Gaussian estimation: Sequence and wavelet models

Draft version, September 8, 2015
Comments welcome

Iain M. Johnstone

Contents

(working) Preface

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

$$y_i = \theta_i + \epsilon z_i, \qquad i \in I, \tag{0.1}$$

where I is finite or countable, $\{\theta_i\}$ is fixed and unknown, $\{z_i\}$ are i.i.d. $N(0,1)$ noise variables and ϵ is a known noise level. If I is finite, this is an old friend, the multivariate normal means model, with independent co-ordinates 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, 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 several responses to these objections. 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). The broad strategy is to apply various methods that one understands in the multivariate normal model to finite submodels, and to argue that often 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 one aim 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 idealisation in our models than was necessary when developing theory for the estimation of one, two or three parameters from modest numbers of observations.

Thirdly, it turns out that model (0.1) is often a reasonable approximation, in large samples, to other nonparametric settings. In parametric statistics, the central limit theorem and asymptotic normality of estimators extends the influence of multivariate normal theory to generalized linear models and beyond. In nonparametric estimation, it has long been observed that similar features 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, notably by Russian theoretical statisticians, led by I. A. Ibragimov and R. Z. Khasminskii. It was somewhat slower to receive wide discussion in the West. However, it received a considerable impetus when it was observed that (0.1) was a natural setting in which to understand the estimation of signals, functions and images in wavelet orthonormal bases. In turn, wavelet bases made it possible to give a linked theoretical and methodological account of function estimation that responded appropriately to spatial inhomogeneties in the data, such as (in an extreme form) discontinuities and cusps.

The goal of this book is to give an introductory account of some of the theory of estimation in the Gaussian sequence model that reflects these ideas.

Estimators are studied and compared using the tools of statistical decision theory, which for us means typically (but not always) comparison of mean squared error over appropriate classes of sets Θ supposed to contain the unknown vector θ. The best-worst-case or minimax principle is used, though deliberately more often in an approximate way than exactly. Indeed, we look for various kinds of approximate *adaptive* minimaxity, namely estimators that are able to come close to the minimax criterion simultaneously over a class of parameter sets. A basic theme is that the geometric characteristics of the parameter sets, which themselves often reflect assumptions on the *type* of smoothness of functions, play a critical role.

In the larger first part of the book, Chapters 1- 9, an effort is made to give "equal time" to some representative linear and non-linear estimation methods. Linear methods, of which kernel estimators, smoothing splines, and truncated series approaches are typical examples, are seen to have excellent properties when smoothness is measured in a sufficiently spatially uniform way. When squared error loss is used, this is geometrically captured by the use of hyperrectangles and ellipsoids. Non linear methods, represented here primarily by thresholding of data in a wavelet transform domain, come to the fore when smoothness of a less uniform type is permitted. To keep the account relatively self-contained, introductions to topics such as Gaussian decision theory, wavelet bases and transforms, and smoothness classes of functions are included. A more detailed outline of topics appears in Section 1.6 after an expanded introductory discussion. Starred sections contain more technical material and can be skipped on a first reading.

The second part of the book, Chapters 10– 15, is loosely organized as a tour of various types of asymptotic optimality in the context of estimation in the sequence model. Thus, one may be satisfied with optimality "up to log terms", or "up to constants" or "with exact constants". One might expect that as the demands on quality of optimality are ratcheted up, so are the corresponding assumptions, and that the tools appropriate to the task change. In our examples, intended to be illustrative rather than exhaustive, this is certainly the case. The other organizing theme of this second part is a parallel discussion of results for simple or "monoresolution" models (which need have nothing to do with wavelets) and conclusions specifically for multiresolution settings.

We often allow the noise level ϵ in (0.1) to depend on the index i –a small enough change to be easily accommodated in many parts of the theory, but allowing a significant expansion in models that are fairly directly convertible to sequence form. Thus, many linear inverse problems achieve diagonal form through a singular value or wavelet-vaguelette decompo-

sition, and problems with correlated Gaussian noise can be diagonalized by the principal compoent or Karhunen-Loève transformation.

Of course much is omitted. To explain some of the choices, we remark that the project began over ten years ago as an account of theoretical properties of wavelet shrinkage estimators based largely on work with David Donoho, Gérard Kerkyacharian and Dominique Picard. Much delay in completion ensued, due to other research and significant administrative distractions. This history has shaped decisions on how to bring the book to light after so much elapsed time. First among the choices has been to cast the work more as a graduate text and less as a current research monograph, which is hopefully especially apparent in the earlier chapters. Second, and consistent with the first, the book does not attempt to do justice to related research in recent years, including for example the large body of work on non-orthogonal regression, sparse linear models and compressive sensing. It is hoped, however, that portions of this book will provide helpful background for readers interested in these areas as well.

The intended readership, then, includes graduate students and others who would like an introduction to this part of the theory of Gaussian estimation, and researchers who may find useful a survey of a part of the theory. Helpful background for reading the book would be familiarity with mathematical statistics at the level of a first year doctoral course in the United States.

The exercises, which are concentrated in the earlier chapters, are rather variable in complexity and difficulty. Some invite verifications of material in the text, ranging from the trivial to the more elaborate, while others introduce complementary material.

Acknowledgements [in progress]

This project has an absurdly long history and a matching list of happy debts of gratitude. The prehistory begins with a DMV seminar in March 1995 at Oberwolfach on wavelets in statistics, jointly with Dave Donoho, and a June 1996 course at Kasteel de Berkct in the Netherlands organized by Piet Groeneboom.

The transition from LaTeX slides to blackboard exposition marks the true beginning of the book, and I am grateful to Lucien Birgé, Olivier Catoni and Pascal Massart for the invitation to give an advanced course at the École Normale Supérieure in Paris in Spring of 1998, and for the scientific and personal welcome extended by them and by Gérard Kerkyacharian, Dominique Picard and Alexander Tsybakov.

I warmly thank my coauthors: particularly Dave Donoho, with whom much of the wavelets in statistics work began, and repeat offenders Gérard Kerkyacharian, Dominique Picard and Bernard Silverman, as well as our friends Felix Abramovich, Yoav Benjamini, Jeff Hoch, Brenda MacGibbon, Alan Stern, and the late Marc Raimondo, who is sorely missed.

For encouragement and thoughtful comments on the manuscript, I'm greatly indebted to Felix Abramovich, Peter Bickel, Larry Brown, Emmanuel Candès, Laurent Cavalier, Shingchang Kou, Yi Lin, Brenda MacGibbon, Stéphane Mallat, Boaz Nadler, Michael Nussbaum, John Rice, Martin Slawski and Cun-Hui Zhang, as well as to the (then) students in courses at Berkeley and Stanford – Ery Arias Castro, Arnab Chakraborty, Bowen Deng, Zhou Fan, Jiashun Jin, Arthur Lu, Zongming Ma, Charles Mathis, Debhashis Paul, Hualin Wang, Johannes Lederer. Some very valuable suggestions came from reviewers commis-

sioned by John Kimmel and Lauren Cowles, especially Anirban Das Gupta, Sam Efromovich and Martin Wainwright.

For the final push, I wish to specially thank Tony Cai, whose encouragement to complete the book took the concrete form of insightful counsel along with organizing further helpful comments from our colleagues Weidong Liu, Mark Low, Lie Wang, Ming Yuan and Harry Zhou. Michael Martin and Terry O'Neill at the Australian National University, and Marta Sanz at the University of Barcelona, hosted a sabbatical leave which gave some discipline to the protracted project. Harry Zhou, Tony Cai, Min Chen and Jianqing Fan invited me to give a June 2013 short course at the Chinese Academy of Sciences in Beijing, where invaluable help was provided by Shaojun Guo. Zhou Fan at Stanford read every chapter; his eagle eye still found an astounding number of infelicities and outright errors. Presumably, therefore, many still remain despite so much help, the responsibility of course lies with the author.

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 and National Institutes of Health, which have supported much of my own research and writing, and to the Australian National University and University of Barcelona which provided space and time for writing.

Chapter dependency graph. A heavy solid line indicates a more than incidental scientific dependence of the higher numbered chapter on the lower numbered one. A dotted line indicates a weaker formal dependence, perhaps at the level of motivation. A more specific indication of cross-chapter dependence at the level of sections can then be found below.

In the first part, Chapters 2 and 3 provide basic material for the book as a whole, while the decision theory of Chapter 4 is important for virtually everything that follows. The linear estimation results, Chapters 5 and 6, form one endpoint in themselves. Chapters 8 and 9 on thresholding and properties of wavelet shrinkage form the other main endpoint in Part I; the wavelet primer Chapter 7 prepares the way.

In the second part, with numbers shown in Courier font, there some independence of the chapters at a formal level: while they lean heavily on Part I, the groups $\{10\}, \{11, 12\}$ and $\{13, 14, 15\}$ can be read separately of one another. The first chapter in each of these three groups (for Ch. 10, the first half) does not require any wavelet/multiresolution ideas.

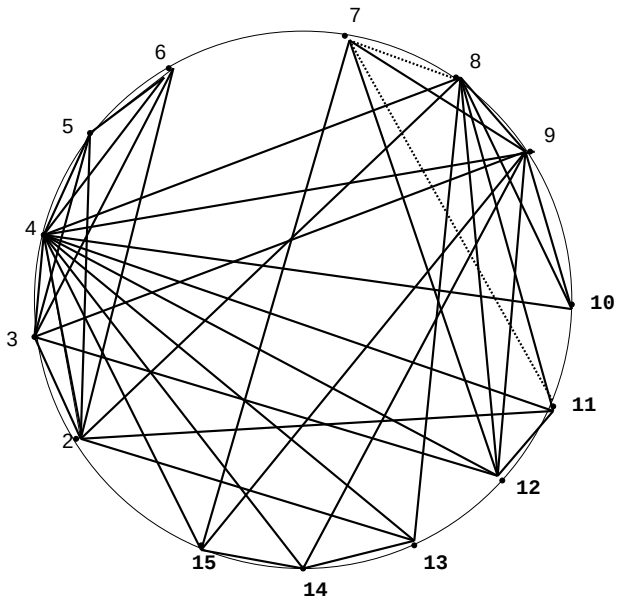

Figure 0.1 Chapter Dependencies

List of Notation

Standard notations.

x_+, positive part; $[x]$, fractional part; $\lfloor x \rfloor$, largest previous integer; $\lceil x \rceil$, smallest following integer.

$\star$, convolution; $\#\{\cdot\}$, cardinality; $a \wedge b = \min(a, b)$; $a \vee b = \max(a, b)$; log, base e logarithm.

$\mathbb{R}$, real numbers; $\mathbb{C}$, complex numbers; $\mathbb{Z}$, integers, $\mathbb{N} = \{1, 2, \ldots\}, \mathbb{N}_0 = \{0, 1, 2, \ldots\}$ natural numbers.

$\mathbb{R}^n, \mathbb{C}^n$ n-dimensional real, complex Euclidean space

$\mathbb{R}^\infty$, (countably infinite) sequences of reals, 58

Sets, Indicator (functions). A^c complement; A_t dilation; (REF?) $I_A, I_A(t)$ indicator function; $\mathrm{sign}(x)$ sign function.

Derivatives. Univariate: g', g'' or $g^{(r)}$, or $D^r g$, rth derivative; Partial: $\partial/\partial t, \partial/\partial x$; Multivariate: $D_i g$ or $\partial g/\partial x_i$; Divergence: ∇^T, (2.56).

Matrices. I_n identity, $n \times n$; A^T transpose; $\mathrm{tr}A$ trace; $\mathrm{rank}(A)$ rank; A^{-1} inverse; $\varrho_1(A) \geq \ldots \geq \varrho_n(A)$ eigenvalues; $A^{1/2}$ non-negative definite square root; $|A| = (A^T A)^{1/2}$ p. 36

Vectors. $\theta = (\theta_i)$ in sequence model; $\mathbf{f} = f(t_l)$ in time domain, 12; e_k has 1 in kth place, 0s elsewhere. Indices in sequence model: i for generic sequences, k for specific concrete bases.

Inner Products. $u^T v = \langle u, v \rangle = \sum_i u_i v_i$, Euclidean inner product; $\langle \cdot, \cdot \rangle_n$ normalized, p. 12; $\langle \cdot, \cdot \rangle_\varrho$ weighted, p. 80.

Norms. Vectors. $\| \cdot \|$, when unspecified, Euclidean norm $(\sum u_i^2)^{1/2}$; $\| \cdot \|_{2,n}$ normalized Euclidean p. 12; $\| \cdot \|_p$, ℓ_p norm, (1.5)

Matrices. $\| \cdot \|_{HS}$ Hilbert-Schmidt, p. (ii), (3.13), (C.5).

Functions. $\| f \|_p$ *Fill in!*; $\|g\|_{2,I}, \|g\|_{\infty,I}$ restricted to interval I, p. 67, (3.30).

Function spaces.

$L_2[0, 1]$	`App Ref.`
$\ell_2 = \ell_2(\mathbb{N})$	`App Ref.`
$\ell_{2,\varrho} = \ell_2(\mathbb{N}, (\varrho_i^{-2}))$	`App Ref.`

Normal/Gaussian Density and Distribution $\phi, \phi_\epsilon, \Phi, \tilde{\Phi}$ p. 20

Distributions, Expectations. (2.7), §4.1. Joint $\mathbb{P}(d\theta, dy), \mathbb{E}$;
Conditional on θ: $P(dy|\theta)$, $P_\theta(dy)$, E_θ;
Conditional on y: $P(d\theta|y)$, $P_y(d\theta)$, E_y;
Marginals: for y, $P_\pi(dy)$, E_{P_π}; for θ, $\pi(d\theta)$, E_π;

Collections: $\mathcal{P}$; supported on Θ, $\mathcal{P}(\Theta)$ before (4.19); convolutions $\mathcal{P}^*$ after (4.21); substochastic $\mathbb{P}_+(\mathbb{R})$, C.19; moment constrained, $\mathcal{M}, \mathcal{M}(C)$, §4.11.

Random Variables. Y (vector of) observations; Z mean zero; iid = independent and identically distributed; $\stackrel{\mathcal{D}}{=}$ equality in distribution.

Stochastic operators. E, expectation; Var, variance; Cov, covariance; Med, median, Exercise 2.4; Bias, bias, p. 35;

Estimators. $\hat{\theta} = \hat{\theta}(y)$, general sequence estimator at noise level ϵ; $\hat{\mu} = \hat{\mu}(x)$, general estimator at noise level 1, $\hat{\theta}_i, \hat{\theta}_{jk}$ ith, or (j,k)th component of θ;

$\hat{\theta}_\pi$, Bayes estimator for prior π, (2.9), p. 106; θ_y, posterior mean (2.14); $\hat{\theta}^*$, minimax estimator,

Specific classes of estimators: $\hat{\theta}_C$, linear estimators with matrices C, (2.45); $\hat{\theta}_c$ diagonal linear esimators with shrinkage constants c (3.14); $\hat{\theta}_\lambda$ threshold estimator, (1.7), (2.5), (2.6); except in Chapter 3, where it is a regularization (spline) estimator, (3.39); $\hat{\theta}_\nu$ truncation estimator, (3.17); $\hat{\theta}_h$ kernel estimator, (3.33);

$\hat{f}(t)$, $\hat{f}_h$, $\tilde{f}_h$, $\hat{f}_\lambda$, function estimator (3.21), (3.44).

Estimators with superscripts. $\hat{\theta}^{JS+}$ Positive part James-Stein estimator. (2.70) `[to index?]`

Decision theory: Loss function $L(a,\theta)$, §2.3, randomized decision rule, $\delta(A|y)$ (A.10). Distance between statistical problems, $\Delta_d(\mathcal{P}_0,\mathcal{P}_1)$, $\Delta(\mathcal{P}_0,\mathcal{P}_1)$, (3.88). L_1-distance, $L_1(\mathcal{P}_0,\mathcal{P}_1)$, (3.89).

Risk functions For estimators.

$r(\hat{\theta},\theta)$,	§2.5;
$r(\delta,\theta)$	of randomized rule, (A.10)
$r_L(\varrho,\eta)$	of linear shrinkage rule, (2.49)
$r_S(\lambda,\mu), r_H(\lambda,\mu)$	of soft (§2.7, §8.2) and hard (§8.2) thresholding.

For priors.

$B(\hat{\theta},\pi)$,	integrated risk (4.2);
$B(\pi), B(\pi,\epsilon)$,	Bayes risk (4.3);
$B(\mathcal{P}), B(\mathcal{P},\Sigma)$	maximum Bayes risk over collection $\mathcal{P}$, (4.14), with covariance matrix Σ, §4.9.

Minimax risks. $R_\mathcal{E}(\Theta,\Sigma)$, $R_\mathcal{E}(\Theta,\epsilon)$, $R_\mathcal{E}(\mathcal{F},\epsilon)$: minimax risks for parameter sets Θ, (3.2); function class $\mathcal{F}$, (3.7); noise covariance matrix Σ, §4.9; for estimator classes $\mathcal{E} =$ N, all estimators; = L, linear estimators, §3.1; = DL, diagonal linear, (4.61).

$R_n = R_N(\mathbb{R}^n,\epsilon)$, (2.50).

$\rho_N(\tau,\epsilon), \rho_L(\tau,\epsilon), \rho_P(\tau,\epsilon)$: univariate minimax risks for bounded interval $[-\tau,\tau]$, for nonlinear (4.26), linear (4.28), and projection (4.32) estimators.

1

Introduction

> And hither am I come, a Prologue armed,... to tell you, fair beholders, that our play leaps o'er the vaunt and firstlings of those broils, beginning in the middle; starting thence away to what may be digested in a play. (Prologue, *Troilus and Cressida* William Shakespeare.)

The study of linear methods, non-linear thresholding and sparsity in the special but central setting of Gaussian data is enlightened by statistical decision theory. This overture chapter introduces these themes and the perspective to be adopted.

Section 1.1 begins with two data examples, in part to emphasize that while this is a theoretical book, the motivation for the theory comes from describing and understanding the properties of commonly used methods of estimation.

A first theoretical comparison follows in Section 1.2, using specially chosen cartoon examples of sparse signals. In order to progress from constructed cases to a plausible theory, Section 1.3 introduces, still in a simple setting, the formal structures of risk function, Bayes rules and minimaxity that are used throughout.

The signal in Gaussian white noise model, the main object of study, makes its appearance in Section 1.4, in both continuous and sequence forms, along with informal connections to finite regression models and spline smoothing estimators. Section 1.5 explains briefly why it is our guiding model; but it is the goal of the book to flesh out the story, and with some of the terms now defined, Section 1.6 provides a more detailed roadmap of the work to follow.

1.1 A comparative example

We use two real data examples to introduce and motivate some of the themes of the book. In the first case, linear methods of estimation seem more or less adequate, while in the second we see substantial improvement by the use of non-linear wavelet thresholding.

The temperature data. Figure 1.1 shows daily minimum temperatures Y_l in degrees Celsius recorded in Canberra, Australia in the leap year 2008. A smoother summary curve might be helpful to see the temperature trend shorn of day to day variability.

We might adopt as a (provisional, approximate) model

$$Y_l = f(t_l) + \sigma Z_l, \qquad l = 1, \ldots, n. \tag{1.1}$$

The observation Y_l is the minimum temperature at a fixed time period t_l, here equally spaced, with $n = 366$, $f(t)$ is an unknown mean temperature function, while Z_l is a noise term,

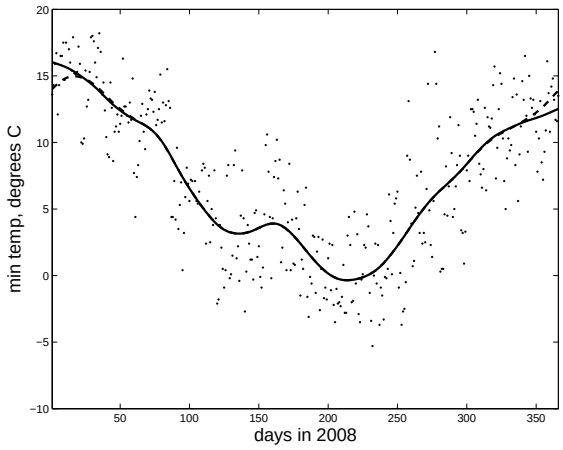

Figure 1.1 Spline smoothing of Canberra temperature data. Solid line: original spline fit, Dashed line: periodic spline, as described in text.

assumed to have mean zero, and variance one—since the standard deviation σ is shown explicitly.

Many approaches to smoothing could be taken, for example using local averaging with a kernel function or using local (linear) regression. Here we briefly discuss two versions of smoothing splines informally—Section 1.4 has formulas and a little more detail. The choice of splines here is merely for definiteness and convenience—what is important, and shared by other methods, is that the estimators are *linear* in the data Y, and depend on a tuning or bandwidth parameter λ.

A least squares approach would seek an estimator $\hat{f}$ to minimize a residual sum of squares $S(f) = n^{-1} \sum_{l=1}^{n} [Y_l - f(t_l)]^2$. In nonparametric estimation, in which f is unconstrained, this would lead to an interpolation, $\hat{f}(t_l) = Y_l$, an overfitting which would usually be too rough to use as a summary. The spline approach brings in a penalty for roughness, for example $P(f) = \int (f'')^2$ in terms of the squared second derivative of f. The spline estimator is then chosen to minimize $S(f) + \lambda P(f)$, where the *regularization parameter* λ adjusts the relative importance of the two terms.

As both S and P are quadratic functions, it is not surprising (and verified in Section 1.4) that the minimizing $\hat{f}_\lambda$ is indeed linear in the data Y for a given value of λ. As λ increases from 0 to ∞, the solution will pass from rough (interpolating the data) to smooth (the linear least squares fit). A subjective choice of λ was made in Figure 1.1, but it is often desirable to have an "automatic" or data-driven choice specified by some algorithm.

Depending on whether one's purpose is to obtain a summary for a given year, namely 2008, or to obtain an indication of an annual cycle, one may or may not wish to specifically require f and $\hat{f}_\lambda$ to be periodic. In the periodic case, it is natural to do the smoothing using Fourier series. If y_k and θ_k denote the kth Fourier coefficient of the observed data and unknown function respectively, then the periodic linear spline smoother takes on the simple

coordinatewise linear form $\hat{\theta}_k = y_k/(1+\lambda w_k)$ for certain known constants w_k that increase with frequency like k^4.

Interestingly, in the temperature example, the periodic and nonperiodic fits are similar, differing noticeably only within a short distance of the year boundaries. This can be understood in terms of an 'equivalent kernel' form for spline smoothing, Section 3.5.

To understand the properties of linear estimators such as $\hat{f}_\lambda$, we will later add assumptions that the noise variables Z_l are Gaussian and independent. A probability plot of residuals in fact shows that these temperature data are reasonably close to Gaussian, though not independent, since there is a clear lag-one sample autocorrelation. However the dependence appears to be short-range and appropriate adjustments for it could be made in a detailed analysis of this example.

The NMR data. Figure 1.2 shows a noisy nuclear magnetic resonance (NMR) signal sampled at $n = 2^J = 1024$ points. Note the presence both of sharp peaks and baseline noise. The additive regression model (1.1) might again be appropriate, this time with $t_l = l/n$ and perhaps with f substantially less smooth than in the first example.

The right hand panel shows the output of wavelet denoising. We give a brief description of the method using the lower panels of the figure—more detail is found in Chapter 7. The noisy signal is transformed, via an orthogonal discrete wavelet transform, into wavelet coefficients y_{jk}, organized by scale (shown vertically, from coarsest level $j = 4$ to finest level $j = J - 1 = 9$) and by location, shown horizontally, with coefficients located at $k2^{-j}$ for $k = 1, \ldots, 2^j$. Correspondingly, the unknown function values $f(t_l)$ transform into unknown wavelet coefficients θ_{jk}. In this transform domain, we obtain estimates $\hat{\theta}_{jk}$ by performing a hard thresholding

$$\hat{\theta}_{jk} = \begin{cases} y_{jk} & \text{if } |y_{jk}| > \hat{\sigma}\sqrt{2\log n}, \\ 0 & \text{otherwise} \end{cases}$$

to retain only the "large" coefficients, setting all others to zero. Here $\hat{\sigma}$ is a robust estimate of the error standard deviation[1]. The factor $\sqrt{2\log n}$ reflects the likely size of the largest of n independent zero mean standard normal random variables—Chapter 8 has a detailed discussion.

The thresholded coefficients, shown in the lower right panel, are then converted back to the time domain by the inverse discrete wavelet transform, yielding the estimated signal $\hat{f}(t_l)$ in the top right panel. The wavelet "denoising" seems to be effective at removing nearly all of the baseline noise, while preserving much of the structure of the sharp peaks.

By contrast, the spline smoothing approach cannot accomplish both these tasks at the same time. The right panel of Figure 1.3 shows a smoothing spline estimate with an automatically chosen[2] value of λ. Evidently, while the peaks are more or less retained, the spline estimate has been unable to remove all of the baseline noise.

An intuitive explanation for the different behaviors of the two estimates can be given using the idea of kernel averaging, in which a function estimate $\hat{f}(t) = n^{-1}\sum_l w_l(t)Y_l$ is obtained by averaging the data Y_l with a weight function

$$w_l(t) = h^{-1}K(h^{-1}(t - t_l)), \tag{1.2}$$

[1] using the median absolute deviation $MAD\{y_{J-1,k}\}/0.6745$, explained in Section 7.5

[2] chosen to minimize an unbiased estimate of mean squared error, Mallows C_L, explained in Section 6.4

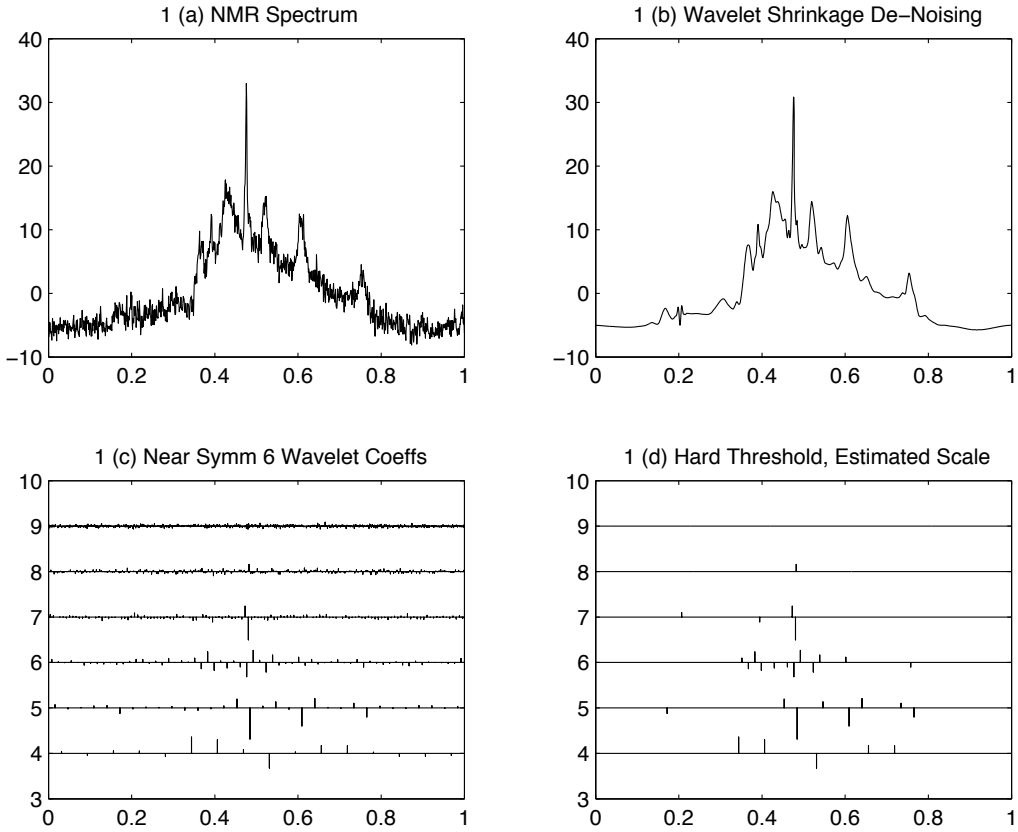

Figure 1.2 Wavelet thresholding of the NMR signal. Data originally via Chris Raphael from the laboratory of Andrew Maudsley, then at UCSF. Signal has $n = 1024$ points, discrete wavelet transform using Symmlet6 filter in `Wavelab`, coarse scale $L = 4$, hard thresholding with threshold $\hat{\sigma}\sqrt{2\log n}$ as in the text.

for a suitable kernel function K, usually non-negative and integrating to 1. The parameter h is the "bandwidth", and controls the distance over which observations contribute to the estimate at point t. (Section 3.3 has more detail.) The spline smoothing estimator, for equally spaced data, can be shown to have approximately this form, with a one-to-one correspondence between h and λ described in Chapter 6.4. A key property of the spline estimator is that the value of h does not vary with t.

By contrast, the kernel average view of the wavelet threshold estimate in Figure 1.2 shows that $h = h(t)$ depends on t strongly–the bandwidth is small in a region of sharp transients, and much larger in a zone of "stationary" behavior in which the noise dominates. This is shown schematically in Figure 1.3, but can be given a more precise form, as is done in Section 7.5.

One of the themes of this book is to explore the reasons for the difference in performance of splines and wavelet thresholding in these examples. An important ingredient can be seen by comparing the lower panels in Figure 1.2. The true signal—assuming that we can speak

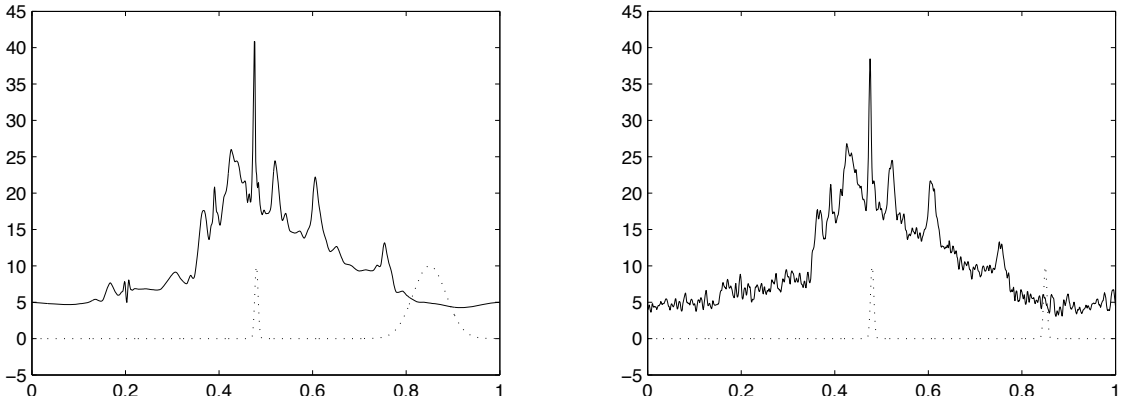

Figure 1.3 Schematic comparison of averaging kernels: The baseline dashed bell curves give qualitative indications of the size of the bandwidth h in (1.2), the equivalent kernel. In the left panel, corresponding to wavelet thresholding, the equivalent kernel depends on position, $h = h(t_l)$, whereas in the right panel, for spline smoothing, it is translation invariant.

of such a thing—appears to be concentrated in a relatively small number of wavelet coefficients, while the noise is scattered about globally and at an apparently constant standard deviation within and across levels. Thus the thresholding can literally clean out most of the noise while leaving the bulk of the signal energy, concentrated as it is in a few coefficients, largely undisturbed. This *sparsity of representation* of the signal in the wavelet transform domain is an essential property.

The example motivates a number of questions:

- *what are the properties of thresholding?* Can we develop expressions for, say, mean squared error and understand how to choose the value of the threshold?
- *when is it effective – e.g. better than linear shrinkage?* Can we compare the mean squared error of linear estimators and thresholding over various classes of functions, representing different amounts and types of smoothness?
- *what is the role of sparsity?* Can we develop quantitative measures of sparsity of representation and describe how they affect the possible mean squared error?
- *are optimality statements possible?* Can we identify assumptions on classes of functions for which it is possible to assert that linear, or threshold, estimators are, in an appropriate sense, nearly best?
- *are extensions to other settings possible?* Are there other nonparametric estimation problems, such as density estimation or linear inverse problems, in which similar phenomena appear?

Our goal will be to develop some theoretical definitions, tools and results to address these issues. A key technique throughout will be to use "sequence models", in which our methods, hypotheses and results are phrased in terms of the coefficients, θ_k or θ_{jk}, that appear when the function f is expanded in an orthogonal basis. In the NMR example, the (wavelet)

coefficients are those in the bottom panels of Figure 1.2, while in the weather data, in the periodic form, they are the Fourier coefficients.

In the next sections we turn to a first discussion of these questions in the simplest sequence model. Exactly *why* sequence models repay detailed study is taken up in Section 1.5.

1.2 A first comparison of linear methods, sparsity and thresholding

We begin with a simple model, with an n-dimensional observation vector $y \sim N_n(\theta, \epsilon^2 I)$ with θ being the unknown mean and ϵ^2 the variance, assumed known.[3] We will study a sequence form of the model,

$$y_k = \theta_k + \epsilon z_k, \qquad z_k \overset{\text{iid}}{\sim} N(0,1), \tag{1.3}$$

which may be obtained by taking coefficients in *any* orthonormal basis. We might call this a "monoresolution" model when we wish to think of what is going on at a single level in the wavelet transform domain, as in the bottom panels of Figure 1.2.

Assume now that the θ_k are random, being drawn independently from a Gaussian prior distribution $N(0, \tau^2)$. The posterior distribution of θ_k given the data y is also Gaussian, and the Bayes estimator is given by the posterior mean

$$\hat{\theta}_k = \frac{\rho}{\rho + 1} y_k, \qquad \rho = \frac{\tau^2}{\epsilon^2}. \tag{1.4}$$

The constant ρ is the squared signal-to-noise ratio. The estimator, sometimes called the Wiener filter, is optimal in the sense of minimizing the posterior expected squared error.

This analysis has two important features. First, the assumption of a Gaussian prior distribution produces an optimal estimator which is a *linear* function of the data y. Second, the estimator does not depend on the choice of orthonormal basis: both the model (1.3) 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.

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 ℓ_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".

Figure 1.4 shows part of a reconstructed signal represented in two different bases: panel a) is a subset of 2^7 wavelet coefficients θ^W, while panel b) is a subset of 2^7 Fourier coefficients θ^F. Evidently θ^W has a much sparser representation than does θ^F.

The sparsity of the coefficients in a given basis may be quantified using ℓ_p norms [4]

$$\|\theta\|_p = \left(\sum_{k=1}^{n} |\theta_k|^p \right)^{1/p}, \tag{1.5}$$

[3] The use of ϵ in place of the more common σ already betrays a later focus on "low noise" asympotics!

[4] in fact, only a *quasi*-norm for $p < 1$, Appendix C.1.

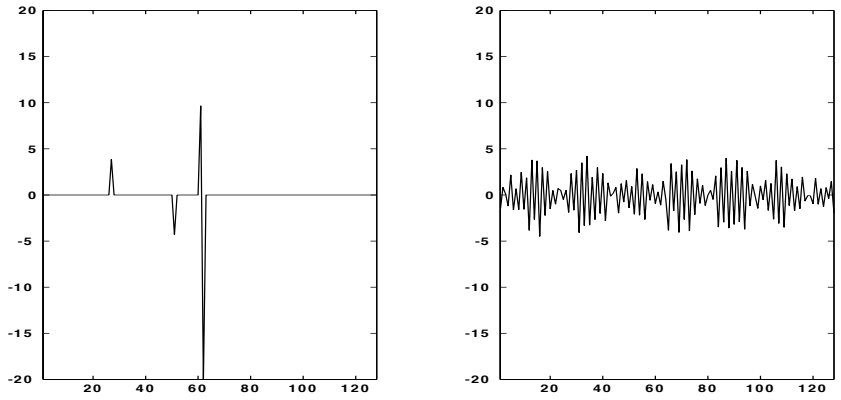

Figure 1.4 Panel (a): θ_k^W = level 7 of estimated NMR reconstruction $\hat{f}$ of Figure 1.2, while in panel (b): θ_k^F = Fourier coefficients of $\hat{f}$ at frequencies $65 \ldots 128$, both real and imaginary parts shown. While these do not represent exactly the same projections of f, the two overlap and $\|\theta^F\|_2 = 25.3 \approx 23.1 = \|\theta^W\|_2$.

which track sparsity for $p < 2$, with smaller p giving more stringent measures. Thus, while the ℓ_2 norms of our two representations are roughly equal:

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

the ℓ_1 norm of the sparser representation θ^W is smaller by a factor of 6.5:

$$\|\theta^F\|_1 = 246.5 \gg 37.9 = \|\theta^W\|_1.$$

Figure 1.5 shows that the ℓ_p-norm level 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 ℓ_p ball to have large energy (i.e. ℓ_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 ℓ_p norm.

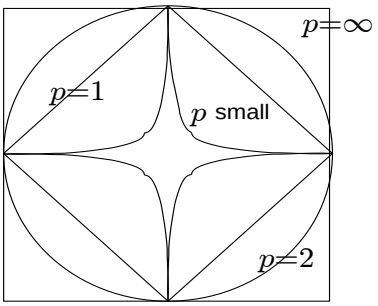

Figure 1.5 Contours of ℓ_p balls

Thus, we will use sets $\{\|\theta\|_p \le C\}$ as quantitative models for *a priori* constraints that the signal θ has an approximately sparse representation in the given basis.

How might we exploit this sparsity information in order to estimate θ better: in other words, can we estimate θ^W better than θ^F? We quantify the quality of estimator $\hat{\theta}(y)$ using Mean Squared Error (MSE):

$$E_\theta \|\hat{\theta} - \theta\|^2 = \sum_{k=1}^{n} E(\hat{\theta}_k - \theta_k)^2, \tag{1.6}$$

in which the expectation averages over the distribution of y given θ, and hence over the noise $z = (z_k)$ in (1.3).

Figure 1.6 shows an idealized case in which all θ_k are zero except for two spikes, each of size $1/2$. Assume, for simplicity here, that $\epsilon = \epsilon_n = 1/\sqrt{n}$ and that $p = C = 1$: it is thus supposed that $\sum_1^n |\theta_k| \le 1$. Consider the class of linear estimators $\hat{\theta}_c(y) = cy$, which have per co-ordinate variance $c^2\epsilon_n^2$ and squared bias $(1-c)^2\theta_k^2$. Consequently, the mean squared error (1.6)

$$\text{MSE} = \sum_1^n c^2\epsilon_n^2 + (1-c)^2\theta_k^2 = c^2 + (1-c)^2/2 = \begin{cases} 1 & c = 1 \\ 1/2 & c = 0. \end{cases}$$

The upper right panel shows the unbiased estimate with $c = 1$; this has no bias and only variance. The lower left panels shows $c = 0$ with no variance and only bias. The MSE calculation shows that no value of c leads to a linear estimate with much better error–the minimum MSE is 1/3 at $c = 1/3$. As an aside, if we were interested instead in the absolute, or ℓ_1 error $\sum_k |\hat{\theta}_k - \theta_k|$, we could visualize it using the vertical lines—again this is relatively large for all linear estimates.

In the situation of Figure 1.6, thresholding is natural. As in the preceding section, define the *hard threshold* estimator by its action on coordinates:

$$\hat{\theta}_{\lambda,k}(y) = \begin{cases} y_k & \text{if } |y_k| \ge \lambda\epsilon_n, \\ 0 & \text{otherwise.} \end{cases} \tag{1.7}$$

The lower right panel of Figure 1.6 uses a threshold of $\lambda\epsilon_n = 2.4\epsilon_n = 0.3$. For the particular configuration of true means θ_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 coefficients to zero except for the small fraction of noise that exceeds the threshold.

As is verified in more detail in Exercise 1.2, the MSE of $\hat{\theta}_\lambda$ consists essentially of two variance contributions each of ϵ_n^2 from the two spikes, and $n-2$ squared bias contributions of $2\epsilon_n^2\lambda\phi(\lambda)$ from the zero components, where $\phi(\lambda) = (2\pi)^{-1/2}e^{-\lambda^2/2}$ denotes the standard Gaussian density. Hence, in the two spike setting,

$$\begin{aligned} E_\theta\|\hat{\theta}_\lambda - \theta\|^2 &\approx 2\epsilon_n^2 + 2(n-2)\epsilon_n^2\lambda\phi(\lambda) \\ &\approx 2n^{-1} + 2\lambda\phi(\lambda) \approx 0.139 \end{aligned} \tag{1.8}$$

when $n = 64$ and $\lambda = 2.4$. This mean squared error is of course much better than for any of the linear estimators.

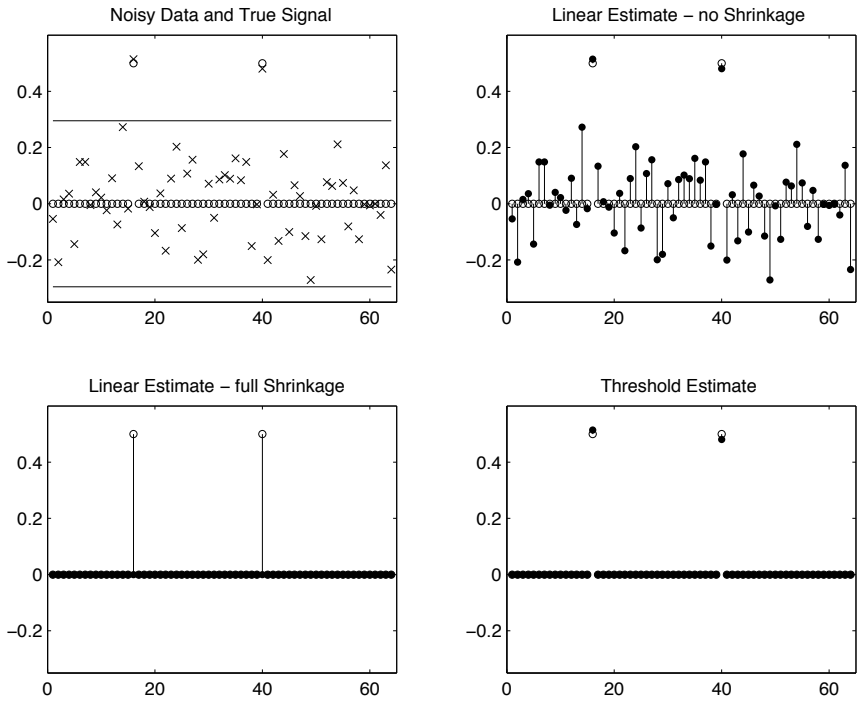

Figure 1.6 (a) Visualization of model (1.3): open circles are unknown values θ_k, crosses are observed data y_k. In the other panels, solid circles show various estimators $\hat{\theta}$, for $k = 1, \ldots, n = 64$. Horizontal lines are thresholds at $\lambda = 2.4\epsilon_n = 0.3$. (b) Vertical lines indicate absolute errors $|\hat{\theta}_{1,k} - \theta_k|$ made by leaving the data alone: $\hat{\theta}_1(y) = y$. (c) Corresponding absolute errors for the zero estimator $\hat{\theta}_0(y) = 0$. (d) Much smaller errors due to hard thresholding at $\lambda = 0.3$.

1.3 A game theoretic model and minimaxity

The skeptic will object that the configuration of Figure 1.6 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 a probability distribution π for θ subject only to the sparsity constraint that $E_\pi \|\theta\|_1 \leq 1$.

(iii) The payoff–the loss to the statistician–is calculated as the expected mean squared error of $\hat{\theta}(y)$ when θ is chosen according to π and then the observed data y is drawn from model (1.3): $y = \theta + \epsilon_n z$ for $z \sim N_n(0, I)$. Thus the expected loss, or *risk*, now averages

over *both* θ *and* y:

$$B(\hat{\theta}, \pi) = E_\pi E_\theta \|\hat{\theta}(y) - \theta\|_2^2. \tag{1.9}$$

(Here, E_θ denotes expectation over y given θ, and E_π expectation over $\theta \sim \pi$, Section 4.1.) Of course, the Statistician tries to minimize the risk and Nature to maximize it.

Classical work in statistical decision theory (Wald, 1950; Le Cam, 1986), Chapter 4 and Appendix A, 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*:

$$R_n = \inf_{\hat{\theta}} \sup_{\pi} B(\hat{\theta}, \pi) = \sup_{\pi} \inf_{\hat{\theta}} B(\hat{\theta}, \pi). \tag{1.10}$$

An estimator $\hat{\theta}^*$ attaining the left hand infimum in (1.10) is called a *minimax* strategy or *estimator* for player I, while a prior distribution π^* 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 1.7: if Nature uses π^*, 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 π^*.

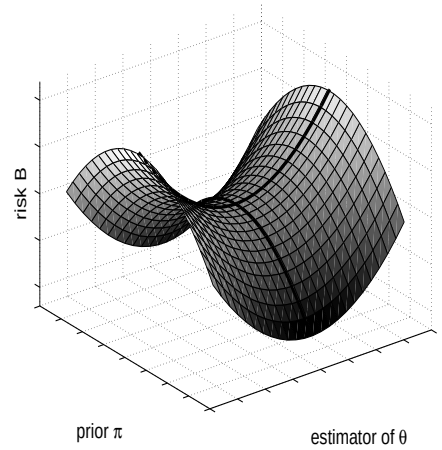

Figure 1.7 Left side lower axis: strategies π for Nature. Right side lower axis: strategies $\hat{\theta}$ for the Statistician. Vertical axis: payoff $B(\hat{\theta}, \pi)$ from the Statistician to Nature. The saddlepoint indicates a pair $(\hat{\theta}^*, \pi^*)$ of optimal strategies.

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. Indeed, as will be seen in Section 13.5, an *approximately* least favorable distribution is given by drawing the individual coordinates $\theta_k, k = 1, \ldots, n$

	Prior Constraint	
	traditional (ℓ_2)	*sparsity* (ℓ_1)
minimax estimator	linear	thresholding
least favorable π	Gaussian	sparse
minimax MSE	$= 1/2$	$\sim \sqrt{\frac{\log n}{n}}$

Table 1.1 *Comparison of structure of optimal strategies in the monoresolution game under traditional and sparsity assumptions.*

independently from a *two point* distribution with

$$\theta_k = \begin{cases} \epsilon_n \sqrt{\log n} & \text{with probability } \alpha_n \doteq 1/\sqrt{n \log n} \\ 0 & \text{otherwise.} \end{cases} \tag{1.11}$$

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 $n\alpha_n = \sqrt{n/\log n}$ 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 remaining, more numerous, $n - \sqrt{n/\log n}$ observations that are pure noise. Of course, what makes this distribution difficult for Player I, the Statistician, is that the *locations* of the non-zero components are random as well.

It can also be shown, Chapter 13, that an approximately minimax estimator for this setting is given by the hard thresholding rule described earlier, but with threshold given roughly by $\lambda_n = \epsilon_n \sqrt{\log(n \log n)}$. This estimate asymptotically achieves the minimax value

$$R_n \sim \sqrt{\log n / n}$$

for MSE. [Exercise 1.3 bounds the risk $B(\hat{\theta}_{\lambda_n}, \pi)$, (1.9), for this prior, hinting at how this minimax value arises]. It can also be verified that no *linear* estimator can achieve a risk less than $1/2$ if Nature chooses a suitably uncooperative probability distribution for θ, Theorem 9.5 and (9.29).

In the setting of the previous section with $n = 64$ and $\epsilon_n = 1/\sqrt{n}$, we find that the non-zero magnitudes $\epsilon_n \sqrt{\log n} = 0.255$ and the expected non-zero number $n\alpha_n = 3.92$. Finally, the threshold value $\epsilon_n \sqrt{\log(n \log n)} = .295$.

This—and any—statistical decision problem make a large number of assumptions, including values of parameters that typically are not known in practice. We will return later to discuss the virtues and vices of the minimax formulation. For now, it is perhaps the qualitative features of this solution that most deserve comment. Had we worked with simply a signal to noise constraint, $E_\pi \|\theta\|_2^2 \le 1$, say, we would have obtained a Gaussian prior distribution $N(0, \tau_n^2)$ as being approximately least favorable and the linear Wiener filter (1.4) with $\epsilon_n^2 = \tau_n^2 = 1/n$ as an approximately minimax estimator. As may be seen from the summary in Table 1.1, the imposition of a sparsity constraint $E_\pi \|\theta\|_1 \le 1$ reflects additional *a priori* information and yields great improvements in the quality of possible estimation, and produces optimal strategies that take us far away from Gaussian priors and linear methods.

1.4 The Gaussian Sequence Model

In this section we introduce the general sequence model, an extension of (1.3) that will be our main focus of study. The observed data are $y = (y_i)$ for i in a discrete index set $\mathcal{I}$ such as the positive integers $\mathbb{N}$. It is assumed that the components y_i are statistically independent of one another, and follow Gaussian, or normal, distributions with unknown means θ_i and known positive standard deviations $\epsilon\varrho_i$. Thus the sequence model may be written as

$$y_i = \theta_i + \epsilon\varrho_i z_i, \qquad z_i \overset{i.i.d}{\sim} N(0,1), \qquad i \in \mathcal{I}. \tag{1.12}$$

The index set will typically be a singleton, $\mathcal{I} = \{1\}$, finite $\mathcal{I} = \{1,\dots,n\}$, or infinite $\mathcal{I} = \mathbb{N}$. Multidimensional index sets, such as $\{1,\dots,n\}^d$ or $\mathbb{N}^d$ are certainly allowed, but will appear only occasionally. The scale parameter ϵ sets the level of the noise, and in some settings will be assumed to be small.

In particular, we often focus on the model with $\mathcal{I} = \{1,\dots,n\}$. Although this model is finite dimensional, it is actually non-parametric in character since the dimension of the unknown parameter equals that of the data. In addition, we often consider asymptotics as $n \to \infty$.

We turn to a first discussion of models motivating, or leading to, (1.12)—further examples and details are given in Chapters 2 and 3.

Nonparametric regression. In the previous two sections, θ was a vector with no necessary relation among its components. Now we imagine an unknown function $f(t)$. The independent variable t is thought of as low dimensional (1 for signals, 2 for images, 3 for volumetric fields etc.); indeed we largely confine attention to functions of a single variable, say time, in a bounded interval, say $[0,1]$. In a sampled-data model, we might have points $0 \le t_1 \le \cdots \le t_n \le 1$, and

$$Y_l = f(t_l) + \sigma Z_l, \qquad Z_l \overset{iid}{\sim} N(0,1). \tag{1.13}$$

This is the model for the two examples of Section 1.1 with the i.i.d. Gaussian assumption added.

We can regard Y, Z and $\mathbf{f} = (f(t_l))$ as vectors in $\mathbb{R}^n$, viewed as the "time domain" and endowed with a normalized inner product $\langle \mathbf{a}, \mathbf{b}\rangle_n = (1/n)\sum_{l=1}^n a_l b_l$, and corresponding norm $\|\cdot\|_{2,n}$. Let $\{\boldsymbol{\varphi}_i\}$ be an arbitrary orthonormal basis with respect to $\langle\cdot,\cdot\rangle_n$. For example, if the t_l were equally spaced, this might be the discrete Fourier basis of sines and cosines. In general, form the inner products

$$y_k = \langle Y, \boldsymbol{\varphi}_k\rangle_n, \qquad \theta_k = \langle \mathbf{f}, \boldsymbol{\varphi}_k\rangle_n \qquad z_k = \sqrt{n}\langle Z, \boldsymbol{\varphi}_k\rangle_n. \tag{1.14}$$

One can check easily that under model (1.13), the z_k are iid $N(0,1)$, so that (y_k) satisfies (1.3) with $\epsilon = \sigma/\sqrt{n}$.

We illustrate the reduction to sequence form with the smoothing spline estimator used in Section 1.1, and so we suppose that an estimator $\hat{f}$ of f in (1.13) is obtained by minimizing the penalized sum of squares $S(f) + \lambda P(f)$, or more explicitly

$$Q(f) = n^{-1}\sum_{l=1}^n [Y_l - f(t_l)]^2 + \lambda \int_0^1 (f'')^2. \tag{1.15}$$

The account here is brief; for much more detail see Green and Silverman (1994) and the chapter notes.

It turns out that a unique minimizer exists and belongs to the space S of "natural cubic splines"–twice continuously differentiable functions that are formed from cubic polynomials on each interval $[t_l, t_{l+1}]$ and are furthermore linear on the outermost intervals $[0, t_1]$ and $[t_n, 1]$. Equally remarkably, the space S has dimension exactly n, and possesses a special orthonormal basis, the *Demmler-Reinsch* basis. This basis consists of functions $\varphi_k(t)$—and associated vectors $\boldsymbol{\varphi}_k = (\varphi_k(t_l))$—that are simultaneously orthogonal both on the set of sampling points and on the unit interval:

$$\langle \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_k \rangle_n = \delta_{jk} \qquad \text{and} \qquad \int_0^1 \varphi_j'' \varphi_k'' = w_k \delta_{jk}. \tag{1.16}$$

[The Kronecker delta $\delta_{jk} = 1$ if $j = k$ and 0 otherwise.] 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 $k \geq 3$, it can be shown that φ_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 sort of Fourier basis that depends on the knot locations $\{t_l\}$.

This double orthogonality allows us to rewrite the criterion $Q(f)$, for $f \in \mathsf{S}$, in terms of coefficients in the Demmler-Reinsch basis:

$$Q(\theta) = \sum_1^n (y_k - \theta_k)^2 + \lambda \sum_1^n w_k \theta_k^2. \tag{1.17}$$

(Exercise 1.4.) The charm is that this can now readily be minimized term by term to yield the sequence model expression for the smoothing spline estimate $\hat{\theta}_{SS}$:

$$\hat{\theta}_{SS,k} = c_{\lambda k} y_k = \frac{1}{1 + \lambda w_k} y_k. \tag{1.18}$$

The estimator is thus linear in the data and operates *co-ordinatewise*. It achieves its smoothing aspect by shrinking the higher "frequencies" by successively larger amounts dictated by the increasing weights λw_k. In the original time domain,

$$\hat{\mathbf{f}} = \sum_k \hat{\theta}_{SS,k} \boldsymbol{\varphi}_k = \sum_k c_{\lambda k} y_k \boldsymbol{\varphi}_k. \tag{1.19}$$

There is no shrinkage on the constant and linear terms: $c_{\lambda 1} = c_{\lambda 2} = 1$, but for $k \geq 3$, the shrinkage factor $c_{\lambda k} < 1$ and decreases with increasing frequency. Large values of smoothing parameter λ 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 $U = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n]/\sqrt{n}$. Then $Y = \sqrt{n} U y$ and $\mathbf{f} = \sqrt{n} U \theta$, and so

$$\hat{\mathbf{f}} = \sqrt{n} U \hat{\theta} = U c_\lambda U^T Y, \qquad c_\lambda = \text{diag}\,(c_{\lambda k}). \tag{1.20}$$

Notice that the change of basis matrix U does not depend on λ. Thus, many important

aspects of the spline smoothing problem, such as the issue of choosing λ well from data, can be studied in the diagonal sequence form that the quasi-Fourier basis provides.

Software packages, such as `spline.smooth` in `R`, 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 (1.20).

We have so far emphasized structure that exists whether or not the points t_l are equally spaced. If, however, $t_l = l/n$ *and* it is assumed that f is periodic, then everything in the approach above has an explicit form in the Fourier basis—Section 3.4.

Continuous Gaussian white noise model. Instead of sampling a function at a discrete set of points, we might suppose that it can be observed—with noise!—throughout the entire interval. This leads to the central model to be studied in this book:

$$Y(t) = \int_0^t f(s)ds + \epsilon W(t), \qquad 0 \le t \le 1, \tag{1.21}$$

which we will sometimes write in an equivalent form, in terms of instantaneous increments

$$dY(t) = f(t)dt + \epsilon dW(t), \qquad 0 \le t \le 1. \tag{1.22}$$

The observational noise consists of a standard Brownian motion W, scaled by the known noise level ϵ. For an arbitrary square integrable function g on $[0, 1]$, we therefore write

$$\int_0^1 g(t)dY(t) = \int_0^1 g(t)f(t)dt + \epsilon \int_0^1 g(t)dW(t). \tag{1.23}$$

The third integral features a deterministic function g and a Brownian increment dW and is known as a Wiener integral. We need only a few properties of standard Brownian motion and Wiener integrals, which are recalled in Appendix C.13.

The function Y is observed, and we seek to recover the unknown function f, assumed to be square integrable: $f \in L_2[0, 1]$, for example using the integrated squared error loss

$$\|\hat{f} - f\|_{L_2}^2 = \int_0^1 (\hat{f} - f)^2.$$

To rewrite the model in sequence form, we may take any orthonormal basis $\{\varphi_i(t)\}$ for $L_2[0, 1]$. Examples include the Fourier basis, or any of the classes of orthonormal wavelet bases to be discussed later. To set notation for the coefficients, we write

$$y_i = Y(\varphi_i) = \int_0^1 \varphi_i dY, \qquad \theta_i = \langle f, \varphi_i \rangle = \int_0^1 f\varphi_i, \qquad z_i = W(\varphi_i) = \int_0^1 \varphi_i dW. \tag{1.24}$$

From the stationary and independent increments properties of Brownian motion, the Wiener integrals z_i are Gaussian variables that have mean 0 and are uncorrelated:

$$\mathrm{Cov}(z_i, z_j) = E\Big[\int_0^1 \varphi_i dW \cdot \int_0^1 \varphi_j dW\Big] = \int_0^1 \varphi_i\varphi_j \, dt = \delta_{ij}.$$

As a result, the continuous Gaussian model is entirely equivalent to the constant variance

sequence model (1.3). The Parseval relation, (C.1), converts squared error in the function domain to the analog in the sequence setting:

$$\int_0^1 (\hat{f} - f)^2 = \sum_i (\hat{\theta}_i - \theta_i)^2. \tag{1.25}$$

Linking regression and white noise models. Heuristically, the connection between (1.13) and (1.21) arises by forming the partial sum process of the discrete data, now assumed to be equally spaced, $t_l = l/n$:

$$Y_n(t) \triangleq \frac{1}{n}\sum_1^{[nt]} Y_l = \frac{1}{n}\sum_1^{[nt]} f(l/n) + \frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_1^{[nt]} Z_l. \tag{1.26}$$

The signal term is a Riemann sum approximating $\int_0^t f$, and the error term $n^{-\frac{1}{2}}\sum^{[nt]} Z_l$ 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 (1.21), we see that, informally, the processes $Y_{\epsilon(n)}(t)$ and $Y_n(t)$ merge as $n \to \infty$. A formal statement and proof of this result is given in Chapter 3.11, using the notion of asymptotic equivalence of statistical problems, which implies closeness of risks for all decision problems with bounded loss. Here we simply observe that heuristically there is convergence of mean average squared errors. Indeed, for fixed functions $\hat{f}$ and $f \in L_2[0,1]$:

$$\|\hat{f} - f\|_{2,n}^2 = n^{-1}\sum_1^n [\hat{f}(l/n) - f(l/n)]^2 \to \int_0^1 [\hat{f} - f]^2.$$

Non white noise models. So far we have discussed only the constant variance subclass of models (1.12) in which $\lambda_i \equiv 1$. The scope of (1.12) is considerably broadened by allowing unequal $\lambda_i > 0$. Here we make only a few remarks, deferring further discussion and examples to Chapters 2 and 3.

When the index set I is finite, say $\{1, \ldots, n\}$, two classes of multivariate Gaussian models lead to (1.12):

(i) $Y \sim N(\theta, \epsilon^2\Sigma)$, by transforming to an orthogonal basis that diagonalizes Σ, so that (ϱ_i^2) are the eigenvalues of Σ, and

(ii) $Y \sim N(A\theta, \epsilon^2 I)$, by using the singular value decomposition of $A = \sum_i b_i u_i v_i^T$ and setting $y_i = b_i^{-1} Y_i$, so that $\varrho_i = b_i^{-1}$ are the inverse singular values.

When the index set I is countably infinite, case (i) corresponds to a Gaussian process with unknown mean function f and the sequence form is obtained from the Karhunen-Loève transform (Section 3.10). Case (ii) corresponds to observations in a linear inverse problem with additive noise, $Y = Af + \epsilon Z$, in which we do not observe f but rather its image Af after the action of a linear operator A, representing some form of integration, smoothing or blurring. The conversion to sequence form is again obtained using a singular value decomposition, cf. Chapter 3.

1.5 Why study the sequence model?

While the sequence models (1.3) and (1.12) are certainly idealizations, there are several reasons why they repay detailed study.

(i) simplicity. By focusing on sequences of independent Gaussian variables, we can often do exact calculations. Generally, it turns out that all the issues are fundamental rather than merely technical. In parametric statistics, the analogy would be with study of the multivariate normal model after use of the central limit theorem and other asymptotic approximations.

(ii) depth. The model makes it possible to focus directly on important and profound phenomena, such as the Stein effect, in which maximum likelihood estimates of three or more mean parameters can be (often significantly) improved by shrinkage toward a point or subspace. Similarly, the "concentration of measure" phenomenon for product measures in high dimensional spaces (such as our Gaussian error distributions) plays an important role.

(iii) relevance. The sequence models and estimators used in them turn out to be close *enough* to actual methods to yield useful insights. Thus the contrast between linear estimators and thresholding is able to explain more or less fully some practically important phenomena in function estimation.

The finite dimensional multivariate normal model is the foundation of parametric statistical theory. For nonparametric statistics, the continuous 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 and Khasminskii (1980); (n.d., for example) have argued, the difficulties thrown up by the "signal+noise" model are essential rather than technical in nature.

1.6 Plan of the book

In the Canberra temperature and NMR data examples we saw that linear spline and nonlinear wavelet threshold estimators respectively were reasonably satisfactory, at least so far as one can tell without knowledge of the "ground truth". The examples illustrate a basic point that, in function estimation, as elsewhere in statistics, an optimal or at least good choice of method will depend on the circumstances of the problem.

The theory to be developed in this book will formulate classes of assumptions under which linear estimators can perform well, and then move to circumstances in which coordinatewise thresholding is optimal, either in "monoresolution" or "multiresolution" settings.

The chapters are grouped into two parts. In the first, chapters 2–9 contain a sampling of material of broadest interest. In the second, Chapters 10 –15 then go into greater detail about optimality results for thresholding-type estimators in both 'monoresolution' and multiresolution models.

We use the ideas and tools of statistical decision theory, particularly Bayes rules and minimaxity, throughout; introductory material appears in Chapters 2–4 and especially in Chapter 4. Chapters 5–6 focus primarily on optimality properties of linear estimators, especially using geometric properties of parameter spaces such as hyperrectangles and ellipsoids. Pinsker's theorem on the asymptotic optimality of linear rules over ellipsoids is discussed in Chapter 5. Chapter 6 introduces the notion of adaptive optimality—the ability of an estimaor to perform 'optimally' over a scale of parameter spaces without having to depend on *a priori* assumptions about parameters of those spaces. The James-Stein estimator is seen to lead to a class of adaptively minimax estimators that is quite similar to certain smoothing spline or kernel estimators that are commonly used in practice.

The focus then turns to the phenomena of sparsity and non-linear estimation via coordinatewise thresholding. To set the stage, Chapter 7 provides a primer on orthonormal wavelet bases and wavelet thresholding estimation. Chapter 8 focuses on the properties of thresholding estimators in the "sparse normal means" model: $y \sim N_n(\theta, \sigma^2 I)$ and the unknown vector θ is assumed to be sparse. Chapter 9 explores the consequences of these thresholding results for wavelet shrinkage estimation, highlighting the connection between sparsity, non-linear approximation and statistical estimation.

Part II is structured around a theme already implicit in Chapters 8 and 9: while wavelet bases are specifically designed to analyze signals using multiple levels of resolution, it is helpful to study initially what happens with thresholding etc. at a single resolution scale both for other applications, and before assembling the results across several scales to draw conclusions for function estimation.

Thus Chapters 10–14 are organized around two strands: the first strand works at a single or mono-resolution level, while the second develops the consequences in multiresolution models. Except in Chapter 10, each strand gets its own chapter. Three different approachs are explored—each offers a different tradeoff between generality, sharpness of optimality, and complexity of argument. We consider in turn

(i) optimal recovery and 'universal' thresholds (Ch. 10)
(ii) penalized model selection (Chs. 11, 12)
(iii) minimax-Bayes optimal methods (Chs. 13, 14)

The Epilogue, Chapter 15 has two goals. The first is to provide some detail on the comparison between discrete and continuous models. The second is to mention some recent related areas of work not covered in the text. The Appendices collect background material on the minimax theorem, functional classes, smoothness and wavelet decompositions.

1.7 Notes

§1.4. Although our main interest in the Demmler and Reinsch (1975) [DR] basis lies in its properties, for completeness, we provide a little more information on its construction. More detail for our penalty $P(f) = \int (f'')^2$ appears in Green and Silverman (1994) [GS]; here we make more explicit the connection between the two discussions. Indeed, [GS] describe tridiagonal matrices Q and R, built respectively from divided differences and from inner products of linear B-splines. The Demmler-Reinsch weights w_k and basis vectors $\boldsymbol{\varphi}_k$ are given respectively by eigenvalues and vectors of the matrix $K = QR^{-1}Q^T$. The functions $\varphi_k(t)$ are derived from $\boldsymbol{\varphi}_k$ using the natural interpolating spline ($A^+\boldsymbol{\varphi}_k$ in DR) given in [GS] Section 2.4.2.

Related books and monographs. The book of Ibragimov and Khasminskii (1981), along with their many research papers has had great influence in establishing the central role of the signal in Gaussian noise model. Textbooks on nonparametric estimation include Efromovich (1999) and Tsybakov (2009), which include coverage of Gaussian models but range more widely, and Wasserman (2006) which is even broader, but omits proofs.

Closer to the research level are the St. Flour courses by Nemirovski (2000) and Massart (2007). Neither are primarily focused on the sequence model, but do overlap in content with some of the chapters of this book. Ingster and Suslina (2003) focuses largely on hypothesis testing in Gaussian sequence models. References to books focusing on wavelets and statistics are collected in the notes to Chapter 7.

Exercises

1.1 *(Limiting cases of ℓ_p norms.)* Show that

$$\|\theta\|_\infty := \max_k |\theta_k| \qquad = \lim_{p\to\infty} \Big(\sum_{k=1}^n |\theta_k|^p \Big)^{1/p}, \tag{1.27}$$

$$\|\theta\|_0 := \#\{k : \theta_k \neq 0\} = \lim_{p\to 0} \quad \sum_{k=1}^n |\theta_k|^p. \tag{1.28}$$

[$\|\theta\|_\infty$ is a legitimate norm on $\mathbb{R}^n$, while $\|\theta\|_0$ is not: note the absence of the pth root in the limit. Nevertheless it is often informally called the ℓ_0 norm.]

1.2 *(Approximate MSE of thresholding for two spike signal.)* Suppose that $y \sim N_n(\theta, \epsilon^2 I)$, compare (1.3), and that $\hat\theta_\lambda = (\hat\theta_{\lambda,k})$ denotes hard thresholding, (1.7).

(a) Verify the MSE decomposition

$$E(\hat\theta_{\lambda,k} - \theta_k)^2 = E\{(y_k - \theta_k)^2, |y_k| \geq \lambda\epsilon_n\} + \theta_k^2 P\{|y_k| \leq \lambda\epsilon_n\}. \tag{1.29}$$

(b) When $\theta_k = 0$, so that $y_k = \epsilon_n z_k$, show that, as $\lambda \to \infty$, the MSE

$$E\hat\theta_{\lambda,k}^2 = \epsilon_n^2 E\{z_k^2, |z_k| > \lambda\} \sim 2\epsilon_n^2 \lambda \phi(\lambda), \tag{1.30}$$

in the usual sense that the ratio of the two sides approaches one.

(c) When θ_k is large relative to $\lambda\epsilon_n$, show that the MSE is approximately $E(y_k - \theta_k)^2 = \epsilon_n^2$.

(d) Conclude that (1.8) holds.

1.3 *(Risk bound for two point prior.)* Let $y \sim N_n(\theta, \epsilon_n^2 I)$ and $\hat\theta_\lambda$ denote the hard thresholding rule (1.7). Let $r(\lambda, \theta_k; \epsilon_n) = E(\theta_{\lambda,k} - \theta_k)^2$ denote the risk (mean squared error) in a single co-ordinate.

(i) for the two point prior given in (1.11), express the Bayes risk $B(\hat\theta_\lambda, \pi) = E_\pi E_\theta \|\hat\theta_\lambda - \theta\|_2^2$ in terms of the risk function $\theta \to r(\lambda, \theta; \epsilon_n)$.

(ii) Using (1.29), derive the bound

$$r(\lambda, \mu\epsilon_n; \epsilon_n) \leq (1 + \mu^2)\epsilon_n^2.$$

(iii) Using also (1.30), verify that for $\lambda = \epsilon_n \sqrt{\log(n \log n)}$,

$$B(\hat\theta_\lambda, \pi) \leq \sqrt{\log n/n} \cdot (1 + o(1)).$$

[This gives the risk for a 'typical configuration' of θ drawn from the least favorable prior (1.11). It does not yet show that the minimax risk R_n satisfies this bound. For a simple, but slightly suboptimal, bound see Theorem 8.1; for the actual argument, Theorems 13.7, 13.9 and 13.17].

1.4 *(Sequence form of spline penalized sum of squares.)* Take as given the fact that the minimizer of (1.15) belongs to the space S and hence has a representation $f(t) = \sum_{k=1}^n \theta_k \varphi_k(t)$ in terms of the Demmler-Reinsch basis $\{\varphi_k(t)\}_{k=1}^n$. Use the definitions (1.14) and orthogonality relations (1.16) to verify that

(i) $\mathbf{f} = (f(t_l))$ equals $\sum_k \theta_k \boldsymbol{\varphi}_k$ and $\|Y - \mathbf{f}\|_{2,n}^2 = \sum_{k=1}^n (y_k - \theta_k)^2$.

(ii) $\int (f'')^2 = \sum_1^n w_k \theta_k^2$ and hence that $Q(f) = Q(\theta)$ given by (1.17).

2

The multivariate normal distribution

> We know not to what are due the accidental errors, and precisely because we do not know, we are aware they obey the law of Gauss. Such is the paradox. (Henri Poincaré, *The Foundations of Science.*)

Estimation of the mean of a multivariate normal distribution, $y \sim N_n(\theta, \sigma_0^2 I)$, is the elemental estimation problem of the theory of statistics. In parametric statistics it is sometimes plausible as a model in its own right, but more often occurs–perhaps after transformation–as a large sample approximation to the problem of estimating a finite dimensional parameter governing a smooth family of probability densities.

In nonparametric statistics, it serves as a building block for the study of the infinite dimensional Gaussian sequence model and its cousins, to be introduced in the next chapter. Indeed, a recurring theme in this book is that methods and understanding developed in the finite dimensional Gaussian location model can be profitably transferred to nonparametric estimation.

It is therefore natural to start with some definitions and properties of the finite Gaussian location model for later use. Section 2.1 introduces the location model itself, and an extension to known diagonal covariance that later allows a treatment of certain correlated noise and linear inverse problem models.

Two important methods of generating estimators, regularization and Bayes rules, appear in Sections 2.2 and 2.3. Although both approaches can yield the same estimators, the distinction in point of view is helpful. Linear estimators arise from quadratic penalties/Gaussian priors, and the important conjugate prior formulas are presented. Non-linear estimators arise from ℓ_q penalties for $q < 2$, including the soft and hard thresholding rules, and from sparse mixture priors that place atoms at 0, Section 2.4.

Section 2.5 begins the comparative study of estimators through their mean squared error properties. The bias and variance of linear estimators are derived and it is shown that sensible linear estimators in fact *must* shrink the raw data. The James-Stein estimator explodes any hope that we can get by with linear methods, let alone the maximum likelihood estimator. Its properties are cleanly derived using Stein's unbiased estimator of risk; this is done in Section 2.6.

Soft thresholding consists of pulling each co-ordinate y_i towards, but not past, 0 by a threshold amount λ. Section 2.7 develops some of its properties, including a simple oracle inequality which already shows that thresholding outperforms James-Stein shrinkage on sparse signals, while James-Stein can win in other 'dense' settings.

Section 2.8 turns from risk comparison to probability inequalities on the tails of Lipschitz functions of a multivariate normal vector. This "concentration" inequality is often useful in high dimensional estimation theory; the derivation given has points in common with that of Stein's unbiased risk estimate.

Section 2.9 makes some remarks on more general linear models $Y = A\beta + \sigma e$ with correlated Gaussian errors e, and how some of these can be transformed to diagonal sequence model form.

2.1 Sequence models

The simplest finite white Gaussian sequence model has

$$y_i = \theta_i + \epsilon z_i, \qquad i = 1, \ldots, n. \tag{2.1}$$

Here (y_i) represents the observed data. The signal (θ_i) is unknown—there are n unknown parameters. The (z_i) are independent $N(0,1)$ noise or 'error' variables, and ϵ is the noise level, which for simplicity we generally assume to be known. The model is called *white* because the noise level ϵ is the same at all indices, which often represent increasing frequencies. Typically we will be interested in estimation of θ.

Equation (2.1) can also be 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—one justifications is recalled in Exercise 2.26. We write $\phi_\epsilon(y-\theta) = \prod_i \phi_\epsilon(y_i-\theta_i)$ for the joint density of (y_i) with respect to Lebesgue measure. The univariate densities $\phi_\epsilon(y_i) = (2\pi\epsilon^2)^{-1/2}\exp\{-y_i^2/2\epsilon^2\}$. We put $\phi = \phi_1$ and $\Phi(y) = \int_{-\infty}^{y}\phi(s)ds$ for the standard normal density and cumulative distribution function, and $\tilde{\Phi}(y) = 1 - \Phi(y)$ for the right tail complement.

Two generalizations considerably extend the scope of the finite sequence model. In the first, corresponding to indirect or inverse estimation,

$$y_i = \alpha_i\theta_i + \epsilon z_i, \qquad i = 1, \ldots, n, \tag{2.2}$$

the constants α_i are known and positive. In the second, relevant to correlated noise,

$$y_i = \theta_i + \epsilon\varrho_i z_i, \qquad i = 1, \ldots, n. \tag{2.3}$$

Here again the constants ϱ_i are known and positive. Of course these two models are equivalent in the sense that dividing by α_i in the former and setting $\varrho_i = 1/\alpha_i$ and $y_i' = y_i/\alpha_i$ yields the latter. In this sense, we may regard (2.3) as describing the general case. In Section 2.9, we review some Gaussian linear models that can be reduced to one of these sequence forms.

Among the issues to be addressed are

(i) we imagine (θ_i) to be "high dimensional". In particular, as ϵ decreases, the number of parameters $n = n(\epsilon)$ may increase. This makes the problem fundamentally *nonparametric*.
(ii) what are the effects of (α_i) or (ϱ_i), i.e. the consequences of indirect estimation, or correlated noise, on the ability to recover θ?
(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 attainable error of estimation and estimators that (more or less) achieve these bounds?

Before starting in earnest, we briefly introduce the Stein effect, a phenomenon mentioned already in Section 1.5 as basic to high-dimensional estimation, as motivation for much of the work of this chapter.

Perhaps the obvious first choice of estimator of θ in model (2.1) is $\hat{\theta}_I(y) = y$. It is the least squares and maximum likelihood estimator. It is unbiased, $E_\theta \hat{\theta}_I = \theta$, and its mean squared error, (1.6), is constant: $E_\theta \|\hat{\theta}_I - \theta\|^2 = n\epsilon^2 = R_n$, say.

However it is easy to greatly improve on the MLE when the dimension n is large. Consider first the linear shrinkage estimators $\hat{\theta}_c(y) = cy$ for $c < 1$, introduced in Section 1.2: we saw that the MSE

$$E_\theta \|\hat{\theta}_c - \theta\|^2 = c^2 n\epsilon^2 + (1-c)^2 \|\theta\|^2.$$

This MSE is less than R_n if $\|\theta\|^2 < \gamma_c R_n$ for $\gamma_c = (1+c)/(1-c)$ and can be *much* smaller at $\theta = 0$, compare Figure 2.1.

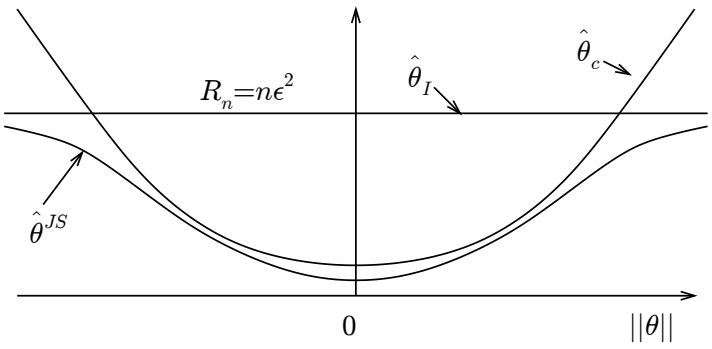

Figure 2.1 Schematic comparison of mean squared error functions for the unbiased estimator (MLE) $\hat{\theta}_I$, a linear shrinkage estimator $\hat{\theta}_c$ and James-Stein estimator $\hat{\theta}^{JS}$.

Of course, to be assured of improvement, we must know in advance that $\|\theta\|$ is small, otherwise the estimator may be (much) worse, so the example is not entirely convincing.

The great suprise of the Stein effect is that by allowing c to depend on y, namely $\hat{\theta} = c(y)y$, we can obtain MSE improvement *for all* $\theta \in \mathbb{R}^n$. Indeed,James and Stein (1961), building on Stein (1956), showed that if $c^{JS}(y) = (n-2)\epsilon^2/\|y\|^2$, then $\hat{\theta}^{JS} = c^{JS}(y)y$ satisfies

$$E_\theta \|\hat{\theta}^{JS} - \theta\|^2 < n\epsilon^2 \qquad \text{for all } \theta \in \mathbb{R}^n.$$

A proof is given in Section 2.6.

The *magnitude* of the improvement depends strongly on $\|\theta\|$: for $\theta = 0$, the MSE is less than $2\epsilon^2$, offering a huge reduction from $n\epsilon^2$. More generally, Section 2.6 shows that

$$E_\theta \|\hat{\theta}^{JS} - \theta\|^2 \le 2\epsilon^2 + \frac{(n-2)\epsilon^2 \|\theta\|^2}{(n-2)\epsilon^2 + \|\theta\|^2}.$$

Thus, like the linear shrinkage estimators, $\hat{\theta}^{JS}$ offers great MSE improvement near 0, but unlike the linear estimator, the improvement persists, albeit of small magnitude, even if $\|\theta\|$ is large. This is summarized qualitatively in Figure 2.1.

These improvements offered by linear and James-Stein estimators, along with those of the threshold estimators introduced in Section 1.3, motivate the more systematic study of wide classes of estimators using shrinkage and thresholding in the sequence models (2.1)– (2.3).

2.2 Penalized Least Squares, Regularization and thresholding

Two common, and related, methods of deriving and motivating estimators are via penalized least squares and via Bayes rules. We discuss the first here and the second in the next section.

We begin with model (2.2), which for a moment we write in matrix form $Y = A\theta + \epsilon z$, with $A = \mathrm{diag}(\alpha_i)$. The unbiased and least squares estimate of θ is found by minimizing $\theta \to \|Y - A\theta\|_2^2$. If θ is high dimensional, we may wish to *regularize* the solution by introducing a *penalty function* $P(\theta)$, and minimizing instead the penalized least squares criterion

$$Q(\theta) = \|Y - A\theta\|_2^2 + \lambda P(\theta).$$

The reason for the names "regularize" and "penalty function" becomes clearer in the general linear model setting, Section 2.9. Here we explore the special consequences of diagonal structure. Indeed, since A is diagonal, the "data term" is a sum of individual components and so it is natural to assume that the penalty also be additive: $P(\theta) = \sum p_i(\theta_i)$, so that

$$Q(\theta) = \sum_i (y_i - \alpha_i\theta_i)^2 + \lambda p_i(\theta_i).$$

Two simple and commonly occurring penalty functions are *quadratic*: $P(\theta) = \sum \omega_i\theta_i^2$ for some non-negative constants ω_i, and q^{th} *power*: $P(\theta) = \|\theta\|_q^q = \sum_{i=1}^n |\theta_i|^q$.

The crucial *regularization parameter* λ determines the relative weight given to the sum of squared error and penalty terms: much more will be said about this later. As λ varies from 0 to $+\infty$, we may think of the penalized estimates $\hat{\theta}_\lambda$ as forming a path from the roughest, least squares solution vector $\hat{\theta}_0 = (y_i/\alpha_i)$ to the smoothest solution vector $\hat{\theta}_\infty = 0$.

Since $Q(\theta)$ has an additive structure, it can be minimized term by term, leading to a univariate optimization for each coefficient estimate $\hat{\theta}_i$. This minimization can be done explicitly in each of three important cases.

(i) ℓ_2 penalty: $p_i(\theta_i) = \omega_i\theta_i^2$. By differentiation, we obtain a co-ordinatewise linear shrinkage estimator

$$\hat{\theta}_i(y) = \frac{\alpha_i}{\alpha_i^2 + \lambda\omega_i} y_i. \tag{2.4}$$

(ii) ℓ_1 penalty: $p(\theta_i) = 2|\theta_i|$. We take $\alpha_i \equiv 1$ here for convenience. Considering only a single co-ordinate and dropping subscripts i, we have

$$Q(\theta) = (y - \theta)^2 + 2\lambda|\theta|.$$

Note that $Q(\theta)$ is convex and

$$\tfrac{1}{2}Q'(\theta) = \begin{cases} \theta - (y-\lambda) & \theta > 0 \\ \theta - (y+\lambda) & \theta < 0 \end{cases}$$

is piecewise linear with positive slope except for an upward jump of 2λ at $\theta = 0$. Hence $Q'(\theta)$ has exactly one sign change (from negative to positive) at a single point $\theta = \hat{\theta}_\lambda$ which must therefore be the minimizing value of $Q(\theta)$. Depending on the value of y, this crossing point is positive, zero or negative, indeed

$$\hat{\theta}_\lambda(y) = \begin{cases} y-\lambda & y > \lambda \\ 0 & |y| \le \lambda \\ y+\lambda & y < -\lambda. \end{cases} \tag{2.5}$$

This is called *soft thresholding* at threshold λ. As is evident from Figure 2.2, the estimator $\hat{\theta}_\lambda$ is characterized by a threshold zone $y \in [-\lambda, \lambda]$, in which all data is set to 0, and by shrinkage toward 0 by a fixed amount λ whenever y lies outside the threshold zone: $|y| > \lambda$. The thresholding is called 'soft' as it is a continuous function of input data y. When applied to vectors $y = (y_i)$, it typically produces sparse fits, with many co-ordinates $\hat{\theta}_{\lambda,i} = 0$, with larger values of λ producing greater sparsity.

(iii) ℓ_0 penalty. $p(\theta_i) = I\{\theta_i \neq 0\}$. The total penalty counts the number of non-zero coefficients:

$$P(\theta) = \sum_i p(\theta_i) = \#\{i : \theta_i \neq 0\}.$$

(Exercise 1.1 explains the name ℓ_0-penalty). Again considering only a single coordinate, and writing the regularization parameter as λ^2,

$$Q(\theta) = (y-\theta)^2 + \lambda^2 I\{\theta \neq 0\}.$$

By inspection,

$$\min_\theta Q(\theta) = \min\{y^2, \lambda^2\},$$

and the ℓ_0-penalized least squares estimate is given by

$$\hat{\theta}_\lambda(y) = \begin{cases} y & |y| > \lambda \\ 0 & |y| \le \lambda. \end{cases} \tag{2.6}$$

This is called *hard thresholding* at threshold λ: The estimator 'keeps' or 'kills' the data y according as it lies outside or inside the threshold zone $[-\lambda, \lambda]$. Again $\hat{\theta}_\lambda$ produces sparse fits (especially for large λ), but with the difference that there is no shrinkage of retained coefficients. In particular, the estimate is no longer a continuous function of the data.

2.3 Priors, posteriors and Bayes estimates

Throughout the book, we will make heavy use of the Bayesian machinery and concepts of priors and posteriors and of the decision theoretic ideas of loss functions and Bayes estimators. The ideas and notation are introduced here; we will see that there are close connections

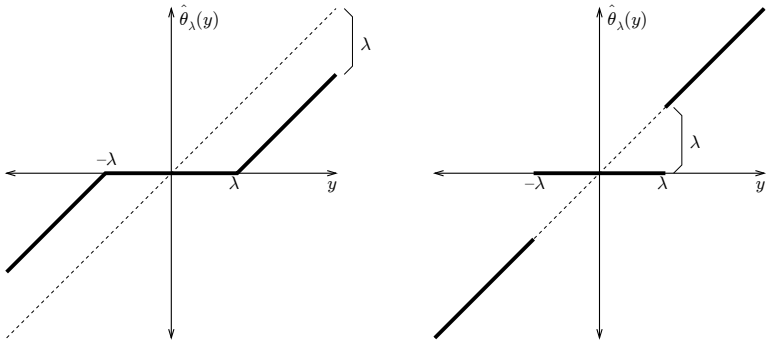

Figure 2.2 Left panel: soft thresholding at λ, showing threshold zone and shrinkage by λ towards 0 outside threshold zone. Dashed line is 45 degree line. Right panel: hard thresholding, with no shrinkage outside the threshold zone.

between the form of Bayes estimators and the penalized estimators of the last section. The more decision theoretic detail, is postponed to Chapter 4.

Suppose we have a prior probability distribution $\pi(d\theta)$ on $\mathbb{R}^n$, and a family of sampling distributions $P(dy|\theta)$, namely a collection of probability measures indexed by θ on the sample space $\mathcal{Y} = \mathbb{R}^n$. Then there is a joint distribution $\mathbb{P}$, say, on $\Theta \times \mathcal{Y}$ and two factorizations into marginal and conditional distributions:

$$\mathbb{P}(d\theta, dy) = \pi(d\theta)P(dy|\theta) = P_\pi(dy)\pi(d\theta|y). \tag{2.7}$$

Here $P_\pi(dy)$ is the marginal distribution of y and $\pi(d\theta|y)$ the posterior for θ given y.

Now suppose that all sampling distributions have densities with respect to Lebesgue measure, $P(dy|\theta) = p(y|\theta)dy$. Then the marginal distribution also has a density with respect to Lebesgue measure, $P_\pi(dy) = p(y)dy$, with

$$p(y) = \int p(y|\theta)\pi(d\theta), \tag{2.8}$$

and we arrive at *Bayes' formula* for the posterior distribution

$$\pi(d\theta|y) = \frac{p(y|\theta)\pi(d\theta)}{p(y)}.$$

In part, this says that the posterior distribution $\pi(d\theta|y)$ is absolutely continuous with respect to the prior $\pi(d\theta)$, and applies equally well whether the prior is discrete (for example, as at (2.26) below) or continuous. We denote by E_y expectation with respect to the posterior distribution given y; thus $E_y h(\theta) = \int h(\theta)\pi(d\theta|y)$.

A *loss function* associates a loss $L(a, \theta) \geq 0$ with each pair (a, θ) in which $a \in \mathbb{R}^n$ denotes an action, or estimate, chosen by the statistician, and $\theta \in \mathbb{R}^n$ denotes the true parameter value. Typically $L(a, \theta) = w(a - \theta)$ is a function $w(\cdot)$ of $a - \theta$. Our main examples here will be quadratic and qth power losses:

$$w(t) = t^T Q t, \qquad w(t) = \|t\|_q^q = \sum_{i=1}^n |t_i|^q.$$

Here Q is assumed to be a positive definite matrix. Given a prior distribution π and observed data y, the *posterior expected loss* (or *posterior risk*)

$$E_y L(a, \theta) = \int L(a, \theta)\pi(d\theta|y)$$

is a function of a (and y). The *Bayes estimator* corresponding to loss function L is obtained by minimizing the posterior expected loss:

$$\hat{\theta}_\pi(y) = \text{argmin}_a E_y L(a, \theta). \tag{2.9}$$

For now, we assume that a unique minimum exists, and ignore measure theoretic questions (see the Chapter Notes).

The *Bayes risk* corresponding to prior π is the expected value—with respect to the marginal distribution of y—of the posterior expected loss of $\hat{\theta}_\pi$:

$$B(\pi) = E_{P_\pi} E_y L(\hat{\theta}_\pi(y), \theta). \tag{2.10}$$

Remark. The *frequentist* definition of risk function begins with the first factorization in (2.7), thus

$$r(\hat{\theta}, \theta) = E_\theta L(\hat{\theta}(y), \theta). \tag{2.11}$$

This will be taken up in Section 2.5 and beyond, and also in Chapter 4, where it is seen to lead to an alternate, but equivalent definition of the Bayes rule $\hat{\theta}_\pi$ in (2.9).

Example 1. Quadratic loss and posterior mean. Suppose that $L(a, \theta) = (a-\theta)^T Q(a-\theta)$ for some positive definite matrix Q. Then $a \to E_y L(a, \theta)$ has a unique minimum, given by the zero of

$$\nabla_a E_y L(a, \theta) = 2Q[a - E_y\theta],$$

and so the Bayes estimator for a quadratic loss function is just the posterior mean

$$\hat{\theta}_\pi(y) = E_y\theta = E(\theta|y). \tag{2.12}$$

Note, in particular, that this result does not depend on the particular choice of $Q > 0$. The posterior expected loss of $\hat{\theta}_\pi$ is given by

$$E[L(\hat{\theta}_\pi, \theta)|y] = E[\theta - E(\theta|y)]^T Q[\theta - E(\theta|y)] = \text{tr}[Q\text{Cov}(\theta|y)]. \tag{2.13}$$

Conjugate priors for the multivariate normal. Suppose that the sampling distribution $P(dy|\theta)$ is multivariate Gaussian $N_n(\theta, \Sigma)$ and that the prior distribution $\pi(d\theta)$ is also Gaussian: $N_n(\theta_0, T)$ [1]. Then the marginal distribution $P_\pi(dy)$ is $N_n(\theta_0, \Sigma + T)$ and the posterior distribution $\pi(d\theta|y)$ is also multivariate normal $N_n(\theta_y, \Sigma_y)$–this is the conjugate prior property. Perhaps most important are the formulas for the posterior mean and covariance matrix:

$$\theta_y = (\Sigma^{-1} + T)^{-1})^{-1}(\Sigma^{-1}y + T^{-1}\theta_0), \qquad \Sigma_y = (\Sigma^{-1} + T^{-1})^{-1} \tag{2.14}$$

[1] Here T is mnemonic for upper case τ.

and the equivalent forms

$$\theta_y = T(T+\Sigma)^{-1}y + \Sigma(T+\Sigma)^{-1}\theta_0, \qquad \Sigma_y = T - T(T+\Sigma)^{-1}T. \tag{2.15}$$

Before the derivation, some remarks:

The posterior mean θ_y is a weighted average of the data y and the prior mean θ_0: the first formula shows that the weights are given by the data and prior *precision* matrices Σ^{-1} and T^{-1} respectively. The posterior precision Σ_y^{-1} is the sum of the prior and data precision matrices, and notably, does not depend on the data y! Hence, in this case, using (2.13), the Bayes risk (2.10) is just $B(\pi) = \mathrm{tr} Q\Sigma_y$.

In the important special case in which the prior mean $\theta_0 = 0$, then $\theta_y = Cy$ is a linear shrinkage rule, shrinking toward 0.

The quadratic regularization estimates discussed in the previous section can be interpreted as Bayes estimates for suitable priors. In the orthogonal setting, $A = I$, estimate (2.4) corresponds to posterior mean (2.14) for a prior $\theta \sim N(0, \lambda^{-1}\Omega^{-1})$ with $\Omega = \mathrm{diag}(\omega_i)$ and sampling variance $\Sigma = I$. See Exercise 2.25 for a more general connection between regularization and Bayes' rule.

Proof Recall the basic formula for conditional distributions in the multivariate normal setting (e.g. (Mardia et al., 1979, p. 63). Namely, if

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right]$$

with $\Sigma_{21} = \Sigma_{12}^T$, then

$$\begin{aligned} y_1|y_2 &\sim N(\theta_{1|2}, \Sigma_{1|2}) \\ \theta_{1|2} &= \theta_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \theta_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

Apply this to the joint distribution that is implied by the assumptions on sampling distribution and prior, after noting that $\mathrm{Cov}(\theta, y) = T$,

$$\begin{pmatrix} \theta \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix}, \begin{pmatrix} T & T \\ T & T+\Sigma \end{pmatrix}\right]$$

Since $I - T(T+\Sigma)^{-1} = \Sigma(T+\Sigma)^{-1}$ this yields formulas (2.15) for the posterior mean and variance. Formulas (2.14) may then be recovered by matrix algebra, using the identity

$$T - T(T+\Sigma)^{-1}T = (T^{-1} + \Sigma^{-1})^{-1}. \qquad \square$$

Remark 2.1 (i) Exercise 2.1 suggests an alternate, direct derivation of formulas (2.14).

(ii) The properties of the posterior mean and variance are sufficiently strong that they lead to characterizations of Gaussian priors. For example, if $\hat{\theta}(y) = cy + b$ 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 π 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 and Ylvisaker, 1979). In addition, the constancy of posterior variance characterizes Gaussian priors, see Exercise 2.3.

Product priors and posteriors. Suppose that the components of the prior are independent, so that we may form the product measure $\pi(d\theta) = \prod_i \pi_i(d\theta_i)$, *and* suppose that the sampling distributions are independent, each depending on only one θ_i, so that $P(dy|\theta) = \prod_i P(dy_i|\theta_i)$. Then from Bayes' formula the posterior distribution factorizes also:

$$\pi(d\theta|y) = \prod_i \pi(d\theta_i|y_i). \tag{2.16}$$

In this situation, then, calculations can be done co-ordinatewise, and are hence generally much simpler.

Additive Loss Functions take the special form

$$L(a,\theta) = \sum_i \ell(a_i,\theta_i). \tag{2.17}$$

Under the assumption of product joint distributions, we have just seen that the posterior distribution factorizes. In this case, the $i-$th component of the posterior expected loss

$$E_y \ell(a_i,\theta_i) = \int \ell(a_i,\theta_i)\pi(d\theta_i|y_i)$$

can be computed based on (a_i, y_i) alone. As a result, the posterior expected loss $E_y L(a,\theta)$ can be minimized term by term, and so the Bayes estimator

$$\hat{\theta}_\pi(y) = \text{argmin}_{(a_i)} E_y \sum_i \ell(a_i,\theta_i) = (\hat{\theta}_{\pi_i}(y_i)) \tag{2.18}$$

is *separable*: the $i-$th component of the estimator depends only on y_i.

Posterior mean-median-mode. Consider in particular the $q-$th power loss functions $L_q(a,\theta) = \sum_i |a_i - \theta_i|^q$. The preceding discussion on separability allows us to focus on a single co-ordinate, and

$$\hat{\theta}_{\pi_1}(y_1) = \text{argmin}_a \int |a-\theta_1|^q \pi(d\theta_1|y_1).$$

The posterior expected loss on the right side is strictly convex if $q > 1$, and so has a unique minimizer if, for example, $E_y|\theta_1|^q < \infty$.

Some particular cases are familiar: $q = 2$ corresponds to the posterior mean, and $q = 1$ to the posterior *median*.

Indeed, for $q = 1$ recall the standard fact that $a \to \int |a-\theta| F(d\theta)$ is minimized at any median a_0, namely a point a_0 of F for which $F((-\infty, a_0]) \geq \frac{1}{2}$ *and* $F([a_0,\infty)) \geq \frac{1}{2}$.

Remark 2.2 Finally, $q = 0$ corresponds to the posterior *mode* (for discrete π). Indeed, we may think of $L_0(a,\theta) = \sum_i I\{i : a_i \neq \theta_i\}$ as counting error, compare Exercise 1.1, so that

$$E[L_0(a,\theta)|y] = \sum_i P(\{\theta_i \neq a_i\}|y_i).$$

Again we can focus on a single co-ordinate–say choosing a_1 to estimate θ_1. A *discrete* prior has the form $\pi(d\theta_1) = \sum_1^r p_l \delta_{t_l}(d\theta_1)$, where δ_{t_l} denotes a unit probability mass at

t_l. Now $a_1 \to P(\{\theta_1 \neq a_1\}|y_1)$ is minimized by choosing the posterior mode, namely $a_1 = \text{argmax}_{t_l} P(\theta = t_l|y_1)$, the most likely discrete value of θ given the observed data.

Examples. For the remainder of this section, we return to squared error loss and consider the Gaussian sequence model. In the next section, we look at some examples involving the posterior median.

Suppose, consistent with (2.3), that the sampling distributions of $y_i|\theta_i$ are independently $N(\theta_i, \sigma_i^2)$, for $i = 1, \ldots, n$. Assume independent conjugate priors $\theta_i \sim N(\theta_{0i}, \tau_i^2)$. This is just the diagonal form of the multivariate Gaussian model considered earlier. Putting $\Sigma = \text{diag}(\sigma_i^2)$ and $T = \text{diag}(\tau_i^2)$ into the earlier formulas (2.14)–(2.15) yields the marginal distribution $y_i \sim N(\theta_{0i}, \sigma_i^2 + \tau_i^2)$. The posterior law has $\theta_i|y_i \sim N(\theta_{y,i}, \sigma_{y,i}^2)$, with the two formulas for the posterior mean given by

$$\theta_{y,i} = \frac{\sigma_i^{-2} y_i + \tau_i^{-2}\theta_{0i}}{\sigma_i^{-2} + \tau_i^{-2}} = \frac{\tau_i^2 y_i + \sigma_i^2 \theta_{0i}}{\tau_i^2 + \sigma_i^2}, \tag{2.19}$$

and the forms for the posterior variance being

$$\sigma_{y,i}^2 = \frac{1}{\sigma_i^{-2} + \tau_i^{-2}} = \frac{\tau_i^2 \sigma_i^2}{\tau_i^2 + \sigma_i^2}. \tag{2.20}$$

Thus, for example, the posterior mean

$$\theta_{y,i} \approx \begin{cases} \theta_{0,i} & \text{if} \quad \sigma_i^2 \gg \tau_i^2, \\ y_i & \text{if} \quad \tau_i^2 \gg \sigma_i^2, \end{cases}$$

corresponding to very concentrated and very vague prior information about θ respectively.

Remark. On notation: formulas are often simpler in the case of unit noise, $\epsilon = 1$, and we reserve a special notation for this setting: $x \sim N_n(\mu, I)$, or equivalently

$$x_i = \mu_i + z_i, \qquad z_i \overset{iid}{\sim} N(0,1), \tag{2.21}$$

for $i = 1, \ldots, n$. It is usually easy to recover the formulas for general ϵ by rescaling. Thus, if $y = \epsilon x$ and $\theta = \epsilon\mu$, then $y \sim N_n(\theta, \epsilon^2 I)$ and so if $\hat{\theta} = \epsilon\hat{\mu}$, then for example

$$E\|\hat{\theta}(y) - \theta\|^2 = \epsilon^2 E\|\hat{\mu}(x) - \mu\|^2. \tag{2.22}$$

Examples. 1. There is a useful analytic expression for the posterior mean in the Gaussian shift model $x \sim N_n(\mu, I)$. First we remark that in this case the marginal density (2.8) has the convolution form $p(x) = \pi \star \phi(x) = \int \phi(x - \mu)\pi(d\mu)$. Since $p(x)$ is finite everywhere–it has integral 1 and is continuous–it follows from a standard exponential family theorem (Lehmann and Romano, 2005, Theorem 2.7.1) that $p(x)$ is actually an analytic function of x, and so in particular is infinitely differentiable everywhere.

Now, the Bayes estimator can be written

$$\hat{\mu}_\pi(x) = \int \mu\phi(x - \mu)\pi(d\mu)/p(x).$$

The standard Gaussian density satisfies

$$\frac{\partial}{\partial x_i}\phi(x) = -x_i\phi(x),$$

and so by rewriting $\mu = x + (\mu - x)$, we arrive at

$$\hat{\mu}_\pi(x) = x + \frac{\nabla p(x)}{p(x)} = x + \nabla \log p(x), \tag{2.23}$$

which represents the Bayes rule as the perturbation of the maximum likelihood estimator $\hat{\mu}_0(x) = x$ by a logarithmic derivative of the marginal density of the prior.

We illustrate how this representation allows one to deduce shrinkage properties of the estimator from assumptions on the prior. Suppose that the prior $\pi(d\mu) = \gamma(\mu)d\mu$ has a continuously differentiable density that satisfies, for all μ,

$$\|\nabla \log \gamma(\mu)\| \le \Lambda. \tag{2.24}$$

This forces the prior tails to be at least as heavy as exponential: it is easily verified that

$$\gamma(0)e^{-\Lambda\|\mu\|} \le \gamma(\mu) \le \gamma(0)e^{\Lambda\|\mu\|},$$

so that Gaussian priors, for example, are excluded.

Representation (2.23) shows that $\hat{\mu}_\pi(x)$ has *bounded shrinkage*: $\|\hat{\mu}_\pi(x) - x\| \le \Lambda$ for all x. Indeed, observing that $(\partial/\partial x_i)\phi(x-\mu) = -(\partial/\partial\mu_i)\phi(x-\mu)$, we have

$$(\partial p/\partial x_i)(x) = \int -(\partial\phi/\partial\mu_i)(x-\mu)\gamma(\mu)d\mu = \int (\partial\gamma/\partial\mu_i)\phi(x-\mu)d\mu$$

where we used (2.24) to conclude that $\gamma(\mu)\phi(x-\mu) \to 0$ as $\mu \to \infty$. Consequently,

$$\|\nabla \log p(x)\| \le \int \|\nabla \log \gamma(\mu)\|\phi(x-\mu)\gamma(\mu)d\mu \big/ p(x) \le \Lambda. \tag{2.25}$$

2. Discrete priors will play an important role at several points in this book. Here consider the simplest case, a symmetric two point prior concentrated on $\{-\tau, \tau\}$:

$$\pi_\tau = \tfrac{1}{2}(\delta_\tau + \delta_{-\tau}). \tag{2.26}$$

The posterior also concentrates on $\{-\tau, \tau\}$, but with posterior probabilities given by

$$\pi(\{\tau\}|x) = \frac{\frac{1}{2}\phi(x-\tau)}{\frac{1}{2}\phi(x-\tau) + \frac{1}{2}\phi(x+\tau)} = \frac{e^{x\tau}}{e^{x\tau} + e^{-x\tau}}, \tag{2.27}$$

so that

$$\pi(\{\tau\}|x) > \pi(\{-\tau\}|x) \quad \text{if and only if} \quad x > 0. \tag{2.28}$$

The posterior mean lies between $-\tau$ and $+\tau$:

$$\hat{\mu}_\tau(x) = E(\mu|x) = \tau \tanh \tau x, \tag{2.29}$$

and the posterior variance is found (try it!) to be

$$E\big[\big(\mu - E_\pi(\mu|x)\big)^2|x\big] = \frac{\tau^2}{\cosh^2 \tau x},$$

and the Bayes risk

$$B(\pi_\tau) = \tau^2 e^{-\tau^2/2} \int_{-\infty}^{\infty} \frac{\phi(x)dx}{\cosh \tau x}. \tag{2.30}$$

2.4 Sparse mixture priors and thresholding

In this section, we look at some estimators that can be derived from "sparse priors". For simplicity, we continue to assume that the observation x has variance one.

A simple model for a sparse high dimensional vector has components drawn i.i.d. from

$$\pi(d\mu) = (1-w)\delta_0(d\mu) + w\gamma(\mu)d\mu. \tag{2.31}$$

Thus a (large) fraction $1-w$ of co-ordinates are 0, while a (small) fraction w are drawn from a prior probability distribution γ. Such *sparse mixture priors* will occur in several later chapters. Later we will consider simple discrete priors for $\gamma(d\mu)$, but for now we assume that $\gamma(d\mu)$ has a density $\gamma(\mu)d\mu$ which is *symmetric* about 0 and *unimodal*.

In this section, our main interest in these priors is that their posterior *medians* generate threshold rules in which the threshold zone depends naturally on the sparsity level w.

Proposition 2.3 *Suppose that the prior has mixture form* (2.31) *for* $w > 0$ *and that the non-zero density* $\gamma(\mu)$ *is symmetric and unimodal. The posterior median* $\hat{\mu}(x) = \hat{\mu}_\pi(x)$ *is*

(a) monotone in x *and antisymmetric:* $\hat{\mu}(-x) = -\hat{\mu}(x)$,

(b) a shrinkage rule: $0 \le \hat{\mu}(x) \le x$ *for* $x \ge 0$,

(c) a threshold rule: there exists $t(w) > 0$ *such that*

$$\hat{\mu}(x) = 0 \quad \text{if and only if} \quad |x| \le t(w).$$

(d) Finally, the threshold $t(w)$*, as a function of* w*, is continuous and strictly decreasing from* $t = \infty$ *at* $w = 0$ *to* $t = 0$ *at* $w = 1$.

Some remarks: we focus on the posterior *median* since it turns out (REF?) that the posterior mean must be a smooth, even analytic, function of x, and so cannot have a threshold zone. Unimodality means that $\gamma(\mu)$ is decreasing in μ for $\mu \ge 0$; this assumption facilitates the proof that the posterior median $\hat{\mu}_\pi$ is a shrinkage rule.

The behavior of $t(w)$ with w is intuitive: with smaller w, a greater fraction of the data $x_i = \mu_i + z_i$ are pure noise, and so we might seek a higher threshold $t(w)$ in order to screen out that noise, knowing that there is a smaller chance of falsely screening out true signal. Compare Figure 2.3.

Before beginning the proof, we explore the structure of the posterior corresponding to prior (2.31). First, the marginal density for x is

$$p(x) = \int \phi(x-\mu)\pi(d\mu) = (1-w)\phi(x) + wg(x),$$

where the convolution density $g(x) = \phi \star \gamma(x) = \int \phi(x-\mu)\gamma(\mu)d\mu$. For later use, it is helpful to split up $g(x)$ into parts $g_{\mathrm{p}}(x)$ and $g_{\mathrm{n}}(x)$ corresponding to integrals over $\mu > 0$ and $\mu < 0$ respectively. Note that $g_{\mathrm{p}}(x)$ and $g_{\mathrm{n}}(x)$ respectively satisfy

$$(g_{\mathrm{p/n}}/\phi)(x) = \int_0^\infty e^{\pm x\mu - \mu^2/2}\gamma(\mu)d\mu, \tag{2.32}$$

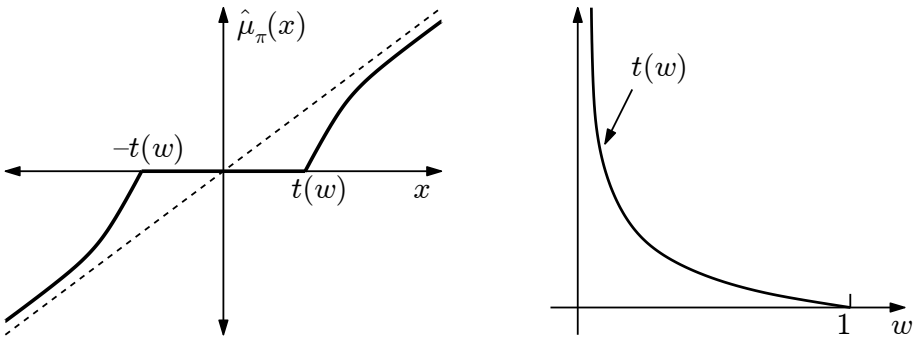

Figure 2.3 Left: posterior median estimator $\hat{\mu}_\pi(x)$ showing threshold zone $x \in [-t(w), t(w)]$. Right: Threshold $t(w)$ decreases as w increases.

where the formula for g_n also uses the symmetry of γ about 0.

Turning now to the form of the posterior, we note that since the prior has an atom at 0, so must also the posterior, and hence

$$\pi(A|x) = \pi(\{0\}|x) I_A(0) + \int_A \pi(\mu|x) d\mu, \tag{2.33}$$

with

$$\pi(\{0\}|x) = \frac{(1-w)\phi(x)}{p(x)}, \qquad \pi(\mu|x) = \frac{w\gamma(\mu)\phi(x-\mu)}{p(x)}. \tag{2.34}$$

We can rewrite the posterior nicely using a slightly non-standard choice of dominating measure: $\nu(d\mu) = \delta_0(d\mu) + d\mu$. We then have $\pi(A|x) = \int_A \pi(\mu|x)\nu(d\mu)$, with

$$\pi(\mu|x) = w(\mu)\phi(x-\mu)/p(x), \tag{2.35}$$

for *all* $\mu, x \in \mathbb{R}$. Here $w(\mu) = 1 - w$ for $\mu = 0$ and $w\gamma(\mu)$ for $\mu \neq 0$.

Proof of Proposition 2.3. (a) Since γ is assumed unimodal and symmetric about 0, its support is an interval $[-a, a]$ for some $a \in (0, \infty]$. Consequently, the posterior density has support $[-a, a]$ and $\pi(\mu|x) > 0$ for $\mu \in (-a, a)$ and all x. In particular, the posterior median $\hat{\mu}_\pi(x)$ is uniquely defined.

We will show that $x < x'$ implies that for $m \in \mathbb{R}$,

$$\pi(\mu > m|x) < \pi(\mu > m|x'), \tag{2.36}$$

from which it follows that the posterior distribution is stochastically increasing and in particular that the posterior median is increasing in x. The product form representation (2.35) suggests an argument using ratios: if $\mu < \mu'$ then cancellation and properties of the Gaussian density yield

$$\frac{\pi(\mu'|x')\pi(\mu|x)}{\pi(\mu|x')\pi(\mu'|x)} = \exp\{(\mu'-\mu)(x'-x)\} > 1.$$

Now move the denominator to the right side and integrate with respect to the dominating

measure ν over $\mu' \in R = (m, \infty)$ and $\mu \in R^c = (-\infty, m]$ to get

$$\pi(R|x')\pi(R^c|x) > \pi(R^c|x')\pi(R|x),$$

or, using the notion of $\text{Odds}(A) = P(A)/P(A^c)$,

$$\text{Odds}(R|x') > \text{Odds}(R|x),$$

which is equivalent to (2.36), which we sought to establish. The anti-symmetry of the posterior median is immediate from the symmetry of the prior and the Gaussian error density.

(b) The property $\hat{\mu}(x) \leq x$ will follow from unimodality of $\gamma(\mu)$, while the property $\hat{\mu} \geq 0$ will use a similar argument, using instead symmetry of $\gamma(\mu)$. The shrinkage property will follow if we show that $\pi(\mu \leq x|x) \geq 1/2$. Now from (2.34),

$$\pi(\mu > x|x) = w \int_x^\infty \phi(x-\mu)\gamma(\mu)d\mu/p(x).$$

Since $p(x) > wg(x)$, clearly a sufficient condition for $\pi(\mu > x|x) < 1/2$ is that

$$\int_x^\infty \phi(x-\mu)\gamma(\mu)d\mu \leq \int_{-\infty}^x \phi(x-\mu)\gamma(\mu)d\mu,$$

or equivalently that

$$\int_0^\infty \phi(\mu')\gamma(x+\mu')d\mu' \leq \int_0^\infty \phi(\mu')\gamma(x-\mu')d\mu'$$

which indeed follows from the unimodality hypothesis (combined with symmetry for the case when $\mu' > x$).

For later use, we use (2.34) and the definition of g_{p} to write

$$\pi(\mu > 0|x) = \frac{wg_{\text{p}}(x)}{(1-w)\phi(x) + wg(x)}. \tag{2.37}$$

If $x < 0$, then $g_{\text{p}}(x) < g(x)/2$ using the symmetry of γ, and so $\pi(\mu > 0|x) < 1/2$ and hence the posterior median $\hat{\mu}(x) \leq 0$. By antisymmetry, we conclude that $\hat{\mu}_\pi(x) \geq 0$ for $x \geq 0$.

(c) Now we turn to existence of the threshold zone. If $w < 1$, we have $\pi(\{0\}|x) > 0$ and by symmetry $\pi(\mu < 0 \mid x = 0) = \pi(\mu > 0 \mid x = 0)$, so it must be that

$$\pi(\mu < 0 \mid x = 0) < \tfrac{1}{2} < \pi(\mu \leq 0 \mid x = 0) \tag{2.38}$$

so that $\hat{\mu}_\pi(0) = 0$, which is also clear by reason of symmetry. More importantly, the functions $x \to \pi(\mu > 0|x)$ and $\pi(\mu \geq 0|x)$ are continuous (e.g. from (2.37)) and strictly increasing (from (2.36) proved above). Consequently, (2.38) remains valid on an *interval*: $-t(w) \leq x \leq t(w)$, which is the threshold zone property. Compare Figure 2.4.

(d) From Figure 2.4, the threshold $t = t(w)$ satsifies $\pi(\mu > 0|t) = 1/2$, and rearranging (2.37) we get the equation

$$2wg_{\text{p}}(t) = (1-w)\phi(t) + wg(t).$$

Dividing by $w\phi(t)$, rearranging and then using representations (2.32), we get

$$w^{-1} = 1 + (g_{\text{p}} - g_{\text{n}})/\phi = 1 + 2\int_0^\infty \sinh(t\mu)e^{-\mu^2/2}\gamma(\mu)d\mu.$$

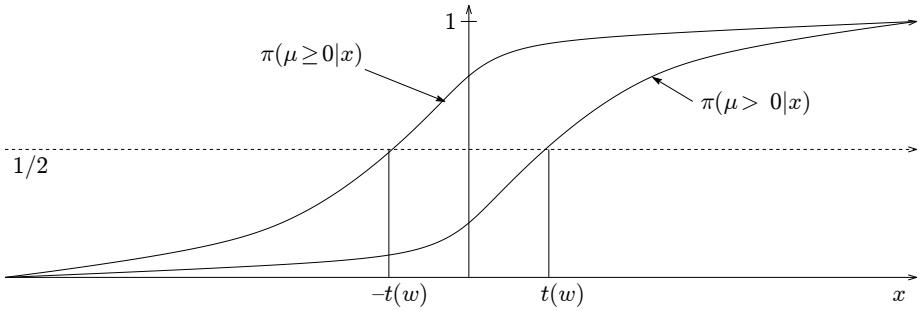

Figure 2.4 The threshold zone arises for the set of x for which both $\pi(\mu \geq 0|x) \geq 1/2$ and $\pi(\mu > 0|x) \leq 1/2$.

Since the right side is continuous and monotone in t, we conclude that w is a continuous and strictly decreasing function of t, from $w = 1$ at $t = 0$ to $w = 0$ at $t = \infty$. □

The tails of the prior density γ have an important influence on the amount of shrinkage of the posterior median. Consider the following univariate analog of (2.24):

$$(\log \gamma)(\mu) \text{ is absolutely continuous, and } |(\log \gamma)'| \leq \Lambda \text{ a.e.} \tag{2.39}$$

Exercise 2.5 outlines the proof of

Proposition 2.4 *Assume that the prior density has logarithmic derivative bounded by* Λ, (2.24). *Then the posterior median* $\hat{\mu}$ *has bounded shrinkage: for all* x,

$$|\hat{\mu}(x; w) - x| \leq t(w) + \Lambda + 2. \tag{2.40}$$

Remark. The condition (2.39) implies, for $u > 0$, that $\log \gamma(u) \geq \log \gamma(0) - \Lambda u$ and so, for all u, that $\gamma(u) \geq \gamma(0)e^{-\Lambda|u|}$. Hence, for bounded shrinkage, the assumption requires the tails of the prior to be exponential or heavier. Gaussian priors do not satisfy (2.40), and indeed the shrinkage is then proportional to x for large x. Heuristically, this may be seen by arguing that the effect of the atom at 0 is negligible for large x, so that the posterior is essentially Gaussian, so that the posterior median equals the posterior mean, and is given, from (2.19) by

$$\tau^2 y/(\tau^2 + 1) = y - y/(\tau^2 + 1).$$

For actual calculations, it is useful to have a more explicit expression for the posterior median. From (2.34) and the succeeding discussion, we may rewrite

$$\pi(\mu|x) = w(x)\gamma(\mu|x),$$

where $w(x) = wg(x)/p(x) = \pi\{\mu \neq 0|x\}$. Let $\tilde{\Gamma}(\tilde{\mu}|x) = \int_{\tilde{\mu}}^{\infty} \gamma(\mu|x)d\mu$. If $x \geq t(w)$, then the posterior median $\hat{\mu} = \hat{\mu}_\pi(x)$ is defined by the equation

$$w(x)\tilde{\Gamma}(\hat{\mu}|x) = 1/2. \tag{2.41}$$

Example. A prior suited to numerical calculation in software is the Laplace density

$$\gamma_a(\mu) = \tfrac{1}{2}ae^{-a|\mu|},$$

which satisfies (2.39). The following formulas may be verified (Exercise 2.6 fills in some details.) First, define

$$\beta(x) = \frac{g(x)}{\phi(x)} - 1 = \frac{a}{2}\left[\frac{\Phi}{\phi}(x-a) + \frac{\tilde{\Phi}}{\phi}(x+a)\right] - 1.$$

Then, for the posterior median, using (2.41),

$$\hat{\mu}(x) = \max\{0, x - a - \Phi^{-1}(z_0)\}, \tag{2.42}$$

with $z_0 = a^{-1}\phi(x-a)[w^{-1} + \beta(x)]$. One can verify that as $x \to \infty$,

$$\beta(x) \sim \tfrac{1}{2}a/\phi(x-a), \qquad z_0 \sim \tfrac{1}{2}, \qquad \hat{\mu}(x) \sim x - a. \tag{2.43}$$

In particular, we see the bounded shrinkage property—for large x, the data is pulled down by about a. The threshold $t = t(w)$ and the weight $w = w(t)$ are related by

$$w(t)^{-1} = a(\Phi/\phi)(t-a) - \beta(t). \tag{2.44}$$

See Figure 2.5.

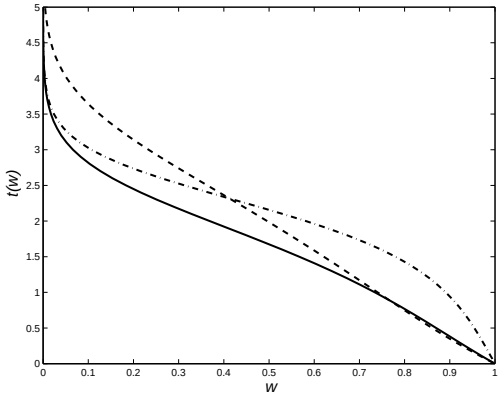

Figure 2.5 Threshold $t(w)$ as a function of non-zero prior mass w for the Laplace density for three values of scale parameter a: Dash-dot: $a = 0.1$, Solid: $a = 0.5$, Dashed $a = 2$. Increasing sparsity (smaller w) corresponds to larger thresholds.

2.5 Mean squared error and linear estimators

We have described a large class of estimators that can be obtained using priors and regularization penalties and so it is natural to ask: how might we compare their properties? The simplest and most common approach is to study the mean squared error

$$r(\hat{\theta}, \theta) = E_\theta \|\hat{\theta} - \theta\|^2 = E_\theta \sum_{i=1}^{n} \left[\hat{\theta}_i(y) - \theta_i\right]^2.$$

Let us begin with the sequence model $y \sim N_n(\theta, \epsilon^2 I)$ and the class of *linear* estimators

$$\hat{\theta}_C(y) = Cy \tag{2.45}$$

for some $n \times n$ matrix C. The class of linear estimators includes smoothing splines, seen in Chapter 1, kernel estimators (Chapter 3) and other frequently used methods.

For any estimator $\hat{\theta}$ with a finite variance, linear or not, the mean square error splits into variance and (squared) bias terms, yielding the *variance-bias decomposition*:

$$\begin{aligned} E\|\hat{\theta} - \theta\|^2 &= E\|\hat{\theta} - E\hat{\theta}\|^2 + \|E\hat{\theta} - \theta\|^2 \\ &= \mathrm{var}(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta}). \end{aligned} \tag{2.46}$$

More specifically, since $\|\hat{\theta} - E\hat{\theta}\|^2 = \mathrm{tr}(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T$, we have

$$\mathrm{var}(\hat{\theta}) = \mathrm{tr}[\mathrm{Cov}(\hat{\theta})].$$

For linear estimators $\hat{\theta}_C$, clearly $\mathrm{Cov}(Cy) = \epsilon^2 CC^T$ and so

$$\mathrm{var}(\hat{\theta}_C) = \epsilon^2 \mathrm{tr}\, CC^T = \epsilon^2 \mathrm{tr}\, C^T C.$$

The bias $E\hat{\theta}_C - \theta = (C - I)\theta$, and hence the mean squared error becomes

$$r(\hat{\theta}_C, \theta) = \epsilon^2 \mathrm{tr}\, C^T C + \|(I - C)\theta\|^2. \tag{2.47}$$

[Note that only second order distributional assumptions are used here, namely that $Ez = 0$ and $\mathrm{Cov}(z) = I$.]

The mean squared error is a quadratic function of θ, and the squared bias term is unbounded except in the case $C = I$. In this case $\hat{\theta}_I(y) = y$ is the *maximum likelihood estimator* (MLE), it is exactly unbiased for θ and the MSE of the MLE is constant,

$$r(\hat{\theta}_I, \theta) \equiv n\epsilon^2.$$

Thus, with linear estimators we already see the fundamental issue: there is no single estimator with uniformly best mean squared error, compare Figure 2.1.

One way to exclude poor estimators is through the notion of *admissibility*. We say that estimator $\hat{\theta}$ is *inadmissible* if there exists another estimator $\hat{\theta}'$ such that $R(\hat{\theta}', \theta) \leq R(\hat{\theta}, \theta)$ for all θ, with strict inequality occurring for *some* θ. Such an estimator $\hat{\theta}'$ is said to *dominate* $\hat{\theta}$. And if no such dominating $\hat{\theta}'$ exists, then the original estimator $\hat{\theta}$ is called admissible. Admissibility itself is a rather weak notion of optimality, but the concept is useful because–in principle, if not always in practice–one would not want to use an inadmissible estimator. Thus, typically, inadmissibility results are often of more interest than admissibility ones.

The most important (and surprising) fact about admissibilty is that the MLE $\hat{\theta}_I$ is itself **in**admissible exactly when $n \geq 3$. Indeed, as indicated in Section 2.1, the James-Stein estimator $\hat{\theta}^{JS}$ dominates the MLE everywhere: $r(\hat{\theta}^{JS}, \theta) < n\epsilon^2 = r(\hat{\theta}_I, \theta)$ for all $\theta \in \mathbb{R}^n, n \geq 3$. A short proof is given in the next section.

We can now describe a nice result on inadmissibility for linear estimators. We saw in Chapter 1.4 that cubic smoothing splines shrink all frequencies except for a two dimensional subspace on which no shrinkage occurs. This turns out to be admissible, and in fact, all reasonable, i.e. admissible, linear estimators must behave in this general manner.

Theorem 2.5 *Suppose that $y \sim N_n(\theta, \epsilon^2 I)$. The linear estimator $\hat{\theta}_C(y) = Cy$ is admissible (for squared error loss) if and only if C*

(i) *is symmetric,*

(ii) *has eigenvalues* $0 \le \varrho_i(C) \le 1$, *and*
(iii) *has at most two* $\varrho_i(C) = 1$.

Proof We show only that each of these conditions is necessary for admissibility: if the condition fails we show how to construct a dominating estimator. (i) We use the notation $|A| = (A^T A)^{1/2}$ and the fact (Exercise 2.7) that $\mathrm{tr}A \le \mathrm{tr}|A|$, with equality only if A is symmetric, $A^T = A$.

Let D be defined via the identity $I - D = |I - C|$; clearly D is symmetric, and we use the variance-bias decomposition (2.46) to show that the MSE of $\hat{\theta}_D$ is everywhere better than that of $\hat{\theta}_C$ if C is not symmetric. Since

$$(I - D)^T (I - D) = |I - C|^2 = (I - C)^T (I - C),$$

the two estimators have the same (squared) bias. Turning to the variance terms, write

$$\mathrm{tr}\, D^T D = \mathrm{tr}\, I - 2\mathrm{tr}(I - D) + \mathrm{tr}(I - D)^T (I - D). \tag{2.48}$$

Comparing with the corresponding variance term for $\hat{\theta}_C$, we see that $\mathrm{tr}\, D^T D < \mathrm{tr}\, C^T C$ if and only if

$$\mathrm{tr}(I - D) = \mathrm{tr}|I - C| > \mathrm{tr}(I - C)$$

which occurs if and only if C fails to be symmetric.

(ii) As we may now assume that C is symmetric, we can find a decomposition $C = URU^T$ with U orthogonal and $R = \mathrm{diag}(\varrho_i)$ containing the (real) eigenvalues of C. Now change variables to $\eta = U^T \theta$ and $x = U^T y \sim N(\eta, \epsilon^2 I)$. Orthogonality of U implies that $E\|Cy - \theta\|^2 = E\|Rx - \eta\|^2$, so we have

$$r(\hat{\theta}_C, \theta) = r(\hat{\eta}_R, \eta) = \sum_i \epsilon^2 \varrho_i^2 + (1 - \varrho_i)^2 \eta_i^2 = \sum_i r_L(\varrho_i, \eta_i), \tag{2.49}$$

say. Clearly, if any eigenvalue $\varrho_i \notin [0, 1]$, a strictly better MSE results by replacing ϱ_i by 1 if $\varrho_i > 1$ and by 0 if $\varrho_i < 0$.

(iii) Now suppose that $\varrho_1 = \ldots = \varrho_d = 1 > \varrho_i$ for $i > d \ge 3$, and let $x^d = (x_1, \ldots, x_d)$. We have noted that the James-Stein estimator is everywhere better than $\hat{\eta}_I(x^d) = x^d$. So if we define a new estimator $\hat{\eta}$ to use $\hat{\eta}^{JS}$ on x^d and to continue to use $\varrho_i x_i$ for $i > d$, then

$$r(\hat{\eta}, \eta) = r(\hat{\eta}^{JS}, \eta^d) + \sum_{i>d} r_L(\varrho_i, \eta_i) < r(\hat{\eta}_R, \eta),$$

and so $\hat{\eta}$ dominates $\hat{\eta}_R$ and hence $\hat{\theta}_C$.

For the converse, that conditions (i)-(iii) imply that $\hat{\theta}_C$ is admissible, see Cohen (1966). For the special case of the univariate MLE, see Remark 4.3 below. □

This still leaves a lot of linear estimators, to say nothing of the non-linear ones. To choose among the many admissible and other reasonable rules, other criteria are needed. Two approaches are commonly used. The average case approach requires specification of a prior distribution; this is discussed in Chapter 4. The worst case approach compares estimators by their maximum risk, seeking to find estimators whose maximum risk is as small as possible:

$$R_n = \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^n} E_\theta \|\hat{\theta} - \theta\|^2. \tag{2.50}$$

Here the infimum is taken over all estimators, linear or non-linear. We take up the systematic study of minimaxity in Chapter 4. For now, we mention the classical fact that the MLE $\hat{\theta}_I(y) = y$ is minimax:

$$R_n = n\epsilon^2 = \sup_{\theta \in \mathbb{R}^n} E_\theta \|y - \theta\|^2. \tag{2.51}$$

(This is proved, for example, using Corollary 4.10 and Proposition 4.16).

Mallows' C_L and C_p. There is a simple and useful unbiased estimate of the MSE of linear estimators $\hat{\theta}_C$. To derive it, observe that the residual $y - \hat{\theta}_C = (I - C)y$, and that the mean residual sum of squares (RSS) satisfies

$$E\|y - \hat{\theta}_C\|^2 = E\|(I - C)(\theta + \epsilon z)\|^2 = \epsilon^2 \mathrm{tr}\,(I - C)^T(I - C) + \|(I - C)\theta\|^2. \tag{2.52}$$

Consequently the *C_L-statistic*, denoted here by U,

$$U(y) := \|y - \hat{\theta}_C\|^2 - n\epsilon^2 + 2\epsilon^2 \mathrm{tr}\, C$$

is found, by combining (2.52) and (2.47), to be an unbiased estimate of MSE:

$$EU(y) = E\|\hat{\theta}_C - \theta\|^2.$$

Here is one application. If the matrix $C = C(\lambda)$ depends on a 'shrinkage' or 'bandwidth' parameter λ, and if ϵ^2 is known (or can be estimated), then one possibility is to choose λ to minimize the C_L estimate of MSE:

$$\begin{aligned} \hat{\lambda} &= \mathrm{argmin}_\lambda\, U_\lambda(y) \\ U_\lambda(y) &= \|y - C(\lambda)y\|^2 - n\epsilon^2 + 2\epsilon^2 \mathrm{tr}\, C(\lambda). \end{aligned} \tag{2.53}$$

If $C = P_K$ represents orthogonal projection onto the subspace spanned by the coordinates in a subset $K \subset \{1, \ldots, n\}$ of cardinality n_K, then $\mathrm{tr} P_K = n_K$. One might then choose the subset to minimize

$$U_K(y) = \|y - P_K y\|^2 + 2\epsilon^2 n_K - n\epsilon^2. \tag{2.54}$$

This version of the criterion is called Mallows' C_p. In applying it, one may wish to restrict the class of subsets K, for example to initial segments $\{1, \ldots, k\}$ for $1 \le k \le n$.

2.6 The James-Stein estimator and Stein's Unbiased Risk Estimate

We have seen that Mallows' C_L provides an unbiased estimate of the risk of a linear rule $\hat{\theta}_C(y) = Cy$. In fact, there is a wide-ranging generalization: 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.

Suppose that g is a nice function of a single variable $z \in \mathbb{R}$. Integration by parts and the rapid decay of the Gaussian density tails show that

$$\int g(z) z \phi(z) dz = \int g(z)\big[-\frac{d}{dz}\phi(z)\big] dz = \int g'(z)\phi(z) dz.$$

If $Z \sim N_n(0, I)$ and $g : \mathbb{R}^n \to \mathbb{R}$, the formula becomes

$$E[Z_i g(Z)] = E[D_i g(Z)], \tag{2.55}$$

where $D_i g = \partial g / \partial x_i$ is another notation for partial derivative.

Suppose now that g is vector valued, $g : \mathbb{R}^n \to \mathbb{R}^n$, that $X \sim N_n(\mu, I)$ and define the *divergence* of g

$$\nabla^T g = \sum_i D_i g_i = \sum_i \frac{\partial}{\partial x_i} g_i. \tag{2.56}$$

We may then rewrite (2.55) as

$$E\,(X - \mu)^T g(X) = E\,\nabla^T g(X), \tag{2.57}$$

Regularity conditions *do* need attention here: some counterexamples are given below. It is, however, enough in (2.55) and (2.57) to assume that g is *weakly differentiable*: i.e. that g is absolutely continuous on all line segments parallel to the co-ordinate axes, and its partial derivatives (which consequently exist almost everywhere) are integrable on compact sets. Appendix C.22 gives the conventional definition of weak differentiability and the full proof of (2.57) and the following important consequence.

Proposition 2.6 *Suppose that $g : \mathbb{R}^n \to \mathbb{R}^n$ is weakly differentiable, that $X \sim N_n(\mu, I)$ and that for $i = 1, \ldots, n$, $E_\mu |X_i g_i(X)| + E|D_i g_i(X)| < \infty$. Then*

$$E_\mu \|X + g(X) - \mu\|^2 = E_\mu \{n + 2\nabla^T g(X) + \|g(X)\|^2\}. \tag{2.58}$$

Remarks. (i) The expression

$$U(x) = n + 2\nabla^T g(x) + \|g(x)\|^2$$

is called *Stein's unbiased risk estimate* (SURE). In the particular case of a linear estimator $\hat{\mu}(x) = Cx$, it reduces to Mallows' C_L. Indeed $g(x) = (C - I)x$ and so $\nabla^T g(x) = \operatorname{tr} C - n$ and so

$$U(x) = -n + 2\operatorname{tr} C + \|(I - C)x\|^2.$$

(ii) Soft thresholding satisfies the weak differentiability condition. Indeed, writing $\hat{\mu}_S(x) = x + g_S(x)$, we see from (2.5) that

$$g_{S,i}(x) = \begin{cases} -\lambda & x_i > \lambda \\ -x_i & |x_i| \le \lambda \\ \lambda & x_i < -\lambda \end{cases} \tag{2.59}$$

is absolutely continuous as a function of each x_i, with derivative bounded by 1.

(iii) By contrast, hard thresholding has $\hat{\mu}_H(x) = x + g_H(x)$ which is not even continuous, $g_{H,i}(x) = -x_i I\{|x_i| \le \lambda\}$, and so the unbiased risk formula cannot be applied.

(iv) Generalization to noise level ϵ and more generally to $Y \sim N_n(\theta, V)$ is straightforward (see Exercise 2.8).

The James-Stein estimate. For $X \sim N_n(\mu, I)$, the James-Stein estimator is defined by

$$\hat{\mu}^{JS}(x) = \left(1 - \frac{n-2}{\|x\|^2}\right)x, \tag{2.60}$$

and was used by James and Stein (1961) to give a more explicit demonstration of the inadmissibility of the maximum likelihood estimator $\hat{\mu}^{MLE}(x) = x$ in dimensions $n \geq 3$. [The MLE is known to be admissible for $n = 1, 2$, see e.g. Lehmann and Casella (1998, Ch. 5, Example 2.5 and Problem 4.5).] Later, Stein (1981) showed that the inadmissibility may be verified immediately from the unbiased risk formula (2.58). Indeed, if $n \geq 3$, $g(x) = -(n-2)\|x\|^{-2}x$ is weakly differentiable, and

$$D_i g_i(x) = -(n-2)\left(\frac{1}{\|x\|^2} - \frac{2x_i^2}{\|x\|^4}\right)$$

so that $\nabla^T g(x) = -(n-2)^2\|x\|^{-2}$ and so the unbiased risk estimator

$$U(x) = n - (n-2)^2\|x\|^{-2}.$$

Consequently

$$r(\hat{\mu}^{JS}, \mu) = n - (n-2)^2 E_\mu \|X\|^{-2}, \tag{2.61}$$

which is finite and everywhere smaller than $r(\hat{\mu}^{MLE}, \mu) = E_\mu\|X - \mu\|^2 \equiv n$ *so long as* $n \geq 3$. We need $n \geq 3$ for finiteness of $E\|X\|^{-2}$, see (2.64) below.

Remarks. (i) The James-Stein rule may be derived from a linear Bayes shrinkage estimator by estimating the shrinkage constant from the data. This "empirical Bayes" interpretation, due to Efron and Morris (1973), is given in Exercise 2.11.

(ii) Where does the factor $n-2$ come from? A partial explanation: the estimator $\hat{\mu}(x) = (1 - \beta/\|x\|^2)x$ has unbiased risk estimate $U_\beta(x) = n - \{2\beta(n-2) - \beta^2\}/\|x\|^2$, and this quantity is minimized, for each x, by the choice $\beta = n-2$. Note that $\beta = 2(n-2)$ gives the same risk as the MLE.

(iii) The *positive part* James-Stein estimator

$$\hat{\mu}^{JS+}(x) = \left(1 - \frac{n-2}{\|x\|^2}\right)_+ x \tag{2.62}$$

has necessarily even better MSE than $\hat{\mu}^{JS}$ (Exercise 2.12), and hence better than $\hat{\mu}^{MLE}$.

The unbiased risk estimate leads to an informative bound on the mean squared error of the James-Stein rule.

Proposition 2.7 *If $X \sim N_n(\mu, I)$, then the James-Stein rule satisfies*

$$E_\mu\|\hat{\mu}^{JS} - \mu\|^2 \leq 2 + \frac{(n-2)\|\mu\|^2}{(n-2) + \|\mu\|^2}. \tag{2.63}$$

Proof For general μ, the sum of squares $\|X\|^2$ follows a non-central chisquared distribution with non-centrality parameter $\|\mu\|^2$. The non-central distribution may be realized as a mixture of central chi-squared distributions χ^2_{n+2N}, where N is a Poisson variate with mean $\|\mu\|^2/2$. (cf. e.g. Johnson and Kotz (1970, p. 132)). Recall also the formula

$$E\,[1/\chi^2_n] = 1/(n-2). \tag{2.64}$$

Hence, by conditioning first on N, and then using (2.64) and Jensen's inequality,

$$E[1/\chi^2_{n+2N}] = E[1/(n-2+2N)] \geq 1/(n-2+\|\mu\|^2).$$

Substituting into the unbiased risk formula (2.61), we obtain

$$r(\hat{\mu}^{JS}, \mu) \leq 2 + (n-2) - \frac{(n-2)^2}{n-2+\|\mu\|^2},$$

which yields the desired result after rearrangement. □

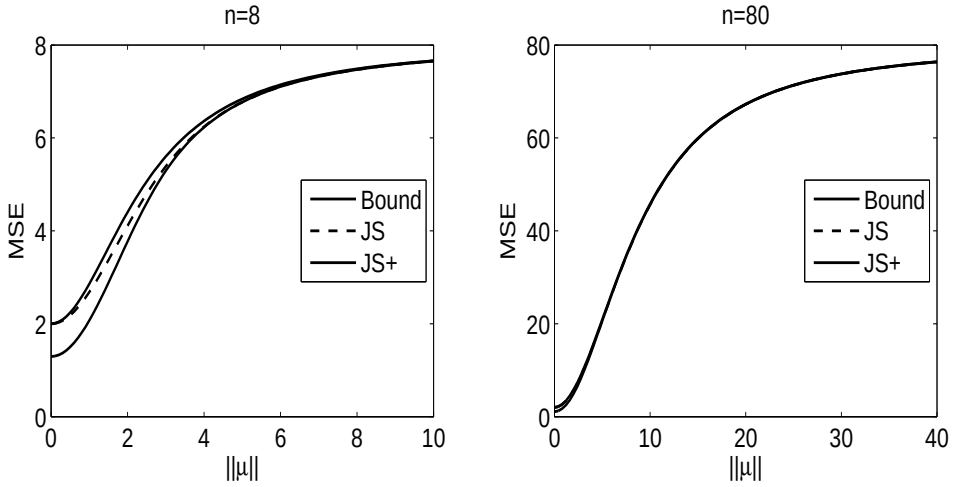

Figure 2.6 Exact risk functions of James-Stein rule $\hat{\mu}^{JS}$ (dashed) and positive part James-Stein $\hat{\mu}^{JS+}$ (solid) compared with upper bound from right side of (2.63). In the right panel ($n = 80$) the three curves are nearly indistinguishable.

Figure 2.6 illustrates several important aspects of the risk of the James-Stein estimator. First, the improvement offered by James-Stein relative to the MLE can be very large. For $\mu = 0$, we see from (2.61) and (2.64) that $r(\hat{\mu}^{JS}, 0) = 2$ while $r(\hat{\mu}^{MLE}, \mu) \equiv n$.

Second, the region of significant savings can be quite large as well. For $\|\mu\|^2 \leq \beta n$, the upper bound (2.63) is less than $(1+\beta n)/(1+\beta)$ so that, for example, if $\|\mu\|^2 \leq 4n$, then the savings is (roughly) at least 20 %. See also Exercise 2.14.

Third, the additional improvement offered by the positive part estimator can be significant when both $\|\mu\|$ and n are small, but otherwise the simple upper bound (2.63) gives a picture of the risk behavior that is accurate enough for most purposes.

Remark. Exercise 2.18 provides details on the exact risk formulas for $\hat{\mu}^{JS+}$ used in Figure 2.6. It is known, e.g. Lehmann and Casella (1998, Example 5.7.3), that the positive part James-Stein rule cannot be admissible. While dominating estimators have been found, (Shao and Strawderman, 1994), the actual amount of improvement over $\hat{\mu}^{JS+}$ seems not to be of practical importance.

Direct use of Jensen's inequality in (2.61) yields a simpler bound inferior to (2.63), Exercise 2.13.

Corollary 2.8 *Let $\hat{\mu}_c(x) = cx$ be a linear shrinkage estimate. Then*

$$r(\hat{\mu}^{JS}, \mu) \le 2 + \inf_c r(\hat{\mu}_c, \mu). \tag{2.65}$$

Proof The MSE of a linear shrinkage estimator $\hat{\mu}_c$ is

$$E\|cX - \mu\|^2 = c^2 n + (1-c)^2\|\mu\|^2. \tag{2.66}$$

In an idealized situation in which $\|\mu\|$ is known, the *ideal* shrinkage factor $c = c^{IS}(\mu)$ would be chosen to minimize this MSE, so that

$$c^{IS}(\mu) = \frac{\|\mu\|^2}{n + \|\mu\|^2}, \tag{2.67}$$

and

$$\inf_c r(\hat{\mu}_c, \mu) = \frac{n\|\mu\|^2}{n + \|\mu\|^2} \ge \frac{(n-2)\|\mu\|^2}{n - 2 + \|\mu\|^2}, \tag{2.68}$$

so that we need only refer to the preceding proposition. □

This is an example of an *oracle inequality*:

$$r(\hat{\mu}^{JS}, \mu) \le 2 + r(\hat{\mu}^{IS}, \mu), \tag{2.69}$$

the risk of a bona fide estimator $\hat{\mu}^{JS}$ is bounded by the risk of the ideal estimator $\hat{\mu}^{IS}(x) = c^{IS}(\mu)x$, (unrealizable in practice, of course) plus an additive constant. If one imagines the ideal shrinkage factor $c^{IS}(\mu)$ as being provided by an 'oracle' with supernatural knowledge, then (2.68) says that the James-Stein estimator can almost mimic the oracle.

In high dimensions, the constant 2 is small in comparison with the risk of the MLE, which is everywhere equal to n. On the other hand the bound (2.69) is sharp: at $\mu = 0$, the unbiased risk equality (2.61) shows that $r(\hat{\mu}^{JS}, 0) = 2$, while the ideal risk is zero.

The James-Stein estimator $\hat{\mu}^{JS}$ can be interpreted as an adaptive linear estimator, that is, an estimator that while itself not linear, is derived from a linear estimator by estimation of a tuning parameter, in this case the shrinkage constant. The ideal shrinkage constant $c^{IS}(\mu) = 1 - n/(n + \|\mu\|^2)$ and we can seek to estimate this using X. Indeed, $E\|X\|^2 = n + \|\mu\|^2$ and so $E\|X\|^{-2} \ge 1/(n + \|\mu\|^2)$, with approximate equality for large n. Consider therefore estimates of the form $\hat{c}(x) = 1 - \beta/\|x\|^2$ and note that we may determine β by observing that for $\mu = 0$, we have $E\hat{c} = 1 - \beta/(n-2) = 0$. Hence $\beta = n - 2$, and in this way, we recover precisely the James-Stein estimator.

For use in the next section, we record a version of (2.69) for arbitrary noise level. Define

$$\hat{\theta}^{JS+}(y) = \left(1 - \frac{(n-2)\epsilon^2}{\|y\|^2}\right)_+ y \tag{2.70}$$

Corollary 2.9 *Let $Y \sim N_n(\theta, \epsilon^2 I)$. The James-Stein estimate $\hat{\theta}^{JS+}(y)$ in* (2.70) *satisfies*

$$E\|\hat{\theta}^{JS+} - \theta\|^2 \le 2\epsilon^2 + \frac{n\epsilon^2\|\theta\|^2}{n\epsilon^2 + \|\theta\|^2}.$$

2.7 Risk of soft thresholding

A brief study of the mean squared error properties of soft threshold estimators both illustrates some of the preceding ideas and allows for a first comparison of thresholding with James-Stein shrinkage. Chapter 8 has a more systematic discussion.

Initially we adopt the unit noise setting, $X \sim N_n(\mu, I)$ and evaluate Stein's unbiased risk estimate for $\hat{\mu}_\lambda(x) = x + g_S(x)$, where the form of $g_S(x)$ for soft thresholding was given in (2.59). We have $(\partial g_{S,i}/\partial x_i)(x) = -I\{|x_i| \le \lambda\}$ a.e. and so

$$\begin{aligned} E_\mu \|\hat{\mu}_\lambda(x) - \mu\|^2 &= E_\mu[U_\lambda(x)] \\ U_\lambda(x) &= n - 2\sum_1^n I\{|x_i| \le \lambda\} + \sum_1^n \min(x_i^2, \lambda^2). \end{aligned} \tag{2.71}$$

Since $U_\lambda(x)$ depends only on λ and the observed x, it is natural to consider minimizing $U_\lambda(x)$ over λ to get a threshold estimate $\hat{\lambda}_{SURE}$.

Consider the one dimensional case with $X \sim N(\mu, 1)$. Let the (scalar) risk function $r_S(\lambda, \mu) = E_\mu[\hat{\mu}_\lambda(x) - \mu]^2$. By inserting the definition of soft thresholding and then changing variables to $z = x - \mu$, we obtain

$$r_S(\lambda, \mu) = \mu^2 \int_{-\lambda-\mu}^{\lambda-\mu} \phi(z)dz + \int_{\lambda-\mu}^{\infty} (z-\lambda)^2\phi(z)dz + \int_{-\infty}^{-\lambda-\mu} (z+\lambda)^2\phi(z)dz.$$

Several useful properties follow from this formula. First, after some cancellation, one finds that

$$\frac{\partial}{\partial \mu} r_S(\lambda, \mu) = 2\mu[\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)] \le 2\mu, \tag{2.72}$$

which shows in particular that the risk function is monotone increasing for $\mu \ge 0$ (and of course is symmetric about $\mu = 0$).

The risk at $\mu = 0$ has a simple form

$$r_S(\lambda, 0) = 2\int_\lambda^\infty (z-\lambda)^2\phi(z)dz = 2(\lambda^2 + 1)\tilde{\Phi}(\lambda) - 2\lambda\phi(\lambda)$$

and, using the bound for Mills ratio $\tilde{\Phi}(\lambda) \le \lambda^{-1}\phi(\lambda)$ valid for $\lambda > 0$, (C.15),

$$r_S(\lambda, 0) \le 2\lambda^{-1}\phi(\lambda) \le e^{-\lambda^2/2},$$

with the final inequality true for $\lambda > 2\phi(0) \doteq 0.8$.

Hence the risk increases from a typically small value at $\mu = 0$, to its value at $\mu = \infty$,

$$r_S(\lambda, \infty) = 1 + \lambda^2,$$

(which follows, for example, by inspection of (2.71)). See Figure 2.7.

Some useful risk bounds are now easy consequences. Indeed, from (2.72) we have $r_S(\lambda, \mu) - r_S(\lambda, 0) \le \mu^2$. Using also the bound at ∞, we get

$$r_S(\lambda, \mu) \le r(\lambda, 0) + \min(\mu^2, 1 + \lambda^2).$$

Making a particular choice of threshold, $\lambda_U = \sqrt{2\log n}$, and noting that $r_S(\lambda_U, 0) \le$

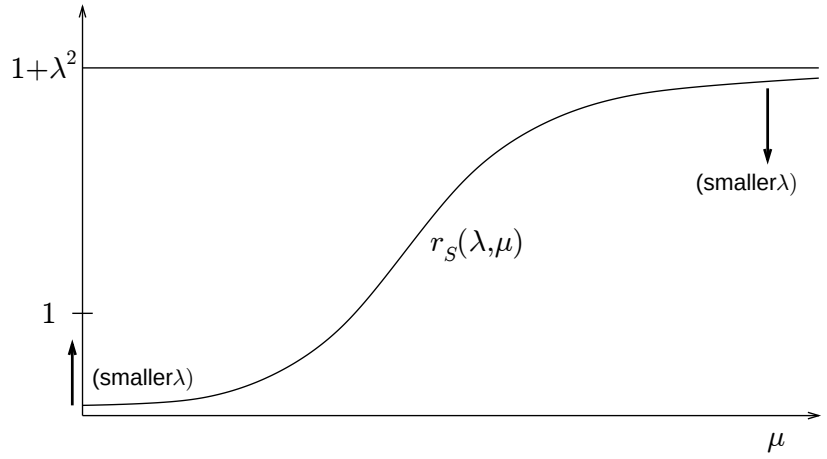

Figure 2.7 Qualitative behavior of risk function for soft thresholding. Arrows show how the risk function changes as the threshold λ is decreased.

$e^{-\lambda_U^2/2} = 1/n$, we arrive at

$$r_S(\lambda_U, \mu) \le (1/n) + (2\log n + 1)\min(\mu^2, 1).$$

Returning to noise level ϵ, and a vector observation $Y \sim N_n(\theta, \epsilon^2 I)$, and adding over the n coordinates, we can summarize our conclusions.

Lemma 2.10 *Let $Y \sim N_n(\theta, \epsilon^2 I)$ and $\hat{\theta}_\lambda$ denote soft thresholding with $\lambda = \epsilon\sqrt{2\log n}$. Then for all θ,*

$$E\|\hat{\theta}_\lambda - \theta\|^2 \le \epsilon^2 + \sum_{i=1}^n \min(\theta_i^2, \epsilon^2 + \lambda^2) \tag{2.73}$$

$$\le \epsilon^2 + (2\log n + 1)\sum_{i=1}^n \min(\theta_i^2, \epsilon^2).$$

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. Using the bound $\frac{1}{2}\min(a,b) \le ab/(a+b) \le \min(a,b)$, we find that the main term in the James-Stein bound Corollary 2.9 is

$$\frac{n\epsilon^2\|\theta\|^2}{n\epsilon^2 + \|\theta\|^2} \in [\tfrac{1}{2}, 1]\min(\sum \theta_i^2, n\epsilon^2).$$

For thresholding, looking at the main term in Lemma 2.10, we see that thresholding dominates (in terms of mean squared error) if

$$(2\log n)\sum_i \min(\theta_i^2, \epsilon^2) \ll \min\left(\sum \theta_i^2, n\epsilon^2\right).$$

For example, with $\epsilon = 1/\sqrt{n}$, and if θ is highly sparse, as for example in the case of a spike

such as $\theta = (1, 0, \ldots, 0)$, then the left side equals $(2 \log n)/n$ which is much smaller than the right side, namely 1.

Conversely, James-Stein dominates if all $|\theta_i|$ are nearly equal—recall, for example, the "comb" $\theta = \epsilon(1, \ldots, 1)$, where now the left side equals $(2 \log n) \cdot n\epsilon^2$ which is now *larger* than the right side, namely $n\epsilon^2 = 1$.

While thresholding has a smaller risk by a factor proportional to $\log n/n$ in our example, note that it can never be more than a multiplicative factor $2 \log n$ worse than James-Stein, since $\sum \min(\theta_i^2, \epsilon^2) \leq \min\left(\sum \theta_i^2, n\epsilon^2\right)$ (Exercise 2.19).

2.8 A Gaussian concentration inequality

A property of the multivariate normal model that finds frequent use in high dimensional estimation is the concentration of the distribution of Lipschitz functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be Lipschitz(L) if

$$|f(x) - f(y)| \leq L\|x - y\|$$

for all $x, y \in \mathbb{R}^n$. Here $\|x\|$ is the usual Euclidean norm on $\mathbb{R}^n$. If f is differentiable, then we can take $L = \sup \|\nabla f(x)\|$.

Proposition 2.11 *If $Z \sim N_n(0, I)$, and $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz(L), then*

$$P\{f(Z) \geq Ef(Z) + t\} \leq e^{-t^2/(2L^2)}, \tag{2.74}$$

$$P\{f(Z) \geq Med f(Z) + t\} \leq \tfrac{1}{2} e^{-t^2/(2L^2)}. \tag{2.75}$$

This property is sometimes expressed by saying that the tails of the distribution of a Lipschitz function of a Gaussian vector are *subgaussian*.

Note that the dimension n 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!

Some statistically relevant examples of Lipschitz functions include

(i) Order statistics. If $z_{(1)} \geq z_{(2)} \geq \cdots \geq z_{(n)}$ are the order statistics of a data vector z, then $f(z) = z_{(k)}$ has Lipschitz constant $L = 1$. The same is true for the absolute values $|z|_{(1)} \geq \cdots \geq |z|_{(n)}$. Section 8.9 has results on the maxima of Gaussian noise variates.

(ii) Ordered eigenvalues of symmetric matrices. Let A be an $n \times n$ symmetric matrix with eigenvalues $\varrho_1(A) \geq \varrho_2(A) \geq \cdots \geq \varrho_n(A)$. If E is also symmetric, then from Weyl's inequality (e.g. (Golub and Van Loan, 1996, p. 56 and 396))

$$|\varrho_k(A + E) - \varrho_k(A)| \leq \|E\|_{HS},$$

where $\|E\|_{HS}^2 = \sum_{i,j} e_{i,j}^2$ denotes the square of the *Hilbert-Schmidt*, or *Frobenius* norm, which is the Euclidean norm on $n \times n$ matrices. This is of statistical relevance, for example, if A is a sample covariance matrix, in which case $\varrho_1(A)$ is the largest principal component variance.

(iii) Orthogonal projections. If S is a linear subspace of $\mathbb{R}^n$, then $f(z) = \|P_S z\|$ has Lipschitz constant 1. If $\dim S = k$, then $\|P_S z\|^2 \stackrel{\mathcal{D}}{=} \chi^2_{(k)}$ and so

$$E\|P_S z\| \le \{E\|P_S z\|^2\}^{1/2} = \sqrt{k}$$

and so the inequality implies

$$P\{\, \|P_S z\| \ge \sqrt{k} + t \,\} \le e^{-t^2/2}. \tag{2.76}$$

These bounds play a key role in the oracle inequalities of Chapter 11.3.

(iv) Linear combinations of χ^2 variates. Suppose that $\alpha_i \ge 0$. Then $f(z) = (\sum \alpha_i z_i^2)^{1/2}$ is differentiable and Lipschitz: $\|\nabla f(z)\|^2 \le \|\alpha\|_\infty$. Then a fairly direct consequence of (2.74) is the tail bound

$$P\{\sum \alpha_j (Z_j^2 - 1) > t\} \le \exp\{-t^2/(32\|\alpha\|_1\|\alpha\|_\infty)\} \tag{2.77}$$

for $0 < t \le \|\alpha\|_1$ (Exercise 2.23). This is used for Pinsker's theorem in Chapter 5.4. The form $\sum \alpha_j (Z_j^2 - 1)$ also arises as the limiting distribution of degenerate U-statistics of order 1, e.g. Serfling (1980, Sec. 5.5.2).

(v) Exponential sums. The function $f(z) = \log \sum_1^n \exp(\beta z_k)$ is Lipschitz(β). It appears, for example, in the study of Gaussian likelihood ratios of sparse signals, Section 2.10.

The two concentration inequalities of Proposition 2.11 have a number of proofs. We give an analytic argument for the first that builds on Stein's integration by parts identity (2.55). For the second, we shall only indicate how the result is reduced to the isoperimetric property of Gaussian measure—see e.g. Ledoux (2001) for a more complete discussion.

We begin with a lemma that bounds covariances in terms of derivatives. Let $\gamma = \gamma^n$ denote the canonical Gaussian measure on $\mathbb{R}^n$ corresponding to $Z \sim N_n(0, 1)$.

Lemma 2.12 *Assume that $Y, Z \sim N_n(0, I)$ independently and set $Y_\theta = Y \cos\theta + Z \sin\theta$ for $0 \le \theta \le \pi/2$. Suppose that f and g are differentiable real valued functions on $\mathbb{R}^n$ with ∇f and $\nabla g \in L_2(\gamma)$. Then*

$$\mathrm{Cov}\{f(Y), g(Y)\} = \int_0^{\pi/2} E[\nabla f(Y)^T \nabla g(Y_\theta)] \sin\theta d\theta. \tag{2.78}$$

An immediate corollary of (2.78) is the Gaussian Poincaré inequality:

$$\mathrm{Var} f(Y) \le E\|\nabla f(Y)\|^2. \tag{2.79}$$

Proof Since Y and Z are independent, our covariance $Ef(Y)g(Y) - Ef(Y)Eg(Z)$ may be written $Ef(Y)[g(Y) - g(Z)]$. We exploit the path Y_θ from $Y_0 = Y$ to $Y_{\pi/2} = Z$, writing

$$g(Y) - g(Z) = -\int_0^{\pi/2} (d/d\theta) g(Y_\theta) d\theta.$$

We calculate $(d/d\theta)g(Y_\theta) = Z_\theta^T \nabla g(Y_\theta)$, where $Z_\theta = dY_\theta/d\theta = -Y \sin\theta + Z\cos\theta$. We arrive at

$$Ef(Y)[g(Y) - g(Z)] = -\int_0^{\pi/2} E[f(Y) Z_\theta^T \nabla g(Y_\theta)] d\theta. \tag{2.80}$$

The vectors Y_θ and Z_θ are independent and $N_n(0, I)$, being a rotation through angle θ of the original Y and Z, Lemma C.11. Inverting this rotation, we can write $Y = Y_\theta \cos\theta - Z_\theta \sin\theta$. Considering for now the ith term in the inner product in (2.80), we therefore have

$$\begin{aligned} E[f(Y)Z_{\theta,i} D_i g(Y_\theta)] &= E[f(Y_\theta \cos\theta - Z_\theta \sin\theta) Z_{\theta,i} D_i g(Y_\theta)] \\ &= -\sin\theta \cdot E[D_i f(Y) D_i g(Y_\theta)], \end{aligned}$$

where the second equality uses Stein's identity (2.55) applied to the $(n+i)$th component of the $2n$-dimensional spherical Gaussian vector (Y_θ, Z_θ). Adding over the n co-ordinates i and inserting into (2.80), we recover the claimed covariance formula. □

Proof of Concentration inequality (2.74). This uses an exponential moment method. By rescaling and centering, we may assume that $L = 1$ and that $Ef(Y) = 0$. We will first show that for all $t > 0$,

$$E[e^{tf(Y)}] \le e^{t^2/2}. \tag{2.81}$$

Make the temporary additional assumption that f is differentiable. The Lipschitz bound on f entails that $\|\nabla f\| \le 1$. We are going to apply the identity of Lemma 2.12 with the functions f and $g = e^{tf}$. First, observe that

$$E[\nabla f(Y)^T \nabla g(Y_\theta)] = tE[e^{tf(Y)} \nabla f(Y)^T \nabla f(Y_\theta)] \le tE e^{tf(Y)}.$$

Introduce the notation $e^{u(t)} = Ee^{tf(Y)}$, differentiate with respect to t and then use (2.78) along with the previous inequality:

$$u'(t)e^{u(t)} = E[f(Y)e^{tf(Y)}] \le \int_0^{\pi/2} te^{u(t)} \sin\theta d\theta = te^{u(t)}.$$

Hence $u'(t) \le t$ for $t > 0$ and $u(0) = 0$, from which we get $u(t) \le t^2/2$ and so (2.81). The assumption that f is differentiable can be removed by smoothing: note that the sequence $f_n = f \star \phi_{1/n}$ is Lipschitz(1) and converges to f uniformly (Exercise 2.24), so that (2.81) follows by Fatou's lemma.

Now we conclude by using Markov's inequality and (2.81). For each $t > 0$,

$$\begin{aligned} P(f(X) \ge u) &= P(e^{tf(X)} \ge e^{tu}) \\ &\le e^{-tu} Ee^{tf(X)} \le e^{-tu+t^2/2}. \end{aligned}$$

The minimizing choice of t is $t = u$, and this yields our concentration inequality. □

We finish with a remark on (2.75). If A is a subset of $\mathbb{R}^n$ and $t > 0$, the dilation $A_t = \{z \in \mathbb{R}^n : d(z, A) < t\}$. We appeal to the Gaussian isoperimetric inequality, e.g. Ledoux (2001, (2.9)), which states that if γ is canonical Gaussian measure on $\mathbb{R}^n$ and A is a Borel set such that $\gamma(A) = \Phi(a)$ for some $a \in \mathbb{R}$, then $\gamma(A_t) \ge \Phi(a + t)$ for every $t > 0$.

In particular, if we take $A = \{z : f(z) \le \mathrm{Med} f\}$, then $a = 0$ and if f is Lipschitz(1) , we have $A_t \subset \{z : f(z) \le \mathrm{Med} f + t\}$. Consequently, using the isoperimetric inequality,

$$P(f(Z) > \mathrm{Med} f + t) \le \gamma(A_t^c) \le \tilde{\Phi}(t) \le \tfrac{1}{2} e^{-t^2/2},$$

where the final inequality is (2.98) in Exercise 2.21.

2.9 Some more general linear models

In this section we briefly describe some more general Gaussian models that can be reduced to sequence form, and review some approaches to regularization. As the emphasis is on sequence models, we do not discuss recent research areas such as the lasso or compressed sensing (see Chapter 16 for some references).

Some models that reduce to sequence form. A fairly general Gaussian linear model for estimation of means in correlated noise might be described in vector notation as $Y = A\beta + \sigma Z$, or equivalently $Y \sim N(A\beta, \sigma^2\Sigma)$. Some frequently occurring subclasses of this model can be reduced to one of the three sequence forms (2.1) - (2.3).

First, when $Y \sim N_n(\beta, \epsilon^2 I)$, one can take co-ordinates in *any* orthonormal basis $\{u_i\}$ for $\mathbb{R}^n$, yielding

$$y_i = \langle Y, u_i\rangle, \qquad \theta_i = \langle \beta, u_i\rangle, \qquad z_i = \langle Z, u_i\rangle. \tag{2.82}$$

An essentially equivalent situation arises when $Y \sim N_n(A\beta, \sigma^2 I)$, and the matrix A itself has orthogonal columns: $A^TA = mI_n$. 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). The model can be put in the form (2.1) simply by premultiplying by $m^{-1}A^T$: define $y = m^{-1}A^TY, z = m^{-1/2}A^TZ$, and note especially the noise calibration $\epsilon = \sigma/\sqrt{m}$.

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

$$Y_l = f(l/n) + \sigma Z_l, \qquad l = 1, \ldots, n \tag{2.83}$$

with $Z_l \overset{iid}{\sim} N(0,1)$ becomes an example of (2.1) if one uses as design matrix an inverse discrete orthogonal wavelet (or Fourier) transform W^T to express $\mathbf{f} = (f(l/n)) = \sqrt{n}W^T\theta$. Thus here $A = \sqrt{n}W^T$ and $z = WZ$. The components of y and θ are wavelet (or Fourier) coefficients of Y and $\mathbf{f}$ respectively. Compare the discussion around (1.20) and (7.22).

If we drop the requirement (2.83) that the errors be normally distributed, keeping only the first and second moment requirements that Z have mean 0 and covariance I, then the same will be true of the transformed errors z. If the matrix W is in some sense 'dense', so that $z_i = \sum_k w_{ik}Z_k$ has many non-zero terms of similar size, then by a central limit theorem for independent summands such as Lyapunov's or Lindeberg's, the z_i will be approximately normally distributed.

Second, assume that $Y \sim N(A\beta, \epsilon^2 I)$, with A an $N \times M$ matrix. This can be converted into model (2.2) using the *singular value decomposition* $A = \sum_{i=1}^n \alpha_i u_i v_i^T$, where we assume that $\alpha_i > 0$ for $i = 1, \ldots, n = \mathrm{rank}(A)$. We obtain

$$A\beta = \sum_i \alpha_i \theta_i u_i, \qquad \theta_i = \langle v_i, \beta\rangle, \tag{2.84}$$

so that $y_i = [Y, u_i] = [A\beta, u_i] + \epsilon[Z, u_i] = \alpha_i\theta_i + \epsilon z_i$ satisfies (2.2). Here we use notation $\langle\cdot,\cdot\rangle$ and $[\cdot,\cdot]$ to distinguish inner products in domain and range spaces $\mathbb{R}^M$ and $\mathbb{R}^N$ respectively.

If one is specifically interested in the components of β, this transformation is not espe-

cially helpful. However, if the main focus is on the vector β, then the expansion $\beta = \sum \theta_i v_i$ may be useful, as can occur in the study of linear inverse problems, Chapter 3.

Interest in estimation of $\theta = A\beta$ can also arise in certain prediction problems. For example, in the "in-sample" setting, one assesses a predictor $\hat{\theta} = A\hat{\beta}$ of a new observation vector $Y^* = A\beta + \sigma Z^*$ via the mean squared error $E\|A\hat{\beta} - Y^*\|^2 = E\|A(\hat{\beta} - \beta) - \sigma Z^*\|^2 = E\|\hat{\theta} - \theta\|^2 + N\sigma^2$.

Thirdly, assume that $Y \sim N(\beta, \epsilon^2\Sigma)$, with positive definite covariance matrix Σ, with eigenvalues and eigenvectors

$$\Sigma u_i = \varrho_i^2 u_i, \qquad \varrho_i > 0,$$

so that with definitions (2.82), we recover the third sequence model (2.3), after noting that $\mathrm{Cov}(y_i, y_j) = \epsilon^2 u_i^T \Sigma u_j = \epsilon^2 \varrho_i^2 \delta_{ij}$. Here $\delta_{ij} = 1$ if $i = j$ and 0 otherwise is the usual Kronecker delta. This version arises as the limiting Gaussian model in the large sample local asymptotic normal approximation to a smooth parametric model of fixed dimension, Exercise 2.26.

Infinite sequence model analogs of the last two models are discussed in Sections 3.9 and 3.10 respectively.

In the most general setting $Y \sim N(A\beta, \epsilon^2\Sigma)$, however, a simple sequence version will typically only be possible if AA^T and Σ have the same sets of eigenvectors (including multiplicities). This does occur, for example, if AA^T and Σ are circulant matrices [2], and so are diagonalized by the discrete Fourier transform, (e.g. Gray (2006, Ch. 3)), or more generally if AA^T and Σ commute.

Penalization and regularization. The least squares estimate of β is found by minimizing $\beta \to \|Y - A\beta\|_2^2$. When $Y \sim N(A\beta, \sigma_0^2 I)$, this is also the maximum likelihood estimate. If β is high dimensional, or if A has a smoothing character with many small singular values α_i, then the least squares solution for β is often ill-determined. See below for a simple example, and Section 3.9 for more in the setting of linear inverse problems.

A commonly used remedy is to *regularize* the solution by introducing a *penalty function* $P(\beta)$, and minimizing instead the penalized least squares criterion

$$Q(\beta) = \|Y - A\beta\|_2^2 + \lambda P(\beta). \tag{2.85}$$

A minimizer of (2.85) might be called a regularized least squares estimator. If in addition $Y \sim N(A\beta, \sigma_0^2 I)$, it can also be interpreted as a posterior mode, Exercise 2.25. In this sense, then, choice of a penalty term corresponds to choice of a prior.

Two simple and commonly occurring penalty functions are *quadratic*: $P(\beta) = \beta^T \Omega \beta$ for some non-negative definite matrix Ω, and q^{th} *power*: $P(\beta) = \|\beta\|_q^q = \sum_{i=1}^n |\beta_i|^q$. If P is strictly convex, or if P is convex and $A^T A > 0$, then Q is strictly convex and so the penalized criterion has at most one global minimum. Typically a minimum exists, and we denote it $\hat{\beta}(\lambda)$.

The *kernel* of the penalty, $\ker P = \{\beta : P(\beta) = 0\}$, typically consists of "very smooth" β. In our examples, if $\Omega > 0$ is positive definite, or if $q > 0$, then necessarily $\ker P = \{0\}$. More generally, if the penalty uses, say, squared second differences, then $P_2(\beta) =$

[2] A matrix $\boldsymbol{C}$ is circulant if each row is obtained by cyclically shifting the previous row to the right by one; it is thus determined by its first row.

$\sum_{i=2}^{n-1}(\beta_{i+1} - 2\beta_i + \beta_{i-1})^2$ and $\ker P_2 = \{\beta : \beta_k = c_0 + c_1 k,\ c_0, c_1 \in \mathbb{R}\}$ consists of linear functions.

The crucial *regularization parameter* λ determines the relative weight given to the sum of squared error and penalty terms: more will be said about this later, for example in Section 3.6 and Chapter 11. As λ varies from 0 to $+\infty$, we may think of the penalized estimates $\hat{\beta}(\lambda)$ as forming a path from the roughest, least squares solution $\hat{\beta}(0) = \hat{\beta}_{LS}$ to the smoothest solution $\hat{\beta}(\infty)$ which necessarily belongs to $\ker P$.

We consider three especially important examples. First, the quadratic penalty $P(\beta) = \beta^T \Omega \beta$ is nice because it allows explicit solutions. The penalized criterion is itself quadratic:

$$Q(\beta) = \beta^T (A^T A + \lambda\Omega)\beta - 2Y^T A\beta + Y^T Y.$$

Let us assume, for convenience, that at least one of $A^T A$ and Ω is positive definite. In that case, $\partial^2 Q/\partial\beta^2 = 2(A^T A + \lambda\Omega)$ is positive definite and so there is a unique minimizer

$$\hat{\beta}(\lambda) = (A^T A + \lambda\Omega)^{-1} A^T Y. \tag{2.86}$$

This is the classical *ridge regression* or *Tikhonov regularization* estimate (see chapter notes for some references), with *ridge* matrix Ω. For each λ, the estimate is a *linear* function $S(\lambda)Y$ of the data, with smoother matrix $S(\lambda) = (A^T A + \lambda\Omega)^{-1} A^T$. The trajectory $\lambda \to \hat{\beta}(\lambda)$ shrinks from the least squares solution $\hat{\beta}(0) = (A^T A)^{-1} A^T Y$ down to $\hat{\beta}(\infty) = 0$.

Second, consider ℓ_1 penalties, which are used to promote sparsity in the solution. If the penalty is imposed after transformation to a sequence form such as (2.2) or (2.3), so that $P(\theta) = \sum |\theta_i|$, then the co-ordinatewise thresholding interpretation of Section 2.1 is available. When imposed in the original variables, so that $P(\beta) = \sum_1^n |\beta_i|$, the resulting estimator is known as the *lasso* – for **l**east **a**bsolute **s**election and **s**hrinkage **o**perator, introduced by Tibshirani (1996), see also Chen et al. (1998). There is no explicit solution, but the optimization problem is convex and many algorithms and a huge literature exist. See for example Bühlmann and van de Geer (2011) and Hastie et al. (2012).

Third, the ℓ_0 penalty $P(\beta) = \|\beta\|_0 = \#\{i : \beta_i \neq 0\}$ also promotes sparsity by penalizing the number of non-zero coefficients in the solution. As this penalty function is not convex, the solution is in general difficult to compute. However, in sufficiently sparse settings, the ℓ_0 and ℓ_1 solutions can coincide, and in certain practical settings, successful heuristics exist. (e.g. Donoho and Huo (2001), Candès and Romberg (2007), Bühlmann and van de Geer (2011)).

Example. Convolution furnishes a simple example of ill-posed inversion and the advantages of regularization. Suppose that $A = (a_{k-j}, 1 \le j, k \le n)$ so that $A\beta = a \star \beta$ represents convolution with the sequence (a_k). Figure 2.8 shows a simple example in which $a_0 = 1, a_{\pm 1} = 1/2$ and all other $a_k = 0$. Although A is formally invertible, it is nearly singular, since for $\beta_{\text{osc}} = (+1, -1, +1, \ldots, \pm 1)$, we have $A\beta_{\text{osc}} \doteq 0$, indeed the entries are exactly zero except at the boundaries. The instability of A^{-1} can be seen in the figure: the left panel shows both $y = A\beta$ and $y' = A\beta + \sigma Z$ for a given signal β and a small added noise with $\sigma = .005$ and Z being a draw from $N_n(0, I)$. Although the observations y and y' are nearly identical, the least squares estimator $\hat{\beta}_{LS} = (A^T A)^{-1} A^T y = A^{-1} y$ is very different from $\hat{\beta}'_{LS} = A^{-1} y'$. Indeed A is poorly conditioned, its smallest singular value is $\alpha_n \doteq 0.01$, while the largest singular value $\alpha_1 \doteq 2$.

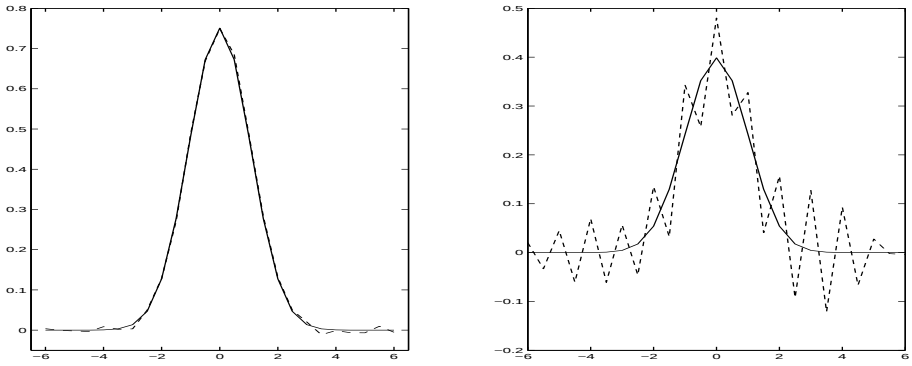

Figure 2.8 Left: Observed data $y = A\beta$, solid line, and $y' = A\beta + \sigma Z$, dashed line, for $\beta_l = \phi(t_l)$, the standard normal density with $t_l = (l/n) - 6$ and $n = 13, \sigma = 0.005$ and Z a draw from $N_n(0, I)$. Right: reconstructions $\hat{\beta}_{LS} = A^{-1}y'$, dashed line, and regularized $\hat{\beta}(\lambda)$, solid line, from (2.86) with $\lambda = 0.01 = 2\epsilon = 2\sigma$.

Regularization with the squared second difference penalty P_2 removes the difficulty: with $\lambda = 0.01$, the reconstruction $\hat{\beta}(\lambda)$ from (2.86) is visually indistinguishable from the true β.

This may be understood in the sequence domain. If the banded matrices A and Ω are lightly modified in their corners to be circulant matrices, then both are diagonalized by the (orthogonal) discrete Fourier transform, and in the Fourier coefficient domain, the effect of regularization is described by the co-ordinatewise formula (2.4). Indeed, substituting the frequency domain observation model $y_i = \alpha_i \theta_i + \epsilon z_i$, where here $\epsilon = \sigma$, we have

$$\hat{\theta}_i(y) = \frac{\alpha_i^2}{\alpha_i^2 + \lambda\omega_i}\theta_i + \frac{\epsilon\alpha_i}{\alpha_i^2 + \lambda\omega_i} z_i.$$

The sequence α_i decreases with increasing frequency i, while the regularizer constants ω_i increase. Thus at high frequencies, when $\lambda = 0$ the noise is amplified to $(\epsilon/\alpha_i)z_i$ (causing the jagged features in the figure), while when λ is positive ($= 2\epsilon$ in the figure), the term $\lambda\omega_i \gg \epsilon\alpha_i$ at high frequencies and the noise is successfully damped down.

2.10 Notes

Some of the material in this chapter is classical and can be found in sources such as Lehmann and Casella (1998).

§2. The connection between regularization with the ℓ_1 penalty and soft thresholding was exploited in Donoho et al. (1992), but is likely much older.

The soft thresholding estimator is also called a "limited translation rule" by Efron and Morris (1971, 1972).

§3. Measure theoretic details may be found, for example, in Lehmann and Romano (2005). Measurability of estimators defined by extrema, such as the Bayes estimator (2.9), requires care: see for example Brown and Purves (1973).

The characterizations of the Gaussian distribution noted in Remark 2.1(ii) are only two of many, see for example DasGupta (2011a, p157) and the classic treatise on characterizations Kagan et al. (1973).

Identity (2.23) is sometimes called Tweedie's formula (by Efron (2011) citing Robbins (1956)), and

more usually Brown's formula, for the extensive use made of it in Brown (1971). The posterior variance formula of Exercise 2.3 appears in Srinivasan (1973).

§4. Priors built up from sparse mixture priors such as (2.31) are quite common in Bayesian variable selection problems; see for example George and McCulloch (1997). The connection with posterior median thresholding and most of the results of this section come by specialization from Johnstone and Silverman (2004a), which also has some remarks on the properties of the posterior mean for these priors. Full details of the calculations for the Laplace example and a related "quasi-Cauchy" prior may be found in Johnstone and Silverman (2005a, §6).

§5. Basic material on admissibility is covered in Lehmann and Casella (1998, Ch. 5). Inadmissibility of the MLE was established in the breakthrough paper of Stein (1956). The James-Stein estimator and positive part version were introduced in James and Stein (1961), for more discussion of the background and significance of this paper see Efron (1993). Inadmissibility of the best invariant estimator of location in $n \geq 3$ dimensions is a very general phenomenon, see e.g. Brown (1966). Theorem 2.5 on eigenvalues of linear estimators and admissibility is due to Cohen (1966). Mallows' C_L and its relative C_p are discussed in Mallows (1973).

§6. Stein (1981) presented the unbiased estimate of risk, Proposition 2.6 and, among much else, used it to give the quick proof of dominance of the MLE by the James Stein estimator presented here. The Stein identity characterizes the family of normal distributions: for example, if $n = 1$ and (2.57) holds for C^1 functions of compact support, then necessarily $X \sim N(\mu, 1)$, (Diaconis and Zabell, 1991).

Many other estimators dominating the MLE have been found–one classic paper is that of Strawderman (1971). There is a large literature on extensions of the James-Stein inadmissibility result to spherically symmetric distributions and beyond, one example is Evans-Stark (1996).

The upper bound for the risk of the James-Stein estimator, Proposition 2.7 and Corollary 2.8 are based on Donoho and Johnstone (1995).

We have also not discussed confidence sets – for example, Brown (1966) and Joshi (1967) show inadmissibility of the usual confidence set and Hwang and Casella (1982) show good properties for recentering the usual set at the positive part James-Stein estimate.

§7. The unbiased risk estimate for soft thresholding was exploited in Donoho and Johnstone (1995), while Lemma 2.10 is from Donoho and Johnstone (1994a).

§8. The median version of the Gaussian concentration inequality (2.75) is due independently to Borell (1975) and Sudakov and Cirel′son (1974). The expectation version (2.74) is due to Cirel'son et al. (1976). Systematic accounts of the (not merely Gaussian) theory of concentration of measure are given by Ledoux (1996, 2001).

Our approach to the analytic proof of the concentration inequality is borrowed from Adler and Taylor (2007, Ch. 2.1), who in turn credit Chaumont and Yor (2003, Ch. 3.10), which has further references. The proof of Lemma 2.12 given here is lightly modified from Chatterjee (2009, Lemma 5.3) where it is used to prove central limit theorems by Stein's method. Tao (2011) gives a related but simpler proof of a weaker version of (2.74) with $\frac{1}{2}$ replaced by a smaller value C. An elegant approach via the semi-group of the Ornstein-Uhlenbeck process is described in Ledoux (1996, Ch. 2), this also involves an integration by parts formula.

Sharper bounds than (2.76) for the tail of χ^2 random variables are available (Laurent and Massart (1998), Johnstone (2001), Birgé and Massart (2001), `[CHECK!]`) . The constant 32 in bound (2.77) can also be improved to 8 by working directly with the chi-squared distribution.

The Hermite polynomial proof of the univariate Gaussian-Poincaré inequality (2.79) was written down by Chernoff (1981); for some historical remarks on and extensions of the inequality, see Beckner (1989).

§9. For more on Hadamard matrices and weighing designs, see for example Hedayat and Wallis (1978) and its references and citing articles.

Traditional references for ridge regression and Tikhonov regularization are, respectively, Hoerl and Kennard (1970) and Tikhonov and Arsenin (1977). A more recent text on inverse problems is Vogel (2002).

Exercises

2.1 *(Gaussian priors.)* Suppose that $\theta \sim N_n(\theta_0, T)$ and that $y|\theta \sim N_n(\theta, I)$. Let $p(\theta, y)$ denote the joint density of (θ, y). Show that

$$-2 \log p(\theta, y) = \theta^T B \theta - 2\gamma^T \theta + r(y).$$

Identify B and γ, and conclude that $\theta|y \sim N(\theta_y, \Sigma_y)$ and evaluate θ_y and Σ_y.

2.2 *(Posterior variance formula.)* Suppose that $x \sim N(\mu, 1)$ and that μ is drawn from a prior $\pi(d\mu)$ with marginal density $p(x) = (\pi \star \phi)(x)$. We saw that the posterior mean $\hat{\mu}_\pi(x) = E(\mu|x)$ is given by (the scalar version of) of (2.23). Show that the posterior *variance*

$$\mathrm{Var}(\mu|x) = \frac{d}{dx}\hat{\mu}_\pi(x) = 1 + (\log p)''(x).$$

2.3 *(Some characterizations of Gaussian priors.)* Again suppose that $x \sim N(\mu, 1)$ and that μ is drawn from a proper prior $\pi(d\mu)$ with marginal density $p(x) = (\pi \star \phi)(x)$.
(a) Suppose that p is log-quadratic, specifically that $(\log p)(x) = \alpha x^2 + \beta x + \gamma$ with $\alpha \neq 0$. Show that necessarily $\alpha < 0$ and that $\pi(d\mu)$ is Gaussian.
(b) Instead, suppose that the posterior mean $E(\mu|x) = cx + b$ for all x and some $c > 0$. Show that $\pi(d\mu)$ is Gaussian.
(c) Now, suppose only that the posterior variance $\mathrm{Var}(\mu|x) = c > 0$ for all x. Show that $\pi(d\mu)$ is Gaussian.

2.4 *(Minimum L_1 property of median.)* Let F be an arbitrary probability distribution function on $\mathbb{R}$. A median of F is any point a_0 for which

$$F(-\infty, a_0] \geq \tfrac{1}{2} \qquad \text{and} \qquad F[a_0, \infty) \geq \tfrac{1}{2}.$$

Show (without calculus!) that

$$a \to M(a) = \int |a - \theta| dF(\theta)$$

is minimized at any median a_0.

2.5 *(Bounded shrinkage for the posterior median).* Establish Proposition 2.4, for example using the steps outlined below.
(a) Show using (2.24) that

$$\mathrm{Odds}(\mu > c|X = x, \mu \neq 0) \geq \frac{\int_c^\infty e^{-\Lambda\mu}\phi(x-\mu)d\mu}{\int_{-\infty}^c e^{-\Lambda\mu}\phi(x-\mu)d\mu} \geq \frac{P(Z > -t-2)}{P(Z < -t-2)} \geq 3,$$

if $c = x - (\Lambda + t + 2)$ and Z is standard Gaussian.
(b) Show that

$$\mathrm{Odds}(\mu \neq 0|X = x) \geq \frac{(g/\phi)(x)}{(g/\phi)(x-2)} \cdot \frac{1-w}{w}(g/\phi)(t) = \frac{(g/\phi)(x)}{(g/\phi)(x-2)}.$$

(c) Using (2.25), show that

$$\frac{(g/\phi)(x)}{(g/\phi)(x-2)} \geq \exp \int_{x-2}^x (t - \Lambda)dt \geq \exp(2t + 2) \geq 2,$$

the last inequality holding if $x \geq t + \Lambda + 2$.
(d) Show that if $x \geq t + \Lambda + 2$, then $P(\mu \geq x - (t + \Lambda + 2)|X = x) \geq (3/4)(2/3) = 1/2$.

2.6 *(Posterior median formulas for Laplace priors.)* For the prior $\gamma_a(\mu) = \frac{1}{2}ae^{-a|\mu|}$, show that

$$g(x) = \tfrac{1}{2}a\exp(\tfrac{1}{2}a^2)\{e^{-ax}\Phi(x-a) + e^{ax}\tilde{\Phi}(x+a)\},$$
$$\tilde{\Gamma}(\mu|x) = \int_{\tilde{\mu}}^\infty \gamma(\mu|x)d\mu = \frac{e^{-ax}\tilde{\Phi}(\mu - x + a)}{e^{-ax}\Phi(x-a) + e^{ax}\tilde{\Phi}(x+a)}.$$

Use these expressions to verify the posterior median formula (2.42) and the threshold relation (2.44).

2.7 *(Properties of $|A|$.)* Let A be a square matrix and $|A| = (A^T A)^{1/2}$.
(i) Show how the polar decomposition $A = U|A|$, for suitable orthogonal U, can be constructed from the SVD (2.84) of A.
(ii) Let (μ_i, e_i) be eigenvalues and eigenvectors of $|A|$. Show that $\operatorname{tr} A \le \operatorname{tr} |A|$.
(iii) If equality holds in (ii), show that $Ae_i = |A|e_i$ for each i, and so that A must be symmetric.

2.8 *(Unbiased risk estimator for correlated data.)* (i) Suppose that $Y \sim N_d(\theta, V)$. For a linear estimator $\hat{\theta}_C(y) = Cy$, show that

$$r(\hat{\theta}_C, \theta) = \operatorname{tr} CVC^T + \|(I - C)\theta\|^2.$$

(ii) If, in addition, $g : \mathbb{R}^n \to \mathbb{R}^n$ is smooth and satisfies $E\{|Y_i g_i(Y)| + |D_i g_j(Y)|\} < \infty$ for all i, j, show that

$$E_\theta \|Y + g(Y) - \theta\|^2 = E_\theta\{\operatorname{tr} V + 2\operatorname{tr}[VDg(Y)] + \|g(Y)\|^2\}. \tag{2.87}$$

(iii) Suppose that all variances are equal, $V_{ii} = \sigma^2$. Show that the unbiased risk estimate for soft thresholding in this correlated case is still given by (2.71), after inserting a rescaling to noise level σ.

2.9 *(A large class of estimators dominating the MLE.)* Suppose that $X \sim N_p(\mu, I)$, $S = \sum X_i^2$, and consider estimators of the form

$$\hat{\mu}_\gamma(X) = \left(1 - \frac{\gamma(S)(p-2)}{S}\right) X.$$

Suppose that $\gamma(S) \in [0, 2]$ is non-decreasing and absolutely continuous. Show that $\hat{\mu}_\gamma$ is at least as good as the MLE: for all μ, we have $E_\mu\|\hat{\mu}_\gamma - \mu\|^2 \le p$. [This is Baranchik's (1970) theorem, and describes a class of minimax estimators.].

2.10 *(Risk of spherical shrinkage.)* Suppose that $Y \sim N_n(\theta, \epsilon^2 I)$ and that the estiator $\hat{\theta}$ is a "spherical shrinker", $\hat{\theta}(y) = c(\|y\|)y$, for some nice function $c : \mathbb{R}^+ \to \mathbb{R}$, which may depend on ϵ. (Typically $0 \le c \le 1$.) Show that the mean squared error $E_\theta\|\hat{\theta} - \theta\|^2$ is a function of $\|\theta\|$.

2.11 *(An empirical Bayes derivation of James-Stein.)* Suppose that $y_i|\theta_i \overset{\text{ind}}{\sim} N(\theta_i, \sigma_0^2)$ for $i = 1, \ldots, n$ with σ_0^2 known. Let θ_i be drawn independently from a $N(\mu, \tau^2)$ prior.
(i) Show that the (posterior mean) Bayes estimator is

$$\hat{\theta}_i = \mu + \lambda(y_i - \mu), \qquad \lambda = \tau^2/(\sigma_0^2 + \tau^2).$$

(ii) Let $\bar{y} = n^{-1}\sum_1^n y_i$ and $S^2 = \sum_1^n (y_i - \bar{y})^2$. Show that $\bar{y}$ and $(n-3)\sigma_0^2/S^2$ are unbiased for μ and $1 - \lambda$ in the marginal distribution of y_i given hyperparameters μ and τ^2.
(iii) This yields the unknown-mean form of the James-Stein estimator:

$$\hat{\theta}_i^{JS} = \bar{y} + \left(1 - \frac{(n-3)\sigma_0^2}{S^2}\right)(y_i - \bar{y}).$$

Explain how to modify the argument to recover the 'known-mean' form (2.60).

2.12 *(Positive part helps.)* Show that the positive part James-Stein estimator (2.62) has MSE smaller than the original rule (2.60): $E\|\hat{\mu}^{JS+} - \mu\|^2 < E\|\hat{\mu}^{JS} - \mu\|^2$ for all $\mu \in \mathbb{R}^n$.

2.13 *(A simpler, but weaker, version of* (2.63)*.)* Use Jensen's inequality in (2.61) to show that

$$r(\hat{\mu}_{JS}, \mu) \le 4 + n\|\mu\|^2/(n + \|\mu\|^2).$$

2.14 *(Probable risk improvements under a Gaussian prior.)*
Suppose that $\mu \sim N_n(0, \tau^2 I)$. Show that if $\|\mu\|^2 \le E(\|\mu\|^2) + kSD(\|\mu\|^2)$, then

$$r(\hat{\mu}^{JS}, \mu) \le 2 + \frac{\tau^2}{1+\tau^2}(n + k\sqrt{2n}).$$

2.15 *(Central χ^2 facts.)* Write $f_d(w) = e^{-w/2} w^{d/2-1}/[2^{d/2}\Gamma(d/2)]$ for the probability density function of χ^2_d. Verify the relations

$$w f_d(w) = d f_{d+2}(w), \tag{2.88}$$

$$(\partial/\partial w) f_{d+2}(w) = \tfrac{1}{2}[f_d(w) - f_{d+2}(w)], \tag{2.89}$$

$$P(\chi^2_d \le d) \ge 1/2. \tag{2.90}$$

and finally, for $\tau \ge 1$,

$$\tfrac{2}{5}\tau^{-1} d^{-1/2}(\tau^{-1}e^{\tau-1})^{-d/2} \le P(\chi^2_d \ge \tau d) \le \tfrac{1}{2}(\tau^{-1}e^{\tau-1})^{-d/2}. \tag{2.91}$$

(Cai, 1999). [For (2.91), show that

$$(\tau^{-1}e^{\tau-1})^{d/2} P(\chi^2_d \ge \tau d) = \frac{e^{-d/2}}{\Gamma(d/2)} \int_{d/2}^{\infty} e^{-\tau(u-d/2)} u^{d/2-1} du.$$

For the two bounds use Stirling's formula and (2.90) respectively.]

2.16 *(Poisson mixture representation for noncentral χ^2.)* Let $X \sim N_d(\mu, I)$ and define the noncentrality parameter $\xi = \|\mu\|^2$. The noncentral $\chi^2_d(\xi)$ distribution refers to the law of $W_d = \|X\|^2$. This exercise offers two verifications of the representation of its density $f_{\xi,d}$ as a Poisson$(\xi/2)$ mixture of central χ^2_{d+2j} distributions:

$$f_{\xi,d}(w) = \sum_{j=0}^{\infty} p_{\xi/2}(j) f_{d+2j}(w), \tag{2.92}$$

where $p_\lambda(j) = e^{-\lambda}\lambda^j/j!$ is the Poisson density.
(i) Show that $\chi^2_d(\xi) \overset{\mathcal{D}}{=} (Z + \sqrt{\xi})^2 + \chi^2_{d-1}$ for independent $Z \sim N(0,1)$ and χ^2_{d-1}, and hence that it suffices to do the case $d = 1$.
(ii) [Direct argument.] Write the density of $Y = X^2$ and use the Taylor series for $\cosh\sqrt{y\xi}$.
(iii) [Moment generating function.] Show that for $Z \sim N(0,1)$ and suitable s,

$$E e^{s(Z+\sqrt{\xi})^2} = (1-2s)^{-1/2} e^{\xi s(1-2s)^{-1}}.$$

(iv) Consider the difference operator $\Delta f_d = f_{d+2} - f_d$, and define the operator exponential, as usual, via $e^{\lambda\Delta} = \sum_{j\ge 0}(\lambda\Delta)^j/j!$. Show that (2.92) can be rewritten as

$$f_{\xi,d}(w) = e^{\xi\Delta/2} f_d(w). \tag{2.93}$$

2.17 *(Noncentral χ^2 facts.)* Let $S^2 \sim \chi^2_d(\xi)$ be a noncentral χ^2 variate, having density $f_{\xi,d}(w)$ and distribution function $F_{\xi,d}(w)$ Use (2.93) to show that

$$\frac{\partial}{\partial\xi} f_{\xi,d}(w) = -\frac{\partial}{\partial w} f_{\xi,d+2}(w) = \tfrac{1}{2}[f_{\xi,d+2}(w) - f_{\xi,d}(w)], \tag{2.94}$$

and hence that

$$F_{\xi,d+2}(w) = F_{\xi,d}(w) - 2 f_{\xi,d+2}(w). \tag{2.95}$$

Derive a noncentral analog of (2.88):

$$f_{\xi,d+2}(w) \le (w/d) f_{\xi,d}(w). \tag{2.96}$$

2.18 *(Exact MSE for the positive part James-Stein estimator.)*
(i) Show that the unbiased risk estimator for $\hat{\mu}^{JS+}$ is

$$U(x) = \begin{cases} n - (n-2)^2 \|x\|^{-2}, & \|x\| > n-2 \\ \|x\|^2 - n, & \|x\| < n-2. \end{cases}$$

(ii) Let $F(t;k) = P(\chi^2_k \le t)$ and $\tilde{F}(t;k) = 1 - F(t;k)$. Show that for $t \ge 0$,

$$E[\chi^2_k, \chi^2_k \le t] = kF(t;k+2)$$
$$E[\chi^{-2}_k, \chi^2_k \le t] = (k-2)^{-1} F(t;k-2).$$

(iii) If $X \sim N_n(\mu, I)$, then let $K \sim \text{Poisson}(\|\mu\|^2/2)$ and $D = n + 2K$. Show that

$$r(\hat{\mu}^{JS}, \mu) = n - E_\mu (n-2)^2/(D-2)$$
$$r(\hat{\mu}^{JS+}, \mu) = n - E_\mu \Big\{ \frac{(n-2)^2}{D-2} \tilde{F}(n-2; D-2) + 2nF(n-2; D) - DF(n-2; D+2) \Big\}.$$

[which can be evaluated using routines for $F(t;k)$ available in many software packages.]

