + \beta)$ are orthonormal on $[a, b]$. The $L_2[a, b]$-orthogonal projection of $f$ onto the space of polynomials of degree at most $r - 1$ is given by

$$\Pi_{r-1} f(x) = \int \Pi_{r-1}(x, y) f(y) dy,$$

where

$$\Pi_{r-1}(x, y) = \sum_{k=0}^{r-1} p_k(x) p_k(y).$$

Young's inequality provides easy bounds on the error of these approximations for functions in $\mathcal{W}_p(I)$.

D.15. COROLLARY. Let $1 \leq p \leq \infty$, $I = [a, b]$ and $f \in \mathcal{W}_p(I)$.

(i) If $T_{r-1} f$ is the Taylor polynomial of degree $r - 1$ approximating $f$ at $c \in I$, then

$$\| f - T_{r-1} f \|_p(I) \leq \| f^{(r)} \|_p(I).$$  

(ii) If $\Pi_{r-1} f$ is the $L_2(I)$-orthogonal projection of $f$ on $\mathcal{P}_{r-1}$, then

$$\| f - \Pi_{r-1} f \|_p(I) \leq \gamma_r(I) \| f^{(r)} \|_p(I).$$

PROOF. For (i) apply Young's inequality to $\mu = \nu = \text{Lebesgue measure}$ on $I$ and the kernel in (D.37), namely

$$K(x, y) = \begin{cases} \frac{(x-y)^{r-1}}{(r-1)!} \text{sgn}(x-y) & \text{if } c < y < x \text{ or } x < y < c, \\ 0 & \text{otherwise.} \end{cases}$$

From this it follows that both $M_1$ and $M_2 \leq |I|^r / r!$.

For (ii), we note that since $\Pi_{r-1}(T_{r-1} f) = T_{r-1} f$,

$$f - \Pi_{r-1} f = (I - \Pi_{r-1})(f - T_{r-1} f) = (I - \Pi_{r-1})K f^{(r)},$$
where $K$ is given by (D.40). Consequently
\[ \| f - \Pi_{r-1} f \|_p \leq \| I - \Pi_{r-1} \|_p (|I| r! / r!) \| f^{(r)} \|_p. \]

To bound $\| \Pi_{r-1} \|_p$, we again use (D.33). Since the Legendre polynomials satisfy $|P_k(x)| \leq 1$ for $x \in [-1, 1]$, and using Cauchy-Schwartz to bound $\int |p_k| \leq |I|^{1/2}$, we find that
\[ \int_I |\Pi_{r-1}(x, y)| dy \leq \sum_k |p_k(x)| \int_I |p_k(y)| dy \leq |I| \sum_{k=0}^{r-1} \sqrt{2k+1} = |I| a_r, \]

say. From this we have that $\| I - \Pi_{r-1} \|_p \leq 1 + |I| a_r$, completing the proof. \qed
APPENDIX E

Glossary

Exponential Inequalities.

E.1. LEMMA. Suppose that \( Z_i \overset{i.i.d.}{\sim} N(0,1) \) and \( \alpha_i \in \mathbb{R}, i \in I. \) Then for \( t > 0, \)
\[
(\text{E.1}) \quad P\{ \sum \alpha_j (Z_j^2 - 1) > t \} \leq \exp\{ -t/8 \min(t/\|\alpha\|_2^2, 1/\|\alpha\|_\infty) \}.
\]
\[
(\text{E.2}) \quad \leq \exp\{ -t^2/(8\|\alpha\|_1 \|\alpha\|_\infty) \} \quad \text{if } t \leq \|\alpha\|_1.
\]

PROOF. Let \( Y_j = Z_j^2 - 1. \) Using Markov’s inequality and independence of the \( Y_j, \)
\[
P = P\{ \sum \alpha_j Y_j > t \} \leq e^{-st} E e^{s \sum \alpha_j Y_j} = e^{-st} \Pi E e^{s\alpha_j Y_j}.
\]
For \( u = s\alpha_j \leq 1/4, \)
\[
E e^{u(Z^2-1)} = e^{-u} \int_{-\infty}^{\infty} e^{u z^2 - z^2/2} dz/\sqrt{2\pi} = e^{-u}/(1 - 2u)^{1/2} \leq e^{2u^2}.
\]
where the inequality is equivalent to \( \log(1 - v) \geq -v - v^2 \) for \( -\infty \leq v = 2u \leq 1/2. \) Thus
\[
(\text{E.3}) \quad P \leq \exp\{ -st + 2s^2 \|\alpha\|_2^2 \} \quad \text{for } 0 \leq s \leq s_u = 1/4 \|\alpha\|_\infty.
\]
Minimizing the quadratic in \( s \) at \( s = s_\ast, \) and accounting separately for the cases when the minimum \( s_\ast \leq s_u \) or not yields (E.1). The second inequality follows from \( \sum \alpha_i^2 \leq \sup |\alpha_i| \sum |\alpha_i|. \) \( \square \)