2.19 *(Comparison of shrinkage and thresholding.)* As in Section 2.7, use ideal risk $n\epsilon^2\|\theta\|^2/(n\epsilon^2 + \|\theta\|^2)$ as a proxy for the risk of the James-Stein estimator. Show that

$$\sum_{i=1}^n \min(\theta_i^2, \epsilon^2) \le 2 \frac{n\epsilon^2\|\theta\|^2}{n\epsilon^2 + \|\theta\|^2},$$

and identify sequences (θ_i) for which equality occurs. Thus, verify the claim in the last paragraph of Section 2.7.

2.20 *(Shrinkage and thresholding for an approximate form of sparsity.)* Suppose that $\epsilon = n^{-1/2}$ and $p < 2$. Compare the the large n behavior of the MSE of James-Stein estimation and soft thresholding at $\lambda = \epsilon\sqrt{2\log n}$ on the weak-ℓ_p-extremal sequences

$$\theta_k = k^{-1/p}, \qquad k = 1, \dots, n.$$

2.21 *(Simple Gaussian tail bounds.)* (a) Let $\tilde{\Phi}(t) = \int_t^\infty \phi(s)ds$ and show that for $t > 0$,

$$\tilde{\Phi}(t) \le \phi(t)/t. \tag{2.97}$$

(b) By differentiating $e^{t^2/2}\tilde{\Phi}(t)$, show also that for $t > 0$,

$$\tilde{\Phi}(t) \le \tfrac{1}{2} e^{-t^2/2}. \tag{2.98}$$

2.22 *(Median and mean for maxima.)* If $Z \sim N_n(0, I)$ and M_n equals either $\max_i Z_i$ or $\max_i |Z_i|$, then use (2.74) to show that

$$|EM_n - \text{Med}M_n| \le \sqrt{2\log 2}. \tag{2.99}$$

(Massart (2007)).

2.23 *(Chi-squared tail bound.)* Use the inequality $(1+x)^{1/2} \geq 1 + x/4$ for $0 \leq x \leq 1$ to verify (2.77).

2.24 *(Easy approximate delta function results.)* (a) Let ϕ_ϵ denote the $N_n(0, \epsilon^2 I)$ density in $\mathbb{R}^n$. Suppose that f is Lipschitz(L) and define $f_\epsilon = f \star \phi_\epsilon$. By writing the convolution in two ways, show that f_ϵ is still Lipschitz(L), but also differentiable, even C^∞, and that $f_\epsilon(x) \to f(x)$ uniformly on $\mathbb{R}^n$ as $\epsilon \to 0$.

(b) Now suppose that $\psi \geq 0$ is a C^∞ function with $\int \psi = 1$ and support contained in the unit ball $\{x \in \mathbb{R}^n : \|x\| \leq 1\}$. [Such functions exist and are known as *mollifiers*.] Let $\psi_\epsilon = \epsilon^{-n}\psi(x/\epsilon)$. Suppose that f is continuous on $\mathbb{R}^n$ and define $f_\epsilon = f \star \psi_\epsilon$. Use the same arguments as in (a) to show that f_ϵ is differentiable, even C^∞), and that $f_\epsilon \to f$ uniformly on compact sets in $\mathbb{R}^n$.

[Part (a) is used in the proof of the concentration inequality, Proposition (2.74), while part (b) is a key component of the proof of the approximation criterion for weak differentiability used in the proof of Stein's unbiased risk estimate, Proposition 2.6 – see C.22]

2.25 *(Regularization and Bayes' rule.)* Suppose that $Y \sim N_n(A\beta, \sigma^2 I)$. Show that the minimizer $\hat{\beta}$ of the penalized least squares criterion (2.85) can be interpreted as the posterior *mode* of a suitable prior $\pi(d\beta)$ and identify π.

2.26 *(Local asymptotic normality and the Gaussian model.)* Suppose $X_1, \ldots, X_n \overset{\text{ind}}{\sim} f_\theta(x)\nu(dx)$ for $\theta \in \Theta \subset \mathbb{R}^p$. Let the loglikelihood for a single observation be $\ell_\theta = \log f_\theta(x)$. Write ∂_i for $\partial/\partial\theta_i$ and set $\dot{\ell}_\theta = (\partial_i \ell_\theta)$ and $\ddot{\ell}_\theta = (\partial_{ij}\ell_\theta)$. Under common regularity conditions, the Fisher information matrix $I_\theta = E_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T = -E_\theta \ddot{\ell}_\theta$. The following calculations make it plausible that the models $\mathcal{P}_n = (P^n_{\theta_0 + h/\sqrt{n}}, h \in \mathbb{R}^p)$ and $\mathcal{Q} = (N(h, I_{\theta_0}^{-1}), h \in \mathbb{R}^p)$ have similar statistical properties for n large. Full details may be found, for example, in van der Vaart (1998, Ch. 7).

(i) Show that

$$\log \prod (f_{\theta_0 + h/\sqrt{n}}/f_{\theta_0})(X_l) = h^T \Delta_{n,\theta_0} - \tfrac{1}{2} h^T I_{n,\theta_0} h + o_p(1) \tag{2.100}$$

and that under $P^n_{\theta_0}$,

$$\Delta_{n,\theta_0} = n^{-1/2} \sum \dot{\ell}_{\theta_0}(X_l) \Rightarrow N(0, I_{\theta_0}), \qquad I_{n,\theta_0} = -n^{-1} \sum \ddot{\ell}_{\theta_0}(X_l) \to I_{\theta_0}.$$

(ii) Let $g_h(y)$ denote the density of $N(h, I_{\theta_0}^{-1})$ and show that

$$\log(g_h/g_0)(y) = h^T I_{\theta_0} y - \tfrac{1}{2} h^T I_{\theta_0} h. \tag{2.101}$$

(iii) The two log likelihood ratios look similar if we define $Y_n = I_{\theta_0}^{-1}\Delta_{n,\theta_0}$. Argue heuristically that under $P^n_{\theta_0 + h/\sqrt{n}}$, we have $Y_n \Rightarrow N(h, I_{\theta_0}^{-1})$.

3

The infinite Gaussian sequence model

> It was agreed, that my endeavors should be directed to persons and characters supernatural, or at least romantic, yet so as to transfer from our inward nature a human interest and a semblance of truth sufficient to procure for these shadows of imagination that willing suspension of disbelief for the moment, which constitutes poetic faith. (Samuel Taylor Coleridge, *Biographia Literaria*, 1817)

For the first few sections, we focus on the infinite white Gaussian sequence model

$$y_i = \theta_i + \epsilon z_i \qquad i \in \mathbb{N}. \tag{3.1}$$

For some purposes and calculations this is an easy extension of the finite model of Chapter 2, while in other respects important new issues emerge. For example, the unbiased estimator $\hat{\theta}(y) = y$ has infinite mean squared error, and bounded parameter sets are no longer necessarily compact, with important consequences that we will see.

Right away, it must be remarked that we are apparently attempting to estimate an infinite number of parameters on the basis of what must necessarily be a finite amount of data. This calls for a certain suspension of disbelief which the theory attempts to reward.

Essential to the effort is some assumption that most of the θ_i are small in some sense. In this chapter we require θ to belong to an ellipsoid. In terms of functions expressed in a Fourier basis, this corresponds to mean-square smoothness. This and some consequences for mean squared error of linear estimators over ellipsoids are developed in Section 3.2, along with a first rate of convergence result, for a truncation estimator that ignores all high frequency information.

We have seen already in the introductory Section 1.4 that (3.1) is equivalent to the continuous Gaussian white noise model. This connection, along with the heuristics also sketched there, allow us to think of this model as approximating the equispaced nonparametric regression model $Y_l = f(l/n) + \sigma Z_l$, compare (1.13). This opens the door to using (3.1) to gain insight into frequently used methods of nonparametric estimation. Thus, kernel and smoothing spline estimators are discussed in Sections 3.3 and 3.4 respectively, along with their bias and variance properties. In fact, a smoothing spline estimator is a kernel method in disguise and in the sequence model it is fairly easy to make this explicit, so Section 3.5 pauses for this detour.

Mean squared error properties return to the agenda in Section 3.6. The worst case MSE of a given smoothing spline over an ellipsoid (i.e. smoothness class) is calculated. This depends on the regularization parameter of the spline estimator, which one might choose to minimize

the worst case MSE. With this choice, standard rate of convergence results for smoothing splines can be derived.

The rest of the chapter argues that the splendid simplicity of the sequence model (3.1) actually extends to a variety of other settings. Two approaches are reviewed: transformation and approximation. The transformation approach looks at models that can be put into the independent Gaussian sequence form $y_i = \theta_i + \epsilon \varrho_i z_i$ for $i \in \mathbb{N}$ and known positive constants ϱ_i. This can be done for linear inverse problems with white Gaussian noise via the singular value decomposition, Section 3.9, and for processes with correlated Gaussian noise via the Karhunen-Loève transform (aka principal components), Section 3.10.

The approximation approach argues that with sufficient data, more concrete nonparametric function estimation problems such as density and spectral density estimation and flexible regression models "look like" the Gaussian sequence model. Methods and results can in principle, and sometimes in practice, be transferred from the simple white noise model to these more applications oriented settings. Section 3.11 gives a brief review of these results, in order to provide further motivation for our detailed study of the Gaussian sequence model in later chapters.

3.1 Parameter spaces and ellipsoids

An informal description of the Gaussian white noise model, in both continuous and discrete forms, was already given in Chapter 1.4. Some of the interesting features of this infinite dimensional model leap out as soon as we try to describe it more formally and attempt to define parameter spaces and minimax risks.

Begin with the sequence form, (1.3), which puts $y_i = \theta_i + \epsilon z_i$ for $i \in \mathbb{N}$ and $z_i \overset{\text{iid}}{\sim} N(0,1)$. Denote the probability measure corresponding to $y = (y_i, i \in \mathbb{N})$ by P_θ. It is defined on the sample space $\mathbb{R}^\infty$ equipped with the Borel σ-field – see the Chapter Notes for the topological fine print. A first important feature is that the measures P_θ are *not* all mutually absolutely continuous, in contrast with the finite-dimensional case.

To explain this, we recall some definitions. Two probability measures P and Q are called *equivalent* if each of P and Q are mutually absolutely continuous relative to the other. They are called *orthogonal*, or *mutually singular*, if there exists a set A with $P(A) = 0$ and $Q(A) = 1$.

It turns out, in this Gaussian setting, that P_θ is either equivalent or orthogonal to the pure noise model P_0 in which $\theta = 0$. It is equivalent to P_0 if and only if $\theta \in \ell_2$, with likelihood ratio given by

$$\frac{dP_\theta}{dP_0}(y) = \exp\left\{\frac{\langle y, \theta\rangle}{\epsilon^2} - \frac{\|\theta\|^2}{2\epsilon^2}\right\}.$$

All this follows from Kakutani's theorem, which will be recalled in detail in the more general setting of Section 3.8. Note that the random variable $\langle y, \theta\rangle$ appearing in the likelihood ratio has a $N(0, \|\theta\|^2)$ distribution under P_0 and in particular is finite P_0-almost surely.

As an example of a case in which P_θ is orthogonal to P_0 we may suppose that $\theta_i \equiv 1$ for all i, and let $A = \{(y_i) \in \mathbb{R}^\infty : n^{-1}\sum_{i=1}^n y_i \to 1\}$. It follows from the strong law of large numbers that $P_\theta(A) = 1$ while $P_0(A) = 0$.

The continuous form (1.21) puts $Y(t) = \int_0^t f(t)dt + \epsilon W(t), 0 \le t \le 1$. The sample

space is taken to be $C[0,1]$, the space of continuous functions on $[0,1]$ equipped with the norm $\|f\|_\infty = \sup\{|f(x)|, x \in [0,1]\}$ and the Borel σ-field. We denote the probability measure corresponding to $\{Y(t), 0 \le t \le 1\}$ by P_f. The sequence form is recovered from the continuous form by taking coefficients in an orthonormal basis $\{\varphi_i\}$, compare (1.24).

Turning to the loss function, we focus in this chapter on squared error, except in Section 3.11. Thus $L(\hat\theta, \theta) = \|\hat\theta - \theta\|_2^2 = \sum_{i\in\mathbb{N}}(\hat\theta_i - \theta_i)^2$ and the mean squared error

$$r(\hat\theta, \theta) = E_\theta L(\hat\theta(y), \theta) = E_\theta \|\hat\theta(y) - \theta\|_2^2.$$

This can be expressed in terms of functions and in the continuous time domain using the Parseval relation (1.25), yielding $r(\hat f, f)$.

Suppose that θ is restricted to lie in a *parameter space* $\Theta \subset \ell_2$ and compare estimators through their worst case risk over Θ. A particular importance attaches to the best possible worst-case risk, called the *minimax risk* over Θ:

$$R_N(\Theta, \epsilon) = \inf_{\hat\theta} \sup_{\theta\in\Theta} E_\theta L(\hat\theta(y), \theta). \tag{3.2}$$

The subscript "N" is a mnemonic for "non-linear" estimators, to emphasise that no restriction is placed on the class of estimators $\hat\theta$. 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. Note also that we will often drop explicit reference to the noise level ϵ, writing simply $R_N(\Theta)$ or $R_\mathcal{E}(\Theta)$.

This is an extension of the notion of minimax risk over $\mathbb{R}^n$, introduced in Section 2.5. Indeed, in (3.2) we are *forced* to consider proper subsets Θ of $\ell_2(\mathbb{N})$–this is a second new feature of the infinite dimensional model. To see this, recall the classical minimax result quoted at (2.51), namely 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 we must have $R_N(\ell_2(\mathbb{N}), \epsilon) = \infty$, and in particular for *any* estimator $\hat\theta$

$$\sup_{\theta\in\ell_2(\mathbb{N})} E_\theta \|\hat\theta - \theta\|_2^2 = \infty. \tag{3.3}$$

Thus, a fundamental feature of non-parametric estimation is that some *a priori* restriction on the class of signals θ is required in order to make meaningful comparisons of estimators.

Fortunately, a great variety of such classes is available:

Lemma 3.1 *If Θ is compact in ℓ_2, then for ℓ_2 error, $R_N(\Theta, \epsilon) < \infty$.*

Proof Just consider the zero estimator $\hat\theta_0 \equiv 0$: then $\theta \to r(\hat\theta_0, \theta) = \|\theta\|_2^2$ is continuous on the compact Θ and so attains its maximum: $R_N(\Theta) \le \sup_\Theta r(\hat\theta_0, \theta) < \infty$. □

Two important classes of parameter spaces are the *ellipsoids* and *hyperrectangles*:

$$\Theta(a, C) = \{\theta : \sum_{i=1}^{\infty} a_i^2\theta_i^2 \le C^2\}, \tag{3.4}$$

$$\Theta(\tau) = \{\theta : |\theta_i| \le \tau_i \text{ for all } i\} \tag{3.5}$$

We will see that each class can be used to encode different types of smoothness for functions $f \in L_2[0,1]$. For now, we record criteria for compactness, with the proofs as Exercise 3.1.

Lemma 3.2 *The ellipsoid $\Theta(a, C)$ is ℓ_2-compact if and only if $a_i > 0$ and $a_i \to \infty$. The hyperrectangle $\Theta(\tau)$ is ℓ_2-compact if and only if $\sum \tau_i^2 < \infty$.*

Compactness is not necessary for finiteness of the minimax risk, as the classical finite-dimensional result (2.51) already shows. Indeed, Lemma 3.1 extends to sets of direct product form $\Theta = \mathbb{R}^r \times \Theta'$, where $r < \infty$ and Θ' is compact. We will need this in the next paragraph; the easy proof is Exercise 3.2. 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 Θ being $\|\cdot\|$-compact.

Remark - LABEL?. Lemma 3.1 remains true if we assume only that Θ is *bounded* in ℓ_2–norm, see Exercise 3.3. So the reader might wonder why we focused immediately on the stronger condition of compactness. The reason comes from the small noise limit $\epsilon \to 0$: if Θ is norm bounded and closed but not compact (e.g. any norm ball in ℓ_2), then Section 5.5 shows that *no* estimator can be consistent: $R_N(\Theta, \epsilon) \not\to 0$. Thus, such parameter sets Θ are in some sense "too large".

Ellipsoids and mean square smoothness. Ellipsoids furnish one of the most important and interpretable classes of examples of parameter spaces Θ. Consider first the continuous form of the Gaussian white noise model (1.21). For the moment, we restrict attention to the subspace $L_{2,\mathrm{per}}[0, 1]$ of square integrable *periodic* functions on $[0, 1]$. For integer $\alpha \geq 1$, let $f^{(\alpha)}$ denote the αth derivative of f and

$$\mathcal{F} = \mathcal{F}(\alpha, L) = \Big\{ f \in L_{2,\mathrm{per}}[0,1] : f^{(\alpha-1)} \text{ is absolutely continuous, and } \int_0^1 [f^{(\alpha)}(t)]^2 dt \leq L^2 \Big\}. \tag{3.6}$$

Thus, the average L_2 norm of $f^{(\alpha)}$ is required not merely to be finite, but also to be less than a quantitative bound L–as we will shortly see, this guarantees finiteness of risks. Historically, considerable statistical interest focused on the behavior of the minimax estimation risk

$$R_N(\mathcal{F}, \epsilon) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E \int_0^1 [\hat{f} - f]^2 \tag{3.7}$$

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

The ellipsoid interpretation of the parameter spaces $\mathcal{F}(\alpha, L)$ comes from the sequence form of the white noise model. Consider the orthonormal trigonometric basis for $L_2[0, 1]$, in which

$$\varphi_1(t) \equiv 1, \qquad \begin{cases} \varphi_{2k}(t) &= \sqrt{2}\sin 2\pi k t \\ \varphi_{2k+1}(t) &= \sqrt{2}\cos 2\pi k t, \end{cases} \qquad k = 1, 2, \ldots \tag{3.8}$$

Let $\theta_k = \langle f, \varphi_k \rangle = \int_0^1 f\varphi_k$ denote the Fourier coefficients of f. Let $\check{\Theta}_2^\alpha(C)$ denote the ellipsoid (3.4) with semiaxes

$$a_1 = 0, \qquad a_{2k} = a_{2k+1} = (2k)^\alpha. \tag{3.9}$$

Thus $\check{\Theta}_2^\alpha(C)$ has the form $\mathbb{R} \times \Theta'$ with Θ' compact. Furthermore, it exactly captures the notion of smoothness in mean square.

Lemma 3.3 *Suppose $\alpha \in \mathbb{N}$. For $f \in L_{2,\mathrm{per}}[0,1]$, let $\theta = \theta[f] \in \ell_2$ denote coefficients in the Fourier basis* (3.8). *Let $\mathcal{F}(\alpha, L)$ be given by* (3.6) *and $\check{\Theta}_2^\alpha(C)$ the ellipsoid with semiaxes* (3.9). *Then $f \in \mathcal{F}(\alpha, \pi^\alpha C)$ if and only if $\theta \in \check{\Theta}_2^\alpha(C)$.*

Proof Outline (For full details, see e.g. Tsybakov (2009, pp. 196-8)). Differentiation takes a simple form in the Fourier basis: if $l = 2k$ or $2k+1$, then $\varphi_l^{(\alpha)} = \pm(2\pi k)^\alpha \varphi_{l'}$, with $l' = 2k$ or $2k+1$ also. In addition, $l' = l$ iff α is even. Hence, if $f = \sum \theta_l \varphi_l$, and $f^{(\alpha-1)}$ is absolutely continuous then

$$\int [f^{(\alpha)}]^2 = \pi^{2\alpha} \sum a_l^2 \theta_l^2, \tag{3.10}$$

so that if (3.6) holds, then $\theta \in \check{\Theta}_2^\alpha(C)$. For the converse, one shows first that finiteness of $\sum a_l^2 \theta_l^2$ implies that $f^{(\alpha-1)}$ exists and is absolutely continuous. Then for $\theta \in \check{\Theta}_2^\alpha(C)$, the previous display shows that (3.6) holds. □

The statistical importance of this result is that the function space minimax risk problem (3.7) is equivalent to a sequence space problem (3.2) under squared ℓ_2 loss. In the sequence version, the parameter space is an ellipsoid. Its simple geometric form was exploited by Pinsker (1980) to give a complete solution to the description of minimax risk and estimators. We shall give Pinsker's solution in Chapter 5 as an illustration of tools that will find use for other parameter sets Θ in later chapters.

We often simplify the semiaxes (3.9) to $a_k = k^\alpha$ and set

$$\Theta_2^\alpha(C) = \{\theta : \sum_{k=1}^{\infty} k^{2\alpha} \theta_k^2 \le C^2\}. \tag{3.11}$$

The two definitions are quite similar, indeed $\Theta_2^\alpha(C) \subset \check{\Theta}_2^\alpha(C) \subset \mathbb{R} \times \Theta_2^\alpha(C)$, compare Exercise REF, and so their minimax risks are very close,

$$R_N(\Theta_2^\alpha(C), \epsilon) \le R_N(\check{\Theta}_2^\alpha(C), \epsilon) \le R_N(\Theta_2^\alpha(C), \epsilon) + \epsilon^2. \tag{3.12}$$

Remark. The ellipsoid view of mean-square smoothness extends to non-integer α. Finiteness of $\sum k^{2\alpha}\theta_k^2$ can then be taken as a definition of finiteness of the Sobolev seminorm $\|f^{(\alpha)}\|_2$ even for non-integer α. Appendices B and C.25 contain further details and references.

3.2 Linear estimators and truncation

Linear estimators are simple and widely used, and so 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. Some of these will be explored in more detail in the next sections, but we begin with some general remarks applicable to all linear estimators. Then we do a first example of a rate of convergence calculation using a simple class of estimators which truncate high frequencies.

In the sequence model, all linear estimates 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. It is therefore easy to extend the discussion of linear estimators in Section 2.5 to the infinite case. Thus the

mean squared error of $\hat{\theta}_C$ is still given by (2.47), $r(\hat{\theta}, \theta) = \epsilon^2 \mathrm{tr}\, C^T C + \|(I - C)\theta\|^2$, one must only pay attention now to the convergence of infinite sums. In particular, for $r(\hat{\theta}_C, \theta)$ to be finite, C needs to have finite Hilbert-Schmidt, or Frobenius, norm

$$\|C\|_{HS}^2 = \mathrm{tr}\, C^T C = \sum_{i,j=1}^{\infty} c_{ij}^2 < \infty. \tag{3.13}$$

Thus, C must be a bounded linear operator on ℓ_2 with square summable singular values. In particular, in the infinite sequence case, the maximum likelihood estimator $C = I$ must be excluded! Hence the bias term is necessarily unbounded over all of ℓ_2, namely $\sup_{\theta \in \ell_2} r(\hat{\theta}_C, \theta) = \infty$, as is expected anyway from the general result (3.3).

The convergence in (3.13) implies that most c_{ij} will be small, corresponding at least heuristically, to the notion of shrinkage. Thus, familiar smoothing methods such as the Wiener filter and smoothing splines are indeed linear shrinkers except possibly for a low dimensional subspace on which no shrinkage is done. Recall, for example, formula (1.18) for the smoothing spline estimator in the Demmler-Reinsch basis from Section 1.4, in which $w_1 = w_2 = 0$ and w_k increases for $k \geq 3$. This shrinks all co-ordinates but the first two.

More generally, in the infinite sequence model it is again true that reasonable linear estimators must shrink in all but at most two eigendirections. Indeed Theorem 2.5 extends to the infinite sequence model (3.1) in the natural way: a linear estimator $\hat{\theta}_C(y) = Cy$ is admissible for squared error loss if and only if C is symmetric with finite Hilbert-Schmidt norm (3.13) and eigenvalues $\varrho_i(C) \in [0, 1]$ with at most two $\varrho_i(C) = 1$ (Mandelbaum, 1984).

A particularly simple class of linear estimators is given by diagonal shrinkage, $C = \mathrm{diag}(c_k)$ for a sequence of constants $c = (c_k)$. In this case, we write $\hat{\theta}_c$ and the MSE decomposition simplifies to

$$r(\hat{\theta}_c, \theta) = \epsilon^2 \sum_k c_k^2 + \sum_k (1 - c_k)^2 \theta_k^2. \tag{3.14}$$

This form is easy to study because it is additive in the co-ordinates. Thus, it can be desirable that the basis $\{\varphi_k\}$ be chosen so that the estimators of interest have diagonal shrinkage form. We will see how this can happen with kernel and spline estimators in Sections 3.3 and 3.4.

Maximum risk over ellipsoids. We illustrate by deriving an expression for the maximum risk of a diagonal linear estimator over an ellipsoid.

Lemma 3.4 *Assume the homoscedastic white noise model $y_k = \theta_k + \epsilon z_k$. Let $\Theta = \Theta(a, C) = \{\theta : \sum a_k^2 \theta_k^2 \leq C^2\}$ and consider a diagonal linear estimator $\hat{\theta}_c(y) = (c_k y_k)$. Then the maximum risk*

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

Proof The diagonal linear estimator has variance-bias decomposition (3.14). The worst case risk over Θ has a corresponding form

$$\bar{r}(\hat{\theta}_c; \epsilon) = \sup_{\theta \in \Theta} r(\hat{\theta}_c, \theta) = \bar{V}(\epsilon) + \bar{B}^2(\Theta). \tag{3.15}$$

The max variance term $\bar{V}(\epsilon) = \epsilon^2 \sum_k c_k^2$ does not depend on Θ. On the other hand, the

max bias term $\bar{B}^2(\Theta) = \sup_\Theta \sum_k (1-c_k)^2\theta_k^2$ does not depend on the noise level ϵ. It does depend on Θ, but can be easily evaluated on ellipsoids.

We remark that if $a_k = 0$ for some k, then $\bar{B}(\Theta) = \infty$ unless $c_k = 1$. In the latter case, the kth index does not contribute to $\bar{B}(\Theta)$. So without loss of generality, we assume that $a_k > 0$ for all k and make new variables $s_k = a_k^2\theta_k^2/C^2$. The linear function $(s_k) \to \sum d_k s_k$ has maximum value $\sup d_k$ over the non-negative simplex $\sum s_k \le 1$. Hence,

$$\bar{B}^2(\Theta) = C^2 \sup_k \; a_k^{-2}(1-c_k)^2. \tag{3.16}$$

and the lemma follows from (3.15). □

Next, we apply this expression to the simplest of diagonal shrinkage methods and do a first rates of convergence calculation.

Truncation estimators and Rates of Convergence

A particularly simple class of linear estimators is given by projection onto a subset of the co-ordinate axes: $(P_{\mathcal{I}} y)_i = y_i$ if and only if $i \in \mathcal{I}$. If the indices i correspond to frequency and the focus is on smoothing, it may be reasonable to restrict attention to nested sets of low frequencies $\mathcal{I}_\nu = \{i : i \le \nu\}$. We might call such a rule

$$\hat{\theta}_{\nu,i}(y) = \begin{cases} y_i & i \le \nu, \\ 0 & i > \nu \end{cases} \tag{3.17}$$

a *truncation*, or *spectral cutoff* estimator, as it discards frequencies above ν. *Caution*: a truncation estimator is quite different from a threshold estimator, e.g. (2.6)—the truncation estimator decides in advance, based on index i, and is linear, while the threshold estimator uses the data y_i and is *non*linear.

It is then natural to ask how to choose ν. The MSE at a given θ arises from variance at low frequencies and from bias at high ones:

$$r(\hat{\theta}_\nu, \theta) = \sum_i E(\hat{\theta}_{\nu,i} - \theta_i)^2 = \nu\epsilon^2 + \sum_{i>\nu} \theta_i^2.$$

This follows from the MSE formula (3.14) for diagonal estimators, by noting that c corresponds to a sequence beginning with ν ones followed by zeros.

Of course θ is unknown, but adopting the minimax approach, one might suppose that a particular ellipsoid $\Theta(a, C)$ is given, and then find that value of ν which minimizes the maximum MSE over that ellipsoid. Using Lemma 3.4, for an ellipsoid with $k \to a_k^2$ increasing, we have for the maximum risk

$$\bar{r}(\hat{\theta}_\nu; \epsilon) := \sup_{\theta \in \Theta(a,C)} r(\hat{\theta}_\nu, \theta) = \nu\epsilon^2 + C^2 a_{\nu+1}^{-2}.$$

Now specialize further to the mean-square smoothness classes $\breve{\Theta}_2^\alpha(C)$ in the trigonometric basis (3.8) in which the semi-axes a_i follow the polynomial growth (3.9). If we truncate at frequency k, then $\nu = 2k+1$ (remember the constant term!) and

$$\bar{r}(\hat{\theta}_\nu; \epsilon) = (2k+1)\epsilon^2 + C^2(2k+2)^{-2\alpha}.$$

[Note that the result is the same for both $\check{\Theta}_2^\alpha(C)$ and $\Theta_2^\alpha(C)$.]

As the cut-off frequency k increases, there is a trade-off of increasing variance with decreasing bias–here and quite generally, this is a characteristic feature of linear smoothers indexed by a model size parameter. The maximum risk function is convex in k, and the optimal value is found by differentiation[1]:

$$2k_* + 2 = (2\alpha C^2/\epsilon^2)^{1/(2\alpha+1)}.$$

Substituting this choice into the previous display and introducing the *rate of convergence* index $r = 2\alpha/(2\alpha + 1)$, we find

$$\begin{aligned}\bar{r}_*(\epsilon) &= \min_\nu \max_{\theta\in\Theta_2^\alpha(C)} r(\hat{\theta}_\nu, \theta) \\ &= (2\alpha)^{1/(2\alpha+1)} C^{2(1-r)}\epsilon^{2r} + C^2(2\alpha C^2/\epsilon^2)^{-r} + O(\epsilon^2) \\ &\sim b_\alpha C^{2(1-r)}\epsilon^{2r}, \end{aligned} \tag{3.18}$$

as $\epsilon \to 0$, where the constant $b_\alpha = (2\alpha)^{1/(2\alpha+1)}(1 + 1/(2\alpha))$, and as usual the notation $a(\epsilon) \sim b(\epsilon)$ means that $a(\epsilon)/b(\epsilon) \to 1$ as $\epsilon \to 0$.

The calculation uncovers some important properties:

- The optimum cutoff frequency depends on the *signal to noise ratio* C/ϵ and the amount of smoothness α that is assumed—indeed k_* increases with C/ϵ and typically decreases with α.
- The 'rate of convergence' as $\epsilon \to 0$ is ϵ^{2r}. If one thinks of ϵ^2 as a proxy for inverse sample size $1/n$, then the rate [2] becomes $r = 2\alpha/(2\alpha + 1)$.
- The rate r increases with smoothness α: for twice differentiable functions $r = 4/5$, and r increases to 1 as $\alpha \nearrow \infty$.

While we have established the performance at the best choice of truncation frequency, we do not yet know whether better rates of convergence might be attained by other estimators over the same smoothness classes $\Theta_2^\alpha(C)$. It turns out that this rate is indeed optimal, and the development of lower bounds, applicable to *all* estimators, is a major task that begins in Chapter 4. For this example in particular, see Section 4.7, especially Proposition 4.23.

3.3 Kernel Estimators

Kernel estimators form an important and widely used class of linear methods in nonparametric regression and density estimation problems and beyond. We give a definition in the continuous Gaussian white noise model, discuss the connection with certain non-parametric regression settings, and then begin to look at bias, variance and MSE properties. Finally, the sequence space form of a kernel estimator is derived in the Fourier basis, and shown to have diagonal shrinkage form.

[1] We ignore the fact that k_* should be an integer: as $\epsilon \to 0$, it turns out that using say $[k_*]$ would add a term of only $O(\epsilon^2)$, which will be seen to be negligible. See also the discussion on page 75.

[2] it would be more correct to write "rate index" to refer to r, but here and throughout we simply say "rate".

A *kernel* $K(u)$ is a real valued, square integrable function with $\int K(u)du = 1$, *not necessarily* non-negative. The kernel is scaled to have *bandwidth* h

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right).$$

Some common kernels include

$$K(t) = \begin{cases} (2\pi)^{-1/2}e^{-t^2/2} & \text{Gaussian} \\ (1/2)I_{[-1,1]}(t) & \text{Uniform} \\ (3/4)(1-t^2)I_{[-1,1]}(t) & \text{Quadratic/Epanechnikov} \\ (15/16)(1-t^2)^2 I_{[-1,1]}(t) & \text{Biweight.} \end{cases} \tag{3.19}$$

These are all symmetric and non-negative; all but the first also have compact support.

Our theory will be developed for estimation of functions f periodic on $[0,1]$, that is $f(t+j) = f(t)$ for all $j \in \mathbb{Z}$. We also form the periodized kernel

$$\mathring{K}_h(t) = \sum_{j\in\mathbb{Z}} K_h(t+j). \tag{3.20}$$

For simplicity we assume that K has compact support $[-t_0, t_0]$, which guarantees convergence of (3.20), though the results hold more generally. With observations in the continuous white noise model, $dY(t) = f(t)dt + \epsilon\, dW(t)$, $0 \le t \le 1$, the kernel estimator of f is

$$\hat{f}_h(s) = \int_0^1 \mathring{K}_h(s-t)dY(t). \tag{3.21}$$

The integral is interpreted as in (1.23) and it follows from the compact support assumption that $(s,t) \to \mathring{K}_h(s-t)$ is a square integrable kernel, so that $\hat{f}_h(s)$ is periodic, has finite variance, and belongs to $L_2[0,1]$ almost surely–for further details see C.13.

Remark. To help in motivating this estimator, we digress briefly to consider the nonparametric regression model $Y_l = f(t_l) + \sigma Z_l$, for ordered t_l in $[0,1]$ and $l = 1,\dots,n$. A locally weighted average about s would estimate $f(s)$ via

$$\hat{f}(s) = \sum_l w_l(s)Y_l \Big/ \sum_l w_l(s). \tag{3.22}$$

A typical choice of weights might use a kernel $K(u)$ and set $w_l(s) = K_h(s-t_l)$. This is sometimes called the Nadaraya-Watson estimator, see the Chapter Notes for references. A difficulty with (3.22) is that it 'runs out' of data on one side of s when s is near the boundaries 0 and 1. Since we assume that f is periodic, we may handle this by extending the data periodically:

$$Y_{l+jn} = Y_l \qquad \text{for } j \in \mathbb{Z}. \tag{3.23}$$

We also simplify by supposing that the design points t_l are equally spaced, $t_l = l/n$. Then we may make an integral approximation to the denominator in (3.22),

$$\sum_l w_l(s) = \sum_l K_h(s-l/n) \doteq n\int K_h(s-t)dt = n.$$

This leads to the Priestley-Chao form

$$\hat{f}_{h,n}(s) = n^{-1} \sum_{l \in \mathbb{Z}} K_h(s - t_l) Y_l. \tag{3.24}$$

Now use the assumed periodicity (3.23) to rewrite the right hand sum as

$$\sum_{l=1}^{n} \sum_{j \in \mathbb{Z}} K_h(s - j - l/n) Y_l = \sum_{l=1}^{n} \mathring{K}_h(s - l/n) Y_l.$$

To link this with the continuous form (3.21), we recall the partial sum process approximation $Y_n(t)$ from (1.26), and view Y_l as the scaled increment given by $n(\Delta_{1/n} Y_n)(l/n) = n[Y_n(l/n) - Y_n((l-1)/n)]$. We obtain

$$\hat{f}_{h,n}(s) = \sum_{1}^{n} \mathring{K}_h(s - l/n)\, (\Delta_{1/n} Y_n)(l/n).$$

Now approximate $Y_n(t)$ by the limiting process $Y(t)$ and the sum by an integral to arrive at formula (3.21) for $\hat{f}_h(s)$.

Now we return to the white noise model and derive the bias and variance properties of $\hat{f}_h$.

Lemma 3.5 *In the continuous white noise model on* $[0,1]$*, assume that* f *is periodic. Let* $\hat{f}_h$*,* (3.21)*, denote convolution with the periodized kernel* (3.20)*. Then*

$$E\, \hat{f}_h(s) = \int_{-\infty}^{\infty} K_h(s - t) f(t) dt = (K_h f)(s).$$

If also

$$supp(K) \subset [-t_0, t_0], \quad \textit{and} \quad h < 1/(2t_0), \tag{3.25}$$

then

$$\mathrm{Var}\, \hat{f}_h(s) = \epsilon^2 \int_{-\infty}^{\infty} K_h^2(t) dt = \frac{\epsilon^2}{h} \|K\|_2^2.$$

From the first formula, one sees that $\hat{f}_h$ estimates a smoothed version of f given by convolution with the kernel of bandwidth h. The smaller the value of h, the more narrowly peaked is the kernel K_h and so the local average of f more closely approximates $f(s)$. One calls K_h an "approximate delta-function". Thus as h decreases so does the bias $E\, \hat{f}_h(s) - f(s)$, but inevitably at the same time the variance increases, at order $O(1/h)$,

Here we use operator notation $K_h f$ for the convolution $K_h \star f$ over $\mathbb{R}$, consistent with later use in the book, compare e.g. Section 3.9 and (C.6).

Proof The first display is obtained by observing that

$$E\, \hat{f}_h(s) = \int_0^1 \mathring{K}_h(s - t) f(t) dt = \int_{-\infty}^{\infty} K_h(s - u) f(u) du.$$

For the second, we use formula (1.23) and the Wiener integral identity (C.19) to write

$$\mathrm{Var}\, \hat{f}_h(s) = \epsilon^2 \int_0^1 \mathring{K}_h^2(t) dt.$$

When $h < 1/(2t_0)$ and $j \neq j'$, the supports of $u \to K_h(u - j)$ and $K_h(u - j')$ do not overlap, and so

$$\mathring{K}_h^2(u) = \Big[\sum_j K_h(u + j)\Big]^2 = \sum_j K_h^2(u + j).$$

This yields the first equality for $\mathrm{Var}\,\hat{f}_h(s)$ and the second follows by rescaling. □

Local MSE. The mean squared error of $\hat{f}_h$ as an estimator of f *at the point* s has a decomposition into variance and squared bias terms in the same manner as (2.46):

$$E[\hat{f}_h(s) - f(s)]^2 = \mathrm{Var}\,\hat{f}_h(s) + [E\,\hat{f}_h(s) - f(s)]^2.$$

Using the assumptions and conclusions of the previous lemma, we have

$$E[\hat{f}_h(s) - f(s)]^2 = \frac{\epsilon^2}{h}\|K\|_2^2 + (K_h f - f)^2(s). \tag{3.26}$$

Global MSE. The global, or integrated, mean squared error of a kernel estimator is defined by $E\int_0^1[\hat{f}_h(s) - f(s)]^2 ds$. It may easily be evaluated by integrating (3.26) over $s \in [0, 1]$ (using Fubini's theorem). To summarize results, we distinguish between L_p norms $\|g\|_{p,I}$ on the observation interval $I = [0, 1]$, and on all of $\mathbb{R}$, namely $\|g\|_p$. Then, under the assumptions of Lemma 3.5 and (3.25),

$$E\|\hat{f}_h - f\|_{2,I}^2 = \frac{\epsilon^2}{h}\|K\|_2^2 + \|(I - K_h)f\|_{2,I}^2. \tag{3.27}$$

Again this is a decomposition into variance and bias terms. The result holds even without (3.25) if we replace K by $\mathring{K}$ on the right side.

Notice the similarity of this mean squared error expression to (2.47) for a linear estimator in the sequence model. This is no surprise, given the sequence form of $\hat{f}_h$ to be described later in this section.

$q-$th order kernels and bias reduction. A kernel is said to be of *$q-$th order* if it has vanishing moments of order 1 through $q - 1$:

$$\mu_p = \int_{-\infty}^{\infty} v^p K(v)dv = \begin{cases} 1 & p = 0 \\ 0 & p = 1, \ldots, q-1 \\ q!c_q \neq 0 & p = q. \end{cases} \tag{3.28}$$

Observe that if K is symmetric about zero then necessarily $q \geq 2$. However, if K is symmetric and *non-negative*, then $c_2 > 0$ and so $q = 2$. We will see shortly that to obtain fast rates of convergence, kernels of order $q > 2$ are required. It follows that such kernels must necessarily have 'negative sidelobes'.

To see the bias reduction afforded by a $q-$th order kernel, assume that f has q continuous derivatives on $[0, 1]$. Then the Taylor series approximation to f at s takes the form

$$f(s - hv) = f(s) + \sum_{p=1}^{q-1} \frac{(-hv)^p}{p!} f^{(p)}(s) + \frac{(-hv)^q}{q!} f^{(q)}(s(v)),$$

for suitable $s(v)$ between $s - hv$ and s. The bias of $\hat{f}_h$ at s becomes

$$K_h f(s) - f(s) = \int K(v)[f(s-hv) - f(s)]dv = \frac{(-h)^q}{q!} \int v^q K(v) f^{(q)}(s(v))dv \quad (3.29)$$

after using the vanishing moments (3.28).

As a result, the maximal bias of a $q-$th order kernel is uniformly $O(h^q)$:

$$\| K_h f - f \|_{\infty,I} = \sup_{0 \le s \le 1} |K_h f(s) - f(s)| \le c_q h^q \| f^{(q)} \|_{\infty,I}. \quad (3.30)$$

Thus, other things being equal, (which they may not be, see Section 6.5), higher order kernels might seem preferable due to their bias reduction properties for smooth functions. Exercise 3.8 has an example of an infinite order kernel. We will see this type of argument in studying the role of vanishing moments for wavelets in Chapter 7.

In summary, if K is a $q-$th order kernel, if (3.25) holds, and if f is C^q and periodic on $[0, 1]$, then as $h \to 0$ we have the local and global MSE approximations

$$\begin{aligned} E[\hat{f}_h(s) - f(s)]^2 &= \frac{\epsilon^2}{h} \|K\|_2^2 + c_q^2 h^{2q} [D^q f(s)]^2 [1 + o(1)] \\ E\| \hat{f}_h - f \|_{2,I}^2 &= \frac{\epsilon^2}{h} \|K\|_2^2 + c_q^2 h^{2q} \| D^q f \|_{2,I} [1 + o(1)]. \end{aligned} \quad (3.31)$$

The Variance-Bias Lemma. The approximate MSE expressions just obtained have a characteristic form, with a variance term decreasing in h balanced by a bias term that grows with h. The calculation to find the minimizing value of h occurs quite frequently, so we record it here once and for all.

Lemma 3.6 (Variance-Bias) *The function $G(h) = vh^{-1} + bh^{2\rho}$, defined for $h \ge 0$ and positive constants v, b and ρ, has minimizing value and location given by*

$$G(h_*) = e^{H(r)} b^{1-r} v^r \qquad \text{and} \qquad h_* = r^{-1} e^{-H(r)} (v/b)^{1-r}.$$

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

For example, with kernel estimates based on a kernel K of order q, (3.31), shows that h can be thought of as a bandwidth and v as a variance factor (such as n^{-1} or ϵ^2), while b is a bias factor (for example involving $c(K, q) \int (D^q f)^2$, with $\rho = q$).)

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 $(2\rho)^{-1}$ at the optimum, so that in the typical case $\rho > \frac{1}{2}$, the bias contribution is the smaller of the two at the optimum h_*.

Sequence space form of kernel estimators. Our kernels have a translation invariant form, $K_h(s, t) = K_h(s - t)$, and so in the Fourier basis the corresponding estimators should correspond to diagonal shrinkage introduced in the last section and to be analyzed further in later chapters. To describe this, let $\varphi_k(s)$ denote the trigonometric basis (3.8), and recall that

the correspondence between the continuous model (1.21) and sequence form (3.1) is given by formulas (1.24) for y_k, θ_k etc. Thus we have

$$\hat{f}_h(t) = \sum_{k=1}^{\infty} \hat{\theta}_{h,k} \varphi_k(t). \tag{3.32}$$

To describe the coefficients $\hat{\theta}_{h,k}$, define the Fourier transform of an integrable kernel by

$$\widehat{K}(\xi) = \int_{-\infty}^{\infty} K(s) e^{-i\xi s} ds.$$

Some of its properties are recalled in Appendix C.9.

Lemma 3.7 *Assume that the kernel $K(s)$ is symmetric and has Fourier transform $\widehat{K}(\xi)$. Then in the sine-cosine form of the trigonometric basis, the kernel estimator $\hat{f}_h$ has the sequence form given, for $k \geq 1$, by* [3]

$$\hat{\theta}_{h,k} = c_{h,k} y_k \qquad \textit{with} \qquad c_{h,2k-1} = c_{h,2k} = \widehat{K}(2\pi k h). \tag{3.33}$$

Thus the diagonal shrinkage constants in estimator $\hat{\theta}_h$ are given by the Fourier transform of kernel K, which is real-valued since K is symmetric. Their behavior for small bandwidths is determined by that of $\widehat{K}$ near zero. Indeed, the r-th derivative of $\widehat{K}(\xi)$ at zero involves the r-th moment of K, namely $\widehat{K}^{(r)}(0) = (-i)^r \int t^r K(t) dt$. Hence an equivalent description of a $q-$th order kernel, (3.28), says

$$\widehat{K}(\xi) = 1 - b_q \xi^q + o(\xi^q) \qquad \text{as} \quad \xi \to 0 \tag{3.34}$$

where $b_q = -(-i)^q c_q \neq 0$. Typically $b_q > 0$, reflecting the fact that the estimator usually shrinks coefficients toward zero.

For the first three kernels listed at (3.19), we have

$$\widehat{K}(\xi) = \begin{cases} e^{-\xi^2/2} & \text{Gaussian} \\ \sin \xi / \xi & \text{Uniform} \\ (3/\xi^2)(\sin \xi / \xi - \cos \xi) & \text{Quadratic/Epanechnikov.} \end{cases} \tag{3.35}$$

Proof We begin with the orthobasis of complex exponentials $\varphi_k^C(s) = e^{2\pi i k s}$ for $k \in \mathbb{Z}$. The complex Fourier coefficients of the kernel estimator $\hat{f}_h$ are found by substituting the periodized form of (3.21) and interchanging orders of integration:

$$\int_0^1 \hat{f}_h(s) e^{-2\pi i k s} ds = \int_0^1 \mathring{K}_h(s-t) e^{-2\pi i k (s-t)} ds \cdot \int_0^1 e^{-2\pi i k t} dY(t).$$

In other words, we have the diagonal form $\hat{\theta}_{h,k}^C(y) = \gamma_{h,k}^C y_k^C$ for $k \in \mathbb{Z}$. Now using first the periodicity of $\mathring{K}_h$, and then its expression (3.20) in terms of K, we find that

$$\begin{aligned} \gamma_{h,k}^C &= \int_0^1 \mathring{K}_h(u) e^{-2\pi i k u} du = \int_{-\infty}^{\infty} K_h(u) e^{-2\pi i k u} du \\ &= \widehat{K_h}(2\pi k) = \widehat{K}(2\pi k h). \end{aligned} \tag{3.36}$$

[3] The reader should note an unfortunate clash of two established conventions: the hats in $\hat{\theta}$, $\hat{f}$ denoting estimators should not be confused with the wider ones in $\widehat{K}$ denoting Fourier transform!

Observe that since K is symmetric we have $\widehat{K}(-\xi) = \widehat{K}(\xi)$ and so $\gamma^C_{h,-k} = \gamma^C_{h,k}$.

It remains to convert this to the real trigonometric basis. The relation between Fourier coefficients $\{f^C_k, k \in \mathbb{Z}\}$ in the complex exponential basis and real coefficients $\{f_k, k \geq 1\}$ in trigonometric basis (3.8) is given by

$$f_1 = f^C_0, \qquad f_{2k} = (1/i\sqrt{2})(f^C_k - f^C_{-k}), \qquad f_{2k+1} = (1/\sqrt{2})(f^C_k + f^C_{-k}).$$

The diagonal form (3.33) now follows from this and (3.36) since $\gamma^C_{h,-k} = \gamma^C_{h,k}$. For example

$$\begin{aligned}\hat{\theta}_{h,2k+1} &= (\hat{\theta}^C_{h,k} + \hat{\theta}^C_{h,-k})/\sqrt{2} = (\gamma^C_{h,k} y^C_k + \gamma^C_{h,-k} y^C_{-k})/\sqrt{2} \\ &= \widehat{K}(2\pi kh)(y^C_k + y^C_{-k})/\sqrt{2} = \widehat{K}(2\pi kh) y_{2k+1}. \qquad \square\end{aligned}$$

3.4 Periodic spline estimators

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. As seen in Section 1.4, 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 periodic f, 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 and Tibshirani (1990) and Green and Silverman (1994).

Finite equispaced regression model. Suppose therefore that we observe

$$Y_l = f(l/n) + \epsilon Z_l, \qquad Z_l \overset{\text{iid}}{\sim} N(0,1) \tag{3.37}$$

for $l = 1, \ldots, n$, and unknown f assumed periodic on $[0,1]$. Since the observation points are equally spaced, we can use the Fourier basis $\varphi_k(t)$, (3.8). For convenience in notation, we consider only $n = 2n_d + 1$ odd. Let S_n now denote the linear space of trigonometric polynomials of degree n_d: thus $\mathsf{S}_n = \{f : f(t) = \sum_{k=1}^n c_k \varphi_k(t), t \in [0,1]\}$. For $m \in \mathbb{N}$, the mth order periodic smoothing spline will be the minimum $\hat{f}_{\lambda,n}(t)$ in S_n of

$$Q_n(f) = n^{-1} \sum_{l=1}^n [Y_l - f(t_l)]^2 + c_m \lambda \int_0^1 (D^m f)^2.$$

Here we allow a more general mth derivative penalty; the constant c_m is specified below.

The discrete sine and cosine vectors will be $\boldsymbol{\varphi}_k = (\varphi_k(t_l))$. The key point is that for the Fourier basis, the vectors $\boldsymbol{\varphi}_k$ are discrete orthogonal on $\{1/n, \ldots, (n-1)/n, 1\}$ and *at the same time* the functions $\varphi_k(t)$ are continuous orthonormal on $[0,1]$, see Exercise 3.7 and Appendix C.9. Using the properties of differentiation in the Fourier basis, as in the proof of

Lemma 3.3, these double orthogonality relations take the form

$$n^{-1}\sum_{l=1}^{n}\varphi_j(t_l)\varphi_k(t_l) = \delta_{jk} \qquad \text{and} \qquad \int_0^1 D^m\varphi_j \cdot D^m\varphi_k = \pi^{2m}w_k\delta_{jk},$$

where now the weights are explicit:

$$w_1 = 0, \qquad w_{2k} = w_{2k+1} = (2k)^{2m}. \tag{3.38}$$

We now convert the objective $Q_n(f)$ into sequence space form. Let $y_k = n^{-1}\sum_{l=1}^{n} Y_l\varphi_k(t_l)$ be the empirical Fourier coefficients of $\{Y_l\}$, for $k = 1,\dots,n$. Then, using the double orthogonality relations, for any function $f \in \mathsf{S}_n$, we have

$$Q_n(f) = Q_n(\theta) = \sum_{k=1}^{n}(y_k - \theta_k)^2 + \lambda\sum_{k=1}^{n} w_k\theta_k^2.$$

[We have set $c_m = \pi^{-2m}$.] This quadratic polynomial has a unique minimum

$$\hat{\theta}_{\lambda,k} = c_{\lambda,k}y_k = (1 + \lambda w_k)^{-1}y_k, \tag{3.39}$$

for $k = 1,\dots,n$. The corresponding estimator of f,

$$\hat{f}_{\lambda,n}(t) = \sum_{k=1}^{n}\hat{\theta}_{\lambda,k}\varphi_k(t), \tag{3.40}$$

might be called a periodic smoothing spline estimator based on $\{Y_1,\dots,Y_n\}$. The periodic spline problem therefore has many of the qualitative features of general spline smoothing seen in Section 1.4, along with a completely explicit description.

Remark. It is *not* true that the minimizer of $Q(f)$ over all functions lies in S_n, as was the case with cubic splines. The problem lies with *aliasing*: the fact that when $0 < r \le n$ and $l \in \mathbb{N}$, we have $\varphi_r = \varphi_{r+2ln}$ when restricted to $t_1,\dots,t_n$. See Exercise 3.9.

Infinite sequence model. The periodic spline estimate (3.40) in the finite model (3.37) has a natural analogue in the infinite white noise model, which we recall has the dual form

$$\begin{aligned} Y_t &= \int_0^t f(s)ds + \epsilon W_t && t \in [0,1], \\ \Leftrightarrow \qquad y_k &= \theta_k + \epsilon z_k && k \in \mathbb{N}. \end{aligned} \tag{3.41}$$

with the second row representing coefficients in the trigonometric basis (3.8). The smoothing spline estimate $\hat{\theta}_\lambda$ in the infinite sequence model is the minimizer[4] of

$$Q(\theta) = \sum_{1}^{\infty}(y_k - \theta_k)^2 + \lambda\sum_{1}^{\infty} w_k\theta_k^2. \tag{3.42}$$

We again use weights w_k given by (3.38) corresponding to the mth order penalty $P(f) =$

[4] In the Gaussian white noise model, $\sum_1^\infty (y_k - \theta_k)^2 = \infty$ with probability one, but this apparent obstacle may be evaded by minimizing the equivalent criterion $\tilde{Q}(\theta) = Q(\theta) - \sum y_k^2$.

$\int (D^m f)^2$. The estimator $\hat{\theta}_\lambda = (\hat{\theta}_{\lambda,k})$ minimizing $Q(\theta)$ has components again given by (3.39), only now the index $k \in \mathbb{N}$. The corresponding estimate of f is

$$\hat{f}_\lambda(t) = \sum_1^\infty \hat{\theta}_{\lambda,k}\varphi_k(t) = \sum_1^\infty c_{\lambda,k}\, y_k \varphi_k(t) \tag{3.43}$$

Studying $\hat{f}_\lambda$ in the infinite model rather than $\hat{f}_{\lambda,n}$ in the finite model amounts to ignoring discretization, which does not have a major effect on the principal results (we return to this point in Chapter 15).

We may interpret the m-th order smoothing spline as a Bayes estimator. Indeed, if the prior makes the co-ordinates θ_k independently $N(0, \tau_k^2)$ with $\tau_k^2 = b/w_k$, then the posterior mean, according to (2.19), is linear with shrinkage factor

$$c_{2k} = c_{2k+1} = \frac{b(2k)^{-2m}}{b(2k)^{-2m} + \epsilon^2} = \frac{1}{1 + \lambda(2k)^{2m}},$$

after adopting the calibration $\lambda = \epsilon^2/b$. Section 3.10 interprets this prior in terms of $(m-1)$-fold integrated Brownian motion.

Some questions we aim to address using $\hat{\theta}_\lambda$ include

(a) what is the MSE of $\hat{\theta}_\lambda$, or rather the worst case MSE of $\hat{\theta}_\lambda$ over mean square smoothness classes such as $\Theta_2^\alpha(C)$?

(b) what is the best (i.e. minimax) choice of regularization parameter λ, and how does it and the resulting minimax MSE depend on α, C and ϵ?

After a digression that relates spline and kernel estimators, we take up these questions in Section 3.6.

3.5 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 (3.37) and the infinite sequence model (3.41).

	Finite regression model (3.37)	Infinite sequence model (3.41)
Kernel	$\hat{f}_{h,n}(s) = n^{-1}\sum_{l=1}^n \mathring{K}_h(s-t_l)Y_l$	$\hat{f}_h(s) = \int_0^1 \mathring{K}_h(s-t)dY(t)$
		$\Updownarrow$
Spline	$\hat{f}_{\lambda,n}(s) = \sum_{k=1}^n c_{\lambda,k}\, y_k\, \varphi_k(s)$	$\hat{f}_\lambda(s) = \sum_{k=1}^\infty c_{\lambda,k}\, y_k\, \varphi_k(s)$

Table 3.1 *The analogy between spline smoothing and regression goes via versions of each method in the infinite sequence model.*

As we have just seen, in terms of functions, the spline estimate is given by the series in the lower row of Table 3.1, with shrinkage constants $c_{\lambda,k}$ given by (3.39).

We can now derive the kernel representation of the infinite sequence spline estimate.

Substituting (1.24), $y_k = \int \varphi_k dY$ into $\hat{f}_\lambda = \sum_k c_{\lambda,k} y_k \varphi_k$, we get

$$\hat{f}_\lambda(s) = \int_0^1 C_\lambda(s,t) dY(t), \qquad C_\lambda(s,t) = \sum_1^\infty c_{\lambda,k} \varphi_k(s) \varphi_k(t). \tag{3.44}$$

Now specialize to the explicit weights for periodic splines in (3.38). Then $c_{\lambda,2k} = c_{\lambda,2k+1}$, and from (3.8) and the addition formula for sines and cosines,

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

Hence the kernel $C_\lambda(s,t)$ has translation form $C_\lambda(s-t)$, and from (3.39) has formula

$$C_\lambda(u) = 1 + \sum_{k=1}^\infty \frac{2\cos 2\pi k u}{1 + \lambda(2k)^{2m}}.$$

But we can describe C_λ more explicitly! First, recall from the previous section that a function f on $\mathbb{R}$ can be made periodic with period 1 by periodizing: $\overset{\circ}{f}(t) = \sum_{j\in\mathbb{Z}} f(t+j)$.

Theorem 3.8 *The periodic smoothing spline is the Fourier series of a kernel estimator:*

$$\hat{f}_\lambda(s) = \int_0^1 C_\lambda(s-t) dY(t).$$

With $\lambda = (\pi h)^{2m}$, kernel $C_\lambda(u) = \overset{\circ}{K}_h(u)$ is the periodized version of $K_h(u) = (1/h)K(u/h)$. The equivalent kernel is given for $m = 1$ by $K(u) = (1/2)e^{-|u|}$ and for $m = 2$ by

$$K(u) = \frac{1}{2} e^{-|u|/\sqrt{2}} \sin\Big(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\Big). \tag{3.45}$$

For general m, K is a $(2m)-$th order kernel, described at (3.48) *below.*

The kernel K_h has exponential decay, and is essentially negligible for $|u| \geq 8h$ for $m = 1, 2$ and for $|u| \geq 10h$ for $m = 3, 4$–compare Figure 3.1. The wrapped kernel $\overset{\circ}{K}_h$ is therefore effectively identical with K_h on $[-\frac{1}{2}, \frac{1}{2}]$ when h is small: for example $h < 1/16$ or $h < 1/20$ respectively will do.

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.

Proof One approach is to use the Poisson summation formula, recalled at (C.13) and (C.14). Thus, inserting $\lambda = (\pi h)^{2m}$ we may rewrite C_λ as

$$C_\lambda(u) = \sum_{k\in\mathbb{Z}} \frac{e^{2\pi i k u}}{1 + (2\pi k h)^{2m}} = \sum_j K_h(u+j) \tag{3.46}$$

where the second equality uses the Poisson formula (C.14). Since equality holds for all u, we read off that K_h has Fourier transform

$$\widehat{K_h}(\xi) = (1 + h^{2m}\xi^{2m})^{-1}. \tag{3.47}$$

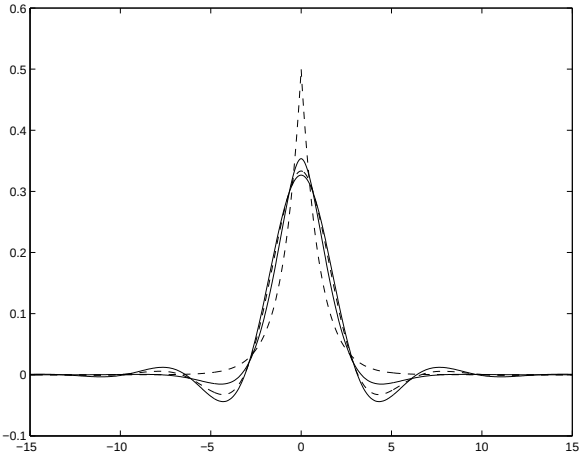

Figure 3.1 Equivalent kernels for spline smoothing: dashed lines show $m = 1, 3$ and solid lines $m = 2, 4$. Only $m = 1$ is non-negative, the "side lobes" are more pronounced for increasing m.

Corresponding to the rescaling $K_h(u) = (1/h)K(u/h)$, we have $\widehat{K_h}(\xi) = \widehat{K}(h\xi)$. From $\widehat{K}(\xi) = (1 + \xi^2)^{-1}$ one can verify that $K(u) = e^{-|u|}/2$, and a little more work yields the $m = 2$ result. More generally, from Erdélyi et al. (1954, (Vol. 1), p.10), with $r_k = (2k - 1)\pi/(2m)$,

$$K(u) = (2m)^{-1} \sum_{k=1}^{m} e^{-|u| \sin r_k} \sin(|u| \cos r_k + r_k), \tag{3.48}$$

which reduces to the cited expressions for $m = 1$ and $m = 2$. □

Remark. We mention two other approaches. Exercise 3.10 outlines a direct derivation via contour integration. Alternately, by successively differentiating (3.46), it is easily seen that

$$h^4 K_h^{(4)} + K_h = \sum_l \delta_l, \tag{3.49}$$

where δ_l is the delta function at l. The solution of $h^4 K_h^{(4)} + K_h = \delta_0$ on $\mathbb{R}$ may be found by Fourier transformation, and yields the $m = 2$ case of (3.47), and then this is converted into a solution of (3.49) by periodization.

3.6 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, as in the discussion of truncation estimators in Section 3.2 is to select convenient parameter spaces Θ, to evaluate the worst case MSE of $\hat{\theta}_\lambda$ over Θ, and

then to choose the value of λ that minimizes this maximum error. This yields information on the rate of convergence of $\hat{\theta}_\lambda$ to θ as $\epsilon \to 0$: we shall see that such rates of convergence, although crude tools, already yield useful information about estimators.

Spline estimators for fixed λ. We adopt the infinite sequence model (3.41) and the diagonal linear estimator $\hat{\theta}_\lambda$ of (3.39), with $w_0 = 0$ and $w_{2k} = w_{2k+1} = (2k)^{2m}$ for $k \geq 1$. For $m \in \mathbb{N}$, it is the Fourier sequence form of the mth order spline with roughness penalty $\int (D^m f)^2$, but the sequence form actually makes sense for any $m > 0$. We also specialize to the Sobolev ellipsoid $\check{\Theta}_2^\alpha(C)$ with semi-axes given by (3.9), i.e. $a_{2k} = a_{2k+1} = (2k)^\alpha$, which in the trigonometric basis, corresponds to α-fold differentiability in mean square, Lemma 3.3.

Proposition 3.9 *Suppose that $m > 1/4$. The worst case mean squared error for an m-th order spline estimate $\hat{\theta}_\lambda$ over a Sobolev ellipsoid $\check{\Theta}_2^\alpha(C)$ is, for small λ, approximately*

$$\bar{r}(\hat{\theta}_\lambda, \epsilon) = \sup_{\theta \in \check{\Theta}_2^\alpha(C)} r(\hat{\theta}_{\lambda_*}, \theta; \epsilon) \doteq v_m \epsilon^2 \lambda^{-1/2m} + b_{\alpha m} C^2 \lambda^{2 \wedge (\alpha/m)}. \tag{3.50}$$

The worse case configuration is, approximately[5], given by $\theta^* = C k_*^{-\alpha} e_{k_*}$, where

$$k_* = \begin{cases} \big((2m-\alpha)/\alpha\big)^{1/(2m)} \lambda^{-1/(2m)} & \text{if } \alpha < 2m, \\ 2 & \text{if } \alpha \geq 2m. \end{cases}$$

Values for the constants v_m and $b_{\alpha m}$ appear in the proof. Remarks on the approximate equality symbol $\doteq$ are given below. The relative error in each term in (3.50) is $O(\lambda^{1/m})$.

Proof As seen in Lemma 3.4, the maximum risk of a diagonal linear estimator $\hat{\theta}_c$ over an ellipsoid is given by

$$\bar{r}(\hat{\theta}_c; \epsilon) = \epsilon^2 \sum_{k=1}^{\infty} c_k^2 + C^2 \sup_{k \geq 1} a_k^{-2} (1 - c_k)^2 = \bar{V}(\epsilon) + \bar{B}^2,$$

say. In order to evaluate these maximum variance and bias terms for the spline choice $c_k = (1 + \lambda w_k)^{-1}$, we pause for some remarks on approximating the sums and minima involved.

Aside on discretization approximations. Often a simpler expression results by replacing a sum by its (Riemann) integral approximation, or by replacing a minimization over non-negative integers by an optimization over a continuous variable in $[0, \infty)$. We use the special notation $\doteq$ to denote the approximate inequality in such cases. For the sum

$$S(\lambda) = \sum_{k=0}^{\infty} k^p (1 + \lambda k^q)^{-r} \doteq \kappa \lambda^{-\mu}, \qquad \mu = (p+1)/q, \tag{3.51}$$

with p, q and $r \geq 0$ and convergence if and only if $qr > p + 1$, and

$$\kappa = \kappa(p, r; q) = \int_0^\infty v^p (1 + v^q)^{-r} dv = \Gamma(r - \mu)\Gamma(\mu)/(q\Gamma(r)). \tag{3.52}$$

The approximation becomes an equality as $\lambda \to 0$, so that $S(\lambda)/\kappa\lambda^{-\mu} \to 1$.

[5] since k_* should be replaced by an even integer, being whichever of the two values closest to k_* leads to the larger squared bias.

For minimization, we observe that, if $0 < \alpha < \gamma$,

$$\bar{S}(\lambda) = \min_{k\in\mathbb{N}} \lambda k^{\alpha} + k^{\alpha-\gamma} \doteq \inf_{x>0} \lambda x^{\alpha} + x^{\alpha-\gamma} = \bar{\kappa}\lambda^{1-\bar{\mu}}, \tag{3.53}$$

with $\bar{\kappa} = e^{H(\alpha/\gamma)}$ and $\bar{\mu} = \alpha/\gamma$. The final equality uses the Variance-Bias Lemma 3.6 with $v = \lambda, h = x^{-\alpha}$, etc. Again we have asymptotic equality, $\bar{S}(\lambda)/\bar{\kappa}\lambda^{1-\bar{\mu}} \to 1$ as $\lambda \to 0$.

The errors in these discretization approximations[6] are quadratic in the size of the discretization step, and so can be expected often to be fairly small. Briefly, for the integral approximation, if G has, for example, $G(0) = 0$ and $\int_0^\infty |G''| < \infty$, then the difference between $\sum_{k=0}^{\infty} G(k\delta)\delta$ and $\int_0^\infty G(x)dx$ is $O(\delta^2)$, as follows from the standard error analysis for the trapezoid rule. Similarly, if G is C^2, then the difference between $\min_{k\in\mathbb{N}} G(k\delta)$ and $\inf_{x>0} G(x)$ is $O(\delta^2)$, as follows from the usual Taylor expansion bounds.

Finally, for later use we record, for $p \geq 0$ and $a > 1$, that

$$\sum_{k=1}^{K}(2[k/2])^p \doteq \sum_{k=1}^{K} k^p \doteq (p+1)^{-1}K^{p+1}, \quad \text{and} \quad \sum_{k=1}^{K} a^k \doteq \gamma_a a^K \tag{3.54}$$

which means for the first sum, that the relative error in the integral approximation is $O(K^{-2})$. In the second sum, $\gamma_a = a/(a-1)$, and the relative error is geometrically small, $O(a^{-K})$.

Continuation of proof of Proposition 3.9. For the variance term, use the integral approximation (3.51) with $p = 0, q = 2m$ and $r = 2$:

$$\epsilon^{-2}\bar{V}(\epsilon) = 1 + 2\sum_{k=1}^{\infty}[1 + \lambda(2k)^{2m}]^{-2} \doteq \sum_{k=0}^{\infty}(1 + \lambda k^{2m})^{-2} \doteq v_m \lambda^{-1/2m},$$

so long as $m > 1/4$. To evaluate v_m, use (3.52) with $r = 2$ and $\mu = \mu_m = 1/(2m)$,

$$v_m = \mu_m \Gamma(2-\mu_m)\Gamma(\mu_m) = (1-\mu_m)/\mathrm{sinc}(\mu_m), \tag{3.55}$$

using Euler's reflection formula $\Gamma(1-\mu)\Gamma(\mu) = \pi/\sin(\pi\mu)$, and the normalized sinc function $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$. In the case $m = 2$ (cubic splines), $v_2 = 3\sqrt{2}\pi/16$.

For the squared bias term, note first that $\bar{B}^2 = C^2 \sup_{k\in 2\mathbb{N}} a_k^{-2}(1-c_k)^2$ since in the sine-cosine basis $a_{2k} = a_{2k+1}$, $c_{2k} = c_{2k+1}$, and $c_1 = 1$. Now for $k \in 2\mathbb{N}$, we have $a_k^{-1} = k^{-\alpha}$ and $1 - c_k = [1 + \lambda^{-1}k^{-2m}]^{-1}$, so that (3.16) becomes

$$\bar{B} = C\lambda\{\min_{k\in 2\mathbb{N}} \lambda k^{\alpha} + k^{\alpha-2m}\}^{-1}.$$

If $\alpha \geq 2m$, then the minimum in $\bar{B}$ occurs at $k_* = 2$, with $\bar{B}^2 \sim b_{\alpha m}C^2\lambda^2$ for small λ, and $b_{\alpha m} = 2^{2(2m-\alpha)}$. If $\alpha \leq 2m$, then by differentiation, the minimum in $\bar{B}$ occurs at the claimed value of k_* and to evaluate the minimum value, apply (3.53) with $\gamma = 2m$ and $\bar{\mu} = \alpha/(2m)$ to obtain

$$\bar{B}^2 \doteq C^2\lambda^2(\bar{\kappa}\lambda^{1-\bar{\mu}})^{-2} = b_{\alpha m}C^2\lambda^{\alpha/m},$$

with $b_{\alpha m} = e^{-2H(\alpha/2m)}$. Evaluating this, we can summarize the two cases via

$$b_{\alpha m} = \begin{cases} (2m)^{-2}\alpha^{\alpha/m}(2m-\alpha)^{2-\alpha/m} & \alpha \leq 2m \\ 2^{2(2m-\alpha)} & \alpha > 2m. \end{cases} \tag{3.56}$$

[6] Actually, we work in the reverse direction, from discrete to continuous!

Note that $b_{\alpha m} = 1$ if $\alpha = 2m$. Combining the variance and bias terms yields (3.50). □

Remarks. 1. The exponent of λ in the bias term shows that high smoothness, namely $\alpha \geq 2m$, has no effect on the worst-case mean squared error.

2. The 'degrees of freedom' of the smoother $\hat{\theta}_\lambda = S_\lambda y$ is approximately

$$\operatorname{tr} S_\lambda = \sum_k c_k = 1 + \sum_{k \geq 2} (1 + \lambda(2[k/2])^{2m})^{-1} \doteq c\lambda^{-1/(2m)}.$$

In the equivalent kernel of the Section 3.5, we saw that λ corresponded to h^{2m}, and so the degrees of freedom $\operatorname{tr} S_\lambda$ is approximately proportional to h^{-1}. In addition, if $\alpha \leq 2m$, $\operatorname{tr} S_\lambda$ is also proportional to the least favorable frequency k_*.

3. The same proof works for $\Theta_2^\alpha(C)$ in place of $\breve{\Theta}_2^\alpha(C)$, using $k \in \mathbb{N}$ in place of $k \in 2\mathbb{N}$. Going forward, we state results for $\Theta_2^\alpha(C)$.

Minimax λ for the spline estimator. Our interest now turns to the value of λ that minimizes the maximum risk (3.50). This is called the *minimax* λ for the parameter space Θ.

Theorem 3.10 *Consider an m-th order periodic spline estimator $\hat{\theta}_\lambda$ as in* (3.39)*, for $m \in \mathbb{N}$ and $\lambda > 0$, and its maximum risk over a Sobolev ellipsoid $\Theta_2^\alpha(C), \alpha > 0$. Let $r(\alpha) = 2\alpha/(2\alpha + 1)$, and set $r = r(\alpha \wedge 2m)$. Then the minimax λ_* leads to*

$$\bar{r}(\hat{\theta}_{\lambda_*}; \epsilon) = \sup_{\theta \in \Theta_2^\alpha(C)} r(\hat{\theta}_{\lambda_*}, \theta; \epsilon) = c_1(\alpha, m) C^{2(1-r)} \epsilon^{2r} + O(\epsilon^2), \tag{3.57}$$

as $\epsilon \to 0$, with

$$\lambda_* \sim c_2(\alpha, m)(\epsilon^2/C^2)^{2m(1-r)}.$$

Proof Formula (3.50) shows that there is a variance-bias tradeoff, with small λ corresponding to small 'bandwidth' h and hence high variance and low bias, with the converse being true for large λ. To find the optimal λ, apply the Variance-Bias Lemma 3.6 with the substitutions

$$h = \lambda^{1/(2m)}, \qquad v = v_m \epsilon^2, \qquad b = b_{\alpha m} C^2, \qquad \rho = \alpha \wedge 2m,$$

where v_m and $b_{\alpha m}$ are the variance and bias constants (3.55) and (3.56) respectively.

Again from the Variance-Bias lemma, one can also identify the constants explicitly:

$$\begin{aligned} c_1(\alpha, m) &= e^{H(r)} b_{\alpha m}^{1-r} v_m^r, \\ c_2(\alpha, m) &= \left(v_m/(2\rho b_{\alpha m})\right)^{2m/(2\rho+1)}. \end{aligned}$$

The $O(\epsilon^2)$ error term follows from the remarks on discretization error. □

Remarks. 1. The rate of convergence $r = r(\alpha \wedge 2m)$ increases with α until $\alpha = 2m$, but does not improve further for functions with smoothness greater than α. We say that the rate *saturates* at $2m$, or that $r(2m)$ is a "speed limit" for m-th order splines.

2. In particular, for the typical choice $m = 2$, the rate of convergence saturates at speed limit $r(4) = 8/9$. For this rate functions of smoothness of order $\alpha = 4$ are required, and as seen in the last section, the kernel for $m = 2$ will have negative sidelobes. However, if one uses a *non-negative* kernel, the 'generic' rate of convergence for a kernel estimator (at the optimal h) is $n^{-4/5} \asymp (\epsilon^2)^{4/5}$, at least for less smooth f with 2 continuous derivatives.

3. We will see in Chapter 5 that $r(\alpha)$ is the best possible rate of convergence, in the minimax sense, over $\Theta_2^\alpha(C)$. Thus m-th order splines can attain the optimal rate for all smoothness indices $\alpha \le 2m$.

An important point is that the optimal choice of λ_* needed to achieve this optimal rate depends on (C, α) (as well as on m and ϵ^2). These values are unlikely to be known in practice, so the problem of *adaptation* consists, in this case, in finding estimators that achieve the optimal rate *without* having to specify values for C and α. Chapter 6 has more on this.

4. If $\alpha = m$, then $b_{\alpha m} = 1/4$, which leads to the useful special case

$$\bar{r}(\hat{\theta}_{\lambda_*}; \epsilon) \sim e^{H(r)} (C^2/4)^{1-r} (v_m \epsilon^2)^r. \tag{3.58}$$

In particular, for cubic splines over ellipsoids of twice differentiable functions in mean square, we get that $\lambda_* \sim (v_2\epsilon^2/C^2)^{4/5}$. For a fixed function f, recall from (3.10) 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 λ is desired, one might set $C^2 = \pi^{-4} \int f''^2$ to arrive at the proposal

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

5. We may compare the minimax-λ MSE for splines given in (3.57) with the minimax-ν MSE for truncation estimators (3.18). By comparing the constant terms, it can be verified for $m > 1/2$ and $0 \le \alpha \le 2m$ that the spline estimators have asymptotically at least as good MSE performance as the truncation estimator, Exercise 3.12. We will see in Section 5.2 just how close to optimal the spline families actually come.

6. We have assumed periodic f, equispaced sampling points t_l and Gaussian errors to allow a concrete analysis. All of these assumptions can be relaxed, so long as the design points t_l are reasonably regularly spaced. A selection of relevant references includes Cox (1983); Speckman (1985); Cox (1988); Carter et al. (1992).

3.7 Back to kernel-type estimators

For periodic smoothing splines, we have just found the rate of convergence for the optimal (minimax) choice of regularization parameter λ, at least for smoothness classes up to a speed limit determined by the order of the penalty. Broadly similar results hold for two other popular classes of linear estimators defined by a bandwidth or tuning parameter. The proofs follow the pattern just developed, so we describe the results only briefly, leaving some details to Exercises 3.13 and 3.14.

Kernel estimators. Suppose that K is a symmetric kernel of compact support and order q. Let kernel estimator $\hat{f}_h$ be given by (3.21) and let $\hat{\theta}_h$ denote its sequence form provided in Lemma 3.7. The order q provides a speed limit: so long as $\alpha \le q$, then a result entirely analogous to the spline Theorem 3.10 holds:

$$\inf_h \sup_{\theta \in \Theta_2^\alpha(C)} r(\hat{\theta}_h, \theta; \epsilon) \sim c_{\alpha,K} C^{2(1-r)} \epsilon^{2r},$$

where the exact form of $c_{\alpha,K}$ is given in Exercise 3.13: it has a structure entirely analogous to the periodic spline case.

Local polynomial regression. Consider the finite equispaced regression model (3.37) for periodic f, with data extended periodically as in (3.23). Let K be a kernel of compact support and let $\hat{\beta}$ minimize

$$\sum_{l\in\mathbb{Z}}\Big(Y_l - \sum_{j=0}^{p}\beta_j(t_l-t)^j\Big)^2 K_h(t_l-t). \tag{3.59}$$

Then the *local polynomial estimator* of degree p puts $\hat{f}_{p,h}(t) = \hat{\beta}_0$. This may be motivated by the Taylor expansion of $f(s)$ about $s = t$, in which $\beta_j = f^{(j)}(t)/j!$. The kernel K_h serves to localize the estimation in $\hat{f}_{p,h}(t)$ to data within a distance of order h of t.

It can be shown, Exercise 3.14, that the local polynomial regression estimator has an equivalent kernel estimator form

$$\hat{f}_{p,h}(t) = n^{-1}\sum_{l\in\mathbb{Z}} K_h^*(t_l-t)Y_l,$$

compare (3.24), where the equivalent kernel K^* is a kernel of order (at least) $p+1$, even if the "starting" kernel K has order 2. Consequently, the rates of convergence results described for higher order kernel estimators also apply to local polynomial regression. A comprehensive discussion of local polynomial regression is given by Fan and Gijbels (1996).

3.8 Non-white Gaussian sequence models

So far in this chapter, we have focused on the white infinite sequence model (3.1) and its cousins. Many of the methods of this book extend to a 'non-white' sequence model

$$y_i = \theta_i + \epsilon\varrho_i z_i, \qquad i \in \mathbb{N}, \tag{3.60}$$

where the z_i are again i.i.d. $N(0,1)$, but the ϱ_i are known positive constants.

In the next two sections, we explore two large classes of Gaussian models which can be transformed into (3.60). These two classes parallel those discussed for the finite model in Section 2.9. The first, linear inverse problems, studies models of the form $Y = Af + \epsilon Z$, where A is a linear operator, and the singular value decomposition (SVD) of A is needed to put the model into sequence form. The second, correlated data, considers models of the form $Y = f + \epsilon Z$, where Z is a correlated Gaussian process. In this setting, it is the Karhunen-Loève transform (KLT, also called principal component analysis) that puts matters into sequence form (3.60). The next two sections develop the SVD and KLT respectively, along with certain canonical examples that illustrate the range of possibilities for (ϱ_i).

First, some preliminaries about model (3.60). For example, when is it well defined? We pause to recall the elegant Kakutani dichotomy for product measures (e.g. Williams (1991, Ch. 14), Durrett (2010, Ch. 5)). Let P and Q be probability measures on a measurable space $(\mathcal{X},\mathcal{B})$, absolutely continuous with respect to a probability measure ν. (For example, $\nu = (P+Q)/2$.) Write $p = dP/d\nu$ and $q = dQ/d\nu$. The *Hellinger affinity*

$$\rho(P,Q) = \int\sqrt{pq}\,d\nu \tag{3.61}$$

does not depend on the choice of ν. Now let $\{P_i\}$ and $\{Q_i\}$ 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 $P = \prod P_i$ and $Q = \prod Q_i$. The affinity behaves well for products: $\rho(P, Q) = \prod \rho(P_i, Q_i)$.

Kakutani's dichotomy says that if the components $P_i \sim Q_i$ for $i = 1, 2, \ldots$ then the products P and Q are either equivalent or orthogonal. And there is an explicit criterion:

$$P \sim Q \qquad \text{if and only if} \qquad \prod_{i=1}^{\infty} \rho(P_i, Q_i) > 0.$$

In either case $L_\infty = \lim \prod_1^n dP_i/dQ_i$ exists Q-a.s. And when $P \sim Q$, the likelihood ratio dP/dQ is given by the product $L_\infty = \prod_{i=1}^\infty dP_i/dQ_i$, whereas if P is orthogonal to Q, then $Q(L_\infty = 0) = 1$.

The criterion is easy to apply for Gaussian sequence measures. A little calculation shows that the univariate affinity

$$\rho(N(\theta, \sigma^2), N(\theta', \sigma^2)) = \exp\{-(\theta - \theta')^2/(8\sigma^2)\}.$$

Let P_θ denote the product measure corresponding to (3.60). The dichotomy says that for two different mean vectors θ and θ', the measures P_θ and $P_{\theta'}$ are equivalent or orthogonal. [See Exercise 3.16 for an implication for statistical classification]. The product affinity

$$\rho(P_\theta, P_{\theta'}) = \exp\{-D^2/8\}, \qquad D^2 = \sum_i (\theta_i - \theta_i')^2/(\varrho_i \epsilon)^2. \tag{3.62}$$

Thus P_θ is absolutely continuous relative to P_0 if and only if $\sum \theta_i^2/\varrho_i^2 < \infty$, in which case the density is given in terms of the inner product $\langle y, \theta\rangle_\varrho = \sum_i y_i\theta_i/\varrho_i^2$ by

$$\frac{dP_\theta}{dP_0}(y) = \exp\left\{\frac{\langle y, \theta\rangle_\varrho}{\epsilon^2} - \frac{\|\theta\|_\varrho^2}{2\epsilon^2}\right\}.$$

Here $\theta_i/(\varrho_i\epsilon)$ might be interpreted as the signal-to-noise ratio of the i-th co-ordinate.

We will again be interested in evaluating the quality of estimation of θ that is possible in model (3.60). An important question raised by the extended sequence model is the effect of the constants (ϱ_i) on quality of estimation—if ϱ_i increases with i, we might expect, for example, a decreased rate of convergence as $\epsilon \to 0$.

We will also be interested in the comparison of linear and non-linear estimators in model (3.60). For now, let us record the natural extension of formula (2.47) for the mean squared error of a linear estimator $\hat\theta_C(y) = Cy$. Let $R = \mathrm{diag}(\varrho_i^2)$, then

$$r(\hat\theta_C, \theta) = \epsilon^2 \mathrm{tr}\, CRC^T + \|(C - I)\theta\|^2. \tag{3.63}$$

Hellinger and L_1 distances. We conclude this section by recording some facts about distances between (Gaussian) measures for use in Section 3.11 on asymptotic equivalence. A more systematic discussion may be found in Lehmann and Romano (2005, Ch. 13.1).

Let P and Q be probability measures on $(\mathcal{X}, \mathcal{B})$ and ν a dominating measure, such as $P + Q$. Let p and q be the corresponding densities. The *Hellinger* distance $H(P, Q)$ and

L_1 or *total variation* distance between P and Q are respectively given by [7]

$$H^2(P,Q) = \tfrac{1}{2}\int (\sqrt{p}-\sqrt{q})^2 d\nu,$$
$$\|P-Q\|_1 = \int |p-q| d\nu.$$

Neither definition depends on the choice of ν. Expanding the square in the Hellinger distance, we have $H^2(P,Q) = 1-\rho(P,Q)$, where ρ is the affinity (3.61). The Hellinger distance is statistically useful because the affinity behaves well for products (i.e. independence), as we have seen. The L_1 distance has a statistical interpretation in terms of the sum of errors of the likelihood ratio test between P and Q:

$$1-\tfrac{1}{2}\|P-Q\|_1 = P(p \le q) + Q(q < p),$$

as is easily checked directly. Thus the sum of errors is small if and only if the L_1 distance is large. The measures are related (Lehmann and Romano, 2005, Th. 13.1.2) by

$$H^2(P,Q) \le \tfrac{1}{2}\|P-Q\|_1 \le [1-\rho^2(P,Q)]^{1/2}. \tag{3.64}$$

It is instructive to compute these distances when $P = P_0$ and $Q = P_\theta$ are Gaussian measures with means 0 and θ, and with common variances. Then $\rho(P_\theta, P_0)$ is given by (3.62) with $\theta' = 0$. To calculate the L_1 distance, observe that the likelihood ratio

$$q/p = \exp(W - D^2/2), \qquad W = \sum_i y_i\theta_i/(\varrho_i\epsilon)^2.$$

Under P_0 and P_θ respectively, $W \sim N(0, D^2)$ and $W \sim N(D^2, D^2)$ and we find

$$\|P_\theta - P_0\|_1 = 2[1 - 2\tilde{\Phi}(D/2)], \tag{3.65}$$

where as usual $\tilde{\Phi}(u) = \int_u^\infty \phi$ is the right Gaussian tail probability. We can now compare the quantities in (3.64) assuming that D is small. Indeed

$$H^2(P_\theta, P_0) \approx D^2/8, \quad \tfrac{1}{2}\|P_\theta - P_0\|_1 \approx \phi(0)D, \quad \text{and} \quad [1-\rho^2(P_\theta,P_0)]^{1/2} \approx D/2.$$

In the continuous Gaussian *white* noise model ($\varrho_i \equiv 1$), we can re-interpret D^2 using Parseval's identity, so that $\|P_f - P_{\bar{f}}\|_1$ is given by (3.65) with

$$D^2 = \epsilon^{-2}\int_0^1 (f-\bar{f})^2. \tag{3.66}$$

3.9 Linear inverse problems

The continuous signal in Gaussian noise model led to a homoscedastic version of the basic sequence model (1.12). The more general form with unequal variances can arise when we do not observe f—ignoring the noise for now—but rather its image Af after the action of an operator A, representing some form of integration, smoothing or blurring. The recovery of f from the indirect observations Af is called an *inverse problem* and has a rich literature which

[7] Caution: definitions in the literature can differ by a factor of 2.

we barely touch. We consider only linear operators A and settings which lend themselves to expression in sequence model form.

We begin with an idealized extension of the continuous white noise model (1.21) and then pass to examples. Suppose, then, that the unknown function f is defined and square integrable on some domain $T \subset \mathbb{R}^d$.

The linear operator A is assumed to be bounded as a transformation from $\mathcal{H} = L_2(T, \mu_1)$ to $\mathcal{K} = L_2(U, \mu_2)$. Let the inner products on $\mathcal{H}$ and $\mathcal{K}$ be denoted $\langle \cdot, \cdot \rangle$ and $[\cdot, \cdot]$ respectively. The observations are given by

$$Y = Af + \epsilon Z, \tag{3.67}$$

a process on U, interpreted to mean that for each $\psi \in L_2(U, \mu_2)$, the observation is a scalar

$$Y(\psi) = [Af, \psi] + \epsilon Z(\psi), \tag{3.68}$$

and $Z = \{Z(\psi)\}$ is a Gaussian process with mean zero and covariance function

$$\mathrm{Cov}(Z(\psi), Z(\psi')) = \int_U \psi \psi' \, d\mu_2. \tag{3.69}$$

The setting of *direct* estimation, in which $A = I$, is a special case in which $\mathcal{H} = \mathcal{K} = L_2[0, 1]$. With $\psi = I_{[0,t]}$, we write $Y(t)$ for $Y(\psi)$ and recover the signal in continuous white Gaussian noise model (1.21). Also included is the case of matrices, considered in Section 2.9, with $T = \{1, \ldots, p\}$ and $U = \{1, \ldots, n\}$, with μ_1 and μ_2 given by counting measure and with A an $n \times p$ design matrix.

To arrive at the sequence form of (3.67), we employ the singular value decomposition (SVD) of the operator A. For definitions and details, see Appendix C.2 and the references given there. Suppose that $A : \mathcal{H} \to \mathcal{K}$ is a compact linear operator between Hilbert spaces, with null space $N(A) = \{f \in \mathcal{H} : Af = 0\}$ and adjoint $A^* : \mathcal{K} \to \mathcal{H}$. The *singular value decomposition* of A consists of two sets of singular functions

(i) $\{\varphi_k\}$, an orthonormal set in $\mathcal{H}$ whose closed linear span equals the orthogonal complement of $N(A)$,
(ii) $\{\psi_k\}$, an orthonormal set in $\mathcal{K}$, and
(iii) singular values $b_k > 0$, such that

$$A\varphi_k = b_k \psi_k, \qquad A^* \psi_k = b_k \varphi_k.$$

From $[Af, \psi] = \langle f, A^* \psi \rangle$ and this last display, we have

$$[Af, \psi_k] = b_k \langle f, \varphi_k \rangle. \tag{3.70}$$

Suppose now that A is one-to-one, so that $\{\varphi_k\}$ is an orthonormal basis for $\mathcal{H}$. We can expand $f = \sum_k \theta_k \varphi_k$ where $\theta_k = \langle f, \varphi_k \rangle$. The *representer equations* (3.70) show that f can be represented in terms of Af through $[Af, \psi_k]$. Indeed,

$$f = \sum b_k^{-1} [Af, \psi_k] \varphi_k. \tag{3.71}$$

As observables, introduce $Y_k = Y(\psi_k)$. From our model (3.68), $Y_k = [Af, \psi_k] + \epsilon Z(\psi_k)$. The representer equations say that $[Af, \psi_k] = b_k \langle f, \varphi_k \rangle = b_k \theta_k$, so that Y_k / b_k is unbiased

for θ_k. Now introduce $z_k = Z(\psi_k)$; from the covariance formula (3.69) and orthogonality in $L_2(U)$, it is evident that the z_k are i.i.d. $N(0,1)$. We arrive at the sequence representation

$$Y_k = b_k\theta_k + \epsilon z_k. \tag{3.72}$$

As with the regression model of Chapter 2, we set $y_k = Y_k/b_k$ and $\varrho_k = \epsilon/b_k$ to recover our basic sequence model (1.12). After next describing some examples, we return later in this section to the question of building an estimator of f from the $\{Y_k\}$, or equivalently the $\{y_k\}$. However, it is already clear that the rate of variation inflation, i.e. the rate of decrease of b_k with k, plays a crucial role in the analysis.

Examples

(i) Deconvolution. Smoothing occurs by convolution with a known integrable function a:

$$Af(u) = (a \star f)(u) = \int_0^1 a(u-t)f(t)dt,$$

and the goal is to reconstruct f. The two-dimensional analog is a natural model for image blurring.

The easiest case for describing the SVD occurs when a is a periodic function on $\mathbb{R}$ with period 1. It is then natural to use the Fourier basis for $\mathcal{H} = \mathcal{K} = L_2[0,1]$. In the complex form, $\varphi_k^C(u) = e^{2\pi kiu}$, and with $f_k = \langle f, \varphi_k^C\rangle, a_k = \langle a, \varphi_k^C\rangle$, the key property is

$$(a \star f)_k = a_k f_k, \quad \leftrightarrow \quad A\varphi_k^C = a_k\varphi_k^C.$$

If a is also even, $a(-t) = a(t)$, then a_k is real valued, the singular values $b_k = |a_k|$, and the singular functions $\psi_k = \mathrm{sign}(a_k)\varphi_k^C$.

If $a(u) = I\{|u| \le u_0\}$ is the "boxcar" blurring function, then $a_k = \sin(2\pi k u_0)/(\pi k)$, so that the singular values $b_k \approx O(k^{-1})$. For a smooth, say with r continuous derivatives, then $b_k = O(k^{-r})$ (e.g. Katznelson (1968, Ch. 1.4)).

(ii) Differentiation. We observe $Y = g + \epsilon Z$, with g assumed to be 1-periodic and seek to estimate the derivative $f = g'$. We can express g as the output of integration: $g(u) = Af(u) = \int_0^u f(t)dt + c(f)$, where $c(f)$ is the arbitrary constant of integration. We suppose that $\mathcal{H} = \mathcal{K} = L_2[0,1]$. Roughly speaking, we can take the singular functions φ_k and ψ_k to be the trigonometric basis functions, and the singular values $b_k \approx 1/|\pi k| = O(k^{-1})$.

A little more carefully, consider for example the real trignometric basis (3.8). Choose the constants of integration $c(\varphi_k)$ so that $A\varphi_{2k} = -(2\pi k)^{-1}\varphi_{2k+1}$ and $A\varphi_{2k+1} = (2\pi k)^{-1}\varphi_{2k}$. Since the observed function is assumed periodic, it is reasonable to set $A\varphi_1 = 0$. So now A is well defined on $L_2[0,1]$ and one checks that $A^* = -A$ and hence, for $k \ge 1$ that the singular values $b_{2k} = b_{2k+1} = 1/|2\pi k|$.

More generally, we might seek to recover $f = g^{(m)}$, so that, properly interpreted, g is the m-th iterated integral of f. Now the singular values $b_k \approx |\pi k|^{-m} = O(k^{-m})$ for $k \ne 1$.

(iii) The Abel equation $Af = g$ has

$$(Af)(u) = \frac{1}{\sqrt{\pi}}\int_0^u \frac{f(t)}{\sqrt{u-t}}dt$$

and goes back to Abel (1826), see Keller (1976) for an engaging elementary discussion and Gorenflo and Vessella (1991) for a list of motivating applications, including Abel's original tautochrone problem.

The singular value decomposition in this and the next example is not so standard, and the derivation is outlined in Exercise 3.18. To describe the result, let $\mathcal{H} = L_2[0,1]$ with $\{\varphi_k\}$ given by normalized Legendre polynomials $\varphi_k(u) = \sqrt{2k+1}P_k(1-2u)$, $k \geq 0$. On the other side, again for $k \geq 0$, let

$$\psi_k(u) = \sqrt{2/\pi}\sin(k+\tfrac{1}{2})\theta, \qquad u = \sin^2(\theta/2)$$

for $0 \leq \theta \leq \pi$. Setting $\tilde{\psi}_k(\theta) = \psi_k(u)$, the functions $\tilde{\psi}_k$ are orthonormal in $L_2[0,\pi]$ (and $\psi_k(u)$ can be expressed in terms of modified Jacobi polynomials $\sqrt{u}P_k^{1/2,-1/2}(1-2u)$, see (3.75) below). It is shown in Exercise 3.18 that $A\varphi_k = b_k\psi_k$ with singular values

$$b_k = (k+1/2)^{-1/2}.$$

Thus, in terms of decay of singular values, A behaves like half-order integration.

Remark. It is perhaps not immediately clear that A is a bounded linear operator on $L_2[0,1]$ (although of course it follows from the SVD). The kernel $A(u,t) = (u-t)^{-1/2}I\{u \geq t\}$ is not square integrable on $[0,1]^2$, so the simplest criterion, finiteness of the Hilbert-Schmidt norm (C.7), doesn't apply. See the chapter Notes for further remarks.

(iii') Wicksell problem. Following Wicksell (1925) and Watson (1971), suppose that spheres are embedded in an opaque medium and one seeks to estimate the density of the sphere radii, p_S, by taking a planar cross-section through the medium and estimating the density p_O of the observed circle radii.

Assume that the centers of the spheres are distributed at random according to a homogeneous Poisson process. Then p_O and p_S are related by (Watson, 1971, eq. (5))

$$p_O(y) = \frac{y}{\mu}\int_y^b \frac{p_S(s)}{\sqrt{s^2-y^2}}ds, \qquad \mu = \int_0^b sp_S(s)ds. \tag{3.73}$$

We may put this into Abel equation form. Suppose, by rescaling, that $b = 1$ and work on the scale of squared radii, letting g be the density of $u = 1-y^2$ and p be the density of $t = 1-s^2$. Setting $\kappa = 2\mu/\sqrt{\pi}$, we get

$$g(u) = \frac{1}{2\mu}\int_0^u \frac{p(t)}{\sqrt{u-t}}dt = \frac{1}{\kappa}(Ap)(u).$$

Thus we can use observations on g and the SVD of A to estimate $f = p/\kappa$. To obtain an estimate of p we can proceed as follows. Since $\varphi_0 \equiv 1$ and p is a probability density, we have $\langle p, \varphi_0\rangle = 1$. Thus from (3.70)

$$1 = \kappa\langle f, \varphi_0\rangle = \kappa b_0^{-1}[Af, \psi_0]$$

and so $\kappa = b_0/[g, \psi_0]$ and hence

$$p = \kappa f = \sum_k \frac{b_0}{b_k}\frac{[g,\psi_k]}{[g,\psi_0]}\varphi_k$$

expresses p in terms of observable functions $[g, \psi_k]$.

(iv) Fractional order integration. For $\delta > 0$, let

$$(A_\delta f)(u) = \frac{1}{\Gamma(\delta)} \int_0^u \frac{f(t)}{(u-t)^{1-\delta}} dt = (f \star \Psi_\delta)(u) \tag{3.74}$$

where $\Psi_\delta(x) = x_+^{\delta-1}/\Gamma(\delta)$ and $x_+ = \max(x, 0)$. Gel'fand and Shilov (1964, §5.5) explain how convolution with Ψ_δ and hence operator A_δ can be interpreted as integration of (fractional) order δ. Of course, $(A_1 f)(u) = \int_0^u f(t)dt$ is ordinary integration and $\delta = 1/2$ yields the Abel operator.

The SVD of A_δ can be given in terms of Jacobi polynomials $P_k^{a,b}(1-2x), k \geq 0$, and their normalization constants $g_{a,b;k}$, Appendix C.31 and Exercise 3.18:

$$\begin{aligned} \varphi_k(u) &= \sqrt{2k+1} P_k(1-2u) && \text{on } L_2([0,1], du) \\ \psi_k(u) &= g_{\delta,-\delta;k}^{-1} u^\delta P_k^{\delta,-\delta}(1-2u) && \text{on } L_2([0,1], u^{-\delta}(1-u)^{-\delta} du), \\ b_k &= (\Gamma(k-\delta+1)/\Gamma(k+\delta+1))^{1/2}. \end{aligned} \tag{3.75}$$

Thus, consistent with previous examples, the singular values $b_k \sim k^{-\delta}$ as $k \to \infty$, and so decay at a rate corresponding to the order (integer or fractional) of integration.

(v) Heat equation. The classical one dimensional heat equation describes the diffusion of heat in a rod. If $u(x,t)$ denotes the temperature at position x in the rod at time t, then in appropriate units, u satisfies the equation [8]

$$\frac{\partial}{\partial t} u(x,t) = \frac{\partial^2}{\partial x^2} u(x,t). \tag{3.76}$$

For our discussion here, we will assume that the initial temperature profile $u(x,0) = f(x)$ is unknown, and that the boundary conditions are periodic: $u(0,t) = u(1,t)$. We make noisy observations on the temperature in the rod at a time $T > 0$.

$$Y(x) = u(x,T) + \epsilon Z(x),$$

and it is desired to estimate the initial condition $f(x)$. See Figure 3.2.

The heat equation (3.76) is a *linear* partial differential equation, having a unique solution which is a linear transform of the intial data f:

$$u(x,T) = (A_T f)(x).$$

This can be expressed in terms of the Gaussian heat kernel, but we may jump directly to the SVD of A_T by recalling that (3.76) along with the given boundary conditions can be solved by separation of variables. If we assume that the unknown, periodic f has Fourier sine expansion

$$f(x) = \sqrt{2} \sum_{k=1}^{\infty} \theta_k \sin \pi k x,$$

[8] In this and the next example, we use notation conventional for these settings.

then it is shown in introductory books on partial differential equations that the solution

$$u(x,T) = \sqrt{2}\sum_{k=1}^{\infty} \theta_k e^{-\pi^2 k^2 T} \sin \pi k x.$$

Thus $\varphi_k(x) = \psi_k(x) = \sqrt{2}\sin \pi k x$, and the singular values $b_k = e^{-\pi^2 k^2 T}$.

The very rapid decay of b_k shows that the heat equation is extraordinarily ill-posed.

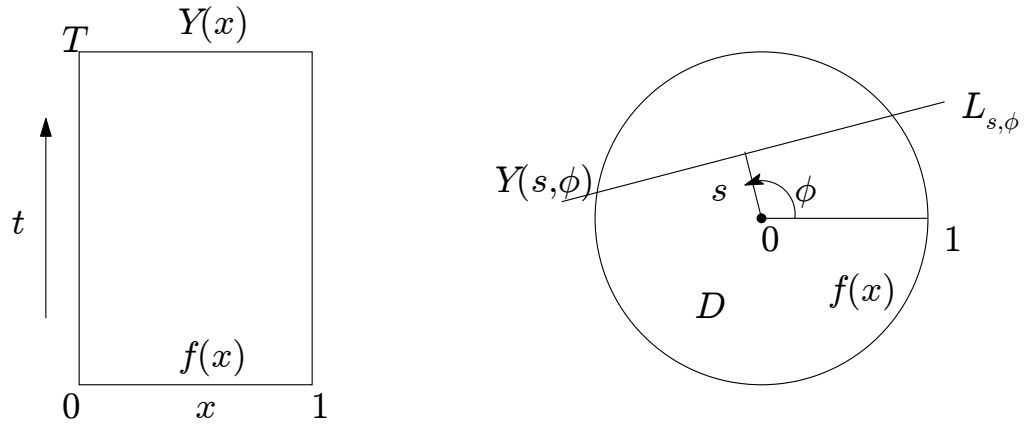

Figure 3.2 Left panel: domain for the heat equation. We observe $u(x,T)$ plus noise (top line) and wish to recover the initial data $f(x) = u(x,0)$, (bottom line). Right panel: domain for computed tomography example. We observe line integrals $(Af)(s,\phi)$ along lines $L_{s,\phi}$ plus noise, and wish to recover $f(x), x \in D$.

(vi) Radon transform and 2-d computed tomography (CT). In a two-dimensional idealization, this is the problem of reconstructing a function from its line integrals. Thus, let $T = D$ be the unit disc in $\mathbb{R}^2$, and suppose that the unknown $f \in \mathcal{H} = L^2(D, \pi^{-1}dx)$.

A line at angle ϕ from the vertical and distance s from the origin is given by $t \to (s\cos\phi - t\sin\phi, s\sin\phi + t\cos\phi)$ and denoted by $L_{s,\phi}$, compare Figure 3.2. The corresponding line integral is

$$\begin{aligned}(Af)(s,\phi) &= \text{Ave}\,[f\,|L_{s,\phi} \cap D] \\ &= \frac{1}{2\sqrt{1-s^2}} \int_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} f(s\cos\phi - t\sin\phi, s\sin\phi + t\cos\phi)dt.\end{aligned}$$

Here $(s,\phi) \in R = \{0 \le s \le 1, 0 \le \phi \le 2\pi\}$. The observations are noisy versions of the line integrals

$$Y(s,\phi) = Af(s,\phi) + \epsilon W(s,\phi), \qquad (s,\phi) \in R.$$

The SVD of A was derived in the optics and tomography literatures (Marr, 1974; Born and Wolf, 1975); we summarize it here as a two-dimensional example going beyond the Fourier basis. There is a double index set $\mathcal{N} = \{(l,m) : m = 0, 1, \ldots; l = m, m-2, \ldots, -m\}$,

where m is the "degree" and l the "order". For $\nu = (l, m)$, the singular functions are

$$\varphi_\nu(r,\theta) = \sqrt{m+1}Z_m^{|l|}(r)e^{il\theta}, \qquad \psi_\nu(s,\phi) = U_m(s)e^{il\phi},$$

and the singular values $b_\nu = 1/\sqrt{m+1}$. Here $U_m(\cos\theta) = \sin(m+1)\theta/\sin\theta$ are Chebychev polynomials of the second kind, and the *Zernike* polynomials are characterized by the orthogonality relation $\int_0^1 Z_{k+2s}^k(r)Z_{k+2t}^k(r)rdr = ((k+2s+1)/2)\delta_{st}$.

The main point here is that the singular values b_ν decay slowly: the reconstruction problem is only mildly ill-posed, consistent with the now routine use of CT scanners in medicine.

Ill-posedness and Estimation

Return to observation model (3.67) and its sequence form (3.72). In view of the representation (3.71) and the fact that $EY_k = [Af, \psi_k]$, it is natural to consider a class of weighted diagonal linear estimators

$$\hat{f}(t) = \sum_k \hat{\theta}_k \varphi_k(t) = \sum_k c_k b_k^{-1} Y_k \varphi_k(t),$$

where the shrinkage constants $c_k \in [0,1]$ are chosen to counteract the variance inflation effects of the small singular values b_k. Examples that fall within this class include

(a) Truncated SVD, also known as a projection or "spectral cut-off" estimator:

$$\hat{\theta}_{\nu,k} = \begin{cases} b_k^{-1}Y_k & k \le \nu, \\ 0 & k > \nu. \end{cases}$$

Equivalently, $c_k = I\{k \le \nu\}$, which projects onto the lowest ν generalized frequencies, compare (3.17).

(b) Tikhonov(-Phillips) regularization. We assume that there is a sequence of positive increasing, constants w_k^2 such that for each value of the *regularization parameter* $\lambda > 0$,

$$\hat{\theta}_{\lambda,k} = \frac{b_k}{b_k^2 + \lambda w_k} Y_k.$$

In this case the shrinkage constants $c_k = b_k^2/(b_k^2 + \lambda w_k)$. For direct estimation, $b_k \equiv 1$, and in the Fourier basis with $w_{2k} = w_{2k+1} = (2k)^{2m}$, this reduces to the mth order smoothing spline estimator. In the general case, it arises from a penalized least squares problem

$$\min_f \|Y - Af\|^2 + \lambda\|\Omega f\|^2,$$

if the singular functions φ_k also satisfy $\Omega^*\Omega\varphi_k = w_k\varphi_k$. This occurs, for example, in the trigonometric basis, with $\Omega = D^m$, namely mth order differentiation.

Rates of convergence. We use the example of the truncated SVD to give a brief discussion of the connection between the decay of the singular values b_k and rates of convergence, following the approach of Section 3.2. A similar analysis is possible for the regularized Tikhonov estimates, along the lines of Section 3.6.

We carry out the analysis in model (3.60), so that $\varrho_k = 1/b_k$. Then the truncated SVD

estimator is identical to (3.17) and the analysis of maximum risk over ellipsoids $\Theta_2^\alpha(C)$ can be patterned after (3.14). As was done there, let $\hat{\theta}_\nu$ be given by (3.17). Its mean squared error

$$r(\hat{\theta}_\nu, \theta) = E\|\hat{\theta}_\nu - \theta\|_2^2 = \epsilon^2 \sum_{k=1}^{\nu} \varrho_k^2 + \sum_{k>\nu} \theta_k^2,$$

and as before, let $\bar{r}(\hat{\theta}_\nu; \epsilon)$ denote its maximum risk over parameter space Θ.

Proposition 3.11 *Consider the non-white sequence model $\varrho_k = k^\beta$ for $\beta \geq 0$. Introduce the rate parameter $r = 2\alpha/(2\alpha + 2\beta + 1)$ and let $s = r/(2\alpha) = 1/(2\alpha + 2\beta + 1)$. Let $\hat{\theta}_\nu$ be the truncation estimator* (3.17) *and choose $\nu_* = \lceil \gamma C^2/\epsilon^2)^s \rceil$ for some $\gamma > 0$. Then*

$$R_N(\Theta_2^\alpha(C), \epsilon) \leq \bar{r}(\hat{\theta}_{\nu_*}) \leq c_{\alpha\beta\gamma} C^{2(1-r)} \epsilon^{2r}(1 + O(\epsilon^{2s})). \tag{3.77}$$

This rate r is in fact optimal, as is shown in Proposition 4.23–more refined results at the level of constants come with Pinsker's theorem in Chapter 5. We see that the effect of the degree β of decay of the singular values is to degrade the rate of convergence of maximum MSE from $2\alpha/(2\alpha + 1)$ to $2\alpha/(2\alpha + 2\beta + 1)$. Thus the faster the decay of the singular values, the slower the rate of convergence, and also the smaller the frequency ν_* at which the data is truncated. For this reason β is sometimes called an index of the ill-posedness.

Proof For any ellipsoid $\Theta(a, C)$ we have $\bar{r}(\hat{\theta}_\nu; \epsilon) = \epsilon^2 \sum_{k=1}^{\nu} \varrho_k^2 + C^2 a_{\nu+1}^{-2}$, so long as a_k^2 is increasing with k. Inserting $a_k = k^\alpha$ and $\varrho_k = k^\beta$ and using an integral approximation,

$$\bar{r}(\hat{\theta}_\nu; \epsilon) \leq (2\beta + 1)^{-1} \epsilon^2 (\nu + 1)^{2\beta+1} + C^2 (\nu + 1)^{-2\alpha}.$$

Setting $\nu_* + 1 = (\gamma C^2/\epsilon^2)^s$, we obtain the leading term in (3.77), and converting ν_* to an integer yields a relative error $1 + O(\epsilon^{2s})$. □

Remark. If $\varrho_k = 1/b_k$ grows exponentially fast, or worse, then the inversion problem might be called *severely* ill-posed. In this case, the attainable rates of convergence over Sobelev smoothness classes $\Theta_2^\alpha(C)$ are much slower: Exercise 3.20, for the heat equation, derives rates that are algebraic in $\log \epsilon^{-1}$. One can recover rates of convergence that are "polynomial" in ϵ by assuming much greater smoothness, for example by requiring Θ to be an ellipsoid of analytic functions, see Section 5.2 and Exercise 5.2.

3.10 Correlated noise

The Karhunen-Loève transform. Let $T = [a, b]$ or more generally, a compact set in $\mathbb{R}^d$. Suppose that $\{Z(t), t \in T\}$ is a zero mean Gaussian random process on an index set T. [That is, all finite-dimensional distributions $(Z(t_1), \ldots, Z(t_k))$ are Gaussian for all $(t_1, t_2, \ldots, t_k) \in T^k$ and positive integer k.] Assume also that Z is continuous in quadratic mean, or equivalently (Ash and Gardner, 1975, Ch 1.3) that the covariance function (or kernel)

$$R(s, t) = EZ(s)Z(t)$$

is jointly continuous in $(s,t) \in T^2$. The operator $Rf(s) = \int R(s,t)f(t)dt$ is nonnegative definite because it arises from a covariance kernel:

$$\langle Rf, f\rangle = \iint f(s)\mathrm{Cov}(Z(s), Z(t))f(t)dsdt = \mathrm{Var}\left(\int f(s)Z(s)ds\right) \geq 0.$$

Under these conditions it follows (Appendix C.4 has some details and references) that R is a compact operator on $L^2(T)$, and so it has, by the Hilbert-Schmidt theorem, a complete orthonormal basis $\{\varphi_k\}$ of eigenfunctions with eigenvalues $\varrho_k^2 \geq 0$,

$$\int R(s,t)\varphi_k(t)dt = \varrho_k^2\varphi_k(s), \qquad s \in T.$$

In addition, by Mercer's Theorem C.5, the series

$$R(s,t) = \sum \varrho_k^2\varphi_k(s)\varphi_k(t)$$

converges uniformly and in mean square on $T \times T$.

Define Gaussian variables (for k such that $\varrho_k > 0$)

$$z_k = \varrho_k^{-1}\int \varphi_k(t)Z(t)dt.$$

The z_k are i.i.d. $N(0,1)$: this follows from the orthonormality of eigenfunctions:

$$\mathrm{Cov}\left(\int \varphi_k Z, \int \varphi_{k'} Z\right) = \int_{T\times T} \varphi_k R\varphi_{k'} = \langle \varphi_k, R\varphi_{k'}\rangle = \varrho_k^2\delta_{kk'}. \tag{3.78}$$

The sum

$$Z(t) = \sum_k \varrho_k z_k \varphi_k(t)$$

converges in mean-square uniformly in t. Indeed, for a tail sum $r_{mn} = \sum_m^n \langle Z, \varphi_k\rangle\varphi_k$ we have, using (3.78), $Er_{mn}^2 = \sum_{i=m}^n \varrho_k^2\varphi_k^2(t) \to 0$ uniformly in t as $m, n \to \infty$ by Mercer's theorem.

If the eigenfunctions φ_k corresponding to $\varrho_k > 0$ 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 orthobasis for $L_2(T)$. Since R is symmetric, these φ_k correspond to $\varrho_k = 0$.

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

$$Y(t) = \theta(t) + \epsilon Z(t), \qquad t \in T.$$

Such a model occurs commonly in functional data analysis, in which $Y(t)$ models a smooth curve and there is smoothness also in the noise process due to correlation. See for example Ramsay and Silverman (2005); Hall and Hosseini-Nasab (2006), and Exercise 3.16.

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

$$y_k = \langle Y, \varphi_k\rangle, \qquad \theta_k = \langle \theta, \varphi_k\rangle,$$

to obtain exactly the sequence model (3.60). [Of course, co-ordinates corresponding to $\varrho_k = 0$ are observed perfectly, without noise.] From the discussion in Section 3.8, it follows that model P_θ is equivalent to P_0 if and only if $\sum_k \theta_k^2/\varrho_k^2 < \infty$.

To summarize: for our purposes, the Karhunen-Loève transform gives (i) a diagonalization of the covariance operator of a mean-square continuous Gaussian process and (ii) an example of the Gaussian sequence model. As hinted at in the next subsection it also provides a way to think about and do computations with Gaussian priors in the sequence model.

Connection to Principal Components Analysis. Constructing the KLT is just the stochastic process analog of finding the principal components of a sample covariance matrix. Indeed, suppose that the sample data is $\{x_{ij}\}$ for $i = 1, \ldots, n$ cases and $j = 1, \ldots, p$ variables. Let $\bar{x}_j = n^{-1} \sum_i x_{ij}$ denote the sample mean for variable j. Set $z_{ij} = x_{ij} - \bar{x}_j$ and make the correspondence $Z(\omega, t) \leftrightarrow z_{ij}$, identifying the realization ω with i, and the "time" t with j. Then $R(t_1, t_2) = EZ(t_1)Z(t_2)$ corresponds to an entry in the sample covariance matrix $S_{j_1 j_2} = n^{-1} \sum_i (x_{ij_1} - \bar{x}_{j_1})(x_{ij_2} - \bar{x}_{j_2})$.

Example: Integrated Wiener process priors.

In Section 3.4, it was seen that (periodic) smoothing spline estimators could be viewed as a posterior mean Bayes estimator for a suitable Gaussian prior. We show here that the prior can be interpreted in terms of the Karhunen-Loève transform of integrated Brownian motion.

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

$$Z_m^0(t) = \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} dW(u), \qquad t \in [0, 1].$$

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:

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

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

$$Z_2^\sigma(t) = \sigma\xi_0 + \sigma\xi_1 t + \int_0^t (t-u) dW(u). \tag{3.79}$$

Wahba (1978, 1983, 1990) has advocated the use of Z_m^σ 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. She showed (Wahba, 1990, Th. 1.5.3), for the nonparametric regression setting (1.13), that the smoothing spline based on the roughness penalty $\int (D^m f)^2$ arises as the limit of posterior means calculated from the Z_m^σ priors as $\sigma \to \infty$.)

This prior distribution has some curious features, so we explore its Karhunen-Loève transform. The key conclusion: for each $\sigma \geq 0$, and in the $\sigma \to \infty$ limit, the eigenvalues satisfy

$$\varrho_k \sim (\pi k)^{-m}, \qquad \text{as } k \to \infty.$$

Recall that the spline estimator (3.40) in the Gaussian sequence model arises as a Bayes estimator for the prior with independent $N(0, \tau_k^2)$ co-ordinates with $\tau_k \propto k^{-m}$.

We discuss only the cases $m = 1, 2$ here. For general m, the same behavior (for Z_m^0) is established by Gao et al. (2003).

It is simpler to discuss the $m = 1$ situation first, with $Z_1^\sigma(t) = \sigma\xi_0 + W(t)$, and covariance

kernel $R_\sigma(s,t) = \mathrm{Cov}\,(Z_1^\sigma(s), Z_1^\sigma(t)) = \sigma^2 + s \wedge t$. The eigenvalue equation $R_\sigma \varphi = \varrho^2 \varphi$ becomes

$$\sigma^2 \int_0^1 \varphi(t)dt + \int_0^s t\varphi(t)dt + s\int_s^1 \varphi(t)dt = \varrho^2\varphi(s). \tag{3.80}$$

Differentiating with respect to s yields

$$\int_s^1 \varphi(t)dt = \varrho^2\varphi'(s) \tag{3.81}$$

and differentiating a second time yields the second order ordinary differential equation

$$-\varphi(s) = \varrho^2\varphi''(s) \qquad 0 \le s \le 1. \tag{3.82}$$

The homogeneous equation $\varrho^2\varphi'' + \varphi = 0$ has two linearly independent solutions given by trigonometric functions

$$\varphi(t) = a\sin(t/\varrho) + b\cos(t/\varrho). \tag{3.83}$$

The equations (3.80) and (3.81) impose boundary conditions which non-zero eigenfunctions must satisfy:

$$\varphi'(1) = 0, \qquad \varphi'(0) = \varphi(0)/\sigma^2.$$

[The first condition is evident from (3.81) while the second follows by combining the two equations: $\varrho^2\varphi'(0) = \int \varphi = \varrho^2\varphi(0)/\sigma^2$.]

Let us look first at the $\sigma \to \infty$ limit advocated by Wahba. In this case the boundary conditions become simply $\varphi'(0) = \varphi'(1) = 0$. Substituting into (3.83), the first condition implies that $a = 0$ and the second that $\sin(1/\varrho) = 0$. Consequently the eigenvalues and eigenfunctions are given by

$$\varrho_k = 1/k\pi, \qquad \varphi_k(s) = \sqrt{2}\cos k\pi s, \qquad k = 1, 2, \ldots.$$

Equation (3.82) arises in traditional mathematical physics by separation of variables in the 'vibrating string' equation, e.g. Courant and Hilbert (1953, Sec. 5.3). The boundary condition $\varphi'(1) = 0$ corresponds 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 $\varphi(0) = 0$, when the left end of the string is fixed at 0–recall that $W(0) = 0$ almost surely. [9] The condition for general σ, $\varphi'(0) = \varphi(0)/\sigma^2$, corresponds to an 'elastically attached' endpoint.

Table 3.2 shows the eigenvalues ϱ_k and eigenfunctions corresponding to these various natural boundary conditions - all are easily derived from (3.83).

To describe the stochastic process, or "prior distribution" associated with *periodic* boundary conditions, recall that the Brownian Bridge $\tilde{W}(t) = W(t) - tW(1)$ satisfies $\tilde{W}(1) = \tilde{W}(0) = 0$ and has $\mathrm{Cov}(\tilde{W}(s), \tilde{W}(t)) = s \wedge t - st$. Proceeding as before, define a "free" Brownian Bridge

$$\tilde{Z}^\sigma(t) = \sigma\xi_0 + \tilde{W}(t),$$

and verify that it has covariance kernel $\tilde{R}_\sigma(s,t) = \sigma^2 + s \wedge t - st$. Equations (3.80) and

[9] In this case, the eigenfunctions $\sqrt{2}\sin(k + \frac{1}{2})\pi t$ happen to coincide with the left singular functions of the Abel transform of the previous section.

	Boundary Conditions	Eigenvalues	Eigenfunctions
$\sigma = \infty$	$\varphi'(0) = \varphi'(1) = 0$	$\varrho_k^{-1} = k\pi$	$\sqrt{2}\cos k\pi t$
$\sigma = 0$	$\varphi(0) = \varphi'(1) = 0$	$\varrho_k^{-1} = (k + \frac{1}{2})\pi$	$\sqrt{2}\sin(k + \frac{1}{2})\pi t$
$0 < \sigma < \infty$	$\varphi'(0) = \varphi(0)/\sigma^2$, $\varphi'(1) = 0$	$\varrho_k^{-1} \in (k\pi, (k + \frac{1}{2})\pi)$	$c_k \sin \varrho_k^{-1} t + \ldots$ $c_k \sigma^2 \varrho_k^{-1} \cos \varrho_k^{-1} t$
Periodic	$\varphi(0) = \varphi(1)$, $\varphi'(0) = \varphi'(1)$	$\varrho_{2k-1}^{-1} = \varrho_{2k}^{-1} = 2\pi k$	$\sqrt{2}\sin 2\pi k t$, $\sqrt{2}\cos 2\pi k t$

Table 3.2 *Effect of Boundary Conditions for the vibrating string equation*

(3.81) change in an obvious way, but the differential equation (3.82) remains the same. The boundary conditions become

$$\varphi(0) = \varphi(1), \qquad \varphi'(0) = \sigma^{-2}\varphi(0) + \varphi'(1),$$

and so the standard periodic boundary conditions and the usual sine and cosine eigenfunctions, as used in Section 3.4, emerge in the $\sigma \to \infty$ limit. See the final row of Table 3.2.

In all cases summarized in Table 3.2, the eigenfunctions show increasing oscillation with increasing k, 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 and 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: $\varrho_k^{-1} \sim k\pi$.

The analysis of the integrated Wiener prior (3.79), corresponding to cubic smoothing splines, then proceeds along the same lines, with most details given in Exercise 3.11 (see also Freedman (1999, Sec. 3)). The eigenvalue equation is a *fourth* order differential equation:

$$\varphi(s) = \varrho^2 \varphi^{(4)}(s).$$

This equation is associated with the vibrating *rod* (Courant and 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 $\sigma = \infty$ limit become

$$\varphi''(0) = \varphi'''(0) = 0, \qquad \varphi''(1) = \varphi'''(1) = 0,$$

corresponding to "free ends" at both limits.

3.11 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 using Le Cam's notion of asymptotic equivalence of experiments. In each such 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.

This section has two parts. In the first, we give the proof of the simplest case of the equivalence result of Brown and Low (1996a), which shows that nonparametric regression on [0, 1] is asymptotically equivalent with the Gaussian white noise model. Some heuristics for this convergence were given in Chapter 1.4.

In the second part, essentially independent of the first, we give an informal, heuristic account of some of the other results in the growing list of equivalence results. The reader primarily interested in heuristcs can jump there directly.

Brown and Low's equivalence theorem

Outline of approach. We consider three statistical problems, each indexed by n, and having a common parameter space $f \in \Theta$.

$$(\mathcal{P}_n) \qquad dY_n(t) = f(t)dt + \sigma n^{-1/2} dW(t), \qquad 0 \le t \le 1, \tag{3.84}$$

$$(\bar{\mathcal{P}}_n) \qquad d\bar{Y}_n(t) = \bar{f}_n(t)dt + \sigma n^{-1/2} dW(t), \qquad 0 \le t \le 1, \tag{3.85}$$

$$(\mathcal{Q}_n) \qquad Y_l = f(l/n) + \sigma Z_l \qquad l = 1, \dots, n. \tag{3.86}$$

In problem $(\bar{\mathcal{P}}_n)$, the function $\bar{f}_n$ is a step function approximation to f, being piecewise constant on intervals $[(l-1)/n, l/n)$. $\mathcal{Q}_n$ is the nonparametric regression problem (1.13) with sample size n, while $\mathcal{P}_n$ is the continuous Gaussian white noise model at noise level $\epsilon = \sigma/\sqrt{n}$. We will define a distance $\Delta(\mathcal{P}_n, \mathcal{Q}_n)$ between statistical problems and show that it converges to zero in two steps. First, problems $\mathcal{P}_n$ and $\bar{\mathcal{P}}_n$ are on the same sample space, and so a convenient criterion in terms of L_1 distance shows that $\Delta(\mathcal{P}_n, \bar{\mathcal{P}}_n) \to 0$ under suitable conditions on Θ. Second, a reduction by sufficiency will show that in fact $\Delta(\bar{\mathcal{P}}_n, \mathcal{Q}_n) = 0$.

Before implementing this agenda, we need some definitions (due to Le Cam) to formalize the notion of distance between statistical problems. (See Le Cam (1986) and Le Cam and Yang (2000); also Nussbaum (2004) for an introduction and van der Vaart (2002) for historical perspective.)

Consider a regular statistical problem $\mathcal{P}$, taken to be a collection of probability measures $\{P_\theta, \theta \in \Theta\}$ on a sample space $\mathcal{Y}$. [10] Let $\mathcal{A}$ be an action space and $L : \mathcal{A} \times \Theta \to [0, \infty)$ a loss function. The risk function of a (randomized) decision rule $\delta(A|y)$ is denoted by

$$r_L(\delta, \theta) = \iint L(a, \theta)\delta(da|y)P_\theta(dy), \tag{3.87}$$

compare (A.10) and the surrounding discussion for more detail. If $\delta(\cdot|y)$ is a point mass at $\hat{\theta}(y)$, then this definition reduces to (2.11).

Now consider two regular statistical problems $\mathcal{P}_0, \mathcal{P}_1$ with sample spaces $\mathcal{Y}_0, \mathcal{Y}_1$ but the

[10] "Regular" means that it is assumed that the sample space $\mathcal{Y}$ is a complete separable metric space, equipped with the associated Borel σ-field, and that the family $\{P_\theta\}$ is dominated by a σ-finite measure. These assumptions hold for all cases we consider.

same parameter space Θ. Let the two corresponding families of distributions be denoted by $\{P_{i,\theta}, \theta \in \Theta\}$ for $i = 0, 1$. The *deficiency* $\Delta_d(\mathcal{P}_0, \mathcal{P}_1)$ of $\mathcal{P}_0$ with respect to $\mathcal{P}_1$ is the smallest number $\epsilon \in [0, 1]$ such that for every arbitrary loss function L with $0 \le L(a, \theta) \le 1$ and every decision rule δ_1 in problem $\mathcal{P}_1$, there is a decision rule δ_0 in problem $\mathcal{P}_0$ such that $r_{0,L}(\delta_0, \theta) \le r_{1,L}(\delta_1, \theta) + \epsilon$ for all $\theta \in \Theta$. To obtain a distance on statistical problems, we symmetrize and set

$$\Delta(\mathcal{P}_0, \mathcal{P}_1) = \max\{\Delta_d(\mathcal{P}_0, \mathcal{P}_1), \Delta_d(\mathcal{P}_1, \mathcal{P}_0)\}. \tag{3.88}$$

The definition of distance is quite elaborate because it requires that performance in the two problems be similar regardless of the choice of estimand (action space) and measure of performance (loss function). In particular, since the loss functions need not be convex, randomized decision rules must be allowed (cf. (A.10)–(A.13) in Appendix A).

A simplification can often be achieved when the problems have the same sample space.

Proposition 3.12 *If $\mathcal{Y}_0 = \mathcal{Y}_1$ and $\mathcal{P}_0$ and $\mathcal{P}_1$ have a common dominating measure ν, then*

$$\Delta(\mathcal{P}_0, \mathcal{P}_1) \le L_1(\mathcal{P}_0, \mathcal{P}_1),$$

where the maximum L_1 distance is defined by

$$L_1(\mathcal{P}_0, \mathcal{P}_1) = \sup_{\theta \in \Theta} \int |p_{0,\theta}(y) - p_{1,\theta}(y)| \nu(dy). \tag{3.89}$$

Proof In the definition of deficiency, when the sample spaces agree, we can use the same decision rule in $\mathcal{P}_0$ as in $\mathcal{P}_1$, and if we write $\|L\|_\infty = \sup |L(a, \theta)|$, then from (3.87)

$$|r_{0,L}(\delta, \theta) - r_{1,L}(\delta, \theta)| \le \|L\|_\infty \int |p_{0,\theta}(y) - p_{1,\theta}(y)| \nu(dy).$$

In the definition of deficiency, we only consider loss functions with $\|L\|_\infty \le 1$. Maximizing over θ shows that $r_{0,L}(\delta, \theta) \le r_{1,L}(\delta, \theta) + L_1(\mathcal{P}_0, \mathcal{P}_1)$. Repeating the argument with the roles of $\mathcal{P}_0$ and $\mathcal{P}_1$ reversed completes the proof. □

A sufficient statistic causes no loss of information in this sense.

Proposition 3.13 *Let $\mathcal{P}$ be a regular statistical problem with sample space $\mathcal{Y}$. Suppose that $S : \mathcal{Y} \to \mathcal{S}$ is a sufficient statistic, and let $\mathcal{Q} = \{Q_\theta; \theta \in \Theta\}$ denote the problem in which $S = S(Y)$ is observed. Then $\Delta(\mathcal{P}, \mathcal{Q}) = 0$.*

Proof Since $S = S(Y)$ is sufficient for Y, there is a kernel $K(C|s)$ defined for (Borel) subsets $C \subset \mathcal{Y}$ such that $P_\theta(C) = \int K(C|s) Q_\theta(ds)$. This formalizes[11] the notion that the distribution of Y given S is free of θ. Given a decision rule δ for problem $\mathcal{P}$, we define a rule δ' for $\mathcal{Q}$ by $\delta'(A|s) = \int \delta(A|y) K(dy|s)$. By chasing the definitions, it is easy to verify, given a loss function L, that $r_{\mathcal{Q},L}(\delta', \theta) = r_{\mathcal{P},L}(\delta, \theta)$, where the subscripts indicate the statistical problem. Hence $\Delta_d(\mathcal{Q}, \mathcal{P}) = 0$. Since a rule for $\mathcal{Q}$ is automatically a rule for $\mathcal{P}$, we trivially have also $\Delta_d(\mathcal{P}, \mathcal{Q}) = 0$, and hence $\Delta(\mathcal{P}, \mathcal{Q}) = 0$. □

[11] The existence of such a kernel, specifically a regular conditional probability distribution, is guaranteed for a regular statistical problem, see. e.g. Schervish (1995, Appendix B.3) or Breiman (1968).

We are now ready to formulate and prove a special case of the Brown-Low theorem. Consider parameter spaces of Hölder continuous functions of order α. The case $0 < \alpha < 1$ is of most interest here—Appendix C gives the definitions for $\alpha \geq 1$. We set

$$\Theta_H^\alpha(C) = \{f \in C([0,1]) : |f(x) - f(y)| \leq C|x-y|^\alpha, \text{for all } x, y \in [0,1]\}. \quad (3.90)$$

Theorem 3.14 *Let $\mathcal{P}_n$ and $\mathcal{Q}_n$ denote the continuous Gaussian white noise model* (3.84) *and the discrete regression model* (3.86) *respectively. Let the parameter space Θ for both models be the Hölder function class $\Theta_H^\alpha(C)$. Then, so long as $\alpha > 1/2$, the two problems are asymptotically equivalent:*

$$\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0.$$

Proof We pursue the two step approach outlined earlier. Given a function $f \in \Theta_H^\alpha(C)$, define a piecewise constant step function approximation to it from the values $f(l/n)$. Set

$$\bar{f}_n(t) = f(l/n) \qquad \text{if } (l-1)/n \leq t < l/n,$$

and put $\bar{f}_n(1) = f(1)$. [This type of interpolation from sampled values occurs again in Chapter 15.] As indicated at (3.85), let $\overline{\mathcal{P}}_n$ denote the statistical problem in which $\bar{f}_n$ is observed in continuous white noise. Since both $\mathcal{P}_n$ and $\overline{\mathcal{P}}_n$ have sample space $\mathcal{Y} = C([0,1])$ and are dominated, for example by P_0, the distribution of Y_n under $f = 0$, we have $\Delta(\mathcal{P}_n, \overline{\mathcal{P}}_n) \leq L_1(\mathcal{P}_n, \overline{\mathcal{P}}_n)$. The L_1 distance between P_f and $P_{\bar{f}_n}$ can be calculated fairly easily; indeed from (3.65) and (3.66),

$$\|P_f - P_{\bar{f}_n}\|_1 = 2[1 - 2\tilde{\Phi}(D_n(f)/2)],$$

$$\sigma^2 D_n^2(f) = n\int_0^1 [\bar{f}_n(t) - f(t)]^2 dt.$$

From the Hölder assumption $|f(t) - f(l/n)| \leq C|t - l/n|^\alpha$ for $t \in [(l-1)/n, l/n)$. [If $\alpha \geq 1$, it is enough to use $\alpha = 1$ and the Lipschitz property]. Consequently

$$\sigma^2 D_n^2(f) \leq n^2 C^2 \int_0^{1/n} u^{2\alpha} du = (2\alpha+1)^{-1} C^2 n^{1-2\alpha},$$

and this holds *uniformly* for all $f \in \Theta_H^\alpha(C)$. Since $1 - 2\tilde{\Phi}(\delta) \sim 2\phi(0)\delta$ as $\delta \to 0$, we conclude that $L_1(\mathcal{P}_n, \overline{\mathcal{P}}_n) \to 0$ so long as $\alpha > 1/2$.

For the second step, reduction by sufficiency, define

$$S_{n,l}(\bar{Y}_n) = n\left[\bar{Y}_n(l/n) - \bar{Y}_n((l-1)/n)\right], \qquad l = 1, \ldots, n. \quad (3.91)$$

The variables $S_{n,l}$ are independent Gaussians with mean $f(l/n)$ and variance σ^2. Hence the vector $S_n = (S_{n,l})$ is an instance of statistical problem $\mathcal{Q}_n$. In addition, $S_n = S_n(\bar{Y}_n)$ is sufficient for $f \in \Theta$ in problem $\overline{\mathcal{P}}_n$ (Exercise 3.19 prompts for more detail), and so $\Delta(\overline{\mathcal{P}}_n, \mathcal{Q}_n) = 0$. Combining the two steps using the triangle inequality for metric Δ, we obtain $\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0$. □

Remarks. 1. Let us describe how to pass from a procedure in one problem to a corresponding procedure in the other. Given a rule δ_n in regression problem $\mathcal{Q}_n$, we define a rule $\delta'_n(Y_n)$ in the white noise problem $\mathcal{P}_n$ simply by forming $S_n(Y_n)$ as in (3.91) and setting $\delta'_n(Y_n) = \delta_n(S_n)$. In the other direction we use the construction in the proof of Proposition

3.13. Given a rule δ_n in white noise problem $\mathcal{P}_n$, we may equally well use it in problem $\overline{\mathcal{P}}_n$ which has the same sample space as $\mathcal{P}_n$. So we may define δ'_n in the regression problem by

$$\delta'_n(A|s_n) = E\left[\delta_n(A|\bar{Y}_n) \mid S_n(\bar{Y}_n) = s_n\right].$$

The conditional expectation is well defined as an estimator (free of f) by sufficiency, though of course it may in general be hard to evaluate. The evaluation is easy however in the case of a linear estimator $\delta_n(Y_n)(u) = \int_0^1 c(u,t)dY_n(t)$: one can check that

$$\delta'_n(S_n)(u) = \sum_{l=1}^{n} c_{nl}(u)S_{n,l}, \qquad c_{nl}(u) = \int_{(l-1)/n}^{l/n} c(u,t)dt.$$

2. Theorem 3.14 extends to a regression model with unequally spaced and heteroscedastic observations: instead of (3.86), suppose that $\mathcal{Q}_n$ becomes

$$Y_l = f(t_{nl}) + \sigma(t_{nl})Z_l, \qquad l = 1,\dots,n.$$

If $t_{nl} = H^{-1}(l/(n+1))$ for a strictly increasing and absolutely continuous distribution function H and if $\sigma(t)$ is well-behaved, then after suitably modifying the definition (3.91), Brown and Low (1996a) show that $\mathcal{Q}_n$ is still asymptotically equivalent to $\mathcal{P}_n$.

3. An example shows that equivalence fails when $\alpha = 1/2$. Define $\epsilon_n(t) = \sqrt{t}$ on $[0, 1/(2n)]$ and then reflect it about $1/(2n)$ to extend to $[1/(2n), 1/n]$. Then extend ϵ_n by translation to each interval $[(l-1)/n, l/n]$ so as to obtain a sawtooth-like function on $[0,1]$ which is Hölder continuous with $\alpha = 1/2$, and for which $\sqrt{n}\int_0^1 \epsilon_n = \sqrt{2}/3$. Now consider estimation of the linear functional $Lf = \int_0^1 f(t)dt$. In problem $\mathcal{P}_n$, the normalized difference $\sqrt{n}(Y_n(1) - Lf) \sim N(0,\sigma^2)$ exactly for all f and n. However, in model $\mathcal{Q}_n$, the observation vector $\mathbf{Y} = (Y_l)$ has the same distribution whether $f = f_0 \equiv 0$ or $f = f_{1n} = \epsilon_n$, since $\epsilon_n(l/n) = 0$. Thus there can be no estimator $\delta_n(Y)$ in $\mathcal{Q}_n$ for which $\sqrt{n}(\delta_n(\mathbf{Y}) - Lf) \to N(0,1)$ in distribution uniformly over $f \in \Theta_H^{1/2}(1)$, since $\sqrt{n}Lf_0 = 0$ while $\sqrt{n}Lf_{1n} = \sqrt{2}/3$.

Some other examples

Density Estimation. Suppose that $X_1,\dots,X_n$ 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

$$dY(t) = \sqrt{f}(t)dt + \frac{1}{2\sqrt{n}}dW(t), \qquad 0 \le t \le 1. \tag{3.92}$$

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

Here is a heuristic argument, in the spirit of (1.26), that leads to (3.92). 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 kth 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

$X_1, \ldots, X_{N_n}$ of i.i.d. from f, with $N_n \sim \text{Poisson}(n)$. Using the Poisson thinning property, the number of observations, $N_n(I_{kn})$ say, falling in the kth bin I_{kn} will be Poisson with mean $n\int_{I_{kn}} f \approx nf(t_k)h_n$. The square root transformation is variance stabilizing for the Poisson family and so $Y_{kn} := \sqrt{N_n(I_{kn})} \sim N(\sqrt{f(t_k)nh_n}, 1/4)$ approximately for large n. Thus $Y_{kn} \approx \sqrt{f}(t_k)\sqrt{nh_n} + \frac{1}{2}Z_{kn}$ with Z_{kn} independent and approximately standard Gaussian. Now form a partial sum process as in (1.26), and premultiply by $\sqrt{h_n/n}$ to obtain

$$Y_n(t) = \sqrt{\frac{h_n}{n}} \sum_{k=1}^{[m_n t]} Y_{kn} \approx \frac{1}{m_n} \sum_{k=1}^{[m_n t]} \sqrt{f}(t_k) + \frac{1}{2\sqrt{n}} \frac{1}{\sqrt{m_n}} \sum_{k=1}^{[m_n t]} Z_{kn}.$$

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 (3.92).

A non-constructive proof of equivalence was given by Nussbaum (1996) under the assumption that f is $\alpha-$Hölder continuous for $\alpha > 1/2$, (3.90), and uniformly bounded below, $f(t) \geq \epsilon > 0$. A constructive argument was given by Brown et al. (2004) under a variety of smoothness conditions, including the Hölder condition with $\alpha > 1/2$. While the heuristic argument given above can be formalized for $\alpha > 1$, Brown et al. (2004) achieve $\alpha > 1/2$ via a conditional coupling argument that can be traced back to Komlós et al. (1975).

Nonparametric Generalized Linear Models. This is an extension of model (3.86) 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. Kolaczyk (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 θ 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. (Lehmann and Casella (1998, Ch. 1) or Brown (1986) have more background on exponential families). All the standard examples – Poisson, Bernoulli, Gaussian mean, Gaussian variance, exponential – are included. We will describe a form of the equivalence result in the mean value parameterization, given by $\mu(\theta) = \psi'(\theta) = E_\theta U(X)$. Let $t_l = l/n$, $l = 1, \ldots, n$ and g be a sufficiently smooth function, typically with Hölder smoothness greater than 1/2. Assume that we have observations (t_l, X_l) in which X_l is drawn from $P_{\theta_l}(dx)$ with $\mu_l = \mu(\theta_l) = g(t_l)$; call this model $\mathcal{P}_n$. [In the usual generalised linear model setting with canonical link function, one models $\theta = (\theta_l) = X\beta$ in terms of p predictors with coefficients $\beta_1, \ldots \beta_p$. If the predictor had the form of an expansion in (say) polynomials in t, so that $(X\beta)_l = \sum_k \beta_k p_k(t_l)$, then we would be replacing $\mu_l = \mu(\sum_k \beta_k p_k(t_l))$ by the nonparametric $g(t_l)$.]

Recall that $\psi''(\theta) = \text{Var}_\theta U(X)$, and let $V(\mu)$ be the *variance stabilizing* transformation for $\{P_\theta\}$ defined through $V'(\mu(\theta)) = 1/\sqrt{\psi''(\theta)}$. Then Grama and Nussbaum (1998) show that this experiment is asymptotically equivalent to

$$(\mathcal{Q}_n) \qquad dY(t) = V(g(t))dt + n^{-1/2} dW(t), \qquad 0 \leq t \leq 1,$$

in the sense that $\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0$. The Poisson case, with $V(\mu) = 2\sqrt{\mu}$, is closely related to the density estimation setting discussed earlier. For a second example, if X_l are independent $N(0, g(t_l))$, then we are in the Gaussian scale family and the corresponding exponential family form for $N(0, \sigma^2)$ has natural parameter $\theta = -1/\sigma^2$, mean parameter $\mu(\theta) =$

$-1/(2\theta)$ and variance stabilising transformation $V(\mu) = 2^{-1/2}\log\mu$. So the corresponding white noise problem has $dY(t) = 2^{-1/2}\log g(t) + n^{-1/2}dW(t)$, for $t \in [0,1]$.

Spectral density estimation. Suppose that $X^n = (X_1, \ldots, X_n)$ is a sample from a stationary Gaussian random process with mean zero and spectral density function $f(\xi)$ on $[-\pi, \pi]$, related to the covariance function $\gamma(k) = EX_jX_{j+k}$ via $f(\xi) = (2\pi)^{-1}\sum_{-\infty}^{\infty} e^{i\xi k}\gamma(k)$. Estimation of the spectral density f was the first nonparametric function estimation model to be studied asymptotically – see for example Grenander and Rosenblatt (1957).

Observe that $X^n \sim N(0, \Gamma_n(f))$ where the covariance matrix is Toeplitz: $\Gamma_n(f)_{jk} = \gamma(k-j)$. A classical approximation in time series analysis replaces the Toeplitz covariance matrix by a *circulant* matrix $\tilde{\Gamma}_n(f)$ in which the rows are successive shifts by one of a single periodic function on $\{0, 1, \ldots, n-1\}$. [12] The eigenvalues of a circulant matrix are given by the discrete Fourier transform of the top row, and so the eigenvalues of $\tilde{\Gamma}_n(f)$ are *approximately* $f(\xi_j)$ where ξ_j are equispaced points on $[-\pi, \pi]$. After an orthogonal transformation to diagonalize $\tilde{\Gamma}_n(f)$, one can say heuristically that the model $X^n \sim N(0, \Gamma_n(f))$ is approximately equivalent to

$$Z_j \sim N(0, f(\xi_j)), \qquad j = 1, \ldots, n.$$

This is the Gaussian scale model discussed earlier, and so one expects that both statistical problems will be asymptotically equivalent with

$$dZ(\xi) = \log f(\xi) + 2\pi^{1/2}n^{-1/2}dW(\xi), \qquad \xi \in [-\pi, \pi] \tag{3.93}$$

for f in a suitable function class, such as the Hölder function class $\Theta_H^\alpha(C)$ on $[-\pi, \pi]$ with $\alpha > 1/2$ and restricted also to bounded functions $\epsilon \le f(\xi) \le 1/\epsilon$. Full proofs are given in Golubev et al. (2010).

This by no means exhausts the list of examples where asymptotic equivalence has been established; one might add random design nonparametric regression and estimation in diffusion processes. For further references see the bibliography of Carter (2011).

Some cautions are in order when interpreting these results. First, there are significant regularity conditions, for example concerning the smoothness of the unknown f. Thus, Efromovich and Samarov (1996) have a counterexample for estimation of $\int f^2$ at very low smoothness. Meaningful error measures for spectral densities may not translate into, say, squared error loss in the Gaussian sequence model. See Cai and Zhou (2009a) for some progress with unbounded loss functions. 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.

3.12 Notes

§1. 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)

[12] Indeed, set $\tilde{\gamma}_n(l) = \gamma(l)$ for $0 \le l \le (n-1)/2$, make $\tilde{\gamma}_n$ periodic by reflection about $n/2$ and define $\tilde{\Gamma}_n(f)_{jk} = \tilde{\gamma}_n(k-j)$.

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. [Terminology from point-set topology here and below may be found in analysis texts, e.g. Folland (1999), or the appendix of Bogachev (1998)]. 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 $\mathbb{R}$. 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.

One can formally extend the infinitesimal representation (1.22) to a compact set $D \subset \mathbb{R}^n$ if $t \to W_t$ is $d-$parameter Brownian sheet (Hida, 1980). If φ_i is an orthonormal basis for $L_2(D)$, then the operations (1.24) again yield data in the for of model (3.1).

For more discussion (and citations) for kernel nonparametric regression estimators such as Nadaraya-Watson and Priestley-Chao, and discussion of the effect of boundaries in the nonperiodic case, we refer to books on nonparametric smoothing such as Wand and Jones (1995); Simonoff (1996).

There is some discussion of orthogonal series methods in Hart (1997), though the emphasis is on lack-of-fit tests. Eubank (1999) has a focus on spline smoothing.

Rice and Rosenblatt (1981) show that in the non-periodic case, the rate of convergence of the MSE is determined by the boundary behavior of f.

Cogburn and Davis (1974) derived the equivalent kernel corresponding to periodic smoothing splines for equally spaced data, see also Cox (1983).

We have not discussed the important question of data-determined choices of the regularization parameter λ in spline-like smoothers, in part because a different approach based on the James-Stein estimator is studied in Chapter 6. Some popular methods include C_p, (generalized) cross validation, and (generalized) maximum likelihood. Some entries into the literature looking at theoretical properties include Speckman (1985); Wahba (1985); Efron (2001).

There is a large literature on the matching of posterior and frequentist probabilities in parametric models - the Bernstein-von Mises phenomenon. The situation is more complicated for non-parametric models. Some simple examples are possible with Gaussian sequence models and Gaussian priors—Johnstone (2010) develops three examples to illustrate some possibilities.

We have given only a brief introduction to linear inverse problems in statistics, with a focus on the singular value decomposition. A broader perspective is given in the lecture notes by Cavalier (2011), including other ways of imposing smoothness such as through "source conditions". `Other books.`

L_2 boundedness of the fractional integration operator A_δ is a consequence of classical results of Hardy and Littlewood (1928), see also Gorenflo and Vessella (1991, pp. 64–67). Indeed, for $\delta \le 1/2$, the operator A_δ is bounded from $L_2[0,1]$ to $L_s[0,1]$ for a value $s = s(\delta) > 2$, while for $\delta > 1/2$, it is bounded from $L_2[0,1]$ to $C^{\delta-1/2}([0,1])$.

There is a large literature on the Wicksell problem – representative examples include Hall and Smith (1988), which introduces the transformation to squared radii, and Groeneboom and Jongbloed (1995), who study an isotonic estimation. See also Feller (1971, Ch. 3.11) and the lectures on inverse problems by Groeneboom (1996).

For upper and lower bounds on rates of convergence in statistical inverse problems see Koo (1993), `and...`

For more on the singular value decomposition of the Radon transform given in (v), see Johnstone and Silverman (1990).

Exercises

3.1 *(Compactness criteria.)* Here ℓ_2 denotes square summable sequences with the norm $\|\theta\|^2 = \sum \theta_i^2$.

(a) Suppose $a_k \ge 0$. The ellipsoid $\Theta = \{\theta : \sum_{k\ge1} a_k^2\theta_k^2 \le C^2\}$ is ℓ_2-compact if and only if $a_k > 0$ and $a_k \to \infty$.

(b) The hyperrectangle $\Theta = \prod_{k\ge1}[-\tau_k, \tau_k]$ is ℓ_2-compact if and only if $\sum_{k\ge1} \tau_k^2 < \infty$.

3.2 *(Extended compactness criterion.)* If $\Theta = \mathbb{R}^r \times \Theta'$, where $r < \infty$ and Θ' is compact in ℓ_2, show that for squared error loss, $R_N(\Theta, \epsilon) < \infty$.

3.3 *(Bounded parameter spaces.)* Use (3.14) to show that Lemma 3.2 remains true if Θ is assumed only to be norm bounded.

3.4 *(Affinity and L_1 distance for Gaussians.)* (a) Let ρ denote Hellinger affinity (3.61) and show

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

(b) Verify (3.65).

3.5 *(Equivalence for marginals?)* In the Gaussian sequence model $y_k = \theta_k + \epsilon z_k$, consider priors $\theta_k \sim N(0, \tau_k^2)$, independently with $\tau_k^2 = bk^{-2m}$. Under what conditions on m is the *marginal* distribution $P_\pi(dy)$ equivalent to $P_0(dy)$, the distribution conditional on $\theta = 0$?

3.6 *(Complex exponentials.)* Let $W(t)$ be a real-valued Brownian motion for $t \in [0, 1]$, and consider the complex exponentials $e_l(t) = e^{2\pi i l t}$, for $l \in \mathbb{Z}$. Let $g \in L_2[0, 1]$ be real valued. Let $z_l = \int_0^1 e_l dW$ and $g_l = \int_0^1 g e_l$. Show that $\mathrm{Var}(\sum g_l z_l) = \sum_{l\in\mathbb{Z}} |g_l|^2$ (even though $z_{-l} = \bar{z}_l$ and z_l contain dependencies).

3.7 *(Discrete orthogonality relations).* Let $\mathbf{e}_k$ denote the vector in $\mathbb{C}^n$ obtained by sampling the k-th complex exponential at $t_j = j/n$. Thus $\mathbf{e}_k = \{\exp(2\pi i k j/n), j = 0, 1, \ldots, n-1\}$. For $\mathbf{f}, \mathbf{g} \in \mathbb{C}^n$, use the usual inner product $\langle \mathbf{f}, \mathbf{g}\rangle_n = n^{-1} \sum_1^n f_k \bar{g}_k$. Show that for $k, l \in \mathbb{Z}$,

$$\langle \mathbf{e}_k, \mathbf{e}_l\rangle_n = \begin{cases} 1 & \text{if } k - l \in n\mathbb{Z} \\ 0 & \text{otherwise.} \end{cases}$$

Turn now to the real case. For $k \geq 0$, let $\mathbf{c}_k = \{\cos(2\pi k j/n), j = 0, 1, \ldots, n-1\}$ and define $\mathbf{s}_k$ analogously using the $k-$th sine frequency. If $n = 2m+1$ is odd, then take $\{\mathbf{c}_0, \mathbf{s}_1, \mathbf{c}_1, \ldots \mathbf{s}_m, \mathbf{c}_m\}$ as the basis B_n for $\mathbb{R}^n$. If $n = 2m + 2$ is even, then adjoin $\mathbf{c}_{n/2}$ to the previous set to form B_n. Show that the following orthogonality relations hold for basis vectors in B_n:

$$\langle \mathbf{c}_k, \mathbf{c}_l\rangle_n = \langle \mathbf{s}_k, \mathbf{s}_l\rangle_n = \frac{1}{2}\delta_{kl}, \qquad \langle \mathbf{c}_k, \mathbf{s}_l\rangle_n = 0,$$

with the exception of

$$\langle \mathbf{c}_0, \mathbf{c}_0\rangle_n = \langle \mathbf{c}_{n/2}, \mathbf{c}_{n/2}\rangle_n = 1,$$

where the last equation is only needed if n is even.
Hint. Derive the real relations from the complex by writing $\mathbf{e}_k = \mathbf{c}_k + i\mathbf{s}_k$ and using the complex orthogonality relations for pairs (k, l) and $(k, -l)$.

3.8 *(Infinite order kernels.)* Let $c \in (0, 1)$ and $h_c(\xi) = 1/(|\xi| - c)^2$ and show that the function $e^{h_0(\xi)} I\{\xi \geq 0\}$ is C^∞. Define

$$\widehat{K}(\xi) = \begin{cases} 1 & \text{if } |\xi| \leq c \\ \exp\{-bh_1(\xi)\exp(-bh_c(\xi))\} & \text{if } c \leq |\xi| \leq 1 \\ 0 & \text{if } |\xi| \geq 1 \end{cases}$$

and show that $K(s) = (2\pi)^{-1} \int e^{is\xi} \widehat{K}(\xi) d\xi$ is a C^∞ kernel of infinite order (i.e. satisfies (3.28) with $q = \infty$) that decays faster than $|s|^{-m}$ for any $m > 0$. (McMurry and Politis, 2004)

3.9 *(Aliasing example.)* Consider equispaced model (3.37) with $n = 5$, and as in Section 3.4, let S_5 be the linear span of trigonometric polynomials of degree $n_d = 2$. Let f° minimize $Q(f)$ given below (3.38), and let $f_\epsilon^\circ = f^\circ + \epsilon(\varphi_{11} - \varphi_1)$. Show, under an appropriate condition, that $Q(f_\epsilon^\circ) < Q(f^\circ)$ for ϵ small. Hence the minimum of $Q(f)$ does not lie in S_5.

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

$$\frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{e^{i\gamma x}}{x-\alpha}dx = \begin{cases} e^{i\alpha\gamma} & \text{if } \gamma > 0 \\ 0 & \text{if } \gamma < 0. \end{cases}$$

Use the partial fraction expansion

$$\prod_{k=1}^{r}(x-\beta_k)^{-1} = \sum_{k=1}^{r} c_k(x-\beta_k)^{-1}, \qquad 1/c_k = \prod_{j\neq k}(\beta_k - \beta_j),$$

to compute the equivalent kernel $L(t)$ given that $\hat{L}(\xi) = (1+\xi^4)^{-1}$.

3.11 *(Wahba's prior for cubic splines.)* Show that

$$Z_2^\sigma(t) = \sigma\xi_1 + \sigma\xi_2 t + \int_0^t (t-u)dW(u),$$

the integrated (free) Wiener process, has covariance function

$$R_\sigma(s,t) = \sigma^2(1+st) + R_0(s,t),$$
$$R_0(s,t) = \begin{cases} \frac{1}{2}s^2t - \frac{1}{6}s^3 & 0 \le s \le t \\ \frac{1}{2}st^2 - \frac{1}{6}t^3 & 0 \le t \le s. \end{cases}$$

By differentiating the eigenvalue equation

$$\int_0^1 R_\sigma(s,t)\varphi(t)dt = \varrho^2\varphi(s)$$

four times, show that φ satisfies

$$\varphi(s) = \varrho^2\varphi^{(4)}(s),$$

with boundary conditions

$$\varphi''(0) = \sigma^{-2}\varphi'(0), \varphi'''(0) = \sigma^{-2}\varphi(0) \qquad \varphi''(1) = \varphi'''(1) = 0.$$

With $\sigma = 0$, show that the boundary conditions imply the equation $\cos\varrho^{-1/2}\cosh\varrho^{-1/2} = -1$ for the eigenvalues. In the $\sigma = \infty$ limit, show that the corresponding equation is $\cos\varrho^{-1/2}\cosh\varrho^{-1/2} = 1$. In either case, show that the eigenvalues satisfy, for large n

$$\varrho_n \sim \frac{1}{(n+\frac{1}{2})^2\pi^2} \sim \frac{1}{n^2\pi^2}.$$

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

3.12 *(Splines dominate truncation estimators.)*
(a) Let $H(r) = -r\log r - (1-r)\log(1-r)$ be the binary entropy function and verify that the coefficient b_α in the truncation maximum MSE (3.18) satisfies $b_\alpha = e^{H(r)}$, $r = 2\alpha/(2\alpha+1)$.
(b) Conclude for $m > 1/2$ and $0 \le \alpha \le 2m$ that the spline maximum MSE (3.57) is asymptotically no larger than (3.18).

3.13 *(Minimax rates of convergence for kernel estimators.)*
Suppose that K is a symmetric kernel of compact support and order q. As in Section 3.3 and Lemma 3.7, let $\hat{f}_h(t) = \int_0^1 \mathring{K}_h(s-t)Y(dt)$ be a periodized kernel estimator with (sine-cosine basis) sequence form $\hat{\theta}_h$.
(a) Let $v_K = \int K^2$ and $b_{\alpha,K} = \pi^{2\alpha} \sup_\xi [1-\hat{K}(\xi)]^2/\xi^{2\alpha}$. Show that $b_{\alpha,K} < \infty$ iff $\alpha \le q$.
(b) Show that if $\alpha \le q$ and $h = h(\epsilon) \to 0$ as $\epsilon \to 0$ then in the Gaussian white noise model, the worst case mean squared error over α-smoothness ellipsoids satsifies

$$\bar{r}(\hat{f}_h, \epsilon) = \sup_{\theta \in \Theta_2^\alpha(C)} r(\hat{\theta}_h, \theta; \epsilon) \doteq v_K \epsilon^2 h^{-1} + b_{\alpha,K} C^2 h^{2\alpha}.$$

(c) Let the right side of the previous display be $r_a(h;\epsilon)$. Show that for $r = 2\alpha/(2\alpha+1)$ and $\alpha \le q$ that

$$\inf_h r_a(h;\epsilon) = c_{\alpha,K} C^{2(1-r)} \epsilon^{2r},$$

with $c_{\alpha,K} = e^{H(r)} b_{\alpha,K}^{1-r} v_K^r$. This indicates that kernel estimators (of sufficiently high order) also attain the "optimal" rate of convergence corresponding to α mean-square derivatives.

3.14 *(Local polynomial regression and its equivalent kernel.)* Consider the finite equispaced regression model (3.37) for periodic f, with data extended periodically as in (3.23). Let K be a kernel of compact support and let $\hat{f}_{p,h}(t)$ be the local polynomial estimator of degree p defined at (3.59).
(a) Show that $\hat{\beta}$ can be written in the weighted least squares form

$$\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}.$$

(b) Let the moments of the kernel $\mu_k = \int v^k K(v) dv$. Define the moment matrix $S = (S_{jk})_{j,k=0,\dots,p}$ by $S_{jk} = \mu_{j+k}$, and write $S^{-1} = (S^{jk})$. Show that the local polynomial estimator, for large n, has the approximate form

$$\hat{f}(t) = n^{-1} \sum_{l \in \mathbb{Z}} K_h^*(t_l - t) Y_l,$$

where the *equivalent kernel* $K_h^*(t) = K^*(t/h)/h$ and

$$K^*(t) = \Big(\sum_{k=0}^{p} S^{0k} t^k\Big) K(t).$$

(c) Show that the kernel K^* satisfies

$$\int v^r K^*(v) dv = \delta_{0r}, \qquad 0 \le r \le p.$$

(d) Suppose that K is symmetric, $K(-t) = K(t)$, and evaluate K^* in the case $p = 2$.

3.15 *(Computational comparison.)* Consider two functions on $[0,1]$:

$$f_1(t) = \sin 4\pi t^2, \qquad f_2(t) = (e^{4t} - 1 - t)(1-t)^2,$$

and consider the model

$$Y_i = f(i/n) + \sigma z_i, \qquad z = 1, \dots, n,$$

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

$$\min Q(f) = n^{-1}\sum [Y_i - f(i/n)]^2 + \lambda \int_0^1 f''^2$$

among cubic splines and trignometric polynomials respectively. Note that $\hat{f}_{SS,\lambda}$ can be computed in `R` 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 λ, use the value suggested by the ellipsoid considerations in class:

$$\lambda = (\pi/2)^4 (6\sqrt{2})^{4/5} (n \int f''^2)^{-4/5}.$$

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

$$\frac{\text{ave} \| \hat{f}_{SS} - \hat{f}_{PER} \|_2^2}{\text{ave} \| \hat{f}_{SS} - f \|_2^2}.$$

3.16 *(Perfect classification.)* Consider the two class classification problem in which y is observed in the heteroscedastic sequence model (3.60) and it is desired to decide whether $\theta = \theta^0$ or $\theta = \theta^1$ obtains. Then consider a loss function $L(a, \theta) = I\{a \neq \theta\}$ with $a \in \{\theta^0, \theta^1\}$ and a prior distribution putting mass π_0 on θ^0 and $\pi_1 = 1 - \pi_0$ on θ^1.
(a) Let $\mu = \theta^0 - \theta^1$ and show that the optimal classifier (i.e. the Bayes rule in the above set up) is the Fisher linear discriminant, using $T(y) = \langle y - \theta^1, \mu \rangle_Q - \|\mu\|_Q^2/2$.
(b) Show that perfect classification–i.e. incurring zero error probabilities–occurs if and only if $D^2 = \sum \mu_i^2/(\varrho_i \epsilon)^2 = \infty$. (Modified from (Delaigle and Hall, 2011).)

3.17 *(Maximum risk of the Pinsker estimator.)* Consider a slightly different family of shrinkage rules, to appear in Pinsker's theorem, and also indexed by a positive parameter:

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

Show that the maximum risk over a Sobolev ellipsoid $\Theta_2^\alpha(C)$ is approximated by

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

where

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

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

$$\begin{aligned} \mu_* &\sim (2mC^2/\bar{v}_m \epsilon^2)^{r/2} \\ \bar{r}(\hat{\theta}_{\mu_*}; \epsilon) &\sim e^{H(r)} C^{2-2r} (\bar{v}_m \epsilon^2)^r. \end{aligned} \tag{3.94}$$

3.18 *(SVD for fractional integration.)* Let A_δ be the operator of fractional order integration (3.74). This exercise outlines the derivation of the singular value decomposition for a class of domain spaces, based on identites for Gauss' hypergeometric function and Jacobi polynomials that are recalled in Appendix C.30. Let $\rho_n(a, \delta) = \Gamma(a+n+1)/\Gamma(a+\delta+n+1) \sim n^{-\delta}$ as $n \to \infty$.
(a) Interpret identities (C.34) and (C.35) in terms of the operator A_δ and Jacobi polynomials:

$$A_\delta [w^a P_n^{a,b}(1-2w)](x) = \rho_n(a,\delta) x^{a+\delta} P_n^{a+\delta, b-\delta}(1-2x).$$

(b) Let $g_{a,b;n}$ denote the normalizing constants for Jacobi polynomials in (C.36); show that

$$\varphi_{a,b;n}(x) := g_{a,b;n}^{-1} x^a P_n^{a,b}(1-2x)$$

are orthonormal in $H^2_{-a,b} := L_2\big([0,1], x^{-a}(1-x)^b dx\big)$.

(c) Verify that the singular value decomposition of $A_\delta : H^2_{-a,b} \to H^2_{-a-\delta,b-\delta}$ is given by

$$\varphi_n = \varphi_{a,b;n}, \quad \psi_n = \varphi_{a+\delta,b-\delta;n}, \quad b_n^2 = \rho_n(a,\delta)\rho_n(b-\delta,\delta) \sim n^{-2\delta}, \quad n \to \infty.$$

(d) Set $a = 0$ and $b = 0$ to recover the SVD of A_δ as given in (3.75).

(e) Set $a = 0, \delta = 1/2$ and use the formula (Szegö, 1967, (4.1.8))

$$P_n^{1/2,-1/2}(x) = \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n} \frac{\sin((2n+1)\theta/2)}{\sin(\theta/2)}, \quad x = \cos\theta$$

to recover the SVD of $A_{1/2}$ as given in Section 3.9 part (iii).

3.19 *(Sufficiency part of the Brown-Low convergence theorem.)* Provide the details in the claim that $S_n = S_n(\bar{Y}_n)$ is sufficient for $f \in \Theta$ in problem $\overline{\mathcal{P}}_n$. Specifically, if $\{t_i,\ i = 1,\dots,r\} \subset [0,1]$ is a finite set, and $S_n = (S_{n,l},\ l = 1,\dots,n)$ with $S_{n,l} = n[\bar{Y}_n(l/n) - \bar{Y}_n((l-1)/n)]$, then show that $\mathcal{L}(\{\bar{Y}_n(t_i)\}|S_n)$ is free of f.

3.20 *(Rates of convergence in a severely ill-posed problem.)* Assume model (3.60) with $\varrho_k = e^{\gamma k^2}$. [In the case of the heat equation (3.76) $\gamma = \pi^2 T$.] Let $\hat{\theta}_\nu$ be the truncation estimator (3.17) and choose ν_* to approximately minimize $\bar{r}(\hat{\theta}_\nu) = \sup_{\theta \in \Theta_2^\alpha(C)} r(\hat{\theta}_\nu, \theta)$. Show that, as $\epsilon \to 0$,

$$R_N(\Theta_2^\alpha(C), \epsilon) \le \bar{r}(\hat{\theta}_{\nu_*}) \le c_{\alpha,\rho} C^2 [\log(C/\epsilon)]^{-\alpha}(1 + o(1)).$$

3.21 *(Transformations of white noise model.)* Show how to transform a model

$$dY(t) = \gamma f(t)dt + \epsilon dW(t) \qquad 0 \le t \le 1$$

into one of the form

$$d\tilde{Y}(s) = \tilde{f}(s)ds + c\epsilon d\tilde{W}(s) \qquad a \le s \le b,$$

where $\tilde{W}(s)$ is again a standard Brownian motion, and evaluate c. [This transformation connects (3.93) with (3.92) for $t \in [0,1]$.

4

Gaussian decision theory

> In addition to those functions studied there are an infinity of others, and unless some principle of selection is introduced we have nothing to look forward to but an infinity of test criteria and an infinity of papers in which they are described. (G. E. P. Box, discussion in *J. R. S. S. B., 1956*)

In earlier chapters we have formulated the Gaussian sequence model and indicated our interest in comparisons of estimators through their maximum risks, typically mean squared error, over appropriate parameter spaces. It is now time to look more systematically at questions of optimality.

Many powerful tools and theorems relevant to our purpose have been developed in classical statistical decision theory, often in far more general settings than used here. This chapter introduces some of these ideas, tailored for our needs. We focus on comparison of properties of estimators rather than the explicit taking of decisions, so that the name "decision theory" is here of mostly historical significance.

Our principle of selection—comparison, really—is minimaxity: look for estimators whose worst case risk is (close to) as small as possible for the given parameter space, often taken to encode some relevant prior information. This principle is open to the frequent and sometimes legitimate criticism that the worst case may be an irrelevant case. However, we aim to show that by appropriate choice of parameter space, and especially of *families* of parameter spaces, that sensible estimators emerge both blessed and enlightened from examination under the magnifying glass of the minimax prinicple.

A minimax estimator is exactly or approximately a Bayes estimator for a suitable "least favorable" prior. [1] It is then perhaps not surprising that the properties of Bayes rules and risks play a central role in the study of minimaxity. Section 4.1 begins therefore with Bayes estimators, now from a more frequentist viewpoint than in Chapter 2. Section 4.2 goes more deeply than Chapter 2 into some of the elegant properties and representations that appear for squared error loss in the Gaussian model.

The heart of the chapter lies in the development of tools for evaluating, or approximating $R_N(\Theta)$, the minimax risk when the parameter is assumed to belong to Θ. Elementary lower bounds to minimax risk can often be derived from Bayes rules for priors supported on the parameter space, as discussed in Section 4.3. For upper bounds and actual evaluation of the minimax risk, the minimax theorem is crucial. This is stated in Section 4.4, but an overview of its proof, even in this Gaussian setting, must be deferred to Appendix A.

[1] For a more precise statement, see Proposition 4.9.

Statistical independence and product structure of parameter spaces plays a vital role in "lifting" minimax results from simpler component spaces to their products, as shown in Section 4.5.

A theme of this book is that conclusions about function estimation can sometimes be built up from very simple, even one dimensional, parametric constituents. As an extended example of the techniques introduced, we will see this idea at work in Sections 4.6 - 4.8. We start with minimaxity on a bounded interval in a single dimension and progress through hyperrectangles—products of intervals—to ellipsoids and more complex quadratically convex sets in $\ell_2(\mathbb{N})$.

Byproducts include conclusions on optimal (minimax) rates of convergence on Hölder, or uniform, smoothness classes, and the near mean square optimality of linear estimators over all quadratically convex sets.

With notation and terminology established and some examples already discussed, Section 4.10 gives an overview of the various methods used for obtaining lower bounds to minimax risks throughout the book.

A final Section 4.11 outlines a method for the exact asymptotic evaluation of minimax risks using classes of priors with appropriately simple structure. While this material is used on several later occasions, it can be omitted on first reading.

4.1 Bayes Estimators

The setting for this chapter is the heteroscedastic Gaussian sequence model

$$y_i = \theta_i + \epsilon \varrho_i z_i \tag{4.1}$$

for $i \in I \subset \mathbb{N}$, with z_i i.i.d. $N(0,1)$ and ϵ and ϱ_i known positive constants. The parameter space is the collection of θ for which $\sum \theta_i^2/\varrho_i^2 < \infty$, denoted $\ell_2(\mathbb{N}, (\varrho_i^{-2}))$, as explained in Section 3.8. Of course, many of our initial remarks about decision theoretic definitions hold for more general statistical models $\{P_\theta, \theta \in \Theta\}$.

In Section 2.3 we approached Bayes rules via calculations with the posterior distribution, for example using the posterior mean for squared error loss. In this chapter we largely adopt a different, though mathematically equivalent, approach, which considers instead the average of (frequentist) risk functions with respect to a prior distribution. Thus, if π is a probability distribution on $\ell_2(I)$, the *integrated risk* of an estimator $\hat{\theta}$ is defined by

$$\begin{aligned} B(\hat{\theta}, \pi) &= \int r(\hat{\theta}, \theta) \pi(d\theta) \\ &= E_\pi r(\hat{\theta}, \theta) = E_\pi E_\theta L(\hat{\theta}(y), \theta). \end{aligned} \tag{4.2}$$

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

$$B(\pi) = \inf_{\hat{\theta}} B(\hat{\theta}, \pi). \tag{4.3}$$

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

Remark 4.1 One reason 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.* This is useful for the minimax theorem, Appendix A. Representation (4.3) also shows that the Bayes risk $B(\pi)$ is a concave function of π, which helps in studying least favorable distributions (e.g. Proposition 4.14).

The decidedly frequentist definition of Bayes estimators fortunately agrees with the Bayesian definition given at (2.9), under mild regularity conditions (see the Chapter Notes). We saw that the joint distribution $\mathbb{P}$ of the pair (θ, y) may be decomposed two ways:

$$\mathbb{P}(d\theta, dy) = \pi(d\theta)P(dy|\theta) = P_\pi(dy)\pi(d\theta|y),$$

where $P_\pi(dy)$ is the marginal distribution of y and $\pi(d\theta|y)$ is the posterior distribution of θ given y. The integrated risk of (4.2), 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_π} 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 indeed obtained by minimizing the posterior expected loss (2.9), $\hat{\theta}_\pi(y) = \operatorname{argmin}_a E_y L(a, \theta)$.

As seen in Chapter 2.3, 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, $\hat{\theta}_\pi(y) = E_\pi(\theta|y)$.

If the loss function $L(a, \theta)$ is strictly convex in a, then the Bayes estimator $\hat{\theta}_\pi$ is unique (a.e. P_θ for each θ) if both $B(\pi) < \infty$, and $P_\pi(A) = 0$ implies $P_\theta(A) = 0$ for each θ. (Lehmann and Casella, 1998, Corollary 4.1.4)

Remark 4.2 *Smoothness of risk functions.* We digress briefly to record for later use some information about the smoothness of the risk functions of general estimators $\hat{\theta}$. For $y \sim N_n(\theta, \epsilon^2\Lambda)$, with Λ diagonal and quadratic loss, the risk function $\theta \to r(\hat{\theta}, \theta)$ is analytic, i.e. has a convergent power series expansion, on the interior of the set on which it is finite. This follows, for example, from Lehmann and Romano (2005, Theorem 2.7.1), since $r(\hat{\theta}, \theta) = \int \|\hat{\theta}(y) - \theta\|^2 \phi_\epsilon(\Lambda^{-1/2}(y - \theta))dy$ can be expressed in terms of Laplace transforms.

Example. Univariate Gaussian. We revisit some earlier calculations to illustrate the two perspectives on Bayes risk. If $y|\theta \sim N(\theta, \epsilon^2)$ and the prior $\pi(d\theta)$ sets $\theta \sim N(0, \tau^2)$ then the posterior distribution $\pi(d\theta|y)$ was found in Section 2.3 to be Gaussian with mean $\hat{\theta}_\pi(y) = \tau^2 y/(\tau^2 + \epsilon^2)$, which is *linear* in y, and constant posterior variance $\tau^2/(\tau^2 + \epsilon^2)$. Turning now to the frequentist perspective,

$$B(\pi_\tau) = \inf_{\hat{\theta}} B(\hat{\theta}, \pi_\tau).$$

Suppose that, for the moment arbitrarily, we restrict the minimization to linear estimators $\hat{\theta}_c(y) = cy$. Formula (2.47) showed that the risk of $\hat{\theta}_c$

$$r(\hat{\theta}_c, \theta) = c^2\epsilon^2 + (1 - c)^2\theta^2$$

so that the integrated risk

$$B(\hat{\theta}_c, \pi_\tau) = c^2\epsilon^2 + (1-c)^2\tau^2.$$

Minimizing this over c yields the linear minimax choice $c_{\mathrm{LIN}} = \tau^2/(\tau^2+\epsilon^2)$ and value $B(\pi_\tau) = \tau^2\epsilon^2/(\tau^2+\epsilon^2)$, which agrees with the result of the posterior calculation.

Remark 4.3 If $y|\theta \sim N(\theta, \epsilon^2)$, then the univariate MLE $\hat{\theta}_1(y) = y$ is admissible for squared error loss. As promised in Theorem 2.5, we indicate a proof: the result is also used at Corollary 4.10 below. It suffices to take $\epsilon = 1$. The argument is by contradiction: supposing $\hat{\theta}_1$ to be inadmissible, we can find a dominating estimator $\tilde{\theta}$ and a parameter value θ_0 so that $r(\tilde{\theta}, \theta) \le 1$ for all θ, with $r(\tilde{\theta}, \theta_0) < 1$. The risk function $r(\tilde{\theta}, \theta)$ is continuous by Remark 4.2, so there would exist $\delta > 0$ and an interval I of length $L > 0$ containing θ_0 for which $r(\tilde{\theta}, \theta) \le 1-\delta$ when $\theta \in I$. Now bring in the conjugate priors π_τ. From the example above, $1 - B(\pi_\tau) \sim \tau^{-2}$ as $\tau \to \infty$. However, the definition (4.2) of integrated risk implies that

$$1 - B(\tilde{\theta}, \pi_\tau) \ge \delta\pi_\tau(I) \sim c_0\delta\tau^{-1}$$

as $\tau \to \infty$, with $c_0 = L/\sqrt{2\pi}$. Consequently, for τ large, we must have $B(\tilde{\theta}, \pi_\tau) < B(\pi_\tau)$, contradicting the very definition of the Bayes risk $B(\pi_\tau)$. Hence $\hat{\theta}_1$ must be admissible.

4.2 Univariate Bayes rules for squared error loss

A number of formulas for Bayes estimators take especially convenient, even elegant, forms when squared error loss is used. We concentrate on the univariate setting, $n = 1$, needed for our applications. However some of the results hold in higher dimensions as well.

We begin with a simple quadratic identity relating a prior π, its marginal density p, defined at (2.8), and the corresponding Bayes estimator $\hat{\theta}_\pi$.

Proposition 4.4 (Brown) *Suppose that $y \sim N(\theta, \epsilon^2)$ and that $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. For any estimator $\hat{\theta}$ and prior distribution $\pi(d\theta)$,*

$$B(\hat{\theta}, \pi) - B(\pi) = \int (\hat{\theta} - \hat{\theta}_\pi)^2 p. \tag{4.4}$$

Proof Write $\mathbb{E}$ for expectation with respect to the joint distribution of (θ, y) when $\theta \sim \pi$. The left side above can be rewritten

$$\mathbb{E}(\hat{\theta} - \theta)^2 - \mathbb{E}(\hat{\theta}_\pi - \theta)^2 = \mathbb{E}(\hat{\theta} - \hat{\theta}_\pi)(\hat{\theta} + \hat{\theta}_\pi - 2\theta).$$

As we have seen, with squared error loss the Bayes estimator is given by the posterior mean $\hat{\theta}_\pi(y) = \mathbb{E}(\theta|y)$ and so by conditioning on y,

$$\mathbb{E}(\hat{\theta} - \hat{\theta}_\pi)\theta = \mathbb{E}(\hat{\theta} - \hat{\theta}_\pi)\hat{\theta}_\pi.$$

Substitute this into the previous display and (4.4) falls out. □

We apply Brown's identity and some facts about Fisher information, reviewed here and in Appendix C.20 to obtain some useful bounds on Bayes risks. If P is a probability measure on $\mathbb{R}$ with absolutely continuous density $p(y)dy$, the Fisher information is defined by

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

This agrees with the definition of Fisher information for parametric families when $p(y;\theta) = p(y-\theta)$ is a location family. If $P_\tau(dy) = p(y/\tau)dy/\tau$ is a scaled version of p, then it is immediate that $I(P_\tau) = I(P_1)/\tau^2$.

Fisher information is bounded below by precision: for any distribution P,

$$I(P) \geq 1/\text{Var } P. \tag{4.5}$$

with equality if and only if P is Gaussian. [For a location family, this is the Cramér-Rao bound.] The proof is just the Cauchy-Schwarz inequality. Indeed, we may 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 $(p'/p)(y) = (\log p)'(y) = c(y-\mu)$, so that p is Gaussian.

Now we return to Brown's identity. We also need the Tweedie/Brown formula (2.23) for a Bayes estimator, which for noise level ϵ and dimension $n = 1$ takes the form

$$\hat{\theta}_\pi(y) = y + \epsilon^2 p'(y)/p(y). \tag{4.6}$$

Recalling the unbiased estimator $\hat{\theta}_0(y) = y$, we might write this as $\hat{\theta}_0 - \hat{\theta}_\pi = -\epsilon^2(p'/p)(y)$. Of course, $B(\hat{\theta}_0, \pi) = E_\pi E_\theta (y-\theta)^2 = \epsilon^2$, regardless of the prior π. If now in (4.4), we insert $\hat{\theta}_0$ for $\hat{\theta}$, we have

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

Since p is the absolutely continuous density of the marginal distribution $\pi \star \Phi_\epsilon$, we arrive at a formula that is also sometimes called Brown's identity:

Proposition 4.5 (Brown) *For $y \sim N(\theta, \epsilon^2)$ and squared error loss,*

$$B(\pi, \epsilon) = \epsilon^2[1 - \epsilon^2 I(\pi \star \Phi_\epsilon)]. \tag{4.7}$$

Inserting the information bound (4.5) in the previous display, we arrive at

Corollary 4.6

$$B(\pi, \epsilon) \leq \frac{\epsilon^2 \text{Var } \pi}{\epsilon^2 + \text{Var } \pi}, \tag{4.8}$$

with equality if and only if π is Gaussian.

Proof Indeed, convolution of probability measures corresponds to addition of independent random variables, hence $\text{Var}(\pi \star \Phi_\epsilon) = \text{Var } \pi + \epsilon^2$, and so (4.5) yields the inequality. Also from (4.5), equality occurs only if the convolution $\pi \star \Phi_\epsilon$ is Gaussian, which implies, for example using characteristic functions (C.10) that π itself is Gaussian. □

Finally, we give a matching lower bound for $B(\pi, \epsilon)$ that is sometimes easier to use than (4.7). It is a version of the van Trees inequality (Van Trees, 1968) (see Exercise 4.2)

$$B(\pi, \epsilon) \geq \epsilon^2/(1 + \epsilon^2 I(\pi)). \tag{4.9}$$

We give two further applications of Brown's identity: to studying continuity and (directional) derivatives of Bayes risks.

Continuity of Bayes risks. This turns out to be a helpful property in studying Bayes minimax risks, e.g. in Section 8.7.

Lemma 4.7 *If π_n converges weakly to π, then $B(\pi_n) \to B(\pi)$.*

Note that definition (4.3) itself implies only upper semicontinuity for $B(\pi)$.

Proof It suffices to consider unit noise $\epsilon = 1$. Let $p_n(y) = \int \phi(y-\theta)\, d\pi_n$ and define $p(y)$ correspondingly from π. From (4.7), it is enough to show that

$$I(\pi_n \star \Phi) = \int \frac{p_n'^2}{p_n} \to \int \frac{p'^2}{p} = I(\pi \star \Phi). \tag{4.10}$$

Weak convergence says that $\int g\, d\pi_n \to \int g\, d\pi$ for every g bounded and continuous, and so p_n, p_n' and hence $p_n'^2/p_n$ converge respectively to p, p' and p'^2/p pointwise in $\mathbb{R}$. We construct functions G_n and G such that

$$0 \le \frac{p_n'^2}{p_n} \le G_n, \qquad 0 \le \frac{p'^2}{p} \le G,$$

and $\int G_n \to \int G$, and use the extended version of the dominated convergence theorem, Theorem C.7, to conclude (4.10). Indeed, from representation (4.6)

$$\frac{p_n'}{p_n}(y) = \hat{\theta}_{\pi_n}(y) - y = E_{\pi_n}[\theta - y|y],$$

and so $(p_n'/p_n)^2 \le E_{\pi_n}[(\theta-y)^2|y]$, or equivalently

$$\left(\frac{p_n'}{p_n}\right)^2(y) \le G_n(y) := \int (\theta-y)^2 \phi(y-\theta)\, \pi_n(d\theta).$$

A corresponding bound holds with π_n and p_n replaced by π and p and yields a bounding function $G(y)$. To complete the verification, note also that

$$\int G_n(y)\, dy = \iint (y-\theta)^2 \phi(y-\theta)\, dy\, \pi_n(d\theta) = 1 = \int G(y) dy. \qquad \square$$

Remark. The smoothing effect of the Gaussian density is the key to the convergence (4.10). Indeed, in general Fisher information is only lower semicontinuous: $I(\pi) \le \liminf I(\pi_n)$, see also Appendix C.20. For a simple example in which continuity fails, take discrete measures π_n converging weakly to Φ, so that $I(\pi_n)$ is infinite for all n.

Derivatives of Bayes Risk. Brown's identity also leads to an interesting formula for the directional or Gateaux derivative for the Bayes risk. We use it later, Proposition 4.13, to exhibit saddle points.

Lemma 4.8 *For priors π_0 and π_1, let $\pi_t = (1-t)\pi_0 + t\pi_1$ for $t \in [0,1]$. For all $\epsilon > 0$,*

$$\frac{d}{dt} B(\pi_t)|_{t=0} = B(\hat{\theta}_{\pi_0}, \pi_1) - B(\pi_0). \tag{4.11}$$

Formula (4.11), which involves a "change of prior", should be compared with (4.4), which involves a "change of estimator", from $\hat{\theta}$ to $\hat{\theta}_\pi$.

Proof Write $P_t = \Phi_\epsilon \star \pi_t$ for the marginal distributions. Since $I(P_t) < \infty$, the density $p_t(y)$ of P_t exists, along with its derivative $p_t' = (d/dy)p_t$ a.e. Introduce

$$\psi_0(y) = -(p_0'/p_0)(y) = [y - \hat{\theta}_{\pi_0}(y)]/\epsilon^2,$$

where the final equality uses the Bayes estimator representation (4.6). From Brown's identity (4.7), $(d/dt)B(\pi_t) = -\epsilon^4(d/dt)I(P_t)$. Differentiating $I(P_t) = \int p_t'^2/p_t$ under the integral sign, we obtain (see Appendix C.20 for details)

$$-\epsilon^4 \frac{d}{dt} I(P_t)|_{t=0} = \epsilon^4 \int [2\psi_0 p_1' + \psi_0^2 p_1]dy + \epsilon^4 I(P_0). \tag{4.12}$$

Since $p_1 = \phi_\epsilon \star \pi_1$ is the marginal density of π_1 and $\epsilon^2 p_1'(y) = \int -(y-\theta)\phi_\epsilon(y-\theta)\pi_1(d\theta)$, we can write the previous integral as

$$\begin{aligned}\iint \big[-2(y - \hat{\theta}_{\pi_0})(y-\theta) + (y - \hat{\theta}_{\pi_0})^2\big]\phi_\epsilon(y-\theta)\pi_1(d\theta)dy \\ = -\epsilon^2 + E_{\pi_1} E_\theta(\theta - \hat{\theta}_{\pi_0})^2 = -\epsilon^2 + B(\hat{\theta}_{\pi_0}, \pi_1).\end{aligned}$$

Since $B(\pi_0) = \epsilon^2 - \epsilon^4 I(P_0)$ by Brown's identity (4.7), we arrive at formula (4.11). □

4.3 A lower bound for minimax risk

We now take up the study of minimax estimators and the minimax theorem. Curiously, although minimaxity is a frequentist notion, in some sense at an opposite extreme from Bayesian estimation, its study is heavily dependent, at a technical level, on Bayesian calculations. Return to the general sequence model (4.1). Recall from Section 3.1 the definition of the minimax risk over parameter set Θ:

$$R_N(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} r(\hat{\theta}, \theta).$$

We begin with an elementary, but very useful, lower bound for $R_N(\Theta)$ that may be derived using Bayes risks of priors supported in Θ. Indeed, if $\operatorname{supp} \pi \subset \Theta$, then

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

We sometimes write $\bar{r}(\hat{\theta})$ for $\sup_{\theta\in\Theta} r(\hat{\theta}, \theta)$. Minimizing over $\hat{\theta}$, we have

$$B(\pi) \le \inf_{\hat{\theta}} \sup_{\Theta} r(\hat{\theta}, \theta) = R_N(\Theta). \tag{4.13}$$

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

$$B(\mathcal{P}) = \sup_{\pi \in \mathcal{P}} B(\pi). \tag{4.14}$$

A prior that attains the supremum will be called *least favorable*. A sequence of priors for which $B(\pi_n)$ approaches the supremum is called a least favorable sequence. Letting $\operatorname{supp}\mathcal{P}$ denote the union of all $\operatorname{supp}\pi$ for π in $\mathcal{P}$, we obtain the lower bound

$$\operatorname{supp}\mathcal{P} \subset \Theta \quad \Longrightarrow \quad R_N(\Theta) \geq B(\mathcal{P}). \tag{4.15}$$

Implicit in these remarks is a classical sufficient condition for minimaxity.

Proposition 4.9 *An estimator $\hat{\theta}_0$ is minimax if there exists a sequence of priors π_n with $B(\pi_n) \to \bar{r} = \sup_\theta r(\hat{\theta}_0, \theta)$.*

Proof Indeed, from (4.13) we have $\bar{r} \leq R_N(\Theta)$, which says that $\hat{\theta}_0$ is minimax. □

Corollary 4.10 *If $y|\theta \sim N(\theta, \epsilon^2)$, then $\hat{\theta}_1(y) = y$ is minimax for squared error loss. In addition, $\hat{\theta}_1$ is the unique minimax estimator.*

Proof Indeed, using the conjugate priors π_τ, we have $\bar{r}(\hat{\theta}_1) = \epsilon^2 = \lim_{\tau\to\infty} B(\pi_\tau)$. To establish uniqueness, suppose that $\hat{\theta}_1'$ is another minimax estimator with $P_\theta(\hat{\theta}_1 \neq \hat{\theta}_1') > 0$ for some and hence every θ. Then strict convexity of the loss function implies that the new estimator $\tilde{\theta} = (\hat{\theta}_1 + \hat{\theta}_1')/2$ satisfies, for all θ, $r(\tilde{\theta}, \theta) < (r(\hat{\theta}_1, \theta) + r(\hat{\theta}_1', \theta))/2 \leq \epsilon^2$ which contradicts the admissibility of $\hat{\theta}_1$, Remark 4.3. □

Example 4.11 Bounded normal mean. Suppose that $y \sim N(\theta, 1)$ and that it is known *a priori* that $|\theta| \leq \tau$, so that $\Theta = [-\tau, \tau]$. This apparently very special problem will be an important building block later in this chapter. We use the notation $\rho_N(\tau, 1)$ for the minimax risk $R_N(\Theta)$ in this case, in order to highlight the interval endpoint τ and the noise level, here equal to 1.

Let V_τ denote the prior on $[-\tau, \tau]$ with density $(3/(2\tau^3))(\tau - |\theta|)_+^2$; from the discussion above $\rho_N(\tau, 1) \geq B(V_\tau)$. The van Trees inequality (4.9) and $I(V_\tau) = I(V_1)/\tau^2$ implies that

$$\rho_N(\tau, 1) \geq \frac{1}{1 + I(V_\tau)} = \frac{\tau^2}{\tau^2 + I(V_1)}. \tag{4.16}$$

From this one learns that $\rho_N(\tau, 1) \nearrow 1$ as $\tau \to \infty$, indeed at rate $O(1/\tau^2)$. An easy calculation shows that $I(V_1) = 12$. For the exact asymptotic behavior of $1 - \rho_N(\tau, 1)$, see the remark following (4.43) in Section 4.6.

4.4 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.

We state a version of the minimax theorem suited to the Gaussian sequence model. Even in this setting, the proof is elaborate, and so we defer to Appendix A a discussion of its assumptions and proof, and of its connections with the classical minimax theorems of game theory.

A function $f : T \to \mathbb{R}$ on a metric space T is said to be lower semicontinuous at t if

$f(t) \le \liminf_{s\to t} f(s)$. 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 $\mathcal{A} = (\bar{\mathbb{R}})^\infty$, equipped with the (metrizable) topology of pointwise convergence: $a^n \to a$ if and only if $a_i^n \to a_i$ for each i.

Theorem 4.12 *Consider the Gaussian sequence estimation problem* (4.1) *with parameter space* $\ell_{2,\varrho} = \ell_2(\mathbb{N}, (\varrho_i^{-2}))$ *and suppose that for each* θ *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* (4.2). *Let* $\mathcal{P}$ *be a convex set of probability measures on* $\ell_{2,\varrho}$. *Then*

$$\inf_{\hat\theta}\sup_{\pi\in\mathcal{P}} B(\hat\theta,\pi) = \sup_{\pi\in\mathcal{P}}\inf_{\hat\theta} B(\hat\theta,\pi) = B(\mathcal{P}) \tag{4.17}$$

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

The theorem, and identity (4.17) in particular, allows us to refer to $B(\mathcal{P})$, defined in (4.14), as a minimax quantity, and so for $B(\mathcal{P})$ we use the term *Bayes minimax risk* (also known as Γ-minimax risk, see Chapter Notes).

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

$$B(\hat\theta^*, \pi) \le B(\hat\theta^*, \pi^*) \le B(\hat\theta, \pi^*).$$

If a saddlepoint exists, then $\hat\theta^*$ is a Bayes rule for π^* (from the right side), and π^* is a least favorable distribution (since the left side implies $B(\pi) \le B(\pi^*)$ for all π). See Figure 1.7. Proposition 4.14 below gives one setting in which a saddlepoint is guaranteed.

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

$$\sup_{\theta\in\Theta} r(\hat\theta,\theta) \le \sup_{\pi\in\mathcal{P}} B(\hat\theta,\pi).$$

If $\mathcal{P}$ is also convex, then minimizing over all estimators $\hat\theta$ and using the minimax theorem (4.17) gives an upper bound on minimax risk that we will use frequently:

$$R_N(\Theta) \le B(\mathcal{P}). \tag{4.18}$$

The bound is useful because the Bayes-minimax risk $B(\mathcal{P})$ is often easier to evaluate than the minimax risk $R_N(\Theta)$. We can often (see Section 4.11) show that the two are comparable in the low noise limit as $\epsilon \to 0$:

$$R_N(\Theta,\epsilon) \sim B(\mathcal{P},\epsilon).$$

3. In some cases, we may combine the lower and upper bounds (4.15) and (4.18). For example, if $\mathcal{P} = \mathcal{P}(\Theta) = \{\pi : \operatorname{supp}\pi \subset \Theta\}$, then $\mathcal{P}$ is convex and so

$$R_N(\Theta) = B(\mathcal{P}(\Theta)). \tag{4.19}$$

In this case, if $\hat\theta^*$ is minimax for (4.17), then it is minimax for ordinary risk:

$$\sup_{\theta\in\Theta} r(\hat\theta^*,\theta) = R_N(\Theta).$$

Example 4.11 continued. In the bounded normal mean problem of the last section, we have $\Theta = [-\tau, \tau]$ and so

$$\rho_N(\tau, 1) = \sup\{B(\pi) \ : \ \text{supp}\, \pi \subset [-\tau, \tau]\}. \tag{4.20}$$

4. 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 $\|a^{(\infty)} - \theta\|_p^p \le \liminf_n \|a^{(n)} - \theta\|_p^p$. See also Exercise 4.6.

Univariate Bayes Minimax Problems

Return to the univariate setting $n = 1$ with $\epsilon = 1$. Suppose that $\mathcal{P} \subset \mathcal{P}(\mathbb{R})$ is a convex set of probability measures. From the Fisher information representation (4.7).

$$B(\mathcal{P}) = \sup_{\pi \in \mathcal{P}} B(\pi) = 1 - \inf_{P \in \mathcal{P}^\star} I(P), \tag{4.21}$$

where $\mathcal{P}^\star = \{\Phi \star \pi, \pi \in \mathcal{P}\}$. We can again exploit properties of Fisher information $I(P)$ to understand better the Bayes minimax problem $B(\mathcal{P})$. We take advantage also of the fact that convolution with the normal distribution makes every $P \in \mathcal{P}^\star$ smooth. The results find application in Sections 4.6, 8.7 and 13.3. We will refer to Appendix C.19 - C.20 for some properties of $\mathcal{P}(\mathbb{R})$ and Fisher information.

Let $\check{\pi}$ be the distribution of $-\theta$ when $\theta \sim \pi$; call $\mathcal{P}$ *symmetric* if $\pi \in \mathcal{P}$ implies $\check{\pi} \in \mathcal{P}$.

Proposition 4.13 *If $\mathcal{P} \subset \mathcal{P}(\mathbb{R})$ is convex and weakly compact, then there is a unique least favorable distribution $\pi_0 \in \mathcal{P}$. If $\mathcal{P}$ is symmetric, then so is π_0.*

Proof Since $B(\pi)$ is weakly upper semicontinuous on a weakly compact set $\mathcal{P}$, it attains its maximum at some π_0, and correspondingly $P_0 = \Phi \star \pi_0$ minimizes $I(P)$ over $\mathcal{P}^\star$. Since $p_0 = \phi \star \pi_0$ is positive on all of $\mathbb{R}$, we conclude from criterion C.21 that P_0 is the unique minimizer of $I(P)$ on $\mathcal{P}^\star$, so that π_0 is also unique. Since $I(\check{\pi} \star \Phi) = I(\pi \star \Phi)$ for any π, we conclude from the uniqueness just shown that if $\mathcal{P}$ is symmetric, so must be π_0. □

Remark. For Section 8.7 we need an extension of Proposition 4.13. Let $\mathcal{P}_+(\mathbb{R})$ denote the collection of (sub-stochastic) measures π on $\mathbb{R}$ with $0 < \pi(\mathbb{R}) \le 1$, endowed with the topology of vague convergence, C.19. Then Proposition 4.13 also holds if $\mathcal{P} \subset \mathcal{P}_+(\mathbb{R})$ is convex and vaguely compact. The same proof works, since $I(P)$ is vaguely upper semicontinuous, and as $\pi_0(\mathbb{R}) > 0$, we have $p_0 > 0$ on all of $\mathbb{R}$.

Finally, we show that a least favorable distribution generates a saddle point in the Bayes minimax problem.

Proposition 4.14 *Let $\mathcal{P} \subset \mathcal{P}(\mathbb{R})$ be convex. Given $\pi_0, \pi_1 \in \mathcal{P}$, let $\pi_t = (1-t)\pi_0 + t\pi_1$, and let $\hat{\theta}_{\pi_0}$ be the Bayes rule for π_0. Then the following are equivalent:*

$$\begin{aligned}
&(i) && \pi_0 \in \mathcal{P} \quad \textit{is least favorable,} \\
&(ii) && (d/dt)B(\pi_t)|_{t=0+} \le 0 && \textit{for every } \pi_1 \in \mathcal{P}, \\
&(iii) && B(\hat{\theta}_{\pi_0}, \pi_1) \le B(\pi_0) && \textit{for every } \pi_1 \in \mathcal{P},
\end{aligned} \tag{4.22}$$

so that $(\hat{\theta}_{\pi_0}, \pi_0)$ is a saddle point for the Bayes minimax problem.

Proof Convexity of $\mathcal{P}$ says that $\pi_t = (1-t)\pi_0 + t\pi_1$ also belongs to $\mathcal{P}$ for $0 \le t \le 1$. If π_0 is least favorable, then concavity of $B(\pi)$ on $\mathcal{P}$ implies that $(d/dt)B(\pi_t)|_{t=0} \le 0$ for each $\pi_1 \in \mathcal{P}$, so (i) $\Rightarrow$ (ii). Lemma 4.8 exactly says that (ii) $\Leftrightarrow$ (iii). Finally (iii) $\Rightarrow$ (i) because then $B(\pi_1) \le B(\hat{\theta}_{\pi_0}, \pi_1) \le B(\pi_0)$. □

4.5 Product Priors and Spaces

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

Suppose that the prior π makes the groups independent: $\pi(d\theta) = \prod_j \pi_j(d\theta_j)$. In (2.16) we saw that the posterior factorizes, and if in addition the loss function is additive, (2.17), then the Bayes rule is separable (2.18). In such cases, the risk functions are additive

$$r(\hat{\theta}_\pi, \theta) = \sum_j EL(\hat{\theta}_{\pi_j}(y_j), \theta_j) = \sum_j r(\hat{\theta}_{\pi_j}, \theta_j) \tag{4.23}$$

and in consequence, so are the Bayes risks

$$B(\pi) = \int r(\hat{\theta}_\pi, \theta)\pi(d\theta) = \sum_j B(\pi_j). \tag{4.24}$$

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

Lemma 4.15 $B(\bar{\pi}) \ge B(\pi)$.

Proof Because of the independence structure, the $\bar{\pi}-$posterior distribution of θ_j given y in fact depends only on y_j—compare (2.16). 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 θ, (4.23),

$$r(\hat{\theta}_{\bar{\pi}}, \theta) = \sum_j r(\hat{\theta}_{\bar{\pi},j}, \theta_j).$$

The π-average of the rightmost term therefore depends only the marginals π_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 more intuitively why the product marginal prior $\bar{\pi}$ is harder than π, consider squared error loss: conditioning on all of y has to be better—lower variance—than conditioning on just y_j:

$$\begin{aligned}\mathbb{E}_\pi[E_\pi(\theta_j|y) - \theta_j]^2 &= E_\pi \mathrm{Var}\,(\theta_j|y) \\ &\le E_\pi \mathrm{Var}\,(\theta_j|y_j) = \mathbb{E}_{\bar{\pi}}[E_{\bar{\pi}}(\theta_j|y_j) - \theta_j]^2.\end{aligned}$$

Indeed, the inequality above follows from the identity $\mathrm{Var}X = E[\mathrm{Var}(X|Y)]+\mathrm{Var}[E(X|Y)]$ using for X the conditional distribution of $\theta_j|y_j$ and for Y the set $\{y_k : k \neq j\}$.

Product Spaces. Suppose that $\Theta \subset \ell_2(I)$ is a product space $\Theta = \Pi_{j\in J}\Theta_j$. Again the index j may refer to individual coordinates of $\ell_2(I)$ or to a cluster of coordinates. If the loss function is additive and convex, then the minimax risk for Θ can be built from the minimax risk for each of the subproblems Θ_j.

Proposition 4.16 *Suppose that $\Theta = \Pi_{j\in J}\Theta_j$ 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 θ_j. Then*

$$R_N(\Pi_j\Theta_j,\epsilon) = \sum_j R_N(\Theta_j,\epsilon). \tag{4.25}$$

If $\theta_j^*(y_j)$ is separately minimax for each Θ_j, then $\theta^*(y) = \left(\theta_j^*(y_j)\right)$ is minimax for Θ.

Remarks: 1. There *is* something to prove here: among estimators $\hat\theta$ competing in the left side of (4.25), each coordinate $\hat\theta_j(y)$ may depend on *all* components y_j, $j \in J$. The result says that a minimax estimator need not have such dependencies: $\theta_j^*(y)$ 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–Exercise 4.9.

Proof By the minimax theorem (4.12):

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

where $\mathcal{P}(\Theta)$ denotes the collection of all probability measures supported in Θ. Given any such prior π, construct a new prior $\bar\pi$ as the product of the marginal distributions π_j of θ_j under π. Lemma 4.15 shows that $\bar\pi$ is more difficult than π : $B(\bar\pi) \geq B(\pi)$. Because of the product structure of Θ, each π_j is supported in Θ_j and $\bar\pi$ still lives on Θ. Thus the maximization can be restricted to priors with independent coordinates. Bayes risk is then additive, by (4.24), 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. □

4.6 Single Bounded Normal Mean

In this section and the next two, we confine attention to squared error loss.

If $y \sim N(\theta,\epsilon^2)$ and there is no constraint on θ, then we have seen, for example at (2.51), that the minimax mean squared error for estimation of θ based on y equals the variance ϵ^2. Suppose now that θ is known to lie in a *bounded* interval of length 2τ, which without any real loss of generality we may assume to be centered about 0, so that we write $\Theta(\tau) = [-\tau,\tau]$. It is clear that any estimator $\hat\theta$, whether linear or not, can be improved simply by enforcing

the interval constraint: if $\tilde{\theta} = [\hat{\theta}]_{-\tau}^{\tau} = \max\{\min\{\hat{\theta}, \tau\}, -\tau\}$, then $r(\tilde{\theta}, \theta) \le r(\hat{\theta}, \theta)$. This section asks how much better is the nonlinear minimax risk

$$\rho_N(\tau, \epsilon) = \inf_{\hat{\theta}} \sup_{\theta \in [-\tau, \tau]} E_\theta(\hat{\theta} - \theta)^2 \tag{4.26}$$

than $\rho_N(\infty, \epsilon) = \epsilon^2$ and than the corresponding *linear* minimax risk $\rho_L(\tau, \epsilon)$ obtained by restricting $\hat{\theta}$ to linear estimators of the form $\hat{\theta}_c(y) = cy$?

Linear Estimators. Applying the variance-biase decomposition of MSE, (2.46), to a *linear* estimator $\hat{\theta}_c(y) = cy$, we obtain $E(\hat{\theta}_c - \theta)^2 = c^2\epsilon^2 + (1-c)^2\theta^2$. If the parameter is known to lie in a bounded interval $[-\tau, \tau]$, then the maximum risk occurs at the endpoints:

$$\sup_{\theta \in [-\tau, \tau]} E(\hat{\theta}_c - \theta)^2 = c^2\epsilon^2 + (1-c)^2\tau^2 = r(\hat{\theta}_c, \tau). \tag{4.27}$$

The minimax linear estimator is thus found by minimizing the quadratic function $c \to r(\hat{\theta}_c, \tau)$. It follows that

$$\rho_L(\tau, \epsilon) = \inf_c r(\hat{\theta}_c, \tau) = \frac{\epsilon^2\tau^2}{\epsilon^2 + \tau^2}. \tag{4.28}$$

The minimizer $c_* = \tau^2/(\epsilon^2 + \tau^2) \in (0, 1)$ and the corresponding minimax linear estimator

$$\hat{\theta}_{\mathrm{LIN}}(y) = \frac{\tau^2}{\epsilon^2 + \tau^2} y. \tag{4.29}$$

Thus, if the prior 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 to be used.

Of course, $\hat{\theta}_{\mathrm{LIN}}$ is also Bayes for the prior $\pi_\tau(d\theta) = N(0, \tau^2)$ and squared error loss. Indeed, from (2.19) we see that the posterior is Gaussian, with mean (4.29) and variance equal to the linear minimax risk (4.28). Note that this prior is *not* concentrated on $\Theta(\tau)$: only a moment statement is possible: $E_\pi\theta^2 = \tau^2$.

There is a simple but important scale invariance relation

$$\rho_L(\tau, \epsilon) = \epsilon^2 \rho_L(\tau/\epsilon, 1). \tag{4.30}$$

Writing $\nu = \tau/\epsilon$ for the *signal-to-noise* ratio, we have

$$\rho_L(\nu, 1) = \nu^2/(1+\nu^2) \sim \begin{cases} \nu^2 & \nu \to 0 \\ 1 & \nu \to \infty. \end{cases} \tag{4.31}$$

These results, however simple, are nevertheless a first quantitative indication of the importance of prior information, here quantified through ν, on possible quality of estimation.

Projection Estimators. Orthogonal projections form an important and simple subclass of linear estimators. The particular case of projections orthogonal to the co-ordinate axes was defined and discussed in Section 3.2. In one dimension the situation is almost trivial, with only two possibilities. Either $\hat{\theta}_0(y) \equiv 0$ 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

$$\rho_P(\tau, \epsilon) = \inf_{c \in \{0,1\}} \sup_{\theta \in [-\tau, \tau]} E(\hat{\theta}_c - \theta)^2 = \min(\tau^2, \epsilon^2). \tag{4.32}$$

The choice is to "keep or kill": if the signal to noise ratio τ/ϵ exceeds 1, use $\hat{\theta}(y) = y$, otherwise use $\hat{\theta}(y) = 0$. The inequalities

$$\tfrac{1}{2}\min(\tau^2, \epsilon^2) \le \frac{\tau^2\epsilon^2}{\tau^2+\epsilon^2} \le \min(\tau^2, \epsilon^2) \tag{4.33}$$

imply immediately that $\frac{1}{2}\rho_P(\tau,\epsilon) \le \rho_L(\tau,\epsilon) \le \rho_P(\tau,\epsilon)$, so that the best projection estimator is always within a factor of 2 of the best linear estimator.

Non-linear estimators. The non-linear minimax risk $\rho_N(\tau,\epsilon)$, (4.26), cannot be evaluated analytically in general. However, the following properties are easy enough:

$$\rho_N(\tau,\epsilon) \le \rho_L(\tau,\epsilon), \tag{4.34}$$

$$\rho_N(\tau,\epsilon) = \epsilon^2\rho_N(\tau/\epsilon, 1), \tag{4.35}$$

$$\rho_N(\tau,\epsilon) \text{ is increasing in } \tau, \tag{4.36}$$

$$\lim_{\tau\to\infty} \rho_N(\tau,\epsilon) = \epsilon^2. \tag{4.37}$$

Indeed, (4.34) is plain since more estimators are allowed in the nonlinear competition, while (4.35) follows by rescaling, and (4.36) is obvious. Turning to (4.37), we recall that the classical result (2.51) says that the minimax risk for θ unconstrained to any interval, $\rho_N(\infty,\epsilon) = \epsilon^2$. Thus (4.37) asserts continuity as τ increases without bound—and this follows immedately from the example leading to (4.16): $\rho_N(\tau,1) \ge \tau^2/(\tau^2 + I(V_1))$.

In summary so far, we have the bounds $\rho_N \le \rho_L \le \rho_P$, as illustrated in Figure 4.1, from which we might guess that the bounds are relatively tight, as we shall shortly see.

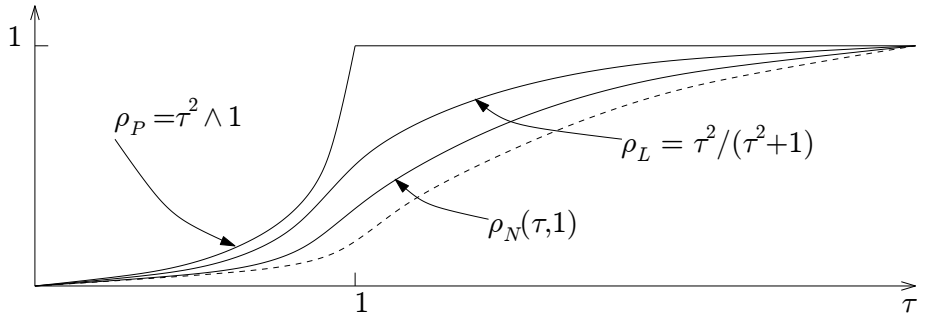

Figure 4.1 Schematic comparison of risk functions ρ_P, ρ_L and ρ_N, dotted line is the lower bound (4.16): $\rho_N(\tau,1) \ge \tau^2/(\tau^2 + I(V_1)) = \tau^2/(\tau^2+12)$.

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.

Theorem 4.17

$$\mu^* := \sup_{\tau,\epsilon} \frac{\rho_L(\tau,\epsilon)}{\rho_N(\tau,\epsilon)} \le 1.25. \tag{4.38}$$

Thus, regardless of signal bound τ and noise level ϵ, linear rules are within 25% of optimal for mean squared error. The bound $\mu^* < \infty$ is due to Ibragimov and Khasminskii (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 show only a weaker result: that μ^* is finite and bounded by 2.22, which says that linear rules are within 122% of optional. For the extra work to get the much better bound of 1.25 we refer to Donoho et al. (1990). Our approach uses projection estimators and the Bayes risk identity (2.30) for the symmetric two point priors $\pi_\tau = (1/2)(\delta_\tau + \delta_{-\tau})$ to give a short and instructive demonstration that $\mu^* \le 1/B(\pi_1)$. Numerical evaluation of the integral (2.30) then shows the latter bound to be approximately 2.22.

First, it is enough to take $\epsilon = 1$, in view of the scaling invariances (4.30) and (4.35). We may summarize the argument by the inequalities:

$$\frac{\rho_L(\tau,1)}{\rho_N(\tau,1)} \le \frac{\tau^2 \wedge 1}{\rho_N(\tau,1)} \le \frac{1}{B(\pi_1)}. \tag{4.39}$$

Indeed, the first bound reflects a reduction to projection estimators, (4.32). For the second inequality, consider first $\tau \ge 1$, and use monotonicity (4.36) and the minimax risk lower bound (4.15) to obtain

$$\rho_N(\tau,1) \ge \rho_N(1,1) \ge B(\pi_1).$$

For $\tau \le 1$, again $\rho_N(\tau,1) \ge B(\pi_\tau)$ and then from (2.30) $\tau^2/B(\pi_\tau)$ is increasing in τ. □

An immediate corollary, using also (4.28) and (4.33), is a bound for ρ_N:

$$(2\mu^*)^{-1} \min(\tau^2, \epsilon^2) \le \rho_N(\tau,\epsilon) \le \min(\tau^2, \epsilon^2). \tag{4.40}$$

The proof also gives sharper information for small and large τ: indeed, the linear minimax risk is then essentially equivalent to the non-linear minimax risk:

$$\mu(\tau) = \rho_L(\tau,1)/\rho_N(\tau,1) \to 1 \qquad \text{as } \tau \to 0, \infty. \tag{4.41}$$

Indeed, for small τ, the middle term of (4.39) is bounded by $\tau^2/B(\pi_\tau)$, which approaches 1, as may be seen from (2.30). For large τ, the same limit results from (4.37). Thus, as $\tau \to 0$, $\hat\theta_0(y) = 0$ is asymptotically optimal, while as $\tau \to \infty, \hat\theta(y) = y$ is asymptotically best. These remarks will play a role in the proof of Pinsker's theorem in the next chapter.

Least favorable priors are discrete.*

The fine structure of minimax rules is in general complicated, although some interesting and useful information is available. First, a property of analytic functions which plays a key role, both here and in Section 8.7.

Lemma 4.18 *Let ν be a probability measure and $K(\nu)$ the smallest interval containing the support of ν. Suppose that $r(\theta)$ is analytic on an open interval containing $K(\nu)$ and satsifies*

$$r(\theta) \le r_\nu = \int r(\theta')\nu(d\theta'), \qquad \theta \in K(\nu). \tag{4.42}$$

Then either $r(\theta)$ is constant on $K(\nu)$, or ν is a discrete measure whose support has no points of accumulation.

Proof Property (4.42) implies that the set $K = \{\theta \in K(\nu) : r(\theta) = r_\nu\}$ has ν-probability equal to 1. Since K is a closed set, it follows from the definition (C.15) that the support of ν is contained in K. Now we recall, e.g. C.8, that if the set of zeros of an analytic function, here $r(\theta) - r_\nu$, has an accumulation point θ_0 inside its domain D, then the function is identically zero on the connected component of D containing θ_0. □

Now to the minimax rules. Let $\bar{r}(\hat{\theta}) = \max_{|\theta|\le\tau} r(\hat{\theta}, \theta)$. Given a prior distribution π, let $M(\pi)$ denote the set of points where the Bayes rule for π attains its maximum risk:

$$M(\pi) = \big\{\theta \in [-\tau, \tau] \,:\, r(\hat{\theta}_\pi, \theta) = \bar{r}(\hat{\theta}_\pi)\big\}.$$

Proposition 4.19 *For the non-linear minimax risk $\rho_N(\tau, \epsilon)$ given by (4.26), a unique least favorable distribution π_τ exists and $(\hat{\theta}_{\pi_\tau}, \pi_\tau)$ is a saddlepoint. The distribution π_τ is symmetric, supp$(\pi_\tau) \subset M(\pi_\tau)$ and $M(\pi_\tau)$ is a finite set. Conversely, if a prior π satisfies supp$(\pi) \subset M(\pi)$, then $\hat{\theta}_\pi$ is minimax.*

Proof We apply Propositions 4.13 and 4.14 to the symmetric set $\mathcal{P}_\tau$ of probability measures supported on $[-\tau, \tau]$, which is weakly compact. Consequently a unique least favorable distribution $\pi_\tau \in \mathcal{P}_\tau$ exists, it is symmetric, and the corresponding Bayes rule $\hat{\theta}_{\pi_\tau}$ satisfies

$$r(\hat{\theta}_{\pi_\tau}, \theta) \le B(\pi_\tau) = \int r(\hat{\theta}_{\pi_\tau}, \theta)\pi_\tau(d\theta),$$

as we see by considering the point masses $\pi = \delta_\theta$ for $\theta \in [-\tau, \tau]$.

The risk function $\theta \to r(\hat{\theta}_{\pi_\tau}, \theta)$ is finite and hence analytic on $\mathbb{R}$, Remark 4.2 of Section 2.5, and not constant (Exercise 4.1). The preceding lemma shows that supp$(\pi_\tau) \subset M(\pi_\tau)$, which can have no points of accumulation and (being also compact) must be a finite set.

Finally, if supp$(\pi) \subset M(\pi)$, then $r(\hat{\theta}_\pi, \theta) = \bar{r}(\hat{\theta}_\pi)$ and so $\hat{\theta}_\pi$ must be minimax:

$$\bar{r}(\hat{\theta}_\pi) = B(\hat{\theta}_\pi, \pi) = \inf_{\hat{\theta}} B(\hat{\theta}, \pi) \le \inf_{\hat{\theta}} \bar{r}(\hat{\theta}). \qquad \square$$

In general, this finite set and the corresponding minimax estimator can only be determined numerically, see Kempthorne (1987); Donoho et al. (1990); Gourdin et al. (1994). Nevertheless, one can still learn a fair amount about these least favorable distributions. Since the posterior distribution of π_τ must also live on this finite set, and since the root mean squared error of $\hat{\theta}_\tau$ must be everywhere less than ϵ, one guesses heuristically that the support points of π_τ will be spaced at a distance on the scale of the noise standard deviation ϵ. See Exercise 4.13 and Figure 4.2.

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

1. Proposition 4.19 says that the symmetric two point prior $\pi_\tau = (1/2)(\delta_\tau + \delta_{-\tau})$ is minimax if $\{-\tau, \tau\} \subset M(\pi_\tau)$. For this two point prior, the posterior distribution and mean $\hat{\theta}_{\pi_\tau}$ were given in Chapter 2, (2.27) – (2.29), and we recall that the Bayes risk satisfies (2.30).

2. Since the posterior distribution concentrates on $\pm\tau$, one guesses from monotonicity and symmetry considerations that $M(\pi_\tau) \subset \{-\tau, 0, \tau\}$ for all τ. The formal proof uses a sign change argument linked to total positivity of the Gaussian distribution – see Casella and Strawderman (1981).

3. A second sign change argument shows that there exists τ_2 such that for $|\tau| < \tau_2$,

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

Thus $\mathrm{supp}(\pi) = \{-\tau, \tau\} = M(\pi_\tau)$ and so $\hat{\theta}_\tau$ is minimax for $|\tau| < \tau_2$, and numerical work shows that $\tau_2 \doteq 1.057$.

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

$$\pi_\tau = (1-\alpha)\delta_0 + (\alpha/2)(\delta_\tau + \delta_{-\tau}).$$

This three point prior appears again in Chapters 8 and 13.

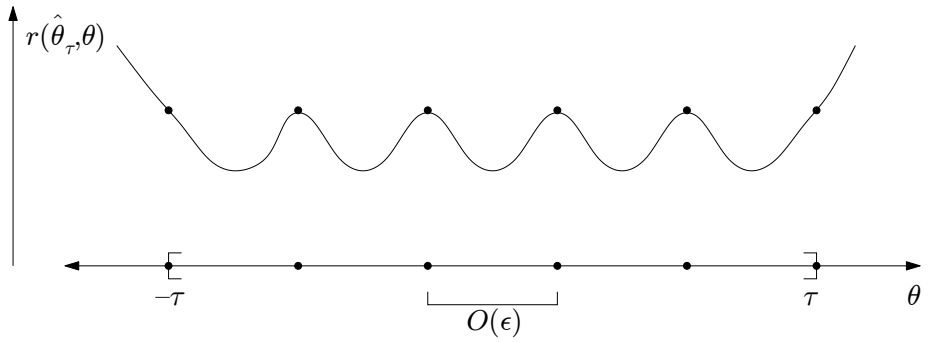

Figure 4.2 as the interval $[-\tau, \tau]$ grows, the support points of the least favorable prior spread out, and a risk function reminiscent of a standing wave emerges.

As $|\tau|$ increases, prior support points are added successively and we might expect a picture such as Figure 4.2 to emerge. Numerical calculations may be found in Gourdin et al. (1994). An interesting phenomenon occurs as τ gets large: if the least favorable distributions π_τ are rescaled to $[-1, 1]$ by setting $\nu_\tau(A) = \pi_\tau(\tau A)$, then Bickel (1981) derives the weak limit $\nu_\tau \Rightarrow \nu_\infty$, with

$$\nu_\infty(ds) = \cos^2(\pi s/2)ds, \tag{4.43}$$

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

4.7 Hyperrectangles

In this section, we 'lift' the results for intervals to hyperrectangles, and obtain some direct consequences for nonparametric estimation over Hölder classes of functions.

The set $\Theta \subset \ell_2(I)$ is said to be a hyperrectangle if

$$\Theta = \Theta(\tau) = \{\theta : |\theta_i| \le \tau_i \text{ for all } i \in I\} = \prod_i [-\tau_i, \tau_i].$$

For $\Theta(\tau)$ to be compact, it is necessary and sufficient that $\sum \tau_i^2 < \infty$, Lemma 3.2. Algebraic

and exponential decay provide natural examples for later use:

$$|\theta_k| \le Ck^{-\alpha}, \qquad k \ge 1, \alpha > 0, C > 0, \tag{4.44}$$

$$|\theta_k| \le Ce^{-ak}, \qquad k \ge 1, a > 0, C > 0. \tag{4.45}$$

We suppose that data y from the heteroscedastic Gaussian model (3.60) is observed, but for notational ease here, we set $\epsilon_i = \varrho_i \epsilon$, so that

$$y_i = \theta_i + \epsilon_i z_i, \qquad i \in I. \tag{4.46}$$

We seek to compare the linear and non-linear minimax risks $R_N(\Theta(\tau), \epsilon) \le R_L(\Theta(\tau), \epsilon)$. The notation emphasizes the dependence on scale parameter ϵ, for later use in asymptotics.

Proposition 4.16 says that the non-linear minimax risk over a hyperrectangle decomposes into the sum of the one-dimensional component problems:

$$R_N(\Theta(\tau), \epsilon) = \sum \rho_N(\tau_i, \epsilon_i). \tag{4.47}$$

Minimax *linear* estimators have a similar structure:

Proposition 4.20 *(i) If $\hat{\theta}_C(y) = Cy$ is minimax linear over a hyperrectangle $\Theta(\tau)$, then necessarily C must be diagonal. (ii) Consequently,*

$$R_L(\Theta(\tau), \epsilon) = \sum_i \rho_L(\tau_i, \epsilon_i) \tag{4.48}$$

Before proving this, we draw an immediate and important consequence: by applying Theorem 4.17 term by term, $\rho_L(\tau_i, \epsilon_i) \le \mu^* \rho_N(\tau_i, \epsilon_i)$, it follows that the Ibragimov-Hasminski theorem lifts from intervals to hyperrectangles:

Corollary 4.21 *In model* (4.46)*,*

$$R_L(\Theta(\tau), \epsilon) \le \mu^* R_N(\Theta(\tau), \epsilon). \tag{4.49}$$

Proof of Proposition 4.20. First note that a diagonal linear estimator $\hat{\theta}_c(y) = (c_i y_i)$ has mean squared error of additive form:

$$r(\hat{\theta}_c, \theta) = \sum_i \epsilon_i^2 c_i^2 + (1 - c_i)^2 \theta_i^2. \tag{4.50}$$

Let $\bar{r}(\hat{\theta}_C) = \sup\{r(\hat{\theta}_C, \theta), \theta \in \Theta(\tau)\}$ and write $d(C) = \mathrm{diag}(C)$ for the matrix (which is infinite if I is infinite) obtained by setting the off-diagonal elements to 0. We show that this always improves the estimator over a hyperrectangle:

$$\bar{r}(\hat{\theta}_{d(C)}) \le \bar{r}(\hat{\theta}_C). \tag{4.51}$$

Recall formula (3.63) for the mean squared error of a linear estimator. The variance term is easily bounded—with $\Delta = \mathrm{diag}(\epsilon_i^2)$, we have, after dropping off-diagonal terms,

$$\mathrm{tr}\, C^T \Delta C = \sum_{ij} c_{ij}^2 \epsilon_i^2 \ge \sum_i c_{ii}^2 \epsilon_i^2 = \mathrm{tr}\, d(C)^T \Delta d(C).$$

For the bias term, $\|(C - I)\theta\|^2$, we employ a simple but useful *random signs* technique. Let $\sigma \in \Theta(\tau)$ and denote by $V(\sigma)$ the vertex set of the corresponding hyperrectange $\Theta(\sigma)$:

thus $V(\sigma) = \{(s_i\sigma_i) : i \in I, s_i = \pm 1\}$. Let π_σ be a probability measure that makes θ_i independently equal to $\pm\sigma_i$ with probability $1/2$. Then we may bound the maximum squared bias from below by an average, and then use $E\theta_i\theta_j = \sigma_i^2\delta_{ij}$ to obtain

$$\begin{aligned}\sup_{\theta\in V(\sigma)} \|(C-I)\theta\|^2 &\geq E\theta^T(C-I)^T(C-I)\theta \\ &= \text{tr}\,(C-I)\Lambda(C-I)^T \geq \|(d(C)-I)\sigma\|^2,\end{aligned}$$

where $\Lambda = \text{diag}(\sigma_k^2)$ and the final inequality simply drops off-diagonal terms in evaluating the trace.

The risk of a diagonal linear estimator is identical at all the vertices of $V(\sigma)$—compare (4.50)—and so for all vertex sets $V(\sigma)$ we have shown that

$$\sup_{\theta\in V(\sigma)} r(\hat{\theta}_C, \theta) \geq \sup_{\theta\in V(\sigma)} r(\hat{\theta}_{d(C)}, \theta).$$

Since $\sigma \in \Theta(\tau)$ is arbitrary, we have established (4.51) and hence part (i).

Turning to part (ii), we may use this reduction to diagonal linear estimators to write

$$R_L(\Theta(\tau), \epsilon) = \inf_{(c_i)} \sup_{\theta\in\Theta(\tau)} \sum_i E(c_i y_i - \theta_i)^2.$$

Now, by the diagonal form $c_i y_i$ and the product structure of $\Theta(\tau)$, the infimum and the supremum can be performed term by term. Doing the supremum first, and using (4.27),

$$R_L(\Theta(\tau), \epsilon) = \inf_c r(\hat{\theta}_c, \tau). \tag{4.52}$$

Now minimizing over c, we get the right side of (4.48). □

Remarks. 1. It is evident from the proof that we only improve the maximum risk by restricting each c_i to the interval $[0, 1]$.

2. For the admissibility result, Theorem 2.5, 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.

The remainder of this section comprises three examples illustrating the usefulness of hyperrectangles, both in their own right, and as handy tools for lower bounds.

Hyperrectangles and smoothness.

If (θ_i) represent the coefficients of a function f in an appropriate orthonormal basis, then the rate of decay of τ_i in a hyperrectangle condition can correspond to smoothness information about f. For periodic functions on $[0, 1]$, the Fourier basis is natural. If f is C^α, in the sense of Hölder continuity (see Appendix C.23), then the Fourier coefficients satisfy (4.44) for some constant C (e.g. Katznelson (1968, p. 25) for α integer-valued and Zygmund (1959, p. 46) for $0 < \alpha < 1$.) 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 C and a so that (4.45) holds. (e.g. Katznelson (1968, p. 26)). The size of the domain of analyticity grows with a, Exercise 4.14. However,

analyticity conditions are less often used in nonparametric theory than are constraints on a finite number of derivatives.

From this perspective, the situation is much better for wavelet bases, to be discussed in Chapter 7 and Appendix B, since Hölder smoothness is exactly characterized by hyperrectangle conditions, at least for non-integer α.

To describe this, we introduce doubly indexed vectors (θ_{jk}) and hyperrectangles

$$\Theta^{\alpha}_{\infty}(C) = \{(\theta_{jk}) : |\theta_{jk}| \le C 2^{-(\alpha+1/2)j},\ j \in \mathbb{N},\ k = 1, \dots, 2^j\}. \tag{4.53}$$

Let (θ_{jk}) be coefficients of f in an orthonormal wavelet basis for $L_2[0,1]$ of regularity m. Then, according to Remark 9.7, f is C^{α} for $\alpha < m$ and $\alpha \notin \mathbb{N}$, if and only if for some constant C, the coefficients $(\theta_{jk}) \in \Theta^{\alpha}_{\infty}(C)$ defined in (4.53). The subscript ∞ indicates that the bounds hold for all (j,k) and emphasizes that Hölder continuity measures uniform smoothness.

Proposition 4.22 *Assume a Gaussian sequence model $y_{jk} = \theta_{jk} + \epsilon_j z_{jk}$, with $\epsilon_j = 2^{\beta j}\epsilon$, $\beta > -1/2$, and θ assumed to belong to a Hölder ball $\Theta^{\alpha}_{\infty}(C)$ defined at* (4.53)*. Then*

$$R_N(\Theta^{\alpha}_{\infty}(C), \epsilon) \asymp C^{2(1-r)}\epsilon^{2r}, \qquad r = 2\alpha/(2\alpha + 2\beta + 1). \tag{4.54}$$

The notation shows the explicit dependence on both C and ϵ. The expression $a(\epsilon) \asymp b(\epsilon)$ means that there exist positive constants $\gamma_1 < \gamma_2$ depending only on α and β, but not on C or ϵ, such that for all ϵ, we have $\gamma_1 \le a(\epsilon)/b(\epsilon) \le \gamma_2$. The constants γ_i may not be the same at each appearance of $\asymp$.

While the wavelet interpretation is not needed to state and prove this result (which is why it can appear in this chapter!) its importance derives from the smoothness characterization. Indeed, this result exhibits the same rate of convergence as we saw for *mean square* smoothness, i.e. for Θ an ellipsoid, in the upper bound of (3.18). Note that here we also have a lower bound.

The noise level $\epsilon_j = 2^{\beta j}\epsilon$ is allowed to depend on 'level' j, but not on k—the parameter β corresponds to that in the ill-posed linear inverse problems of Section 3.9, and appears again in the discussion of the wavelet-vaguelette decomposition and the 'correlated levels' model of Section 12.4.

Proof Using (4.47), we can reduce to calculations based on the single bounded normal mean problem:

$$R_N(\Theta, \epsilon) = \sum_j 2^j \rho_N(C 2^{-(\alpha+1/2)j}, 2^{\beta j}\epsilon).$$

Using (4.40), $\rho_N(\tau, \epsilon) = \gamma(\tau^2 \wedge \epsilon^2)$, where $\gamma = \gamma(\tau/\epsilon) \in [1/(2\mu^*), 1]$. So let $j_* \in \mathbb{R}$ solve

$$C 2^{-(\alpha+1/2)j_*} = 2^{\beta j_*}\epsilon.$$

For $j < j_*$, the variance term $2^{2\beta j}\epsilon^2$ is active in the bound for ρ_N, while for $j > j_*$ it is the squared bias term $C^2 2^{-(2\alpha+1)j}$ which is the smaller. Hence, with $j_0 = [j_*]$,

$$R_N(\Theta, \epsilon) \asymp \sum_{j \le j_0} 2^j \cdot 2^{2\beta j}\epsilon^2 + C^2 \sum_{j > j_0} 2^{-2\alpha j}.$$

These geometric sums are dominated by their leading terms, multiplied by constants depending only on α and β. Consequently,

$$R_N(\Theta, \epsilon) \asymp 2^{(2\beta+1)j_*}\epsilon^2 + C^2 2^{-2\alpha j_*} \asymp (C^2)^{(2\beta+1)/(2\alpha+2\beta+1)}(\epsilon^2)^{2\alpha/(2\alpha+2\beta+1)}$$

which becomes (4.54) on substituting for r. □

Hyperrectangles and lower bounds for rates

For a general parameter space Θ, it is clear that if a hyperrectangle $\Theta(\tau) \subset \Theta$, then we have a lower bound $R_N(\Theta) \geq R_N(\Theta(\tau))$. As we have seen, the minimax risk of a hyperrectangle has such simple structure that it is tempting to use this as a technique for lower bounds for general Θ. We will discuss this approach more systematically in the next section and in Section 9.3. Here we do a simple example with ellipsoids in order to obtain simple lower bounds on rates of convergence for some of the problems considered in Chapter 3.

Consider model (4.46) for $I = \mathbb{N}$ and $\epsilon_k = \epsilon k^\beta$, which covers some important examples from the inverse problems discussion in Section 3.9 as well as, of course, the white noise case $\beta = 0$. Let Θ be the ellipsoid $\Theta_2^\alpha(C) = \{\theta : \sum a_k^2\theta_k^2 \leq C^2\}$ for $a_k = k^\alpha$, which corresponds (in the Fourier basis at least) to mean square smoothness of order α.

Proposition 4.23 *Assume model* (4.46) *with* $\epsilon_k = \epsilon k^\beta$ *for* $\beta \geq 0$ *and ellipsoidal parameter space* $\Theta_2^\alpha(C)$ *for* $\alpha, C > 0$. *Let the rate parameter* $r = 2\alpha/(2\alpha + 2\beta + 1)$. *Then*

$$R_N(\Theta_2^\alpha(C), \epsilon) \geq c_{\alpha\beta} C^{2(1-r)}\epsilon^{2r},$$

for $\epsilon/C \leq d_{\alpha\beta}$ *where the constants* $c_{\alpha\beta}, d_{\alpha\beta}$ *depend only on* α, β.

Thus, as discussed in Section 3.9, β reflects the ill-posedness of the estimation problem.

Proof Let $\Theta(\tau) = [-\tau, \tau]^m$ denote the hypercube in which the first m coordinates satisfy $|\theta_k| \leq \tau$ and all subsequent coordinates $\theta_{m+j} \equiv 0$. From the hyperrectangle structure, (4.47), and the univariate bound, (4.40), we obtain

$$R_N(\Theta(\tau), \epsilon) = \sum_{k=1}^m \rho_N(\tau, \epsilon_k) = \gamma \sum_1^m \min(\tau^2, \epsilon_k^2),$$

where $\gamma \in [1/(2\mu^*), 1]$. Since $\epsilon_k = \epsilon k^\beta$ is increasing, we may set $\tau = \epsilon_m$ and obtain

$$R_N(\Theta(\epsilon_m), \epsilon) \geq \gamma\epsilon^2 \sum_1^m k^{2\beta} \geq \gamma c_\beta \epsilon^2 m^{2\beta+1},$$

where $c_\beta = (2\beta + 1)^{-1}$. Now, for $\Theta(\epsilon_m)$ to be contained in $\Theta_2^\alpha(C)$ it suffices that

$$\epsilon_m^2 \sum_1^m a_k^2 = \epsilon_m^2 \sum_1^m k^{2\alpha} \leq c_\alpha \epsilon^2 (m+1)^{2\alpha+2\beta+1} \leq C^2.$$

Let m_1 be the largest integer such that $c_\alpha(m+1)^{2\alpha+2\beta+1} \leq C^2/\epsilon^2$ and $m_0 \in \mathbb{R}$ the solution to $c_\alpha m_0^{2\alpha+2\beta+1} = C^2/\epsilon^2$. It is easy to check that if $m_0 \geq 4$, say, then $m_1/m_0 \geq 1/2$. We may therefore conclude from the previous two displays that for $\epsilon/C \leq d_{\alpha\beta}$,

$$R_N(\Theta, \epsilon) \geq R_N(\Theta(\epsilon_{m_1}), \epsilon) \geq \gamma c_\beta \epsilon^2 m_1^{2\beta+1} \geq c_{\alpha\beta}\epsilon^2 m_0^{2\beta+1} \geq c_{\alpha\beta}\epsilon^2 (C^2/\epsilon^2)^{1-r}. \quad \square$$

Hyperrectangles and discrete loss functions

Suppose again that $y_i \overset{\text{ind}}{\sim} N(\theta_i, \epsilon^2)$ for $i = 1, \ldots, n$ and consider the product prior

$$\theta_i \overset{\text{ind}}{\sim} \tfrac{1}{2}(\delta_{\tau_i} + \delta_{-\tau_i}).$$

We take a brief break from squared error loss functions to illustrate the discussion of product priors, additive loss functions and posterior modes of discrete priors (cf. Section 2.3) in the context of three related *discrete* loss functions

$$\begin{aligned} L_0(a, \theta) &= \sum_i I\{a_i \neq \theta_i\}, \\ N(a, \theta) &= \sum_i I\{\operatorname{sgn} a_i \neq \operatorname{sgn} \theta_i\} \quad \text{and} \\ N_c(a, \theta) &= I\{N(a, \theta) \geq c\}. \end{aligned}$$

Here L_0 is counting error, while N counts *sign* errors and N_c, which is *not* additive, is the indicator of a tail event for N.

In each case, the Bayes rule for π, in accordance with (2.9), is found by minimizing, over a, the posterior expected loss. Since the prior has independent coordinates, so does the posterior, which is given by the noise level ϵ version of (2.27). Hence the distribution of θ_i given y is concentrated on $\pm\tau_i$, and by (2.28), it follows that for all three losses $E[L(a, \theta)|y]$ is minimized by the same Bayes rule

$$\hat{\theta}_{\pi,i}(y) = \tau_i \operatorname{sgn}(y_i),$$

and observe that

$$N(\hat{\theta}_\pi, \theta) = \sum_i I\{\operatorname{sgn} y_i \neq \operatorname{sgn} \theta_i\}$$

counts sign errors in the data.

Using the equivalent frequentist view of Bayes estimators, $B(\hat{\theta}, \pi) \geq B(\hat{\theta}_\pi, \pi)$, cf. Section 4.1, we have therefore shown, using loss N_c as an example, that for all estimators $\hat{\theta}$, and in the joint distribution $\mathbb{P}$ of (θ, y), that

$$\mathbb{P}\{N(\hat{\theta}, \theta) \geq c\} \geq \mathbb{P}\{N(\hat{\theta}_\pi, \theta) \geq c\}.$$

Consider now a hypercube situation, in which all $\tau_i \equiv \tau$. Then in the joint distribution $\mathbb{P}$, we have $N(\hat{\theta}_\pi, \theta) \overset{\mathcal{D}}{\sim} \text{Bin}(n, \pi_1)$, where $\pi_1 = P\{N(\tau, \epsilon^2) < 0\} = \Phi(-\tau/\epsilon)$. Hence, for loss function N_c, the Bayes risk is a binomial probability tail event, $P\{\text{Bin}(n, \pi_1) \geq c\}$.

These remarks will be used later for lower bounds in the optimal recovery approach to thresholding, Section 10.4.

4.8 Orthosymmetry and Hardest rectangular subproblems

Although the minimax structure of hyperrectangles is, as we have just seen, essentially straightforward, it is a key tool for obtaining deeper results on minimax risks for more general sets satisfying certain symmetry and convexity properties that we now define.

Θ is said to be *solid* and *orthosymmetric* if $\theta \in \Theta$ and $|\xi_i| \leq |\theta_i|$ for all i implies

that $\xi \in \Theta$. If a solid, orthosymmetric Θ contains a point τ, then it contains the entire hyperrectangle that τ defines: $\Theta(\tau) \subset \Theta$.

Examples of solid orthosymmetric sets:

- Sets defined by the contours of symmetric increasing functions. Thus, if ψ is increasing on $\mathbb{R}^+$, then $\{\theta : \sum a_i \psi(\theta_i^2) \le 1\}$ is solid and orthosymmetric.
- ℓ_p *bodies*: defined by $\sum_i a_i^p |\theta_i|^p \le C^p$ for $p > 0$, and
- *Besov bodies*: defined by $\sum_j 2^{jsq} (\sum_k |\theta_{jk}|^p)^{q/p} \le C^q$ for $0 < p, q \le \infty$, Section 9.6.

Since Θ contains $\Theta(\tau)$ for each $\tau \in \Theta$, it is clear that $R_N(\Theta) \ge R_N(\Theta(\tau))$. In the last section we saw how inequalities for hypercubes could give lower bounds for rates of convergence. Here, however, we will see that remarkably sharp information can sometimes be obtained by optimizing over the full class of hyperrectangles. Thus we consider the *hardest rectangular subproblem* of Θ:

$$R_N(\Theta) \ge \sup_{\tau \in \Theta} R_N(\Theta(\tau)). \tag{4.55}$$

For linear estimation, we show that the minimax risk for Θ may be found among *diagonal* linear estimators.

Lemma 4.24 *Let $\hat{\theta}_c(y) = (c_i y_i)$ denote a diagonal linear estimator with $c \in \ell_2(\mathbb{N}, (\epsilon_i^2))$, Suppose that Θ is solid and orthosymmetric. Then*

$$R_L(\Theta) = \inf_c \sup_{\theta \in \Theta} r(\hat{\theta}_c, \theta). \tag{4.56}$$

Proof Indeed, we first observe that according to the proof of Proposition 4.20, the maximum risk of any linear estimator $\hat{\theta}_C$ over any hyperrectangle can be reduced by discarding off-diagonal terms:

$$\sup_{\theta \in \Theta(\tau)} r(\hat{\theta}_C, \theta) \ge \sup_{\theta \in \Theta(\tau)} r(\hat{\theta}_{\mathrm{diag}(C)}, \theta).$$

The previous display holds for every hyperrectangle, and Θ is orthosymmetric, so

$$\begin{aligned} \inf_C \sup_{\theta \in \Theta} r(\hat{\theta}_C, \theta) &\ge \inf_C \sup_{\tau \in \Theta} \sup_{\theta \in \Theta(\tau)} r(\hat{\theta}_{\mathrm{diag}(C)}, \theta) \\ &= \inf_c \sup_{\theta \in \Theta} r(\hat{\theta}_c, \theta) \ge \inf_C \sup_{\theta \in \Theta} r(\hat{\theta}_C, \theta), \end{aligned}$$

and we must have equality throughout, completing the proof. □

Quadratic convexity. To fully relate the linear minimax risk of Θ to that of the rectangular subproblems $\Theta(\tau)$, we need an extra convexity property. Θ is said to be *quadratically convex* if $\Theta_+^2 = \{(\theta_i^2) : \theta \in \Theta\}$ is convex. Examples include sets of the form $\{\theta : \sum a_i \psi(\theta_i^2) \le 1\}$ for ψ a convex function. This makes it clear that quadratic convexity is a more restrictive property than ordinary (linear) convexity. Particular examples include

- ℓ_p *bodies*: for $2 \le p \le \infty$, and
- *Besov bodies*: for $2 \le p \le q \le \infty$.

Just as in (4.55) the *linear* minimax risk over Θ is clearly bounded below by that of the hardest rectangular subproblem. However, for quadratically convex Θ, and squared error loss (to which it is adapted), the *linear* difficulties are actually *equal*:

Theorem 4.25 *(Donoho et al., 1990) Consider the heteroscedastic Gaussian sequence model* (4.46)*. If Θ is compact, solid, orthosymmetric and quadratically convex, then*

$$R_L(\Theta) = \sup_{\tau \in \Theta} R_L(\Theta(\tau)). \tag{4.57}$$

Combining (4.57), (4.49) and (4.55), we immediately obtain a large class of sets for which the linear minimax estimator is almost as good as the non-linear minimax rule.

Corollary 4.26 *If Θ is compact, solid, orthosymmetric and quadratically convex, then $R_L(\Theta) \le \mu^* R_N(\Theta)$.*

This collection includes ℓ_p bodies for $p \ge 2$ – and so certainly ellipsoids, solid spheres, etc. and the Besov bodies just discussed.

Remark 4.27 The results of preceding Theorem and Corollary extend easily, cf. Exercise 4.16, to parameter spaces with a Euclidean factor: $\Theta = \mathbb{R}^k \times \Theta'$, where Θ' is compact (and solid, orthosymmetric and quadratically convex). This brings in useful examples such as Sobolev ellipsoids in the Fourier basis, $\check{\Theta}_2^\alpha(C)$, recall (3.9).

Proof of Theorem 4.25. First we observe that (4.57) can be formulated as a minimax theorem. Indeed, (4.56) displays the left side as an inf sup. From (4.52) we see that the right side of (4.57) equals $\sup_{\tau\in\Theta} \inf_c r(\hat{\theta}_c, \tau)$. Thus, we need to show that

$$\inf_c \sup_{\theta\in\Theta} r(\hat{\theta}_c, \theta) = \sup_{\theta\in\Theta} \inf_c r(\hat{\theta}_c, \theta). \tag{4.58}$$

Using expression (4.50), we may rewrite $r(\hat{\theta}_c, \theta) = f(c, s)$, where

$$f(c, s) = \sum_i c_i^2 \epsilon_i^2 + (1 - c_i)^2 s_i,$$

and $s = (\theta_i^2)$. Now we verify that we may apply the Kneser-Kuhn minimax theorem, Corollary A.4, to $f(c, s)$. Clearly f is convex-concave – indeed, even linear in the second argument. By Remark 1 following Proposition 4.20, we may assume that the vector $c \in \ell_2(\mathbb{N}, (\epsilon_i^2)) \cap [0, 1]^\infty$, while $s \in \Theta_+^2 \subset \ell_1$. The latter set is convex by assumption and ℓ_1-compact by the assumption that Θ is ℓ_2-compact. (Check this, using the Cauchy-Shwarz inequality.) Finally, $f(c, s)$ is trivially ℓ_1-continuous in s for fixed c in $[0, 1]^\infty$. □

Example. Let $\Theta_{n,2}(C)$ denote an ℓ_2 ball of radius C in $\mathbb{R}^n$: $\{\theta : \sum_1^n \theta_i^2 \le C^2\}$. Theorem 4.25 says, in the homoscedastic case $\epsilon_i \equiv \epsilon$, that

$$R_L(\Theta_{n,2}(C), \epsilon) = \sup \left\{ \epsilon^2 \sum_1^n \frac{\tau_i^2}{\epsilon^2 + \tau_i^2} : \sum_1^n \tau_i^2 \le C^2 \right\},$$

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,

$$R_L(\Theta_{n,2}(C), \epsilon) = n\epsilon^2 \cdot \frac{C^2}{n\epsilon^2 + C^2}, \tag{4.59}$$

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 ∞.

While the norm ball in infinite sequence space, $\Theta_2(C) = \{\theta \in \ell_2 : \|\theta\|_2 \le C\}$ is *not* compact, the preceding argument does yield the lower bound

$$R_L(\Theta_2(C), \epsilon) \ge C^2,$$

which already shows that no linear estimate can be uniformly consistent as $\epsilon \to 0$ over all of $\Theta_2(C)$. Section 5.5 contains an extension of this result.

Remark. We pause to preview how the various steps taken in this chapter and the next can add up to a result of some practical import. Let $\hat{\theta}_{PS,\lambda}$ denote the periodic smoothing spline with regularization parameter λ in the white noise model, Section 3.4. If it is agreed to compare estimators over the mean square smoothness classes $\Theta^\alpha = \Theta_2^\alpha(C)$, Section 3.1, (3.9), it will turn out that one cannot improve very much over smoothing splines from the worst-case MSE point of view.

Indeed, borrowing some results from the next chapter (§5.1, §5.2), the best mean squared error for such a smoothing spline satisfies

$$R_{PS}(\Theta^\alpha, \epsilon) = \inf_\lambda \sup_{\theta \in \Theta^\alpha} r(\hat{\theta}_{PS,\lambda}, \theta; \epsilon) \le (1 + c(\alpha, \epsilon))\, R_L(\Theta^\alpha, \epsilon),$$

along with the bound $\lim_{\epsilon \to 0} c(\alpha, \epsilon) \le 0.083$ if $\alpha \ge 2$. In combination with this chapter's result bounding linear minimax risk by a small multiple of non-linear minimax risk, Corollary 4.26, we can conclude that

$$R_{PS}(\Theta_2^\alpha(C), \epsilon) \le (1.10)(1.25)\, R_N(\Theta_2^\alpha(C), \epsilon)$$

for all $\alpha \ge 2$ and at least all sufficiently small ϵ. Thus even arbitrarily complicated non-linear esimators cannot have worst-case mean squared error much smaller than that of the relatively humble linear smoothing spline.

4.9 Correlated Noise*

For this section we consider a modification of Gaussian sequence model (3.1),

$$y_i = \theta_i + \epsilon z_i, \qquad i \in \mathbb{N}, \qquad \mathrm{Cov}(z) = \Sigma, \tag{4.60}$$

in which the components z_i may be correlated. In contrast to Section 3.10, we may not necessarily wish to work in a basis that exactly diagonalizes Σ. This will be of interest in the later discussion of linear inverse problems with a wavelet-vaguelette decomposition, Chapter 12.

Make the obvious extensions to the definition of minimax risk among all non-linear and among linear estimators. Thus, for example, $R_N(\Theta, \Sigma) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} EL(\hat{\theta}(y), \theta)$ when y follows (4.60). For suitable classes of priors $\mathcal{P}$, we similarly obtain Bayes minimax risks $B(\mathcal{P}, \Sigma)$. The first simple result captures the idea that adding independent noise can only make estimation harder. Recall the non-negative definite ordering of covariance matrices or operators: $\Sigma' \preceq \Sigma$ means that $\Sigma - \Sigma'$ is non-negative definite.

Lemma 4.28 *Consider two instances of model* (4.60) *with* $\Sigma' \preceq \Sigma$. *Suppose that the loss function* $a \to L(a, \theta)$ *is convex. Then*

$$R_N(\Theta, \Sigma') \le R_N(\Theta, \Sigma), \qquad \textit{and} \qquad R_L(\Theta, \Sigma') \le R_L(\Theta, \Sigma).$$

Similarly, $B(\mathcal{P}, \Sigma') \le B(\mathcal{P}, \Sigma)$.

Proof A conditioning argument combined with Jensen's inequality is all that is needed. Indeed, let y follow (4.60) and in parallel let $y' = \theta + \epsilon z'$ with $\mathrm{Cov}(z') = \Sigma'$. Since $\Sigma \succeq \Sigma'$, we can find a zero mean Gaussian vector w with covariance $\epsilon(\Sigma - \Sigma')$, independent of z', so that $y = y' + w$. Let $\hat{\theta}(y)$ be an arbitrary estimator for noise Σ; we define

$$\tilde{\theta}(y') = E_\theta[\hat{\theta}(y)|y'] = E[\hat{\theta}(y' + w)|y']$$

which has risk function at least as good as $\hat{\theta}(y)$. Indeed, using convexity of the loss function,

$$E_\theta L(\tilde{\theta}(y'), \theta) \le E_\theta E[L(\hat{\theta}(y' + w), \theta)|y'] = E_\theta E_\theta[L(\hat{\theta}(y), \theta)|y'] = E_\theta L(\hat{\theta}(y), \theta).$$

Since this holds for arbitrary $\hat{\theta}$, the statements for nonlinear minimax risk and for Bayes minimax risk follow.

If $\hat{\theta}(y) = Cy$ is linear, then $\tilde{\theta}(y') = E[C(y' + w)|y'] = Cy'$ is also linear and so the preceding display also establishes the linear minimax inequality result. □

Corollary 4.29 *In white noise model* (3.1), *if* $\epsilon' \le \epsilon$, *then* $R_N(\Theta, \epsilon') \le R_N(\Theta, \epsilon)$.

When the noise is independent in each coordinate and Θ is orthosymmetric, we have seen at (4.56) that the minimax linear estimator can be found among diagonal estimators. When the noise is correlated, however, diagonal estimation can be quite poor. First some notation: for covariance matrix Σ, let $\Sigma_d = \mathrm{diag}(\Sigma)$ be the diagonal matrix with entries taken from the diagonal of Σ. When considering only *diagonal* linear estimators, $\hat{\theta}_{c,i}(y) = c_i y_i$, let

$$R_{DL}(\Theta, \Sigma) = \inf_c \sup_{\theta \in \Theta} r(\hat{\theta}_c, \theta). \tag{4.61}$$

Of course, $R_{DL}(\Theta, \Sigma) = R_{DL}(\Theta, \Sigma_d)$ since R_{DL} involves only the variances of z. Finally, let the correlation matrix corresponding to Σ be

$$\varrho(\Sigma) = \Sigma_d^{-1/2} \Sigma \Sigma_d^{-1/2}.$$

Proposition 4.30 *Suppose that* y *follows the correlated Gaussian model* (4.60). *Let* $\varrho_{\min}$ *denote the smallest eigenvalue of* $\varrho(\Sigma)$. *Suppose that* Θ *is orthosymmetric and quadratically convex. Then*

$$R_L(\Theta, \Sigma) \le R_{DL}(\Theta, \Sigma) \le \varrho_{\min}^{-1} R_L(\Theta, \Sigma).$$

If Σ is diagonal, then $\varrho_{\min} = 1$ and $R_{DL} = R_L$. This happens, for example, in the Karhunen-Loève basis, Section 3.10. If Σ is near-diagonal—in a sense to be made more precise in Chapter 12—then not much is lost with diagonal estimators. For general Σ, it can happen that $\varrho_{\min}$ is small and the bound close to sharp, see the example below.

Proof Only the right hand side bound needs proof. It is easily verified that $\Sigma \succeq \varrho_{\min} \Sigma_d$ and that $\varrho_{\min} \le 1$ and hence using Lemma 4.28 that

$$R_L(\Theta, \Sigma) \ge R_L(\Theta, \varrho_{\min} \Sigma_d) \ge \varrho_{\min} R_L(\Theta, \Sigma_d).$$

By (4.56), in the independent co-ordinates model, $R_L(\Theta, \Sigma_d) = R_{DL}(\Theta, \Sigma_d)$. But as noted above, $R_{DL}(\Theta, \Sigma_d) = R_{DL}(\Theta, \Sigma)$. □

Example 4.31 Consider a p-variate "intra-class" correlation model in which $z_k = \sigma\eta + w_k$ is built from a common variable η and from w_k, all assumed to be i.i.d $N(0, 1)$. Then one checks that $\Sigma_{jk} = \sigma^2 + \delta_{jk}$ and then that $\varrho_{\min}(\varrho(\Sigma)) = 1/(\sigma^2 + 1)$.

Suppose that $\sigma^2 > 1$, and $\tau^2 = \tau_p^2 \to \infty$ but $\tau_p^2 = o(p)$. Then for the hypercube $\Theta(\tau) = \{\theta = \sum_1^p \theta_k e_k : |\theta_k| \le \tau\}$, it can be shown, Exercise 4.19, that

$$R_{DL}(\Theta(\tau)) \sim (\sigma^2 + 1) R_L(\Theta(\tau)), \tag{4.62}$$

as $p \to \infty$, so that the bound of Proposition 4.30 is essentially sharp.

4.10 Lower Bounds Overview*

We are much concerned with evaluation of minimax risks $R_N(\Theta, \epsilon)$. In some cases, as we have seen, the risk can be evaluated exactly. More often one must resort to approximations and bounds. In such cases, upper bounds are typically obtained by focusing on a specific estimator or class of estimators and evaluating or bounding the maximum risk.

The development of lower bounds, of necessity applicable to all estimators, must take into account the structure of the estimation model, parameter set and loss function. While there are many strategies available, the construction of appropriate lower bounds is still driven as much by art as by algorithm.

We provide, for reference, a brief list of the strategies used for lower bounds at various points in this book. This overview necessarily reflects the special structure of the Gaussian sequence model–independence of co-ordinates, geometric structure of parameter spaces and Gaussian likelihoods–and our focus on global estimation rather than linear or nonlinear functionals. For other problems in nonparametric estimation, such as regression or density estimation, and/or estimation of functionals, consult for example the more detailed account in Tsybakov (2009, Ch. 2).

(a) Exact evaluation of the minimax risk via the Bayes minimax risk: $R_N(\Theta) = B(\mathcal{P})$, in the case of product spaces $\Theta = \prod_i \Theta_i$, after using independence to write $R_N(\Theta) = \sum_i R_N(\Theta_i)$. This is used for intervals and hyperrectangles in Sections 4.6 and 4.7.

(b) Approximation of the minimax risk, for example by a Bayes minimax risk, in the low noise limit: $R_N(\Theta, \epsilon) \sim B(\mathcal{M}, \epsilon)$ as $\epsilon \to 0$. Here $\mathcal{M}$ is a suitable class of priors, often defined by moment constraints. A sequence of priors in $\mathcal{M}$ that is asymptotically least favorable *and* asymptotically concentrated on Θ is used. This strategy is described in the next section and used repeatedly: Remark 4.33 has a detailed list. An essential building block in some cases is the asymptotic analysis of the Bayes risk behavior of sparse two point priors, Section 8.5.

(c) Bounds using specific priors: $R_N(\Theta, \epsilon) \ge B(\pi_{\gamma(\epsilon)})$. Sometimes the structure is simple enough (for example, extreme sparsity with a fixed number of spikes) that asymptotic concentration fails, but one can work with a specific prior sequence $\pi_{\gamma(\epsilon)}$ supported on Θ. See Section 8.6 and 13.2.

(d) Containment: clearly, if $\Theta \supset \Theta_0$ then $R_N(\Theta) \ge R_N(\Theta_0)$ and Θ_0 may have simpler structure so that one can use evaluation (a). Such an approach is non-asymptotic, i.e. can

yield bounds for all ϵ. See, for example, Sections 4.7 and 9.3 in which Θ_0 is a hypercube. This can be enough if all that is sought is a bound on the rate of convergence.

(e) Containment with optimization. Given a family of spaces $\Theta_\gamma \subset \Theta$, optimize the choice of γ: $R_N(\Theta) \geq \sup_\gamma R_N(\Theta_\gamma)$. This is used for Besov bodies (Section 9.9), ℓ_p balls (Section 11.6), and in the 'Besov shell' method of Section 10.8.

(f) Comparison of loss functions or models. If $L(a,\theta) \geq L'(a,\theta)$ either everywhere or with π-probability one for suitable priors π, then it may be easier to develop bounds using L', for example via Bayes risk calculations. This strategy is used in Section 10.4 with ℓ_q loss functions and in Section 8.6 below (8.64). A variant of the comparison strategy appears in Section 4.9 in which an ordering of covariance matrices implies an ordering of risks, and in Section 15.3 in which a discrete sampling model is compared with a continuous white noise model.

(g) Generic reduction to testing/classification. In this approach, a finite subset $\Theta_F \subset \Theta$ is chosen so that every pair of points satisfies $\|\theta_i - \theta_j\| \geq 2\delta$. If $w(\cdot)$ is increasing, then

$$\inf_{\tilde\theta} \sup_{\Theta} E_\theta w(\|\tilde\theta - \theta\|) \geq w(\delta) \inf_{\hat\theta} \max_{\theta_i \in \Theta_F} P_{\theta_i}(\hat\theta \neq \theta_i),$$

where the right-side infimum is over estimators taking values in Θ_F. Thus the estimation problem has been reduced to a classification error problem, which might, for example, be bounded by a version of Fano's lemma (e.g. Cover and Thomas (1991)). This common strategy is used here only in Section 5.5–where more details are given–where there is no special structure on Θ.

4.11 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 ϵ, which is easy after exploiting 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 Θ in the sense that point masses $\delta_\theta \in \mathcal{M}$ for $\theta \in \Theta$. Then, as we have seen at (4.18) and (4.17),

$$R_N(\Theta,\epsilon) \leq B(\mathcal{M},\epsilon) = \sup_{\pi\in\mathcal{M}} B(\pi). \tag{4.63}$$

We call the right side the *Bayes-minimax* risk. Often $\mathcal{M}$ is defined by constraints on marginal moments and in general $\mathcal{M}$ will *not* be supported on Θ. For example, if $\Theta(C)$ is the ellipsoid defined by $\sum a_i^2\theta_i^2 \leq C^2$, then we might use $\mathcal{M}(C) = \{\pi(d\theta) : \sum a_i^2 E_\pi \theta_i^2 \leq C^2\}$.

The idea is that a judiciously chosen relaxation of the constraints defining Θ may make the problem easier to evaluate, and yet still be asymptotically equivalent to Θ 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 ν_ϵ *supported in* Θ, that is nearly least favorable: $B(\nu_\epsilon) \sim B(\mathcal{M},\epsilon)$. Then asymptotic equivalence would follow from the chain of inequalities

$$B(\nu_\epsilon) \leq R_N(\Theta,\epsilon) \leq B(\mathcal{M},\epsilon) \sim B(\nu_\epsilon). \tag{4.64}$$

(b) Asymptotic Concentration. Often it is inconvenient to work directly with priors supported on Θ. 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 eventually *concentrates* on Θ:

$$\pi_\epsilon(\Theta) \to 1. \tag{4.65}$$

If one then constructs the conditioned prior $\nu_\epsilon = \pi_\epsilon(\cdot|\Theta)$ and additionally shows that

$$B(\nu_\epsilon) \sim B(\pi_\epsilon), \tag{4.66}$$

then asymptotic equivalence follows by replacing the last similarity in (4.64) by $B(\mathcal{M}, \epsilon) \sim B(\pi_\epsilon) \sim B(\nu_\epsilon)$.

There are significant details to fill in, which vary with the specific application. We try to sketch some of the common threads of the argument here, noting that some changes may be needed in each setting. There is typically a nested *family* of minimax problems with parameter space $\Theta(C)$ depending on C, so that $C < C'$ implies that $\Theta(C) \subset \Theta(C')$. Often, but not always, C will be a scale parameter: $\Theta(C) = C\Theta(1)$. We assume also that the corresponding prior family is similarly nested. Let $R(C, \epsilon) \le B(C, \epsilon)$ denote the frequentist and Bayes minimax risks over $\Theta(C)$ and $\mathcal{M}(C)$ respectively. We exploit the nesting structure by taking π_ϵ as the least favorable prior for $B(\gamma C, \epsilon)$ for some $\gamma < 1$. Although π_ϵ 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. The setting is ℓ_2 loss, but the argument can easily be generalized, at least to other additive norm based loss functions. Since C remains fixed, set $\Theta = \Theta(C)$. Let π_ϵ be a prior distribution with $B(\pi_\epsilon) \ge \gamma B(\gamma C, \epsilon)$ and $\pi_\epsilon(\Theta) > 0$. Set $\nu_\epsilon = \pi_\epsilon(\cdot|\Theta)$, and let $\hat\theta_{\nu_\epsilon}$ be the Bayes estimator of θ for the conditioned prior ν_ϵ. The task is to relate $B(\nu_\epsilon)$ to $B(\pi_\epsilon)$. From the frequentist definition of Bayes risk $B(\pi_\epsilon) \le B(\hat\theta_{\nu_\epsilon}, \pi_\epsilon)$, and so

$$\begin{aligned} B(\pi_\epsilon) &\le E_{\pi_\epsilon}\{||\hat\theta_{\nu_\epsilon} - \theta||^2|\Theta\}\pi_\epsilon(\Theta) + \mathbb{E}_{\pi_\epsilon}\{||\hat\theta_{\nu_\epsilon} - \theta||^2, \Theta^c\} \\ &\le B(\nu_\epsilon)\pi_\epsilon(\Theta) + 2\mathbb{E}_{\pi_\epsilon}\{||\hat\theta_{\nu_\epsilon}||^2 + ||\theta||^2, \Theta^c\}. \end{aligned} \tag{4.67}$$

Here and below we use $\mathbb{E}_{\pi_\epsilon}$ to denote expectation over the joint distribution of (θ, y) when prior π_ϵ is used. Since ν is concentrated on Θ, $B(\nu_\epsilon) \le R(C, \epsilon)$, and on putting everything together, we have

$$\gamma B(\gamma C, \epsilon)(1 + o(1)) \le B(\pi_\epsilon) \le R(C, \epsilon)\pi_\epsilon(\Theta) + 2\mathbb{E}_{\pi_\epsilon}\{\|\hat\theta_{\nu_\epsilon}\|^2 + ||\theta||^2, \Theta^c\}.$$

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

Lemma 4.32 *Suppose that for each $\gamma < 1$ one chooses $\pi_\epsilon \in \mathcal{M}(\gamma C)$ such that, as $\epsilon \to 0$,*

$$B(\pi_\epsilon) \ge \gamma B(\gamma C, \epsilon)(1 + o(1)), \tag{4.68}$$

$$\pi_\epsilon(\Theta) \to 1, \tag{4.69}$$

$$\mathbb{E}_{\pi_\epsilon}\{\|\hat\theta_{\nu_\epsilon}\|^2 + ||\theta||^2, \Theta^c\} = o(B(\gamma C, \epsilon)). \tag{4.70}$$

Then for each such γ,

$$R(C, \epsilon) \ge \gamma B(\gamma C, \epsilon)(1 + o(1)). \tag{4.71}$$

Often the function $B(\gamma C, \epsilon)$ will have sufficient regularity that one can easily show

$$\lim_{\gamma \nearrow 1} \liminf_{\epsilon \to 0} \frac{B(\gamma C, \epsilon)}{B(C, \epsilon)} = 1. \tag{4.72}$$

See, for example, Exercise 4.8 for the scale family case. In general, combining (4.71) with (4.72), it follows that $R(C, \epsilon) \sim B(C, \epsilon)$.

Remark 4.33 Versions of this approach appear

(i) in the discussion of Pinsker's theorem, where Θ is an ellipsoid, Chapter 5,
(ii) in estimation of η-sparse signals, where Θ is an ℓ_0-ball, Chapter 8,
(iii) and of *approximately* sparse signals, where Θ is an ℓ_p ball, Chapter 13,
(iv) and estimation of functions with spatial inhomogeneity, in which Θ is a Besov ball, Chapter 14.

4.12 Notes

Brown (1971) cites James and Stein (1961) for identity (4.4) but it is often called Brown's identity for the key role it plays in the former paper.

Aside: The celebrated paper of Brown (1971) uses (4.4) and (2.23) (the n-dimensional version of (4.6)) 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).

A careful measure theoretic discussion of conditional distributions is given in Schervish (1995, Appendix B.3). Broad conditions for the Borel measurability of Bayes rules found by minimizing posterior expected loss are given in Brown and Purves (1973).

Brown et al. (2006) gives an alternative proof of the Bayes risk lower bound (4.9), along with many other connections to Stein's identity (2.58). Improved bounds on the Bayes risk are given by Brown and Gajek (1990).

§4. The Bayes minimax risk $B(\mathcal{P})$ introduced here is also called the Γ-minimax risk (where Γ refers to the class of prior distributions) in an extensive literature; overviews and further references may be found in Berger (1985) and Ruggeri (2006).

The primary reference for the second part of this chapter is Donoho et al. (1990), where Theorems 4.17, 4.25 and 9.5 (for the case $\epsilon_i \equiv \epsilon$) may be found. The extension to the heteroscedastic setting given here is straightforward. The short proof of Theorem 4.17 given here relies on a minimax theorem; Donoho et al. (1990) give a direct argument.

More refined bounds in the spirit of the Ibragimov-Hasminskii bound of Theorem 4.17, valid for all $\epsilon > 0$, were derived and applied by Levit (2010a,b).

`[J and MacGibbon?]` A Bayesian version of the I-H bound is given by Vidakovic and DasGupta (1996), who show that the linear Bayes minimax risk for all symmetric and unimodal priors on $[-\tau, \tau]$ as at most 7.4% worse than the exact minimax rule. `[make exercise?]`

It is curious that the limiting least favorable distribution (4.43) found by Bickel (1981), after the transformation $x = \sin(\pi s/2)$, becomes $(2/\pi)\sqrt{1-x^2}dx$, the Wigner semi-circular limiting law for the (scaled) eigenvalues of a real symmetric matrix with i.i.d. entries (e.g. Anderson et al. (2010, Ch. 2)). Local repulsion—of prior support points, and of eigenvalues—is a common feature.

Levit (1980, 1982, 1985) and Berkhin and 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 π^2), could be interpreted as twice the principal eigenvalue of the Laplacian (here $= -2d^2/dt^2$) 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 ϵ or sample size n.

The overview of lower bound methods for nonparametric estimation in Tsybakov (2009, Ch. 2) is accompanied by extensive historical bibliography.

Exercises

4.1 *(Qualitative features of risk of proper Bayes rules.)* Suppose that $y \sim N(\theta, \epsilon^2)$, that θ has a proper prior distribution π, and that $\hat{\theta}_\pi$ is the squared error loss Bayes rule.
(a) Show that $r(\hat{\theta}_\pi, \theta)$ cannot be constant for $\theta \in \mathbb{R}$. [Hint: Corollary 4.10.]
(b) If $E_\pi|\theta| < \infty$, then $r(\hat{\theta}_\pi, \theta)$ is at most quadratic in θ: there exist constants a, b so that $r(\hat{\theta}_\pi, \theta) \le a + b\theta^2$. [Hint: apply the covariance inequality (C.8) to $E_\pi[|\theta - x|\phi(\theta - x)]$.]
(c) Suppose in addition that π is supported in a bounded interval I. Show that $P_\theta(\hat{\theta}_\pi \in I) = 1$ for each θ and hence that $r(\hat{\theta}_\pi, \theta)$ is unbounded in θ, indeed $r(\hat{\theta}_\pi, \theta) \ge c\theta^2$ for suitable $c > 0$.

4.2 *(Proof of van Trees inequality.)* Suppose that $X \sim N(\theta, 1)$ and that the prior π has density $p(\theta)d\theta$. Let $\mathbb{E}$ denote expectation with respect to the joint distribution of (x, θ). Let $A = \hat{\theta}(y) - \theta$; and $B = (\partial/\partial\theta)\log[\phi(y - \theta)\,p(\theta)]$. Show that $\mathbb{E}AB = 1$, and then use the Cauchy-Schwarz inequality to establish (4.9). (Belitser and Levit, 1995)

4.3 *(Fisher information for priors on an interval.)* (a) Consider the family of priors $\pi_\beta(d\theta) = c_\beta(1 - |\theta|)^\beta$. For what values of β is $I(\pi_\beta) \le \infty$?
(b) What is the minimum value of $I(\pi_\beta)$?
(c) Show that ν_∞ in (4.43) minimizes $I(\pi)$ among probability measures supported on $[-1, 1]$.

4.4 *(Cramer-Rao bound and the uncertainty principle.)* Suppose that f is a differentiable, possibly complex valued function with $\int |f|^2 = 1$. Show that the Fisher information bound (4.5) implies

$$\int x^2|f(x)|^2dx \int \xi^2|\widehat{f}(\xi)|^2d\xi \ge \frac{\pi}{2}, \tag{4.73}$$

where $\widehat{f}(\xi) = \int e^{-i\xi x} f(x)dx$. [This is Heisenberg's inequality: see Dym and McKean (1972, p.116-122) for the extension to all $f \in L_2(\mathbb{R})$, not necessarily differentiable, and for some information on the connection of (4.73) with the uncertainty principle of quantum mechanics.]

4.5 *(Truncation of (near) least favorable priors.)* (a) Given a probability measure $\pi(d\theta)$ on $\mathbb{R}$, and M sufficiently large, define the restriction to $[-M, M]$ by $\pi^M(A) = \pi(A||\theta| \le M)$. Show that π^M converges weakly to π as $M \to \infty$.
(b) If π satisfies $\int |\theta|^p d\pi \le \eta^p$, show that π^M does also, for $M \ge \eta$.
(c) Given a class of probability measures $\mathcal{P}$ and $\gamma < 1$, show using Lemma 4.7 that there exists $\pi \in \mathcal{P}$ and M large so that $B(\pi^M) \ge \gamma B(\mathcal{P})$.

4.6 *(continuity properties of ℓ_p loss.)* Consider the loss function $L(a, \theta) = \|a - \theta\|_p^p$ as a function of $\theta \in \ell_2(\mathbb{N})$. Show that it is continuous for $p \ge 2$, while for $p < 2$ it is lower semi-continuous but not continuous.

4.7 *(Pathologies of risk functions.)* For $y \sim N(\theta, 1)$, and squared error loss, show that the (otherwise absurd) estimator $\hat{\theta}(y) = e^{y^2/4}/(1 + y)I\{y > 0\}$ has a risk function which is discontinuous at 0, but still lower semicontinuous.

4.8 *(Scaling and 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 (i) $R_N(\Theta(C); \epsilon)$, and (ii) $R_L(\Theta(C); \epsilon)$.

(a) Suppose that $\epsilon' \neq \epsilon$ and set $C' = (\epsilon'/\epsilon)C$. For each definition of $R(C, \epsilon)$ show that

$$R(C, \epsilon) = (\epsilon/\epsilon')^2 R(C', \epsilon').$$

In particular, of course, if $\epsilon' = 1$,

$$R(C, \epsilon) = \epsilon^2 R(C/\epsilon, 1). \tag{4.74}$$

(b) If both C, ϵ vary, we only get bounds. In each case, show that if $C' \leq C$ and $\epsilon' \leq \epsilon$, then

$$R(C, \epsilon) \leq (C/C')^2 (\epsilon/\epsilon')^2 R(C', \epsilon'),$$

and that if $\mathcal{P}(C) = C\mathcal{P}(1)$ is a scale family of priors, that the same result holds for $B(C, \epsilon) = B(\mathcal{P}(C); \epsilon)$.
(c) Conclude that

$$\lim_{\gamma \to 1} \liminf_{\epsilon \to 0} \frac{B(\gamma C, \epsilon)}{B(C, \epsilon)} = 1.$$

4.9 *(Direct argument for minimaxity on products).* In the setting of Proposition 4.16, suppose that $(\hat{\theta}_j^*, \theta_j^\circ)$ is a saddle-point in the j-th problem. Let $\hat{\theta}^*(y) = (\hat{\theta}_j^*(y_j))$ and $\theta^\circ = (\theta_j^\circ)$. Show without using priors that $(\hat{\theta}^*, \theta^\circ)$ is a saddle-point in the product problem.

4.10 (*Taking the interval constraint literally.*) Recall that if $Y \sim N(\theta, \epsilon^2)$, we defined $\rho_L(\tau, \epsilon) = \inf_{\hat{\theta}_c} \sup_{\theta \in [-\tau, \tau]} E[\hat{\theta}_c(Y) - \theta]^2$, for linear estimators $\hat{\theta}_c(Y) = cY$. An awkward colleague complains "it is nonsensical to study $\rho_L(\tau, \epsilon)$ since no estimator $\hat{\theta}_c$ in the class is sure to satisfy the constraint $\hat{\theta} \in [-\tau, \tau]$." How might one reply?

4.11 *(Bounded normal mean theory for L_1 loss.)* Redo the previous question for $L(\theta, a) = |\theta - a|$. In particular, show that

$$\hat{\theta}_\pi(y) = \tau \text{ sgn } y, \qquad \text{and} \qquad B(\pi_\tau) = 2\tau\tilde{\Phi}(\tau),$$

where, as usual $\tilde{\Phi}(\tau) = \int_\tau^\infty \phi(s)ds$. In addition, show that

$$\mu^* = \sup_{\tau, \epsilon} \frac{\rho_L(\tau, \epsilon)}{\rho_N(\tau, \epsilon)} \leq \frac{1}{B(\pi_1)} \doteq 1/.32 < \infty.$$

Hint: show that $\rho_L(\tau, 1) \leq \rho_P(\tau, 1) = \min(\tau, \sqrt{2/\pi})$.

4.12 *(Continued.)* For L_1 loss, show that (a) $\rho_N(\tau, \epsilon) = \epsilon\rho_N(\tau/\epsilon, 1)$ is increasing in τ, and (b) $\lim_{\tau \to \infty} \rho_N(\tau, \epsilon) = \epsilon\gamma_0$, where $\gamma_0 = E_0|z| = \sqrt{2/\pi}$.
[Hint: for (b) consider the uniform prior on $[-\tau, \tau]$.]

4.13 *(Discrete prior spacing and risk functions.)* This exercise provides some direct support for the claim before Figure 4.2 that a risk function bounded by ϵ^2 forces a discrete prior to have atoms spaced at most $O(\epsilon)$ apart. To simplify, consider $\epsilon = 1$.
(a) Show that for any estimator $\hat{\mu}(x) = x + g(x)$ that if $|g(x)| \geq M$ for $x \in K$, then

$$r(\hat{\mu}, \mu) \geq M^2 P_\mu(K) - 2\sqrt{M}.$$

(b) Again for simplicity consider a two point prior, which may as well be taken as $\pi(d\mu) = \pi_0\delta_{-\mu_0} + \pi_1\delta_{\mu_0}$ with $\pi_1 = 1 - \pi_0$. Show that the posterior mean

$$\hat{\mu}_\pi(x) = \mu_0 \frac{\pi_0 e^{\mu_0 x} - \pi_1 e^{-\mu_0 x}}{\pi_0 e^{\mu_0 x} + \pi_1 e^{-\mu_0 x}}.$$

(c) Consider first $\pi_0 = 1/2$ and argue that there exists $a > 0$ such that if μ_0 is large, then for some $|\mu| < \mu_0$

$$r(\hat{\mu}_\pi, \mu) > a\mu_0^2. \tag{4.75}$$

(d) Now suppose that $\pi_0 = 1/2 - \gamma$ and show that (4.75) still holds for $a = a(\gamma)$.

4.14 *(Hyperrectangles, exponential decay and domain of analyticity.)* Suppose $f(t) = \sum_{-\infty}^{\infty} \theta_k e^{2\pi i k t}$ and consider the associated function $g(z) = \sum_{-\infty}^{\infty} \theta_k z^k$ of the complex variable $z = re^{2\pi i t}$. If $|\theta_k| = O(e^{-a|k|})$, show that g is analytic in the annulus $A_a = \{z : e^{-a} < |z| < e^a\}$. A near converse also holds: if g is analytic in a domain containing $\overline{A_a}$, then $|\theta_k| = O(e^{-a|k|})$. Thus, the larger the value of a, the greater the domain of analyticity.

4.15 *(Minimax affine implies diagonal.)* An *affine* estimator has the form $\hat{\theta}_{C,b}(y) = Cy + b$. Prove the following extension of Proposition 4.20: if $\hat{\theta}_{C,b}$ is minimax among affine estimators over a hyperrectangle $\Theta(\tau)$, then necessarily $b = 0$ and C must be diagonal.

4.16 *(Linear minimaxity on products with a Euclidean factor.)* Adopt the setting of Section 4.8: the model $y_i = \theta_i + \epsilon_i z_i$, (4.46), with $\theta \in \Theta$ orthosymmetric and with squared error loss.
(a) Suppose first that $\Theta = \Theta^\circ \times \Theta'$ with both Θ° and Θ' being solid orthosymmetric. Show that $R_L(\Theta) = R_L(\Theta^\circ) + R_L(\Theta')$. [Hint: start from (4.56).]
(b) If Θ' satisfies the assumptions of Theorem 4.25, i.e. is compact, solid, orthosymmetric and quadratically convex, then show that the conclusion of that theorem applies to $\Theta = \mathbb{R}^k \times \Theta'$: namely $R_L(\Theta) = \sup_{\tau \in \Theta} R_L(\Theta(\tau))$.

4.17 *(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, γ and η are extended to periodic functions of $k \in \mathbb{Z}$, that is $x(k+n) = x(k)$, and so on. Define the *shift* of γ by $(S\gamma)_k = \gamma_{k+1}$. The set Γ 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 $\mathrm{e} = e^{2\pi i/n}$ and note that the discrete Fourier transform $y = \mathcal{F}x$ can be written

$$y_k = \sum_{l=0}^{n-1} \mathrm{e}^{kl} x_l, \qquad k = 0, \ldots, n-1.$$

Similarly, let $\theta = \mathcal{F}\gamma, z = \mathcal{F}\eta$ and $\Theta = \mathcal{F}\Gamma$.
(b) Show that shift-invariance of Γ means that $\theta = (\theta_k) \in \Theta$ implies $M^l\theta = (\mathrm{e}^{lk}\theta_k) \in \Theta$ for $l \in \mathbb{Z}$. In particular, we have $\mathcal{F}S = M^{-1}\mathcal{F}$.
(c) Let $V(\tau) = \{M^l\tau, l \in \mathbb{Z}\}$ denote the *orbit* of τ 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

$$\sup_{\theta \in V(\tau)} r(\hat{\theta}_{C^0,0}, \theta) \le \sup_{\theta \in V(\tau)} r(\hat{\theta}_{C,b}, \theta).$$

Thus, on a translation invariant set Γ, an estimator that is minimax among affine estimators must have diagonal linear form when expressed in the discrete Fourier basis.

4.18 *(No minimax result for projection estimators.)* Show by example that the equality (4.58) fails if c is restricted to $\{0,1\}^I$, the class of projections onto subsets of the co-ordinates.

4.19 *(Linear and diagonal minimax risk in intra-class model.)*
Consider the setting of Example 4.31.

(a) Show that in the basis of the Karhunen-Loève transform, the variances are

$$\varepsilon_1^2 = p\sigma^2 + 1, \qquad \varepsilon_k^2 = 1, \quad k \geq 2.$$

(b) Show that $R_L(\Theta(\tau)) = \sum_i \epsilon_i^2 \tau^2/(\epsilon_i^2+\tau^2)$, and $R_{DL}(\Theta(\tau)) = p(1+\sigma^2)\tau^2/(1+\sigma^2+\tau^2)$.
(c) Derive conclusion (4.62).

5

Linear Estimators and Pinsker's Theorem

Compared to what an ellipse can tell us, a circle has nothing to say. (E. T. Bell).

Under appropriate assumptions, linear estimators have some impressive optimality properties. This chapter uses the optimality tools we have developed to study optimal linear estimators over ellipsoids, which as we have seen capture the notion of mean-square smoothness of functions. In particular, the theorems of Pinsker (1980) are notable for several reasons. The first gives an exact evaluation of the linear minimax risk in the Gaussian sequence model for quadratic loss over general ellipsoids in ℓ_2. The second 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. The results apply to ellipsoids generally, and thus to all levels of Hilbert-Sobolev smoothness, and also to varying noise levels in the co-ordinates, and so might be considered as a crowning result for linear estimation.

The linear minimax theorem can be cast as a simple Lagrange multiplier calculation, Section 5.1. Section 5.2 examines some examples in the white noise model. Ellipsoids of mean square smoothness and of analytic function lead to very different rates of convergence (and constants!). Fractional integration illustrates the use of the linear minimax theorem for inverse problems. Finally, a concrete comparison shows that the right smoothing spline is actually very close in performance to linear minimax rule.

Section 5.3 states the "big" theorem on asymptotic minimax optimality of linear estimators among *all* estimators in the low noise limit. In this section we give a proof for the white noise model with polynomial ellipsoid constraints – this allows a simplified argument in which Gaussian priors are nearly least favorable. The Bayes rules for these Gaussian priors are linear, and are essentially the linear minimax rules, which leads to the asymptotic efficiency.

Section 5.4 gives the proof for the more general case, weaving in ideas from Chapter 4 in order to combine the Gaussian priors with other priors needed for co-ordinates that have especially 'large' or 'small' signal to noise ratios.

The chapter concludes with a diversionary interlude, Section 5.5, that explains why the infinite sequence model requires a compactness assumption for even as weak a conclusion as consistency to be possible in the low noise limit.

5.1 Exact evaluation of linear minimax risk.

In this chapter we consider the non-white Gaussian sequence model,

$$y_i = \theta_i + \epsilon_i z_i, \qquad \epsilon_i > 0, \qquad i \in \mathbb{N}. \tag{5.1}$$

[Recall (3.60), where $\epsilon_i = \varrho_i \epsilon$.] Suppose that Θ is an ellipsoid in $\ell_2(\mathbb{N})$:

$$\Theta = \Theta(a, C) = \{\theta : \sum a_i^2 \theta_i^2 \le C^2\}. \tag{5.2}$$

We will call $\{a_i\}$ a *proper* semi-axis sequence if the a_i are positive and nondecreasing, with $a_i \to \infty$. A pleasant surprise is that there is an explicit solution for the minimax linear estimator over such ellipsoids.

Proposition 5.1 ***[Pinsker]*** *Suppose that the observations follow sequence model* (5.1) *and that* Θ *is an ellipsoid* (5.2) *with proper semi-axis sequence* a_i*. Then the minimax linear risk*

$$R_L(\Theta) = \sum_i \epsilon_i^2 (1 - a_i/\mu)_+, \tag{5.3}$$

where $\mu = \mu(C)$ *is determined by*

$$\sum \epsilon_i^2 a_i (\mu - a_i)_+ = C^2. \tag{5.4}$$

The linear minimax estimator is given by

$$\hat{\theta}_i^*(y) = c_i^* y_i = (1 - a_i/\mu)_+ y_i, \tag{5.5}$$

and is Bayes for a Gaussian prior π_C *having independent components* $\theta_i \sim N(0, \tau_i^2)$ *with*

$$\tau_i^{*2} = \epsilon_i^2 (\mu/a_i - 1)_+. \tag{5.6}$$

If some $a_i = 0$, the result remains true with the proviso that for such co-ordinates $c_i^* = 1$ and θ_i has an improper flat prior.

Some characteristics of the linear minimax estimator (5.5) deserve note. Since the ellipsoid weights a_i are increasing, the shrinkage factors c_i decrease with i and hence downweight the higher "frequencies" more. In addition, there is a *cutoff* at the first index i such that $a_i \ge \mu$: the estimator is zero at frequencies above the cutoff. Finally, the optimal linear estimator depends on all the parameters $C, (\epsilon_i)$, and (a_i)—as they vary, so does the optimal estimator. In particular, the least favorable distributions, determined by the variances τ_i^2 change with changing noise level.

Proof The set Θ is solid, orthosymmetric and quadratically convex. Since $\sup a_i = \infty$ it is also compact. Thus the minimax linear risk is determined by the hardest rectangular subproblem, and from Theorem 4.25,

$$R_L(\Theta) = \sup_{\tau \in \Theta} R_L(\Theta(\tau)) = \sup \Big\{ \sum_i \frac{\epsilon_i^2 \tau_i^2}{\epsilon_i^2 + \tau_i^2} \ : \ \sum a_i^2 \tau_i^2 \le C^2 \Big\}. \tag{5.7}$$

This maximum may be evaluated by forming the Lagrangian

$$\mathcal{L} = \sum_i \Big\{ \epsilon_i^2 - \frac{\epsilon_i^4}{\epsilon_i^2 + \tau_i^2} \Big\} - \frac{1}{\mu^2} \sum_i a_i^2 \tau_i^2.$$

Simple calculus shows that the maximum is attained at τ_i^{*2} given by (5.6). The positive part constraint arises because τ_i^2 cannot be negative. The Lagrange multiplier parameter μ is uniquely determined by the equation $\sum a_i^2 \tau_i^2 = C^2$, which on substitution for τ_i^{*2} yields (5.4). This equation has a unique solution since the left side is a continuous, strictly increasing, unbounded function of μ. The corresponding maximum is then (5.3).

We have seen that the hardest rectangular subproblem is $\Theta(\tau^*)$, with τ^* given by (5.6). The minimax linear estimator for $\Theta(\tau^*)$, recalling (4.28), is given by $\hat{\theta}_i^* = c_i^* y_i$ with

$$c_i^* = \frac{\tau_i^{*2}}{\epsilon_i^2 + \tau_i^{*2}} = \Big(1 - \frac{a_i}{\mu}\Big)_+. \tag{5.8}$$

We now show that $\hat{\theta}^*$ is minimax linear for *all of* Θ. Lemma 3.4 (generalized in the obvious way to model (5.1)) evaluates the maximum risk of $\hat{\theta}^*$ over an ellipsoid (5.2) as

$$\sup_{\theta \in \Theta} r(\hat{\theta}^*, \theta) = \sum_i \epsilon_i^2 c_i^{*2} + C^2 \sup_i a_i^{-2}(1 - c_i^*)^2.$$

From (5.8) it is clear that $a_i^{-1}(1 - c_i^*)$ equals μ^{-1} for all $a_i \le \mu$ and is less than μ^{-1} for $a_i > \mu$. Consequently, using (5.4) and (5.8),

$$C^2 \sup_i a_i^{-2}(1 - c_i^*)^2 = C^2/\mu^2 = \sum_i \epsilon_i^2 \frac{a_i}{\mu}\Big(1 - \frac{a_i}{\mu}\Big)_+ = \sum_i \epsilon_i^2 c_i^*(1 - c_i^*).$$

Combining the last two displays, and recalling (5.3),

$$\sup_{\theta \in \Theta} r(\hat{\theta}^*, \theta) = \sum \epsilon_i^2 c_i^* = R_L(\Theta),$$

which shows that $\hat{\theta}^*$ is indeed minimax linear over Θ. Finally, from (5.8) it is evident that $\hat{\theta}^*$ is Bayes for a prior with independent $N(0, \tau_i^{*2})$ components. □

Remark. The proof of the proposition first uses (5.7), which corresponds to a minimax theorem for the payoff function $r(c, \theta) = r(\hat{\theta}_c, \theta)$, as seen at (4.58). The proof then goes further than the minimax statement (4.58) to exhibit $(\hat{\theta}^*, \tau^*) = (\hat{\theta}_{c^*}, \tau^*)$ as a saddlepoint: the extra argument in the second paragraph shows that for all c and for all $\theta \in \Theta$,

$$r(\hat{\theta}_{c^*}, \theta) \le r(\hat{\theta}_{c^*}, \tau^*) \le r(\hat{\theta}_c, \tau^*).$$

Comparison of ellipsoids. We will want to evaluate approximations to $R_L(\Theta)$ when $\epsilon_i = \epsilon \varrho_i$ and ϵ is small. The next result says that when two ellipsoids are 'close', the corresponding risk approximations are close also.

Proposition 5.2 *Suppose $\{\bar{a}_i\}$ and $\{a_i\}$ are proper semi-axis sequences satisfying $\bar{a}_i \le a_i$ with $\bar{a}_i/a_i \to 1$ as $i \to \infty$. Then for all $C > 0$*

$$R_L(\Theta(\bar{a}, C), \epsilon) \sim R_L(\Theta(a, C), \epsilon) \qquad as \quad \epsilon \to 0.$$

Proof Define $\bar{\theta}_i$ as the positive square root of $\bar{\theta}_i^2 = \epsilon_i^2(\mu/\bar{a}_i - 1)_+$. From Proposition 5.1, this is a least favorable sequence for $\bar{\Theta} = \Theta(\bar{a}, C)$. Hence

$$R_L(\bar{\Theta}, \epsilon) = \sum \epsilon_i^2 (1 - \bar{a}_i/\mu)_+ \le \epsilon^2 [R_\kappa + S_\kappa(\epsilon)] \tag{5.9}$$

for each κ, if we define $R_\kappa = \sum_1^\kappa \varrho_i^2$ and $S_\kappa(\epsilon) = \sum_{\kappa+1}^\infty \varrho_i^2 (1 - \bar{a}_i/\mu_\epsilon)_+$.

Now turn to $\Theta = \Theta(a, C)$, let $r_i = \bar{a}_i/a_i$ and introduce $\theta_i = r_i\bar{\theta}_i$. Since $\bar{\theta} \in \bar{\Theta}$, we have $\theta \in \Theta$. Since $r_i \le 1$, and from (5.8), we have $\bar{\theta}_i^2/(\epsilon^2 + \bar{\theta}_i^2) = (1 - \bar{a}_i/\mu)_+$, so that

$$R_L(\Theta, \epsilon) \ge \sum_k \frac{\epsilon_i^2\theta_i^2}{\epsilon_i^2 + \theta_i^2} \ge \sum_k \frac{\epsilon_i^2 r_i^2 \bar{\theta}_i^2}{\epsilon_i^2 + \bar{\theta}_i^2} = \epsilon^2 \sum_i r_i^2 \varrho_i^2 (1 - \bar{a}_i/\mu_\epsilon)_+. \tag{5.10}$$

Now fix $\delta > 0$ and note that $r_i^2 \ge 1 - \delta$ for $i \ge \kappa = \kappa(\delta)$. Since $\Theta \subset \bar{\Theta}$, we have, after combining (5.9) and (5.10),

$$1 \ge \frac{R_L(\Theta, \epsilon)}{R_L(\bar{\Theta}, \epsilon)} \ge \frac{(1-\delta)S_\kappa(\epsilon)}{R_\kappa + S_\kappa(\epsilon)} \ge \frac{1-\delta}{1+\delta}$$

for all ϵ sufficiently small, since (5.4) implies that $S_\kappa(\epsilon) \to \infty$ as $\epsilon \to 0$. Since δ is arbitrary, the proof is done. □

5.2 Some Examples

Sobolev Ellipsoids.

Consider the white noise case, $\sigma_k^2 \equiv \epsilon^2$ and the mean-square smoothness parameter spaces $\Theta = \check{\Theta}_2^\alpha(C)$ in the trigonometric basis[1] considered in Section 3.1. Thus, the ellipsoid weights satisfy $a_1 = 0$ and $a_{2k} = a_{2k+1} = (2k)^\alpha$ for $\alpha > 0$. Equivalently, $a_k = (2[k/2])^\alpha$. From Proposition 5.2, we get the same asymptotic evaluation of minimax linear risk by considering $\Theta = \Theta_2^\alpha(C)$ with $a_k = k^\alpha$. So, rewrite the condition (5.4) determining μ_ϵ as

$$\mu \sum_{k \in N} a_k - \sum_{k \in N} a_k^2 = C^2/\epsilon^2. \tag{5.11}$$

Here, $N = N(\mu) = \{k : a_k < \mu\} = \{1, \dots, k_\mu\}$ where $k_\mu = [\mu^{1/\alpha}]$. For $p = 0, 1, 2$ we then have integral approximations (compare (3.54))

$$\sum_{k \in N} a_k^p = \sum_{k=1}^{k_\mu} k^\alpha = \frac{\mu^{p+1/\alpha}}{\alpha p + 1} + O(\mu^p).$$

Substituting into (5.11) and solving for μ_ϵ, we obtain

$$\mu_\epsilon^{1/\alpha} = \left(\frac{(\alpha+1)(2\alpha+1)}{\alpha}\frac{C^2}{\epsilon^2}\right)^{1-r} + O(1), \tag{5.12}$$

where, in the usual rate of convergence notation, $r = 2\alpha/(2\alpha+1)$. We finally have

$$\begin{aligned} R_L(\Theta) &= \epsilon^2 \sum_{k \in N} \left(1 - \frac{a_k}{\mu}\right) = \epsilon^2 \left(\mu^{1/\alpha} - \frac{1}{\mu}\frac{\mu^{1+1/\alpha}}{\alpha+1} + O(1)\right) \\ &= \frac{\alpha}{\alpha+1}\epsilon^2\mu_\epsilon^{1/\alpha} + O(\epsilon^2) = \left(\frac{\alpha\epsilon^2}{\alpha+1}\right)^r \left((2\alpha+1)C^2\right)^{1-r} + O(\epsilon^2) \\ &= P_r C^{2(1-r)}\epsilon^{2r} + O(\epsilon^2), \end{aligned} \tag{5.13}$$

[1] For concrete examples we index co-ordinates by k rather than i used in the general theory, in part to avoid confusion with $i = \sqrt{-1}$!

where the *Pinsker constant*

$$P_r = \left(\frac{\alpha}{\alpha+1}\right)^r (2\alpha+1)^{1-r} = \left(\frac{r}{2-r}\right)^r (1-r)^{r-1}.$$

Remarks. 1. As in previous chapters, the rate of convergence ϵ^{2r} depends on the assumed smoothness α: the greater the smoothness, the closer is the rate to the parametric rate ϵ^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 ϵ^2/C^2.

3. The shrinkage weights, $c_k \approx (1-k^\alpha/\mu)_+$ in (5.5) assign weight close to 1 for low frequencies, and cut off at $k \approx \mu^{1/\alpha} \propto (C^2/\epsilon^2)^{1/(2\alpha+1)}$. Thus, the number of frequencies retained is an algebraic power of C/ϵ, decreasing as the smoothness α increases.

Fractional integration

We turn to an example of inverse problems that leads to increasing variances ϵ_k^2 in the sequence model. Consider the noisy indirect observations model

$$Y = Af + \epsilon Z, \tag{5.14}$$

introduced at (3.67), When A is β-fold integration, examples (ii)-(iv) of Section 3.9 showed that the singular values $b_k \sim c_\beta^{-1} k^{-\beta}$ as $k \to \infty$, with relative error $O(1/k)$. The constant $c_\beta = \pi^\beta$ in the trigonometric basis, and equals 1 for the Legendre polynomial basis. So we obtain an example of sequence model (5.1) with $a_k \approx k^\alpha$ as before and $\epsilon_k \approx c_\beta k^\beta \epsilon$. Proposition 5.1 allows evaluation of the minimax mean squared error over $\Theta_2^\alpha(C)$. A calculation similar to that done earlier in this section yields a straightforward extension of (5.13):

$$R_L(\Theta^\alpha(C), \epsilon) \sim P_{r,\beta} C^{2(1-r_\beta)} (c_\beta \epsilon)^{2r_\beta},$$

with $r_\beta = 2\alpha/(2\alpha+2\beta+1)$ and

$$P_{r,\beta} = \left(\frac{\alpha}{\alpha+2\beta+1}\right)^{r_\beta} \frac{(2\alpha+2\beta+1)^{1-r_\beta}}{2\beta+1}.$$

The index β of ill-posedness leads to a reduction in the rate of convergence from $r = 2\alpha/(2\alpha+1)$ in the direct case to $r_\beta = 2\alpha/(2\alpha+2\beta+1)$. When β is not too large, the degradation is not so serious.

Remarks. 1. When the Legendre polynomial basis is used, the reader may ask whether the ellipsoid $\Theta_2^\alpha(C)$ has an interpretation in terms of smoothness. The answer is yes, if *weighted* Hilbert-Sobolev spaces are used, in which the mean square smoothness condition in (3.6) is replaced by

$$\int_0^1 [f^{(\alpha)}(t)]^2 t^\alpha (1-t)^\alpha dt \le L^2,$$

and an analog of Lemma 3.3 holds, with different constants. For details see Domínguez et al. (2011, Thm. 5.1) and references therein.

2. When the trigonometric basis is used one must be careful with the definition of A due to the arbitrary constant of integration. Following Zygmund (2002), consider periodic functions with integral 0. Then one can set $A_1(f)(t) = \int_0^t f(s)ds$, and define A_β for

$\beta \in \mathbb{N}$ by iteration, $A_\beta = A_1^\beta$. For $\beta > 0$ non-integer, if $e_k(t) = e^{2\pi i k t}$ for $k \in \mathbb{Z}$ and $f(t) \sim \sum c_k e_k(t)$ with $c_0 = 0$, one can define $(A_\beta f)(t) \sim \sum_k (2\pi i k)^{-\beta} c_k e_k(t)$, and Zygmund (2002, Vol. II, p. 135) shows that

$$(A_\beta f)(t) = \frac{1}{\Gamma(\beta)} \int_{-\infty}^{t} f(s)(t-s)^{\beta-1} ds.$$

Ellipsoids of analytic functions.

Return to the white noise setting $\epsilon_k^2 \equiv \epsilon^2$. Again consider the trigonometric basis for periodic functions on $[0,1]$, but now with $a_0 = 1$ and $a_{2k} = a_{2k+1} = e^{\alpha k}$, so that $\Theta(a, C) = \{\theta : \sum e^{2\alpha k}(\theta_{2k-1}^2 + \theta_{2k}^2) \le C^2\}$. Since the semiaxes decay exponentially with frequency, these ellipsoids contain only infinitely differentiable (actually, analytic) functions, which are thus much smoother than typical members of the Sobolev classes. See also Exercise 4.14.

We turn to interpretation of the linear minimax solution of Proposition 5.1. For given μ, the sum in (5.4) involves geometric sums like $\sum_1^r e^{\alpha k p} \doteq c_{\alpha,p} e^{\alpha r p}$ for $p = 1$ and 2, compare (3.54), which unlike the Sobolev case are dominated by a single leading term.

To solve for μ, set $\mu = e^{\alpha r}$ and note that the constraint (5.4) may be rewritten as

$$F(r) = \sum_{k \ge 1} e^{\alpha k}(e^{\alpha r} - e^{\alpha k})_+ = C^2/(2\epsilon^2).$$

Restricting r to positive integers, we have $F(r) \doteq e^{2\alpha r}\gamma_\alpha$, with $\gamma_\alpha = c_{\alpha,1} - c_{\alpha,2} > 0$, from which we may write our sought-after solution as $\mu = \beta e^{\alpha r_0}$ for $\beta \in [1, e^\alpha)$ with

$$r_0 = \left\lceil \frac{1}{2\alpha} \log \frac{C^2}{2\gamma_\alpha \epsilon^2} \right\rceil.$$

Now we may write the minimax risk (5.3) as $\epsilon \to 0$ in the form

$$R_L(\Theta, \epsilon) = \epsilon^2 + 2\epsilon^2 \sum_{k=1}^{r_0} \left(1 - \beta^{-1} e^{-\alpha(r_0 - k)}\right).$$

Thus it is apparent that the number of retained frequencies r_0 is logarithmic in signal to noise—as opposed to algebraic, in the Sobolev case—and the smoothing weights $c_k = 1 - \beta^{-1} e^{-\alpha(r_0-k)}$ are very close to 1 except for a sharp decline that occurs near r_0. In particular, the minimax linear risk

$$R_L(\Theta, \epsilon) \sim 2\epsilon^2 r_0 \sim \frac{\epsilon^2}{\alpha} \log \epsilon^{-2}$$

is only logarithmically worse than the parametric rate ϵ^2, and the dependence on $\Theta(a, C)$ comes, at the leading order term, only through the analyticity range α and not via the scale factor C.

The minimax estimator compared with smoothing splines.

Still in the white noise setting, we return to the Sobolev ellipsoid setting to suggest that information derived from study of the minimax linear estimate and its asymptotic behavior

is quite relevant to the smoothing spline estimates routinely computed in applications by statistical software packages. The following discussion, which expands on remarks at the end of Section 4.8, is inspired by Carter et al. (1992).

We have seen in Chapter 3 that the Lagrange multiplier form of smoothing spline problem in the sequence model has form (3.42) with solution

$$\hat{\theta}^{SS}_{\lambda,k} = (1 + \lambda a_k^2)^{-1} y_k,$$

if we choose weights $w_k = a_k^2$ corresponding to the ellipsoid (5.2). This should be compared with the linear minimax solution of (5.5), which we write as

$$\hat{\theta}^{M}_{\mu,k} = (1 - a_k/\mu)_+ y_k.$$

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:

$$\bar{r}(\hat{\theta}; \epsilon) = \sup_{\theta \in \Theta_2^\alpha(C)} r(\hat{\theta}, \theta; \epsilon).$$

Thus we will take $\hat{\theta}$ to be either $\hat{\theta}^{SS}_\lambda$ or θ^M_μ. First, however, it is necessary to specify the order of smoothing spline: we take the weights equal to the (squared) ellipsoid weights: $w_k = a_k^2$, thus $w_{2k} = w_{2k+1} = (2k)^{2\alpha}$. When α is a positive integer m, this corresponds to a roughness penalty $\int (D^m f)^2$. We also need to specify the value of the regularization parameter λ, respectively μ, to be used in each case. A reasonable choice is the optimum, or *minimax* value

$$\lambda_* = \operatorname*{argmin}_\lambda \bar{r}(\hat{\theta}_\lambda; \epsilon).$$

Here λ_* is shorthand for, respectively, λ_{SS}, the minimax value for the spline family, and $\lambda_M = \mu_M^{-2}$, that for the minimax family. This is exactly the calculation done in Chapter 3 at (3.58) and (3.94), p. 103, for the spline $\hat{\theta}^{SS}_\lambda$ and minimax $\hat{\theta}^M_\mu$ families respectively . [Of course, the result for the minimax family must agree with (5.13)!] In both cases, the solutions took the form, again with $r = 2\alpha/(2\alpha + 1)$,

$$\lambda_* \sim (c_1 \epsilon^2 / C^2)^r, \qquad \bar{r}(\hat{\theta}_{\lambda_*}, \epsilon) \sim c_2 e^{H(r)} C^{2(1-r)} \epsilon^{2r}, \tag{5.15}$$

where the binary entropy function $H(r) = -r \log r - (1 - r) \log(1 - r)$ and

$$\begin{aligned} c_1^{SS} &= 2v_\alpha/\alpha, & c_2^{SS} &= v_\alpha^r / 4^{1-r}, & v_\alpha &= (1 - 1/2\alpha)/\operatorname{sinc}(1/2\alpha), \\ c_1^{M} &= \tfrac{1}{2}\bar{v}_\alpha/\alpha, & c_2^{M} &= \bar{v}_\alpha^r, & \bar{v}_\alpha &= 2\alpha^2/(\alpha + 1)(2\alpha + 1). \end{aligned}$$

Thus the methods have the same dependence on noise level ϵ and scale C, with differences appearing only in the coefficients. We may therefore summarize the comparison through the ratio of maximum mean squared errors. Remarkably, the low noise smoothing spline maximal MSE turns out to be only negligibly larger than the minimax linear risk of the

Pinsker estimate. Indeed, for $\Theta = \Theta_2^\alpha(C)$, using (5.15), we find that as $\epsilon \to 0$,

$$\frac{R_{SS}(\Theta,\epsilon)}{R_L(\Theta,\epsilon)} \sim \left(\frac{v_\alpha}{\bar{v}_\alpha}\right)^r \left(\frac{1}{4}\right)^{1-r} \doteq \begin{cases} 1.083 & \alpha = 2 \\ 1.055 & \alpha = 4 \\ \to 1 & \alpha \to \infty. \end{cases} \tag{5.16}$$

Similarly, we may compare the asymptotic choices of the smoothing parameter:

$$\frac{\lambda_{SS}}{\lambda_M} \sim \left(\frac{4v_\alpha}{\bar{v}_\alpha}\right)^r \doteq \begin{cases} 4.331 & \alpha = 2 \\ 4.219 & \alpha = 4 \\ \to 4 & \alpha \to \infty, \end{cases}$$

and so λ_{SS} is approximately four times λ_M and this counteracts the lesser shrinkage of smoothing splines noted earlier.

Furthermore, in the discrete smoothing spline setting of Section 3.4, Carter et al. (1992) present small sample examples in which the efficiency loss of the smoothing spline is even less than these asymptotic values, see also Exercise 5.4. 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.

5.3 Pinsker's Asymptotic Minimaxity Theorem

We return to the general sequence model $y_i = \theta_i + \epsilon_i z_i$, where, for asymptotic analysis, we introduce a small parameter ϵ via

$$\epsilon_i = \epsilon \varrho_i .$$

We make two assumptions on the ellipsoid weights (a_i) and noise variances (ϵ_i^2):

(i) a_i are positive and nondecreasing with $\sup_i a_i = \infty$, and
(ii) as $\mu \to \infty$, the ratio

$$\eta^2(\mu) = \max_{a_i \le \mu} \epsilon_i^2 \Big/ \sum_{a_i \le \mu/2} \epsilon_i^2 \to 0. \tag{5.17}$$

Theorem 5.3 (Pinsker) *Assume that (y_i) follows sequence model* (5.1) *with noise levels (ϵ_i). Let $\Theta = \Theta(a, C)$ be an ellipsoid* (5.2) *defined by weights (a_i) and radius $C > 0$. Assume that the weights satisfy conditions (i) and (ii). Then, as $\epsilon \to 0$,*

$$R_N(\Theta,\epsilon) = R_L(\Theta,\epsilon)(1 + o(1)). \tag{5.18}$$

Thus the linear minimax estimator (5.5) *is asymptotically minimax among all estimators.*

Remarks. 1. The hardest rectangular subproblem results of Section 4.8 say that $R_L(\Theta;\epsilon) \le 1.25 R_N(\Theta;\epsilon)$, but this theorem asserts that, in the low noise limit, linear estimates cannot be beaten over ellipsoids, being fully efficient.

2. The condition that $\sup a_i = \infty$ is equivalent to compactness of Θ in ℓ_2, recall Exercise 3.1. In Section 5.5, it is shown for the white noise model that if Θ is not compact, then $R_N(\Theta,\epsilon)$ does not even approach 0 as $\epsilon \to 0$.

3. If $a_1, \ldots, a_r = 0$, the result still holds: $\Theta = \mathbb{R}^r \times \Theta'$ for an ellipsoid Θ', and the factor $\mathbb{R}^r$ adds an asymptotically neglible term $\epsilon^2 \sum_1^r \varrho_i^2$ to both minimax risks.

4. In the white noise model, $\epsilon_i = \epsilon\lambda$, condition (ii) follows from (i). More generally, condition (ii) rules out exponential growth of ϵ_i^2, however it is typically satisfied if ϵ_i^2 grows polynomially with i.

5. Pinsker's proof is actually for an even more general situation. We aim to give the essence of Pinsker's argument in somewhat simplified settings.

The approach and a special case

The approach is to construct a family of priors, indexed by ϵ, that has Bayes risk comparable to the minimax linear risk as $\epsilon \to 0$. Indeed, dropping explicit reference to Θ, we know from Chapter 4 that

$$R_L(\epsilon) \geq R_N(\epsilon) = \sup\{B(\pi) : \text{supp } \pi \subset \Theta\},$$

so that if we can construct a family of priors $\pi_\epsilon \subset \Theta$ for which

$$\liminf_{\epsilon \to 0} \; B(\pi_\epsilon)/R_L(\epsilon) \geq 1, \tag{5.19}$$

then it must be that $R_N(\epsilon) \sim R_L(\epsilon)$ as $\epsilon \to 0$.

We give first a proof under some relatively restricted conditions:

(i') (polynomial growth) $b_1 i^\alpha \leq a_i \leq b_2 i^\alpha$ for positive constants b_1, b_2 and α.
(ii') (white noise) $\epsilon_i \equiv \epsilon$,

Under these conditions we give a proof based on Gaussian priors. Although special, the conditions do cover the Sobolev ellipsoid and spline examples in the previous section, and give the flavor of the argument in the more general setting.

Pinsker's *linear* minimax theorem provides, for each ϵ, a Gaussian prior with independent co-ordinates $\theta_i \sim N(0, \tau_{i\epsilon}^2)$ where $\tau_{i\epsilon}^2 = \epsilon^2(\mu_\epsilon/a_i - 1)_+$ and the Lagrange multiplier μ_ϵ satisfies $\sum_i a_i(\mu_\epsilon - a_i)_+ = C^2/\epsilon^2$. Since the sequence $(\tau_{i\epsilon}^2)$ maximizes (5.7), we might call this the least favorable *Gaussian* prior. It cannot be least favorable among all priors, in the sense of Section 4.3, for example because it is not supported on Θ. Indeed, for this prior

$$E \sum a_i^2 \theta_i^2 = \sum a_i^2 \tau_{i\epsilon}^2 = C^2, \tag{5.20}$$

so that the ellipsoid constraint holds only in mean. However, we will show under our restricted conditions that a modification is indeed asymptotically concentrated on Θ, and implements the heuristics described above. The modification is made in two steps. First, define a Gaussian prior with slightly shrunken variances:

$$\pi_\epsilon^G : \;\; \theta_i \sim N\big(0, (1-\kappa_\epsilon)\tau_{i\epsilon}^2\big)\,, \tag{5.21}$$

with $\kappa_\epsilon \searrow 0$ to be specified. We will show that $\pi_\epsilon^G(\Theta) \to 1$ and so for the second step it makes sense to obtain a prior supported on Θ by conditioning

$$\pi_\epsilon(A) := \pi_\epsilon^G(A|\theta \in \Theta).$$

The idea is then to show that these π_ϵ satisfy (5.19).

Comparing Gaussian and conditioned priors. We can do calculations easily with π_ϵ^G since it is Gaussian, but we are ultimately interested in $\pi_\epsilon(\cdot)$ and its Bayes risk $B(\pi_\epsilon)$. We need to show that they are close, which we expect because $\pi_\epsilon^G(\Theta) \approx 1$.

Let $\mathbb{E}$ denote expectation under the joint distribution of (θ, y) when $\theta \sim \pi_\epsilon^G$. Let $\hat{\theta}_\epsilon$ denote the Bayes rule for prior π_ϵ, so that $\hat{\theta}_\epsilon = \mathbb{E}[\theta|\Theta, y]$. The connection between π_ϵ^G and π_ϵ is captured in

Lemma 5.4

$$(1-\kappa_\epsilon)R_L(\epsilon) \le B(\pi_\epsilon) + \mathbb{E}[\|\hat{\theta}_\epsilon - \theta\|^2, \Theta^c]. \tag{5.22}$$

Proof The argument is like that leading to (4.67) in Section 4.11. Since π_ϵ^G consists of co-ordinates θ_i independently distributed as $N(0, (1-\kappa_\epsilon)\tau_i^2)$, the Bayes risk is a sum of univariate terms:

$$B(\pi_\epsilon^G) = \sum \rho_L(\sqrt{1-\kappa_\epsilon}\tau_i, \epsilon) \ge (1-\kappa_\epsilon)\sum \rho_L(\tau_i, \epsilon) = (1-\kappa_\epsilon)R_L(\epsilon). \tag{5.23}$$

Therefore any estimator $\hat{\theta}$ satisfies

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 = B(\hat{\theta}, \pi_\epsilon^G) \ge B(\pi_\epsilon^G) \ge (1-\kappa_\epsilon)R_L(\epsilon). \tag{5.24}$$

There is also a decomposition

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 = \mathbb{E}[\|\hat{\theta} - \theta\|^2|\Theta]\pi_\epsilon^G(\Theta) + \mathbb{E}[\|\hat{\theta} - \theta\|^2, \Theta^c].$$

If for $\hat{\theta}$ we take the Bayes rule for π_ϵ, namely $\hat{\theta}_\epsilon$, then by definition $\mathbb{E}[\|\hat{\theta}_\epsilon - \theta\|^2|\Theta] = B(\pi_\epsilon)$. Now, simply combine this with the two previous displays to obtain (5.22). □

Now a bound for the 'remainder' term in (5.22), using the ellipsoid structure.

Lemma 5.5 $\mathbb{E}[\|\hat{\theta}_\epsilon - \theta\|^2, \Theta^c] \le c\pi_\epsilon^G(\Theta^c)^{1/2}C^2$.

Proof First observe that by definition $\hat{\theta}_{\epsilon,i} = \mathbb{E}[\theta_i|\Theta, y]$ and so

$$\mathbb{E}\hat{\theta}_{\epsilon,i}^4 \le \mathbb{E}\mathbb{E}[\theta_i^4|\Theta, y] = \mathbb{E}\theta_i^4 = 3\tau_i^4,$$

since $\theta_i \sim N(0, \tau_i^2)$. Consequently $\mathbb{E}(\hat{\theta}_{\epsilon,i} - \theta_i)^4 \le 2^4\mathbb{E}\theta_i^4 = c^2\tau_i^4$. Applying the Cauchy-Schwarz inequality term by term to $\|\hat{\theta}_\epsilon - \theta\|^2 = \sum_i(\hat{\theta}_{\epsilon,i} - \theta_i)^2$, we find

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}_\epsilon - \theta\|^2, \Theta^c] &\le \pi_\epsilon^G(\Theta^c)^{1/2}\sum_i[\mathbb{E}(\hat{\theta}_{\epsilon,i} - \theta_i)^4]^{1/2} \\ &\le c\pi_\epsilon^G(\Theta^c)^{1/2}\sum_i \tau_i^2.\end{aligned} \tag{5.25}$$

Let $a_{\min} = \min a_i$, and the bound $\sum \tau_i^2 \le a_{\min}^{-2}\sum_i a_i^2\tau_i^2 = a_{\min}^{-2}C^2$ completes the proof. □

Putting together the two lemmas, we have

$$B(\pi_\epsilon) \ge (1-\kappa_\epsilon)R_L(\epsilon) - c\pi_\epsilon^G(\Theta^c)^{1/2}C^2, \tag{5.26}$$

and so for (5.19) it remains to show that for suitable $\kappa_\epsilon \to 0$, we also have $\pi_\epsilon^G(\Theta^c) \to 0$ sufficiently fast.

π_ϵ^G *concentrates on* Θ. Under the Gaussian prior (5.21), $E\sum a_i^2\theta_i^2 = (1-\kappa_\epsilon)C^2$ and so the complementary event

$$\Theta^c = \Big\{\theta : \sum a_i^2(\theta_i^2 - E\theta_i^2) > \kappa_\epsilon C^2\Big\}.$$

We may write $a_i^2\theta_i^2$ as $\beta_i Z_i^2$ in terms of independent standard Gaussians Z_i with

$$\beta_i = (1-\kappa_\epsilon)a_i^2\tau_{i\epsilon}^2 = (1-\kappa_\epsilon)\epsilon^2 a_i(\mu_\epsilon - a_i)_+.$$

Now apply the concentration inequality for weighted χ^2 variates, (2.77), to obtain

$$\pi_\epsilon^G(\Theta^c) \le \exp\{-t^2/(32\|\beta\|_1\|\beta\|_\infty\},$$

with $t = \kappa_\epsilon C^2$. Now use (5.20) and the bound $x(1-x) \le 1/4$ to obtain

$$\|\beta\|_1 \le \sum a_i^2\tau_{i\epsilon}^2 = C^2, \qquad \|\beta\|_\infty \le \epsilon^2\mu_\epsilon^2/4.$$

Consequently the denominator $32\|\beta\|_1\|\beta\|_\infty \le 8C^2(\epsilon\mu_\epsilon)^2$.

We now use the polynomial growth condition (i') to bound μ_ϵ. The calculation is similar to that in the Sobolev ellipsoid case leading to (5.12); Exercise 5.5 fills in the details. The result is that, for a constant c_α now depending on α, b_1 and b_2, and $r = 2\alpha/(2\alpha+1)$ and changing at each appearance,

$$\epsilon\mu_\epsilon \le c_\alpha C^r\epsilon^{1-r},$$

We conclude from the concentration inequality that

$$\pi_\epsilon^G(\Theta^c) \le \exp\{-c_\alpha\kappa_\epsilon^2(C/\epsilon)^{2(1-r)}\},$$

and hence that if κ_ϵ is chosen of somewhat larger order than ϵ^{1-r} while still approaching zero, say $\kappa_\epsilon \propto \epsilon^{1-r-\delta}$, then $\pi_\epsilon^G(\Theta^c)^{1/2} \le \exp(-c_\alpha/\epsilon^{2\delta}) = o(R_L(\epsilon))$ as required.

Remark 5.6 Our special assumptions (i') and (ii') were used only in the concentration inequality argument to show that $\pi_\epsilon^G(\Theta^c) \ll R_L(\epsilon)$. The "bad bound" comes at the end of the proof of Lemma 5.5: if instead we were able to replace $\sum\tau_i^2 \le a_{\min}^{-2}C^2$ by a bound of the form $\sum\tau_i^2 \le cR_L(\epsilon)$, then it would be enough to show $\pi_\epsilon^G(\Theta^c) \to 0$. For the bound to be in terms of $R_L(\epsilon)$, it is necessary to have the τ_i comparable in magnitude, see (5.34) below. This can be achieved with a separate treatment of the very large and very small values of τ_i, and is the approach taken in the general case, to which we now turn.

5.4 General case proof*

This section establishes Pinsker's theorem 5.3 under the original assumptions (i) and (ii), following the original argument of Pinsker.

There are three ways in which asymptotic equivalence of linear and non-linear estimates can occur. The first two are essentially univariate, and rely on the equivalence established at (4.41):

$$\frac{\rho_N(\tau,\epsilon)}{\rho_L(\tau,\epsilon)} \to 1 \qquad \text{as } \tau/\epsilon \to 0 \text{ or } \infty.$$

The third situation, covering intermediate values of τ/ϵ, exploits high-dimensionality in

a critical way. It uses a Gaussian prior, for which the optimal estimator is linear. As we have seen in the special case considered in the last section, a concentration of measure property guarantees, as dimensionality grows, that such a prior is essentially supported on an appropriate ellipsoid.

Pinsker's proof handles the three modes simultaneously. The first step is to define a partition of indices $i \in \mathbb{N}$ into three sets N_s, N_g and N_b (with the mnemonics "**s**mall", "**g**aussian" and "**b**ig"), with the co-ordinate signal-to-noise ratios $\tau_{i\epsilon}^2/\epsilon_i^2$ determined by (5.6). The partition depends on a parameter $q > 1$ and declares that

$$i \in \quad N_s, \qquad N_g, \qquad N_b,$$

according as

$$\tau_{i\epsilon}^2/\epsilon_i^2 \in \quad [0, q^{-1}], \qquad (q^{-1}, q), \qquad [q, \infty), \tag{5.27}$$

which is seen for $\tau_i^2/\epsilon_i^2 = (\mu_\epsilon/a_i - 1)_+$ to be equivalent to

$$a_i \in \quad \Big[\frac{q\mu_\epsilon}{q+1}, \infty\Big], \quad \Big(\frac{\mu_\epsilon}{q+1}, \frac{q\mu_\epsilon}{q+1}\Big), \quad \Big(0, \frac{\mu_\epsilon}{q+1}\Big]. \tag{5.28}$$

Of course, the sets N_m, for $m \in \{s, g, b\}$, depend on ϵ and q.

Example: Sobolev ellipsoids (white noise case) continued. It turns out that each of the regimes "**b**", "**g**" and "**s**" occurs for a large range of indices i even in this canonical case. Indeed, recall from (5.12) that $\mu_\epsilon = c_\alpha(C/\epsilon)^{2\alpha/(2\alpha+1)}$. If we use the fact that $a_k \sim k^\alpha$, it is easy to see, for example, that

$$|N_g| \doteq \frac{q^{1/\alpha} - 1}{(q+1)^{1/\alpha}} c_\alpha^{1/\alpha} (C^2/\epsilon^2)^{1-r} \to \infty,$$

with similar expressions for $|N_b|$ and $|N_s|$ that also increase proportionally to $(C^2/\epsilon^2)^{1-r}$.

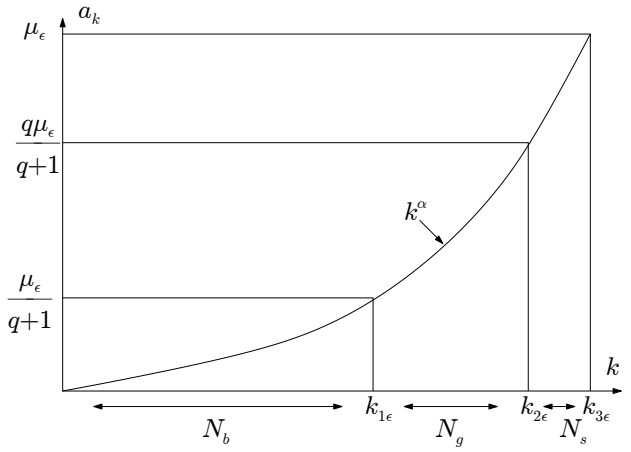

Figure 5.1 The "big", "gaussian" and "small" signal to noise regimes for Sobolev ellipsoids

Definition of priors $\pi = \pi(\epsilon, q)$. A key role is played by the minimax prior variances $\tau_{i\epsilon}^2$ found in Proposition 5.1. We first use them to build sub-ellipsoids Θ_s, Θ_b and Θ_g, defined for $m \in \{s, b, g\}$ by

$$\Theta_m = \Theta_m(\epsilon, q) = \{(\theta_i, i \in N_m) : \sum_{N_m} a_i^2 \theta_i^2 \le \sum_{N_m} a_i^2 \tau_{i\epsilon}^2\}.$$

Since $\sum a_i^2 \tau_{i\epsilon}^2 = C^2$, we clearly have $\Theta_s \times \Theta_g \times \Theta_b \subset \Theta$. We now define priors $\pi_{m\epsilon} = \pi_m(\epsilon, q)$ supported on Θ_m, see also Figure 5.2:

$\pi_{s\epsilon}$: for $i \in N_s$, set $\theta_i \overset{ind}{\sim} \pi_{\tau_{i\epsilon}}$, the two point priors at $\pm\tau_{i\epsilon}$,

$\pi_{b\epsilon}$: for $i \in N_b$, set $\theta_i \overset{ind}{\sim} \pi^V_{\tau_{i\epsilon}}$, cosine priors on $[-\tau_{i\epsilon}, \tau_{i\epsilon}]$, with density $\tau_{i\epsilon}^{-1} \cos^2(\pi\theta_i/2\tau_{i\epsilon})$, recall (4.43),

$\pi_{g\epsilon}$: for $i \in N_g$, first define π^G_ϵ, which sets $\theta_i \overset{ind}{\sim} N(0, (1-\kappa)\tau_{i\epsilon}^2)$ for some fixed $\kappa \in (0, 1)$. Then define $\pi_{g\epsilon}$ by conditioning:

$$\pi_{g\epsilon}(A) = \pi^G_\epsilon(A|\theta \in \Theta_g).$$

While the "Gaussian" components prior π^G_ϵ is not supported in Θ_g, we shall see that it *nearly* is (for a suitable choice of κ that is not too small), and so it makes sense to define π_g by conditioning. The full prior $\pi_\epsilon = \pi_{s\epsilon} \times \pi_{g\epsilon} \times \pi_{b\epsilon}$ and clearly π_ϵ is supported on Θ.

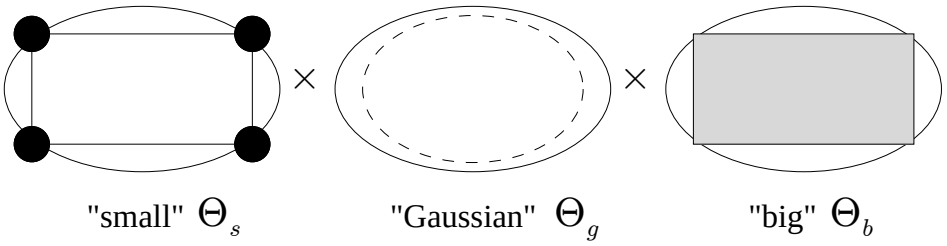

Figure 5.2 The "small" components prior is supported on the extreme points of a hyperrectangle in Θ_s; the "big" component prior lives on a solid hyperrectangle in Θ_b. The "Gaussian" components prior is mostly supported on Θ_g, cf. (5.36), note that the density contours do not match those of the ellipsoid.

Observe that the minimax risk $R_L(\epsilon) = R_s(\epsilon) + R_g(\epsilon) + R_b(\epsilon)$, where for $m = s, l, g$

$$R_m(\epsilon) = \sum_{i \in N_m} \rho_L(\tau_{i\epsilon}, \epsilon_i). \tag{5.29}$$

We show that the priors $\pi_{m\epsilon} = \pi_m(\epsilon, q)$ have the following properties:

(a) $B(\pi_{s\epsilon}) \ge r_s(q^{-1/2}) R_s(\epsilon)$ for all ϵ, and $r_s(q^{-1/2}) \to 1$ as $q \to \infty$,

(b) $B(\pi_{b\epsilon}) \ge r_b(q^{1/2}) R_b(\epsilon)$ for all ϵ, and $r_b(q^{1/2}) \to 1$ as $q \to \infty$, and

(c) If $\delta > 0$ and $q = q(\delta)$ are given, and if $R_g(\epsilon) \ge \delta R_L(\epsilon)$, then for $\epsilon < \epsilon(\delta)$ sufficiently small, $B(\pi_{g\epsilon}) \ge (1 - \delta) R_g(\epsilon)$.

Assuming these properties to have been established, we conclude the proof as follows. Fix $\delta > 0$ and then choose $q(\delta)$ large enough so that both $r_s(q^{-1/2})$ and $r_b(q^{1/2}) \geq 1 - \delta$. We obtain

$$B(\pi_{m\epsilon}) \geq (1-\delta)R_m(\epsilon), \qquad \text{for } m \in \{s, b\}. \tag{5.30}$$

Now, if $R_g(\epsilon) \geq \delta R_L(\epsilon)$, then the previous display holds also for $m = g$ and ϵ sufficiently small, by (c), and so adding, we get $B(\pi_\epsilon) \geq (1-\delta)R_L(\epsilon)$ for ϵ sufficiently small. On the other hand, if $R_g(\epsilon) \leq \delta R_L(\epsilon)$, then, again using (5.30),

$$B(\pi_\epsilon) \geq (1-\delta)[R_b(\epsilon) + R_s(\epsilon)] = (1-\delta)[R_L(\epsilon) - R_g(\epsilon)] \geq (1-\delta)^2 R_L(\epsilon).$$

Either way, we establish (5.19), and are done. So it remains to prove (a) - (c).

Proofs for (a) and (b). These are virtually identical and use the fact that two point and cosine priors are asymptotically least favorable as $\tau_{i\epsilon}/\epsilon_i \to 0$ and ∞ respectively. We tackle $B(\pi_{s\epsilon})$ first. For a scalar problem $y_1 = \theta_1 + \epsilon_1 z_1$ with univariate prior $\pi(d\theta)$ introduce the notation $B(\pi, \epsilon_1)$ for the Bayes risk. In particular, consider the two-point priors π_τ needed for the small signal case. By scaling, $B(\pi_\tau, \sigma) = \sigma^2 B(\pi_{\tau/\sigma}, 1)$, and the explicit formula (2.30) for $B(\pi_{\tau/\sigma}, 1)$ shows that when written in the form

$$B(\pi_\tau, \sigma) = \rho_L(\tau, \sigma) g(\tau/\sigma), \tag{5.31}$$

we must have $g(t) \to 1$ as $t \to 0$. Now, using this along with the additivity of Bayes risks, and then (5.27) and (5.29), we obtain

$$B(\pi_{s\epsilon}) = \sum_{N_s} B(\pi_{\tau_{i\epsilon}}, \epsilon_i) = \sum_{N_s} g(\tau_{i\epsilon}/\epsilon_i)\rho_L(\tau_{i\epsilon}, \epsilon_i) \geq r_s(q^{-1/2}) R_s(\epsilon), \tag{5.32}$$

if we set $r_s(u) = \inf_{0 \leq t \leq u} g(t)$. Certainly $r_s(u) \to 1$ as $u \to 0$, and this establishes (a).

For the large signal case (b), we use the cosine priors π_τ^V, the van Trees bound (4.9), and scaling properties of Fisher information, so that the analog of (5.31) becomes

$$B(\pi_\tau^V, \sigma) \geq \rho_L(\tau, \sigma) h(\tau/\sigma),$$

with $h(t) = (t^2 + 1)/(t^2 + I(\pi_1^V)) \to 1$ as $t \to \infty$. The analog of (5.32), $B(\pi_{g\epsilon}) \geq r_b(q^{1/2}) R_b(\epsilon)$ follows with $r_b(q) = \inf_{t \geq q} h(t) \to 1$ as $t \to 1$. Note that this argument actually works for any scale family of priors with finite Fisher information.

Proof of (c): This argument builds upon that given in the special white noise setting in the previous section. Let $\hat{\theta}_g = \mathbb{E}[\theta | \Theta_g, y]$ denote the Bayes rule for $\pi_{g\epsilon}$. With the obvious substitutions, the argument leading to (5.22) establishes that

$$(1-\kappa)R_g(\epsilon) \leq B(\pi_{g\epsilon}) + \mathbb{E}[\|\hat{\theta}_g - \theta\|^2, \Theta_g^c]. \tag{5.33}$$

Now we estimate $\mathbb{E}[\|\hat{\theta}_g - \theta\|^2, \Theta_g^c]$ by a small, but crucial, modification of Lemma 5.5, as foreshadowed in Remark 5.6. For indices i in the Gaussian range N_g, we use (5.28) to observe that $\tau_{i\epsilon}^2/\rho_L(\tau_{i\epsilon}, \epsilon_i) = \mu/a_i \leq 1 + q$, so that

$$\sum_{N_g} \tau_{i\epsilon}^2 \leq (1+q) \sum_{N_g} \rho_L(\tau_{i\epsilon}, \epsilon_i) = (1+q) R_g(\epsilon). \tag{5.34}$$

For $i \in N_g$ the variables are Gaussian and we may reuse the proof of Lemma 5.5 up through the inequality (5.25). Now substituting the bound above, we obtain the important bound

$$\mathbb{E}\{\|\hat{\theta}_g - \theta\|^2, \Theta_g^c\} \le c(q)\pi_\epsilon^G(\Theta_g^c)^{1/2} R_g(\epsilon). \tag{5.35}$$

Now we show that π_ϵ^G asymptotically concentrates on Θ_g, specifically

$$\pi_\epsilon^G(\Theta_g^c) \le 2q(\kappa^{-1} - 1)^2 \eta^2(N_g), \tag{5.36}$$

where

$$\eta^2(N) = \max_{i \in N} \epsilon_i^2 \Big/ \sum_{i \in N} \epsilon_i^2.$$

The bound (5.36) reflects three necessary quantities, and hence shows why the method works. First q governs the signal to noise ratios $\tau_{i\epsilon}^2/\epsilon_i^2$, while κ governs the 'slack' in the expectation ellipsoid. Finally $\eta^2(N_g)$ is a surrogate for the number of components $1/|N_g|$ in the unequal variance case. (Indeed, if all ϵ_i^2 are equal, this reduces to $1/|N_g|$).

Proof of (5.36). Indeed, let $S = \sum_{N_g} a_i^2\theta_i^2$ and $C_g^2 = \sum_{N_g} a_i^2\tau_{i\epsilon}^2$. Noting first that $E\theta_i^2 = (1-\kappa)\tau_{i\epsilon}^2$ and $\mathrm{Var}\,\theta_i^2 = 2(1-\kappa)^2\tau_{i\epsilon}^4$, we have

$$\begin{aligned} ES &= (1-\kappa)C_g^2, \qquad \text{and} \\ \mathrm{Var}S &\le 2(1-\kappa)^2 C_g^2 \max\{a_i^2\tau_{i\epsilon}^2\}. \end{aligned}$$

Now from Chebychev's inequality,

$$\begin{aligned} \pi^G(\Theta_g^c) &= P\{S - \mathbb{E}S > \kappa C_g^2\} \le \kappa^{-2} C_g^{-4} \mathrm{Var}\, S \\ &\le 2(\kappa^{-1}-1)^2 \max_i a_i^2\tau_{i\epsilon}^2 \Big/ \Big(\sum_{N_g} a_i^2\tau_{i\epsilon}^2\Big). \end{aligned}$$

From definition (5.6) of $\tau_{i\epsilon}^2$ and bounds (5.28) defining the Gaussian range N_g:

$$a_i^2\tau_{i\epsilon}^2 = \epsilon_i^2 a_i(\mu_\epsilon - a_i)_+ \in \epsilon_i^2\mu^2[q(q+1)^{-2}, 1/4],$$

and so

$$\frac{\max a_i^2\tau_{i\epsilon}^2}{\sum a_i^2\tau_{i\epsilon}^2} \le \frac{(q+1)^2}{4q} \frac{\max \epsilon_i^2}{\sum \epsilon_j^2} \le q\,\eta^2(N_g). \qquad \square$$

Inserting bound (5.36) into (5.35), we obtain

$$\mathbb{E}\{\|\hat{\theta}_g - \theta\|^2, \Theta_g^c\} \le c(q)(\kappa^{-1}-1)\eta(N_g)R_g(\epsilon). \tag{5.37}$$

We now use the hypothesis $R_g(\epsilon) \ge \delta R_L(\epsilon)$ to obtain a bound for $\eta(N_g)$. Indeed, using the definition of R_g and (5.7), we have

$$\begin{aligned} \sum_{N_g} \epsilon_i^2 \le R_g(\epsilon) \ge \delta R_L(\epsilon) &= \delta \sum_i \epsilon_i^2(1 - a_i/\mu)_+ \\ &\ge (\delta/2) \sum_{a_i \le \mu/2} \epsilon_i^2, \end{aligned}$$

and since (5.28) says that $N_g \subset \{i : a_i \le \mu_\epsilon\}$,

$$\eta^2(N_g) = \max_{i\in N_g} \epsilon_i^2 \Big/ \sum_{i\in N_g} \epsilon_i^2 \le (2/\delta) \max_{a_i\le\mu_\epsilon} \epsilon_i^2 \Big/ \sum_{a_i\le\mu/2} \epsilon_i^2 = (2/\delta)\eta^2(\mu_\epsilon).$$

Combining this last bound with (5.37), we obtain

$$\mathbb{E}[\|\hat\theta_g - \theta\|^2, \Theta_g^c] \le f(q,\kappa,\delta)\eta(\mu_\epsilon)R_g(\epsilon),$$

where $f(q,\kappa,\delta) = c(q)(\kappa^{-1}-1)\sqrt{2/\delta}$. We may now rewrite (5.33) to get

$$B(\pi_{g\epsilon}) \ge R_g(\epsilon)[1-\kappa - f(q,\kappa,\delta)\eta(\mu_\epsilon)].$$

Recall that $\delta > 0$ and $q = q(\delta)$ are given. We now set $\kappa = \delta/2$, and note that $\eta(\mu_\epsilon) \to 0$ as $\epsilon \to 0$. Indeed, that $\mu_\epsilon \to \infty$ follows from condition (5.4), here with $\epsilon_i^2 = \epsilon^2\varrho_i^2$, along with the assumption (i) that $a_i \nearrow \infty$ monotonically. Our assumption (ii) then implies that $\eta(\mu_\epsilon) \to 0$. Consequently, for $\epsilon < \epsilon(\delta, q(\delta))$, we have $f(q(\delta),\delta/2,\delta)\eta(\mu_\epsilon) < \delta/2$. Thus $B(\pi_{g\epsilon}) \ge (1-\delta)R_g(\epsilon)$ and this completes the proof of (c).

5.5 *Interlude:* Compactness and Consistency

This section, a digression, is included for variety, and 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, for quite general sets Θ, compactness is both necessary and sufficient for the existence of a uniformly consistent estimator, so long as we use the ℓ_2 norm to define both the error measure and the topology on Θ.

Theorem 5.7 *In the homoscedastic Gaussian sequence model* (3.1)*, assume that Θ is bounded in $\ell_2(\mathbb{N})$. Then as $\epsilon \to 0$, $R_N(\Theta,\epsilon) \to 0$ if and only if $\overline{\Theta}$ 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 again illustrates 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 $\{\theta \in \ell_2(\mathbb{N}) : \|\theta\|_2 \le 1\}$, or indeed on any open set in the norm topology.

This result is about the *infinite dimensional* nature of the sequence model (3.1). Of course if $\Theta \subset \mathbb{R}^r$ is bounded (and closed) then it is automatically compact, and the result is anyway trivial, since $R_N(\Theta,\epsilon) \le R_N(\mathbb{R}^r,\epsilon) = r\epsilon^2 \to 0$.

The boundedness condition is not necessary: it is a simple exercise to extend the theorem to sets of the form $\Theta = \mathbb{R}^r \times \Theta'$ with Θ' bounded in ℓ_2.

Because there are no longer any geometric assumptions on Θ, the tools used for the proof change: indeed methods from testing, classification and from information theory now appear. While the result 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 Θ determines the possible rates of convergence of $R_N(\Theta)$.

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 $P_{\theta_1}, \ldots, P_{\theta_m}$ be a corresponding set of probability distributions on $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$. For convenience, assume that the P_{θ_i} are mutually absolutely continuous, and so have positive densities p_i with respect to some dominating measure ν. Then, the Kullback-Leibler divergence between two probability measures P and Q having densities p, q relative to ν is

$$K(P, Q) = \int \log \frac{dP}{dQ}\, dP = \int \log \frac{p}{q}\, p\, d\nu. \tag{5.38}$$

The following lower bound is a formulation by Birgé (1983, Lemma 2.7) of Fano's lemma as presented in Ibragimov and Khasminskii (1981, pages 324-5).

Lemma 5.8 *With the above definitions, let $\hat{\theta} : \mathcal{Y} \to \{\theta_1, \ldots, \theta_m\}$ be an arbitrary estimator. Then*

$$ave_i\, P_{\theta_i}\{\hat{\theta} \neq \theta_i\} \geq 1 - \frac{ave_{i,j}\, K(P_{\theta_i}, P_{\theta_j}) + \log 2}{\log(m-1)}. \tag{5.39}$$

Remark. Both averages in inequality (5.39) can of course be replaced by maxima over i and (i, j) respectively.

Proof The first step is to show the inequality

$$n^{-1} \sum_i P_{\theta_i}(\hat{\theta} \neq \theta_i) \geq 1 - n^{-1} \int \max_i p_i d\nu. \tag{5.40}$$

[This is a multiple-hypothesis version of the Neyman-Pearson lemma.] To see this, view the left side as the integrated risk $B(\hat{\theta}, \pi)$ of $\hat{\theta}$ for the classification error loss function $L(a, \theta) = I(a \neq \theta)$ and the uniform prior π placing mass $1/n$ on each θ_i. It is therefore bounded below by the Bayes risk $B(\pi)$ of the Bayes estimator $\hat{\theta}_\pi = \text{argmax}_{\theta_i} \pi(\theta_i | y)$ derived from the posterior probabilities, cf. Remark 2.2. We have

$$B(\pi) = n^{-1} \sum_i P_{\theta_i}(\hat{\theta}_\pi \neq \theta_i) = 1 - n^{-1} \int \sum_i I(\hat{\theta}_\pi = \theta_i) p_i d\nu$$

and equality of the two right sides above follows from the definition of the Bayes rule.

Suppose now that k is chosen so that $p_k = \max_i p_i$. Here k depends on y, but the argument is carried out pointwise. Let $q_j = p_j/(1 - p_k)$. Some algebra then shows

$$\begin{aligned} \sum_j p_j \log p_j &= p_k \log p_k + (1 - p_k) \log(1 - p_k) + (1 - p_k) \sum_{j \neq k} q_j \log q_j \\ &\geq -\log 2 - (1 - p_k) \log(n - 1) \end{aligned} \tag{5.41}$$

where we applied twice the entropy bound $\sum_1^m \pi_l \log \pi_l^{-1} \leq \log m$ for a discrete probability vector $(\pi_1, \ldots, \pi_m)$.

Now make the particular choice $\nu = \sum_{k=1}^n P_{\theta_k}$ for dominating measure, and integrate (5.41) with respect to the probability measure $n^{-1}\nu$ and then rearrange to get

$$\begin{aligned} n^{-1}\log(n-1)\int \max_i p_i \, d\nu &\le \log 2 + \log(n-1) + n^{-1}\sum_j \int p_j \log p_j \, d\nu \\ &\le \log 2 + \mathrm{ave}_{j,k} K(P_{\theta_j}, P_{\theta_k}), \end{aligned} \tag{5.42}$$

where we have used the explicit form of ν and convexity of $-\log x$ to get

$$\int p_j \log p_j \, d\nu = -\int \log\Big\{n \cdot n^{-1}\sum_k \frac{dP_{\theta_k}}{dP_{\theta_j}}\Big\} dP_{\theta_j} \le -\log n + n^{-1}\sum_k K(P_{\theta_j}, P_{\theta_k}).$$

Now insert (5.42) into (5.40) to complete the proof. □

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: Θ is totally bounded if and only if for every δ, there is a finite set $\{\theta_i, \ldots, \theta_m\}$ such that the open balls $B(\theta_i, \delta)$ of radius δ centered at θ_i cover $\overline{\Theta}$: so that $\overline{\Theta} \subset \cup_{i=1}^m B(\theta_i, \delta)$. Also, since Θ is bounded, it has a finite *diameter* $\Delta = \sup\{\|\theta_1 - \theta_2\| : \theta_1, \theta_2 \in \Theta\}$.

Let $\delta > 0$ be given. Since $R_N(\Theta, \epsilon) \to 0$, there exists a noise level ϵ and an estimator $\tilde{\theta}_\delta$ such that

$$E_{\theta,\epsilon}\|\tilde{\theta}_\delta - \theta\|^2 \le \delta^2/2 \qquad \text{for all } \theta \in \Theta. \tag{5.43}$$

Let Θ_δ be a finite and $2\delta-$*discernible* subset of Θ: each distinct pair θ_i, θ_j in Θ_δ satisfies $\|\theta_i - \theta_j\| > 2\delta$. From $\tilde{\theta}_\delta(y)$ we build an estimator $\hat{\theta}_\delta(y)$ with values confined to Θ_δ by choosing a closest $\theta_i \in \Theta_\delta$ to $\tilde{\theta}_\delta(y)$: of course, whenever $\hat{\theta}_\delta \ne \theta_i$, it must follow that $\|\tilde{\theta}_\delta - \theta_i\| \ge \delta$. Consequently, from Markov's inequality and (5.43), we have for all i

$$P_{\theta_i}\{\hat{\theta}_\delta \ne \theta_i\} \le P_{\theta_i}\{\|\tilde{\theta}_\delta - \theta_i\| \ge \delta\} \le \delta^{-2} E\|\tilde{\theta}_\delta - \theta_i\|^2 \le 1/2. \tag{5.44}$$

On the other hand, the misclassification inequality (5.39) provides a lower bound to the error probability: for the noise level ϵ Gaussian sequence model, one easily evaluates

$$K(P_{\theta_i}, P_{\theta_j}) = \|\theta_i - \theta_j\|^2/2\epsilon^2 \le \Delta^2/2\epsilon^2,$$

where Δ is the diameter of Θ, and so

$$\max_i P_{\theta_i}\{\hat{\theta}_\delta \ne \theta_i\} \ge 1 - \frac{\Delta^2/2\epsilon^2 + \log 2}{\log(|\Theta_\delta| - 1)}.$$

Combining this with (5.44) gives a uniform upper bound for the cardinality of Θ_δ:

$$\log(|\Theta_\delta| - 1) \le \Delta^2\epsilon^{-2} + 2\log 2.$$

We may therefore speak of a $2\delta-$discernible subset $\Theta_\delta \subset \Theta$ of *maximal* cardinality, and for such a set, it is easily checked that $\overline{\Theta}$ is covered by closed balls of radius 4δ centered at the points of Θ_δ. Since δ was arbitrary, this establishes that $\overline{\Theta}$ is totally bounded, and so compact.

Sufficiency of Compactness

Given $\delta > 0$, we will construct an estimator $\hat{\theta}_\epsilon$ such that $E_\theta \|\hat{\theta}_\epsilon - \theta\|^2 \le 20\,\delta^2$ on Θ for all sufficiently small ϵ. Indeed, compactness of Θ supplies a finite set $\Theta_\delta = \{\theta_1, \ldots, \theta_m\}$ such that $\overline{\Theta} \subset \cup_{i=1}^m B(\theta_i, \delta)$, and we will take $\hat{\theta}_\epsilon$ to be the maximum likelihood estimate on the *sieve* Θ_δ. Thus we introduce the (normalized) log-likelihood

$$L(\theta) = \epsilon^2 \log dP_{\theta,\epsilon}/dP_{0,\epsilon} = \langle y, \theta\rangle - \tfrac{1}{2}\|\theta\|^2, \tag{5.45}$$

and the maximum likelihood estimate

$$\hat{\theta}_\epsilon = \arg\max_{\theta_i \in \Theta_\delta} L(\theta).$$

Since Θ has diameter Δ, we have for any $\theta \in \Theta$ the simple MSE bound

$$E_\theta \|\hat{\theta}_\epsilon - \theta\|^2 \le (4\delta)^2 + \Delta^2 \sum_{i:\|\theta_i - \theta\| \ge 4\delta} P_\theta\{\hat{\theta}_\epsilon = \theta_i\}. \tag{5.46}$$

We now show that the terms in the second sum are small when ϵ is small. Let $\theta \in \Theta$ be fixed, and choose a point in Θ_δ, renumbered to θ_1 if necessary, so that $\theta \in B(\theta_1, \delta)$. To have $\hat{\theta}_\epsilon = \theta_i$ certainly implies that $L(\theta_i) \ge L(\theta_1)$, and from (5.45)

$$L(\theta_i) - L(\theta_1) = \langle y - \tfrac{1}{2}(\theta_i + \theta_1), \theta_i - \theta_1\rangle.$$

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 $L(\theta_i) \ge L(\theta_1)$ implies

$$\epsilon Z \ge \langle \tfrac{1}{2}(\theta_i + \theta_1) - \theta, u\rangle \ge \tfrac{1}{2}\|\theta_i - \theta_1\| - \delta \ge \delta/2,$$

where in the second inequality we used $|\langle \theta_1 - \theta, u\rangle| \le \|\theta_1 - \theta\| < \delta$, and in the third $\|\theta_i - \theta_1\| \ge 3\delta$. Thus $P_\theta\{\hat{\theta}_\epsilon = \theta_i\} \le \tilde{\Phi}(\delta/(2\epsilon))$, and so from (5.46)

$$E_\theta \|\hat{\theta}_\epsilon - \theta\|^2 \le (4\delta)^2 + m\Delta^2 \tilde{\Phi}(\delta/(2\epsilon)) \le 20\delta^2,$$

whenever ϵ is sufficiently small.

5.6 Notes and Exercises

Pinsker's paper inspired a considerable literature. Here we mention only two recent works which contain, among other developments, different proofs of the original result: Belitser and Levit (1995) and Tsybakov (1997), and the examples given in Sections 5.2– 5.2.

As noted in the proof of Theorem 4.25, identity (5.7) is itself a minimax theorem, indeed Pinsker gave a direct proof; see also the account in Tsybakov (2009, Sec. 3.2).

Efromovich (1996) gives an extension of the sharp minimax constant results to a variety of nonparametric settings including binary, binomial, Poisson and censored regression models.

For further discussion of minimax estimation in the fractional integration setting, see Cavalier (2004).

The use of parameter spaces of analytic functions goes back to Ibragimov and Khasminskii (1983) and Ibragimov and Khasminskii (1984), see also Golubev and Levit (1996).

The consistency characterization, Theorem 5.7, is a special case of a result announced by Ibragimov and Has′minskiĭ (1977), and extended in Ibragimov and Khasminskii (1997). The approach of maximizing likelihood over subsets that grow with sample size was studied as the "method of sieves" in Grenander (1981).

Exercises

5.1 *Pinsker constant: tracking the error terms.* Consider the fractional integration setting of Section 5.2 in which $\epsilon_k = \epsilon/b_k$ and the singular values $b_k = c_\beta^{-1} k^{-\beta}(1 + O(k^{-1}))$. This of course includes the special case of direct estimation with $b_k \equiv 1$.
(a) Consider the ellipsoids $\check{\Theta}_2^\alpha(C)$ with $a_{2k} = a_{2k+1} = (2k)^\alpha$. Let $N = N(\mu) = \{k : a_k < \mu\}$ and show that

$$k_\mu = |N(\mu)| = \mu^{1/\alpha} + O(1) = \mu^{1/\alpha}(1 + O(k_\mu^{-1})).$$

(b) For $p = 0, 1, 2$, let $S_p = \sum_{k \in N} b_k^{-2} a_k^p$ and show that

$$S_p = (2\beta + p\alpha + 1)^{-1} c_\beta^2 k_\mu^{2\beta + p\alpha + 1}(1 + O(k_\mu^{-1})).$$

(c) Verify that $R_L(\Theta) = \epsilon^2(S_0 - \mu S_1)$ where μ satisfies $\mu S_1 - S_2 = C^2/\epsilon^2$, and hence show that (CHECK)

$$R_L(\Theta^\alpha(C), \epsilon) = P_{r,\beta} C^{2(1-r_\beta)}(c_\beta \epsilon)^{2r_\beta}(1 + O(k_\mu^{-1})).$$

5.2 *Polynomial rates in severely ill-posed problems.* Consider ellipsoids with $a_{2k} = a_{2k-1} = e^{\alpha k}$ corresponding to analytic functions as in Section 5.2. Suppose that $\epsilon_k = e^{\beta k}\epsilon$ with $\beta > 0$, so that the estimation problem is "severely-ill-posed". Show that the linear minimax risk

$$R_L(\Theta, \epsilon) \sim Q C^{2(1-\rho)} \epsilon^{2\rho}$$

with $\rho = \alpha/(\alpha + \beta)$ and $Q_{\alpha\beta} = Q_{\alpha\beta}(C/\epsilon)$ is a continuous, positive, periodic function of $\log(C/\epsilon)$ which therefore takes values in an interval $[Q_{\alpha\beta}^-, Q_{\alpha\beta}^+] \subset (0, \infty)$.

5.3 *(Pinsker theorem for commuting operator and covariance.)* Consider a more general ellipsoid $\Theta = \{\theta : \theta^T A \theta \le C^2\}$ for a positive definite matrix A. Suppose $Y \sim N(\theta, \Sigma)$ and that A and Σ commute: $A\Sigma = \Sigma A$. Show that there is a linear transformation of Y for which the Pinsker theorems hold.
[The situation appears to be less simple if A and Σ do not commute.]

5.4 (*Non asymptotic bound for efficiency of smoothing splines.*) This exercise pursues the observations of Carter et al. (1992) that the efficiency of smoothing splines is even better for "non-asymptotic" values of ϵ.
(i) Revisit the proof of Proposition 3.9 and show that for $\alpha = m$ and the trigonometric basis

$$\bar{r}(\hat{\theta}_\lambda; \epsilon) \le v_\alpha \epsilon^2 \lambda^{-1/(2\alpha)} + (C^2/4)\lambda + \epsilon^2.$$

(ii) Revisit the evaluation of $R_L(\Theta, \epsilon)$ prior to (8.10) and show that

$$R_L(\Theta, \epsilon) \ge \frac{\alpha}{\alpha + 1} \epsilon^2 \mu_\epsilon^{1/\alpha}.$$

(iii) Let A be the set of values (α, δ) for which

$$\delta \sum_{k \ge 0} (\delta k)^\alpha [1 - (\delta k)^\alpha]_+ \le \int_0^1 v^\alpha (1 - v^\alpha) dv.$$

[It is *conjectured* that this holds for most or all $\alpha > 0, \delta > 0$]. Show that

$$\mu_\epsilon \ge \bar{\mu}_\epsilon = \left(\frac{(\alpha + 1)(2\alpha + 1)}{\alpha} \frac{C^2}{\epsilon^2} \right)^{\alpha/(2\alpha+1)}.$$

so long as $(\alpha, \bar{\mu}_\epsilon^{-1/\alpha}) \in A$.

(iv) Conclude that in these circumstances,

$$\frac{R_{SS}(\Theta;\epsilon)}{R_L(\Theta;\epsilon)} \le e_\alpha + c_\alpha(\epsilon/C)^{2(1-r)}$$

for all $\epsilon > 0$. [Here e_α is the constant in the limiting efficiency (5.16).]

5.5 *(Pinsker proof: bounding μ_ϵ.)* Adopt assumptions (i'), (ii') in the special case of Pinsker's theorem in Section 5.3. Let k_ϵ be the smallest integer so that $b_2(k_\epsilon + 1)^\alpha \ge \mu_\epsilon/2$. Show that

$$\sum_i a_i(\mu_\epsilon - a_i)_+ \ge b_1(\mu_\epsilon/2)\sum_{i=1}^{k_\epsilon} i^\alpha.$$

Conclude from this, with $c_\alpha = c_\alpha(\alpha, b_1, b_2)$ and $r = 2\alpha/(2\alpha+1)$, that $\epsilon\mu_\epsilon \le c_\alpha C^r \epsilon^{1-r}$.

6

Adaptive Minimaxity over Ellipsoids

> However beautiful the strategy, you should occasionally look at the results. (Winston Churchill)

An estimator that is exactly minimax for a given parameter set Θ will depend, often quite strongly, on the details of that parameter set. While this is informative about the effect of assumptions on estimators, it is impractical for the majority of applications in which no single parameter set comes as part of the problem description.

In this chapter, we shift perspective in order to study the properties of estimators that can be defined without recourse to a fixed Θ. Fortunately, it turns out that certain such estimators can come close to being minimax over a whole *class* of parameter sets. We exchange exact optimality for a single problem for approximate optimality over a range of circumstances. The resulting 'robustness' is usually well worth the loss of specific optimality.

The example developed in this chapter is the use of the James-Stein estimator on *blocks* of coefficients to approximately mimic the behavior of linear minimax rules for particular ellipsoids.

The problem is stated in more detail for ellipsoids in Section 6.1. The class of linear estimators that are constant on blocks is studied in Section 6.2, while the blockwise James-Stein estimator appears in Section 6.3. The adaptive minimaxity of blockwise James-Stein is established; the proof boils down to the ability of the James-Stein estimator to mimic the ideal linear shrinkage rule appropriate to each block, as already seen in Section 2.6.

While the blockwise shrinkage approach may seem rather tied to the details of the sequence model, in fact it accomplishes its task in a rather similar way to kernel smoothers or smoothing splines in other problems. This is set out both by heuristic argument and in a couple of concrete examples in Section 6.4.

Looking at the results of our blockwise strategy (and other linear methods) on one of those examples sets the stage for the focus on non-linear estimators in following chapters: linear smoothing methods, with their constant smoothing bandwidth, are ill-equipped to deal with data with sharp transitions, such as step functions. It will be seen later that the adaptive minimax point of view still offers useful insight, but now for a different class of estimators (wavelet thresholding) and wider classes of parameter spaces.

Section 6.5 is again an interlude, containing some remarks on "fixed θ" versus worst case asymptotics and on superefficiency. Informally speaking, superefficiency refers to the possibility of exceptionally good estimation performance at isolated parameter points. In parametric statistics this turns out, fortunately, to be usually a peripheral issue, but examples

given here show that points of superefficiency are endemic in nonparametric estimation. The dangers of over-reliance on asymptotics based on a single θ are illustrated in an example where nominally optimal bandwidths are found to be very sensitive to aspects of the function that are difficult to estimate at any moderate sample size.

In this chapter we focus on the white noise model–some extensions of blockwise James Stein to linear inverse problems are cited in the Notes.

6.1 The problem of adaptive estimation

We again suppose that we are in the white noise Gaussian sequence model $y_i = \theta_i + \epsilon z_i$, and consider the family of ellipsoids $\Theta_2^\alpha(C)$, (3.11) which correspond to smoothness constraints $\int (D^\alpha f)^2 \le L^2$ on periodic functions in $L_2[0,1]$ when represented in the Fourier basis (3.8): To recall, for $\alpha, C > 0$, we have

$$\Theta_2^\alpha(C) = \{\theta \in \ell_2 : \sum_{k=1}^{\infty} a_k^2 \theta_k^2 \le C^2\}, \tag{6.1}$$

with $a_k = k^\alpha$ for $k \ge 1$. As we have seen in previous chapters, Pinsker's theorem delivers a linear estimator $\hat{\theta}_\epsilon(\alpha, C)$, given by (5.5), which is minimax linear for all $\epsilon > 0$, and asymptotically minimax among *all* estimators as $\epsilon \to 0$.

As a practical matter, the constants (α, C) are generally unknown, and even if one believed a certain value (α_0, C_0) to be appropriate, there is an issue of robustness of MSE performance of $\hat{\theta}_\epsilon(\alpha_0, C_0)$ to misspecification of (α, C). One possible way around this problem is to construct an estimator family $\hat{\theta}_\epsilon^*$, whose definition does not depend on (α, C), such that if θ is in fact restricted to some $\Theta_2^\alpha(C)$, then $\hat{\theta}_\epsilon^*$ has MSE appropriate to that space:

$$\sup_{\theta \in \Theta^\alpha(C)} r(\hat{\theta}_\epsilon^*, \theta) \le c_\epsilon(\Theta) R_N(\Theta^\alpha(C), \epsilon) \qquad \text{as } \epsilon \to 0, \tag{6.2}$$

where $c_\epsilon(\Theta)$ is a bounded sequence. Write $\mathcal{T}_2$ for the collection of all ellipsoids $\{\Theta_2^\alpha(C) : \alpha, C > 0\}$. One then calls $\hat{\theta}_\epsilon$ *rate-adaptive*: it "learns" the right rate of convergence for all $\Theta \in \mathcal{T}_2$. In later chapters, we also consider a weaker notion of rate adaptivity in which $c_\epsilon(\Theta)$ may grow at some power of $\log \epsilon^{-1}$.

If $\hat{\theta}_\epsilon^*$ has the stronger property that $c_\epsilon(\Theta) \to 1$ for each $\Theta \in \mathcal{T}_2$ as $\epsilon \to 0$, 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 Efromovich and Pinsker (1984), and this chapter presents their blockwise estimator approach, lightly modified with use of the James-Stein method. We will see that good non-asymptotic bounds are also possible, and that the blockwise James-Stein estimator is a plausible estimator for practical use in appropriate settings.

6.2 Blockwise Estimators

Consider first an abstract countable index set $\mathcal{I}$ partitioned into an ordered sequence $\mathcal{B}$ of blocks B_j of finite cardinality n_j. We make the notational convention in this chapter that a subscript j *always* indexes a block, and that y_j, z_j and θ_j denote *vectors* of coefficients,

thus for example $y_j = \{y_i, i \in B_j\}$. This chapter mostly focuses on the case $\mathcal{I} = \mathbb{N}$ and with the blocks defined by a strictly increasing sequence $\{l_j, j \geq 0\} \subset \mathbb{N}$ with $l_0 = 1$ and

$$B_j = \{l_j, l_j + 1, \ldots, l_{j+1} - 1\}, \qquad n_j = l_{j+1} - l_j. \tag{6.3}$$

Often $l_j = [L_j]$ for $L_j \in \mathbb{R}_+$. In some cases, the sequence $l_{j\epsilon}$ and associated blocks $B_{j\epsilon}$ might depend on noise level ϵ.

Particular examples might include $L_j = (j+1)^\beta$ for some $\beta > 0$, or $L_j = e^{\sqrt{j}}$. An ϵ-dependent example is given by "weakly geometric blocks", with $\ell_\epsilon = \log \epsilon^{-1}$ and $L_{j\epsilon} = \ell_\epsilon(1 + 1/\ell_\epsilon)^{j-1}$. However, we will devote particular attention to the case of *dyadic blocks.* in which $l_j = 2^j$, so that block B_j has cardinality $n_j = 2^j$.

The block James-Stein estimator. We construct an estimator which on each block B_j (or $B_{j\epsilon}$) applies the positive part James-Stein estimator (2.62):

$$\hat{\theta}_j^{JS}(y_j) = \left(1 - \frac{(n_j - 2)\epsilon^2}{\|y_j\|^2}\right)_+ y_j. \tag{6.4}$$

A key benefit of the James-Stein estimate is the good bounds for its MSE. Proposition 2.7, or rather its Corollary 2.9 for noise level ϵ, shows that when $n_j \geq 3$,

$$r_\epsilon(\hat{\theta}_j^{JS}, \theta_j) \leq 2\epsilon^2 + \sum_j \frac{n_j \epsilon^2 \|\theta_j\|^2}{n_j \epsilon^2 + \|\theta_j\|^2}. \tag{6.5}$$

The full blockwise estimator, $\hat{\theta}^{BJS}$, is then defined by

$$\hat{\theta}_j^{BJS}(y) = \begin{cases} y_j & j < L \\ \hat{\theta}_j^{JS}(y_j) & L \leq j < J_\epsilon \\ 0 & j \geq J_\epsilon \end{cases}. \tag{6.6}$$

For the 'earliest' blocks, specified by L, no shrinkage is performed. This may be sensible 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_ϵ. Usually J_ϵ is chosen so that $l_\epsilon = l_{J_\epsilon} = [\epsilon^{-2}]$, which is proportional to 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.

Block ellipsoids. Along with a set of blocks B_j, it is natural to introduce *block ellipsoids*. If the ellipsoid semiaxes a_k are constant on blocks, $a_k \equiv b_j$ for $k \in B_j$, then we have

$$\sum a_k^2 \theta_k^2 = \sum b_j^2 \|\theta_j\|^2, \tag{6.7}$$

and we denote the corresponding ellipsoid

$$\Theta(b, C) = \{\theta : \sum b_j^2 \|\theta_j\|^2 \leq C^2\}.$$

A useful example occurs with dyadic blocks, in which we consider a variant of the ellipsoids (6.1) that is defined using weights that are constant on the dyadic blocks: $a_l \equiv 2^{j\alpha}$ if $l \in B_j = \{2^j, \ldots, 2^{j+1} - 1\}$. The corresponding *dyadic Sobolev ellipsoids*

$$\Theta_D^\alpha(C) = \{\theta : \sum_{j \geq 0} 2^{2j\alpha} \sum_{l \in B_j} \theta_l^2 \leq C^2\}. \tag{6.8}$$

Let $\mathcal{T}_{D,2}$ denote the class of such dyadic ellipsoids $\{\Theta_D^\alpha(C), \alpha, C > 0\}$.

The two approaches are norm-*equivalent*: write $\|\theta\|_{F,\alpha}^2$ for the squared norm corresponding to (6.1) and $\|\theta\|_{D,\alpha}^2$ for that corresponding to (6.8). Exercise 6.1 fills in the details. It is then easily seen that for all $\theta \in \ell_2$:

$$\|\theta\|_{D,\alpha} \le \|\theta\|_{F,\alpha} \le 2^\alpha \|\theta\|_{D,\alpha}. \tag{6.9}$$

As a result, the minimax risks $R_{\mathcal{E}}(\Theta, \epsilon)$ for $\Theta = \Theta_D^\alpha(C)$ and for $\Theta = \Theta_2^\alpha(C)$ are within a multiplicative factor $2^{2\alpha}$ for any estimator class $\mathcal{E}$.

Remark. For wavelet bases, ellipsoid weights that are constant on dyadic blocks are the natural way to represent mean-square smoothness–see Section 9.6. In this case, the index $I = (j,k)$, with $j \ge 0$ and $k \in \{0, \dots, 2^j - 1\}$. There is a simple mapping of doubly indexed coefficients $\theta_{j,k}$ onto a single sequence θ_l by setting $l = 2^j + k$, including the special case $\theta_{-1,0} \leftrightarrow \theta_0$, compare (9.42).

Block diagonal linear estimators. This term refers to the subclass of diagonal linear estimators in which the shrinkage factor is constant within blocks: for all blocks j:

$$\hat{\theta}_{j,c_j}(y) = c_j y_j, \qquad c_j \in \mathbb{R}.$$

The mean squared error on the jth block has a simple form, directly or from (3.14),

$$r_\epsilon(\hat{\theta}_{j,c_j}, \theta_j) = n_j \epsilon^2 c_j^2 + (1 - c_j)^2 \|\theta_j\|^2. \tag{6.10}$$

The corresponding minimax risk among *block linear* diagonal estimators is then

$$R_{BL}(\Theta, \epsilon; \mathcal{B}) = \inf_{(c_j)} \sup_{\Theta} \sum_j r_\epsilon(\hat{\theta}_{j,c_j}, \theta_j).$$

The final argument $\mathcal{B}$, which we often omit, reminds that the definition depends on the particular block sequence used. Although this chapter focuses on ellipsoids, the definition is meaningful for more general Θ. For example, the minimax theorem for diagonal linear estimators, Theorem 4.25, has an analog in the block case. Indeed, if Θ is compact, solid-orthosymmetric and quadratically convex, then (Exercise 6.4)

$$R_{BL}(\Theta, \epsilon) = \sup_{\Theta} \inf_{(c_j)} \sum_j r_\epsilon(\hat{\theta}_{j,c_j}, \theta_j). \tag{6.11}$$

The right side of (6.11) has an interpretation as an ideal shrinkage risk. The minimization over (c_j) can be carried out term by term in the sum. The ideal shrinkage factor on the jth block is found by minimizing (6.10). This yields shrinkage factor $c^{IS}(\theta_j) = \|\theta_j\|^2/(n_j\epsilon^2 + \|\theta_j\|^2)$ and the corresponding ideal estimator $\hat{\theta}_j^{IS}(y) = c^{IS}(\theta_j) y_j$ has *ideal risk*

$$r_\epsilon(\hat{\theta}_j^{IS}, \theta_j) = \frac{n_j \epsilon^2 \|\theta_j\|^2}{n_j \epsilon^2 + \|\theta_j\|^2}, \tag{6.12}$$

and we may write

$$R_{BL}(\Theta, \epsilon) = \sup_{\Theta} \sum_j \frac{n_j \epsilon^2 \|\theta_j\|^2}{n_j \epsilon^2 + \|\theta_j\|^2}. \tag{6.13}$$

Block Linear versus Linear. Certainly, $R_L(\Theta,\epsilon) \le R_{BL}(\Theta,\epsilon)$. However, in two cases, more can be said:

(i) if $\Theta = \Theta(b, C)$ is a block ellipsoid associated with blocks B_j then

$$R_L(\Theta,\epsilon) = R_{BL}(\Theta,\epsilon) \qquad \text{for all } \epsilon > 0. \tag{6.14}$$

The dyadic Sobolev ellipsoids $\Theta^\alpha_D(C)$ are block symmetric and so are an example for (6.14).

(ii) For general ellipsoids $\Theta(a, C)$ as in (5.2), and a block scheme (6.3), measure the *oscillation* of the weights a_k within blocks by

$$\mathrm{osc}_a(B_j) = \max_{k,k' \in B_j} \frac{a_k}{a_{k'}}.$$

If $a_k \to \infty$ and $\mathrm{osc}_a(B_j) \to 1$, then we have

$$R_L(\Theta,\epsilon) \sim R_{BL}(\Theta,\epsilon) \qquad \text{as } \epsilon \to 0. \tag{6.15}$$

In the ϵ-dependent case, (6.15) holds if we require that $\max_{k \ge j} \mathrm{osc}(B_{k\epsilon}) \to 1$ as $j \to \infty$ and $\epsilon \to 0$ jointly.

In the Fourier ellipsoid case, $\mathrm{osc}_a(B_j) = (L_{j+1}/L_j)^\alpha + O(L_j^{-1})$. Relation (6.15) applies to all $\Theta^\alpha_2(C)$ if one uses blocks B_j defined by either $L_j = (j+1)^\beta$ for $\beta > 0$, or $L_j = e^{\sqrt{j}}$ – in either case $\mathrm{osc}_a(B_j) \to 1$. For weakly geometric blocks, $\mathrm{osc}(B_{j\epsilon}) \sim (1 + 1/\ell_\epsilon)^\alpha \to 1$ as $j \to \infty, \epsilon \to 0$. However, the block sizes must necessarily be subgeometric in growth: for dyadic blocks, $l_j = 2^j$, the condition fails: $\mathrm{osc}_a(B_j) \to 2^\alpha$.

Proof of (6.14). We saw at (5.7) that for any ellipsoid

$$R_L(\Theta,\epsilon) = \sup\Big\{\sum_i \frac{\epsilon^2\theta_i^2}{\epsilon^2+\theta_j^2}, \sum_i a_i^2\theta_i^2 \le C^2\Big\}.$$

In the present setting, a_i is constant on each block. We can thus make use of the identity

$$\sup\Big\{\sum_i \frac{\epsilon^2\theta_i^2}{\epsilon^2+\theta_j^2} : \sum_i \theta_i^2 \le c^2\Big\} = \frac{n\epsilon^2c^2}{n\epsilon^2+c^2}, \tag{6.16}$$

which follows from concavity of $u \to \epsilon^2 u/(\epsilon^2+u)$, cf. (4.59). On block B_j, then, we substitute n_j for n and $\|\theta_j\|^2$ for c^2, and the result follows from (6.7) and (6.16). □

Proof of (6.15). Define a 'blocked' ellipsoid $\bar\Theta \supset \Theta$ by setting $\bar a_i = \min\{a_i, i \in B_j\}$. Since $\bar a_i$ is constant on blocks, $R_{BL}(\bar\Theta,\epsilon) = R_L(\bar\Theta,\epsilon)$ by (6.14). In addition, $\bar a_i \le a_i$ and $\bar a_i/a_i \to 1$ by the oscillation condition, and so $R_L(\bar\Theta,\epsilon) \sim R_L(\Theta,\epsilon)$ from the risk comparison Proposition 5.2. The claim now follows from the chain of inequalities

$$R_L(\Theta,\epsilon) \le R_{BL}(\Theta,\epsilon) \le R_{BL}(\bar\Theta,\epsilon) = R_L(\bar\Theta,\epsilon) \sim R_L(\Theta,\epsilon)$$

The ϵ-dependent case is left as Exercise 6.3. □

Remark. The identity (6.14) for block linear minimax risk holds for more general sets Θ under the assumption of *block symmetry*: namely if Θ is invariant to permutations of indices l within blocks, see Exercise 6.5 for a more precise statement.

6.3 Adaptivity of Blockwise James Stein Estimation

Our main result provides some classes of block sequences for which the blockwise James Stein estimator is rate adaptive. Recall that $\mathcal{T}_2$ and $\mathcal{T}_{D,2}$ denote the classes of Sobolev and dyadic Sobolev ellipsoids defined below (6.2) and (6.8) respectively.

Theorem 6.1 *In the homoscedastic white noise model, let $\hat{\theta}^{BJS}$ denote the block James-Stein estimator* (6.6).
(i) For dyadic blocks, let $J_\epsilon = \log_2 \epsilon^{-2}$. Then for each $\Theta \in \mathcal{T}_{D,2}$, the estimator $\hat{\theta}^{BJS}$ is adaptive minimax as $\epsilon \to 0$:

$$\sup_{\theta \in \Theta} r_\epsilon(\hat{\theta}^{BJS}, \theta) \sim R_N(\Theta, \epsilon). \tag{6.17}$$

(ii) For more general choices of blocks, for each $\Theta \in \mathcal{T}_2$ assume that $\mathrm{osc}_a(B_j) \to 1$ as $j \to \infty$ or $\max_{k \geq j} \mathrm{osc}_a(B_{k\epsilon}) \to 1$ as $j \to \infty$ and $\epsilon \to 0$ jointly. Define the block index J_ϵ in (6.6) *using the equation $l_{J_\epsilon} = \epsilon^{-2}$, and suppose that it satisfies $J_\epsilon = o(\epsilon^{-\eta})$ for all $\eta > 0$. Then adaptive minimaxity* (6.17) *holds also for each $\Theta \in \mathcal{T}_2$.*

For known noise level we see that, unlike the Pinsker linear minimax rule, which depends on C and details of the ellipsoid weight sequence (here α), the block James-Stein estimator has no adjustable parameters (other than the integer limits L and J_ϵ), and yet it can achieve asymptotically the exact minimax rate and constant for a range of values of C and α.

Some remarks on the assumptions in case (ii): A definition of J_ϵ such as through $l_{J_\epsilon} = \epsilon^{-2}$ means that necessarily $J_\epsilon \to \infty$ as $\epsilon \to 0$. The oscillation condition prevents the block sizes from being too large while the bound $\epsilon^\eta J_\epsilon \to 0$ means that the block sizes cannot be too small. Some further discussion of qualifying block sizes may be found after the proof.

Proof We decompose the mean squared error by blocks,

$$r_\epsilon(\hat{\theta}^{BJS}, \theta) = \sum_j r_\epsilon(\hat{\theta}_j^{JS}, \theta_j)$$

and employ the structure of $\hat{\theta}^{BJS}$ given in (6.6). On low frequency blocks, $j < L$, the estimator is unbiased and contributes only variance terms $n_j \epsilon^2$ to MSE. On high frequency blocks, $j \geq J_\epsilon$, only a bias term $\|\theta_j\|^2$ is contributed. On the main frequency blocks, $L \leq j < J_\epsilon$, we use the key bound (6.5). Assembling the terms, we find

$$r_\epsilon(\hat{\theta}^{BJS}, \theta) \leq (l_L + 2J_\epsilon - 2L)\epsilon^2 + \sum_{j=L}^{J_\epsilon - 1} r_\epsilon(\hat{\theta}_j^{IS}, \theta_j) + \sum_{l \geq l_\epsilon} \theta_l^2. \tag{6.18}$$

In view of (6.13), the first right-side sum is bounded above by the block linear minimax risk $R_{BL}(\Theta, \epsilon)$. Turning to the second sum, for any ellipsoid $\Theta(a, C)$ with $a_l \nearrow \infty$, define the (squared) *maximal tail bias*

$$\Delta_\epsilon(\Theta) = \sup \left\{ \sum_{l \geq l_\epsilon} \theta_l^2 : \sum a_l^2 \theta_l^2 \leq C^2 \right\} = C^2 a_{l_\epsilon}^{-2}. \tag{6.19}$$

We therefore conclude that for such ellipsoids

$$\sup_\Theta r_\epsilon(\hat{\theta}^{BJS}, \theta) \leq (l_L + 2J_\epsilon)\epsilon^2 + R_{BL}(\Theta, \epsilon) + \Delta_\epsilon(\Theta). \tag{6.20}$$

Under either assumption (i) or (ii), we have as $\epsilon \to 0$ that

$$R_{BL}(\Theta, \epsilon) \sim R_L(\Theta, \epsilon) \sim R_N(\Theta, \epsilon),$$

where the first relation follows from (6.14) or (6.15) respectively, and the second relation follows from Pinsker's theorem, Theorem 5.3 and Remark 3.

Since the left side of (6.20) is, by definition, larger than $R_N(\Theta, \epsilon)$, we will be done if we show that the first and third right side terms in (6.20) are of smaller order than $R_N(\Theta, \epsilon) \asymp \epsilon^{2r}$ (with, as usual, $r = 2\alpha/(2\alpha + 1)$).

For the first term, note that l_L is fixed, and that $J_\epsilon \epsilon^2 = o(\epsilon^{2-\eta})$ for each $\eta > 0$ by assumption (ii), which is also satisfied by $J_\epsilon = \log_2 \epsilon^{-2}$ in the dyadic blocks case (i). Clearly we can choose η small enough so that $\epsilon^{2-\eta} = O(\epsilon^{2r})$.

For the third term, since $a_k \asymp k^\alpha$ and $2\alpha > r$,

$$\Delta_\epsilon(\Theta) \le c C^2 l_\epsilon^{-2\alpha} \asymp C^2 (\epsilon^2)^{2\alpha} \ll \epsilon^{2r}. \qquad \square$$

Remarks. 1. *Smaller blocks.* For traditional Sobolev ellipsoids, dyadic blocks are too large, since with $a_l \sim l^\alpha$, $\operatorname{osc}(B_j) \to 2^\alpha$, and so one has only rate adaptivity: indeed $R_{BL}(\Theta, \epsilon) \le 2^{2\alpha} R_N(\Theta, \epsilon)(1 + o(1))$. However, part (ii) of the previous theorem shows that exact adaptation *can* be achieved with smaller block sizes, for which $\operatorname{osc} B_j \to 1$. Thus $L_j = e^{\sqrt{j}}$ works, for example, as do weakly geometric blocks, $L_j = \ell_\epsilon (1 + 1/\ell_\epsilon)^{j-1}$, with $\ell_\epsilon = \log \epsilon^{-1}$. However, the sequence $L_j = (j+1)^\beta$ is less satisfactory, since $l_{J_\epsilon} = [\epsilon^{-2}]$ implies that $J_\epsilon \sim \epsilon^{-2/\beta}$ and so $\epsilon^2 J_\epsilon$ is not $o(\epsilon^{2r})$ in the smoother cases, when $2\alpha + 1 \ge \beta$.

In fact, this last problem arises from the bound $2\epsilon^2$ in (6.5), and could be reduced by using a modified estimator

$$\hat{\theta}_j = \left(1 - \frac{\lambda_j n_j \epsilon^2}{\|y_j\|^2}\right)_+ y_j, \qquad j \le J_\epsilon, \tag{6.21}$$

with $\lambda_j = 1 + t_j$ with $t_j > 0$. This reduces the error at zero to essentially a large deviation probability, see e.g. Brown et al. (1997), who use dyadic blocks and $t_j = 1/2$, or Cavalier and Tsybakov (2001) who use smaller blocks and $t_j \sim (n_j^{-1} \log n_j)^{1/2}$.

2. Depending on the value of λ_j–close to 1 or larger–one might prefer to refer to (6.21) as a block *shrinkage* or a block *thresholding* estimator. As just noted, the value of λ_j determines the chance that $\hat{\theta}_j \ne 0$ given that $\theta_j = 0$, and this chance is small in the block thresholding regime.

The use of block thresholding in conjunction with smaller sized blocks of wavelet coefficients has attractive MSE properties, even for function spaces designed to model spatial inhomogeneity. For example, Cai (1999) uses blocks of size $\log n = \log \epsilon^{-2}$ and $\lambda_j = 4.505$. We return to an analysis of a related block threshold approach in Chapters 8 and 9.

3. The original Efromovich and Pinsker (1984) estimator set

$$\hat{\theta}_j = \begin{cases} (1 - \frac{n_j \epsilon^2}{\|y_j\|^2}) y_j, & \|y_j\|^2 \ge (1 + t_j) n_j \epsilon^2, \\ 0 & \text{otherwise} \end{cases} \tag{6.22}$$

for $t_j > 0$ and $j \le J_\epsilon$. To prove adaptive minimaxity over a broad class of ellipsoids (5.2), they required in part that $n_{j+1}/n_j \to 1$ and $t_j \to 0$, but slowly enough that $\sum_j 1/(t_j^3 n_j) < \infty$. The class of estimators (6.21) is smoother, being continuous and weakly differentiable. Among these, the Block James-Stein estimator (6.4) makes the particular choice $\lambda_j = (n_j -$

$2)/n_j < 1$ and has the advantage that the oracle bound (6.5) deals simply with the events $\{\hat{\theta}_j = 0\}$ in risk calculations.

4. Theorem 6.1 is an apparently more precise result than was established for Hölder classes in the white noise case of Proposition 4.22, where full attention was not given to the constants. In fact the preceding argument goes through, since $\Theta^{\alpha}_{\infty}(C)$ defined in (4.53) satisfies all the required conditions, including block symmetry. However, $\Theta = \Theta^{\alpha}_{\infty}(C)$ lacks a simple explicit value for $R_N(\Theta, \epsilon)$, even asymptotically, though some remarks can be made. Compare Theorem 14.2 and Section 14.4.

6.4 Comparing adaptive linear estimators

We now give some examples to make two points: first, that many linear smoothing methods, with their tuning parameters chosen from the data, behave substantially similarly, and second, that the Block James Stein shrinkage approach leads to one such example, whether conducted in blocks of Fourier frequencies or in a wavelet domain.

Consider the continuous Gaussian white noise model (1.21) 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

$$\hat{\theta}_k = \kappa(hk) y_k. \tag{6.23}$$

As examples, we cite

1. Weighted Fourier series. The function κ 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. We saw in Section 3.3 that 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. Representation (6.23) follows after taking Fourier coefficients. Compare Lemma 3.7 and the examples given there.

3. Smoothing splines. We saw in Sections 1.4 and 3.4 that the estimator $\hat{\theta}_k$ minimizes

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

where the penalty term viewed in the time domain takes the form of a derivative penalty $\int (D^r f)^2$ for some integer r. In this case, $\hat{\theta}_k$ again has the representation (6.23) with $\lambda = h^{2r}$ and $\kappa(hk) = [1 + (hk)^{2r}]^{-1}$.

In addition, many methods of choosing h or λ from the data y have been shown to be asymptotically equivalent to first order —these include cross validation, Generalized cross validation, Rice's method based on unbiased estimates of risk, final prediction error, Akaike information criterion—see e.g. Härdle et al. (1988) for details and literature references. In this section we use a method based on an unbiased estimate of risk.

The implication of the adaptivity result Theorem 6.1 is that appropriate forms of the block James-Stein estimator should perform approximately as well as the best linear (or nonlinear) estimators, whether constructed by Fourier weights, kernels or splines, and without the need for an explicit choice of smoothing parameter from the data.

We will see this in examples below, but first we give an heuristic explanation of the close connection of these linear shrinkage families with the block James-Stein estimator (6.4).

Consider a Taylor expansion of $\kappa(s)$ about $s = 0$. If the time domain kernel $K(t)$ corresponding to κ is even about 0, then the odd order terms vanish and $\kappa(s) = 1 + \kappa_2 s^2/2 + \kappa_4 s^4/4! + \ldots$, so that for h small and a positive even integer q we have $\kappa(hk) \approx 1 - b_q h^q k^q$, compare (3.34).

Now consider grouping the indices k into blocks B_j–for example, dyadic blocks $B_j = \{k : 2^j \le k < 2^{j+1}\}$. Then the weights corresponding to two indices $k, \bar{k}$ in the same block are essentially equivalent: $k^{2r}/\bar{k}^{2r} \in [2^{-2r}, 2^{2r}]$ so that we may approximately write

$$\hat{\theta}_k \approx (1 - c_j) y_k, \qquad k \in B_j. \tag{6.24}$$

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 estimate c_j using an unbiased risk criterion, as described in Sections 2.5 and 2.6. Putting $C = (1 - c_j) I_{n_j}$ in the Mallows's C_L criterion (2.53) yields

$$U_{c_j}(y) = n_j \epsilon^2 - 2 n_j \epsilon^2 c_j + c_j^2 |y_j|^2. \tag{6.25}$$

[As noted below (2.58), this formula also follows from Stein's unbiased risk estimator applied to $\hat{\theta}_j(y) = y_j - c_j y_j$.] The value of c_j that minimizes (6.25) is $\hat{c}_j = n_j \epsilon^2 / \|y_j\|^2$, which differs from the James-Stein estimate (6.4) 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 (6.24). In the figures below, we compare four methods:

(i) *LPJS*: apply block James-Stein estimate (6.6) on each dyadic block in the Fourier frequency domain: $\hat{\theta}^{LPJS}(y) = (\hat{\theta}_j^{LPJS}(y_j))$. Dyadic blocking in the frequency domain is a key feature of Littlewood-Paley theory in harmonic analysis.