Note that if all \( \alpha_i \leq 0, \) as would arise when considering the left tail of the distribution, then (E.3) holds for all \( s \geq 0 \) and the bound (E.1) simplifies to

E.2. COROLLARY. Suppose that \( Z_i \overset{i.i.d.}{\sim} N(0,1), \) and \( \beta_i \geq 0. \) Then for \( t > 0, \)
\[
(\text{E.4}) \quad P\{ \sum \beta_j (Z_j^2 - 1) < -t \} \leq \exp\{ -t^2/8 \|\beta\|_2^2 \}.
\]
E.0.18. The Ornstein-Uhlenbeck semigroup. There are analytical and probabilistic proofs of Proposition 13.1 - we describe the analytical one, in part because of its tangential connection with Stein's Unbiased Risk formula.

Let \( Z \sim N_n(0, I) \) and \( f \) be a Lipschitz function of \( x \in \mathbb{R}^n \). For \( t \geq 0, \) let
\[
u_t = e^{-t}, \quad v_t = (1 - e^{-2t})^{1/2},
\]
and let
\[
P_t \ f(x) = E f(u_t x + v_t Z),
\]
or equivalently—using \( \phi_\sigma \) to denote the \( N(0, \sigma^2) \) density—
\[
P_t f(x) = (f * \phi_\sigma)(u_t x).
\]
Thus \( P_t f(x) = x \) and \( P_\infty f(x) \equiv E f(Z) \), while for \( t > 0 \), (E.6) shows that \( x \to P_t f(x) \) is \( C^\infty \) in \( x \). Another consequence of the smoothing effect of \( P_t \) is seen by using (E.5) and the Lipschitz property to see that \( |P_t f(x) - P_t f(y)| \) is bounded by \( u_t \| f \|_{\text{Lip}} \| x - y \| \), so that
\[
\| P_t f \|_{\text{Lip}} \leq e^{-t} \| f \|_{\text{Lip}}.
\]
Similarly \( |P_t f(0) - f(0)| \) is bounded by \( v_t \| f \|_{\text{Lip}} E \| Z \| \): together with the previous sentence, this yields a uniform domination
\[
|P_t f(x)| \leq |f(0)| + \| f \|_{\text{Lip}}(\sqrt{\pi} + \| x \|).
\]

The important semigroup property says that \( P_s P_t f = P_{s+t} f \) for all \( s, t \geq 0 \). To see this, use (E.5) and independent standard Gaussians \( Z_s, Z_t \) and \( (Z_{s+t}) \) to write
\[
P_s P_t f(x) = E f(u_s u_t x + u_s v_s Z_s + v_t Z_t)
\]
and then note that \( u_t u_s = u_{t+s} \) and \( u_t v_s Z_s + v_t Z_t \equiv u_{t+s} Z_{s+t} \).

E.3. Lemma. Suppose that \( f \) and \( g \) are both \( C^2 \) and Lipschitz functions on \( \mathbb{R}^n \). Then
\[
d_{dt} P_t f|_{t=0} = L f := \Delta f - \langle x, \nabla f \rangle,
\]
and
\[
E(L f) g = -E(\nabla f, \nabla g).
\]

Remarks. 1. Formula (E.9) defines the generator \( L \) of the semigroup \( \{P_t, t \geq 0\} \) and identifies it with the Ornstein-Uhlenbeck operator.

2. The semigroup property extends (E.9) to all \( t \geq 0 \):
\[
d_{dt} P_t f = LP_t f.
\]

3. Formula (E.10) is the analog for Gaussian measure of the partial integration formula \( \int f \Delta g = -\int (\nabla f, \nabla g) \) for Lebesgue measure. When \( g \equiv 1 \), (E.10) reduces to Stein's formula (9.1) applied to \( h = \nabla f \):
\[
E(\nabla f, z) = E \Delta f.
\]
PROOF. We begin with (E.10). Integration by parts shows that for $C^1$ and bounded $h$,
\[- \int h(D_t g) \phi = \int D_t (h \phi) g = \int (D_t h - x_i h) g \phi.\]
Now set $h = D_t f$ and sum over $i$ to obtain (E.10). For (E.9), we have from (E.5)
\[\frac{d}{dt} P_t f(x) = E(\nabla f(u_t x + v_t Z), \dot{u}_t x + \dot{v}_t Z),\]
where $\dot{u} = (du/dt)$ etc. After using Stein’s formula (E.12), this becomes
\[\dot{u}_t E(\nabla f(u_t x + v_t Z), x) + v_t \dot{v}_t E \Delta f(u_t x + v_t Z).
\]
Since $\dot{u}_0 = -1$ and $v_t \dot{v}_t = \frac{1}{2}D_t v_t^2 \to 1$ as $t \to 0$, formula (E.9) drops out. □

**E.0.19. Proof of concentration inequality.** Proposition 13.1 will follow from an inequality for the moment generating function of $f(Z)$.

**E.4. Proposition.** Suppose that $Z \sim N_n(0, I)$ and that $Eg(Z) = 0$ with $\|g\|_{Lip} \leq 1$. Then for all real $\lambda$,
\[Ee^{\lambda g(Z)} \leq e^{\lambda^2/2}.
\]
Indeed, if $f$ is our function in Proposition 13.1, apply this result to $g = f - Ef$:
\[P\{f - Ef \geq t\} \leq e^{-M Ee^{\lambda|f - Ef|}}\]
\[\leq e^{-M + \lambda^2 \|f\|_{Lip}^2/2}.
\]
Minimizing over $\lambda$ yields the conclusion of Proposition 13.1.

**Proof of Proposition E.4.** Suppose first that $g$ is smooth, say $C^2$. Since $P_t g$ is Lipschitz, the function $G(t) = E\exp(\lambda P_t g)$ is well defined. Using the generator formula (E.11) and then the partial integration (E.10),
\[G'(t) = \lambda E L(P_t g) e^{\lambda P_t g} = -\lambda^2 E(\nabla P_t g, \nabla P_t g) e^{\lambda P_t g}.\]
Now the Euclidean norm $\|\nabla P_t g\| \leq \|P_t g\|_{Lip} \leq e^{-t} \|g\|_{Lip} \leq e^{-t}$ using (E.7), and so
\[-G'(t) \leq \lambda^2 e^{-2t} G(t).
\]
Hence $H(t) = \log G(t)$ satisfies $-H'(t) \leq \lambda^2 e^{-2t}$ along with $H(\infty) = 0$, (since $G(\infty) = Ee^{\lambda M}$ = 1), and so
\[\log Ee^{\lambda M} = H(0) = \int_0^\infty -H'(t) dt \leq \lambda^2/2.
\]
For general Lipschitz $g$, note that $P_t g$ is $C^\infty$ and is Lipschitz(1) (from (E.7)) with mean 0. Hence $Ee^{\lambda P_t g} \leq e^{\lambda^2/2}$ and the dominated convergence theorem (permitted by (E.8)) yields $Ee^{\lambda g} = \lim_{t \to 0} Ee^{\lambda P_t g}$ and hence the conclusion. □

[Refer to Stochastic argument, Ledoux]
**Bounded operators on Hilbert space.** [For further details on this material, see texts on functional analysis such as Reed & Simon (1980, Chapter 6).]

$\mathcal{H}$ denotes a Hilbert space - for us usually $\ell^2(\mathbb{N})$. $\mathcal{L}(\mathcal{H})$ denotes the space of bounded linear operators on $\mathcal{H}$.

An operator $A \in \mathcal{L}(\mathcal{H})$ is **compact** if $A$ takes bounded sets to sets with compact closure. Equivalently, $A$ is compact if and only if for every bounded sequence $\{x_n\} \in \mathcal{H}$, the sequence $\{Ax_n\}$ has a convergent subsequence.

The Hilbert-Schmidt theorem. Let $A \in \mathcal{L}(\mathcal{H})$ be compact and self-adjoint. There exists a complete orthonormal basis $\{e_n\}$ for $\mathcal{H}$ such that $Ae_n = \lambda_n e_n$ with $\lambda_n \in \mathbb{R}$ and $\lambda_n \to 0$ as $n \to \infty$.

An operator $B \in \mathcal{L}(\mathcal{H})$ is **positive** if $\langle Bx, x \rangle \geq 0$ for all $x \in \mathcal{H}$. [On a complex Hilbert space, a positive operator is necessarily self-adjoint, but not on real $\mathcal{H}$.

If $A \in \mathcal{L}(\mathcal{H})$ is positive, there exists a unique square root $B \in \mathcal{L}(\mathcal{H})$ which is positive and satisfies $B^2 = A$. One writes $B = \sqrt{A}$ or $A^{1/2}$.

The **absolute value** of $A \in \mathcal{L}(\mathcal{H})$ is defined by $|A| = \sqrt{A^*A}$.