(ii) *WaveJS*: apply the James-Stein estimate (6.6) on each dyadic block in a wavelet coefficient domain: the blocks $y_j = (y_{jk}, k = 1, \ldots, 2^j)$.

(iii) *AutoSpline*: Apply a smoothing spline for the usual energy penalty $\int (f'')^2$ using a regularization parameter $\hat{\lambda}$ chosen by minimizing an unbiased estimator of risk.

(iv) *AutoTrunc*: In the Fourier frequency domain, use a cutoff function: $\hat{\kappa}(hl) = I\{l \le [h^{-1}]\}$ and choose the location of the cutoff by an unbiased risk estimator.

Implementation details. Let the original time domain data be $Y = (Y(l), l = 1, \ldots, N)$ for $N = 2^J$. The discrete Fourier transform (DFT), e.g. as implemented in MATLAB, sets

$$y(\nu) = \sum_{l=1}^{N} Y(l) e^{2\pi i (l-1)(\nu-1)/N}, \qquad \nu = 1, \ldots, N. \tag{6.26}$$

If the input Y is real, the output $y \in \mathbb{C}^N$ must have only N (real) free parameters. Indeed $y(1) = \sum_1^N Y(l)$ and $y(N/2+1) = \sum_1^N (-1)^l Y(l)$ are real, and for $r = 1, \ldots, N/2 - 1$, we have conjugate symmetry

$$y(N/2 + 1 + r) = \overline{y(N/2 + 1 - r)}. \tag{6.27}$$

Thus, to build an estimator, one can specify how to modify $y(1), \ldots, y(N/2+1)$ and then impose the constraints (6.27) before transforming back to the time domain by the inverse DFT.

1. (LPJS). Form dyadic blocks

$$y_j = \{\mathrm{Re}(y(\nu)), \mathrm{Im}(y(\nu)) : 2^{j-1} < \nu \le 2^j\}$$

for $j = 2, \ldots, J-1$. Note that $n_j = \#(y_j) = 2^j$. Apply the James Stein estimator (6.4) to each y_j, while leaving $y(\nu)$ unchanged for $\nu = 1, 2$. Thus $L = 2$, and we take $\epsilon^2 = (N/2)\sigma^2$, in view of (6.41).

2. (WaveJS). Now we use a discrete wavelet transform instead of the DFT. Anticipating the discussion in the next chapter, Y is transformed into wavelet coefficients $(y_{jk}, j = L, \dots, J-1, k = 1, \dots, 2^j)$ and scaling coefficients $(\tilde{y}_{Lk}, k = 1, \dots, 2^L)$. We use $L = 2, J = J_\epsilon$ and the Symmlet 8 wavelet, and apply Block James Stein to the blocks $y_j = (y_{jk} : k = 1, \dots, 2^j)$, while leaving the scaling coefficients $\tilde{y}_L$ unchanged.

3. (Autospline). We build on the discussion of periodic splines in Section 3.4. There is an obvious relabeling of indices so that in the notation of this section, $\nu = 1$ corresponds to the constant term, and each $\nu > 1$ to a pair of indices $2(\nu-1)-1$ and $2(\nu-1)$. Hence, linear shrinkage takes the form $\hat{\theta}_\lambda(\nu) = c_\nu(\lambda) y(\nu)$ with

$$c_\nu(\lambda) = [1 + \lambda(\nu-1)^4]^{-1}.$$

Note that $c_\nu(\lambda)$ is real and is the same for the "cosine" and "sine" terms. We observe that $c_1(\lambda) = 1$ and decree, for simplicity, that $c_{N/2+1}(\lambda) = 0$. Then, on setting $d_\nu = 1 - c_\nu$ and applying Mallow's C_L formula (2.53), we get an unbiased risk criterion to be minimized over λ:

$$U(\lambda) = N\epsilon^2 + \sum_{\nu=2}^{N/2} d_\nu(\lambda)^2 |y(\nu)|^2 - 4 d_\nu(\lambda)\epsilon^2,$$

4. (AutoTruncate). The estimator that cuts off at frequency ν_0 is, in the frequency domain,

$$\hat{\theta}_{\nu_0}(\nu) = \begin{cases} y(\nu) & \nu \le \nu_0 \\ 0 & \nu > \nu_0. \end{cases}$$

Using Mallows C_p, noting that each frequency ν corresponds to *two* real degrees of freedom, and neglecting terms that do not change with ν_0, we find that the unbiased risk criterion has the form

$$U_{\nu_0}(y) = -N\epsilon^2 + 4\nu_0\epsilon^2 + \sum_{\nu_0+1}^{N/2} |y(\nu)|^2, \qquad \nu_0 \in \{1, \dots, N/2\}.$$

These are applied to two examples: (a) the minimum temperature data introduced in Section 1.1, and (b) a 'blocky' step function with simulated i.i.d. Gaussian noise added. For simplicity in working with dyadic blocks, we have chosen a subset of $N = 256$ days [1]. The temperature data has correlated noise, so our theoretical assumptions don't hold exactly. Indeed, one can see the different noise levels in each wavelet band (cf Chapter 7.5). We used an upper bound of $\hat{\sigma} = 5$ in all cases. Also, the underlying function is not periodic over this range and forcing the estimator to be so leads to somewhat different fits than in Figure 1.1; the difference is not central to the discussion in this section.

The qualitative similarity of the four smoothed temperature fits is striking: whether an unbiased risk minimizing smoothing parameter is used with splines or Fourier weights, or whether block James-Stein shrinkage is used in the Fourier or wavelet domains. The similarity of the linear smoother and block James-Stein fits was at least partly explained near (6.24).

The similarity of the Fourier and wavelet James-Stein reconstructions may be explained as follows. The estimator (6.24) 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

[1] For non-dyadic sample sizes, see Section 7.8 for some references discussing wavelet transforms. In the Fourier domain, one might simply allow the block of highest frequencies to have dimension $N - 2^{[\log_2 N]}$.

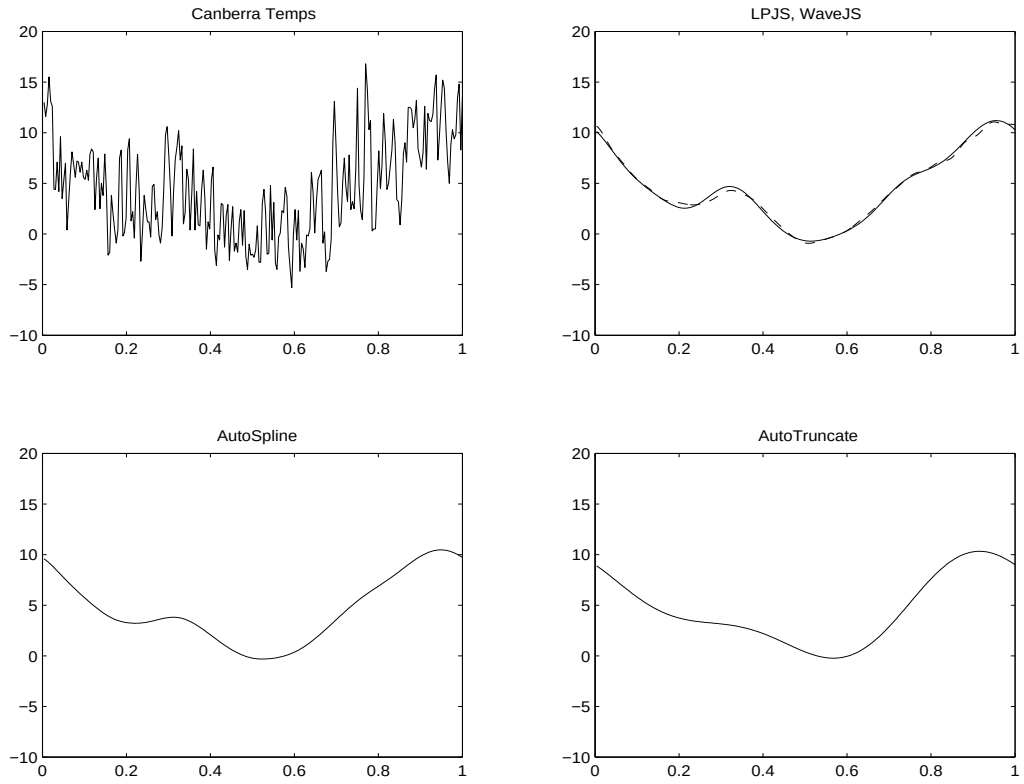

Figure 6.1 Top left: Canberra temperature data from Figure 1.1. Top right: block James-Stein estimates in the Fourier (solid) and wavelet (dashed) domains. Bottom panels: linear spline and truncation smoothers with bandwidth parameter chosen by minimizing an unbiased risk criterion.

belonging to that octave. The James-Stein estimator (6.4) is invariant to such orthogonal basis changes.

The (near) linear methods that agree on the temperature data also give similar, but now unsatisfactory, results on the 'blocky' example, Figure 6.2. Note that none of the methods are effective at simultaneously removing high frequency noise *and* maintaining the sharpness of jumps and peaks.

It will be the task of the next few chapters to explain why the methods fail, and how wavelet thresholding can succeed. For now, we just remark that the blocky function, which evidently fails to be differentiable, does not belong to any of the ellipsoidal smoothing classes $\Theta_2^\alpha(C)$ for $\alpha \geq 1/2$, based on the expectation that the Fourier coefficients decay at rate $O(1/k)$. The theorems of this and the previous chapter offer only slow convergence rate guarantees when $\alpha < 1/2$, which is consistent with the poor performance in Figure 6.2.

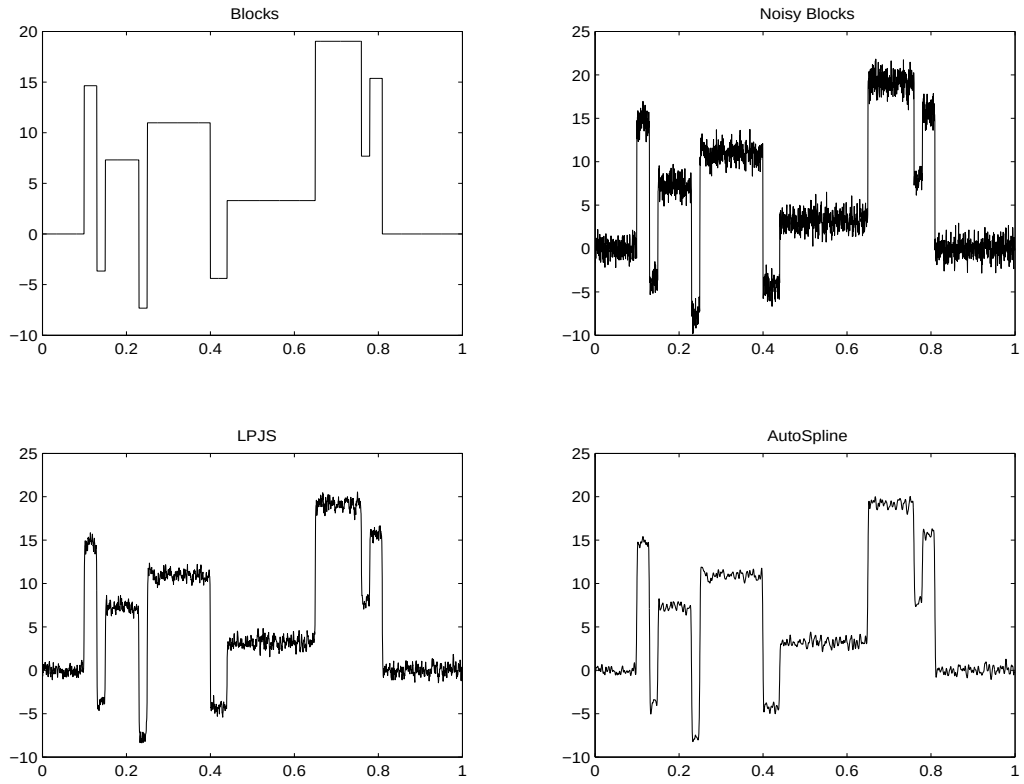

Figure 6.2 Top panels: A "blocky" step function with i.i.d Gaussian noise added, N = 2048. Bottom panels: selected reconstructions by block James-Stein and by smoothing spline (with data determined λ) fail to remove all noise.

Discussion

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 τ_i given by (5.6). Recalling that the θ_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 6.3 shows two such sample paths in the trigonometric basis (3.8) corresponding to smoothness $m = 1$ and $m = 2$ in (3.9). Despite the different levels of smoothness, notice that the spatial homogeneity in each case—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. Consider the NMR signal in Figure 1.2 or the plethysmograph signal in Figure 6.4. One sees regions of great "activity" or "oscillation" in the signal, and other regions of relative smoothness.

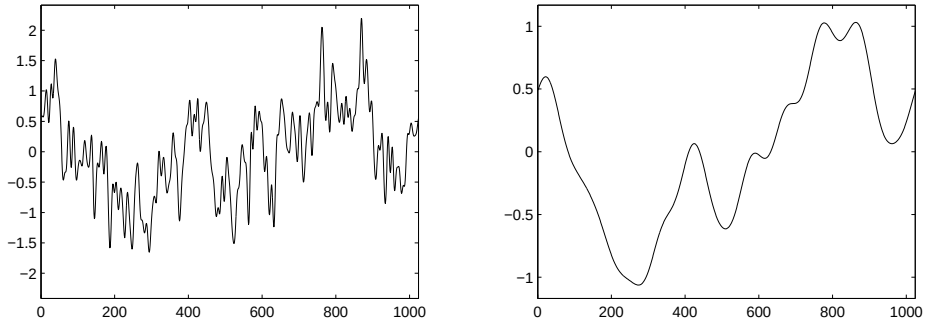

Figure 6.3 Sample paths from two Gaussian priors corresponding to (5.6) in Pinsker's theorem, which are near least favorable for ellipsoids in the trigonometric basis. In both cases $\epsilon = 0.5$ and $C = 500$. Left: mean square derivatives $\alpha = 1$. Right $\alpha = 2$. The corresponding computed values of $\mu(\epsilon, C)$ are 228.94 and 741.25 respectively.

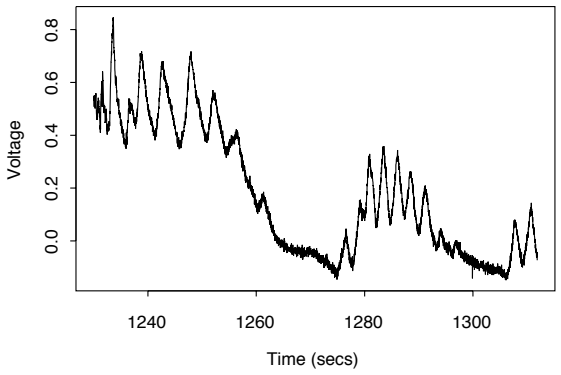

Figure 6.4 Sample signal from an inductance plethysmograph used to measure airflow in a patient. The larger scale oscillations correspond to normal breathing and the gap in the center to a vomiting episode. For more details, see Nason (2010). This figure is from Johnstone and Silverman (2005b).

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 asks 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 (5.5) comes from the appearance of smoothing methods with locally varying bandwidth such as LO(W)ESS, Cleveland (1979). We will see the locally varying bandwidth aspect of wavelet shrinkage in Section 7.5.

Commentary on the minimax approach. One may think of minimax decision theory as a method for evaluating the consequences of assumptions—the sampling model, loss function, and particularly the structure of the postulated parameter space Θ. The results of a minimax solution consist, of course, of the minimax strategy, the least favorable prior, the value, and also, information gained in the course of the analysis.

The minimax method can be successful if the structure of Θ is intellectually 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. For some datasets, as we have seen, this is a satisfactory description.

The clarity of the solution, paradoxically, also reveals some limitations of the formulation. The juxtaposition of the Pinsker priors and some other particular datasets suggests that for some scientific problems, one needs richer models of parameter spaces than ellipsoids. This is one motivation for the introduction of Besov bodies in Chapter 9.6 below.

6.5 *Interlude:* Superefficiency

This section looks at *one* of the motivations that underlies the use of worst-case and minimax analyses: a desire for a robust alternative to "fixed θ" asymptotics. In fixed θ asymptotics, the unknown function θ is kept fixed, and the risk behavior of an estimator sequence $\hat{\theta}_\epsilon$ is analysed 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 been effective in large sample analysis of finite dimensional models. Problems such as superefficiency are not serious enough to affect the practical implications widely drawn from Fisher's asymptotic theory of maximum likelihood.

In nonparametric problems with infinite dimensional parameter spaces, however, fixed θ asymptotics is more fragile. Used with care, it yields useful information. However, if optimization is pushed too far, it can suggest conclusions valid only for implausibly large sample sizes, and misleading for actual practice. In nonparametrics, superefficiency is more pervasive: even 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, we illustrate these points, along with concluding remarks about worst-case and minimax analysis.

Parametric Estimation: the Hodges example.

Suppose that $y \sim N(\theta, \epsilon^2)$ is a single scalar observation with ϵ small. A rather special case of Fisherian parametric asymptotics asserts that if $\hat{\theta}_\epsilon$ is an asymptotically normal and unbiased estimator sequence, $\epsilon^{-1}(\hat{\theta}_\epsilon - \theta) \xrightarrow{\mathcal{D}} N(0, v(\theta))$ when θ is true, *then* necessarily

$v(\theta) \geq 1$. A consequence for mean squared error would then be that

$$\liminf_{\epsilon \to 0} \epsilon^{-2} E_\theta(\hat{\theta}_\epsilon - \theta)^2 = \liminf r_\epsilon(\hat{\theta}, \theta)/R_N(\Theta, \epsilon) \geq 1.$$

[For this subsection, $\Theta = \mathbb{R}$.] Hodges' counterexample modifies the MLE $\hat{\theta}(y) = y$ in a shrinking neighborhood of a single point:

$$\hat{\theta}_\epsilon(y) = \begin{cases} 0 & |y| < \sqrt{\epsilon} \\ y & \text{otherwise.} \end{cases}$$

Since $\sqrt{\epsilon} = \frac{1}{\sqrt{\epsilon}} \cdot \epsilon$ is many standard deviations in size, it is clear that *if* $\theta = 0$, this estimator has MSE equal to $2\epsilon^2 \int_{\epsilon^{-1/2}}^{\infty} y^2 \phi(y) dy << \epsilon^2$. On the other hand, if $\theta \neq 0$ and ϵ 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 θ, and so

$$\epsilon^{-2} E_\theta(\hat{\theta}_\epsilon - \theta)^2 \to \begin{cases} 0 & \text{if } \theta = 0, \\ 1 & \text{otherwise,} \end{cases}$$

in clear violation of the Fisherian program. A fuller introduction to this and related superefficiency issues appears in Lehmann and Casella (1998, Section 6.2). Here we 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, for almost all θ,

$$\liminf_{\epsilon \to 0} \frac{r_\epsilon(\hat{\theta}_\epsilon, \theta)}{R_N(\Theta, \epsilon)} \geq 1. \tag{6.28}$$

(ii) Superefficiency entails poor performance at nearby points. For Hodges' example, consider $\theta_\epsilon = \sqrt{\epsilon}/2$. Since the threshold zone extends $1/(2\sqrt{\epsilon})$ standard deviations to the right of θ_ϵ, it is clear that $\hat{\theta}_\epsilon$ makes a squared error of $(\sqrt{\epsilon}/2)^2$ with high probability, so $\epsilon^{-2} r(\hat{\theta}_\epsilon, \sqrt{\epsilon}/2) \doteq \epsilon^{-2}(\sqrt{\epsilon}/2)^2 \to \infty$. Consequently

$$\sup_{|\theta| \leq \sqrt{\epsilon}} \frac{r(\hat{\theta}_\epsilon, \theta)}{R_N(\Theta, \epsilon)} \to \infty. \tag{6.29}$$

Le Cam, Huber and Hajek showed that more generally, superefficiency at θ_0 forces poor properties in a neighborhood of θ_0. Since broadly efficient estimators such as maximum likelihood are typically available with good risk properties, superefficiency has less relevance in parametric settings.

Remark. Hodges' estimator is an example of hard thresholding, to be discussed in some detail for wavelet shrinkage in non-parametric estimation. It is curious that the points of superefficiency that are unimportant for the one-dimensional theory become essential for sparse estimation of high dimensional signals.

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. (6.6) and Theorem 6.1(i) is

a theoretically and practically promising method. Nevertheless, every fixed θ is a point of superefficiency in the sense of (6.28):

Proposition 6.2 *(Brown et al., 1997) Consider the white Gaussian sequence model. Let $\Theta = \Theta_2^\alpha(C)$ be a Sobolev ellipsoid* (6.1)*, and let $\hat{\theta}_\epsilon^{BJS}$ be the block James-Stein estimator* (6.6) *corresponding to dyadic blocks* (6.3)*. Then for every $\theta \in \Theta$, as $\epsilon \to 0$,*

$$\frac{r_\epsilon(\hat{\theta}_\epsilon^{BJS}, \theta)}{R_N(\Theta, \epsilon)} \to 0. \tag{6.30}$$

Thus, if Θ 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 $\epsilon^{8/5}$, corresponding to $n^{-4/5}$ in sample size terms. Brown et al. (1997) also show that convergence cannot, in general, be very much faster – at best of logarithmic order in ϵ^{-1} – but the fixed θ 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 The proof uses dyadic blocks for concreteness; for extension to other blocking schemes, see Exercise 6.8. Fix $\Theta = \Theta_2^\alpha(C)$ and recall from (5.13) that $R_N(\Theta, \epsilon) \asymp \epsilon^{2r}$ as $\epsilon \to 0$, with $r = 2\alpha/(2\alpha + 1)$. A "fixed θ" bound for the risk of $\hat{\theta}^{BJS}$ follows from (6.18) and (6.12): indeed, with $L = 2$ and $ab/(a + b) \le \min(a, b)$, we may write

$$r_\epsilon(\hat{\theta}^{BJS}, \theta) \le 2J_\epsilon\epsilon^2 + \sum_j \min(n_j\epsilon^2, \|\theta_j\|^2) + \sum_{l > \epsilon^{-2}} \theta_l^2.$$

The proof of Theorem 6.1 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_ϵ, use the variance component below j_ϵ and the bias term thereafter:

$$R_1(\theta, \epsilon) \le 2^{j_\epsilon}\epsilon^2 + 2^{-2\alpha j_\epsilon} \sum_{j \ge j_\epsilon} 2^{2\alpha j} \|\theta_j\|^2.$$

To show that $R_1(\theta, \epsilon) = o(\epsilon^{2r})$, first fix a $\delta > 0$ and then choose j_ϵ so that $2^{j_\epsilon}\epsilon^2 = \delta\epsilon^{2r}$. [Of course, j_ϵ should be an integer, but there is no harm in ignoring this point.] It follows that $2^{-2\alpha j_\epsilon} = \delta^{-2\alpha}\epsilon^{2r}$, and so

$$\epsilon^{-2r} R_1(\theta, \epsilon) \le \delta + \delta^{-2\alpha} \sum_{j \ge j_\epsilon} 2^{2\alpha j} \|\theta_j\|^2 = \delta + o(1),$$

since the tail sum vanishes as $\epsilon \to 0$, for $\theta \in \Theta^\alpha(C)$. Since $\delta > 0$ is arbitrary, this shows that $R_1(\theta, \epsilon) = o(\epsilon^{2r})$ and establishes (6.30). □

The next result shows that for every consistent estimator sequence, and every parameter point $\theta \in \ell_2$, there exists a *shrinking* ℓ_2 neighborhood of θ over which the worst case risk of the estimator sequence is arbitrarily worse than it is at θ itself. Compare (6.29). 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 ℓ_2 norm: $B(\theta_0, \eta) = \{\theta : \|\theta - \theta_0\|_2 < \eta\}$. Such neighborhoods do not have compact closure in ℓ_2, and fixed θ asymptotics does not give any hint of the perils that lie arbitrarily close nearby.

Proposition 6.3 *Suppose that $\hat{\theta}_\epsilon$ is any estimator sequence such that $r_\epsilon(\hat{\theta}_\epsilon, \theta_0) \to 0$. Then there exists $\eta_\epsilon \to 0$ such that as $\epsilon \to 0$,*

$$\sup_{\theta \in B(\theta_0, \eta_\epsilon)} \frac{r_\epsilon(\hat{\theta}_\epsilon, \theta)}{r_\epsilon(\hat{\theta}_\epsilon, \theta_0)} \to \infty. \tag{6.31}$$

Remark. The result remains true if the neighborhood $B(\theta_0, \eta_\epsilon)$ is replaced by its intersection with any dense set: for example, the class of infinitely differentiable functions.

Proof Let $\gamma_\epsilon^2 = r_\epsilon(\hat{\theta}_\epsilon, \theta_0)$: we show that $\eta_\epsilon = \sqrt{\gamma_\epsilon}$ will suffice for the argument. The proof is a simple consequence of the fact that $\overline{B(1)} = \{\theta : \|\theta\|_2 \le 1\}$ is not compact (compare Theorem 5.7 or the example following Theorem 4.25), so that $R_N(B(1), \epsilon) \ge c_0 > 0$ even as $\epsilon \to 0$. All that is necessary is to rescale the estimation problem by defining $\bar{\theta} = \eta_\epsilon^{-1}(\theta - \theta_0)$, $\bar{y} = \eta_\epsilon^{-1}(y - \theta_0)$, $\bar{\epsilon} = \eta_\epsilon^{-1}\epsilon$, and so on. Then $\bar{y} = \bar{\theta} + \bar{\epsilon} z$ is an instance of the original Gaussian sequence model, and $\overline{B(\theta_0, \eta_\epsilon)}$ corresponds to the unit ball $\overline{B(1)}$. Rescaling the estimator also via $\hat{\bar{\theta}}_\epsilon(\bar{y}) = \eta_\epsilon^{-1}\big(\hat{\theta}_\epsilon(y) - \theta_0\big)$,

$$\gamma_\epsilon^{-2} E\|\hat{\theta}_\epsilon - \theta\|^2 = \eta_\epsilon^2 \gamma_\epsilon^{-2} E_{\bar{\epsilon}} \|\hat{\bar{\theta}}_\epsilon(\bar{y}) - \bar{\theta}\|^2,$$

and so, writing S_ϵ for the left side of (6.31), we obtain

$$S_\epsilon \ge \gamma_\epsilon^{-1} R_N(B(1), \epsilon) \ge c_0 \gamma_\epsilon^{-1} \to \infty. \qquad \square$$

Ultra-asymptotic bandwidth selection

Here is a "fixed-f" argument often encountered in asymptotics. Consider kernel estimators and the equispaced regression model discussed in Section 3.4. Using a qth order kernel, (3.28), in estimate $\hat{f}_h$, (3.21), leads to an approximate MSE expression, (3.31), of the form

$$r_a(h) = c_0(K)(nh)^{-1} + c_1(K)h^{2q} \int (D^q f)^2. \tag{6.32}$$

Then $r_a(h)$ is minimized at a bandwidth $h = h_n(f)$, and the minimum value $r_a(h_n(f))$ 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 $\tilde{f}_n$ of f, and uses $h_n(\tilde{f}_n)$ in the kernel estimate, such as (3.21) or (3.24). This approach goes back at least to Woodroofe (1970), for further references and discussion see Brown et al. (1997).

We study 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. We use the Fourier basis with $\mathbb{Z}$ as index, and let positive integers l label cosine terms of frequency l and negative l label the sine terms, so that

$$f(t) = \sum_{l \ge 0} \theta_l \cos 2\pi l t + \sum_{l < 0} \theta_l \sin 2\pi l t \tag{6.33}$$

As in Section 3.3 and 6.4, represent a kernel estimator in the Fourier domain by diagonal shrinkage

$$\hat{\theta}_{h,l} = \kappa(2\pi hl)y_l, \tag{6.34}$$

where $\kappa(s) = \int e^{-ist}K(t)dt$ is the Fourier transform of kernel K. The qth order moment condition becomes a statement about derivatives at zero, cf. (3.34). To simplify calculations, we use a specific choice of qth order kernel:

$$\kappa(2\pi s) = (1 - |s|^q)_+. \tag{6.35}$$

With this kernel, the mean squared error of (6.34) can be written explicitly as

$$r_\epsilon(\hat{\theta}_h, \theta) = \sum_{|l|\le[h^{-1}]} \epsilon^2(1 - |hl|^q)^2 + |hl|^{2q}\theta_l^2 + \sum_{|l|>[h^{-1}]} \theta_l^2. \tag{6.36}$$

We will do exact MSE calculations with this expression, but it is also helpful to use an integral approximations to the variance term and to approximate the squared bias term by $b_q(\theta) = \sum l^{2q}\theta_l^2$. This yields an approximate form

$$r_{a,\epsilon}(\hat{\theta}_h, \theta) = a_q\epsilon^2 h^{-1} + b_q(\theta)h^{2q},$$

which is exactly analogous to (6.32). Here $a_q = 4q^2(2q+1)^{-1}(q+1)^{-1}$, and $b_q(\theta)$ 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 asymptotically MSE-optimal bandwidth is found by minimizing $h \to r_{a,\epsilon}(\hat{\theta}_h, \theta)$. The Variance-Bias Lemma 3.6 gives

$$h_\epsilon = h_\epsilon(\theta) = \left[\frac{a_q\epsilon^2}{2qb_q(\theta)}\right]^{1/(2q+1)}, \tag{6.37}$$

and corresponding MSE

$$r_\epsilon(\hat{\theta}_{h_\epsilon(\theta)}, \theta) \sim c_q \left(2qb_q(\theta)\right)^{1/(2q+1)} \left(a_q\epsilon^2\right)^{2q/(2q+1)}, \tag{6.38}$$

with $c_q = 1 + (2q)^{-1}$. Thus the rate of convergence, $2q/(2q+1)$, 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 made arbitrarily close to 1, by using a kernel of sufficiently high order q.

However, this is an over literal use of fixed θ asymptotics – a hint of the problem is already suggested by the constant term in (6.38), which depends on $b_q(\theta)$ and could grow rapidly with q. However, we may go further and do exact MSE calculations with formula (6.36) using kernel (6.35). As specific test configurations in (6.33) we take

$$\theta_l = c(l_1, l_2)\begin{cases} |l|^{-3} & l \text{ even}, l \in [l_1, l_2] \\ |l|^{-3} & l \text{ odd}, -l \in [l_1, l_2] \\ 0 & \text{otherwise}, \end{cases} \tag{6.39}$$

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 choices are

$$\begin{aligned} &\text{(I)} \quad l_1 = 4, \quad l_2 = 20, \quad C = 60, \\ &\text{(II)} \quad l_1 = 4, \quad l_2 = 400, \quad C = 60. \end{aligned}$$

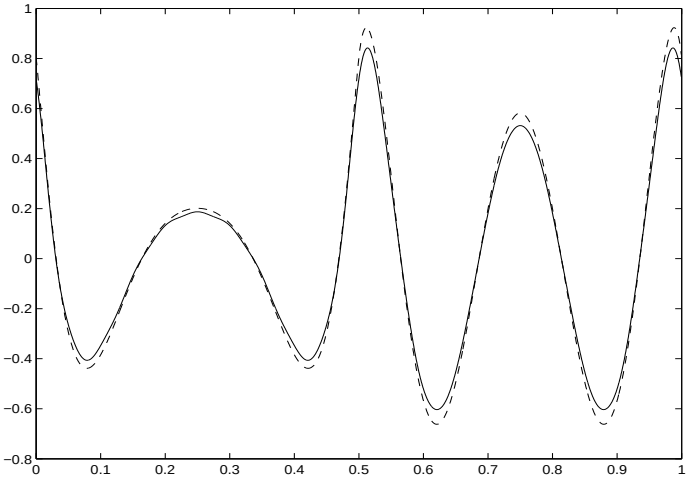

Figure 6.5 Two C^∞ functions, defined at (6.33)–(6.39). Solid line is θ^{I}, containing frequencies only through $l = 20$, dashed line is θ^{II}, with frequences up to $l = 400$.

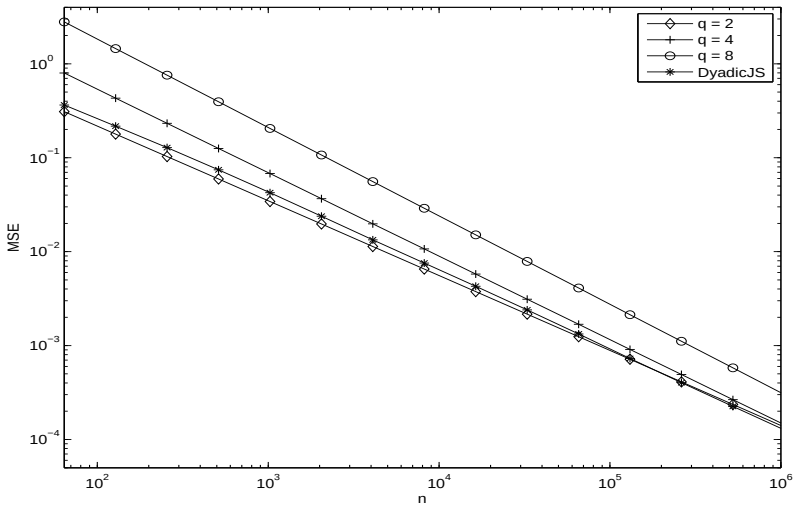

Figure 6.6 MSE of ideal bandwidth choice for θ^{II} : $r_\epsilon(\hat{\theta}_{h_\epsilon(\theta^{\mathrm{II}})}, \theta^{\mathrm{II}})$ resulting from qth order optimal bandwidth (6.37) for $q = 2, 4, 8$ with exact risks calculated using (6.36). Also shown is the upper bound (6.18) for the risk of the dyadic blocks James Stein estimator (6.6).

which differ only in the number of high frequency terms retained and are visually close, Figure 6.5.

Figure 6.6 shows the MSE $r_\epsilon(\hat{\theta}_{h_\epsilon(\theta^{\mathrm{II}})}, \theta^{\mathrm{II}})$ produced by using the qth order optimal bandwidth (6.37) for $q = 2, 4, 8$ with exact risks calculated using (6.36). Clearly the 8th order

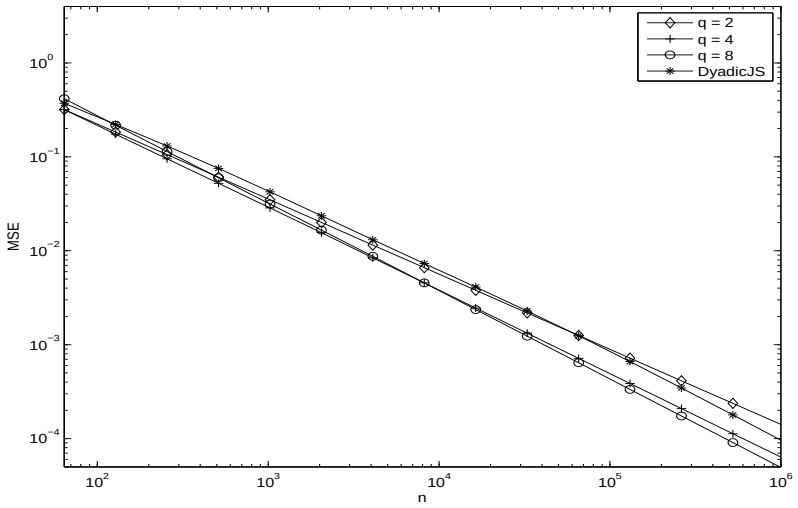

Figure 6.7 Corresponding plot of MSEs and James-Stein bound for ideal bandwidth choice for θ^{I}.

kernel is always several times worse than the 2nd order kernel for $n = \epsilon^{-2}$ 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 6.7 shows that the situation is not so bad in the case of curve I : because the higher frequencies are absent, the variance term in (6.36) is not so inflated in the $q = 8$ case.

However, with moderate noise levels ϵ, a test would not be able to discriminate beween θ^{I} and θ^{II}. This is an instance of the nearby instability of MSE seen earlier in this section.

We can also use (6.37) to compute the relative size of optimal bandwidths for the two functions, using $R_q = h_{\epsilon,q}(\theta_1)/h_{\epsilon,q}(\theta_2)$ as a function of q. Indeed, for $q = 2, 4, 8$, one computes that $R_q = 1, 2.6$ and 6.8.

Thus, at least for $q > 2$, both $h_\epsilon(\theta)$ and $r(\hat{\theta}_{h_\epsilon}, \theta)$ are very sensitive to aspects of the function that are difficult or impossible to estimate at small sample sizes. The fixed θ expansions such as (6.32) and (6.38) are potentially unstable tools.

Remarks. 1. Block James Stein estimation. Figures 6.6 and 6.7 also show the upper bounds (6.18) 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 6.8 compares the ratio $r_\epsilon(\hat{\theta}^{BJS}, \theta)/r_\epsilon(\hat{\theta}^q, \theta)$ of the Block JS mean squared error to the qth order kernel MSE over a much larger range of $n = \epsilon^{-2}$. The James Stein MSE *bound* is never much worse than the MSE of the qth order optimal bandwidth, and in many cases is much better.

2. Smoothness assumptions. Since θ^{I} and θ^{II} have finite Fourier expansions, they are certainly C^∞, but here they behave more like functions with about *two* square summable derivatives. Thus from the adaptivity Theorem 6.1, for α large, one expects that Block JS should eventually improve on the $q = 4$ and $q = 8$ kernels, and this indeed occurs in Figure 6.8 on the right side of the plot. However, the huge sample sizes show this "theoretical"

to be impractical. Such considerations point toward the need for quantitative measures of smoothness—such as Sobolev or Besov norms—that combine the *sizes* of the individual coefficients rather than qualitative hypotheses such as the mere existence of derivatives.

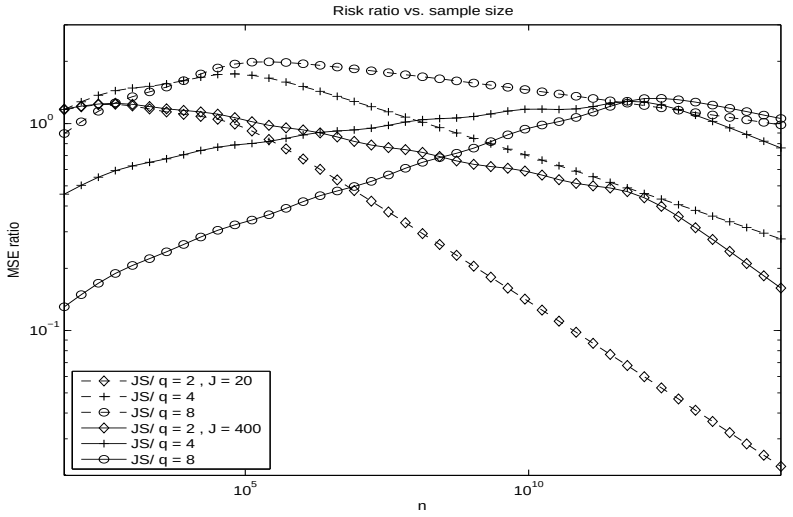

Figure 6.8 Ratio of James Stein MSE bound to actual MSE for kernels of order $q = 2, 4, 8$ at $\theta = \theta^{\mathrm{I}}$ (dotted) and θ^{II} (solid) over a wide range of sample sizes $n = \epsilon^{-2}$.

3. Speed limits. There is a uniform version of (6.38) that says that over ellipsoids of functions with α mean-square derivatives, the uniform rate of convergence using the qth order kernel is at best $(\epsilon^2)^{2q/(2q+1)}$, no matter how large α is. By contrast, the adaptivity results of Theorem 6.1 (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 way, the antithesis of fixed θ analysis. The least favorable configuration—whether parameter point θ_ϵ or prior distribution π_ϵ—will generally change with noise level ϵ. This is natural, since the such configurations represent the "limit of resolution" attainable, which improves as the noise diminishes.

The choice of the space Θ to be maximized over is certainly critical, and greatly affects the least favorable configurations found. This at least has the virtue of making clearer the consequences of assumptions—far more potent in nonparametrics, even if hidden. It might be desirable to have some compromise between the local nature of fixed θ asymptotics, and the global aspect of minimax analysis—perhaps in the spirit of the local asymptotic 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.

6.6 Notes

§2 and 3. The first results in the adaptive minimax setting are due to Efromovich and Pinsker (1984), who pioneered the use of estimator (6.22), and Golubev (1987). The approach of §2 and §3 follows that of Donoho and Johnstone (1995). Our definition of oscillation of ratios is not the standard additive one: usually $\omega_f(B) = \sup_{x\in B} f - \inf_{x\in B} f$ and so our definition $\mathrm{osc}_a(B) = \exp(\omega_{\log a}(B))$.

Cavalier and Tsybakov (2001) introduce the term 'penalized blockwise Stein rule' for the variant (6.21), and use it to establish sharp oracle inequalities and sharp asymptotic minimaxity results for very general classes of ellipsoids, along with near optimal results for Besov spaces. They also emphasize the use of weakly geometric blocks, which were studied by Nemirovski (2000). Cavalier and Tsybakov (2002) extend the penalized blockwise Stein approach to linear inverse problems. Efromovich (2004b) establishes a similar oracle inequality for the Efromovich-Pinsker estimator (6.22) under weaker assumptions on the noise model. In the spirit of the extension of these results to other nonparametric models, as discussed in Section 3.11, we mention the sharp adaptivity results of Efromovich and Pinsker (1996) for nonparametric regression with fixed or random design and heteroscedastic errors. Rigollet (2006) has a nice application of these ideas to adaptive density estimation on $\mathbb{R}$.

We have focused on functions of a single variable: Efromovich (2010) gives an example of use of thresholding and dyadic blocking for a series estimator in a fairly flexible multivariate setting.

§4. The comparison of linear methods draws from Donoho and Johnstone (1995) and Donoho et al. (1995). Johnstone (1994) has more on drawing sample paths from least favorable and near least favorable priors on ellipsoids and Besov balls.

§5. The first part of this section borrows from Brown et al. (1997), in particular Proposition 6.2 is a version of Theorem 6.1 there. van der Vaart (1997) gives a review of the history and proofs around superefficiency. These articles contain full references to the work of Le Cam, Huber and Hajek.

The exact risk analysis is inspired by the study of density estimation in Marron and Wand (1992), which in turn cites Gasser and Müller (1984). 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 try to consider the behavior of "plug-in" methods that attempt to estimate $h_\epsilon(\theta)$ – variability in the data based estimates of $h_\epsilon(\theta)$ 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 $q = 2$.

While the choice $q = 8$ may seem extreme in the setting of traditional density estimation, it is standard to use wavelets with higher order vanishing moments – for example, the Daubechies Symmlet 8 discussed in Daubechies (1992, p. 198-199) or Mallat (1998, p. 252), see also Chapter 7.1. Analogs of (6.32) and (6.38) for wavelet based density estimates appear in Hall and Patil (1993), though of course these authors do not use the expansions for bandwidth selection.

Exercises

6.1 *(Equivalence of Fourier and dyadic Sobolev norms.)* Fix $\alpha > 0$. In the Fourier ellipsoid case, let $a_k = k^\alpha$ for $k \geq 1$. For the dyadic case, let $\tilde{a}_l = 2^{j\alpha}$ if $l = 2^j + k$ for $j \geq 0$ and $k = 0, 1, \ldots, 2^j - 1$. Verify for $l \geq 1$ that

$$2^{-\alpha} a_l \leq \tilde{a}_l \leq a_l.$$

and hence obtain the inequalities (6.9).

6.2 *(Speed limits for qth 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 $\hat{\theta}_{h,l} = \kappa(hl)y_l$.

(a) Explain why it is reasonable to express the statement "K is a qth 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 $\Theta_2^\alpha(C) = \{\theta : \sum a_l^2 \theta_l^2 \leq C^2\}$ with $a_l = l^\alpha$ be, as usual, an ellipsoid of $\alpha-$mean square differentiable functions. If K is a qth order kernel in the sense of part (a), show that for each $\alpha > q$,

$$\inf_{h>0} \sup_{\theta\in\Theta^\alpha(C)} r_\epsilon(\hat{\theta}_h, \theta) \geq c(\alpha, q, C)(\epsilon^2)^{2q/(2q+1)}.$$

[Thus, for a second order kernel, the (uniform) rate of convergence is $n^{-4/5}$, 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.]

6.3 *(Oscillation within blocks.)* Let $\Theta(a, C)$ be an ellipsoid $\{(\theta_i) : \sum a_i^2\theta_i^2 \le C^2\}$. Assume that $a_i \nearrow \infty$. Let blocks $B_{j\epsilon}$ be defined as in (6.3) and the oscillation of a_i within blocks by

$$\mathrm{osc}(B_{j\epsilon}) = \max_{l,l' \in B_{j\epsilon}} \frac{a_l}{a_{l'}}.$$

Show that if $\max_{k\ge j} \mathrm{osc}(B_{k\epsilon}) \to 1$ as $j \to \infty$ and $\epsilon \to 0$ jointly, then

$$R_L(\Theta, \epsilon) \sim R_{BL}(\Theta, \epsilon) \qquad \text{as } \epsilon \to 0.$$

The next two exercises consider sets Θ more general than ellipsoids.

6.4 *(Minimax theorem for block linear estimators).* Show that the minimax theorem, (6.13), holds if Θ is compact, solid-orthosymmetric and quadratically convex.

6.5 *(Block linear minimaxity.)* This exercise shows that if Θ is compact, solid-orthosymmetric and block-symmetric, then

$$R_L(\Theta, \epsilon) = R_{BL}(\Theta, \epsilon) \qquad \text{for all } \epsilon > 0. \tag{6.40}$$

(i) Suppose first that Θ is also quadratically convex. Define a vector $\bar\theta(\theta)$ from θ by replacing θ_i by $\|\theta_j\|\sqrt{n_j}$, that is, θ_i^2 is replaced by its average on the block in which i lies. Show that $\theta \in \Theta$ implies $\bar\theta(\theta) \in \Theta$.

(ii) Establish (6.40) assuming that Θ is also quadratically convex.

(iii) Show, using Theorem 9.5, that the assumption of quadratic convexity can be removed.

6.6 *(White noise in frequency domain).* Consider the discrete Fourier transform (6.26). Suppose in addition that the $Y(l)$ are i.i.d. mean zero, variance σ^2 variables and N is even. Show that

$$\mathrm{Var}(\mathrm{Re}(y(v))) = \mathrm{Var}(\mathrm{Im}(y(v))) = (N/2)\sigma^2. \tag{6.41}$$

6.7 *(Time domain form of kernel* (6.35)*).* Let $L(t) = \sin t/(\pi t)$, and assume, as in (6.35), that $\kappa(s) = (1 - |s|^q)_+$. Show that the corresponding time domain kernel

$$K(t) = L(t) - (-i)^q L^{(q)}(t).$$

Make plots of K for $q = 2, 4$ and compare with Figure 3.1. Why is the similarity not surprising?

6.8 *(Superefficiency for Block James-Stein.)* In Proposition 6.2, suppose that $\Theta_2^\alpha(C)$ is given. Show that conclusion (6.30) holds for *any* blocking scheme (6.3) for which $J_\epsilon\epsilon^2 = o(\epsilon^{2r})$.

6.9 *(Exact risk details.)* This exercise records some details leading to Figures 6.5–6.8.
(i) For vectors $x, X \in \mathbb{C}^N$, the inverse discrete Fourier transform `x = ifft(X)` sets $x(j) = N^{-1}\sum_{k=1}^N X(k)e^{-2\pi i(j-1)(k-1)/N}$, $j = 1, \dots, N$. Suppose now that

$$X(1) = N\theta_0, \qquad \mathrm{Re}[X(l+1)] = N\theta_l, \quad \mathrm{Im}[X(l+1)] = N\theta_{-l}$$

for $1 \le l < N/2$ and $X(k) = 0$ for $k > N/2$. Also, set $t_j = j/N$. Verify that

$$\mathrm{Re}[x(j)] = f(t_{j-1}) = \theta_0 + \sum_{l=1}^{N/2} \theta_l \cos 2\pi l t_{j-1} + \theta_{-l} \sin 2\pi l t_{j-1}, \qquad j = 1, \dots, N.$$

(ii) Consider the sequence model in the form $y_l = \theta_l + \epsilon z_l$ for $l \in \mathbb{Z}$. For the coefficients specified by (6.39) and below, show that the risk function (6.36) satisfies

$$r(\hat{\theta}_h, \theta) = \epsilon^2 + 2\epsilon^2 \sum_1^{l_h} [1 - (hl)^q]^2 + h^{2q} C_{12}^2 \sum_{l=l_1}^{l_2 \wedge l_h} j^{2q-6} + C_{12}^2 \sum_{l_h+1}^{l_2} j^{-6},$$

where $l_h = [h^{-1}]$ and $C_{12}^2 = C^2 / \sum_{l=l_1}^{l_2} j^{-2}$.
(iii) Introduce functions (which also depend on l_1, l_2 and C)

$$V(m, n; h, q) = \sum_{l=m}^{n} [1 - (hl)^q]^2, \qquad B(m, n; p) = C_{12}^2 \sum_{l=m \vee l_1}^{n \wedge l_2} j^{p-6}.$$

Confirm that in terms of V and B,

$$b_q(\theta) = C_{12}^2 \sum_{l_1}^{l_2} j^{2q-6} = B(l_1, l_2; 2q)$$

$$r(\hat{\theta}_h, \theta) = \epsilon^2 + 2\epsilon^2 V(1, l_h; h, q) + h^{2q} B(1, l_h; 2q) + B(l_h + 1, l_2; 0).$$

The figures use a vector of values of ϵ^2 and hence of $h = h_\epsilon$ in (6.37) and l_h; these representations facilitate the vectorization of the calculations.
(iv) For the block James-Stein estimator, define blocks $y_b = (y_l, 2^{b-1} < |l| \le 2^b)$, so that $n_b = 2^b$. Choose $n_\epsilon = \epsilon^{-2}$ so that $J_\epsilon = \log_2 n_\epsilon$ is an integer. Show that (6.18) becomes

$$r_\epsilon(\hat{\theta}^{BJS}, \theta) \le (2J_\epsilon + 1)\epsilon^2 + \sum_{b=2}^{J_\epsilon - 1} \frac{n_b B_b}{n_b + B_b n_\epsilon} + B_\epsilon,$$

where $B_b = B(2^{b-1} + 1, 2^b; 0)$ and $B_\epsilon = B(2^{J_\epsilon - 1} + 1, l_2; 0)$.

7

A Primer on Estimation by Wavelet Shrinkage

> When I began to look at what Meyer had done, I realized it was very close to some ideas in image processing. Suppose you have an image of a house. If you want to recognize simply that it is a house, you do not need most of the details. So people in image processing had the idea of approaching the images at different resolutions. (Stéphane Mallat, quoted in *New York Times*.)

The essence of the wavelet transform is to decompose a signal or image into subparts of increasing detail or resolution. In the presence of noisy data, and when combined with thresholding, this *multiresolution* approach provides a powerful tool for estimating the underlying object.

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 signal or regression data observed in noise. 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. Both this introductory account and the later theory will show how the shortcomings of linear estimators can be overcome by appropriate use of simple non-linear thresholding. We do not attempt to be encyclopedic in coverage of what is now a large area, rather we concentrate on orthogonal wavelet bases and the associated multiresolution analyses for functions of a single variable.

The opening quote hints at the interplay between disciplines that is characteristic of wavelet theory and methods, and so is reflected in the exposition here.

Section 7.1 begins with the formal definition of a multiresolution analysis (MRA) of square integrable functions, and indicates briefly how particular examples are connected with important wavelet families. We consider decompositions of $L_2(\mathbb{R})$ and of $L_2([0,1])$, though the latter will be our main focus for the statistical theory.

This topic in harmonic analysis leads directly into a signal processing algorithm: the "two-scale" relations between neighboring layers of the multiresolution give rise in Section 7.2 to filtering relations which, in the case of wavelets of compact support, lead to the fast $O(n)$ algorithms for computing the direct and inverse wavelet transforms on discrete data.

Section 7.3 explains in more detail how columns of the discrete wavelet transform are related to the continuous wavelet and scaling function of the MRA, while Section 7.4 describes the changes needed to adapt to finite data sequences.

Finally in Section 7.5 we are ready to describe wavelet thresholding for noisy data using the discrete orthogonal wavelet transform of $n = 2^J$ equally spaced observations. The

'hidden sparsity' heuristic is basic: the wavelet transform of typical 'true' signals is largely concentrated in a few co-ordinates while the noise is scattered throughout, so thresholding will retain most signal while suppressing most noise.

How the threshold itself is set is a large question we will discuss at length. Section 7.6 surveys some of the approaches that have been used, and for which theoretical support exists. The discussion in these two sections is informal, with numerical examples. Corresponding theory is developed in later chapters.

7.1 Multiresolution analysis

A wavelet is of course a little wave, the name being chosen to indicate two key properties: oscillation and short duration. The remarkable feature is that by stretching and shifting the wavelet one can, under suitable conditions, obtain a system capable of representing an arbitrary (square integrable) function. When that system is an orthonormal basis–the case of main interest for us–we refer to the generator ψ as an orthonormal wavelet.

This is not an *ab initio* exposition of wavelet ideas and theorems: some authoritative books include Meyer (1990), Daubechies (1992), Mallat (2009) and others listed in the chapter notes. Rather we present, without proofs, some 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.

It is a striking fact that the fast algorithms for *discrete* orthogonal wavelet transforms have their origin in change of basis operations on square integrable functions of a *continuous* variable. We therefore begin with the notion of a multiresolution analysis of $L_2(\mathbb{R})$. We concentrate on the univariate case, though the ideas extend to $L_2(\mathbb{R}^d)$. Constructions in the frequency domain play an important role, but these are largely deferred to a sketch in Appendix B.1 and especially the references given there.

Definition 7.1 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) $V_j \subset V_{j+1}$,
(ii) $f(x) \in V_j$ if and only if $f(2x) \in V_{j+1}, \forall j \in \mathbb{Z}$,
(iii) $\cap_{j\in\mathbb{Z}} V_j = \{0\}, \qquad \overline{\cup_{j\in\mathbb{Z}} V_j} = L_2(\mathbb{R})$.
(iv) there 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 φ in (iv) is called the *scaling function* of the given MRA. Set $\varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k)$. One says that φ_{jk} has scale 2^{-j} and location $k2^{-j}$. Properties (ii) and (iv) imply that $\{\varphi_{jk}, k \in \mathbb{Z}\}$ is an orthonormal basis for V_j. The orthogonal projection from $L_2(\mathbb{R}) \to V_j$ 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 (iii) implies that $P_j f \to f$ in $L_2(\mathbb{R})$ as $j \to \infty$.

Example. Haar MRA. Set $I_{jk} = [2^{-j}k, 2^{-j}(k+1)]$. The "Haar multiresolution analysis" is defined by taking $\varphi = I_{[0,1]}$ and

$$V_j = \{f \in L_2(\mathbb{R}) : f|_{I_{jk}} = c_{jk}\},$$

with $c_{jk} \in \mathbb{R}$. Thus V_j consists of piecewise constant functions on intervals of length 2^{-j}, and $P_j f(x)$ is the average of f over the interval I_{jk} that contains x.
Example. Box spline MRA. Given $r \in \mathbb{N}$, set

$$V_j = \{f \in L_2 \cap C^{r-1} \text{and } f|_{I_{jk}} \text{ is a polynomial of degree } r\}.$$

If $r = 0$, this reduces to the Haar MRA. If $r = 1$, we get continuous, piecewise linear functions and if $r = 3$, cubic splines. For more on the construction of the scaling function φ, see Appendix B.1.

A key role in wavelet analysis is played by a pair of *two scale equations*, (7.1) and (7.3), and their associated discrete filter sequences. Given an MRA with scaling function φ, since $V_{-1} \subset V_0$, one may express $\varphi_{-1,0}$ in terms of $\varphi_{0,k}$ using the first of the two scale equations

$$\frac{1}{\sqrt{2}}\varphi\Big(\frac{x}{2}\Big) = \sum_k h[k]\varphi(x-k). \tag{7.1}$$

The sequence $\{h[k]\}$ is called the *discrete filter* associated with φ. For the Haar MRA example, $h[0] = h[1] = 1/\sqrt{2}$, while all other $h[k]$ vanish.

Now take Fourier transforms, (C.10), of both sides: since $\widehat{\varphi_{0k}}(\xi) = e^{-ik\xi}\widehat{\varphi}(\xi)$, the two scale equation has the re-expression

$$\widehat{\varphi}(2\xi) = 2^{-1/2}\widehat{h}(\xi)\widehat{\varphi}(\xi), \tag{7.2}$$

where the *transfer function*

$$\widehat{h}(\xi) = \sum h[k]e^{-ik\xi}.$$

The MRA conditions (i)–(iv) imply important structural constraints on $\widehat{h}(\xi)$. These in turn lead to theorems describing how to construct scaling functions φ. Some of these are reviewed, with references, in Appendix B.1.

Now we turn to the 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 for a wavelet $\psi \in W_{-1} \subset V_0$ must satisfy a second two scale equation

$$\frac{1}{\sqrt{2}}\psi(\frac{x}{2}) = \sum_k g[k]\varphi(x-k). \tag{7.3}$$

Again, taking the Fourier transform of both sides and defining $\widehat{g}(\xi) = \sum g_k e^{-ik\xi}$,

$$\widehat{\psi}(2\xi) = 2^{-1/2}\widehat{g}(\xi)\widehat{\varphi}(\xi). \tag{7.4}$$

Define recentered and scaled wavelets $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. Suppose that it is possible to define ψ using (7.4) so that $\{\psi_{jk}, k \in \mathbb{Z}\}$ form an orthonormal basis for W_j. Then it may be shown from property (iii) of the MRA that the full collection $\{\psi_{jk}, (j,k) \in \mathbb{Z}^2\}$ forms an orthonormal basis for $L_2(\mathbb{R})$.

Thus we have two decompositions

$$L_2(\mathbb{R}) = \oplus_{j\in\mathbb{Z}} W_j = V_J \oplus \bigoplus_{j\geq J} W_j,$$

for each J, with corresponding expansions

$$f = \sum_{j,k} \langle f, \psi_{jk}\rangle \psi_{jk} = \sum_k \langle f, \varphi_{Jk}\rangle \varphi_{Jk} + \sum_{j\geq J}\sum_k \langle f, \psi_{jk}\rangle \psi_{jk}. \tag{7.5}$$

The first is called a *homogeneous* expansion, while the second is said to be inhomogeneous since it combines only the detail spaces at scales finer than J.

Figure 7.1(left) shows some examples of ψ_{jk} for a few values of j, k: as elements of an orthonormal basis, they are mutually orthogonal with L_2-norm equal to 1.

A key heuristic idea is that for typical functions f, the wavelet coefficients $\langle f, \psi_{jk}\rangle$ are large only at low frequencies or wavelets located close to singularities of f. This heuristic notion is shown schematically in Figure 7.1(right) and is quantified in some detail in Section 9.6 and Appendix B.

Here is a simple result describing the wavelet coefficients of piecewise constant functions.

Lemma 7.2 *Suppose ψ has compact support $[-S, S]$ and $\int \psi = 0$. Suppose f is piecewise constant with d discontinuities. Then at level j at most $(2S-1)d$ of the wavelet coefficients $\theta_{jk} = \int f\psi_{jk}$ are non-zero, and those are bounded by $c2^{-j/2}$, with $c = \|\psi\|_1\|f\|_\infty$.*

Proof Let the discontinuities of f occur at $x_1, \ldots, x_d$. Since $\int \psi = 0$,

$$\theta_{jk} = \int f\psi_{jk} = 2^{-j/2}\int f(2^{-j}(t+k))\psi(t)dt$$

vanishes unless some x_i lies in the interior of $\mathrm{supp}(\psi_{jk})$. In this latter case, we can use the right hand side integral to bound $|\theta_{jk}| \leq \|f\|_\infty\|\psi\|_1 2^{-j/2}$. The support of ψ_{jk} is $k2^{-j} + 2^{-j}[-S, S]$, and the number of k for which $x_i \in \mathrm{int}(\mathrm{supp}(\psi_{jk}))$ is at most $2S-1$. So the total number of non-zero θ_{jk} at level j is at most $(2S-1)d$. □

The construction of some celebrated pairs (φ, ψ) of scaling function and wavelet is sketched, with literature references, in Appendix B.1. Before briefly listing some of the well known families, we discuss several properties that the pair (φ, ψ) might possess.

Support size. Suppose that the support of ψ is an interval of length S, say $[0, S]$. Then ψ_{jk} is supported on $k2^{-j} + 2^{-j}[0, S]$. Now suppose also that f has a singularity at x_0. The size of S determines the range of influence of the singularity on the wavelet coefficients $\theta_{jk}(f) = \int f\psi_{jk}$. Indeed, at level j, the number of coefficients that 'feel' the singularity at x_0 is just the number of wavelet indices k for which supp ψ_{jk} covers x_0, which by rescaling is equal to S (or $S-1$ if x_0 lies on the boundary of $\mathrm{supp}\psi_{jk}$).

It is therefore in principle desirable to have small support for ψ and φ. These are in turn determined by the support of the filter h, by means of the two scale relations (7.1) and (7.3). 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, Mallat (1999, Chapter 7) shows that

(i) $\mathrm{supp}\,\varphi = \mathrm{supp}\,h$ if one of the two is compact, and

(ii) if $\mathrm{supp}\,\varphi = [N_1, N_2]$, then $\mathrm{supp}\,\psi = [\frac{N_1-N_2+1}{2}, \frac{N_2-N_1+1}{2}]$.

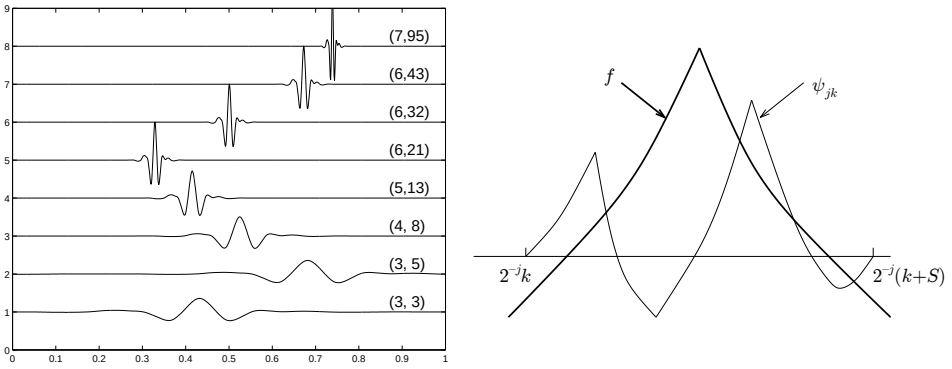

Figure 7.1 Left panel: Wavelets (from the Symmlet-8 family), the pair (j,k) indicates wavelet ψ_{jk}, at resolution level j and approximate location $k2^{-j}$. Right panel: Schematic of a wavelet ψ_{jk} of compact support "hitting" a singularity of function f.

Vanishing moments. The wavelet ψ is said to have r *vanishing moments* if

$$\int x^l \psi(x)dx = 0 \qquad l = 0, 1, \ldots r-1. \tag{7.6}$$

Thus ψ is orthogonal to all polynomials of degree $r-1$. As a result, the rate of decay of wavelet coefficients of a smooth function is governed by the number of vanishing moments of the wavelet ψ. For example, in Appendix B.1 we prove:

Lemma 7.3 *If f is C^α on $\mathbb{R}$ and ψ has $r \geq \lceil \alpha \rceil$ vanishing moments, then*

$$|\langle f, \psi_{jk} \rangle| \leq c_\psi C 2^{-j(\alpha+1/2)}.$$

If α is a positive integer, then the C^α assumption is just the usual notion that f has α continuous derivatives, and the constant $C = \|D^\alpha f\|_\infty / \alpha!$. For $\alpha > 0$ non-integer, we use the definition of Hölder smoothness of order α, given in Appendix C.23. Note the parallel with the definition (3.28) of vanishing moments for an averaging kernel K, and the expression (3.29) for the approximation error of a qth order kernel.

Daubechies (1988) showed that existence of p vanishing moments for an orthogonal wavelet implied a support length for h, and hence for φ, ψ, of at least $2p-1$. Thus, for such wavelets, there is a tradeoff between short support and large numbers of vanishing moments. A resolution of this tradeoff is perhaps best made according to the context of a given application.

Regularity. Given an estimate $\hat{f}(x)$ of function f, we see by writing out the wavelet expansion in (7.5) as $\hat{f}(x) = \sum \hat{\theta}_{jk} \psi_{jk}(x)$ that the smoothness of $x \to \psi_{jk}(x)$ can impact the visual appearance of a reconstruction. However it is the number of vanishing moments that affects the size of wavelet coefficients at fine scales, at least in regions where f is smooth. So both properties are in general relevant. For the common wavelet families [to be reviewed below], it happens that regularity increases with the number of vanishing moments.

For orthonormal wavelet bases, regularity of ψ implies that a corresponding number of moments vanish. We refer to Daubechies (1992, §5.5) for the proof of

Proposition 7.4 *If $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ is an orthonormal basis for $L_2(\mathbb{R})$, and if ψ is C^r, with $\psi^{(k)}$ bounded for $k \le r$ and $|\psi(x)| \le C(1+|x|)^{-r-1-\epsilon}$, then $\int x^k \psi(x)dx = 0$ for $k = 0, \dots, r$.*

Some wavelet families. The common constructions of instances of (φ, ψ) use Fourier techniques deriving from the two scale equations (7.2) and (7.4) and the filter transfer function $\widehat{h}(\xi)$. Many constructions generate a family indexed by the number of vanishing moments p. For some further details see the examples in Appendix B.1, and wavelet texts, such as Mallat Ch.7 or Daubechies (1992). Figure 7.2 shows some examples of the families to follow.

Haar. The simplest and only rarely best: $\varphi = I_{[0,1]}$ and $\psi = I_{[0,1/2]} - I_{[1/2,1]}$. It has a single vanishing moment, and of course no smoothness.

Meyer. $\widehat{\varphi}(\xi), \widehat{\psi}(\xi)$ have compact support in frequency ξ, and so $\varphi(x)$ and $\psi(x)$ are C^∞, but do not have compact support in x – in fact they have only polynomial decay for large x. The wavelet has infinitely many vanishing moments.

Battle-Lemarié spline. These are wavelets derived from the spline MRA. The pair $\varphi(x), \psi(x)$ are polynomial splines of degree m and hence are C^{m-1} in x. They have exponential decay in x, and are symmetric (resp. anti-symmetric) about $x = 1/2$ for m odd (resp. even). The wavelet has $m+1$ vanishing moments.

Compact support wavelets. Daubechies constructed several sets of compactly supported wavelets and scaling functions, indexed by the number of vanishing moments p for ψ.

(a) "Daubechies" family – the original family of wavelets D_{2p} in which ψ has minimum support length $2p-1$, on the interval $[-p+1, p]$. The wavelets are quite asymmetric, and have regularity that grows roughly at rate $0.2p$, though better regularity is known for small p – e.g. just over C^1 for $p = 3$.

(b) "Symmlet" family – another family with minimum support $[-p+1, p]$, but with filter h chosen so as to make ψ as close to symmetric (about $\frac{1}{2}$) as possible.

(c) "Coiflet" family – a family with $K = 2p$ vanishing moments for ψ and also for φ:

$$\int \varphi = 1, \qquad \int t^k \varphi = 0, \quad 1 \le k < K.$$

This constraint forces a larger support length, namely $3K - 1$.

Wavelets on the interval $[0, 1]$.

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 extensions by reflection or folding, 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

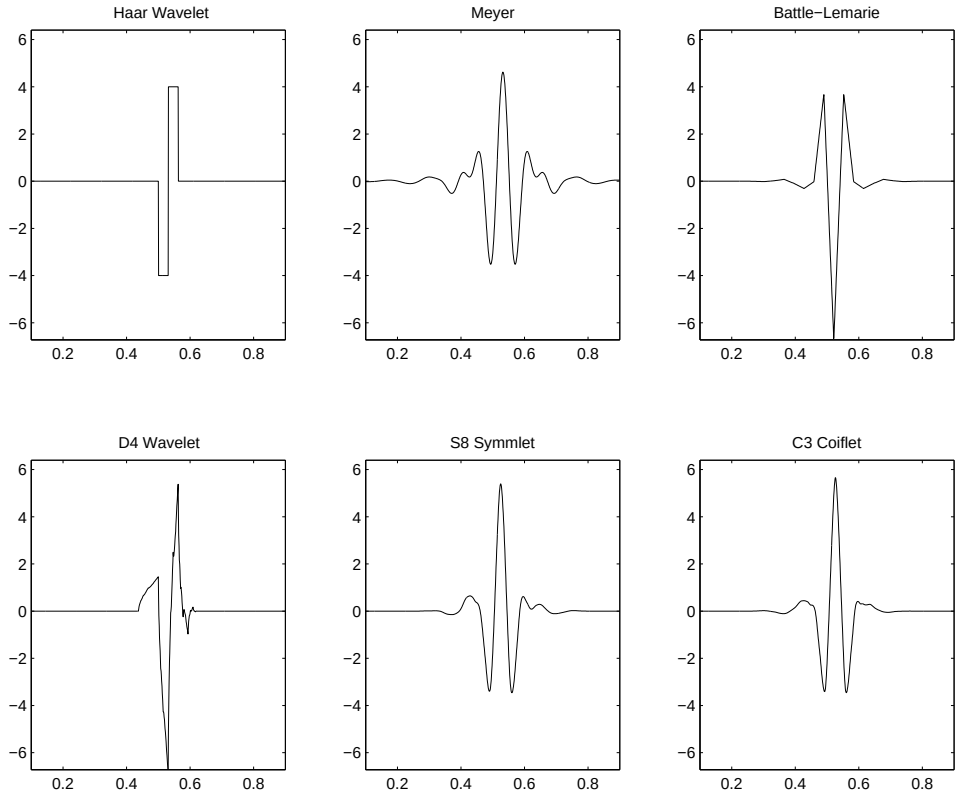

Figure 7.2 The wavelet $\psi_{4,8}(x)$ from the members of several common wavelet families. The Battle-Lemarié case uses linear splines, $m = 1$. For the Daubechies, Symmlet and Coiflet cases, $p = 2, 8$ and 3 respectively, yielding 2, 8 and 6 vanishing moments. Produced using the function `MakeWavelet.m` in `WaveLab`.

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. [The approach of "folding" across boundaries, is dicussed 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 $L_{2,\mathrm{per}}(I)$ by periodization. Suppose that φ and ψ are nice orthonormal scaling and wavelet functions for $L_2(\mathbb{R})$ and define

$$\varphi^{\mathrm{per}}_{j,k}(x) = \sum_{\ell\in\mathbb{Z}} \varphi_{j,k}(x+\ell), \qquad \psi^{\mathrm{per}}_{j,k}(x) = \sum_{\ell\in\mathbb{Z}} \psi_{j,k}(x+\ell).$$

The definition implies that $\varphi^{\mathrm{per}}_{j,k+r2^j}(x) = \varphi^{\mathrm{per}}_{j,k}(x)$ and $\psi^{\mathrm{per}}_{j,k+r2^j}(x) = \psi^{\mathrm{per}}_{j,k}(x)$ for any integers k and r and $j \geq 0$. If φ, ψ have compact support, then for j larger than some j_1, these sums reduce to a single term for each $x \in I$. [Again, this is analogous to the discussion of periodization of kernels at (3.20), (3.25) and the proof thereafter.]

Define $V^{\mathrm{per}}_j = \mathrm{span}\,\{\varphi^{\mathrm{per}}_{jk}, k \in \mathbb{Z}\}$, and $W^{\mathrm{per}}_j = \mathrm{span}\,\{\psi^{\mathrm{per}}_{jk}, k \in \mathbb{Z}\}$: this yields an

orthogonal decomposition

$$L_{2,per}(I) = V_L^{\text{per}} \oplus \bigoplus_{j \geq L} W_j^{\text{per}},$$

with $\dim V_j^{\text{per}} = \dim W_j^{\text{per}} = 2^j$ for $j \geq 0$. Meyer makes 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 $[0, 1]$ For non-periodic functions on $[0, 1]$, one must take a different approach. We summarize results of the "CDJV construction", described in detail in Cohen et al. (1993b), which builds on Meyer (1991) and Cohen et al. (1993a). The construction begins with a Daubechies pair (φ, ψ) having p vanishing moments and minimal support $[-p + 1, p]$. For j such that $2^j \geq 2p$ and for $k = p, \ldots, 2^j - p - 1$, the scaling functions $\varphi_{jk}^{\text{int}} = \varphi_{jk}$ have support contained wholly in $[0, 1]$ and so are left unchanged. At the boundaries, for $k = 0, \ldots, p - 1$, construct orthonormal functions φ_k^L with support $[0, p + k]$ and φ_k^R with support $[-p - k, 0]$, and set

$$\varphi_{jk}^{\text{int}} = 2^{j/2}\varphi_k^L(2^j x), \qquad \varphi_{j,2^j-k-1}^{\text{int}} = 2^{j/2}\varphi_k^R(2^j(x-1)).$$

The $2p$ functions φ_k^L, φ_k^R are finite linear combinations of scaled and translated versions of the original φ and so have the same smoothness as φ. We can now define the multiresolution spaces $V_j^{\text{int}} = \text{span}\{\varphi_{jk}^{\text{int}}, k = 0, \ldots, 2^j - 1\}$. It is shown that $\dim V_j^{\text{int}} = 2^j$, and that they have two key properties:

(i) in order that $V_j^{\text{int}} \subset V_{j+1}^{\text{int}}$, it is required that the boundary scaling functions satisfy two scale equations. For example, on the left side

$$\frac{1}{\sqrt{2}}\varphi_k^L\left(\frac{x}{2}\right) = \sum_{l=0}^{p-1} H_{kl}^L \varphi_l^L(x) + \sum_{m=p}^{p+2k} h_{km}^L \varphi(x-m).$$

(ii) each V_j^{int} contains, on $[0, 1]$, all polynomials of degree at most $p - 1$.

Turning now to the wavelet spaces, W_j^{int} is defined as the orthogonal complement of V_j^{int} in V_{j+1}^{int}. Starting from a Daubechies wavelet ψ with support in $[-p + 1, p]$ and with p vanishing moments, construct orthonormal ψ_k^L with support in $[0, p + k]$ and ψ_k^R with support in $[-p - k, 0]$ and define ψ_{jk}^{int} as for $\varphi_{jk}^{\text{int}}$ replacing $\varphi, \varphi_k^L, \varphi_k^R$ by ψ, ψ_k^L and ψ_k^R. It can be verified that $W_k^{\text{int}} = \text{span}\{\psi_{jk}^{\text{int}}, k = 0, \ldots, 2^{j-1}\}$ and that for each L with $2^L \geq 2p$,

$$L_2([0, 1]) = V_L^{\text{int}} \oplus \bigoplus_{j \geq L} W_j^{\text{int}}, \tag{7.7}$$

and hence $f \in L_2[0, 1]$ has an expansion

$$f(x) = \sum_{k=0}^{2^L-1} \beta_k \varphi_{Lk}^{\text{int}}(x) + \sum_{j \geq L} \sum_{k=0}^{2^j-1} \theta_{jk} \psi_{jk}^{\text{int}}(x),$$

where all indicated functions $\varphi_{Lk}^{\text{int}}$ and ψ_{jk}^{int} are orthonormal, and $\beta_k = \langle f, \varphi_{Lk}^{\text{int}} \rangle$ and $\theta_{jk} = \langle f, \psi_{jk}^{\text{int}} \rangle$. Note especially from property (ii) that since V_L^{int} contains polynomials of degree $\leq p - 1$, it follows that all ψ_{jk}^{int} have vanishing moments of order p.

7.2 The Cascade algorithm for the Discrete Wavelet Transform

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 φ and wavelet ψ 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 (B.6) and (B.11). 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.

Suppose, then, that we have a multiresolution analysis $\{V_j\}$ generated by an orthonormal scaling function φ, and with detail spaces W_j generated by an orthonormal wavelet ψ so that the collection $\{\psi_{jk}, j, k \in \mathbb{Z}\}$ forms an orthonormal basis for $L_2(\mathbb{R})$.

Analysis and Synthesis operators. Consider a function $f \in V_j$. Let $a_j = \{a_j[k]\}$ denote the coefficients of f in the orthobasis $\mathcal{B}_j = \{\varphi_{jk}, k \in \mathbb{Z}\}$, so that

$$a_j[k] = \langle f, \varphi_{jk} \rangle.$$

Since $V_j = V_{j-1} \oplus W_{j-1}$, we can also express f in terms of the basis

$$\mathcal{B}'_j = \{\varphi_{j-1,k}, k \in \mathbb{Z}\} \cup \{\psi_{j-1,k}, k \in \mathbb{Z}\}$$

with coefficients

$$a_{j-1}[k] = \langle f, \varphi_{j-1,k} \rangle, \qquad d_{j-1}[k] = \langle f, \psi_{j-1,k} \rangle, \tag{7.8}$$

and mnemonics "a" for approximation and "d" for detail.

Since $\mathcal{B}$ and $\mathcal{B}'$ are orthonormal bases for the same space, the change of basis maps

$$\begin{aligned} A_j: &\quad a_j \to \{a_{j-1}, d_{j-1}\} && \text{("analysis")} \\ S_j: &\quad \{a_{j-1}, d_{j-1}\} \to a_j && \text{("synthesis")} \end{aligned}$$

must be orthogonal, and transposes of one another:

$$A_j A_j^T = A_j^T A_j = I, \qquad S_j = A_j^{-1} = A_j^T.$$

To derive explicit expressions for A_j and S_j, rewrite the two-scale equations (7.1) and (7.3) in terms of level j, in order to express $\varphi_{j-1,k}$ and $\psi_{j-1,k}$ in terms of φ_{jk}, using the fact that V_{j-1} and W_{j-1} are contained in V_j. Rescale by replacing x by $2^j x - 2k$ and multiply both equations by $2^{j/2}$. Recalling the notation $\varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k)$, we have

$$\varphi_{j-1,k}(x) = \sum_l h[l]\varphi_{j,2k+l}(x) = \sum_l h[l-2k]\varphi_{jl}(x). \tag{7.9}$$

The corresponding relation for the coarse scale wavelet reads

$$\psi_{j-1,k}(x) = \sum_l g[l]\varphi_{j,2k+l}(x) = \sum_l g[l-2k]\varphi_{jl}(x). \tag{7.10}$$

Taking inner products with f as in (7.8) yields the representation of A_j:

$$\begin{aligned} a_{j-1}[k] &= \sum_l h[l-2k]a_j[l] = Rh \star a_j[2k] \\ d_{j-1}[k] &= \sum_l g[l-2k]a_j[l] = Rg \star a_j[2k], \end{aligned} \tag{7.11}$$

where R denotes the *reversal* operator $Ra[k] = a[-k]$, and $\star$ denotes discrete convolution $a \star b[k] = \sum a[k-l]b[l]$. Introducing also the *downsampling* operator $Da[k] = a[2k]$, we could write, for example, $a_{j-1} = D(Rh \star a_j)$. Thus the analysis, or "fine-to-coarse" step $A_j : a_j \to (a_{j-1}, d_{j-1})$ can be described as "filter with Rh and Rg and then downsample".

Synthesis step S_j. Since $\varphi_{j-1,k} \in V_{j-1} \subset V_j$, we can expand $\varphi_{j-1,k}$ as $\sum_l \langle \varphi_{j-1,k}, \varphi_{jl} \rangle \varphi_{jl}$, along with an analogous expansion for $\psi_{j-1,k} \in W_{j-1} \subset V_j$. Comparing the coefficients (7.9) and (7.10) yields the identifications

$$\langle \varphi_{j-1,k}, \varphi_{jl} \rangle = h[l-2k], \qquad \langle \psi_{j-1,k}, \varphi_{jl} \rangle = g[l-2k].$$

Since $\varphi_{jl} \in V_j = V_{j-1} \oplus W_{j-1}$, we may use the previous display to write

$$\varphi_{jl} = \sum_k h[l-2k]\varphi_{j-1,k} + g[l-2k]\psi_{j-1,k}. \tag{7.12}$$

[Note that this time the sums are over k (the level $j-1$ index), not over l as in the analysis step!]. Taking inner products with f in the previous display leads to the synthesis rule

$$a_j[l] = \sum_k h[l-2k]a_{j-1}[k] + g[l-2k]d_{j-1}[k]. \tag{7.13}$$

To write this in simpler form, introduce the *zero-padding* operator $Za[2k] = a[k]$ and $Za[2k+1] = 0$, so that

$$a_j[l] = h \star Za_{j-1}[l] + g \star Zd_{j-1}[l].$$

So the sythesis or *coarse-to-fine* step $S_j : (a_{j-1}, d_{j-1}) \to a_j$ can be described as "zero-pad, then filter with h (and g), and then add".

Computation. If the filters h and g have support length S, the analysis steps (7.11) each require S multiplys and adds to compute each coefficient. The synthesis step (7.13) similarly needs S multiplys and adds per coefficient.

The Cascade algorithm. We may represent the successive application of analysis steps beginning at level J and continuing down to a coarser level L by means of a cascade diagram

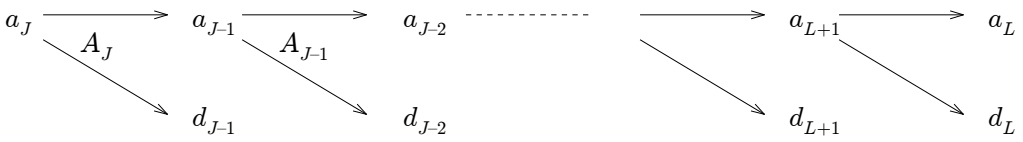

Figure 7.3 The cascade algorithm

Composition of each of these orthogonal transformations produces an orthogonal transformation $W = A_{L+1} \cdots A_{J-1} A_J$:

$$a_J \longleftrightarrow \{d_{J-1}, d_{J-2}, \ldots, d_L, a_L\}. \tag{7.14}$$

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, $W^T = S_J S_{J-1} \cdots S_{L+1}$.

W as a 'matrix'. W represents a change of basis from $V_J = \text{span}\{\varphi_{Jk}, k \in \mathbb{Z}\}$ to

$$V_L \oplus W_L \oplus \cdots \oplus W_{J-1} = \text{span}\{\{\varphi_{Lk}\} \cup \{\psi_{jk}\}, L \le j \le J-1, k \in \mathbb{Z}\}.$$

Define index sets $\mathcal{D} = \{I = (j,k) : L \le j \le J-1;\ k \in \mathbb{Z}\}$ and $\mathcal{A} = \{I = (L,k) : k \in \mathbb{Z}\}$. If we write $W = (W_{Ik})$ for $I \in \mathcal{D} \cup \mathcal{A}$ and $k \in \mathbb{Z}$, then we have

$$W_{Ik} = \begin{cases} \langle \psi_I, \varphi_{Jk} \rangle & I \in \mathcal{D} \\ \langle \varphi_{Lk'}, \varphi_{Jk} \rangle & I = (L,k') \in \mathcal{A}. \end{cases}$$

7.3 Discrete and Continuous Wavelets

Our goal now is to describe more explicitly how the rows W_I of the wavelet transform matrix W are related to the $L_2(\mathbb{R})$ wavelets $\psi_{jk}(x) = 2^{j/2}\psi(2^j - k)$. For simplicity, we ignore boundary effects and remain in the setting of $\ell_2(\mathbb{Z})$.

The discrete filtering operations of the cascade algorithm make no explicit use of the wavelet ψ and scaling function φ. Yet they are derived from the multiresolution analysis generated by (φ, ψ), and it is our goal in this subsection to show more explicitly how the orthonormal rows of the discrete wavelet transform are approximations to the orthobasis functions φ_{jk} and ψ_{jk}.

Approximating φ and ψ from the filter cascade. So far, the cascade algorithm has been described implicitly, by iteration. We now seek a more explicit representation. Let $h^{(r)} = h \star Zh \star \cdots \star Z^{r-1}h$ and $g^{(r)} = h^{(r-1)} \star Z^{r-1}g$ for $r \ge 2$ and $h^{(1)} = h, g^{(1)} = g$.

Lemma 7.5

$$a_{j-r}[k] = \sum_n h^{(r)}[n - 2^r k]\, a_j[n] = Rh^{(r)} \star a_j[2^r k].$$
$$d_{j-r}[k] = \sum_n g^{(r)}[n - 2^r k]\, a_j[n] = Rg^{(r)} \star a_j[2^r k].$$

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. The proof of this and all results in this section is deferred to the end of the chapter.

To describe the approximation of φ and ψ it is helpful to consider the sequence of nested lattices $2^{-r}\mathbb{Z}$ for $r = 1, \ldots$. Define functions $\varphi^{(r)}, \psi^{(r)}$ on $2^{-r}\mathbb{Z}$ using the r-fold iterated filters:

$$\varphi^{(r)}(2^{-r}n) = 2^{r/2}h^{(r)}[n], \qquad \psi^{(r)}(2^{-r}n) = 2^{r/2}g^{(r)}[n]. \tag{7.15}$$

Clearly $\varphi^{(1)}$ and $\psi^{(1)}$ are essentially the original filters h and g, and we will show that $\varphi^{(r)} \to \varphi, \psi^{(r)} \to \psi$ in an appropriate sense. Indeed, interpret the function $\varphi^{(r)}$ on $2^{-r}\mathbb{Z}$ as a (signed) measure $\mu_r = \mu[\varphi^{(r)}]$ that places mass $2^{-r}\varphi^{(r)}(2^{-r}n)$ at $2^{-r}n$. Also interpret the function φ on $\mathbb{R}$ as the density with respect to Lebesgue measure of a signed measure $\mu = \mu[\varphi]$. Then weak convergence of μ_r to μ means that $\int f d\mu_r \to \int f d\mu$ for all bounded continuous functions f.

Proposition 7.6 *The measures $\mu[\varphi^{(r)}]$ and $\mu[\psi^{(r)}]$ converge weakly to $\mu[\varphi]$ and $\mu[\psi]$ respectively as $r \to \infty$.*

The left panel of Figure 7.1 illustrates the convergence for the Daubechies D4 filter.

We now describe the columns of the discrete wavelet transform in terms of these approximate scaling and wavelet functions. To do so, recall the indexing conventions $\mathcal{D}$ and $\mathcal{A}$ used in describing (W_{Ii}). In addition, for $x \in 2^{-(j+r)}\mathbb{Z}$, define

$$\varphi_{jk}^{(r)}(x) = 2^{j/2}\varphi^{(r)}(2^j x - k), \qquad \psi_{jk}^{(r)}(x) = 2^{j/2}\psi^{(r)}(2^j x - k). \tag{7.16}$$

Proposition 7.7 *Suppose that $N = 2^J$. The discrete wavelet transform matrix (W_{Ii}) with $I = (j,k)$ and $i \in \mathbb{Z}$ is given by*

$$W_{Ii} = \begin{cases} \langle \psi_I, \varphi_{Ji} \rangle = N^{-1/2}\psi_{jk}^{(J-j)}(i/N) & I = (jk) \in \mathcal{D}, \\ \langle \varphi_{Lk}, \varphi_{Ji} \rangle = N^{-1/2}\varphi_{Lk}^{(J-L)}(i/N) & I \in \mathcal{A}. \end{cases}$$

Thus, the Ith row of the wavelet transform matrix looks like $\psi_I^{(J-j)}$ (where $I = (j,k)$), and the greater the separation between the detail level j and the original sampling level J, the closer the corresponding function $\psi_{jk}^{(J-j)}$ is to the scaled wavelet $\psi_{jk}(x)$.

Cascade algorithm on sampled data. We have developed the cascade algorithm assuming that the input sequence $a_J[k] = \langle f, \varphi_{Jk} \rangle$. What happens if instead we feed in as inputs $a_J[k]$ a sequence of sampled values $\{f(k/N)\}$?

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

$$\langle f, g \rangle_N = N^{-1} \sum_{k \in \mathbb{Z}} f(N^{-1}k) g(N^{-1}k). \tag{7.17}$$

Proposition 7.8 *If $a_J[k] = N^{-1/2} f(N^{-1}k)$, and $N = 2^J$, then for $j \le J$,*

$$a_j[k] = \langle \varphi_{jk}^{(J-j)}, f \rangle_N, \qquad d_j[k] = \langle \psi_{jk}^{(J-j)}, f \rangle_N, \qquad k \in \mathbb{Z}. \tag{7.18}$$

Thus, when applied on sampled data, the cascade algorithm produces discrete wavelet coefficients which approximate the true wavelet coefficients of the underlying functions in two steps: 1) the integral is approximated by a sum over an equally spaced grid, and 2) the functions φ_{jk} and ψ_{jk} are approximated by $\varphi_{jk}^{(J-j)}$ and $\psi_{jk}^{(J-j)}$.

Formulas (7.18) 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 (7.18) 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 $\langle \varphi_{jk}, f \rangle$.

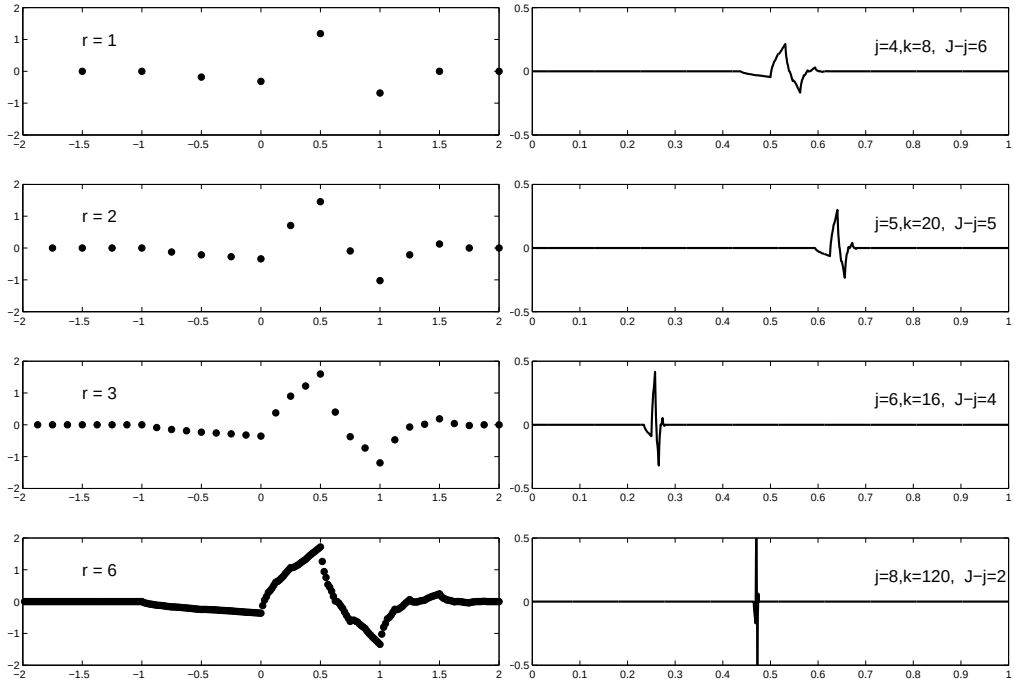

Figure 7.4 Left: The function $\psi^{(r)}$ on $2^{-r}\mathbb{Z}$ for the Daubechies $D4$ filter for various values of r. Right: rows of the wavelet transform matrix, $N = 1024$, for the Daubechies D4 filter, showing scale j, location k and iteration number $J - j$.

Continuous world	Discrete World
$a_J[k] = \langle \varphi_{Jk}, f \rangle$	$a_J[k] = N^{-1/2} f(kN^{-1})$
$\downarrow$	$\downarrow$
$a_j[k] = \langle \varphi_{jk}, f \rangle$	$a_j[k] = \langle \varphi_{jk}^{(J-j)}, f \rangle_N$
$d_j[k] = \langle \psi_{jk}, f \rangle$	$d_j[k] = \langle \psi_{jk}^{(J-j)}, f \rangle_N$

Table 7.1 *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.*

Table 7.1 highlights a curious parallel between the "continuous" and "discrete" worlds: the discrete filtering operations represented by the cascade algorithm, through the DWT matrix W, are the same in both cases!

7.4 Finite data sequences.

So far we have worked with infinite sequences a_j and $d_j \in \ell_2(\mathbb{Z})$. We turn to 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 (e.g. Cohen et al. (1993b)) that preserve orthogonality.

In either case, the detail vectors d_j in (7.14) 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 $2^{J-1} + 2^{J-2} + \ldots + 2^L + 2^L = N$ 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 $\{d_{J-1}, d_{J-2}, \ldots, d_L, a_L\}$. Compare again Figure 7.3. More specifically, we index $w = (w_I)$ with $I = (j, k)$ and

$$\begin{aligned} w_{jk} &= d_{jk} \qquad j = L, \ldots, J-1 \text{ and } k = 1, \ldots, 2^j \\ w_{L-1,k} &= a_{Lk} \qquad k = 1, \ldots, 2^L. \end{aligned}$$

With this notation, we may write $y = W^T w$ in the form

$$y = \sum w_I \boldsymbol{\psi}_I \tag{7.19}$$

with $\boldsymbol{\psi}_I$ denoting the columns of the inverse discrete wavelet transform matrix W^T. [The bolding is used to distinguish the *vector* $\boldsymbol{\psi}_I$ arising in the finite transform from the *function* $\psi_I \in L_2(\mathbb{R})$.] If we set $t_l = l/N$ and adopt the suggestive notation

$$\boldsymbol{\psi}_I(t_l) := \boldsymbol{\psi}_{I,l},$$

then we may write the forward transform $w = Wy$ in the form

$$w_I = \sum_l \boldsymbol{\psi}_I(t_l) y_l. \tag{7.20}$$

7.5 Wavelet shrinkage estimation

Basic model. Observations are taken at equally spaced points $t_l = l/n$, $l = 1, \ldots, n = 2^J$, and are assumed to satisfy

$$Y_l = f(t_l) + \sigma z_l, \qquad z_l \overset{i.i.d}{\sim} N(0, 1). \tag{7.21}$$

It is assumed, for now, that σ is known. The goal is to estimate f, at least at the observation points t_l. The assumption that the observation points are equally spaced is quite important–see the chapter notes for references–whereas the specific form of the error model and knowledge of σ are less crucial.

Basic strategy. The outline is simply described. First, the *tranform* step, which uses a finite orthogonal wavelet transform W as described in the previous section. Second, a *processing*

step in the wavelet domain, and finally an inverse transform, which is accomplished by W^T, since W is orthogonal.

$$\begin{array}{ccc} (n^{-1/2}Y_l) & \xrightarrow{W} & (w_I) \\ & & \downarrow \eta \\ \left(n^{-1/2}\hat{f}(t_l)\right) & \xleftarrow{W^T} & (\hat{w}_I) \end{array} \tag{7.22}$$

Transform step. Being an orthogonal transform, W is non-redundant, and given $n = 2^J$ data values (y_l) 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 in Section 7.2.

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 in Section 7.1 and in Mallat (2009, Chapter 7). Common choices in the `Matlab` library `Wavelab` include (boundary adjusted) versions of $D4$ or the symmlet $S8$.

Processing Step. Generally the estimated coefficients $\hat{w} = \eta(w)$ are found by the recipe

$$\hat{w}_I = \begin{cases} \eta(w_I; t) & I \in \mathcal{D} \\ w_I & I \in \mathcal{A}. \end{cases}$$

The index $I = (j, k)$ belongs to a set corresponding to *details* or *approximations*:

$$\mathcal{D} = \{I : L \le j \le J-1; k = 1, \ldots, 2^j\}, \qquad \mathcal{A} = \{I = (L, k) : k = 1, \ldots, 2^L\}.$$

The transformation $\eta(w_I; t)$ is a scalar function of the observed coefficient w_I, usually non-linear and depending on a parameter t. We say that η operates *co-ordinatewise*. Often, the 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 $I \in \mathcal{I}_j = \{(j, k) : k = 1, \ldots, 2^j\}$. In some cases, the function η itself may depend on the coefficient index I or level j. Common examples include (compare Figure 2.2) *hard thresholding*:

$$\eta_H(w_I; t) = w_I I\{|w_I| \ge t\},$$

and *soft thresholding*:

$$\eta_S(w_I; t) = \begin{cases} w_I - t & w_I > t \\ 0 & |w_I| \le t \\ w_I + t & w_I < -t. \end{cases}$$

These may be regarded as special cases of a more general class of *threshold shrinkage rules*, which are defined by the properties

odd:	$\eta(-x, t) = -\eta(x, t)$,
shrinks:	$\eta(x, t) \le x$ if $x \ge 0$,
bounded:	$x - \eta(x, t) \le t + b$ if $x \ge 0$, (some $b < \infty$),
threshold:	$\eta(x, t) = 0$ iff $\lvert x\rvert \le t$.

Here are some examples. All but the first depend on an additional tuning parameter.

1. $\eta(x,t)(x-t^2/x)_+$ suggested by Gao (1998) based on the "garotte" of Breiman (1995),

2. Soft-hard thresholding (Gao and Bruce, 1997): This is a compromise between soft and hard thresholding defined by

$$\eta(x;t_1,t_2) = \begin{cases} 0 & \text{if } |x| \le t_1 \\ \operatorname{sgn}(x)\frac{t_2(|x|-t_1)}{t_2-t_1} & \text{if } t_1 < |x| \le t_2 \\ x & \text{if } |x| > t_2. \end{cases}$$

3. The smooth clipped absolute deviation (SCAD) penalty threshold function of Fan and Li (2001), and

4. $\eta(x;t,a)$ 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), as discussed in Section 2.4 and below.

Methods for estimating t from data will be discussed in the next section.

Another possibility is to threshold *blocks* of coefficients. One example is James-Stein shrinkage applied to the whole j-th level of coefficients:

$$\eta_{JS}(w_I; s(w_j)) = s(w_j)w_I,$$
$$s(w_j) = (1-(2^j-2)\sigma^2/|w_j|^2)_+.$$

The entire signal is set to zero if the total energy is small enough, $|w_j|^2 < (2^j-2)\sigma^2$, otherwise a common, data-determined *linear* shrinkage applies to all co-ordinates. When the true signal is sparse, this is less effective than thresholding, because the shrinkage factor either causes substantial error in the large components, or fails to shrink the noise elements - it cannot avoid both problems simultaneously. An effective remedy is to use *smaller* blocks of coefficients, as discussed in the next section and Chapters 8 and 9.

The estimator. Writing $\hat{f}$ for the vector $\left(N^{-1/2}\hat{f}(t_l)\right)$ and y for $(N^{-1/2}y_l)$, we may summarize the process as

$$\hat{f} = W^T\eta(Wy).$$

This representation makes the important point that the scaling and wavelet functions φ and ψ are not required or used in the calculation. So long as the filter h is of finite length, and the wavelet coefficient processing $w \to \hat{w}$ is $O(N)$, then so is the whole calculation.

Nevertheless, the iteration that occurs within the cascade algorithm generates approximations to the wavelet, cf. Section 7.3. Thus, we may write the estimator more explicitly as

$$\begin{aligned} \hat{f}(t_l) &= \sum_I \eta_I(w)\boldsymbol{\psi}_I(t_l) \qquad (7.23) \\ &= \sum_{I\in\mathcal{A}} w_I\boldsymbol{\varphi}_I(t_l) + \sum_{I\in\mathcal{D}} \eta(w_I)\boldsymbol{\psi}_I(t_l), \end{aligned}$$

Thus, $\boldsymbol{\psi}_I = N^{-1/2}\psi_{jk}^{(J-j)}$ 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, after being scaled and located to match ψ_{jk}, compare (7.16) and Proposition 7.7.

The (I, l)th entry in the discrete wavelet transform matrix W is given by $N^{-1/2}\psi_{jk}^{(J-j)}(N^{-1}l)$ and in terms of the columns $\boldsymbol{\psi}_I$ of W, we have $y_l = \sum_I w_I \boldsymbol{\psi}_I(N^{-1}l)$.

First examples are given by the NMR data shown in Figure 1.2 and the simulated 'Bumps' example in Figure 7.5. The panels in Figure 1.2 correspond to the vertices of the processing diagram (7.22) (actually transposed!). The simulated example allows a comparison of soft and hard thresholding with the true signal and shows that hard thresholding here preserves the peak heights more accurately.

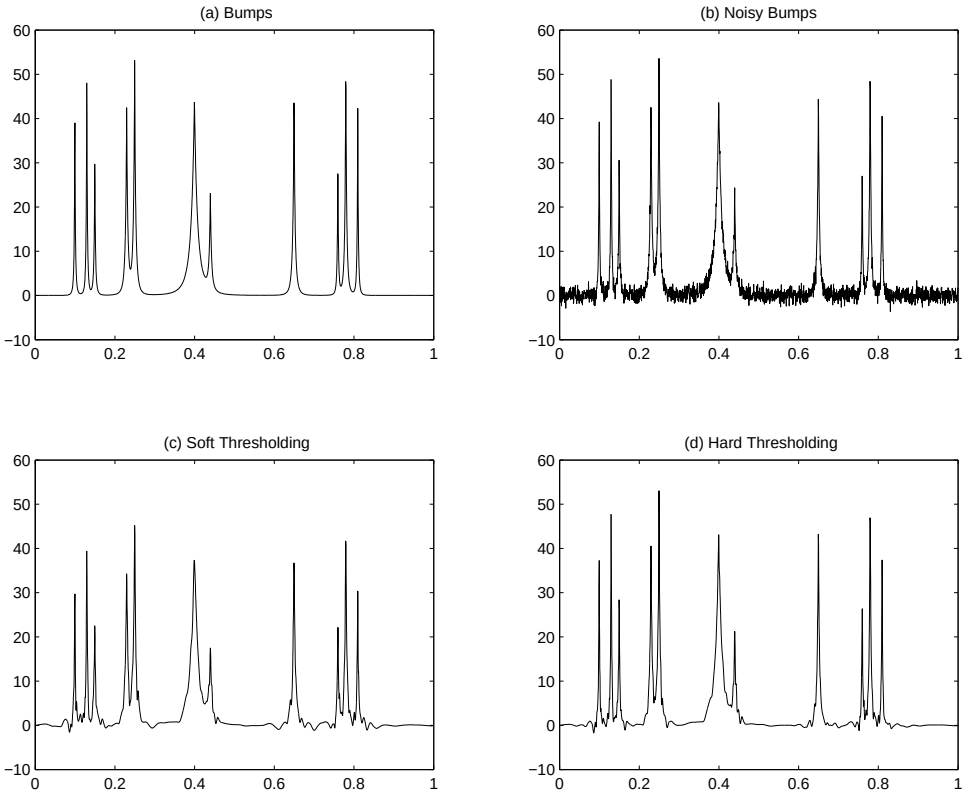

Figure 7.5 Panel (a): artificial 'Bumps' signal constructed to resemble a spectrum, formula in Donoho and Johnstone (1994a), $\|f\|_N = 7$ and $N = 2048$ points. In (b) i.i.d. $N(0, 1)$ noise is added, so the signal to noise ratio is 7. Bottom panels show the result of soft (c) and hard (d) thresholding with threshold $t = \sqrt{2\log n} \approx 3.905$, using a discrete wavelet transform with Symmlet8 filter and coarse scale $L = 5$.

The thresholding estimates have three important properties. They are *simple*, based on co-ordinatewise operations, *non-linear*, and yet *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 Figure 6.2 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 (7.21) is converted by the orthogonal wavelet transform into

$$w_I = \theta_I + \epsilon \tilde{z}_I, \qquad \epsilon = \sigma/\sqrt{n},\ \tilde{z}_I \overset{i.i.d}{\sim} N(0,1). \tag{7.24}$$

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 ϵ^2. On the other hand, in our examples, 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 Chapters 8, 9 and beyond. 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. This too is taken up in Chapter 9.

Estimation of σ. 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^{J-1}) such coefficients, one can estimate σ^2 well using a robust estimator

$$\hat{\sigma}^2 = MAD\{w_{J-1,k}, k \in \mathcal{I}_{J-1}\}/0.6745,$$

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 The choice of the threshold shrinkage rule η and the selection of threshold t are somewhat separate issues. The choice of η is problem dependent. For example, hard thresholding exactly preserves the data values above the threshold, and as such can be good for preserving peak heights (say in spectrum estimation), whereas soft thresholding forces a substantial shrinkage. The latter leads to smoother visual appearance of reconstructions, but this property is often at odds with that of good fidelity – as measured for example by average squared error between estimate and truth.

Correlated data. If the noise z_l in (7.21) is stationary and correlated, then the wavelet transform has a decorrelating effect. (Johnstone and Silverman (1997) has both a heuristic and more formal discussion). In particular, the levelwise variances $\sigma_j^2 = \mathrm{Var}(w_{jk})$ are independent of k. Hence it is natural to apply *level-dependent* thresholding

$$\hat{w}_{jk} = \eta(w_{jk}, t_j).$$

For example, one might take $t_j = \hat{\sigma}_j \sqrt{2\log n}$ with $\hat{\sigma}_j = MAD_k\{w_{jk}\}/0.6745$.

Figure 7.6 shows an ion channel example from Johnstone and Silverman (1997) known to have a stationary correlated noise structure. Two different level dependent choices of thresholds are compared. Consistent with remarks in the next section, and later theoretical results, the $\sqrt{2\log n}$ choice is seen to be too high.

Wavelet shrinkage as a spatially adaptive kernel method. We may write the result of thresholding using (7.19) and (7.25) in the form

$$\hat{f}(t_l) = \sum_I \hat{w}_I \boldsymbol{\psi}_I(t_l) \qquad \hat{w}_I = c_I(y) w_I \tag{7.25}$$

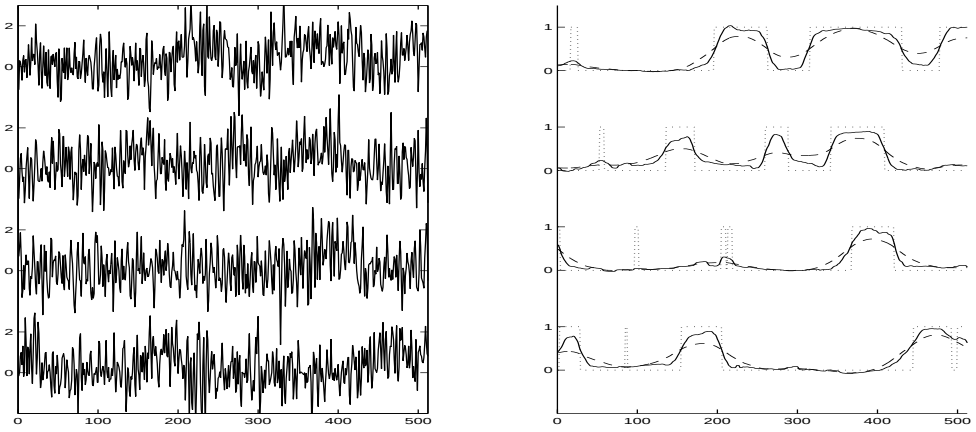

Figure 7.6 Ion channel data. Panel (a) sample trace of length 2048. Panel (b) Dotted line: true signal, Dashed line: reconstruction using translation invariant (TI) thresholding at $\hat{\sigma}_j \sqrt{2 \log n}$. Solid line: reconstruction using TI thresholding at data determined thresholds (a combination of SURE and universal).Further details in Johnstone and Silverman (1997).

where we have here written $\eta_I(w)$ in the data-dependent "linear shrinkage" form $c_I(w)y_I$.

Inserting the wavelet transform representation (7.20) into (7.25) leads to a kernel representation for $\hat{f}(t_l)$:

$$\hat{f}(t_l) = \sum_I \sum_m c_I(y)\boldsymbol{\psi}_I(t_l)\boldsymbol{\psi}_I(t_m)y_m = \sum_m \hat{K}(t_l, t_m)y_m,$$

where the *kernel*

$$\hat{K}(s,t) = \sum_I c_I(y)\boldsymbol{\psi}_I(s)\boldsymbol{\psi}_I(t), \qquad s,t \in \{t_l = l/N\}. \tag{7.26}$$

The hat in this kernel emphasizes that it depends on the data through the coefficients $c_I(y)$. The individual component kernels $K_I(t,s) = \boldsymbol{\psi}_I(t)\boldsymbol{\psi}_I(s)$ 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_l is of order $2^{-j(t_l)}$, where

$$j(t_l) = \max\{j : c_I(y)\boldsymbol{\psi}_I(t_l) \neq 0, \text{ some } I \in \mathcal{I}_j\}.$$

In other words, t_l 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 $2^{-j}k$. Compare Figure 7.7 and Exercise 7.2. By separating the terms in (7.26) 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 $K_A(t_l, t_m) = \sum_{I \in \mathcal{A}} \boldsymbol{\varphi}_I(t_l)\boldsymbol{\varphi}_I(t_m)$ does not depend on the observed data y.

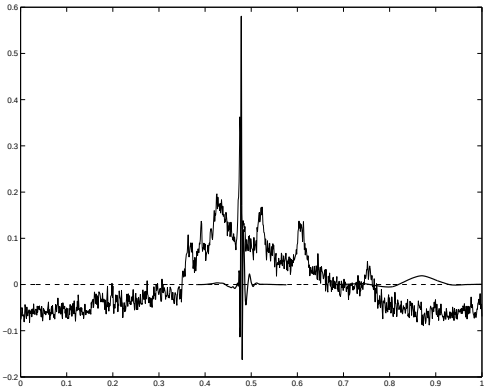

Figure 7.7 Spatially adaptive kernel corresponding to hard thresholding of the NMR signal as in Figure 1.2. The kernel $t_m \to \hat{K}(t_l, t_m)$, compare (7.26), is shown for $t_{l,1} \approx 0.48$ and $t_{l,2} \approx 0.88$. The bandwidth at 0.88 is broader because $j(t_{l,2}) < j(t_{l,1})$.

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.

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

$$\hat{f}^{TI} = \mathrm{Ave}_{1 \le k \le N} \, (S^{-k} \circ \hat{f} \circ S^k). \tag{7.27}$$

The translation invariant (TI) estimator averages over all N shifts, and so would appear to involve at least $O(N^2)$ calculation. However, the proposers of this method, Coifman and Donoho (1995), describe how the algorithm can in fact be implemented in $O(N \log N)$ operations.

It can be seen from Figure 7.8 that the extra averaging implicit in $\hat{f}^{TI}$ reduces artifacts considerably–compare the bottom panels of Figure 7.5. 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.

Software. The wavelet shrinkage figures in this book were produced in `Matlab` using the public domain library `WaveLab` (version 850) available at `stat.stanford.edu`. `Matlab` also has a proprietary wavelet toolbox. In `R`, the `WaveThresh` package is available at `cran.r-project.org` and is described in the book by Nason (2008).

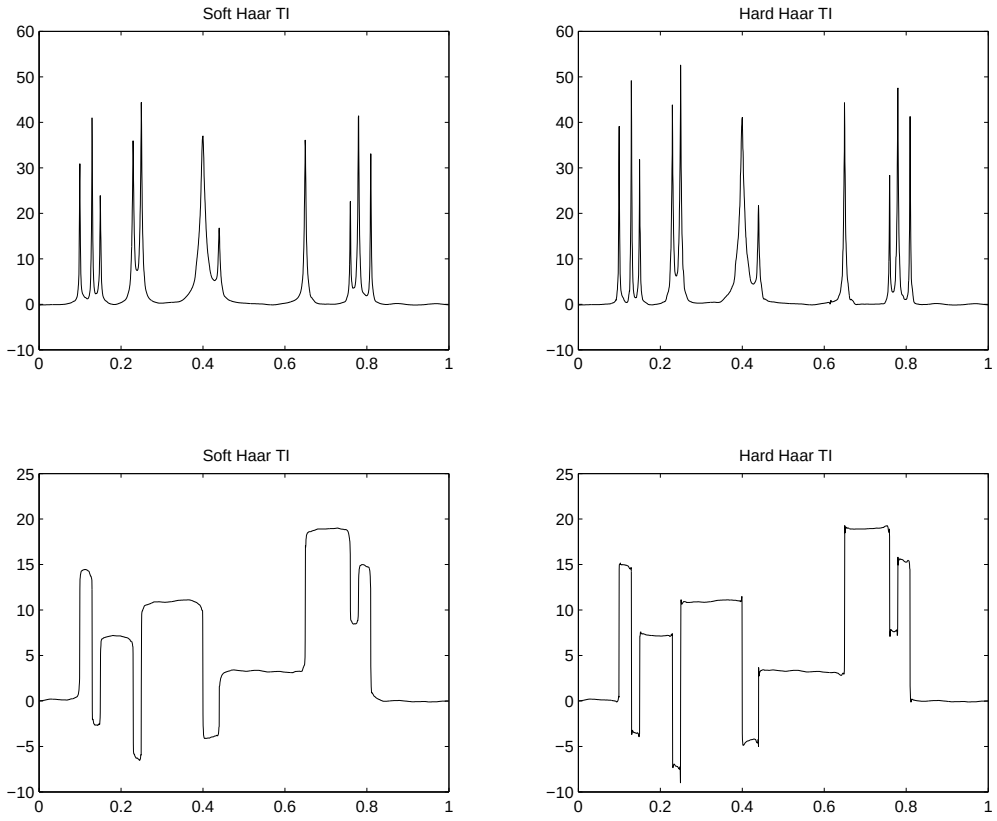

Figure 7.8 A comparison of translation invariant thresholding (7.27) applied to $\hat{f}$ given by soft and hard Haar wavelet thresholding, at $t = \sqrt{2 \log n}$, for $n = 2048$, for the test signals Bumps of Figure 7.5 and Blocks of Figure 6.2. For direct comparisons of thresholding with and without TI-averaging, see Coifman and Donoho (1995).

7.6 Choice of threshold.

We give only a partial discussion of this large topic here, and focus only on 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 threshold function as $|x| \to \infty$ is also significant. More specifically, the growth of $x - \eta(x)$, whether approaching zero or a constant or diverging, influences the bias properties of the estimate, particularly for large signal components.

Often, one may know from previous experience or subjective belief that a particular choice of threshold (say 3σ or 5σ) is appropriate. On the other hand, one may seek an *automatic* method for setting a threshold; this will be the focus of our discussion.

'Automatic' thresholding methods can be broadly divided into *fixed* versus *data-dependent*.

"Fixed" methods set a threshold in advance of observing data. One may use a fixed number of standard deviations $k\sigma$, or a more conservative limit, such as the *universal* threshold $t = \sigma\sqrt{2\log n}$.

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 (compare (8.31)) that for $n \geq 2$,

$$P_n = P\{\max_{1\leq i\leq n} |Z_i| > \sqrt{2\log n}\} \leq \frac{1}{\sqrt{\pi \log n}}.$$

Similarly, it can be shown [1] that the expected number of $|Z_i|$ that exceed the threshold will satisfy the same bound. For a wide range of values of n, including $64 = 2^6 \leq n \leq 2^{20}$, the expected number of exceedances will be between 0.15 and 0.25, so only in at most a quarter of realizations will *any* pure noise variables exceed the threshold.

Since the wavelet transform is orthogonal, it follows from (7.24) that

$$P\{\hat{f}_n \equiv 0 | f \equiv 0\} = P\{w \equiv 0 | \theta \equiv 0\} = 1 - P_n \to 1.$$

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. When combined with the soft thresholding non-linearity, the universal threshold leads to visually smooth reconstructions, but at the cost of considerable bias and relatively high mean squared error. This shows up in the theory as extra logarithmic terms in the rate of convergence of this estimator, e.g. Theorem 10.10.

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 $t_k = \sigma z(q/2 \cdot k/n)$. Let $\hat{k}_F = \max\{k : |y|_{(k)} \geq t_k\}$, and set $\hat{t}_F = t_{\hat{k}_F}$ and use this as the hard threshold

$$\hat{\theta}_k(y) = y_k I\{|y_k| \geq \hat{t}_F\}. \tag{7.28}$$

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.) Figure 7.9 shows the method on two signals of different sparsity levels: the threshold $\hat{t}_F$ chosen is higher for the sparser signal.

As is shown in Abramovich et al. (2006), the FDR estimator has excellent mean squared error performance in sparse multinormal mean situations–for example being asymptotically adaptive minimax over ℓ_p balls. In addition (unpublished), it achieves the "right" rates of

[1] For more detail on these remarks, see the proof of (8.31) and Table 8.1 in the next chapter.

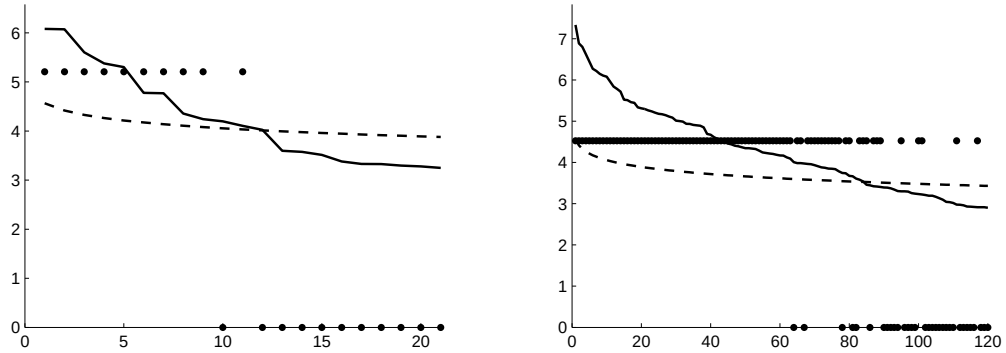

Figure 7.9 Illustration of FDR thresholding at different sparsities (a) 10 out of 10,000. $\mu_i = \mu_0 \doteq 5.21$ for $i = 1, \ldots, n_0 = 10$ and $\mu_i = 0$ if $i = 11, 12, \ldots, n = 10,000$. Data y_i from model (1.3), $\epsilon = 1$. Solid line: ordered data $|y|_{(k)}$. Solid circles: true unobserved mean value μ_i corresponding to observed $|y|_{(k)}$. Dashed line: FDR quantile boundary $t_k = z(q/2 \cdot k/n)$, $q = 0.05$. Last crossing at $\hat{k}_F = 12$ produces threshold $\hat{t}_F = 4.02$. Thus $|y|_{(10)}$ and $|y|_{(12)}$ are false discoveries out of a total of $\hat{k}_F = 12$ discoveries. (b) 100 out of 10,000. $\mu_i = \mu_0 \doteq 4.52$ for $i = 1, \ldots, n_0 = 100$; otherwise zero. Same FDR quantile boundary, $q = 0.05$. Now there are $\hat{k}_F = 84$ discoveries, yielding $\hat{t}_F = 3.54$. (redrawn from Abramovich et al. (2006).)

convergence over Besov function classes - thus removing the logarithmic terms present when the $\sqrt{2 \log n}$ threshold is used. In Chapter 11, we will see that a related estimator arises from penalized least squares model selection, and yields the correct rate of convergence results both in the single sequence model and for wavelet function estimation, Section 12.1.

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 σ_{kk} of the covariance matrix are constant and equal to σ^2. This situation arises, for example, if in the wavelet domain, $k \to y_{jk}$ is a stationary process.

At (2.71) and Exercise 2.8, we derived the unbiased risk criterion for soft thresholding, and found that $E_\theta \|\hat{\theta} - \theta\|^2 = E_\theta \hat{U}(t)$, where (putting in the noise level σ^2)

$$\hat{U}(t) = \sigma^2 n + \sum_k \min(y_k^2, t^2) - 2\sigma^2 \sum_k I\{|y_k| \le t\}.$$

Now set

$$\hat{t}_{SURE} = \underset{0 \le t \le \sigma\sqrt{2 \log n}}{\operatorname{argmin}} \hat{U}(t).$$

The criterion $\hat{U}(t)$ does not depend on details of the correlation (σ_{jk}, $j \ne k$) and so can be used in correlated data settings when the correlation structure is unknown, without the need of estimating it. Also, $\hat{U}(t)$ is piecewise quadratic with jumps at $|y_k|$, so 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. See Donoho and Johnstone (1995); Johnstone (1999); Cai and Zhou (2009b) and Exercise 12.3.

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. As in Section 2.4, allow θ_i to independently be drawn from a mixture prior distribution π:

$$\theta_i \sim (1-w)\delta_0 + w\gamma_a.$$

Here w is the probability that θ_i is non-zero, and $\gamma_a(d\theta)$ is a family of distributions with scale parameter $a > 0$, for example the double exponential, or Laplace, density

$$\gamma_a(d\theta) = (a/2)e^{-a|\theta|}d\theta.$$

Using L_1 loss $\|\hat\theta - \theta\|_1 = \sum_1^n |\hat\theta_i - \theta_i|$, it was shown in Section 2.4 that the Bayes rule for this prior is the *median* $\hat\theta_{EB}(y)$ of the posterior distribution of θ given y :

$$\hat\theta_{EB,i}(y) = \eta(y_i; w, a),$$

and that the posterior *median* η has *threshold* structure:

$$\eta(y; w, a) = 0 \quad \text{if} \quad |y| \le \sigma t(w, a),$$

while for large $|y|$, it turns out, (2.43), that $|y - \eta(y)| \sim \sigma a$.

The hyperparameters (w, a) can be estimated by maximizing the marginal likelihood of (w, a) given data (y_i). Indeed, the marginal of y_i

$$m(y_i|w,a) = \int \phi_\sigma(y_i - \theta_i)\pi(d\theta) = (1-w)\phi_\sigma(y_i) + w\int \phi_\sigma(y_i - \theta_i)\gamma_a(d\theta_i)$$

and the corresponding likelihood $\ell(w, a) = \prod_i m(y_i|w, a)$.

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 (Johnstone and Silverman, 2005b).

A numerical comparison. Table 7.2 is an extract from two larger tables in Johnstone and Silverman (2004a) summarizing results of a simulation comparison of 18 thresholding methods. The observations $x = \mu_0 I_S + z$ are of length 1000 with I_S denoting the indicator function of a set $S \subset \mathcal{I} = \{1, \ldots, 1000\}$, and with noise z_i being i.i.d. standard normal. The non-zero set S is a random subset of $\mathcal{I}$ for each noise realization, and each of three sizes $K = |S| = 5, 50, 500$ corresponding to 'very sparse', 'sparse' and 'dense' signals respectively. Four signal strengths $\mu_0 = 3, 4, 5$ and 7 were used, though only two are shown here. There are thus $3 \times 4 = 12$ configurations. One hundred replications were carried out for each of the values of K and μ_0, with the same 100,000 noise variables used for each set of replications.

Among the 18 estimators, we select here: 'Universal' soft and hard thresholding at level $\sqrt{2\log n} \approx 3.716$, FDR thresholding with $q = 0.1$ and 0.01, SURE thresholding, and

finally empirical Bayes thresholding first with $a = 0.2$ fixed and w estimated, and second with (a, w) estimated, in both cases by marginal maximum likelihood.

For each estimation method $\hat{\theta}_m$ and configuration θ_c, the average total squared error was recorded over the $n_r = 100$ replications:

$$r(\hat{\theta}_m, \theta_c) = n_r^{-1} \sum_{r=1}^{n_r} \|\hat{\theta}_m(\theta_c + z^{(r)}) - \theta_c\|_2^2.$$

Some results are given in Table 7.2 and the following conclusions can be drawn:

- thresholding with the universal threshold particularly with moderate or large amounts of moderate sized signal, can give disastrous results, with soft even worse than hard,
- Estimating the scale parameter a is probably preferable to using a fixed value, though it does lead to slower computations. In general, the automatic choice is quite good at tracking the best fixed choice, especially for sparse and weak signal.
- SURE is a competitor when the signal size is small ($\mu_0 = 3$) but performs poorly when μ_0 is larger, particularly in the sparser cases.
- If q is chosen appropriately, FDR can outperform exponential in some cases, but in the original larger tables, it is seen that the choice of q is crucial and varies from case to case.

Table 7.2 *Left columns: Average of total squared error of estimation of various methods on a mixed signal of length 1000. Right columns: Selected quantiles of ineff* $(\hat{\theta}_m)$.

Number nonzero	5		50		500		ineff $(\hat{\theta}_m)$		
Value nonzero	3	5	3	5	3	5	**med**	**10th**	**max**
$a = 0.2$	38	18	299	95	1061	665	*18*	*30*	*48*
exponential	36	17	214	101	857	783	7	30	52
SURE	38	42	202	210	829	835	35	151	676
FDR q=0.01	43	26	392	125	2568	656	44	91	210
FDR q=0.1	40	19	280	113	1149	651	18	39	139
universal soft	42	73	417	720	4156	7157	529	1282	1367
universal hard	39	18	370	163	3672	1578	50	159	359

An alternative way to compare methods is through their *inefficiency*, which compares the risk of $\hat{\theta}_m$ for a given configuration θ_c with the best over all 18 methods:

$$\text{ineff}(\hat{\theta}_m, \theta_c) = 100 \left[\frac{r(\hat{\theta}_m, \theta_c)}{\min_m r(\hat{\theta}_m, \theta_c)} - 1 \right].$$

The inefficiency vector $\text{ineff}(\hat{\theta}_m)$ for a given method has 12 components (corresponding to the configurations θ_c) and Table 7.2 also records three upper quantiles of this vector: median, and 10th and 12th largest. Minimizing inefficiency has a minimax flavor—it turns out that the empirical Bayes methods have the best inefficiencies in this experiment.

5. Block Thresholding

In Chapter 6 we saw the advantages of blocking in adaptive estimation over ellipsoids. It is natural to ask if the use of (smaller) blocks in conjunction with thresholding could also be advantageous. Suppose that $y \in \mathbb{R}^n$ is partitioned into B blocks each of size L, thus we assume $n = BL$. While other groupings of co-ordinates are possible, for simplicity we take contiguous blocks

$$y_b = (y_{b(L-1)+1}, \ldots, y_{bL}), \qquad b = 1, \ldots, B.$$

Let $S_b^2 = \sum_{k \in b} y_k^2$. We can then define a block thresholding rule via the general prescription

$$\hat{\theta}_b(y) = c(S_b/\epsilon) y_b,$$

where the function $c(\cdot)$ has a thresholding character. Three natural choices are

$$c^H(s) = I\{s > \lambda\}, \qquad c^S(s) = \Big(1 - \frac{\lambda\sqrt{L}}{s}\Big)_+, \qquad c^{JS}(s) = \Big(1 - \frac{\lambda^2 L}{s^2}\Big)_+,$$

corresponding to block hard, soft and James-Stein thresholding respectively. Each of these may be thought of as an extension of a univariate threshold rule to blocks of size L. Thus, for $\epsilon = 1$ write $\hat{\mu}(x)$ for $\hat{\theta}(xe_1)$ and note that the three cases reduce to ordinary hard, soft, and garotte thresholding (Gao (1998), see also Section 8.2) respectively.

Hard thresholding of blocks was studied by Hall et al. (1999a,b), who took $L = (\log n)^\kappa$ for $\kappa > 1$. Block James-Stein thresholding was investigated by Cai (1999) with $L = \log n$ and $\lambda^2 = 4.505$. In Chapter 8 we will study Block soft thresholding, which has monotonicity properties that make it easier to handle, and we will recover analogs of most of Cai's results, including the motivation for choice of L and λ.

One may wish to estimate both block size L and threshold λ from data, level by level, for example by minimizing unbiased estimate of risk. In this way one might obtain larger thresholds and smaller blocks for sparser signals. This is studied at length by Cai and Zhou (2009b), see also Efromovich and Valdez-Jasso (2010).

7.7 Further Details

Proof of Lemma 7.5. We write this out for a_j; there is a parallel argument for d_j. The argument is by induction. The case $r = 1$ is the analysis step (7.11). For general r, (7.11) gives

$$a_{j-r}[k] = Rh \star a_{j-r+1}[2k],$$

and using the induction hypothesis for $r - 1$, we obtain

$$\begin{aligned} a_{j-r}[k] &= \sum_l h[l-2k] \sum_n h^{(r-1)}[n - 2^{r-1}l]\, a_j[n] \\ &= \sum_n a_j[n] \sum_l h^{(r-1)}[n - 2^{r-1}l]\, h[l-2k]. \end{aligned}$$

Now $h[l-2k] = Z^{r-1}h[2^{r-1}l - 2^r k]$ and since $Z^{r-1}h[m] = 0$ unless $m = 2^{r-1}l$, and so the inner sum equals

$$\sum_m h^{(r-1)}[n-m] Z^{r-1} h[m - 2^r k] = h^{(r-1)} \star Z^{r-1}h[n - 2^r k] = h^{(r)}[n - 2^r k]. \qquad \square$$

Proof of Proposition 7.6. *Relating $h^{(r)}$ to φ.* Recall from (B.6) that the scaling function φ was defined by the Fourier domain formula $\widehat{\varphi(\xi)} = \prod_{j=1}^{\infty} \frac{\widehat{h}(2^{-j}\xi)}{\sqrt{2}}$. This suggests that we look at the Fourier transform of $h^{(r)}$. First note that the transform of zero padding is given by

$$\widehat{Zh}(\omega) = \sum_l e^{-il\omega} Zh[l] = \sum_k e^{-i2k\omega} h[k] = \widehat{h}(2\omega),$$

so that $\widehat{h^{(r)}}(\omega) = \prod_{p=0}^{r-1} \widehat{h}(2^p\omega)$. Making the substitution $\omega = 2^{-r}\xi$, we are led to define an r^{th} approximation to φ as a distribution $\varphi^{(r)}$ having Fourier transform

$$\widehat{\varphi^{(r)}}(\xi) = 2^{-r/2}\widehat{h^{(r)}}(2^{-r}\xi) = \prod_{j=1}^{r} \frac{\widehat{h}(2^{-j}\xi)}{\sqrt{2}}. \tag{7.29}$$

Observe that $\widehat{\varphi^{(r)}}(\xi)$ has period $2^{r+1}\pi$. This suggests, and we now verify, that $\varphi^{(r)}$ can be thought of as a function (or more precisely, a measure) defined on $2^{-r}\mathbb{Z}$. Indeed, a discrete measure $\mu = \sum_n m[n]\delta_{2^{-r}n}$ supported on $2^{-r}\mathbb{Z}$ has Fourier transform

$$\widehat{\mu}(\xi) = \int e^{-i\xi x}\mu(dx) = \sum_n m[n]e^{-i\xi 2^{-r}n} = \widehat{m}(2^{-r}\xi).$$

Thus, the quantity $2^{-r/2}\widehat{h^{(r)}}(2^{-r}\xi)$ in (7.29) is the Fourier transform of a measure $\sum_n 2^{-r/2}h^{(r)}[n]\delta_{2^{-r}n}$. Secondly, a real valued function $g(2^{-r}n)$ defined on $2^{-r}\mathbb{Z}$ is naturally associated to the measure $\mu[g] = \sum_n 2^{-r}g(2^{-r}n)\delta_{2^{-r}n}$, (the normalizing multiple 2^{-r} can be motivated by considering integrals of functions against $\mu[g]$). Combining these two remarks shows that $\varphi^{(r)}$ is indeed a function on $2^{-r}\mathbb{Z}$, with

$$2^{-r}\varphi^{(r)}(2^{-r}n) = 2^{-r/2}h^{(r)}[n]. \tag{7.30}$$

Furthermore, the measure $\mu_r = \mu[\varphi^{(r)}]$ has Fourier transform $\widehat{\varphi^{(r)}}(\xi)$. Since $\widehat{\varphi^{(r)}}(\xi) \to \widehat{\varphi}(\xi)$ for all ξ and $\widehat{\varphi}(\xi)$ is continuous at 0, it follows from the Lévy-Cramér theorem C.18, appropriately extended to signed measures, that $\mu[\varphi^{(r)}]$ converges weakly to $\mu[\varphi]$.

The weak convergence for $\mu[\psi^{(r)}]$ to $\mu[\psi]$ follows similarly from the analog of (7.29)

$$\widehat{\psi^{(r)}}(\xi) = 2^{-r/2}\widehat{g^{(r)}}(2^{-r}\xi) = \frac{\widehat{g}(2^{-1}\xi)}{\sqrt{2}} \prod_{j=2}^{r} \frac{\widehat{h}(2^{-j}\xi)}{\sqrt{2}}.$$

Indeed, the product converges to $\widehat{\varphi}(\xi)$, so (7.4) shows that $\widehat{\psi^{(r)}}(\xi) \to \widehat{\psi}(\xi)$. □

PROOF OF PROPOSITION 7.7. We first re-interpret the results of Lemma 7.5. Suppose $j < J$. Since $\varphi_{jk} \in V_J$, we have

$$\varphi_{jk} = \sum_n \langle \varphi_{jk}, \varphi_{Jn}\rangle \varphi_{Jn},$$

(and similarly for $\psi_{jk} \in W_j \subset V_J$.) If $f \in V_J$ and as before we set $a_j[k] = \langle f, \varphi_{jk}\rangle$, and $d_j[k] = \langle f, \psi_{jk}\rangle$, then by taking inner products with f in the previous display,

$$a_j[k] = \sum_n \langle \varphi_{jk}, \varphi_{Jn}\rangle a_J[n].$$

Replacing j with $J - r$ and comparing the results with those of the Lemma, we conclude that

$$\langle \varphi_{J-r,k}, \varphi_{Jn}\rangle = h^{(r)}[n - 2^r k], \qquad \langle \psi_{J-r,k}, \varphi_{Jn}\rangle = g^{(r)}[n - 2^r k].$$

Comparing the first of these with (7.30) and replacing $r = J - j$, we get

$$\langle \varphi_{jk}, \varphi_{Jn}\rangle = 2^{(j-J)/2}\varphi^{(J-j)}(2^{j-J}n - k) = N^{-1/2}\varphi_{jk}^{(J-j)}(n/N),$$

which is the second equation of Proposition 7.7. The first follows similarly.

PROOF OF PROPOSITION 7.8. Let $r = J - j$, so that $a_j = a_{J-r}$ and, using Lemma 7.5, $a_j[k] = \sum_n h^{(r)}[n - 2^r k]\, a_J[n]$. From (7.15),

$$h^{(r)}[n - 2^r k] = 2^{-r/2}\varphi^{(r)}(2^{-r}n - k) = N^{-1/2}\varphi^{(r)}_{jk}(N^{-1}n),$$

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.

7.8 Notes

§1. In addition to the important books by Meyer (1990), Daubechies (1992) and Mallat (2009) already cited, we mention a selection of books on wavelets from various perspectives: Hernández and Weiss (1996), Chui (1997), Wojtaszczyk (1997), Jaffard et al. (2001), Walter and Shen (2001), Cohen (2003), Pinsky (2009), Starck et al. (2010). Heil and Walnut (2006) collects selected important early papers in wavelet theory.

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 orthobases, we jump directly to the definition of multiresolution analysis due to Mallat and Meyer here in a unidimensional form given by Hernández and Weiss (1996):

1. Warning: many authors use the opposite convention $V_{j+1} \subset V_j$!

Conditions (i) - (iv) are not mutually independent -see for example Theorem 2.1.6 in Hernández and Weiss (1996).

Unequally spaced data? `[TC & LW: fill in!]`

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.

Books specifcially focused on wavelets in statistics include Ogden (1997), Härdle et al. (1998), Vidakovic (1999), Percival and Walden (2000), Jansen (2001) and Nason (2008). The emphasis in these books is more on describing methods and software and less on theoretical properties. Härdle et al. (1998) is a more theoretically oriented treatment of wavelets, approximation and statistical estimation, and has considerable overlap in content with the later chapters of this book, though with a broader focus than the sequence model alone.

SURE thresholding is discussed in Donoho and Johnstone (1995), which includes details of the $O(n \log n)$ computational complexity.

Exercises

7.1 *(Plotting approximations to Daubechies wavelets.)* Let $\psi^{(r)}$ be the approximation to a wavelet ψ defined at (7.15). The left panel of Figure 7.4 shows plots of $m \to \psi^{(r)}(2^{-r}m)$ as r varies. As in Proposition 7.7, let $W_{In} = \langle \psi_I, \varphi_{Jn}\rangle$. In particular if $I = (J - r, k)$ show that

$$\psi^{(r)}(2^{-r}m) = 2^{r/2} W_{I,m+2^r k}$$

and hence that it is possible to plot $\psi^{(r)}(2^{-r}m)$ by evaluating $W^T e_I$.

7.2 *(Extracting the adaptive kernel.)* 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 in (7.25) and let $\delta_l \in \mathbb{R}^N$ have zero entries except for a 1 in the l-th place. Show that the adaptive kernel at t_l, namely the vector $\hat{K}_l = \{\hat{K}(t_l, t_m)\}_{m=1}^N$, may be calculated using the wavelet transform via $\hat{K}_l = W^t C W \delta_l$.

8

Thresholding and Oracle inequalities

Less is more. (*Anon.*)

Oracle, n. something regarded as an infallible guide or indicator, esp. when its action is viewed as recondite or mysterious; a thing which provides information, insight, or answers. (Oxford English Dictionary)

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.

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.

Section 8.1 begins with a simple univariate mean squared error bound for hard thresholding. As a first application, this is immediately used to show much faster rates of convergence over ℓ_1 balls in $\mathbb{R}^n$ than are possible with linear estimators.

A more systematic comparison of soft and hard thresholding begins in Section 8.2, with univariate upper and lower bounds for mean squared error that differ only at the level of constants. Soft thresholding is easier to study theoretically, but is not always better in practice. Some attention is also paid to thresholding of *blocks* of coefficients by an extension of soft thresholding.

Turning to data in n dimensions, in Section 8.3 we look at the properties of thresholding at or near $\lambda = \epsilon\sqrt{2\log n}$, a value closely connected with the maximum (absolute) value of n independent normal variates of standard deviation ϵ, here thought of as pure noise. Its mean squared error, for *any* signal θ in white Gaussian noise, is within a logarithmic factor of that achievable by an oracle who knows which co-ordinates exceed the noise level. If

the n co-ordinates are grouped into blocks of size L, there arises a family of such oracle inequalities with thresholds λ varying with L.

Without further information on the nature or size of θ, this logarithmic factor cannot be improved. In the remainder of this chapter, we focus on the consequences of assuming that we do have such information, namely that the signal is sparse.

A simple class of models for a sparse signal says that at most a small number of co-ordinates can be non-zero, k out of n say, though we do not know *which* ones. The minimax risk for estimation of θ in such cases is studied in Sections 8.4–8.8, and is shown, for example, to be asymptotic to $2\epsilon_n^2 k_n \log(n/k_n)$ if the non-zero fraction $k_n/n \to 0$. Thresholding rules are asymptotically minimax in this case, and the upper bound is an easy consequence of earlier results in this chapter.

The lower bound requires more preparation, being given in Section 8.6. It is based on construction of a nearly least favorable sparse prior. We consider in parallel two different models of sparsity. In the first, univariate model, we observe $x \sim N(\mu, 1)$ and give μ a prior π and study Bayes risks $B(\pi)$, for example over models of sparse priors. In the multivariate model, $y \sim N_n(\theta, \epsilon_n I)$, we consider a high dimensional vector θ_n with a large proportion of components $\theta_i = 0$ and seek estimators that minimize the maximum risk $r(\hat{\theta}, \theta)$. Of course, the two models are related, as one method of generating a sparse multivariate mean vector is to draw i.i.d. samples from a sparse prior π in the univariate model.

Sections 8.5 and 8.6 are devoted to sparse versions of the univariate and multivariate models ($k_n/n \to 0$) respectively. The former introduces sparse two point priors, supported mostly on 0 but partly on a single value $\mu > 0$, this value being set up precisely so that observed data near μ will, in the posterior, still be construed as most likely to have come from the atom at 0! The latter section looks at a similar heuristic in the multivariate model, studying independent copies of a "single spike" prior on a collection of k_n blocks, and arriving at a proof of the lower bound half of the $2\epsilon_n^2 k_n \log(n/k_n)$ limiting minimax risk claim. The single spike prior approach can in particular handle the "highly sparse" case in which k_n remains bounded as n grows, whereas the approach based on i.i.d draws from a univariate prior requires the extra assumption that $k_n \to \infty$.

Sections 8.7 and 8.8 consider respectively univariate and multivariate models in which the non-zero fraction can take any positive value not necessarily approaching zero. The univariate results provide a foundation for a comprehensive statement of the limiting minimax risk properties over multivariate models of exact sparsity, Theorem 8.20.

Notation. We continue to write $y \sim N_n(\theta, \epsilon^2 I)$ for the Gaussian model with noise level ϵ, and use a distinguished notation $x \sim N(\mu, 1)$ when focusing on a single observation with noise level one.

8.1 A crude MSE bound for hard thresholding.

Consider a single observation $y \sim N(\theta, \epsilon^2)$. The hard 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 θ is smaller than the noise ϵ 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.

Proposition 8.1 *If $y \sim N(\theta, \epsilon^2)$, there exists a constant M such that if $\lambda \geq 4$*

$$r_H(\lambda, \theta) \leq \begin{cases} M[\theta^2 + \lambda\phi(\lambda - 1)\epsilon^2] & \text{if } |\theta| \leq \epsilon \\ M\lambda^2\epsilon^2 & \text{if } |\theta| > \epsilon. \end{cases} \tag{8.1}$$

[As usual, ϕ denotes the standard normal density function.]

Proof Consider first the small signal case $|\theta| < \epsilon$. Arguing crudely,

$$E_\theta[yI_E - \theta]^2 \leq 2E_\theta y^2 I_E + 2\theta^2.$$

The first term is largest when $|\theta| = \epsilon$. In this case, if we set $x = y/\epsilon \sim N(1, 1)$ then

$$E_\theta y^2 I_E \leq \epsilon^2 \cdot 2\int_\lambda^\infty x^2\phi(x - 1)dx \leq 4\lambda\phi(\lambda - 1)\epsilon^2, \tag{8.2}$$

where we used the fact that for $y \geq 3$, $(y+1)^2\phi(y) \leq 2(y^2-1)\phi(y) = 2(d/dy)[-y\phi(y)]$.

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

$$(yI_E - \theta)^2 \leq 2\epsilon^2(z^2 + \lambda^2).$$

Taking expectations gives the result, for example with $M = 8$. We have however de-emphasized the explicit constants (which will be improved later anyway in Lemma 8.5 and (8.19)) to emphasise the structure of the bound, which is the most important point here. □

Exercise 8.2 shows how the condition $\lambda > 4$ can be removed.

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 standard deviation of order ϵ, except that errors of order $\lambda\epsilon$ are produced about half the time when $\theta = \lambda\epsilon$!

Example 8.2 Let us see how (8.1) yields rough but useful information in an n-dimensional estimation problem. Suppose, as in the introductory example of Section 1.3, that $y \sim N_n(\theta, \epsilon_n^2 I)$ with $\epsilon_n = n^{-1/2}$ and that θ is assumed to be constrained to lie in an ℓ_1-ball $\Theta_{n,1} = \{\theta \in \mathbb{R}^n : \sum |\theta_i| \leq 1\}$. On this set, the minimax risk for linear estimation equals $1/2$ (shown at (9.29) in the next chapter), but thresholding does much better. 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,

$\theta_i^2 \le n^{-1/2}|\theta_i|$, so $\sum_{S_n} \theta_i^2 \le n^{-1/2}$. Now using (8.1)

$$\begin{aligned}\sum r_H(\lambda,\theta_i) &\le M\sum_{B_n}\lambda^2\epsilon^2 + M\sum_{S_n}[\theta_i^2 + \lambda\phi(\lambda-1)\epsilon^2] \\ &\le M\lambda^2 n^{-1/2} + M[n^{-1/2} + \lambda\phi(\lambda-1)].\end{aligned}$$

Choosing, for now, $\lambda = 1 + \sqrt{\log n}$, so that $\phi(\lambda-1) = \phi(0)n^{-1/2}$, we finally arrive at

$$E\|\hat\theta_\lambda - \theta\|^2 \le M' \log n/\sqrt{n}.$$

While this argument does not give exactly the right rate of convergence, which is $(\log n/n)^{1/2}$, let alone the correct constant, compare (13.39) and Theorem 13.17, it already shows clearly that thresholding is much superior to linear estimation on the ℓ_1 ball.

8.2 Properties of Thresholding Estimators

In this section we look in detail at two types of thresholding—soft and hard—in the simplest case: univariate data with known noise level equal to 1. We adopt special notation both for the data, $x \sim N(\mu, 1)$, and for the threshold estimators $\hat\delta(x)$. Both estimators can be described directly, or as the solution of a penalized least squares problem. This was discussed in Section 2.2; for convenience and setting notation, we repeat the results here. Indeed, using $\hat\delta_S$ and $\hat\delta_H$ for soft and hard thresholding respectively, we have

$$\begin{aligned}\hat\delta_S(x,\lambda) &= \begin{cases} x-\lambda & x > \lambda \\ 0 & |x| \le \lambda \\ x+\lambda & x < -\lambda. \end{cases} \\ &= \arg\min_\mu\ (x-\mu)^2 + 2\lambda|\mu|,\end{aligned} \tag{8.3}$$

while

$$\begin{aligned}\hat\delta_H(x,\lambda) &= \begin{cases} x & |x| > \lambda \\ 0 & |x| \le \lambda. \end{cases} \\ &= \arg\min_\mu\ (x-\mu)^2 + \lambda^2 I\{\mu \ne 0\}.\end{aligned} \tag{8.4}$$

Similarities. These two estimators are both *non-linear,* and in particular have in common the notion of a *threshold region* $|x| \le \lambda$, where the signal is estimated to be zero. Of course, hard thresholding is discontinuous, while soft thresholding is constructed to be continuous, which explains the names. Compare Figure 2.2. The threshold parameter in principle can vary over the entire range $(0, \infty)$, so the family includes as limiting cases 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, which 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 $\mu \geq 0$.

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 (e.g. Chapter 10). 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 and Morris (1971), who term it a "limited-translation" rule.

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. Some examples were given in Section 7.5. While these and other proposals can offer useful advantages in practice, we concentrate here on soft and hard thresholding, because of their simplicity and the fact that they well illustrate the main theoretical phenomena.

Soft thresholding.

We begin with soft thresholding as it is somewhat easier to work with mathematically. The explicit risk function $r_S(\lambda, \mu) = E[\hat{\delta}_S(x, \lambda) - \mu]^2$ can be calculated by considering the various zones separately – explicit formulas are given in Section 8.10. Here we focus on qualitative properties and bounds. We first restate for completeness some results already proved in Section 2.7. 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 of soft thresholding is increasing in $\mu > 0$:

$$\frac{\partial}{\partial \mu} r_S(\lambda, \mu) = 2\mu\Phi([I_\lambda - \mu]) \leq 2\mu, \tag{8.5}$$

while

$$r_S(\lambda, \infty) = 1 + \lambda^2, \tag{8.6}$$

which shows the effect of the bias due to the shrinkage by λ, and

$$r_S(\lambda, 0) = 2\int_\lambda^\infty (z-\lambda)^2\phi(z)dz \quad \begin{cases} \leq e^{-\lambda^2/2} & (\text{all } \lambda) \\ \leq 4\lambda^{-1}\phi(\lambda) & (\lambda \geq \sqrt{2}). \\ \sim 4\lambda^{-3}\phi(\lambda) & (\lambda \text{ large}). \end{cases} \tag{8.7}$$

(compare Exercise 8.3). A sharper bound is sometimes useful (also Exercise 8.3)

$$r_S(\lambda, 0) \leq 4\lambda^{-3}(1 + 1.5\lambda^{-2})\phi(\lambda), \tag{8.8}$$

valid for all $\lambda > 0$. 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:

Lemma 8.3 *Let* $\bar{r}_S(\lambda,\mu) = \min\{r_S(\lambda,0)+\mu^2, 1+\lambda^2\}$. *For all* $\lambda > 0$ *and* $\mu \in \mathbb{R}$,

$$\tfrac{1}{2}\bar{r}_S(\lambda,\mu) \le r_S(\lambda,\mu) \le \bar{r}_S(\lambda,\mu). \tag{8.9}$$

The risk bound $\bar{r}_S(\lambda,\mu)$ has the same qualitative flavor as the crude bound (8.1) derived earlier for hard thresholding, only now the constants are correct. In fact, the bound is sharp when μ is close to 0 or ∞.

Figure 8.1 gives a qualitative picture of these bounds. Indeed, the (non-linear) soft thresholding rule can be thought of as the result of 'splining' together three linear estimators, namely $\hat{\delta}_0(x) \equiv 0$ when x is small and $\hat{\delta}_{1,\pm\lambda}(x) = x \mp \lambda$ when $|x|$ is large. Compare Figure 2.2. The risk of $\hat{\delta}_0$ is μ^2, while that of $\hat{\delta}_{1,\pm\lambda}$ is $1+\lambda^2$. Thus our risk bound is essentially the minimum of these two risk functions of linear estimators, with the addition of the $r_S(\lambda,0)$ term. We see therefore that $r_S(\lambda,0)+\mu^2$ is a 'small signal' bound, most useful when $|\mu|$ is small, while $1+\lambda^2$ is useful as a 'large signal' bound.

Proof Symmetry of the risk function means that we may assume without loss that $\mu \ge 0$. By (8.5), the partial derivative $(\partial/\partial\mu) r_S(\lambda,\mu') \le 2\mu'$, and so

$$r_S(\lambda,\mu) - r_S(\lambda,0) = \int_0^\mu (\partial/\partial\mu) r_S(\lambda,\mu') d\mu' \le \int_0^\mu 2\mu'\, d\mu' = \mu^2. \tag{8.10}$$

The upper bound follows from this and (8.6). For the lower bound, observe that if $x > \lambda$, then $\hat{\delta}_S(x,\lambda) - \mu = x - \lambda - \mu$, while if $x \le \lambda$, then $\hat{\delta}_S(x,\lambda) \le 0$ and so $|\hat{\delta} - \mu| \ge \mu$. Writing $x = \mu + z$, we have

$$r_S(\lambda,\mu) \ge E[(z-\lambda)^2 I\{z+\mu > \lambda\}] + \mu^2 P(z+\mu < \lambda). \tag{8.11}$$

If $\mu \le \lambda$, the right side is bounded below by

$$E[(z-\lambda)^2 I\{z > \lambda\}] + \mu^2/2 = (r_S(\lambda,0)+\mu^2)/2,$$

using (8.7). If $\mu \ge \lambda$, then from monotonicity of the risk function, $r_S(\lambda,\mu) \ge r_S(\lambda,\lambda)$, and applying (8.11) at $\mu = \lambda$,

$$r_S(\lambda,\mu) \ge E[(z-\lambda)^2 I\{z > 0\}] + \lambda^2/2 = \lambda^2 - 2\lambda\phi(0) + 1/2 \ge (\lambda^2+1)/2$$

with the last inequality valid if and only if $\lambda \ge \sqrt{8/\pi}$. In this case, the right sides of the last two displays both exceed $\bar{r}(\lambda,\mu)/2$ and we are done. The proof of the lower bound for $\lambda < \sqrt{8/\pi}$ is deferred to Section 8.10. □

Consequences of (8.9) are well suited to showing the relation between sparsity and quality of estimation. As was also shown in Section 2.7, using elementary properties of minima, one may write

$$r_S(\lambda,\mu) \le r_S(\lambda,0) + (1+\lambda^2) \wedge \mu^2. \tag{8.12}$$

In conjunction with the bound $r_S(\lambda,0) \le e^{-\lambda^2/2}$, (8.7), and rescaling as in (2.22), we obtain

Corollary 8.4 *Suppose* $y \sim N(\theta,\epsilon^2)$. *Let* $\delta > 0$ *and* $\lambda_\delta = \sqrt{2\log\delta^{-1}}$. *Then*

$$r_S(\epsilon\lambda_\delta,\theta) \le \delta\epsilon^2 + (1+2\log\delta^{-1})(\theta^2 \wedge \epsilon^2). \tag{8.13}$$

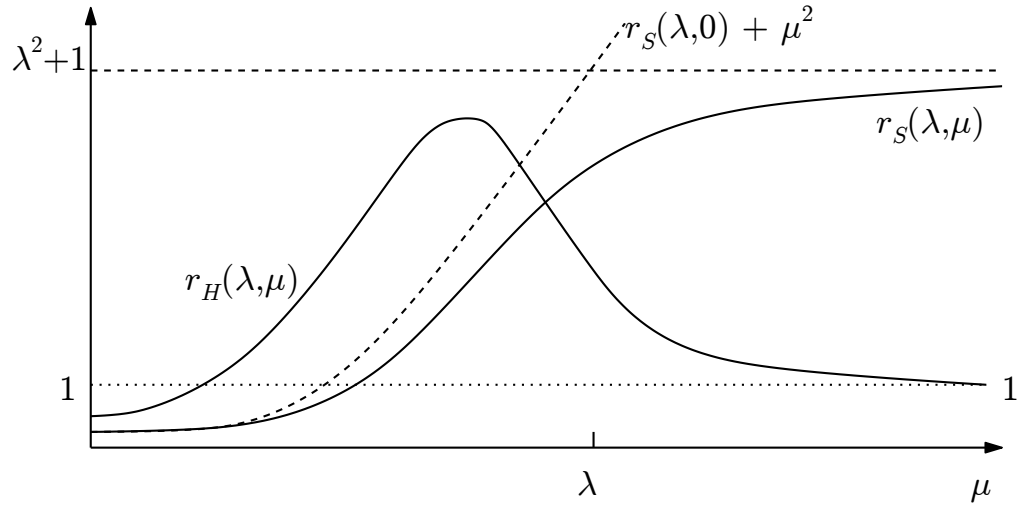

Figure 8.1 Schematic diagram of risk functions of soft and hard thresholding when noise level $\epsilon = 1$ and the threshold λ is moderately large. Dashed lines indicate upper bounds for soft thresholding of Lemma 8.3. Dotted line is risk of the unbiased estimator $\hat{\delta}_1(x) = x$.

Hard thresholding

The risk function is easily written in the form

$$r_H(\lambda,\mu) = \mu^2\Phi(I_\lambda - \mu) + \int_{|z+\mu|>\lambda} z^2\phi(z)dz. \tag{8.14}$$

The extreme values for small and large μ are:

$$\begin{aligned} r_H(\lambda,\infty) &= 1 \\ r_H(\lambda,0) &= 2\int_\lambda^\infty z^2\phi(z)dz = 2\lambda\phi(\lambda) + 2\tilde{\Phi}(\lambda) \sim 2\lambda\phi(\lambda), \end{aligned} \tag{8.15}$$

as $\lambda \to \infty$. Note that the value at ∞ 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 λ. However (8.14) also shows that there is a large risk near $\mu = \lambda$ when λ is large:

$$r_H(\lambda,\lambda) \sim \lambda^2/2.$$

See Exercise 8.6 for more information near $\mu = \lambda$.

Qualitatively, then, as μ increases, the risk of hard thresholding starts out quite small and has quadratic behavior for μ small, then increases to a maximum of order λ^2 near $\mu = \lambda$, and then falls away to a limiting value of 1 as $\mu \to \infty$. See Figure 8.1.

An analogue of the upper bound of Lemma 8.3 is available for hard thresholding. In this case, define

$$\bar{r}_H(\lambda,\mu) = \begin{cases} \min\{r_H(\lambda,0) + 1.2\mu^2, 1+\mu^2\} & 0 \le \mu \le \lambda \\ 1 + \mu^2\tilde{\Phi}(\mu-\lambda) & \mu \ge \lambda, \end{cases} \tag{8.16}$$

and extend $\bar{r}$ to negative μ by making it an even function.

Lemma 8.5 *(a) For $\lambda > 0$ and $\mu \in \mathbb{R}$,*

$$(5/12)\bar{r}_H(\lambda,\mu) \le r_H(\lambda,\mu) \le \bar{r}_H(\lambda,\mu). \tag{8.17}$$

(b) The large μ component of $\bar{r}$ has the bound

$$\sup_{\mu\ge\lambda} \mu^2\tilde{\Phi}(\mu-\lambda) \le \begin{cases} \lambda^2/2 & \text{if } \lambda \ge \sqrt{2\pi}, \\ \lambda^2 & \text{if } \lambda \ge 1. \end{cases}$$

Proof Again we assume without loss that $\mu \ge 0$. The upper bound for $\mu \ge \lambda$ is a direct consequence of (8.14). For $0 \le \mu \le \lambda$, the approach is as used for (8.10), but for the details of the bound $0 \le (\partial/\partial\mu) r_H(\lambda,\mu) \le 2.4\mu$, we refer to Donoho and Johnstone (1994a, Lemma 1). As a result we obtain, for $0 \le \mu \le \lambda$,

$$r_H(\lambda,\mu) \le r_H(\lambda,0) + 1.2\mu^2.$$

The alternate bound, $r_H(\lambda,\mu) \le 1+\mu^2$, is immediate from (8.14).

The lower bound is actually easier—by checking separately the cases $\mu \ge \lambda$ and $\mu \le \lambda$, it is a direct consequence of an inequality analogous to (8.11):

$$E_\mu[\hat{\delta}_H(x,\lambda)-\mu]^2 \ge E[z^2, z+\mu > \lambda] + \mu^2 P(z+\mu < \lambda).$$

For part (b), set $\alpha = \mu - \lambda \ge 0$ and define $g(\alpha) = (\lambda+\alpha)^2\tilde{\Phi}(\alpha)$. We have

$$g'(\alpha) = (\lambda+\alpha)\phi(\alpha)h(\alpha), \qquad h(\alpha) = 2\big(\tilde{\Phi}(\alpha)/\phi(\alpha)\big) - \lambda - \alpha,$$

and $h(0) = \sqrt{2\pi} - \lambda \le 0$ if $\lambda \ge \sqrt{2\pi}$. Differentiation and the bound $\tilde{\Phi}(\alpha) \le \phi(\alpha)/\alpha$ show that h is decreasing and hence negative on $[0,\infty)$, so that $g(\alpha) \le g(0) = \lambda^2/2$. In the case where we only assume that $\lambda \ge 1$, we have $g(\alpha) \le \lambda^2(1+\alpha)^2\tilde{\Phi}(\alpha) \le \lambda^2$, as may be checked numerically, or by calculus. □

For use in later sections, we record some corollaries of the risk bounds. First, for all λ,

$$r_H(\lambda,\mu) \le 1+\lambda^2, \qquad \text{for all } \mu. \tag{8.18}$$

For $\lambda \ge 1$ this follows from (8.16) and Lemma 8.5(b). For $0 \le \lambda \le 1$, see Exercise 8.5. More specifically, for $\lambda \ge 1$,

$$r_H(\lambda,\mu) \le \begin{cases} r_H(\lambda,0) + 1.2\mu^2 & \mu \le 1 \\ 1+\lambda^2 & \mu > 1. \end{cases} \tag{8.19}$$

Second (Exercise 8.4)

$$r_H(\lambda,0) \le \begin{cases} (2\lambda + \sqrt{2\pi})\phi(\lambda) & \text{all } \lambda > 0 \\ 4\lambda\phi(\lambda) & \lambda > 1. \end{cases} \tag{8.20}$$

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

$\hat{\delta}(x)$ is any estimator vanishing for $|x| \le \lambda$, then simply by considering the error made by estimating 0 when $\mu = \lambda$, we find that

$$E_\lambda(\hat{\delta}(x) - \lambda)^2 \ge \lambda^2 P_\lambda\{|x| \le \lambda\} \approx \lambda^2/2 \qquad \text{for large } \lambda. \tag{8.21}$$

Block soft thresholding

In this subsection, we turn to the version of block thresholding, Section 7.6, that is easiest to analyze, namely block soft thresholding. We will see that its properties nicely extend those of univariate soft thresholding.

Here we consider a single block of size d: the blocks will be combined in the next section. And, again for simplicity, we take unit level noise $\epsilon = 1$. Hence, suppose that $X \sim N_d(\mu, I)$ and set $S^2 = \|X\|_2^2$ and define

$$\hat{\mu}(X, \lambda) = \left(1 - \frac{\lambda\sqrt{d}}{S}\right)_+ X.$$

If the observation vector X is large, $S \ge \lambda\sqrt{d}$, then $\hat{\mu}$ pulls it toward the origin by an amount $\lambda\sqrt{d}$: this clearly generalizes soft thresholding in dimension $d = 1$. The corresponding mean squared error

$$r_{S,d}(\lambda, \mu) = E_\mu \|\hat{\mu}(X, \lambda) - \mu\|_2^2$$

is a radial function, i.e. depends only on $\|\mu\|^2$, compare Exercise 2.10.

The results of the next proposition form an almost exact d-dimensional extension of those obtained for univariate soft thresholding in Section 2.7.

Proposition 8.6 *The risk function $r_{S,d}(\lambda, \mu)$ is monotone increasing in $\xi = \|\mu\|^2$, with*

$$r_{S,d}(\lambda, 0) \le 2P(\chi_d^2 \ge \lambda^2 d), \qquad (\textit{for } \lambda \ge 1) \tag{8.22}$$

$$r_{S,d}(\lambda, \infty) = (1 + \lambda^2)d, \qquad (\textit{for } \lambda > 0). \tag{8.23}$$

In terms of the ξ-derivative, there is the bound

$$(\partial/\partial\xi) r_{S,d}(\lambda, \mu) \le 1. \tag{8.24}$$

Let $F(t) = t - 1 - \log t$. For the tail probability (8.22)*, with $\lambda \ge 1$, we have*

$$2P(\chi_d^2 \ge \lambda^2 d) \le \exp\{-F(\lambda^2)d/2\}. \tag{8.25}$$

Proof Block soft thresholding $\hat{\mu}(x, \lambda)$ is weakly differentiable in x, so we can derive an unbiased estimate of risk U using Proposition 2.6. As might be expected from the fact that the risk depends only on $\|\mu\|^2$, we find that U depends only on $\|X\|^2$, which has a noncentral chisquared distribution $S^2 \sim \chi_d^2(\xi)$ with noncentrality $\xi = \|\mu\|^2$. Writing $f_{\xi,d}(w)$ and $F_{\xi,d}(w)$ for the density and distribution functions of $W = S^2$ and setting $\tau = \lambda^2 d$, and $r(\tau, \xi)$ for the block soft thresholding risk $r_{S,d}(\lambda, \mu)$, and we arrive at

$$r(\tau, \xi) = \int_0^\tau U_1(w) f_{\xi,d}(w) dw + \int_\tau^\infty U_2(w) f_{\xi,d}(w) dw, \tag{8.26}$$

where the pieces of the unbiased risk estimate are given by

$$U_1(w) = w - d, \qquad U_2(w) = d + \tau - 2(d-1)(\tau/w)^{1/2}. \tag{8.27}$$

Since $U_2(\infty) = d + \tau = d(1+\lambda^2)$, we easily obtain (8.23), essentially from the monotone convergence theorem as $\xi \to \infty$.

To compute the derivative of $\xi \to r(\tau, \xi)$, we will need some identities for noncentral χ^2 densities from the exercises of Chapter 2. Indeed, using $(\partial/\partial\xi) f_{\xi,d} = -(\partial/\partial w) f_{\xi,d+2}$, (2.94), followed by partial integration and observing that $(U_2 - U_1)(\tau) = 2$, we obtain

$$(\partial/\partial\xi) r(\tau,\xi) = 2 f_{\xi,d+2}(\tau) + \int_0^\tau U_1'(w) f_{\xi,d+2}(w) dw + \int_\tau^\infty U_2'(w) f_{\xi,d+2}(w) dw.$$

Since $U_1'(w) = 1$ and using $F_{\xi,d+2} = F_{\xi,d} - 2 f_{\xi,d+2}$, (2.95), we arrive at

$$(\partial/\partial\xi) r(\tau,\xi) = F_{\xi,d}(\tau) + (d-1)\tau^{1/2} \int_\tau^\infty w^{-3/2} f_{\xi,d+2}(w) dw \geq 0 \tag{8.28}$$

which shows the monotonicity. Notice that when $d = 1$, the second right side term drops out and we recover the derivative formula (8.5) for scalar thresholding. Borrowing from Exercise 2.17 the inequality $f_{\xi,d+2}(w) \leq (w/d) f_{\xi,d}(w)$, we find that the second term in the previous display is bounded above by

$$\frac{d-1}{d} \int_\tau^\infty \left(\frac{\tau}{w}\right)^{1/2} f_{\xi,d}(w) dw \leq 1 - F_{\xi,d}(\tau),$$

which completes the verification that $(\partial/\partial\xi) r(\tau,\xi) \leq 1$.

For the risk at zero, rewrite (8.26) as

$$r(\tau,\xi) = \xi + \int_\tau^\infty (U_2 - U_1)(w) f_{\xi,d}(w) dw,$$

and note that for $w \geq \tau$ we have $(U_2 - U_1)' = -1 + (d-1)\tau^{1/2} w^{-3/2} \leq 0$, so long as $\lambda \geq 1$ (and $d \geq 1$). Consequently $r(\tau, 0) \leq 2[1 - F_d(\tau)]$ as was claimed.

The inequality (8.25) is part of (2.91) in Exercise 2.15. □

It is now easy to establish upper and lower bounds for the risk of block soft thresholding that have the flavor of Lemma 8.3 in the univariate case.

Proposition 8.7 *The mean squared error of block soft thresholding satisfies*

$$r_{S,d}(\lambda,\mu) \leq r_{S,d}(\lambda,0) + \min\{\|\mu\|^2, (1+\lambda^2)d\},$$

and, for $\lambda \geq \sqrt{2}$,

$$r_{S,d}(\lambda,\mu) \geq r_{S,d}(\lambda,0) + \tfrac{1}{4}\min\{\|\mu\|^2, \lambda^2 d/2\}.$$

Proof The upper bound is immediate from (8.24) and (8.23). [Of course, a stronger bound $\min\{r_{S,d}(\lambda,0) + \|\mu\|^2, (1+\lambda^2)d\}$ holds, but the form we use later is given.] For the lower bound, again put $\tau = \lambda^2 d$ and $r(\tau,\xi) = r_{S,d}(\lambda,\mu)$ and use representation (8.28) to write

$$r(\tau,\xi) - r(\tau,0) \geq \int_0^\xi F_{\xi',d}(\tau) d\xi'.$$

Suppose that $\xi \leq \tau/2$. Exercise 8.8 shows that for $\tau \geq 2d$, we have $F_{\xi',d}(\tau) \geq 1/4$, and so

the display is bounded below by $\xi/4 = \|\mu\|^2/4$. When $\xi \geq \tau/2$, simply use monotonicity of the risk function and the bound just proved to get $r(\tau, \xi) \geq r(\tau, \tau/2) \geq r(\tau, 0) + \tau/8$. □

8.3 Thresholding in $\mathbb{R}^n$ and Oracle Inequalities

Let us turn now to the vector setting in which we observe n co-ordinates, $y_i = \theta_i + \epsilon z_i$, with as usual, z_i being i.i.d. $N(0, 1)$. A leading example results from the discrete equispaced regression model (7.21) after applying a discrete orthogonal wavelet transform, compare (7.24).

Consider an estimator built from soft (or hard) thresholding applied co-ordinatewise, at threshold $\lambda_n = \epsilon\sqrt{2\log n}$:

$$\hat{\theta}^S_{\lambda_n,i} = \hat{\delta}_S(y_i, \epsilon\sqrt{2\log n}), \tag{8.29}$$

and let $\hat{\theta}^H_{\lambda_n}$ denote hard thresholding at the same level.

Remark. Here is one reason for the specific choice $\lambda_n = \epsilon\sqrt{2\log n}$ (other choices will be discussed later.) We show that this threshold level is conservative, in the sense that

$$P\{\hat{\theta} = 0 | \theta = 0\} \to 1 \tag{8.30}$$

as $n \to \infty$, so that with high probability, $\hat{\theta}$ does not assert the presence of "spurious structure". Indeed, note that if each y_i is distributed independently as $N(0, \epsilon^2)$, then the chance that at least one observation exceeds threshold λ_n equals the extreme value probability

$$\varpi_n = P\{\max_{i=1,\ldots,n} |Z_i| \geq \sqrt{2\log n}\} = 1 - \left[1 - 2\tilde{\Phi}\left(\sqrt{2\log n}\right)\right]^n \leq \frac{1}{\sqrt{\pi\log n}}, \tag{8.31}$$

valid for $n \geq 2$ (see 3° in Section 8.10).

Table 8.1 compares the exact value ϖ_n of the extreme value probability with the upper bound ϖ'_n given in (8.31). Also shown is the expectation of the number N_n of values Z_i that exceed the $\sqrt{2\log n}$ threshold. It is clear that the exceedance probability converges to zero rather slowly, but also from the expected values that the *number* of exceedances is at most one with much higher probability, greater than about 97%, even for n large. Compare also Exercise 8.10. And looking at the ratios ϖ'_n/ϖ_n, one sees that while the bound ϖ'_n is not fully sharp, it does indicate the (slow) rate of approach of the exceedence probability to zero.

The classical extreme value theory result Galambos (1978, p. 69) for the maximum of n i.i.d. $N(0, 1)$ variables Z_i, namely $M_n = \max_{i=1,\ldots,n} Z_i$ states that

$$b_n^{-1}[M_n - a_n] \xrightarrow{\mathcal{D}} W, \qquad P(W \leq t) = \exp\{-e^{-t}\}, \tag{8.32}$$

where $a_n = \sqrt{2\log n} - (\log\log n + \log 4\pi)/(2\sqrt{2\log n})$ and $b_n = 1/\sqrt{2\log n}$. Section 8.9 has some more information on the law of M_n.

Here we are actually more interested in $\max_{i=1,\ldots,n} |Z_i|$, but this is described quite well by M_{2n}. (Exercise 8.12 explains why). Thus the exceedance probability ϖ_n might be approximated by $\varpi_n^W = P(W \leq c_{2n})$ where $c_{2n} = (\sqrt{2\log n} - a_{2n})/b_{2n})$. Although the convergence to the extreme value distribution in (8.32) is slow, of order $1/\log n$ (e.g. Hall

n	$\sqrt{2\log n}$	ϖ_n	ϖ_n^W	$\mathbb{E}N_n$	ϖ_n'
32	2.63	0.238	0.248	0.271	0.303
64	2.88	0.223	0.231	0.251	0.277
128	3.12	0.210	0.217	0.235	0.256
256	3.33	0.199	0.206	0.222	0.240
512	3.53	0.190	0.196	0.211	0.226
1024	3.72	0.182	0.188	0.201	0.214
2048	3.91	0.175	0.180	0.193	0.204
4096	4.08	0.169	0.174	0.186	0.196

Table 8.1 *For i.i.d. Gaussian noise: sample size n, threshold* $\sqrt{2\log n}$*, exceedance probability* ϖ_n*, extreme value theory approximation* ϖ_n^W *expected number of exceedances* $\mathbb{E}N_n$*, upper bound* ϖ_n' *of* (8.31)

(1979), Galambos (1978, p. 140)). Table 8.1 shows the extreme value approximation to be better than the direct bound (8.31).

A simple non-asymptotic bound follows from the Tsirelson-Sudakov-Ibragimov bound Proposition 2.11 for a Lipschitz(1) function $f : \mathbb{R}^n \to \mathbb{R}$ of a standard Gaussian $n-$vector $Z \sim N_n(0, I)$:

$$P\{|f(Z) - Ef(Z)| \geq t\} \leq 2e^{-t^2/2}.$$

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 suggests more: that the limiting distribution has standard deviation $O(1/\sqrt{\log n})$ about a_n.

Ideal Risk. Suppose that $y_i = \theta_i + \epsilon z_i, \quad i = 1, \ldots n$, with, as usual z_i being i.i.d. $N(0, 1)$. Given a fixed value of θ, an *ideal* linear estimator $\theta_{c,i}^* = c_i^* y_i$ would achieve the best possible mean squared error among linear estimators for the given θ:

$$\min_{c_i} r(\theta_{c,i}^*, \theta) = \frac{\theta_i^2 \epsilon^2}{\theta_i^2 + \epsilon^2} \in [\tfrac{1}{2}, 1] \cdot \theta_i^2 \wedge \epsilon^2.$$

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

$$\theta_i^*(y) = \begin{cases} y_i & \text{if } \theta_i^2 \geq \epsilon^2 \\ 0 & \text{if } \theta_i^2 < \epsilon^2. \end{cases}$$

One can imagine an "oracle", who has partial, but valuable, information about the unknown θ: for example, which co-ordinates are worth estimating and which can be safely ignored. Thus, with the aid of a "projection oracle", the best mean squared error attainable by a projection estimator is the *ideal risk*:

$$\mathcal{R}(\theta, \epsilon^2) = \sum_i \min(\theta_i^2, \epsilon^2), \tag{8.33}$$

In Chapter 9 we will discuss further the significance of the ideal risk, and especially its interpretation in terms of sparsity.

Of course, the statistician does not normally have access to such oracles, but we now show that it is nevertheless possible to mimic the ideal risk with threshold estimators, at least up to a precise logarithmic factor.

Proposition 8.8 *Suppose that* $y \sim N_n(\theta, \epsilon^2)$. *For the soft thresholding estimator* (8.29) *at threshold* $\lambda_n = \epsilon\sqrt{2\log n}$,

$$E\|\hat{\theta}^S_{\lambda_n} - \theta\|_2^2 \le (2\log n + 1)\Big[\epsilon^2 + \sum_1^n \min(\theta_i^2, \epsilon^2)\Big]. \tag{8.34}$$

A similar result holds for $\hat{\theta}^H_{\lambda_n}$, *with the multiplier* $(2\log n + 1)$ *replaced by* $(2\log n + 1.2)$.

The factor $2\log n$ *is optimal without further restrictions on* θ, *as* $n \to \infty$,

$$\inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}^n} \frac{E\|\hat{\theta} - \theta\|^2}{\epsilon^2 + \sum_1^n \min(\theta_i^2, \epsilon^2)} \ge (2\log n)(1 + o(1)). \tag{8.35}$$

Results of this type help to render the idea of ideal risk statistically meaningful: a genuine estimator, depending only on available data, and not upon access to an oracle, can achieve the ideal risk $\mathcal{R}(\theta, \epsilon)$ up to the (usually trivial) additive factor ϵ^2 and the multiplicative factor $2\log n + 1$. In turn, the lower bound (8.9) shows that the ideal risk is also a lower bound to the mean squared error of thresholding, so that

$$\tfrac{1}{2}\mathcal{R}(\theta, \epsilon) \le E\|\hat{\theta}^S_{\lambda_n} - \theta\|_2^2 \le (2\log n + 1)[\epsilon^2 + \mathcal{R}(\theta, \epsilon)].$$

This logarithmic penalty can certainly be improved if we add extra constraints on θ: for example that θ belong to some ℓ_p ball, weak or strong (Chapter 13). However, lower bound (8.35) shows that the $2\log n$ factor is optimal for unrestricted θ, at least asymptotically.

Note that the upper bounds are non-asymptotic, holding for all $\theta \in \mathbb{R}^n$ and $n \ge 1$.

The upper bound extends trivially to correlated, heteroscedastic data, since thresholding depends only on the univariate marginal distributions of the data. The only change is to replace ϵ^2 by ϵ_i^2, the variance of the ith coordinate, in the ideal risk, and to modify the additive factor to ave $_{1\le i\le n}\epsilon_i^2$. There is also a version of the lower bound under some conditions on the correlation structure: for details see Johnstone and Silverman (1997).

Proof *Upper bound.* For soft thresholding, a slightly stronger result was already established as Lemma 2.10. For hard thresholding, we first set $\epsilon = 1$ and use (8.19) to establish the bound, for $\lambda_n = \sqrt{2\log n}$

$$r_H(\lambda, \mu) \le (2\log n + 1.2)(n^{-1} + \mu^2 \wedge 1).$$

This is clear for $\mu > 1$, while for $\mu < 1$, one verifies that $r_H(\lambda, 0) = 2\lambda\phi(\lambda) + 2\tilde{\Phi}(\lambda) \le (2\log n + 1.2)n^{-1}$ for $n \ge 2$. Finally, add over co-ordinates and rescale to noise level ϵ.

Lower bound. The proof is deferred till Section 8.6, since it uses the sparse two point priors to be discussed in the next section. □

Remark. Alan Miller's variable selection scheme. A method of Miller (1984, 1990) offers a nice perspective on $\sqrt{2\log n}$ thresholding. Consider a traditional linear regression model

$$y = X\beta + \sigma^2 z,$$

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 $\tilde{\beta}^t = [\beta^t\ \beta^{*t}]$ and fit the enlarged regression model $y = \tilde{X}\tilde{\beta}$ by a forward stepwise method. Let the method stop just before the first 'fake' variable x_i^* enters the model. The new variables x_i^* are approximately orthonormal among themselves and approximately orthogonal to each x_i (see Exercise 8.15), so the estimated coefficients $\hat{\beta}_i^*$ are nearly i.i.d. $N(0, 1)$, and so the stopping criterion amounts to "enter variables above the threshold given by $\max_{i=1,\ldots,n} |\hat{\beta}_i^*| \doteq \sqrt{2 \log n}$".

Smaller thresholds. It is possible to obtain a bound of the form (8.34) for all $\theta \in \mathbb{R}^n$

$$E\|\hat{\theta}^S_{\lambda_n^*} - \theta\|_2^2 \le \Lambda_n^* \Big[\epsilon^2 + \sum_1^n \min(\theta_i^2, \epsilon^2)\Big]. \tag{8.36}$$

valid for thresholds λ_n^* and bounds Λ_n^* notably smaller than $\epsilon\sqrt{2 \log n}$ and $2 \log n + 1$ respectively. The details for (8.36) are in Section 8.10; it is shown there that for $n \ge 4$, $\lambda_n^* \in (0, \infty)$ is the unique solution of

$$(n + 1) r_S(\lambda, 0) = (1 + \lambda^2), \tag{8.37}$$

and that we may take $\Lambda_n^* = 1 + \lambda_n^{*2}$. This univariate equation is easily solved numerically and Table 8.2 shows results for some practically relevant dyadic powers. These values hold for finite values of n in the typical ranges of practical interest and so do not contradict the asymptotic result (8.35). However, as the focus is now on optimizing a bound for MSE, the conservative property (8.30) is lost.

A serviceable empirical approximation to λ_n^* for all $25 \le n \le 25,000$ is given by

$$\tilde{\lambda}_n = -1.3 + \sqrt{1.9 \log n}, \tag{8.38}$$

yielding a bound that is within 10% of Λ_n^* and also between one third and one half of $2 \log n + 1$, see Exercise 8.7.

Table 8.2 *Minimax MSE threshold λ_n^* and bound Λ_n^* in* (8.36) *compared to 'universal' threshold $\sqrt{2 \log n}$ and bound $2 \log n + 1$ in* (8.34).

n	λ_n^*	$\sqrt{2 \log n}$	Λ_n^*	$2 \log n + 1$
32	1.276	2.633	2.549	7.931
64	1.474	2.884	3.124	9.318
128	1.669	3.115	3.755	10.704
256	1.859	3.330	4.439	12.090
512	2.045	3.532	5.172	13.477
1024	2.226	3.723	5.950	14.863
2048	2.403	3.905	6.770	16.249
4096	2.575	4.079	7.629	17.636
8192	2.743	4.245	8.522	19.022
16384	2.906	4.405	9.446	20.408

Block thresholding

We indicate the extension of Proposition 8.8 to block (soft) thresholding. Suppose that $\theta \in \mathbb{R}^n$ is partitioned into B blocks each of size L, thus we assume $n = BL$. While other groupings of co-ordinates are possible, for simplicity we take contiguous blocks

$$\theta_b = (\theta_{b(L-1)+1}, \ldots, \theta_{bL}), \qquad b = 1, \ldots, B.$$

Let y be partitioned similarly; we sometimes abuse notation and write $y_b = (y_k, k \in b)$. As in Chapter 6.2, we might consider block diagonal estimators $\hat{\theta}_{c,b} = (c_b y_b)$. For simplicity, we focus on *projections*, with $c_b = 0$ or 1. The mean squared error of $\hat{\theta}_{c,b}$ is then either entirely bias, $\|\theta_b\|^2$ when $c_b = 0$, or entirely variance, $L\epsilon^2$, when $c_b = 1$. The *ideal* projection chooses the minimum of the two and is given by

$$\theta_b^*(y) = \begin{cases} y_b & \text{if } \|\theta_b\|^2 \geq L\epsilon^2 \\ 0 & \text{if } \|\theta_b\|^2 < L\epsilon^2. \end{cases}$$

Of course, this "projection oracle" requires knowledge of the block norms $\|\theta_b\|^2$, and it achieves the *block ideal risk*

$$\mathcal{R}(\theta, \epsilon; L) = \sum_b \min(\|\theta_b\|^2, L\epsilon^2).$$

Block soft thresholding can mimic the projection oracle. Let $S_b^2 = \sum_{k \in b} y_k^2$ and define $\hat{\theta}_\lambda^B = (\hat{\theta}_{\lambda,b})$ by

$$\hat{\theta}_{\lambda,b}^B(y) = \eta_{S,L}(y_b, \lambda\epsilon) = \left(1 - \frac{\lambda\epsilon\sqrt{L}}{S_b}\right)_+ y_b, \qquad b = 1, \ldots, B. \tag{8.39}$$

With these definitions, and after rescaling to noise level ϵ, we can rewrite the conclusion of Proposition 8.7 as follows.

Proposition 8.9 *Suppose that $y \sim N_n(\theta, \epsilon^2 I)$ and that $n = BL$. The block soft thresholding estimator $\hat{\theta}_\lambda^B$,* (8.39)*, satisfies*

$$E\|\hat{\theta}_\lambda^B - \lambda\|_2^2 \leq B\epsilon^2 r_{S,L}(\lambda, 0) + \mathcal{R}(\theta, \bar{\lambda}\epsilon; L),$$

where $r_{S,L}(\lambda, 0)$ is bounded at (8.22) *and $\bar{\lambda}^2 = 1 + \lambda^2$. If λ, L are chosen such that $r_{S,L}(\lambda, 0) \leq n^{-1}$, then*

$$E\|\hat{\theta}_\lambda^B - \lambda\|_2^2 \leq \epsilon^2 + \sum_{b=1}^B \min(\|\theta_b\|^2, L\bar{\lambda}^2\epsilon^2).$$

We turn to choice of block size and threshold. The factor $L\bar{\lambda}^2 = L(1 + \lambda^2)$ in the risk bound should in principle be as small as possible consistent with a small value for $r_{S,L}(\lambda, 0)$, say $r_{S,L}(\lambda, 0) \leq O(B^{-1})$. For simplicity, we strengthen this slightly to $r_{S,L}(\lambda, 0) \leq n^{-1}$. From (8.25) with $\lambda \geq 1$, we have $r_{S,L}(\lambda, 0) \leq \exp\{-F(\lambda^2)L/2\} \leq n^{-1}$ so long as $F(\lambda^2) \geq (2\log n)/L$. We restrict $F(\cdot)$ to $[1, \infty)$, on which it is monotone increasing, and solve

$$F(\lambda^2) = \lambda^2 - 1 - \log\lambda^2 = (2\log n)/L.$$

In several cases of interest, we obtain

$$\begin{aligned} L &= \log n & \lambda_L &= \sqrt{4.50524} \\ L &= 1 & \lambda_L &\sim \sqrt{2\log n} \\ L &= (\log n)^{(1+\delta)} & \lambda_L &\sim 1. \end{aligned} \tag{8.40}$$

As a function of block size L, the factor $L(1+\lambda^2)$ may be written as $L(1+F^{-1}((2\log n)/L))$ and since $F^{-1}(x) \geq \max(1,x)$, we find that $L(1+\lambda^2) \geq L+\max(L,\log n)$. From this perspective, then, there is little advantage to choosing block sizes of order larger than $L = \log n$.

8.4 Models for sparsity and upper bounds

In the remainder of this chapter, we will work out some consequences of an explicit quantification of sparsity in terms of the ℓ_0 norm. In this section, we introduce this (and related) sparsity models and suggest the form of the results to be established. As usual, we suppose a Gaussian white noise model

$$y_i = \theta_i + \epsilon_n z_i, \qquad i = 1,\dots,n \tag{8.41}$$

Models for sparsity. A natural measure of the sparsity of a vector $\theta \in \mathbb{R}^n$ is obtained by simply counting the number of nonzero components,

$$\|\theta\|_0 = \#\{\, i : \theta_i \neq 0 \,\}. \tag{8.42}$$

The subscript 0 acknowledges that this measure is sometimes called the ℓ_0-norm[1]. The set of k-sparse vectors in $\mathbb{R}^n$ will be denoted by

$$\Theta_{n,0}[k] = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq k\}, \tag{8.43}$$

though we often just abbreviate this as $\Theta_n[k]$. If $k \ll n$ and the components of θ represent pixel intensities then $\Theta_n[k]$, perhaps with an additional constraint that $\theta_i \geq 0$, models the collection of "nearly black" images (Donoho et al., 1992).

The set (8.43) is sometimes called a model for *exact sparsity* since even small departures from zero are forbidden in more than k components. One might alternatively consider a weaker notion, that of *approximate sparsity*, which requires only that *most* $|\theta_i|$ are small in an appropriate sense. For example, one might stipulate that the ordered absolute values $|\theta|_{(1)} \geq |\theta|_{(2)} \geq \dots$ decrease at some rate: for given $C > 0$ and $p > 0$, that

$$|\theta|_{(k)} \leq C k^{-1/p},$$

known as a *weak–ℓ_p* condition. Alternatively, one might adopt a *strong* ℓ_p assumption

$$\sum_1^n |\theta_i|^p \leq C^p.$$

Results for these notions of approximate sparsity will be given in Chapters 9, 11 and 13.

[1] –somewhat inaccurately as it is not homogeneous.

In this chapter we concentrate on exact sparsity, an important case which is also technically easiest to work with. Exact sparsity may also be viewed as a limiting case of approximate sparsity as $p \to 0$, in the sense that

$$\|\theta\|_p^p = \sum_{i=1}^n |\theta_i|^p \to \#\{i : \theta_i \neq 0\} = \|\theta\|_0.$$

Minimax risk over $\Theta_n[k]$. When θ is restricted to a parameter set Θ, the minimax risk for mean squared error is defined, as in earlier chapters, by

$$R_N(\Theta, \epsilon) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} E_\theta \|\hat{\theta} - \theta\|_2^2.$$

We are interested in the advantage that accrues—as measured by reduction in minimax risk—if we may make a sparsity assumption that restricts θ to $\Theta_n[k]$. It is relatively easy to derive an upper bound for the approximate minimax risk over $\Theta_n[k]$ using the results on the mean squared error $r(\lambda, \mu)$ of soft thresholding obtained in Section 8.2 in the univariate, unit noise setting, so we begin with this.

Assume model (8.41) and $\epsilon_n = 1$, by rescaling if necessary. On parameter space $\Theta_n[k]$ at most k coordinates are non-zero, and so

$$\sum_{i=1}^n r(\lambda, \mu_i) \leq (n-k) r(\lambda, 0) + k \sup_\mu r(\lambda, \mu).$$

Now come the bounds for soft thresholding obtained in Section 8.2. The risk at zero is bounded by (8.7), where we use, for example, the middle expression for $\lambda \geq \sqrt{2}$. In addition, for *all* values of μ, the risk is bounded by $1 + \lambda^2$, compare (8.6). Thus, the previous display is bounded by

$$4n\lambda^{-1}\phi(\lambda) + k(\lambda^2 + 1).$$

Set the threshold at $\lambda_n = \sqrt{2\log(n/k)}$ and observe that $\phi(\lambda_n) = \phi(0)e^{-\lambda_n^2/2} = \phi(0)k/n$. Consequently,

$$\sum_{i=1}^n r(\lambda_n, \mu_i) \leq \frac{2k}{\sqrt{\log(n/k)}} + k[2\log(n/k) + 1].$$

Asymptotic model. Our object is to study the asymptotic behavior of the minimax risk R_N as the number of parameters n increases. We regard the noise level $\epsilon = \epsilon_n$ and number of non-zero components $k = k_n$ as known functions of n. This framework accomodates a common feature of statistical practice: as the amount of data increases—here thought of as a decreasing noise level ϵ_n per parameter—so too does the number of parameters that one may contemplate estimating. To simplify the theory, however, we mostly set $\epsilon = 1$. Since it is a scale parameter, it is easy to put it back into the statement of results, for example as in Theorems 8.10 and 8.20 below.

Consider first the case in which $k_n/n \to 0$ (the situation when $k_n/n \to \eta > 0$ is deferred to Sections 8.7 and 8.8). Here the contribution from the risk at 0 in the previous display is of smaller order than that from the risk bound for the k_n nonzero components, and we arrive at

the upper bound

$$\sup_{\theta \in \Theta_n[k]} r(\hat{\theta}_{\lambda_n}, \theta) \le 2k_n \log(n/k_n)(1 + o(1)). \tag{8.44}$$

The leading term is proportional to the number of non-zero components k_n, while the multiplier $2\log(n/k_n)$ can be interpreted as the per-component cost of not knowing the locations of these non-zero components.

This upper bound for minimax risk over $\Theta_n[k_n]$ turns out to be asymptotically optimal. The result is formulated in the following theorem, to be finally proved in Section 8.6.

Theorem 8.10 *Assume model* (8.41) *and parameter space* (8.43). *If* $k_n/n \to 0$ *as* $n \to \infty$, *then*

$$R_N(\Theta_n[k_n], \epsilon_n) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta_n[k_n]} r(\hat{\theta}, \theta) \sim 2k_n \epsilon_n^2 \log(n/k_n). \tag{8.45}$$

Of course, this is much smaller than the minimax risk for the unconstrained parameter space $R_N(\mathbb{R}^n) \equiv n\epsilon_n^2$, which as noted at (2.51), is attained by the MLE $\hat{\theta}(y) = y$. The assumption of sparsity $\Theta_n[k_n]$ entails a huge reduction in minimax risk, from $n\epsilon_n^2$ to $2k_n\epsilon_n^2 \log(n/k_n)$, and this reduction can, for example, be achieved using soft thresholding at $\lambda_n = \epsilon_n \sqrt{2\log(n/k_n)}$.

To establish Theorem 8.10, we need lower bounds on the minimax risk. These will be obtained by computing the Bayes risks of suitable nearly least favorable priors π, compare the discussion in Chapter 4.3.

We consider two approaches to constructing these priors, to be outlined here and developed in detail over the next three sections. In the first, which we now call the 'multivariate problem', we work with model (8.41) and the n-dimensional mean squared error $r(\hat{\theta}, \theta) = \sum_1^n E_\theta[\hat{\theta}_i(y) - \theta_i]^2$. In the second, the 'univariate Bayes problem', we consider a scalar observation $y_1 = \theta_1 + \epsilon z_1$, but in addition suppose that θ_1 is random with distribution π_1, and that an estimator $\delta(y_1)$ is evaluated through its integrated MSE:

$$B(\delta, \pi) = E_\pi E_\theta[\delta(y_1) - \theta_1]^2 = E_\pi r(\delta, \theta_1).$$

An obvious connection between the two approaches runs as follows: suppose that an estimator $\hat{\theta}(y)$ in the multivariate problem is built by co-ordinatewise application of a univariate estimator δ, so that $\hat{\theta}_i(y) = \delta(y_i)$, and that to a vector $\theta = (\theta_i)$ we associate a univariate (discrete) distribution $\pi_n^e = n^{-1} \sum_{i=1}^n \nu_{\theta_i}$, where ν_θ denotes a unit point mass at θ. Then the multivariate and univariate Bayes mean squared errors are related by

$$r(\hat{\theta}, \theta) = \sum_1^n r(\delta, \theta_i) = n B(\delta, \pi_n^e).$$

The sparsity condition $\Theta_n[k]$ in the multivariate problem, cf. (8.43), corresponds to requiring that the prior $\pi = \pi_n^e$ in the univariate problem satisfy

$$\pi\{\theta_1 \neq 0\} \le k/n.$$

We will see that the univariate problem is easier to analyze, and that sometimes, but not always, the multivariate minimax risk may be evaluated via the univariate Bayes approach.