An operator $U \in \mathcal{L}(\mathcal{H})$ is an **isometry** if $\|Ux\| = \|x\|$ for all $x \in \mathcal{H}$. $U$ is a **partial isometry** if $U$ is an isometry when restricted to the closed subspace $(\text{Ker}U)^\perp$.

**Polar decomposition.** If $A \in \mathcal{L}(\mathcal{H})$, there exists a partial isometry $U$ such that $A = U|A|$. [And in particular, $\text{Ran} \subset (\text{Ker}U)^\perp$.

An operator $A \in \mathcal{L}(\mathcal{H})$ is **Hilbert-Schmidt** if for some orthonbasis $\{e_i\}$

$$\|A\|^2_{HS} = \sum_{i,j} |\langle e_i, Ae_j \rangle|^2 < \infty.$$  

The value of $\|A\|^2_{HS}$ does not depend on the orthonbasis chosen: regarding $A$ as an infinite matrix, $\|A\|^2_{HS} = \text{tr} A^*A$.

If $A$ is Hilbert-Schmidt on $\mathcal{H}$ and $B \in \mathcal{L}(\mathcal{H})$ then $AB$ and $BA$ are both Hilbert-Schmidt.

[Real vs. complex Hilbert space?]
[Convex Set]
[1.s.c. and max on compact]
[metric space: seq cty = cty]

A subset \( K \) of a metric space is compact if every covering of \( K \) by open sets has a finite subcover.

A subset \( K \) of a metric space is totally bounded if it can be covered by finitely many balls of radius \( \epsilon \) for every \( \epsilon > 0 \).

[Ref: Rudin FA p 369] If \( K \) is a closed subset of a complete metric space, then the following three properties are equivalent: (a) \( K \) is compact, (b) Every infinite subset of \( K \) has a limit point in \( K \), (c) \( K \) is totally bounded.

[Orthogonality]
[Support of \( \mu \)]

Let \( P \) and \( Q \) be probability measures on a measurable space \( (\mathcal{X}, \mathcal{B}) \), absolutely continuous with respect to a probability measure \( \lambda \). (For example, \( \lambda = (P + Q)/2 \).) Write \( p = dP/d\lambda \) and \( q = dQ/d\lambda \). The Hellinger affinity

\[
h(P, Q) = \int \sqrt{pq} d\lambda
\]

do not depend on the choice of \( \lambda \).

Let \( \{P_n\} \) and \( \{Q_n\} \) be two sequences of probability measures on \( \mathbb{R} \). Define product measures on sequence space \( \mathbb{R}^\infty \), with the product Borel \( \sigma \)-field, by \( P = \prod P_n \) and \( Q = \prod Q_n \).

E.5. Theorem. If \( P_n \sim Q_n \) for \( n = 1, 2, \ldots \) then \( P \) and \( Q \) are either equivalent or orthogonal. Moreover,

\[
P \sim Q \text{ if and only if } \prod_{k=1}^\infty h(P_k, Q_k) > 0.
\]

In case \( P \sim Q \),

\[
\frac{dP}{dQ} = \prod_{k=1}^\infty \frac{dP_k}{dQ_k}.
\]
There is a useful and simple criterion for the convergence of a series of Gaussian random variables. It is relatively easy to prove using, for example, the convergence theorem for square integrable martingales (see e.g. Bogachev (1998, page 2) or Williams (1991, page 111)).

**E.6. Theorem.** Let \( \xi_n \) be a sequence of independent Gaussian random variables with mean zero and variance \( \sigma_n^2 \). Then the following conditions are equivalent:

(i) The series \( \sum_{n=1}^{\infty} \xi_n \) converges almost surely;
(ii) The series in (i) converges in probability;
(iii) The series in (i) converges in \( L_2 \);
(iv) \( \sum_{n=1}^{\infty} \sigma_n^2 < \infty \).

A process \( \{ Z(t), t \in T \} \) is Gaussian if all finite-dimensional distributions \( \{ Z(t_1), \ldots, Z(t_k) \} \) have Gaussian distributions for all \( (t_1, t_2, \ldots, t_k) \in T^k \) and positive integer \( k \). It is said to be continuous in quadratic mean if \( E[|Z(t + h) - Z(t)|^2] \to 0 \) as \( h \to 0 \) at all \( t \).

Standard Brownian motion on the interval \( [0, 1] \) is defined as a Gaussian process \( \{ W_t \} \) with mean zero and covariance function \( \text{Cov}(W_s, W_t) = s \wedge t \).