Candidates for least favorable priors. In the univariate Bayes problem, we build a prior π^{IID} by n iid draws from a univariate mixture: for η small, such as near k/n, suppose that

$$\theta_i \overset{\text{iid}}{\sim} (1-\eta)\nu_0 + \eta\nu. \tag{8.46}$$

Here ν_0 represents point mass at 0 and ν an arbitrary probability distribution on $\mathbb{R}$. Such mixture priors are considered systematically later, Section 8.7. At first it will be enough to set $\nu = \nu_\mu$, a point mass at a relatively large value μ, as will be discussed in the next section.

In (8.46), the number of non-zero components θ_i follows a binomial distribution, so that $N_n = \#\{\theta_i \neq 0\} \sim \mathrm{Bin}(n, \eta)$; and

$$\pi^{\mathrm{IID}}(\Theta_n[k_n]) = \Pr(N_n \leq k_n).$$

Thus, the simplicity of the 'univariate' prior π^{IID} is offset by the fact that it is not supported on $\Theta_n[k]$. When n is large and $k_n \to \infty$, the prior is nearly supported, or 'concentrated' on $\Theta_n[k_n]$: this is the basis of the minimax Bayes method sketched in Section 4.11.

In the multivariate problem, by contrast, we will build a prior supported on $\Theta_n[k]$ with probability one. We start with some definitions.

Given $\tau > 0$, the *single spike prior* $\pi_S(\tau; m)$ chooses an index $I \in \{1, \ldots, m\}$ at random and then sets $\theta = \tau e_I$. [2] If ν_μ denotes unit point mass at μ, write

$$\pi_S(\tau; m) = m^{-1} \sum_{i=1}^{m} \nu_{\tau e_i}. \tag{8.47}$$

The *independent blocks* prior π^{IB} on $\Theta_n[k]$ is built as follows. Fix k and divide $\{1, \ldots, n\}$ into k contiguous blocks B_j, $j = 1, \ldots, k$, each of length $m = [n/k]$. Set $\tau = \sqrt{2\log(n/k)}$ and draw components θ_i in each block B_j according to an independent copy of $\pi_S(\tau; m)$. Finally, set $\theta_i = 0$ for the remaining $n - km$ components.

Informally, π^{IB} picks a single spike of height τ in each of k blocks, with the location of the spike within each block being independent and uniformly distributed. The value of τ will be chosen to be close to, but smaller than, $\lambda_n = \sqrt{2\log n}$, compare Section 8.6.

The two priors π^{IID} and π^{IB} are clearly related. If we set $\eta = k/n$, then under π^{IID}, $EN_n = n\eta = k$ and if $k = k_n \to \infty$, then N_n *concentrates*, and in particular we have $P\{N_n \leq k_n(1+\epsilon)\} \to 0$. When $k_n \to \infty$, it therefore turns out that both π^{IID} and π^{IB} can be shown to be asymptotically least favorable and hence both yield (8.45).

If however $k_n \equiv k$ remains fixed as n grows, we are in what might be called a *highly sparse* situation. [As examples for motivation, one might think of terrorists on airline passenger lists or locations of stars in a night sky image.] Here π^{IID} does not concentrate, but we will see that π^{IB} may still be used to establish (8.45).

Thus, the advantage of π^{IB} compared with π^{IID} is that it is supported on $\Theta_n[k]$. The price we pay is that the co-ordinates within a block are dependent, and this leads to extra effort in evaluating the behavior of the posterior distribution. This extra effort is rewarded with some greater generality: the highly sparse case is covered as well.

We begin, in the next section, with a discussion of sparse two point univariate priors of the form (8.46). The succeeding section takes up the single spike prior (8.47) and the independent blocks prior constructed from it.

[2] It might seem more natural to allow $\theta = \pm\tau e_I$, but this leads to slightly messier formulas, e.g. in (8.59).

8.5 Sparse univariate two point priors

We first study the curious properties of the two point prior

$$\pi_{\alpha,\mu} = (1-\alpha)\nu_0 + \alpha\nu_\mu, \qquad \mu > 0, \tag{8.48}$$

which we call a *sparse* prior in the case when α is small.

The univariate prior results of this section are used later, in Sections 8.7, 8.8, in the study of the $k_n/n \to \eta > 0$ limit, and they also provide a 'warm-up' for the multivariate spike priors. However they are not strictly needed for the proof of the $k_n/n \to 0$ case, (8.45), and so may be skipped by the reader focusing only on this case.

The joint distribution $x|\mu' \sim N(\mu', 1)$ and $\mu' \sim \pi_{\alpha,\mu}$ is a simple idealized model for studying estimation in sparse situations. For example, one might imagine that a source transmits a signal μ_i that on the ith transmission is independently either zero with high probability $1-\alpha$ or of strength μ with low probability α. The signal is received as $x_i = \mu_i + z_i$ with i.i.d. standard Gaussian noise added, and the task is to decide/estimate what was sent.

The posterior distribution is also concentrated on $\{0, \mu\}$, and

$$P(\{\mu\}|x) = \frac{\alpha\phi(x-\mu)}{\alpha\phi(x-\mu) + (1-\alpha)\phi(x)} = \frac{1}{1+m(x)},$$

where the posterior probability ratio

$$m(x) = \frac{P(\{0\}|x)}{P(\{\mu\}|x)} = \frac{(1-\alpha)}{\alpha}\frac{\phi(x)}{\phi(x-\mu)} = \exp(\tfrac{1}{2}\lambda^2 - x\mu + \tfrac{1}{2}\mu^2) \tag{8.49}$$

is decreasing in x: $m(x)/m(y) = e^{-\mu(x-y)}$. We have set $\lambda = \lambda_\alpha = (2\log(1-\alpha)/\alpha)^{1/2}$.

The *posterior indifference point* is that value of x at which the posterior is indifferent between 0 and μ, so that $P(\{0\}|x) = P(\{\mu\}|x)$, or $m(x) = 1$. We focus on the apparently peculiar situation in which this indifference point lies to the *right* of μ. Inserting $x = \mu + a$ into (8.49), we are led to the

Definition. The two point prior $\pi_{\alpha,\mu}$ has *sparsity* $\alpha > 0$ and *overshoot* $a > 0$ if μ satisfies

$$\mu^2 + 2a\mu = \lambda_\alpha^2. \tag{8.50}$$

The prior probability on 0 is so large that even if x is larger than μ, but smaller than $\mu + a$, the posterior distribution places more weight on 0 than μ. [3] See Figure 8.2.

The Bayes rule for squared error loss is as usual the posterior mean, which becomes

$$\delta_\pi(x) = \mu P(\{\mu\}|x) = \frac{\mu}{1+m(x)}. \tag{8.51}$$

Substituting (8.50) into (8.49), we obtain $m(x) = \exp\{-\mu(x-\mu-a)\}$ and

$$\delta_\pi(\mu + z) = \frac{\mu}{1+e^{-\mu(z-a)}}. \tag{8.52}$$

In particular, observe that $\delta_\pi(\mu)$ is small, and even $\delta_\pi(\mu + a) = \mu/2$ is far from μ.

[3] Fire alarms are rare, but one may not believe that a ringing alarm signifies an actual fire without further evidence.

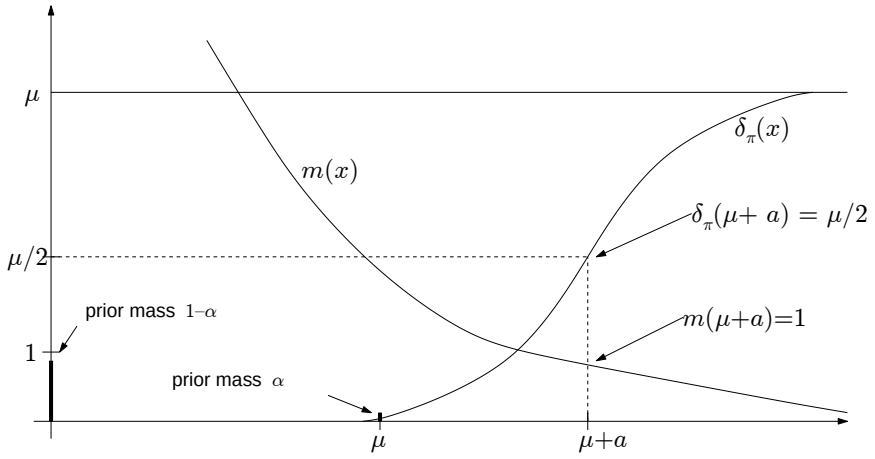

Figure 8.2 Two point priors with sparsity α and overshoot a: posterior probability ratio $m(x)$ and posterior mean $\delta_\pi(x)$

Now consider asymptotics as $\alpha \to 0$ and choose $a_\alpha \to \infty$ with $a_\alpha = o(\lambda_\alpha)$. The solution μ_α of (8.50) satisfies

$$\mu_\alpha = (\lambda_\alpha^2 + a_\alpha^2)^{1/2} - a_\alpha \tag{8.53}$$

$$\sim \lambda_\alpha \sim \sqrt{2\log\alpha^{-1}}. \tag{8.54}$$

In this case, there is a simple and important asymptotic approximation to the Bayes risk of a sparse two point prior $\pi_\alpha = \pi_{\alpha,\mu_\alpha}$.

Lemma 8.11 *Let π_α have sparsity α and overshoot $a_\alpha = (2\log\alpha^{-1})^\gamma$, for $0 < \gamma < 1/2$. Then, as $\alpha \to 0$,*

$$B(\pi_\alpha) \sim \alpha\mu_\alpha^2.$$

Proof By definition, we have

$$B(\pi_\alpha) = (1-\alpha)r(\delta_\pi, 0) + \alpha r(\delta_\pi, \mu_\alpha). \tag{8.55}$$

Thus, a convenient feature of two point priors is that to study the Bayes risk, the frequentist risk function of δ_π only needs to be evaluated at two points. We give the heuristics first. When μ_α is large and the overshoot a_α is also large (though of smaller order), then (8.52) shows that for $x \sim N(\mu_\alpha, 1)$, the Bayes rule δ_π essentially estimates 0 with high probability, thus making an error of about μ_α^2. *A fortiori*, if $x \sim N(0,1)$, then δ_π estimates 0 (correctly) with even higher probability. More concretely, we will show that, as $\alpha \to 0$,

$$r(\delta_\pi, \mu_\alpha) \sim \mu_\alpha^2, \qquad r(\delta_\pi, 0) = o(\alpha\mu_\alpha^2). \tag{8.56}$$

Inserting these relations into the Bayes risk formula (8.55) yields the result. The primary contribution comes from the risk at $\mu = \mu_\alpha$, and the large error μ_α^2 that is made there.

The first relation is relatively easy to obtain. Using (8.52), we may write

$$r(\delta_\pi, \mu_\alpha) = \mu_\alpha^2 \int_{-\infty}^{\infty} \frac{\phi(z)dz}{[1+e^{\mu(z-a)}]^2} \sim \mu_\alpha^2, \tag{8.57}$$

as $\alpha \to 0$, since the integral converges to 1 as both μ_α and $a_\alpha \to \infty$ by the dominated convergence theorem. The second relation takes a little extra work, see Section 8.10. In fact, it is not needed if the goal is simply to establish a lower bound for $B(\pi_\alpha)$. □

8.6 Sparse multivariate block priors

This section turns attention to multivariate sparse priors and their Bayes risks, and in particular priors which pick a single spike in each of several disjoint blocks, with the location of the spike within each block being independent and uniform. The main work lies in establishing a good lower bound for the single spike case; insights from the univariate two point priors just studied provide useful guidance.

Single spike prior

We begin, then, with the simplest possible sparse setting: signals in $\mathbb{R}^n$ with at most one non-zero co-ordinate. We suppose that the index of this nonzero co-ordinate is unknown and evaluate the cost of that ignorance. In this section, we concentrate on the unit noise model, $\epsilon = 1$, thus $y \overset{\mathcal{D}}{\sim} N_n(\theta, I)$.

To develop lower bounds—our primary interest here—let e_I denote the unit vector with 1 in the Ith slot and 0 elsewhere and define a 'bounded single spike' parameter set by

$$\Theta_n(\tau) = \{\theta \in \mathbb{R}^n : \theta = \gamma e_I \text{ for some } I \in \{1, \dots, n\}\ ,\ |\gamma| \le \tau\}. \tag{8.58}$$

Thus, $\Theta_n(\tau)$ is the union of n orthogonal needles, each corresponding to a 1−dimensional bounded interval $[-\tau, \tau]$. We will often let τ depend on n, $\tau = \tau_n$.

Recall the single spike prior $\pi_S(\tau; n)$ which chooses $I \in \{1, \dots, n\}$ at random and sets $\theta = \tau e_I$, compare (8.47). The posterior distribution of I is [4]

$$p_{in}(y) = P(I = i|y) = \frac{\phi(y - \tau e_i)}{\sum_j \phi(y - \tau e_j)} = \frac{e^{\tau y_i}}{\sum_j e^{\tau y_j}}. \tag{8.59}$$

The posterior mean of θ has components given, for example, by

$$\hat{\theta}_{\pi,1} = E(\theta_1|y) = \tau P(I = 1|y) = \tau p_{1n}(y). \tag{8.60}$$

Here and below, in subscripts we abbreviate $\pi_S(\tau; n)$ by π.

The main goal of this subsection is to motivate and establish an asymptotic lower bound for the Bayes risk of such single spike priors under quadratic loss.

Proposition 8.12 *Let $y \sim N_n(\theta, I)$. Let $\pi_n = \pi_S(\tau_n, n)$ be the sparse prior* (8.47)*. Then the n-variate Bayes risk*

$$B(\pi_n) \ge (\tau_n^2 \wedge 2\log n)(1 + o(1)).$$

The proof requires a sequence of lemmas. The first exploits the heuristic used for the sparse univariate prior in Lemma 8.11 in the previous section.

[4] This is a slight abuse of terminology, since the (now discrete valued) parameter is really $\theta \in \{\tau e_1, \dots, \tau e_n\}$ and not I per se.

Lemma 8.13 *The prior $\pi_n = \pi_S(\tau; n)$ has Bayes risk, for squared error loss, bounded as follows*

$$B(\pi_n) \geq \tau^2 E_{\tau e_1}[1 - p_{1n}(y)]^2.$$

Proof Write the Bayes risk in terms of the joint distribution of (θ, y) when $\theta \sim \pi_S(\tau; n)$, and exploit the symmetry with respect to co-ordinates to reduce to the first component:

$$B(\pi_n) = \mathbb{E}\sum_{i=1}^{n}[\hat{\theta}_{\pi,i} - \theta_i]^2 = n\mathbb{E}[\hat{\theta}_{\pi,1} - \theta_1]^2.$$

Now decompose according to the marginal distribution of θ_1 to obtain an analog of (8.55):

$$\mathbb{E}[\hat{\theta}_{\pi,1} - \theta_1]^2 = \frac{1}{n}E_{\tau e_1}[\hat{\theta}_{\pi,1} - \tau]^2 + \frac{1}{n}\sum_{i=2}^{n} E_{\tau e_i}[\hat{\theta}^2_{\pi,1}].$$

With an appropriate choice of τ, to be made later, we might expect the second term on the right side to be of smaller order than the first, compare (8.56) in the previous section. We therefore drop the second term, and using (8.60), find that

$$B(\pi_n) \geq E_{\tau e_1}[\hat{\theta}_{\pi,1} - \tau]^2 = \tau^2 E_{\tau e_1}[p_{1n}(y) - 1]^2. \tag{8.61}$$

□

We turn now to the choice of τ. Whether $P(I = 1|y)$ is close to 1 or 0 depends on whether y_1 is larger than $\max_{j\geq 2} y_j$ or not. Since $y_1 \sim N(\tau, 1)$ and $\max_{j\geq 2} y_j \approx \lambda_n = \sqrt{2\log n}$, one suspects that $P(I = 1|y)$ will typically be close to 1 if $\tau \gg \lambda_n$ and close to 0 if $\tau \ll \lambda_n$. [This is also consistent with the previous section if we take $\alpha = 1/n$ there, and use (8.54) to conjecture that τ_n should be of rough order $\sqrt{2\log n}$.] While these heuristics are basically correct, some subtleties emerge as we argue more rigorously.

We first study the behavior of exponential Gaussian sums such as appear in the denominator of (8.59). Initially, consider the case in which all variables have mean zero. Define

$$W_n = n^{-1}e^{-\tau_n^2/2}\sum_{k=1}^{n} e^{\tau_n z_k}.$$

Since $Ee^{\tau_n z_k} = e^{\tau_n^2/2}$ we have $EW_n = 1$. We might expect a law of large numbers to hold, at least if τ_n is not too big. However, if τ_n is as large as λ_n then W_n fails to be consistent.

Lemma 8.14 *Let $z_1, \ldots, z_n \overset{iid}{\sim} N(0,1)$ and $\lambda_n = \sqrt{2\log n}$. Then*

$$W_n \xrightarrow{p} \begin{cases} 1 & \text{if} \quad \tau_n - \lambda_n \to -\infty \\ 1 - \Phi(v) & \text{if} \quad \tau_n = \lambda_n + v. \end{cases}$$

Remark. Part 6° of Section 8.10 briefly connects the behavior of W_n to results in the random energy model of statistical physics.

Proof If τ_n is small enough, the ordinary weak law of large numbers applies. Fix $\gamma \in (1/2, 1)$: if $\tau_n < \sqrt{\gamma\log n}$, then $\mathrm{Var}\, W_n = n^{-1}(e^{\tau_n^2} - 1) \to 0$, and $W_n \to 1$ in probability by Chebychev's inequality. However, if $\tau_n \geq \sqrt{\gamma\log n}$, the variance can be large and we must truncate, as is done in the triangular array form of the weak law of large numbers,

recalled in Proposition C.14. Put $X_{nk} = e^{\tau_n z_k}$ and $b_n = e^{\tau_n \lambda_n}$, and then introduce $\bar{X}_{nk} = X_{nk} I\{|X_{nk}| \le b_n\} = e^{\tau_n z_k} I\{z_k \le \lambda_n\}$. We must verify the truncation conditions (i) and (ii) of Proposition C.14. As a preliminary, rewrite $n = e^{\lambda_n^2/2} = c_0/\phi(\lambda_n)$, where $c_0 = \phi(0)$. A short calculation shows that for any r,

$$E\bar{X}_{nk}^r = Ee^{r\tau_n z} I\{z \le \lambda_n\} = e^{r^2\tau_n^2/2}\Phi(\lambda_n - r\tau_n).$$

Thus $\sum_1^n P(X_{nk} > b_n) = n\tilde{\Phi}(\lambda_n) \le c_0/\lambda_n \to 0$, using the Mills ratio bound (8.88) and writing $c_0 = \phi(0)$. This establishes condition (i). For $r = 2$, we use the previous display and (8.88) again to obtain

$$b_n^{-2} \sum E\bar{X}_{nk}^2 = c_0\tilde{\Phi}(2\tau_n - \lambda_n)/\phi(2\tau_n - \lambda_n) \le c_0/(2\tau_n - \lambda_n).$$

If $\tau_n \ge \sqrt{\gamma \log n}$ with $\gamma > 1/2$, then $2\tau_n - \lambda_n \to \infty$ and condition (ii) holds.

Now set $a_n = \sum_1^n E\bar{X}_{nk} = ne^{\tau_n^2/2}\Phi(\lambda_n - \tau_n)$. The weak law Proposition C.14 says that $\sum_1^n e^{\tau_n z_k} = a_n + o_p(b_n)$, or equivalently that

$$W_n = \Phi(\lambda_n - \tau_n) + o_p(b_n n^{-1} e^{-\tau_n^2/2}).$$

Now $b_n n^{-1} e^{-\tau_n^2/2} = \exp\{-(\lambda_n - \tau_n)^2/2\}$ and hence $W_n \xrightarrow{p} 1$ if $\lambda_n - \tau_n \to \infty$. If, instead, $\tau_n = \lambda_n + v$, then $W_n = \tilde{\Phi}(v) + o_p(e^{-v^2/2})$ and the second case follows. □

We can now describe how τ_n must be chosen to ensure that the posterior probability $p_{1n}(y) \to 0$ in (8.61).

Lemma 8.15 *Let P_θ denote the law of $y \sim N_n(\theta, I)$ and suppose that $\theta \sim \pi_n$ where $\theta = \tau_n e_I$ for I chosen uniformly on $\{1, \dots, n\}$. If $\lambda_n = \sqrt{2\log n}$ and $\lambda_n - \tau_n \to \infty$, then*

$$p_{1n}(y) = P(I = 1|y) \to 0 \quad \textit{in} \quad P_{\tau_n e_1} - \textit{probability}.$$

Proof Under $P_{\tau_n e_1}$, we have $y_j = \tau_n \delta_{j1} + z_j$ and so from (8.59), we arrive at

$$p_{1n}(y) = [1 + V_n W_{n-1}]^{-1},$$

where $W_{n-1} = (n-1)^{-1} e^{-\tau_n^2/2} \sum_2^n e^{\tau_n z_i}$ and $V_n = (n-1)e^{-\tau_n^2/2 - \tau_n z_1}$.

Then by Lemma 8.14, since $\lambda_n - \tau_n \to \infty$, $W_n \to 1$ in probability as $n \to \infty$. For V_n, observe that

$$\lambda_n^2 - \tau_n^2 - 2\tau_n z \ge (\lambda_n - \tau_n - z_+)(\lambda_n + \tau_n) \to \infty, \tag{8.62}$$

again because $\lambda_n - \tau_n \to \infty$. Consequently $V_n \to \infty$ for each fixed z_1 and so $p_{1n}(y) \to 0$ in probability. □

Proof of Proposition 8.12. First suppose that $\lambda_n = \sqrt{2\log n}$ and that $\lambda_n - \tau_n \to \infty$. We then directly apply Lemmas 8.13 and 8.15: since $x \to (1-x)^2$ is bounded and continuous for $x \in [0,1]$ and $p_{1n}(y) \to 0$ in $P_{\tau_n e_1}$–probability, we have $B(\pi_n) \ge \tau_n^2(1 + o(1))$.

If instead it is not the case that $\lambda_n - \tau_n \to \infty$, then choose $\tau_n' \le \tau_n$ which satisfies both $\tau_n' \sim \tau_n \wedge \lambda_n$ and also $\lambda_n - \tau_n' \to \infty$. For example $\tau_n' = \tau_n \wedge \lambda_n - \log \lambda_n$ will do. Then use τ_n' in the argument of the previous paragraph to conclude that

$$B(\pi_n) \ge \tau_n'^2(1 + o(1)) \sim \tau_n^2 \wedge \lambda_n^2. \quad \square$$

Remark. Let us say a little more about the connections between these calculations and those of the previous section. We may identify λ_n with λ_α for $\alpha = 1/(n+1)$ and then think of τ_n as corresponding to the support point μ_α in the two point prior π_α. The overshoot condition $a = (2\log\alpha^{-1})^\gamma$ for $0 < \gamma < 1/2$ combined with (8.53) shows that $\lambda_\alpha - \mu_\alpha \to \infty$, which is the analog of $\lambda_n - \tau_n \to \infty$. The overshoot condition also shows that the posterior probability $P(\{\mu_\alpha\}|x = \mu_\alpha + w) \to 0$ for each $w \in \mathbb{R}$ as $\alpha \to 0$, which corresponds to $p_{1n}(y) \to 0$ in Lemma 8.15.

Lifting to independent blocks

Proof of Theorem 8.10. The upper bound to the minimax risk $R_N(\Theta_n[k_n])$ was established at (8.44) using soft thresholding. The lower bound, following the approach of Chapter 4, uses a suitable least favorable prior. To guess at a form for this prior, note that the upper bound may be rewritten as $k_n \cdot 2\log(n/k_n)$. This suggests the use of k_n disjoint blocks of dimension n/k_n, each with a single spike of size about $\sqrt{2\log(n/k_n)}$. If the prior makes these blocks independent, then the Bayes risk is the sum of the Bayes risk for the k_n blocks, compare (4.24), and the result will follow from the single spike case, Proposition 8.12.

Now to the details: divide the indices $\{1,\dots,n\}$ into k_n blocks, each of size $m = m_n = [n/k_n]$ and on each such block to use a single spike prior $\pi_S(\tau_m; m)$ with τ_m chosen so that $\tau_m \sim \lambda_m = (2\log m)^{1/2}$ and $\lambda_m - \tau_m \to \infty$ (for example $\tau_m = \lambda_m - \log\lambda_m$ will do). Note that the assumption $k_n/n \to 0$ guarantees that $m_n \to \infty$, as is needed for Proposition 8.12. The product prior π_n^{IB} obtained by making the k_n blocks independent, is supported in $\Theta_n[k_n]$ since it chooses exactly k_n spikes with probability one. Consequently, from (4.13), independence (compare (4.24)) and Proposition 8.12, we have

$$\begin{aligned} R_N(\Theta_n[k_n]) &\geq B(\pi_n^{\mathrm{IB}}) = k_n B(\pi_S(\tau_m; m)) \\ &\geq k_n \tau_m^2 (1 + o(1)) \\ &\sim 2k_n \log(n/k_n)(1 + o(1)). \end{aligned} \tag{8.63}$$

Optimality of $\sqrt{2\log n}$ risk bound. We are now also able establish the minimax lower bound (8.35). Set $\epsilon = 1$ without loss of generality and bring in a non-standard loss function

$$\tilde{L}(\hat{\theta}, \theta) = \frac{\|\hat{\theta} - \theta\|^2}{1 + \sum_i \min(\theta_i^2, 1)}. \tag{8.64}$$

Let $\tilde{r}(\hat{\theta}, \theta)$ and $\tilde{B}(\hat{\theta}, \pi) = \int \tilde{r}(\hat{\theta}, \theta)\pi(d\theta)$ respectively denote risk and integrated risk for the new loss function. The left hand side of (8.35) is the minimax risk for the new loss function, and arguing as in Section 4.3, compare (4.13)–(4.15), we obtain a lower bound

$$\tilde{R}_N = \inf_{\hat{\theta}} \sup_{\theta} \tilde{r}(\hat{\theta}, \theta) \geq \inf_{\hat{\theta}} \sup_{\pi} \tilde{B}(\hat{\theta}, \pi) \geq \sup_{\pi} \tilde{B}(\pi).$$

A nearly least favorable prior is again given by the independent block spike prior $\pi_n = \pi_n^{\mathrm{IB}}$ with $k_n = [\log n]$ blocks eacn of length $m_n = [n/\log n]$. Again, the remaining indices are ignored and choose $\tau_m = \lambda_m - \log\lambda_m$ so that $\lambda_m - \tau_m \to \infty$. Since exactly one

coefficient is non-zero in each block, we have with probability one under π_n that

$$1 + \sum_i \min(\theta_i^2, 1) = 1 + k_n.$$

Thus the modified Bayes risk $\tilde{B}(\pi_n) = B(\pi_n)/(1 + k_n)$, and, arguing again as at (8.63), $B(\pi_n) \geq k_n \tau_m^2 (1 + o(1))$. Putting the pieces together

$$\tilde{R}_N \geq \tilde{B}(\pi_n) \geq (k_n/(1 + k_n))\, \tau_m^2 (1 + o(1)) = (2 \log n)(1 + o(1)).$$

Some other bounds for single spike priors

Using the single spike prior, one can derive lower bounds of different flavors for various purposes. We illustrate with two more examples, each of which will find application later in the book.

First a bound that applies for the whole scale of ℓ_p error measures. It is phrased in terms of the *probability* of a large norm error rather than via an expected p-th power error—this is appropriate for the application to optimal recovery in Chapter 10.

Proposition 8.16 *Fix $\eta > 0$. There exists a function $\pi_\eta(n) \to 1$ as $n \to \infty$ such that for any $\tau_n \leq \sqrt{(2-\eta)\log n}$ and all $p > 0$*

$$\inf_{\hat{\theta}} \sup_{\Theta_n(\tau_n)} P_\theta\{\|\hat{\theta} - \theta\|_p \geq \tau_n/2\} \geq \pi_\eta(n). \tag{8.65}$$

Proof Since the spike prior π_n concentrates on $\Theta_n(\tau)$, we have $\sup_{\theta \in \Theta_n(\tau)} P_\theta(A) \geq \mathbb{P}(A)$, where $\mathbb{P}$ denotes the joint distribution of (θ, y) for $\theta \sim \pi_n$.

The argument makes use of the maximum a posteriori estimator for the spike prior π_n, given by $\hat{\theta}_\pi^{MAP} = \tau e_{\hat{I}}$, where $\hat{I} = \operatorname{argmax}_i P(I = i|y) = \operatorname{argmax}_i y_i$. It is the Bayes estimator for the spike prior π_n and loss function $L(a, \theta) = I\{a \neq \theta\}$, so that for any estimator $\hat{\theta}$, $\mathbb{P}(\hat{\theta} \neq \theta) \geq \mathbb{P}(\hat{\theta}_\pi^{\mathrm{MAP}} \neq \theta)$.

Let $\hat{\theta}$ be a given arbitrary estimator and let $\hat{\theta}^*(y)$ be the estimator defined from it by choosing a point from the set $\{\tau e_1, \ldots, \tau e_n\}$ that is closest to $\hat{\theta}(y)$ in (quasi-)norm $\|\cdot\|_p$. Therefore, if $\|\hat{\theta} - \tau e_i\|_p < \tau/2$ then $\hat{\theta}^* = \tau e_i$— this is obvious for $p \geq 1$, while for $p < 1$ it follows from the triangle inequality for $\|\cdot\|_p^p$. Hence

$$\mathbb{P}(\|\hat{\theta} - \theta\|_p \geq \tau/2) \geq \mathbb{P}(\hat{\theta}^* \neq \theta) \geq \mathbb{P}(\hat{\theta}_\pi^{\mathrm{MAP}} \neq \theta) = \mathbb{P}(\hat{I} \neq I).$$

By symmetry, and recalling that $y_i = \theta_i + z_i$,

$$\begin{aligned} \mathbb{P}\{\hat{I} \neq I\} &= P_{\tau e_1}\{y_1 \neq \max_i y_i\} \\ &= P\{z_1 + \tau < \max_{i=2,\ldots,n} z_i\} = P\{M_{n-1} - Z > \tau\} \end{aligned} \tag{8.66}$$

where $M_n := \max_{i=1,\ldots,n} z_i$ is the maximum of n independent standard Gaussian variates and Z is another, independent, standard Gaussian.

Now appeal to (8.66) and the hypothesis $\tau_n \leq \sqrt{(2-\eta)\log n}$ to conclude that the minimax error probability is bounded below by

$$\pi_\eta(n) = P\{M_{n-1} - Z \geq \sqrt{(2-\eta)\log n}\}.$$

It is intuitively plausible from (8.32) that $\pi_\eta(n) \to 1$ as $n \to \infty$ for fixed η. One possible proof, admittedly crude, goes as follows. Set $a_n = \sqrt{(2-\eta)\log n}$ and $a_n' = \sqrt{(2-\eta')\log n}$ for some $\eta' < \eta$. We have

$$P(M_{n-1} - Z \geq a_n) \geq P(M_{n-1} \geq a_n')P(Z \leq a_n' - a_n).$$

For any $\eta' > 0$, we have $P(M_{n-1} \geq a_n') \to 1$, for example by (8.81) in Section 8.9. A little algebra shows that $a_n' - a_n \geq \sqrt{2\gamma \log n}$ for some $\gamma(\eta, \eta') > 0$ and hence $P(Z \leq a_n' - a_n) \to 1$ also. □

The second result is for MSE and offers an example of a non-asymptotic bound, that is one valid for all finite n. It prepares for further non-asymptotic bounds in Section 11.4. As might be expected, the non-asymptotic bounds are less sharp than their asymptotic cousins. To state them, recall that for a single bounded normal mean in $[-\tau, \tau]$, Section 4.6 showed that the minimax risk $\rho_N(\tau, 1) = R_N([-\tau, \tau], 1)$ satisfies

$$c_0(\tau^2 \wedge 1) \leq \rho_N(\tau, 1) \leq \tau^2 \wedge 1,$$

for a suitable $0 < c_0 < 1$.

Proposition 8.17 *Suppose that $y \sim N_n(\theta, I)$. There exists $c_1 > 0$ such that for all $n \geq 2$,*

$$c_1[\tau^2 \wedge (1 + 2\log n - 2\log\log n))] \;\leq\; R_N(\Theta_n(\tau)) \;\leq\; (\log n)^{-1/2} + \tau^2 \wedge (1 + 2\log n).$$

Proof For the upper bound, consider the maximum risk of soft thresholding at $\lambda_n = \sqrt{2\log n}$. Bound (8.12) says that

$$\sup_{\Theta_n(\tau)} r(\hat{\theta}_{\lambda_n}, \theta) \leq (n-1)r_S(\lambda_n, 0) + r_S(\lambda_n, \tau) \leq n r_S(\lambda_n, 0) + \tau^2 \wedge (1 + \lambda_n^2).$$

We now employ a risk bound for threshold $\lambda_n = \sqrt{2\log n}$ at 0 for $n \geq 2$

$$r_S(\lambda_n, 0) \leq n^{-1}(\log n)^{-1/2}.$$

Indeed, this follows from (8.8) for $n \geq 3$ since then $\lambda_n \geq \sqrt{2}$, while for $n = 2$ we just evaluate risk (8.83) numerically. The upper bound of the proposition is now immediate.

For the lower bound, this time we seek a bound for $B(\pi_n)$ valid for all n. Introduce $\ell_n = \sqrt{1 + 2\log(n/\log n)}$ and $\tau_n = \tau \wedge \ell_n$. We start from (8.61) and note that on the event $E_n = \{y_1 \neq \max_j y_j\}$ we have $p_{1n}(y) \leq 1/2$ and so $B(\pi_n) \geq (\tau_n^2/4)P_{\tau_n e_1}(E_n)$. From (8.66)

$$P(E_n) \geq P\{Z < 0, M_{n-1} > \tau_n\} = \tfrac{1}{2}P\{M_{n-1} \geq \tau_n\} \geq \tfrac{1}{2}P\{M_{n-1} \geq \ell_n\}$$

We leave it as Exercise 8.13 to verify that $P(M_{n-1} > \ell_n) \geq c_0$ for $n \geq 2$. □

8.7 Minimax sparse estimation—univariate model

In this section and the next we formalize the notion of classes of η-sparse signals and consider minimax estimation over such classes. We begin with a univariate model and then show how it leads to results for sparse estimation in $\mathbb{R}^n$. Suppose that $Y = \theta + \epsilon Z$ with

$Z \sim N(0,1)$ and that θ is drawn from a distribution π which assigns probability at most η to the nonzero value. Thus let $\mathcal{P}(\mathbb{R})$ be the collection of probability measures on $\mathbb{R}$ and

$$\mathfrak{m}_0(\eta) = \{\, \pi \in \mathcal{P}(\mathbb{R}) : \ \pi(\{0\}) \geq 1 - \eta \,\}.$$

Equivalently, $\mathfrak{m}_0(\eta)$ consists of those probability measures having a represention

$$\pi = (1-\eta)\delta_0 + \eta\nu, \tag{8.67}$$

where δ_0 is a unit point mass at 0 and ν an arbitrary probability distribution on $\mathbb{R}$. To avoid trivial cases, assume that $0 < \eta < 1$.

Given π, the integrated risk, using squared error loss, for an estimator $\hat{\theta}(y)$ of θ is then $B(\hat{\theta}, \pi) = E_\pi(\hat{\theta}(Y) - \theta)^2$. We study the Bayes minimax risk

$$\beta_0(\eta, \epsilon) := \inf_{\hat{\theta}} \sup_{\pi \in \mathfrak{m}_0(\eta)} B(\hat{\theta}, \pi) = \sup\{\, B(\pi) : \ \pi \in \mathfrak{m}_0(\eta) \,\}, \tag{8.68}$$

where the second equality uses the minimax theorem 4.12. From the scale invariance

$$\beta_0(\eta, \epsilon) = \epsilon^2 \beta_0(\eta, 1),$$

it will suffice to study the unit noise quantity $\beta_0(\eta, 1)$, which we now write as $\beta_0(\eta)$.

Proposition 8.18 *The univariate Bayes risk $\beta_0(\eta)$ is concave and increasing (and hence continuous) for $0 \leq \eta \leq 1$, with $\beta_0(\eta) \geq \eta$ and $\beta_0(1) = 1$. As $\eta \to 0$, the minimax risk*

$$\beta_0(\eta) \sim 2\eta \log \eta^{-1},$$

and an asymptotically minimax rule is given by soft thresholding at $\lambda = (2\log\eta^{-1})^{1/2}$.

Proof First, monotonicity is obvious. Concavity of β_0 follows from concavity of Bayes risk $\pi \to B(\pi)$, Remark 4.1, together with convexity of the constraint defining $\mathfrak{m}_0(\eta)$: if both $\pi_0(\{0\})$ and $\pi_1(\{0\}) \geq \eta$, then of course $((1-\alpha)\pi_0 + \alpha\pi_1)(\{0\}) \geq \eta$.

When $\eta = 1$, there is no constraint on the priors, so by (4.20), $\beta_0(1) = \rho_N(\infty, 1) = 1$. Concavity of the Bayes risk also implies that $B((1-\eta)\delta_0 + \eta\nu) \geq (1-\eta)B(\delta_0) + \eta B(\nu)$, and since $B(\delta_0) = 0$, maximizing over ν shows that $\beta_0(\eta) \geq \eta$.

Finally, we consider behavior as $\eta \to 0$. For soft thresholding δ_λ, we have $r(\lambda, \mu) \leq 1 + \lambda^2$, compare Lemma 8.3. Since $\pi = (1-\eta)\delta_0 + \eta\nu$, we have

$$B(\delta_\lambda, \pi) = (1-\eta)r(\lambda, 0) + \eta \int r(\lambda, \mu)\nu(d\mu) \leq r(\lambda, 0) + \eta(1+\lambda^2).$$

For $\lambda = (2\log\eta^{-1})^{1/2}$ large, recall from (8.7) that $r(\lambda, 0) \sim 4\lambda^{-3}\phi(\lambda) = o(\eta)$, so that

$$\beta_0(\eta) \leq \sup_{\pi \in \mathfrak{m}_0(\eta)} B(\delta_\lambda, \pi) \leq 2\eta \log \eta^{-1} + O(\eta).$$

For a lower bound we choose a sparse prior π as in Lemma 8.11 with sparsity η and overshoot $a = (2\log\eta^{-1})^{1/4}$. Then, from that lemma and (8.54), we obtain

$$\beta_0(\eta) \geq B(\pi_{\eta,\mu(\eta)}) \sim \eta\mu^2(\eta) \sim 2\eta \log \eta^{-1}. \qquad \square$$

The existence and nature of the least favorable distribution for $\mathfrak{m}_0(\eta)$ is of some interest, and will be used later for the pth moment case, Proposition 13.5. The proof may be skipped at a first reading without loss of continuity.

Proposition 8.19 *Assume* $0 < \eta < 1$. *The Bayes minimax problem associated with* $\mathfrak{m}_0(\eta)$ *and* $\beta_0(\eta)$ *has a unique least favorable distribution* π_η. *The measure* π_η *is proper, symmetric and has countably infinite support with* $\pm\infty$ *as the only accumulation points.*

Of course, symmetry means that $\pi_\eta(B) = \pi_\eta(-B)$ for measurable sets $B \subset \mathbb{R}$.

Proof of Proposition 8.19 The set $\mathfrak{m}_0(\eta)$ is not weakly compact; instead we regard it as a subset of $\mathcal{P}_+(\mathbb{R})$, the substochastic measures on $\mathbb{R}$ with positive mass on $\mathbb{R}$, with the vague topology. [For more detail on vague convergence, see appendix C.19, and for example Huber and Ronchetti (2009).] Since $\mathfrak{m}_0(\eta)$ is then vaguely compact, we can apply Proposition 4.13 (via the remark immediately following it) to conclude the existence of a unique least favorable prior $\pi_\eta \in \mathcal{P}_+(\mathbb{R})$. Since $\eta < 1$, we know that $\pi_\eta(\mathbb{R}) > 0$. In addition, π_η is symmetric.

A separate argument is needed to show that π_η is proper, $\pi_\eta(\mathbb{R}) = 1$. Suppose on the contrary that $\alpha = 1 - \pi_\eta(\mathbb{R}) > 0$. From the Fisher information representation (4.4) and (4.21), we know that $P_0 = \Phi \star \pi_\eta$ minimizes $I(P)$ for P varying over the convolution set $\mathfrak{m}_0(\eta)^\star = \{P = \Phi \star \pi : \pi \in \mathfrak{m}_0(\eta)\}$. We may therefore use the variational criterion in the form given at (C.23). Thus, let $P_1 = P_0 + \alpha\Phi \star \mu$ for an arbitrary (prior) probability measure μ on $\mathbb{R}$. Let the corresponding densities be p_1 and p_0, and set $\psi_0 = -p_0'/p_0$. Noting that $p_1 - p_0 = \alpha\phi \star \mu$, we may take $\mu = \delta_\theta$ for each $\theta \in \mathbb{R}$, and (C.23) becomes

$$E_\theta[-2\psi_0' + \psi_0^2] \le 0.$$

Stein's unbiased risk formula (2.58) applied to $d_{\pi_0}(x) = x - \psi_0(x)$ then shows that $r(d_{\pi_0}, \theta) \le 1$ for all θ. Since $d_0(x) = x$ is the *unique* minimax estimator of θ when $x \sim N(\theta, 1)$, Corollary 4.10, we have a contradiction and so it must be that $\pi_\eta(\mathbb{R}) = 1$.

As π_η is proper and least favorable, Proposition 4.14 yields a saddle point $(\hat\theta_{\pi_\eta}, \pi_\eta)$. Using the mixture representation (8.67), with $\nu = \nu_\eta$ corresponding to π_η, well defined because $\eta > 0$, we obtain from (4.22) applied to point masses $\nu = \delta_\theta$ that for all θ

$$r(\hat\theta_{\pi_\eta}, \theta) \le \int r(\hat\theta_{\pi_\eta}, \theta')\nu_\eta(d\theta').$$

In particular, $\theta \to r(\hat\theta_{\pi_\eta}, \theta)$ is uniformly bounded for all θ, and so is an analytic function on $\mathbb{R}$, Remark 4.2. It cannot be constant (e.g. Exercise 4.1) and so we can appeal to Lemma 4.18 to conclude that ν_η is a discrete measure with no points of accumulation in $\mathbb{R}$. The support of ν_η must be (countably) infinite, for if it were finite, the risk function of $\hat\theta_{\pi_\eta}$ would necessarily be unbounded (again, Exercise 4.1). □

8.8 Minimax sparse estimation—multivariate model

This section presents a more comprehensive result for the multivariate minimax risk under exact sparsity, $R_N(\Theta_n[k_n], \epsilon_n)$. We show that in the large n limit, it can be expressed in terms of the univariate Bayes risk $\beta_0(\eta)$.

Theorem 8.20 *Assume model* (8.41) *and parameter space* (8.43) *with* $\eta_n = k_n/n \to \eta \ge 0$. *Then the minimax risk* (8.45) *satisfies*

$$R_N(\Theta_n[k_n], \epsilon_n) \sim n\epsilon_n^2\beta_0(k_n/n).$$

In particular, if $k_n/n \to 0$,

$$R_N(\Theta_n[k_n], \epsilon_n) \sim 2\epsilon_n^2 k_n \log(n/k_n),$$

and the (soft or hard) thresholding estimators $\hat{\theta}_i(y) = \hat{\delta}(y_i, \epsilon_n\sqrt{2\log(n/k_n)})$ *are asymptotically minimax.*

The "highly sparse" case in which the number of spikes k remains fixed is included:

$$R_N(\Theta_n[k], \epsilon_n) \sim 2\epsilon_n^2 k \log n.$$

Proof The case $k_n/n \to 0$ shown in the second display has essentially been proved in previous sections. Indeed, the upper bound was established at (8.44) using soft thresholding at level $\epsilon_n\sqrt{2\log(n/k_n)}$. The same argument works for hard thresholding, now using global risk bound (8.18) and bound (8.20) for the risk at zero. The lower bound was obtained with the independent blocks prior in the paragraph concluding with (8.63). Finally, the equivalence of the two displays $n\beta_0(k_n/n) \sim 2k_n \log(n/k_n)$ was shown in Proposition 8.18.

We turn now to the proof when $\eta_n \to \eta > 0$. The proof uses the Bayes minimax method sketched in Chapter 4 with both upper and lower bounds derived in terms of priors built from i.i.d. draws from univariate priors in $\mathfrak{m}_0(\eta_n)$. As an intermediate, we need the class of priors supported *on average* on $\Theta_n[k_n]$, recall (8.42),

$$\mathcal{M}_n = \mathcal{M}_n[k_n] = \{\pi \in \mathcal{P}(\mathbb{R}^n) : \ E_\pi \|\theta\|_0 \le k_n\}$$

and the subclass $\mathcal{M}_n^e = \mathcal{M}_n^e[k_n] \subset \mathcal{M}_n[k_n]$ of *exchangeable* or permutation-invariant priors.

The upper bound can now be outlined in a single display,

$$R_N(\Theta_n[k_n], \epsilon_n) \le B(\mathcal{M}_n, \epsilon_n) = B(\mathcal{M}_n^e, \epsilon_n) = n\epsilon_n^2\beta_0(k_n/n). \tag{8.69}$$

To explain, recall that $B(\mathcal{M}, \epsilon) = \sup\{B(\pi), \pi \in \mathcal{M}\}$. The first inequality follows because $\mathcal{M}_n$ contains all point masses δ_θ for $\theta \in \Theta_n[k_n]$, compare (4.18). If we start with a draw from prior π and then permute the coordinates randomly with a permutation σ, where the average is taken over permutations of the n coordinates, we obtain a new, exchangeable prior $\pi^e = \mathrm{ave}(\pi \circ \sigma)$. Concavity of the Bayes risk, Remark 4.1, guarantees that $B(\pi) \le B(\pi^e)$; this implies the second equality, since $\mathcal{M}_n^e \subset \mathcal{M}_n$.

The univariate marginal π_1 of an exchangeable prior $\pi \in \mathcal{M}_n^e[k_n]$ belongs to $\mathfrak{m}_0(k_n/n)$, and the independence trick of Lemma 4.15 says that if we make all coordinates independent with marginal π_1, then the product prior π_1^n is less favorable than π—recall that the posterior variance of each θ_i in π_1^n depends on only y_i and so is larger than the posterior variance of θ_i in π, which may depend on all of y. As a result,

$$B(\pi) \le B(\pi_1^n) = nB(\pi_1).$$

Rescaling to noise level one and maximizing over $\pi_1 \in \mathfrak{m}_0(k_n/n)$, we obtain the equality in the third part of (8.69). Note that the upper bound holds for *all* $k_n \le n$.

The idea for the lower bound runs as follows. Using the arguments of Section 4.3,

$$R_N(\Theta_n[k_n], \epsilon_n) \ge \sup\{B(\pi), \mathrm{supp}\ \pi \subset \Theta_n[k_n]\}.$$

An approximately least favorable prior for the right side might be constructed as $\pi = \pi_1^n$, corresponding to taking n i.i.d. draws from a univariate prior $\pi_1 \in \mathfrak{m}(k_n/n)$ with π_1 chosen to be nearly least favorable for $\mathfrak{m}(k_n/n)$. This is a version of the prior π^{IID} described in Section 8.4. The same technical difficulty arises: let $N_n = \#\{i : \theta_i \neq 0\}$. Even though $EN_n \leq k_n$, we don't have $\pi(N_n \leq k) = 1$ and so it is not guaranteed that supp $\pi \subset \Theta_n[k_n]$.

The Bayes minimax method of Section 4.11 patches this up by modifying the definition of π so that $\pi(N_n \leq k) \to 1$ as $n \to \infty$. The family of parameter spaces will be $\Theta_n[k]$, nested by k. The sequence of problems will be indexed by n, so that the noise level ϵ_n and sparsity k_n depend on n. We use the exchangeable classes of priors $\mathcal{M}_n^e$ defined above, with Bayes minimax risk given by $n\epsilon_n^2\beta_0(k_n/n)$, compare (8.69). We introduce the notation

$$B_n(\kappa, \epsilon_n) = n\epsilon_n^2\beta_0(\kappa/n), \tag{8.70}$$

which is equally well defined for non-integer κ, compare definition (8.68). For each fixed $\gamma < 1$, then, we will construct a sequence of priors $\pi_n \in \mathcal{M}_n^e[\gamma k_n]$. which are built from i.i.d. draws from a suitable one-dimensional distribution π_{1n}. With Θ_n denoting $\Theta_n[k_n]$, we will show that π_n has the properties, as $n \to \infty$,

$$B(\pi_n) \geq \gamma B_n(\gamma k_n, \epsilon_n)(1 + o(1)), \tag{8.71}$$

$$\pi_n(\Theta_n) \to 1, \tag{8.72}$$

$$\mathbb{E}_{\pi_n}\{\|\hat{\theta}_{\nu_n}\|^2 + \|\theta\|^2, \Theta_n^c\} = o(B_n(\gamma k_n, \epsilon_n)) \tag{8.73}$$

where $\nu_n(\cdot) = \pi_n(\cdot|\Theta_n)$, and $\hat{\theta}_{\nu_n}$ is the Bayes estimator for the conditioned prior ν_n and that

$$\lim_{\gamma \to 1} \liminf_{n \to \infty} \frac{B_n(\gamma k_n, \epsilon_n)}{B_n(k_n, \epsilon_n)} = 1. \tag{8.74}$$

It then follows from Lemma 4.32 and the discussion after (4.72) that $R_N(\Theta_n[k_n], \epsilon_n) \geq B_n(k_n, \epsilon_n)(1 + o(1))$. In conjunction with the upper bound in (8.69) this will complete the proof of Theorem 8.20.

For $\gamma < 1$, we may choose M and a univariate prior $\pi_M \in \mathfrak{m}_0(\gamma\eta)$ with support contained in $[-M, M]$ and satisfying $B(\pi_M) \geq \gamma\beta_0(\gamma\eta)$, compare Exercise 4.5. The corresponding prior π_n in the noise level ϵ_n problem is constructed as $\theta_i = \epsilon_n\mu_i$, where $\mu_1, \ldots, \mu_n$ are i.i.d. draws from π_M. By construction and using $\beta_0(\eta_n) \sim \beta_0(\eta)$, we then have

$$B(\pi_n) = n\epsilon_n^2 B(\pi_M) \geq n\epsilon_n^2\gamma\beta_0(\gamma\eta) \sim \gamma B_n(\gamma k_n, \epsilon_n), \tag{8.75}$$

where the final equivalence uses (8.70) and the fact that $\beta_0(\gamma\eta_n) \sim \beta_0(\gamma\eta)$, a consequence of Proposition 8.18.

Since $\pi_M\{\mu_1 \neq 0\} \leq \gamma\eta$, we may bound $\|\theta\|_0$ above stochastically by a Binomial$(n, \gamma\eta)$ variable, N_n say, so that

$$\pi_n\{\Theta_n^c\} \leq P\{N_n - EN_n > k_n - n\gamma\eta\} = O(n^{-1}),$$

for example by Chebychev's inequality, since Var $N_n \leq \gamma\eta n$ and $k_n - \gamma\eta n = (\eta_n - \gamma\eta)n \geq \delta n$ for n large.

For the technical condition (8.73), observe that under π_n, we have $\|\theta\|^2 \leq n\epsilon_n^2 M^2$ with probability one, so that the same is true for $\|\hat{\theta}_{\nu_n}\|^2$, and so the left side of (8.73) is bounded

by $2nM\epsilon_n^2\pi_n(\Theta_n^c)$. On the other hand $B(\gamma k_n, \epsilon_n) \sim n\epsilon_n^2\beta_0(\gamma\eta)$, so that (8.73) also follows from $\pi_n(\Theta_n^c) \to 0$.

Property (8.74) follows from the continuity of $\beta_0(\gamma\eta)$ as $\gamma \to 1$. □

Remark. This Bayes minimax approach also works when $\eta_n \to 0$, so long as $k_n \to \infty$. For further discussion of this in the case of ℓ_p balls, see Section 13.5.

8.9 The distribution of $M_n = \max Z_i$

Simple bounds follow from the concentration inequalities (2.74) and (2.75). Since $z \to \max z_i$ is a Lipschitz(1) function, we have for $t > 0$

$$P\{|M_n - \mathrm{Med}M_n| \ge t\} \le e^{-t^2/2} \tag{8.76}$$
$$P\{|M_n - EM_n| \ge t\} \le 2e^{-t^2/2}.$$

Both $\mathrm{Med}M_n$ and EM_n are close to $L_n = \sqrt{2\log n}$. Indeed

$$|EM_n - \mathrm{Med}M_n| \le \sqrt{2\log 2}, \tag{8.77}$$
$$L_n - 1 \le \mathrm{Med}M_n \le L_n, \tag{8.78}$$
$$L_n - 1 - \sqrt{2\log 2} \le EM_n \le L_n. \tag{8.79}$$

Here the upper bounds hold for all n, and the lower ones for $n \ge 14$. Indeed, the bound (8.77) is Exercise 2.22. The right bound of (8.78) follows from (8.80) below, and for the left bound see Exercise 8.14. The upper bound of (8.79) is Proposition C.12 and the lower bound then follows from (8.77) and (8.78). Of course, asymptotic expressions for $\mathrm{Med}M_n$ and EM_n follow with some extra work from the extreme value theorem (8.32).

In fact the distribution of M_n is confined largely to a shrinking interval of width $2\log^2 L_n/L_n$, mostly below L_n. Indeed, arguing analogously to (8.31), we have for $n \ge 2$,

$$P\{M_n \ge L_n\} \le 1/(\sqrt{2\pi}L_n). \tag{8.80}$$

while Exercise 8.14 shows that for $L_n \ge 3$,

$$P\{M_n \le L_n - 2L_n^{-1}\log^2 L_n\} \le \exp\{-\tfrac{1}{3}\exp(\log^2 L_n)\}. \tag{8.81}$$

Numerics. Finding the quantiles of M_n, defined by $P\{M_n \le x_\alpha\} = \alpha$, is easily done, and yields the central columns in the table below. EM_n is found as in Exercise 8.11. We abbreviate the lower bound by $\tilde{L}_n = L_n - 2L_n^{-1}\log^2 L_n$

n	$\tilde{L}_n$	$x_{.10}$	$x_{.50}$	EM_n	$x_{.90}$	L_n
32	1.92	1.48	2.02	2.07	2.71	2.63
128	2.29	2.10	2.55	2.59	3.15	3.11
1024	2.79	2.84	3.20	3.25	3.71	3.72
4096	3.11	3.26	3.58	3.63	4.05	4.08

8.10 Further details

1°. The mean squared error of a thresholding rule $\hat{\delta}(x,\lambda)$ (either hard or soft) is found by breaking the range of integration into regions $(-\infty,-\lambda), [-\lambda,\lambda]$ and (λ,∞) to match the thresholding structure. For example, with soft thresholding

$$\begin{aligned} r(\lambda,\mu) &= E_\mu[\hat{\delta}(x,\lambda)-\mu]^2 \qquad (8.82)\\ &= \int_{-\infty}^{-\lambda}(x+\lambda-\mu)^2\phi(x-\mu)dx + \int_{-\lambda}^{\lambda}\mu^2\phi(x-\mu)dx + \int_{\lambda}^{\infty}(x-\lambda-\mu)^2\phi(x-\mu)dx \end{aligned}$$

One obtains the following basic mean squared error formulas:

$$\begin{aligned} r_S(\lambda,\mu) &= 1+\lambda^2+(\mu^2-\lambda^2-1)[\Phi(\lambda-\mu)-\Phi(-\lambda-\mu)] \qquad (8.83)\\ &\quad -(\lambda-\mu)\phi(\lambda+\mu)-(\lambda+\mu)\phi(\lambda-\mu),\\ r_H(\lambda,\mu) &= \mu^2[\Phi(\lambda-\mu)-\Phi(-\lambda-\mu)]+\tilde{\Phi}(\lambda-\mu)+\tilde{\Phi}(\lambda+\mu) \qquad (8.84)\\ &\quad +(\lambda-\mu)\phi(\lambda-\mu)+(\lambda+\mu)\phi(\lambda+\mu) \end{aligned}$$

where ϕ and Φ denote the standard Gaussian density and cumulative distribution functions respectively, and $\tilde{\Phi}(x)=1-\Phi(x)$.

2°. *Proof of lower bound in Lemma 8.3 for* $0\le\lambda\le 2$. Let μ_λ be the solution in μ of $r(\lambda,0)+\mu^2 = 1+\lambda^2$. Since $r(\lambda,0)\le e^{-\lambda^2/2}<1$, (compare (8.7)), it is clear that $\mu_\lambda>\lambda$. For $\mu\le\mu_\lambda$ we may write, using (8.5),

$$R(\lambda,\mu)=\frac{r(\lambda,\mu)}{\bar{r}(\lambda,\mu)}=\frac{r(\lambda,0)+\int_0^\mu 2s\Phi(I_\lambda-s)ds}{r(\lambda,0)+\mu^2}.$$

We first verify that $R(\lambda,\mu)$ is decreasing in $\mu\le\mu_\lambda$. Indeed $\mu\to[c+f_1(\mu)]/[c+f_2(\mu)]$ is decreasing if both $f_1'(\mu)\le f_2'(\mu)$ 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 $\mu\ge\mu_\lambda$, we also have $R(\lambda,\mu)\ge R(\lambda,\mu_\lambda)$ since $r(\lambda,\mu)\ge r(\lambda,\mu_\lambda)$ while $\bar{r}(\lambda,\mu)\equiv 1+\lambda^2$. Consequently, for all μ

$$R(\lambda,\mu)\ge r(\lambda,\mu_\lambda)/[r(\lambda,0)+\mu_\lambda^2],$$

and numerical evaluation for $0\le\lambda\le 2$ shows the right side to be bounded below by .516, with the minimum occurring for $\lambda\in[.73,.74]$.

3°. *Proof of (8.31).* We have that $\varpi_n = 1-(1-\delta)^n\le n\delta$, with

$$\delta = 2\tilde{\Phi}(\sqrt{2\log n})\le\frac{2\phi(\sqrt{2\log n})}{\sqrt{2\log n}}=\frac{1}{n\sqrt{\pi\log n}}.$$

4°. *Proof of* (8.36). Let us first observe that, so long as $nr_S(\lambda,0)\ge 1$,

$$\sup_\mu\frac{r_S(\lambda,\mu)}{n^{-1}+\mu^2\wedge 1}\le\frac{n}{n+1}\max\{(n+1)r_S(\lambda,0),1+\lambda^2\}=:\Lambda_n(\lambda), \qquad (8.85)$$

say. To see this, consider two cases separately. For $\mu\ge 1$, the risk $\mu\to r_S(\lambda,\mu)$ increases to $1+\lambda^2$ at $\mu=\infty$. For $\mu\le 1$, the ratio on the left side is bounded using (8.10) by

$$\frac{r_S(\lambda,0)+\mu^2}{n^{-1}+\mu^2}\le nr_S(\lambda,0).$$

Thus, for λ satisfying $nr_S(\lambda,0)\ge 1$,

$$r_S(\lambda,\mu)\le\Lambda_n(\lambda)\{n^{-1}+\min(\mu^2,1)\},$$

and the bound (8.36) follows by adding over co-ordinates and rescaling to noise level ϵ.

Now we seek the minimum value of $\Lambda_n(\lambda)$. Since $\lambda\to 1+\lambda^2$ is strictly increasing and $\lambda\to r_S(\lambda,0)$ is

strictly decreasing (as is seen from the integral in (8.7)), it follows that when $nr_S(\lambda, 0) \geq 1$, the minimum value of $\Lambda_n(\lambda)$ is attained for the (unique) solution $\lambda_n^* \in (0, \infty)$ of the equation (8.37). So we must check that $nr_S(\lambda_n^*, 0) \geq 1$, which from (8.37) is the same as $\lambda_n^* \geq n^{-1/2}$. So now it suffices to check that $(n+1)r_S(n^{-1/2}, 0) \geq 1 + n^{-1}$, that is, $nr_S(n^{-1/2}, 0) \geq 1$. But, using (8.7), the left side increases with n, and is seen numerically to exceed 1 already at $n = 4$. Finally, we set $\Lambda_n^* = \Lambda_n(\lambda_n^*) = 1 + \lambda_n^{*2}$.

5°. *Proof of second half of* (8.56). Combining (8.49) and (8.50), we have $m(x) = e^{\mu(\mu+a-x)}$. Using formula (8.51) for δ_π, then changing variables to $z = x - \mu - a$ and finally exploiting (8.50), we find that

$$(1-\alpha)E_0\delta_\pi^2 = (1-\alpha)\mu^2 \int \frac{\phi(x)dx}{[1+e^{\mu(\mu+a-x)}]^2} = \mu^2\alpha\phi(a)\int_{-\infty}^{\infty} \frac{e^{-(\mu+a)z-z^2/2}dz}{[1+e^{-\mu z}]^2}.$$

We now verify that

$$\frac{(1-\alpha)E_0\delta_\pi^2(z)}{\alpha\mu^2(\alpha)} \leq \phi(a)\int_0^\infty e^{-(\mu+a)z-z^2/2}dz + \int_{-\infty}^{\infty} \frac{\phi(w)dw}{1+e^{\mu(w+a)}}. \tag{8.86}$$

Consider the final integral in the antepenultimate display, first over $(0, \infty)$: we may replace the denominator by 1 to obtain the first term in (8.86). Over $(-\infty, 0)$, we have $e^{-\mu z}/[1+e^{-\mu z}] \leq 1$, and with $v = -z$ this part of the integral is bounded by

$$\mu^2\alpha\int_0^\infty \frac{\phi(v-a)dv}{1+e^{\mu v}},$$

which with $w = v - a$ leads to the second term in (8.86). By dominated convergence, both right hand side terms converge to zero as μ and $a \to \infty$.

6°. The variable $\sum_{k=1}^n e^{\tau_n z_k}$ is the basic quantity studied in the *random energy model* of statistical physics, where it serves as a toy model for spin glasses, e.g. Mézard and Montanari (2009, Ch. 5). In the current notation, it exhibits a phase transition at $\tau_n = \lambda_n = \sqrt{2\log n}$, with qualitatively different behavior in the "high temperature" ($\tau_n < \lambda_n$) and "low temperature" ($\tau_n > \lambda_n$) regimes.

Here is a little more detail on the phase transition. Write $S_n(\beta) = \sum_1^n e^{\beta z_k}$ for $z_1, \ldots, z_n \overset{\text{iid}}{\sim} N(0,1)$. If β is small enough, then heuristically the sum behaves like n times its expectation and

$$\log\big(\sum e^{\beta z_k}\big) \approx \log(nEe^{\beta z_1}) = (\lambda_n^2 + \beta^2)/2,$$

with $\lambda_n = \sqrt{2\log n}$. However, for large β, it is better to approximate the sum by the dominant term

$$\log\big(\sum e^{\beta z_k}\big) \approx \beta z_{(1)} \approx \beta\lambda_n.$$

The crossover in behavior occurs for β near λ_n, see in part Proposition C.12, and is formalized in the following statement, which may be proved directly from the discussion of the random energy model in Talagrand (2003, Ch. 1.1, 2.2). If $\lambda_n = \sqrt{2\log n}$ and $a_n \to a > 0$, then

$$\frac{\log S_n(a_n\lambda_n)}{\log n} \xrightarrow{\ p\ } \begin{cases} 1+a^2 & a \leq 1 \\ 2a & a > 1. \end{cases} \tag{8.87}$$

8.11 Notes

§8.2, 8.3. Many of the ideas and bounds for soft and hard thresholding in Sections 8.2 and 8.3, and in particular the oracle inequality Proposition 8.8, come from Donoho and Johnstone (1994a), see also Donoho and Johnstone (1996) In classical antiquity an oracle, such as the priestess Pythia at the temple of Apollo at Delphi, was a person or agency believed to convey wise counsel inspired by the gods.

Soft thresholding appeared earlier as one of the class of limited translation rules of Efron and Morris (1971) and also in Bickel (1983), as discussed below.

The discussion of block soft thresholding leading to Propositions 8.6 and 8.9 is inspired by the study of block *James-Stein* thresholding in Cai (1999): the definitions were compared in Section 7.6.5. Block

soft thresholding is somewhat easier to analyze because of the monotonicity property of its mean squared error, cf. (8.28), but leads to similar results: Proposition 8.9 and the asymptotic dependence of threshold λ_L on block size L in (8.40) essentially match Cai's conclusions. The paper of Cavalier and Tsybakov (2001), already mentioned in the notes to Chapter 6, shows the broad adaptivity properties of penalized blockwise Stein methods, and the interaction between block sizes and penalties. Some of the properties of block soft thresholding developed here appear also in Donoho et al. (2012) where they are used to study phase transitions in compressed sensing.

Some of the methods used in Proposition 8.6 are derived from a distinct but related problem studied in Johnstone (2001), namely threshold estimation of the noncentrality parameter based on $W \sim \chi^2_d(\xi)$.

§8.4. The discussion paper Donoho et al. (1992) identified ℓ_0 sparsity—the term 'nearly black' was used there for sparse non-negative signals—as a property leading to significant reductions in minimax MSE: Theorem 8.10 is established there in the case $k_n \to \infty$.

§8.5. The discussion of sparse univariate two point priors builds on Donoho et al. (1992), which was in turn influenced by Bickel (1983, 1981), where priors with atoms at 0 appear in the study of minimax estimation of a normal mean subject to good risk properties at a point, such as $\mu = 0$.

§8.6. There is some discussion of single spike priors in Donoho et al. (1997), for example a version of Proposition 8.16, but most of the development leading to the k_n-fixed part of Theorem 8.10 is new to this volume. Parallel results were obtained independently by Zhang (2012b).

§8.7. The study of Fisher information over classes of distributions with a sparse convolution component, such as $F = \Phi \star \pi$ for $\pi \in \mathfrak{m}_0(\eta)$ was stimulated by Mallows (1978). The key parts of Proposition 8.18 are established in Bickel (1983). Bickel and Collins (1983) studied the minimization of Fisher information over classes of distributions. Via Brown's identity, Proposition 4.5, this leads to results for the Bayes minimax risk (8.68). In particular Proposition 8.19 is proven there.

§8.8. Theorem 8.8 is an ℓ_0 or exact sparsity version of a corresponding result for ℓ_p balls 13.17, a version of which was first established in Donoho and Johnstone (1994b).

§8.9. An alternative reference for standard extreme value theory results for M_n and $|M|_n$ is de Haan and Ferreira (2006, Theorem 2.2.1). DasGupta (2011b) gives a detailed nonasymptotic analysis of the mean and median of M_n, and explores use of $\tilde{\lambda}_n = \Phi^{-1}(1-e^{-\gamma}/n)$, γ = Euler's constant, as an improvement upon the $\lambda_n = \sqrt{2\log n}$ threshold.

Lai and Robbins (1976) show that $EM_n \le (2\log n - \log\log n)^{1/2}$ for $n \ge 3$ and that this holds whatever be the interdependence among the $z_i \sim_d N(0,1)$.

Exercises

8.1 *(Mills ratio and Gaussian tails.)* The function $R(\lambda) = \tilde{\Phi}(\lambda)/\phi(\lambda)$ is sometimes called Mills ratio. Show that the modified form

$$M(\lambda) = \frac{\lambda\tilde{\Phi}(\lambda)}{\phi(\lambda)} = \int_0^\infty e^{-v-v^2/(2\lambda^2)}dv,$$

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

$$\tilde{\Phi}_l(\lambda) = \lambda^{-1}\phi(\lambda)\sum_{k=0}^{l}\frac{(-1)^k}{k!}\frac{\Gamma(2k+1)}{2^k\lambda^{2k}}.$$

Show that for each $k \ge 0$ and all $\lambda > 0$ that

$$\tilde{\Phi}_{2k+1}(\lambda) \le \tilde{\Phi}(\lambda) \le \tilde{\Phi}_{2k}(\lambda).$$

[Hint: induction shows that $(-1)^{l-1}[e^{-x} - \sum_0^l (-1)^k x^k/k! \ge 0$ for $x \ge 0$.]
As consequences, we obtain, for example, the bounds

$$\lambda^{-1}\phi(\lambda)(1-\lambda^{-2}) \le \tilde{\Phi}(\lambda) \le \lambda^{-1}\phi(\lambda), \tag{8.88}$$

and the expansion, for large λ,

$$\tilde{\Phi}(\lambda) = \lambda^{-1}\phi(\lambda)[1 - \lambda^{-2} + 3\lambda^{-4} - 15\lambda^{-6} + O(\lambda^{-8})]. \tag{8.89}$$

Show that the general term in brackets $[\cdot]$ is $(-1)^k(2k-1)!!/\lambda^{2k}$ where $(2k-1)!! = (2k-1)\times(2k-3)\times\cdots\times 3\times 1$.

8.2 *(alternate hard threshold bound.)* Show how the proof of Proposition 8.1 can be modified so as to show that for all $\lambda > 0$,

$$r_H(\lambda,\theta) \le \begin{cases} 2[\theta^2 + 2(\lambda+15)\phi(\lambda-1)\epsilon^2] & \text{if } |\theta| \le \epsilon \\ 2(\lambda^2+1)\epsilon^2 & \text{if } |\theta| > \epsilon. \end{cases}$$

8.3 *(Risk of soft thresholding at 0.)* Let $z \sim N(0,1)$, and $r_S(\lambda,0) = E\hat{\delta}_S^2(z)$ denote the mean squared error of soft thresholding at $\lambda = 0$, compare (8.7).
(a) Use (8.88) and (8.89) to show that

$$\begin{aligned} r_S(\lambda,0) &\le 4\lambda^{-3}(1+1.5\lambda^{-2})\phi(\lambda) && \lambda > 0, \\ r_S(\lambda,0) &\sim 4\lambda^{-3}\phi(\lambda) && \lambda \to \infty. \end{aligned}$$

(b) Conclude that $r_S(\lambda,0) \le 4\lambda^{-1}\phi(\lambda)$ if, say, $\lambda \ge \sqrt{2}$.
(c) Let $\delta(\lambda) = e^{-\lambda^2/2} - r_S(\lambda,0)$. Use (8.88) to show that $\delta(\lambda) > 0$ for $\lambda \ge \lambda_0 = 2\phi(0)$.
(d) Show that $\delta(\lambda)$ is concave for $\lambda \in [0,1]$, and conclude that $r(\lambda,0) \le e^{-\lambda^2/2}$ for all $\lambda \ge 0$.

8.4 Derive the following inequalities for hard thresholding, which are sharper than direct application of the bounds in (8.17):

$$\begin{aligned} r_H(\lambda,\lambda) &\ge (\lambda^2+1)/2, \\ r_H(\lambda,0) &\ge (2\lambda \vee \sqrt{2\pi})\phi(\lambda), \\ r_H(\lambda,0) &\le (2\lambda + \sqrt{2\pi})\phi(\lambda) \\ r_H(\lambda,0) &\le 2(\lambda + 1/\lambda)\phi(\lambda). \end{aligned}$$

(Birgé and Massart, 2001)

8.5 *(Hard thresholding risk as function of λ.)*
(a) Use the hard thresholding analog of (8.82) to show that

$$\partial r_H(\lambda,\mu)/\partial\lambda = \lambda(2\mu-\lambda)\phi(\lambda-\mu) - \lambda(2\mu+\lambda)\phi(\lambda+\mu).$$

(b) Verify that $\phi(x) \le \frac{1}{2}$ and $|x|\phi(x) \le \frac{1}{4}$ for all x, and hence that for $\mu \ge 0$,

$$\partial r_H(\lambda,\mu)/\partial\lambda \le \lambda \qquad \text{for } 0 \le \lambda \le 1.$$

(c) Verify (8.18) for $\le \lambda \le 1$.

8.6 *(risk behavior near threshold.)* In the notation of Section 8.2, show that
(i) for soft thresholding, as $\lambda \to \infty$,

$$r_S(\lambda,\lambda) = \lambda^2 - \sqrt{2/\pi}\lambda + 1/2 + \tilde{\Phi}(2\lambda) \sim \lambda^2.$$

(ii) for hard thresholding, as $\lambda \to \infty$,

$$\begin{aligned} r_H(\lambda, \lambda - 2\sqrt{\log\lambda}) &= (\lambda - 2\sqrt{\log\lambda})^2 + O((\log\lambda)^{-1/2}), \\ r_H(\lambda,\lambda) &\sim \lambda^2/2, \\ r_H(\lambda, \lambda + 2\sqrt{\log\lambda}) &\le 1 + (2\pi\log\lambda)^{-1/2}. \end{aligned}$$

8.7 *(Empirical approximation to minimax MSE threshold.)*
This exercise proposes a numerical check on the properties of the empirical approximation (8.38) to the minimax MSE threshold λ_n^* defined in Section 8.3.
(a) Using the values in Table 8.2, do a linear regression of λ_n^* on $\sqrt{\log n}$ and compare the resulting coefficients with those in the approximation (8.38).
(b) For a grid of values of $25 \le n \le 25,000$ evaluate numerically λ_n^*, Λ_n^* and $\Lambda_n(\tilde{\lambda}_n)$–the latter is defined at (8.85). Now, for this grid, evaluate

$$\Lambda_n(\tilde{\lambda}_n)/\Lambda_n^*, \qquad \Lambda_n(\tilde{\lambda}_n)/(2\log n + 1)$$

and so substantiate the claims that $\Lambda_n(\tilde{\lambda}_n)$ is within 10% of Λ_n^* and also between one third and one half of $2\log n + 1$. Verify also that the condition $n r_S(\tilde{\lambda}_n, 0) \ge 1$ is satisfied.

8.8 *(Crude bound for noncentral χ^2.)* If $\xi \le \tau/2$ and $\tau \ge 2d$, show that

$$P(\chi_d^2(\xi) \le \tau) \ge 1/4.$$

[One approach: write $\chi_d^2(\xi) \overset{\mathcal{D}}{=} \chi_{d-1}^2 + (Z + \sqrt{\xi})^2$ with Z an independent standard normal variate and exploit $\{\chi_d^2(\xi) \le \tau\} \supset \{\chi_{d-1}^2 + Z^2 + \xi \le \tau, Z < 0\}$ along with (2.90).

8.9 *(Unbiased risk estimate for block soft thresholding.)* The vectors of the vector field $x \to x/\|x\|$ in $\mathbb{R}^d$ have constant length for $x \ne 0$. Nevertheless, show that its divergence $\nabla^T(x/\|x\|) = (d-1)/\|x\|$. Verify the unbiased risk formula (8.26)– (8.27).

8.10 *(Number of exceedances of universal threshold.)* Let $N_n = \sum_1^n I\{|Z_i| \ge \sqrt{2\log n}\}$.
(a) If Z_i are i.i.d. $N(0,1)$, show that $N_n \sim \mathrm{Bin}(n, p_n)$ with $p_n = 2\tilde{\Phi}(\sqrt{2\log n})$.
(b) Show that $P(N_n \ge 2) \le (np_n)^2 \le 1/(\pi \log n)$.
(c) Show that the total variation distance between the distribution of N_n and that of a Poisson(np_n) variate converges to 0 as $n \to \infty$.

8.11 *(Expected value of maximum.)* If $Z_i \overset{\text{ind}}{\sim} N(0,1)$ and $M_n = \max_{1 \le i \le n} Z_i$, show that

$$EM_n = \int_{-\infty}^0 \Phi^n(z)dz + \int_0^\infty [1 - \Phi^n(z)]dz.$$

8.12 *(Maximum of absolute values of Gaussian noise mimics M_{2n}.)* Let $h_i, i = 1, \ldots$ be independent half-normal variates (i.e. $h_i = |Z_i|$ for $Z_i \sim N(0,1)$), and ϵ_i be independent ± 1 variates, independent of $\{h_i\}$. Let $Z_i = h_i\epsilon_i$ and T_n be the random time at which the number of positive ϵ_i reaches n. Show that the Z_i are independent standard normal and that

$$\max_{i=1,\ldots,n} |Z_i| \overset{\mathcal{D}}{=} \max_{i=1,\ldots,n} h_i = \max_{i=1,\ldots,T_n} Z_i = M_{T_n},$$

and that T_n is close to $2n$ in the sense that

$$(T_n - 2n)/\sqrt{2n} \Rightarrow N(0,1).$$

8.13 *(Lower bound for maximum of Gaussians.)* Let $z_i \overset{i.i.d.}{\sim} N(0,1)$ and $M_n = \max z_i$. Let $\ell_n = \sqrt{1 + 2\log(n/\log n)}$. Show that for some $c_1 > 0$, for all $n \ge 2$,

$$P(M_{n-1} \ge \ell_n) \ge c_1.$$

Hint. Use (8.88) and $(1-x)^m \le e^{-mx}$.

8.14 *(Left tail bound for M_n.)* Let $L_n = \sqrt{2\log n}$, and as above $M_n = \max z_i$.
(a) Show that $P\{M_n \le \lambda\} \le \exp\{-n\tilde{\Phi}(\lambda)\}$.

(b) Establish the left side of bound (8.78) for $n \geq 14$ by using (8.88) and numerical evaluation.
(c) Again use (8.88) to show that

$$P\{M_n \leq L_n - 2L_n^{-1}\log^2(L_n)\} \leq \exp\{-H(L_n)\exp(\log^2 L_n)\}$$

where $H(x) = \phi(0)(\lambda^{-1} - \lambda^{-3})\exp\left(\log^2 x - 2x^{-2}\log^4 x\right)$ and $\lambda = x - 2x^{-1}\log^2 x$. Verify numerically that $H(x) \geq 1/3$ for $x \geq 3$ and hence conclude (8.81).

8.15 *(Properties of Miller's selection scheme.)* 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 $1/\sqrt{N-1}$.

8.16 *(Miller's selection scheme requires many components.)* Suppose that $x_1 = (1,-1,0)^T/\sqrt{2}$ and $x_2 = (0,-1,1)^T/\sqrt{2}$. Consider the random permutations x_1^* and x_2^* described in A. Miller's selection method. Compute the distribution of $\langle x_1^*, x_2^*\rangle$ and show in particular that it equals 0 with zero probability.

8.17 *(Plotting risk functions for sparse two point priors.)* Consider the two point prior (8.48), and the associated version (8.50) having sparsity α and overshoot a. At (8.56) we computed the approximate risk function at two points $\mu' = 0$ and $\mu' = \mu(\alpha)$. Here, make numerical plots of the risk function $\mu' \to r(\delta_\pi, \mu') = E_{\mu'}[\delta_\pi(x) - \mu']^2$,
(a) for some sparse prior choices of (μ, α) in (8.48),
(b) for some choices of sparsity and overshoot (α, a) (so that μ is determined by (8.50)).

8.18 *(Lower bound in Theorem 8.20, sparse case)* Adopt the setting of Section 8.8. Suppose that $\eta_n \to 0$ and that $k_n \to \infty$. Let $\gamma < 1$ be given, and build π_n from n i.i.d draws (scaled by ϵ_n) from the univariate sparse prior $\pi_{\gamma\eta_n}$ with sparsity $\gamma\eta_n$ and overshoot $(2\log(\gamma\eta_n)^{-1})^{1/4}$, compare Section 8.5. Show that

(a) The number N_n of non-zero components in a draw from π_n is distributed as Binomial$(n, \gamma\eta_n)$, and hence that $\pi_n(\Theta_n) \to 1$ if and only if $k_n \to \infty$,
(b) on Θ_n, we have $\|\theta\|^2 \leq \gamma^{-1}\epsilon_n^2\mu_n^2 E N_n$ `(define` μ_n`)`, and
(c) for all y, show that $\|\hat{\theta}_{v_n}\|^2 \leq \gamma^{-1}\epsilon_n^2\mu_n^2 E N_n$.

As a result, verify that the sequence π_n satisfies conditions (8.71) – (8.74), and hence that $R_N(\Theta_n(k_n), \epsilon_n) \geq B_n(k_n, \epsilon_n)(1 + o(1))$.

9

Sparsity, adaptivity and wavelet thresholding

> The guiding motto in the life of every natural philosopher should be, "Seek simplicity and distrust it." (*The Concept of Nature*, Alfred North Whithead)

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 then draw consequences for adaptive minimax estimation, first in the single sequence model, and then in multiresolution settings. The simplicity lies in the sparsity of representation and the distrust in the quantification of error.

In Section 9.1, traditional linear approximation is contrasted with a version of non-linear approximation that greedily picks off the largest coefficients in turn. Then a more explicitly statistical point of view relates the size of *ideal risk* to the non-linear approximation error. Thirdly, we look at the decay of individual ordered coefficients: this is expressed in terms of a weak ℓ_p condition. The intuitively natural connections between these viewpoints can be formalized as an equivalence of (quasi-)norms in Section 9.2.

Consequences for estimation now flow quite directly. Section 9.3 gives a lower bound for minimax risk using hypercubes, and the oracle inequalities of the last chapter in terms of ideal risk combined with the quasi-norm equivalences lead to upper bounds for $\sqrt{2\log n}$ thresholding over weak ℓ_p balls that are only a logarithmic factor worse than the hypercube lower bounds. When $p < 2$, these are algebraically better rates than can be achieved by any linear estimator–this is seen in Section 9.5 using some geometric ideas from Section 4.8.

Up to this point, the discussion applies to any orthonormal basis. To interpret and extend these results in the setting of function estimation we need to relate sparsity ideas to smoothness classes of functions, and it is here that wavelet bases play a role.

The fundamental idea may be expressed as follows. 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 reflect the average rather than worst case smoothness.

Thus, a key idea is the degree of uniformity of smoothness that is assumed, and this is measured in an L_p sense. Section 9.6 introduces this topic in more detail by comparing three examples, namely uniform ($p = \infty$), mean-squre ($p = 2$) and average ($p = 1$) smoothness conditions, and then working up to the definition of Besov classes as a systematic framework covering all the cases.

Focusing on the unit interval $[0, 1]$, it turns out that many Besov classes of smoothness

α are contained in weak $\ell_{p(\alpha)}$ balls, see Section 9.7. After some definitions for estimation in the continuous Gaussian white noise problem in Section 9.8, the way is paved for earlier results in this chapter to yield, in Section 9.9, broad adaptive near-minimaxity results for $\sqrt{2\log n}$ thresholding over Besov classes.

These results are for integrated mean squared error over all $t \in [0, 1]$; Section 9.10 shows that the same estimator, and similar proof ideas, lead to rate of convergence results for estimating $f(t_0)$ at a single point t_0.

The final Section 9.11 gives an overview of the topics to be addressed in the second part of the book.

In this chapter, in order to quantify sparsity and smoothness, we need two conventional weakenings of the notion of a norm on a linear space: namely quasi-norms, which satisfy a weakened triangle inequality, and semi-norms, which are not necessarily positive definite. The formal definitions are recalled in Appendix C.1.

9.1 Approximation, Ideal Risk and Weak ℓ_p Balls

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 $K \subset \mathbb{N}$:

$$P_K f = \sum_{i \in K} \theta_i \psi_i .$$

(In particular, $P_\varnothing f = 0$.) 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

$$\| f - P_K f \|_2^2 = \sum_{i \notin K} \theta_i^2 .$$

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 a subset K of size k: we have

$$c_k^2(f) = \inf \left\{ \| f - P_K f \|_2^2 : \#(K) \le k \right\},$$

or what is the same

$$c_k^2(\theta) = \inf \left\{ \sum_{i \notin K} \theta_i^2 : \#(K) \le k \right\}. \tag{9.1}$$

Let $|\theta|_{(1)} \ge |\theta|_{(2)} \ge \ldots$ denote the amplitudes of θ in decreasing order. Then $c_k^2(f)$ is what remains after choosing the k largest coefficients, and so

$$c_k^2(f) = c_k^2(\theta) = \sum_{l > k} |\theta|_{(l)}^2,$$

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 this set

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

$$\|\theta\|_{c,\alpha}^2 = \sup_{k \geq 0} k^{2\alpha} \sum_{l>k} |\theta|_{(l)}^2, \tag{9.2}$$

with the convention that $k^{2\alpha} = 1$ when $k = 0$. The subscript 'c' is mnemonic for 'compression'. In other words, $\|\theta\|_{c,\alpha} = C$ means that $(\sum_{l>k} |\theta|_{(i)}^2)^{1/2} \leq C k^{-\alpha}$ for all k and that C is the smallest constant with this property. Exercise 9.2 shows the quasi-triangle inequality.

So far, the index set has been $\mathbb{N}$. The expression (9.1) for $c_k^2(\theta)$ is well defined for any finite or countable index set I, and hence so is $\|\theta\|_{c,\alpha}^2$, if the supremum is taken over $k = 0, 1, \ldots, |I|$.

Ideal Risk

Return to estimation in a Gaussian white sequence model,

$$y_i = \theta_i + \epsilon z_i, \qquad i \in I,$$

thought of, as usual, as the coefficients of the continuous Gaussian white noise model (1.21) in the orthonormal basis $\{\psi_i\}$.

Suppose that $K \subset I$ indexes a finite subset of the variables and that P_K is the corresponding orthogonal projection. The variance-bias decomposition of MSE is given by

$$E\|P_K y - f\|^2 = \#(K)\epsilon^2 + \|P_K f - f\|^2,$$

compare (2.47). The subset minimizing MSE depends on f; to characterize this 'ideal' subset and its associated ideal risk it is again helpful to organize the minimization by size of subset:

$$\begin{aligned} \mathcal{R}(f,\epsilon) &:= \inf_{K \subset \mathbb{N}} E\|P_K y - f\|^2 && (9.3)\\ &= \inf_{k \geq 0} \Big\{ k\epsilon^2 + \inf_{K:\#(K)=k} \|P_K f - f\|^2 \Big\} && (9.4)\\ &= \inf_{k \geq 0} \big\{ k\epsilon^2 + c_k^2(\theta) \big\}. && (9.5) \end{aligned}$$

The second and third forms show an important connection between ideal estimation and non-linear approximation. They hint at the manner in which approximation theoretic results have a direct implication for statistical estimation.

Write $S_k = k\epsilon^2 + c_k^2(\theta)$ for the best MSE for model size k. The differences

$$S_k - S_{k-1} = \epsilon^2 - |\theta|_{(k)}^2$$

are increasing with k, and so the largest value of k minimizing $k \to S_k$ occurs as $k \to |\theta|_{(k)}^2$ 'crosses' the level ϵ^2, or more precisely, at the index k given by

$$N(\epsilon) = N(\theta,\epsilon) = \#\{i : |\theta_i| \geq \epsilon\}, \tag{9.6}$$

Compare Figure 9.1. [in approximation theory, this is called the distribution function of $|\theta|$, a usage related to, but not identical with, the standard statistical term.]

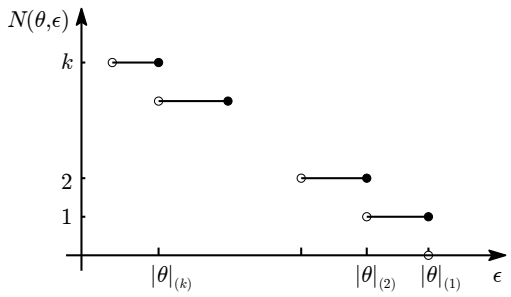

Figure 9.1 $\epsilon \to N(\epsilon)$, the number of coordinates of θ greater than or equal to ϵ in magnitude, is left continuous, with jumps at $|\theta|_{(k)}, k = 1, \ldots, n$.

Thus, (9.5) and (9.6) yield the decomposition

$$\mathcal{R}(\theta, \epsilon) = N(\theta, \epsilon)\epsilon^2 + c^2_{N(\epsilon)}(\theta). \tag{9.7}$$

It is also apparent that, in an orthonormal basis, the ideal subset estimation risk coincides with our earlier notion of ideal risk, Section 8.3:

$$\mathcal{R}(f, \epsilon) = \mathcal{R}(\theta, \epsilon) = \sum \min(\theta_i^2, \epsilon^2).$$

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 $\{i : |\theta_i| \geq \epsilon\}$.

The ideal risk is small precisely when both $N(\epsilon)$ and $c_{N(\epsilon)}$ are. This has the following interpretation: suppose that $N(\theta, \epsilon) = k$ and let $K_k(\theta)$ be the best approximating set of size k. Then the ideal risk consists of a variance term $k\epsilon^2$ 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 $K_k(\theta) = \{i : |\theta_i| > \epsilon\}$, the bias term is as small as it can be for any projection estimator estimating only k coefficients.

The rate of decay of $\mathcal{R}(\theta, \epsilon)$ with ϵ measures the rate of estimation of θ (or $f[\theta]$) using the ideal projection estimator for the given basis. Again to quantify this, we define a second sequence quasi-norm, for $0 < r < 1$, by

$$\|\theta\|_{IR,r}^{2(1-r)} = \sup_{\epsilon > 0} \epsilon^{-2r} \sum_i \min(\theta_i^2, \epsilon^2), \tag{9.8}$$

where 'IR' is mnemonic for 'ideal risk'. In other words, $\|\theta\|_{IR,r} = B$ guarantees that $\mathcal{R}(\theta, \epsilon) \leq B^{2(1-r)}\epsilon^{2r}$ for all $\epsilon > 0$, and that B is the smallest constant for which this is true. The exponent in $\|\theta\|_{IR,r}^{(1-r)}$ ensures the scale invariance $\|c\theta\|_{IR,r} = |c|\|\theta\|_{IR,r}$ and further that it is a quasi-norm, Exercise 9.2(c).

Identity (9.7) says that good estimation is possible precisely when θ compresses well in basis $\{\psi_i\}$, in the sense that both the number of large coefficients $N(\epsilon)$ and the compression number $c^2_{N(\epsilon)}$ are small. Proposition 9.1 below uses (9.7) to show that the compression number and ideal risk sequence quasi-norms are equivalent.

Weak ℓ_p and Coefficient decay

A further natural measure of the "compressibility" of θ is the rate at which the *individual* magnitudes $|\theta_i|$ decay. More formally, we say that $\theta = (\theta_i, i \in I) \in w\ell_p$, if the *decreasing rearrangement* $|\theta|_{(1)} \geq |\theta|_{(2)} \geq \ldots$ satisfies, for some C and all $l = 1, \ldots, |I|$,

$$|\theta|_{(l)} \leq C l^{-1/p},$$

and we set $\|\theta\|_{w\ell_p}$ equal to the smallest such C. Thus

$$\|\theta\|_{w\ell_p} = \max_k k^{1/p} |\theta|_{(k)}.$$

Here $\|\theta\|_{w\ell_p}$ is a quasi-norm, since instead of the triangle inequality, it satisfies only

$$\|\theta + \theta'\|^p_{w\ell_p} \leq 2^p (\|\theta\|^p_{w\ell_p} + \|\theta'\|^p_{w\ell_p}), \qquad (p > 0). \tag{9.9}$$

See 3° below for the proof, and also Exercise 9.1. We write $w\ell_p(C)$ for the (quasi-)norm ball of radius C, or $w\ell_{n,p}(C)$ if we wish to emphasize that $I = \{1, \ldots, n\}$.

Smaller values of p correspond to faster decay for the components of θ. We will be especially interested in cases where $p < 1$, since these correspond to the greatest sparsity.

We note some relations satisfied by $w\ell_p(C)$.

1°. $\ell_p(C) \subset w\ell_p(C)$. This follows from

$$[k^{1/p} |\theta|_{(k)}]^p \leq k \cdot (1/k) \sum_{l=1}^{k} |\theta|^p_{(l)} \leq \|\theta\|^p_{\ell_p}.$$

2°. $w\ell_p \subset \ell_{p'}$ for all $p' > p$, since if $\theta \in w\ell_p$, then

$$\sum_1^\infty |\theta|^{p'}_{(k)} \leq C^{p'} \sum_1^\infty k^{-p'/p} = C^{p'} \zeta(p'/p),$$

where $\zeta(s) = \sum_1^\infty k^{-s}$ is Riemann's zeta function.

3°. A plot of $N(\theta, \epsilon)$ versus ϵ, Figure 9.1, shows that the maximum of $\epsilon \to \epsilon^p N(\theta, \epsilon)$ may be found among the values $\epsilon = |\theta|_{(k)}$. Hence we obtain

$$\|\theta\|^p_{w\ell_p} = \sup_{\epsilon > 0} \epsilon^p N(\theta, \epsilon). \tag{9.10}$$

This representation makes it easy to establish the quasi-norm property. Indeed, since

$$N(\theta + \theta', \epsilon) \leq N(\theta, \epsilon/2) + N(\theta', \epsilon/2),$$

we obtain (9.9) immediately. Let $N(\Theta, \epsilon) = \sup_{\theta \in \Theta} N(\theta, \epsilon)$. Equation (9.10) also yields the implication

$$\epsilon^p N(\Theta, \epsilon) \leq C^p \text{ for all } \epsilon \implies \Theta \subset w\ell_p(C). \tag{9.11}$$

A note: Figure 9.1 shows that in (9.10), the supremum can be restricted to $0 < \epsilon \leq |\theta|_{(1)}$. Then the previous display can be modified as follows. If $\|\theta\|_\infty \leq a$ on Θ, then

$$\epsilon^p N(\Theta, \epsilon) \leq C^p \text{ for } 0 < \epsilon \leq a \implies \Theta \subset w\ell_p(C). \tag{9.12}$$

9.2 Quasi-norm equivalences

In preceding subsections, we have defined three quantitative measures of the sparseness of a coefficient vector θ.

(a) $\|\theta\|_{c,\alpha}$ as a measure of the rate α of non-linear ℓ_2 approximation of θ using a given number of coefficients,
(b) $\|\theta\|_{IR,r}$ 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) $\|\theta\|_{w\ell_p}$ as a measure of the rate $1/p$ of decay of the individual coefficients $|\theta|_{(l)}$.

We now show that these measures are actually equivalent, if one makes the calibrations:

$$r = 2\alpha/(2\alpha+1), \qquad p = 2/(2\alpha+1), \qquad \Longrightarrow \qquad p = 2(1-r). \tag{9.13}$$

Proposition 9.1 *Let $\alpha > 0$, and suppose that $r = r(\alpha)$ and $p = p(\alpha)$ are given by* (9.13). *Then, with $c_p = [2/(2-p)]^{1/p}$,*

$$3^{-1/p}\|\theta\|_{w\ell_p} \le \|\theta\|_{c,\alpha} \le \|\theta\|_{IR,r} \le c_p\|\theta\|_{w\ell_p}. \tag{9.14}$$

Proof The proof goes from right to left in (9.14). Since all the measures depend only on the absolute values of (θ_i), by rearrangement we may suppose without loss of generality that θ is positive and decreasing, so that $\theta_k = |\theta|_{(k)}$.

1°. Suppose first that $C = \|\theta\|_{w\ell_p}$, so that $\theta_k \le Ck^{-1/p}$. Hence

$$\begin{aligned}\sum_k \min(\theta_k^2, t^2) &\le \sum_{k=1}^{\infty} \min(C^2k^{-2/p}, t^2) \le \int_0^{\infty} (Cu^{-1/p})^2 \wedge t^2\, du \\ &= u_* t^2 + \frac{p}{2-p} C^2 u_*^{1-2/p} = \left(1 + \frac{p}{2-p}\right) C^p t^{2r}.\end{aligned}$$

Here $u_* = C^p t^{-p}$ is the point of balance in the pairwise minimum. Hence

$$\|\theta\|_{IR}^{2(1-r)} = \sup_{t\ge 0} t^{-2r} \sum \min(\theta_k^2, t^2) \le [2/(2-p)]\|\theta\|_{w\ell_p}^p.$$

2°. Now let $C = \|\theta\|_{IR,r}$, so that for all positive t, we have $t^{-2r}\sum \min(\theta_k^2, t^2) \le C^{2(1-r)}$. In particular, when $t = \theta_k$, we obtain, for all $k \ge 1$,

$$\theta_k^{-2r}[k\theta_k^2 + c_k^2(\theta)] \le C^{2(1-r)}.$$

Hence $\theta_k \le k^{-1/p}C$ and so

$$c_k^2(\theta) \le \theta_k^{2r} C^{2-2r} \le k^{-2r/p} C^2.$$

Since $2r/p = 2\alpha$, we conclude for every $k \ge 1$, that $k^{2\alpha}c_k^2(\theta) \le C^2 = \|\theta\|_{IR}^2$. It remains to consider the exceptional case $k = 0$: putting $t = \theta_1$ in the definition of $\|\theta\|_{IR,r}$, we find $c_0^2(\theta) \le C^{2(1-r)}\theta_1^{2r}$ and also that $\theta_1^2 \le C^{2(1-r)}\theta_1^{2r}$. Hence $\theta_1^p \le C^p$ and so $c_0^2(\theta) \le C^2$, which completes the verification.

3°. Let $C = \|\theta\|_{c,\alpha}$, so that $c_k^2(\theta) \le C^2k^{-2\alpha}$ for $k \ge 1$ and $c_0^2(\theta) \le C^2$. This implies that $\theta_1^2 \le C^2$, and for $k \ge 2$ and $1 \le r < k$ that

$$\theta_k^2 \le r^{-1} \sum_{k-r+1}^{k} \theta_j^2 \le r^{-1} c_{k-r}^2(\theta) \le C^2/[r(k-r)^{2\alpha}] \le C^2(3/k)^{1+2\alpha},$$

where for the last inequality we set $r = [k/2] \geq k/3$. Consequently, for all $k \geq 1$,

$$\|\theta\|^2_{w\ell_p} = \sup_k k^{2/p}\theta_k^2 \leq 3^{2/p}C^2. \qquad \square$$

9.3 A Risk Lower Bound via Embedding of hypercubes.

We have just seen that $N(\theta, \epsilon)$, the number of coefficients with modulus larger than ϵ, 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 Θ under squared error loss.

Suppose Θ is solid and orthosymmetric, as defined in Section 4.8. For each $\theta \in \Theta$ and $\epsilon > 0$, the definition shows that Θ contains a hypercube $\Theta(\epsilon)$ with center 0, side length 2ϵ and dimension $N(\theta, \epsilon)$. The *ϵ-hypercube dimension*

$$N(\Theta, \epsilon) := \sup_{\theta\in\Theta} N(\theta, \epsilon) \tag{9.15}$$

denotes the maximal dimension of a zero-centered ϵ-hypercube embedded in Θ.

In the white Gaussian sequence model at noise level ϵ, the minimax risk for a p-dimensional ϵ-hypercube $[-\epsilon, \epsilon]^p$ is given, from (4.47) and (4.35) by

$$R_N([-\epsilon, \epsilon]^p, \epsilon) = p\epsilon^2\rho_N(1, 1),$$

where $c_0 = \rho_N(1, 1)$ is the minimax risk in the unit noise univariate bounded normal mean problem on $[-1, 1]$. Since Θ contains the hypercube $\Theta(\epsilon)$, we arrive at a lower bound for the minimax risk

$$R_N(\Theta, \epsilon) \geq c_0\epsilon^2 N(\Theta, \epsilon). \tag{9.16}$$

Examples. 1. ℓ_p balls. In Chapters 11 and 13, we study at length estimation over

$$\ell_{n,p}(C) = \Theta_{n,p}(C) = \{\theta \in \mathbb{R}^n : \sum_1^n |\theta_i|^p \leq C^p\}. \tag{9.17}$$

We clearly have $\#\{i : |\theta_i| \geq \epsilon\} \leq \sum_1^n |\theta_i|^p/\epsilon^p$, and so the ϵ-hypercube dimension

$$N(\ell_{n,p}(C), \epsilon) = \min(n, [C^p/\epsilon^p]). \tag{9.18}$$

Hence, if $C > \epsilon$, we find from this and (9.16) that

$$R_N(\ell_{n,p}(C), \epsilon) \geq c_1 \min(n\epsilon^2, C^p\epsilon^{2-p}), \tag{9.19}$$

where $c_1 = c_0/2$. Since $\ell_{n,p}(C) \subset w\ell_{n,p}(C)$, the same lower bound applies also to the weak ℓ_p ball.

2. Products. Since $N((\theta_1, \theta_2), \epsilon) = N(\theta_1, \epsilon) + N(\theta_2, \epsilon)$, we have

$$N(\Theta_1 \times \Theta_2, \epsilon) = N(\Theta_1, \epsilon) + N(\Theta_2, \epsilon). \tag{9.20}$$

3. Sobolev ellipsoids. Suppose, following Section 3.1, that $\Theta = \Theta_2^\alpha(C)$ is the ellipsoid $\{\theta : \sum_1^\infty a_k^2\theta_k^2 \leq C^2\}$ with $a_k = k^\alpha$. Since a_k is increasing with k, the hypercube $[-\epsilon, \epsilon]^p$ is contained in Θ if $\epsilon^2 \sum_1^p k^{2\alpha} \leq C^2$. Thus we may bound $N_\epsilon = N(\Theta, \epsilon)$ from the equation

$\epsilon^2 \sum_1^{N_\epsilon} k^{2\alpha} = C^2$. This was done carefully in Proposition 4.23 (our white noise case here corresponds to $\beta = 0$ there), and with $r = 2\alpha/(2\alpha + 1)$ led to the conclusion

$$R_N(\Theta_2^\alpha, \epsilon) \geq c(\alpha) C^{2(1-r)} \epsilon^{2r}.$$

If we only seek a lower bound on rates of convergence, this is certainly simpler than the more refined arguments of Chapters 4.8 and 5.

Remark. In (9.15), we chose the side ϵ of the hypercube equal to the noise level ϵ. More generally, if the side length τ of the hypercube is arbitrary, then

$$R_N(\Theta, \epsilon) \geq \epsilon^2 N(\Theta, \tau) \rho_N(\tau/\epsilon, 1),$$

which can, if desired, be optimized over τ to yield a sharper bound than (9.16): see, for example, Section 11.4.

9.4 Near Adaptive Minimaxity for weak ℓ_p balls

We are now ready to combine upper and lower bounds to arrive at an adaptive minimaxity result, up to logarithmic terms, for $\sqrt{2 \log n}$ thresholding on ℓ_p balls, both strong and weak. More precise results will be given in later chapters, but the charm of this version lies in the relatively simple proof given the tools we have developed.

To state the result, introduce a "control function" $r^\circ_{n,p}(C, \epsilon) = \min(n\epsilon^2, C^p \epsilon^{2-p})$, and the constant $\gamma_p = 2/(2-p)$.

Theorem 9.2 *Suppose that $y \sim N_n(\theta, \epsilon^2 I)$. Then for $0 < p < 2$ and all $\epsilon < C$,*

$$\begin{aligned} c_1 r^\circ_{n,p}(C, \epsilon) \leq R_N(\ell_{n,p}(C), \epsilon) &\leq R_N(w\ell_{n,p}(C), \epsilon) \\ &\leq (2 \log n + 1)[\epsilon^2 + \gamma_p r^\circ_{n,p}(C, \epsilon)]. \end{aligned}$$

The final bound is attained for all ϵ *by $\hat\theta^U$, soft thresholding at $\epsilon\sqrt{2 \log n}$.*

Proof The first inequality is the hypercube bound (9.19), and the second follows from $\ell_{n,p}(C) \subset w\ell_{n,p}(C)$. It remains to assemble the upper bound for $\hat\theta^U$. From the soft thresholding oracle inequality, Proposition 8.8, we have

$$r_\epsilon(\hat\theta^U, \theta) \leq (2 \log n + 1)[\epsilon^2 + \mathcal{R}(\theta, \epsilon)]. \tag{9.21}$$

Define r through $p = 2(1 - r)$. Then Proposition 9.1 provides the bound

$$\mathcal{R}(\theta, \epsilon) \leq \|\theta\|^{2(1-r)}_{IR,r} \epsilon^{2r} \leq \gamma_p \|\theta\|^p_{w\ell_p} \epsilon^{2-p}. \tag{9.22}$$

Of course, in addition $\mathcal{R}(\theta, \epsilon) = \sum_1^n \min(\theta_i^2, \epsilon^2) \leq n\epsilon^2$, and so we are done:

$$\sup_{\theta \in w\ell_{n,p}(C)} r_\epsilon(\hat\theta^U, \theta) \leq (2 \log n + 1)[\epsilon^2 + \gamma_p \min(n\epsilon^2, C^p \epsilon^{2-p})]. \tag{9.23}$$

□

The minimax risks depend on parameters p, C and ϵ, whereas the threshold estimator $\hat\theta^U$ requires knowledge only of the noise level ϵ—which, if unknown, can be estimated as described in Chapter 7.5. Nevertheless, estimator $\hat\theta^U$ comes within a logarithmic factor of the minimax risk over a wide range of values for p and C. In the next section, we shall see how much of an improvement over linear estimators this represents.

The upper bound in Theorem 9.2 can be written, for $\epsilon < C$ and $n \geq 2$, as

$$c_2 \log n \cdot r^{\circ}_{n,p}(C, \epsilon)$$

if one is not too concerned about the explicit value for c_2. Theorem 11.6 gives upper and lower bounds that differ by constants rather than logarithmic terms.

Exercise 9.4 extends the weak ℓ_p risk bound (9.23) to general thresholds λ.

A variant for strong ℓ_p balls

The inequalities for ideal risk developed via Proposition 9.1 are based on comparisons of $\mathcal{R}(\theta, \epsilon)$ to ϵ^{2r} over the entire range of ϵ, compare the definition in (9.8).

We now develop a related bound by thinking of ϵ as fixed and allowing the ball radius C to grow. A helpful byproduct is an explicit description of the least favorable vectors in different ranges of C.

Lemma 9.3 *Let $0 < p < 2$, $\gamma = C/\epsilon$ and $[\cdot]$ and $\{\cdot\}$ be integer and fractional part. Then*

$$\sup_{\|\theta\|_p \leq C} \sum_{i=1}^{n} \min(\theta_i^2, \epsilon^2) = \begin{cases} C^2 & \text{if } \gamma \leq 1 \\ \epsilon^2\big([\gamma^p] + \{\gamma^p\}^{2/p}\big) & \text{if } 1 \leq \gamma \leq n^{1/p} \\ n\epsilon^2 & \text{if } \gamma > n^{1/p}. \end{cases} \tag{9.24}$$

$$= \epsilon^2 \min\{R(\gamma^p), n\}, \tag{9.25}$$

where $R(t) = [t] + \{t\}^{2/p}$. The least favorable configurations are given, from top to bottom above, by permutations and sign changes of

$$(C, 0, \ldots, 0), \qquad (\epsilon, \ldots, \epsilon, \mu\epsilon, 0, \ldots, 0) \qquad \text{and} \quad (\epsilon, \ldots, \epsilon).$$

In the middle vector, there are $[\gamma^p]$ coordinates equal to ϵ, and $\mu < 1$ is given by $\mu^p = \{\gamma^p\} = \gamma^p - [\gamma^p]$.

Proof First observe that we may rewrite the left side of (9.24) as

$$\sup\left\{\sum_{1}^{n} \theta_i^2 \;:\; \theta \in \ell_{n,p}(C) \cap \ell_{n,\infty}(\epsilon)\right\}. \tag{9.26}$$

If $C \leq \epsilon$, then the ℓ_p ball is entirely contained in the ℓ_∞ cube of side ϵ, and the maximum of $\sum \theta_i^2$ over the ℓ_p ball is attained at the spike $\theta^* = C(1, 0, \ldots, 0)$ or permutations. This yields the first bound in (9.24). At the other extreme, if $C \geq n^{1/p}\epsilon$, then the ℓ_∞ cube is contained entirely within the ℓ_p ball and the maximum of $\sum \theta_i^2$ is attained at the dense configuration $\theta^* = \epsilon(1, \ldots, 1)$.

If $\epsilon < C < n^{1/p}\epsilon$, the worst case vectors are subject to the ℓ_∞ constraint and are then permutations of the vector $\theta^* = (\epsilon, \ldots, \epsilon, \mu\epsilon, 0, \ldots, 0)$ with n_0 components of size ϵ and the remainder μ determined by the ℓ_p condition:

$$n_0\epsilon^p + \mu^p\epsilon^p = C^p.$$

To verify that this is indeed the worst case configuration, change variables to $u_i = \theta_i^p$ in (9.26): the problem is then to maximize the convex function $u \to \sum u_i^{2/p}$ subject to the

convex constraints $\|u\|_1 \le C^p$ and $\|u\|_\infty \le \epsilon^p$. This forces an extremal solution to occur on the boundary of the constraint set and to have the form described.

Thus $n_0 = [C^p/\epsilon^p]$ and $\mu^p = \{C^p/\epsilon^p\}$. Setting $\gamma^p = C^p/\epsilon^p$, we obtain

$$\begin{aligned}\sum \min(\theta_i^2, \epsilon^2) &= n_0\epsilon^2 + \mu^2\epsilon^2 \\ &= \epsilon^2[\gamma^p] + \epsilon^2\{\gamma^p\}^{2/p}.\end{aligned}$$

□

A simpler, if slightly weaker, version of (9.24) is obtained by noting that the first two rows of the right side are bounded by $\epsilon^2\gamma^p = C^p\epsilon^{2-p}$, so that for all $C > 0$

$$\sup_{\|\theta\|_p \le C} \mathcal{R}(\theta, \epsilon) \le \min(n\epsilon^2, C^p\epsilon^{2-p}) = r^\circ_{n,p}(C, \epsilon). \tag{9.27}$$

An immediate corollary is a sharper version of Theorem 9.2 which, due to the restriction to strong ℓ_p balls, comes without the constant $\gamma_p = 2/(2-p)$.

Corollary 9.4 *For $0 < p < 2$,*

$$R_N(\ell_{n,p}(C), \epsilon) \le (2\log n + 1)[\epsilon^2 + r^\circ_{n,p}(C, \epsilon)].$$

Proof Simply insert the new bound (9.27) into the oracle inequality (9.21). □

9.5 The woes of linear estimators.

We make some remarks about the maximum risk of linear estimators. While the techniques used are those of Section 4.8, the statistical implications are clearer now that we have established some properties of non-linear thresholding.

For any set $\Theta \subset \ell_2(I)$, we recall the notation for the "square" of Θ, namely $\Theta_+^2 = \{(\theta_i^2) : \theta \in \Theta\}$. The *quadratically convex hull* of Θ is then defined as

$$\mathrm{QHull}(\Theta) = \{\theta : (\theta_i^2) \in \mathrm{Hull}(\Theta_+^2)\}, \tag{9.28}$$

where $\mathrm{Hull}(S)$ denotes the closed convex hull of S. Of course, if Θ is closed orthosymmetric and quadratically convex, then $\mathrm{QHull}(\Theta) = \Theta$. However, for ℓ_p-bodies $\Theta_p(a, C) = \{\theta : \sum_i a_i^p|\theta_i|^p \le C^p\}$ with $p < 2$,

$$\mathrm{QHull}(\Theta_p(a, C)) = \{\theta : \sum a_i^2\theta_i^2 \le C^2\}$$

is an *ellipsoid*. (Checking this is a good exercise in the definitions.) The key property of quadratic convexification is that it preserves the maximum risk of *linear* estimators.

Theorem 9.5 *Let Θ be solid orthosymmetric and compact. Then*

$$R_L(\Theta, \epsilon) = R_L(\mathrm{QHull}(\Theta), \epsilon).$$

Proof Since Θ is orthosymmetric, (4.56) shows that linear minimax estimators may be found that are diagonal, with risk functions given by (4.50). Such risk functions are linear in $s = (\theta_i^2)$ and hence have the same maximum over $\mathrm{Hull}(\Theta_+^2)$ as over Θ_+^2. □

Remark. Combining Theorems 4.25 and 9.5 , we observe that the minimax linear risk of Θ is still determined by the hardest rectangular subproblem, but now of the *enlarged* set $\mathrm{QHull}(\Theta)$. Of course, $\mathrm{QHull}(\Theta)$ may be much larger that Θ, and so (in contrast to Corollary

4.26) it could certainly happen now that $R_L(\Theta) \gg R_N(\Theta)$: we will see examples in the later discussion of ℓ_p balls and Besov spaces.

For a key example, let $p < 2$ and consider the ℓ_p ball $\ell_{n,p}(C) = \Theta_{n,p}(C)$ of (9.17). Since $\mathrm{QHull}(\Theta_{n,p}(C)) = \Theta_{n,2}(C)$, the previous theorem, along with (4.59) and a constant $\gamma \in [1/2, 1]$ yields

$$\begin{aligned} R_L(\ell_{n,p}(C), \epsilon) = R_L(\Theta_{n,2}(C), \epsilon) &= \frac{n\epsilon^2 C^2}{n\epsilon^2 + C^2} \\ &= \gamma \min(n\epsilon^2, C^2) = \gamma r^\circ_{n,2}(C, \epsilon). \end{aligned} \tag{9.29}$$

Combining this with Theorem 9.2 or Corollary 9.4, which we may do simply by contrasting $r^\circ_{n,p}$ with $r^\circ_{n,2}$, we see that $C^p \epsilon^{2-p} \ll C^2$ exactly when $\epsilon \ll C$. Hence for $p < 2$, the non-linear minimax risk is an algebraic order of magnitude smaller than the linear minimax risk. Furthermore, $\sqrt{2 \log n}$ thresholding captures almost all of this gain, giving up only a factor logarithmic in n.

9.6 Function spaces and wavelet coefficients

To draw consequences of these results for function estimation, we need to relate sparsity ideas to smoothness classes of functions. We have seen when smoothness of functions is measured in, say, a mean-square sense—corresponding to L_2 integrals $\int (D^\alpha f)^2$—that linear estimators are close to optimal for mean-square error. [Recall, for example, Corollary 4.26 and Lemma 3.3.] On the other hand, it is apparent that non-linear estimators, for example using thresholding of wavelet coefficients, can greatly outperform linear estimators. In order to have a mathematical framework to describe this, we measure smoothness using other L_p measures, typically for $p < 2$ when the estimation error is measured in mean-square. It might at first seem simplest, then, to consider L_p integrals of derivatives $\int |D^\alpha f|^p$, the Sobolev (semi-)norms. However, when working with wavelet bases $\{\psi_{jk}\}$, it turns out to be helpful to have the flexibility to sum separately over location k with an ℓ_p index and over scale j with an ℓ_q index. For this purpose it has proved helpful to formulate the notion of smoothness using Besov spaces.

This section gives some motivation for the definition of Besov measures of smoothness of functions. More systematic discussion can be found in the books by Meyer (1990), Frazier et al. (1991) and Triebel (1983). Instead the approach here is

- first, to give some heuristic remarks on L_p measures of smoothness and the tradeoff between worst-case, $p = \infty$, and average case, $p = 1$, measures,
- then, to explore the use of magnitudes of wavelet coefficients to describe smoothness of functions in examples with $p = 1, 2$ and ∞, and
- then to give a definition of Besov norms on sequences of wavelet coefficients that encompasses the three examples,
- and finally to introduce one definition of Besov norm on functions and to indicate the equivalence to the sequence norm definition.

This approach is somewhat roundabout, in that we do not begin directly with Besov smoothness measures on functions. There are two reasons for this. The first is pragmatic: it

is the sequence form that is most heavily used for the statistical theory. The second is to simplify exposition—while the rich theory of Besov spaces $B^\alpha_{p,q}(\Omega)$ on domains and $B^\alpha_{p,q}(\mathbb{R}^n)$ on Euclidean space can be approached in various, largely equivalent, ways, it does take some work to establish equivalence with the sequence form in terms of wavelet coefficients. To keep the treatment relatively self-contained, Appendix B gives the definition of $B^\alpha_{p,q}([0,1])$ in terms of moduli of smoothness and shows the equivalence with the sequence form using classical ideas from approximation theory.

Some Heuristics

Some traditional measures of smoothness use L_p norms to measure the size of derivatives of the function. Hence, for functions f for which $D^{k-1}f$ is absolutely continuous, define the *semi-norm*

$$|f|_{W^k_p} = \left(\int |D^k f|^p \right)^{1/p}, \qquad 1 \le p \le \infty.$$

When $p = \infty$, the integral is replaced by a supremum $\sup_x |D^k f(x)|$. These semi-norms vanish on polynomials of degree less than k, and it is customary to add the L_p norm of the function to obtain an actual norm. Thus (the p^{th} power of) the Sobolev norm is defined by

$$\|f\|^p_{W^k_p} = \int |f|^p + \int |D^k f|^p.$$

The Sobolev space W^k_p of functions with k derivatives existing a.e. 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 seminorm is modified to yield the *Hölder* norms

$$\|f\|_{C^k} = \|f\|_\infty + \|D^k f\|_\infty.$$

Figure 9.2 shows two examples of how smaller p corresponds to a more averaged and less worst-case measure of smoothness. For the function in the first panel,

$$\|f'\|_1 = 2, \qquad \|f'\|_2 = \sqrt{1/a + 1/b}, \qquad \|f'\|_\infty = 1/a.$$

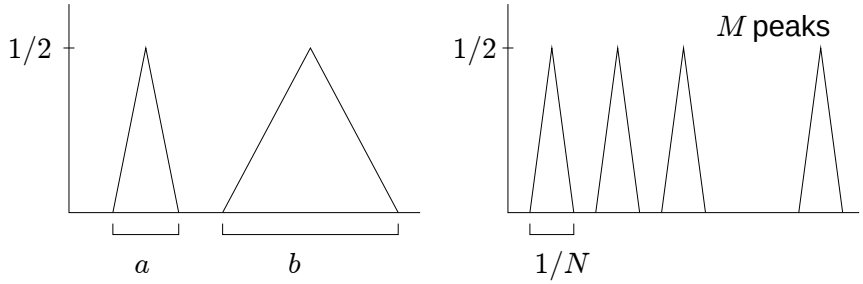

Figure 9.2 Two piecewise differentiable functions f for comparison of different L_p measures of smoothness for f'.

In the 1−norm the peaks have equal weight, while in the 2−norm the narrower peak dominates, and finally in the ∞−norm, the wider peak has no influence at all. The second panel compares the norms of a function with M peaks each of width $1/N$:

$$\|f'\|_1 = M, \qquad \|f'\|_2 = \sqrt{MN}, \qquad \|f'\|_\infty = N.$$

The 1−norm is proportional to the number of peaks, while the ∞−norm measures the slope of the narrowest peak and so is unaffected by the number of spikes. The 2−norm is a compromise between the two. Thus, again smaller values of p are more forgiving of inhomegeneity. If, as in much of this work, the estimation error is measured as a global average (for example, as in mean integrated squared error), then we should be able to accomodate some degree of such inhomogeneity in smoothness.

Decay of wavelet coefficients–some examples

A basic idea is to use the relative magnitude of wavelet coefficients across scales to describe the smoothness of functions. We explore this in the cases $p = 1$, $p = 2$ and $p = \infty$ before showing how Besov sequence norms provide a unifying framework. To avoid boundary issues, we work with an orthonormal wavelet basis for $L_2(\mathbb{R})$, and so assume that a square integrable function f has expansion

$$f(x) = \sum_k \beta_{Lk}\varphi_{Lk}(x) + \sum_{j\geq L}\sum_k \theta_{jk}\psi_{jk}(x). \tag{9.30}$$

Hölder smoothness, $p = \infty$. We consider only $0 < \alpha < 1$, for which $|f(x) - f(y)| \leq C|x - y|^\alpha$ for all x, y.

Theorem 9.6 *Suppose that $0 < \alpha < 1$ and that (φ, ψ) are C^1 and have compact support. Then $f \in C^\alpha(\mathbb{R})$ if and only if there exists $C > 0$ such that*

$$|\beta_{Lk}| \leq C, \qquad |\theta_{jk}| \leq C2^{-(\alpha+1/2)j}, \quad j \geq L. \tag{9.31}$$

Reflecting the uniformity in x, the conditions on the wavelet coefficients are uniform in k, with the decay condition applying to the scales j.

Proof Assume first that $f \in C^\alpha$, so that, Appendix C.23, $\|f\|_\alpha = \|f\|_\infty + |f|_\alpha < \infty$, where $|f|_\alpha = \sup |f(x) - f(x')|/|x - x'|^\alpha$. For the coarse scale coefficients

$$|\beta_{Lk}| \leq 2^{-L/2}\|f\|_\infty\|\varphi\|_1,$$

as is easily verified. For the wavelet coefficients, although this is a special case of Lemma 7.3, we give the details here. What we rely on is that $\int \psi = 0$—this follows from Proposition 7.4 since ψ is C^1—which allows the wavelet coefficient to be rewritten as

$$\langle f, \psi_{jk}\rangle = 2^{-j/2}\int [f(x_k + 2^{-j}v) - f(x_k)]\psi(v)dv \tag{9.32}$$

for $x_k = k2^{-j}$. The Hölder smoothness now provides the claimed bound

$$|\langle f, \psi_{jk}\rangle| \leq 2^{-j/2}|f|_\alpha 2^{-j\alpha}\int |v|^\alpha\psi(v)dv = c_{\psi,\alpha}|f|_\alpha 2^{-j(\alpha+1/2)}, \tag{9.33}$$

and we conclude the first half by taking $C = \max(2^{-L/2}\|f\|_\infty\|\varphi\|_1, c_{\psi,\alpha}|f|_\alpha)$.

In the reverse direction, we wish to use (9.31) to show that $\|f\|_\infty + |f|_\alpha \le cC$ for a constant c depending only on α and properties of φ and ψ. For $|f|_\alpha$, we use (9.30) to decompose the difference $f(x) - f(x')$ into terms $\Delta_\beta(f) + \Delta_\theta(f)$, where, for example,

$$\Delta_\theta(f) = \sum_{j,k} \theta_{jk}[\psi_{jk}(x) - \psi_{jk}(x')].$$

We focus on $\Delta_\theta(f)$ here, since the argument for $\Delta_\beta(f)$ is similar and easier. Using the decay (9.31) of the coefficients θ_{jk},

$$|\Delta_\theta(f)| \le C \sum_{j \ge L} 2^{-(\alpha+1/2)j} \sum_k 2^{j/2}|\psi(2^j x - k) - \psi(2^j x' - k)|.$$

If the length of the support of ψ is S, then at most $2S$ terms in the sum over k are non-zero. In addition, the difference can be bounded using $\|\psi'\|_\infty$ when $|2^j x - 2^j x'| \le 1$, and using simply $2\|\psi\|_\infty$ otherwise. Hence

$$|\Delta_\theta(f)| \le c_\psi C \sum_{j \ge L} 2^{-\alpha j} \min\{2^j|x - x'|, 1\},$$

where $c_\psi = 2S \max\{2\|\psi\|_\infty, \|\psi'\|\}$. Let $j_* \in \mathbb{R}$ satisfy $2^{-j_*} = |x - x'|$. The summands above increase geometrically for $j < j_*$ (using the assumption that $\alpha < 1$!), and decrease geometrically for $j > j_*$. Consequently

$$|\Delta_\theta(f)| \le c_\alpha c_\psi C 2^{-\alpha j_*} \le c' C |x - x'|^\alpha,$$

which, together with the bound for $\Delta_\beta(f)$ gives the Hölder bound for $|f|_\alpha$ we seek.

The bound for $\|f\|_\infty$ is much easier. Indeed, from (9.30),

$$|f(x)| \le SC\|\varphi\|_\infty + SC\|\psi\|_\infty \sum_j 2^{-\alpha j} \le cC. \qquad \square$$

Remark 9.7 We mention the extension of this result to $\alpha \ge 1$. Let $r = \lceil \alpha \rceil$. Assume that φ and ψ are C^r with compact support, and that ψ has at least r vanishing moments. If $f \in C^\alpha(\mathbb{R})$, then there exists positive C such that inequalities (9.31) hold. Conversely, if $\alpha > 0$ is not an integer, these inequalities imply that $f \in C^\alpha(\mathbb{R})$. The proof of these statements are a fairly straightforward extension the arguments given above (Exercise 9.8.)

When α is an integer, to achieve a characterization, a slight extension of C^α is needed, see Section B.3 in the Appendix for some extra detail.

Remark 9.8 In the preceding proof, we see a pattern that recurs often with multiresolution models: a count or error that is a function of level j increases geometrically up to some critical level j_0 and decreases geometrically above j_0. The total count or error is then determined up to a constant by the value at the critical level. While it is often easier to compute the bound in each case as needed, we give a illustrative statement here. If $\beta, \gamma > 0$, then on defining $r = \gamma/(\beta + \gamma)$, $c_\beta = (1 - 2^{-\beta})^{-1}$ and c_γ similarly, we have

$$\sum_{j \in \mathbb{Z}} \min(\delta 2^{\beta j}, C 2^{-\gamma j}) \le (c_\beta + c_\gamma) C^{1-r} \delta^r. \tag{9.34}$$

The critical level may be taken as $j_0 = [j_*]$, where j_* is the solution to $\delta 2^{\beta j_*} = C 2^{-\gamma j_*}$.

Mean square smoothness, $p = 2$. Already in Chapter 3 we studied smoothness in the mean square sense, with norms $\|f\|_{W_2^r}^2 = \int f^2 + \int (D^r f)^2$. Mean square smoothness also has a very natural expression in terms of wavelet coefficients. Suppose that (φ, ψ) are C^r. Then we may formally differentiate the homogeneous wavelet expansion $f = \sum_{jk} \theta_{jk} \psi_{jk}$ to obtain

$$D^r f(x) = \sum_{jk} 2^{rj} \theta_{jk} \psi_{jk}^{(r)}(x).$$

The system $\{\psi_{jk}^{(r)}\}$ is no longer orthonormal, but it turns out that it is the next best thing, namely a *frame*, meaning that there exist constants C_1, C_2 such that for all $f \in W_2^r$,

$$C_1 \sum_{jk} 2^{2rj} \theta_{jk}^2 \le \Big\| \sum_{jk} 2^{rj} \theta_{jk} \psi_{jk}^{(r)}(x) \Big\|_2^2 \le C_2 \sum_{jk} 2^{2rj} \theta_{jk}^2. \tag{9.35}$$

These remarks render plausible the following result, proved in Appendix B.4.

Theorem 9.9 *If (ϕ, ψ) are C^r with compact support and ψ has $r+1$ vanishing moments, then there exist constants C_1, C_2 such that*

$$C_1 \|f\|_{W_2^r}^2 \le \sum_k \beta_{Lk}^2 + \sum_{j \ge L, k} 2^{2rj} \theta_{jk}^2 \le C_2 \|f\|_{W_2^r}^2. \tag{9.36}$$

Average smoothness, $p = 1$. We consider functions in W_1^1, for which the norm measures smoothness in an L_1 sense: $\|f\|_{W_1^1} = \|f\|_1 + \int |f'|$. This is similar to, but not identical with, the notion of bounded variation of a function, cf. Appendix C.24: if f lies in W_1^1 then $|f|_{TV} = \int |f'|$.

We show that membership in W_1^1 can be *nearly* characterized by ℓ_1-type conditions on wavelet coefficients. To state the result, adopt the notation $\theta_{j\cdot}$ for the coefficients (θ_{jk}) at the jth level and similarly for $\beta_{L\cdot}$ at coarse scale L.

Theorem 9.10 *Suppose that (φ, ψ) are C^1 with compact support. Then there exist constants C_1 and C_2 such that*

$$C_1 \Big[\|\beta_{L\cdot}\|_1 + \sup_{j \ge L} 2^{j/2} \|\theta_{j\cdot}\|_1 \Big] \le \|f\|_{W_1^1} \le C_2 \Big[\|\beta_{L\cdot}\|_1 + \sum_{j \ge L} 2^{j/2} \|\theta_{j\cdot}\|_1 \Big].$$

The same bounds hold for $\|f\|_{TV}$.

Proof Begin with the right hand bound. Observing that $|\psi_{jk}|_{W_1^1} = |\psi_{jk}|_{TV} = 2^{j/2} \|\psi'\|_1$ and applying the triangle inequality to the wavelet expansion (9.30), we get with $|f|_*$ denoting either $|f|_{TV}$ or $|f|_{W_1^1}$

$$|f|_* \le 2^{L/2} \|\varphi'\|_1 \sum_k |\beta_{Lk}| + \|\psi'\|_1 \sum_j 2^{j/2} \sum_k |\theta_{jk}|$$

with a similar expression for $\|f\|_1$, with $\|\psi'\|_1$ and $2^{j/2}$ in the second right side term replaced by $\|\psi\|_1$ and $2^{-j/2}$ and with analagous changes in the first right side term.

For the left hand inequality, we suppose that $f \in W_1^1$. Since ψ is C^1, we conclude as before that $\int \psi = 0$, and it follows from integration by parts that if supp $\psi \subset I$, then

$$\left| \int_I f\psi \right| \le \tfrac{1}{2}\|\psi\|_1 \int_I |Df|$$

Suppose that ψ has support contained in $[-S+1, S]$. Applying the previous bound to wavelet coefficient $\theta_{jk} = \int f\psi_{jk}$ yields a bound $|\theta_{jk}| \le c_\psi 2^{-j/2} \int_{I_{jk}} |Df|$, where the interval $I_{jk} = 2^{-j}[k-S+1, k+S]$. For j fixed, as k varies, any given point x falls in at most $2S$ intervals I_{jk}, and so adding over k yields, for each $j \ge L$,

$$2^{j/2} \sum_k |\theta_{jk}| \le 2S \cdot c_\psi \cdot |f|_{W_1^1}.$$

A similar but easier argument shows that we also have $\|\beta_{L\cdot}\|_1 \le 2^{L/2} \cdot 2S\|\varphi\|_\infty \cdot \|f\|_1$. Adding this to the last display yields the left bound. The extension of this argument to $f \in TV$ is left to Exercise 9.9. □

Besov sequence norms

Comparing the three cases, we may contrast how the coefficients at a given level j are weighted and combined over k:

Hölder:	$p = \infty$,	$2^{(\alpha+1/2)j}\|\theta_{j\cdot}\|_\infty$
Mean square:	$p = 2$,	$2^{\alpha j}\|\theta_{j\cdot}\|_2$
Average:	$p = 1$,	$2^{(\alpha-1/2)j}\|\theta_{j\cdot}\|_1$

[In the last two cases, we are extrapolating from $\alpha = r \in \mathbb{N}$ and $\alpha = 1$ respectively.]

Introducing the index $a = \alpha + 1/2 - 1/p$, we can see each case as a particular instance of a weighted ℓ_p norm $c_j = 2^{aj}\|\theta_{j\cdot}\|_p$. To combine the information in c_j across levels j, we use ℓ_q norms $(\sum_{j\ge L} |c_j|^q)^{1/q}$, which spans a range of measures from worst case, $q = \infty$, to average case, $q = 1$.

We use θ as an abbreviation for $\{\beta_{Lk}\} \cup \{\theta_{jk}, j \ge L, k \in \mathbb{Z}\}$, and define

$$\|\theta\|_{b^\alpha_{p,q}} = \|\beta_{L\cdot}\|_p + \Big(\sum_{j\ge L} 2^{ajq} \|\theta_{j\cdot}\|_p^q \Big)^{1/q}, \tag{9.37}$$

where again, $a = \alpha + 1/2 - 1/p$. In the case $q = \infty$, this is interpreted as

$$\|\theta\|_{b^\alpha_{p,\infty}} = \|\beta_{L\cdot}\|_p + \sup_{j\ge L} 2^{aj}\|\theta_{j\cdot}\|_p.$$

Written out in full indicial glory, (9.37) becomes

$$\|\theta\|_{b^\alpha_{p,q}} = \Big(\sum_k |\beta_{Lk}|^p \Big)^{1/p} + \Big(\sum_{j\ge L} 2^{ajq} \big(\sum_k |\theta_{jk}|^p\big)^{q/p} \Big)^{1/q}.$$

Thus, the three parameters may be interpreted as follows:

$\alpha > 0$	smoothness
$p \in (0, \infty]$	averaging (quasi-)norm over locations k
$q \in (0, \infty]$	averaging (quasi-)norm over scales j.

The notation $\|\theta\|_b \asymp \|f\|_{\mathcal{F}}$ is used for equivalence of norms: it means that there exist constants C_1, C_2, not depending on θ (or f) such that

$$C_1\|\theta\|_b \le \|f\|_{\mathcal{F}} \le C_2\|\theta\|_b.$$

Armed with the Besov index notation, we may summarize the inequalities described in the three function class examples considered earlier as follows:

Hölder smoothness, $p = \infty$.	$\|\theta\|_{b^\alpha_{\infty,\infty}} \asymp \|f\|_{C^\alpha}$,	$\alpha > 0$,
Mean-square smoothness, $p = 2$.	$\|\theta\|^2_{b^\alpha_{2,2}} \asymp \int \|f\|^2 + |D^\alpha f|^2$,	$\alpha \in \mathbb{N}$,
Average smoothness/TV, $p = 1$.	$C_1\|\theta\|_{b^1_{1,\infty}} \le \int \|f\| + |Df| \le C_2\|\theta\|_{b^1_{1,1}}$.	

In the Hölder case, we use the Zygmund class interpretation of C^α when $\alpha \in \mathbb{N}$, Appendix B.11. The average smoothness/TV result corresponds only to $\alpha = 1$.

Example 9.11 Consider $f(x) = A|x|^\beta g(x)$. Here g is just a window function included to make f integrable; for example suppose that g is equal to 1 for $|x| \le 1/2$ and vanishes for $|x| \ge 1$ and is C^∞ overall. Assume that $\beta > -1/p$ so that $f \in L_p$. Suppose that the wavelet ψ has compact support, and $r > \beta + 1/p$ vanishing moments. Then it can be shown (Exercise 9.10) that $\|\theta\|_{b^\alpha_{p,\infty}} \le c_{\alpha\beta p}A < \infty$ whenever $\alpha \le \beta + 1/p$. Thus one can say that f has smoothness of order $\beta + 1/p$ when measured in L_p. Again, smaller p is more forgiving of a local singularity.

Besov function space norms

Our discussion here is brief; see Appendix B for more detail and references. To test if a function $f(x)$ belongs to function space $B^\alpha_{p,q}$, one starts with an integer $r > \alpha$ and the r-th order differences of $\Delta^r_h(f,x)$ of step length h, averaged over x in L_p. The largest such average for $h \le t$ defines the integral modulus of smoothness $\omega_r(f,t)_p$. The function $f \in B^\alpha_{p,\infty}$ if the ratio $\omega_r(f,t)_p/t^\alpha$ is uniformly bounded in $t > 0$. If instead the ratio belongs to $L_q((0,\infty), dt/t)$ then $f \in B^\alpha_{p,q}$. In each case the L_q norm of $\omega_r(f,t)_p/t^\alpha$ defines the semi-norm $|f|_{B^\alpha_{p,q}}$ and then the norm $\|f\|_{B^\alpha_{p,q}} = \|f\|_p + |f|_{B^\alpha_{p,q}}$.

The discussion in Appendix B is tailored to Besov spaces on a finite interval, say $[0,1]$. It is shown there, Theorem B.9, that if (φ, ψ) are a C^r scaling function and wavelet of compact support giving rise to an orthonormal basis for $L_2[0,1]$ by the CDJV construction, then the sequence norm (9.37) and the function norm are equivalent

$$C_1\|f\|_{b^\alpha_{p,q}} \le \|f\|_{B^\alpha_{p,q}} \le C_2\|f\|_{b^\alpha_{p,q}}. \tag{9.38}$$

The constants C_i may depend on $(\varphi, \psi, \alpha, p, q, L)$ but *not* on f. The proof is given for $1 \le p, q \le \infty$ and $0 < \alpha < r$.

Relations among Besov spaces. The parameter q in the Besov definitions for averaging across scale plays a relatively minor role. It is easy to see, for example from (9.37), that

$$B^\alpha_{p,q_1} \subset B^\alpha_{p,q_2}, \qquad \text{for } q_1 < q_2$$

so that $B^{\alpha}_{p,q} \subset B^{\alpha}_{p,\infty}$ for all q, [1] and so we mainly focus on the $B^{\alpha}_{p,\infty}$ or more precisely the $b^{\alpha}_{p,\infty}$ norm in our discussion.

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)

Proposition 9.12 *If $\alpha' < \alpha$ and $p' > p$ are related by $\alpha' - 1/p' = \alpha - 1/p$, then*

$$B^{\alpha}_{p,q} \subset B^{\alpha'}_{p',q}.$$

In fact, the proof becomes trivial using the sequence space form (9.37).

The situation can be summarized in Figure 9.3, which represents smoothness α 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 (i.e. $p' = \infty$) of *lower* degree $\alpha' = \alpha - 1/p'$. Indeed $B^{\alpha}_{p,q} \subset B^{\alpha'}_{\infty,q} \subset B^{\alpha'}_{\infty,\infty} \equiv C^{\alpha'}$ if $\alpha' \notin \mathbb{N}$.

The line $\alpha = 1/p$ represents the boundary of continuity. If $\alpha > 1/p$, then functions in $B^{\alpha}_{p,q}$ 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 $B^1_{1,1} \subset TV \subset B^1_{1,\infty}$.

Finally, for $B^{\alpha}_{p,q}([0,1])$, the line $\alpha = 1/p - 1/2$ represents the boundary of L_2 compactness - if $\alpha > 1/p - 1/2$, then $B^{\alpha}_{p,q}$ norm balls are compact in L_2: this observation is basic to estimation in the L_2 norm, as noted in Section 5.5. Exercise 9.11 outlines one way to verify this, leaning on the sequence space form (9.37) and the norm equivalence (9.38).

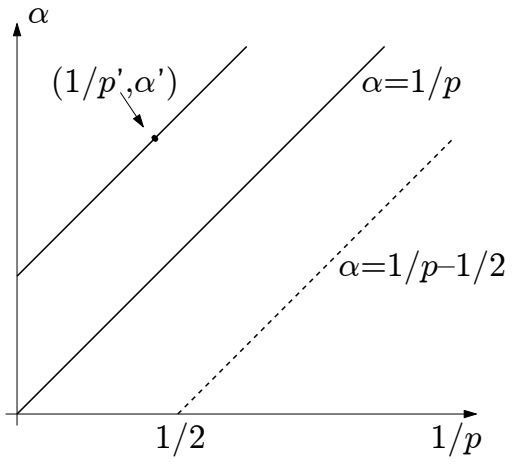

Figure 9.3 Summarizes the relation between function spaces through the primary parameters α (smoothness) and $1/p$ (integration in L_p). The middle line is the 'boundary of continuity' and the bottom, dashed, line is the 'boundary of compactness'.

[1] If $(B_1, \|\cdot\|_1)$ and $(B_2, \|\cdot\|_2)$ are normed linear spaces, $B_1 \subset B_2$ means that for some constant C, we have $\|f\|_2 \leq C\|f\|_1$ for all $f \in B_1$.

Besov and Sobolev norms. While the Besov family does not match the Sobolev family precisely, we do have the containment, for $r \in \mathbb{N}$,

$$W_p^r \subset B_{p,\infty}^r.$$

In addition, when $p \le 2$ we have

$$B_{p,p}^r \subset W_p^r.$$

We can write these embedding statements more explicitly. For $r \in \mathbb{N}$, there exists a constant C such that

$$\|f\|_{B_{p,\infty}^r}^p \le C \int_0^1 |f|^p + |D^r f|^p. \tag{9.39}$$

In the other direction, for $0 < p \le 2$ and $r \in \mathbb{N}$, there exists a constant C such that

$$\int_0^1 |f|^p + |D^r f|^p \le C \|f\|_{b_{p,p}^r}^p. \tag{9.40}$$

A proof of (9.39) appears in Appendix B after (B.27), while for (9.40), see Johnstone and Silverman (2005b), though the case $p \le 1$ is elementary.

More generally, $W_p^r = F_{p,2}^r$ belongs to the Triebel class of spaces, in which the order of averaging over scale and space is reversed relative to the Besov class, see e.g. Frazier et al. (1991) or Triebel (1983). In particular, this approach reveals an exceptional case in which $W_2^r = B_{2,2}^r$, cf Theorem 9.9.

Simplified notation

Consider a multiresolution analysis of $L_2[0,1]$ of one of the forms discussed in Section 7.1. For a fixed coarse scale L, we have the decomposition $L_2([0,1]) = V_L \oplus W_L \oplus W_{L+1} \oplus \cdots$, and associated expansion

$$f(x) = \sum_{k=0}^{2^L-1} \beta_k \varphi_{Lk}(x) + \sum_{j \ge L} \sum_{k=0}^{2^j-1} \theta_{jk} \psi_{jk}(x). \tag{9.41}$$

For the statistical results to follow, we adopt a simplified notation for the Besov sequence norms, abusing notation slightly. To this end, for $j < L$, define coefficients θ_{jk} to 'collect' all the entries of (β_k):

$$\begin{aligned} \theta_{jk} &= \beta_{2^j+k}, \qquad 0 \le j < L, 0 \le k < 2^j, \\ \theta_{-1,0} &= \beta_0. \end{aligned} \tag{9.42}$$

If we now write, again with $a = \alpha + 1/2 - 1/p$,

$$\|\theta\|_{b_{p,q}^\alpha}^q = \sum_j 2^{ajq} \|\theta_{j\cdot}\|_p^q,$$

then we have an equivalent norm to that defined at (9.37). Indeed, since L is fixed and all norms on a fixed finite dimensional space, here $\mathbb{R}^{2^L}$, are equivalent, we have

$$\|\beta_\cdot\|_p \asymp \left(\sum_{j=-1}^{L-1} 2^{ajq} \|\theta_{j\cdot}\|_p^q \right)^{1/q}.$$

In the case of Besov spaces on [0, 1], we will therefore often write $\Theta^\alpha_{p,q}$ instead of $b^\alpha_{p,q}$.

Notation for norm balls. For $C > 0$, let

$$\Theta^\alpha_{p,q}(C) = \Big\{\theta : \sum_j 2^{ajq}\|\theta_{j\cdot}\|_p^q \le C^q\Big\}.$$

Note that $\Theta^\alpha_{p,q}(C) \subset \Theta^\alpha_{p,\infty}(C)$, where

$$\Theta^\alpha_{p,\infty}(C) = \{\theta : \|\theta_{j\cdot}\|_p \le C2^{-aj}, \text{for all } j \ge -1\}. \tag{9.43}$$

9.7 Besov Bodies and weak ℓ_p Balls

We have seen that the weak ℓ_p quasi-norm measures the sparsity of a coefficient sequence θ, with smaller p corresponding to greater sparsity. If a parameter set Θ is contained within $w\ell_p$, then all elements $\theta \in \Theta$ satisfy the same decay estimate. We now describe some relationships between the Besov and weak ℓ_p norms for the Besov spaces on [0, 1]. [As a matter of notation, we note that $c_{p\alpha}$ will denote a constant depending only on α and p, and not necessarily the same at each appearance.]

Proposition 9.13 *Suppose that $\alpha > 1/p - 1/2$, or equivalently that $p > p_\alpha = 2/(2\alpha+1)$. Then*

$$\Theta^\alpha_{p,q} \subset w\ell_{p_\alpha},$$

but $\Theta^\alpha_{p,q} \not\subset w\ell_s$ for any $s < p_\alpha$.

Recall that the notation $B_1 \subset B_2$ for (quasi-)normed linear spaces means that there exists a constant c such that $\|x\|_{B_2} \le c\|x\|_{B_1}$ for all x. See Figure 9.4.

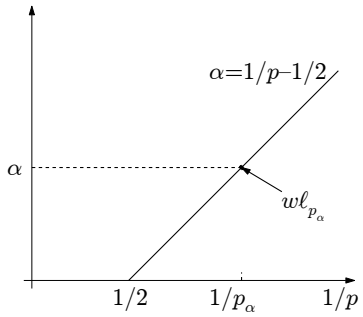

Figure 9.4 Besov spaces $\Theta^\alpha_{p,q}$ on the dotted line are included in $w\ell_{p_\alpha}$.

Proof Using the simplified notation for Besov norm balls, we need to show that, for some constant c_1 allowed to depend on α and p,

$$\Theta^\alpha_{p,q}(C) \subset w\ell_{p_\alpha}(c_1 C) \tag{9.44}$$

for $p > p_\alpha$, but that no such constant exists for $w\ell_s$ for $s < p_\alpha$.

Since $\Theta^\alpha_{p,q}(C) \subset \Theta^\alpha_{p,\infty}(C)$, it suffices to establish (9.44) for $\Theta^\alpha_{p,\infty}(C)$, which in view of

(9.43) is just a product of ℓ_p balls $\ell_{2^j,p}(C2^{-aj})$. Hence, using (9.20) and (9.18) to calculate dimension bounds for products of ℓ_p balls, and abbreviating $\Theta = \Theta^\alpha_{p,q}(C)$, we arrive at

$$N(\Theta,\epsilon) \le 1 + \sum_{j\ge 0} \min\{2^j, (C\epsilon^{-1}2^{-aj})^p\}.$$

According to (9.12), we may restrict attention to $\epsilon \le C$, since $\|\theta\|_\infty \le C$ for all $\theta \in \Theta^\alpha_{p,q}(C)$. The terms in the sum have geometric growth up to and decay away from the maximum j_* defined by equality between the two terms: thus $2^{j_*(\alpha+1/2)} = C/\epsilon$, independent of $p > p_\alpha$. Since $\epsilon \le C$, we have $j_* \ge 0$. Hence $N(\Theta,\epsilon) \le c_{\alpha p}2^{j_*}$ where we may take $c_{\alpha p} = 3 + (1-2^{-ap})^{-1} < \infty$ for $ap > 0$, which is equivalent to $p > p_\alpha$. Now, from the definition of j_*, we have $\epsilon^{p_\alpha}2^{j_*} = C^{p_\alpha}$, and so

$$\epsilon^{p_\alpha} N(\Theta,\epsilon) \le c_{\alpha p} C^{p_\alpha} \tag{9.45}$$

and so, using the criterion (9.11) for weak ℓ_p, we obtain (9.44) with $c_1 = c_{\alpha p}^{1/p_\alpha}$.

For the second part, consider the Besov shells $\Theta^{(j_0)}$ defined as the collection of those $\theta \in \Theta^\alpha_{p,q}(C)$ for which $\theta_{jk} = 0$ unless $j = j_0$. Note then that $\Theta^{(j_0)} \equiv \ell_{2^j,p}(C2^{-ja})$. Consider the shell corresponding to level $j = [j_*]$ with j_* determined above: since this shell belongs to $\Theta = \Theta^\alpha_{p,q}(C)$ for all q, we have, from (9.18)

$$N(\Theta,\epsilon) \ge \min\{2^j, [(C2^{-ja}/\epsilon)^p]\} \ge \tfrac{1}{2}2^{j_*} = \tfrac{1}{2}(C/\epsilon)^{p_\alpha}, \tag{9.46}$$

and hence that $\epsilon^s N(\Theta,\epsilon) \ge \frac{1}{2}C^{p_\alpha}\epsilon^{s-p_\alpha}$ is unbounded in ϵ if $s < p_\alpha$. □

Remarks. 1. Note that in the case $\alpha = 1/p - 1/2$, we have $a = 0$, and so

$$\Theta^\alpha_{p,p}(C) = \{\theta : \sum_j \sum_k |\theta_{jk}|^p \le C^p\} = \ell_p(C).$$

Note that there is no compactness here!

2. What happens to the embedding results when $p = p_\alpha$? For $q \le p_\alpha$ we have

$$\Theta^\alpha_{p_\alpha,q}(C) \subset \Theta^\alpha_{p_\alpha,p_\alpha}(C) = \ell_{p_\alpha}(C) \subset w\ell_{p_\alpha}(C)$$

It can also be seen that $\ell_{p_\alpha}(C) \subset \Theta^\alpha_{p_\alpha,\infty}(C)$.

3. However, there is no containment relation between $w\ell_{p_\alpha}(C)$ and $\Theta^\alpha_{p_\alpha,\infty}(C)$:

(i) The vector θ defined by $\theta_{jk} = C\delta_{k0} \in \Theta^\alpha_{p_\alpha,\infty}(C)$ but is not in $w\ell_{p_\alpha}(C')$ for any C'.

(ii) The vectors θ^{j_0} defined by $\theta^{j_0}_{jk} = \delta_{jj_0}Ck^{-1/p_\alpha}$ for $k = 1,\dots 2^j$ are each in $w\ell_{p_\alpha}(C)$, but $\|\theta^{j_0}\|_{b^\alpha_{p_\alpha,\infty}} \asymp Cj_0^{1/p_\alpha}$ is unbounded in j_0.

9.8 A framework for wavelet shrinkage results

As always our setting is the continuous Gaussian white noise model (1.21). This can be converted into a sequence model by taking coefficients in any orthonormal basis, as described in (1.24) - (1.25). Let us repeat this now explicitly in the context of an orthonormal wavelet basis adapted to $L_2[0,1]$.

Given a fixed coarse scale L, suppose that we are given an orthonormal wavelet basis $\{\varphi_{Lk}, k = 0,\dots,2^L-1\} \cup \{\psi_{jk}, k = 0,\dots,2^j-1, j \ge L\}$ leading to expansion (9.41)

for any $f \in L_2[0,1]$. In parallel with the convention (9.42) for scaling coefficients, we will for abbreviation adopt the symbols

$$\begin{aligned} \psi_{jk} &= \varphi_{L,2^j+k} \qquad 0 \le j < L, 0 \le k < 2^j \\ \psi_{-1,0} &= \varphi_{L,0}. \end{aligned}$$

With these conventions we may define the index set

$$\mathcal{I} = \{(jk) : j \ge 0, k = 0, \dots, 2^j - 1\} \cup \{(-1,0)\}.$$

As in Sections 7.4, 7.5, we write $I = (jk)$ when convenient. With this understanding, our wavelet sequence model becomes

$$y_I = \theta_I + \epsilon z_I, \qquad I \in \mathcal{I}, \tag{9.47}$$

with observed coefficients $y_I = \langle \psi_I, dY \rangle$, true coefficients $\theta_I = \langle f, \psi_I \rangle$, and noise $z_I = \langle \psi_I, dW \rangle$. We emphasize that our convention implies that these are re-indexed scaling function coefficients for $j < L$ and genuine wavelet coefficients for $j \ge L$.

Every function $f \in L_2[0,1]$ has the expansion $f = \sum \theta_I \psi_I$, and the Parseval relation $\int f^2 = \sum_I \theta_I^2$ shows that the mapping from f to θ is an isometry, which we sometimes write $\theta[f]$. Thus $\theta[f]_I = \langle f, \psi_I \rangle$ for $I \in \mathcal{I}$. For the inverse mapping, we write $f[\theta]$ for the function defined by $f[\theta](t) = \sum \theta_I \psi_I(t)$.

In the continuous white noise model, we estimate the function f using mean integrated squared error $\int (\hat{f} - f)^2$, and of course

$$\|\hat{f} - f\|_2^2 = \sum_{\mathcal{I}} (\hat{\theta}_I - \theta_I)^2 = \|\hat{\theta} - \theta\|_{\ell_2}^2. \tag{9.48}$$

We can now use the Besov bodies to define function classes

$$\mathcal{F} = \mathcal{F}_{p,q}^{\alpha}(C) = \{f : \theta[f] \in \Theta_{p,q}^{\alpha}(C)\}, \tag{9.49}$$

secure in the knowledge that under appropriate conditions on the multiresolution analysis, these function classes will be equivalent to norm balls in $B_{p,q}^{\alpha}[0,1]$, compare (9.38).

Our choice of definitions has made the continuous white noise estimation problem exactly equivalent to the seqence model. Using the natural definition of minimax risks, we therefore have the identity

$$\begin{aligned} R_{\mathcal{E}}(\mathcal{F}, \epsilon) &= \inf_{\hat{f} \in \mathcal{E}} \sup_{\mathcal{F}} E_f \|\hat{f} - f\|^2 \\ &= \inf_{\hat{\theta} \in \mathcal{E}} \sup_{\Theta} E_\theta \|\hat{\theta} - \theta\|^2 = R_{\mathcal{E}}(\Theta, \epsilon). \end{aligned} \tag{9.50}$$

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 $\hat{\theta}_I(y) = \hat{\delta}_I(y_I)$, where $\hat{\delta}$ belongs to one of the four families in the following table. In the table, v is a *scalar* variable.

Family	Description	Form of $\hat{\delta}_I$
$\mathcal{E}_L$	Diagonal linear procedures in the wavelet domain	$\hat{\delta}_I^L(v) = c_I v$
$\mathcal{E}_S$	Soft thresholding of wavelet coefficients	$\hat{\delta}_I^S(v) = (\|v\| - \lambda_I)_+ \mathrm{sgn}(v)$
$\mathcal{E}_H$	Hard thresholding of wavelet coefficients	$\hat{\delta}_I^H(v) = v1_{\{\|v\| \geq \lambda_I\}}$
$\mathcal{E}_N$	Scalar nonlinearities of wavelet coefficients	Arbitrary $\hat{\delta}_I^N(v)$

The corresponding estimators in classes $\mathcal{E}$ in (9.50) in the continous white noise model are defined by $\hat{f} = f[\hat{\theta}] = \sum_I \hat{\delta}_I(\langle \psi_I, dY \rangle)\psi_I$, where $\hat{\theta} \in \mathcal{E}_S, \mathcal{E}_L$ and so on.

A 'projected' model. On occasion, it will be useful to consider a version of the wavelet sequence model (9.47) in which only the first $n = 2^J$ coefficients are observed. For this purpose define the initial index set

$$\mathcal{I}_{(n)} = \{(jk) \in \mathcal{I} : j < J = \log_2 n\}.$$

Clearly $|\mathcal{I}_{(n)}| = n$, and the term 'projected white noise model' refers to observations

$$y_{jk} = \theta_{jk} + \epsilon z_{jk}, \qquad (jk) \in \mathcal{I}_{(n)}. \tag{9.51}$$

The name 'projected' reflects the fact that the vector $\theta^{(n)}$ defined by

$$\theta_{jk}^{(n)} = \begin{cases} \theta_{jk} & (jk) \in \mathcal{I}_{(n)} \\ 0 & (jk) \in \mathcal{I} \backslash \mathcal{I}_{(n)} \end{cases}$$

can be viewed as the image of θ under orthogonal projection $P_n : L_2 \to V_J$.

The projected model has two uses for us. First, under the calibration $\epsilon = n^{-1/2}$, it provides an n-dimensional submodel of (9.47) that is a natural intermediate step in the white noise model approximation of the Gaussian nonparametric regression model (7.21) with n observations. This theme is developed in more detail in Chapter 15. Second, it is a natural model in which to study properties of $\sqrt{2 \log n}$ thresholding of a set of n white Gaussian observations.

9.9 Adaptive minimaxity for $\sqrt{2 \log n}$ thresholding

We combine the preceding results about Besov bodies and weak ℓ_p with properties of thresholding established in Chapter 8 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 (9.47) and apply soft thresholding to the first $n = \epsilon^{-2} = 2^J$ coefficients, using threshold $\lambda_\epsilon = \sqrt{2 \log \epsilon^{-2}} = \sqrt{2 \log n}$:

$$\hat{\theta}_{jk}^U = \begin{cases} \eta_S(y_{jk}, \lambda_\epsilon \epsilon) & j < J \\ 0 & j \geq J. \end{cases} \tag{9.52}$$

The corresponding function estimate, written using the notational conventions of the last section, is

$$\hat{f}_n(t) = \sum_{(jk)\in\mathcal{I}_n} \hat{\theta}^U_{jk}\psi_{jk}(t). \tag{9.53}$$

Remarks. 1. A variant that more closely reflects practice would spare the coarse scale coefficients from thresholding: $\hat{\theta}_{jk}(y) = y_{jk}$ for $j < L$. In this case, we have

$$\hat{f}_n(t) = \sum_{k=0}^{2^L-1} \tilde{y}_{Lk}\varphi_{Lk}(t) + \sum_{j=L}^{J-1}\sum_{k=0}^{2^j-1} \hat{\theta}^U_{jk}\psi_{jk}(t) \tag{9.54}$$

where $\tilde{y}_{Lk} = \langle \varphi_{Lk}, dY\rangle$. Since L remains fixed (and small), the difference between (9.53) and (9.54) will not affect the asymptotic results below.

2. 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 (2.83). After a discrete orthogonal wavelet transform, we would arrive at the projected white noise model (9.47), with calibration $\epsilon = \sigma n^{-1/2}$. The restriction of thresholding in (9.52) to levels $j < J$ corresponds to what we might do with real data: namely threshold the n empirical discrete orthogonal wavelet transform coefficients.

The next theorem gives an indication of the broad adaptation properties enjoyed by wavelet thresholding.

Theorem 9.14 *Assume model* (9.47)*, and that* $\alpha > (1/p - 1/2)_+$, $0 < p, q \le \infty$, $0 < C < \infty$. *If* $p < 2$, *then assume also that* $\alpha \ge 1/p$. *Let* $\hat{\theta}^U$ *denote soft thresholding at* $\epsilon\sqrt{2\log n}$, *defined at* (9.52). *Let* $r = 2\alpha/(2\alpha+1)$. *Then for any Besov body* $\Theta = \Theta^\alpha_{p,q}(C)$ *and as* $\epsilon \to 0$,

$$\begin{aligned}\sup_\Theta r_\epsilon(\hat{\theta}^U, \theta) &\le c_{\alpha p}(2\log\epsilon^{-2})C^{2(1-r)}\epsilon^{2r}(1+o(1)), \qquad (9.55)\\ &\le c_{\alpha p}(2\log\epsilon^{-2})R_N(\Theta,\epsilon)(1+o(1)).\end{aligned}$$

A key aspect of this theorem is that thresholding "learns" the rate of convergence appropriate to the parameter space Θ. The definition (9.52) of $\hat{\theta}^U$ does not depend on the parameters of $\Theta^\alpha_{p,q}(C)$, and yet, when restricted to such a set, the MSE attains the rate of convergence $r(\alpha)$ appropriate to that set, subject only to extra logarithmic terms.

The constant $c_{\alpha p}$ depends only on (α, p) and may change at each appearance; its dependence on α and p could be made more explicit using the inequalities in the proof.

Proof Let $\theta^{(n)}$ and $\hat{\theta}^{(n)}$ denote the first n coordinates – i.e. (j,k) with $j < J$ – of θ and $\hat{\theta}$ respectively. To compute a bound on the risk (mean squared error) of $\hat{\theta}$, we apply the soft thresholding risk bound (8.34) of Proposition 8.8 to $\hat{\theta}^{(n)}$. Since $\hat{\theta}_{jk} \equiv 0$ except in these first n coordinates, what remains is a "tail bias" term:

$$\begin{aligned} r(\hat{\theta}^U, \theta) &= E_\theta\|\hat{\theta}^{(n)} - \theta^{(n)}\|^2 + \|\theta^{(n)} - \theta\|^2 \qquad (9.56)\\ &\le (2\log\epsilon^{-2}+1)[\epsilon^2 + \mathcal{R}(\theta^{(n)},\epsilon)] + \sum_{j\ge J}\|\theta_{j\cdot}\|^2. \qquad (9.57)\end{aligned}$$

Bound (9.57) is a pointwise estimate – valid for each coefficient vector θ. We now

investigate its consequences for the worst case MSE of thresholding over Besov bodies $\Theta = \Theta^\alpha_{p,q}(C)$. Given α, we set as before,

$$r = 2\alpha/(2\alpha+1), \qquad p(\alpha) = 2/(2\alpha+1) = 2(1-r).$$

Now comes a crucial chain of inequalities. We use first the definition (9.8) of the ideal risk semi-norm, then the bound for ideal risk in terms of weak $\ell_{p(\alpha)}$ (the third inequality of (9.14)), and finally the fact that the Besov balls $\Theta^\alpha_{p,q}(C)$ are embedded in $w\ell_{p(\alpha)}$, specifically (9.44). Thus, we conclude that for any $\theta \in \Theta^\alpha_{p,q}(C)$ and any $\epsilon > 0$,

$$\mathcal{R}(\theta^{(n)}, \epsilon) \le \|\theta\|^{2(1-r)}_{IR,r}\epsilon^{2r} \le c_\alpha \|\theta\|^{p(\alpha)}_{w\ell_{p(\alpha)}}\epsilon^{2r} \le c_{\alpha p} C^{2(1-r)}\epsilon^{2r}. \tag{9.58}$$

Tail bias. First, note the simple bound

$$\sup\{\|\theta\|_2 : \|\theta\|_p \le C, \theta \in \mathbb{R}^n\} = Cn^{(1/2-1/p)_+}. \tag{9.59}$$

which follows from a picture: when $p < 2$, the vectors having largest ℓ_2 norm in an ℓ_p ball are sparse, being signed permutations of the "spike" $C(1,0,\dots,0)$. When $p \ge 2$, the extremal vectors are *dense*, being sign flips of $Cn^{-1/p}(1,\dots,1)$.

Now we combine across levels to obtain a tail bias bound. Suppose that $\theta \in \Theta^\alpha_{p,q}(C) \subset \Theta^\alpha_{p,\infty}(C)$: we have $\|\theta_j\|_p \le C2^{-aj}$. Now use (9.59) and write $\alpha' = \alpha - (1/p - 1/2)_+$ to get $\|\theta_j\|_2 \le C2^{-\alpha' j}$. Clearly then $\sum_{j\ge J}\|\theta_j\|_2^2$ is bounded by summing the geometric series and we arrive at the tail bias bound

$$\sup_{\theta\in\Theta^\alpha_{p,q}(C)} \|\theta^{(n)} - \theta\|^2 \le c_{\alpha'} C^2 2^{-2\alpha' J}. \tag{9.60}$$

Inserting the ideal risk and tail bias bounds (9.58) and (9.60) into (9.57), we get the non-asymptotic bound, valid for $\theta \in \Theta^\alpha_{p,q}(C)$,

$$r(\hat\theta^U, \theta) \le (2\log\epsilon^{-2}+1)[\epsilon^2 + c_{\alpha p}C^{2(1-r)}\epsilon^{2r}] + c_{\alpha p}C^2\epsilon^{4\alpha'}. \tag{9.61}$$

Now suppose that C is fixed and $\epsilon \to 0$. We verify that $\epsilon^{2\alpha'} = o(\epsilon^r)$. This is trivial when $p \ge 2$, since $2\alpha > r$. When $p < 2$, the condition $\alpha \ge 1/p$ implies $2\alpha' = 2a \ge 1 > r$. This completes the proof of (9.57).

Lower Bounds. We saw in the proof of Proposition 9.13 that when $\epsilon \le C$, the set $\Theta^\alpha_{p,q}(C)$ contains ϵ-hypercubes of dimension $N(\Theta,\epsilon) \ge c_0(C/\epsilon)^{p(\alpha)}$. Hence the general hypercube lower bound (9.16) implies that for $\epsilon \le C$,

$$R_N(\Theta,\epsilon) \ge c_1(C/\epsilon)^{p(\alpha)}\epsilon^2 = c_1 C^{2(1-r)}\epsilon^{2r}. \tag{9.62}$$

□

Remarks. 1. *Linear* estimators cannot do nearly as well when $p < 2$. As discussed in Section 9.5, this is because the linear minimax rate r' must be the same for the much larger quadratically convex hull of $\Theta^\alpha_{p,q}(C)$. The slower rate turns out to equal $r' = 2\alpha'/(2\alpha'+1)$, where $\alpha' = \alpha - (1/p - 1/2)_+$. For more detail, see Section 14.6.

2. The condition $\alpha \ge 1/p$ in the $p < 2$ case could be weakened to $\alpha > 1/p - 1/2$ by choosing to threshold, say $(\log_2\epsilon^{-2})^2$ levels rather than $\log_2\epsilon^{-2}$. However, we retain the latter choice in order to stay closer to what one might do with data in practice. The condition $\alpha > 1/p$ implies, by embedding results mentioned in Section 9.6, that the functions

$f[\theta]$ are continuous, which seems a reasonable condition in order to speak sensibly of point evaluation in model (2.83).

*Block Thresholding**

We briefly look at how the adaptation results are modified if block thresholding, considered in Sections 7.6, 8.2 and 8.3, is used instead of thresholding of individual coefficients. We focus on block soft thresholding for simplicity, as the results are then a relatively direct extension of previous arguments for scalar thresholding. With a choice $L = \log n$ for the block size [2], we will obtain improvements in the logarithmic factors that multiply the $n^{-r} = \epsilon^{2r}$ convergence rate in Theorem 9.14. However, our earlier lower bounds on thresholding risk also show that for these estimators, the logarithmic terms cannot be removed.

Consider a dyadic block size $L = 2^{j_0}$ for simplicity, where j_0 will grow slowly with decreasing ϵ, specifically $L = o(\epsilon^{-\eta})$ for any $\eta > 0$. At level $j \geq j_0$, the 2^j indices are gathered into blocks of size 2^{j_0}, thus

$$\theta_{jb} = (\theta_{j,b(L-1)+1}, \ldots, \theta_{j,bL}), \qquad b = 1, \ldots, 2^{j-j_0},$$

and the block data vector y_{jb} is defined similarly. Now define the block soft thresholding estimate on (wavelet) coefficients (y_{jk}) by

$$\hat{\theta}^B_{jb} = \begin{cases} \eta_{S,L}(y_{jb}, \lambda\epsilon) & j_0 \leq j < J \\ 0 & j \geq J, \end{cases} \tag{9.63}$$

where $\eta_{S,L}$ is the block soft threshold rule defined at (8.39). For the coarsest levels $j < j_0$, use the unbiased estimators $\hat{\theta}_{jk} = y_{jk}$.

We choose block size L and threshold parameter λ so that

$$\lambda^2 - \log\lambda^2 - 1 \geq (2\log n)/L. \tag{9.64}$$

The main example has $L = \log n$ and $\lambda = \lambda_* = \sqrt{4.50524}$, for which the left side equals 2. As noted in Section 8.3, when (9.64) holds we have $r_{S,L}(\lambda, 0) \leq n^{-1} = \epsilon^2$ and can apply the block oracle inequality of Proposition 8.9.

Theorem 9.15 *Adopt the assumptions of Theorem 9.14 for α, p, q and C. Let $\hat{\theta}^B$ denote block soft thresholding with λ and L chosen to satsify* (9.64)*, with $n = \epsilon^{-2}$. Suppose also that $L = o(\epsilon^{-\eta})$ for any $\eta > 0$. Let $\kappa = (1/p - 1/2)_+$. Then for any Besov body $\Theta = \Theta^\alpha_{p,q}(C)$ and as $\epsilon \to 0$,*

$$\begin{aligned} c'_{\alpha p}(L^\kappa C)^{2(1-r)}(\lambda\epsilon)^{2r} &\leq \sup_\Theta r_\epsilon(\hat{\theta}^B, \theta) \\ &\leq c_{\alpha p}(L^\kappa C)^{2(1-r)}(\lambda\epsilon)^{2r}(1 + o(1)) \\ &\leq c_{\alpha p}\lambda^{2r} L^{2(1-r)\kappa} R_N(\Theta, \epsilon)(1 + o(1)). \end{aligned} \tag{9.65}$$

The theorem applies even for coordinatewise thresholding, $L = 1$. In this case, with the previous threshold choice $\lambda = \sqrt{2\log n}$, we obtain a slight improvement in the exponent of

[2] For this section we use L for block size, to distinguish from wavelet coarse scale L

the logarithmic term, to $(\log \epsilon^{-2})^r$. However, the lower bound shows that this power is best possible for this threshold choice.

Turning to blocks of size $L = \log n$, with λ now constant, as studied by Cai (1999), one can check that the power of $\log \epsilon^{-2}$, namely $2(1-r)\kappa$, is no larger than r, since by assumption $\alpha \geq (1/p - 1/2)_+$. And the logarithmic term vanishes when $p \geq 2$. So in this sense, using logarithmic block sizes offers an improvement. However, the lower bound shows for $L = \log n$ that a logarithmic term is again necessary if $p < 2$.

The logarithmic terms can be removed by allowing the thresholds to depend on level j and to be estimated from the data. We return to this topic in Chapters 11 and later.

Proof For the upper bound, we follow the approach for Theorem 9.14. Let $\theta^{(n)}$ and $\hat{\theta}^{(n,B)}$ collect the coordinates from levels $j_0 \leq j < L$. We have the risk decomposition

$$r(\hat{\theta}^B, \theta) = L\epsilon^2 + E\|\hat{\theta}^{(n,B)} - \theta^{(n)}\|^2 + \sum_{j \geq J} \|\theta_{j\cdot}\|^2. \tag{9.66}$$

For the blocks appearing in $\theta^{(n,B)}$ we can use the oracle inequality for block thresholding, Proposition 8.9, to obtain the analog of (9.57):

$$E\|\hat{\theta}^{(n,B)} - \theta^{(n)}\|^2 \leq \sum_{j_0 \leq j < J} \sum_b E\|\hat{\theta}_{jb} - \theta_{jb}\|^2 \leq \epsilon^2 + \mathcal{R}(\theta^{(n)}, \bar{\lambda}\epsilon; L).$$

Here $\bar{\lambda}^2 = \lambda^2 + 1$. We need a bound on the block ideal risk analogous to (9.58). This is a consequence of an extension to a block version of the ideas around weak ℓ_p balls. In Exercises 9.5 and 9.6, we sketch the proof of an inequality that states for $\theta \in \Theta^\alpha_{p,q}(C)$,

$$\mathcal{R}(\theta^{(n)}, \bar{\lambda}\epsilon; L) = \sum_{j_0 \leq j < J} \sum_b \min\{\|\theta_{jb}\|^2, L\bar{\lambda}^2\epsilon^2\} \leq c_{\alpha p}(L^\kappa C)^{2(1-r)}(\bar{\lambda}\epsilon)^{2r}.$$

With this inequality in hand, and the assumption $L = o(\epsilon^{-\eta})$ for $\eta > 0$, the rest of the proof follows from (9.66) as for Theorem 9.14.

For the lower bound, we first use the lower bound on risk for block soft thresholding, Proposition 8.7, to obtain

$$r_\epsilon(\hat{\theta}^B, \theta) \geq (1/8) \sum_{j_0 \leq j < J} \sum_b \min(\|\theta_{jb}\|^2, \lambda^2 L\epsilon^2).$$

As in the proof of Proposition 9.13, the space $\Theta^\alpha_{p,q}(C)$ contains a copy of the ℓ_p ball $\Theta^{(j)} = \ell_{2^j,p}(C2^{-aj})$, essentially by setting all $\theta_{j'k} = 0$ for $j' \neq j$. Hence, for each level $j < J$,

$$\sup_\Theta r_\epsilon(\hat{\theta}^B, \theta) \geq (1/8) \sup_{\Theta^{(j)}} \sum_b \min(\|\theta_{jb}\|^2, \lambda^2 L\epsilon^2). \tag{9.67}$$

At this point, we focus on $p < 2$, leaving $p \geq 2$ to Exercise 9.12. We first adapt Lemma 9.3, the evaluation of ideal risk over ℓ_p balls, to this setting. Regard $\ell_{B,p}(C)$ as an ℓ_p ball of block norms; the lemma says in part that

$$\sup_{\ell_{B,p}(C)} \sum_{b=1}^B \min(\|\theta_b\|^2, \epsilon^2) \geq \begin{cases} \epsilon^2[C^p/\epsilon^p] & 1 \leq C/\epsilon \leq B^{1/p} \\ B\epsilon^2 & C/\epsilon \geq B^{1/p}. \end{cases} \tag{9.68}$$

Observe that if $(\|\theta_b\|)_{b=1}^B \in \ell_{B,p}(C)$ and $n = ŁB$, then the vector $(\|\theta_1\|, \ldots, \|\theta_B\|, 0, \ldots, 0)$ with $n - B$ zeros belongs to $\ell_{n,p}(C)$, so that the lower bound above applies to $\ell_{n,p}(C)$ also.

We may now apply the previous display to (9.67), making the assignments

$$B \leftrightarrow 2^j/Ł, \qquad \epsilon \leftrightarrow \lambda\sqrt{Ł}\epsilon, \qquad C \leftrightarrow C2^{-aj}.$$

It will be seen that the resulting bounds from (9.68) increase for $j \le j_*$ and decrease with $j \ge j_*$, where j_* is determined by the equation $C/\epsilon = B^{1/p}$, which with the identifications just given and with $p(\alpha) = 2/(2\alpha + 1)$ becomes

$$2^{j_*/p(\alpha)} = Ł^{1/p-1/2}C/(\lambda\epsilon).$$

At $j = j_*$, the bound

$$B\epsilon^2 \leftrightarrow 2^{j_*}(\lambda\epsilon)^2 = (\lambda\epsilon)^{2r}(Ł^\kappa C)^{2(1-r)},$$

which is the bound claimed. □

9.10 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 dyadic sequence model (9.47). We again consider the wavelet threshold estimator with threshold $\sqrt{2\log n}$, this time without shrinkage of coarse scale coefficients, so that the estimator $\hat{f}_n(t_0)$ is given by (9.54).

In global estimation, we have seen that results are naturally obtained both for average ($p < \infty$) as well as uniform ($p = \infty$) measures of smoothness. For estimation at a point, we need smoothness information locally, near that point, which would not be directly guaranteed by an average measure. For that reason, we adopt a hypothesis of Hölder smoothness here. Recall from (9.49) that $\mathcal{F}^\alpha_{\infty,\infty}(C) = \{f : \theta[f] \in \Theta^\alpha_{\infty,\infty}(C)\}$. We state the result in terms of the sample size parameter $n = \epsilon^{-2}$.

Theorem 9.16 *Suppose that the wavelet ψ is C^α, has compact support and has at least $\lceil\alpha\rceil$ vanishing moments. Let $r = 2\alpha/(2\alpha+1)$ and let $\hat{f}_n(t_0)$ be given by* (9.52) *and* (9.54). *Then*

$$\sup_{f\in\mathcal{F}^\alpha_{\infty,\infty}(C)} E[\hat{f}_n(t_0) - f(t_0)]^2 \le c_{\psi,\alpha}C^{2(1-r)}\Big(\frac{\log n}{n}\Big)^r(1+o(1)). \tag{9.69}$$

Proof Decompose the estimation error over 'coarse', 'mid' and 'tail' scales:

$$\hat{f}_n(t_0) - f(t_0) = \sum_{I\in c} a_I + \sum_{I\in m} a_I + \sum_{I\in t} a_I. \tag{9.70}$$

The main term runs over the mid scales,

$$\sum_{I\in m} a_I = \sum_{j=L}^{J-1}\sum_k (\hat{\theta}_{jk} - \theta_{jk})\psi_{jk}(t_0),$$

and points to the new point in the proof. In global estimation, the error $\|\hat{f} - f\|^2$ is expressed in terms of that of the coefficients, $\sum(\hat{\theta}_I - \theta_I)^2$, by Parseval's equality, using the

orthonormality of the basis functions ψ_{jk}. In estimation at a point t_0, there is no orthogonality in t, and instead we bound the root mean squared (RMS) error of a sum by the sum of the RMS errors:

$$E\Big(\sum_I a_I\Big)^2 = \sum_{I,J} Ea_I a_J \le \sum_{I,J} \sqrt{Ea_I^2}\sqrt{Ea_J^2} = \Big(\sum_I \sqrt{Ea_I^2}\Big)^2. \tag{9.71}$$

We can use previous results to bound the individual terms Ea_I^2. Indeed, recall from (8.9) the mean squared error bound for a soft threshold estimator with threshold λ, here given for noise level ϵ and $\bar\lambda^2 = 1 + \lambda^2$:

$$r_S(\lambda\epsilon, \theta; \epsilon) \le \epsilon^2 r(\lambda, 0) + \theta^2 \wedge \bar\lambda^2\epsilon^2 \tag{9.72}$$

Since $\lambda = \sqrt{2\log n}$, we have from (8.7) that $r(\lambda, 0) \le n^{-1}$. We use the Hölder continuity assumption and Lemma 7.3 to bound $|\theta_{jk}| \le cC2^{-(\alpha+1/2)j}$. In conjunction with $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we obtain

$$\begin{aligned}\sqrt{Ea_I^2} &\le |\psi_I(t_0)|\,[\epsilon\sqrt{r(\lambda,0)} + |\theta_I| \wedge \bar\lambda\epsilon] \\ &\le c_\psi 2^{j/2}\ \ [1/n + C2^{-(\alpha+1/2)j} \wedge \delta_n]\end{aligned}$$

where $\delta_n = \bar\lambda\epsilon$ can be taken as $\sqrt{2\log n/n}$ by increasing c_ψ slightly.

In the sum over I, to control the number of terms we use the compact support assumption on ψ: suppose that it has length S. Then for a given level j, at most S terms $\psi_{jk}(t_0)$ are non-zero. Hence

$$\begin{aligned}\sum_{I\in m}\sqrt{Ea_I^2} &\le cS2^{J/2}/n + cS\sum_{j<J} 2^{j/2}\min(C2^{-(\alpha+1/2)j}, \delta_n) \\ &\le c/\sqrt{n} + c_{\alpha,\psi}C^{1-r}\delta_n^r,\end{aligned} \tag{9.73}$$

where we have used geometric decay bound (9.34).

To organize the rest of the proof, combine (9.70) and (9.71); we obtain

$$E[\hat f_n(t_0) - f(t_0)]^2 = E\Big(\sum_{I\in c\cup m\cup t} a_I\Big)^2 \le \Big(\sum_{I\in c\cup m\cup t}\sqrt{Ea_I^2}\Big)^2.$$

In the coarse scale sum over $I \in c$, the terms $a_I = (y_{Lk} - \beta_{Lk})\varphi_{Lk}(t_0)$ for $k = 0, \dots, 2^L - 1$. We have $Ea_I^2 \le c_\varphi^2 n^{-1}$ and so

$$\sum_{I\in c}\sqrt{Ea_I^2} \le 2^L c_\varphi n^{-1/2}. \tag{9.74}$$

In the tail sum over $I \in t$, we have $a_I = \theta_I\psi_I(t_0)$ for $I = (jk)$ and $j \ge J$. Using again the Hölder coefficient decay bound and the compact support of ψ,

$$\sum_{I\in t}|a_I| \le c_\psi S\sum_{j\ge J} C2^{-(\alpha+1/2)j}\cdot 2^{j/2} \le cC2^{-\alpha J} = cCn^{-\alpha}. \tag{9.75}$$

Combining the coarse, mid and tail scale bounds (9.74), (9.73) and (9.75), we complete the proof:

$$E[\hat f_n(t_0) - f(t_0)]^2 \le (c_1 n^{-1/2} + c_2 C^{1-r}\delta_n^r + c_3 Cn^{-\alpha})^2 \le c_2^2 C^{2(1-r)}\delta_n^{2r}(1 + o(1)).$$

□

Remarks. 1. The corresponding *lower* bound for estimation at a point over Hölder classes is of order n^{-r}, *without* the log term. More precisely

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}^{\alpha}_{\infty,\infty}(C)} E[\hat{f}(t_0) - f(t_0)]^2 \geq c\, C^{2(1-r)} n^{-r}.$$

We will not give a proof as we have not discussed estimation of linear functionals, such as $f \to f(t_0)$, in detail. However, an argument in the spirit of Chapter 4 can be given relatively easily using the method of hardest *one*-dimensional subfamilies, see Donoho and Liu (1991, Sec. 2). The dependence of the minimax risk on n and C can also be nicely obtained by a renormalization argument, Donoho and Low (1992). For Besov balls, see Exercise 9.14.

2. If we knew both α and C, then we would be able to construct a linear minimax estimator $\hat{f}_n^{\alpha,C} = \sum_I c_I y_I$ where the (c_I) are the solution of a quadratic programming problem depending on C, α, n (Ibragimov and Khasminskii (1982); Donoho and Liu (1991); Donoho (1994)). This estimator has worst case risk over $\mathcal{F}^{\alpha}_{\infty,\infty}(C)$ asymptotic to the correct rate bound $c_\alpha C^{2(1-r)} n^{-r}$. However, 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 (9.53) does not depend on the parameters (C, α), and yet achieves nearly the optimal rate of convergence – up to a factor $\log^r n$ – over all the Hölder classes.

Lepskii (1991) and Brown and Low (1996b) have shown that this rate penalty $\log^r n$ is in fact optimal: even if the correct Hölder class is one of two, specified by pairs (α_0, C_0) and (α_1, C_1) with $\alpha_0 < \alpha_1$, then

$$\inf_{\hat{f}_n} \max_{i=0,1} \sup_{\mathcal{F}^{\alpha_i}_{\infty,\infty}(C_i)} C_i^{2(r_i-1)} n^{r_i} E[\hat{f}_n(t_0) - f(t_0)]^2 \geq c_2 \log^{r_0} n.$$

3. It is evident both intuitively and also from Lemma 7.3 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. Indeed Lemma 7.3 is only needed for indices I with $\psi_I(t_0) \neq 0$.

9.11 *Outlook:* Overview of remaining chapters.

The statistical results which conclude this first part of the book, Theorems 9.2 and 9.14, make quite informative statements about co-ordinatewise 'universal' thresholding. For example, the class of parameter spaces is broad enough to decisively distinguish thresholding from any linear estimator. The results do however raise or leave open a number of related questions, some of which are explored in more detail in the second part of the book, and are outlined here.

One basic theme, already apparent in the structure of this chapter, recurs in each setting. A result or technique is first formulated in a 'single sequence' model, as for example in Theorem 9.2. The same technique can then be carried over to function estimation by regarding each level j in the wavelet transform as an instance of the sequence model, and then combining over levels, as for example in Theorem 9.14

Other loss functions (Chapter 10). In Theorems 9.2 and 9.14, as in most of the rest

of this book, the focus has been on the squared error loss function. We give an analog of the near-minimaxity result Theorem 9.14 for loss functions $\|\hat{\theta} - \theta\|_{b^{\alpha'}_{p',q'}}$ from the class of Besov norms. Wavelet thresholding, at threshold $\epsilon\sqrt{2\log n}$, is simultaneously near asymptotic minimax (up to at most a logarithmic factor) for all these loss functions. The technique is borrowed from the deterministic optimal recovery model of numerical analysis. The early sections do the preparatory work in the single sequence model.

Losing the log term: optimal rates (Chapters 11, 12). It is of both theoretical and practical interest to understand whether it is possible to remove the logarithmic gap ($\log n$ in Theorem 9.2 and $\log \epsilon^{-2}$ in Theorem 9.14) between upper and lower bounds, while still using adaptive estimators of threshold type. (Recall, for example, Figure 7.6, in which the threshold $\sqrt{2\log n}$ was too large).

This question is intimately linked with the use of data-dependent thresholds. We sketch a heuristic argument that suggests that an estimator using a constant threshold $\lambda\epsilon$ (even if λ depends on n) cannot be simultaneously minimax over ℓ_p balls $\ell_{n,p}(C)$ as p and C vary.

Suppose $y \sim N_n(\theta, \epsilon^2 I)$ and $\hat{\theta}_{\delta,i}(y) = \eta_S(y_i, \epsilon\lambda_\delta)$ where $\lambda_\delta = \sqrt{2\log \delta^{-1}}$. Using Corollary 8.4 and adding over co-ordinates yields[3]

$$r(\hat{\theta}_\delta, \theta) \le 2\delta n\epsilon^2 + (1 + 2\log\delta^{-1}) \sum_{i=1}^{n} \min(\theta_i^2, \epsilon^2).$$

Now maximize over $\theta \in \ell_{n,p}(C)$–it can be shown, e.g. Lemma 9.3–that for $1 \le (C/\epsilon)^p \le n$, we have $\sum \min(\theta_i^2, \epsilon^2) \le C^p\epsilon^{2-p}$, and so

$$\sup_{\theta\in\ell_{n,p}(C)} r(\hat{\theta}_\delta, \theta) \le 2\delta n\epsilon^2 + (1 + 2\log\delta^{-1})C^p\epsilon^{2-p}.$$

We might select δ to minimize the right side bound: this immediately leads to a proposed choice $\delta = n^{-1}(C/\epsilon)^p$ and threshold $\lambda = \sqrt{2\log n(\epsilon/C)^p}$. Observe that as the signal to noise ratio C/ϵ increases from 1 to $n^{1/p}$, the nominally optimal threshold decreases from $\sqrt{2\log n}$ to 0, and no single threshold value appears optimal for anything other than a limited set of situations.

A number of approaches to choosing a data dependent threshold were reviewed in Section 7.6. In Chapter 11 we explore another alternative, based on complexity penalized model selection. Informally it may be described as imposing a penalty of order $2k\log(n/k)$ on models of size k. If we denote by $\hat{k}$ the size of the selected model, the associated threshold is often close to $\epsilon(2\log n/\hat{k})^{1/2}$, so that larger or 'denser' selected models correspond to smaller thresholds and 'sparser' models to larger ones. A virtue of the complexity penalized approach is the existence of oracle inequalities analogous to Proposition 8.8, but without the multiplicative log term—loosely, one may say that the logarithm was incorporated instead into the penalty. The corresponding estimator is defined adaptively, i.e. without reference to p and C, and yet satisfies *non-asymptotic* upper and lower bounds for MSE over the range of ℓ_p balls, that differ only at the level of constants.

The complexity penalized bounds have implications for wavelet shrinkage estimation of functions when applied separately at each level of a multiresolution analysis. In Chapter

[3] the factor $2\delta n\epsilon^2$, while a looser bound than given by (8.13), leads to cleaner heuristics here.

12, we show that this leads to estimators that are rate-adaptive over a wide range of Besov spaces: essentially an analog of Theorem 9.14 without the $\log n$ multiplicative term. In this chapter we also return to the theme of linear inverse problems used as a class of examples in earlier chapters: the wavelet-vaguelette decomposition (WVD) allows one to construct adaptive rate-optimal wavelet shrinkage estimators for a class of inverse problems possessing a WVD.

Exact constants (Chapters 13, 14). In discussing adaptive minimaxity, we have emphasized the practical importance of estimators which do not depend on the indices of parameter spaces such as $\ell_p(C)$ and $\Theta^\alpha_{p,q}(C)$. However, in order to calibrate the performance of these estimators, and to more fully understand the structure of these estimation settings, it is also of interest to evaluate exactly or asymptotically the minimax risk for specific parameter sets such as the ℓ_p balls or Besov bodies. Such an evaluation should be accompanied by a description of the (approximately) minimax estimator and their corresponding least favorable priors.

Thus, in Chapter 13, the optimality results for ℓ_p balls are summarized, and the thresholds $\lambda = \sqrt{2 \log n(\epsilon/C)^p}$ derived heuristically above are shown in fact to be asymptotically minimax for $\ell_{n,p}(C)$. In particular, thresholding rules are found to be asymptotically optimal among *all* esimators in the limit $n^{-1}(C_n/\epsilon_n)^p \to 0$.

In Chapter 14 these considerations are extended to Besov bodies. A key structural result is that *separable* rules, one for which $\hat{\theta}_i(y)$ depends on y_i alone, can be found which are asymptotically minimax, and the corresponding least favorable priors make individual wavelet coefficients independent. Of course, these estimators and priors depend strongly on the indices α, p, q and C.

Epilogues.
A. Continuous versus discrete ...
B. Some related topics. ...

9.12 Notes

§1. DeVore (1998) is an excellent survey article on basic ideas of non-linear approximation. The equivalence of the compression, ideal risk and weak ℓ_p quasi-norms was shown by Donoho (1993). [The definition of $\|\theta\|_{IR,r}$ is slightly modified here.] Using absolute values $|\theta|_{(i)}$ (rather than squares) to define a compression norm $\sup_k k^{-1+1/p} \sum_{i=1}^k |\theta|_{(i)}$ works for the more restricted range $1 < p$, e.g. DeVore and Lorentz (1993, Ch. 2, Prop. 3.3).

§3. The construction of lower bounds using subsets of growing cardinality has a long history reviewed in Tsybakov (2009, Ch. 2); important papers on the use of hypercubes include Bretagnolle and Huber (1979) and Assouad (1983).

(Remark on $p/((2-p)$ as difference between weak and strong ℓ_p norm minimax risks. Also FDR connections?).

Meyer (1990, Section 6.4) explains that it is not possible to characterize the integer Hölder classes $C^m(\mathbb{R})$ in terms of moduli of wavelet coefficients.

Theorem 9.6 and Remark 9.7 extend to $C^\alpha([0,1])$ with the same proof, so long as the boundary wavelets satisfy the same conditions as ψ.

§6. Meyer (1990, Chapter 3) establishes a more general form of Theorem 9.9: using a Fourier definition of W_2^α and the notion of an r-regular multiresolution analysis, he establishes the equivalence (9.36) for all *real* α with $|\alpha| < r$.

Diagrams using the $(\alpha, 1/p)$ plane are used by Devore, for example in the survey article on nonlinear approximation DeVore (1998).

§9. Efromovich (2004a, 2005) uses lower bounds to risk for specific signal to provide insight on block and threshold choice.

§10. The pointwise estimation upper bound of Theorem 9.16 appears in Donoho and Johnstone (1996) along with discussion of optimality of the $\log^r n$ penalty in adaptation over α. Cai (2002) shows that $\log n$ is the optimal block size choice to achieve simultaneously optimal global and local adaptivity.

Exercises

9.1 *(Quasi-norm properties.)* (a) Give an example of θ and θ' for which

$$\|\theta + \theta'\|_{w\ell_p} > \|\theta\|_{w\ell_p} + \|\theta'\|_{w\ell_p}.$$

(b) Verify that for $a, b \geq 0$ and $p > 0$,

$$2^{(1-p)_+}(a^p + b^p) \leq (a+b)^p \leq 2^{(p-1)_+}(a^p + b^p).$$

9.2 *(More quasi-norm properties.)*
(a) Show that $|\theta + \theta'|_{(2k-1)} \leq |\theta|_{(k)} + |\theta'|_{(k)}$ for each $k \geq 1$.
(b) Hence show that $\|\theta\|_{c,\alpha}$ defined at (9.2) is a quasi-norm.
(c) Show also that $\|\theta\|_{IR,r}$ defined at (9.8) is a quasi-norm.

9.3 *(ℓ_p constraints in step function basis.)* Suppose ϕ is the indicator of the unit interval $[0, 1]$ and $\phi_{n,k}(t) = n^{1/2}\phi(nt - k)$. If $f = \sum_1^n \theta_k \phi_{n,k}$, show that

$$\int_0^1 |f|^p = n^{p/2-1} \sum_1^n |\theta_k|^p.$$

9.4 *(Fixed thresholds on weak ℓ_p balls).* Suppose that $y \sim N_n(\theta, \epsilon^2 I)$, and let $c_p = 2/(2-p)$.
(i) Let $\hat{\theta}^\lambda$ denote soft thresholding at $\lambda\epsilon$. Assume that $p < 2$. Show that

$$\bar{r}_\epsilon(\hat{\theta}^\lambda, w\ell_{n,p}(C)) = \sup_{\theta \in w\ell_{n,p}(C)} r_\epsilon(\hat{\theta}^\lambda, \theta) \leq n\epsilon^2 r_S(\lambda, 0) + c_p(1+\lambda^2)^{1-p/2} C^p \epsilon^{2-p}.$$

This should be compared with bound (9.23) for $\lambda = \epsilon\sqrt{2\log n}$.
(ii) Let C_n, ϵ_n depend on n and define the normalized radius $\eta_n = n^{-1/p}(C_n/\epsilon_n)$. If $\eta_n \to 0$ as $n \to \infty$, set $\lambda_n = \sqrt{2\log \eta_n^{-p}}$ and show that

$$\bar{r}_\epsilon(\hat{\theta}^\lambda, w\ell_{n,p}(C)) \leq c_p \cdot n\epsilon_n^2 \cdot \eta_n^p (2\log \eta_n^{-p})^{1-p/2}(1 + o(1)).$$

[This turns out to be the minimax risk for weak ℓ_p; compare the corresponding result for strong ℓ_p in (13.47).]

9.5 *(Block weak ℓ_p norms.)* Suppose the elements of $\theta = (\theta_k, k \in \mathbb{N})$ are grouped into successive blocks of size L, so $\theta_b = (\theta_{b(L-1)+1}, \ldots, \theta_{bL})$. Let $\|\theta_b\|$ be the ℓ_2 norm $\left(\sum_{k\in b} \theta_k^2\right)^{1/2}$ and with slight abuse of notation write $\|\theta\|_{(b)}$ for the bth largest of the ordered values of $\|\theta_b\|$, thus $\|\theta\|_{(1)} \geq \|\theta\|_{(2)} \geq \cdots$. Then say that θ belongs to block weak-ℓ_p if $\|\theta\|_{(b)} \leq Cb^{-1/p}$, and let $\|\theta\|_{w\ell_{p,L}}$ denote the smallest such C. Let $N(\theta, \delta; L) = \#\{b : \|\theta_b\| \geq \delta\}$. Show that

$$\|\theta\|^p_{w\ell_{p,L}} = \sup_b b\|\theta\|^p_{(b)} = \sup_{\delta>0} \delta^p N(\theta, \delta; L),$$

9.6 *(Besov bodies and block weak ℓ_p balls.)* (a) By analogy with (9.15), define an extension to blocks of hypercube dimension: $N(\Theta, \delta; L) = \sup_{\theta\in\Theta} N(\theta, \delta; L)$. Show that (if L divides n),

$$N(\ell_{n,p}(C), \epsilon; L) \leq \min\left(\frac{n}{L}, L^{(p/2-1)_+}\frac{C^p}{\epsilon^p}\right).$$

(b) Suppose that $L = 2^{j_0}$ for some j_0. Now consider a segment of the Besov body $\Theta = \Theta^{\alpha}_{p,q}(C) \cap \{\theta : \theta_{j\cdot} = 0, j < j_0\}$. Show that

$$N(\Theta, \epsilon\sqrt{L}; L) \le \sum_{j=j_0}^{\infty} \min\left(\frac{2^j}{L}, L^{(p/2-1)_+}\left(\frac{C}{\epsilon\sqrt{L}}2^{-aj}\right)^p\right).$$

Let $p_\alpha = 2/(2\alpha+1)$. Show that for $p > p_\alpha$ and some $c = c_{p\alpha}$, for all $\theta \in \Theta$ we have

$$\|\theta\|_{w\ell_{p_\alpha,L}} \le cL^{1/(p\wedge 2)-1/p_\alpha}C,$$

thus generalizing (9.44).

(c) Conclude that with $r = 2\alpha/(2\alpha+1) = 1 - p_\alpha/2$, we have

$$\sup_{\theta\in\Theta^{\alpha}_{p,q}(C)} \sum_{j=j_0}^{\infty}\sum_{b} \min(\theta_{jb}^2, \delta^2 L) \le c\delta^{2r}(L^{(1/p-1/2)_+}C)^{2(1-r)}.$$

(d) Consider now all of $\Theta^{\alpha}_{p,q}(C)$, instead of the segment Θ. Show that now there is a vector $\theta \in \Theta^{\alpha}_{p,q}(C)$ with

$$\|\theta\|_{w\ell_{p_\alpha,L}} \ge cC.$$

[This is the reason for considering only blocks at level j_0 and above.]

9.7 *(James-Stein and thresholding on a sparse signal.)* Suppose that $X \sim N_n(\mu_n, I)$, let $\hat{\mu}^{JS}$ denote the James-Stein estimator (2.60), and $\hat{\mu}^{\lambda}$ soft thresholding at λ.

(i) Suppose that $\|\mu_n\|_2^2 \sim \gamma n$ as $n \to \infty$. Show that $r(\hat{\mu}^{JS}, \mu_n) \sim [\gamma/(\gamma+1)]n$.

(ii) Let $\mu_{n,k} = n^{1/2}k^{-1/p}, k = 1, \dots, n$ be the weak ℓ_p extremal vector, with $0 < p < 2$. Show that with $\lambda_n = \sqrt{(2-p)\log n}$,

$$r(\hat{\mu}^{\lambda_n}, \mu_n) \le c_p n^{p/2}(\log n)^{1-p/2}, \qquad \text{while} \qquad r(\hat{\mu}^{JS}, \mu_n) \sim c'_p n.$$

9.8 *(Hölder smoothness and wavelet coefficients.)* Assume the hypotheses of Remark 9.7 and in particular that smoothness α satisfies $m < \alpha < m+1$ for $m \in \mathbb{N}$. Show that the bounds

$$|\beta_{Lk}| \le C, \qquad |\theta_{jk}| \le C2^{-(\alpha+1/2)j},$$

imply that

$$|D^m f(x) - D^m f(y)| \le C'|x-y|^{\alpha-m}.$$

9.9 *(Wavelet coefficients of BV functions.)* Show that if $\int \psi = 0$ and supp $\psi \subset I$, then for $f \in TV$, we have

$$\int_I f\psi \le \tfrac{1}{2}\|\psi\|_1 |f|_{TV}.$$

[Hint: begin with step functions.] Thus, complete the proof of the upper bound in Theorem 9.10.

9.10 *(Besov norm of a singularity.)* Verify Example 9.11, for example as follows. Let $S(\psi_{jk})$ denote the support of wavelet ψ_{jk}. Establish the bounds

$$|\theta_{jk}| \le \begin{cases} C2^{-j(\beta+1/2)}|k|^{-(r-\beta)} & 0 \notin S(\psi_{jk}) \\ C2^{-j(\beta+1/2)} & 0 \in S(\psi_{jk}), \end{cases}$$

and hence show that $2^{ja}\|\theta_{j\cdot}\|_p \le c2^{j(\alpha-\beta-1/p)}$.

9.11 *(Compactness criterion.)* (a) Show, using the total boundedness criterion C.16, that $\Theta^\alpha_{p,q}(C) \subset \Theta^\alpha_{p,\infty}(C)$ is ℓ_2-compact when $\alpha > (1/p - 1/2)_+$.
(b) Show also that if $\alpha = 1/p - 1/2$ then $\Theta^\alpha_{p,p}(C)$ is not compact.

9.12 *(Lower bound, $p \geq 2$ case in Theorem 9.15.)* Recalling that $\Theta^{(j)} = \ell_{2^j,p}(C2^{-aj})$ and $p_\alpha = 2/(2\alpha+1)$, show that

$$\sup_{\Theta^{(j)}} \sum_b \min(\|\theta_{jb}\|^2, \lambda^2 L\epsilon^2) = 2^j [C^2 2^{-2j/p_\alpha} \wedge \lambda^2\epsilon^2],$$

and hence, for suitable choice of $j_* \in \mathbb{R}$, that the right side takes the value $(\lambda\epsilon)^{2r} C^{2(1-r)}$.

9.13 *(Thresholding at very fine scales.)* We wish to weaken the condition $\alpha \geq 1/p$ in Theorem 9.14 to $\alpha > 1/p - 1/2$. Instead of setting everything to zero at levels J and higher (compare (9.52)), one possibility for controlling tail bias better is to apply soft thresholding at very high scales at successively higher levels:

$$\hat\theta_{jk} = \begin{cases} \delta_S(y_{jk}, \lambda_j\epsilon), & j < J^2 \\ 0 & j \geq J^2 \end{cases}$$

where for $l = 0, 1, \ldots, J-1$,

$$\lambda_j = \sqrt{2(l+1)\log\epsilon^{-2}} \qquad \text{for} \quad lJ \leq j < (l+1)J.$$

Show that if, now $\alpha > 1/p - 1/2$, then the upper risk bound in Theorem 9.14 continues to hold with $\log\epsilon^{-2}$ replaced by, say, $(\log\epsilon^{-2})^3$.

9.14 *(Pointwise estimation over Besov classes.)*
(a) Show that point evaluation—the mapping $f \to f(t_0)$ for a fixed $t_0 \in (0,1)$—is a continuous functional on $B^\alpha_{p,q}$ so long as $\alpha > 1/p$.
(b) Assume then that $\alpha > 1/p$. Show that if we use a Besov ball $\mathcal{F}^\alpha_{p,q}(C)$ in place of the Hölder ball $\mathcal{F}^\alpha_{\infty,\infty}(C)$, then the pointwise estimation bound (9.69) holds with the *slower* rate $r' = 2\alpha'/(2\alpha'+1)$, where $\alpha' = \alpha - 1/p$, in contrast with the rate for global estimation $r = 2\alpha/(2\alpha+1)$ of Theorem 9.14. [The optimality of this slower rate for $\mathcal{F}^\alpha_{p,q}(C)$ follows, for example, from the renormalization argument of Donoho and Low (1992).]

10

The optimal recovery approach to thresholding.

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 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 Θ and *loss functions* $\|\hat{\theta} - \theta\|$, in contrast with our previous focus mainly on squared error loss. We again 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.

The setting is the projected version of the white noise model with $n = 2^J$ observations, (9.51), restated here for convenience:

$$y_I = \theta_I + \epsilon z_I, \qquad I \in \mathcal{I}_{(n)}, \tag{10.1}$$

where $I = (jk)$, and $\mathcal{I}_{(n)} = \{(jk) : 0 \le j < J, k = 0, 1, \ldots, 2^j - 1\} \cup \{(-1, 0)\}$ is the collection of the first n wavelet coefficients. As usual ϵ is known and $z_I \overset{iid}{\sim} N(0, 1)$.

We continue our study of asymptotic properties of thresholding at a level $\delta_n = \epsilon\sqrt{2\log n}$, already begun in Sections 9.9 and 9.10 which focused on adaptation results for global and pointwise squared error respectively. In this chapter we focus on global error measures (and parameter spaces) drawn from the Besov scale and derive two types of result.

First, the function estimates $\hat{f} = f[\hat{\theta}]$ corresponding to (9.52) 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 10.6). Second, the threshold estimator (9.52) is *simultaneously* near minimax (Theorem 10.10).

The proofs of these two properties 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.

Section 10.1 begins therefore with a description of the near minimax properties of soft thresholding in the deterministic optimal recovery model. It introduces the modulus of continuity of the error norm with respect to the parameter space, which later plays a key role in evaluating rates of convergence.

The statistical consequences are developed in two steps: first in a general n-dimensional 'monoresolution' Gaussian white noise model, in Sections 10.2–10.4, which makes no special mention of wavelets, and later in Sections 10.5– 10.8 for the multiresolution wavelet sequence model (10.1).

In both cases, when phrased in terms of moduli of continuity, upper bounds are direct consequences of the deterministic results: this is set out in the monoresolution setting in Section 10.2.

Actual evaluation of the modulus of continuity is taken up in Section 10.3, for the setting of error norms and parameter sets defined by ℓ_p norms. As we seek to cover models of sparsity, we include the cases $0 < p < 1$, for which the ℓ_p measure is only a quasi-norm.

The main finding is that for Θ an ℓ_p-ball and $\|\cdot\|$ an $\ell_{p'}$-norm, the behavior of the modulus depends on whether $p \geq p'$, corresponding to dense least favorable configurations, or whether $p < p'$, corresponding to sparse configurations.

Lower bounds in the statistical model do not flow directly from the deterministic one, and so Section 10.4 collects the arguments in the monoresolution setting, with separate results for sparse and dense cases.

Section 10.5 takes up the multiresolution model, beginning with the important fact that wavelets provide *unconditional bases* for the Besov scale of spaces: this may be seen as a formalization of the idea that shrinkage of coefficients—as in linear estimation or in thresholding—is a stable operation. The property of preservation of smoothness under thresholding, highlighted earlier, is a direct consequence.

Section 10.6 begins with the multiresolution analog of Section 10.2: drawing consequences of a deterministic observation model, now incorporating the notion of 'tail bias', which is introduced to deal with estimation of a full sequence vector (θ_I) with the cardinality of $\mathbb{N}$ based on only n observations. The main results for estimation in Besov norms over Besov balls in statistical model (10.1) are formulated. A new phenomenon appears: a distinct, and slower, rate of convergence for parameter combinations $\mathbf{p}$ in a 'logarithmic' zone. (The reason for the name appears after the detailed statement of Theorem 10.10.)

The details of the calculation of the modulus of continuity for Besov norms are taken up in Section 10.7. The modulus provides a convenient summary describing the rate of convergence corresponding to $\|\cdot\|_{b'}$ and Θ_b. An important tool is the use of 'Besov shells', which consist in looking at signals θ whose only non-zero components lie in the j-th shell. Focusing on the j-th shell alone reduces the calculations to an ℓ_p ball. By studying the modulus as the shell index j varies, we see again the pattern of geometric decay away from a critical level $j_* = j_*(\mathbf{p})$.

Finally, Section 10.8 presents lower bounds for the multiresolution setting. The Besov shell device, after appropriate calibration, reduces the lower bound arguments to previous results for ℓ_p-balls and error measures presented in Section 10.4.

10.1 A Deterministic Optimal Recovery Model

Consider the following *deterministic* version of the sequence model. Data $x = (x_I)$ is observed that satisfies

$$x_I = \theta_I + \delta u_I \qquad |u_I| \leq 1 \qquad I \in \mathcal{I}. \tag{10.2}$$

It is desired to recover the unknown vector θ, 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 δ is assumed known. The worst case error suffered by an estimator $\hat{\theta}$ is then

$$e(\hat{\theta}, \theta; \delta) = \sup_{|u_I| \le 1} \|\hat{\theta}(x) - \theta\|. \tag{10.3}$$

We will see that a number of conclusions for the statistical (Gaussian) sequence model can be drawn, after appropriate calibration, from the deterministic model (10.2).

Assumptions on loss function and parameter space. Throughout this chapter we will assume:

(i) $\Theta \subset \ell_2(\mathcal{I})$ is solid and orthosymmetric, and

(ii) The error norm $\|\cdot\|$ is also solid and orthosymmetric, in the sense that

$$|\xi_I| \le |\theta_I| \ \forall I \quad \Rightarrow \quad \|\xi\| \le \|\theta\|.$$

The error norm can be convex, as usual, or at least ρ-convex, $0 < \rho \le 1$, in the sense that $\|\theta + \xi\|^\rho \le \|\theta\|^\rho + \|\xi\|^\rho$.

The Uniform Shrinkage Property of Soft Thresholding. Soft thresholding at threshold λ can be used in the optimal recovery setting: $\hat{\theta}_\lambda = (\hat{\theta}_{\lambda,I})$ is, as usual,

$$\hat{\theta}_{\lambda,I}(x_I) = \operatorname{sgn}(x_I)(|x_I| - \lambda)_+. \tag{10.4}$$

The shrinkage aspect of *soft* thresholding has the simple but important consequence that the estimate remains confined to the parameter space:

Lemma 10.1 *If Θ is solid orthosymmetric and $\lambda \ge \delta$, then $\theta \in \Theta$ implies that $\hat{\theta}_\lambda \in \Theta$.*

Proof Since soft thresholding shrinks each data coordinate x_I towards 0 (but not past 0!) by an amount λ that is greater than the largest possible noise value δ that could be used to expand θ_I in generating x_I, it is clear that $|\hat{\theta}_{\lambda,I}| \le |\theta_I|$. Since Θ is solid orthosymmetric, this implies $\hat{\theta}_\lambda \in \Theta$. □

Minimax Error. The minimax error of recovery in the determinstic model is

$$E(\Theta, \delta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} e(\hat{\theta}, \theta; \delta),$$

where $e(\hat{\theta}, \theta) = e(\hat{\theta}, \theta; \delta)$ is given by (10.3). Good bounds on this minimax error can be found in terms of a *modulus of continuity* defined by

$$\Omega(\delta) = \Omega(\delta; \Theta, \|\cdot\|) = \sup_{(\theta_0, \theta_1) \in \Theta \times \Theta} \{\|\theta_0 - \theta_1\| : \|\theta_0 - \theta_1\|_\infty \le \delta\}. \tag{10.5}$$

Thus, the modulus measures the error norm $\|\cdot\|$ of differences of sequences in the parameter space Θ that are separated by at most δ in uniform norm.

Theorem 10.2 *Suppose that Θ is solid and orthosymmetric, and that the error norm $\|\cdot\|$ is solid, orthosymmetric and ρ-convex. Then*

$$(1/2^{1/\rho})\Omega(\delta) \le E(\Theta, \delta) \le 2\Omega(\delta).$$

In addition, soft thresholding $\hat{\theta}_\delta$ is near minimax simultaneously for all such parameter spaces and error norms.

Proof For each noise vector $u = (u_I)$ under model (10.2), and $\theta \in \Theta$, we have $\hat{\theta}_\delta \in \Theta$ by the uniform shrinkage property. In addition, for each u,

$$\|\hat{\theta}_\delta - \theta\|_\infty \le \|\hat{\theta}_\delta - x\|_\infty + \|x - \theta\|_\infty \le 2\delta.$$

Hence $(\hat{\theta}_\delta, \theta)$ is a feasible pair for the modulus, and so it follows from the definition that $e(\hat{\theta}_\delta, \theta) \le \Omega(2\delta)$. Since $\Theta/2 \subset \Theta$ by solid orthosymmetry, we also have $\Omega(2\delta) \le 2\Omega(\delta)$.

Turning now to a *lower bound*, suppose that the pair $(\theta_0, \theta_1) \in \Theta \times \Theta$ attains the value $\Omega(\delta)$ defining the modulus.[1] The data sequence $x = \theta_1$ is potentially observable under (10.2) if either $\theta = \theta_0$ or $\theta = \theta_1$, and so for any estimator $\hat{\theta}$ and ρ-convex $\|\cdot\|$,

$$\sup_{\theta\in\Theta} e(\hat{\theta}, \theta) \ge \sup_{\theta\in\{\theta_0,\theta_1\}} \|\hat{\theta}(\theta_1) - \theta\| \ge \Omega(\delta)/2^{1/\rho},$$

because, if not, $\|\theta_1 - \theta_0\|^\rho \le \|\theta_1 - \hat{\theta}(\theta_1)\|^\rho + \|\hat{\theta}(\theta_1) - \theta_0\|^\rho < \Omega(\delta)^\rho$. □

We now define a modified modulus of continuity which is more convenient for calculations with ℓ_p and Besov norm balls.

$$\Omega^\circ(\delta; \Theta, \|\cdot\|) = \sup\{\|\theta\| : \theta \in \Theta, \|\theta\|_\infty \le \delta\}.$$

In fact, $\Omega(\delta; \Theta, \|\cdot\|) = \Omega^\circ(\delta; \Theta - \Theta, \|\cdot\|)$, where $\Theta - \Theta = \{\theta_1 - \theta_2 : \theta_i \in \Theta\}$ is the Minkowski sum of the sets Θ and $-\Theta$. If Θ is a norm ball $\Theta(C) = \{\theta : \|\theta\| \le C\}$ (so that $0 \in \Theta$), and if $\|\cdot\|$ is ρ-convex, then the modified modulus is equivalent to the original one:

$$\Omega^\circ(\delta) \le \Omega(\delta) \le 2^{1/\rho}\Omega^\circ(2^{-1/\rho}\delta). \tag{10.6}$$

Indeed, the left inequality follows by taking pairs of the form $(\theta, 0)$ in (10.5). For the right inequality, let (θ_0, θ_1) be any feasible pair for (10.5) with $\Theta = \Theta(C)$. Then the scaled difference $\theta = 2^{-1/\rho}(\theta_0 - \theta_1) \in \Theta(C)$ and satisfies $\|\theta\|_\infty \le 2^{-1/\rho}\delta$, so

$$\|\theta_0 - \theta_1\| = 2^{1/\rho}\|\theta\| \le 2^{1/\rho}\Omega^\circ(2^{-1/\rho}\delta).$$

The right inequality follows after maximizing over feasible pairs (θ_0, θ_1).

Note that Ω° (and Ω) satisfy the bounds

$$(c \wedge 1)\Omega^\circ(\delta) \le \Omega^\circ(c\delta) \le (c \vee 1)\Omega^\circ(\delta). \tag{10.7}$$

10.2 Monoresolution stochastic model: upper bounds

In the deterministic model of optimal recovery, Theorem 10.2 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

$$y_i = \theta_i + \epsilon z_i \qquad z_i \overset{iid}{\sim} N(0,1), \quad i = 1, \dots, n. \tag{10.8}$$

The connection with the optimal recovery model, with $\mathcal{I} = \{1, \dots, n\}$, is made by considering the event

$$A_n = \{\sup_{I\in\mathcal{I}} |z_I| \le \sqrt{2\log n}\}, \tag{10.9}$$

[1] If the supremum in (10.5) is not attained, the argument above can be repeated for an approximating sequence.

which because of the properties of maxima of i.i.d. Gaussians (c.f. Section 8.9) has probability approaching one:

$$P(A_n) = \varpi_n \geq 1 - 1/\sqrt{\pi \log n} \nearrow 1 \qquad \text{as } n \to \infty.$$

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 $\hat{\theta}_{\delta_n}$ at level $\epsilon\sqrt{2\log n}$, compare (8.29) (which uses notation λ_n in place of δ_n here). We therefore obtain immediately

Proposition 10.3 *Consider the Gaussian model* (10.8) *with n observations and* $\Theta \subset \mathbb{R}^n$. *If* $(\Theta, \|\cdot\|)$ *is solid, orthosymmetric then*

$$\sup_{\theta\in\Theta} P\{\|\hat{\theta}_{\delta_n} - \theta\| \leq 2\Omega(\epsilon\sqrt{2\log n})\} \geq \varpi_n \to 1.$$

In the next two sections, we explore the implications for estimation over ℓ_p-balls in $\mathbb{R}^n$ using error measured in $\ell_{p'}$ norms. We need first to evaluate the modulus Ω for this class of Θ and $\|\cdot\|$, and then to investigate lower bounds to match the upper bounds just proved.

Remark. In most of the book we have been concerned with statements about expected losses: $r(\hat{\theta}, \theta) = E_\theta L(\hat{\theta}, \theta)$. The optimal recovery approach leads more naturally to results about probabilities for losses: $P_\theta\{L(\hat{\theta}, \theta) > ct_n\}$. At least for upper bounds, the latter is weaker than the former, though they are related via $r(\hat{\theta}, \theta) = \int_0^\infty P_\theta\{L(\hat{\theta}, \theta) > t\}dt$, which follows from the identity $EX = \int_0^\infty P(X > t)dt$ for integrable random variables $X \geq 0$.

10.3 Modulus of continuity for ℓ_p balls

In the definition of the modulus $\Omega(\delta)$, we take $\Theta = \Theta_{n,p}(C) = \{\theta \in \mathbb{R}^n : \sum_1^n |\theta_i|^p \leq C^p\}$ and $\|\cdot\|$ equal to the (quasi-)norm of $\ell_{p',n}$ for $0 < p' < \infty$. While the leading case for $\|\cdot\|$ is perhaps $p' = 2$, the method works equally well for more general p', and it is instructive to see the dependence on p'. We introduce a new notation

$$\begin{aligned} W_{n;p',p}(\delta, C) &= \Omega^\circ(\delta;\ \Theta_{n,p}(C), \|\cdot\|_{p'}) \\ &= \sup\{\|\theta\|_{p'} :\ \|\theta\|_\infty \leq \delta, \|\theta\|_p \leq C\}. \end{aligned}$$

Usually we write more simply just $W_n(\delta, C)$, and sometimes just W. Equivalently,

$$W_n^{p'}(\delta, C) = \sup\Big\{\sum_{i=1}^n \min(|\theta_i|^{p'}, \delta^{p'}) \ :\ \sum_{i=1}^n |\theta_i|^p \leq C^p\Big\}.$$

We show that

$$W_n^{p'}(\delta, C) \approx n_0\delta_0^{p'}, \tag{10.10}$$

with the *least favorable* configurations being given up to permutations and sign changes by

$$\theta^* = (\delta_0, \ldots, \delta_0, 0, \ldots, 0), \qquad \delta_0 \leq \delta, \tag{10.11}$$

with n_0 non-zero coordinates and $1 \leq n_0 \leq n$. The explicit values of (n_0, δ_0) are shown in

Figure 10.1. The approximate equality $\approx$ occurs only if $1 < n_0 < n$ and is interpreted as in (10.12) below.

The result is a generalization of that for $p' = 2$ given in Lemma 9.3. We will therefore be more informal: the verification is mostly by picture—compare Figure 10.1. First, however, set $x_i = |\theta_i|^p$, so that we may rewrite

$$W^{p'} = \sup\Big\{ \sum x_i^{p'/p} \ : \ \sum x_i \le C^p, \|x\|_\infty \le \delta^p \Big\}.$$

The function $f(x) = \sum x_i^{p'/p}$ is concave for $p' \le p$ and strictly convex for $p' > p$, in both cases over a convex constraint set. We take the two cases in turn.

(i) $p \ge p'$. Let $\bar{x} = \text{ave } x_i$, and $\tilde{x} = (\bar{x}, \dots, \bar{x})$. By concavity, $f(\tilde{x}) \ge f(x)$, and so the maximum of f occurs at some vector $c(1, \dots, 1)$. In this case, equality occurs in (10.10).

(ii) $p < p'$. Convexity implies that the maximum occurs at extreme points of the constraint set. For example, if $Cn^{-1/p} \le \delta \le C$, then

$$\theta^* = (\delta, \dots, \delta, \eta, 0, \dots, 0), \qquad \text{with } n_0\delta^p + \eta^p = C^p, \quad \eta \le \delta.$$

Hence $W_n^{p'}(\delta) = n_0\delta_0^{p'} + \eta^{p'}$ with $\delta_0 = \delta$, and we have $W_n^{p'}(\delta, C) \approx C^p\delta^{p'-p}$, or more precisely

$$\tfrac{1}{2}C^p\delta^{p'-p} \le W_n^{p'}(\delta, C) \le 2C^p\delta^{p'-p}.$$

Indeed, using the equation $n_0\delta^p + \eta^p = C^p$ with $n_0 \ge 1$, we find

$$\frac{W^{p'}}{C^p\delta^{p'-p}} = \frac{n_0\delta^{p'} + \eta^{p'}}{(n_0\delta^p + \eta^p)\delta^{p'-p}} \in \left[\frac{n_0}{n_0+1}, \frac{n_0+1}{n_0}\right] \subset [\tfrac{1}{2}, 2]. \tag{10.12}$$

Thus n_0, or the ratio n_0/n, measures the *sparsity* of the least favorable configuration. When $p \ge p'$, the least favorable configurations are *always* dense, since the contours of the $\ell_{p'}$ loss touch those of the ℓ_p norm along the direction $(1, \dots, 1)$. On the other hand, when $p < p'$, the maximum value of $\ell_{p'}$ error over the intersection of the ℓ_p ball and $\delta-$cube is always attained on the boundary of the cube, which leads to sparser configurations when $C < \delta n^{1/p}$.

For later use, note the special case when there is no constraint on $\|\theta\|_\infty$:

$$W_{n;p',p}(\infty, C) = \sup\{\|\theta\|_{p'} : \|\theta\|_p \le C\} = Cn^{(1/p'-1/p)_+}. \tag{10.13}$$

10.4 Lower Bounds for ℓ_p balls

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.

For now, we restrict to parameter spaces which are ℓ_p balls, defined as in earler chapters by $\Theta_{n,p}(C) = \{\theta \in \mathbb{R}^n : \sum_1^n |\theta_i|^p \le C^p\}$. In stating lower bounds for the statistical model over ℓ_p balls, we need to use the structure of extremal configurations for the modulus $\Omega(\epsilon) = \Omega^\circ(\epsilon; \Theta_{n,p}(C), \|\cdot\|_{p'})$. Indeed, for given (p, C, ϵ, n), let n_0 and δ_0 be the number

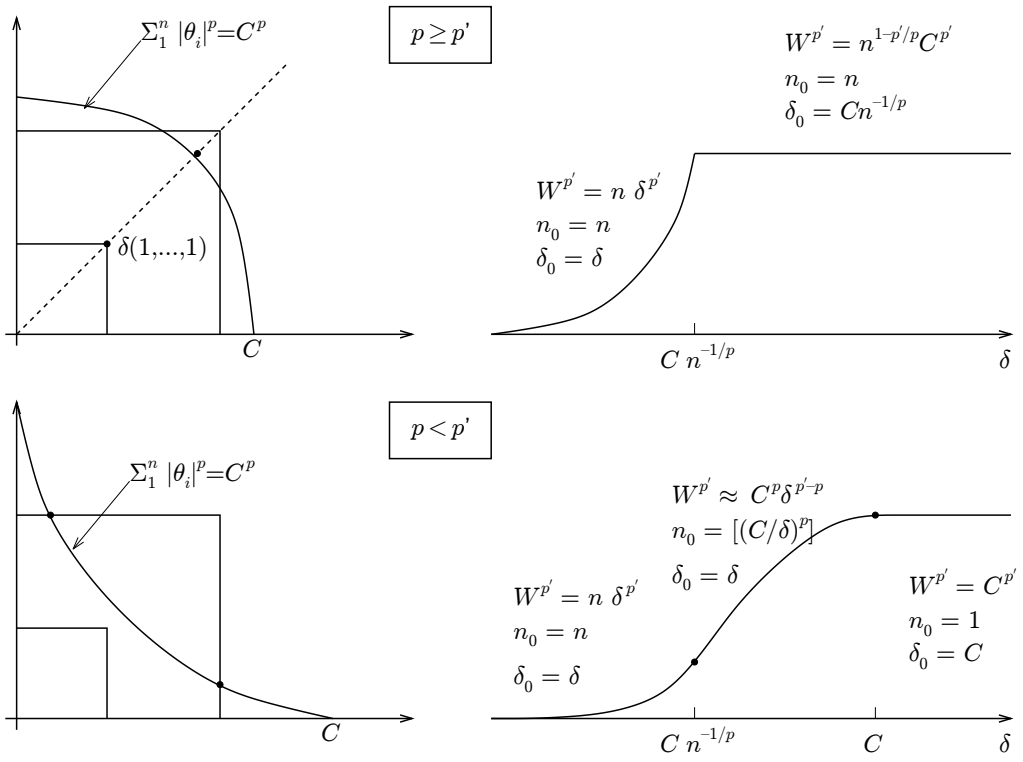

Figure 10.1 Top panel: Concave case $p \geq p'$, Bottom panel: Convex case $p < p'$. Left column shows a schematic view of the $\|\cdot\|_p$ and $\|\cdot\|_\infty$ constraints. Right column plots $W_n^{p'}(\delta, C)$ against δ, with annotations showing the definitions of n_0, δ_0 in each zone. The approximate inequality $\approx$ is interpreted as in (10.12).

and magnitude of non-zero components in the extremal vectors θ_{n_0,δ_0} of (10.11) and Figure 10.1. We develop two bounds, the first intended for dense cases (n_0 large), and the second for sparse ones($n_0 = 1$), though this is not formally part of the hypotheses.

Proposition 10.4 *Assume data are taken from model* (10.8)*.*

(i) (Dense case). Let $n_0 = n_0(p, n, C, \epsilon)$ be the number of components of size δ_0 in the least favorable configuration for $\Theta_{n,p}(C)$. Let $\pi_0 = \Phi(-1)/2$. Then

$$\inf_{\hat\theta} \sup_{\Theta_{n,p}(C)} P\{\|\hat\theta - \theta\|_{p'} \geq (\pi_0/2)^{1/p'} W_n(\epsilon, C)\} \geq 1 - e^{-2n_0\pi_0^2}. \tag{10.14}$$

(ii) (Sparse case). Fix $\eta > 0$ small. There exist functions $\pi_\eta(n) \to 1$ as $n \to \infty$ such that for any $C_n \leq \epsilon\sqrt{(2-\eta)\log n}$, then, as $n \to \infty$,

$$\inf_{\hat\theta} \sup_{\Theta_{n,p}(C_n)} P\{\|\hat\theta - \theta\|_{p'} \geq \tfrac{1}{2} C_n\} \geq \pi_\eta(n). \tag{10.15}$$

Remarks. 1. We apply this to a sequence of problems indexed by n with $C = C_n$ and $\epsilon = \epsilon_n$. In the dense case, $p \geq p'$, we always have $n_0 = n$, compare Figure 10.1. Again

from the figure, in the sparse case $p < p'$, now $n_0 \to \infty$ so long as $C_n/\epsilon_n \to \infty$. The improved lower bound of part (ii) applies so long as $C_n/\epsilon_n \le \sqrt{(2-\eta)\log n}$.

2. Thus, in the statistical model in Proposition 10.3, the lower bound in the dense case (10.14) is of order $\Omega(\epsilon_n)$, whereas an upper bound for estimation over $\Theta_{n,p}(C)$ is given, on a set of high probability, by $\Omega(\epsilon_n\sqrt{2\log n})$, using $\sqrt{2\log n}$ soft thresholding. Thus there is a gap between the two bounds that is (at most) of logarithmic order, compare (10.7). However, the near optimality of soft thresholding holds quite generally: the result holds for all $\ell_{p'}$ losses, and over all ℓ_p balls $\Theta_{n,p}(C)$.

3. In the sparse case, $p < p'$, one can rewrite the lower bound in terms of the modulus Ω by setting $c_\eta = (1-\eta/2)^{1/2}$, and observing from Figure 10.1 that if $C_n \le c_\eta \epsilon_n \sqrt{2\log n}$, then the lower bound $C_n/2 = \Omega(\epsilon\sqrt{2\log n})/2$. Thus in the sparse case 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.

Proof *Sparse Case.* This follows immediately from Proposition 8.16, because the single spike set $\Theta_n(C_n)$ of (8.58) is contained in each $\Theta_{n,p}(C_n)$.

Dense Case. The argument uses a version of the hypercube method seen in Sections 4.7 and 9.3. Let (n_0, δ_0) be parameters of the worst case configuration for $W_n(\epsilon, C)$: from the figures

$$\delta_0 = \begin{cases} \min\{\epsilon, Cn^{-1/p}\} & \text{if } p \ge p' \\ \min\{\epsilon, C\} & \text{if } p < p'. \end{cases}$$

from which it is clear that $\delta_0 \le \epsilon$. Let π be the distribution on θ which makes θ_i independently equal to $\pm\delta_0$ with probability $\frac{1}{2}$ for $i = 1, \dots, n_0$, and all other co-ordinates 0. Since supp $\pi \subset \Theta$, we have for any $(\theta, y)-$measurable event A,

$$\sup_{\theta\in\Theta} P_\theta(A) \ge P_\pi(A). \tag{10.16}$$

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 $\hat\theta$, summing over the first n_0 coordinates. Under P_π,

$$\|\hat\theta - \theta\|_{p'}^{p'} \ge \delta_0^{p'} N(\hat\theta(y),\theta). \tag{10.17}$$

Combining (10.16) and (10.17), we conclude that

$$\sup_{\theta\in\Theta} P_\theta\{\|\hat\theta-\theta\|_{p'}^{p'} \ge c\delta_0^{p'}\} \ge P_\pi\{N(\hat\theta,\theta)\ge c\}.$$

It was shown in Section 4.7 that the right side probability is minimized over $\hat\theta$ by the rule $\hat\theta_{\pi,i}(y) = \delta_0 \mathrm{sgn}(y_i)$. Hence $N(\hat\theta_\pi,\theta) = \sum_{i=1}^{n_0} I\{\mathrm{sgn}(y_i)\theta_i < 0\}$ counts sign errors in the data. Since the first n_0 co-ordinates are i.i.d., this is a binomial variable with n_0 trials and with 'success' probability

$$\pi_1 = P_\pi\{y_1\theta_1 < 0\} = P\{\delta_0 + \epsilon z < 0\} = \Phi(-\delta_0/\epsilon).$$

Consequently our minimax error probability

$$S(c) = \inf_{\hat\theta}\sup_{\Theta} P_\theta\{\|\hat\theta-\theta\|_{p'}^{p'} \ge c\delta_0^{p'}\} \ge P\{\mathrm{Bin}(n_0,\pi_1)\ge c\}.$$

Let $c = n_0\pi_0$; and suppose that $\pi_1 > \pi_0$. Write $K(\pi_0, \pi_1)$ for the Kullback-Leibler divergence $\pi_0 \log(\pi_0/\pi_1) + (1-\pi_0)\log((1-\pi_0)/(1-\pi_1))$. At the end of the chapter we recall the Cramér-Chernoff large deviations principle

$$P\{\text{Bin}\,(n_0, \pi_1) < n_0\pi_0\} \le e^{-n_0 K(\pi_0,\pi_1)},$$

along with the inequality $K(\pi_0, \pi_1) \ge 2(\pi_1 - \pi_0)^2$. If $\delta_0 \le \epsilon$, then $\pi_1 \ge 2\pi_0$ and so we conclude that

$$1 - S(n_0\pi_0) \le e^{-2n_0\pi_0^2},$$

and since $n_0\delta_0^{p'} \ge (1/2)W_n^{p'}(\epsilon, C)$, this establishes (10.14). □

10.5 Multiresolution model: unconditional bases

We now turn to estimation in the multiresolution model (10.1), which as we have seen is intimately related to estimation of a function $f(t)$ on $[0, 1]$ in the continuous Gaussian white noise model (1.21). As in Chapter 9, we are interested in parameter spaces for f defined by quantitative measures of smoothness such as (quasi-)norm balls in Besov spaces.

We describe here a key property of wavelet bases that allows us to establish strong properties for co-ordinatewise soft thresholding. An *unconditional basis* $\{\psi_I\}$ for a Banach space B can be defined by two conditions. The first is that $\{\psi_I\}$ is a *Schauder basis*, meaning every $v \in B$ has a unique representation, that is a unique sequence $\{\theta_I\} \subset \mathbb{C}$ such that $v = \sum_1^\infty \theta_I\psi_I$. The second is a *multiplier* property: there exists a constant C such that for every N and all sequences $\{m_I\} \subset \mathbb{C}$ with $|m_I| \le 1$, we have

$$\|\sum_1^N m_I\theta_I\psi_I\| \le C\|\sum_1^N \theta_I\psi_I\|. \tag{10.18}$$

Several equivalent forms and interpretations of the definition are given by Meyer (1990, I, Ch. VI). Here we note only that (10.18) 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 9.6 is that *wavelets form unconditional bases for the Besov scale of function spaces*. Indeed, when viewed in terms of the sequence norms

$$\|f\|_{B^\alpha_{p,q}} \asymp \|\theta\|_{b^\alpha_{p,q}} = \|\beta_{L\cdot}\|_p + \Big(\sum_{j\ge L} 2^{ajq}\|\theta\|_p^q\Big)^{1/q},$$

recall (9.38) and (9.39), the multiplier property is trivially satisfied, since $\|\theta[f]\|$ depends on θ_{jk} *only through* $|\theta_{jk}|$. Donoho (1993, 1996) has shown that unconditional bases are in a certain sense optimally suited for compression and statistical estimation.

Definition 10.5 Suppose that the orthonormal wavelet ψ is C^R and has D vanishing moments. Consider a *scale* of functional spaces

$$\mathcal{C}(R, D) = \{B^\alpha_{p,q}[0,1] : 0 < p, q \le \infty, 1/p < \alpha < \min(R, D)\}. \tag{10.19}$$

As seen in Section 9.6 after Proposition 9.12, these spaces are all embedded in $C[0,1]$, since $\alpha > 1/p$. The wavelet system $\{\psi_{jk}\}$ forms an unconditional basis for each of the spaces in the scale, since $\alpha < \min(R, D)$, (Donoho, 1992b).

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 deterministic model (10.2) can be used to construct an estimator of $f = \sum \theta_I \psi_I$ by setting $\hat{f} = \sum \hat{\theta}_{\lambda,I} \psi_I$, where estimator $\hat{\theta}_\lambda$ is given by (10.4). The uniform shrinkage property combined with the multiplier property (10.18) imply that whatever be the noise u,

$$\|\hat{f}\|_{\mathcal{F}} \le C \|f\|_{\mathcal{F}}.$$

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 ψ with D vanishing moments, this property holds *simultaneously* for all spaces $\mathcal{F}$ in the scale $\mathcal{C}(R, D)$ of (10.19).

Statistical model. We may immediately draw conclusions for the statistical model (10.1). On the event A_n of (10.9), the uniform shrinkage property Lemma 10.1 implies that the estimator $\hat{\theta}_{\delta_n} \in \Theta$ whenever $\theta \in \Theta$. Here $\delta_n = \epsilon\sqrt{2\log n}$. Consequently, for function spaces in the scale $\mathcal{C}(R, D)$, we have on A_n that $\|\hat{f}_n\|_{\mathcal{F}} \le C(\mathcal{F})\|f\|_{\mathcal{F}}$. Hence

Theorem 10.6 *Assume model* (10.1). *For each function space $\mathcal{F} \in \mathcal{C}(R, D)$ there exists a constant $C(\mathcal{F})$ such that*

$$P\{\|\hat{f}_n\|_{\mathcal{F}} \le C(\mathcal{F})\|f\|_{\mathcal{F}} \ \ \forall \mathcal{F} \in \mathcal{C}\} \ge \varpi_n \to 1.$$

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 ϖ_n so that one can assert that $\hat{f}_n$ does not find "spurious structure".

Remark. In general, Fourier series do not behave nearly so stably under shrinkage of coefficients as do wavelet series. Indeed, Kahane et al. (1977) showed that given any periodic $f \in L_2[0,1]$, there exists a continuous periodic function g on $[0,1]$ such that the respective Fourier coefficients $\{f_k\}$ and $\{g_k\}$ satisfy $|f_k| \le |g_k|$. Thus, shrinkage of Fourier coefficents can make a function rougher.

10.6 Statistical Upper and Lower Bounds

We now turn to the statement and proof of results for the statistical model (10.1) that are valid simultaneously for parameter spaces and error measures based on norms from the scale $\mathcal{C}(R, D)$. As in the monoresolution case of Section 10.4, we apply deterministic optimal recovery results to a high probability set in the statistical model.

A projected optimal recovery model. Our statistical model is based on $n = \epsilon^{-2} = 2^J$ observations, while the estimand f, or equivalently the sequence $(\theta_I[f], I \in \mathcal{I})$ is indexed by all of $\mathbb{N}$. We therefore begin with an extension of the deterministic results of Section 10.1 to a 'projected' model with a finite number n of observations:

$$x_I = \theta_I + \delta u_I \qquad I \in \mathcal{I}_{(n)}, \qquad |\mathcal{I}_{(n)}| = n.$$

We write $x^{(n)} = (x_I; I \in \mathcal{I}_{(n)})$. Again, one still attempts to recover the entire object θ, and the corresponding minimax recovery error is

$$E(\Theta, \delta; n) = \inf_{\hat{\theta}(x^{(n)})} \sup_{\Theta} e(\hat{\theta}(x^{(n)}), \theta; \delta).$$

Clearly $E(\Theta, \delta; n) \geq E(\Theta, \delta)$ since esitmators that use only $x^{(n)}$ are a subclass of those allowed in $E(\Theta, \delta)$. Projection onto the $n-$data model is defined by

$$(P_n\theta)_I = \begin{cases} \theta_I & I \in \mathcal{I}_{(n)} \\ 0 & \text{otherwise.} \end{cases}$$

Even when the noise level $\delta = 0$, there is still an error of recovery due to the attempt to infer the full vector θ from only n components. Indeed

$$e(\hat{\theta}(x^{(n)}), \theta; 0) = \|\hat{\theta}(\theta^{(n)}) - \theta\|.$$

Let $\Theta_n^{\perp} = \{\theta \in \Theta : \theta^{(n)} = 0\}$. We make the

Definition. The *tail $n-$width* of Θ in norm $\|\cdot\|$ is

$$\Delta(n; \Theta, \|\cdot\|) = \sup_{\theta \in \Theta}\{\|\theta\| : \theta \in \Theta_n^{\perp}\}. = \sup_{\theta \in \Theta}\{\|\theta\| : P_n\theta = 0\}. \tag{10.20}$$

Lemma 10.7 *If Θ and the error norm $\|\cdot\|$ are solid and orthosymmetric,*

$$\Delta(n; \Theta, \|\cdot\|) = E(\Theta, 0; n). \tag{10.21}$$

Proof Let $\hat{\theta}_{\mathrm{id}}(\theta^{(n)}) = (\theta^{(n)}, 0)$ be the 'estimator' that concatenates zeros after the indices in $\mathcal{I}_n$. Clearly $e(\hat{\theta}_{\mathrm{id}}(\theta^{(n)}), \theta; 0) = \|(I - P_n)\theta\|$ so that $E \leq \Delta(n)$. In the other direction,

$$E(\Theta, 0; n) \geq \inf_{\hat{\theta}} \sup_{\theta \in \Theta_n^{\perp}} \|\hat{\theta}(0^{(n)}) - \theta\| \geq \sup_{\theta \in \Theta_n^{\perp}} \|\theta\| = \Delta(n),$$

where the second inequality follows since both Θ and $\|\cdot\|$ are solid and orthosymmetric. □

It is then straightforward to establish the following finite data analog of Theorem 10.2.

Proposition 10.8 *Suppose that Θ is solid and orthosymmetric, and that the error norm $\|\cdot\|$ is solid, orthosymmetric and ρ-convex. Then*

$$\max\{\Omega(\delta)/2^{1/\rho}, \Delta(n)\} \leq E(\Theta, \delta; n) \leq c_\rho[2\Omega(\delta) + \Delta(n)].$$

In addition, soft thresholding $\hat{\theta}_\delta$ is near minimax simultaneously for all such parameter spaces and error norms.

Proof Since $E(\Theta, \delta; n)$ is bounded below by both $E(\Theta, \delta)$ and $E(\Theta, 0; n)$, the lower bound follows from Theorem 10.2 and Lemma 10.7. For the upper bound, consider the first n and the remaining co-ordinates separately and use ρ-convexity:

$$\|\hat{\theta}_\delta - \theta\| \leq c_\rho[\|\hat{\theta}_\delta^{(n)} - \theta^{(n)}\| + \|\theta^{(n)} - \theta\|] \leq c_\rho[2\Omega(\delta) + \Delta(n)].$$ □

Global Estimation Bounds

In a similar manner, we can immediately convert the upper-bound part of Proposition 10.8 to a statement in the projected Gaussian model with $\delta_n = \epsilon_n \sqrt{2 \log n}$: for the soft threshold estimator $\hat{\theta}_{\delta_n}$, we have for all solid, orthosymmetric Θ that

$$\sup_{\Theta} P\{\|\hat{\theta}_{\delta_n} - \theta\| \le 2c_\rho[\Omega(\delta) + \Delta(n)]\} \ge \varpi_n \to 1.$$

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), 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 Θ and error (quasi-)norm $\|\cdot\|$ taken from the Besov scale. Thus, recalling the sequence based definition (9.37), we use one Besov norm $\|\cdot\|_b = \|\cdot\|_{b^\alpha_{p,q}}$ to define a parameter space $\Theta(C) = \{\theta : \|\theta\|_{b^\alpha_{p,q}} \le C\}$, and a typically different Besov norm $\|\cdot\|_{b'} = \|\cdot\|_{b^{\alpha'}_{p',q'}}$ for the error measure. This of course represents a substantial extension of the class of error measures: the squared error loss considered in most of the rest of the book corresponds to $\alpha' = 0$, $p' = q' = 2$. We remark that the norm $\|\cdot\|_{b'}$ is ρ-convex with $\rho = \min(1, p', q')$, Exercise 10.1.

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. We then formulate the statistical conclusions in terms of the modulus functions–this is the main result, Theorem 10.10, of this chapter.

An interesting feature is the appearance of distinct zones of parameters $\mathbf{p} = (\alpha, p, q, \alpha', p', q')$:

Regular	$\mathcal{R} = \{p' \le p\}$	$\cup$	$\{p' > p,\ (\alpha + 1/2)p > (\alpha' + 1/2)p'\}$
Logarithmic	$\mathcal{L} =$		$\{p' > p,\ (\alpha + 1/2)p < (\alpha' + 1/2)p'\}$

In the "critical case" $(\alpha + 1/2)p = (\alpha' + 1/2)p'$, the behavior is more complicated and is discussed in Donoho et al. (1997).

We recall that the notation $a(\epsilon) \asymp b(\epsilon)$ means that there exist constants c_1, c_2 and c_3, here allowed to depend on $\mathbf{p}$ but not ϵ or C, such that for all $\epsilon < c_3$ we have the pair of bounds $c_1 a(\epsilon) \le b(\epsilon) \le c_2 a(\epsilon)$.

Theorem 10.9 *Let $\Theta = \Theta^\alpha_{p,q}(C)$ and $\|\cdot\| = \|\cdot\|_{b^{\alpha'}_{p',q'}}$. Assume that*

$$\tilde{\alpha} = \alpha - \alpha' - (1/p - 1/p')_+ > 0.$$

(a) Then the modulus $\Omega(\delta; \Theta, \|\cdot\|)$ given by (10.5) *satisfies*

$$\Omega(\delta) \asymp C^{1-r}\delta^r \qquad \text{as } \delta \to 0. \tag{10.22}$$

where the rate exponent is given by $r =$

$$r_R = \frac{(\alpha - \alpha')}{\alpha + 1/2}, \qquad \text{for } \mathbf{p} \in \mathcal{R},$$

$$r_L = \frac{\tilde{\alpha}}{\alpha + 1/2 - 1/p}, \qquad \text{for } \mathbf{p} \in \mathcal{L}.$$

(b) the tail bias satisfies, with $c_2 = (2^{\tilde{\alpha}q'} - 1)^{-1/q'}$,

$$\Delta(n) \le c_2 C n^{-\tilde{\alpha}}. \tag{10.23}$$

If in addition $\alpha > 1/p$, *then* $\Delta(n) = o(\Omega(n^{-1/2}))$.

Part (b) shows that the condition $\tilde{\alpha} > 0$ is needed for the tail bias to vanish with increasing n; we refer to it as a *consistency* condition. In particular, it forces $\alpha' < \alpha$. In the logarithmic zone, the rate of convergence is reduced, some simple algebra shows that for $\mathbf{p} \in \mathcal{L}$ we have $r_L < r_R$.

Some understanding of the regular and logarithmic zones comes from the smoothness parameter plots introduced in Chapter 9.6. For given values of the error norm parameters α' and p', Figure 10.2 shows corresponding regions in the $(1/p, \alpha)$ plane. The regular/logarithmic boundary is given by the dashed line $\alpha = \omega/p - 1/2$ having slope $\omega = (\alpha' + 1/2)p'$. The *consistency boundary* corresponding to condition $\alpha > \alpha' + (1/p - 1/p')+$ is given by the broken line with inflection at $(1/p', \alpha')$. Note that the two lines in fact intersect exactly at $(1/p', \alpha')$.

If $\omega > 1$, or what is the same, if $a' = \alpha' + 1/2 - 1/p' > 0$, then there is a logarithmic zone. In this case, the consistency boundary lies wholly on or above the continuity boundary $\alpha = 1/p$ so long as $\alpha' \geq 1/p'$, otherwise the condition $\alpha > 1/p$ imposes an additional constraint.

On the other hand, if $\omega \leq 1$ or $a' \leq 0$, the zone boundary line is tangent to the consistency line and there is no logarithmic zone. This explains why there is no logarithmic zone for traditional squared error loss, corresponding to $\alpha' = 0$, $p' = 2$. In this case the continuity boundary $\alpha = 1/p$ implies a further constraint to ensure negligibility of the tail bias.

As particular examples, one might contrast the error measure $\int |D^2 \hat{f} - D^2 f|$, with $\alpha' = 2$, $p' = 1$ and $\omega = 5/2$, which has a logarithmic zone, with the measure $\int |\hat{f} - f|$, with $\alpha' = 0$, $p' = 1$ and $\omega = 1/2$, which does not.

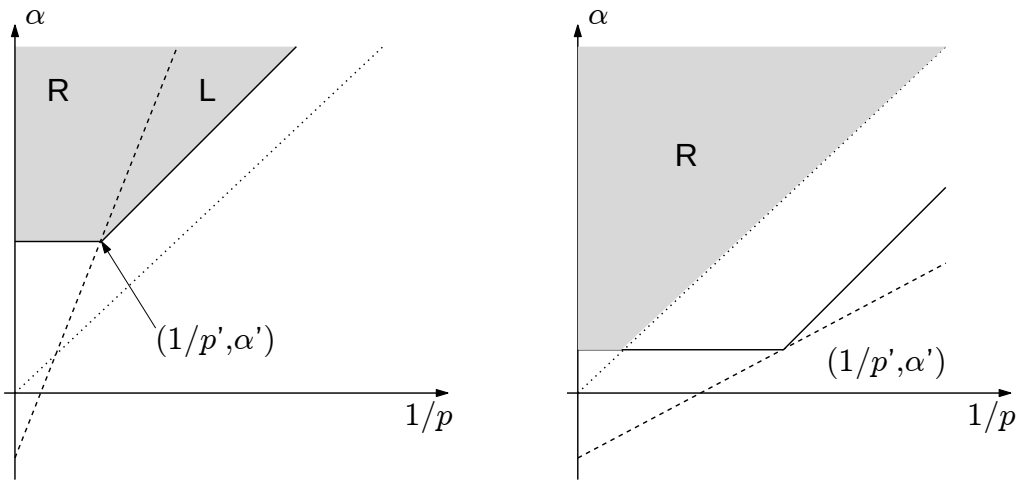

Figure 10.2 Schematic representation of regular $\mathcal{R}$ and logarithmic $\mathcal{L}$ zones in in two cases: left panel when $\omega = (\alpha' + 1/2)p' > 1$, and right panel with $\omega < 1$ and no logarithmic zone. In both cases, solid line is consistency boundary $\alpha = \alpha' + (1/p - 1/p')_+$, dashed line is the regular/logarithmic boundary $\alpha = \omega/p - 1/2$ and dotted line is the continuity boundary $\alpha = 1/p$.

Make the normalization $\epsilon = n^{-1/2}$. Using the bounds derived for the Besov modulus and for the tail bias in Theorem 10.9 we obtain the first display in

Theorem 10.10 *Assume model* (10.1) *with* $\epsilon = n^{-1/2}$. *Let* $\Theta = \Theta^{\alpha}_{p,q}(C)$ *and* $\|\cdot\| = \|\cdot\|_{b^{\alpha'}_{p',q'}}$. *Assume that* $\tilde{\alpha} = \alpha - \alpha' - (1/p - 1/p')_+ > 0$ *and that* $\alpha > 1/p$. *Then soft thresholding,* (8.29)*, satisfies*

$$\sup_{\theta \in \Theta(C)} P\{\|\hat{\theta}_{\delta_n} - \theta\| \le c\Omega(n^{-1/2}\sqrt{\log n})\} \ge \varpi_n \to 1.$$

There exists a constant $c = c(\boldsymbol{p})$ *such that*

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \ge c\Omega(n^{-1/2})\} \to 1. \tag{10.24}$$

In the logarithmic case, the lower bound can be strengthened to $\Omega(n^{-1/2}\sqrt{\log n})$.

As in Section 10.4, the lower bound (10.24) is established by separate arguments, in Section 10.8.

Thus, soft thresholding at $\delta_n = \epsilon_n\sqrt{2\log n}$ is simultaneously nearly minimax (up to a logarithmic term) over all parameter spaces and loss functions—indexed by a total of 7 parameters(!)—in the scale $\mathcal{C}(R, D)$, 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 $\Theta^{\alpha}_{2,2}$ and using it on another space $\Theta^{\alpha}_{p,q}$ with $p < 2$ in which linear estimators are known (e.g. Chapter 9.9) to have suboptimal rates of convergence.

10.7 Besov Modulus and Tail Bias

In this section prove Theorem 10.9. We evaluate the asymptotic order of the modulus of continuity $\Omega(\delta)$ when both parameter space $\Theta^{\alpha}_{p,q}$ and error measure $\|\cdot\|_{b'} = \|\cdot\|_{b^{\alpha'}_{p',q'}}$ are taken from the Besov scale. The approach is to reduce the optimization defining the modulus to a hardest resolution level j, where one is effectively dealing with scaled versions of ℓ_p norms in both the error measure and in the reduced parameter space.

As in Section 7.5, write $\mathcal{I}_j = \{(jk) : 0 \le k \le 2^j - 1\}$ for the indices at level j, and define the *Besov shells*

$$\Theta^{(j)} = \{\theta \in \Theta : \theta_I = 0, I \notin \mathcal{I}_j\}.$$

If $\theta^{(j)}$ is derived from θ by setting to zero all components θ_I with $I \notin \mathcal{I}_j$, then

$$\|\theta^{(j)}\|_b = \|\theta^{(j)}\|_{b^{\alpha}_{p,q}} = 2^{aj}\|\theta_{j\cdot}\|_p \tag{10.25}$$

where, again, $a = \alpha + 1/2 - 1/p$. This shows that $\Theta^{(j)}$ is isomorphic to a scaled ℓ_p ball: $\Theta^{(j)} \cong \Theta_{2^j,p}(C2^{-aj})$. The modified modulus of continuity, when restricted to the jth shell,

reduces in turn to a scaled form of the ℓ_p-modulus:

$$\begin{aligned}\Omega_j(\delta) &:= \Omega^\circ(\delta; \Theta^{(j)}, \|\cdot\|_{b'}) \\ &= 2^{a'j} W_{2^j}(\delta, C2^{-aj}) = W_{2^j}(2^{a'j}\delta, C2^{-(a-a')j}),\end{aligned} \tag{10.26}$$

where we have used the invariance $bW_n(\delta, C) = W_n(b\delta, bC)$. It is easy to verify that nothing essential (at the level of rates of convergence) is lost by considering the shell moduli: with $\rho = p' \wedge q' \wedge 1$ and $c_\rho = 2^{1/\rho}$,

$$\|(\Omega_j(\delta))_j\|_{\ell_\infty} \le \Omega(\delta) \le c_\rho \|(\Omega_j(\delta/c_\rho))_j\|_{\ell_{q'}}. \tag{10.27}$$

[*Proof of* (10.27). The lower bound is easy: first $\Omega(\delta) \ge \Omega^\circ(\delta)$ and then restrict the supremum over θ to the j-th shell, so that $\Omega^\circ(\delta) \ge \Omega_j(\delta)$ for each j. For the upper bound, first use (10.6) to reduce to showing $\Omega^\circ(\delta) \le \|(\Omega_j(\delta))_j\|_{\ell_{q'}}$. Then using the definition of $\theta^{(j)}$,

$$\begin{aligned}\Omega^\circ(\delta)^{q'} &= \sup\Big\{\sum_j \|\theta^{(j)}\|_{b'}^{q'} : \sum_j \|\theta^{(j)}\|_b^q \le C^q, \|\theta^{(j)}\|_\infty \le \delta\Big\} \\ &\le \sum_j \sup\big\{\|\theta^{(j)}\|_{b'}^{q'} : \|\theta^{(j)}\|_b^q \le C^q, \|\theta^{(j)}\|_\infty \le \delta\big\}\end{aligned}$$

since doing the maximizations separately can only increase the supremum. The final expression is just $\sum_j \Omega_j^{q'}(\delta)$ and so the upper bound follows.]

In view of (10.26) we can use the ℓ_p-modulus results to compute $\Omega_j(\delta)$ by making the substitutions

$$n_j = 2^j, \qquad \delta_j = 2^{a'j}\delta, \qquad C_j = C2^{-(a-a')j}.$$

We will now show that $\Omega_j(\delta)$ decays geometrically away from a single critical level, i.e. there exists $j_* \in \mathbb{R}$ and $\kappa = \kappa(\alpha, \alpha', p, p') > 0$ such that

$$\Omega_j(\delta) \le \delta^r C^{1-r} 2^{-\kappa|j-j_*|}. \tag{10.28}$$

'Sparse' case $p < p'$. We use the lower panel of Figure 10.1: as $\delta = \delta_j$ increases, the three zones for W translate into three zones for $j \to \Omega_j$, illustrated in the top panel of Figure 10.3.

Zone (i): $\delta_j < C_j n_j^{-1/p}$. This corresponds to

$$2^{(a+1/p)j} = 2^{(\alpha+1/2)j} < C/\delta,$$

so that the zone (i)/(ii) boundary occurs at j_0 satisfying $2^{(\alpha+1/2)j_0} = C/\delta$. In zone (i),

$$\Omega_j^{p'} = n_j \delta_j^{p'} = \delta^{p'} 2^{(1+p'a')j},$$

and with $n_0 = 2^j$, the maximum possible, this is a 'dense' zone.

At the boundary j_0, on setting $r_0 = (\alpha - \alpha')/(\alpha + 1/2)$, we have

$$\Omega_{j_0} = \delta 2^{j_0(a'+1/p')} = \delta(C/\delta)^{(\alpha'+1/2)/(\alpha+1/2)} = C^{1-r_0}\delta^{r_0}.$$

Zone (ii): $C_j n_j^{-1/p} < \delta_j < C_j$. The right inequality corresponds to $\delta < C 2^{-aj}$, so that the zone (ii)/(iii) boundary occurs at j_1 satisfying $2^{aj_1} = C/\delta$. In zone (ii),

$$\Omega_j^{p'} \approx C_j^p \delta_j^{p'-p} = C^p \delta^{p'-p} 2^{-(pa-p'a')j},$$

and observe using $a = \alpha + 1/2 - 1/p$ etc., that

$$pa - p'a' = p(\alpha + 1/2) - p'(\alpha' + 1/2)$$

is positive in the regular zone and negative in the logarithmic zone, so that Ω_j is geometrically decreasing in the regular zone and geometrically increasing in the logarithmic zone. The least favorable configuration has non-zero cardinality

$$n_0 = (C_j/\delta_j)^p = (C/\delta)^p 2^{-paj} = 2^{pa(j_1-j)},$$

decreasing from 2^{j_0} at $j = j_0$ to 1 at $j = j_1$, so this is a zone of increasing sparsity.

Zone (iii): $C_j < \delta_j$. In this sparse zone, $n_0 = 1$ and

$$\Omega_j^{p'} = C_j^{p'} = C^{p'} 2^{-p'(a-a')j},$$

where we note that for $p < p'$,

$$a - a' = \alpha - \alpha' - (1/p - 1/p') = \tilde{\alpha} > 0,$$

by our hypothesis. Define also

$$r_1 = 1 - a'/a = \tilde{\alpha}/(\alpha + 1/2 - 1/p) = r_L. \tag{10.29}$$

At the boundary j_1 we then have

$$\Omega_{j_1} = C 2^{-(a-a')j_1} = C(\delta/C)^{(a-a')/a} = C^{1-r_1} \delta^{r_1}.$$

The dense case, $p \geq p'$ is simpler. We refer to the bottom panel of Figure 10.3.

Zone (i) $\delta_j < C_j n_j^{-1/p}$. This zone is the same as in the sparse case, so for $j \leq j_0$ defined by $2^{(\alpha+1/2)j_0} = C/\delta$, we have

$$\Omega_j^{p'} = \delta^{p'} 2^{(1+p'a')j} = \delta^{p'} 2^{(\alpha+1/2)p'j}$$

and at the boundary level j_0, again $\Omega_{j_0} = C^{1-r_0} \delta^{r_0}$ with r_0 as before.

Zone (ii) $C_j n_j^{-1/p} < \delta_j$. We now have

$$\Omega_j^{p'} = n_j^{1-p'/p} C_j^{p'} = C^{p'} 2^{-(\alpha-\alpha')p'j}$$

and $\Omega_j = \Omega_{j_0} 2^{-(\alpha-\alpha')(j-j_0)}$.

Again we see that the geometric decay property (10.28) holds, with $j_* = j_0$ and $r = r_0$, and as at all levels j, the least favorable configuration at level j_0 is dense, $n_0 = 2^{j_0}$.

To summarize, we have established the geometric decay (10.28), under the assumptions of the Theorem 10.10, and outside the critical case $(\alpha + 1/2)p = (\alpha' + 1/2)p'$. In the regular case, $j_* = j_0$ and $r = r_0$ and the least favorable configuration at level j_0 is dense,

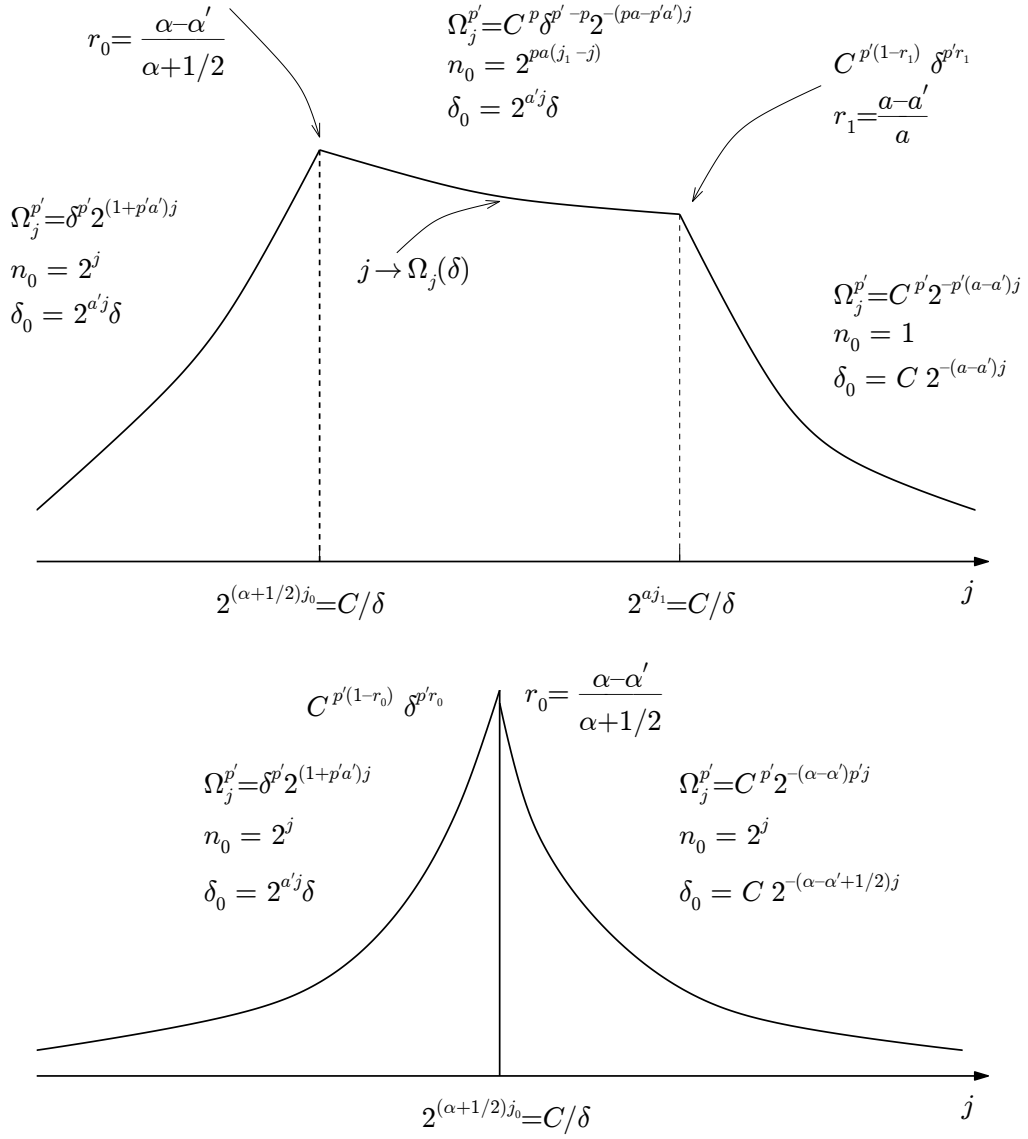

Figure 10.3 Schematic of the Besov modulus $\Omega_j(\delta)$, defined by (10.26), when viewed as a function of level j, with δ, C held fixed. Top panel is 'sparse' case, $p < p'$ (in the regular zone), bottom is 'dense' case $p \geq p'$

$n_0 = 2^{j_0}$. In the logarithmic case, $j_* = j_1$ and $r = r_1$, and the least favorable configuration at level j_1 is sparse, $n_0 = 1$.

The evaluation (10.22) follows from this and (10.27).

Evaluation of Besov tail widths These can be reduced to calculations on Besov shells by

the same approach as used to prove (10.28). If we set

$$\Delta_j = \sup\{\|\theta^{(j)}\|_{b'} : \|\theta^{(j)}\|_b \le C\},$$

then the full tail width is related to these shell widths by

$$\Delta_{J+1} \le \Delta(2^J; \Theta, \|\cdot\|_{b'}) \le \|(\Delta_j)_{j>J}\|_{\ell_{q'}}. \tag{10.30}$$

Using Besov shell identity (10.25),

$$\begin{aligned}\Delta_j &= 2^{ja'} \sup\{\|\theta_j\|_{p'} : \|\theta_j\|_p \le C2^{-aj}\} \\ &= 2^{ja'} W_{2^j;p',p}(\infty, C2^{-aj}).\end{aligned}$$

Substituting the identity (10.13), $W_n(\infty, C) = n^{(1/p'-1/p)_+}C$, we find

$$\Delta_j = 2^{ja'}2^{j(1/p'-1/p)_+}C2^{-aj} = C2^{-j\tilde{\alpha}}.$$

In view of (10.30), the full tail bias $\Delta(2^J;\Theta)$ is equivalent to $\Delta_J = C2^{-J\tilde{\alpha}} = Cn^{-\tilde{\alpha}}$. Indeed $\sum_{j>J}\Delta_j^{q'} \le C^{q'}\sum_{j>J}2^{-\tilde{\alpha}q'j}$, and so $\Delta(2^J) \le C2^{-\tilde{\alpha}J}(2^{\tilde{\alpha}q'}-1)^{-1/q'}$. This completes the proof of (10.23).

We now verify that the assumption $\alpha > 1/p$ (continuity) guarantees negligibility of the tail bias term: $\Delta(n) = o(\Omega(n^{-1/2}))$. From (10.23), $\Delta(n) = O(n^{-\tilde{\alpha}})$, while from (10.22), $\Omega(n^{-1/2}) \asymp n^{-r/2}$, so it is enough to verify that $\tilde{\alpha} > r/2$. If $\mathbf{p}$ is in the logarithmic zone, this is immediate when $\alpha > 1/p$.

If $\mathbf{p}$ is in the regular zone, the condition $\tilde{\alpha} > r/2$ becomes $\alpha - \alpha' - (1/p - 1/p')_+ > (\alpha-\alpha')/(2\alpha+1)$. If $p' \le p$ this is trivial, while for $p' > p$ it is the same as

$$\frac{2\alpha}{2\alpha+1}(\alpha-\alpha') > (1/p - 1/p').$$

Now the condition for $\mathbf{p}$ to be regular, namely $(2\alpha'+1)/(2\alpha+1) < p/p'$, is equivalent to the previous display with the right side replaced by $\alpha p(1/p - 1/p')$. So, again using $\alpha > 1/p$, we are done.

10.8 Lower Bounds

We again use the device of Besov shells to reduce to previous results obtained for ℓ_p balls and their associated least favorable configurations.

For a shell at any level j, we have $\|\theta\|_{b'} \ge \|\theta^{(j)}\|_{b'}$ and also $\Theta^{(j)} \subset \Theta$, and so

$$\sup_{\Theta} P\{\|\hat{\theta} - \theta\|_{b'} \ge \gamma\} \ge \sup_{\Theta^{(j)}} P\{\|\hat{\theta}^{(j)} - \theta^{(j)}\|_{b'} \ge \gamma\}. \tag{10.31}$$

Now since $\|\theta^{(j)}\|_{b'} = 2^{a'j}\|\theta_{j\cdot}\|_{p'}$ and since $\theta^{(j)} \in \Theta^{(j)}$ if and only if $\|\theta_{j\cdot}\|_p \le C2^{-aj}$, the right hand side above equals

$$\sup_{\Theta_{2^j,p}(C2^{-aj})} P\{\|\hat{\theta}_{j\cdot} - \theta_{j\cdot}\|_{p'} \ge \gamma 2^{-a'j}\}. \tag{10.32}$$

Regular case. The Besov shell we use corresponds to the critical level $j_0 = p(\alpha)\log_2(C/\delta)$,

where $p(\alpha) = 2/(2\alpha + 1)$ and we set $\delta = \epsilon = n^{-1/2}$. The setting is 'dense' because (cf. top panel of Figure 10.3) there are $n_0 = 2^{j_0}$ non-zero components with size $\delta_0 = 2^{j_0 a'}\epsilon$.

Hence, we apply the dense ℓ_p-ball modulus lower bound, Proposition 10.4, to $\Theta_{2^{j_0},p}(C2^{-aj_0})$. Hence, comparing (10.32) and (10.14), we are led to equate

$$\gamma 2^{-a' j_0} = c_{p'} W_{2^{j_0}}(\epsilon, C2^{-aj_0}),$$

after putting $c_{p'} = (\pi_0/2)^{1/p'}$. Recalling the definition of the shell modulus, (10.26), we get

$$\gamma = c_{p'} \Omega_{j_0}(\epsilon).$$

Because of the geometric decay of the shell modulus away from j_0, compare (10.28), there exists $c_1 = c_1(\mathbf{p})$ for which

$$\Omega(\epsilon) \le c_1 \Omega_{j_0}(\epsilon). \tag{10.33}$$

Combining the prior two displays, we can say that $\gamma \ge c_2 \Omega(\epsilon)$ and hence

$$\inf_{\hat\theta} \sup_{\Theta} P\{\|\hat\theta - \theta\|_{b'} \ge c_2 \Omega(\epsilon)\} \ge 1 - e^{-2n_0\pi_0^2}.$$

Here $n_0 = 2^{j_0} = (C/\epsilon)^{p(\alpha)} = (C\sqrt{n})^{p(\alpha)} \to \infty$ as $n \to \infty$, and so the regular case part of Theorem 10.10 is proven.

Logarithmic case. From the modulus calculation, we expect the least favorable configurations to be at shells near j_1 and to be highly sparse, perhaps a single spike. We therefore use the lower bounds derived for the 'bounded single spike' parameter spaces $\Theta_n(\tau)$ introduced at (8.58).

First we note that if $\delta_j \le C2^{-aj}$, then $\Theta_{2^j}(\delta_j) \subset \Theta_{2^j,p}(C2^{-aj})$. Fix $\eta > 0$. If also $\delta_j \le \epsilon\sqrt{(2-\eta)\log 2^j}$, then from Proposition 10.4(ii) we can say that

$$\inf_{\hat\theta} \sup_{\Theta_{2^j,p}(C2^{-aj})} P\{\|\hat\theta_{j\cdot} - \theta_{j\cdot}\|_{p'} \ge \delta_j/2\} \ge \pi_\eta(2^j).$$

Bearing in mind the two conditions on δ_j, it is clear that the largest possible value for δ_j is

$$\bar\delta_j = \min\{\epsilon\sqrt{(2-\eta)\log 2^j}, C2^{-aj}\}.$$

The implied best bound in (10.31) that is obtainable using the j-th shell is then given by the solution to $\gamma_j 2^{-a' j} = \bar\delta_j/2$, namely

$$\gamma_j = \tfrac{1}{2} 2^{a' j} \bar\delta_j.$$

Let $\bar\jmath_1 = \max\{j : \epsilon\sqrt{(2-\eta)\log 2^j} \le C2^{-aj}\}$. It is clear that γ_j is increasing for $j \le \bar\jmath_1$ and (since $a > a'$) decreasing for $j > \bar\jmath_1$, so our best shell bound will be derived from $\gamma_{\bar\jmath_1}$. Since we only observe data for levels $j < \log_2 n = \log_2 \epsilon^{-2}$, we also need to check that $\bar\jmath_1 < \log_2 \epsilon^{-2}$, and this is done below. To facilitate the bounding of $\gamma_{\bar\jmath_1}$, we first observe that from the definition of $\bar\jmath_1$, it follows that

$$2^{-a-1} \cdot C2^{-a\bar\jmath_1} \le \bar\delta_{\bar\jmath_1} \le C2^{-a\bar\jmath_1}, \tag{10.34}$$

and, after inserting again $\bar{\delta}_{\bar{j}_1} = c_\eta \epsilon \sqrt{\bar{j}_1}$,

$$c_3 \left(\frac{\epsilon \sqrt{\bar{j}_1}}{C} \right)^{1/a} \le 2^{-\bar{j}_1} \le c_4 \left(\frac{\epsilon \sqrt{\bar{j}_1}}{C} \right)^{1/a}, \tag{10.35}$$

where c_3, c_4 depend on a and η. After taking logarithms in the right bound, we obtain

$$\bar{j}_1 + (2a)^{-1} \log_2 \bar{j}_1 \ge \log_2 (C^{1/a} c_4^{-1}) + a^{-1} \log_2 \epsilon^{-1}.$$

Since $\log_2 \epsilon^{-1} = (\log_2 e) \log \epsilon^{-1} > \log \epsilon^{-1}$, it follows that for $\epsilon < \epsilon_1(a, C)$,

$$\bar{j}_1 \ge a^{-1} \log \epsilon^{-1}. \tag{10.36}$$

From the left bound in (10.35), we have $\bar{j}_1 \le (2a)^{-1} \log_2 \epsilon^{-2} + \log_2 (C^{1/a} c_3^{-1}) < \log_2 \epsilon^{-2}$ for $\epsilon < \epsilon_2(a, C)$ since $2a > 1$. Hence, as claimed, $\bar{j}_1 < \log_2 n$ for ϵ small.

Using (10.34), (10.35) and (10.36) in turn, along with definition (10.29) for r_L, we find

$$\gamma_{\bar{j}_1} \ge 2^{-a-2} C 2^{-(a-a')\bar{j}_1} \ge c C \left(\frac{\epsilon \sqrt{\bar{j}_1}}{C} \right)^{\frac{a-a'}{a}} \ge c C^{1-r_L} [\epsilon \sqrt{\log \epsilon^{-1}}]^{r_L} \ge c \Omega(\epsilon \sqrt{\log \epsilon^{-1}}),$$

where the constant $c = c(\mathbf{p})$ may differ each time.

Returning to (10.31) and inserting $\gamma_{\bar{j}_1}$, we have

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\|_{b'} \ge c \Omega(\epsilon \sqrt{\log \epsilon^{-1}})\} \ge \pi_\eta(2^{\bar{j}_1})$$

for $\epsilon < \epsilon(a, C)$; i.e. for $n > n(a, C)$. From (10.36) it is clear that $\bar{j}_1 \to \infty$ as $n \to \infty$ so that $\pi_\eta(2^{\bar{j}_1}) \to 1$.

10.9 Further Details

Here is a proof of the special case of the Cramér-Chernoff result that we need, using a standard change of measure argument. For much more, see texts on large deviations such as Dembo and Zeitouni (2010). Let P_π denote a binomial distribution Bin (n_0, π), and let B denote the corresponding random variable. The likelihood ratio

$$\frac{dP_{\pi_0}}{dP_{\pi_1}} = \left(\pi_0/\pi_1\right)^B \left(\bar{\pi}_0/\bar{\pi}_1\right)^{n_0 - B}.$$

Defining $\lambda = \log \pi_0/\pi_1$ and $\bar{\lambda} = \log \bar{\pi}_0/\bar{\pi}_1$, rewrite the loglikelihood 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 $\{B \le E_{\pi_0} B\} = \{L \ge E_{\pi_0} L\}$, while $E_{\pi_0} L = n_0 K(\pi_0, \pi_1)$. Consequently, using Markov's inequality along with $E_{\pi_1} e^L = 1$,

$$P_{\pi_1}\{B \le n_0 \pi_0\} = P_{\pi_1}\{e^L \ge e^{n_0 K}\} \le e^{-n_0 K} E_{\pi_1} e^L = e^{-n_0 K}.$$

For the bound on the divergence

$$\begin{aligned} K(\pi_0, \pi_1) &= \pi_0 \log \frac{\pi_0}{\pi_1} + (1-\pi_0) \log \frac{1-\pi_0}{1-\pi_1} \\ &= \pi_0 \int_{\pi_0}^{\pi_1} \frac{-du}{u} + (1-\pi_0) \int_{\pi_0}^{\pi_1} \frac{du}{1-u} \\ &= \int_{\pi_0}^{\pi_1} \frac{u-\pi_0}{u(1-u)} du \geq 4 \int_{\pi_0}^{\pi_1} (u-\pi_0) du = 2(\pi_1-\pi_0)^2. \end{aligned}$$

This is a very special case of the Pinsker inequality $K(P,Q) \geq \frac{1}{2}\|P-Q\|_1^2$, see e.g. Tsybakov (2009, Lemma 2.5).

10.10 Notes

The literature on optimal recovery goes back to Golomb and Weinberger (1959) and a 1965 Moscow dissertation of Smolyak. See also Micchelli (1975); Micchelli and Rivlin (1977) and Donoho (1994); the last cited 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 θ.

The material in this chapter is drawn from Donoho (1992a); Donoho et al. (1995) and Donoho et al. (1997). The second of these includes historical and bibliographical material and a contributed discussion.

While the Besov shell structure emerges naturally here in the study of $\sqrt{2\log n}$ thresholding, it provides a basic point of reference for studying properties of other threshold selection schemes over the same range of **p**. For example, this structure is used heavily in Johnstone and Silverman (2005b) to study wavelet shrinkage using an empirical Bayes choice of threshold, introduced in Section 7.6.

Exercises

10.1 *(ρ-convexity of Besov sequence norm.)* Show that $\|\cdot\|_{b^\alpha_{p,q}}$ is ρ-convex with $\rho = \min(1,p,q)$. [Hint: consider first $p < 1 \wedge q$ and use ℓ_r-norm triangle inequalities for $r = q/p$. Modify the approach slightly for the case $q < 1 \wedge p$.]

11

Penalization and Oracle Inequalities

The investigation of sparsity in previous chapters has been satisfied with demonstrating the optimality of estimators by showing that they achieve minimax risks or rates up to terms logarithmic in sample size or noise level. In this chapter and the next, our ultimate goal is to obtain sharper bounds on rates of convergence - in fact exactly optimal rates, rather than ones with spurious log terms.

The tools used for this purpose are of independent interest. These include model selection via penalized least squares, where the penalty function is not ℓ_2 or even ℓ_1 but instead a function of the *number* of terms in the model. We will call these *complexity penalties*.

Some of the arguments work for general (i.e. non-orthogonal) linear models, so we begin with this important framework. We do not use this extra generality in this book, nor pursue the now substantial literature on sparsity based oracle inequalities for linear and non-linear models (see the Chapter Notes for some references). Instead, we derive a bound adequate for our later results on sharp rates of convergence.

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 "$k \log(p/k)$" penalties, in which, roughly speaking, the penalty to enter the k^{th} variable is a function that decreases with k approximately like $2\zeta \log(p/k)$, for $\zeta \geq 1$.

Section 11.1 begins in the linear model setting with all subsets regression and introduces penalized least squares estimation with penalties that depend on the size of the subset or model considered.

Section 11.2 pauses to specialize to the case of orthogonal design–equivalent to the sequence model–in order to help motivate the class of penalties to be studied. We show the connection to thresholding, importantly with the thresholds $\hat{t}_{\text{pen}}$ now depending on the data, and decreasing as the size $\hat{k}$ of selected subset increases. The $k \log(p/k)$ class of penalties is motivated by connection to the expected size of coefficients–Gaussian order statistics–in a null model.

In Section 11.3, we present the main oracle inequality for a class of penalties including $\text{pen}(k) = \zeta k[1+\sqrt{2\log(p/k)}]^2$ for $\zeta > 1$. The Gaussian concentration of measure inequality of Section 2.8 plays an important role. Indeed, in considering all subsets of p variables, 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.

Section 11.4 applies the oracle inequality to the Gaussian sequence model to obtain non-asymptotic upper bounds for minimax risk over ℓ_p balls $\Theta_{n,p}(C)$. Lower bounds are ob-

tained via embedded product spaces. Both bounds are expressed in terms of a 'control function' $r_{n,p}(C)$, which when $p < 2$, clearly exhibits the transition from a zone of 'sparse' least favorable configurations to a 'dense' zone. This is the second main theorem of the chapter, and these conclusions are basic for the sharp rate results on estimation over Besov classes in Chapter 12.

The remaining sections contain various remarks on and extensions of these results. Section 11.5 provides more detail on the connection between the complexity penalty functions and thresholding, and on several equivalent forms of the theoretical complexity in the orthogonal case.

Section 11.6 remarks on the link between traditional forward and backward stepwise model selection criteria and the class of penalties considered in this chapter.

Section 11.7 prepares for results in the next chapter on sharp rates for linear inverse problems by presenting a modified version of the main oracle inequality.

11.1 All subsets regression and complexity penalized least squares

We begin with the usual form of the general linear model with Gaussian errors:

$$y = X\beta + \epsilon z = \mu + \epsilon z, \qquad z \sim N_n(0, I). \tag{11.1}$$

There are n observations y and p unknown parameters β, connected by an $n \times p$ design matrix X with columns

$$X = [x_1, \cdots, x_p].$$

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 ϵ 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 white noise model (1.21), $dY(t) = f(t)dt + \epsilon dW(t)$, and suppose that the observed data are inner products of Y with n orthonormal functions $\psi_1, \ldots, \psi_n$. Thus

$$y_i = \langle f, \psi_i \rangle + \epsilon z_i, \qquad i = 1, \ldots, n.$$

Now consider the possibility of approximating f by elements from a *dictionary* $\mathcal{D} = \{\phi_1, \phi_2, \ldots, \phi_p\}$. 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. As a simple illustration, the ψ_i might be sinusoids at the first n frequencies, while the dictionary elements might allow a much finer sampling of frequencies

$$\phi_\kappa(t) = \sin(2\pi\kappa t/p), \qquad \kappa = 1, \ldots, p = n^\beta \gg n.$$

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 $f = \sum_{\kappa=1}^{p} \beta_\kappa \phi_\kappa$, then these observation equations become an instance of the general linear model (11.1) with

$$X_{i\kappa} = \langle \psi_i, \phi_\kappa \rangle.$$

Again, the hope is that one can find an estimate $\hat{\beta}$ for which only a small number of components $\hat{\beta}_\kappa \neq 0$.

All subsets regression. To each subset $K \subset \{1, \ldots, p\}$ of cardinality $n_K = |K|$ corresponds a regression model which fits only the variables x_κ for $\kappa \in K$. [1] The possible fitted vectors μ that could arise from these variables lie in the model space

$$S_K = \text{span}\{x_\kappa : \kappa \in K\}.$$

The dimension of S_K is at most n_K, and could be less in the case of collinearity.

Let P_K denote orthogonal projection onto S_K: the least squares estimator $\hat{\mu}_K$ of μ is given by $\hat{\mu}_K = P_K y$. We include the case $K = \varnothing$, writing $n_\varnothing = 0$, $S_\varnothing = \{0\}$ and $\hat{\mu}_\varnothing(y) \equiv 0$. The issue in all subsets regression consists in deciding how to select a subset $\hat{K}$ on the basis of data y: the resulting estimate of μ is then $\hat{\mu} = P_{\hat{K}} y$.

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 $\hat{\mu} = X\hat{\beta}$ that it generates:

$$E\|X\hat{\beta} - X\beta\|^2 = E\|\hat{\mu} - \mu\|^2.$$

The mean of a projection estimator $\hat{\mu}_K$ is just the projection of μ, namely $E\hat{\mu}_K = P_K\mu$, while its total variance is $\epsilon^2 \text{tr} P_K = \epsilon^2 \dim S_K$. From the variance-bias decomposition of MSE,

$$E\|\hat{\mu}_K - \mu\|^2 = \|P_K\mu - \mu\|^2 + \epsilon^2 \dim S_K.$$

A *saturated* model arises from any subset with $\dim S_K = n$, so that $\hat{\mu}_K = y$ "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 μ lies close to a low rank subspace — $\mu \approx \sum_{\kappa \in K} \beta_\kappa x_\kappa$ for $|K|$ small—then $\hat{\mu}_K$ offers substantial risk savings over a saturated model. Thus, it seems that one would wish to expand the dictionary $\mathcal{D}$ 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 also leads to a natural generalization of the notion of ideal risk introduced at (8.33) in Chapter 8.3, and which in this chapter we denote by $\mathcal{R}_1(\mu, \epsilon)$. For each mean vector μ, there will be an optimal model subset $K = K(\mu)$ which attains the ideal risk

$$\mathcal{R}_1(\mu, \epsilon) = \min_K \|\mu - P_K\mu\|^2 + \epsilon^2 \dim S_K.$$

[1] We use κ to denote the index of a variable x_κ to distinguish from the use of k in $k \log(p/k)$.

[2] 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 $\hat{\beta}$: $y^* = X\beta + \epsilon z^*$. A natural prediction of y^* is $X\hat{\beta}$, and its mean squared error, averaging over the distributions of both z and z^*, is

$$E\|y^* - X\hat{\beta}\|^2 = E\|X\beta - X\hat{\beta}\|^2 + n\epsilon^2,$$

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.]

Of course, this choice $K(\mu)$ is not available to the statistician, since μ is unknown. The challenge, taken up below, is to see to what extent penalized least squares estimators can "mimic" ideal risk, in a fashion analagous to the mimicking achieved by threshold estimators in the orthogonal setting.

Complexity penalized least squares. The residual sum of squares (RSS) of model K is

$$\|y - \hat{\mu}_K\|^2 = \|y - P_K y\|^2,$$

and clearly decreases as the model K increases. To discourage simply using a saturated model, or more generally to discourage overfitting, we introduce a penalty on the size of the model, $\text{pen}(n_K)$, that is increasing in n_K, and then define a complexity criterion

$$C(K, y) = \|y - P_K y\|^2 + \epsilon^2 \text{pen}(n_K). \tag{11.2}$$

The complexity penalized RSS estimate $\hat{\mu}_{\text{pen}}$ is then given by orthogonal projection onto the subset that minimizes the penalized criterion:

$$\hat{K}_{\text{pen}} = \text{argmin}_K \, C(K, y), \qquad \hat{\mu}_{\text{pen}} = P_{\hat{K}_{\text{pen}}} y. \tag{11.3}$$

Corresponding to the 'empirical' complexity $C(K, y)$ is a theoretical complexity $C(K, \mu)$ based on the 'true' value μ. We will be interested in the extent to which the estimator $\hat{\mu}_{\text{pen}}$ can mimick the minimal theoretical complexity $\min_K C(K, \mu)$.

The simplest penalty function grows linearly in the number of variables in the model:

$$\text{pen}_0(k) = \lambda^2 k, \tag{11.4}$$

where we will take $\lambda^2 = \lambda_p^2$ to be roughly of order $2 \log p$. The well known AIC criterion would set $\lambda^2 = 2$. In our Gaussian setting, it is equivalent to Mallows' C_p, compare (2.54). This is effective for selection among a nested sequence of models, but is known to overfit in all-subsets settings, e.g. Nishii (1984); Foster and George (1994) and Exercise 11.1. The BIC criterion (Schwarz, 1978) puts $\lambda^2 = \log n$. Foster and George (1994) took $\lambda^2 = 2 \log p$, dubbing it RIC for Risk Inflation Criterion.

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:

$$\begin{aligned} \min_K C(K, \mu) &= \min_K \left[\|\mu - P_K \mu\|^2 + \epsilon^2 \text{pen}_0(n_K) \right] \\ &\le \lambda_p^2 \min \left[\|\mu - P_K \mu\|^2 + \epsilon^2 n_K \right] = \lambda_p^2 \, \mathcal{R}_1(\mu, \epsilon). \end{aligned}$$

Let $\lambda_p = \zeta(1 + \sqrt{2 \log p})$ for $\zeta > 1$. Then for penalty function (11.4) and arbitary μ, it will be shown that

$$E \|\hat{\mu}_{\text{pen}} - \mu\|^2 \le a(\zeta) \lambda_p^2 \mathcal{R}_1(\mu, \epsilon) + b(\zeta) \epsilon^2,$$

where bounds for $a(\zeta), b(\zeta)$ are given in Theorem 11.3 below, in particular $a(\zeta)$ is decreasing in ζ. 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, $\text{pen}(\dim S_K)$, instead of $\text{pen}(n_K)$. However, when $k \to \text{pen}(k)$ is strictly monotone, this will yield the same models as minimizing (11.2), since a collinear model will always be rejected in favor of a sub-model with the same span.

11.2 Orthogonal Case

For this section we specialize to the n-dimensional white Gaussian sequence model:

$$y_i = \mu_i + \epsilon z_i, \qquad i = 1, \ldots, n, \qquad z_i \overset{iid}{\sim} N(0,1). \tag{11.5}$$

This is the canonical form of the more general orthogonal regression setting $Y = X\beta + \epsilon Z$, with N dimensional response and n dimensional parameter vector β linked by an orthogonal design matrix X satisfying $X^T X = I_n$, and with the noise $Z \sim N_n(0, I)$. This reduces to (11.5) after premultiplying by X^T and setting $y = X^T Y$, $\mu = \beta$ and $z = X^T Z$.

We will see in this section that, in the orthogonal regression setting, the penalized least squares estimator can be written in terms of a penalty on the number of non-zero elements (Lemma 11.1). There are also interesting connections to hard thresholding, in which the threshold is data dependent. We then use this connection to help motivate the form of penalty function to be used in the oracle inequalities of the next section.

The columns of the design matrix implicit in (11.5) 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 $K \subset \{1, \ldots, n\}$ is simply given by co-ordinate projection P_K:

$$(P_K y)_i = \begin{cases} y_i & i \in K \\ 0 & i \notin K. \end{cases}$$

The complexity criterion (11.2) becomes

$$C(K, y) = \sum_{i \notin K} y_i^2 + \epsilon^2 \text{pen}(n_K),$$

where $n_K = |K|$ still. Using $|y|_{(l)}$ to denote the order statistics of $|y_i|$, in decreasing order, we can write

$$\min_K C(K, y) = \min_{0 \le k \le n} \sum_{l > k} |y|_{(l)}^2 + \epsilon^2 \text{pen}(k). \tag{11.6}$$

There is an equivalent form of the penalized least squares estimator in which the model selection aspect is less explicit, being replaced by a minimization over μ. Let $N[\mu] = \#\{i : \mu_i \neq 0\}$ be the number of non-zero components of μ.

Lemma 11.1 *Suppose that $k \to pen(k)$ is monotone increasing. In orthogonal model* (11.5)*, the penalized least squares estimator* (11.3) *can be written*

$$\hat{\mu}_{pen}(y) = \operatorname*{argmin}_{\mu} \|y - \mu\|^2 + \epsilon^2 \text{pen}(N[\mu]).$$

Proof The model space S_K corresponding to subset K consists of vectors μ whose components μ_i vanish for $i \notin K$. Let $S_K^+ \subset S_K$ be the subset on which the components $\mu_i \neq 0$ for every $i \in K$. The key point is that on S_K^+ we have $N[\mu] = n_K$. Since $\mathbb{R}^n$ is the disjoint union of all S_K^+—using $\{0\}$ in place of $S_\varnothing^+$—we get

$$\min_{\mu} \|y - \mu\|^2 + \epsilon^2 \text{pen}(N[\mu]) = \min_K \min_{\mu \in S_K^+} \|y - \mu\|^2 + \epsilon^2 \text{pen}(n_K).$$

The minimum over $\mu \in S_K^+$ can be replaced by a minimum over $\mu \in S_K$ without changing the value because if $\mu \in S_K \backslash S_K^+$ there is a smaller subset K' with $\mu \in S_{K'}^+$—here we use

monotonicity of the penalty. So we have recovered precisely the model selection definition (11.3) of $\hat{\mu}_{\text{pen}}$. □

Remark. Essentially all the penalties considered in this chapter are monotone increasing in k. Our shorthand terminology "$2k \log(p/k)$ penalties" has the minor defect that $k \to k \log(p/k)$ is decreasing for $k \geq p/e$. However this is inessential and easily fixed, for example, by using $k \to k(1 + \log(p/k))$ which *is* increasing for $0 \leq k \leq p$.

Connection with thresholding. When $\text{pen}_0(k) = \lambda^2 k$, we recover the ℓ_0 penalty and the corresponding estimator is hard thresholding at $\epsilon\lambda$, as seen in Section 2.3. To explore the connection with thresholding for more general penalties, consider the form $\text{pen}(k) = \sum_{l=1}^{k} t_{n,l}^2$. Then the optimal value of k in (11.6) is

$$\hat{k} = \underset{k}{\text{argmin}} \sum_{l>k} |y|_{(l)}^2 + \epsilon^2 \sum_{l=1}^{k} t_{n,l}^2. \tag{11.7}$$

We show that $\hat{\mu}_{\text{pen}}$ corresponds to hard thresholding at a *data-dependent* value $\hat{t}_{\text{pen}} = t_{n,\hat{k}}$.

Proposition 11.2 *If $k \to t_{n,k}$ is strictly decreasing, then*

$$|y|_{(\hat{k}+1)} < \epsilon t_{n,\hat{k}} \leq |y|_{(\hat{k})}, \tag{11.8}$$

and

$$\hat{\mu}_{\text{pen},i}(y) = \begin{cases} y_i & |y_i| \geq \epsilon t_{n,\hat{k}} \\ 0 & \textit{otherwise}. \end{cases} \tag{11.9}$$

Figure 11.1 illustrates the construction of estimated index $\hat{k}$ and threshold $\hat{t}_{\text{pen}}$.

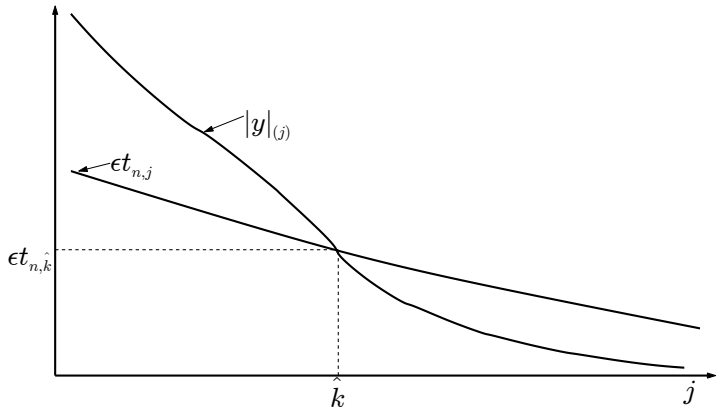

Figure 11.1 Schematic showing construction of data dependent threshold from the sequence $t_{n,l}$ and ordered data magnitudes $|y|_{(l)}$.

Proof Let $S_k = \epsilon^2 \sum_{l=1}^{k} t_{n,l}^2 + \sum_{l>k} |y|_{(l)}^2$. For notational simplicity, we write t_k instead of $t_{n,k}$. We have

$$S_k - S_{k-1} = \epsilon^2 t_k^2 - |y|_{(k)}^2.$$

Now $\hat{k}$ minimizes $k \to S_k$, so in particular we have both $S_{\hat{k}} \le S_{\hat{k}-1}$ and $S_{\hat{k}} \le S_{\hat{k}+1}$, which respectively imply that

$$|y|_{(\hat{k})} \ge \epsilon t_{\hat{k}}, \qquad \text{and} \qquad |y|_{(\hat{k}+1)} \le \epsilon t_{\hat{k}+1} < \epsilon t_{\hat{k}},$$

where at the last strict inequality we used the assumption on t_k. Together, these inequalities yield (11.8) and also the set identity

$$\{i : |y_i| \ge \epsilon t_{\hat{k}}\} = \{i : |y_i| \ge |y|_{(\hat{k})}\}.$$

Since the set on the right side is $\hat{K}$, we have shown (11.9). □

Gaussian order statistics and $2k \log(n/k)$ *penalties.* The "z-test" for $\mu_i = 0$ is based on $|z_i|/\epsilon$. [If ϵ^2 were unknown and estimated by an independent χ^2 variate, this would be a t-statistic $z_i/\hat{\epsilon}$.] Under the null model $\mu = 0$, it is natural to ask for the magnitude of the k-th largest test statistic $|z|_{(k)}/\epsilon$ as a calibration for whether to enter the k-th variable into the model. It can be shown that if $k_n = o(n)$, then as $n \to \infty$,

$$E|z|_{(k_n)} = \sqrt{2\log(n/k_n)}(1 + o(1)), \tag{11.10}$$

so that a plausible threshold $t^2_{n,k}$ for entry of the k-th variable is of order $2\log(n/k)$. Hence $\mathrm{pen}(k)$ itself is of order $2k\log(n/k)$. [The justification of this is similar to that for (11.11) and Exercise 11.3.

A heuristic justification for (11.10) comes from the equivalence of the event $\{|z|_{(k)} \ge t\}$ with $\{\#\{i : |z_i| \ge t\} \ge k\}$. Under the null model $\mu = 0$, the latter is a binomial event, so

$$P\{|z|_{(k)} \ge t\} = P\{\mathrm{Bin}(n, 2\tilde{\Phi}(t)) \ge k\}.$$

Setting the mean value $2n\tilde{\Phi}(t)$ of the binomial variate equal to k yields $t_{n,k} \sim \sqrt{2\log(n/k)}$. Exercise 11.4 has a more formal demonstration.

Example. FDR estimation. In Chapter 7.6, (7.28) described a data dependent threshold choice that is closely related to penalized estimation as just described with $t_{n,k} = z(kq/2n)$. Indeed, let $\hat{k}_F = \max\{k : |y|_{(k)} \ge \epsilon t_{n,k}\}$ denote the last crossing, and consider also the first crossing $\hat{k}_G + 1 = \min\{k : |y|_{(k)} < \epsilon t_{n,k}\}$. If $\hat{k}_{\mathrm{pen}}$ denotes the penalized choice (11.7), then Section 11.6 shows that

$$\hat{k}_G \le \hat{k}_{\mathrm{pen}} \le \hat{k}_F$$

and in simulations it is often found that all three agree.

In Exercise 11.3, it is verified that if k, possibly depending on n, is such that $k/n \to 0$ as $n \to \infty$, then

$$t^2_{n,k} \le (1/k)\sum_1^k t^2_{n,l} \le t^2_{n,k} + 2 \sim 2\log(n/k \cdot 2/q) \tag{11.11}$$

and hence that

$$\mathrm{pen}(k) \sim 2k\log(n/k \cdot 2/q). \tag{11.12}$$

11.3 Oracle Inequalities

Consider a penalty of the form

$$\text{pen}(k) = \zeta k(1 + \sqrt{2L_k})^2 \qquad (\zeta > 1, L_k \geq 0). \tag{11.13}$$

This form is chosen both to approximate the $2k \log(n/k)$ class just introduced in the orthogonal case, in which $p = n$, and to be convenient for theoretical analysis. The penalty reduces to pen_0 of (11.4) if L_k is identically constant. Typically, however, the sequence $L_k = L_{p,k}$ is chosen so that $L_{p,k} \geq \log(p/k)$ and is decreasing in k. We will see in Section 11.4 and the next chapter that this property is critical for removing logarithmic terms in convergence rates. As a concession to our theoretical analysis, we need $\zeta > 1$ and the extra "1" in (11.13) for the technical arguments. The corresponding thresholds are then a bit larger than would otherwise be desirable in practice.

We abuse notation a little and write L_K for L_{n_K}. Associated with the penalty is a constant

$$M = \sum_K e^{-L_K n_K}, \tag{11.14}$$

where the sum is taken over all subsets of $\{1, \ldots, p\}$.

Here are a couple of examples of penalty functions and associated evaluations of M.

(i) Penalty (11.4), namely $\text{pen}_0(k) = \lambda_p^2 k$, takes the form (11.13) if λ_p is written as $\lambda_p = \sqrt{\zeta}(1 + \sqrt{2\alpha \log p})$ and we set $L_k \equiv \alpha \log p$. Since there are at most $\binom{p}{k} \leq p^k/k!$ subsets $K \subset \{1, \ldots, p\}$ having cardinality $n_K = k$,

$$M = \sum_K e^{-n_K \alpha \log p} = \sum_{k=0}^{p} \binom{p}{k} e^{-k\alpha \log p} \leq \sum_{k=0}^{\infty} \frac{(p \cdot p^{-\alpha})^k}{k!} \leq \exp(p^{1-\alpha}).$$

The last term is uniformly bounded in p so long as $\alpha \geq 1$. Thus, convergence of (11.14) and the theorem below require that $\lambda_p^2 \sim \zeta \cdot (2 \log p)$ or larger when p is large.

(ii) Now suppose that $L_k = \log(p/k) + \alpha'$, with $L_0 = \log p$. Proceeding much as above,

$$M = \sum_{k=0}^{p} \binom{p}{k} e^{-kL_k} \leq \sum_{0}^{\infty} \frac{p^k}{k!} \left(\frac{k}{p}\right)^k e^{-\alpha' k} \leq 1 + \sum_{k=1}^{\infty} \frac{1}{\sqrt{2\pi k}} e^{-(\alpha'-1)k}, \tag{11.15}$$

using Stirling's formula, $k! = \sqrt{2\pi k} k^k e^{-k+\theta}$, with $(12k)^{-1} \leq \theta \leq (12k+1)^{-1}$. The last sum converges so long as $\alpha' > 1$.

The first main result of this chapter is an oracle inequality for the penalized least squares estimator.

Theorem 11.3 *In model* (11.1)*, let $\hat{\mu}$ be a penalized least squares estimator (11.2)–(11.3) for a penalty pen(k) depending on $\zeta > 1$ and constant M defined at* (11.13) *and* (11.14)*. Then there exist constants $a = a(\zeta), b = b(\zeta)$ such that for all μ,*

$$E\|\hat{\mu}_{\text{pen}} - \mu\|^2 \leq a \min_K C(K, \mu) + bM\epsilon^2. \tag{11.16}$$

The constants may be taken respectively as $a(\zeta) = (3\zeta + 1)(\zeta + 1)^2/(\zeta - 1)^{-3}$ and as $b(\zeta) = 4\zeta(\zeta + 1)^3/(\zeta - 1)^{-3}$.

The constants a and b are not sharp; note however that $a(\zeta)$ is decreasing in ζ with limit 3 as $\zeta \to \infty$. Section 11.7 has a variant of this result designed for (mildly) correlated noise and inverse problems.

Proof 1°. Writing $y = \mu + \epsilon z$ and expanding (11.2), we have

$$C(\hat{K}, y) = \|\hat{\mu}_{\hat{K}} - \mu\|^2 + 2\epsilon\langle \mu - \hat{\mu}_{\hat{K}}, z\rangle + \epsilon^2\|z\|^2 + \epsilon^2 \mathrm{pen}(n_{\hat{K}}).$$

We aim to use the minimizing property, $C(\hat{K}, y) \le C(K, y)$, to get an upper bound for $\|\hat{\mu}_{\hat{K}} - \mu\|^2$. To this end, for an arbitrary index K, writing $P_K^\perp = I - P_K$ and $\mu_K = P_K\mu$, we have

$$\begin{aligned}\|P_K^\perp y\|^2 &= \|P_K^\perp \mu\|^2 + 2\epsilon\langle P_K^\perp \mu, P_K^\perp z\rangle + \epsilon^2\|P_K^\perp z\|^2 \\ &\le \|P_K^\perp \mu\|^2 + 2\epsilon\langle \mu - \mu_K, z\rangle + \epsilon^2\|z\|^2.\end{aligned}$$

Consequently

$$C(K, y) = \|P_K^\perp y\|^2 + \epsilon^2 \mathrm{pen}(n_K) \le C(K, \mu) + 2\epsilon\langle \mu - \mu_K, z\rangle + \epsilon^2\|z\|^2.$$

By definition, $C(\hat{K}, y) \le C(K, y)$, so combining the corresponding equations and cancelling terms yields a bound for $\hat{\mu}_{\hat{K}} - \mu$:

$$\|\hat{\mu}_{\hat{K}} - \mu\|^2 \le C(K, \mu) + 2\epsilon\langle \hat{\mu}_{\hat{K}} - \mu_K, z\rangle - \epsilon^2 \mathrm{pen}(n_{\hat{K}}). \tag{11.17}$$

The merit of this form is that we can hope to appropriately apply the Cauchy-Schwarz inequality, (11.21) below, to the linear term $\langle \hat{\mu}_{\hat{K}} - \mu_K, z\rangle$, and take a multiple of $\|\hat{\mu}_{\hat{K}} - \mu\|^2$ over to the left side to develop a final bound.

2°. We outline the strategy based on (11.17). We construct an increasing family of sets Ω_x for $x > 0$, with $P(\Omega_x^c) \le Me^{-x}$ and then show for each $\eta \in (0, 1)$ that there are constants $a_0(\eta), b_0(\eta)$ for which we can bound the last two terms of (11.17): when $\omega \in \Omega_x$,

$$2\epsilon\langle \hat{\mu}_{\hat{K}} - \mu_K, z\rangle - \epsilon^2 \mathrm{pen}(n_{\hat{K}}) \le (1-\eta^2)\|\hat{\mu}_{\hat{K}} - \mu\|^2 + a_0(\eta)C(K, \mu) + b_0(\eta)\epsilon^2 x. \tag{11.18}$$

Assuming for now the truth of (11.18), we can insert it into (11.17) and move the squared error term on the right side to the left side of (11.17). We get

$$\|\hat{\mu}_{\hat{K}} - \mu\|^2 \le \eta^{-2}(1 + a_0(\eta))C(K, \mu) + \eta^{-2}b_0(\eta)\epsilon^2 X, \tag{11.19}$$

where $X(\omega) = \inf\{x : \omega \in \Omega_x\}$. Clearly $X(\omega) > x$ implies that $\omega \notin \Omega_x$, and so using the bound on $P(\Omega_x^c)$ gives $EX = \int_0^\infty P(X > x)dx \le M$. Hence, taking expectations, then minimizing over K, and setting $a_1(\eta) = \eta^{-2}(1 + a_0(\eta))$ and $b_1(\eta) = \eta^{-2}b_0(\eta)$, we get

$$E\|\hat{\mu}_{\hat{K}} - \mu\|^2 \le a_1(\eta) \min_K C(K, \mu) + b_1(\eta)\epsilon^2 M. \tag{11.20}$$

3°. We turn to the derivation of (11.18). Consider a pair of subsets K, K': we imagine K as fixed, and K' as being variable (it will later be set to $\hat{K}$.) To effectively bound the inner product term, introduce random variables

$$\chi_{K,K'} = \sup\{\langle u, z\rangle/\|u\|, u \in S_K \oplus S_{K'}\} = \|P_{K\cup K'} z\|,$$

where $P_{K\cup K'}$ is orthogonal projection on $S_K \oplus S_{K'}$. Hence

$$\langle \hat{\mu}_{K'} - \mu_K, z\rangle \le \|\hat{\mu}_{K'} - \mu_K\| \cdot \chi_{K,K'}, \tag{11.21}$$

since $z \sim N_n(0, I)$ and clearly $\chi^2_{K,K'} \sim \chi^2_{(d)}$ with degrees of freedom $d = \dim(S_K \oplus S_{K'}) \le n_K + n_{K'}$.

Now use the Lipschitz concentration of measure bound (2.76), which says here that $P\{\chi_{(d)} > \sqrt{d} + t\} \le e^{-t^2/2}$ for all $t \ge 0$, and, crucially, for all non-negative integer d. (If $d = 0$, then $\chi_{(0)} = 0$.) For arbitrary $x > 0$, let $E_{K'}(x)$ be the event

$$\chi_{K,K'} \le \sqrt{n_K + n_{K'}} + \sqrt{2(L_{K'} n_{K'} + x)}, \tag{11.22}$$

and in the concentration bound set $t^2 = 2(L_{K'} n_{K'} + x)$. Let $\Omega_x = \cap_{K'} E_{K'}(x)$, so that

$$P(\Omega_x^c) \le e^{-x} \sum_{K'} e^{-L_{K'} n_{K'}} = M e^{-x}.$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ twice in (11.22) and then combining with (11.21), we conclude that on the set Ω_x,

$$\langle \hat{\mu}_{K'} - \mu_K, z \rangle \le \|\hat{\mu}_{K'} - \mu_K\| \cdot [\sqrt{n_{K'}}(1 + \sqrt{2L_{K'}}) + \sqrt{n_K} + \sqrt{2x}].$$

The key to extracting $\|\hat{\mu}_{K'} - \mu_K\|^2$ with a coefficient less than 1 is to use the inequality $2\alpha\beta \le c\alpha^2 + c^{-1}\beta^2$, valid for all $c > 0$. Thus, for $0 < \eta < 1$ and $c = 1 - \eta$,

$$\begin{aligned} &2\epsilon \langle \hat{\mu}_{K'} - \mu_K, z \rangle \\ &\quad \le (1-\eta)\|\hat{\mu}_{K'} - \mu_K\|^2 + \frac{\epsilon^2}{1-\eta}\Big[\sqrt{n_{K'}}(1 + \sqrt{2L_{K'}}) + \sqrt{n_K} + \sqrt{2x}\Big]^2. \end{aligned} \tag{11.23}$$

Now use this trick again, now in the form $(\alpha+\beta)^2 \le (1+\eta)\alpha^2 + (1+\eta^{-1})\beta^2$, on each of the right side terms. In the first term, use $\|\hat{\mu}_{K'} - \mu_K\| \le \|\hat{\mu}_{K'} - \mu\| + \|\mu_K - \mu\|$ and get

$$(1-\eta^2)\|\hat{\mu}_{K'} - \mu\|^2 + (\eta^{-1} - \eta)\|\mu_K - \mu\|^2.$$

In the second, use $\mathrm{pen}(n_{K'}) = \zeta n_{K'}(1 + \sqrt{2L_{K'}})^2$ and get

$$\frac{1+\eta}{1-\eta}\zeta^{-1}\epsilon^2 \mathrm{pen}(n_{K'}) + \frac{1+\eta^{-1}}{1-\eta}\epsilon^2(2n_K + 4x).$$

Now, choose η so that $(1+\eta)/(1-\eta) = \zeta$, and then move the resulting $\epsilon^2 \mathrm{pen}(n_{K'})$ term to the left side of (11.23). To bound the rightmost terms in the two previous displays, set

$$a_0(\eta) = \max\left\{\eta^{-1} - \eta, \frac{1+\eta^{-1}}{1-\eta}\frac{2}{\zeta}\right\}, \qquad b_0(\eta) = \frac{4(1+\eta^{-1})}{1-\eta}, \tag{11.24}$$

and note that $\zeta n_K \le \mathrm{pen}(n_K)$. Finally, setting $K' = \hat{K}$, we recover the desired inequality (11.18) and hence (11.20). Inserting $\eta = (\zeta - 1)/(\zeta + 1)$ gives the values for $a(\zeta) = a_1(\eta)$ and $b(\zeta) = b_1(\eta)$ quoted in the Theorem. □

Orthogonal Case. An important simplification occurs in the theoretical complexity $C(K, \mu)$ in the orthogonal case. As in Section 11.2, but now using μ rather than y,

$$C(K, \mu) = \sum_{i \notin K} \mu_i^2 + \epsilon^2 \mathrm{pen}(n_K)$$

The minimum theoretical complexity is denoted by

$$\mathcal{R}(\mu, \epsilon) = \min_{K} C(K, \mu). \tag{11.25}$$

Then, as at (11.6) we have

$$\mathcal{R}(\mu, \epsilon) = \min_{0 \le k \le n} \sum_{l>k} \mu_{(l)}^2 + \epsilon^2 \mathrm{pen}(k). \tag{11.26}$$

Let us note some interesting special cases, for which we write the penalty in the form

$$\mathrm{pen}(k) = k\lambda_k^2.$$

First, with $\lambda_k \equiv \lambda$, so that $\mathrm{pen}(k) = \lambda^2 k$ is proportional to k, we verify that

$$\mathcal{R}(\mu, \epsilon) = \sum_k \min(\mu_k^2, \lambda^2 \epsilon^2), \tag{11.27}$$

and the ideal risk $\mathcal{R}_1(\mu, \epsilon)$ of Chapter 8 corresponds to choice $\lambda \equiv 1$. In addition, the oracle inequalities of Sections 2.7 and 8.3, in the specific form (2.73), can be seen to have the form (11.29).

Second, if $k \to \lambda_k$ is monotone, there is a co-ordinatewise upper bound for theoretical complexity with a form generalizing (11.27).

Lemma 11.4 *If pen$(k) = k\lambda_k^2$ with $k \to \lambda_k^2$ is non-increasing, then*

$$\mathcal{R}(\mu, \epsilon) \le \sum_{k=1}^{n} \min(\mu_{(k)}^2, \lambda_k^2 \epsilon^2).$$

Proof Without loss of generality, put $\epsilon = 1$. Let $k' = \max\{k \ge 1 : \lambda_k \epsilon \le |\mu|_{(k)}\}$ if such an index exists, otherwise set $k' = 0$. Let $M_k = \sum_{j>k} \mu_{(j)}^2$. Since both $k \to \lambda_k$ and $k \to |\mu|_{(k)}$ are non-increasing, we have

$$\begin{aligned} \sum_{k=1}^{n} \min(\mu_{(k)}^2, \lambda_k^2) &= \sum_{1}^{k'} \min(\mu_{(k)}^2, \lambda_k^2) + M_{k'} \ge k'(\mu_{(k')}^2 \wedge \lambda_{k'}^2) + M_{k'} \qquad (11.28) \\ &= k'\lambda_{k'}^2 + M_{k'} \ge \min_k M_k + k\lambda_k^2. \end{aligned}$$

□

Corollary 11.5 *In the special case of orthogonal model* (11.5)*, the bound of Theorem 11.3 becomes*

$$\begin{aligned} E\|\hat{\mu}_{\mathrm{pen}} - \mu\|^2 &\le a\mathcal{R}(\mu, \epsilon) + bM\epsilon^2 \qquad (11.29) \\ &\le a \sum_{k=1}^{n} \min(\mu_{(k)}^2, \lambda_k^2 \epsilon^2) + bM\epsilon^2, \end{aligned}$$

where the second inequality assumes also pen$(k) = k\lambda_k^2$ with $k \to \lambda_k^2$ non-increasing.

11.4 Non-asymptotic bounds for ℓ_p-balls

Suppose that we observe data from the n-dimensional Gaussian signal plus noise model (11.5), and that μ is constrained to lie in a ball of radius C defined by the ℓ_p norm:

$$\Theta = \Theta_{n,p}(C) = \{\mu \in \mathbb{R}^n : \sum_{i=1}^{n} |\mu_i|^p \le C^p\}. \tag{11.30}$$

We seek 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 later see that these lead to the optimal rates of convergence for these classes of parameter spaces.

The non-asymptotic bounds will have a number of consequences. We will again see a sharp transition between the sparse case $p < 2$, in which non-linear methods clearly outperform linear ones, and the more traditional setting of $p \ge 2$.

The upper bounds will illustrate the use of the $2k\log(n/k)$ type oracle inequalities established in the last section. They will also be used in the next 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 and products of "spikes".

While the non-asymptotic bounds have the virtue of being valid for finite $\epsilon > 0$, their disadvantage is that the upper and lower bounds may be too conservative. The optimal constants can be found from a separate asymptotic analysis as $\epsilon \to 0$, see Chapter 13 below.

A control function. The non-asymptotic bounds will be expressed in terms of a control function $r_{n,p}(C,\epsilon)$ defined separately for $p \ge 2$ and $p < 2$. The control function captures key features of the minimax risk $R_N(\Theta_{n,p}(C),\epsilon)$ but is more concrete, and is simpler in form. As with the minimax risk, it can be reduced by rescaling to a unit noise version

$$r_{n,p}(C,\epsilon) = \epsilon^2 r_{n,p}(C/\epsilon). \tag{11.31}$$

For $p < 2$, the control function is given by

$$r_{n,p}(C) = \begin{cases} C^2 & \text{if } C \le \sqrt{1+\log n}, \\ C^p\left[1+\log(n/C^p)\right]^{1-p/2} & \text{if } \sqrt{1+\log n} < C \le n^{1/p}, \\ n & \text{if } C \ge n^{1/p}. \end{cases} \tag{11.32}$$

See Figure 11.2. As will become evident from the proof, the three zones correspond to situations where the least favorable signals are 'near zero', 'sparse' and 'dense' respectively. A little calculus shows that $C \to r_{n,p}(C)$ is monotone increasing in C for $0 < p < 2$, except at the discontinuity at $\check{C} = \sqrt{1+\log n}$. This discontinuity is not serious; for example we have the simple bound

$$\frac{r_{n,p}(\check{C}-)}{r_{n,p}(\check{C}+)} \le 2, \tag{11.33}$$

valid for all n and $p \in (0,2)$. Here $r(\check{C}\pm)$ denotes the limit of $r(C)$ as $C \searrow \check{C}$ and $C \nearrow \check{C}$ respectively. Indeed, the left side is equal to

$$[1-(p/2)\check{C}^{-2}\log\check{C}^2]^{p/2-1} \le 2^{1-p/2} \le 2$$

using the crude bound $x^{-1}\log x \le 1/2$ valid for $x \ge 1$. Numerical work would show that the bound is actually considerably less than 2, especially for n large.

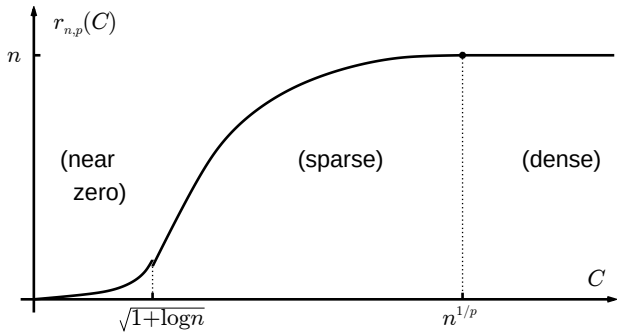

Figure 11.2 Schematic of the control function (11.32) for $p < 2$, showing the three zones for C, and the discontinuity at $C = \sqrt{1+\log n}$.

For $p \ge 2$, the control function is simpler:

$$r_{n,p}(C) = \begin{cases} n^{1-2/p}C^2 & \text{if } C \le n^{1/p}, \\ n & \text{if } C \ge n^{1/p}. \end{cases} \tag{11.34}$$

To show that the bounds provided by the control function can be attained, we use a penalized least squares estimator $\hat{\mu}_P$ for a specific choice of penalty of the form (11.13). Thus $\mathrm{pen}(k) = k\lambda_k^2$ with $\lambda_k = \sqrt{\zeta}(1+\sqrt{2L_{n,k}})$, and

$$L_{n,k} = (1+2\beta)\log(n\gamma/k). \tag{11.35}$$

[By convention, set $L_{n,0} = L_{n,1}$ and $\lambda_0 = \lambda_1$, in any case $\mathrm{pen}(0) = 0$.]

The parameter β is included for applications to inverse problems in Chapter 12; for most other purposes we can take $\beta = 0$. The constant γ is included to obtain convergence of the sum defining the constant M: when $\beta = 0$ we need $\gamma > e$ (compare (11.15)).

Here is the main result of this section, saying that the minimax MSE for ℓ_p-balls is described, up to constants, by the control function $r_{n,p}(C)$, and that penalized least squares estimation can globally mimic the control function.

Theorem 11.6 *For $n \ge 1$, and $0 < p \le \infty, 0 < C < \infty$, there exist constants a_1 and $c_1(\zeta,\beta,\gamma)$ so that*

$$a_1 r_{n,p}(C,\epsilon) \le R_N(\Theta_{n,p}(C)) \tag{11.36}$$

$$\le \sup_{\Theta_{n,p}(C)} E\|\hat{\mu}_P - \mu\|^2 \le c_1[\epsilon^2 + r_{n,p}(C,\epsilon)]. \tag{11.37}$$

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.

Constants convention. In the statement and proof, we use c_i to denote constants that depend on (ζ, β, γ) and a_i to stand for absolute constants. While information is available about each such constant, we have not tried to assemble this into the final constants a_1 and c_1 above, as they would be far from sharp.

Proof 1°. *Upper Bounds.* We may assume, by scaling, that $\epsilon = 1$. As we are in the orthogonal setting, the oracle inequality of Theorem 11.3 combined with (11.25) and (11.26) takes the form

$$E\|\hat{\mu}_P - \mu\|^2 \le a\mathcal{R}(\mu) + bM,$$

where $a = a(\zeta), b = b(\zeta), M = M(\beta, \gamma)$ and

$$\mathcal{R}(\mu) = \min_{0\le k\le n} \sum_{j>k} \mu_{(j)}^2 + k\lambda_k^2. \tag{11.38}$$

For the upper bound, then, we need then to show that when $\mu \in \Theta_{n,p}(C)$,

$$\mathcal{R}(\mu) \le c(\zeta, \beta)(\log\gamma)\, r_{n,p}(C), \tag{11.39}$$

where the dependence on γ is made explicit for use in Section 12.4.

We might guess that worst case bounds for (11.38) occur at gradually increasing values of k as C increases. In particular, the extreme zones for C will correspond to $k = 0$ and n. It turns out that these two extremes cover most cases, and then the main interest in the proof lies in the sparse zone for $p < 2$.

Now to the details. Let us first note the simple bound

$$\lambda_k^2 \le 2\zeta(1 + c_\beta)(\log\gamma)(1 + \log(n/k)), \qquad 1 \le k \le n. \tag{11.40}$$

First put $k = n$ in (11.38). Using the previous display, we obtain

$$\mathcal{R}(\mu) \le c_3(\log\gamma)n \tag{11.41}$$

for $c_3 = 2\zeta(1 + c_\beta)$, valid for all C (and all p), but useful in the dense zone $C \ge n^{1/p}$.

For $p \ge 2$, simply by choosing $k = 0$ in (11.38), we also have

$$\mathcal{R}(\mu) \le n \cdot n^{-1}\sum \mu_k^2 \le n\left(n^{-1}\sum |\mu_k|^p\right)^{2/p} \le n^{1-2/p}C^2. \tag{11.42}$$

Combining the last two displays suffices to establish (11.39) in the $p \ge 2$ case.

For $p < 2$, note that $\sum |\mu_l|^p \le C^p$ implies that $|\mu|_{(l)} \le Cl^{-1/p}$, and hence that

$$\sum_{l>k}^{n} \mu_{(l)}^2 \le C^{2-p}(k+1)^{1-2/p}\sum_{l>k} |\mu|_{(l)}^p \le C^2(k+1)^{1-2/p}.$$

We can now dispose of the extreme cases. Putting $k = 0$, we get $\mathcal{R}(\mu) \le C^2$, as is needed for $C \le \sqrt{1 + \log n}$. For $C \ge n^{1/p}$, again use bound (11.41) corresponding to $k = n$.

We now work further on bounding $\mathcal{R}(\mu)$ for the range $C \in [\sqrt{1+\log n}, n^{1/p}]$. Inserting the last display into (11.38) and ignoring the case $k = n$, we obtain

$$\mathcal{R}(\mu) \le \min_{0 \le k < n} C^2(k+1)^{1-2/p} + k\lambda_k^2. \tag{11.43}$$

Now observe from (11.40) that $\lambda_{k-1}^2 \le c(\zeta,\beta)(\log\gamma)(1+\log(n/k))$ for $1 \le k \le n$. Putting this into (11.43), we arrive at

$$\mathcal{R}(\mu) \le c(\zeta,\beta)(\log\gamma) \min_{1 \le k \le n} \{C^2 k^{1-2/p} + k(1+\log(n/k))\}. \tag{11.44}$$

We now pause to consider the lower bounds, as the structure turns out to be similar enough that we can finish the argument for both bounds at once in part 3° below.

2°. *Lower Bounds.* For $p \ge 2$, we use a hypercube lower bound. Since $\Theta_{n,p}(C)$ contains the cube $[-Cn^{-1/p}, Cn^{-1/p}]^n$, we have by (4.25) and (4.40), with $a_2 = 2/5$,

$$R_N(\Theta) \ge n\rho_N(Cn^{-1/p}, 1) \ge a_2 n \min(C^2 n^{-2/p}, 1).$$

For $p < 2$, we will use products of the single spike parameter sets $\Theta_m(\tau)$ consisting of a single non-zero component in $\mathbb{R}^m$ of magnitude at most τ, compare (8.58). Proposition 8.17 gave a lower bound for minimax mean squared error over such single spike sets.

Working in $\mathbb{R}^n$, for each fixed number k, one can decree that each block of $[n/k]$ successive coordinates should have a single spike belonging to $\Theta_{[n/k]}(\tau)$. Since minimax risk is additive on products, Proposition 4.16, we conclude from Proposition 8.17 that for each k

$$R_N(\Pi_1^k \Theta_{[n/k]}(\tau)) \ge a_3 k(\tau^2 \wedge (1+\log[n/k]).$$

Now $\Theta_{n,p}(C)$ contains such a product of k copies of $\Theta_{[n/k]}(\tau)$ if and only if $k\tau^p \le C^p$, so that we may take $\tau = Ck^{-1/p}$ in the previous display. Therefore

$$R_N(\Theta_{n,p}(C)) \ge a_4 \max_{1 \le k \le n} C^2 k^{1-2/p} \wedge (k + k\log(n/k)), \tag{11.45}$$

where we also used $1 + \log[x] \ge (1+\log x)/(1+\log 2)$ for $x \ge 1$.

Again we draw two quick conclusions: for $C \le \sqrt{1+\log n}$, the choice $k = 1$ yields the bound C^2, while for $C \ge n^{1/p}$, the choice $k = n$ gives the lower bound n.

3°. *Completion of proof.* Let us summarize the remaining task. Define two functions

$$g(x) = C^2 x^{1-2/p}, \qquad h(x) = x + x\log(n/x).$$

Then with $p < 2$, and for $\sqrt{1+\log n} \le C \le n^{1/p}$, and abbreviating $r_{n,p}(C)$ in this range by $r(C) = C^p\big[1+\log(n/C^p)\big]^{1-p/2}$, we seek absolute constants a_5 and a_6 so that

$$a_5 r(C) \le \max_{1 \le k \le n} g(k) \wedge h(k), \qquad \min_{1 \le k \le n} g(k) + h(k) \le a_6 r(C). \tag{11.46}$$

Since g is decreasing and h is increasing for $0 \le x \le n$, it is natural to look for $x_\star = x_\star(C) \in \mathbb{R}$ at which $g(x_\star) = h(x_\star)$, compare Figure 11.3. At the point of intersection,

$$x_\star = C^p\big[1+\log(n/x_\star)\big]^{-p/2}, \tag{11.47}$$

$$g(x_\star) = C^p\big[1+\log(n/x_\star)\big]^{1-p/2}. \tag{11.48}$$

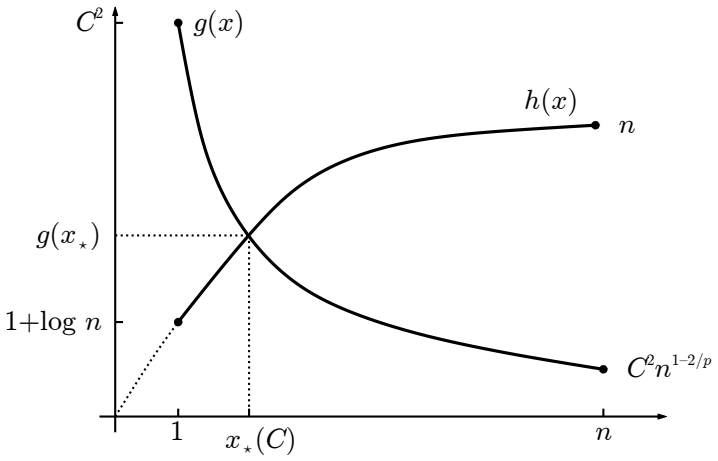

Figure 11.3 Diagram of functions g and h and their intersection, when $p < 2$ and $\sqrt{1+\log n} \leq C \leq n^{1/p}$.

It is clear from Figure 11.3 that $C \to x_\star(C)$ is strictly increasing, with

$$x_\star(\sqrt{1+\log n}) = 1, \quad \text{and} \quad x_\star(n^{1/p}) = n.$$

Hence $1 \leq x_\star \leq n$ if and only if $\sqrt{1+\log n} \leq C \leq n^{1/p}$; this explains the choice of transition points for C in the definition of $r(C)$.

We now relate the intersection value $g(x_\star(C))$ to $r(C)$; we will show that

$$r(C) \leq g(x_\star(C)) \leq 2r(C). \tag{11.49}$$

One direction is easy: putting $x_\star \leq n$ into (11.47) shows that $x_\star \leq C^p$, and hence from (11.48) that $g(x_\star) \geq r(C)$. For the other direction, make the abbreviations

$$s = 1 + \log(n/x_\star), \qquad \text{and} \qquad t = 1 + \log(n/C^p).$$

Now taking logarithms in equation (11.47) shows that $s \leq t + \log s$. But $\log s \leq s/2$ (since $s \geq 1$ whenever $x_\star \leq n$), and so $s \leq 2t$. Plugging this into (11.48), we obtain (11.49).

We are not quite done since the extrema in the bounds (11.46) 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

$$\tfrac{1}{2} h(\lceil x \rceil) \leq h(x) \leq 2h(\lfloor x \rfloor). \tag{11.50}$$

Indeed, h is concave and $h(0) = 0$, and so for x positive, $h(x)/2 \leq h(x/2)$. Since h is increasing for $0 \leq y \leq n$, it follows that if $x \leq 2y$, then $h(x) \leq 2h(y)$. Since $x \geq 1$ implies both $x \leq 2\lfloor x \rfloor$ and $\lceil x \rceil \leq 2x$, the bounds (11.50) follow.

For the upper bound in (11.46), take $k = \lceil x_\star \rceil$: since g is decreasing, and using (11.50) and then (11.49), we find

$$\min_{1 \leq k \leq n} g + h \leq (g+h)(\lceil x_\star \rceil) \leq g(x_\star) + 2h(x_\star) = 3g(x_\star) \leq 6r(C).$$

For the lower bound, take $k = \lfloor x_\star \rfloor$, and again from the same two displays,

$$\max_{1\le k\le n} g \wedge h \ge (g \wedge h)(\lfloor x_\star \rfloor) = h(\lfloor x_\star \rfloor) \ge \tfrac{1}{2} h(x_\star) = \tfrac{1}{2} g(x_\star) \ge \tfrac{1}{2} r(C). \qquad \square$$

11.5 Penalties, thresholds, and theoretical complexities

In this section we make some further remarks on the special structure of penalized least squares estimation in the orthogonal case. We write the penalties studied in the last two sections in the form

$$\mathrm{pen}(k) = k\lambda_k^2, \qquad \lambda_k = \sqrt{\zeta}(1 + \sqrt{2L_k}).$$

We first describe a sense in which $\hat{\mu}_{\mathrm{pen}}$ corresponds to thresholding at a data-determined threshold $\lambda_{\hat{k}}$. Defining $t_k^2 = k\lambda_k^2 - (k-1)\lambda_{k-1}^2$, we can rewrite $\mathrm{pen}(k)$ in the form $\sum_1^k t_l^2$ needed for the thresholding result Proposition 11.2, which interprets $\hat{\mu}_{\mathrm{pen}}$ as hard thresholding at $\hat{t} = t_{\hat{k}}$ where $\hat{k} = |\hat{J}|$ is the size of the selected model.

It is then heuristically plausible that $t_k \approx \lambda_k$, but here is a more precise bound.

Lemma 11.7 *Suppose that the function $k \to L_k$ appearing in λ_k is decreasing, and for some constant $b \ge 0$ satisfies*

$$L_k \ge \max(\tfrac{1}{2}, 2b), \qquad k(L_{k-1} - L_k) \le b. \tag{11.51}$$

Then we have the bounds

$$\lambda_k - 4\zeta b/\lambda_k \le t_k \le \lambda_k.$$

Note in particular that if L_k is constant, then we can take $b = 0$ and $t_k = \lambda_k$. More generally, if $L_k = (1 + 2\beta)\log(n\gamma/k)$ for $\beta \ge 0$, then condition (11.51) holds with $b = 1 + 2\beta$ so long as $\gamma \ge e^2$. In sparse cases, $k = o(n)$, we have $\lambda_k \asymp \sqrt{\log n}$ and t_k gets closer to λ_k as n grows.

Proof From the definition of t_k^2 and the monotonicity of λ_k^2 we have

$$t_k^2 - \lambda_k^2 = (k-1)(\lambda_k^2 - \lambda_{k-1}^2) \le 0,$$

so that $t_k \le \lambda_k$. For the other bound, again use the definition of t_k^2, now in the form

$$\lambda_k - t_k \le \lambda_{k-1} - t_k = \frac{\lambda_{k-1}^2 - t_k^2}{\lambda_{k-1} + t_k} = k\frac{\lambda_{k-1} + \lambda_k}{\lambda_{k-1} + t_k}(\lambda_{k-1} - \lambda_k). \tag{11.52}$$

Setting $\delta = \lambda_k - t_k$ and $\Lambda = \lambda_{k-1} + \lambda_k$, this takes the form

$$\frac{\delta}{\Lambda} \le \frac{k(\lambda_{k-1} - \lambda_k)}{\lambda_{k-1} + t_k} = \frac{k(\lambda_{k-1} - \lambda_k)}{\Lambda - \delta}. \tag{11.53}$$

Using now the definition of λ_k, then the bounds $L_{k-1} \ge \frac{1}{2}$ and $k(L_{k-1} - L_k) \le b$, we find

$$k(\lambda_{k-1} - \lambda_k) = \sqrt{2\zeta}\frac{k(L_{k-1} - L_k)}{\sqrt{L_{k-1}} + \sqrt{L_k}} \le \frac{2\zeta \cdot b}{\sqrt{\zeta}(1 + \sqrt{2L_k})} = \frac{2\zeta b}{\lambda_k}. \tag{11.54}$$

The bound $L_k \geq 2b$ implies $\lambda_k^2 \geq 4\zeta b$, and if we return to first inequality in (11.53) and simply use the crude bound $t_k \geq 0$ and $\lambda_{k-1} \geq \lambda_k$ along with (11.54), we find that

$$\delta/\Lambda \leq 2\zeta b/\lambda_k^2 \leq 1/2.$$

Returning to the second inequality in (11.53), we now have $\delta/\Lambda \leq 2k(\lambda_{k-1} - \lambda_k)/\Lambda$, and again using (11.54), we get $\delta \leq 4\zeta b/\lambda_k$, which is the bound we claimed. □

Some equivalences. We have been discussing several forms of minimization that turn out to be closely related. To describe this, we use a modified notation. We consider

$$\mathcal{R}_S(s,\tau) = \min_{0\leq k\leq n} \sum_{l=1}^{k} s_l + \sum_{l=k+1}^{n} \tau_l, \tag{11.55}$$

$$\mathcal{R}_C(s,\tau) = \min_{0\leq k\leq n} k s_k + \sum_{l=k+1}^{n} \tau_l, \tag{11.56}$$

$$\mathcal{R}(s,\tau) = \sum_{k=1}^{n} \min(s_k, \tau_k). \tag{11.57}$$

With the identifications $s_k \leftrightarrow t_{n,k}^2$ and $\tau_k \leftrightarrow |y|_{(k)}^2$, the form $\mathcal{R}_S$ recovers the objective function in the thresholding formulation of penalization, (11.7). When using a penalty of the form $\mathrm{pen}(k) = k\lambda_k^2$, compare (11.26), we use a measure of the form $\mathcal{R}_C$. Finally, the co-ordinatewise minimum is perhaps simplest.

Under mild conditions on the sequence $\{s_k\}$, these measures are equivalent up to constants. To state this, introduce a hypothesis:

(H) The values $s_k = \sigma(k/n)$ for $\sigma(u)$ a positive decreasing function on $[0,1]$ with

$$\lim_{u\to 0} u\sigma(u) = 0, \qquad \sup_{0\leq u\leq 1} |u\sigma'(u)| \leq c_1.$$

For such a function, let $c_\sigma = 1 + c_1/\sigma(1)$.

A central example is given by $\sigma(u) = 2\log(\gamma/u)$, with $c_1 = 2$ and $c_\sigma = 1 + (\log\gamma)^{-1}$.

Proposition 11.8 *Let the sequence $\{s_k\}$ satisfy hypothesis (H). Let $\mathcal{R}_S, \mathcal{R}_C$ and $\mathcal{R}$ be the minima defined in* (11.55)–(11.56) *above. Then the measures are equivalent: for all non-negative, decreasing sequences $\tau \in \mathbb{R}^n$,*

$$c_\sigma^{-1}\mathcal{R}_S(s,\tau) \leq \mathcal{R}_C(s,\tau) \leq \mathcal{R}(s,\tau) \leq \mathcal{R}_S(s,\tau) \leq c_\sigma \mathcal{R}_C(s,\tau).$$

Remark. The central two inequalities, in which c_σ does not appear, are valid for any positive decreasing sequence $\{s_k\}$, without any need for hypothesis *(H)*.

Proof Consider first the bounds not involving the constant c_σ. The bound $\mathcal{R}_C \leq \mathcal{R}$ is precisely Lemma 11.4, while $\mathcal{R} \leq \mathcal{R}_S$ is immediate since each sum appearing in (11.55) is bounded below by $\sum \min(s_k, \tau_k)$. The bounds with c_σ will follow if we show that *(H)* implies $\sum_1^k s_l \leq c_\sigma k s_k$ for $k = 0, \ldots, n$. But

$$\sum_1^k s_l = \sum_1^k \sigma(l/n) \leq n \int_0^{k/n} \sigma(u)du,$$

and by partial integration

$$\int_0^v \sigma(u)du = v\sigma(v) + \int_0^v u|\sigma'(u)|du \le v[\sigma(v) + c_1] \le c_\sigma v\sigma(v).$$

Combining the previous two displays gives the bound we need. □

11.6 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:

$$\hat{J}_k = \text{argmax}_J\{\|P_J y\|^2 : n_J = k\},$$

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

$$\hat{k}_G = \min\{k : \|Q_{k+1}y\|^2 - \|Q_k y\|^2 \le \epsilon^2 t^2_{p,k+1}\}. \tag{11.58}$$

Note that we allow the threshold to depend on k: in practice it is often constant, but we wish to allow $k \to t^2_{p,k}$ 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

$$\hat{k}_F = \max\{k : \|Q_k y\|^2 - \|Q_{k-1}y\|^2 \ge \epsilon^2 t^2_{p,k}\}. \tag{11.59}$$

Remarks. 1. In the orthogonal case, $y_i = \mu_i + \epsilon z_i, i = 1, \ldots, n$ with order statistics $|y|_{(1)} \ge |y|_{(2)} \ge \ldots \ge |y|_{(n)}$, we find that

$$\|Q_k y\|^2 = \sum_{l=1}^{k} |y|^2_{(l)},$$

so that

$$\hat{k}_F = \max\{k : |y|_{(k)} \ge \epsilon t_{p,k}\}, \tag{11.60}$$

and that $\hat{k}_F$ agrees with the FDR definition (7.28) with $t_{p,k} = z(qk/2n)$. 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

$$\begin{aligned} S(k) &= \|y - Q_k y\|^2 + \epsilon^2 \sum_{l=1}^{k} t_{p,l}^2, \\ \hat{k}_2 &= \operatorname{argmin}_{0 \le k \le n} S(k). \end{aligned} \tag{11.61}$$

Thus the associated penalty function is $\mathrm{pen}(k) = \sum_1^k t_{p,l}^2$ and the corresponding estimator is given by (11.2) and (11.3).

The optimal model size for $\mathrm{pen}(k)$ is bracketed between the stepwise quantities.

Proposition 11.9 *Let $\hat{k}_G, \hat{k}_F$ be the forward and backward stepwise variable numbers defined at (11.58) and (11.59) respectively, and let $\hat{k}_2$ be the global optimum model size for pen(k) defined at (11.61). Then*

$$\hat{k}_G \le \hat{k}_2 \le \hat{k}_F.$$

Proof Since $\|y - Q_k y\|^2 = \|y\|^2 - \|Q_k y\|^2$,

$$S(k+1) - S(k) = \|Q_k y\|^2 - \|Q_{k+1} y\|^2 + \epsilon^2 t_{p,k+1}^2.$$

Thus

$$S(k+1) \begin{Bmatrix} < \\ = \\ > \end{Bmatrix} S(k) \quad \text{according as} \quad \|Q_{k+1} y\|^2 - \|Q_k y\|^2 \begin{Bmatrix} > \\ = \\ < \end{Bmatrix} \epsilon^2 t_{p,k+1}^2.$$

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 $S(\hat{k}_2 + 1) < S(\hat{k}_2)$. □

11.7 An oracle inequality for use in inverse problems

This section prepares the way for the use of the oracle inequalities and wavelet-vaguelette decomposition to provide sharp rates for a class of linear inverse problems in the next chapter. Two modifications in the basic result of Theorem 11.3 are needed: first, an extension to moderately correlated noise, and second a variation that treats the null model $J = \varnothing$ differently in the 'variance' term.

Suppose then that $y = \theta + \epsilon z$, now with z assumed to be zero-mean Gaussian, but *weakly correlated*: i.e.

$$\xi_0 I \preceq \mathrm{Cov}(z) \preceq \xi_1 I, \tag{11.62}$$

where $\xi_0 \le 1 \le \xi_1$ and $A \preceq B$ means that $B - A$ is non-negative definite. We continue to use the penalty $\mathrm{pen}(k) = \zeta k(1 + \sqrt{2L_k})^2$. In order to handle the "variance inflation" aspect of inverse problems, we want to replace the constant M in the variance term in (11.16) by one that excludes the zero model:

$$M' = \sum_{J \ne \varnothing} e^{-L_J n_J}. \tag{11.63}$$

[This is explained further at (11.66) and (12.38) below.]

Theorem 11.10 *Consider observations in the weakly correlated model* (11.62). *Let* $\hat{\mu}_{\text{pen}}$ *be a penalized least squares estimator of (11.2)- (11.3) for a penalty* (11.13) *and constant* M' *defined at* (11.63). *Then there exist constants* $a' = a'(\zeta), b = b(\zeta)$ *such that for all* μ

$$E\|\hat{\mu}_{\text{pen}} - \mu\|^2 \le a'(\zeta) \inf_J C(J,\mu) + b(\zeta)\xi_1 M'\epsilon^2. \tag{11.64}$$

The constant $a'(\zeta)$ *may be taken as* $4\zeta(\zeta+1)^2/(\zeta-1)^3$ *and* $b(\zeta)$ *as in Theorem 11.3.*

Remark. An alternative would be to modify the penalty to $\text{pen}(k) = \zeta\xi_1 k(1+\sqrt{2L_k})^2$, which would lead to larger thresholds corresponding to the largest noise direction in $\text{Cov}(z)$. In that case, the constant ζ in $a'(\zeta)$ and $b'(\zeta)$ would be replaced by $\xi_1\zeta$.

Proof 1°. We modify the proof of the previous theorem in two steps. First fix J and assume that $\text{Cov}(z) = I$. Let $E_{J'}(x)$ be defined as in (11.22), and then let $\Omega'_x = \cap_{J'\neq\varnothing} E_{J'}(x)$ and $X' = \inf\{x : \omega \notin \Omega'_x\}$. On the set $\hat{J} \neq \varnothing$, we have, as before,

$$\|\hat{\mu}_{\hat{J}} - \mu\|^2 \le a_1(\eta)C(J,\mu) + b_1(\eta)\epsilon^2 X'.$$

Now consider the event $\hat{J} = \varnothing$. First, note that if $\|\mu\|^2 \le \epsilon^2\text{pen}(1)$, we have on $\hat{J} = \varnothing$ that, for all J

$$\|\hat{\mu}_{\hat{J}} - \mu\|^2 = \|\mu\|^2 \le C(J,\mu).$$

Suppose, instead, that $\|\mu\|^2 \ge \epsilon^2\text{pen}(1)$, so that $C(J,\mu) \ge \epsilon^2\text{pen}(1)$ for all J—here we use the monotonicity of $k \to \text{pen}(k)$. Pick a J' with $n_{J'} = 1$; on Ω'_x we have

$$\langle z, -\mu_J\rangle \le \|\mu_J\| \cdot \chi_{J,J'} \le \|\mu_J\| \cdot [(1+\sqrt{2L_1}) + \sqrt{n_J} + \sqrt{2x}].$$

We now proceed as in the argument from (11.23) to (11.24), except that we bound $\epsilon^2\text{pen}(1) \le C(J,\mu)$, concluding that on Ω'_x and $\hat{J} = \varnothing$, we may use in place of (11.18),

$$\begin{aligned} 2\epsilon\langle z, -\mu_J\rangle &\le (1-\eta^2)\|\mu\|^2 + C(J,\mu) \\ &\quad + a_0(\eta)C(J,\mu) + b_0(\eta)\epsilon^2 x. \end{aligned}$$

Consequently, combining all cases

$$\|\hat{\mu}_{\hat{J}} - \mu\|^2 \le \eta^{-2}(2+a_0(\eta))C(J,\mu) + \eta^{-2}b_0(\eta)\epsilon^2 X',$$

which might be compared with (11.19). Taking expectations, then minimizing over J, we obtain again (11.20), this time with $a_2(\eta) = \eta^{-2}(2+a_0(\eta))$ and $b_1(\eta)$ unchanged, but with M' in place of M. Inserting $\eta = (\zeta-1)/(\zeta+1)$ gives $a'(\zeta) = a_2(\eta)$ and $b(\zeta) = b_1(\eta)$ as before.

2°. The extension to weakly correlated z is straightforward. We write $y = \mu + \epsilon_1 z_1$, where $\epsilon_1 = \sqrt{\xi_1}\epsilon$ and $\Sigma = \text{Cov}(z_1) \preceq I$. We apply the previous argument with ϵ, z replaced by ϵ_1 and z_1. The only point where the stochastic properties of z_1 are used is in the concentration inequality that is applied to $\chi_{J,J'}$. In the present case, if we put $z_1 = \Sigma^{1/2}Z$ for $Z \sim N(0,I)$, we can write

$$\chi_{J,J'} = \|P\Sigma^{1/2}Z\|,$$

where P denotes orthoprojection onto $S_J \oplus S_{J'}$. Since $\lambda_1(\Sigma^{1/2}) \le 1$, the map $Z \to$

$\chi_{J,J'}(Z)$ is Lipschitz with constant at most 1, so that the concentration bound applies. We remark also that

$$[E\chi_{J,J'}(Z)]^2 \le E\|P\Sigma^{1/2}Z\|^2 = E\mathrm{tr}\Sigma^{1/2}P\Sigma^{1/2}ZZ^{\mathrm{T}} = \mathrm{tr}P\Sigma \le \lambda_1(\Sigma)\mathrm{tr}P \le \mathrm{tr}P,$$

where the second last inequality uses von Neumann's trace inequality (e.g. Mirsky (1975), or via general results for unitarily invariant matrix norms, e.g. Bhatia (1997, Prop. IV.2.4)). □

In particular, we will in Chapter 12 make use of penalties for which

$$L_{n,k} = (1+2\beta)\log(\gamma_n n/k) \tag{11.65}$$

with $\gamma_n = \gamma\ell(n;\beta,\epsilon)$ where $\gamma > e$ and the function $\ell(n) \ge 1$ and may depend on β and the noise level ϵ. For this choice, the constant M' in (11.64) satisfies (after using the Stirling formula bound $k! > \sqrt{2\pi k}\,k^k e^{-k}$),

$$\begin{aligned} M' \le \sum_{k=1}^{n} \frac{n^k}{k!}\Big(\frac{k}{n\gamma_n}\Big)^{k(1+2\beta)} &\le \sum_{k=1}^{n} \frac{1}{\sqrt{2\pi k}}\left(\frac{k^{2\beta}}{n^{2\beta}}\frac{e}{\gamma_n^{1+2\beta}}\right)^k \\ &\le \frac{1}{n^{2\beta}\gamma_n}\sum_{k\ge 1}\frac{k^{2\beta}e}{\sqrt{2\pi k}}\left(\frac{e}{\gamma_n^{1+2\beta}}\right)^{k-1} \le \frac{C_{\beta,\gamma}}{n^{2\beta}\gamma_n}, \end{aligned} \tag{11.66}$$

for example if $\gamma_n \ge \gamma > e$.

The factor $n^{2\beta}$ in the denominator is crucial, and is the reason for the exclusion of $J = \varnothing$ in the definition (11.63) of M'.

11.8 Notes

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 and Stine (1997) $\mathrm{pen}(k) = \epsilon^2\sum_1^k 2\log(n/j)$ via information theory.
- George and Foster (2000) Empirical Bayes approach. [$\mu_i \overset{i.i.d.}{\sim} (1-w)\delta_0 + wN(0,C)$ 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 and Knight (1999) in the orthogonal case leads to $\mathrm{pen}(k) = 2\epsilon^2\sum_1^k 2\log(n/j)$.
- FDR - discussed above (?).
- Birgé and 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.

`Add refs to Tsybakov oracle ineqs.`

The formulation and proof of Theorem 11.3 is borrowed from Birgé and Massart (2001). Earlier versions in `[D-J, fill in.]`

2. The formulation and methods used for Theorem 11.6 are inspired by Birgé and Massart (2001). See also the St. Flour course Massart (2007).

§6. Some bounds for $\hat{k}_F - \hat{k}_G$ in sparse cases are given in Abramovich et al. (2006).

Exercises

11.1 *(Overfitting of AIC.)* Consider the penalized least squares setting (11.2)–(11.3) with penalty $\mathrm{pen}_0(k) = 2k$ along with $n = p$ and orthogonal design matrix $X = I$. Show that the estimator $\hat{\mu}_{\mathrm{pen}}(x) = \hat{\delta}_H(x, \lambda\epsilon)$ is given by hard thresholding with $\lambda = \sqrt{2}$.
(a) Show that the MSE at $\mu = 0$ is approximately $c_0 n\epsilon^2$ and evaluate c_0.
(b) With $\mathrm{pen}_0(k) = 2k \log n$ and hence $\lambda = \sqrt{2 \log n}$, show that the MSE at $\mu = 0$ is approximately $c_1 \sqrt{\log n}\epsilon^2$ and evaluate c_1.

11.2 *(Why Proposition 11.2 isn't simpler.)* Consider the orthogonal case as in Section 11.2. Let λ_k be a sequence of positive constants. Suppose $\mathrm{pen}(k) = k\lambda_k^2$ is increasing in k. Let $\hat{k}$ be a minimizing value of k in (11.6). It is tempting to conclude that $\hat{\mu}_{\mathrm{pen}}(y)$ is given by hard thresholding at $\lambda_{\hat{k}}$, namely

$$\hat{\mu}_{\mathrm{pen},i}(y) = y_i I(|y_i| \geq \lambda_{\hat{k}}).$$

Show by counterexample that this is false in general.

11.3 *(Gaussian quantiles and $2k \log(n/k)$ penalties.)* Define the Gaussian quantile $z(\eta)$ by the equation $\tilde{\Phi}(z(\eta)) = \eta$.
(a) Use (8.88) to show that

$$z^2(\eta) = 2 \log \eta^{-1} - \log \log \eta^{-1} - r(\eta),$$

and that when $\eta \leq 0.01$, we have $1.8 \leq r(\eta) \leq 3$ (Abramovich et al., 2006).
(b) Show that $z'(\eta) = -1/\phi(z(\eta))$ and hence that if $0 < \eta_1 < \eta_2 < \frac{1}{2}$, then

$$z(\eta_1) - z(\eta_2) \leq \frac{\eta_2 - \eta_1}{\eta_1 z(\eta_1)}.$$

(c) Verify (11.11) and (11.12).

11.4 *(Approximation for expected kth order statistic.)* Verify (11.10) for $k_n = o(n)$. *Hint.* Use $EX = \int_0^\infty P(X \geq t)dt$ along with suitable tail bounds for binomial random variables.

11.5 *(A 'small signal' bound for $\mathcal{R}(\mu)$.)* Suppose that $p \leq 2$, that $k \to k\lambda_k^2$ is increasing, and that $\mu \in \Theta_{n,p}(C)$ for $C \leq \lambda_1$. Show that $|\mu|_{(k)} \leq \lambda_k$ for all k, and hence in the orthogonal case that $\mathcal{R}(\mu) = \sum_{k=1}^n \mu_k^2$.

11.6 *(Monotonicity of penalty.)* If $\lambda_k = \sqrt{\zeta}(1 + \sqrt{2L_k})$ with $L_k = (1 + 2\beta)\log(n\gamma/k)$ for $k \geq 1$ (and $L_0 = L_1$) and $\gamma > e$, verify that $k \to k\lambda_k^2$ is monotone increasing for $0 \leq k \leq n$.

11.7 *(Inadequacy of $(2 \log n)k$ penalty.)* Suppose $p < 2$. If $\mathrm{pen}(k) = k\lambda_k^2$ has the form $\lambda_k^2 \equiv 2 \log n$, use (11.27) with $\epsilon = 1$ to show that

$$\sup_{\mu \in \Theta_{n,p}(C)} \mathcal{R}(\mu) \leq \bar{r}_{n,p}(C),$$

where, with $\zeta(r) = \sum_{k=1}^\infty k^{-r}$,

$$\bar{r}_{n,p}(C) = \begin{cases} C^2 & C < \sqrt{2 \log n} \\ C^p (2 \log n)^{1-p/2} & \sqrt{2 \log n} \leq C < n^{1/p}\sqrt{2 \log n} \\ 2n \log n & C \geq n^{1/p}\sqrt{2 \log n}. \end{cases}$$

Especially for C near $n^{1/p}$ or larger, this is inferior by a log term to the control function $r_{n,p}(C)$ obtained with penalty (11.35).

12

Exact rates for estimation on Besov spaces

We return to function estimation, for example in the continuous Gaussian white noise model (1.21), viewed in the context of the sequence model corresponding to coefficients in an orthonormal wavelet basis. We return also to the estimation framework of Section 9.8 with the use of Besov bodies $\Theta^{\alpha}_{p,q}(C)$ to model different degrees of smoothness and sparsity. The plan is to apply the results on penalized estimation from the last chapter separately to each level of wavelet coefficients.

This chapter has two main goals. The first is to remove the logarithmic terms that appear in the upper bounds of Theorem 9.14 (and also in Chapter 10) while still using adaptive estimators of threshold type. The reader may wish to review the discussion in Section 9.11 for some extra context for this goal.

The second aim of this chapter is finally to return to the theme of linear inverse problems, introduced in Chapter 3 with the goal of broadening the class of examples to which the Gaussian sequence model applies. We now wish to see what advantages can accrue through using thresholding and wavelet bases, to parallel what we have studied at length for direct estimation in the white noise model.

In the first section of this chapter, we apply the $2k\log(n/k)$ oracle inequality of Chapter 11 and its ℓ_p ball consequences to show that appropriate penalized least squares estimates (which have an interpretation as data dependent thresholding) adapt exactly to the correct rates of convergence over essentially all reasonable Besov bodies. Thus, we show that for an explicit $\hat{\theta}^P$,

$$\sup_{\Theta} E\|\hat{\theta}^P - \theta\|^2 \le cR_N(\Theta,\epsilon)(1+o(1))$$

as $\epsilon \to 0$ simultaneously for all $\Theta = \Theta^{\alpha}_{p,q}(C)$ in a large set of values for (α, p, q, C), although the constant c does depend on these values.

Our approach is based on the inequalities of Chapter 11.4, which showed that the ℓ_p-ball minimax risk could, up to multiplicative constants, be described by the relatively simple control functions $r_{n,p}(C,\epsilon)$ defined there. The device of "Besov shells"—consisting of sets of vectors $\theta \in \Theta$ that vanish except on level j, and hence are equivalent to ℓ_p-balls—allows the study of minimax risks on Θ to be reduced to the minimax risks and hence control functions $R_j = r_{n_j,p}(C_j,\epsilon_j)$ where the parameters $(n_j = 2^j, C_j, \epsilon_j)$ vary with j. Accordingly, a study of the shell bounds $j \to R_j$ yields our sharp rate results.

We describe an alternative to the singular value decomposition, namely the *wavelet-vaguelette* decomposition (WVD), for a class of linear operators. The left and right singular function systems of the SVD are replaced by wavelet-like systems which still have mul-

tiresolution structure and yield sparse representations of functions with discontinuities. The function systems are not exactly orthogonal, but they are *nearly* orthogonal, in the sense of 'frames', and are in fact a sufficient substitute for analyzing the behavior of threshold estimators.

In Section 12.2, then, we indicate some drawbacks of the SVD for object functions with discontinuities and introduce the elements of the WVD.

Section 12.3 lists some examples of linear operators A having a WVD, including integration of integer and fractional orders, certain convolutions and the Radon transform. The common feature is that the stand-in for singular values, the *quasi-singular* values, decay at a rate algebraic in the number of coefficients, $\kappa_j \approx 2^{-\beta j}$ at level j.

Section 12.4 focuses on a particular idealisation, motivated by the WVD examples, that we call the "correlated levels model", cf (12.32). This generalizes the white noise model by allowing noise levels $\epsilon_j = 2^{\beta j}\epsilon$ that grow in magnitude with resolution level j, a key feature in inverting data in ill-posed inverse problems. In addition, the model allows for the kind of near-independence correlation structure of noise that appears in problems with a WVD.

Using co-ordinatewise thresholding–with larger thresholds chosen to handle the variance inflation with level–we easily recover the optimal rate of convergence up to a logarithmic factor. This analysis already makes it possible to show improvement in the rates of convergence, compared to use of the SVD, that are attainable by exploiting sparsity of representation in the WVD.

By returning to the theme of penalized least squares estimation with $2n \log n/k$ penalties, we are again able to dispense with the logarithmic terms in the rates of convergence in the correlated levels model. The proof is begun in Section 12.4 up to the point at which the argument is reduced to study of ℓ_p control functions on Besov shells. This topic is taken up in Section 12.5.

12.1 Direct estimation

We consider the sequence model

$$y_{jk} = \theta_{jk} + \epsilon z_{jk}, \qquad j \in \mathbb{N},\ k = 1, \dots, 2^j; \tag{12.1}$$

with $z_{jk} \sim N(0,1)$ independently. Although it is a special case of the correlated levels model discussed later in Section 12.4, we begin with this setting for simplicity and because of the greater attention we have given to the direct estimation model. As in previous chapters, the single subscript j refers to a vector: $y_j = (y_{jk}), \theta_j = (\theta_{jk})$ etc.

We use a penalized least squares estimator on each level j, with the penalty term allowed to depend on j, so that

$$\hat{\theta}_P(y_j) = \underset{\theta_j}{\operatorname{argmin}}\ \|y_j - \theta_j\|^2 + \epsilon^2 \mathrm{pen}_j(N[\theta_j]), \tag{12.2}$$

Here $N[\theta_j]$ denotes the number of non-zero entries in θ_j, so that we are considering a complexity penalized estimator of the form studied in Chapter 11, compare in particular Lemma 11.1. The penalty term will be of the type used to obtain oracle inequalities in

Section 11.3, thus

$$\mathrm{pen}_j(k) = k\lambda_{j,k}^2, \qquad \lambda_{j,k} = \sqrt{\zeta}\big(1 + \sqrt{2\log(2^j\gamma/k)}\big).$$

As discussed there, we assume that $\zeta > 1$ and $\gamma > e$ so that the oracle inequality of Theorem 11.3 may be applied, with $M = M(\gamma)$ guaranteed to be finite for $\gamma > e$ by virtue of (11.15).

As in earlier chapters, compare Sections 9.9 and 10.6, we define a cutoff level $J = \log_2 \epsilon^{-2}$ and use the penalized least squared estimate only on levels $j < J$. As noted there, in the calibration $\epsilon = n^{-1/2}$, this corresponds to estimating the first n wavelet coefficients $\theta[f]$ of a function f based on n observations in a discrete regression model such as (1.13) (with $\sigma = 1$ there).

We put these levelwise estimates together to get a wavelet penalized least squares estimate $\hat\theta^P = (\hat\theta_j^P)$:

$$\hat\theta_j^P(y) = \begin{cases} \hat\theta_P(y_j) & j < J \\ 0 & j \ge J. \end{cases}$$

Remark 4 below discusses what happens if we estimate at all levels j.

The estimator is equivalent to hard thresholding. Indeed, let $\hat k_j = N[\hat\theta_P(y_j)]$ be the number of non-zero entries in $\hat\theta_P(y_j)$ and set $t_k^2 = k\lambda_k^2 - (k-1)\lambda_{k-1}^2$. Then Proposition 11.2 says that $\hat\theta_j^P$ is is equivalent to hard thresholding at $\hat t_j = t_{\hat k_j}$, and then Lemma 11.7 confirms that $t_{\hat k_j} \approx \lambda_{j,\hat k_j}$. Observe that the term $[2\log(2^j\gamma/\hat k_j)]^{1/2}$ may be rather smaller than the universal threshold $(2\log\epsilon^{-2})^{1/2}$, both because $j < J$, which corresponds to $2^j < \epsilon^{-2}$, and also because $\hat k_j$ may be large. This reflects a practically important phenomenon: $\sqrt{2\log n}$ thresholds can be too high in some settings, for example in Figure 7.6, and lower choices of threshold can yield much improved reconstructions and MSE performance. The extra factor $\sqrt{\zeta} > 1$ and the extra constant 1 in the definition of $\lambda_{j,k}$ are imposed by the theoretical approach taken here, but should not obscure the important conceptual point.

Theorem 12.1 *Assume model* (12.1) *and let* $\hat\theta^P$ *be the wavelet penalized least squares estimate described above, and assume that* $\gamma > e$ *and* $\zeta > 1$. *For* $\alpha > (1/p - 1/2)_+$ *along with* $0 < p, q \le \infty$, *there exist constants* $c_0, \dots, c_3$ *such that*

$$\begin{aligned} c_0 C^{2(1-r)}\epsilon^{2r} &\le R_N(\Theta_{p,q}^\alpha(C), \epsilon) \\ &\le \sup_{\Theta_{p,q}^\alpha(C)} E\|\hat\theta^P - \theta\|^2 \le c_1 C^{2(1-r)}\epsilon^{2r} + c_2 C^2(\epsilon^2)^{2\alpha'} + c_3\epsilon^2\log\epsilon^{-2}. \end{aligned}$$

The lower bound holds for all $C \ge \epsilon$, *and the upper bound for all* $C > 0$. *Here* $r = 2\alpha/(2\alpha+1)$, *and with* $a = \alpha + 1/2 - 1/p$,

$$\alpha' = \begin{cases} \alpha & \text{if } p \ge 2 \\ a & \text{if } p < 2. \end{cases}$$

Remarks. 1. The dependence of the constants on the parameters defining the estimator and Besov space is given by $c_1 = c_1(\zeta, \gamma, \alpha, p)$, $c_2 = c_2(\alpha, p)$ and $c_3 = c_3(\zeta, \gamma)$, while c_0 is an absolute constant.

2. Let us examine when the $C^{2(1-r)}\epsilon^{2r}$ term dominates as $\epsilon \to 0$. Since $r < 1$, the $\epsilon^2\log\epsilon^{-2}$ term is always negligible. If $p \ge 2$, then $2\alpha' = 2\alpha > r$ and so the tail bias term

$C^2(\epsilon^2)^{2\alpha'}$ is also of smaller order. If $p < 2$, a convenient condition is that $\alpha \geq 1/p$, for then $\alpha' = a \geq 1/2 > r/2$, and again $C^2(\epsilon^2)^{2\alpha'}$ is of smaller order.

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 ϵ the tail bias $C^2(\epsilon^2)^{2\alpha'} < C^{2(1-r)}\epsilon^{2r}$. Simple algebra shows that this occurs when

$$\epsilon < C^{-r/(2\alpha'-r)},$$

showing the key role of the radius C.

4. The estimator $\hat{\theta}^P$ truncates at level J, setting all higher levels to zero. If instead one estimates θ_j at *all* levels, then it is possible to remove the tail bias term $C^2(\epsilon^2)^{2\alpha'}$. In order that the term $c_3\epsilon^2 \log \epsilon^{-2}$ not increase in order of magnitude, it is necessary to increase the penalty at levels $j \geq J$. For details we refer to Section 12.4, where the current model corresponds to the special case $\beta = 0$.

Proof We give here the part of the argument that reduces the bounds to analysis of ℓ_p-ball control functions corresponding to Besov shells; that analysis is deferred to Section 12.5. *Upper bound.* The levelwise structure of $\hat{\theta}^P$ yields the MSE decomposition

$$E\|\hat{\theta}^P - \theta\|^2 = \sum_{j<J} E\|\hat{\theta}_P(y_j) - \theta_j\|^2 + \Delta_J(\theta), \tag{12.3}$$

where $\Delta_J(\theta) = \sum_{j\geq J} \|\theta_j\|^2$ is the "tail bias" due to not estimating beyond level J. The maximum tail bias over $\Theta^\alpha_{p,q}(C)$ was evaluated at (9.60) and yields the bound $c_2C^2 2^{-2\alpha' J}$, with $c_2 = c_2(\alpha, p)$. Since $2^{-J} = \epsilon^2$, we recover the tail bias bound.

To bound the mean squared error of $\hat{\theta}_P(y_j)$, we appeal to the oracle inequality Theorem 11.3. Since model (12.1) is orthogonal, we in fact use Corollary 11.5. Using (11.29), then, we obtain

$$E\|\hat{\theta}_P(y_j) - \theta_j\|^2 \leq c_3\epsilon^2 + c_3\mathcal{R}_j(\theta_j, \epsilon), \tag{12.4}$$

where $c_3(\zeta, \gamma) = \max\{a(\zeta), b(\zeta)M(\gamma)\}$, and in accordance with (11.26), the level j theoretical complexity is given by

$$\mathcal{R}_j(\theta_j, \epsilon) = \min_{0\leq k\leq n_j} \sum_{l>k} \theta^2_{j(l)} + \epsilon^2 k\lambda^2_{j,k}, \tag{12.5}$$

where $\theta^2_{j(l)}$ denotes the l-th largest value among $\{\theta^2_{jk}, j = 1, \ldots, 2^k\}$.

Summing over $j < J = \log_2 \epsilon^{-2}$, the first term on the right side of (12.4) yields the $c_3\epsilon^2 \log \epsilon^{-2}$ term in the upper bound of Theorem 12.1.

To bound $\sum_j \mathcal{R}_j(\theta_j, \epsilon)$ we use the Besov shells $\Theta^{(j)} = \{\theta \in \Theta : \theta_I = 0 \text{ for } I \notin \mathcal{I}_j\}$ introduced in Section 10.7. The maximum of $\mathcal{R}_j(\theta_j, \epsilon)$ over Θ can therefore be obtained by maximizing over $\Theta^{(j)}$ alone, and so

$$\sup_\Theta \sum_j \mathcal{R}_j(\theta_j, \epsilon) \leq \sum_j \sup_{\Theta^{(j)}} \mathcal{R}_j(\theta_j, \epsilon). \tag{12.6}$$

We also recall the interpretation of $\Theta^{(j)}$ as ℓ_p–balls:

$$\Theta^{(j)} \equiv \Theta_{n_j,p}(C_j) \qquad \text{for} \qquad n_j = 2^j, \;\; C_j = C2^{-aj}.$$

The maximization of theoretical complexity over ℓ_p-balls was studied in detail in Section 11.4. Let $r_{n,p}(C, \epsilon)$ be the control function for minimax mean squared error at noise level ϵ. The proof of Theorem 11.6 yields the bound

$$\mathcal{R}_j(\theta_j, \epsilon) \le c_4 r_{n_j,p}(C_j, \epsilon)$$

for $\theta_j \in \Theta_{n_j,p}(C_j)$ with $c_4 = c_4(\zeta, \gamma)$, compare (11.39). The previous display together with (12.6) shows that it remains to bound $\sum_j r_{n_j,p}(C_j, \epsilon)$.

In Section 12.5, we show that the shell bounds $R_j = r_{n_j,p}(C_j, \epsilon)$ peak at a critical level j_*, and decay geometrically away from the value R_{j_*} at this least favorable level, so that the series is indeed summable. So the final bound we need, namely

$$\sup_{\Theta} \sum_j \mathcal{R}_j(\theta_j, \epsilon) \le c_1 C^{2(1-r)} \epsilon^{2r},$$

follows from Proposition (12.41) to be proved there. The constant $c_1 = c_1(\zeta, \gamma, \alpha, p)$ since it depends both on $c_4(\zeta, \gamma)$ and the parameters (α, p) of $\Theta = \Theta^{\alpha}_{p,q}(C)$.

Lower bound. For $\epsilon \le C$, we saw already in Theorem 9.14 that $R_N(\Theta, \epsilon) \ge c C^{2(1-r)} \epsilon^{2r}$, but we can also rewrite the argument using Besov shells and control functions for ℓ_p balls. Since each shell $\Theta^{(j)} \subset \Theta$, we have

$$R_N(\Theta, \epsilon) \ge R_N(\Theta^{(j)}, \epsilon) \ge R_N(\Theta_{n_j,p}(C_j), \epsilon) \ge a_1 r_{n_j,p}(C_j, \epsilon),$$

by the lower bound part of Theorem 11.6. Consequently $R_N(\Theta, \epsilon) \ge a_1 \max_j R_j$, and that this is bounded below by $c_0 C^{2(1-r)} \epsilon^{2r}$ is also shown in Proposition 12.6. □

12.2 Wavelet-Vaguelette Decomposition

We return to the model for linear inverse problems adopted in Section 3.9. The focus there was on use of the singular value decomposition (SVD), linear estimators and the effect of the index of ill-posedness β on the resulting rates of convergence over function spaces of mean square smoothness type ($p = 2$). Here we turn to defects of the SVD for functions with singularities or spatially varying smoothness (as captured by Besov bodies with $p < 2$), and the construction of an alternative decomposition for certain linear operators that is better adapted to wavelet bases.

Stochastic observation model. Let A be a linear operator from $\mathcal{D}(A) \subset L_2(T, \langle \cdot, \cdot \rangle)$ to $\mathcal{R}(A) \subset L_2(U, [\cdot, \cdot])$. We consider an idealized model in which Af is observed in additive Gaussian noise. Assume that we observe

$$Y = Af + \epsilon Z, \tag{12.7}$$

which is interpreted to mean that, for all $g \in L_2(U)$, we have

$$Y(g) = [Af, g] + \epsilon Z(g), \tag{12.8}$$

and the process $g \to Z(g)$ is Gaussian, with zero mean and covariance

$$\mathrm{Cov}(Z(g), Z(h)) = [g, h]. \tag{12.9}$$

From (12.9), we have $Z(\sum_k \alpha_k g_k) = \sum_k \alpha_k Z(g_k)$ a.s. when g_k and $\sum_k \alpha_k g_k \in L_2(U)$.

A defect of the Singular Value Decomposition. Suppose that $A : L_2(T) \to L_2(U)$ is a linear operator with singular value decomposition $Ae_k = b_k h_k$ in terms of orthogonal singular systems $\{e_k\}$ for $L_2(T)$ and $\{h_k\}$ for $L_2(U)$. In the examples of Section 3.9, and more generally, the singular functions are 'global' functions, supported on all of T and U respectively. Consequently, the representation of a smooth function with isolated singularities may not be sparse.

Consider a simple example in which $\{e_k\}$ is a trigonometric basis on $[0, 1]$ and f is a (periodic) step function, such as $I_{[1/4,3/4]}(t)$. If A is a convolution with a periodic kernel $a(t)$ with coefficients $b_k = \langle a, e_k \rangle$, then in Section 3.9 we derived the sequence model $y_k = \theta_k + \epsilon_k z_k$ with $\epsilon_k = \epsilon/b_k$. The coefficients $\theta_k = \langle f, e_k \rangle$ would typically have slow decay with frequency k, of order $|\theta_k| \asymp O(1/k)$. The (ideal) risk of the best linear estimator of form $\hat{\theta}_c = (c_k y_k)$ for the given θ has the form

$$\inf_c r(\hat{\theta}_c, \theta) = \sum_k \frac{\theta_k^2 \epsilon_k^2}{\theta_k^2 + \epsilon_k^2} \asymp \sum_k \min(\theta_k^2, \epsilon^2/b_k^2). \tag{12.10}$$

For a typical convolution operator A, the singular values b_k decrease quite quickly, while the coefficients θ_k do not. Hence even the ideal linear risk for a step function in the Fourier basis is apt to be uncomfortably large.

We might instead seek to replace the SVD bases by wavelet bases, in order to take advantage of wavelets' ability to achieve sparse representations of smooth functions with isolated singularities.

Example I. As a running example for exposition, suppose that A is given by integration on $\mathbb{R}$:

$$(Af)(u) = f^{(-1)}(u) = \int_{-\infty}^{u} f(t)dt. \tag{12.11}$$

Let $\{\psi_I\}$ be a nice orthonormal wavelet basis for $L_2(\mathbb{R})$: as usual we use I for the double index (j, k), so that $\psi_I(t) = 2^{j/2}\psi(2^j t - k)$. We may write

$$\begin{aligned} A\psi_I(u) &= \int_{-\infty}^{u} 2^{j/2}\psi(2^j t - k)dt = 2^{-j} \cdot 2^{j/2}(\psi^{(-1)})(2^j u - k) \\ &= 2^{-j}(\psi^{(-1)})_I(u). \end{aligned}$$

The initial difficulty is that $\{u_I := (\psi^{(-1)})_I\}$ is not orthonormal in the way that $\{\psi_I\}$ is.

Suppose initially that we consider an arbitrary orthonormal basis $\{e_k\}$ for $L_2(T)$, so that $f = \sum \langle f, e_k \rangle e_k$. Suppose also that we can find *representers* $g_k \in L_2(U)$ for which

$$\langle f, e_k \rangle = [Af, g_k].$$

According to Proposition C.6, this occurs when each $e_k \in \mathcal{R}(A^*)$. In model (12.8), the corresponding sequence of observations $Y_k = Y(g_k)$ has mean $[Af, g_k] = \langle f, e_k \rangle$ and covariance $\epsilon^2 \Sigma_{kl}$ where $\Sigma_{kl} = \mathrm{Cov}(Z(g_k), Z(g_l)) = [g_k, g_l]$. We might then consider using estimators of the form $\hat{f} = \sum_k c_k(Y_k)e_k$ for co-ordinatewise functions $c_k(Y_k)$, which might be linear or threshold functions. However, Proposition 4.30 shows that in the case of diagonal linear estimators and suitable parameter sets, the effect of the correlation of the Y_k on the efficiency of estimation is determined by $\lambda_{\min}(\rho(\Sigma))$, the minimum eigenvalue of the

correlation matrix corresponding to covariance Σ. In order for this effect to remain bounded even as the noise level $\epsilon \to 0$, we need the representers g_k to be nearly orthogonal in an appropriate sense. To see this, set $u_k = g_k/\|g_k\|_2$, and observe that

$$\begin{aligned}\lambda_{\min}(\rho(\Sigma)) &= \inf\{\alpha^T \rho(\Sigma)\alpha : \|\alpha\|_2 = 1\} \\ &= \inf\left\{\mathrm{Var}\left(\sum \frac{\alpha_k}{\|g_k\|} Z(g_k)\right) : \|\alpha\|_2 = 1\right\} \\ &= \inf\left\{\Big\|\sum \alpha_k u_k\Big\|^2 : \|\alpha\|_2 = 1\right\}.\end{aligned}$$

Hence, we obtain the necessary control if the normalized representers satisfy a bound

$$\Big\|\sum \alpha_k u_k\Big\|_2 \geq c\|\alpha\|_2 \qquad \text{for all } \alpha \in \ell_2. \tag{12.12}$$

We will see that this is indeed often possible if one starts with a wavelet basis $\{\psi_I\}$ for $L_2(T)$.

Remark. In developing the WVD, it is convenient initially to take $T = \mathbb{R}$ to avoid boundary effects, and to exploit translation invariance properties of $\mathbb{R}$. In such cases, it may be that the operator A is only defined on a dense subset $\mathcal{D}(A)$ of $L_2(T)$. For example, with integration, (12.11), the Fourier transform formula $\widehat{Af}(\xi) = (i\xi)^{-1}\hat{f}(\xi)$ combined with the Parseval relation (C.12) shows that $Af \in L_2(\mathbb{R})$ if and only if f belongs to the subset of $L_2(\mathbb{R})$ defined by $\int |\xi|^{-2}|\hat{f}(\xi)|^2 d\xi < \infty$. Similarly, using $A^*g = \int_u^\infty g(t)dt$, it follows that $\mathcal{R}(A^*)$ is the subset of L_2 corresponding to $\int |\xi|^2|\hat{f}(\xi)|^2 d\xi < \infty$.

Let us turn again to wavelet bases. Suppose that $\mathcal{D}(A)$ is dense in $L_2(T)$ and that A is one to one. Suppose that $\{\psi_I\}$ is an orthonormal wavelet basis for $L_2(T)$ such that $\psi_I \in \mathcal{D}(A) \cap \mathcal{R}(A^*)$ for every I. Proposition C.6 provides a representer g_I such that

$$\langle f, \psi_I\rangle = [Af, g_I]. \tag{12.13}$$

Suppose, in addition, that $\|g_I\| = c\kappa_j^{-1}$ is independent of k. Define two systems $\{u_I\}$, $\{v_I\} \in L_2(U)$ by the equations

$$u_I = \kappa_j g_I, \qquad v_I = \kappa_j^{-1} A\psi_I. \tag{12.14}$$

Since for every $f \in \mathcal{D}(A)$ we have $\langle f, A^*u_I\rangle = [Af, \kappa_j g_I] = \langle f, \kappa_j \psi_I\rangle$, we may conclude that

$$A^*u_I = \kappa_j \psi_I, \qquad A\psi_I = \kappa_j v_I. \tag{12.15}$$

In addition, the $\{u_I\}$ and $\{v_I\}$ systems are *biorthogonal*:

$$[v_I, u_{I'}] = \kappa_j^{-1}\kappa_{j'}[A\psi_I, g_{I'}] = \kappa_j^{-1}\kappa_{j'}\langle \psi_I, \psi_{I'}\rangle = \delta_{II'}. \tag{12.16}$$

Since $\langle f, \psi_I\rangle = [Af, g_I] = \kappa_j^{-1}[Af, u_I]$, we have the formal reproducing formula

$$f = \sum \langle f, \psi_I\rangle \psi_I = \sum \kappa_j^{-1}[Af, u_I]\psi_I. \tag{12.17}$$

Example I continued. Let A again correspond to integration. Suppose that the wavelet

ψ is C^1, with compact support and $\int \psi = 0$, so that $\psi \in \mathcal{D}(A) \cap \mathcal{R}(A^*)$. Then formula (12.13) and integration by parts shows that the representer

$$g_I = -(\psi_I)' = -2^j(\psi')_I.$$

Since $\|g_I\|_2 = c_\psi 2^j$, with $c_\psi = \|\psi'\|_2$, we can set $\kappa_j = 2^{-j}$, and then from (12.14),

$$u_I = -(\psi')_I, \qquad v_I = (\psi^{(-1)})_I.$$

We now turn to showing that the (non-orthogonal) systems $\{u_I\}$ and $\{v_I\}$ satisfy (12.12).

To motivate the next definition, note that members of both systems $\{u_I\}$ and $\{v_I\}$ have, in our example, the form $w_I(t) = 2^{j/2}w(2^j t - k)$. If we define a rescaling operator

$$(S_I w)(x) = 2^{-j/2}w(2^{-j}(x + k)), \tag{12.18}$$

then in our example above, but not in general, $(S_I w_I)(t) = w(t)$ is free of I.

Definition 12.2 A collection $\{w_I\} \subset L_2(\mathbb{R})$ is called a system of *vaguelettes* if there exist positive constants C_1, C_2 and exponents $0 < \eta < \eta' < 1$ such that for each I, the rescaled function $\tilde{w} = S_I w_I$ satisfies

$$\tilde{w}(t) \le C_1(1 + |t|)^{-1-\eta'}, \tag{12.19}$$

$$\int \tilde{w}(t)dt = 0 \tag{12.20}$$

$$|\tilde{w}(t) - \tilde{w}(s)| \le C_2|t - s|^\eta \tag{12.21}$$

for $s, t \in \mathbb{R}$.

In some cases, the three vaguelette conditions can be verified directly. Exercise 12.2 gives a criterion in the Fourier domain that can be useful in some other settings.

The following is a key property of a vaguelette system, proved in Appendix B.4. We use the abbreviation $\|\alpha\|_2$ for $\|(\alpha_I)\|_{\ell_2}$

Proposition 12.3 *(i) If $\{w_I\}$ is a system of vaguelettes satisfying* (12.19)– (12.21)*, then there exists a constant C, depending on (C_1, C_2, η, η') such that*

$$\Big\| \sum_I \alpha_I w_I \Big\|_2 \le C \|\alpha\|_2 \tag{12.22}$$

(ii) If $\{u_I\}, \{v_I\}$ are biorthogonal systems of vaguelettes, then there exist positive constants c, C such that

$$c\|\alpha\|_2 \le \Big\| \sum_I \alpha_I u_I \Big\|_2 \ , \ \Big\| \sum_I \alpha_I v_I \Big\|_2 \le C \|\alpha\|_2. \tag{12.23}$$

The second part is a relatively straightforward consequence of the first key conclusion; it shows that having two vaguelette systems that are biorthogonal allows extension of bound (12.22) to a bound in the opposite direction, which we have seen is needed in order to control $\lambda_{\min}(\rho(\Sigma))$.

Thus, if we have two biorthogonal systems of vaguelettes, then each forms a *frame*: up to multiplicative constants, we can compute norms of linear combinations using the coefficients alone.

Definition 12.4 (Donoho (1995)) Let $\{\psi_I\}$ be an orthonormal wavelet basis for $L_2(T)$ and $\{u_I\}, \{v_I\}$ be systems of vaguelettes for $L_2(U)$. Let A be a linear operator with domain $\mathcal{D}(A)$ dense in $L_2(T)$ and taking values in $L_2(U)$. The systems $\{\psi_I\}, \{u_I\}, \{v_I\}$ form a *wavelet vaguelette decomposition* of A if they enjoy the following properties:
(1) quasi-singular values: (12.15)
(2) biorthogonality: (12.16)
(3) near-orthogonality: (12.23).

Note that the quasi-singular values κ_j depend on j, but not on k.

Example I continued. Suppose again that $Af(u) = \int_{-\infty}^{u} f(t)dt$ and that ψ is a C^2 orthonormal wavelet with compact support and two vanishing moments, so that $\int \psi = \int t\psi = 0$. We saw that $\{u_I = -(\psi')_I\}$ and $\{v_I = (\psi^{(-1)})_I\}$ satisfy property (1) with $\kappa_j = 2^{-j}$, and property (2). In order to obtain the frame bounds for property (3), we verify conditions (12.19)–(12.21) for $\psi^{(-1)}$ and ψ', and then appeal to Proposition 12.3. Indeed, ψ' and $\psi^{(-1)}$ have compact support, the latter because ψ does and $\int \psi = 0$. So (12.19) holds trivially. Turning to (12.20), we have $\int \psi' = 0$ by compact support of ψ, and integration by parts shows (using compact support of $\psi^{(-1)}$) that $\int \psi^{(-1)} = -\int t\psi(t)dt = 0$. Finally ψ' is C^1 and $\psi^{(-1)}$ is C^3 so the Hölder property (12.21) follows again from compact support.

12.3 Examples of WVD

1. *r-fold integration.* If $(Af)(u) = \int_{-\infty}^{u} f(t)dt$ and r is a positive integer, we may define the r-fold iterated integral by $A_r f = A(A_{r-1} f)$. We also write $f^{(-r)}$ for $A_r f$. The WVD follows by extending the arguments used for $r = 1$. Suppose that ψ is a C^r orthonormal wavelet with compact support and $r + 1$ vanishing moments, then the WVD is given by

$$\kappa_j = 2^{-rj}, \qquad u_I = (-1)^r (\psi^{(r)})_I, \qquad v_I = (\psi^{(-r)})_I.$$

In particular, for later use we note that $\{\psi_I^{(r)}\}$ forms a system of vaguelettes and satisfies the frame bounds (12.23).

2. *Fractional Integration.* Suppose that A is the fractional integration operator

$$(Af)(u) = \frac{1}{\Gamma(\beta)} \int_{-\infty}^{u} \frac{f(t)}{(u-t)^{1-\beta}} dt = (\Psi_\beta \star f)(u) \tag{12.24}$$

for $0 < \beta < 1$ and $\Psi_\beta(u) = u_+^{\beta-1}/\Gamma(\beta)$. Define the order β fractional derivative and integral of ψ by $\psi^{(\beta)}$ and $\psi^{(-\beta)}$ respectively. The WVD of A is then obtained by setting

$$\kappa_j = 2^{-j\beta}, \qquad u_I = (\psi^{(\beta)})_I, \qquad v_I = (\psi^{(-\beta)})_I. \tag{12.25}$$

To justify these definitions, note that the Fourier transform of Ψ_β is given by (e.g. Gel'fand and Shilov (1964, p. 171)

$$\widehat{\Psi_\beta}(\xi) = \widehat{\Omega}(\xi)|\xi|^{-\beta},$$

where $\widehat{\Omega}(\xi)$ equals $c_\beta = ie^{i(\beta-1)\pi/2}$ for $\xi > 0$ and $c_\beta e^{-i\beta\pi}$ for $\xi < 0$. We use the Parseval

formula (C.12) to express the representer equation (12.13) in the form $\int \widehat{f}\overline{\widehat{\psi_I}} = \int \widehat{f}\widehat{\Psi_\beta}\overline{\widehat{g_I}}$ from which one formally obtains

$$\widehat{g_I}(\xi) = \widehat{\psi_I}/\widehat{\Psi_\beta}(\xi) = |\xi|^\beta \widehat{\psi_I}(\xi)/\overline{\widehat{\Omega}}(\xi).$$

It is easy to check that $\|\widehat{g_I}\|^2 = \|\widehat{g_0}\|^2 2^{2j\beta}$ so that we may take $\kappa_j = 2^{-j\beta}$ and $u_I = \kappa_j g_I$, and, as in (12.14) set $v_I = \kappa_j^{-1} A\psi_I$.

Thus $\{u_I\}$ and $\{v_I\}$ are biorthogonal, and one checks that both systems are obtained by translation and dilation of $\psi^{(-\beta)}$ and $\psi^{(\beta)}$ in (12.25), with

$$\widehat{\psi^{(\beta)}} = |\xi|^\beta \widehat{\psi}(\xi)/\overline{\widehat{\Omega}}(\xi), \qquad \widehat{\psi^{(-\beta)}} = |\xi|^{-\beta}\widehat{\Omega}(\xi)\widehat{\psi}(\xi). \tag{12.26}$$

The biorthogonality relations for $\{u_I\}$ and $\{v_I\}$ will then follow if we verify that $\psi^{(\beta)}$ and $\psi^{(-\beta)}$ satisfy (12.19)–(12.21). The steps needed for this are set out in Exercise 12.2.

3. *Convolution.* The operator

$$(Af)(u) = \int_{-\infty}^{\infty} a(u-t)f(t)dt = (a \star f)(u)$$

is bounded on $L_2(\mathbb{R})$ if $\int |a| < \infty$, by (C.31), so we can take $\mathcal{D}(A) = L_2(\mathbb{R})$. The adjoint A^* is just convolution with $\tilde{a}(u) = a(-u)$, and so in the Fourier domain, $\hat{\tilde{a}}(\xi) = \hat{a}(-\xi)$, and the representer g_I is given by

$$\widehat{g_I} = \widehat{\psi_I}/\hat{\tilde{a}}. \tag{12.27}$$

As simple examples, we consider

$$a_1(x) = e^x I\{x < 0\}, \qquad a_2(x) = \tfrac{1}{2}e^{-|x|}. \tag{12.28}$$

It is easily checked that

$$\hat{a}_1(\xi) = (1 - i\xi)^{-1}, \qquad \hat{a}_2(\xi) = (1 + \xi^2)^{-1},$$

and hence that

$$g_I = \psi_I - (\psi_I)', \qquad g_I = \psi_I - (\psi_I)''. \tag{12.29}$$

Either from representation (12.27), or more directly from (12.29), one finds that with $\beta = 1$ and 2 in the two cases, that

$$\|g_I\|_2^2 \sim \begin{cases} 2^{2j\beta} & \text{as } j \to \infty, \\ 1 & \text{as } j \to -\infty. \end{cases}$$

This is no longer homogeneous in j in the manner of fractional integration, but we can still set $\kappa_j = \min(1, 2^{-j\beta})$.