Suppose that \( \{ \phi_i, i = 1, \ldots \} \) is an arbitrary orthonormal basis for \( L_2[0, 1] \) and that \( z_i \) is an i.i.d. sequence of standard Gaussian variables. Then Shepp (1966) showed that

\[
W_t \overset{D}{=} \sum_{i=1}^{\infty} z_j \int_0^t \delta_j(s)ds
\]

with the series converging almost surely. Particular examples for which this representation was known earlier include the trigonometric basis \( \phi_k(t) = \sqrt{2} \cos(k - \frac{1}{2})\pi t \) (Wiener) and the Haar basis \( \phi_{jk}(t) = 2^{j/2}h(2^j t - k) \) for \( h(t) \) equal to 1 on \( [0, \frac{1}{2}] \) and to \(-1\) on \( [\frac{1}{2}, 1] \) (Lévy).
The Fourier transform of an integrable function on \( \mathbb{R} \) is defined by
\[
\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-i\xi x} \, dx.
\]
If \( f \) is sufficiently nice (SEE REFS), it may be recovered from the inversion formula
\[
f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{i\xi x} \, d\xi.
\]
Parseval's formula: (check normalization with Gaussians)
\[
\int f^2(x) \, dx = \frac{1}{2\pi} \int |\hat{f}(\xi)|^2 \, d\xi.
\]
The Poisson summation formula (Dym & McKean 1972, for example) states that if \((1 + x^2)[|f(x)| + |f'(x)| + |f''(x)|] \) is bounded (or if the same condition holds for \( \hat{f} \)), then
\[
\sum_{k \in \mathbb{Z}} f(k) = \sum_{k \in \mathbb{Z}} \hat{f}(2\pi k).
\]
When applied to \( f(x) = g(x + t) \), this yields a representation for the periodization of \( g \) (REF??): 
\[
\sum_{k} g(t + k) = \sum_{k} e^{2\pi ik t} \hat{g}(2\pi k), \quad t \in \mathbb{R}.
\]
Notation.
\[ \delta_{jk} = 1 \text{ if } i = j \text{ and } 0 \text{ otherwise.} \]

E.7. Theorem (Minkowski’s integral inequality). Let \((X,B_X,\mu)\) and \((Y,B_Y,\nu)\) be \(\sigma\)-finite measure spaces, and let \(f(x,y)\) be a jointly measurable function. Then for \(1 \leq p \leq \infty\),

\[
(E.13) \quad \left( \int \left( \int |f(x,y)|^p \nu(dy) \right)^{1/p} \mu(dx) \right)^{1/p} \leq \int \left( \int |f(x,y)|^p \mu(dx) \right)^{1/p} \nu(dy).
\]

Bibliography


decomposition methods for bounded parameter minimax risk evaluation’,

Gradshteyn, I. S. & Ryzhik, I. M. (1980), Table of Integrals, Series, and

metric generalized linear models’, Probability Theory and Related Fields
111, 167-214.

Green, P. & Silverman, B. (1994), Nonparametric Regression and General-
ized Linear Models, Chapman and Hall, London.


Hall, P. & Patil, P. (1993), Formulae for mean integrated squared error
of nonlinear wavelet-based density estimators, Technical Report CMA-

Härdle, W., Hall, P. & Marron, S. (1988), ‘How far are automatically chosen
regression smoothing parameters from their minimum? (with discussion)’,

Härdle, W., Kerkyacharian, G., Picard, D. & Tsybakov, A. (1998), Wavelets,
Approximation and Statistical Applications, Springer.

Hastie, T. J. & Tibshirani, R. J. (1990), Generalized Additive Models, Chap-
man and Hall.


Hotelling, H. (1933), ‘Analysis of a complex of statistical variables into prin-
cipal components’, J. Educational Psychology 24, 417-441,498-520.


Ibragimov, I. & Khas’minskii, R. (1981), Statistical estimation : asymptotic

Ibragimov, I. A. & Khas’minskii, R. Z. (1980), Asymptotic properties of
some nonparametric estimates in gaussian white noise, in ‘Proceedings
of Third International Summer School in Probability and Mathematical

Ibragimov, I. A. & Khas’minskii, R. Z. (1982), ‘Bounds for the risks of
non-parametric regression estimates’, Theory of Probability and its Applica-
tions 27, 84-99.

Ibragimov, I. A. & Khas’minskii, R. Z. (1984), ‘On nonparametric estimation
of the value of a linear functional in gaussian white noise’, Theory of
Probability and its Applications 29, 18-32.

James, W. & Stein, C. (1961), Estimation with quadratic loss, in ‘Pro-
cedings of Fourth Berkeley Symposium on Mathematical Statistics and

Kuhn, H. (1953), ‘Review of kneser (1952)’.