The biorthogonal systems $\{u_I\}$ and $\{v_I\}$ are given by (12.14). In the case of $u_I = \kappa_j g_I$, the rescaling $S_I u_I$ can be found directly from (12.29), yielding $2^{-j}\psi - \psi^{(\beta)}$ in the case $j > 0$. The vaguelette properties (12.19)– (12.21) then follow from those of the wavelet ψ. For $v_I = \kappa_j^{-1} A\psi_I$, it is more convenient to work in the Fourier domain, see Exercise 12.2.

4. *Radon transform.* For the Radon transform in $\mathbb{R}^2$—compare Section 3.9 for a version on the unit disk—Donoho (1995) develops a WVD with quasi-singular values $\kappa_j = 2^{j/2}$.

The corresponding systems $\{u_I\}$, $\{v_I\}$ are localized to certain curves in the (s, ϕ) plane rather than to points, so they are not vaguelettes, but nevertheless they can be shown to have the near-orthogonality property.

Here is a formulation of the indirect estimation problem when a WVD of the operator is available, building on the examples presented above. Suppose that we observe A in the stochastic observation model (12.7)–(12.9), and that $\{\psi_I, u_I, v_I\}$ form a wavelet-vaguelette decomposition of A. Consider the observations

$$Y(u_I) = [Af, u_I] + \epsilon Z(u_I).$$

Writing $Y_I = Y(u_I), z_I = Z(u_I)$ and noting that $[Af, u_I] = \kappa_j \langle f, \psi_I \rangle = \kappa_j \theta_I$, say, we arrive at

$$Y_I = \kappa_j \theta_I + \epsilon z_I. \tag{12.30}$$

Let Σ be the covariance matrix of $z = (z_I)$. Since $\Sigma_{II'} = \mathrm{Cov}(Z(u_I), Z(u_{I'})) = [u_I, u_{I'}]$,

$$\beta^T \Sigma \beta = \| \sum \beta_I u_I \|_2^2,$$

the near orthogonality property guarantees that

$$\xi_0 I \le \Sigma \le \xi_1 I, \tag{12.31}$$

where the inequalities are in the sense of non-negative definite matrices. We say that the noise z is *nearly independent*.

We are now ready to consider estimation of f from observations on Y. The reproducing formula (12.17) suggests that we consider estimators of f of the form

$$\hat{f} = \sum_I \eta_I(\kappa_j^{-1} Y_I) \psi_I$$

for appropriate univariate estimators $\eta_I(\cdot)$. The near-independence property makes it plausible that restricting to estimators in this class will not lead to great losses in estimation efficiency; this is borne out by results to follow. Introduce $y_I = \kappa_j^{-1} Y_I \sim N(\theta_I, \kappa_j^{-2}\epsilon^2)$. We have $\hat{f} - f = \sum_I [\eta_I(y_I) - \theta_I]\psi_I$ and so, for the mean squared error,

$$E\| \hat{f} - f \|_2^2 = \sum_I E[\eta_I(y_I) - \theta_I]^2 = \sum_I r(\eta_I, \theta_I; \kappa_j^{-1}\epsilon).$$

Notice that if $\kappa_j \sim 2^{-\beta j}$, then the noise level $\kappa_j^{-1}\epsilon \sim 2^{\beta j}\epsilon$ grows rapidly with level j. This is the noise amplification characteristic of linear inverse problems and seen also in Chapter 3.9. In the next section, we study in detail the consequences of using threshold estimators to deal with this amplification.

12.4 The correlated levels model

For the main estimation results of this chapter we adopt the following weakly correlated version of the Gaussian sequence model. For $j \in \mathbb{N}$ and $k = 1, \ldots, 2^j$, let $z = (z_{jk})$ be

jointly normally distributed with mean 0 and covariance matrix Σ. We assume that

$$\begin{aligned} &y_{jk} = \theta_{jk} + \epsilon_j z_{jk}, \qquad\qquad \epsilon_j = 2^{\beta j}\epsilon, \quad \beta \geq 0 \\ &\xi_0 I \preceq \Sigma \preceq \xi_1 I. \end{aligned} \tag{12.32}$$

with the inequalities on the covariance matrix understood in the sense of non-negative definite matrices.

This is an extension of the Gaussian white noise model (12.1) in two significant ways: (i) level dependent noise $\epsilon_j = 2^{\beta j}\epsilon$ with index of ill-posedness β, capturing the noise amplification inherent to inverting an operator A of smoothing type, and (ii) the presence of correlation among the noise components, although we make the key assumption of near-independence.

Motivation for this model comes from the various examples of linear inverse problems in the previous section: when a wavelet-vaguelette decomposition exists, we have both properties (i) and (ii). The model is then recovered from (12.30)–(12.31) when $I = (jk)$ has the standard index set, $\kappa_j = 2^{-\beta j}$ and $y_{jk} = \kappa_j^{-1} Y_{jk}$. Our use of the index set $k = 1, \ldots, 2^j$ corresponds to estimation on the unit interval $[0, 1]$, as in earlier chapters.

The goals for this section are first to explore the effects of the level dependent noise $\epsilon_j = 2^{\beta j}\epsilon$ on choice of threshold and the resulting mean squared error. Then we indicate the advantages in estimation accuracy that accrue with use of the WVD in place of the SVD via an heuristic calculation with coefficients θ corresponding to a piecewise constant function. Finally we introduce an appropriate complexity penalized estimator for the correlated levels model. We formulate a result on minimax rates of estimation and begin the proof by using oracle inequalities to reduce the argument to the analysis of risk control functions over Besov shells.

Let us first examine what happens in model (12.32) when on level j we use soft thresholding with a threshold $\lambda_j \epsilon_j$ that depends on the level j but is otherwise fixed and non-random. Thus $\hat{\theta}^S_{jk} = \eta_S(y_{jk}, \lambda_j \epsilon_j)$. Decomposing the mean squared error by levels, we have

$$r(\hat{\theta}^S, \theta) = \sum E\|\hat{\theta}^S_j - \theta_j\|^2,$$

and if $\lambda_j = \sqrt{2 \log \delta_j^{-1}}$, we have from the soft thresholding risk bound (8.13) that

$$E\|\hat{\theta}^S_j - \theta_j\|^2 \leq 2^j \delta_j \epsilon_j^2 + (\lambda_j^2 + 1) \sum_k \min(\theta_{jk}^2, \epsilon_j^2).$$

The noise term $2^j \delta_j \epsilon_j^2 = \delta_j 2^{(1+2\beta)j} \epsilon^2$, which shows the effect of the geometric inflation of the variances, $\epsilon_j^2 = 2^{2\beta j}\epsilon^2$. To control this term, we might take $\delta_j = 2^{-(1+2\beta)j} = n_j^{-(1+2\beta)}$. This corresponds to threshold

$$\lambda_j = \sqrt{2(1 + 2\beta) \log n_j},$$

which is higher than the 'universal' threshold $\lambda_j^U = \sqrt{2 \log n_j}$ when the ill-posedness index $\beta > 0$. With this choice we arrive at

$$E\|\hat{\theta}^S_j - \theta_j\|^2 \leq \epsilon^2 + c_\beta j \sum_k \min(\theta_{jk}^2, 2^{2\beta j}\epsilon^2). \tag{12.33}$$

At this point we can do a heuristic calculation to indicate the benefits of using the sparse representation provided by the WVD. This will also set the stage for more precise results to follow. Now suppose that the unknown function f is piecewise constant with at most d discontinuities. Then the wavelet tranform of f is sparse, and in particular, if the suppport of ψ is compact, there are at most a bounded number of non-zero coefficients θ_{jk} at each level j, and those coefficients are bounded by $c2^{-j/2}$ by Lemma 7.2. Hence

$$\sum_k \min(\theta_{jk}^2, 2^{2\beta j}\epsilon^2) \le c_{d\psi f} \min(2^{-j}, 2^{2\beta j}\epsilon^2).$$

To find the worst level, we solve for $j = j_*$ in the equation $2^{-j} = 2^{2\beta j}\epsilon^2$, so that $2^{(1+2\beta)j_*} = \epsilon^{-2}$. On the worst level, this is bounded by $2^{-j_*} = (\epsilon^2)^{1/(1+2\beta)}$. The maxima on the other levels decay geometrically in $|j - j_*|$ away from the worst level, and so the sum converges and as a bound for the rate of convergence for this function (12.33) yields

$$j_* 2^{-j_*} \asymp (\log \epsilon^{-2})(\epsilon^2)^{1/(1+2\beta)}.$$

Comparison with SVD. For piecewise constant f, we can suppose that the coefficients in the singular function basis, $\theta_k = \langle f, e_k \rangle$ decay as $O(1/k)$. Suppose that the singular values $b_k \asymp k^{-\beta}$. Then from (12.10),

$$\sum_k \min(\theta_k^2, \epsilon^2/b_k^2) \asymp \sum \min(k^{-2}, k^{2\beta}\epsilon^2) \asymp k_*^{-1},$$

where k_* solves $k^{-2} = k^{2\beta}\epsilon^2$, so that $k_*^{-1} = (\epsilon^2)^{1/(2+2\beta)}$. Hence, the rate of convergence using linear estimators with the singular value decomposition is $O((\epsilon^2)^{1/(2+2\beta)})$, while we can achieve the distinctly faster rate $O(\log \epsilon^{-2}(\epsilon^2)^{1/(1+2\beta)})$ with thresholding and the WVD.

In fact, as the discussion of the direct estimation case (Section 12.1) showed, the $\log \epsilon^{-2}$ term can be removed by using data-dependent thresholding, and it will be the goal of the rest of this chapter to prove such a result.

Main result. We will see that the minimax rate of convergence over $\Theta_{p,q}^{\alpha}(C)$ is $C^{2(1-r)}\epsilon^{2r}$, with $r = 2\alpha/(2\alpha + 2\beta + 1)$, up to constants depending only on (α, p, q) and β.

This is the rate found earlier in Proposition 4.22 for the case of Hölder smoothness ($p = q = \infty$) and in Pinsker's Theorem 5.3 for Hilbert-Sobolev smoothness ($p = q = 2$). The result will be established here for $0 < p, q \le \infty$, thus in particular extending the result to cover sparse cases with $p < 2$.

A further goal of our approach is to define an estimator that achieves the exact rate of convergence ϵ^{2r} without the presence of extra logarithmic terms in the upper bounds, as we have had to accept in previous chapters (8, 9, 10). In addition, we seek to do this with an adaptive estimator, that is, one that does not use knowledge of the parameter space constants (α, p, q, C) in its construction.

These goals can be achieved using a complexity penalized estimator, constructed levelwise, in a manner analogous to the direct case, Section 12.1, but allowing for the modified noise structure. Thus, at level j, we again use a penalized least squares estimator $\hat{\theta}_P(y_j)$, (12.2), with $\mathrm{pen}_j(k) = k\lambda_{j,k}^2$. However, now

$$\lambda_{j,k} = \sqrt{\zeta}(1 + \sqrt{2L_{n_j,k}}), \qquad L_{n_j,k} = (1 + 2\beta)\log(\gamma_{n_j} n_j / k), \tag{12.34}$$

where $n_j = 2^j$ and, with $\gamma > e$,

$$\gamma_{n_j} = \begin{cases} \gamma & \text{if } j \le j_\epsilon = \log_2 \epsilon^{-2} \\ \gamma[1 + (j - j_\epsilon)]^2 & \text{if } j > j_\epsilon. \end{cases} \tag{12.35}$$

The larger penalty constants γ_{n_j} at levels $j > j_\epsilon$ are required to ensure convergence of a sum leading to the $\epsilon^2 \log \epsilon^{-2}$ term in the risk bound below, compare (12.38).

The penalized least squares estimator is equivalent to hard thresholding with level and data dependent threshold $\hat{t}_j = t_{n_j,\hat{k}_j}$ where $t^2_{n,k} = k\lambda^2_k - (k-1)\lambda^2_{k-1} \approx \lambda^2_k$ and $\hat{k}_j = N(\hat{\theta}_P(y_j))$ is the number of non-zero entries in $\hat{\theta}_P(y_j)$. Compare Proposition 11.2 and Lemma 11.7.

The levelwise estimators are combined into an overall estimator $\hat{\theta}^P = (\hat{\theta}^P_j)$ with $\hat{\theta}^P_j(y) = \hat{\theta}_P(y_j)$ for $j \ge 0$. Note that in this model there is no cutoff at a fixed level J.

Theorem 12.5 *Assume the correlated blocks model* (12.32) *and that*

$$\alpha > (2\beta + 1)(1/p - 1/2)_+. \tag{12.36}$$

For all such $\alpha > 0$ and $0 < p, q \le \infty$, for the penalized least squares estimator just described, there exist constants c_i such that if $C \in [\epsilon, \epsilon^{-2(\alpha+\beta)}]$, then

$$\begin{aligned} c_0 C^{2(1-r)}\epsilon^{2r} &\le R_N(\Theta^\alpha_{p,q}(C), \epsilon) \\ &\le \sup_{\Theta^\alpha_{p,q}(C)} E\|\hat{\theta}_P - \theta\|^2 \le c_1 C^{2(1-r)}\epsilon^{2r} + c_2\epsilon^2 \log \epsilon^{-2}. \end{aligned} \tag{12.37}$$

with $r = 2\alpha/(2\alpha+2\beta+1)$. The constants $c_1 = c_1(\alpha, \beta, \gamma, p, \zeta, \xi_1)$ and $c_2 = c_2(\beta, \gamma, \zeta, \xi_1)$.

First a comment about the restrictions on C. In fact, the first inequality holds if $C \ge \epsilon$, and the third inequality holds if $C \le \epsilon^{-2(\alpha+\beta)}$. When $C > \epsilon^{-2(\alpha+\beta)}$, the third inequality holds if $c_1 C^{2(1-r)}\epsilon^{2r} = c_1(C/\epsilon)^{2(1-r)}\epsilon^2$ is replaced by $c_1'(C/\epsilon)^{2(1-r)}\epsilon^2 \log_2(C/\epsilon)$, compare (12.44).

The key point of this theorem is that the estimator $\hat{\theta}_P$ achieves the correct rate of convergence without having to specify any of (α, p, q, C) in advance, subject only to smoothness condition (12.36). The range of validity of the bound for C increases as $\epsilon \to 0$ to include all positive values.

This is essentially a generalization of Theorem 12.1, to which it reduces if $\beta = 0$ and $\xi_0 = \xi_1 = 1$. We could modify $\hat{\theta}_P$ to cut off at a level $J = \log_2 \epsilon^{-2}$ as in that theorem; the result would be an additional tail bias term $cC^2(\epsilon^2)^{2\alpha'}$ in (12.37). In that case, we could also use $\gamma_n \equiv \gamma > e$ rather than the definition (12.35).

The proof has the same structure as in the direct case, Section 12.1. The lower bound, after bounding the effect of correlation, is found from the worst Besov shell. The upper bound uses a penalized least squares estimator, after a key modification to the oracle inequality, Section 11.7, to control the effect of noise inflation with level j. With these–not unimportant–changes, the argument is reduced to the analysis of the ℓ_p-ball control functions $r_{n_j,p}(C_j, \epsilon_j)$; this is deferred to the following section.

Proof We begin with the lower bound. It follows from the covariance comparison Lemma 4.28 that the minimax risk in correlated model (12.32) is bounded below by the risk in a

corresponding independence model in which the z_{jk} are i.i.d. $N(0, \xi_0)$. We may then restrict attention to the Besov shell $\Theta^{(j)} \cong \Theta_{n_j,p}(C_j)$ and conclude that

$$R_N(\Theta^{\alpha}_{p,q}(C), \epsilon) \geq R_N(\Theta_{n_j,p}(C_j), \xi_0\epsilon_j) \geq a_1 r_{n_j,p}(C_j, \xi_0\epsilon_j),$$

by the lower bound part of Theorem 11.6. It will be shown in the next section that this is bounded below by $c_1\xi_0^{2r}C^{2(1-r)}\epsilon^{2r}$.

Turning now to the upper bound, the levelwise structure of $\hat{\theta}_P$ implies that

$$E\|\hat{\theta}_P - \theta\|^2 = \sum_j E\|\hat{\theta}_{P,j} - \theta_j\|^2,$$

and we will apply at each level j the inverse problem variant, Theorem 11.10, of the oracle inequality for complexity penalized estimators. Indeed, at level j, from (12.32) we may assume a model $y_j = \theta_j + \epsilon_j z_j$ with $\dim(y_j) = n_j = 2^j$ and

$$\xi_0 I_{n_j} \leq \mathrm{Cov}(z_j) \leq \xi_1 I_{n_j}.$$

The level j penalized least squares estimator $\hat{\theta}_{P,j}(y_j)$ is as described at (12.34). Theorem 11.10 implies the existence of constants $a'(\zeta), b(\zeta)$ and $M'_j = M'_j(\beta, \gamma_j)$ such that

$$E\|\hat{\theta}_{P,j} - \theta_j\|^2 \leq b(\zeta)\xi_1 M'_j\epsilon_j^2 + a'(\zeta)\mathcal{R}_j(\theta_j, \epsilon_j)$$

where the level j minimum theoretical complexity

$$\mathcal{R}_j(\theta_j, \epsilon_j) = \min_{J\subset\{1,\ldots,2^j\}} C_j(J, \theta),$$

compare (11.25), is defined in terms of the theoretical complexity

$$C_j(J, \theta) = \sum_{k\notin J} \theta_{jk}^2 + \epsilon_j^2 \mathrm{pen}_j(n_J).$$

Consequently,

$$E\|\hat{\theta}_P - \theta\|^2 \leq b(\zeta)\xi_1 \sum_j M'_j\epsilon_j^2 + a'(\zeta)\sum_j \mathcal{R}_j(\theta_j, \epsilon_j).$$

For penalties of the form (12.34), we have using (11.66) a bound

$$M'_j \leq \gamma^{-1}c_{\beta,\gamma}2^{-2\beta j}\begin{cases}1 & j \leq j_\epsilon \\ [1 + (j - j_\epsilon)]^{-2} & j > j_\epsilon\end{cases}$$

which has the desired rapid decay with level j so that

$$\sum_j M'_j\epsilon_j^2 \leq \gamma^{-1}c_{\beta,\gamma}\epsilon^2\Big(j_\epsilon + \sum_{j>j_\epsilon}[1 + (j - j_\epsilon)]^{-2}\Big) \leq c_{\beta,\gamma}\epsilon^2\log\epsilon^{-2}, \tag{12.38}$$

where we have used the specific choice (12.35) of γ_{n_j}.

Since the observation model for y_j is orthogonal, we may argue as at (11.6) that

$$\mathcal{R}_j(\theta_j, \epsilon_j) = \min_{1\leq k\leq 2^j} \sum_{l>k} \theta_{j(l)}^2 + \epsilon_j^2 k\lambda_{j,k}^2.$$

Therefore, using (11.39), the minimum theoretical complexity satisfies

$$\sup_{\theta\in\Theta} \mathcal{R}_j(\theta_j, \epsilon_j) \le c(\log \gamma_{n_j}) r_{n_j,p}(C_j, \epsilon_j),$$

with $c = c(\zeta, \xi, \beta)$, and we note from (12.35) the bound

$$\log \gamma_{n_j} \le \log \gamma + 2(j - j_\epsilon)_+.$$

With the abbreviation $R_j = r_{n_j,p}(C_j, \epsilon_j)$, we arrive at

$$\sum_j \mathcal{R}_j(\theta_j, \epsilon_j) \le c \log \gamma \sum_j R_j + 2c \sum_{j>j_\epsilon} (j - j_\epsilon) R_j.$$

We have reduced our task to that of analyzing the 'shell bounds' R_j, to which we devote the next section. □

12.5 Taming the shell bounds

In the previous chapter, we saw that the minimax risk over ℓ_p balls could be precisely described, up to constant factors, by control functions $r_{n,p}(C, \epsilon)$ of relatively simple form. We recall the scale-ϵ versions here. In the simpler case, $p \ge 2$,

$$r_{n,p}(C, \epsilon) = \begin{cases} n^{1-2/p} C^2 & C \le \epsilon n^{1/p} \\ n\epsilon^2 & C \ge \epsilon n^{1/p}. \end{cases} \tag{12.39}$$

while for $p < 2$,

$$r_{n,p}(C, \epsilon) = \begin{cases} C^2 & C \le \epsilon \sqrt{1 + \log n} \\ C^p \epsilon^{2-p} [1 + \log\left(\frac{n\epsilon^p}{C^p}\right)]^{1-p/2} & \epsilon\sqrt{1+\log n} \le C \le \epsilon n^{1/p} \\ n\epsilon^2 & C \ge \epsilon n^{1/p}. \end{cases} \tag{12.40}$$

We may refer to these cases, from top to bottom, as the 'small signal', 'sparse' and 'dense' zones respectively, corresponding to the structure of the least favorable configurations in the lower bound proof of Theorem 11.6.

We have seen that the ℓ_p ball interpretation of Besov shells $\Theta^{(j)}$ leads, for level j, to the choices

$$n_j = 2^j, \qquad C_j = C 2^{-aj}, \qquad \epsilon_j = 2^{\beta j} \epsilon, \tag{12.41}$$

with $a = \alpha + 1/2 - 1/p$. Let us make the abbreviation

$$R_* = C^{2(1-r)} \epsilon^{2r}. \tag{12.42}$$

Proposition 12.6 *Suppose that $0 < p \le \infty$, $\beta \ge 0$ and $\alpha > (2\beta + 1)(1/p - 1/2)_+$. Let $R_j = r_{n_j,p}(C_j, \epsilon_j)$ denote the control functions* (12.39) *and* (12.40) *evaluated for the shell and noise parameters (n_j, C_j, ϵ_j) defined at* (12.41). *Define $r = 2\alpha/(2\alpha + 2\beta + 1)$. Then there exist constants $c_i(\alpha, \beta, p)$ such that*

$$c_1 C^{2(1-r)} \epsilon^{2r} = c_1 R_* \le \max_{j\ge 0} R_j \le \sum_{j\ge 0} R_j \le c_2 R_* = c_2 C^{2(1-r)} \epsilon^{2r}, \tag{12.43}$$

where the leftmost bound requires also that $C \geq \epsilon$. In addition, if $j_\epsilon = \log_2 \epsilon^{-2}$, then

$$\sum_{j>j_\epsilon} (j - j_\epsilon) R_j \leq \begin{cases} c_2' R_* & C \leq \epsilon^{-2(\alpha+\beta)} \\ c_2'' R_* \log_2(C/\epsilon) & C > \epsilon^{-2(\alpha+\beta)} \end{cases} \tag{12.44}$$

Proof Upper Bounds. We first consider the main case, $C \geq \epsilon$. The essence of the proof is to show that the shell bounds $j \to R_j$ peak at a critical level j_* (≥ 0 when $C \geq \epsilon$), and decay geometrically away from the value R_* at this least favorable level, so that the series in (12.43) are summable. Note that for these arguments, j is allowed to range over non-negative real values, with the results then specialized to integer values for use in (12.43). The behavior for $p < 2$ is indicated in Figure 12.1; the case $p \geq 2$ is similar and simpler.

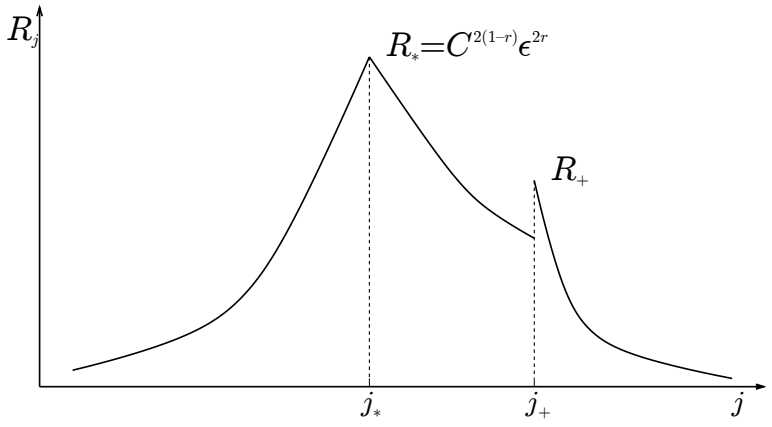

Figure 12.1 Schematic behavior of 'shell risks' R_j; with j treated as a real variable.

More specifically, in the case $p \geq 2$, we show that

$$R_j = \begin{cases} R_* 2^{(2\beta+1)(j-j_*)} & j \leq j_* \\ R_* 2^{-2\alpha(j-j_*)} & j \geq j_* \end{cases} \tag{12.45}$$

with the critical level $j_* \in \mathbb{R}$ being defined by

$$2^{(\alpha+\beta+1/2)j_*} = C/\epsilon, \tag{12.46}$$

and the maximum shell bound being given by a multiple of R_* in (12.42).

In the case $p < 2$, by contrast, there are three zones to consider and we show that

$$R_j = \begin{cases} R_* 2^{(2\beta+1)(j-j_*)} & j \leq j_* \\ R_* 2^{-p\rho(j-j_*)} [1 + \tau(j - j_*)]^{1-p/2} & j_* \leq j < j_+ \\ R_+ 2^{-2a(j-j_+)} & j \geq j_+ \end{cases} \tag{12.47}$$

where R_* is as before and $\rho = \alpha - (2\beta + 1)(1/p - 1/2) > 0$ in view of smoothness assumption (12.36). The values of j_+, τ and R_+ are given below; we also show that always $R_+ \leq c_{\alpha\beta p} R_*$, though typically it is of of smaller order than R_*.

To complete the proof, we establish the geometric shell bounds in (12.45) and (12.47), starting with the simpler case $p \geq 2$. Apply control function (12.39) level by level. Thus, on shell j, the boundary between small C_j and large C_j zones in the control function is given by the equation $(C_j/\epsilon_j)n_j^{-1/p} = 1$. Inserting the definitions from (12.41), we obtain an equation for j. Looking for a solution among real numbers, we obtain the formula (12.46) for the critical level j_*. One sees from (12.46) that $j_* \geq 0$ if and only if $C \geq \epsilon$.

In the large signal zone, $j \leq j_*$, the shell risks grow geometrically: $R_j = n_j\epsilon_j^2 = 2^{(2\beta+1)j}\epsilon^2$. The maximum is attained at $j = j_*$, and on substituting the definition of the critical level j_*, we obtain (12.42).

In the small signal zone, $j \geq j_*$, the shell bounds $R_j = C^2 2^{-2\alpha j}$ and it follows from (12.46) that $C^2 2^{-2\alpha j_*} = R_*$. We have established (12.45).

We turn to the case $p < 2$ and control function (12.40). Since $C_j/\epsilon_j = (C/\epsilon)2^{-(a+\beta)j}$ with $a + \beta > 0$ from (12.36), it is easily verified that the levels j belonging to the dense, sparse and small signal zones in fact lie in intervals $[0, j_*]$, $[j_*, j_+]$ and $[j_+, \infty)$ respectively, where j_* is again defined by (12.46) and $j_+ > j_*$ is the solution of

$$2^{(a+\beta)j_+}[1 + j_+ \log 2]^{1/2} = C/\epsilon.$$

First, observe that the sparse/dense boundary, the definition of j_* and the behavior for $j \leq j_*$ correspond to the small/large signal discussion for $p \geq 2$.

In the sparse zone, $j \in [j_*, j_+)$, the shell risks $R_j = C_j^p \epsilon_j^{2-p}[1 + \log(n_j\epsilon_j^p C_j^{-p})]^{1-p/2}$. Using (12.41), the leading term

$$C_j^p \epsilon_j^{2-p} = C^p \epsilon^{2-p} 2^{-p(a-2\beta/p+\beta)j}$$

decays geometrically for $j \geq j_*$, due to the smoothness assumption (12.36); indeed we have $a - 2\beta(1/p - 1/2) = \alpha - (2\beta + 1)(1/p - 1/2) > 0$. The logarithmic term can be rewritten using the boundary equation (12.46):

$$\log(n_j\epsilon_j^p C_j^{-p}) = p(\alpha + \beta + 1/2)(j - j_*)\log 2.$$

Set $\tau = p(\alpha + \beta + 1/2)\log 2$, we have shown for $j_* \leq j < j_+$ that

$$R_j = C^p \epsilon^{2-p} 2^{-p\rho j}[1 + \tau(j - j_*)]^{1-p/2}.$$

Putting $j = j_*$ gives $R_{j_*} = C^p \epsilon^{2-p} 2^{-p\rho j_*} = C_{j_*}^p \epsilon_{j_*}^{2-p} = n_{j_*}\epsilon_{j_*}^2 = R_*$ and yields the middle formula in (12.47).

In the highly sparse zone $j \geq j_+$, the shell risks $R_j = C_j^2 = C^2 2^{-2aj}$ decline geometrically from the maximum value $R_+ = C^2 2^{-2aj_+}$.

Having established bounds (12.47), we turn to establishing bounds (12.43) and (12.44). The upper bound in (12.43) will follow from the geometric decay in (12.47) once we establish a bound for R_+ in terms of R_*. Let $r_{n,p}^{(1)}$ and $r_{n,p}^{(2)}$ denote the first two functions in (12.40) and set $n+ = n_{j+}$. Then define

$$R_+ = \lim_{j \searrow j_+} R_j = r_{n+,p}^{(1)}(C_{j+}, \epsilon_{j+})$$
$$R'_+ = \lim_{j \nearrow j_+} R_j = r_{n+,p}^{(2)}(C_{j+}, \epsilon_{j+}).$$

We have

$$R'_+ = R_* 2^{-p\rho(j_+ - j_*)}[1 + \tau(j_+ - j_*)]^{1-p/2} \le c_{\alpha\beta p} R_*. \tag{12.48}$$

We saw at (11.33) the discontinuity in $C \to r_{n,p}(C)$ at $C = \sqrt{1 + \log n}$ was bounded, and so if we abbreviate $\check{C}_j = C_{j+}/\epsilon_{j+} = \sqrt{1 + \log n_{j+}}$, then

$$\frac{R_+}{R'_+} = \frac{r_{n_{j+},p}(\check{C}_j -)}{r_{n_{j+},p}(\check{C}_j +)} \le 2.$$

Consequently $R_+ \le 2R'_+ \le 2cR_*$ and the upper bound in (12.43) follows.

For (12.44), we observe that the condition $C \le \epsilon^{-2(\alpha+\beta)}$ implies that $j_* \le j_\epsilon$. Bounding $j - j_\epsilon$ by $j - j_+ + j_+ - j_*$, using (12.47) and $R_+ \le 2R'_+$, we have

$$\sum_{j > j_\epsilon} (j - j_\epsilon) R_j \le \sum_{j_* \le j < j_+} (j - j_*) R_j + \sum_{j > j_+} (j - j_+) R_j + 2(j_+ - j_*) R'_+ \sum_{j > j_+} 2^{-2a(j - j_+)}.$$

Each of the terms on the right side may be bounded by $c_2 R_*$ by use of geometric decay bounds, (12.47) and (12.48). Exercise 12.4 treats the remainin case $C > \epsilon^{-2(\alpha+\beta)}$.

Finally consider the case $C < \epsilon$. From definitions (12.41), note that $C_j/\epsilon_j = (C/\epsilon)2^{-(a+\beta)j} < 1$ for all $j \ge 0$, since $a + \beta > 0$. Consequently, using the small signal cases of (12.39) and (12.40), $R_j = C^2 2^{-2\alpha j}$ and $R_j = C^2 2^{-2aj}$ respectively, so that $\sum_{j \ge 0} R_j \le c_2 C^2$. When $\epsilon < C$, this is trivially and crudely bounded by $c_2 C^{2(1-r)} \epsilon^{2r}$, yielding again (12.43).

Lower Bound. $C \ge \epsilon$ implies $j_* \ge 0$ and it is enough now to observe from (12.47) that

$$\max R_j = \max(R_{\lfloor j_* \rfloor}, R_{\lceil j_* \rceil}, R_{\lceil j_+ \rceil}) \ge c_1 R_*.$$

[When $C < \epsilon$, we have $C_j/\epsilon_j \le C/\epsilon < 1$ for $j \ge 0$, and so from the first cases of (12.39) and (12.40), $\max R_j = R_0 = C^2$ which may be much less than $C^{2(1-r)}\epsilon^{2r}$.] □

12.6 Notes

Remark on critical/sparse regions in STAT paper.

The use of a larger threshold $\sqrt{2(1 + 2\beta)\log n}$ for dealing with noise amplification in inverse problems was advocated by Abramovich and Silverman (1998); these authors also studied a variant of the WVD in which the image function Af rather than f is explanded in a wavelet basis.

Exercises

12.1 *(Simple Fourier facts)*

Recall or verify the following.

(a) Suppose that ψ is C^L with compact support. Then $\hat{\psi}(\xi)$ is infinitely differentiable and

$$|\hat{\psi}^{(r)}(\xi)| \le C_r |\xi|^{-L} \qquad \text{for all } r.$$

(b) Suppose that ψ has K vanishing moments and compact support. Then for $r = 0, \ldots, K-1$, we have $\hat{\psi}^{(r)}(\xi) = O(|\xi|^{K-r})$ as $\xi \to 0$.

(c) For this and the next part, assume that f and $\hat{f}$ are integrable. Show that

$$|f(t)| \le (2\pi)^{-1} \int |\hat{f}(\xi)| d\xi$$

and

$$|f(t) - f(s)| \le (2\pi)^{-1} |t - s| \int |\xi| |\hat{f}(\xi)| d\xi.$$

(d) If $\hat{f}(\xi)$ is C^2 for $0 < |\xi| < \infty$ and if $\hat{f}(\xi)$ and $\hat{f}'(\xi)$ vanish as $|\xi| \to 0, \infty$, then

$$|f(t)| \le (2\pi)^{-1} t^{-2} \int |\hat{f}''(\xi)| d\xi,$$

12.2 *(Vaguelette properties for convolution examples)* (a) Let S_I be the rescaling operator (12.18), and suppose that the system of functions w_I, assumed integrable, can be represented in the Fourier domain via $\widehat{s_I} = \widehat{S_I w_I}$. Show that vaguelette conditions (12.19)–(12.21) are in turn implied by the existence of constants M_i, not depending on λ, such that

$$\text{(i)} \int |\hat{s}_I(\xi)| d\xi \le M_0, \qquad \int |\hat{s}_I''(\xi)| d\xi \le M_1,$$

$$\text{(ii)} \ \hat{s}_I(0) = 0, \qquad \text{and} \qquad \text{(iii)} \int |\xi| \, |\hat{s}_I(\xi)| d\xi \le M_2,$$

with $\widehat{s_I}(\xi)$ and $\widehat{s_I}'(\xi)$ vanishing at $0, \pm\infty$.

(b) Show that if $Af = a \star f$, then for the two systems

$$v_I = \kappa_j^{-1} A \psi_I, \qquad \widehat{s_I}(\xi) = \kappa_j^{-1} \hat{a}(2^j \xi) \hat{\psi}(\xi),$$

$$u_I = \kappa_j g_I, \qquad \widehat{s_I}(\xi) = [\kappa_j / \hat{a}(-2^j \xi)] \hat{\psi}(\xi).$$

(c) Suppose A is given by fractional integration, (12.24), for $0 < \beta < 1$. Suppose that ψ is C^3, of compact support and has $L = 2$ vanishing moments. Show that $\{u_I\}$ and $\{v_I\}$ are vaguelette systems.

(d) Suppose that A is given by convolution with either of the kernels in (12.28). Let $\beta = 1$ for a_1 and $\beta = 2$ for a_2. Suppose that ψ is $C^{2+\beta}$, of compact support and has $L = 2 + \beta$ vanishing moments. Show that $\{u_I\}$ and $\{v_I\}$ are vaguelette systems.

12.3 *(Comparing DeVore diagrams for adaptive estimators.)* Draw $(\alpha, 1/p)$ diagrams, introduced in Section 9.6, to compare regularity conditions for exact minimax rate convergence for some of the estimators in the literature recalled below. For these plots, ignore the regularity condition on the wavelet ψ, and the third Besov space parameter q.

(i) The SUREShrink estimator of Donoho and Johnstone (1995) assumes

$$\alpha > \max\left(1/p, 2(1/p - 1/2)_+\right), \qquad 1 \le p \le \infty.$$

(ii) The SureBlock method of Cai and Zhou (2009b) requires

$$\alpha > 4(1/p - 1/2)_+ + 1/2, \qquad \frac{2\alpha^2 - 1/6}{1 + 2\alpha} > 1/p, \qquad 1 \le p \le \infty.$$

(iii) The penalized least squares estimator of Theorem 12.1 requires

$$\alpha \ge 1/p \quad \text{for } p < 2, \qquad \alpha > 0 \quad \text{for } p \ge 2.$$

12.4 *(Shell bound when C is very large.)* Establish the second bound in (12.44), for example by considering separately j above and below $\tilde{j}_\epsilon := j_* + (j_* - j_\epsilon)$.

13

Sharp minimax estimation on ℓ_p balls

Suppose again that we observe n-dimensional data

$$y_i = \theta_i + \epsilon z_i \qquad\qquad i = 1, \ldots, n, \tag{13.1}$$

with $z_i \overset{\text{iid}}{\sim} N(0,1)$ and θ constrained to lie in a ball of radius C defined by the ℓ_p norm:

$$\Theta = \Theta_{n,p}(C) = \{\theta \in \mathbb{R}^n : \sum_{i=1}^{n} |\theta_i|^p \le C^p\}. \tag{13.2}$$

We seek to estimate θ using squared error loss $\|\hat{\theta} - \theta\|_2^2 = \sum_i (\hat{\theta}_i - \theta_i)^2$. and in particular to evaluate the nonlinear minimax risk

$$R_N(\Theta, \epsilon) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} E_\theta \|\hat{\theta} - \theta\|_2^2, \tag{13.3}$$

and make comparisons with the corresponding linear minimax risk $R_L(\Theta)$.

In previous chapters we have been content to describe the rates of convergence of $R_N(\Theta)$, or non-asymptotic bounds that differ by constant factors. In this chapter, and in the next for a multiresolution setting, we seek an exact, if often implicit, description of the asymptotics of $R_N(\Theta)$. 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

$$\eta_n = n^{-1/p}(C/\epsilon). \tag{13.4}$$

This may be interpreted as the maximum scalar multiple in standard deviation units of the vector $(1, \ldots, 1)$ that is contained within $\Theta_{n,p}(C)$. Alternatively, it is the bound on the average signal to noise ratio measured in ℓ_p-norm: $(n^{-1} \sum |\theta_i/\epsilon|^p)^{1/p} \le n^{-1/p}(C/\epsilon)$.

We also study linear and threshold estimators as two simpler classes that might or might not come close in performance to the full class of non-linear estimators. In each case we also aim for exact asymptotics of the linear or threshold minimax risk.

The ℓ_p- constrained parameter space Θ 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 4.

When $p < 2$, this parameter space may be said to impose a restriction of *approximate sparsity* on θ, as argued in earlier chapters (REFS). It represents a loosening of the requirement of *exact* sparsity studied in Chapter 8 using the ℓ_0 "norm", in the sense that condition (13.2) only requires that most components θ_i are small, rather than exactly zero. Nevertheless, we will see that many of the techniques introduced in Chapter 8 for exact evaluation

of minimax risk under exact sparsity have natural extensions to the setting of approximate sparsity discussed here.

We therefore follow the pattern established in the study of exact sparsity in Sections 8.4–8.8. In sparse cases, here interpreted as $\eta_n \to 0$ for $0 < p < 2$, considered in Section 13.2, thresholding, both soft and hard, again turns out to be (exactly) asymptotically minimax, so long as the threshold is chosen carefully to match the assumed sparsity. Matching lower bounds are constructed using the independent block spike priors introduced in Section 8.4: the argument is similar after taking account of the ℓ_p constraint.

In 'dense' cases, the asymptotic behavior of $R_N(\Theta)$ is described by a Bayes-minimax problem in which the components θ_i of θ are drawn *independently* from an appropriate *univariate* near least favorable distribution $\pi_{1,n}$.

Again, as in Sections 8.5 and 8.7, the strategy is first to study a univariate problem $y = \theta + \epsilon z$, with $z \sim N(0,1)$ and θ having a prior distribution π, now subject to a moment constraint $\int |\theta|^p d\pi \le \tau^p$. In this univariate setting, we can compare linear, threshold and non-linear estimators and observe the distinction between $p \ge 2$, with "dense" least favorable distributions, and $p < 2$, with "sparse" least favorable distributions placing most of their mass at zero.

This is done in Section 13.3, while the following Section 13.4 takes up the properties of the minimax threshold corresponding to the p-th moment constraint. This is used to show that thresholding comes within a (small) constant of achieving the minimax risk over all p and all moment constraints τ – this is an analog of Theorem 4.17 comparing linear and non-linear estimators over bounded intervals $[-\tau, \tau]$.

The second phase in this strategy is to "lift" the univariate results to the n-dimensional setting specified by (13.1)–(13.3). Here the independence of the co-ordinates of y_i in (13.1) and of the θ_i in the least favorable distribution is crucial. The details are accomplished using the Minimax Bayes approach sketched already in Chapter 4.

The Minimax Bayes strategy is not, however, fully successful in extremely sparse cases when the expected number of spikes $n\alpha_n$, remains bounded as n grows–Section 13.5 also compares the i.i.d. univariate priors with the independent block priors used in the sparse case. Finally section 13.6 returns to draw conclusions about near minimaxity of thresholding in the multivariate problem.

13.1 Linear Estimators.

With linear estimators, exact calculations of minimax risk are relatively straightforward and serve as a point of reference for work with non-linear estimators in later sections.

The ℓ_p balls $\Theta_{n,p}(C)$ are solid and orthosymmetric and compact for all $0 < p \le \infty$. However they are quadratically convex only if $p \ge 2$, while for $p < 2$,

$$\mathrm{QHull}[\Theta_{n,p}(C)] = \Theta_{n,2}(C). \tag{13.5}$$

Theorem 9.5 says that the linear minimax risk is determined by the quadratic hull, and so we may suppose that $p \ge 2$. Our first result evaluates the linear minimax risk, and displays the "corner" at $p = 2$.

Proposition 13.1 *Let $\bar{p} = p \vee 2$ and $\bar{\eta} = n^{-1/\bar{p}}(C/\epsilon)$. The minimax linear risk for squared*

error loss is

$$R_L(\Theta_{n,p}(C), \epsilon) = n\epsilon^2\bar{\eta}^2/(1+\bar{\eta}^2),$$

with minimax linear estimator $\hat{\theta}_L$ *given coordinatewise by*

$$\hat{\theta}_{L,i}(y) = [\bar{\eta}^2/(1+\bar{\eta}^2)]y_i.$$

Remark. For large C, and hence large $\bar{\eta}$, the minimax linear risk approaches the unconstrained minimax risk for $\mathbb{R}^n$, namely $n\epsilon^2$.

Proof In view of (13.5), we may suppose that $p = \bar{p} \geq 2$. Theorem 4.25 says that the linear minimax risk is found by looking for the hardest rectangular subproblem:

$$R_L(\Theta_{n,p}(C)) = \sup\Big\{\sum_1^n \epsilon^2\tau_i^2/(\epsilon^2+\tau_i^2) : \sum_1^n \tau_i^p \leq C^p\Big\}.$$

In terms of new variables $u_i = \tau_i^p/C^p$, and a scalar function $\ell(t) = t/(1+t)$, this optimization can be rephrased as that of maximizing

$$f(u) = \epsilon^2 \sum_i \ell(C^2\epsilon^{-2}u_i^{2/p})$$

over the simplex $\sum_1^n u_i \leq 1$ 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 $u = n^{-1}(1,\ldots,1)$. Introducing the normalized radius $\bar{\eta} = n^{-1/(p\vee 2)}(C/\epsilon)$, we may write the corresponding minimax risk as $n\epsilon^2\ell(\bar{\eta}^2)$. From (4.29), the corresponding linear minimax estimate is $\hat{\theta}_L = \ell(\bar{\eta}_n^2)y$. □

Example 13.2 The calibration $\epsilon = 1/\sqrt{n}$ arises frequently in studying sequence model versions of nonparametric problems, compare (1.26). Consider the ℓ_1 ball of radius $C = 1$, namely $\Theta_{n,1} = \{\theta : \sum_1^n |\theta_i| \leq 1\}$. We see that $\bar{\eta} = n^{-1/2}\cdot n^{1/2} = 1$ and that

$$R_L(\Theta_{n,1}) = 1/2, \qquad \hat{\theta}_L(y) = y/2.$$

The proposition just proved shows that $\Theta_{n,1}$ has the same *linear* minimax risk as the solid sphere $\Theta_{n,2}$, though the latter is much larger, for example in terms of volume. We have already seen, in Example 8.2, that non-linear thresholding yields a much smaller maximum risk over $\Theta_{n,1}$ — the exact behavior of $R_N(\Theta_{n,1})$ is given at (13.39) below.

13.2 Asymptotic Minimaxity in the Sparse Case

In this section, we develop exact expressions for the asymptotic behavior of the minimax MSE over ℓ_p balls in the case of approximate sparsity, $p < 2$ and $\eta_n \to 0$. The ℓ_0 case, studied in Sections 8.4– 8.8 can be used to motivate the definitions we need and the resulting formulas. Indeed, we saw that if the number of non-zero components of θ is bounded by k_n^0 out of n, then an asymptotically minimax estimator is given by soft or hard thresholding at $\lambda_n\epsilon_n = \epsilon_n\sqrt{2\log(n/k_n^0)}$. The corresponding asymptotically least favorable prior uses k_n^0 blocks of size $[n/k_n^0]$ and independently within each block chooses a single spike of random location and height approximately $\lambda_n\epsilon_n$ (in fact $(\lambda_n - \log\lambda_n)\epsilon_n$).

In the ℓ_p case, it is not the *number* of non-zero components that is constrained, but rather the ℓ_p-norm. We might then conjecture that the number k_n in a least favorable configuration might be determined by the spike height and the ℓ_p-condition, that is, by a condition of the form

$$k_n(\lambda_n\epsilon_n)^p = k_n\epsilon_n^p\,(2\log(n/k_n))^{p/2} \approx C_n^p. \tag{13.6}$$

To simplify this equation for k_n, replace k_n in the logarithmic term by $C_n^p/\epsilon_n^p = n\eta_n^p$, since this would be the value for k_n given by ignoring the log term completely. More precisely, we use $\max\{C_n^p/\epsilon_n^p, 1\}$, since there must be at least one spike, and this leads to the definition

$$t_n^2 = 2\log(n/(C_n^p\epsilon_n^{-p} \vee 1)) = \min\{2\log\eta_n^{-p}, 2\log n\}. \tag{13.7}$$

This amounts to use of a threshold $t_n\epsilon_n$ of $\epsilon_n\sqrt{2\log\eta_n^{-p}}$ that is capped above at the 'universal' level $\epsilon_n\sqrt{2\log n}$. It is now convenient to define an approximate solution to equation (13.6) for the number of blocks as

$$\kappa_n = C_n^p/(\epsilon_n^p t_n^p) = n\eta_n^p/t_n^p, \tag{13.8}$$

and to observe that when $\eta_n \to 0$, on setting $\check{\kappa}_n = \kappa_n \vee 1$,

$$2\log(n/\check{\kappa}_n) \sim t_n^2. \tag{13.9}$$

To verify this, start from (13.7) and use (13.8), first via the inequality $\eta_n^{-p} \le n/\kappa_n$ and then via the equality $n/\kappa_n = \eta_n^{-p}t_n^p$, we get

$$\begin{aligned} t_n^2 &\le \min\{2\log(n/\kappa_n), 2\log n\} = 2\log n/\check{\kappa}_n \\ &= \min\{2\log\eta_n^{-p} + p\log t_n^2, 2\log n\} \\ &\le t_n^2 + p\log t_n^2, \end{aligned}$$

which immediately yields (13.9).

We also recall a function first encountered in Lemma 9.3:

$$R(t) = [t] + \{t\}^{2/p}. \tag{13.10}$$

where $[\cdot]$ and $\{\cdot\}$ denote integer and fractional parts respectively. See Figure 13.1.

Theorem 13.3 *Let $R_N(C_n, \epsilon_n)$ denote the minimax risk* (13.3) *for estimation over the ℓ_p ball $\Theta_{n,p}(C_n)$ defined at* (13.2). *Define radius η_n by* (13.4), *threshold t_n by* (13.7), *then κ_n by* (13.8) *and $R(t)$ by* (13.10). *Finally, let $\check{\kappa}_n = \kappa_n \vee 1$.*

If $0 < p < 2$ and $\eta_n \to 0$, and if $\kappa_n \to \kappa \in [0, \infty]$, then

$$R_N(C_n, \epsilon_n) \sim R(\kappa_n)\,\epsilon_n^2 \cdot 2\log(n/\check{\kappa}_n). \tag{13.11}$$

An asymptotically minimax rule is given by soft thresholding at $\epsilon_n t_n$ if $\eta_n \ge n^{-1/p}$ and by the zero estimator otherwise.

Remarks. 1. Expression (13.11) is the analog, for approximate ℓ_p sparsity, of the result (8.45) obtained in the case of exact, ℓ_0, sparsity. Indeed, the proof will show that κ_n, or

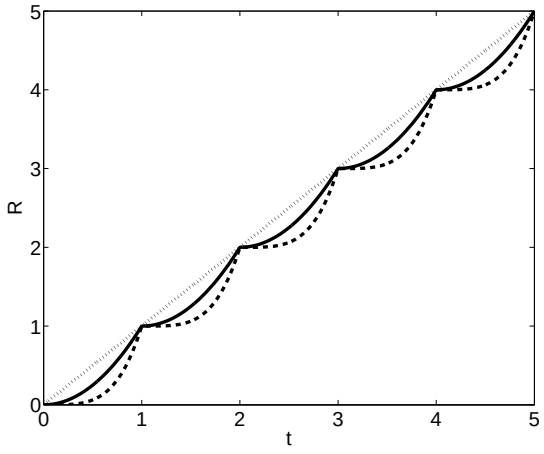

Figure 13.1 The function $R(t) = [t] + \{t\}^{2/p}$ plotted against t for $p = 1$ (solid) and $p = 1/2$ (dashed). The 45° line (dotted) shows the prediction of the Bayes minimax method.

more precisely $[\kappa_n] + 1$, counts the number of non-zero components in a near least favorable configuration. The expression (13.11) simplifies somewhat according to the behavior of κ_n:

$$R_N(C_n, \epsilon_n) \sim \begin{cases} \kappa_n \epsilon_n^2 \cdot 2\log(n/\kappa_n) & \text{if} \quad \kappa_n \to \infty \\ \big([\kappa] + \{\kappa\}^{2/p}\big)\epsilon_n^2 \cdot 2\log n & \text{if} \quad \kappa_n \to \kappa \in (0, \infty) \\ \kappa_n^{2/p} \epsilon_n^2 \cdot 2\log n & \text{if} \quad \kappa_n \to 0. \end{cases} \tag{13.12}$$

When $\kappa_n \to \infty$, (and when $\kappa_n \to \kappa \in \mathbb{N}$), the limiting expression agrees with that found in the case of exact (or ℓ_0) sparsity considered in Theorem 8.20, with κ_n playing the role of k_n there.

When $\eta_n < n^{-1/p}$ it would also be possible to use thresholding at $\epsilon_n \sqrt{2\log \eta_n^{-p}}$, but the condition implies that all $|\theta_i| \le \epsilon_n$ so thresholding is not really needed.

2. The dependence on noise level ϵ_n can be handled by rescaling: directly or from (4.74), $R_N(C_n, \epsilon_n) = \epsilon_n^2 R_N(C_n/\epsilon_n, 1)$. We therefore take $\epsilon_n = 1$ in the proof. Notice that the radius η_n, the threshold parameters λ_n, t_n and the block count parameters $\kappa_n, \check{\kappa}_n$ are all 'scale-free': they depend on C_n and ϵ_n only through C_n/ϵ_n.

Proof (with $\epsilon_n = 1$). *Upper Bound.* When $\eta_n < n^{-1/p}$ expression (13.8) shows that $\kappa_n \le 1$, and so it will be enough to verify that $R_N(C_n, 1) \le C_n^2$. But this is immediate, since for the zero estimator, we have $r(\hat{\theta}_0, \theta) = \|\theta\|_2^2 \le \|\theta\|_p^2 \le C_n^2$.

For $\eta_n \ge n^{-1/p}$ we use soft thresholding at $t_n = \sqrt{2\log \eta_n^{-p}}$ in bound (8.12) to get

$$\sum_i r_S(t_n, \theta_i) \le n\, r_S(t_n, 0) + \sum_i \min(\theta_i^2, t_n^2 + 1).$$

Using bound (8.7) and definitions (13.7) and (13.8), we have

$$nr_S(t_n, 0) \le c_1 n\phi(t_n)/t_n^3 \le c_2 n\eta_n^p/t_n^3 = c_2\kappa_n t_n^{p-3}.$$

For the second term, evaluate the maximum over $\Theta_{n,p}(C_n)$ using Lemma 9.3, with $\epsilon = (t_n^2+1)^{1/2}$ and $\gamma^p = C_n^p/(t_n^2+1)^{p/2} \le \kappa_n$. Since $R(\cdot)$ is monotone, we have

$$\sum_i \min(\theta_i^2, t_n^2+1) \le (t_n^2+1)\min\{R(\gamma^p), n\} \le (t_n^2+1)\min\{R(\kappa_n), n\} \tag{13.13}$$

We now claim that this bound dominates that in the previous display. Consider first the case $\kappa_n \le 1$, in which $C_n^2 \le t_n^2$, and so the middle bound in (13.13) becomes C_n^2. Since $n\eta_n^p = C_n^p \ge 1$, we have $n\eta_n^p/t_n^3 \ll C_n^p \le C_n^2$, and so bound (13.13) indeed dominates. When $\kappa_n \ge 1$, note also that $\eta_n \to 0$ implies $\kappa_n/n \to 0$ and so

$$\kappa_n t_n^{p-3} \ll t_n^2\kappa_n \asymp (t_n^2+1)\min\{R(\kappa_n), n\},$$

so that again bound (13.13) dominates. We arrive at

$$\sup_{\Theta_n} r(\hat\theta, \theta) \le t_n^2 R(\kappa_n)(1+o(1)).$$

To conclude, appeal to the equivalence of t_n^2 with $2\log(n/\check\kappa_n)$, compare (13.9).

Lower Bound. We adapt the approach taken in the ℓ_0-case, building on Proposition 8.12. Divide $\{1,\dots,n\}$ into $[\kappa_n]+1$ contiguous blocks, each being of length m_n, the integer part of $n/([\kappa_n]+1)$. The assumption $\eta_n \to 0$ along with identity (13.8) confirms that $m_n \to \infty$ and that

$$\log m_n \sim \log(n/\kappa_n).$$

The near least favorable prior π_n is built from independent single spike priors on these blocks. The spike height τ_n is set at $\tau_n = t_n - \log t_n$. It is left as (easy) Exercise 13.1 to check that this choice meets the requirements of Proposition 8.12, namely that

$$\sqrt{2\log m_n} - \tau_n \to \infty.$$

For each of the first $[\kappa_n]$ blocks, we use single spike prior $\pi_S(\tau_n; m_n)$. The final block uses instead $\pi_S(\tilde\tau_n; m_n)$, with $\tilde\tau_n = \{\kappa_n\}^{1/p}\tau_n$. To verify that $\pi_n(\Theta_n) = 1$, use the definition of τ_n and (13.8) to observe that

$$\sum_{i=1}^n |\theta_i|^p = [\kappa_n]\tau_n^p + \{\kappa_n\}\tau_n^p = \kappa_n\tau_n^p = n\eta_n^p(t_n - \log t_n)^p/t_n^p < n\eta_n^p.$$

From (4.13)–(4.15), independence across blocks and Proposition 8.12,

$$\begin{aligned} R_N(C_n, 1) \ge B(\pi_n) &= [\kappa_n]B(\pi_S(\tau_n; m_n)) + B(\pi_S(\tilde\tau_n; m_n)) \\ &\sim ([\kappa_n] + \{\kappa_n\}^{2/p})t_n^2 = R(\kappa_n)t_n^2, \end{aligned}$$

and the lower bound half of (13.11) follows from (13.9). □

13.3 Univariate Bayes Minimax Problem

To prepare for the study of 'dense' cases, we begin by considering a univariate p-th moment problem which generalizes both the bounded normal mean problem of Section 4.6 and the sparse normal mean setting of Section 8.7. Suppose that $y \sim N(\theta, \epsilon^2)$, and that θ is distributed according to a prior $\pi(d\theta)$ on $\mathbb{R}$. For $0 < p \le \infty$, assume that π belongs to a class satisfying the p-th moment constraint

$$\mathfrak{m}_p(\tau) = \{\pi(d\theta) : \int |\theta|^p \pi(d\theta) \le \tau^p\},$$

where, interpreting the norm for $p = \infty$ as a supremum, $\mathfrak{m}_\infty(\tau)$ equals the set of priors supported on the bounded interval $[-\tau, \tau]$. With an abuse of notation one can regard the sparse signal model of Section 8.7 as being the $p = 0$ limit of the p-th moment constraint. Indeed, since $\int |\theta|^p d\pi \to \pi\{\theta \neq 0\}$ as $p \to 0$, we can view $\mathfrak{m}_0(t) = \{\pi : \pi\{\theta \neq 0\} \le t\}$ as $\lim_{p\to 0} \mathfrak{m}_p(t^{1/p})$.

The classes $\mathfrak{m}_p(\tau)$ are convex and weakly compact for all $p \le \infty$ and $\tau < \infty$. Such moment constraints are a population version of the "empirical" constraints on $(\theta_1, \ldots \theta_n)$ defining an ℓ_p-ball—compare (13.2).

We study the Bayes minimax risk

$$\beta_p(\tau, \epsilon) = \inf_{\hat\theta} \sup_{\pi \in \mathfrak{m}_p(\tau)} B(\hat\theta, \pi) = \sup\{B(\pi) : \pi \in \mathfrak{m}_p(\tau)\}. \tag{13.14}$$

where the second equality uses the minimax Theorem 4.12 and (4.14) of Chapter 4.

In particular, $\beta_\infty(\tau, \epsilon) = \rho_N(\tau, \epsilon)$, compare (4.26) and (4.19). In addition, the sparse Bayes minimax risk $\beta_0(\tau, \epsilon) = \lim_{p\to 0} \beta_p(\tau^{1/p}, \epsilon)$.

A remark on Notation. We use the lower case letters β and ρ for Bayes and frequentist minimax risk in *univariate* problems, and the upper case letters B and R for the corresponding multivariate minimax risks.

We begin with some basic properties of $\beta_p(\tau, \epsilon)$, valid for all p and τ, and then turn to the interesting case of low signal, $\tau \to 0$, where the distinction between $p < 2$ and $p \ge 2$ emerges clearly.

Proposition 13.4 *The Bayes minimax risk $\beta_p(\tau, \epsilon)$, defined at* (13.14)*, is*

(i) *decreasing in p,*
(ii) *increasing in ϵ,*
(iii) *strictly increasing, concave and continuous in $\tau^p > 0$,*
(iv) *and satisfies (i) $\beta_p(\tau, \epsilon) = \epsilon^2 \beta_p(\tau/\epsilon, 1)$, and*
(ii) $\beta_p(a\tau, \epsilon) \le a^2 \beta_p(\tau, \epsilon)$ for all $a \ge 1$.

Proof First, (1) and 4(i) are obvious, while (2) is Lemma 4.28. Using 4(i), $a^2\beta_p(\tau, \epsilon) = \beta_p(a\tau, a\epsilon)$, so 4(ii) is a consequence of (2). For (3), let $t = \tau^p$: the function $\tilde\beta(t) = \sup\{B(\pi) : \int |\theta|^p d\pi = t\}$ is concave in t because $\pi \to B(\pi)$ is concave and the constraint on π is linear. Monotonicity in τ^p is clear, and continuity follows from monotonicity and 4(ii). Strict monotonicity then follows from concavity. □

The scaling property 4(i) means that it suffices to study the unit noise situation. As in

previous chapters, 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.

Information about the least favorable distribution follows from an extension of our earlier results for $p = \infty$, Proposition 4.19, and $p = 0$, Proposition 8.19. (For the proof, see Exercise 13.5).

Proposition 13.5 *For p and τ in $(0, \infty)$, the Bayes minimax problem associated with $\mathfrak{m}_p(\tau)$ and $\beta_p(\tau)$ has a unique least favorable distribution π_τ. If $p = 2$, then π_τ is Gaussian, namely $N(0, \tau^2)$; while for $p \neq 2$ instead π_τ is proper, symmetric and has discrete support with $\pm\infty$ as the only possible accumulation points. When $p < 2$ the support must be countably infinite.*

Proposition 4.14 then assures us that the Bayes estimator corresponding to π_τ is minimax for $\mathfrak{m}_p(\tau)$.

Thus, the only case in which completely explicit solutions are available is $p = 2$, for which $\beta_2(\tau, \epsilon) = \tau^2\epsilon^2/(\tau^2 + \epsilon^2) = \rho_L(\tau, \epsilon)$, Corollary 4.6 and (4.28). 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 $p > 0$.

$p \geq 2$ *versus* $p < 2$ *in low signal-to noise.*

When $p < 2$ and the moment constraint is small, appropriate choices of two point priors $\pi_{\alpha,\mu} = (1 - \alpha)\delta_0 + \alpha\delta_\mu$ turn out to be approximately least favorable. We build on the discussion of sparse two point priors in Section 8.5. A one parameter family of priors $\pi_{\alpha,\mu(\alpha)}$ was defined there by requiring $\mu(\alpha)$ to satisfy the equation

$$\mu^2/2 + (2\log\alpha^{-1})^{1/4}\mu = \log((1-\alpha)/\alpha), \tag{13.15}$$

and the resulting sparse prior, defined for $\alpha < 1/2$, was said to have sparsity α and overshoot $a = (2\log\alpha^{-1})^{1/4}$.

Definition 13.6 The *sparse ℓ_p prior* $\pi_p[\eta]$ is the sparse prior $\pi_{\alpha,\mu(\alpha)}$ with $\alpha = \alpha_p(\eta)$ determined by the moment condition

$$\alpha\mu^p(\alpha) = \eta^p. \tag{13.16}$$

We write $\mu_p(\eta) = \mu(\alpha_p(\eta))$ for the location of the non-zero support point, and use notation η rather than τ for a small moment constraint.

Exercise 13.2 shows that this definition makes sense for η sufficiently small. Recalling from (8.54) that $\mu(\alpha) \sim \sqrt{2\log\alpha^{-1}}$ for α small, one can verify that as $\eta \to 0$,

$$\alpha_p(\eta) \sim \eta^p(2\log\eta^{-p})^{-p/2} \tag{13.17}$$

$$\mu_p(\eta) \sim (2\log\eta^{-p})^{1/2}. \tag{13.18}$$

Thus, for example,

$$\mu_p^2(\eta) \sim 2\log\alpha_p(\eta)^{-1} = 2\log\eta^{-p} + p\log\mu_p^2(\eta) \sim 2\log\eta^{-p}.$$

We can now state the main result of this subsection.

Theorem 13.7 *As $\eta \to 0$,*

$$\beta_p(\eta) \sim \begin{cases} \eta^2 & 2 \le p \le \infty \\ \eta^p (2\log \eta^{-p})^{1-p/2} & 0 < p < 2. \end{cases} \tag{13.19}$$

If $p \ge 2$, then $\hat{\delta}_0 \equiv 0$ is asymptotically minimax and $\pi = (\delta_{-\eta} + \delta_\eta)/2$ is asymptotically least favorable.

If $p < 2$, then $\hat{\delta}_\lambda$, soft thresholding with threshold $\lambda = \sqrt{2\log \eta^{-p}}$, is asymptotically minimax. The sparse ℓ_p prior $\pi_p[\eta]$ of Definition 13.6 is asymptotically least favorable.

Remarks. 1. In the "nearly black" model of Section 8.7, corresponding to $p = 0$, we found that $\beta_0(\eta) \sim \eta \cdot (2\log \eta^{-1})$ with $\hat{\delta}_\lambda(x)$ being asymptotically minimax with $\lambda = \sqrt{2\log \eta^{-1}}$ and an asymptotically least favorable prior being $\pi_{\eta,\mu(\eta)}$. To see that this ℓ_p theorem is consistent with the $p = 0$ limit, observe that (13.19) implies $\beta_p(\eta^{1/p}) \sim \eta(2\log \eta^{-1})^{1-p/2}$ and recall that $\beta_0(\eta) = \lim_{p\to 0} \beta_p(\eta^{1/p})$.

2. Consider the special choice $\epsilon = n^{-1/2}$. Then $\eta_n^p = n^{-1}(C/\epsilon)^p = n^{-1+p/2}C^p$ 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, $\eta_n^p = 1/n$, we recover the choice $\lambda_n = \sqrt{2\log n}$ discussed in earlier chapters.

3. The threshold estimator $\hat{\delta}_{\sqrt{2\log \eta^{-p}}}$ is also asymptotically minimax when $p \ge 2$, Exercise 13.4.

Proof Consider first $p \ge 2$. For any prior $\pi \in \mathfrak{m}_p(\eta)$,

$$B(\hat{\delta}_0, \pi) = E_\pi \mu^2 \le (E_\pi |\mu|^p)^{2/p} \le \eta^2. \tag{13.20}$$

Consequently $B(\pi) \le \eta^2$, and so also $\beta_p(\eta) \le \eta^2$. In the other direction, consider the symmetric two point prior $\pi_\eta = (1/2)(\delta_\eta + \delta_{-\eta})$; together with (13.14), formula (2.30) for the Bayes risk shows that $\beta_p(\eta) \ge B(\pi_\eta) \sim \eta^2$ as $\eta \to 0$.

Suppose now that $p < 2$. For the lower bound in (13.19), we use the priors $\pi_p[\eta]$ and the asymptotics for their Bayes risks computed in Lemma 8.11. From this and (13.18), we obtain our desired lower bound as $\alpha \to 0$

$$\beta_p(\eta) \ge B(\pi_{\alpha(\eta)}) \sim \alpha\mu^2(\alpha) = \alpha\mu^p(\alpha) \cdot \mu^{2-p}(\alpha) \sim \eta^p (2\log \eta^{-p})^{1-p/2}.$$

For the upper bound, we use an inequality for the maximum integrated risk of soft thresholding:

$$\sup_{\pi \in \mathfrak{m}_p(\eta)} B(\hat{\delta}_\lambda, \pi) \le r_S(\lambda, 0) + \eta^p (1+\lambda^2)^{1-p/2}. \tag{13.21}$$

Assuming this for a moment, we note that $\beta_p(\eta)$ is bounded above by the left side, and in the right side we set $\lambda = \sqrt{2\log \eta^{-p}}$. Recalling from (8.7) that $r_S(\lambda, 0) \sim 4\lambda^{-3}\phi(\lambda) = o(\eta^p)$ as $\eta \to 0$, we see that the second term is dominant and is asymptotically equivalent to $\eta^p(2\log \eta^{-p})^{1-p/2}$ as $\eta \to 0$.

It remains to prove (13.21). We use the risk bound for soft thresholding given at (8.12),

and shown schematically in Figure 13.2. Now, define $\mu_\lambda = (1+\lambda^2)^{1/2}$, and then choose $c = c_\lambda$ so that

$$c\mu_\lambda^p = \mu_\lambda^2 = 1+\lambda^2,$$

that is, $c = (1+\lambda^2)^{1-p/2}$. Compare Figure 13.2. We conclude that

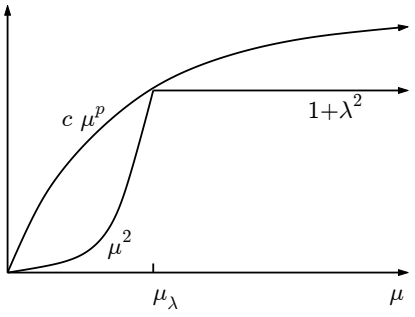

Figure 13.2 Schematic for risk bound: although the picture shows a case with $p < 1$, the argument works for $p < 2$.

$$\begin{aligned} B(\hat\delta_\lambda, \pi) = \int r_S(\lambda,\mu) d\pi &\le r_S(\lambda,0) + c\int \mu^p d\pi \\ &= r_S(\lambda,0) + \eta^p (1+\lambda^2)^{1-p/2}. \end{aligned}$$

As this holds for all $\pi \in \mathfrak{m}_p(\eta)$, we obtain (13.21). Here we used symmetry of $\mu \to r(\lambda,\mu)$ about 0 to focus on those π supported in $[0,\infty)$. □

Remark. There is an alternative approach to bounding $\sup_{\mathfrak{m}_p(\eta)} B_S(\hat\delta_\lambda,\pi)$ which looks for the maximum of the linear function $\pi \to B_S(\hat\delta_\lambda,\pi)$ among the extreme points of the convex $\mathfrak{m}_p(\eta)$ and shows that the maximum is actually of the two point form (8.48). This approach yields (see Exercise 13.8)

Proposition 13.8 *Let $p < 2$ and a threshold λ and moment space $\mathfrak{m}_p(\eta)$ be given. Then*

$$\begin{aligned} \sup\{B(\hat\delta_\lambda,\pi) : \pi \in \mathfrak{m}_p(\eta)\} &= \sup_{\mu\ge\eta} \; r(\lambda,0) + (\eta/\mu)^p [r(\lambda,\mu) - r(\lambda,0)] \\ &\le r(\lambda,0) + \eta^p \mu_\lambda^{2-p} \end{aligned} \tag{13.22}$$

where μ_λ is the unique solution of

$$r(\lambda,\mu_\lambda) - r(\lambda,0) = (\mu_\lambda/p) r_\mu(\lambda,\mu_\lambda). \tag{13.23}$$

The least favorable prior for $\hat\delta_\lambda$ over $\mathfrak{m}_p(\eta)$ is of the two point prior form with α determined from η and $\mu = \mu_\lambda$ by (13.16)*. As $\lambda \to \infty$, we have*

$$\mu_\lambda - \lambda \sim \tilde\Phi^{-1}(p/2). \tag{13.24}$$

Hard thresholding. It is of some interest, and also explains some choices made in the analysis of Section 1.3, to consider when *hard* thresholding $\hat\delta_{H,\lambda}$ is asymptotically minimax.

Theorem 13.9 *If $p < 2$ and $\eta \to 0$, then the hard thresholding estimator $\hat{\delta}_{H,\lambda}$ is asymptotically minimax over $\mathfrak{m}_p(\eta)$ if*

$$\lambda^2 = \begin{cases} 2\log \eta^{-p} & \text{if } 0 < p < 1 \\ 2\log \eta^{-p} + \alpha \log(2\log \eta^{-p}) & \text{if } 1 \le p < 2, \alpha > p - 1. \end{cases} \tag{13.25}$$

The introductory Section 1.3 considered an example with $p = 1$ and $\eta_n = n^{-1/2}$ so that $2\log \eta_n^{-1} = \log n$. In this case the threshold $\lambda = \sqrt{\log n}$ is not asymptotically minimax: the proof below reveals that the risk at 0 is too large. To achieve minimaxity for $p \ge 1$, a slightly larger threshold is needed, and in fact $\lambda_n = \sqrt{\log(n \log^{\alpha} n)}$ works for any $\alpha > 0$.

Proof We adopt a variant of the approach used for soft thresholding. It is left as Exercise 13.3 to use Lemma 8.5 to establish that if $c_\lambda = \lambda^{-p}(1 + \lambda^2)$ and $\lambda \ge \lambda_0(p)$, then

$$r_H(\lambda, \mu) \le r_H(\lambda, 0) + c_\lambda \mu^p. \tag{13.26}$$

Consequently, integrating over any $\pi \in \mathfrak{m}_p(\eta)$, we obtain

$$B(\hat{\delta}_{H,\lambda}, \pi) \le r_H(\lambda, 0) + c_\lambda \eta^p.$$

Since our choices $\lambda(\eta) \to \infty$ as $\eta \to 0$, we may use (8.15), namely $r_H(\lambda, 0) \sim 2\lambda\phi(\lambda)$, to conclude that

$$\sup_{\pi \in \mathfrak{m}_p(\eta)} B(\hat{\delta}_{H,\lambda}, \pi) \le [2\lambda\phi(\lambda) + \lambda^{2-p}\eta^p](1 + o(1)).$$

Since $\lambda \sim \sqrt{2\log \eta^{-p}}$, we obtain minimaxity for hard thresholding so long as the term due to the risk at zero is negligible as $\eta \to 0$:

$$\lambda\phi(\lambda) = o(\lambda^{2-p}\eta^p).$$

It is easily checked that for $0 < p < 1$, this holds true for $\lambda^2 = 2\log \eta^{-p}$, whereas for $1 \le p < 2$, we need the somewhat larger threshold choice in the second line of (13.25). □

For soft thresholding, the risk at zero $r_S(\lambda, 0) \sim 4\lambda^{-3}\phi(\lambda)$ is a factor λ^{-4} smaller than for hard thresholding with the same (large) λ; this explains why larger thresholds are only needed in the hard threshold case.

13.4 Univariate Thresholding

We have seen that thresholding at an appropriate level is minimax in the limit of low signal-to-noise. In this section we look more systematically at the choice of threshold that minimizes mean squared error. We consider the optimal performance of the best threshold rule over the moment space $\mathfrak{m}_p(\tau)$ with the goal of comparing it to the minimax Bayes estimator, which although optimal, is not available explicitly. Define therefore

$$\beta_{S,p}(\tau, \epsilon) = \inf_{\lambda} \sup_{\pi \in \mathfrak{m}_p(\tau)} B(\hat{\delta}_\lambda, \pi), \tag{13.27}$$

where $\hat{\delta}_\lambda$ refers to a soft threshold estimator (8.3) with threshold λ. Throughout this section, we work with soft thresholding, sometimes emphasised by the subscript "S", though some analogous results are possible for hard thresholding (see Donoho and Johnstone (1994b).)

A goal of this section is to establish an analogue of Theorem 4.17, 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 $\mathfrak{m}_p(\tau)$, the preceding sections show that we have to replace linear by threshold estimators. To emphasize that the choice of estimator in (13.27) is restricted to thresholds, we write

$$B(\lambda, \pi) = B(\hat{\delta}_\lambda, \pi) = \int r(\lambda, \mu)\pi(d\mu).$$

Let $B_S(\pi) = \inf_\lambda B(\lambda, \pi)$ denote the best MSE attainable by choice of soft threshold. Our first task is to establish that a unique best $\lambda(\pi)$ exists, Proposition 13.11 below. Then follows a (special) minimax theorem for $B(\lambda, \pi)$. This is used to derive some properties of $\beta_{S,p}(\tau, \epsilon)$ which finally leads to the comparison result, Theorem 13.15.

To begin, we need some preliminary results about how the MSE varies with the threshold.

Dependence on threshold. Let $r_\lambda(\lambda, \mu) = (\partial/\partial\lambda)r(\lambda, \mu)$; from (8.82) and changes of variable one obtains

$$r_\lambda(\lambda, \mu) = 2\int_{-\infty}^{-\mu} w\phi(w - \lambda)dw + 2\int_{-\infty}^{\mu} w\phi(w - \lambda)dw.$$

In particular, for all $\lambda \geq 0$ and μ

$$r_\lambda(0, \mu) = 4\int_{-\infty}^{-|\mu|} w\phi(w)dw < 0, \quad \text{and} \tag{13.28}$$

$$r_\lambda(\lambda, 0) = 4\int_{-\infty}^{0} w\phi(w - \lambda)dw < 0. \tag{13.29}$$

and by subtraction,

$$r_\lambda(\lambda, \mu) - r_\lambda(\lambda, 0) = 2\int_{-|\mu|}^{|\mu|} |w|\phi(w - \lambda)dw. \tag{13.30}$$

After normalizing by $|r_\lambda(\lambda, 0)|$, the threshold risk derivative turns out to be monotone in λ; a result reminiscent of the monotone likelihood ratio property. The proof is given at the end of the chapter.

Lemma 13.10 *For $\mu \neq 0$, the ratio*

$$V(\lambda, \mu) = \frac{r_\lambda(\lambda, \mu)}{|r_\lambda(\lambda, 0)|} \tag{13.31}$$

is strictly increasing in $\lambda \in [0, \infty)$, with $V(0, \mu) < 0$ and $V(\lambda, \mu) \nearrow \infty$ as $\lambda \to \infty$.

Integrated threshold risk. Define $B(\lambda, \pi)$ as above. Since $\mu \to r(\lambda, \mu)$ is a bounded (by $1 + \lambda^2$, from (8.5), (8.6)) analytic function, $B(\lambda, \pi)$ is well defined and differentiable, with

$$(\partial/\partial\lambda)B(\lambda, \pi) = \int r_\lambda(\lambda, \mu)\pi(d\mu). \tag{13.32}$$

Now it can be shown that given π, there is always a unique best, i.e. risk minimizing, choice of threshold.

Proposition 13.11 *If $\pi = \delta_0$, then $\lambda \to B(\lambda, \pi)$ decreases to 0 as $\lambda \to \infty$. If $\pi \neq \delta_0$, then the function $\lambda \to B(\lambda, \pi)$ has a unique minimum $\lambda(\pi)$, $0 < \lambda(\pi) < \infty$, and is strictly decreasing for $\lambda < \lambda(\pi)$ and strictly increasing for $\lambda > \lambda(\pi)$.*

Proof First, $B(\lambda, \delta_0) = r(\lambda, 0)$ is strictly decreasing in λ by (13.29), and that it converges to 0 for large λ is clear from the risk function itself.

For $\pi \neq \delta_0$, it is convenient to normalize by $|r_\lambda(\lambda, 0)|$, and so to use (13.31) and (13.32) to define

$$W(\lambda) = \frac{(\partial/\partial\lambda)B(\lambda, \pi)}{|r_\lambda(\lambda, 0)|} = \int V(\lambda, \mu)\pi(d\mu).$$

From (13.28), it is clear that $W(0) < 0$, while Lemma 13.10 shows that $W(\lambda) \nearrow \infty$ as $\lambda \to \infty$. Hence there exists a zero, $W(\lambda_0) = 0$. Now for any λ

$$W(\lambda) - W(\lambda_0) = \int [V(\lambda, \mu) - V(\lambda_0, \mu)]\pi(d\mu),$$

and so strict monotonicity of $\lambda \to V(\lambda, \mu)$ for $\mu \neq 0$ guarantees that this difference is < 0 or > 0 according as $\lambda < \lambda_0$ or $\lambda > \lambda_0$. Consequently $(\partial/\partial\lambda)B(\lambda, \pi)$ has a single sign change from negative to positive, $\lambda(\pi) = \lambda_0$ is unique and the Proposition follows. □

The best threshold provided by the last proposition has a directional continuity property that will be needed for the minimax theorem below. (For proof, see Further Details).

Lemma 13.12 *If $\pi_0 \neq \delta_0$ and π_1 are probability measures and $\pi_t = (1-t)\pi_0 + t\pi_1$, then $\lambda(\pi_t) \to \lambda(\pi_0)$ as $t \searrow 0$.*

A minimax theorem for thresholding. Just as in the full non-linear case, it is useful to think in terms of least favorable distributions for thresholding. Since the risk function $r(\lambda, \mu)$ is bounded and continuous in μ, the integrated threshold risk $B(\lambda, \pi)$ is linear and weakly continuous in π. Hence

$$B_S(\pi) = \inf_\lambda B(\lambda, \pi)$$

is concave and upper semicontinuous in π. Hence it attains its supremum on the weakly compact set $\mathfrak{m}_p(\tau)$, at a least favorable distribution π_0, say. Necessarily $\pi_0 \neq \delta_0$, as $B_S(\delta_0) = 0$. Let $\lambda_0 = \lambda(\pi_0)$ be the best threshold for π_0, provided by Proposition 13.11.

The payoff function $B(\lambda, \pi)$ is *not* convex in λ, as is shown by consideration of, for example, the risk function $\lambda \to r_S(\lambda, 0)$ corresponding to $\pi = \delta_0$. On the other hand, $B(\lambda, \pi)$ is still linear in π, and this makes it possible to establish the following minimax theorem directly.

Theorem 13.13 *The pair (λ_0, π_0) is a saddlepoint: for all $\lambda \in [0, \infty)$ and $\pi \in \mathfrak{m}_p(\tau)$,*

$$B(\lambda_0, \pi) \leq B(\lambda_0, \pi_0) \leq B(\lambda, \pi_0), \tag{13.33}$$

and hence

$$\inf_\lambda \sup_{\mathfrak{m}_p(\tau)} B(\lambda, \pi) = \sup_{\mathfrak{m}_p(\tau)} \inf_\lambda B(\lambda, \pi)$$

and

$$\beta_{S,p}(\tau, \epsilon) = \sup\{B_S(\pi) : \pi \in \mathfrak{m}_p(\tau)\}. \tag{13.34}$$

Proof This is given as Theorem A.7, in which we take $\mathcal{P} = \mathfrak{m}_p(\tau)$. The hypotheses on $B(\lambda, \pi)$ are satisfied by virtue of Lemma 13.12 and Proposition 13.11. □

With minimax threshold theorem in hand, we turn to understanding the threshold minimax risk $\beta_{S,p}(\tau, \epsilon)$ defined at (13.27).

Proposition 13.14 *The minimax Bayes threshold risk $\beta_{S,p}(\tau, \epsilon)$ also satisfies the properties (1) - (4) of $\beta_p(\tau, \epsilon)$ enumerated in Proposition 13.4.*

Proof The minimax Theorem 13.13 gives, in (13.34), a representation for $\beta_{S,p}(\tau, \epsilon)$ analogous to (13.14) for $\beta_p(\tau, \epsilon)$, and so we may just mimic the proof of Proposition 13.4. except in the case of monotonicity in ϵ, for which we refer to Exercise 13.6. □

We have arrived at the destination for this section, a result showing that, regardless of the moment constraint, there is a threshold rule that comes quite close to the best non-linear minimax rule. It is an analog, for soft thresholding, of the Ibragimov-Has'minskii bound Theorem 4.17.

Theorem 13.15 *(i) For $0 < p \le \infty$,*

$$\sup_{\tau,\epsilon} \frac{\beta_{S,p}(\tau, \epsilon)}{\beta_p(\tau, \epsilon)} = \Lambda(p) < \infty.$$

(ii) For $p \ge 2$, $\Lambda(p) \le 2.22$.

Unpublished 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

$$\mu_p(\tau) = \beta_{S,p}(\tau, 1)/\beta_p(\tau, 1) \to 1 \quad \text{as} \quad \tau \to 0, \infty. \tag{13.35}$$

Proof Most of the ingredients are present in Theorem 13.7 and Proposition 13.14, and we assemble them in a fashion parallel to the proof of Theorem 4.17. The scaling $\beta_{S,p}(\tau, \epsilon) = \epsilon^2 \beta_{S,p}(\tau/\epsilon, 1)$ reduces the proof to the case $\epsilon = 1$. The continuity of both numerator and denominator in $\tau = \tau/\epsilon$ shows that it suffices to establish (13.35).

For small τ, we need only reexamine the proof of Theorem 13.7: the upper bounds for $\beta_p(\tau)$ given there are in fact provided by threshold estimators, with $\lambda = 0$ for $p \ge 2$ and $\lambda = \sqrt{2 \log \tau^{-p}}$ for $p < 2$.

For large τ, use the trivial bound $\beta_{S,p}(\tau, 1) \le 1$, along with the property (1) that $\beta_p(\tau)$ is decreasing in p to write

$$\mu_p(\tau) \le 1/\beta_\infty(\tau) = 1/\rho_N(\tau, 1) \tag{13.36}$$

which decreases to 1 as $\tau \to \infty$, by (4.36) - (4.37). This completes the proof of (i).

For part (ii), use (13.36) to conclude for any p and for $\tau \ge 1$, that $\mu_p(\tau) \le 1/\rho_N(1, 1) \doteq 2.22$. For $\tau \le 1$ and now using $p \ge 2$, we use $\beta_{S,p}(\tau) \le \tau^2$ (compare (13.20)) to write $\mu_p(\tau) \le \tau^2/\beta_\infty(\tau) = \tau^2/\rho_N(\tau, 1)$. The final part of the proof of Theorem 4.17 showed that the right side is bounded above by $1/\rho_N(1, 1) \doteq 2.22$. □

13.5 Minimax Bayes Risk for $n-$dimensional data.

We return to estimation of a $n-$dimensional parameter constrained to an ℓ_p ball and observed in white Gaussian noise–compare model (13.1) and (13.2). Our interest now is in the 'dense' cases in which $\eta_n = n^{-1/p}(C_n/\epsilon_n) \to \eta > 0$ and/or $p \geq 2$. The asymptotics of $R_N(\Theta_{n,p}(C_n))$ will be evaluated by the Bayes minimax approach of Section 4.11. This approach allows reduction to the basic one dimensional Bayes minimax problem studied in the previous section. We choose a collection of priors on $\mathbb{R}^n$

$$\mathcal{M}_n = \{\pi(d\theta) : E_\pi \sum_1^n |\theta_i|^p \leq C_n^p\}. \tag{13.37}$$

which relaxes the ℓ_p-ball constraint of $\Theta_n = \Theta_{n,p}(C_n)$ to an in-mean constraint. The set $\mathcal{M}_n$ contains all point masses δ_θ for $\theta \in \Theta_n$, and is convex, so using (4.18), the minimax risk is bounded above by the Bayes minimax risk

$$R_N(\Theta_{n,p}(C_n)) \leq B(\mathcal{M}_n) = \sup\{B(\pi), \pi \in \mathcal{M}_n\} := B_{n,p}(C_n, \epsilon_n).$$

We first show that that this upper bound is easy to evaluate in terms of a univariate quantity, and later investigate when the bound is asymptotically sharp.

Proposition 13.16 *Let $\beta_p(\eta)$ denote the univariate Bayes minimax risk (13.14) for unit noise, and let $\eta_n = n^{-1/p}C_n/\epsilon_n$ be the dimension normalized radius. Then*

$$B_{n,p}(C_n, \epsilon_n) = n\epsilon_n^2 \beta_p(\eta_n). \tag{13.38}$$

This is the p-th moment analog of the identity (8.69) for the ℓ_0 case. The proofs differ a little since the method used for p-th moments does not preserve the ℓ_0 parameter space.

Proof We use the 'independence trick' of Section 4.5 to show that the maximisation in $B(\mathcal{M}_n)$ can be reduced to univariate priors. Indeed, for any $\pi \in \mathcal{M}_n$, construct a prior $\bar{\pi}$ from the product of the univariate marginals π_i of π. We have the chain of relations

$$B(\pi) \leq B(\bar{\pi}) = \sum_i B(\pi_i) \leq nB(\tilde{\pi}_1).$$

Indeed, Lemma 4.15 says that $\bar{\pi}$ is harder than π, yielding the first inequality. Bayes risk is additive for an independence prior: this gives the equality. For the second inequality, form the average $\tilde{\pi}_1 = n^{-1}\sum_i \pi_i$ and appeal to the concavity of Bayes risk.

The p-th moment of the univariate prior $\tilde{\pi}_1$ is easily bounded:

$$\int |\theta|^p d\tilde{\pi}_1 = n^{-1} \sum_1^n E_{\pi_i}|\theta_i|^p \leq n^{-1}C_n^p,$$

because $\pi \in \mathcal{M}_n$, and so we can achieve the maximization of $B(\mathcal{M}_n)$ by restricting to univariate priors in $\mathfrak{m}_p(\tau)$ with $\tau = n^{-1/p}C_n$. In other words,

$$B_{n,p}(C_n, \epsilon_n) = n\beta_p(n^{-1/p}C_n, \epsilon_n)$$

and now the Proposition follows from the invariance relation 4(i) of Proposition 13.4. □

EXAMPLE 13.2 continued. Let us return to our original example in which $p = 1$, the noise $\epsilon_n = n^{-1/2}$, and the radius $C_n = 1$. Thus $\eta_n = n^{-1} \cdot n^{1/2} = n^{-1/2}$. It follows that

$$R_N(\Theta_{n,1}) \le B_{n,1}(C_n, \epsilon_n) = n \cdot (1/n) \cdot \beta_1(n^{-1/2}) \sim (\log n/n)^{1/2}, \tag{13.39}$$

where the last equivalence uses (13.19). The next theorem will show that this rate and constant are optimal. Recall, for comparison, that $R_L(\Theta_{n,1}, \epsilon_n) = 1/2$.

The main result of this chapter describes the asymptotic behavior of the nonlinear minimax risk $R_N(\Theta)$, and circumstances in which it is asymptotically equivalent to the Bayes minimax risk. In particular, except in the highly sparse settings to be discussed below, 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 $\mathfrak{m}_p(\eta_n)$. We can thus build on the small η results from the previous section.

Theorem 13.17 *Let $R_N(C_n, \epsilon_n)$ denote the minimax risk* (13.3) *for estimation over the ℓ_p ball $\Theta_{n,p}(C_n)$ defined at* (13.2)*, and η_n the normalized signal-to-noise ratio* (13.4)*.*

For $2 \le p \le \infty$, if $\eta_n \to \eta \in [0, \infty]$, then

$$R_N(C_n, \epsilon_n) \sim n\epsilon_n^2 \beta_p(\eta_n). \tag{13.40}$$

For $0 < p < 2$, define threshold t_n by (13.7)*, then κ_n by* (13.8) *and $R(t)$ by* (13.10)*. Finally, let $\check{\kappa}_n = \kappa_n \vee 1$.*

(a) if $\eta_n \to \eta \in (0, \infty]$, then again (13.40) *holds.*

(b) If $\eta_n \to 0$ and $\kappa_n \to \kappa \in [0, \infty]$, then

$$R_N(C_n, \epsilon_n) \sim R(\kappa_n)\epsilon_n^2 \cdot 2\log(n/\check{\kappa}_n). \tag{13.41}$$

(c) If $\eta_n \to 0$ and $\kappa_n \to \infty$, then (13.40) *and* (13.41) *agree.*

The table below displays the nature of the nearly least favorable prior in the various cases covered by the Theorem. We use the word 'dense' when the number of non-zero components N_n in the prior is of exact order n: $EN_n \approx n$. The sparse case, in which EN_n is $o(n)$, occurs only when $p < 2$, and is further subdivided according to whether $EN_n = \kappa_n$ remains finite (the 'highly sparse' case) or not.

	$p \ge 2$	$p < 2$
$\eta_n \to \eta > 0$	dense	dense
$\eta_n \to 0$	dense	$\kappa_n \to \infty$, sparse $\kappa_n \equiv \kappa$, highly sparse

The highly sparse case is noteworthy because, as discussed below, the minimax Bayes approach fails. The practical importance of this case has been highlighted by Mallat in a satellite image deconvolution/denoising application.

Proof The sparse case, namely $0 < p < 2$ and $\eta_n \to 0$, assertion (b) in the theorem, has already been established in Theorem 13.3 and is included in the statement here for completeness. Our main task here is to establish the equivalence (13.40), which, in view of Proposition 13.16, amounts to proving asymptotic equivalence of frequentist and Bayes

minimax risks. The detailed behavior of 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 $\beta_p(\eta, 1)$ and will be described below.

Asymptotic equivalence of R_N and B. To show that the Bayes minimax bound in (13.38) is asymptotically sharp, we construct a series of asymptotically least favorable priors π_n that essentially concentrate on $\Theta_n = \Theta_{n,p}(C_n)$. More precisely, following the recipe of Chapter 4.11, for each $\gamma < 1$ we construct priors π_n satisfying

$$B(\pi_n) \geq \gamma B_{n,p}(\gamma C_n, \epsilon_n)(1 + o(1)), \tag{13.42}$$

$$\pi_n(\Theta_n) \to 1, \text{and} \tag{13.43}$$

$$\mathbb{E}_{\pi_n}\{\|\hat{\theta}_{\nu_n}\|^2 + \|\theta\|^2, \Theta_n^c\} = o(B_{n,p}(\gamma C_n, \epsilon_n)) \tag{13.44}$$

where $\hat{\theta}_{\nu_n}(y) = E_{\pi_n}(\theta|\theta \in \Theta_n, y)$.

In addition, we need the analog of (4.72), which here, using (13.38), becomes

$$\lim_{\gamma \nearrow 1} \lim_{n\to\infty} \frac{B_{n,p}(\gamma C_n, \epsilon_n)}{B_{n,p}(C_n, \epsilon_n)} = \lim_{\gamma \nearrow 1} \lim_{n\to\infty} \frac{\beta_p(\gamma\eta_n)}{\beta_p(\eta_n)} = 1. \tag{13.45}$$

As indicated at Lemma 4.32 and the following discussion, if we verify (13.42) - (13.45) we can conclude that $R_N(C_n, \epsilon_n) \sim B_{n,p}(C_n, \epsilon_n)$.

We will always define π_n by i.i.d rescaled draws from a univariate distribution $\pi_1(d\mu)$ on $\mathbb{R}$ (in some cases $\pi_1 = \pi_{1n}$ depends on n): thus $\pi_n(d\theta) = \pi_{1n}^n(d\theta/\epsilon_n)$. Equivalently, $\theta_i = \epsilon_n \mu_i$ with $\mu_i \sim \pi_{1n}$ drawn i.i.d. for $i = 1, \ldots, n$. Therefore, using (13.38), condition (13.42) can be reexpressed as

$$B(\pi_{1n}) \geq \gamma \beta_p(\gamma\eta_n)(1 + o(1)), \tag{13.46}$$

and property (13.43) may be rewritten as

$$\pi_n(\Theta_n) = P_{\pi_{1n}}\{n^{-1}\sum |\mu_i|^p \leq \eta_n^p\}.$$

We carry out the construction of π_{1n} and π_n in three cases. First, under the assumption that $\eta_n \to \eta \in (0, \infty]$ for all $p \in (0, \infty]$: this is the 'dense' case. Second, we suppose that $\eta_n \to 0$ and $p \geq 2$: this is in fact also a dense case since all components of the least favorable prior are non-zero. Finally, for completeness, we discuss in outline the sparse case $\eta_n \to 0$ with $0 < p < 2$: in this case the i.i.d. prior $\pi_n = \pi_{1n}^n$ establishes (13.41) only when $\kappa_n \to \infty$: this was the reason for using the independent blocks prior in Theorem 13.3.

1°. Suppose first that $\eta_n \to \eta \in (0, \infty]$. Given $\gamma < 1$, there exists $M < \infty$ and a prior π_1 in $\mathfrak{m}_p(\gamma\eta)$ supported on $[-M, M]$ whose Bayes risk satisfies $B(\pi_1) \geq \gamma\beta_p(\gamma\eta)$, compare Exercise 4.5. Property (13.46) follows because $\beta_p(\gamma\eta_n) \to \beta_p(\gamma\eta)$. Noting $E_{\pi_1}|\mu|^p \leq \gamma^p\eta^p$ and that $|\mu_i| \leq M$, property (13.43) follows from the law of large numbers applied to the i.i.d. draws from π_1. Since $|\mu_i| \leq M$ under the prior π_n, both $\|\theta\|^2$ and $\|\hat{\theta}_\nu\|^2$ are bounded by $n\epsilon^2M^2$, the latter because $\|\hat{\theta}_\nu\|^2 \leq E_{\pi_n}\{\|\theta\|^2|\theta \in \Theta_n, y\}$. Hence the left side of (13.44) is bounded by $2n\epsilon_n^2M^2\pi_n(\Theta_n^c)$ while $B_{n,p}(\gamma C_n, \epsilon_n)$ is of exact order $n\epsilon_n^2$, and so (13.44) follows from (13.43). Property (13.45) follows from continuity of β_p, Proposition 13.4.

In summary, $R_N \sim n\epsilon_n^2\beta_p(\eta_n)$ and an asymptotically minimax estimator can be built from the Bayes estimator for a least favorable prior for $\mathfrak{m}_p(\eta)$.

2°. Now suppose that $\eta_n \to 0$. First, observe from (13.19) that $\beta_p(\gamma\eta_n)/\beta_p(\eta_n) \to \gamma^{2\wedge p}$, so that (13.45) holds.

Suppose first that $p \geq 2$. This case is straightforward: we know from the univariate case that the symmetric two point priors $\pi_{1n} = (\delta_{\eta_n} + \delta_{-\eta_n})/2$ are asymptotically least favorable, Theorem 13.7, so π_{1n} satisfies (13.46) for large n. The corresponding measure π_n is already supported on Θ_n, so the remaining conditions are vacuous here.

In summary, $R_N \sim n\epsilon_n^2\eta_n^2$ and $\hat{\theta} = 0$ is asymptotically minimax.

3°. Now suppose that $0 < p < 2$. Although (13.41) was established in Theorem 13.3, we do need to check that formulas (13.40) and (13.41) are consistent. For this, note first that $\check{\kappa}_n = \kappa_n$, and $\kappa_n \to \infty$ implies from (13.8) that $n\eta_n^p \to \infty$ (argue by contradiction), and then also, cf. (13.7), that $t_n^2 = 2\log \eta_n^{-p}$. We have the following chain of relations

$$n\beta_p(\eta_n) \sim n\eta_n^p(2\log\eta_n^{-p})^{1-p/2} = \kappa_n \cdot 2\log\eta_n^{-p} = \kappa_n \cdot 2\log(n/(\kappa_n t_n^p)) \sim \kappa_n \cdot 2\log(n/\kappa_n).$$

The first equivalence uses (13.19), while the second equality is a rewriting of (13.8). The third again uses (13.8), now inside the logarithm, and the fourth applies (13.9) to show that the t_n^p factor is negligible. Finally, of course, $R(\kappa_n) \sim \kappa_n$. In summary, when $\eta_n \to 0$ and $\kappa_n \to \infty$,

$$R_N \sim n\epsilon_n^2\eta_n^p(2\log\eta_n^{-p})^{1-p/2} \tag{13.47}$$

and soft thresholding with $\lambda_n = (2\log\eta_n^{-p})^{1/2}\epsilon_n$ provides an asymptotically minimax estimator. Hard thresholding is also asymptotically minimax so long as the thresholds are chosen in accordance with (13.25). □

As promised, let us look at the i.i.d. prior construction in this sparse case. The univariate prior is chosen as follows. Given $\gamma < 1$, let π_{1n} be the sparse prior $\pi_p[\gamma\eta_n]$ of Definition 13.6 and set

$$\alpha_n = \alpha_p(\gamma\eta_n), \qquad \mu_n = \mu_p(\gamma\eta_n).$$

Thus the least favorable distribution corresponds to vectors θ_i in which most co-ordinates are zero and a small fraction α_n at random positions have magnitude about $\epsilon_n\sqrt{2\log\eta_n^{-p}}$.

From the proof of Theorem 13.7 and Lemma 8.11, we have

$$\beta_p(\gamma\eta_n) \sim B(\pi_{1n}) \sim \alpha_n\mu_n^2. \tag{13.48}$$

This establishes (13.46) and hence (13.42). Everything now turns on the support condition (13.43). Observe that the number N_n of non-zero components in a draw from $\pi_n = \pi_{1n}^n$ is a Binomial(n,α_n) variable, and that $\sum_i |\theta_i|^p = N_n\epsilon_n^p\mu_n^p$. The support requirement becomes

$$\{\theta \in \Theta_n\} = \{N_n \leq C_n^p/(\epsilon_n^p\mu_n^p)\}. \tag{13.49}$$

Rewriting the moment condition

$$\alpha_n\mu_n^p = (\gamma\eta_n)^p = \gamma^p n^{-1} C_n^p/\epsilon_n^p, \tag{13.50}$$

and noting that $EN_n = n\alpha_n$, we find that Chebychev's inequality leads to

$$\pi_n(\Theta^c) = P\{N_n > \gamma^{-p} n\alpha_n\} \leq c_{\gamma p}\mathrm{Var}\, N_n/(EN_n)^2. \tag{13.51}$$

The right side of (13.51) converges to zero exactly when $EN_n = n\alpha_n \to \infty$. We may

verify that $n\alpha_n \to \infty$ is equivalent to $\kappa_n \to \infty$. Indeed, insert (13.8) into the moment condition (13.50) to obtain $\kappa_n/(n\alpha_n) = \gamma^{-p}(\mu_n/t_n)^p$ so that our claim follows from (13.7) and (13.18) so long as $n\eta_n^p \geq 1$. If, instead, $n\eta_n^p \leq 1$ then it follows from (13.8) and (13.17) that both κ_n and $n\alpha_n$ remain bounded.

Thus condition (13.43) holds only on the assumption that $\kappa_n \to \infty$. In this case (13.44) can be verified with a little more work; the details are omitted. Observe that when $\kappa_n \to \kappa \in (1, \infty)$, the minimax Bayes risk approaches $\kappa\epsilon_n^2 \cdot 2\log(n/\kappa)$, whereas the actual minimax risk behaves like $R(\kappa)\epsilon_n^2 \cdot 2\log(n/\kappa)$. Thus Figure 13.1 shows the inefficiency of the minimax Bayes risk for non-integer values of κ.

The assumption that $\kappa_n \to \infty$ ensures that $EN_n \to \infty$. In other words, that Θ_n has large enough radius that the least favorable distribution in the Bayes minimax problem generates an asymptotically unbounded 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 $C_n = n^{-1/2}$, so that $\kappa_n \to 0$. 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*: $R_L \sim n\epsilon_n^2\bar{\eta}^2 \sim n^{-1}$, 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 $\mu_n \sim \sqrt{2\log n}$.

13.6 Near minimaxity of thresholding in $\mathbb{R}^n$.

Let $\hat{\theta}_\lambda(y)$ denote soft thresholding at $\lambda\epsilon$ for data from n-dimensional model (13.1):

$$\hat{\theta}_{\lambda,i} = \hat{\delta}_S(y_i, \epsilon\lambda). \tag{13.52}$$

The minimax risk among soft thresholding estimators over the ℓ_p-ball $\Theta_{n,p}(C)$ is given by

$$R_S(C, \epsilon) = R_S(\Theta_{n,p}(C), \epsilon) = \inf_\lambda \sup_{\theta \in \Theta_{n,p}(C)} E_\theta \|\hat{\theta}_\lambda - \theta\|^2.$$

The next result is a fairly straightforward consequence of Theorems 13.15 and 13.17.

Theorem 13.18 *Adopt the assumptions of Theorem 13.17. If $\eta_n \to \eta \in [0, \infty]$ and, when $p < 2$, if also $\kappa_n \to \infty$, then there exists $\Lambda(p) < \infty$ such that*

$$R_S(C_n, \epsilon_n) \leq \Lambda(p) R_N(C_n, \epsilon_n) \cdot (1 + o(1)). \tag{13.53}$$

If also $\eta_n \to 0$, then

$$R_S(C_n, \epsilon_n) \sim R_N(C_n, \epsilon_n).$$

The proof shows that $\Lambda(p)$ can be taken as the univariate quantity appearing in Theorem 13.15, as so from the remarks there, is likely to be not much larger than 1. Thus, in the high dimensional model (13.1), soft thresholding has bounded minimax efficiency among *all* estimators. In the case when $\eta_n \to 0$, the threshold choice $\lambda_n = \epsilon_n\sqrt{2\log \eta_n^{-p}}$ is asymptotically minimax among all estimators.

Proof For a given vector $\theta = (\theta_i)$, define $\mu_i = \theta_i/\epsilon_n$ and let π_n denote the empirical measure $n^{-1}\sum_i \delta_{\mu_i}$. We can then rewrite the risk of soft thresholding at $\lambda\epsilon_n$, using our earlier notations, respectively as

$$E\sum_i (\hat{\theta}_{\lambda,i} - \theta_i)^2 = \epsilon_n^2 \sum_i r(\lambda, \mu_i) = n\epsilon_n^2 B(\lambda, \pi_n).$$

If $\theta \in \Theta_{n,p}(C_n)$, then the empirical measure satisfies a univariate moment constraint

$$\int |\mu|^p d\pi_n = n^{-1} \sum |\theta_i/\epsilon_n|^p \le n^{-1}(C_n/\epsilon_n)^p = \eta_n^p. \tag{13.54}$$

Consequently $\pi_n \in \mathfrak{m}_p(\eta_n)$, and so

$$\inf_\lambda \sup_\theta E_\theta \|\hat\theta_\lambda - \theta\|^2 \le n\epsilon_n^2 \inf_\lambda \sup_{\pi \in \mathfrak{m}_p(\eta_n)} B(\lambda, \pi).$$

Now recalling definition (13.27) of $\beta_{S,p}(\eta)$ and then Theorem 13.15, the right side equals

$$n\epsilon_n^2 \beta_{S,p}(\eta_n) \le \Lambda(p)\, n\epsilon_n^2 \beta_p(\eta_n) = \Lambda(p) B_{n,p}(C_n, \epsilon_n),$$

where at the last equality we used the minimax Bayes structure Proposition 13.16. Putting this all together, we get

$$R_S(C_n, \epsilon_n) \le \Lambda(p) B_{n,p}(C_n, \epsilon_n)$$

and the conclusion (13.53) now follows directly from Theorem 13.17. If $\eta_n \to 0$, then $\beta_{S,p}(\eta_n) \sim \beta_p(\eta_n)$ by Theorem 13.7 and so we obtain the second statement. □

Remark 13.19 There is a fuller Bayes minimax theory for thresholding, which allows for a different choice of threshold in each co-ordinate. There is a notion of threshold Bayes minimax risk, $B_{S;n,p}(C, \epsilon)$ for priors satisfying (13.37), and a vector version of Theorem 13.15

$$B_{S;n,p}(C, \epsilon) \le \Lambda(p) B_{n,p}(C, \epsilon). \tag{13.55}$$

In this Bayes-minimax threshold theory, there is no advantage to allowing the thresholds to depend on the co-ordinate index: the minimax λ^* has all components the same. This provides some justification for the definition (13.52). Exercise 13.9 has details.

13.7 Appendix: Further details

3°. *Proof of Lemma 13.10.* That $V(0, \mu) < 0$ follows from (13.28). From (13.29) and (13.30), we have

$$V(\lambda, \mu) = \tfrac{1}{2} R(\lambda, \mu) - 1 \tag{13.56}$$

where, after writing ϕ_λ for $\phi(w - \lambda)$,

$$R(\lambda, \mu) = \int_N |w|\phi_\lambda \Big/ \int_D |w|\phi_\lambda = N(\lambda)/D(\lambda),$$

and the intervals $N = (-|\mu|, |\mu|)$ and $D = (-\infty, 0)$. One then checks that

$$\begin{aligned} D(\lambda)^2(\partial/\partial\lambda) R(\lambda, \mu) &= D(\lambda)N'(\lambda) - N(\lambda)D'(\lambda) \\ &= \int_D |w|\phi_\lambda \int_N w|w|\phi_\lambda - \int_D w|w|\phi_\lambda \int_N |w|\phi_\lambda, \end{aligned}$$

after cancellation, and each term on the right side is positive when $\mu \ne 0$ and $\lambda > 0$ since

$$\int_N w|w|\phi_\lambda = \int_0^{|\mu|} w^2[\phi(w - \lambda) - \phi(w + \lambda)]dw > 0,$$

from symmetry and unimodality of ϕ, and $\int_D w|w|\phi_\lambda < 0$ since $D = (-\infty, 0)$. This shows the monotonicity of $V(\lambda, \mu)$ in λ. We turn to the large λ limit: writing μ for $|\mu|$, a short calculation shows that as $\lambda \to \infty$

$$N(\lambda) \geq \int_0^\mu w\phi(w-\lambda)dw = \lambda[\tilde{\Phi}(\lambda-\mu) - \tilde{\Phi}(\lambda)] + \phi(\lambda) - \phi(\mu-\lambda) \sim \frac{\mu}{\lambda}\phi(\lambda-\mu)$$
$$D(\lambda) = -\lambda\tilde{\Phi}(\lambda) + \phi(\lambda) \sim \phi(\lambda)/\lambda^2,$$

so that $R(\lambda, \mu) \geq \lambda\mu e^{\lambda\mu - \mu^2/2}(1 + o(1)) \to \infty$ as $\lambda \to \infty$.

4°· *Proof of Lemma 13.12.* Let $D(\lambda, \pi) = \partial_\lambda B(\lambda, \pi)$; from Proposition 13.11 we know that $\lambda \to D(\lambda, \pi)$ has a single sign change from negative to positive at $\lambda(\pi)$. The linearity of $\pi \to D(\lambda, \pi)$ yields

$$D(\lambda, \pi_t) = D(\lambda, \pi_0) + tD(\lambda, \pi_1 - \pi_0) = D(\lambda) + tE(\lambda),$$

say. Given $\epsilon > 0$, a sufficient condition for $\lambda_t = \lambda(\pi_t)$ to satisfy $|\lambda_t - \lambda_0| < \epsilon$ is that

$$D(\lambda_0 + \epsilon) + tE(\lambda_0 + \epsilon) > 0, \quad \text{and} \quad D(\lambda) + tE(\lambda) < 0$$

for all $\lambda \leq \lambda_0 - \epsilon$. Since $D(\lambda_0 - \epsilon) < 0 < D(\lambda_0 + \epsilon)$ and $\lambda \to E(\lambda)$ is continuous and bounded on $[0, \lambda_0 + 1]$, the condition clearly holds for all $t > 0$ sufficiently small.

13.8 Notes

This chapter is based on Donoho and Johnstone (1994b), which considered the more general case of ℓ_q losses, where the non-linearity phenomena appear in case $q < p$. New to this presentation are the results for the 'highly sparse' case $\kappa_n \to \kappa \in [0, \infty)$ and the minimax theorem for thresholding, Theorem 13.13. The importance of the highly sparse case was emphasized by Mallat at the time of his work on deconvolution of blurred satellite images Kalifa et al. (2003); Kalifa and Mallat (2003). Zhang (2012a) also discusses the highly sparse case and gives a complete discussion for ℓ_q losses.

Least favorable distributions subject to moment constraints for the single normal mean with known variance were studied by Feldman (1991) and shown to be either normal or discrete.

Exercises

13.1 *(Spike heights for least favorable priors.)* In the setting of Theorem 13.3, let $\kappa_n = n\eta_n^p/t_n^p$ and m_n be the integer part of $n/([\kappa_n] + 1)$. Let $\tau_n = t_n - \log t_n$ and show that $\Delta_n = \sqrt{2\log m_n} - \tau_n \to \infty$, for example as follows:
(a) if $\kappa_n < 1$, show that $\Delta_n \geq \log t_n$,
(b) if $\kappa_n \geq 1$, show that $\sqrt{2\log m_n} \geq t_n - c/t_n$.

13.2 *(Sparse priors are well defined.)* Consider the sparse prior $\pi_{\alpha,\mu(\alpha)}$ specified by equation (8.50) with sparsity α and overshoot $a = (2\log\alpha^{-1})^\gamma$ for $0 < \gamma < 1/2$. Let $\eta > 0$ be small and consider the moment constraint equation $\alpha\mu(\alpha)^p = \eta^p$. Show that $m(\alpha) = \alpha\mu(\alpha)^p$ has $m(0+) = 0$ and is increasing for $\alpha > 0$ sufficiently small. Show also, for example numerically, that for some γ, $m(\alpha)$ ceases to be monotone for larger values of α.

13.3 *(Bounds for hard thresholding.)* Use Lemma 8.5 to establish (13.26) by considering in turn $\mu \in [0, \sqrt{5}]$, $\mu \in [\sqrt{5}, \lambda]$ and $\mu \geq \lambda$. Give an expression for $\lambda_0(p)$.

13.4 *(Minimaxity of thresholding for $p \geq 2$.)* In the setting of Theorem 13.7, show that $\hat{\delta}_{\sqrt{2\log \eta^{-p}}}$ is asymptotically minimax when $\eta \to 0$ for $p \geq 2$, for example by using (8.7) and (8.12).

13.5 *(Structure of the p-th moment least favorable distributions.)* Establish Proposition 13.5 by mimicking the proof of Proposition 8.19, allowing for the fact that $\mathfrak{m}_p(\tau)$ is weakly compact.
(a) Let $\nu_\tau(d\theta) = \tau^{-p}|\theta|^p \pi_\tau(d\theta)$ and use strict monotonicity of $\beta_p(\tau)$ to show that ν_τ has total mass 1.
(b) Let $r(\theta) = \tau^p|\theta|^{-p}[r(\hat{\theta}_\tau, \theta) - r(\hat{\theta}_\tau, 0)]$ and verify that for $\theta \neq 0$,

$$r(\theta) \leq \int r(\theta')\nu_\tau(d\theta').$$

(c) Complete the argument using Lemma 4.18 and Exercise 4.1.

13.6 *(Monotonicity of threshold minimax risk.)* Let $r(\lambda, \mu; \epsilon)$ denote the MSE of soft thresholding at λ when $x \sim N(\mu, \epsilon^2)$, and $r(\lambda, \mu) = r(\lambda, \mu; 1)$. Show that the proof of monotonicity of $\epsilon \to \beta_{S,p}(\tau, \epsilon)$ can be accomplished via the following steps:
(a) It suffices to show that if $\epsilon' < \epsilon$, then $r(\lambda\epsilon', \mu; \epsilon') \leq r(\lambda\epsilon, \mu; \epsilon)$ for all μ and λ.
(b) Writing $\mu' = \mu/\rho$, verify that if $\rho \geq 0$,

$$(d/d\rho)r(\rho\lambda, \mu; \rho) = 2\rho r(\lambda, \mu') - \mu r_\mu(\lambda, \mu') = 2\rho E_{\mu'}\{(\delta_\lambda(x) - \mu')^2; |x| \geq \lambda\}.$$

13.7 *(Motivation for minimax threshold value.)* Show that the value $\lambda_p(\eta)$ minimizing the right side of integrated risk bound (13.21) satisfies $\lambda_p(\eta) \sim \sqrt{2\log \eta^{-p}}$ as $\eta \to 0$.

13.8 *(Proof Outline for Proposition 13.8.)* (a) Let $\mu = \kappa^{1/p}$ and $\Phi_\mu(I_\lambda) = \int_{-\lambda}^{\lambda} \phi(x - \mu)dx$ and show that

$$p^2\kappa^2 D_\kappa^2 r(\lambda, \kappa^{1/p}) = 2\mu^3 \Phi_\mu(I_\lambda)\{(2-p)\mu^{-1} + D_\mu \log \Phi_\mu(I_\lambda)\}.$$

(b) For $0 < p < 2$, there exists $\kappa_c > 0$ such that the function $\kappa \to r(\lambda, \kappa^{1/p})$, is convex for $\kappa \in (0, \kappa_c]$ and concave for $\kappa \in [\kappa_c, \infty)$. [Assume, from e.g. Prékopa (1980, Theorem 3 and Sec. 3), that $\mu \to \Phi_\mu(I_\lambda)$ is log-concave on $(0, \infty)$.]
(c) Show that the extreme points of $\mathfrak{m}_p(\tau)$ have the form $(1-\alpha)\delta_{\mu_0} + \alpha\delta_{\mu_1}$, but that it suffices to take $\mu_0 = 0$, and hence recover (13.22).
(d) Use equation (13.23) to establish (13.24).

13.9 *(Bayes minimax theory for thresholding.)* Let $\lambda = (\lambda_i)$ be a vector of thresholds, and define now $\hat{\theta}_\lambda$ by $\hat{\theta}_{\lambda,i}(y) = \hat{\delta}_S(y_i, \lambda_i\epsilon)$. If π is a prior on $\theta \in \mathbb{R}^n$, set $B(\lambda, \pi) = E_\pi E_\theta \|\hat{\theta}_\lambda - \theta\|^2$. Define $\mathcal{M}_n$, the priors satisfying the $\Theta_{n,p}(C)$ constraint in mean, by (13.37) and then define the Bayes-minimax threshold risk by

$$B_{S;n,p}(C, \epsilon) = \inf_\lambda \sup_{\pi \in \mathcal{M}_n} B(\lambda, \pi).$$

(a) Let $B_S(\pi) = \inf_\lambda B(\lambda, \pi)$. Show that $B_S(\pi)$ attains a maximum for some $\pi^* \in \mathcal{M}_n$, and that such a π^* may be chosen to have i.i.d co-ordinates.
(b) Show that a minimax theorem holds:

$$\inf_\lambda \sup_{\mathcal{M}_n} B(\lambda, \pi) = \sup_{\mathcal{M}_n} B_S(\pi),$$

that a saddlepoint (λ^*, π^*) exists, and that the components λ_i^* of the minimax threshold do not depend on i.
(c) Conclude that the vector bound (13.55) holds.

14

Sharp minimax estimation on Besov spaces

14.1 Introduction

In previous chapters, we developed bounds for the behavior of minimax risk $R_N(\Theta(C), \epsilon)$ over Besov bodies $\Theta(C)$. In Chapters 9 and 10, we showed that thresholding at $\sqrt{2 \log \epsilon^{-1}}$ led to asymptotic minimaxity up to logarithmic factors $O(\log \epsilon^{-1})$, while in Chapter 12 we established that estimators derived from complexity penalties achieved asymptotic minimaxity up to constant factors.

In this chapter, we use the minimax Bayes method to study the *exact* asymptotic behavior of the minimax risk, at least in the case of squared error loss. The "price" for these sharper optimality results is that the resulting optimal estimators are less explicitly described and depend on the parameters of Θ.

In outline, we proceed as follows. Section 14.2 replaces the minimax risk $R_N(\Theta(C), \epsilon)$ by an upper bound, the minimax Bayes problem with value $B(C, \epsilon)$, and states the main results of this chapter.

In Section 14.3, we begin study of the optimization over prior probability measures required for $B(C, \epsilon)$, and show that the least favorable distribution necessarily has independent co-ordinates, and hence the corresponding minimax rule is separable, i.e. acts coordinatewise. The $B(C, \epsilon)$ optimization is then expressed in terms of the univariate Bayes minimax risks $\beta_p(\tau, \epsilon)$ studied in Chapter 13.

In Section 14.4, a type of 'renormalization' argument is used to deduce the dependence of $B(C, \epsilon)$ on C and ϵ up to a periodic function of C/ϵ. At least in some cases, this function is almost constant.

In Section 14.5, we show that the upper bound $B(C, \epsilon)$ and minimax risk $R_N(\Theta(C), \epsilon)$ are in fact asymptotically equivalent as $\epsilon \to 0$, by showing that the asymptotically least favorable priors are asymptotically concentrated on $\Theta(C)$.

The minimax risk of *linear* estimators is evaluated in Section 14.6, using notions of quadratic convex hull from Chapter 4—revealing suboptimal rates of convergence when $p < 2$.

In contrast, *threshold* estimators, Section 14.7, can be found that come within a constant factor of $R_N(\Theta(C), \epsilon)$ over the full range of p; these results rely on the univariate Bayes minimax properties of thresholding established in Chapter 13.4.

14.2 Dyadic Sequence Model and Bayes minimax problem

Consider the Gaussian sequence model (9.47) with countable index set, in the dyadic indexing regime

$$y_I = \theta_I + \epsilon z_I \tag{14.1}$$

where I denotes the pair (j,k), supposed to lie in the set $\mathcal{I} = \cup_{j\geq -1}\mathcal{I}_j$, where for $j \geq 0$, $\mathcal{I}_j = \{(j,k) : k = 1,\dots,2^j\}$ and the exceptional $\mathcal{I}_{-1} = \{(-1,0)\}$. For the parameter spaces, we restrict attention, for simplicity of exposition, to a particular class of Besov bodies

$$\Theta = \Theta^\alpha_{p\cdot}(C) = \{\theta = (\theta_I) : \|\theta_{j\cdot}\|_p \leq C2^{-aj} \text{ for all} j\}, \qquad a = \alpha + 1/2 - 1/p.$$

This is the $q = \infty$ case of the Besov bodies $\Theta^\alpha_{p,q}$ considered in earlier chapters. [1] They are supersets of the cases with $q < \infty$, and it turns out that the *rate* of convergence as $\epsilon \to 0$ is the same for all q (Donoho and Johnstone, 1998).

We note that Θ is solid and orthosymmetric, and compact when $\alpha > (1/p - 1/2)_+$, Exercise 14.1(a). The focus will be on global ℓ_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

$$R_N(\Theta,\epsilon) = \inf_{\hat\theta}\sup_{\Theta} E_\theta\|\hat\theta - \theta\|_2^2.$$

In principle, a similar development could be carried out for the ℓ_p loss $\|\hat\theta - \theta\|_p^p = \sum|\hat\theta_I - \theta_I|^p$, or weighted losses of the form $\sum_j 2^{jr}\sum_k|\hat\theta_{jk} - \theta_{jk}|^p$.

The 'hard' constraint that $\|\theta\|_{b^\alpha_{p\infty}} \leq C$ is relaxed to a constraint 'in mean' with respect to a prior π. Define a class of priors

$$\mathcal{M} = \mathcal{M}^\alpha_{p\cdot}(C) = \{\pi(d\theta) : E_\pi\sum_k|\theta_{jk}|^p \leq C^p 2^{-ajp} \text{ for all } j\}.$$

As in earlier chapters, define the integrated risk $B(\hat\theta,\pi) = E_\pi E_\theta\|\hat\theta - \theta\|^2$ and the Bayes minimax risk

$$B(\mathcal{M},\epsilon) = \inf_{\hat\theta}\sup_{\pi\in\mathcal{M}} B(\hat\theta,\pi). \tag{14.2}$$

Since $\mathcal{M}$ contains unit point masses at each $\theta \in \Theta$, we have $R_N(\Theta,\epsilon) \leq B(\mathcal{M},\epsilon)$. We will again see that it is (relatively) easier to study and evaluate the Bayes minimax risk $B(\mathcal{M},\epsilon)$. To emphasize the dependence on C and ϵ, we sometimes write $B(C,\epsilon)$ for $B(\mathcal{M},\epsilon)$.

The results build on the univariate Bayes minimax problem introduced in Section 13.3, with Bayes minimax risk $\beta_p(\tau,\epsilon)$ corresponding to scalar observation $y = \theta + \epsilon z$ and moment constraint $E_\pi|\theta|^p \leq \tau^p$ for the prior π. We use the notation $\beta_p(\eta)$ for the normalized problem with noise $\epsilon = 1$. Let π_η denote the least favorable prior for $\beta_p(\eta)$ and $\delta_\eta = \hat\delta(x;\eta)$ denote the corresponding Bayes-minimax estimator, so that $B(\delta_\eta,\pi_\eta) = B(\pi_\eta) = \beta_p(\eta)$.

A first key property of the Bayes minimax problem is that minimax estimators are *separable* into functions of each individual coordinate:

[1] The more obvious notation $\Theta^\alpha_p(C)$ would clash with our earlier use of $\Theta^\alpha_2(C)$ for Fourier ellipsoids.

Theorem 14.1 *Suppose that* $0 < p \le \infty$ *and* $\alpha > (1/p - 1/2)_+$. *A minimax estimator for* $B(\mathcal{M}, \epsilon)$ *has the form*

$$\hat{\theta}_I^*(y) = \hat{\delta}_j^*(y_I), \qquad I \in \mathcal{I}, \tag{14.3}$$

where $\hat{\delta}_j^*(y)$ *is a scalar non-linear function of the scalar* y. *In fact there is a one parameter family of functions from which the minimax estimator is built: Let* $\hat{\delta}(x;\eta)$ *be the Bayes minimax estimator for the univariate Bayes minimax problem* $\beta_p(\eta)$ *recalled above. Then*

$$\hat{\delta}_j^*(y_I) = \epsilon\hat{\delta}(y_I/\epsilon; \eta_j), \tag{14.4}$$

where $\eta_j = (C/\epsilon)2^{-(\alpha+1/2)j}$.

For $p \neq 2$, the explicit form of $\hat{\delta}(\cdot;\eta)$ is not available, but we will see that useful approximations of $\hat{\delta}(\cdot;\eta)$ by threshold rules are possible.

Second, the exact asymptotic structure of the Bayes minimax risk can be determined.

Theorem 14.2 *Suppose that* $0 < p \le \infty$ *and* $\alpha > (1/p - 1/2)_+$. *Then* $B(C,\epsilon) < \infty$ *and*

$$B(C,\epsilon) \sim P(C/\epsilon) \cdot C^{2(1-r)}\epsilon^{2r}, \qquad \epsilon \to 0,$$

where $r = 2\alpha/(2\alpha+1)$ *and* $P(\cdot) = P(\cdot; \alpha + 1/2, p)$ *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 establish asymptotic equivalence of frequentist and minimax Bayes risk.

Theorem 14.3 *Suppose that* $0 < p \le \infty$ *and* $\alpha > (1/p - 1/2)_+$. *Then*

$$R_N(\Theta,\epsilon) = B(C,\epsilon)(1 + o(1)), \qquad \epsilon \to 0. \tag{14.5}$$

Combining Theorems 14.2–14.3, we conclude that the estimator $\hat{\theta}^*$ is *asymptotically* minimax for R as $\epsilon \to 0$. In short: *a separable nonlinear rule is asymptotically minimax*.

The proofs of these results occupy the next three sections.

14.3 Separable rules

We begin the proof of Theorem 14.1 by noting that $\mathcal{M}$ is convex—this follows immediately from the linearity in π of the expectation constraints. This allows use of the minimax theorem Theorem 4.12 to write that $B(\mathcal{M}) = \sup_{\mathcal{M}} B(\pi)$, so that we may look for a least favorable prior. The optimization is simplified by noting that $\mathcal{M}$ is closed under the operation of replacing π by the levelwise average of marginals. Given a prior $\pi \in \mathcal{M}$, form the univariate marginals π_{jk} and then levelwise averages $\bar{\pi}_j = \mathrm{ave}_k(\pi_{jk})$. Form a new prior $\bar{\pi}$ by making θ_{jk} independent, with $\theta_{jk} \sim \bar{\pi}_j$. By construction

$$\mathrm{ave}_k E_{\bar{\pi}}|\theta_I|^p = \mathrm{ave}_k E_{\pi}|\theta_I|^p,$$

so that $\bar{\pi} \in \mathcal{M}$. As we showed in earlier chapters, e.g. in the proofs of Lemma 4.15 and Proposition 13.16, the prior $\bar{\pi}$ is more difficult for Bayes estimation, so $B(\bar{\pi}) \ge B(\pi)$. Thus it suffices to maximise over priors $\bar{\pi} \in \mathcal{M}$.

The independence structure of $\bar{\pi}$ means that the Bayes estimator $\hat{\theta}_{\bar{\pi}}$ is separable - since prior and likelihood factor, so does the posterior, and so

$$\hat{\theta}_{\bar{\pi},I} = E_{\bar{\pi}_I}(\theta_I | y_I).$$

In addition, the Bayes risk is additive: $B(\bar{\pi}) = \sum_I B(\bar{\pi}_I)$. The constraint for membership in $\mathcal{M}$ becomes, for $\bar{\pi}_j$,

$$E_{\bar{\pi}_j} |\theta_{j1}|^p \le C^p 2^{-(ap+1)j} \qquad \text{for all } j.$$

Let $\omega = \alpha + 1/2$ and note that $ap + 1 = \omega p$. The optimization can now be carried out on each level separately, and, since $\bar{\pi}_j$ is a univariate prior, expressed in terms of the univariate Bayes minimax risk, so that

$$\begin{aligned} B(C,\epsilon) = \sup_{\pi \in \mathcal{M}} B(\pi) &= \sup\{ \sum_{j \ge -1} 2^j B(\bar{\pi}_j) : E_{\bar{\pi}_j} |\theta_{j1}|^p \le C^p 2^{-\omega p j} \} \\ &= \sum_{j \ge -1} 2^j \beta_p(C 2^{-\omega j}, \epsilon). \end{aligned} \tag{14.6}$$

In each case the sum is over $j \ge 1$. Using the scale invariance of $\beta_p(\tau, \epsilon)$, Proposition 13.14, and introducing a parameter ζ through $2^{\omega\zeta} = C/\epsilon$, we have

$$B(C,\epsilon) = \epsilon^2 \sum_{j \ge 0} 2^j \beta_p(2^{-\omega(j-\zeta)}). \tag{14.7}$$

Hence the Bayes-minimax rule must be separable. Recalling the structure of minimax rules for $\beta_p(\eta)$ from Section 13.3, we have

$$\theta_I^*(y) = \epsilon \hat{\delta}(y_I/\epsilon, \eta_j) \qquad \eta_j = (C/\epsilon) 2^{-\omega j}.$$

This completes the proof of Theorem 14.1.

14.4 Exact Bayes minimax asymptotics.

To start the proof of Theorem 14.2, we observe that, since $\beta_p(\eta) \le 1$, we can extend the sum in (14.6) to all $j \in \mathbb{Z}$ at cost of at most ϵ^2:

$$Q(C,\epsilon) = \sum_{j \in \mathbb{Z}} 2^j \beta_p(C 2^{-\omega j}, \epsilon) = B(C,\epsilon) + O(\epsilon^2). \tag{14.8}$$

Since a discrepancy of order ϵ^2 is negligible in non-parametric problems as $\epsilon \to 0$, we may safely study $Q(C,\epsilon)$. Note that $Q(C,\epsilon)$ satisfies the invariances, for $\epsilon > 0, h \in \mathbb{Z}$,

$$Q(C,\epsilon) = \epsilon^2 Q(C/\epsilon, 1), \qquad Q(C 2^{\omega h}, \epsilon) = 2^h Q(C,\epsilon). \tag{14.9}$$

As in (14.7), put $2^{\omega\zeta} = C/\epsilon$. Writing $2^j = 2^{j-\zeta} \cdot 2^\zeta$, we have

$$Q(C,\epsilon) = \epsilon^2 \sum_{j \in \mathbb{Z}} 2^j \beta_p(2^{-\omega(j-\zeta)}) = \epsilon^2 2^\zeta P^\circ(\zeta),$$

where $P^\circ(\zeta)$ is the $1-$periodic function

$$P^\circ(\zeta) = \sum_j 2^{j-\zeta} \beta_p(2^{-\omega(j-\zeta)}) = \sum_{v \in \mathbb{Z}-\zeta} 2^v \beta_p(2^{-\omega v}).$$

Since $2^\zeta = (C/\epsilon)^{1/\omega}$ with $1/\omega = 2/(2\alpha+1) = 2(1-r)$, we get

$$\epsilon^2 2^\zeta = C^{2(1-r)}\epsilon^{2r},$$

yielding the formula in the display in Theorem 14.2, with

$$P(C/\epsilon) = P^\circ(\omega^{-1}\log_2(C/\epsilon)). \tag{14.10}$$

To check convergence of the sum defining $P(\zeta)$, observe that for large negative v, we have $F(v) = 2^v\beta_p(2^{-\omega v}) \asymp 2^v$, while for large positive v, referring to (13.19),

$$F(v) \asymp \begin{cases} 2^v \cdot 2^{-2\omega v} & \text{with} \quad 2\omega - 1 = 2\alpha > 0 \qquad \text{if} \ \ p \geq 2 \\ 2^v \cdot 2^{-p\omega v} v^{1-p/2} & \text{with} \quad p\omega - 1 = p(\alpha+1/2) - 1 > 0 \ \ \text{if} \ \ p < 2. \end{cases}$$

Continuity of $P(\zeta)$ follows from this convergence and the continuity of $\beta_p(\eta)$. This completes the proof of Theorem 14.2.

Remark. How does the location j of the maximum term in $Q(C,\epsilon)$ depend on ϵ? Suppose that v_* is the location of the maximum of the function $v \to 2^v\beta_p(2^{-\omega v})$. Then the maximum in $Q(C,\epsilon)$ occurs at $j_* = v_* + \zeta = v_* + \omega^{-1}\log_2(C/\epsilon)$. Using the calibration $\epsilon = n^{-1/2}$ and $\omega = \alpha + 1/2$, we can interpret this in terms of equivalent sample sizes as

$$j_* = \frac{\log_2 n}{1+2\alpha} + \frac{\log_2 C}{\alpha+1/2} + v_*. \tag{14.11}$$

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 (14.8) converges, the primary contributions to the risk $B(C,\epsilon)$ come from levels below the finest (with $\log_2 n$ corresponding to a sample of size n).

Example. When $p = 2$, explicit solutions are possible because $\beta_2(\eta) = \eta^2/(1+\eta^2)$ and $\hat\delta(x;\eta,2) = wx = [\eta^2/(1+\eta^2)]x$. Recall that $\eta_j = (C/\epsilon)2^{-\omega j} = 2^{-\omega(j-\zeta)}$ decreases rapidly with j above $\zeta = \omega^{-1}\log_2(C/\epsilon)$, so that $\hat\delta_j$ is essentially 0 for such j.

We have $P^\circ(\zeta) = \sum_j g(j-\zeta)$ for

$$g(v) = \frac{2^v}{1+2^{2\omega v}} = \frac{e^{av}}{1+e^{bv}}$$

for $a = \log 2$ and $b = (2\alpha+1)\log 2 > a$. An easy calculation shows that the maximum of g occurs at $v_* = \log_2(1/(2\alpha))/(1+2\alpha)$, compare also (14.11).

Figure 14.1 shows plots of the periodic function $P^\circ(\zeta)$ for several values of α. For small α, the function P° is very close to constant, while for larger α it is close to a single sinusoidal cycle. This may be understood from the Poisson summation formula (C.14). Indeed, since g is smooth, its Fourier transform $\hat g(\xi)$ will decay rapidly, and so the primary contribution in the Poisson formula comes from $P_{0,\alpha} = \hat g(0) = \int_{-\infty}^{\infty} g(t)dt$. The integral may be expressed in terms of the beta function by a change of variables $w = (1+e^{bt})^{-1}$, with the result that $b^{-1}B(c,1-c) = b^{-1}\Gamma(c)\Gamma(1-c)$ for $c = a/b$. Then, from Euler's reflection formula $\Gamma(z)\Gamma(1-z) = \pi/\sin(\pi z)$, and using the normalized sinc function $\mathrm{sinc}\,(x) = \sin(\pi x)/(\pi x)$, we arrive at

$$P_{0,\alpha} = \big(\log 2 \cdot \mathrm{sinc}((2\alpha+1)^{-1})\big)^{-1}. \tag{14.12}$$

Figure 14.1 shows that $P_{0,\alpha}$ provides an adequate summary for $\alpha \leq 2$.

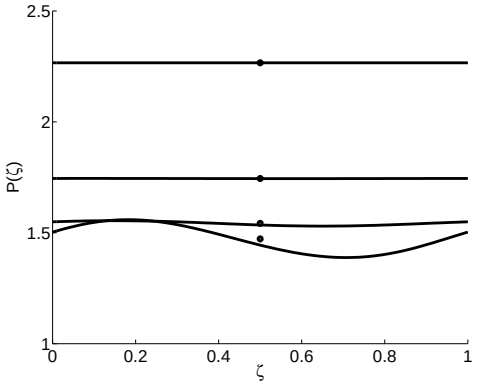

Figure 14.1 Periodic function $P^\circ(\zeta)$ appearing in Bayes-minimax risk, Theorem 14.2, for $p = 2$ and, from bottom to top, $\alpha = 4, 2, 1, 0.5$. Solid circles show the approximation by (14.12) in each case.

14.5 Asymptotic Efficiency

We again use the approach outlined in Chapter 4.11, which involves constructing near least favorable priors π_ϵ that asymptotically concentrate on Θ as $\epsilon \searrow 0$. More specifically, in line with the strategy (4.68)–(4.72), for each $\gamma < 1$, we construct $\pi_\epsilon \in \mathcal{M}^\alpha_{p\cdot}$ such that $B(\pi_\epsilon) \geq \gamma B(\gamma C, \epsilon)$ and verify that $\pi_\epsilon(\Theta) \to 1$, as well as the technical step (4.70).

The idea is to use the renormalized problem $Q(1, 1)$ and $Q(\gamma, 1)$ to build approximately least favorable priors and then to "translate" them to the appropriate sets of resolution levels corresponding to noise level ϵ.

Thus, for each given value $\gamma < 1$, we choose $J = J(\gamma)$ and $M = M(\gamma)$ and then priors π_j, $j = -J, \ldots, J$ such that $\operatorname{supp} \pi_j \subset [-M, M]$ and $E_{\pi_j}|\mu|^p \leq \gamma^p 2^{-\omega j p}$ and together $\{\pi_j\}$ form a near maximizer of $Q(\gamma, 1)$:

$$\sum_{-J}^{J} 2^j B(\pi_j) \geq \gamma Q(\gamma, 1) = \gamma \sum_{-\infty}^{\infty} 2^j \beta_p(\gamma 2^{-\omega j}).$$

To obtain J, we rely on convergence of the sum established in the proof of Theorem 14.2. To obtain M and to construct the individual π_j, we may appeal to Exercise 4.5 as in case (1) of the proof of Theorem 13.17 for ℓ_p balls.

To perform the "translation", we focus on a subsequence of noise levels ϵ_h defined by $C/\epsilon_h = 2^{\omega h}$, for $h \in \mathbb{N}$. [Exercise 14.5 discusses other values of ϵ]. The prior π_{ϵ_h} concentrates on the $2J + 1$ levels $h + j$ centered at $h = \omega^{-1} \log_2 C/\epsilon_h$. Let $\{\mu_{jk}, k \in \mathbb{N}\}$ be an iid sequence drawn from π_j. For $|j| \leq J$, set

$$\theta_{h+j,k} = \epsilon_h \mu_{jk} \qquad k = 1, \ldots, 2^{h+j}. \tag{14.13}$$

Hence, as $\epsilon \to 0$, the near least favorable priors charge a fixed number $2J(\gamma) + 1$ of ever higher frequency bands.

We now verify conditions (4.68) – (4.70) for the sequence π_{ϵ_h}, noting that J and M are

fixed. Working through the definitions and exploiting the invariances (14.9), we have

$$B(\pi_{\epsilon_h}) = \epsilon_h^2 \sum_{j=h-J}^{h+J} 2^j B(\pi_{j-h}) = \epsilon_h^2 2^h \sum_{j=-J}^{J} 2^j B(\pi_j)$$
$$\geq \gamma \epsilon_h^2 2^h Q(\gamma, 1) = \gamma Q(\gamma C, \epsilon_h) \geq \gamma B(\gamma C, \epsilon_h).$$

Recalling the definition of ϵ_h and that $a = \alpha + 1/2 - 1/p = \omega - 1/p$, we have with probability one under the prior π_{ϵ_h} that

$$\theta \in \Theta(C) \Leftrightarrow \sum_k |\theta_{h+j,k}|^p \leq C^p 2^{-a(h+j)p} \quad \text{for } |j| \leq J,$$
$$\Leftrightarrow n_{jh}^{-1} \sum_{k=1}^{n_{jh}} |\mu_{jk}|^p \leq 2^{-\omega j p} \quad \text{for } |j| \leq J,$$

where $n_{jh} = 2^{j+h}$.

Write $X_{jk} = |\mu_{jk}|^p - E|\mu_{jk}|^p$ and set $t_j = (1-\gamma^p)2^{-j\omega p}$. From the moment condition on π_j, it follows that $\{\theta \notin \Theta(C)\} \subset \cup_{j=-J}^{J} \Omega_{jh}$ where

$$\Omega_{jh} = \{n_{jh}^{-1} \sum_{k=1}^{n_{jh}} X_{jk} > t_j\}.$$

Since the probability $P(\Omega_{jh}) \to 0$ as $h \to \infty$ by the law of large numbers, for each of a finite number $2J+1$ of indices j, we conclude that $\pi_{\epsilon_h}(\Theta(C)) \to 1$.

Finally, to check (4.70), observe first that $\|\theta_{v_{\epsilon_h}}\|^2 \leq E_{\pi_{\epsilon_h}}[\|\theta\|^2 | \theta \in \Theta, y]$ and that for π_{ϵ_h} we have, with probability one,

$$\|\theta\|^2 = \epsilon_h^2 \sum_{j=-J}^{J} \sum_{k=1}^{2^{j+h}} |\mu_{jk}|^2 \leq M^2 2^{J+1} C^{2(1-r)} \epsilon_h^{2r}.$$

Consequently,

$$\mathbb{E}\{\|\hat{\theta}_{v_\epsilon}\|^2 + \|\theta\|^2, \Theta^c\} \leq 2c(M, J) B(C, \epsilon_h) \pi_{\epsilon_h}(\Theta^c)$$

and the right side is $o(B(C, \epsilon_h))$ as required, again because $\pi_{\epsilon_h}(\Theta^c) \to 0$.

In conclusion, from Theorem 14.2 and (14.10), we have

$$\frac{B(\gamma C, \epsilon_h)}{B(C, \epsilon_h)} \sim \gamma^{2(1-r)} \frac{P(\gamma C/\epsilon_h)}{P(C/\epsilon_h)} = \gamma^{2(1-r)} \frac{P^\circ(\omega^{-1} \log_2 \gamma)}{P^\circ(0)},$$

and now continuity of P° as $\gamma \to 1$ establishes (4.72).

14.6 Linear Estimates

Using results from Chapter 4, 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 $\Theta = \Theta_{p.}^{\alpha}(C)$ are solid and orthosymmetric, so that

by Theorem 9.5 the linear minimax risk is determined by the quadratic hull of Θ. It follows from the definitions (Exercise 14.2) that

$$\mathrm{QHull}(\Theta^{\alpha}_{p.}) = \Theta^{\alpha'}_{p'.} \qquad p' = p \vee 2, \;\; \alpha' = \alpha - 1/p + 1/p'. \tag{14.14}$$

In particular, $\Theta^{\alpha}_{p.}$ is quadratically convex only if p is at least 2. The Ibragimov-Hasminskii theorem 4.17 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

$$\begin{aligned} R_L(\Theta^{\alpha}_{p.}(C), \epsilon) &\asymp R_N(\Theta^{\alpha'}_{p'.}, \epsilon) \\ &\asymp C^{2(1-r')}\epsilon^{2r'} \qquad\qquad r' = 2\alpha'/(2\alpha'+1). \end{aligned} \tag{14.15}$$

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. This property extends to all $q \leq \infty$ (Donoho and Johnstone, 1998). 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 $\Theta = \Theta^{\alpha}_{p.}(C)$ 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 $\hat{\theta} = (\hat{\theta}_{j,c_j})$, where for each j, $c_j \in [0,1]$ is a scalar and

$$\hat{\theta}_{j,c_j} = c_j y_j \tag{14.16}$$

as vectors in $\mathbb{R}^{2^j}$, and hence that

$$R_L(\Theta, \epsilon) = \inf_{(c_j)} \sup_{\Theta} \sum_j E\|\hat{\theta}_{j,c_j} - \theta_j\|^2. \tag{14.17}$$

A formal verification again uses the observation that $R_L(\Theta) = R_L(\bar{\Theta})$ where $\bar{\Theta} = \mathrm{QHull}(\Theta) = \Theta^{\alpha'}_{p'.}$ as described earlier. Given $\tau \in \bar{\Theta}$, construct $\bar{\tau}$ by setting $\bar{\tau}^2_{jk} \equiv \mathrm{ave}_k \tau^2_{jk}$: since $p' \geq 2$, one verifies that $\bar{\tau} \in \bar{\Theta}$ also. Formula (4.48) shows that $R(\Theta(\tau))$ is a concave function of (τ_i^2), and hence that $R(\Theta(\bar{\tau})) \geq R(\Theta(\tau))$. Consequently, the hardest rectangular subproblem lies among those hyperrectangles that are symmetric within levels j. Since the minimax linear estimator for rectangle $\bar{\tau}$ has the form $\hat{\theta}_{c(\bar{\tau}),I} = [\bar{\tau}_I^2/(\bar{\tau}_I^2 + \epsilon^2)]y_i$, it follows that the minimax linear estimator for $\bar{\Theta}$ has the form (14.16), which establishes (14.17).

14.7 Near Minimaxity of Threshold Estimators

Although described in terms of a two parameter family of co-ordinatewise Bayes estimators, the asymptotic minimax estimators derived at (14.4) 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.

Consider *level dependent* soft thresholding estimators, so that if $\lambda = (\lambda_j)$, we set

$$\hat{\theta}_{\lambda,jk}(y) = \hat{\delta}_S(y_{jk}, \lambda_j \epsilon),$$

where $\hat{\delta}_S(y, \lambda)$ is soft thresholding, cf (8.3). The minimax risk among such soft threshold

estimators over Θ is defined by

$$R_S(\Theta, \epsilon) = \inf_{(\lambda_j)} \sup_{\Theta} E_\theta \|\hat{\theta}_\lambda - \theta\|^2.$$

Over the full range of p, and for a large range of α, thresholding is nearly minimax among all non-linear estimators.

Theorem 14.4 *For $0 < p \le \infty$ and $\alpha > (1/p - 1/2)_+$, with $\Theta = \Theta^\alpha_{p.}(C)$, we have*

$$R_S(\Theta, \epsilon) \le \Lambda(p) R_N(\Theta, \epsilon)(1 + o(1)), \qquad \text{as } \epsilon \to 0.$$

Proof The argument is analogous to that for soft thresholding on ℓ_p balls in $\mathbb{R}^n$, Theorem 13.18. We bound $R_S(\Theta, \epsilon)$ in terms of the Bayes minimax risk $B(C, \epsilon)$ given by (14.2) and (14.6), and then appeal to the equivalence theorem $R_N(\Theta, \epsilon) \sim B(C, \epsilon)$ as $\epsilon \to 0$.

Given $\theta = (\theta_{jk})$, let $\mu_{jk} = \theta_{jk}/\epsilon$. Let π_j denote the empirical measure of $\{\mu_{jk}, k = 1, \ldots, 2^j\}$, so that $\pi_j = 2^{-j} \sum_k \delta_{\mu_{jk}}$. Recalling the definitions of threshold risk $r(\lambda, \mu)$ and Bayes threshold risk $B(\lambda, \pi)$ for unit noise level from Chapter 13, we have

$$E_\theta \|\hat{\theta}_\lambda - \theta\|^2 = \sum_{jk} \epsilon^2 r(\lambda_j, \mu_{jk}) = \sum_j 2^j \epsilon^2 B(\lambda_j, \pi_j).$$

Let $\eta_j = (C/\epsilon) 2^{-\omega j}$; one verifies exactly as at (13.54) that $\theta \in \Theta^\alpha_{p.}(C)$ implies $\pi_j \in \mathfrak{m}_p(\eta_j)$, so that

$$\inf_\lambda \sup_{\Theta^\alpha_{p.}(C)} E_\theta \|\hat{\theta}_\lambda - \theta\|^2 \le \sum_j 2^j \epsilon^2 \beta_{S,p}(\eta_j),$$

since the minimization over thresholds λ_j can be carried out level by level. Now apply Theorem 13.15 to bound $\beta_{S,p}(\eta_j) \le \Lambda(p) \beta_p(\eta_j)$, and so bound the right side of the preceding display by $\Lambda(p) \sum_j 2^j \beta_p(C 2^{-\omega j}, \epsilon)$. Hence, using (14.6)

$$R_S(\Theta, \epsilon) \le \Lambda(p) B(C, \epsilon).$$

Our conclusion now follows from Theorem 14.3. □

Remark. In principle, one could allow the thresholds to depend on location k as well as scale j: $\lambda = (\lambda_{jk})$. Along the lines described in Remark 13.19 and Exercise 13.9, one can define a Bayes minimax threshold risk $B_S(\mathcal{M}, \epsilon)$, show that it is bounded by $\Lambda(p) B(\mathcal{M}, \epsilon)$, and that minimax choices of λ in fact depend only on j and not on k. Further details are in Donoho and Johnstone (1998, §5).

Since $\Lambda(p) \le 2.22$ for $p \ge 2$, 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 (α, p, q, C), however. We recall that special choices of threshold not depending on prior specifications of these parameters were discussed in Chapters 10 and 12.

Similar results may be established for hard thresholding.

14.8 Notes

The results of this chapter are specialized to Besov spaces with $q = \infty$ from those in Donoho and Johnstone (1998), which considers both more general Besov spaces and also the Triebel scale. In these more general settings, the levels j do not decouple in the fashion that led to (14.8), but one may obtain similar asymptotic behavior by using homogeneity properties of the $Q(C, \epsilon)$ problem with respect to scaling and level shifts.

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 π is a probability measure on $\ell_2(\mathcal{I})$, let $\tau_p(\pi)$ denote the sequence of marginal pth moments

$$\tau_p(\pi)_I = (E_\pi |\theta_I|^p)^{1/p}. \qquad I \in \mathcal{I}, \quad p \in (0, \infty].$$

If Θ 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 of each equality gives the notation used in the text, and the right side the notation according to the convention just introduced. The third column specifies the index set $\mathcal{I}$ and the fourth the reference.

(i) Intervals $\Theta = [-\tau, \tau] \subset \mathbb{R}$:	$\mathfrak{m}_p(\tau) = \mathcal{M}_p([-\tau, \tau])$.	$\{1\}$	§13.3
(ii) ℓ_p balls:	$\mathcal{M}_n = \mathcal{M}_p(\Theta_{n,p}(C))$,	$\{1, \ldots, n\}$	(13.37)
(iii) Ellipsoids in Pinsker's Theorem:	$\mathcal{M}(C) = \mathcal{M}_2(\Theta(a, C))$,	$\mathbb{N}$	§5.4
(iv) Besov bodies:	$\mathcal{M}^\alpha_{p\cdot}(C) = \mathcal{M}_p(\Theta^\alpha_{p\cdot}(C))$.	§9.8	§14.2

Exercises

14.1 *(Compactness criteria.)* (a) Show, using the total boundedness criterion C.16, that $\Theta^\alpha_{p\cdot}(C)$ is ℓ_2-compact when $\alpha > (1/p - 1/2)_+$.
(b) Show, using the tightness criterion given in C.18 that $\mathcal{M}^\alpha_{p\cdot}(C)$ is compact in the topology of weak convergence of probability measures on $\mathcal{P}(\ell_2)$ when $\alpha > (1/p - 1/2)_+$.

14.2 *(Quadratic hull of Besov bodies.)* Verify (14.14).

14.3 *(Threshold minimax theorem.)* Formulate and prove a version of the threshold minimax theorem 13.13 in the Bayes minimax setting of this chapter.

14.4 *Completing proof of asymptotic efficiency.*
(a) Show that a function g on $[0, 1]$ satisfies $g(t) \to 1$ as $t \to 0$ if and only if for some $\beta \in (0, 1)$, we have $g(b\beta^j) \to 1$ as $j \in \mathbb{N} \to \infty$, uniformly in $b \in [\beta, 1]$.
(b) Fix $b \in [2^{-\omega}, 1]$ and for $h \in \mathbb{N}$ define ϵ_h by $C/\epsilon_h = b2^{\omega h}$. Modify the argument of Section 14.5 to show that $R_N(\Theta, \epsilon_h) \sim B(C, \epsilon_h)$. [Hint: replace $Q(1, 1)$ by $Q(b, 1)$ in the argument.]

14.5 *Completing proof of asymptotic efficiency.* (a) For general ϵ, write $C/\epsilon = 2^{\omega(h+s)}$ for $h \in \mathbb{N}$ and $s \in [0, 1)$. Verify (4.72) by an appeal to uniform continuity of P°.
(b) Extend the construction of Section 14.5 to general ϵ as follows. Fix $\gamma < 1$. First choose $J = J(\gamma)$ so that $\sum_{j=-J}^{J} 2^j \beta_p(\gamma 2^{-\omega j}) \geq \gamma^{1/3} Q(\gamma, 1)$. Now choose $L = L(\gamma)$ so that if $s_\ell = \ell/L$, then for each $\ell = 0, \ldots, L-1$, $Q(\gamma 2^{\omega s_\ell}) \geq \gamma^{1/3} Q(\gamma 2^{\omega s_{\ell+1}})$. Finally, again using Exercise 4.5, choose $M = M(\gamma)$ and priors $\pi_{j\ell}$ for $(j, \ell) \in \{-J, \ldots, J\} \times \{0, \ldots, L\}$ s.t.

$$\operatorname{supp}(\pi_{j\ell}) \subset [-M, M], \qquad E_{\pi_{j\ell}} |\mu|^p \leq \gamma^p 2^{\omega s_\ell p - \omega j p}, \qquad B(\pi_{j\ell}) \geq \gamma^{1/3} \beta_p(\gamma 2^{\omega s_\ell - \omega j}).$$

From ϵ define h and s and then $\ell = [Ls]$. Define priors π_ϵ using (14.13) with $\pi_{j\ell}$ in place of π_j for $|j| \leq J$. Now verify (4.68) - (4.70) for the priors π_ϵ along the lines of the proof of Section 14.5.

15

Continuous v. Sampled Data

Our theory has been developed so far almost 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*

$$Y_\epsilon(t) = \int_0^t f(s)ds + \epsilon W(t), \qquad t \in [0,1], \tag{15.1}$$

which we have seen is in fact an equivalent representation of (3.1).

Closer to many applications is the *sampled data model* in which one observes

$$\tilde{y}_l = f(t_l) + \sigma \tilde{z}_l, \qquad l = 1, \dots n, \tag{15.2}$$

and it is desired to estimate the function $f \in L_2[0,1]$. Throughout we consider the equally spaced case $t_l = l/n$.

For many purposes, the models (15.1) and (15.2) are very similar, and methods and results developed in one should apply equally well in the other. A general equivalence result of Brown and Low (1996a) implies that for bounded loss function $\ell(.)$ and for collections $\mathcal{F}$ which are bounded subsets of Hölder classes $C^\alpha, \alpha > 1/2$, we have as $\epsilon \to 0$,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} E\ell\Big(\|\hat{f}(Y) - f\|^2_{L^2[0,1]}\Big) \sim \inf_{\hat{f}} \sup_{f \in \mathcal{F}} E\ell\Big(\|\hat{f}(\tilde{y}) - f\|^2_{L^2[0,1]}\Big) \tag{15.3}$$

the expectation on the left-hand side being with respect to white noise observations Y in (15.1) and on the right hand-side being with respect to $\tilde{y}$ in (15.2). However, the general equivalence result fails for $\alpha \le 1/2$ 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 (15.1) while computer algorithms work with empirical wavelet coefficients derived from the sampled data model (15.2). We compare explicitly the sampling operators corresponding to pointwise evaluation and integration against a localized scaling function. The approach taken in this chapter is based on Donoho and Johnstone (1999) and Johnstone and Silverman (2004b).

15.1 The Sampled Data Model: A Wavelet Crime?

The simplest non-parametric regression model (15.2) posits an unknown function observed in homoscedastic Gaussian noise at equally spaced points $t_l = l/n$. We assume that the $\tilde{z}_l$

are i.i.d standard Gaussian variables and that the noise level σ is known. For convenience, suppose throughout that $n = 2^J$ for some integer J.

We have studied at length the white noise model (15.1) which after conversion to wavelet coefficients $y_I = \langle dY_\epsilon, \psi_I \rangle, \theta_I = \langle f, \psi_I \rangle, z_I = \langle dW, \psi_I \rangle$ takes the sequence model form

$$y_I = \theta_I + \epsilon z_I, \qquad I = (j,k),\ j \geq 0, k = 1, \ldots, 2^j. \tag{15.4}$$

This leads to a possibly troubling dichotomy. Much of the theory developed to study wavelet methods is carried out using functions of a continous variable, uses the multiresolution analysis and smoothness classes of functions on $\mathbb{R}$ or $[0,1]$, and the sequence model (15.4). Almost inevitably, most actual data processing is carried out on discrete, sampled data, which in simple cases might be modeled by (15.2).

There is therefore a need to make a connection between the continuous and sampled models, and to show, under appropriate conditions, that conclusions in one model are valid for the other and vice versa. To do this, we compare minimax risks for estimation of f based on sequence data y from (15.4) with that based on sampled data $\tilde{y}$ from (15.2). Hence, set

$$\begin{aligned} R(\mathcal{F}, \epsilon) &= \inf_{\hat{f}(y)} \sup_{f \in \mathcal{F}} E \| \hat{f}(y) - f \|_2^2, \\ \tilde{R}(\mathcal{F}, n) &= \inf_{\hat{f}(\tilde{y})} \sup_{f \in \mathcal{F}} E \| \hat{f}(\tilde{y}) - f \|_2^2. \end{aligned} \tag{15.5}$$

The error of estimation is measured in both cases in the norm of $L_2[0,1]$. The parameter space $\mathcal{F}$ is defined through the wavelet coefficients corresponding to f, as at (9.49):

$$\mathcal{F} = \{ f : \theta[f] \in \Theta^\alpha_{p,q}(C) \}.$$

Remark. One might also be interested in the error measured in the discrete norm

$$\| \hat{f} - f \|_n^2 = (1/n) \sum [\hat{f}(t_l) - f(t_l)]^2. \tag{15.6}$$

Section 15.5 shows this norm is equivalent to $\int_0^1 (\hat{f} - f)^2$ under present assumptions.

Assumption (A) on the wavelet. In this chapter the choice of α, p and q is fixed at the outset, so that we focus on a fixed Besov space $B^\alpha_{p,q}[0,1]$. Given this selection, we choose a Daubechies pair (ϕ, ψ) and an orthonormal wavelet basis (ψ_I) for $L_2[0,1]$ consisting of wavelets of compact support, with elements having R continuous derivatives ($\psi_I \in C^R$) and $(D+1)$ vanishing moments. The basis is chosen so that $\min(R, D) \geq \alpha$, so that it is an unconditional basis of $B^\alpha_{p,q}[0,1]$, and the norm is equivalently given by the Besov sequence norm on the wavelet coefficients. We also assume that the CDJV construction (cf. Section 7.1) is used for wavelets that intersect the boundary of $[0,1]$.

Theorem 15.1 *Let $\alpha > 1/p$ and $1 \leq p, q \leq \infty$; or else $\alpha = p = q = 1$. Then, with $\epsilon_n = \sigma/\sqrt{n}$, we have*

$$\tilde{R}(\mathcal{F}, n) \geq R(\mathcal{F}, \epsilon_n)(1 + o(1)), \quad n \to \infty. \tag{15.7}$$

In words, there is no estimator giving a worst-case performance in the sampled-data-problem (15.2) which is substantially better than what we can get for the worst-case performance of procedures in the white-noise-problem (15.4).

For *upper bounds* to risk in the sampled data problem (15.2), we will specialize to estimators derived by applying certain coordinatewise mappings to the noisy wavelet coefficients.

For the white noise model, this means the estimate is of the form

$$\hat{f} = \sum_I \delta(y_I)\psi_I$$

where each function $\delta_I(y)$ either belongs to one of three specific families – *Linear, Soft Thresholding, or Hard Thresholding* – or else is a general scalar function of a scalar argument. The families are:

- $(\mathcal{E}_L)$ diagonal linear procedures in the wavelet domain, $\delta_I^L(y) = c_I \cdot y$,
- $(\mathcal{E}_S)$ soft thresholding of wavelet coefficients, $\delta_I^S(y) = (|y| - \lambda_I)_+ \mathrm{sgn}(y)$,
- $(\mathcal{E}_H)$ hard thresholding of wavelet coefficients, $\delta_I^H(y) = y1_{\{|y|\geq\lambda_I\}}$, and
- $(\mathcal{E}_N)$ scalar nonlinearities of wavelet coefficients, with arbitrary $\delta_I^N(y)$.

For the sampled-data problem, this means that the estimate is of the form

$$\hat{f} = \sum_I \delta_I(y_I^{(n)})\psi_I, \tag{15.8}$$

where $y_I^{(n)}$ is an empirical wavelet coefficient based on the sampled data $(\tilde{y}_i)$, see Section 15.4 below, and the δ_I belong to one of the families $\mathcal{E}$. Then define the $\mathcal{E}$-minimax risks in the two problems:

$$R_{\mathcal{E}}(\mathcal{F}, \epsilon) = \inf_{\hat{f}\in\mathcal{E}} \sup_{f\in\mathcal{F}} E_{Y_\epsilon} \|\hat{f} - f\|^2_{L^2[0,1]} \tag{15.9}$$

and

$$\tilde{R}_{\mathcal{E}}(\mathcal{F}, n) = \inf_{\hat{f}\in\mathcal{E}} \sup_{f\in\mathcal{F}} E_{\mathbf{y}_n} \|\hat{f} - f\|^2_{L^2[0,1]}. \tag{15.10}$$

With this notation established, we have

Theorem 15.2 *Let $\alpha > 1/p$ and $1 \leq p, q \leq \infty$ or $\alpha = p = q = 1$. Adopt assumption (A) on the wavelet basis. For each of the four classes $\mathcal{E}$ of coordinatewise estimators,*

$$\tilde{R}_{\mathcal{E}}(\mathcal{F}, n) \leq R_{\mathcal{E}}(\mathcal{F}, \epsilon_n)(1 + o(1)), \quad n \to \infty. \tag{15.11}$$

Our approach is to make an explicit construction transforming a sampled-data problem into a quasi-white-noise problem in which estimates from the white noise model can be employed. We then show that these estimates on the quasi-white-noise-model data behave nearly as well as on the truly-white-noise-model data. The observations in the quasi-white-noise problem have constant variance, but may be correlated. The restriction to coordinatewise estimators means that the correlation structure plays no role.

Furthermore, we saw in the last chapter in Theorems 14.1–14.3 that co-ordinatewise nonlinear rules were asymptotically minimax: $R(\mathcal{F}, \epsilon_n) \sim R_{\mathcal{E}_N}(\mathcal{F}, \epsilon_n)$ for the $q = \infty$ cases considered there, and the same conclusion holds more generally for $p \leq q$ (Donoho and Johnstone, 1998).

Remark. The assumptions on (α, p, q) in Theorems 15.1 and 15.2 are needed for the bounds to be described in Section 15.4. Informally, they correspond to a requirement that point evaluation $f \to f(t_0)$ is well defined and continuous, as is needed for model (15.2) to

be stably defined. For example, if $\alpha > 1/p$, then functions in $B^{\alpha}_{p,q}$ are uniformly continuous (by the embedding result Proposition 9.12), while if $\alpha = p = q = 1$, one can use the embedding $B^1_{1,1} \subset TV$ to make sense of point evaluation, by agreeing to use, say, the left continuous version of $f \in TV$. For further discussion, see (Donoho, 1992b, Section 6.1).

15.2 The Projected White Noise Model

Finite dimensional submodels of (15.1) are of interest for a number of reasons. Firstly, when the noise level ϵ is of order $n^{-1/2}$, a model with n observed coefficients is a closer relative of the regression model (15.2). Secondly, for a given parameter space Θ, finite dimensional submodels can be found with dimension $m(\epsilon)$ depending on ϵ 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 ϕ be the scaling function corresponding to the orthonormal wavelet ψ used in the previous section. We consider only projections onto the increasing sequence of multiresolution spaces $V_j = \text{span}\,\{\phi_{ji}, i = 1, \ldots 2^j\}$. Given ϵ, fix a level $J = J(\epsilon)$, set $m = m_\epsilon = 2^{J(\epsilon)}$ and define

$$y_i = \langle \phi_{Ji}, dY \rangle, \qquad z_i = \langle \phi_{Ji}, dW \rangle, \qquad i = 1, \ldots m.$$

The *projected white noise model* refers to observations

$$y_i = \langle f, \phi_{Ji} \rangle + \epsilon z_i, \qquad i = 1, \ldots m. \tag{15.12}$$

Write $y^{[m]}$ for the projected data $y_1, \ldots, y_m$. When $\epsilon = n^{-1/2}$, the choice $J = \log_2 n$ yields an n-dimensional model which is an approximation to (15.2), in a sense to be explored below.

The projected white noise model can be expressed in terms of wavelet coefficients. Indeed, since $V_J = \oplus_{j<J} W_j$, it is equivalent to the 2^J-dimensional submodel of the sequence model given by

$$y_I = \theta_I + \epsilon z_I, \qquad I \in \mathcal{I}^J, \tag{15.13}$$

where we define $\mathcal{I}^J = \cup_{j<J} \mathcal{I}_j$.

Estimation of the unknown coefficients $\langle f, \phi_{Ji} \rangle$ is done in the wavelet basis. Recall that ϕ_{Ji} 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 $\langle f, \phi_{Ji} \rangle$ to $\langle f, \psi_I \rangle = \theta_I$ is called the *discrete wavelet transform* W. Its matrix elements W_{Ii} are just the inner products $\langle \psi_I, \phi_{Ji} \rangle$.

The estimation procedure could then be summarized by the diagram

$$\begin{array}{ccc} (y_l) & \xrightarrow{W} & (y_I) \\ \downarrow & & \downarrow \\ (\hat{f}_{n,l}) & \xleftarrow{W^T} & (\hat{\delta}_I(y_I)) \end{array} \tag{15.14}$$

which is the same as (7.22), except that this diagram refers to observations on inner products, (15.12), whereas the earlier diagram used observations from the sampling model (7.21), here

written in the form (15.2). As noted at the end of Section 7.3, the transformation W is the same in both cases.

Consider now the minimax risk of estimation of $f \in \mathcal{F}$ using data from the projected model (15.12). Because of the Parseval relation (1.25), we may work in the sequence model and wavelet coefficient domain.

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 $\|\theta\|_{2,m^\perp}^2 = \sum_{I \notin \mathcal{I}^J} \theta_I^2$ The following decomposition emphasises the "tail bias" term that results from estimating only up to level J :

$$\|\hat{\theta} - \theta\|^2 = \|\hat{\theta} - \theta\|_{2,m}^2 + \|\theta\|_{2,m^\perp}^2. \tag{15.15}$$

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 $\|\theta\|_{2,m^\perp}^2 = \|f - P_m f\|^2$.

We write $y^{[m]}$ when needed to distinguish data in the projected model from data y in the full sequence model. In the projected model, we consider estimation with loss function

$$L(\hat{\theta}, \theta) = \|\hat{\theta}(y^{[m]}) - \theta\|_{2,m}^2. \tag{15.16}$$

and the projected parameter space

$$\Theta^{[m]}(C) = \{\theta \in \mathbb{R}^m : \|\theta\|_{b_{p,q}^\alpha} \le C\}.$$

The minimax risk in this reduced problem is

$$R_N(\Theta^{[m]}(C); \epsilon) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta^{[m]}(C)} \mathbb{E}\|\hat{\theta}(y^{[m]}) - \theta\|_{2,m}^2.$$

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. For this it is helpful to recall, in current notation, a bound on the maximum tail bias over smoothness classes $\Theta_{p,q}^\alpha$ that was established at (9.60).

Lemma 15.3 *Let $\alpha' = \alpha - (1/p - 1/2)_+ > 0$. Then for a constant $K = K(\alpha')$,*

$$\Delta_m(\Theta) = \sup_{\mathcal{F}_{p,q}^\alpha(C)} \|f - P_{2^J} f\|^2 = \sup_{\Theta_{p,q}^\alpha(C)} \|\theta\|_{2,m^\perp}^2 \le KC^2 2^{-2J\alpha'}.$$

We now establish equivalence of the projected white noise model with the full model.

Proposition 15.4 *Suppose that*

$$J(\epsilon) = \gamma \log_2 \epsilon^{-2}, \qquad \gamma > (1/(2\alpha + 1))(\alpha/\alpha'). \tag{15.17}$$

With $m_\epsilon = 2^{J(\epsilon)}$, we then have

$$R_N(\Theta^{[m_\epsilon]}(C), \epsilon) \sim R_N(\Theta(C), \epsilon) \qquad \epsilon \to 0.$$

Proof An arbitrary estimator $\hat{\theta}(y^{[m]})$ in the projected model can be extended to an estimator in the full sequence model by appending zeros—let $\mathcal{E}^{[m]}$ denote the class so obtained. From (15.15) we obtain

$$\inf_{\hat{\theta} \in \mathcal{E}^{[m]}} \sup_{\Theta(C)} E\|\hat{\theta} - \theta\|^2 \le R_N(\Theta^{[m]}(C), \epsilon) + \Delta_m(\Theta).$$

The left side exceeds $R_N(\Theta(C), \epsilon)$ and Lemma 15.3 shows that, with γ chosen as in (15.17),

$$\Delta_m(\Theta) \le KC^2 2^{-2J\alpha'} = o((\epsilon^2)^{2\alpha/(2\alpha+1)}),$$

so that $\Delta_m(\Theta) = o(R_N(\Theta(C), \epsilon))$, showing that the projected model is asymptotically no easier.

In the reverse direction, we think of $\Theta^{[m]}(C)$ as a product set $\Theta^{[m]}(C) \times \{0\} \subset \Theta(C)$, and apply Proposition 4.16 to conclude that the projected model is not harder:

$$R_N(\Theta^{[m]}(C), \epsilon) = R_N(\Theta^{[m]}(C) \times \{0\}, \epsilon) \le R_N(\Theta(C), \epsilon). \qquad \square$$

Remark. The ratio $(1/(2\alpha + 1))(\alpha/\alpha')$ is certainly less than 1 whenever (i) $p \ge 2$ and $\alpha > 0$, or (ii) $p < 2$ and $\alpha \ge 1/p$.

15.3 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 Theorem 15.1, a lower bound to minimax risk in the sampling problem. The idea 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 15.4 shows that the continous problem, in sequence space form, can be projected to a level $J_{0n} = \gamma \log_2 \epsilon_n^{-2}$ given by (15.17) without loss of difficulty. Let us formulate the sampling problem in a corresponding manner.

In the "sampling problem", we observe data in model (15.2) and seek to estimate f, in principle using loss function $\|\hat{f}(\tilde{y}) - f\|_2^2$. However, we only make our task easier by restricting attention to estimating $P_m f$, the projection of f onto V_{J_0}, and hence to estimation of $\theta = (\theta_I, I \in \mathcal{I}^{J_0})$. The loss function is then

$$L(\hat{\theta}(\tilde{y}), \theta) = \|\hat{\theta}(\tilde{y}) - \theta\|_{2,m}^2. \tag{15.18}$$

When $f \in V_{J_0}$, we have $f = \sum_{I \in \mathcal{I}^{J_0}} \theta_I \psi_I$, and may rewrite (15.2) as

$$\tilde{y}_l = (T\theta)_l + \sigma \tilde{z}_l, \tag{15.19}$$

where T is given in co-ordinates by

$$(T\theta)_l = \sum_I \theta_I \psi_I(t_l). \tag{15.20}$$

We regard this as a map from $(\mathbb{R}^m, \|\cdot\|_{2,m})$ to $(\mathbb{R}^n, \|\cdot\|_n)$, where $\|\cdot\|_n$ is the time domain norm (15.6). It is not a (partial) isometry since the vectors $(\psi_I(t_l) : l = 1, \ldots, n)$ are not orthogonal in the discrete norm. However, it comes close; at the end of the section we establish

Lemma 15.5 *Under assumption (A) on the wavelet system (ψ_I), if T is defined by* (15.20) *for $m = 2^{J_0} < n$, then*

$$\lambda_{\max}(T^{\mathrm{T}}T) \le 1 + c J_0 2^{J_0}/n.$$

As remarked earlier, estimation of $P_m f$ is easier than estimation of f; specifically, from (15.5) and (15.15) we have

$$\tilde{R}(\mathcal{F}, n) \geq \tilde{R}_N(\Theta^{[m]}(C); \epsilon_n).$$

The minimax risk in the sampling problem is, setting $\epsilon_n = \sigma/\sqrt{n}$,

$$\begin{aligned}\tilde{R}_N(\Theta^{[m]}(C); \epsilon_n) &= \inf_{\hat{\theta}(\tilde{y})} \sup_{\theta \in \Theta^{[m]}(C)} \mathbb{E}\|\hat{\theta}(\tilde{y}) - \theta\|_{2,m}^2 \\ &= \sup_{\pi \subset \Theta^{[m]}(C)} \tilde{B}(\pi; \epsilon_n) =: \tilde{B}(C, \epsilon_n), \end{aligned} \tag{15.21}$$

where we used the minimax Theorem 4.12 and $\tilde{B}(\pi, \epsilon_n)$ denotes the Bayes risk in (15.19) and (15.18) when the prior $\pi(d\theta)$ on $\Theta^{[m]}(C)$ is used.

In the projected white noise model, we also use the minimax theorem to reexpress

$$R_N(\Theta^{[m]}(C); \epsilon) = \sup_{\pi \subset \Theta^{[m]}(C)} B(\pi; \epsilon) =: B(C, \epsilon), \tag{15.22}$$

where, as usual, $B(\pi, \epsilon)$ denotes the Bayes risk in (15.13) and (15.16) when the prior $\pi(d\theta)$ on $\Theta^{[m]}(C)$ is used. From Proposition 15.4, we then have

$$R_N(\Theta(C), \epsilon) = R_N(\Theta^{[m_\epsilon]}(C), \epsilon)(1 + o(1)) = B(C, \epsilon_n)(1 + o(1)).$$

With all this notational preparation, we have reduced the "sampling is not easier" result, Theorem 15.1 to proving the statement

$$\tilde{B}(C, \epsilon_n) \geq B(C, \epsilon_n)(1 + o(1)). \tag{15.23}$$

Pushing the sequence model observations (at noise level ϵ_n) through T generates some heteroscedasticity which may be bounded using Lemma 15.5. To see this, we introduce e_l, a vector of zeros except for $\sqrt{n}$ in the lth slot, so that $\|e_l\|_n = 1$ and $(Ty)_l = \sqrt{n}\langle e_l, Ty\rangle_n$. Then

$$\mathrm{Var}(Ty)_l = n\epsilon_n^2 E\langle e_l, Tz\rangle_n^2 = \sigma^2 \|T^{\mathrm{T}} e_l\|_{2,m}^2 \leq \sigma^2 \lambda_n^2$$

where $\lambda_n^2 = \lambda_{\max}(TT^{\mathrm{T}}) = \lambda_{\max}(T^{\mathrm{T}}T)$ is bounded in the Lemma. Now let $\tilde{w}$ be a zero mean Gaussian vector, independent of y, with covariance chosen so that $\mathrm{Var}(Ty + \tilde{w}) = \lambda_n^2 \sigma^2 I_n$. By construction, then, $Ty + \tilde{w} \overset{\mathcal{D}}{=} \tilde{y} = T\theta + \lambda_n \sigma \tilde{z}$.

To implement the basic idea of the proof, let π be a least favorable prior in the sequence problem (15.22) so that $B(\pi, \epsilon_n) = B(C, \epsilon_n)$. Let $\tilde{\theta}_{\pi, \lambda_n \sigma}(\tilde{y})$ denote the Bayes estimator of θ in the sampling model (15.19) and (15.18) *with noise level* $\lambda_n \sigma$.

We construct a *randomized* estimator in the sequence model using the auxiliary variable $\tilde{w}$:

$$\hat{\theta}(y, \tilde{w}) = \tilde{\theta}_{\pi, \lambda_n \sigma}(Ty + \tilde{w}) \overset{\mathcal{D}}{=} \tilde{\theta}_{\pi, \lambda_n \sigma}(\tilde{y})$$

where the equality in distribution holds for the laws of $Ty + \tilde{w}$ and $\tilde{y}$ given θ. Consequently

$$B(\hat{\theta}, \pi; \epsilon_n) = \mathbb{E}_\pi \mathbb{E}_{\theta;\epsilon_n} \|\hat{\theta}(y, \tilde{w}) - \theta\|_{2,m}^2 = \mathbb{E}_\pi \mathbb{E}_{T\theta;\lambda_n\sigma} \|\tilde{\theta}_{\pi,\lambda_n\sigma}(\tilde{y}) - \theta\|_{2,m}^2 = \tilde{B}(\pi; \lambda_n \epsilon_n).$$

Use of randomized rules (with a convex loss function) does not change the Bayes risk $B(\pi)$—see e.g. (A.13) in Appendix A—and so

$$B(C, \epsilon_n) = B(\pi; \epsilon_n) \le B(\hat{\theta}, \pi; \epsilon_n) = \tilde{B}(\pi; \lambda_n \epsilon_n) \le \tilde{B}(C; \lambda_n \epsilon_n),$$

where the last inequality uses (15.21). Appealing to the scaling bounds for Bayes-minimax risks (e.g. Lemma 4.28 and Exercise 4.8) we conclude that

$$\tilde{B}(C; \lambda\epsilon) \le \begin{cases} \lambda^2 \tilde{B}(C/\lambda; \epsilon) \le \lambda^2 \tilde{B}(C; \epsilon) & \text{if } \lambda > 1 \\ \tilde{B}(C; \epsilon) & \text{if } \lambda \le 1. \end{cases}$$

In summary, using again Lemma 15.5,

$$B(C, \epsilon_n) \le (\lambda_n^2 \vee 1)\tilde{B}(C, \epsilon_n) \le \tilde{B}(C, \epsilon_n)(1 + o(1)).$$

This completes the proof of (15.23), and hence of Theorem 15.1.

Proof of Lemma 15.5 The matrix representation $(a_{II'})$ of $A = T^{\mathrm{T}}T$ in the basis $(\psi_I, I \in \mathcal{I}^{J_0})$ is given by

$$a_{II'} = \langle \psi_I, \psi_{I'} \rangle_n = n^{-1} \sum_l \psi_I(t_l)\psi_{I'}(t_l).$$

Exercise 15.1 gives bounds on the distance of these inner products from exact orthogonality:

$$|\langle \psi_I, \psi_{I'} \rangle_n - \delta_{II'}| \le cn^{-1} 2^{(j+j')/2} \chi(I, I'), \tag{15.24}$$

where $\chi(I, I') = 1$ if $\operatorname{supp} \psi_I$ intersects $\operatorname{supp} \psi_{I'}$ and $= 0$ otherwise.

We aim to apply Schur's lemma, Corollary C.28, to A with weights $x_I = 2^{-j/2}$, hence we consider

$$\begin{aligned} S_I = \sum_{I'} |a_{II'}| 2^{-j'/2} &\le 2^{-j/2} + cn^{-1} \sum_{j'} 2^{(j+j')/2} \cdot 2^{(j'-j)_+} \cdot 2^{-j'/2} \\ &\le 2^{-j/2}(1 + cn^{-1} \sum_{j'} 2^{j \vee j'}) \end{aligned}$$

where in the first line we used (15.24) and bounded $\sum_{k'} \chi(I, I')$, the number of $\psi_{j'k'}$ whose supports hits that of ψ_I, by $c2^{(j'-j)_+}$. Now $j \le J_0$ and the sum is over $j' \le J_0$ and hence

$$S_I \le 2^{-j/2}(1 + cn^{-1} J_0 2^{J_0})$$

and the result follows from Schur's lemma. □

15.4 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 does not significantly inflate minimax risk relative to its use on genuine wavelet coefficients.

To do this, we exploit a projected model sequence indexed by dyadic powers of n, using

less than $\log_2 n$ levels, but of full asymptotic difficulty. Indeed, Proposition 15.4 shows that given $\Theta^\alpha_{p,q}$, full asymptotic difficulty can be achieved by choosing $\eta > 0$ such that

$$\gamma = \frac{1}{2\alpha + 1}\frac{\alpha}{\alpha'} + \eta < 1, \tag{15.25}$$

and then setting

$$m_n = 2^{J_{0n}} \qquad J_{0n} = \gamma \log_2 n = \gamma J_n \tag{15.26}$$

Specifically, we prove

Theorem 15.6 *Suppose that $\alpha > 1/p, 1 \le p, q \le \infty$ and that (ϕ, ψ) satisfy Assumption A. Let $\mathcal{E}$ be any one of the four coordinatewise estimator classes of Section 15.1, and let m_n be chosen according to* (15.25) *and* (15.26)*. Then as $n \to \infty$,*

$$\tilde{R}_{\mathcal{E}}(\Theta^{[m]}(C), \epsilon_n) \le R_{\mathcal{E}}(\Theta^{[m]}(C), \epsilon_n)(1 + o(1)).$$

We outline the argument, referring to the literature for full details. A couple of approaches have been used; in each the strategy is to begin with the sampled data model (15.2) and construct from $(\tilde{y}_l)$ a related set of wavelet coefficients $(\tilde{y}_I)$ which satisfy a (possibly correlated) sequence model

$$\tilde{y}_I = \tilde{\theta}_I + \epsilon^{(n)}\tilde{z}_I. \tag{15.27}$$

We then take an estimator $\hat{\theta}(y)$ known to be good in the (projected) white noise model and apply it with the sample data wavelet coefficients $\tilde{y} = (\tilde{y}_I)$ in place of y. The aim then is to show that the performance of $\hat{\theta}(\tilde{y})$ for appropriate Θ and noise level $\epsilon^{(n)}$ is nearly as good as that for $\hat{\theta}(y)$ at original noise level ϵ_n.

(i) Deslauriers-Dubuc interpolation. Define a fundamental function $\tilde{\phi}$ satisfying the interpolation property $\tilde{\phi}(l) = \delta_{l,0}$ and other conditions, and then corresponding scaling functions $\tilde{\phi}_l(t) = \tilde{\phi}(nt - l), l = 1, \ldots, n$. Interpolate the sampled function and data values by

$$\tilde{P}_n f(t) = \sum_{l=1}^{n} f(l/n)\tilde{\phi}_l(t), \qquad \tilde{y}^{(n)}(t) = \sum_{l=1}^{n} \tilde{y}_l\tilde{\phi}_l(t). \tag{15.28}$$

Let $\{\psi_I\}$ be an orthonormal wavelet basis as specified in Assumption A and $\tilde{\theta}_I = \langle \tilde{P}_n f, \psi_I\rangle$ Let $\epsilon^{(n)}$ be the largest standard deviation among the variates $\langle \tilde{y}^{(n)}, \psi_I\rangle$ for $j \le J_0$: it can be shown, in a manner similar to Lemma 15.5, that $\epsilon^{(n)} \sim \epsilon_n$ for n large. Now let $\tilde{y}_I = \langle \tilde{y}^{(n)}, \psi_I\rangle + n_I$, where the n_I are noise inflating Gaussian variates independent of $\tilde{y}^{(n)}$ chosen so that $\mathrm{Var}(\tilde{y}_I) \equiv [\epsilon^{(n)}]^2$. We thus obtain (15.27) though here the variates $\tilde{z}_I$ are in general correlated. This approach is set out in Donoho and Johnstone (1999). Although somewhat more complicated in the processing of the observed data $\tilde{y}_l$ it has the advantage of working for general families of wavelets and scaling functions.

(ii) Coiflets. If the wavelet basis $\{\psi_I\}$ is chosen from a family with sufficient vanishing moments for the scaling function ϕ, then we may work directly with $\tilde{y}_I$ (and $\tilde{\theta}_I$) derived from the discrete wavelet transform of the observations $\tilde{y}_l$ (and $\tilde{\theta}_l$). This approach is set out in Johnstone and Silverman (2004b). While somewhat simpler in the handling of the sampled data $\tilde{y}_l$, it is restricted to scaling functions with sufficient vanishing moments. It has the advantage that, in decomposition (15.27), the interior noise variates $\tilde{z}_I$ are an orthogonal

transformation of the the original noise $\tilde{z}_l$ and hence are independent with $\epsilon^{(n)} = \epsilon_n$. The boundary noise variates $\tilde{z}_I$ may be correlated, but there are at most cJ_0 of these, with uniformly bounded variances $\mathrm{Var}\,\tilde{z}_I \le c\epsilon_n^2$. So in the coiflet case, we could actually take $\mathcal{E}$ to be the class of *all* estimators (scalar or not).

We restrict attention to estimators vanishing for levels $j \ge J_{0n}$, where $2^{J_{0n}} = m = m_n$ is specified in (15.26). It is natural to decompose the error of estimation of θ in terms of $\tilde{\theta}$:

$$\|\hat{\theta}(\tilde{y}) - \theta\|_{2,m} \le \|\hat{\theta}(\tilde{y}) - \tilde{\theta}\|_{2,m} + \|\tilde{\theta} - \theta\|_{2,m}. \tag{15.29}$$

Concerning the second term on the right side, in either the Deslauriers-Dubuc or coiflet settings, one verifies that

$$\sup_{\Theta(C)} \|\tilde{\theta} - \theta\|_{2,m}^2 \le cC^2 2^{-2J_{0n}\alpha'}, \tag{15.30}$$

where $m = 2^{J_{0n}}$ and $\alpha' = \alpha - (1/p - 1/2)+$. [For Deslauriers-Dubuc, this is Lemma 4.1 in Donoho and Johnstone (1999), while for coiflets it follows from Proposition 5 as in the proof of Theorem 2 in Johnstone and Silverman (2004b)].

Turning to the first term on the right side of (15.29), the key remaining issue is to establish that if θ has bounded Besov norm, then the Besov norm of the interpolant coefficients $\tilde{\theta}$ *below level* J_{0n} is not much larger. To emphasise this, we write $P_m\tilde{\theta}$ for the vector whose $(j,k)-$th coefficient is $\tilde{\theta}_{jk}$ if $j < J_{0n}$ and 0 otherwise. The two references just cited show (Lemma 4.2 and Proposition 5 respectively) the existence of constants $\Delta_n = \Delta_n(\phi, \psi, \alpha, p, q) \to 0$ such that

$$\|P_m\tilde{\theta}\|_{b^\alpha_{p,q}} \le (1 + \Delta_n)\|\theta\|_{b^\alpha_{p,q}}. \tag{15.31}$$

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 $\hat{\theta}_n^*$ is asymptotically $\mathcal{E}-$ minimax over $\Theta(C_n)$ – note that we have chosen J_{0n} expressly so that this can be achieved with an estimator that vanishes for $j \ge J_{0n}$. Thus, since we only attempt to estimate the first m components of $\tilde{\theta}$,

$$\begin{aligned}\sup_{\theta\in\Theta(C)} E\|\hat{\theta}_n^*(\tilde{y}) - \tilde{\theta}\|_{2,m}^2 &\le \sup_{\tilde{\theta}\in\Theta(C_n)} E\|\hat{\theta}_n^*(\tilde{y}) - \tilde{\theta}\|_{2,m}^2 \\ &\le R_{\mathcal{E}}(C_n, \epsilon^{(n)})(1 + o(1)).\end{aligned}$$

Lemma 15.7 *If $\epsilon_1 \ge \epsilon_0$ and $C_1 \ge C_0$, then for any of the four estimator classes $\mathcal{E}$*

$$R_{\mathcal{E}}(C_1, \epsilon_1) \le (\epsilon_1/\epsilon_0)^2(C_1/C_0)^2 R_{\mathcal{E}}(C_0, \epsilon_0). \tag{15.32}$$

For the proof, see Donoho and Johnstone (1999). Combining (15.29), (15.30) and (15.32) with the tail bound of Lemma 15.3, we obtain

$$\begin{aligned}\sup_{\theta\in\Theta(C)} E\|\hat{\theta}_n^*(\tilde{y}) - \theta\|^2 &\le (\epsilon^{(n)}/\epsilon_n)^2(C_n/C)^2 R_{\mathcal{E}}(C, \epsilon_n)(1 + o(1)) \\ &= R_{\mathcal{E}}(C, \epsilon_n)(1 + o(1)),\end{aligned}$$

which establishes Theorem 15.2.

Remark. One can rephrase the bound (15.30) in a form useful in the next section. Indeed,

let $\tilde{P}_n f$ be given in the Deslauriers-Dubuc case by (15.28) and in the Coiflet case by $\tilde{P}_n f = n^{-1/2} \sum f(t_l)\phi_{Jl}$. Then the arguments referred to following (15.30) also show that

$$\sup_{\mathcal{F}(C)} \|\tilde{P}_n f - f\|^2 \le c C^2 n^{-2\alpha'} = o(n^{-r}). \tag{15.33}$$

15.5 Estimation in discrete norms

We will now show that the condition (15.33) in fact implies that the quality of estimation in continuous and discrete norms is in fact equivalent:

$$\tilde{R}(\mathcal{F}, n; L_2) \sim \tilde{R}(\mathcal{F}, n; \ell_{2,n}) = \inf_{\hat{f}(\tilde{y})} \sup_{f \in \mathcal{F}} n^{-1} \sum_l E[\hat{f}(t_l) - f(t_l)]^2. \tag{15.34}$$

(and similarly for R.) We describe this in the Coiflet case, but a similar result would be possible in the Deslauriers-Dubuc setting.

Given a continuous function $f \in L_2[0,1]$, we may consider two notions of sampling operator:

$$(S_\phi f)_l = \sqrt{n}\langle f, \phi_{Jl}\rangle, \qquad (S_\delta f)_l = f(t_l).$$

Let P_n denote projection onto $V_J = \text{span}\,\{\phi_{Jl}\}$, with $J = J_{0n}$ given by (15.26), and $\tilde{P}_n$ the "interpolation" operator, so that

$$P_n f = \sum_l \langle f, \phi_{Jl}\rangle \phi_{Jl}, \qquad \text{and} \qquad \tilde{P}_n g = \sum_l n^{-1/2} g(t_l)\phi_{Jl}.$$

From this we obtain Parseval identities like

$$\langle P_n f, \tilde{P}_n g\rangle_2 = \langle S_\phi f, S_\delta g\rangle_n$$

and

$$\|P_n \hat{f} - \tilde{P}_n f\|_2 = \|S_\phi \hat{f} - S_\delta f\|_n. \tag{15.35}$$

First suppose that $\tilde{f} = (\tilde{f}(t_l))$ is a good estimator for $\ell_{2,n}$ loss. Construct the interpolation $\hat{f}(t) = n^{-1/2} \sum_1^n \tilde{f}(t_l)\phi_{Jl}(t)$. From the decomposition

$$\hat{f} - f = \hat{f} - \tilde{P}_n f + \tilde{P}_n f - f$$

and the identity $\|\hat{f} - \tilde{P}_n f\|_2 = \|\tilde{f} - S_\delta f\|_n$, we obtain from (15.33)

$$\|\hat{f} - f\|_2 \le \|\tilde{f} - f\|_n + o(n^{-r/2})$$

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 $\hat{f}(t)$ is a good estimator for L_2 loss. Construct a discrete estimator $\tilde{f}$ using scaling function coefficients $\tilde{f}(t_l) = (S_\phi \hat{f})_l$. From the identity (15.35) and the decomposition

$$P_n \hat{f} - \tilde{P}_n f = P_n(\hat{f} - f) + P_n f - f + f - \tilde{P}_n f$$

we obtain first using (15.35), and then exploiting projection P_n, Lemma 15.3 and (15.33), that

$$\|\tilde{f} - S_\delta f\|_n \le \|\hat{f} - f\|_2 + o(n^{-r/2}).$$

Exercises

15.1 Show that

$$|\psi_I \psi_{I'}(s) - \psi_I \psi_{I'}(t)| \le c 2^{(j+j')/2} 2^{j \vee j'} |s - t|,$$

and that if $\|f\|_L = \sup |f(x) - f(y)|/|x - y|$, then

$$\left| n^{-1} f(t_l) - \int_{t_{l-1}}^{t_l} f \right| \le \tfrac{1}{2} n^{-2} \|f\|_L$$

and hence establish (15.24).

16

Epilogue

Brief mentions of topics of recent activity not discussed:

- Compressed sensing
- sparse non-orthogonal linear models
- covariance matrix estimation
- related non-Gaussian results

Appendix A

Appendix: The Minimax Theorem

The aim of this appendix is to give some justification for the minimax Theorem 4.12, 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 and 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 2.1)) 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).

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

$$\max_j \min_i A(i, j) \le \min_i \max_j A(i, j). \tag{A.1}$$

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

$$A(i_0, j) \le A(i_0, j_0) \le A(i, j_0) \qquad \text{for all } i, j.$$

However, saddlepoints do not exist in general, as is demonstrated already by the matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ The situation is rescued by allowing *mixed* or randomized strategies, which are probability distributions $x = (x(i))_1^m$ and $y = ((y(j))_1^n$ 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

$$f(x, y) = x^T A y = \sum_{i,j} x(i) A(i, j) y(j). \tag{A.2}$$

Write S_m for the simplex of probability vectors $\{x \in \mathbb{R}^n : x_i \ge 0, \sum x_i = 1\}$. The classical minimax theorem of von Neumann states that for an arbitrary $m \times n$ matrix A in (A.2),

$$\min_{x \in S_m} \max_{y \in S_n} f(x, y) = \max_{y \in S_n} \min_{x \in S_m} f(x, y). \tag{A.3}$$

For the payoff matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, it is easily verified that the fair coin tossing strategies $x = y = (\frac{1}{2}\ \frac{1}{2})$ yield a saddlepoint.

We establish below a more general result that implies (A.3).

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-concavity 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 more recent survey is given by Simons (1995).

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) and Peck and Dulmage (1957) observed that the method extends directly to convex-concave f. In addition, it will be useful to allow f to take values in $\mathbb{R} \cup +\infty$. First recall that a function $f : X \to \mathbb{R} \cup \infty$ on a topological space X is *lower semicontinuous* (lsc) iff $\{x : f(x) > t\}$ is open for all t, or equivalently if $\{x : f(x) \le t\}$ is closed for all t. [If X is 1st countable, then these conditions may be rewritten in terms of sequences as $f(x) \le \liminf f(x_n)$ 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 $x_0 \in X$.

Theorem A.1 (Kneser, Kuhn) *Let K, L be convex subsets of real vector spaces and $f : K \times L \to \mathbb{R} \cup \infty$ 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*

$$\inf_{x \in K} \sup_{y \in L} f(x, y) = \sup_{y \in L} \inf_{x \in K} f(x, y). \tag{A.4}$$

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.

Note that if $x \to f(x, y)$ is lower semi-continuous for all $y \in L$, then $x \to \sup_{y \in L} f(x, y)$ is also lower semi-continuous and so the infimum on the left side of (A.4) is attained for some $x_0 \in K$.

Here is an example where f is *not* continuous, and only the semicontinuity condition of the theorem holds. Let $\mathbb{R}^\infty$ denote the space of sequences: a countable product of $\mathbb{R}$ with the product topology: $x^{(n)} \to x$ iff 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 $y \in L = \{y : 0 \leq y_i \leq C \text{ for all } i\}$. Equality (A.4) can easily be checked directly. However, 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 semicontinuity, so we present it here through a couple of lemmas.

Lemma A.2 *Let $f_1, \ldots f_n : K \to \mathbb{R} \cup \infty$ be convex lsc 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 $\sigma \in S_n$,*

$$\sum_1^n \sigma_i f_i(x) > 0 \qquad \textit{for all } x \in K.$$

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 (checking this is a useful exercise). So, with a slight change of notation, assume for all x that $\max\{f(x), g(x)\} > 0$. The sets $M = \{x : f(x) \leq 0\}$ and $N = \{x : g(x) \leq 0\}$ are nonempty (else there is nothing left to prove) and closed (by lower semicontinuity), and hence compact. On M and N respectively, we have $-f/g$ and $-g/f$ defined and usc, and

$$\begin{aligned} &\text{on } M, \quad g > 0 \quad \Rightarrow \quad \frac{-f}{g} \leq \frac{-f}{g}(p) = \alpha \geq 0 \quad \Rightarrow \quad f + \alpha g \geq 0 \\ &\text{on } N, \quad f > 0 \quad \Rightarrow \quad \frac{-g}{f} \leq \frac{-g}{f}(q) = \beta \geq 0 \quad \Rightarrow \quad \beta f + g \geq 0. \end{aligned} \tag{A.5}$$

for some $p \in M, q \in N$ and finite α, β. We seek $\sigma \in [0, 1]$ such that $\sigma f + \bar{\sigma} g > 0$ on K (here $\bar{\sigma} = 1 - \sigma$). The strategy is to show that $\alpha\beta < 1$ and then that α and β can be suitably increased.

First, note that $f(q) = \infty$ implies $\beta = 0$ and that $g(p) = \infty$ forces $\alpha = 0$, so in either case $\alpha\beta < 1$ holds trivially. So we may assume that both $f(q)$ and $g(p)$ are finite. Since $f(p) \leq 0$ and $f(q) > 0$, there exists $\eta > 0$ such that

$$0 = \eta f(p) + \bar{\eta} f(q) \geq f(p_\eta)$$

by convexity, with $p_\eta = \eta p + \bar{\eta} q$ and $\bar{\eta} = 1 - \eta$. But then $\max\{f, g\} > 0$ means that

$$0 < g(p_\eta) \leq \eta g(p) + \bar{\eta} g(q),$$

again using convexity. Using both these displays, along with the definitions of p and q,

$$\eta g(p) > -\bar{\eta} g(q) = \bar{\eta}\beta f(q) = -\eta\beta f(p) = \alpha\beta\eta g(p).$$

Since $\eta > 0$ and $g(p) > 0$, we conclude that $\alpha\beta < 1$, and so in (A.5), we may increase α up to γ and β up to δ in such a way that

$$\gamma\delta = 1, \qquad f + \gamma g > 0 \qquad \text{and} \qquad \delta f + g > 0 \tag{A.6}$$

on M and N respectively. Now $\gamma\delta = 1$ means that

$$\sigma = \frac{1}{1+\gamma} = \frac{\delta}{1+\delta},$$

so on dividing the inequalities in (A.6) by $1+\gamma$ and $1+\delta$ respectively, we get $\sigma f + \bar{\sigma} g > 0$ on $M \cup N$. On $(M \cup N)^c$ we have both $f > 0$ and $g > 0$, so we are done. □

Lemma A.3 *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, $\sum \sigma_i f(x, y_i) > 0$. By concavity, at $y^* = \sum_1^n \sigma_i y_i$, 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. □

Proof of Theorem A.1 That the right side of (A.4) 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,

$$\sup_y \inf_x f \leq c < \inf_x \sup_y f. \tag{A.7}$$

Replacing f by $f - c$ does not harm any of the hypotheses, so we may assume that $c = 0$. The left inequality in (A.7) 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.7)! Hence there must be equality in (A.7). □

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*.

Corollary A.4 *Let K, L be convex subsets of real vector spaces and $f : K \times L \to \mathbb{R} \cup \infty$ 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*

$$\inf_{x \in K} \sup_{y \in L} f(x,y) = \sup_{y \in L} \inf_{x \in K} f(x,y). \tag{A.8}$$

Proof Apply Theorem A.1 to $\bar{f}(y,x) = -f(x,y)$. □

A statistical minimax theorem

First, we state the Gaussian sequence model (4.1) 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. [Terminology from point-set topology here and below may be found in analysis texts, e.g. Folland (1999), or the appendix of Bogachev (1998)]. The space $\mathcal{X}$ is endowed with the Borel $\sigma-$field, and as dominating measure, we take P_0, the centered Gaussian Radon measure (see Bogachev (1998, Example 2.3.6)) defined as the (countable) product of $N(0, \varrho_i^2)$ measures on $\mathbb{R}$. For each $\theta \in \Theta = \ell_{2,\varrho}$, with inner product $\langle \theta, \theta' \rangle_\varrho = \sum \theta_i \theta_i' / \varrho_i^2$, the measure P_θ with mean θ is absolutely continuous (indeed equivalent) to P_0, and has density $f_\theta(x) = dP_\theta/dP_0 = \exp\{\langle \theta, x\rangle_\varrho - \|\theta\|_\varrho^2/2\}$. 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 $\mathcal{A} = (\bar{\mathbb{R}})^\infty$ which with the product topology is compact, 2° 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: $L : \mathcal{A} \times \Theta \to [0, \infty]$.

Theorem A.5 *For the above Gaussian sequence model, we assume (i) that for each θ, the map $a \to L(a, \theta)$ 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 $\ell_{2,\varrho}$. Then*

$$\inf_{\hat\theta} \sup_{\pi \in \mathcal{P}} B(\hat\theta, \pi) = \sup_{\pi \in \mathcal{P}} \inf_{\hat\theta} B(\hat\theta, \pi). \tag{A.9}$$

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 $a_i^{(n)} \to a_i^{(\infty)}$ for each i, then for each fixed m, one has

$$\sum_{i=1}^m |a_i^{(\infty)} - \theta_i|^p = \lim_n \sum_{i=1}^m |a_i^{(n)} - \theta_i|^p \le \liminf_n \|a^{(n)} - \theta\|_p^p.$$

A general framework for statistical decision theory, including minimax and complete class results, has been developed by its chief exponents, including A. Wald, L. Le Cam, C. Stein, and L. Brown, in published and unpublished works. A selection of references includes Wald (1950); Le Cam (1955, 1986); Diaconis and 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 nonparametric 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 general method in the concrete setting of the nonparametric Gaussian sequence model.

Remark. There is a special case (which includes the setting of a bounded normal mean, Section 4.6), in which our statistical minimax theorem can be derived directly from the Kneser-Kuhn theorem. Indeed, if $\Theta \subset \mathbb{R}^n$ is compact, and $\mathcal{P} = \mathcal{P}(\Theta)$, then $\mathcal{P}$ is compact for weak convergence of probability measures. Let K be the class of estimators $\hat{\theta}$ with finite risk functions on Θ, let $L = \mathcal{P}$ and for the payoff function f take $B(\hat{\theta}, \pi) = \int_\Theta r(\hat{\theta}, \theta)\pi(d\theta)$.

Observe that K is convex because $a \to L(a, \theta)$ is; that L is convex *and* compact; and that B is convex-linear. Finally $\pi \to B(\hat{\theta}, \pi)$ is continuous since in the Gaussian model $y_i = \theta_i + \epsilon\lambda_i z_i$, the risk functions $\theta \to r(\hat{\theta}, \theta)$ are continuous and bounded on the compact set Θ. Hence the Kneser-Kuhn Corollary A.4 applies to provide the minimax result.

Randomized decision rules.

The payoff function $B(\hat{\theta}, \pi)$ appearing in Theorem A.5 is linear in π, 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 a given that x is observed, and (ii), for each measurable A, the map $x \to \delta(A|x)$ is measurable. The risk function of a randomized rule δ is

$$r(\delta, \theta) = \int\int L(a, \theta)\delta(da|x)P_\theta(dx), \tag{A.10}$$

and the payoff function we consider is the integrated risk against a probability measure π:

$$B(\delta, \pi) = \int r(\delta, \theta)\pi(d\theta).$$

A major reason for introducing $B(\delta, \pi)$ is that it is bilinear in δ and π. 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

$$\inf_{\delta\in\mathcal{D}} \sup_{\pi\in\mathcal{P}} B(\delta, \pi) = \sup_{\pi\in\mathcal{P}} \inf_{\delta\in\mathcal{D}} B(\delta, \pi), \tag{A.11}$$

Kneser's theorem suggests that we need a topology on decision rules δ with two key properties:

(P1) $\mathcal{D}$ is compact, and
(P2) the risk functions $\delta \to B(\delta, \pi)$ are lower semicontinuous.

Before describing how this is done, we explain how (A.9) follows from (A.11) using the convexity assumption on the loss function. Indeed, given a randomized rule δ, the standard method is to construct a *non-randomized rule* by averaging: $\hat{\theta}_\delta(x) = \int a\delta(da|x)$. Convexity of $a \to L(a, \theta)$ and Jensen's inequality then imply that

$$L(\hat{\theta}_\delta(x), \theta) \le \int L(a, \theta)\delta(da|x).$$

Averaging over $X \sim P_\theta$, and recalling (A.10) shows that $\hat{\theta}_\delta$ is at least as good as δ:

$$r(\hat{\theta}_\delta, \theta) \le r(\delta, \theta) \qquad \text{for all } \theta \in \Theta. \tag{A.12}$$

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 π yields

$$\sup_{\pi} B(\hat{\theta}_\delta, \pi) \leq \sup_{\pi} B(\delta, \pi). \tag{A.13}$$

We then recover (A.9) from (A.11) via a simple chain of inequalities:

$$\begin{aligned}\inf_{\hat{\theta}} \sup_{\pi} B(\hat{\theta}, \pi) &\leq \inf_{\hat{\theta}_\delta} \sup_{\pi} B(\hat{\theta}_\delta, \pi) \leq \inf_{\delta} \sup_{\pi} B(\delta, \pi) \\ = \sup_{\pi} \inf_{\delta} B(\delta, \pi) &\leq \sup_{\pi} \inf_{\hat{\theta}} B(\hat{\theta}, \pi) \leq \inf_{\hat{\theta}} \sup_{\pi} B(\hat{\theta}, \pi),\end{aligned}$$

and since the first and last terms are the same, all terms are equal.

A compact topology for $\mathcal{D}$

We return to establishing properties [P1] and [P2]. The approach of Brown and Le Cam is to identify decision rules δ with bilinear, bicontinuous functionals, and then use the Alaoglu theorem (e.g. Rudin (1973, p. 66)) on weak compactness to induce a topology on $\mathcal{D}$.

For this section, we write $L_\theta(a)$ for the loss function to emphasise the dependence on a. The risk function of a rule δ may then be written

$$r(\delta, \theta) = \int\int L_\theta(a) f_\theta(x) \delta(da|x) P_0(dx) = b_\delta(f_\theta, L_\theta),$$

Here the probability density f_θ is regarded as a non-negative function in the Banach space $L_1 = L_1(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), P_0)$ which is separable as noted earlier. Since $\mathcal{A} = (\bar{\mathbb{R}})^\infty$ 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 $\iota : \delta \to b_\delta$ is a bijection of $\mathcal{D}$ onto

$$\begin{aligned}B_1^+ &= \{b \in B : b \geq 0 \text{ and } b(g, 1) = \|g\|_{L_1} \forall g \geq 0\} \\ &\subset \{b : \|b\|_B \leq 1\},\end{aligned}$$

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 ι 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 $\delta_i \to \delta$ means that

$$b_{\delta_i}(g, c) \to b_\delta(g, c) \qquad \forall (g, c) \in L_1 \times C. \tag{A.14}$$

This topology also satisfies our second requirement: that the maps $\delta \to B(\delta, \pi)$ 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

$$r(\delta, \theta) = \sup_c \{b_\delta(f_\theta, c) : c \leq L_\theta\}.$$

The definition (A.14) says that the maps $\delta \to b_\delta(f_\theta, c)$ 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.

Conclusion

We have now outlined how $B(\delta, \pi)$ may be viewed as a bilinear function on $\mathcal{D} \times \mathcal{P}$ taking values in $[0, \infty]$ which for each fixed π is lsc on the compact $\mathcal{D}$. Kneser's Theorem A.1 now gives us (A.11), which implies, as we have seen, the conclusion (A.9) of Theorem A.5.

A separation theorem

We have now established $B(\delta, \pi)$ as a bilinear function on $\mathcal{D} \times \mathcal{P}$ which for each π 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 Tychonoff's theorem. Now define an upper envelope of the risk functions by setting $\Gamma = \rho(\mathcal{D})$ and then defining

$$\tilde{\Gamma} = \{b \in T : \text{there exists } b' \in \Gamma \text{ with } b' \leq b\}.$$

Brown uses the $\mathcal{D}$ topology constructed above, along with the compactness and lower semicontinuity properties [P1] and [P2] to show that $\tilde{\Gamma}$ is closed and hence compact in T.

Using the separating hyperplane theorem for Euclidean spaces – a consequence of Lemma A.2 – Brown shows

Theorem A.6 *Suppose that $\tilde{\Gamma}$ is convex and closed in T and that $b_0 \in T \backslash \tilde{\Gamma}$. Then there exists $c > 0$, a finite set $(\pi_i)_1^m \subset \mathcal{P}$ and a probability vector $(\xi_i)_1^m$ such that the convex combination $\pi_\xi = \sum \xi_i \pi_i \in \mathcal{P}$ satisfies*

$$b_0(\pi_\xi) < c < b(\pi_\xi) \qquad \text{for all } b \in \tilde{\Gamma}. \tag{A.15}$$

It is now easy to derive the minimax conclusion (A.11). Indeed, write $V = \inf_\delta \sup_\mathcal{P} B(\delta, \pi)$. If $V < \infty$, let $\epsilon > 0$ and choose $b_0 \equiv V - \epsilon$ – clearly $b_0 \notin \tilde{\Gamma}$. Convexity of $\mathcal{D}$ entails convexity of $\tilde{\Gamma}$, 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, $\sup_\pi \inf_\delta B(\delta,\pi) > V - \epsilon$ 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.11).

A.1 A special minimax theorem for thresholding

It is sometimes of interest to restrict the estimator δ in $B(\delta,\pi)$ to a smaller class, for example threshold rules that depend on a single parameter, the threshold λ. We write $B(\lambda,\pi)$ for the payoff function in such cases (for details, see Section 13.4).

In such cases $\lambda \to B(\lambda,\pi)$ need not be convex and so our earlier minimax theorems do not directly apply. In addition, we would like to exhibit a saddle point. In this section, then, we formulate and prove a special minimax theorem tailored to this setting. First, a definition. We call a function $\lambda(\pi)$ defined for π in a convex set $\mathcal{P}$ *Gâteaux continuous* at π_0 if $\lambda\big((1-t)\pi_0 + t\pi_1\big) \to \lambda(\pi_0)$ as $t \to 0$ for each $\pi_1 \in \mathcal{P}$.

Theorem A.7 *Suppose $\Lambda \subset \mathbb{R}$ is an interval and that $\mathcal{P}$ is convex and compact. Suppose that $B : \Lambda \times \mathcal{P} \to \mathbb{R}$ is linear and continuous in π for each $\lambda \in \Lambda$. Then there exists a least favorable π_0.*

Suppose also for each π that $B(\lambda,\pi)$ is continuous in λ, that there is a unique $\lambda(\pi)$ that minimizes B, and that $\lambda(\pi)$ is Gâteaux continuous at π_0. Set $\lambda_0 = \lambda(\pi_0)$.

Then the pair (λ_0,π_0) is a saddlepoint: for all $\lambda \in [0,\infty)$ and $\pi \in \mathcal{P}$,

$$B(\lambda_0,\pi) \le B(\lambda_0,\pi_0) \le B(\lambda,\pi_0), \tag{A.16}$$

and hence

$$\inf_\lambda \sup_{\mathcal{P}} B(\lambda,\pi) = \sup_{\mathcal{P}} \inf_\lambda B(\lambda,\pi) = \sup_{\mathcal{P}} B_S(\pi).$$

Proof First, the least favorable distribution π_0 exists because $\inf_\lambda B(\lambda,\pi)$ is usc on the compact set $\mathcal{P}$. The right side of (A.16) follows from the definition of $\lambda(\pi_0)$. For the left side, given an arbitrary $\pi_1 \in \mathcal{P}$, define $\pi_t = (1-t)\pi_0 + t\pi_1$ for $t \in [0,1]$: by convexity, $\pi_t \in \mathcal{P}$. Let $\lambda_t = \lambda(\pi_t)$ be the best threshold for π_t, so that $B(\pi_t) = B(\lambda_t,\pi_t)$. Heuristically, since π_0 is least favorable, we have $(d/dt)B(\pi_t)|_{t=0} \le 0$, and we want to compute partial derivatives of $B(\lambda_t,\pi_t)$ and then exploit linearity in π.

More formally, for $t > 0$ we have

$$B(\lambda_t,\pi_t) - B(\lambda_0,\pi_0) = B(\lambda_t,\pi_0) - B(\lambda_0,\pi_0) + B(\lambda_0,\pi_t) - B(\lambda_0,\pi_0) + \Delta^2 B,$$

where the left side is ≤ 0 and

$$\Delta^2 B = B(\lambda_t,\pi_t) - B(\lambda_t,\pi_0) - B(\lambda_0,\pi_t) + B(\lambda_0,\pi_0).$$

Now also $B(\lambda_t,\pi_0) \ge B(\lambda_0,\pi_0)$ and by linearity $B(\lambda_0,\pi_t) - B(\lambda_0,\pi_0) = t[B(\lambda_0,\pi_1) - B(\lambda_0,\pi_0)]$ and so

$$0 \ge B(\lambda_0,\pi_1) - B(\lambda_0,\pi_0) + \Delta^2 B/t.$$

Again using the linearity in π,

$$\Delta^2 B/t = [B(\lambda_t,\pi_1) - B(\lambda_0,\pi_1)] - [B(\lambda_t,\pi_0) - B(\lambda_0,\pi_0)] \to 0$$

as $t \to 0$, since $\lambda_t \to \lambda_0$ by Gâteaux continuity of $\lambda(\pi)$, and since $\lambda \to B(\lambda, \pi)$ is continuous. This shows that $B(\lambda_0, \pi_1) \leq B(\lambda_0, \pi_0)$ for any $\pi_1 \in \mathcal{P}$ and completes the proof. □

Remark. Proposition 13.11 shows that $B(\lambda, \pi)$ is quasi-convex in λ, and since it is also linear in π on a convex set, one could appeal to a general minimax theorem, e. g. Sion (1958). However, the general minimax theorems do not exhibit a saddlepoint, which emerges directly from the present more specialized approach.

Exercise. Complete the induction step for the proof of Lemma A.2.

Appendix B

More on Wavelets and Function Spaces

B.1 Building scaling functions and wavelets

This section supplements the account of Section 7.1 by giving more detail on the construction and properties of orthonormal scaling functions φ and wavelets ψ. This is still only a partial account, and we refer to the original sources, as well as Mallat (2009), abbreviated below as [M], where the statements and proofs of Lemmas/Theorems B.1 - B.7 are given.

We sketch two common constructions of a scaling function φ and later the corresponding wavelet ψ: (a) beginning from a Riesz basis, and (b) starting from discrete (especially finite) filters.

(a) Using a Riesz basis. A family $\{e_k\}_{k\in\mathbb{N}}$ is a *Riesz basis* for a Hilbert space H if (i) for all $h \in H$, there is a unique representation $h = \sum \alpha_k e_k$, and (ii) there exist positive absolute constants C_1, C_2 such that for all $h \in H$, $C_1\|h\|^2 \le \sum_k |\alpha_k|^2 \le C_2\|h\|^2$.

It is more common to replace the multiresolution analysis condition (iv) in Definition 7.1 by the weaker condition

(iv') $\quad \exists\, \theta \in V_0$ such that $\{\theta(x-k) : k \in \mathbb{Z}\}$ is a Riesz basis for V_0.[1]

That (iv') is equivalent to (iv) follows from the "orthonormalization trick" discussed below.

A key role in constructions and interpretations is played by the frequency domain and the Fourier transform (C.10). The Plancherel identity (C.12) leads to a frequency domain characterization of the orthnormality and Riesz basis conditions (iv) and (iv'):

Lemma B.1 *Suppose $\varphi \in L_2$. The set $\{\varphi(x-k), k \in \mathbb{Z}\}$ is (i) orthonormal if and only if*

$$\sum_k |\widehat{\varphi}(\xi + 2k\pi)|^2 = 1 \qquad a.e., \tag{B.1}$$

and (ii) a Riesz basis if and only if there exist positive constants C_1, C_2 such that

$$C_1 \le \sum_k |\widehat{\varphi}(\xi + 2k\pi)|^2 \le C_2 \qquad a.e. \tag{B.2}$$

Partial Proof. We give the easy proof of (B.1) since it gives a hint of the role of frequency domain methods. The Fourier transform of $x \to \varphi(x-n)$ is $e^{-in\xi}\widehat{\varphi}(\xi)$. Thus, orthonormality combined with the Plancherel identity gives

$$\delta_{0n} = \int_{-\infty}^{\infty} \varphi(x)\overline{\varphi(x-n)}dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{in\xi}|\widehat{\varphi}(\xi)|^2 d\xi.$$

[1] This use of the θ symbol, local to this appendix, should not be confused with the notation for wavelet coefficients in the main text.

Now partition $\mathbb{R}$ into segments of length 2π, add the integrals, and exploit periodicity of $e^{in\xi}$ to rewrite the right hand side as

$$\frac{1}{2\pi}\int_0^{2\pi} e^{in\xi}\sum_k |\widehat{\varphi}(\xi+2k\pi)|^2 d\xi = \delta_{0n}.$$

The function in (B.1) has as Fourier coefficients the delta sequence $2\pi\delta_{0n}$ and so equals 1 a.e. For part (ii), see e.g. [M, Theorem 3.4]. □

The "orthonormalization trick" creates (B.1) by fiat:

Theorem B.2 *Suppose that $\{V_j\}$ is an MRA, and that $\{\theta(x-k), k\in\mathbb{Z}\}$ is a Riesz basis for V_0. Define*

$$\widehat{\varphi}(\xi) = \widehat{\theta}(\xi)\Big/\Big\{\sum_k |\widehat{\theta}(\xi+2k\pi)|^2\Big\}^{1/2}. \tag{B.3}$$

Then φ is a scaling function for the MRA, and so for all $j\in\mathbb{Z}$, $\{\varphi_{jk} : k\in\mathbb{Z}\}$ is an orthonormal basis for V_j.

Example. Box spline MRA. (See also Chapter 7.1.) Given $r\in\mathbb{N}$, set $\chi = I_{[0,1]}$ and $\theta = \theta_r = \chi\star\cdots\star\chi = \chi^{\star(r+1)}$. Without any loss of generality, we may shift $\theta_r = \chi^{\star(r+1)}$ 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), [M, Sec. 7.1] that

$$\widehat{\theta}_r(\xi) = \left(\frac{\sin\xi/2}{\xi/2}\right)^{r+1} e^{-i\epsilon\xi/2} \qquad \epsilon = \begin{cases} 1 & r \text{ even} \\ 0 & r \text{ odd} \end{cases},$$

$$\sum_k |\widehat{\theta}_r(\xi+2k\pi)|^2 = P_{2r}(\cos\xi/2),$$

where P_{2r} is a polynomial of degree $2r$. For example, in the piecewise linear case $r=1$, $P_2(v) = (1/3)(1+2v^2)$. Using (B.2), this establishes the Riesz basis condition (iv') for this MRA. Thus (B.3) gives an explicit Fourier domain expression for φ which is amenable to numerical calculation. [M, pp 266-268] gives corresponding formulas and pictures for cubic splines.

(b) Using finite filters. The MRA conditions imply important structural constraints on $\widehat{h}(\xi)$: using (B.1) and (7.2) it can be shown that

Lemma B.3 *If φ is an integrable scaling function for an MRA, then*

$$(CMF) \qquad |\widehat{h}(\xi)|^2 + |\widehat{h}(\xi+\pi)|^2 = 2 \qquad \forall\xi\in\mathbb{R} \tag{B.4}$$

$$(NORM) \qquad \widehat{h}(0) = \sqrt{2}. \tag{B.5}$$

(B.4) is called the *conjugate mirror filter* (CMF) condition, while (B.5) is a normalization requirement. Conditions (B.5) and (B.4) respectively imply constraints on the discrete filters:

$$\sum h_k = \sqrt{2}, \qquad \sum h_k^2 = 1.$$

They are the starting point for a unified construction of many of the important wavelet families (Daubechies variants, Meyer, etc.) that begins with the filter $\{h[k]\}$, or equivalently $\widehat{h}(\xi)$. Here is a key result in this construction.

Theorem B.4 *If $\widehat{h}(\xi)$ is $2\pi-$periodic, C^1 near $\xi = 0$ and (a) satisfies* (B.4) *and* (B.5)*, and (b)* $\inf_{[-\pi/2,\pi/2]} |\widehat{h}(\xi)| > 0$*, then*

$$\widehat{\varphi}(\xi) = \prod_{l=1}^{\infty} \frac{\widehat{h}(2^{-l}\xi)}{\sqrt{2}} \tag{B.6}$$

is the Fourier transform of a scaling function $\varphi \in L_2$ that generates an MRA.

That $\widehat{\varphi}$ is generated by an infinite product might be guessed by iteration of the two scale relation (7.2): the work lies in establishing that all MRA properties hold. Condition (b) can be weakened to a necessary and sufficient condition due to Cohen (1990) (see also Cohen and Ryan (1995)).

Building wavelets. The next lemma gives the conditions on g in order that ψ be an orthonormal wavelet, analogous to Lemma B.3. Let z^* denote complex conjugate of z.

Lemma B.5 *[Mallat Lemma 7.1] $\{\psi_{jk}, k \in \mathbb{Z}\}$ is an orthonormal basis for W_j, the orthocomplement of V_j in V_{j+1} if and only if, for all $\xi \in \mathbb{R}$,*

$$|\widehat{g}(\xi)|^2 + |\widehat{g}(\xi + \pi)|^2 = 2 \tag{B.7}$$

$$\widehat{g}(\xi)\widehat{h}^*(\xi) + \widehat{g}(\xi + \pi)\widehat{h}^*(\xi + \pi) = 0. \tag{B.8}$$

One way to satisfy (B.7) and (B.8) is to set

$$\widehat{g}(\xi) = e^{-i\xi}\widehat{h}^*(\xi + \pi), \tag{B.9}$$

and use (B.4). To understand this in the time domain, note that if $\widehat{s}(\xi)$ has (real) coefficients s_k, then conjugation corresponds to time reversal: $\widehat{s}^*(\xi) \leftrightarrow s_{-k}$, while modulation corresponds to time shift: $e^{i\xi}\widehat{s}(\xi) \leftrightarrow s_{k+1}$, and the frequency shift by π goes over to time domain modulation: $\widehat{s}(\xi + \pi) \leftrightarrow (-1)^k s_k$. To summarize, interpreting (B.9) in terms of filter coefficients, one obtains the "mirror" relation

$$g_k = (-1)^{1-k} h_{1-k}. \tag{B.10}$$

Together, (7.4) and (B.9) provide a frequency domain recipe for constructing a candidate wavelet from φ:

$$\widehat{\psi}(2\xi) = 2^{-1/2} e^{-i\xi}\widehat{h}^*(\xi + \pi)\widehat{\varphi}(\xi). \tag{B.11}$$

Of course, there is still work to do to show that this does the job:

Theorem B.6 *If g is defined by* (B.9)*, and ψ by* (7.4)*, then $\{\psi_{jk}, (j,k) \in \mathbb{Z}^2\}$ is an orthonormal basis for $L_2(\mathbb{R})$.*

We illustrate by discussing some of the examples given in Section 7.1.

Example. Box splines again. Given $\widehat{\varphi}$, one constructs $\widehat{h}$ from (7.2), $\widehat{g}$ from (B.9) and $\widehat{\psi}$ from (7.4). 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. See Figure 7.2 for $r = 1$ and [M, p. 281] for $r = 3$.

Example. The class of *Meyer wavelets* (Meyer, 1986) is built from a filter $\widehat{h}(\xi)$ on $[-\pi, \pi]$ satisfying

$$\widehat{h}(\xi) = \begin{cases} \sqrt{2} & |\xi| \le \pi/3 \\ 0 & |\xi| \ge 2\pi/3, \end{cases}$$

the CMF condition (B.4), and that is also required to be C^n at the join points $\pm\pi/3$ and $\pm 2\pi/3$. In fact C^∞ 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, p137-8) and shown in Figure 7.2.

The scaling function $\widehat{\varphi}(\xi) = \prod_1^\infty 2^{-1/2}\widehat{h}(2^{-j}\xi)$ then has support in $[-4\pi/3, 4\pi/3]$, and the corresponding wavelet (defined from (7.4) and (B.9)) has support in the interval $\pm[2\pi/3, 8\pi/3]$. Since $\widehat{\varphi}$ and $\widehat{\psi}$ have compact support, both $\varphi(x)$ and $\psi(x)$ are C^∞ – unlike, say, Daubechies wavelets. However, they cannot have exponential decay in the time domain (which is impossible for C^∞ orthogonal wavelets, according to Daubechies (1992, Corollary 5.5.3)) – at least they are $O(|x|^{-n-1})$ if $\widehat{h}$ is C^n. Finally, since $\widehat{\psi}$ vanishes in a neighborhood of the origin, all its derivatives are zero at 0 and so ψ has an infinite number of vanishing moments.

Figure B.1 shows a schematic of the qualitative frequency domain properties of the squared modulus of $\widehat{\varphi}, \widehat{h}, \widehat{g}$ and finally $\widehat{\psi}$. It can be seen that the space V_0 generated by translates of φ 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 and Weiss, 1996, p.332 and p.61) that φ and the dilations of ψ form a partition of frequency space in the sense that

$$|\widehat{\varphi}(\xi)|^2 + \sum_{j=0}^{\infty} |\widehat{\psi}(2^{-j}\xi)|^2 = 1 \qquad \text{a.e.} \tag{B.12}$$

Vanishing moments. The condition that ψ have r vanishing moments has equivalent formulations in terms of the Fourier transform of ψ and the filter h.

Lemma B.7 *Let ψ be an orthonormal wavelet. If $\widehat{\psi}$ is C^p at $\xi = 0$, then the following are equivalent:*

$$\begin{aligned} &(i) && \int t^j \psi = 0, && j = 0, \dots, p-1. \\ &(ii) && D^j\widehat{\psi}(0) = 0, && j = 0, \dots, p-1. \\ &(iii) && D^j\widehat{h}(\pi) = 0 && j = 0, \dots, p-1. \quad (VM_p) \end{aligned} \tag{B.13}$$

See for example Mallat (1999, Theorem 7.4) or Härdle et al. (1998, Theorem 8.3).

Example. Daubechies wavelets. 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 (2009, Ch. 7) and Meyer (1990, Vol I, Ch. 3)).

The approach is to build a filter $h = \{h_k\}_0^{N-1}$ with $h_k \in \mathbb{R}$ and transfer function $\widehat{h}(\xi) =$

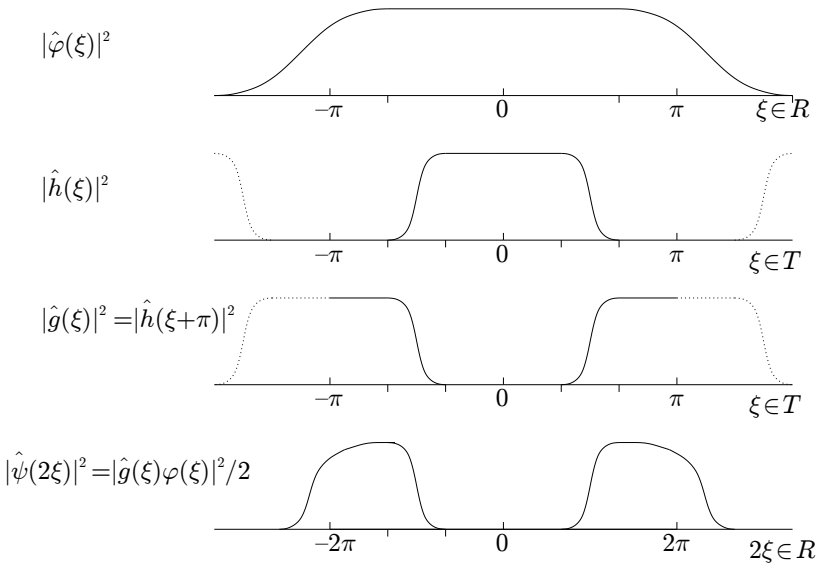

Figure B.1 qualitative frequency domain properties of scaling function $\hat{\varphi}$, transfer functions $\hat{h}, \hat{g}$ and wavelet $\hat{\psi}$ corresponding to the Meyer wavelet; dotted lines show extension by periodicity

$\sum_{k=0}^{N-1} h_k e^{-ik\xi}$ satisfying the conditions of Theorem B.4 and then derive the conjugate filter g and the wavelet ψ from (B.9), (B.11) and Theorem B.6. The vanishing moment condition of order p (VM_p) implies that $\widehat{h}(\xi)$ may be written

$$\widehat{h}(\xi) = \Big(\frac{1+e^{-i\xi}}{2}\Big)^p r(\xi), \qquad r(\xi) = \sum_{0}^{m} r_k e^{-ik\xi},$$

with $N = p + m + 1$ and $r_k \in \mathbb{R}$. Passing to squared moduli, one may write [2]

$$|\widehat{h}(\xi)|^2 = 2(\cos^2 \tfrac{\xi}{2})^p \, P(\sin^2 \tfrac{\xi}{2})$$

for some real polynomial P of degree m. The conjugate mirror filter condition (B.4) then forces, on putting $y = \sin^2 \xi/2$,

$$(1-y)^p P(y) + y^p P(1-y) = 1 \qquad 0 \le y \le 1. \tag{B.14}$$

To have the support length N as small as possible, one seeks solutions of (B.14) of minimal degree m. One solution can be described probabilistically in terms of repeated independent

[2] if $r(\xi) = \sum_0^m r_k e^{-ik\xi}$, with $r_k \in \mathbb{R}$, then $|r(\xi)|^2 = r(\xi)r^*(\xi) = r(\xi)r(-\xi) = \sum_{-m}^m s_k e^{-ik\xi}$ 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, $|(1+e^{-i\xi})/2|^2 = \cos^2(\xi/2)$.

tosses of a coin with Pr(Heads) $= y$. Either p tails occur before p heads or vice versa, so

$$P(y) := \Pr\{p\ T\text{s occur before } p\ H\text{s }\}/(1-y)^p$$
$$= \sum_{k=0}^{p-1} \binom{p+k-1}{k} y^k$$

certainly solves (B.14). Further, it is the *unique* solution of degree $p-1$ or less [3]

To return from the squared modulus scale, appeal to the F. Riesz lemma: if $s(\xi) = \sum_{-m}^{m} s_k e^{-ik\xi} \geq 0$, then there exists $r(\xi) = \sum_0^m r_k e^{-ik\xi}$ 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.

Discussion. Table B.1 sets out some desiderata for a wavelet basis. The last three requirements are in a sense mutually contradictory: it turns out that higher regularity of ψ can only be achieved with longer filters. One advantage of Daubechies' *family* of wavelets ψ_N, indexed by support size N, is to make this tradeoff directly apparent: the smoothness of ψ increases with N at approximate rate $0.2N$ (Daubechies, 1992, §7.1.2).

Table B.1 *Desirable properties of orthonormal wavelet family, together with corresponding conditions on the filter h*

1. Orthonormal wavelet ψ	$\leftrightarrow$ CMF (B.4) and NORM (B.5)
2. p vanishing moments	$\leftrightarrow$ VM$_p$ (B.13)
3. (small) compact support	$\leftrightarrow$ N small
4. (high) regularity of ψ	

More on vanishing moments. We now give the proof that vanishing moments of ψ imply rapid decay of wavelet coefficients, and look at analogs for scaling functions φ and the interval $[0, 1]$.

Proof of Lemma 7.3 We first recall that Hölder functions can be uniformly approximated by (Taylor) polynomials, cf. (C.27). So, let $p(y)$ be the approximating Taylor polynomial of degree $\lceil \alpha \rceil - 1$ at $x_k = k2^{-j}$. Using a change of variable and the vanishing moments property,

$$\int f(x) 2^{j/2} \psi(2^j x - k) dx = 2^{-j/2} \int [f(x_k + 2^{-j} v) - p(2^{-j} v)] \psi(v) dv.$$

[3] If P_1, P_2 are degree $p-1$ solutions of (B.14), 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 $Q \equiv 0$.

Hence, using the Hölder bound (C.27),

$$|\langle f, \psi_{jk}\rangle| \le 2^{-j/2} C 2^{-j\alpha} \int |v|^\alpha |\psi(v)| dv.$$

Setting c_ψ equal to the latter integral yields the result. □

Vanishing moments for the scaling function. The approximation of point values $f(t_i)$ 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 φ. Bearing in mind that the scaling function itself has $\int \varphi = 1$ (e.g. from (B.6)) we say that φ has r vanishing moments if

$$\int x^k \varphi(x) dx = 0 \qquad k = 1, \dots r-1.$$

Lemma B.8 *If f is C^α on $\mathbb{R}$ and φ has at least $r = \lceil \alpha \rceil$ vanishing moments,*

$$|\langle f, \varphi_{jk}\rangle - 2^{-j/2} f(k2^{-j})| \le c_\psi C 2^{-j(\alpha+1/2)}.$$

Proof Modify the proof of Lemma 7.3 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

$$\int f\varphi_{jk} - 2^{-j/2} f(x_k) = 2^{-j/2} \int [f(x_k + 2^{-j}v) - f(x_k) - p_1(2^{-j}v)]\varphi(v) dv$$

and so $|\langle f, \varphi_{jk}\rangle - 2^{-j/2} f(x_k)| \le 2^{-j/2} C 2^{-j\alpha} c_\varphi$, where again $c_\varphi = \int |v|^\alpha |\varphi(v)| dv$. □

Vanishing moments for wavelets on $[0,1]$. Let $\mathcal{P}_p$ denote the space of polynomials of degree p. The vanishing moments theorem (e.g. [M, Ch. 7]) states that if φ and ψ have sufficiently rapid decay, then ψ has p vanishing moments if and only if the Strang-Fix condition is satisfied:

$$\theta_l(t) = \sum_{k=-\infty}^{\infty} k^l \varphi(t-k) \in \mathcal{P}_l \qquad l = 0, 1, \dots, p-1. \tag{B.15}$$

The condition (B.15) says that $\mathcal{P}_{p-1} \subset V_j$ for $j \ge 0$, and further (see Cohen et al. (1993b)) that for $j \ge J_*$, $\mathcal{P}_{p-1} \subset V_j[0,1]$—the multiresolution spaces corresponding to the CDJV construction described at the end of Section 7.1. Consequently $\mathcal{P}_{p-1} \perp W_j[0,1]$ and so for $j \ge J_*$, $k = 1, \dots, 2^j$, we have

$$\int t^l \psi_{jk}^{\text{int}}(t) dt = 0, \qquad l = 0, 1, \dots, p-1.$$

B.2 Further remarks on function spaces and wavelet coefficients

Section 9.6 took an idiosyncratic route, exploring some function spaces on $\mathbb{R}$, then defining Besov sequence norms on $\mathbb{R}$ and finally focusing on Besov sequence and function norms on $[0,1]$. In this section, again without attempting to be comprehensive, we collect some complementary remarks on these topics, and prepare the way for a proof of equivalence of Besov function and sequence norms on $[0,1]$ in the next section.

The Besov and Triebel scales of function spaces on $\mathbb{R}^n$ unify many of the classical spaces of analysis. They form the subject of several books, e.g. Nikol'skii (1975); Peetre (1975); Triebel (1983, 1992) and in particular Frazier et al. (1991), to which we refer for the discussion in this section. We specialize to the case $n = 1$.

Although it is not our main focus, for completeness we give one of the standard definitions of Besov and Triebel spaces on $\mathbb{R}$ that uses Fourier transforms. Let ψ be a "window" function of compact support in the frequency domain: assume, say, that supp $\widehat{\psi} \subset \{1/2 \le |\xi| \le 2\}$ and that $|\widehat{\psi}| \ge c > 0$ on $\{3/5 \le |\xi| \le 5/3\}$.

Given a function f, define "filtered" versions f_j by $\widehat{f}_j(\xi) = \widehat{\psi}(2^{-j}\xi)\widehat{f}(\xi)$: thus $\widehat{f}_j(\xi)$ is concentrated on the double octave $|\xi| \in [2^{j-1}, 2^{j+1}]$. For $\alpha \in \mathbb{R}$ and $0 < p, q \le \infty$, the homogeneous Besov and Triebel semi-norms are respectively defined by

$$|f|_{\dot{B}^\alpha_{p,q}} = \Big(\sum_j (2^{\alpha j} \|f_j\|_{L_p})^q \Big)^{1/q}, \qquad |f|_{\dot{F}^\alpha_{p,q}} = \Big\| \Big(\sum_j (2^{\alpha j} |f_j|)^q \Big)^{1/q} \Big\|_{L_p},$$

with the usual modifications if $p = \infty$ or $q = \infty$; thus $|f|_{\dot{B}^\alpha_{\infty,\infty}} = \sup_j 2^{\alpha j} \|f_j\|_\infty$. Thus the Besov semi-norm integrates over location at each scale and then combines over scale, while the *Triebel* semi-norm reverses this order. They merge if $p = q$: $\dot{B}^\alpha_{p,p} = \dot{F}^\alpha_{p,p}$, and more generally are sandwiched in the sense that $\dot{B}^\alpha_{p,p\wedge q} \subset \dot{F}^\alpha_{p,q} \subset \dot{B}^\alpha_{p,p\vee q}$. Despite the importance of the Triebel scale—$F^k_{p,2}$ equals the Sobolev space W^k_p, for example—we will not focus on them here.

These are the "homogeneous" definitions: if $f_t(x) = f(x/t)/t$, then the semi-norms satisfy a scaling relation: $\|f_t\|_{\dot{B}} = t^{(1/p)-1-\alpha} \|f\|_{\dot{B}}$. These are only semi-norms since they vanish on any polynomial. The "inhomogeneous" versions are defined by bringing in a "low frequency" function φ with the properties that supp $\widehat{\varphi} \subset [-2, 2]$, and $\widehat{\varphi} \ge c > 0$ on $[-5/3, 5/3]$. Then

$$\|f\|_{B^\alpha_{p,q}} = \|\varphi \star f\|_{L_p} + \Big(\sum_{j \ge 1} \big(2^{\alpha j} \|f_j\|_{L_p}\big)^q \Big)^{1/q},$$

with a corresponding definition for $\|f\|_{F^\alpha_{p,q}}$. These are norms for $1 \le p, q \le \infty$, otherwise they are still quasi-norms.

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. A remarkable table may be found in Frazier et al. (1991), from which we extract, in each case for $\alpha > 0$:

Hölder	$C^\alpha \equiv B^\alpha_{\infty,\infty}$	$\alpha \notin \mathbb{N}$
Hilbert-Sobolev	$W^\alpha_2 \equiv B^\alpha_{2,2}$	
Sobolev	$W^\alpha_p \equiv F^\alpha_{p,2}$	$1 < p < \infty$

If the window function ψ also satisfies the wavelet condition $\sum_j |\widehat{\psi}(2^{-j}\xi)|^2 \equiv 1$ a.e., then it is straightforward to verify that $|f|_{\dot{B}^\alpha_{2,2}}$ as defined above satisfies

$$|f|_{\dot{B}^\alpha_{2,2}} \asymp \int |\xi|^{2\alpha} |\widehat{f}(\xi)|^2 d\xi,$$

corresponding with the Fourier domain definition of $\int (D^\alpha f)^2$.

The Besov and Triebel function classes on $\mathbb{R}$ (and $\mathbb{R}^n$) have characterizations in terms of

wavelet coefficients. Using the Meyer wavelet, Lemarié and Meyer (1986) established the characterization for homogeneous Besov norms for $\alpha \in \mathbb{R}$ and $1 \leq p, q \leq \infty$. 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 (φ, ψ) have lower regularity—e.g. the Daubechies families of wavelets—then these characterisations hold for restricted ranges of (α, p, q). By way of example, if φ generates an r–regular MRA (i.e. essentially φ is C^r with rapid decay, see Meyer (1990)) then Meyer's result just cited shows that the equivalence (9.38) holds for $p, q \geq 1, |\alpha| < r$.

B.3 Besov spaces and wavelet coefficients

Let (φ, ψ) be an orthonormal scaling and wavelet function pair, complemented with boundary scaling functions and wavelets to yield an orthonormal basis for $L_2[0, 1]$:

$$f = \sum_k \beta_k \varphi_{Lk} + \sum_{j \geq L} \sum_k \theta_{jk} \psi_{jk}.$$

We have made frequent use of Besov norms on the coefficients $\beta = (\beta_k)$ and $\theta = (\theta_{j\cdot}) = (\theta_{jk})$. To be specific, define

$$\|f\|_{b^\alpha_{p,q}} = \|\beta\|_p + |\theta|_{b^\alpha_{p,q}}, \tag{B.16}$$

where, setting $a = \alpha + 1/2 - 1/p$

$$|\theta|_b^q = |\theta|^q_{b^\alpha_{p,q}} = \sum_{j \geq L} [2^{aj} \|\theta_{j\cdot}\|_p]^q. \tag{B.17}$$

In these definitions, one can take $\alpha \in \mathbb{R}$ and $p, q \in (0, \infty]$ with the usual modification for p or $q = \infty$.

This appendix justifies the term 'Besov norm' by showing that these *sequence* norms are equivalent to standard definitions of Besov norms on *functions* on $L_p(I)$.

We use the term *CDJV multiresolution* to describe the multiresolution analysis of $L_2[0, 1]$ resulting from the construction reviewed in Section 7.1. It is based on a Daubechies scaling function φ and wavelet ψ with compact support. If in addition, ψ is C^r—which is guaranteed for sufficiently large S, we say that the MRA is *r-regular.*

This section aims to give a more or less self contained account of the following result.

Theorem B.9 *Let r be a positive integer and suppose that $\{V_j\}$ is a r-regular CDJV multresolution analysis of $L_2[0, 1]$. Suppose that $1 \leq p, q \leq \infty$ and $0 < \alpha < r$. Let the Besov function space norm $\|f\|_{B^\alpha_{p,q}}$ be defined by (B.27), and the Besov sequence norm $\|f\|_{b^\alpha_{p,q}}$ by (B.16). Then the two norms are equivalent: there exist constants C_1, C_2 depending on (α, p, q) and the functions (φ, ψ), but not on f so that*

$$C_1 \|f\|_{b^\alpha_{p,q}} \leq \|f\|_{B^\alpha_{p,q}} \leq C_2 \|f\|_{b^\alpha_{p,q}}.$$

Equivalences of this type were first described by Lemarié and Meyer (1986) and developed in detail in Meyer (1992, Chapters 6 - 8). for $I = \mathbb{R}$. Their Calderón-Zygmund

operator methods make extensive use of the Fourier transform and the translation invariance of $\mathbb{R}$.

The exposition here, however, 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 and Lorentz (1993). The survey of nonlinear approximation, DeVore (1998), although more general in coverage than needed here, contains much helpful detail.

The conditions on α, p, q are not the most general. For example, Donoho (1992b) develops a class of *interpolating* wavelet transforms using an analog of L_2 multiresolution analysis for continuous functions with coefficients obtained by sampling rather than integration. For this transform, Besov (and Triebel) equivalence results are established for $0 < p, q \le \infty$, but with α now in the range $(1/p, r)$.

An encyclopedic coverage of Besov and Triebel function spaces and their characterizations may be found in the books Triebel (1983, 1992, 2006, 2008).

Outline of approach. One classical definition of the Besov function norm uses a modulus of smoothness based on averaged finite differences. We review this first. The modulus 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^{-rk}, namely $K(f, 2^{-rk})$, with the approximation error due to projection onto V_k,

$$e_k(f) = \|f - P_k f\|_p.$$

This comparison is a consequence of two key inequalities. The 'direct' or 'Jackson' inequality, Corollary B.17 below, bounds the approximation error in terms of the rth derivative

$$\|f - P_k f\|_p \le C 2^{-rk} \|f^{(r)}\|_p.$$

Its proof uses bounds on kernel approximation, along with the key property that each V_j contains $\mathcal{P}_{r-1}$. The 'inverse' or 'Bernstein' inequality, Lemma B.19 below, bounds derivatives of $g \in V_k$:

$$\|g^{(r)}\|_p \le C 2^{rk} \|g\|_p.$$

DeVore (1998) has more on the role of Jackson and Bernstein inequalities.

From this point, it is relatively straightforward to relate the approximation errors $e_k(f)$ with the wavelet coefficient norms (B.17). The steps are collected in the final equivalence result, Theorem B.9, in particular in display (B.48).

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 and 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

$$\|f\|_p(A) = \Big(\int_A |f(x)|^p dx\Big)^{1/p},$$

and, as usual, $\|f\|_\infty(A) = \sup_{x\in A}|f(x)|$.

Let $T_h f(x) = f(x+h)$ denote translation by h. The first difference of a function is

$$\Delta_h(f,x) = f(x+h) - f(x) = (T_h - I)f(x).$$

Higher order differences, for $r \in \mathbb{N}$, are given by

$$\Delta_h^r(f,x) = (T_h - I)^r f(x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x+kh). \tag{B.18}$$

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

$$\omega_r(f,t)_p = \sup_{0\le h\le t} \|\Delta_h^r(f,\cdot)\|_p(A_{rh}).$$

For $p < \infty$, this is a measure of smoothness averaged over A; the supremum ensures monotonicity in t. If $p = \infty$, it is a uniform measure of smoothness, for example

$$\omega_1(f,t)_\infty = \sup\{|f(x) - f(y)|,\ x, y \in A, |x-y| \le t\}.$$

The differences $\Delta_h^r(f,x)$ are linear in f, and so for $p \ge 1$, there is a triangle inequality

$$\omega_r(f+g,t)_p \le \omega_r(f,t)_p + \omega_r(g,t)_p. \tag{B.19}$$

Again by linearity, $\|\Delta_h^r(f,\cdot)\|_p \le 2^r \|f\|_p$ and so also

$$\omega_r(f,t)_p \le 2^r \|f\|_p, \tag{B.20}$$

and more generally, for $0 \le k \le r$,

$$\omega_r(f,t)_p \le 2^{r-k}\omega_k(f,t)_p. \tag{B.21}$$

For $n \in \mathbb{N}$ and $1 \le p \le \infty$ it can be verified that

$$\omega_r(f,nt)_p \le n^r \omega_r(f,t)_p. \tag{B.22}$$

When derivatives exist, the finite difference can be expressed as a kernel smooth of bandwidth h of these derivatives:

Lemma B.10 *Let χ be the indicator of $[0,1]$, and $\chi^{\star r}$ be its r^{th} convolution power. Then*

$$\Delta_h^r(f,x) = h^r \frac{d^r}{dx^r} \int f(x+hu)\chi^{\star r}(u)du \tag{B.23}$$

$$= h^r \int f^{(r)}(x+hu)\chi^{\star r}(u)du, \tag{B.24}$$

the latter inequality holding if $f \in W_p^r$.

The easy proof uses induction. A simple consequence of (B.24) is the bound

$$\omega_r(f,t)_p \le t^r |f|_{W_p^r(I)}, \tag{B.25}$$

valid for all $t \ge 0$. Indeed, rewrite the right side of (B.24) as $h^r \int K(x,v) f^{(r)}(v)dv$, using the kernel

$$K(x,v) = h^{-1}\chi^{\star r}(h^{-1}(v-x))$$

for $x \in I_h$ and $v = x + hu \in I$. Now apply Young's inequality (C.29), which says that the operator with kernel K is bounded on L_p. Note that both M_1 and $M_2 \le 1$ since $\chi^{\star r}$ is a probability density, so that the norm of K is at most one. Hence

$$\|\Delta_h^r(f,\cdot)\|_p(I_{rh}) \le h^r |f|_{W_p^r(I)},$$

and the result follows from the definition of ω_r.

B.11 *Uniform smoothness.* There are two ways to define uniform smoothness using moduli. Consider $0 < \alpha \le 1$. The first is the usual Hölder/Lipschitz definition

$$|f|_{\text{Lip}(\alpha)} = \sup_{t>0} t^{-\alpha}\omega_1(f,t)_\infty,$$

which is the same as (C.26). The second replaces the first-order difference by one of (possibly) higher order. Let $r = [\alpha] + 1$ denote the smallest integer larger than α and put

$$|f|_{\text{Lip}^*(\alpha)} = \sup_{t>0} t^{-\alpha}\omega_r(f,t)_\infty.$$

Clearly these coincide when $0 < \alpha < 1$. When $\alpha = 1$, however, $\text{Lip}^*(1) = Z$ is the Zygmund space, and

$$\|f\|_{\text{Lip}^*(1)} = \|f\|_\infty + \sup_{x, x\pm h \in A} \frac{|f(x+h) - 2f(x) + f(x+h)|}{h}.$$

It can be shown (e.g. DeVore and Lorentz (1993, p. 52)) that $\text{Lip}^*(1) \supset \text{Lip}(1)$ and that the containment is proper, using the classical example $f(x) = x \log x$ on $[0,1]$.

Besov spaces. Let $\alpha > 0$ and $r = \lfloor \alpha \rfloor + 1$. Let $A = \mathbb{R}$, or an interval $[a,b]$. The Besov space $B_{p,q}^\alpha(A)$ is the collection of $f \in L_p(A)$ for which the semi-norm

$$|f|_{B_{p,q}^\alpha} = \Big(\int_0^\infty \Big[\frac{\omega_r(f,t)_p}{t^\alpha} \Big]^q \frac{dt}{t} \Big)^{1/q} \tag{B.26}$$

is finite. If $q = \infty$, we use $|f|_{B_{p,\infty}^\alpha} = \sup_t t^{-\alpha}\omega_r(f,t)_p$. The semi-norm vanishes if f is a polynomial of degree less than r. As norm on $B_{p,q}^\alpha(A)$, we take

$$\|f\|_{B_{p,q}^\alpha} = \|f\|_p + |f|_{B_{p,q}^\alpha}. \tag{B.27}$$

If $p = q = \infty$ and $\alpha < 1$, so that $r = 1$, we recover the Lip(α) or Hölder-α semi-norm. If $\alpha = 1$, then $r = 2$ and $B^1_{\infty,\infty}$ is the Zygmund space.

A simple inequality between Besov and Sobolev norms states that for $m \in \mathbb{N}$,

$$|f|_{B^m_{p,\infty}} \le C \int_I |D^m f|^p.$$

Indeed, take $r = m + 1$ in the definition of the $B^m_{p,\infty}$ norm, then apply (B.21) and (B.25) to get

$$\omega_{m+1}(f,t)_p \le 2\omega_m(f,t)_p \le 2t^m |f|_{W^m_p}$$

so that $|f|_{B^m_{p,\infty}} \le 2|f|_{W^m_p}$ as required.

Remarks. 1. The assumption that $r > \alpha$ is used in at least two places in the equivalence arguments to follow: first in the interpolation space identification of $B^\alpha_{p,q}$, Theorem B.12, and second in Theorem B.20 relating approximation error to the K-functional. This indicates why it is the Zygmund space—and more generally Lip$^*(\alpha)$—that appears in the wavelet characterizations of $B^\alpha_{\infty,\infty}$ for integer α, rather than the traditional C^α spaces.

2. The modulus based definition is equivalent (on $\mathbb{R}^n$) to the earlier Fourier form if $\alpha > n(p^{-1} - 1)_+, 0 < p, q \le \infty$, (e.g. Triebel (1983, p. 110), [For $\alpha > 0, 1 \le p, q \le \infty$, see also Bergh and Löfström (1976, Th. 6.2.5))].

Besov spaces as interpolation spaces

This section shows that Besov spaces are *intermediate* spaces between $L_p(I)$ and $W^r_p(I)$. First we need the notion of *K-functional*, reminiscent of roughness penalized approximations in the theory of splines:

$$K(f,t) = K(f,t; L_p, W^r_p) = \inf\{\|f - g\|_p + t\|D^r g\|_p : g \in W^r_p\}.$$

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 B.13 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 in t. Further, for any f

$$K(f,t) \ge \min(t,1)K(f,1), \tag{B.28}$$

while if $f \in W^r_p$ then by choosing g equal to f or 0 as $t \le 1$ or $t > 1$,

$$K(f,t) \le \min(t,1)\|f\|_{W^r_p}. \tag{B.29}$$

A sort of converse to (B.28) will be useful. We first state a result which it is convenient to prove later, after Proposition B.16. Given $g \in W^r_p$, let $\Pi_{r-1}g$ be the best (in $L_2(I)$) polynomial approximation of degree $r-1$ to g. Then for $C = C(I,r)$,

$$\|g - \Pi_{r-1}g\|_p \le C\|g^{(r)}\|_p. \tag{B.30}$$

Now, let $f \in L_p$ and $g \in W_p^r$ be given. From the definition of K and (B.30),

$$\begin{aligned} K(f,t) \le \|f - \Pi_{r-1} g\|_p &\le \|f - g\|_p + \|g - \Pi_{r-1} g\|_p \\ &\le \|f - g\|_p + C\|g^{(r)}\|_p, \end{aligned}$$

where $C = C(I, r)$. Hence, for all $t \ge a$,

$$K(f,t) \le \max(Ca^{-1}, 1) K(f,a). \tag{B.31}$$

The $K-$functional $K(f,t)$ 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 $0 < \theta < 1$,

$$\rho(f)_{\theta,q} = \Big(\int_0^\infty \Big[\frac{K(f,t)}{t^\theta} \Big]^q \frac{dt}{t} \Big)^{1/q} \qquad 0 < q < \infty \tag{B.32}$$

and, when $q = \infty$, $\rho(f)_{\theta,\infty} = \sup_{0 \le t \le \infty} t^{-\theta} K(f,t)$.

Replacing $K(f,t)$ by $\min(1,t)$ in the integral (B.32) leads to the sum of two integrals $\int_0^1 t^{(1-\theta)q-1} dt$ and $\int_1^\infty t^{-\theta q - 1} dt$, which both converge if and only if $0 < \theta < 1$. Hence property (B.28) shows that in order for $\rho(f)_{\theta,q}$ to be finite for any f other than polynomials, it is necessary that $0 < \theta < 1$.

On the other hand, property (B.29) shows that

$$\rho(f)_{\theta,q} \le c_{\theta q} \|f\|_{W_p^r}. \tag{B.33}$$

We therefore define intermediate, or *interpolation* spaces

$$X_{\theta,q} = (L_p, W_p^r)_{\theta,q} = \{ f \in L_p : \rho(f)_{\theta,q} < \infty \}$$

for $0 < q \le \infty$ and $0 < \theta < 1$, and set $\|f\|_{X_{\theta,q}} = \|f\|_p + \rho(f)_{\theta,q}$.

From the definition and (B.33),

$$W_p^r \subset (L_p, W_p^r)_{\theta,q} \subset L_p.$$

The parameters (θ, q) yield a lexicographic ordering:

$$X_{\theta_1,q_1} \subset X_{\theta_2,q_2} \quad \text{if } \theta_1 > \theta_2, \text{ or if } \theta_1 = \theta_2 \text{ and } q_1 \le q_2.$$

The main reason for introducing interpolation spaces here is that they are in fact Besov spaces.

Theorem B.12 *For $r \in \mathbb{N}$, and $1 \le p \le \infty$, $0 < q \le \infty$, $0 < \alpha < r$,*

$$(L_p, W_p^r)_{\alpha/r,q} = B_{p,q}^\alpha.$$

This follows from the definitions and the next key theorem (Johnen, 1972), which shows that the K-functional is equivalent to the integral modulus of continuity.

Theorem B.13 *Let $A = \mathbb{R}, \mathbb{R}_+$, the unit circle $\mathbb{T}$, or $[0,1]$. For $1 \le p \le \infty$, and $r \in \mathbb{N}$, there exist $C_1, C_2 > 0$ depending only on r, such that for all $f \in L_p$,*

$$C_1 \omega_r(f,t)_p \le K(f, t^r; L_p, W_p^r) \le C_2 \omega_r(f,t)_p, \qquad t > 0. \tag{B.34}$$

Proof We work on the left inequality first: from the triangle inequality (B.19) followed by (B.20) and derivative bound (B.25), we have for arbitrary g,

$$\begin{aligned}\omega_r(f,t)_p &\le \omega_r(f-g,t)_p + \omega_r(g,t)_p \\ &\le 2^r\|f-g\|_p + t^r|g|_{W_p^r}.\end{aligned}$$

Minimizing over g, we obtain the left inequality in (B.34) with $C_1 = 2^{-r}$.

For the right inequality, we only give full details for $A = \mathbb{R}$. Given f, we choose

$$g(x) = f(x) + (-1)^{r+1}\int \Delta_{ut}^r(f,x)\chi^{\star r}(u)du. \tag{B.35}$$

By the Minkowski integral inequality (C.33),

$$\|g-f\|_p \le \int \|\Delta_{ut}^r(f,\cdot)\|_p\chi^{\star r}(u)du \le \omega_r(f,rt)_p \le r^r\omega_r(f,t)_p, \tag{B.36}$$

where the second inequality follows because $\chi^{\star r}$ is a probability density supported on $[0,r]$, and the third uses (B.22).

Now estimate $\|g^{(r)}\|_p$. Use expansion (B.18) for $\Delta_{tu}^r(f,x)$, noting that the $k=0$ term cancels $f(x)$ in (B.35). Differentiate and then use (B.23) to obtain

$$\begin{aligned}g^{(r)}(x) &= \sum_{k=1}^r \binom{r}{k}(-1)^{k+1}\frac{d^r}{dx^r}\int f(x+ktu)\chi^{\star r}(u)du \\ &= \sum_{k=1}^r \binom{r}{k}(-1)^{k+1}(kt)^{-r}\Delta_{kt}^r(f,x).\end{aligned}$$

Again using (B.22), we find

$$t^r\|g^{(r)}\|_p \le \sum_{k=1}^r \binom{r}{k}k^{-r}\omega_r(f,kt)_p \le 2^r\omega_r(f,t)_p.$$

Putting this last inequality and (B.36) into the definition of $K(f,t^r)$ yields the right hand bound with $C_2 = r^r + 2^r$.

If $A = [0,1]$, then g is defined in (B.35) for $x \in I_1 = [0,3/4]$ if $t \le 1/4r^2$. 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 and Lorentz (1993, p. 176, 178). □

For work with wavelet coefficients, we need a discretized version of these measures.

Lemma B.14 *Let $L \in \mathbb{N}$ be fixed. With constants of proportionality depending on I, r, θ, q and L but not on f,*

$$\rho(f)_{\theta,q}^q \asymp \sum_{j=L-1}^{\infty}[2^{\theta rj}K(f,2^{-rj})]^q. \tag{B.37}$$

Proof Since $K(f,t)$ is concave in t with $K(f,0) = 0$, we have $\epsilon K(f,t) \le K(f,\epsilon t)$, and since it is increasing in t, we have for $2^{-r(j+1)} \le t \le 2^{-rj}$,

$$2^{-r}K(f,2^{-rj}) \le K(f,2^{-r(j+1)}) \le K(f,t) \le K(f,2^{-rj}).$$

From this it is immediate that, with $a = 2^{-r(L-1)}$, the sum $S_L(f)$ in (B.37) satisfies

$$S_L(f) \asymp \int_0^a \Big[\frac{K(f,t)}{t^\theta}\Big]^q \frac{dt}{t}$$

with constants of proportionality depending only on (θ, q, r). From (B.31),

$$\int_a^\infty \Big[\frac{K(f,t)}{t^\theta}\Big]^q \frac{dt}{t} \le C[K(f,a)a^{-\theta}]^q$$

where C depends on (I, L, r, θ, q). With $a = 2^{-r(L-1)}$, this last term can be absorbed in the sum $S_L(f)$, completing the proof. □

MRAs on $[0, 1]$

We use the term *CDJV multiresolution* to describe the multiresolution analysis of $L_2[0, 1]$ resulting from the construction reviewed in Section 7.1. It is based on a scaling function φ and wavelet ψ with support in $[-S + 1, S]$ and for which ψ has S vanishing moments. The MRA of $L_2[0, 1]$ is constructed using S left and S right boundary scaling functions $\varphi_k^L, \varphi_k^R, k = 0, \ldots S - 1$.

Choose a coarse level L so that $2^L \ge 2S$. For $j \ge L$, we obtain scaling function spaces $V_j = \text{span}\{\varphi_{jk}\}$ of dimension 2^j. The orthogonal projection operators $P_j : L_2(I) \to V_j$ have associated kernels

$$E_j(x, y) = \sum_k \varphi_{jk}(x)\varphi_{jk}(y),$$

as may be seen by writing

$$P_j f(x) = \sum_k \langle f, \varphi_{jk}\rangle \varphi_{jk}(x) = \int \sum_k \varphi_{jk}(x)\varphi_{jk}(y) f(y) dy.$$

If in addition, ψ is C^r–which is guaranteed for sufficiently large S–we say that the MRA is *r-regular.* Since ψ is C^r it follows (e.g. by Daubechies (1992, Corollary 5.5.2)) that ψ has r vanishing moments. The CDJV construction then ensures that $\mathcal{P}_{r-1}$, the space of polynomials of degree $r - 1$ on $[0, 1]$ is contained in V_L. In fact, we abuse notation and write $V_{L-1} = \mathcal{P}_{r-1}$. The corresponding orthogonal projection operator $P_{L-1} : L_2(I) \to V_{L-1}$ has kernel

$$\Pi_{r-1}(x, y) = \sum_{k=0}^{r-1} p_k(x) p_k(y) \qquad x, y \in I. \tag{B.38}$$

Here $p_k(x)$ are Legendre polynomials of degree k, scaled to be orthnormal on $L_2(I)$. We borrow from Szegö (1967, p. 164) the bound

$$|p_k(x)| \le \sqrt{2k + 1}, \qquad x \in I. \tag{B.39}$$

A simple fact for later use is that P_j have uniformly bounded norms on $L_p[0, 1]$. Define

$$a_q(\varphi) = \max\{\|\varphi\|_q, \|\varphi_k^L\|_q, \|\varphi_k^R\|_q, k = 0, \ldots, S - 1\}. \tag{B.40}$$

Lemma B.15 *Suppose that $\{V_j\}$ is a CDJV multresolution analysis of $L_2[0,1]$. Then for $1 \le p \le \infty$,*

$$\|P_j\|_p \le 2Sa_1(\varphi)a_\infty(\varphi), \tag{B.41}$$

$$\|P_{L-1}\|_p \le C(r). \tag{B.42}$$

Proof We simply apply Young's inequality (C.29). For $j \ge L$, we need the bounds

$$\sum_k |\varphi_{jk}(x)| \le 2S2^{j/2}a_\infty(\varphi), \qquad \int |\varphi_{jk}(y)|dy \le 2^{-j/2}a_1(\varphi)$$

from which it follows that $\int |E_j(x,y)|dy \le 2Sa_1(\varphi)a_\infty(\varphi)$ and similarly for $\int |E_j(x,y)|dx$. We argue similarly for $j = L-1$ using the bounds

$$\sum_{k=0}^{r-1} |p_k(x)| \le Cr^{3/2}, \qquad \int |p_k(y)|dy \le 1. \qquad \square$$

With the addition of boundary wavelets $\psi_k^L, \psi_k^R, k = 0, \dots, S-1$, one obtains detail spaces $W_j = \text{span}\{\psi_{jk}, k = 0, \dots, 2^j - 1\}$ and the decomposition

$$L_2[0,1] = V_L \oplus \oplus_{j \ge L} W_j.$$

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

$$K_h f(x) = \int_I K_h(x,y) f(y) dy \qquad x \in I.$$

In the proof to follow, $\|f\|_p = (\int_I |f|^p)^{1/p}$ is the L_p norm on I.

Proposition B.16 *Suppose that the kernel $K_h(x,y)$ satisfies*

(i) $K_h\pi = \pi$ *for* $\pi \in \mathcal{P}_{r-1}$,

(ii) $K_h(x,y) = 0$ *if* $|y-x| > Lh$,

(iii) $|K_h(x,y)| \le Mh^{-1}$.

on an interval $I \subset \mathbb{R}$. For $p \ge 1$, there exists a constant $C = C(L,M,r)$ such that for $f \in W_p^r(I)$,

$$\|f - K_h f\|_p \le Ch^r \|D^r f\|_p, \qquad h > 0.$$

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.

Proof A function $f \in W_p^r(I)$ has continuous derivatives of order $k = 0, 1, \dots, r-1$. If $x \in I$, we may therefore use the Taylor approximation to f at x by a polynomial π_x of degree $r-1$, so that $f(y) = \pi_x(y) + R_x(y)$ with the integral form of the remainder term

$$R_x(y) = c_{r-1} \int_x^y (D^r f)(u)(y-u)^{r-1} du, \qquad c_{r-1} = 1/(r-1)!$$

Since K_h leaves such polynomials invariant, $K_h f = \pi_x + K_h R_x$, and since $\pi_x(x) = f(x)$,

$$\begin{aligned}(K_h f)(x) - f(x) &= \int_I K_h(x,y) R_x(y) dy \\ &= c_{r-1} \int_I K_h(x,y) \int_x^y (y-u)^{r-1} f^{(r)}(u) du dy \\ &= \int_I \tilde{K}_h(x,u) f^{(r)}(u) du,\end{aligned}$$

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):

$$|\tilde{K}_h(x,u)| \le \begin{cases} cMh^{-1}(Lh)^r & \text{if } |x-u| \le Lh \\ 0 & \text{otherwise .} \end{cases}$$

Since $\int_I |\tilde{K}_h(x,u)| du \le 2cL^{r+1}Mh^r$, with a similar bound for the corresponding integral over $x \in I$, our result follows from Young's inequality (C.29) with $M_1 = M_2 = 2cL^{r+1}Mh^r$. □

A common special case occurs when $K_h(x,y) = h^{-1}K(h^{-1}(x-y))$ is a scaled translation invariant kernel on $\mathbb{R}$. Condition (i) is equivalent to the vanishing moment property $\int t^k K(t) dt = \delta_{k0}$ for $k = 0, 1, \ldots, r-1$. If $K(y)$ is bounded and has compact support, then properties (ii) and (iii) are immediate.

As a second example, consider orthogonal polynomials on $I = [0,1]$ and the associated kernel $\Pi_{r-1}(x,y)$ given in (B.38). Assumptions (i) - (ii) hold for $h = L = 1$. The bound (B.39) shows that (iii) holds with $M = r^2$. Consequently, for $f \in W_p^r(I)$ we obtain the bound $\|f - \Pi_{r-1} f\|_p \le C \|f^{(r)}\|_p$ for $C = C(r)$, which is just (B.30).

Our main use of Proposition B.16 is a Jackson inequality for multiresolution analyses.

Corollary B.17 *Suppose that $\{V_j\}$ is a CDJV multresolution analysis of $L_2[0,1]$. Let P_j be the associated orthogonal projection onto V_j, and assume that $2^j \ge 2S$. Then there exists a constant $C = C(\varphi)$ such that for all $f \in W_p^r(I)$,*

$$\|f - P_j f\|_p \le C 2^{-rj} |f|_{W_p^r}.$$

Proof We claim that assumptions (i)-(iii) hold for the kernel E_j with h taken as 2^{-j}. The CDJV construction guarantees that $\mathcal{P}_{r-1} \subset V_j$ so that (i) holds. In addition the construction implies that (ii) holds with $L = 2S$ and that

$$\#\{k : \varphi_{jk}(x)\varphi_{jk}(y) \ne 0\} \le 2S.$$

It follows that (iii) holds with $M = 2Sa_\infty^2(\varphi)$. □

Bernstein-type Inequalities

First a lemma, inspired by Meyer (1990, p.30), which explains the occurence of terms like $2^{j(1/2-1/p)}$ in sequence norms.

Lemma B.18 *Let $\{\gamma_{jk}, k \in \mathcal{K}\}$ be an orthonormal sequence of functions satisfying*

$$\text{(i)} \qquad \sum_k |\gamma_{jk}(x)| \le b_\infty 2^{j/2}, \qquad \text{and}$$

$$\text{(ii)} \qquad \max_k \int |\gamma_{jk}| \le b_1 2^{-j/2}.$$

Then for all $1 \le p \le \infty$, and any sequence $\lambda = (\lambda_k, k \in \mathcal{K})$,

$$C_1 2^{j(1/2-1/p)} \|\lambda\|_p \le \Big\| \sum_k \lambda_k \gamma_{jk} \Big\|_p \le C_2 2^{j(1/2-1/p)} \|\lambda\|_p. \tag{B.43}$$

Here $C_1 = b_1^{-1}(b_1/b_\infty)^{1/p}$ and $C_2 = b_\infty(b_1/b_\infty)^{1/p}$.

Remarks. 1. If φ is an orthonormal scaling function and $\gamma_{jk}(x) = 2^{j/2}\varphi(2^j x - k)$ for $k \in \mathbb{Z}$, and $|\operatorname{supp}\varphi| \le B$, then (i) and (ii) are trivially satisfied with $b_\infty = B\|\varphi\|_\infty$ and $b_1 = \|\varphi\|_1$.

2. If $\{\gamma_{jk}\} = \{\varphi_{jk}\}$ correspond to a CDJV boundary MRA for $[0,1]$ derived from a scaling function φ with $\operatorname{supp}\varphi \subset [-S+1, S]$, then (i) and (ii) hold with $b_\infty = 2Sa_\infty(\varphi)$ and $b_1 = a_1(\varphi)$, using definitions (B.40). Analogous remarks apply with wavelets, when $\{\gamma_{jk}\} = \{\psi_{jk}\}$.

3. The right side in (B.43) does not require the assumption of orthonormality for $\{\gamma_{jk}\}$.

Proof This is just the extended Young inequality, Theorem C.26. Identify $\mu(dx)$ with Lebesgue measure on $\mathbb{R}$ and $\nu(dy)$ with counting measure on $k \in \mathcal{K}$. Then match $K(x,y)$ with $\gamma_{jk}(x)$ and $f(y)$ with λ_k. Conditions (i) and (ii) imply that $M_1 = b_1 2^{-j/2}$ and $M_2 = b_\infty 2^{j/2}$ suffice for the conditions of the theorem. The right hand inequality above now follows from (C.29). Note that orthonormality of $\{\gamma_{jk}\}$ is not used.

For the left hand inequality, let $g(x) = \sum_k \lambda_k \gamma_{jk}$. The $\{\gamma_{jk}\}$ are orthonormal, so

$$(K^* g)_k = \int \gamma_{jk}(x) g(x) dx = \lambda_k$$

and now the result follows from the adjoint form (C.30) of Young's inequality. □

Now to the variant of the Bernstein inequality that we need. We now require ψ to be C^r.

Lemma B.19 *Suppose that $\{V_j\}$ is a r-regular CDJV multresolution analysis of $L_2[0,1]$. For $g \in V_j$ and $1 \le p \le \infty$, and a constant $c = c(\varphi, r)$,*

$$\|D^r g\|_p \le c 2^{jr} \|g\|_p.$$

Proof Since $g \in V_j$, it has an expansion $g = \sum \lambda_k \varphi_{jk}$, and so

$$D^r g = \sum \lambda_k D^r \varphi_{jk} = 2^{jr} \sum \lambda_k \gamma_{jk},$$

where the functions γ_{jk} are formed from the finite set $\{D^r\varphi, D^r\varphi_k^0, D^r\varphi_k^1\}$ by exactly the same set of linear operations as used to form φ_{jk} from the set $\{\varphi, \varphi_k^0, \varphi_k^1\}$.

Since the $\{\varphi_{jk}\}$ system satisfy the conditions (i) and (ii) of Lemma B.18, the same is true of the $\{\gamma_{jk}\}$ system. From the right side of that Lemma,

$$\|D^r g\|_p = 2^{jr} \|\sum \lambda_k \gamma_{jk}\|_p \le C_2 2^{jr} 2^{j(1/2-1/p)} \|\lambda\|_p.$$

Now apply the left side of the same Lemma to the (orthogonal!) $\{\varphi_{jk}\}$ system to get

$$\|D^r g\|_p \le C_2 C_1^{-1} 2^{jr} \|\sum \lambda_k \varphi_{jk}\|_p = b_1 b_\infty 2^{jr} \|g\|_p. \qquad \square$$

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 $f \in W_p^r(I)$ 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 $K(f, 2^{-rk})$, using the Jackson and Bernstein inequalities, Corollary B.17 and Lemma B.19 respectively. In order to handle low frequency terms, we use the notation V_{L-1} to refer to the space of polynomials of degree at most $r-1$, and adjoin it to the spaces $V_k, k \ge L$ of the multiresolution analysis.

Theorem B.20 *Suppose that $\{V_j\}$ is a r-regular CDJV multresolution analysis of $L_2[0,1]$. Let $r \in \mathbb{N}$ be given. For $1 \le p \le \infty, 0 < q < \infty$ and $0 < \alpha < r$. With constants depending on (α, r, φ), but not on f, we have*

$$\sum_{L-1}^{\infty} [2^{\alpha k} e_k(f)]^q \asymp \sum_{L-1}^{\infty} [2^{\alpha k} K(f, 2^{-rk})]^q. \tag{B.44}$$

Proof 1°. The main work is to show that for $k \ge L-1$

$$C_1 e_k(f) \le K(f, 2^{-kr}) \le C_2 \sum_{j=L-1}^{k} 2^{-(k-j)r} e_j(f), \tag{B.45}$$

with constants $C_i = C_i(\varphi, r)$. For the left hand inequality, let $f \in L_p$ and $g \in W_p^r$ be fixed. Write $f - P_k f$ as the sum of $(I - P_k)(f-g)$ and $g - P_k g$, so that

$$e_k(f) \le \|(I - P_k)(f-g)\|_p + e_k(g).$$

It follows from (B.41) that $\|I - P_k\|_p \le 1 + C(\varphi)$. Together with Jackson inequality Corollary B.17 for $k \ge L$ and (B.30) for $k = L-1$, this yields

$$e_k(f) \le C[\|f-g\|_p + 2^{-rk}|g|_{W_p^r}].$$

Minimizing now over g yields the left side of (B.45).

For the right inequality, set $\psi_j = P_j f - P_{j-1} f \in V_j$ and write $P_k f = \sum_{j=L}^{k} \psi_j + P_{L-1} f$. Now $P_{L-1} f$ is a polynomial of degree at most $r-1$, so $|P_{L-1} f|_{W_p^r} = 0$. For the other terms, apply the Bernstein inequality Lemma B.19 to obtain

$$|P_k f|_{W_p^r} \le \sum_{j=L}^{k} |\psi_j|_{W_p^r} \le c \sum_{L}^{k} 2^{rj} \|\psi_j\|_p \le c \sum_{L}^{k} 2^{rj} [e_{j-1}(f) + e_j(f)].$$

Finally, put this into the $K-$functional definition:

$$K(f, 2^{-kr}) \le \|f - P_k f\|_p + 2^{-kr}|P_k f|_{W_p^r}$$
$$\le (1 + 2^{r+1}c) \sum_{j=L-1}^{k} 2^{-(k-j)r} e_j(f).$$

2°. The left to right bound in (B.44) is immediate from (B.45). For the other inequality, let $b_k = 2^{\alpha k} e_k(f)$ and $c_k = 2^{\alpha k} K(f, 2^{-rk})$ for $k \ge L-1$ and 0 otherwise. Then bound (B.45) says that $c_k \le \sum_{j=L-1}^{\infty} a_{k-j} b_j$ for $k \ge L-1$, where $a_k = C_2 2^{-k(r-\alpha)} I\{k \ge 0\}$. Our bound $\|c\|_q \le c_{r\alpha} C_2 \|b\|_q$ now follows from Young's inequality (C.32). □

Wavelet coefficients, finally

The last step in this chain is now quite easy, namely to relate semi-norms 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

$$Q_j f = \sum_k \theta_{jk} \psi_{jk}, \qquad \theta_{jk} = \langle f, \psi_{jk} \rangle.$$

Let $\|\theta_{j\cdot}\|_p$ denote the ℓ_p-norm of (θ_{jk}), and $a = \alpha + 1/2 - 1/p$.

Lemma B.21 *For $\alpha > 0$ and $1 \le p \le \infty$, and an $r-$regular CDJV multiresolution analysis of $L_2[0,1]$,*

$$\sum_{j \ge L} [2^{\alpha j} \|Q_j f\|_p]^q \asymp \sum_{j \ge L} [2^{aj} \|\theta_{j\cdot}\|_p]^q \asymp \sum_{j \ge L} [2^{\alpha j} e_j(f)]^q$$

Proof The first equivalence follows from Lemma B.18 and the Remark 2 following it:

$$\|Q_j f\|_p \asymp 2^{j(1/2-1/p)} \|\theta_{j\cdot}\|_p, \tag{B.46}$$

For the second equivalence, let $\delta_k = \|Q_k f\|_p$ and $e_k = e_k(f) = \|f - P_k f\|_p$. Clearly $\delta_k \le e_k + e_{k+1}$, which suffices for one of the inequalities. On the other hand, $f - P_j f = \sum_{k \ge j} Q_k f$, and so $e_j \le \sum_{k \ge j} \delta_k$, or equivalently

$$2^{\alpha j} e_j \le \sum_{k \ge j} 2^{-\alpha(k-j)} 2^{\alpha k} \delta_k.$$

The other inequality now follows from Young's inequality (C.32). □

Remark. The same argument as for (B.46) applies also to the projection onto V_L, given by $P_L f = \sum_k \beta_k \varphi_{Lk}$ to show that, with $\beta = (\beta_k)$,

$$\|P_L f\|_p \asymp 2^{L(1/2-1/p)} \|\beta\|_p. \tag{B.47}$$

Summary: norm equivalence

We assemble the steps carried out in earlier subsections to finally establish Theorem B.9.

Proof Combine the definition of the Besov semi-norm (B.26), the equivalence of modulus and $K-$functional (B.34) (with $s = t^r$ and $\theta = \alpha/r$), the dyadic discretization (B.37) and the $(\alpha, q)-$equivalence of $K-$functional and MRA-approximation errors (B.44) to find

$$\begin{aligned} |f|^q_{B^\alpha_{p,q}} &= \int_0^\infty \left[\frac{\omega_r(f,t)_p}{t^\alpha}\right]^q \frac{dt}{t} \\ &\asymp \int_0^\infty \left[\frac{K(f,s)}{s^\theta}\right]^q \frac{ds}{s} \\ &\asymp \sum_{j\geq L-1} [2^{\alpha j} K(f, 2^{-rj})]^q \\ &\asymp \sum_{j\geq L-1} [2^{\alpha j} e_j(f)]^q \end{aligned} \tag{B.48}$$

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 semi-norm on wavelet coefficients:

$$\sum_{j\geq L} [2^{\alpha j} e_j(f)]^q \asymp |\theta|^q_b. \tag{B.49}$$

Although the ranges of summation differ, this is taken care of by inclusion of the L_p norm of f, as we now show. 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, using also (B.47) with (B.41), we get

$$\|f\|_b = \|\beta\|_p + |\theta|_b \leq C\|P_L f\|_p + C|f|_B \leq C(\|f\|_p + |f|_B) = C\|f\|_B.$$

In the other direction, we connect the two chains by writing $|f|_B \leq C[e_{L-1}(f) + |\theta|_b]$ and observing from (B.42) that $e_{L-1}(f) \leq \|I - P_{L-1}\|_p \|f\|_p \leq C\|f\|_p$. Consequently,

$$\|f\|_B = \|f\|_p + |f|_B \leq C(\|f\|_p + |\theta|_b).$$

Now $\|f\|_p \leq e_L(f) + \|P_L f\|_p$ which is in turn bounded by $C(|\theta|_b + \|\beta\|_p)$ by (B.49) and (B.47). Putting this into the last display finally yields $\|f\|_B \leq C\|f\|_b$. □

B.4 Vaguelettes and frames

We rewrite Definition 12.2 without the rescaling operators. A collection $\{w_\lambda\}$ with $\lambda = (j,k)$ and $j \in \mathbb{Z}, k \in \Lambda_j \subset \mathbb{Z}$ is called a system of vaguelettes if there exist constants C_1, C_2 and exponents $0 < \eta < \eta' < 1$ such that

$$|w_\lambda(x)| \leq C_1 2^{j/2}(1 + |2^j x - k|)^{-1-\eta'}, \tag{B.50}$$

$$\int w_\lambda(x)dx = 0, \tag{B.51}$$

$$|w_\lambda(x') - w_\lambda(x)| \leq C_2 2^{j(1/2+\eta)}|x' - x|^\eta. \tag{B.52}$$

Proof of Proposition 12.3. (i) (Meyer and Coifman, 1997, Ch. 8.5) Let $K_{\lambda\lambda'} = \int w_\lambda \bar{w}_{\lambda'}$, our strategy is to use Schur's Lemma C.28 to show that K is bounded on ℓ_2. The ingredients are two bounds for $|K_{\lambda\lambda'}|$. To state the first, use (B.50) to bound $|K_{\lambda\lambda'}| \le C 2^{-|j'-j|/2} L_{\lambda\lambda'}$, where $L_{\lambda\lambda'}$ is the left side of the convolution bound

$$\int \frac{2^{j\wedge j'} dx}{(1+|2^j x - k|)^{1+\eta'}(1+|2^{j'}x - k'|)^{1+\eta'}} \le \frac{C}{(1+2^{j\wedge j'}|k'2^{-j'} - k2^{-j}|)^{1+\eta'}}, \quad \text{(B.53)}$$

verified in Exercise B.1. Denoting the right side by $C M_{\lambda\lambda'}^{1+\eta'}$, the first inequality states

$$|K_{\lambda\lambda'}| \le C_1 2^{-|j'-j|/2} M_{\lambda\lambda'}^{1+\eta'}. \quad \text{(B.54)}$$

For the next inequality, use the zero mean and Hölder hypotheses, (B.51) and (B.52), to argue, just as at (9.32) and (9.33), that for $j' \ge j$,

$$|K_{\lambda\lambda'}| \le C 2^{j(1/2+\eta)} \int |x - k'2^{-j'}|^\eta |w_{\lambda'}(x)| dx.$$

Using again (B.50) to bound $w_{\lambda'}$ and then $\eta < \eta'$ to assure convergence of the integral, we arrive at the second inequality

$$|K_{\lambda\lambda'}| \le C_2 2^{-|j'-j|(1/2+\eta)}. \quad \text{(B.55)}$$

The two bounds are combined by writing $|K_{\lambda\lambda'}|^{1-\theta}|K_{\lambda\lambda'}|^\theta$ and then using (B.54) in the first factor and (B.55) in the second to obtain

$$|K_{\lambda\lambda'}| \le C_3 2^{-|j'-j|(1/2+\delta)} M_{\lambda\lambda'}^{1+\delta} \quad \text{(B.56)}$$

by setting $\delta = \theta\eta$ for $\theta > 0$ sufficiently small that $1 + \delta < (1-\theta)(1+\eta')$.

We apply Schur's Lemma C.28 with weights $p_\lambda = q_\lambda = 2^{-j/2}$ so that, noting the symmetry of $K_{\lambda\lambda'}$, we need to show that $S_\lambda = 2^{j/2}\sum_{\lambda'} |K_{\lambda\lambda'}| 2^{-j'/2}$ is uniformly bounded in $\lambda = (jk)$. From (B.56) we need to bound

$$\sum_{j'} 2^{-(j'-j)/2 - |j'-j|(1/2+\delta)} \sum_{k'} M_{\lambda\lambda'}^{1+\delta}.$$

Consider the sum over k'. If $d = j' - j \ge 0$, then

$$2^{-d}\sum_{k'} M_{\lambda\lambda'}^{1+\delta} = \sum_{k'} \frac{2^{-d}}{(1+|k - 2^{-d}k'|)^{1+\delta}} \le 2^{-d} + \int \frac{dt}{(1+|t|)^{1+\delta}} \le C_\delta,$$

while if $j' < j$ with $\varepsilon = 2^{j'-j}$, the terms $M_{\lambda\lambda'}^{1+\delta} \le C(1+|k' - k\varepsilon|)^{-1-\delta}$ have sum over k' uniformly bounded in k and $\varepsilon \le 1$. Hence in both cases, $\sum_{k'} M_{\lambda\lambda'}^{1+\delta}$ is bounded by $C_\delta 2^{(j'-j)_+}$. Since $u + |u| - 2u_+ = 0$, we have $S_\lambda \le C\sum_j 2^{-\delta|j'-j|} \le C$ uniformly in λ as required.

(ii). The biorthogonality means that $\sum |\alpha_\lambda|^2 = \langle \sum \alpha_\lambda u_\lambda, \sum \alpha_\mu v_\mu \rangle$, and hence by Cauchy-Schwarz that

$$\|\alpha\|^2 \le \|\sum \alpha_\lambda u_\lambda\| \|\sum \alpha_\mu v_\mu\|.$$

From part (i), we have $\|\sum \alpha_\mu v_\mu\| \le C\|\alpha\|$, so it follows that $\|\sum \alpha_\lambda u_\lambda\| \ge C^{-1}\|\alpha\|$. Reverse the roles of u and v to establish the same lower bound for $\|\sum \alpha_\mu v_\mu\|$. □

Proof of Theorem 9.9 We abbreviate $\|f\|_{W_2^r}$ by $\|f\|_r$ and the sequence norm in (9.36) by $|||f|||_r^2$. The approach is to establish $\|f\|_r \le C\,|||f|||_r$ for $f \in V_J$ and then to use a density argument to complete the proof. For $f \in V_J$ we can differentiate term by term to get

$$D^r f = \sum_k \beta_k \varphi_{0k}^{(r)} + \sum_{j=0}^{J} \sum_k 2^{jr} \theta_{jk} \psi_{jk}^{(r)} = D^r f_0 + D^r f_1.$$

Under the hypotheses on ψ, it was shown in Section 12.3, example 1, that $\{(\psi^{(r)})_\lambda\}$ is a system of vaguelettes and hence by Proposition 12.3 satisfies the frame bounds (9.35). Apply the frame bound to conclude that $\|D^r f_1\|_2 \le C\,|||f|||_r$ and Lemma B.18 (for $p = 2$, $j = 0$ with orthogonality not required) to obtain $\|D^r f_0\|_2 \le C \sum \beta_k^2$. Putting these together, we get $\|f\|_r \le C\,|||f|||_r$ for $f \in V_J$. The density argument says that for $f \in W_2^r$, we have $P_J f \to f$ in L_2 and that $D^r P_J f$ is an L_2 Cauchy sequence (since $\|D^r(P_J f - P_K f)\|_2 \le C\,|||P_J f - P_K f|||_r$) so $P_J \to f$ in W_2^r.

In the other direction, for $f \in V_J$, we have $D^r f = \sum_{j \le J,k} 2^{jr} \psi_\lambda^{(r)}$, since the sum converges in L_2 at $J = -\infty$ from the frame bound. Hence

$$\sum_{j \ge 0,k} 2^{2rj} \theta_{jk}^2 \le \sum_{j \le J,k} (2^{rj} \theta_{jk})^2 \le C^2 \|D^r f\|_2^2,$$

while $\sum \beta_k^2 \le \|f\|_2^2$. Add the bounds to get $|||f|||_r^2 \le C^2 \|f\|_r^2$ and extend by density. □

B.5 Notes

Exercises

B.1 *Verification of* (B.53). (a) Set $t = 2^{j'}x - k$, $\rho = 2^{j-j'}$ and $\lambda = k - \rho k'$ and show that the inequality reduces to

$$\int_{-\infty}^{\infty} \frac{dt}{(1 + |\rho t - \lambda|)^\gamma (1 + |t|)^\gamma} \le \frac{C(\gamma)}{(1+\lambda)^\gamma}$$

for $\gamma = 1 + \eta' > 1$ and $0 < \rho \le 1, \lambda \in \mathbb{R}$.
(b) Show that for $\lambda \le 1$ this bound is immediate and for $\lambda \ge 1$ set $g(t) = (1+|\lambda - \rho t|)(1+|t|)$ and obtain the inequality from the bounds

$$g(t) \ge \begin{cases} (1+\lambda)(1+|t|) & t \le 0, \\ (1+\lambda/2)(1+t) & 0 \le t < \lambda/(2\rho), \\ (\lambda/2)(1+|t-\lambda/\rho|) & \lambda/(2\rho) \le t \le \lambda/\rho, \\ \lambda(1+t-\lambda/\rho) & t \ge \lambda/\rho. \end{cases}$$

Appendix C

Background Material

> The reader ... should not be discouraged, if on first reading of §0, he finds that he does not have the prerequisites for reading the prerequisites. (Paul Halmos, *Measure Theory*).

Here we collect bits of mathematical background, with references, that are used in the main text, but are less central to the statistical development (and so, in that important sense, are not prerequisites). Not a systematic exposition, this collection has two aims: initially to save the reader a trip to an authoritative source, and later, if that trip is needed, to point to what is required. References in brackets, like [§1.4], indicate sections of the main text that refer here.

C.1 Norms etc. Basic facts about normed linear spaces and in particular Hilbert spaces are found in many undergraduate analysis texts, e.g. Rudin (1976); Johnsonbaugh and Pfaffenberger (1981). In particular,

A *norm* $\|\cdot\|$ on a real or complex linear space X satisfies three properties: (i) (definiteness) $\|x\| = 0$ if and only if $x = 0$, (ii) (scaling) $\|ax\| = |a|\|x\|$ for any scalar a, and (iii) (triangle inequality) $\|x + y\| \le \|x\| + \|y\|$.

Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on X are called equivalent if there exist $C_1, C_2 > 0$ such that for all $x \in X$,

$$C_1\|x\|_1 \le \|x\|_2 \le C_2\|x\|_1.$$

A *semi-norm* $|\cdot|$ on X satisfies (ii) and (iii) but not necessarily the definiteness condition (i). For a *quasi-norm* $\|\cdot\|$ on X, the triangle inequality is replaced by

$$\|x + y\| \le C(\|x\| + \|y\|),$$

for some constant C, not depending on x or y.

Hilbert spaces etc. [§1.4] If $\{\varphi_i, i \in I\}$ is a complete orthonormal basis for $L_2(T)$, then f can be expanded as $\sum_i c_i\varphi_i$ with coefficients $c_i = \int f\bar{\varphi}_i$ that satisfy the Parseval relation

$$\int_T |f(t)|^2 dt = \sum_{i\in I} |c_i|^2. \tag{C.1}$$

C.2 Compact operators, Hilbert-Schmidt and Mercer theorems. [§3.9]

We begin with some definitions and notation, relying for further detail on Reed and Simon (1980, Ch. VI.5,6) and Riesz and Sz.-Nagy (1955, Ch. VI, §97,98).

Let $\mathcal{H}$ and $\mathcal{K}$ be Hilbert spaces, with the inner product denoted by $\langle \cdot, \cdot \rangle$, with subscripts $\mathcal{H}$ and $\mathcal{K}$ shown as needed. A linear operator $A : \mathcal{H} \to \mathcal{K}$ is bounded if $\|A\| = \sup\{\|Ax\|_{\mathcal{K}} : \|x\|_{\mathcal{H}} \leq 1\} < \infty$. The null space of A is $N(A) = \{x : Ax = 0\}$. The adjoint operator $A^* : \mathcal{K} \to \mathcal{H}$ is defined by the relations $\langle A^* y, x \rangle_{\mathcal{H}} = \langle y, Ax \rangle_{\mathcal{K}}$ for all $x \in \mathcal{H}, y \in \mathcal{K}$. Operator A is self-adjoint if $A^* = A$. We say that a bounded linear operator A is compact if A takes bounded sets to sets with compact closure, or equivalently, if for every bounded sequence $\{x_n\} \subset \mathcal{H}$, the sequence $\{Ax_n\}$ has a convergent subsequence.

Theorem C.3 (Hilbert-Schmidt) *Let A be a compact self-adjoint linear operator on $\mathcal{H}$. There exists a complete orthonormal basis $\{\varphi_n\}$ for $\mathcal{H}$ such that*

$$A\varphi_n = \lambda_n \varphi_n, \qquad \text{with } \lambda_n \in \mathbb{R} \text{ and } \lambda_n \to 0 \text{ as } n \to \infty.$$

The Singular Value Decomposition. Suppose $A : \mathcal{H} \to \mathcal{K}$ is linear and compact. Then $A^*A : \mathcal{H} \to \mathcal{H}$ is self-adjoint and compact, and so the Hilbert-Schmidt theorem yields an orthonormal set $\{\varphi_n\}$ with *positive* eigenvalues

$$A^*A\varphi_n = b_n^2 \varphi_n, \qquad b_n^2 > 0.$$

The set $\{\varphi_n\}$ need *not* be complete! However $A^*A = 0$ on the subspace $N(A) = N(A^*A)$ orthogonal to the closed linear span of $\{\varphi_n\}$. Define

$$\psi_n = \frac{A\varphi_n}{\|A\varphi_n\|} = b_n^{-1} A\varphi_n.$$

The set $\{\psi_n\}$ is orthonormal, and

$$A\varphi_n = b_n \psi_n, \qquad A^*\psi_n = b_n \varphi_n. \tag{C.2}$$

It can be verified that $\{\psi_n\}$ is a complete orthonormal basis for the closure of the range of A, and hence that for any $f \in \mathcal{H}$, using (C.2)

$$Af = \sum_n \langle Af, \psi_n \rangle \psi_n = \sum b_n \langle f, \varphi_n \rangle \psi_n. \tag{C.3}$$

Relations (C.2) and (C.3) describe the *singular value decomposition* of A, and $\{b_n\}$ are the singular values.

We have also

$$f = \sum b_n^{-1} \langle Af, \psi_n \rangle \varphi_n + u, \qquad u \in N(A). \tag{C.4}$$

In (C.3) and (C.4), the series converge in the Hilbert norms of $\mathcal{K}$ and $\mathcal{H}$ respectively.

C.4 *Kernels, Mercer's theorem.* [§3.10, §3.9] An operator $A \in \mathcal{L}(\mathcal{H})$ is *Hilbert-Schmidt* if for some orthobasis $\{e_i\}$

$$\|A\|_{HS}^2 = \sum_{i,j} |\langle e_i, Ae_j \rangle|^2 < \infty. \tag{C.5}$$

The value of $\|A\|_{HS}^2$ does not depend on the orthobasis chosen: regarding A as an infinite matrix, $\|A\|_{HS}^2 = \operatorname{tr} A^*A$. Hilbert-Schmidt operators are compact. An operator A is Hilbert-Schmidt if and only if its singular values are square summable.

Further, if $\mathcal{H} = L^2(T, d\mu)$, then A is Hilbert-Schmidt if and only if there is a square-integrable function $A(s,t)$ with

$$Af(s) = \int A(s,t) f(t) d\mu(t), \tag{C.6}$$

and in that case

$$\|A\|_{HS}^2 = \iint |A(s,t)|^2 d\mu(s) d\mu(t). \tag{C.7}$$

Suppose now that $T = [a,b] \subset \mathbb{R}$ and that $A : L^2(T, dt) \to L^2(T, dt)$ has kernel $A(s,t)$. The kernel $A(s,t)$ is called (i) *continuous* if $(s,t) \to A(s,t)$ is continuous on $T \times T$, (ii) *symmetric* if $A(s,t) = A(t,s)$, and (iii) *non-negative definite* if $(Af, f) \geq 0$ for all f.

These conditions imply that A is square-integrable, $\iint_{T\times T} A^2(s,t) ds dt < \infty$, and hence that A is self-adjoint, Hilbert-Schmidt and thus compact and so, by the Hilbert-Schmidt theorem, A has a complete orthonormal basis $\{\varphi_n\}$ of eigenfunctions with eigenvalues λ_n^2.

Theorem C.5 (Mercer) *If A is continuous, symmetric and non-negative definite, then the series*

$$A(s,t) = \sum_n \lambda_n^2 \varphi_n(s) \overline{\varphi_n(t)}$$

converges uniformly and in $L^2(T \times T)$.

[§12.2] In constructing the WVD in Chapter 12, in some cases it is necessary to consider possibly unbounded linear operators A defined on a dense subset $\mathcal{D}(A) \subset L_2(T)$. See, for example, Reed and Simon (1980, Ch. VIII). We give a useful criterion for the existence of *representers* g for linear functionals $\langle f, \psi \rangle$, in the sense that $[Af, g] = \langle f, \psi \rangle$. Let $\mathcal{R}(A)$ denote the range of A. The following formulation is from Donoho (1995) and Bertero (1989).

Proposition C.6 *Suppose that $A : \mathcal{D}(A) \subset L_2(T) \to L_2(U)$ with $\overline{\mathcal{D}(A)} = L_2(T)$ and that A is one to one. For a given $\psi \in L_2(T)$, the following are equivalent:*

(i) There exists $g \in L_2(U)$ such that

$$\langle f, \psi \rangle = [Af, g] \quad \textit{for all } f \in \mathcal{D}(A).$$

(ii) There exists C such that $\langle f, \psi \rangle \leq C \|Af\|_2$ for all $f \in \mathcal{D}(A)$.

(iii) $\psi \in \mathcal{R}(A^*)$.

Proof We prove (iii) $\Rightarrow$ (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii). If $\psi = A^* g$, then (i) follows from the definition of A^*. Then (i) $\Rightarrow$ (ii) follows from the Cauchy-Schwarz inequality with $C = \|g\|_2$.

(ii) $\Rightarrow$ (iii). The linear functional $Lh = \langle A^{-1} h, \psi \rangle$ is well defined on $\mathcal{R}(A)$ since A is one-to-one. From the hypothesis, for all $h = Af$, we have $|Lh| = |\langle f, \psi \rangle| \leq C \|h\|_2$. Thus L is bounded on $\mathcal{R}(A)$ and so extends by continuity to a bounded linear functional on $\overline{\mathcal{R}(A)}$. The Riesz representation theorem gives a $g \in \overline{\mathcal{R}(A)}$ such that

$$[Af, g] = L(Af) = \langle f, \psi \rangle \qquad \text{for all } f \in \mathcal{D}(A).$$

Since $\langle f, A^* g \rangle = \langle f, \psi \rangle$ for all f on a dense subset of $L_2(T)$, we recover $\psi = A^* g$. □

[§4.2, Lemma 4.7]. An extended form of the dominated convergence theorem, due to Young (1911) and rediscovered by Pratt (1960), has an easy proof, e.g. Bogachev (2007, Vol I, Theorem 2.8.8).

Theorem C.7 *If f_n, g_n and G_n are μ-integrable functions and*

(i) *$f_n \to f$, $g_n \to g$ and $G_n \to G$ a.e. (μ), with g and G integrable,*
(ii) *$g_n \le f_n \le G_n$ for all n, and*
(iii) *$\int g_n \to \int g$ and $\int G_n \to \int G$,*

then f is integrable, and $\int f_n \to \int f$.

Covariance inequality. [Exer. 4.1]. Let Y be a real valued random variable and suppose that $f(y)$ is increasing and $g(y)$ is decreasing. Then, so long as the expectations exist,

$$E[f(Y)g(Y)] \le E[f(Y)]E[g(Y)]. \tag{C.8}$$

For a simple coupling proof, see (Thorisson, 1995, Sec. 2).

Jensen's inequality. We begin with the standard version and then note some extensions. (a) If $f : \mathbb{R} \to \mathbb{R}$ is convex and X is a real valued random variable, then

$$f(EX) \le Ef(X) \tag{C.9}$$

provided both expectations exist.

(b) If f is a convex real valued function on a convex set $C \subset \mathbb{R}^n$, the random vector $X \in C$ almost surely and EX exists, then (C.9) holds (Ferguson, 1967, p.67).

(c) If f is a convex, lower semicontinuous, extended-real valued function on a closed convex set C in a locally convex topological vector space, and the random vector $X \in C$ is (Pettis) integrable and $Ef(X)$ exists, then (C.9) holds (Perlman, 1974, Thm. 3.10).

C.8 *Analytic functions.* A function is analytic if it is given locally by a convergent power series. If the set of zeros of an analytic function has an accumulation point inside its domain, then the function is zero everywhere on the connected component containing the accumulation point. This result is true both for complex and real analytic functions, defined respectively on a domain in $\mathbb{C}$ or $\mathbb{R}$. For real analytic functions, see Krantz and Parks (2002, Ch. 1).

C.9 [§7.1, §12.2, §B.1]. The Fourier transform of an integrable function on $\mathbb{R}$ is defined by

$$\widehat{f}(\xi) = \int_{-\infty}^{\infty} f(x)e^{-i\xi x}dx. \tag{C.10}$$

The Fourier transform of a convolution

$$f \star g(x) = \int f(x-y)g(y)dy$$

is just the product of the Fourier transforms:

$$\widehat{f \star g}(\xi) = \widehat{f}(\xi)\widehat{g}(\xi). \tag{C.11}$$

If f is sufficiently nice, for example if both f and $\widehat{f}$ are integrable, (cf. Folland (1999, Sec. 8.3)), it may be recovered from the inversion formula[1]

$$f(x) = \frac{1}{2\pi}\int_{\infty}^{\infty} \widehat{f}(\xi)e^{i\xi x}d\xi.$$

If $x^k f(x)$ is integrable, then the derivatives of $\widehat{f}(\xi)$ can be expressed in terms of moments: $\widehat{f}^{(k)}(\xi) = (-i)^k \int x^k f(x)e^{-i\xi x}dx$. It follows that the function f has *p vanishing moments*, that is $\int x^k f(x)dx = 0$ for $k = 0, 1, \ldots, p-1$, exactly when the derivatives $\widehat{f}^{(k)}(0) = 0$ for $k = 0, 1, \ldots, p-1$.

The Parseval (or Plancherel) identity states that if $f, g \in L_1 \cap L_2$,

$$\int f(x)\overline{g(x)}dx = \frac{1}{2\pi}\int \widehat{f}(\xi)\overline{\widehat{g}(\xi)}d\xi. \tag{C.12}$$

A periodic function f in $L_2[0,1]$ has Fourier expansion

$$f(x) = \sum_{k\in\mathbb{Z}} c_k e^{2\pi i k x}.$$

The Fourier coefficients $c_k = \int_0^1 f(x)e^{-2\pi i k x}dx$, and satisfy the Parseval relation

$$\int_0^1 |f|^2(x)dx = \sum_{k\in\mathbb{Z}} |c_k|^2.$$

[§3.5, §14.4]. The Poisson summation formula (Folland, 1999, Sec. 8.3) states that if $(1+|x|^2)[|f(x)|$ and $(1+|\xi|^2)|\widehat{f}(\xi)|$ are bounded, then

$$\sum_{j\in\mathbb{Z}} f(j) = \sum_{k\in\mathbb{Z}} \widehat{f}(2\pi k). \tag{C.13}$$

[Dym and McKean (1972, p. 111) gives a sufficient condition on f (or $\widehat{f}$) alone.]
When applied to $f(x) = g(x+t)$, this yields a representation for the periodization of g

$$\sum_j g(t+j) = \sum_k e^{2\pi i k t}\widehat{g}(2\pi k), \qquad t \in \mathbb{R}. \tag{C.14}$$

C.10 The characteristic function of a probability measure is $\widehat{\pi}(\xi) = \int e^{i\xi\theta}d\pi(\theta)$. Note the opposite sign convention from (C.10). The convolution property (C.11) extends to convolution of probability measures: $\widehat{\pi \star \nu}(\xi) = \widehat{\pi}(\xi)\widehat{\nu}(\xi)$.

The characteristic function of an $N(\mu, \sigma^2)$ distributions is $\exp\{i\mu\xi - \sigma^2\xi^2/2\}$. It follows from the convolution property that if the convolution $\pi \star \nu$ of two probability measures is Gaussian, and if one of the factors is Gaussian, then so must be the other factor.

[1] There are several conventions for the placement of factors involving 2π in the definition of the Fourier transform, Folland (1999, p. 278) has a comparative discussion.

Some further properties of the Gaussian distribution. [§2.8].

A standard Gaussian variate, $Z \sim N(0,1)$, has density function

$$\phi(z) = (2\pi)^{-1/2} e^{-z^2/2},$$

and distribution functions

$$\Phi(z) = \int_{-\infty}^{z} \phi(u)du, \quad \tilde{\Phi}(z) = \int_{z}^{\infty} \phi(u)du.$$

From $\phi(u) \le uz^{-1}\phi(u)$, we obtain the simplest bound for *Mills ratio*, (Mills, 1926),

$$\tilde{\Phi}(z)/\phi(z) \le z^{-1} \qquad (z > 0). \tag{C.15}$$

Lemma C.11 *. (a) If $X \sim N_n(\mu, \Sigma)$ and M is an $m \times n$ matrix, then $MX \sim N_m(M\mu, M\Sigma M^T)$.*

(b) If $X \sim N_n(0, \sigma^2 I)$ and U is an $n \times n$ orthogonal matrix, then $UX \sim N_n(0, \sigma^2 I)$ also.

[§8.9, §8.10]. The moment generating function of a standard Gaussian variable is

$$Ee^{\beta z} = e^{\beta^2/2}. \tag{C.16}$$

Proposition C.12 (Talagrand (2003), Proposition 1.1.4.) *Let $z_1, \ldots, z_n \sim N(0,1)$ (not necessarily independent). Then*

$$E \log\Big(\sum_1^n e^{\beta z_i}\Big) \le \begin{cases} \frac{1}{2}\beta^2 + \log n & \text{if } \beta \le \sqrt{2\log n} \\ \beta\sqrt{2\log n} & \text{if } \beta \ge \sqrt{2\log n} \end{cases} \tag{C.17}$$

and, as a consequence,

$$E \max_{i \le n} z_i \le \sqrt{2\log n}. \tag{C.18}$$

C.13 *Brownian motion, Wiener integral.* [§1.4, §3.10]. 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.

The following basic facts about Brownian motion and Wiener integrals may be found, for example, in Kuo (2006, Ch. 2). Standard Brownian motion on the interval $[0,1]$ is defined as a Gaussian process $\{W(t)\}$ with mean zero and covariance function Cov $(W(s), W(t)) = s \wedge t$. It follows that $\{W(t)\}$ has independent increments: if $0 \le t_1 < t_2 < \cdots < t_n$, then the increments $W(t_j) - W(t_{j-1})$ are independent. In addition, the sample paths $t \to W(t, \omega)$ are continuous with probability one.

The *Wiener integral* $X = I(f) = \int_0^1 f(t)dW(t)$ of a deterministic function f is defined first for step functions and then for $f \in L_2[0,1]$ by convergence of random variables in the Hilbert space $L_2(\Omega)$ with inner product $\langle X, Y\rangle = EXY$. We have $EI(f) = 0$ and the identity

$$\langle f, g\rangle_{L_2[0,1]} = EI(f)I(g) \tag{C.19}$$

holds, and $I(f) \sim N(0, \|f\|_2^2)$. If f is continuous and of bounded variation, then $I(f)$ can be interpreted as a Riemann-Stieltjes integral.

If $\{\varphi_i\}$ is an orthonormal basis for $L_2[0,1]$, then $f = \sum \langle f, \varphi_i \rangle \varphi_i$ and

$$I(f) = \sum \langle f, \varphi_i \rangle I(\varphi_i),$$

where the variables $z_i = I(\varphi_i)$ are i.i.d. standard Gaussian, and the series converges almost surely. In particular,

$$W(t) \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\infty} z_i \int_0^t \phi_i(s) ds$$

with the series converging almost surely (Shepp, 1966). Particular examples for which this representation was known earlier include the trigonmetric 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).

If $C(s,t)$ is a square integrable kernel on $L_2([0,1]^2)$, then the Gaussian random function $F(s) = \int_0^1 C(s,t) dW(t) \in L_2[0,1]$ almost surely, having mean zero and finite variance $\int_0^1 C^2(s,t)dt$ for almost all $s \in [0,1]$. If $C(s,t)$ has the expansion $\sum_i c_i \varphi_i(s)\varphi_i(t)$ with square summable coefficients $\sum_i c_i^2 < \infty$, then $F(s) = \sum_i c_i I(\varphi_i)\varphi_i(s)$.

[§8.6]. **Weak law of large numbers for triangular arrays.** Although designed for variables without finite second moment, the truncation method works well for the cases of rapidly growing variances that occur here. The following is taken from Durrett (2010, Thm 2.2.6).

Proposition C.14 *For each n let $X_{nk}, 1 \le k \le n$, be independent. Let $b_n > 0$ with $b_n \to \infty$, and let $\bar{X}_{nk} = X_{nk} I\{|X_{nk}| \le b_n\}$. Suppose that as $n \to \infty$,*
(i) $\sum_{k=1}^n P(|X_{nk}| > b_n) \to 0$, and
(ii) $b_n^{-2} \sum_{k=1}^n E\bar{X}_{nk}^2 \to 0$ as $n \to \infty$.
Let $S_n = X_{n1} + \ldots + X_{nn}$ and put $a_n = \sum_{k=1}^n E\bar{X}_{nk}$. Then

$$S_n = a_n + o_p(b_n).$$

C.15 For the most basic definitions of metric spaces, completeness and separability, we refer to textbooks of analysis and or Wikipedia, or the metric space Appendix M of Billingsley (1999).

Let $\mathcal{X}$ be a separable metric space and μ a probability measure in $\mathcal{X}$. Then there exists a unique closed set F, called the *support of* μ satisfying $\mu(F) = 1$ such that if K is any closed set with $\mu(K) = 1$, then $F \subset K$, e.g. Parthasarathy (1967, Thm. 2.1).

C.16 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 ϵ 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.

[§4.2, §4.4]. 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 1st countable, then these conditions may be rewritten in terms of sequences as $f(x) \leq \liminf f(x_n)$ whenever $x_n \to x$.]

If $\{f_\alpha : \alpha \in A\}$ is a set of lower semicontinous functions, then the pointwise supremum

$$f(x) = \sup_{\alpha \in A} f_\alpha(x)$$

is lower semicontinuous, e.g. Folland (1999, p.218).

A function g is *upper semicontinuous* if $f = -g$ is lsc.

C.17 If X is compact, then an lsc function f attains its infimum: $\inf_{x \in X} f = f(x_0)$ for some $x_0 \in X$, e.g. (Royden, 1988, p.195).

C.18 *Weak convergence of probability measures.* [§4.4]. Let Ω be a complete separable metric space–for us, usually a subset of $\mathbb{R}^n$ for some n. Let $\mathcal{P}(\Omega)$ denote the collection of probability measures on Ω with the Borel $\sigma-$algebra generated by the open sets. We say that $\pi_n \to \pi$ in the weak topology if

$$\int \psi d\pi_n \to \int \psi d\pi \tag{C.20}$$

for all bounded continuous $\psi : \Omega \to \mathbb{R}$.

When $\Omega = \mathbb{R}$ or $\mathbb{R}^d$, the Lévy-Cramér theorem provides a convergence criterion in terms of the characteristic function $\widehat{\pi}(\xi) = \int e^{-i\xi\theta}\pi(d\theta)$, namely that $\pi_n \to \pi$ weakly if and only if $\widehat{\pi}_n(\xi) \to \widehat{\pi}(\xi)$ for all ξ with $\widehat{\pi}(\xi)$ being continuous at 0 (Cramér, 1999, p. 102), (Chung, 1974, p.101).

A collection $\mathcal{P} \subset \mathcal{P}(\Omega)$ is called *tight* if for all $\epsilon > 0$, there exists a compact set $K \subset \Omega$ for which $\pi(K) > 1 - \epsilon$ for every $\pi \in \mathcal{P}$.

Prohorov's theorem (Billingsley, 1999, Ch. 1.5) provides a convenient description of compactness in $\mathcal{P}(\Omega)$: a set $\mathcal{P} \subset \mathcal{P}(\Omega)$ has compact closure if and only if $\mathcal{P}$ is tight.

Thus, if $\Omega = [-\tau, \tau]$ then $\mathcal{P}(\Omega)$ has compact closure. If $\Omega = \mathbb{R}$ and $\mathcal{P} = \{\pi : \int |\theta|^p \pi(d\theta) \leq \eta^p\}$, then Markov's inequality shows that $\pi([-M, M]^c) \leq \eta^p / M^p$ for any $\pi \in \mathcal{P}$, so that $\mathcal{P}$ is tight and hence weakly compact.

C.19 *Vague convergence.* [§4.4]. Let $\Omega = \mathbb{R}$ and $\mathcal{P}_+(\mathbb{R})$ be the collection of *sub-stochastic* measures on $\mathbb{R}$. Equivalently, $\mathcal{P}_+ = \mathcal{P}(\bar{\mathbb{R}})$ for $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, allowing mass at $\pm\infty$. We say that $\pi_n \to \pi$ in the *vague* topology if (C.20) holds for all continuous ψ with compact support, or (equivalently) for all continuous ψ that vanish at $\pm\infty$.

Clearly weak convergence implies vague convergence, and if $\mathcal{P} \subset \mathcal{P}(\mathbb{R})$ is weakly compact, then it is vaguely compact. However $\mathcal{P}(\mathbb{R})$ is not weakly compact (as mass can escape to $\pm\infty$) but $\mathcal{P}_+(\mathbb{R})$ *is* vaguely compact, e.g. from Prohorov's theorem applied to $\mathcal{P}(\bar{\mathbb{R}})$.

C.20 [§4.2, §8.7]. The *Fisher information* for location of a distribution P on $\mathbb{R}$ is

$$I(P) = \sup_{\psi} \frac{\left(\int \psi' dP\right)^2}{\int \psi^2 dP}, \tag{C.21}$$

where the supremum is taken over the set C_0^1 of C^1 functions of compact support for which $\int \psi^2 dP > 0$. For this definition and the results quoted here, we refer to Huber and Ronchetti (2009, Chapter 4), [HR] below.

It follows from this definition that $I(P)$ is a convex function of P. The definition is however equivalent to the usual one: $I(P) < \infty$ if and only if P has an absolutely continuous density p, and $\int p'^2/p < \infty$. In either case, $I(P) = \int p'^2/p$.

Given P_0, P_1 with $I(P_0), I(P_1) < \infty$ and $0 \le t \le 1$, let $P_t = (1-t)P_0 + tP_1$. Differentiating $I(P_t) = \int p_t'^2/p_t$ under the integral sign (which is justified in HR), one obtains

$$\begin{aligned} \frac{d}{dt} I(P_t)|_{t=0} &= \int \frac{2p_0'}{p_0}(p_1' - p_0') - \frac{p_0'^2}{p_0^2}(p_1 - p_0) \\ &= \int [-2\psi_0 p_1' - \psi_0^2 p_1] dx - I(P_0), \end{aligned} \tag{C.22}$$

where we have set $\psi_0 = -p_0'/p_0$ for terms multiplying p_1' and p_1 and observed that the terms involving only p_0' and p_0 collapse to $-I(P_0)$.

Since $I(P)$ is the supremum of a set of vaguely (resp. weakly) continuous functions, it follows that $P \to I(P)$ is vaguely (resp. weakly) lower semicontinuous[2]. Consequently, from C.17, if $\mathcal{P} \subset \mathcal{P}_+(\mathbb{R})$ is vaguely compact, then there is an $P_0 \in \mathcal{P}$ minimizing $I(P)$.

Formula (C.22) yields a helpful variational criterion for characterizing a minimizing P_0. Let $\mathcal{P}_1 = \{P_1 \in \mathcal{P} : I(P_1) < \infty\}$ and for given P_0 and P_1, let $P_t = (1-t)P_0 + tP_1$. Since $I(P)$ is convex in P, a distribution $P_0 \in \mathcal{P}$ minimizes $I(P)$ if and only if $(d/dt)I(P_t) \ge 0$ at $t = 0$ for each $P_1 \in \mathcal{P}_1$.

A slight reformulation of this criterion is also useful. The first term on the right side of (C.22) is $\int -2\psi_0(p_1' - p_0') = \int 2\psi_0'(p_1 - p_0)$ and so P_0 minimizes $I(P)$ over $\mathcal{P}$ if and only if

$$\int [2\psi_0' - \psi_0^2](p_1 - p_0) \ge 0. \tag{C.23}$$

C.21 (Uniqueness). Suppose (i) that $\mathcal{P}$ is convex and $P_0 \in \mathcal{P}$ minimizes $I(P)$ over $\mathcal{P}$ with $0 < I(P_0) < \infty$, and (ii) that the set on which p_0 is positive is an interval and contains the support of every $P \in \mathcal{P}$. Then P_0 is the unique minimizer of $I(P)$ in $\mathcal{P}$.

In our applications, P is typically the marginal distribution $\Phi \star \pi$ for a (substochastic) prior measure π. (For this reason, the notation uses $\mathcal{P}^\star$ for classes of distributions P, which in these applications correspond to classes $\mathcal{P}$ of priors through $\mathcal{P}^\star = \{P = \Phi \star \pi, \pi \in \mathcal{P}\}$.)

[2] indeed, if $V_\psi(P)$ denotes the ratio in (C.21), then $\{P : I(P) > t\}$ is the union of sets of the form $\{P : V_\psi(P) > t, \int \psi^2 dP > 0\}$ and hence is open.

In particular, in the uniqueness result, p_0 is then positive on all of $\mathbb{R}$ and so condition (ii) holds trivially.

C.22 Stein's Unbiased Estimate of Risk. [§2.6]. We provide some extra definitions and details of proof for the unbiased risk identity that comprises Proposition 2.6. 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 and Trudinger (1983, Chapter 7), for omitted details.

A function $g : \mathbb{R}^n \to \mathbb{R}$ is said to be *weakly differentiable* if there exist functions $h_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots n$, such that

$$\int \psi h_i = -\int (D_i \psi) g \qquad \text{for all } \psi \in C_0^\infty,$$

where C_0^∞ denotes the class of C^∞ functions on $\mathbb{R}^n$ 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 $\bar{g}$ that is absolutely continuous on almost all line segments parallel to the co-ordinate axes and whose (classical) partial derivatives (which consequently exist almost everywhere) are locally integrable (e.g. Ziemer (1989, Thm. 2.1.4)).

For approximation arguments, such as in the proof of Proposition 2.6 below, it is convenient to use the following criterion (e.g. Gilbarg and Trudinger (1983, Thm 7.4)): Suppose that g and h are integrable on compact subsets of $\mathbb{R}^n$. Then $h = D_i g$ if and only if there exist C^∞ functions $g_m \to g$ such that also $D_i g_m \to h$ where in both cases the convergence is in L_1 on compact subsets of $\mathbb{R}^n$. [Exercise 2.24 outlines a key part of the proof.]

A C^r *partition of unity* is a collection of C^r functions $\rho_m(x) \geq 0$ of compact support such that for every $x \in \mathbb{R}^n$ we have $\sum_m \rho_m(x) = 1$ and on some neighborhood of x, all but finitely many $\rho_m(x) = 0$. We add the nonstandard requirement that for some $C < \infty$,

$$\sum_m |D_i \rho_m(x)| \leq C \qquad \text{for all } x. \tag{C.24}$$

Exercise C.1 adapts a standard construction (e.g. Rudin (1973, Thm 6.20)) to exhibit an example that suffices for our needs.

Proof of Proposition 2.6 First note that by a simple translation of parameter, it suffices to consider $\mu = 0$. Next, consider scalar C^∞ functions $g : \mathbb{R}^n \to \mathbb{R}$ of compact support. We aim to show that $E[X_i g(X)] = E[D_i g(X)]$, but this is now a simple integration by parts:

$$\begin{aligned} \int x_i g(x)\phi(x)dx &= \int g(x)[-D_i\phi(x)]dx \\ &= \int D_i g(x)\phi(x)dx. \end{aligned} \tag{C.25}$$

Now use the criterion quoted above to extend to weakly differentiable g with compact support: use that fact that for compact $K \subset \mathbb{R}^n$, convergence $f_m \to f$ in $L_1(K)$ implies $f_m h \to f h$, also in $L_1(K)$, for any function h bounded on K (such as x_i or $\phi(x)$).

Finally for extension to weakly differentiable g satisfying $E|X_i g(X)| + |D_i g(X)| < \infty$, let $\{\rho_m\}$ be a C^r partition of unity satsifying (C.24). Let $g_m = g(\rho_1 + \cdots + \rho_m)$. Equality (C.25) extends from the compactly supported g_m to g after a few uses of the dominated convergence theorem.

For a vector function $g : \mathbb{R}^n \to \mathbb{R}^n$, just apply the preceding argument to the components g_i and add. Formula (2.58) follows immediately from (2.57) (since $E_\mu \|X - \mu\|^2 = n$). □

C.23 Hölder spaces. [§4.7, §7.1, §9.6, §B.3]. The Hölder spaces $C^\alpha(I)$ measure smoothness *uniformly* on an interval I, with smoothness parameter α. The norms have the form $\|f\|_{C^\alpha} = \|f\|_{\infty,I} + |f|_\alpha$, with the sup norm added because the seminorm $|f|_\alpha$–which reflects the dependence on α–will typically vanish on a finite dimensional space.

If α is a positive integer, then we require that f have α continuous derivatives, and set $|f|_\alpha = \|D^\alpha f\|_{\infty,I}$.

For $0 < \alpha < 1$, we require finiteness of

$$|f|_\alpha = \sup\left\{\frac{|f(x) - f(y)|}{|x - y|^\alpha},\ x, y \in I\right\}. \tag{C.26}$$

If m is a positive integer and $m < \alpha < m + 1$, then we require both that f have m uniformly continuous derivatives and also finiteness of

$$|f|_\alpha = |D^m f|_{\alpha - m}.$$

We note also that Hölder functions can be uniformly approximated by (Taylor) polynomials. Indeed, we can say that $f \in C^\alpha(I)$ implies that there exists a constant C such that for each $x \in I$, there exists a polynomial $p_x(y)$ of degree $\lceil \alpha \rceil - 1$ such that

$$|f(x + y) - p_x(y)| \le C|y|^\alpha, \qquad \text{if } x + y \in I. \tag{C.27}$$

The constant C can be taken as $|f|_\alpha / c_\alpha$, where c_α equals 1 if $0 < \alpha < 1$ and equals $\prod_{j=0}^{\lceil \alpha \rceil - 1} (\alpha - j)$ if $\alpha \ge 1$.

C.24 Total Variation. [§ 9.6] When $I = [a, b]$, this semi-norm is defined by

$$|f|_{TV(I)} = \sup\{\sum_{i=1}^{n} |f(t_i) - f(t_{i-1})| \ :\ a = t_0 < t_1 < \cdots < t_n = b, n \in \mathbb{N}\}.$$

The corresponding norm $\|f\|_{TV} = \|f\|_1 + |f|_{TV}$. The space TV represents a scientifically interesting enlargement of W_1^1, since when $f \in W_1^1$, we may write

$$|f|_{TV} = \int |Df|, \tag{C.28}$$

but this identity obviously fails for discontinuous piecewise constant functions in TV.

C.25 Sobolev spaces. Let I be a (possibly unbounded) interval of $\mathbb{R}$ and denote by $W_p^k(I)$ the space of functions in $L_p(I)$ for which $D^{k-1}f$ is absolutely continuous on I with $D^k f \in L_p(I)$. The corresponding norm is defined by $\|f\|_{W_p^k} = \|f\|_p + \|D^k f\|_p$, or equivalently by $\|f\|_{W_p^k}^p = \|f\|_p^p + \|D^k f\|_p^p$.

The terminology "f is r-fold differentiable in p-th mean" is sometimes used for functions in Sobolev spaces W_p^r. To explain this, we remark that a function h on $\mathbb{R}$ is called a (strong) L_p derivative of f if $\int |\delta^{-1}[f(x+\delta)-f(x)]-h(x)|^p dx \to 0$ as $\delta \to 0$. The L_p derivative of f, call it h, exists if and only if f is absolutely continuous on bounded intervals and $Df \in L_p$, in which case $h = Df$ (e.g. Folland (1999, p.246).)

[§B.3]. **Background.** For convenience, we record a straightforward extension of Young's inequality for convolutions.

Theorem C.26 *Let $(X, \mathcal{B}_X, \mu)$ and $(Y, \mathcal{B}_Y, \nu)$ be σ–finite measure spaces, and let $K(x, y)$ be a jointly measurable function. Suppose that*

$$(i) \qquad \int |K(x,y)|\mu(dx) \le M_1 \qquad \text{a.e. } (\nu), \qquad \text{and}$$

$$(ii) \qquad \int |K(x,y)|\nu(dy) \le M_2 \qquad \text{a.e. } (\mu).$$

For $1 \le p \le \infty$, the operator

$$(Kf)(x) = \int K(x,y) f(y) \nu(dy)$$

maps $L_p(Y) \to L_p(X)$ with

$$\|Kf\|_p \le M_1^{1/p} M_2^{1-1/p} \|f\|_p. \tag{C.29}$$

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:

$$|Kf(x)| \le \Big[\int |K(x,y)|\nu(dy)\Big]^{1/q} \Big[\int |K(x,y)||f(y)|^p \nu(dy)\Big]^{1/p},$$

so that, using (ii),

$$|Kf(x)|^p \le M_2^{p/q} \int |K(x,y)||f(y)|^p \nu(dy).$$

Now integrate over x, use Fubini's theorem and bound (i) to obtain (C.29). The proof for $p = 1$ is similar and easier. □

Remark. The adjoint $(K^* g)(y) = \int g(x) K(x,y) \mu(dx)$ maps $L_p(X) \to L_p(Y)$ with

$$\|K^* g\|_p \le M_1^{1-1/p} M_2^{1/p} \|g\|_p. \tag{C.30}$$

Two traditional forms of Young's inequality are immediate consequences.

Corollary C.27 (§12.3, §B.3) . *Suppose that* $1 \le p \le \infty$.
(i) If $Kf(x) = \int_{\infty}^{\infty} K(x-y)f(y)dy$, *then*

$$\|Kf\|_p \le \|K\|_1 \|f\|_p. \tag{C.31}$$

(ii) If $c_k = \sum_{j\in\mathbb{Z}} a_{k-j}b_j$, *then*

$$\|c\|_p \le \|a\|_1 \|b\|_p. \tag{C.32}$$

Another consequence, in the L_2 setting, is a version with weights. Although true in the measure space setting of Theorem C.26, we need only the version for infinite matrices.

Corollary C.28 (Schur's Lemma) *[§15.3, §B.4]. Let* $K = (K(i,j))_{i,j\in\mathbb{N}}$ *be an infinite matrix and let* $(p(i))$ *and* $(q(j))$ *be sequences of positive numbers. Suppose that*

$$\text{(i)} \qquad \sum_i p(i)K(i,j) \le M_1 q(j) \qquad j \in \mathbb{N}, \qquad \text{and}$$

$$\text{(ii)} \qquad \sum_j K(i,j)q(j) \le M_2 p(i) \qquad i \in \mathbb{N},$$

Then the operator $(Kb)(i) = \sum_j K(i,j)b(j)$ *is bounded on* ℓ_2 *and*

$$\|Kb\|_2 \le \sqrt{M_1 M_2}\|b\|_2.$$

Proof Use the argument for Theorem C.26, this time expanding $|K(i,j)|$ as

$$|K(i,j)|^{1/2} q(j)^{1/2} \cdot |K(i,j)|^{1/2} q(j)^{-1/2}. \qquad \square$$

Theorem C.29 (Minkowski's integral inequality) *[§B.3]. Let* $(X, \mathcal{B}_X, \mu)$ *and* $(Y, \mathcal{B}_Y, \nu)$ *be* σ*–finite measure spaces, and let* $f(x,y)$ *be a jointly measurable function. Then for* $1 \le p \le \infty$,

$$\left(\int \left|\int f(x,y)\nu(dy)\right|^p \mu(dx)\right)^{1/p} \le \int \left(\int |f(x,y)|^p \mu(dx)\right)^{1/p} \nu(dy). \tag{C.33}$$

See, e.g. Folland (1999, p. 194).

C.30 *Gauss' hypergeometric function* [§3.9]. is defined for $|x| < 1$ by the series

$$F(\alpha,\beta,\gamma;x) = \sum_{n=0}^{\infty} \frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\frac{x^n}{n!},$$

provided that $\gamma \ne 0, -1, -2, \ldots$; and $(\alpha)_n = \alpha(\alpha+1)(\alpha+2)\cdots(\alpha+n-1)$, $(\alpha)_0 = 1$ is the Pochhammer symbol. For $\operatorname{Re}\gamma > \operatorname{Re}\beta > 0$ and $|x| < 1$, Euler's integral representation says that

$$F(\alpha,\beta,\gamma;x) = B(\beta,\gamma-\beta)^{-1}\int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-tx)^{-\alpha}dt,$$

where $B(\beta,\gamma) = \Gamma(\beta)\Gamma(\gamma)/\Gamma(\beta+\gamma)$ is the beta integral. These and most identities given

here may be found in Abramowitz and Stegun (1964, Chs. 15, 22) See also Temme (1996, Chs. 5 and 6) for some derivations. Gel'fand and Shilov (1964, §5.5) show that this formula can be interpreted in terms of differentiation of fractional order

$$\frac{x^{\gamma-1}}{\Gamma(\gamma)}F(\alpha,\beta,\gamma;x) = D^{\beta-\gamma}\left(\frac{x_+^{\beta-1}(1-x)_+^{-\alpha}}{\Gamma(\beta)}\right). \tag{C.34}$$

They then show that the identity $D^{-\delta}D^{\beta-\gamma} = D^{\beta-\gamma-\delta}$ becomes, in integral form

$$x^{\gamma+\delta-1}F(\alpha,\beta,\gamma+\delta;x) = B(\gamma,\delta)^{-1}\int_0^x t^{\gamma-1}F(\alpha,\beta,\gamma;t)(x-t)^{\delta-1}dt. \tag{C.35}$$

C.31 *Jacobi polynomials* arise from the hypergeometric function when the series is finite

$$P_n^{a,b}(1-2x) = \binom{n+a}{n}F(-n,a+b+n+1,a+1;x),$$

where the generalized binomial coefficient is $\Gamma(n+a+1)/\Gamma(n+1)\Gamma(a+1)$. The polynomials $P_n^{a,b}(w), n \geq 0$ are orthogonal with respect to the weight function $(1-w)^a(1+w)^b$ on $[-1,1]$. Special cases include the *Legendre* polynomials $P_n(x)$, with $a = b = 0$, and the *Chebychev* polynomials $T_n(x)$ and $U_n(x)$ of first and second kinds, with $a = b = -1/2$ and $a = b = 1/2$ respectively.

The orthogonality relations, for the corresponding weight function on $[0,1]$, become

$$\int_0^1 P_m^{a,b}(1-2x)P_n^{a,b}(1-2x)\,x^a(1-x)^b dx = g_{a,b;n}^2\delta_{nm},$$

where the Kronecker $\delta_{nm} = 1$ if $n = m$ and 0 otherwise.

$$g_{a,b;n}^2 = \frac{n!}{2n+a+b+1}\frac{\Gamma(a+b+n+1)}{\Gamma(a+n+1)\Gamma(b+n+1)}. \tag{C.36}$$

Exercises

C.1 *(Partition of unity for Proof of Proposition 2.6.)* For $x \in \mathbb{R}^n$, let $\|x\|_\infty = \max|x_k|$.
(a) Exhibit a C^r function $\eta(x) \geq 0$ for which $\eta(x) = 1$ for $\|x\|_\infty \leq 1$ and $\eta(x) = 0$ for $\|x\|_\infty \geq 2$. (Start with $n = 1$.)
(b) Let $p_i, i = 1,2,\ldots$ be an enumeration of the points in $\mathbb{Z}^n$, and set $\eta_i(x) = \eta(x-p_i)$. Let $\rho_1 = \eta_1$ and for $i \geq 1$, $\rho_{i+1} = (1-\eta_1)\cdots(1-\eta_i)\eta_{i+1}$. Show that

$$\rho_1 + \cdots + \rho_m = 1 - (1-\eta_1)\cdots(1-\eta_m),$$

and hence that $\{\rho_i(x)\}$ is a C^r partition of unity and in particular that for all x there exists $C < \infty$ such that

$$|D_i[\rho_1 + \cdots + \rho_m](x)| \leq \sum_{j=1}^m |D_i\eta_j(x)| \leq C.$$

Bibliography

[]

Abel, N. 1826. Resolution dun probleme de mecanique. *J. Reine u. Angew. Math*, **1**, 153–157. [84]

Abramovich, F., and Silverman, B. W. 1998. Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, **85**, 115–129. [347]

Abramovich, F., Sapatinas, T., and Silverman, B. W. 1998. Wavelet thresholding via a Bayesian approach. *J. Royal Statistical Society, Series B.*, **60**, 725–749. [199]

Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. 2006. Adapting to Unknown Sparsity by controlling the False Discovery Rate. *Annals of Statistics*, **34**, 584–653. [205, 206, 327, 328]

Abramowitz, M., and Stegun, I. A. 1964. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. National Bureau of Standards Applied Mathematics Series, vol. 55. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. [442]

Adler, R. J., and Taylor, J. E. 2007. *Random fields and geometry*. Springer Monographs in Mathematics. New York: Springer. [51]

Anderson, G. W., Guionnet, A., and Zeitouni, O. 2010. *An Introduction to Random Matrices*. Cambridge University Press. [134]

Ash, R. B., and Gardner, M. F. 1975. *Topics in Stochastic Processes*. Academic Press. [88]

Assouad, P. 1983. Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.*, **296**(23), 1021–1024. [281]

Beckner, W. 1989. A generalized Poincaré inequality for Gaussian measures. *Proc. Amer. Math. Soc.*, **105**(2), 397–400. [51]

Belitser, E., and Levit, B. 1995. On Minimax Filtering over Ellipsoids. *Mathematical Methods of Statistics*, **4**, 259–273. [135, 157]

Berger, J. O. 1985. *Statistical decision theory and Bayesian analysis*. Second edn. Springer Series in Statistics. New York: Springer-Verlag. [134]

Bergh, J., and Löfström, J. 1976. *Interpolation spaces – An Introduction*. New York: Springer Verlag. [417]

Berkhin, P., and Levit, B. Y. 1980. Asymptotically minimax second order estimates of the mean of a normal population. *Problems of information transmission*, **16**, 60–79. [134]

Bertero, M. 1989. Linear inverse and ill-posed problems. Pages 1–120 of: *Advances in Electronics and Electron Physics*, vol. 75. New York: Academic Press. [431]

Bhatia, R. 1997. *Matrix analysis*. Graduate Texts in Mathematics, vol. 169. Springer-Verlag, New York. [327]

Bickel, P. J., and Collins, J. R. 1983. Minimizing Fisher information over mixtures of distributions. *Sankhyā Ser. A*, **45**(1), 1–19. [246]

Bickel, P. J. 1981. Minimax estimation of the mean of a normal distribution when the parametr space is restricted. *Annals of Statistics*, **9**, 1301–1309. [121, 134, 246]

Bickel, P. J. 1983. Minimax estimation of a normal mean subject to doing well at a point. Pages 511–528 of: Rizvi, M. H., Rustagi, J. S., and Siegmund, D. (eds), *Recent Advances in Statistics*. New York: Academic Press. [245, 246]

Billingsley, P. 1999. *Convergence of probability measures*. Second edn. Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons Inc. A Wiley-Interscience Publication. [435, 436]

Birgé, L. 1983. Approximation dans les éspaces metriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **65**, 181–237. [154, 155]

Birgé, L., and Massart, P. 2001. Gaussian Model Selection. *Journal of European Mathematical Society*, **3**, 203–268. [51, 247, 327]

Birkhoff, G., and Rota, G.-C. 1969. *Ordinary Differential Equations*. Blaisdell. [92]

Bogachev, V. I. 2007. *Measure theory. Vol. I, II*. Berlin: Springer-Verlag. [432]

Bogachev, V. I. 1998. *Gaussian Measures*. American Mathematical Society. [98, 99, 399]

Borell, C. 1975. The Brunn-Minkowski inequality in Gauss space. *Invent. Math.*, **30**(2), 207–216. [51]

Born, M., and Wolf, E. 1975. *Principles of Optics*. 5th edn. New York: Pergamon. [86]

Breiman, L. 1968. *Probability*. Reading, Mass.: Addison-Wesley Publishing Company. [94]

Breiman, L. 1995. Better subset selection using the non-negative garotte. *Technometrics*, **37**, 373–384. [199]

Bretagnolle, J., and Huber, C. 1979. Estimation des densités: risque minimax. *Z. Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **47**, 119–137. [281]

Brown, L., DasGupta, A., Haff, L. R., and Strawderman, W. E. 2006. The heat equation and Stein's identity: connections, applications. *J. Statist. Plann. Inference*, **136**(7), 2254–2278. [134]

Brown, L. D., and Low, M. G. 1996a. Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics*, **3**, 2384–2398. [93, 96, 381]

Brown, L. D., and Purves, R. 1973. Measurable selections of extrema. *Ann. Statist.*, **1**, 902–912. [50, 134]

Brown, L. D. 1971. Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Annals of Mathematical Statistics*, **42**, 855–903. Correction: *Ann. Stat.* **1** 1973, pp 594–596. [51, 134]

Brown, L. D. 1986. *Fundamentals of statistical exponential families with applications in statistical decision theory*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 9. Hayward, CA: Institute of Mathematical Statistics. [97]

Brown, L. D., and Gajek, L. 1990. Information Inequalities for the Bayes Risk. *Annals of Statistics*, **18**, 1578–1594. [134]

Brown, L. D., and Low, M. G. 1996b. A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.*, **24**(6), 2524–2535. [279]

Brown, L. D., Low, M. G., and Zhao, L. H. 1997. Superefficiency in nonparametric function estimation. *Annals of Statistics*, **25**, 2607–2625. [166, 175, 176, 181]

Brown, L. D., Carter, A. V., Low, M. G., and Zhang, C.-H. 2004. Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.*, **32**(5), 2074–2097. [97]

Brown, L. D. 1966. On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist*, **37**, 1087–1136. [51]

Brown, L. 1977. Closure theorems for sequential-design processes. In: Gupta, S., and Moore, D. (eds), *Statistical Decision Theory and Related Topics II*. Academic Press, New York. [399]

Brown, L. 1978. *Notes on Statistical Decision Theory*. Unpublished Lecture Notes. [395, 399, 401, 402]

Bühlmann, P., and van de Geer, S. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer. [49]

Cai, T. T. 1999. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, **27**(3), 898–924. [54, 166, 209, 245, 276]

Cai, T. T. 2002. On block thresholding in wavelet regression: adaptivity, block size, and threshold level. *Statist. Sinica*, **12**(4), 1241–1273. [282]

Cai, T. T., and Zhou, H. H. 2009a. Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann. Statist.*, **37**(6A), 3204–3235. [98]

Cai, T., and Zhou, H. 2009b. A data-driven block thresholding approach to wavelet estimation. *Ann. Statist*, **37**, 569–595. [207, 209, 348]

Candès, E., and Romberg, J. 2007. Sparsity and incoherence in compressive sampling. *Inverse Problems*, **23**(3), 969–985. [49]

Carter, A. V. 2011. *Asymptotic Equivalence of Nonparametric Experiments Bibliography*. webpage at University of California, Santa Barbara, Department of Statistics. [98]

Carter, C., Eagleson, G., and Silverman, B. 1992. A comparison of the Reinsch and Speckman splines. *Biometrika*, **79**, 81–91. [78, 145, 146, 158]

Casella, G., and Strawderman, W. E. 1981. Estimating a bounded normal mean. *Annals of Statistics*, **9**, 870–878. [120]

Cavalier, L. 2004. Estimation in a problem of fractional integration. *Inverse Problems*, **20**(5), 1445–1454. [157]

Cavalier, L., and Tsybakov, A. B. 2001. Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist.*, **10**(3), 247–282. Meeting on Mathematical Statistics (Marseille, 2000). [166, 181, 246]

Cavalier, L. 2011. Inverse problems in statistics. Pages 3–96 of: *Inverse problems and high-dimensional estimation*. Lect. Notes Stat. Proc., vol. 203. Heidelberg: Springer. [99]

Cavalier, L., and Tsybakov, A. 2002. Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, **123**(3), 323–354. [181]

Chatterjee, S. 2009. Fluctuations of eigenvalues and second order Poincaré inequalities. *Probab. Theory Related Fields*, **143**(1-2), 1–40. [51]

Chaumont, L., and Yor, M. 2003. *Exercises in probability*. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 13. Cambridge: Cambridge University Press. A guided tour from measure theory to random processes, via conditioning. [51]

Chen, S. S., Donoho, D. L., and Saunders, M. A. 1998. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**(1), 33–61. [49]

Chernoff, H. 1981. A note on an inequality involving the normal distribution. *Ann. Probab.*, **9**(3), 533–535. [51]

Chui, C. K. 1992. *An Introduction to Wavelets*. San Diego: Academic Press. [407]

Chui, C. K. 1997. *Wavelets: a mathematical tool for signal processing*. SIAM Monographs on Mathematical Modeling and Computation. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). With a foreword by Gilbert Strang. [211]

Chung, K. L. 1974. *A course in probability theory*. Second edn. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London. Probability and Mathematical Statistics, Vol. 21. [436]

Cirel'son, B., Ibragimov, I., and Sudakov, V. 1976. Norm of Gaussian sample function. Pages 20–41 of: *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory*. Lecture Notes in Mathematics, 550. [51]

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836. [172]

Cogburn, R., and Davis, H. T. 1974. Periodic splines and spectral estimation. *Ann. Statist.*, **2**, 1108–1126. [99]

Cohen, A. 1966. All admissible linear estimates of the mean vector. *Annals of Mathematical Statistics*, **37**, 456–463. [36, 51]

Cohen, A., Daubechies, I., Jawerth, B., and Vial, P. 1993a. Multiresolution analysis, wavelets, and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris (A)*, **316**, 417–421. [191]

Cohen, A., Dahmen, W., and Devore, R. 2000. Multiscale Decompositions on Bounded Domains. *Transactions of American Mathematical Society*, **352**(8), 3651–3685. [414]

Cohen, A. 1990. Ondelettes, analyses multirésolutions et filtres miroir en quadrature. *Annales Institut Henri Poincaré, Analyse Non Linéaire*, **7**, 439–459. [407]

Cohen, A. 2003. *Numerical analysis of wavelet methods*. Studies in Mathematics and its Applications, vol. 32. Amsterdam: North-Holland Publishing Co. [211]

Cohen, A., and Ryan, R. 1995. *Wavelets and Multiscale Signal Processing*. Chapman and Hall. [407]

Cohen, A., Daubechies, I., and Vial, P. 1993b. Wavelets on the Interval and Fast Wavelet Transforms. *Applied Computational and Harmonic Analysis*, **1**, 54–81. [191, 197, 411]

Coifman, R., and Donoho, D. 1995. Translation-Invariant De-Noising. In: Antoniadis, A. (ed), *Wavelets and Statistics*. Springer Verlag Lecture Notes. [203, 204]

Courant, R., and Hilbert, D. 1953. *Methods of Mathematical Physics, Volume 1*. Wiley-Interscience. [91, 92]

Cover, T. M., and Thomas, J. A. 1991. *Elements of Information Theory*. Wiley. [132]

Cox, D. D. 1983. Asymptotics for M-type smoothing splines. *Ann. Statist.*, **11**(2), 530–551. [78, 99]

Cox, D. D. 1988. Approximation of method of regularization estimators. *Ann. Statist.*, **16**(2), 694–712. [78]

Cramér, H. 1999. *Mathematical methods of statistics*. Princeton Landmarks in Mathematics. Princeton, NJ: Princeton University Press. Reprint of the 1946 original. [436]

DasGupta, A. 2011a. *Probability for statistics and machine learning*. Springer Texts in Statistics. New York: Springer. Fundamentals and advanced topics. [50]

DasGupta, A. 2011b. *Sharp Nonasymptotic Bounds and Three Term Asymptotic Expansions for the Mean and Median of a Gaussian Sample Maximum*. manuscrip. [246]

Daubechies, I. 1988. Orthonormal Bases of Compactly Supported Wavelets. *Comm. Pure and Applied Math.*, **41**, 909–996. [188, 408]

Daubechies, I. 1992. *Ten Lectures on Wavelets*. CBMS-NSF Series in Applied Mathematics, no. 61. Philadelphia: SIAM. [181, 185, 189, 211, 408, 410, 420]

de Haan, L., and Ferreira, A. 2006. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. New York: Springer. An introduction. [246]

Delaigle, A., and Hall, P. 2011. Achieving near-perfect classification for functional data. *Journal of the Royal Statistical Society, Ser. B.* to appear. [103]

Dembo, A., and Zeitouni, O. 2010. *Large deviations techniques and applications*. Stochastic Modelling and Applied Probability, vol. 38. Berlin: Springer-Verlag. Corrected reprint of the second (1998) edition. [304]

Demmler, A., and Reinsch, C. 1975. Oscillation Matrices with Spline Smoothing. *Numerische Mathematik*, **24**, 375–382. [17]

DeVore, R., and Lorentz, G. 1993. *Constructive Approximation*. Springer Verlag. [281, 414, 416, 419]

DeVore, R. A. 1998. Nonlinear approximation. Pages 51–150 of: *Acta numerica, 1998*. Acta Numer., vol. 7. Cambridge: Cambridge Univ. Press. [281, 414]

Diaconis, P., and Stein, C. 1983. *Lectures on Statistical Decision Theory*. Unpublished Lecture Notes. [399]

Diaconis, P., and Ylvisaker, D. 1979. Conjugate Priors for Exponential Families. *Annals of Statistics*, **7**, 269–281. [26]

Diaconis, P., and Zabell, S. 1991. Closed form summation for classical distributions: variations on a theme of de Moivre. *Statist. Sci.*, **6**(3), 284–302. [51]

Domínguez, V., Heuer, N., and Sayas, F.-J. 2011. Hilbert scales and Sobolev spaces defined by associated Legendre functions. *J. Comput. Appl. Math.*, **235**(12), 3481–3501. [143]

Donoho, D. L., and Johnstone, I. M. 1994a. Ideal Spatial Adaptation via Wavelet Shrinkage. *Biometrika*, **81**, 425–455. [51, 200, 219, 245]

Donoho, D. L., and Johnstone, I. M. 1994b. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory and Related Fields*, **99**, 277–303. [246, 359, 369]

Donoho, D. L., and Johnstone, I. M. 1995. Adapting to unknown smoothness via Wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224. [51, 181, 207, 211, 348]

Donoho, D. L., and Johnstone, I. M. 1996. Neo-classical Minimax Problems, Thresholding, and Adaptive Function Estimation. *Bernoulli*, **2**, 39–62. [245, 282]

Donoho, D. L., and Johnstone, I. M. 1998. Minimax Estimation via Wavelet shrinkage. *Annals of Statistics*, **26**, 879–921. [372, 378, 379, 380, 383]

Donoho, D. L., and Johnstone, I. M. 1999. Asymptotic Minimaxity of Wavelet Estimators with Sampled Data. *Statistica Sinica*, **9**, 1–32. [381, 389, 390]

Donoho, D. L., and Liu, R. C. 1991. Geometrizing Rates of Convergence, III. *Annals of Statistics*, **19**, 668–701. [279]

Donoho, D. L., Liu, R. C., and MacGibbon, K. B. 1990. Minimax risk over hyperrectangles, and implications. *Annals of Statistics*, **18**, 1416–1437. [119, 120, 128, 134]

Donoho, D. L., Johnstone, I. M., Hoch, C., and Stern, A. 1992. Maximum Entropy and the nearly black object. *J. Royal Statistical Society, Ser. B.*, **54**, 41–81. With Discussion. [50, 227, 246]

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. 1995. Wavelet Shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, **57**, 301–369. With Discussion. [181, 305]

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. 1997. Universal Near Minimaxity of Wavelet Shrinkage. Pages 183–218 of: D., P., E., T., and G.L., Y. (eds), *Festschrift for L. Le Cam*. Springer Verlag. [246, 296, 305]

Donoho, D., Johnstone, I., and Montanari, A. 2012. Accurate Prediction of Phase Transitions in Compressed Sensing via a Connection to Minimax Denoising. *IEEE Transactions on Information Theory*. in press. [246]

Donoho, D. L., and Huo, X. 2001. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, **47**(7), 2845–2862. [49]

Donoho, D. 1992a. De-Noising via Soft-Thresholding. *IEEE transactions on Information Theory*, **41**, 613–627. [305]

Donoho, D. 1992b. *Interpolating Wavelet Transforms*. Tech. rept. 408. Department of Statistics, Stanford University. [294, 384, 414]

Donoho, D. 1993. Unconditional bases are optimal bases for data compression and statistical estimation. *Applied and Computational Harmonic Analysis*, **1**, 100–115. [281, 293]

Donoho, D. 1994. Statistical Estimation and Optimal recovery. *Annals of Statistics*, **22**, 238–270. [279, 305]

Donoho, D. 1995. Nonlinear solution of linear inverse problems by Wavelet-Vaguelette Decomposition. *Applied Computational and Harmonic Analysis*, **2**, 101–126. [337, 338, 431]

Donoho, D. 1996. Unconditional Bases and Bit-Level Compression. *Applied Computational and Harmonic Analysis*, **3**, 388–392. [293]

Donoho, D., and Low, M. 1992. Renormalization exponents and optimal pointwise rates of convergence. *Annals of Statistics*, **20**, 944–970. [279, 284]

Dugundji, J. 1966. *Topology*. Allyn and Bacon, Boston. [395]

Durrett, R. 2010. *Probability: theory and examples*. Fourth edn. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. [79, 435]

Dym, H., and McKean, H. P. 1972. *Fourier Series and Integrals*. Academic Press. [135, 433]

Efromovich, S. 1996. On nonparametric regression for IID observations in a general setting. *Ann. Statist.*, **24**(3), 1125–1144. [157]

Efromovich, S. 1999. *Nonparametric curve estimation*. Springer Series in Statistics. New York: Springer-Verlag. Methods, theory, and applications. [17]

Efromovich, S. 2004a. Analysis of blockwise shrinkage wavelet estimates via lower bounds for no-signal setting. *Ann. Inst. Statist. Math.*, **56**(2), 205–223. [282]

Efromovich, S. 2004b. Oracle inequalities for Efromovich-Pinsker blockwise estimates. *Methodol. Comput. Appl. Probab.*, **6**(3), 303–322. [181]

Efromovich, S. 2005. A study of blockwise wavelet estimates via lower bounds for a spike function. *Scand. J. Statist.*, **32**(1), 133–158. [282]

Efromovich, S. 2010. Dimension reduction and adaptation in conditional density estimation. *J. Amer. Statist. Assoc.*, **105**(490), 761–774. [181]

Efromovich, S., and Pinsker, M. 1996. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica*, **6**(4), 925–942. [181]

Efromovich, S., and Samarov, A. 1996. Asymptotic equivalence of nonparametric regression and white noise model has its limits. *Statist. Probab. Lett.*, **28**(2), 143–145. [98]

Efromovich, S., and Valdez-Jasso, Z. A. 2010. Aggregated wavelet estimation and its application to ultra-fast fMRI. *J. Nonparametr. Stat.*, **22**(7), 841–857. [209]

Efromovich, S., and Pinsker, M. 1984. A learning algorithm for nonparametric filtering. *Automat. i Telemeh.*, **11**, 58–65. (in Russian), translated in *Automation and Remote Control*, 1985, p 1434-1440. [161, 166, 181]

Efron, B. 1993. Introduction to "James and Stein (1961) Estimation with Quadratic Loss". Pages 437–442 of: Kotz, S., and Johnson, N. (eds), *Breakthroughs in Statistics: Volume 1: Foundations and Basic Theory*. Springer. [51]

Efron, B. 2001. Selection criteria for scatterplot smoothers. *Ann. Statist.*, **29**(2), 470–504. [99]

Efron, B. 2011. *Tweedie's formula and selection bias*. Tech. rept. Department of Statistics, Stanford University. [50]

Efron, B., and Morris, C. 1971. Limiting the Risk of Bayes and Empirical Bayes Estimators – Part I: The Bayes Case. *J. American Statistical Association*, **66**, 807–815. [50, 216, 245]

Efron, B., and Morris, C. 1972. Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *J. Amer. Statist. Assoc.*, **67**, 130–139. [50]

Efron, B., and Morris, C. 1973. Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.*, **68**, 117–130. [39]

Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. 1954. *Tables of Integral Transforms, Volume 1*. McGraw-Hill. [74]

Eubank, R. L. 1999. *Nonparametric regression and spline smoothing*. Second edn. Statistics: Textbooks and Monographs, vol. 157. New York: Marcel Dekker Inc. [99]

Fan, J., and Gijbels, I. 1996. *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability, vol. 66. London: Chapman & Hall. [79]

Fan, J., and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**(456), 1348–1360. [199]

Fan, K. 1953. Minimax theorems. *Prob. Nat. Acad. Sci. U.S.A.*, **39**, 42–47. [396]

Feldman, I. 1991. Constrained minimax estimation of the mean of the normal distribution with known variance. *Ann. Statist.*, **19**(4), 2259–2265. [369]

Feller, W. 1971. *An introduction to probability theory and its applications, Volume 2*. New York: Wiley. [99]

Ferguson, T. S. 1967. *Mathematical statistics: A decision theoretic approach*. Probability and Mathematical Statistics, Vol. 1. Academic Press, New York-London. [432]

Folland, G. B. 1999. *Real analysis*. Second edn. Pure and Applied Mathematics (New York). New York: John Wiley & Sons Inc. Modern techniques and their applications, A Wiley-Interscience Publication. [99, 399, 433, 436, 440, 441]

Foster, D. P., and George, E. I. 1994. The risk inflation criterion for multiple regression. *Ann. Statist.*, **22**(4), 1947–1975. [309]

Foster, D., and Stine, R. 1997. *An information theoretic comparison of model selection criteria*. Tech. rept. Dept. of Statistics, University of Pennsylvania. [327]

Frazier, M., Jawerth, B., and Weiss, G. 1991. *Littlewood-Paley Theory and the study of function spaces*. NSF-CBMS Regional Conf. Ser in Mathematics, **79**. Providence, RI: American Mathematical Society. [260, 268, 412, 413]

Freedman, D. 1999. On the Bernstein-von Mises Theorem with Infinite Dimensional Parameters. *Annals of Statistics*, **27**, 1119–1140. [92]

Galambos, J. 1978. *The asymptotic theory of extreme order statistics*. John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics. [222, 223]

Gao, F., Hannig, J., and Torcaso, F. 2003. Integrated Brownian motions and exact L_2-small balls. *Ann. Probab.*, **31**(3), 1320–1337. [90]

Gao, H.-Y. 1998. Wavelet Shrinkage DeNoising Using The Non-Negative Garrote. *J. Computational and Graphical Statistics*, **7**, 469–488. [199, 209]

Gao, H.-Y., and Bruce, A. G. 1997. Waveshrink with firm shrinkage. *Statistica Sinica*, **7**, 855–874. [199]

Gasser, T., and Müller, H.-G. 1984. Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, **11**(3), 171–185. [181]

Gel'fand, I. M., and Shilov, G. E. 1964. *Generalized functions. Vol. I: Properties and operations*. Translated by Eugene Saletan. New York: Academic Press. [85, 337, 442]

George, E. I., and McCulloch, R. E. 1997. Approaches for Bayesian Variable Selection. *Statistica Sinica*, **7**, 339–374. [51]

George, E. I., and Foster, D. P. 2000. Calibration and Empirical Bayes Variable Selection. *Biometrika*, **87**, 731–747. [327]

Gilbarg, D., and Trudinger, N. S. 1983. *Elliptic Partial Differential Equations of Second Order*. Second edition edn. Springer-Verlag. [438]

Golomb, M., and Weinberger, H. F. 1959. Optimal approximation and error bounds. Pages 117–190 of: *On Numerical Approximation*. University of Wisconsin Press. [305]

Golub, G. H., and Van Loan, C. F. 1996. *Matrix Computations*. 3rd edn. Johns Hopkins University Press. [44]

Golubev, G. K., and Levit, B. Y. 1996. Asymptotically efficient estimation for analytic distributions. *Math. Methods Statist.*, **5**(3), 357–368. [157]

Golubev, G. K., Nussbaum, M., and Zhou, H. H. 2010. Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Ann. Statist.*, **38**(1), 181–214. [98]

Golubev, G. 1987. Adaptive asymptotically minimax estimates of smooth signals. *Problemy Peredatsii Informatsii*, **23**, 57–67. [181]

Gorenflo, R., and Vessella, S. 1991. *Abel integral equations*. Lecture Notes in Mathematics, vol. 1461. Berlin: Springer-Verlag. Analysis and applications. [84, 99]

Gourdin, E., Jaumard, B., and MacGibbon, B. 1994. Global Optimization Decomposition Methods for Bounded Parameter Minimax Risk Evaluation. *SIAM Journal of Scientific Computing*, **15**, 16–35. [120, 121]

Grama, I., and Nussbaum, M. 1998. Asymptotic Equivalence for Nonparametric Generalized Linear Models. *Probability Theory and Related Fields*, **111**, 167–214. [97]

Gray, R. M. 2006. Toeplitz and Circulant Matrices: A review. *Foundations and Trends in Communications and Information Theory*, **2**, 155–239. [48]

Green, P., and Silverman, B. 1994. *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall. [13, 17, 70]

Grenander, U. 1981. *Abstract inference*. New York: John Wiley & Sons Inc. Wiley Series in Probability and Mathematical Statistics. [157]

Grenander, U., and Rosenblatt, M. 1957. *Statistical Analysis of Stationary Time Series, Second Edition published 1984*. Chelsea. [98]

Groeneboom, P. 1996. Lectures on inverse problems. Pages 67–164 of: *Lectures on probability theory and statistics (Saint-Flour, 1994)*. Lecture Notes in Math., vol. 1648. Berlin: Springer. [99]

Groeneboom, P., and Jongbloed, G. 1995. Isotonic estimation and rates of convergence in Wicksell's problem. *Ann. Statist.*, **23**(5), 1518–1542. [99]

Hall, P., and Patil, P. 1993. *Formulae for mean integrated squared error of nonlinear wavelet-based density estimators*. Tech. rept. CMA-SR15-93. Australian National University. To appear, Ann. Statist. [181]

Hall, P. G., Kerkyacharian, G., and Picard, D. 1999a. On Block thresholding rules for curve estimation using kernel and wavelet methods. *Annals of Statistics*, **26**, 922–942. [209]

Hall, P. 1979. On the rate of convergence of normal extremes. *J. Appl. Probab.*, **16**(2), 433–439. [222]

Hall, P., and Hosseini-Nasab, M. 2006. On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68**(1), 109–126. [89]

Hall, P., and Smith, R. L. 1988. The Kernel Method for Unfolding Sphere Size Distributions. *Journal of Computational Physics*, **74**, 409–421. [99]

Hall, P., Kerkyacharian, G., and Picard, D. 1999b. On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica*, **9**(1), 33–49. [209]

Härdle, W., Hall, P., and Marron, S. 1988. How far are automatically chosen regression smoothing parameters from their minimum? (with discussion). *J. American Statistical Association*, **83**, 86–101. [167]

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. 1998. *Wavelets, approximation, and statistical applications*. Lecture Notes in Statistics, vol. 129. New York: Springer-Verlag. [211, 408]

Hardy, G. H., and Littlewood, J. E. 1928. Some properties of fractional integrals. I. *Math. Z.*, **27**(1), 565–606. [99]

Hart, J. D. 1997. *Nonparametric smoothing and lack-of-fit tests*. Springer Series in Statistics. New York: Springer-Verlag. [99]

Hastie, T., Tibshirani, R., and Wainwright, M. 2012. *L_1 regression?* Chapman and Hall? forthcoming. [49]

Hastie, T. J., and Tibshirani, R. J. 1990. *Generalized Additive Models*. Chapman and Hall. [70]

Hedayat, A., and Wallis, W. D. 1978. Hadamard matrices and their applications. *Ann. Statist.*, **6**(6), 1184–1238. [51]

Heil, C., and Walnut, D. F. 2006. *Fundamental Papers in Wavelet Theory*. Princeton University Press. [211]

Hernández, E., and Weiss, G. 1996. *A First Course on Wavelets*. CRC Press. [211, 408]

Hida, T. 1980. *Brownian Motion*. Springer. [99]

Hoerl, A. E., and Kennard, R. W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55–67. [51]

Huber, P. J., and Ronchetti, E. M. 2009. *Robust Statistics*. Wiley. [240, 437]

Hwang, J. T., and Casella, G. 1982. Minimax confidence sets for the mean of a multivariate normal distribution. *Ann. Statist.*, **10**(3), 868–881. [51]

Ibragimov, I., and Khasminskii, R. 1997. Some estimation problems in infinite-dimensional Gaussian white noise. Pages 259–274 of: *Festschrift for Lucien Le Cam*. New York: Springer. [157]

Ibragimov, I. A., and Has′minskiĭ, R. Z. 1977. Estimation of infinite-dimensional parameter in Gaussian white noise. *Dokl. Akad. Nauk SSSR*, **236**(5), 1053–1055. [157]

Ibragimov, I. A., and Khasminskii, R. Z. 1980. Asymptotic properties of some nonparametric estimates in Gaussian white nose. In: *Proceedings of Third International Summer School in Probability and Mathematical Statistics, (Varna 1978), Sofia*. in Russian. [16]

Ibragimov, I. A., and Khasminskii, R. Z. 1982. Bounds for the risks of non-parametric regression estimates. *Theory of Probability and its Applications*, **27**, 84–99. [279]

Ibragimov, I. A., and Khasminskii, 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. [119, 157]

Ibragimov, I., and Khasminskii, R. 1981. *Statistical estimation : asymptotic theory*. New York: Springer. Khasminskii transliterated as Has'minskii. [17, 155]

Ibragimov, I., and Khasminskii, R. 1983. Estimation of distribution density. *Journal of Soviet Mathematics*, **21**, 40–57. [157]

Ingster, Y. I., and Suslina, I. A. 2003. *Nonparametric goodness-of-fit testing under Gaussian models*. Lecture Notes in Statistics, vol. 169. New York: Springer-Verlag. [17]

Jaffard, S., Meyer, Y., and Ryan, R. D. 2001. *Wavelets*. Revised edn. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). Tools for science & technology. [211]

James, W., and Stein, C. 1961. Estimation with quadratic loss. Pages 361–380 of: *Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability Theory*. University of California Press. [21, 39, 51, 134]

Jansen, M. 2001. *Noise reduction by wavelet thresholding*. Lecture Notes in Statistics, vol. 161. New York: Springer-Verlag. [211]

Johnen, H. 1972. Inequalities connected with the moduli of smoothness. *Mat. Vesnik*, **9(24)**, 289–303. [418]

Johnson, N. L., and Kotz, S. 1970. *Distributions in Statistics: Continuous Univariate Distributions - 2*. Wiley, New York. [39]

Johnsonbaugh, R., and Pfaffenberger, W. E. 1981. *Foundations of mathematical analysis*. Monographs and Textbooks in Pure and Applied Math., vol. 62. Marcel Dekker, Inc., New York. [429]

Johnstone, I. M. 1994. Minimax Bayes, Asymptotic Minimax and Sparse Wavelet Priors. Pages 303–326 of: Gupta, S., and Berger, J. (eds), *Statistical Decision Theory and Related Topics, V*. Springer-Verlag. [181]

Johnstone, I. M. 1999. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statistica Sinica*, **9**, 51–83. [207]

Johnstone, I. M., and Silverman, B. W. 1990. Speed of Estimation in Positron Emission Tomography and related inverse problems. *Annals of Statistics*, **18**, 251–280. [99]

Johnstone, I. M., and Silverman, B. W. 1997. Wavelet Threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B.*, **59**, 319–351. [201, 202, 224]

Johnstone, I. M., and Silverman, B. W. 2004a. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, **32**, 1594–1649. [51, 207]

Johnstone, I. M. 2001. Chi Square Oracle Inequalities. Pages 399–418 of: de Gunst, M., Klaassen, C., and van der Waart, A. (eds), *Festschrift for Willem R. van Zwet*. IMS Lecture Notes - Monographs, vol. 36. Institute of Mathematical Statistics. [51, 246]

Johnstone, I. M. 2010. High dimensional Bernstein-von Mises: simple examples. *IMS Collections*, **6**, 87–98. [99]

Johnstone, I. M., and Silverman, B. W. 2004b. Boundary coiflets for wavelet shrinkage in function estimation. *J. Appl. Probab.*, **41A**, 81–98. Stochastic methods and their applications. [381, 389, 390]

Johnstone, I. M., and Silverman, B. W. 2005a. EbayesThresh: R Programs for Empirical Bayes Thresholding. *Journal of Statistical Software*, **12**(8), 1–38. [51]

Johnstone, I. M., and Silverman, B. W. 2005b. Empirical Bayes selection of wavelet thresholds. *Ann. Statist.*, **33**(4), 1700–1752. [172, 207, 268, 305]

Joshi, V. M. 1967. Inadmissibility of the usual confidence sets for the mean of a multivariate normal population. *Ann. Math. Statist.*, **38**, 1868–1875. [51]

Kagan, A. M., Linnik, Y. V., and Rao, C. R. 1973. *Characterization problems in mathematical statistics*. John Wiley & Sons, New York-London-Sydney. Translated from the Russian by B. Ramachandran, Wiley Series in Probability and Mathematical Statistics. [50]

Kahane, J., de Leeuw, K., and Katznelson, Y. 1977. Sur les coefficients de Fourier des fonctions continues. *Comptes Rendus Acad. Sciences Paris (A)*, **285**, 1001–1003. [294]

Kalifa, J., and Mallat, S. 2003. Thresholding estimators for linear inverse problems and deconvolutions. *Ann. Statist.*, **31**(1), 58–109. [369]

Kalifa, J., Mallat, S., and Rougé, B. 2003. Deconvolution by thresholding in mirror wavelet bases. *IEEE Trans. Image Process.*, **12**(4), 446–457. [369]

Katznelson, Y. 1968. *An Introduction to Harmonic Analysis*. Dover. [83, 123]

Keller, J. B. 1976. Inverse problems. *Amer. Math. Monthly*, **83**(2), 107–118. [84]

Kempthorne, P. J. 1987. Numerical specification of discrete least favorable prior distributions. *SIAM J. Sci. Statist. Comput.*, **8**(2), 171–184. [120]

Kneser, H. 1952. Sur un théorème fondamental de la théorie des jeux. *C. R. Acad. Sci. Paris*, **234**, 2418–2420. [395, 396]

Kolaczyk, E. D. 1997. Nonparametric Estimation of Gamma-Ray Burst Intensities Using Haar Wavelets. *The Astrophysical Journal*, **483**, 340–349. [97]

Komlós, J., Major, P., and Tusnády, G. 1975. An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **32**, 111–131. [97]

Koo, J.-Y. 1993. Optimal rates of convergence for nonparametric statistical inverse problems. *Ann. Statist.*, **21**(2), 590–599. [99]

Kotelnikov, V. 1959. *The Theory of Optimum Noise Immunity*. McGraw Hill, New York. [16]

Krantz, S. G., and Parks, H. R. 2002. *A primer of real analytic functions*. Second edn. Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks]. Boston, MA: Birkhäuser Boston Inc. [432]

Kuhn, H. 1953. Review of Kneser (1952). *Mathematical Reviews*, **14**, 301. [396]

Kuo, H.-H. 1975. *Gaussian Measures in Banach Spaces*. Springer Verlag, Lecture Notes in Mathematics # 463. [98]

Kuo, H.-H. 2006. *Introduction to stochastic integration*. Universitext. New York: Springer. [434]

Lai, T. L., and Robbins, H. 1976. Maximally dependent random variables. *Proceedings of the National Academy of Sciences*, **73**(2), 286–288. [246]

Laurent, B., and Massart, P. 1998. *Adaptive estimation of a quadratic functional by model selection*. Tech. rept. Université de Paris-Sud, Mathématiques. [51]

Le Cam, L. 1986. *Asymptotic Methods in Statistical Decision Theory*. Berlin: Springer. [10, 93, 395, 399]

Le Cam, L. 1955. An extension of Wald's theory of statistical decision functions. *Annals of Mathematical Statistics*, **26**, 69–81. [399]

Le Cam, L., and Yang, G. L. 2000. *Asymptotics in statistics*. Second edn. Springer Series in Statistics. New York: Springer-Verlag. Some basic concepts. [93]

Ledoux, M. 1996. Isoperimetry and Gaussian Analysis. In: Bernard, P. (ed), *Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilities de Saint Flour, 1994*. Springer Verlag. [51]

Ledoux, M. 2001. *The concentration of measure phenomenon*. Mathematical Surveys and Monographs, vol. 89. Providence, RI: American Mathematical Society. [45, 46, 51]

Lehmann, E. L., and Casella, G. 1998. *Theory of Point Estimation*. Second edn. Springer Texts in Statistics. New York: Springer-Verlag. [39, 40, 50, 51, 97, 107, 174]

Lehmann, E. L., and Romano, J. P. 2005. *Testing statistical hypotheses*. Third edn. Springer Texts in Statistics. New York: Springer. [28, 50, 80, 81, 107]

Lemarié, P., and Meyer, Y. 1986. Ondelettes et bases Hilbertiennes. *Revista Matematica Iberoamericana*, **2**, 1–18. [413]

Lepskii, O. 1991. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications*, **35**, 454–466. [279]

Levit, B. 2010a. Minimax revisited. I. *Math. Methods Statist.*, **19**(3), 283–297. [134]

Levit, B. 2010b. Minimax revisited. II. *Math. Methods Statist.*, **19**(4), 299–326. [134]

Levit, B. Y. 1980. On asymptotic minimax estimates of second order. *Theory of Probability and its Applications*, **25**, 552–568. [134]

Levit, B. Y. 1982. Minimax estimation and positive solutions of elliptic equations. *Theory of Probability and its Applications*, **82**, 563–586. [134]

Levit, B. Y. 1985. Second order asymptotic optimality and positive solutions of Schrödinger's equation. *Theory of Probability and its Applications*, **30**, 333–363. [134]

Loader, C. R. 1999. Bandwidth selection: Classical or plug-in? *Annals of Statistics*, **27**, 415–438. [181]

Mallat, S. 1998. *A Wavelet Tour of Signal Processing*. Academic Press. [181]

Mallat, S. 1999. *A Wavelet Tour of Signal Processing*. Academic Press. 2nd, expanded, edition. [187, 190, 408]

Mallat, S. 2009. *A wavelet tour of signal processing*. Third edn. Elsevier/Academic Press, Amsterdam. The sparse way, With contributions from Gabriel Peyré. [185, 198, 211, 405, 408]

Mallows, C. 1973. Some comments on C_p. *Technometrics*, **15**, 661–675. [51]

Mallows, C. 1978. Minimizing an Integral. *SIAM Review*, **20**(1), 183–183. [246]

Mandelbaum, A. 1984. All admissible linear estimators of the mean of a Gaussian distribution on a Hilbert space. *Annals of Statistics*, **12**, 1448–1466. [62]

Mardia, K. V., Kent, J. T., and Bibby, J. M. 1979. *Multivariate Analysis*. Academic Press. [26]

Marr, R. B. 1974. On the reconstruction of a function on a circular domain from a sampling of its line integrals. *J. Math. Anal. Appl.*, **45**, 357–374. [86]

Marron, J. S., and Wand, M. P. 1992. Exact mean integrated squared error. *Ann. Statist.*, **20**(2), 712–736. [181]

Massart, P. 2007. *Concentration inequalities and model selection*. Lecture Notes in Mathematics, vol. 1896. Berlin: Springer. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. [17, 55, 327]

McMurry, T. L., and Politis, D. N. 2004. Nonparametric regression with infinite order flat-top kernels. *J. Nonparametr. Stat.*, **16**(3-4), 549–562. [100]

Meyer, Y. 1986. Principe d'incertitude, bases hilbertiennes et algebres d'operatéurs. *Seminaire Bourbaki*, **662**. [408]

Meyer, Y. 1990. *Ondelettes et Opérateurs, I: Ondelettes, II: Opérateurs de Calderón-Zygmund, III:* (with R. Coifman)*, Opérateurs multilinéaires*. Paris: Hermann. English translations of Vol I. and Vols II-III (combined) published by Cambridge University Press. [185, 190, 211, 260, 281, 293, 406, 408, 413, 422]

Meyer, Y. 1991. Ondelettes sur l'intervalle. *Revista Matematica Iberoamericana*, **7**, 115–133. [191]

Meyer, Y. 1992. *Wavelets and Operators*. Vol. 1. Cambridge University Press. [413]

Meyer, Y., and Coifman, R. 1997. *Wavelets*. Cambridge Studies in Advanced Mathematics, vol. 48. Cambridge: Cambridge University Press. Calderón-Zygmund and multilinear operators, Translated from the 1990 and 1991 French originals by David Salinger. [427]

Mézard, M., and Montanari, A. 2009. *Information, physics, and computation*. Oxford Graduate Texts. Oxford: Oxford University Press. [245]

Micchelli, C. A. 1975. *Optimal estimation of linear functionals*. Tech. rept. 5729. IBM. [305]

Micchelli, C. A., and Rivlin, T. J. 1977. A survey of optimal recovery. Pages 1–54 of: Micchelli, C. A., and Rivlin, T. J. (eds), *Optimal Estimation in Approximation Theory*. New York: Plenum Press. [305]

Miller, A. J. 1984. Selection of subsets of regression variables (with discussion). *J. Roy. Statist. Soc., Series A*, **147**, 389–425. with discussion. [224]

Miller, A. J. 1990. *Subset Selection in Regression*. Chapman and Hall, London, New York. [224]

Mills, J. P. 1926. Table of the Ratio: Area to Bounding Ordinate, for Any Portion of Normal Curve. *Biometrika*, **18**, 395–400. [434]

Mirsky, L. 1975. A trace inequality of John von Neumann. *Monatsh. Math.*, **79**(4), 303–306. [327]
Nason, G. P. 2008. *Wavelet methods in statistics with R*. Use R! New York: Springer. [203, 211]
Nason, G. 2010. *wavethresh: Wavelets statistics and transforms.* R package version 4.5. [172]
Nemirovski, A. 2000. Topics in non-parametric statistics. Pages 85–277 of: *Lectures on probability theory and statistics (Saint-Flour, 1998)*. Lecture Notes in Math., vol. 1738. Berlin: Springer. [17, 181]
Nikol'skii, S. 1975. *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, New York. [412]
Nishii, R. 1984. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**(2), 758–765. [309]
Nussbaum, M. 1996. Asymptotic Equivalence of density estimation and white noise. *Annals of Statistics*, **24**, 2399–2430. [96, 97]
Nussbaum, M. N. 2004. Equivalence asymptotique des Expériences Statistiques. *Journal de la Société Francaise de Statistique*, **145**(1), 31–45. (In French). [93]
Ogden, R. T. 1997. *Essential wavelets for statistical applications and data analysis*. Boston, MA: Birkhäuser Boston Inc. [211]
Parthasarathy, K. 1967. *Probability Measures on Metric Spaces*. Academic Press. [435]
Peck, J., and Dulmage, A. 1957. Games on a compact set. *Canadian Journal of Mathematics*, **9**, 450–458. [396]
Peetre, J. 1975. *New Thoughts on Besov Spaces, I*. Raleigh, Durham: Duke University Mathematics Series. [267, 412, 415]
Percival, D. B., and Walden, A. T. 2000. *Wavelet methods for time series analysis*. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 4. Cambridge: Cambridge University Press. [211]
Perlman, M. D. 1974. Jensen's inequality for a convex vector-valued function on an infinite-dimensional space. *J. Multivariate Anal.*, **4**, 52–65. [432]
Pinsker, M. 1980. Optimal filtration of square-integrable signals in Gaussian noise. *Problems of Information Transmission*, **16**, 120–133. originally in Russian in *Problemy Peredatsii Informatsii* **16** 52-68. [61, 139]
Pinsky, M. A. 2009. *Introduction to Fourier analysis and wavelets*. Graduate Studies in Mathematics, vol. 102. Providence, RI: American Mathematical Society. Reprint of the 2002 original. [211]
Pratt, J. W. 1960. On interchanging limits and integrals. *Annals of Mathematical Statistics*, **31**, 74–77. [432]
Prékopa, A. 1980. Logarithmic concave measures and related topics. Pages 63–82 of: *Stochastic programming (Proc. Internat. Conf., Univ. Oxford, Oxford, 1974)*. London: Academic Press. [370]
Ramsay, J. O., and Silverman, B. W. 2005. *Functional data analysis*. Second edn. Springer Series in Statistics. New York: Springer. [89]
Reed, M., and Simon, B. 1980. *Functional Analysis, Volume 1, revised and enlarged edition*. Academic Press. [429, 431]
Rice, J., and Rosenblatt, M. 1981. Integrated mean squared error of a smoothing spline. *J. Approx. Theory*, **33**(4), 353–369. [99]
Riesz, F., and Sz.-Nagy, B. 1955. *Functional Analysis*. Ungar, New York. [429]
Rigollet, P. 2006. Adaptive density estimation using the blockwise Stein method. *Bernoulli*, **12**(2), 351–370. [181]
Robbins, H. 1956. An empirical Bayes approach to statistics. Pages 157–163 of: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*. Berkeley and Los Angeles: University of California Press. [50]
Royden, H. L. 1988. *Real analysis*. Third edn. Macmillan Publishing Company, New York. [436]
Rudin, W. 1973. *Functional Analysis*. McGraw Hill. [395, 401, 438]
Rudin, W. 1976. *Principles of mathematical analysis*. Third edn. New York: McGraw-Hill Book Co. International Series in Pure and Applied Mathematics. [429]
Ruggeri, F. 2006. Gamma-Minimax Inference. In: *Encyclopedia of Statistical Sciences*. John Wiley & Sons. [134]
Schervish, M. J. 1995. *Theory of statistics*. Springer Series in Statistics. New York: Springer-Verlag. [94, 134]
Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464. [309]

Serfling, R. J. 1980. *Approximation theorems of mathematical statistics*. New York: John Wiley & Sons Inc. Wiley Series in Probability and Mathematical Statistics. [45]

Shao, P. Y.-S., and Strawderman, W. E. 1994. Improving on the James-Stein positive-part estimator. *Ann. Statist.*, **22**(3), 1517–1538. [40]

Shepp, L. A. 1966. Radon-Nikodym derivatives of Gaussian measures. *Annals of Mathematical Statistics*, **37**, 321–354. [90, 435]

Silverman, B. W. 1984. Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, **12**, 898–916. [73]

Simonoff, J. S. 1996. *Smoothing methods in statistics*. Springer Series in Statistics. New York: Springer-Verlag. [99]

Simons, S. 1995. Minimax theorems and their proofs. Pages 1–23 of: Du, D.-Z., and Pardalos, P. (eds), *Minimax and Applications*. Kluwer Academic Publishers. [396]

Sion, M. 1958. On general minimax theorems. *Pacific Journal of Mathematics*, **8**, 171–176. [396]

Speckman, P. 1985. Spline smoothing and optimal rates of convergence in nonparametric regression models. *Annals of Statistics*, **13**, 970–983. [78, 99]

Srinivasan, C. 1973. Admissible Generalized Bayes Estimators and Exterior Boundary Value Problems. *Sankhya*, **43**, 1–25. Ser. A. [51, 134]

Starck, J.-L., Murtagh, F., and Fadili, J. M. 2010. *Sparse image and signal processing*. Cambridge: Cambridge University Press. Wavelets, curvelets, morphological diversity. [211]

Stein, C. 1956. Efficient nonparametric estimation and testing. Pages 187–195 of: *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*. University of California Press, Berkeley, CA. [21, 51]

Stein, C. 1981. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135–1151. [37, 39, 51]

Stoffer, D. 1991. Walsh-Fourier Analysis and Its Statistical Applications. *Journal of the American Statistical Association*, **86**, 461–479. [97]

Stone, C. 1980. Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, **8**, 1348–1360. [296]

Strawderman, W. E. 1971. Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.*, **42**(1), 385–388. [51]

Sudakov, V. N., and Cirel′son, B. S. 1974. Extremal properties of half-spaces for spherically invariant measures. *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, **41**, 14–24, 165. Problems in the theory of probability distributions, II. [51]

Szegö, G. 1967. *Orthogonal Polynomials, 3rd edition*. American Mathematical Society. [104, 420]

Talagrand, M. 2003. *Spin glasses: a challenge for mathematicians*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], vol. 46. Berlin: Springer-Verlag. Cavity and mean field models. [245, 434]

Tao, T. 2011. *Topics in Random Matrix Theory*. draft book mansucript. [51]

Temme, N. M. 1996. *Special functions*. A Wiley-Interscience Publication. New York: John Wiley & Sons Inc. An introduction to the classical functions of mathematical physics. [442]

Thorisson, H. 1995. Coupling methods in probability theory. *Scand. J. Statist.*, **22**(2), 159–182. [432]

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58**(1), 267–288. [49]

Tibshirani, R., and Knight, K. 1999. The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, Series B*, **61**, 529–546. [327]

Tikhonov, A. N., and Arsenin, V. Y. 1977. *Solutions of ill-posed problems*. Washington, D.C.: John Wiley & Sons, New York: V. H. Winston & Sons. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics. [51]

Triebel, H. 1983. *Theory of Function Spaces*. Basel: Birkhäuser Verlag. [260, 268, 412, 414, 415, 417]

Triebel, H. 1992. *Theory of Function Spaces II*. Basel: Birkhäuser Verlag. [412, 414, 415]

Triebel, H. 2006. *Theory of function spaces. III*. Monographs in Mathematics, vol. 100. Basel: Birkhäuser Verlag. [414]

Triebel, H. 2008. *Function spaces and wavelets on domains*. EMS Tracts in Mathematics, vol. 7. European Mathematical Society (EMS), Zürich. [414]

Tsybakov, A. B. 2009. *Introduction to Nonparametric Estimation*. Springer. [17, 61, 131, 135, 157, 281, 305]

Tsybakov, A. 1997. Asymptotically Efficient Signal Estimation in L_2 Under General Loss Functions. *Problems of Information Transmission*, **33**, 78–88. translated from Russian. [157]

van der Vaart, A. W. 1997. Superefficiency. Pages 397–410 of: *Festschrift for Lucien Le Cam*. New York: Springer. [181]

van der Vaart, A. W. 1998. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 3. Cambridge: Cambridge University Press. [56]

van der Vaart, A. 2002. The statistical work of Lucien Le Cam. *Ann. Statist.*, **30**(3), 631–682. Dedicated to the memory of Lucien Le Cam. [93]

Van Trees, H. L. 1968. *Detection, Estimation and Modulation Theory, Part I*. New York: Wiley. [109]

Vidakovic, B. 1999. *Statistical Modelling by Wavelets*. John Wiley and Sons. [211]

Vidakovic, B., and DasGupta, A. 1996. Efficiency of linear rules for estimating a bounded normal mean. *Sankhyā Ser. A*, **58**(1), 81–100. [134]

Vogel, C. R. 2002. *Computational methods for inverse problems*. Frontiers in Applied Mathematics, vol. 23. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). With a foreword by H. T. Banks. [51]

von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. FILL IN. [395]

Wahba, G. 1978. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B.*, **40**, 364–372. [90]

Wahba, G. 1983. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B.*, **45**, 133–150. [90]

Wahba, G. 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, **13**, 1378–1402. [99]

Wahba, G. 1990. *Spline Methods for Observational Data*. Philadelphia: SIAM. [70, 90]

Wald, A. 1950. *Statistical Decision Functions*. Wiley. [10, 399]

Walter, G. G., and Shen, X. 2001. *Wavelets and other orthogonal systems*. Second edn. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton, FL. [211]

Wand, M. P., and Jones, M. C. 1995. *Kernel smoothing*. Monographs on Statistics and Applied Probability, vol. 60. London: Chapman and Hall Ltd. [99]

Wasserman, L. 2006. *All of nonparametric statistics*. Springer Texts in Statistics. New York: Springer. [17]

Watson, G. S. 1971. Estimating Functionals of Particle Size Distribution. *Biometrika*, **58**, 483–490. [84]

Wicksell, S. D. 1925. The corpuscle problem. A mathematical study of a biometric problem. *Biometrika*, **17**, 84–99. [84]

Williams, D. 1991. *Probability with Martingales*. Cambridge University Press, Cambridge. [79]

Wojtaszczyk, P. 1997. *A Mathematical Introduction to Wavelets*. Cambridge University Press. [211, 415]

Woodroofe, M. 1970. On choosing a delta sequence. *Annals of Mathematical Statistics*, **41**, 1665–1671. [176]

Young, W. H. 1911. On semi-integrals and oscillating successions of functions. *Proc. London Math. Soc. (2)*, **9**, 286–324. [432]

Zhang, C.-H. 2012a. Minimax ℓ_q risk in ℓ_p balls. Pages 78–89 of: *Contemporary developments in Bayesian analysis and statistical decision theory: a Festschrift for William E. Strawderman*. Inst. Math. Stat. (IMS) Collect., vol. 8. Inst. Math. Statist., Beachwood, OH. [369]

Zhang, C.-H. 2012b. Minimaxℓ_q risk in ℓ_p balls. Pages 78–89 of: *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*. IMS Collections, vol. 8. Institute of Mathematical Statistics. [246]

Ziemer, W. P. 1989. *Weakly differentiable functions*. Graduate Texts in Mathematics, vol. 120. New York: Springer-Verlag. Sobolev spaces and functions of bounded variation. [438]

Zygmund, A. 1959. *Trigonometric Series, Volume I*. Cambridge University Press, Cambridge. [123]

Zygmund, A. 2002. *Trigonometric series. Vol. I, II*. Third edn. Cambridge Mathematical Library. Cambridge: Cambridge University Press. With a foreword by Robert A. Fefferman. [143, 144]

Index