Optimal Estimation: Prior Free Methods and Physiological Application

by

Martin Raphan

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics

New York University

May 2007

Eero P. Simoncelli—Advisor

Daniel Tranchina—Advisor

To my parents

Acknowledgements

First and foremost, I would like to thank my advisors, Eero Simoncelli and Dan Tranchina. Dan supervised my work on cortical modeling, and his insight and advice were extremely helpful in carrying out the bulk of the work of Chapter 1. He also had many useful comments about the remainder of the material in the thesis. Over the years, I have learned a lot about computational neuroscience in general from discussions with him.

Eero supervised my work on prior-free methods and applications, which make up the substance of Chapters 2-4. His intuition, insight and ideas were crucial in helping me progress in this line of research, and more importantly, in obtaining useful results. I also learned a lot from him about image processing, statistics and computational neuroscience, amongst other things.

I would like to thank my third reader, Charlie Peskin, for his input to my thesis and defense and helpful discussions about the material. I would also like to thank Mehryar Mohri for being on my committee and for some useful discussions about VC type bounds for regression. As well, I would like to thank Francesca Chiaromonte for being on my committee, and for helpful discussions and comments about the material in the thesis. It was good to have a statistician's point of view on the work.

I would like to thank Bob Shapley for his helpful input, and for information about contrast dependent summation area. I would also like to thank him for letting me sit in on his ”new view” class about visual cortex, where I read some very useful papers. I would like to thank members of the Laboratory for Computational

Vision, for helpful comments and discussions along the way. I would also like to thank LCV alumni Liam Paninski and Jonathan Pillow, who both had some particularly useful comments about the prior-free methods. I would also like thank the various people at Courant, too numerous to mention, who have provided help along the way.

I would like to acknowledge New York University (NYU), Howard Hughes Medical Institute (HHMI) and the National Science Foundation (NSF) for their support during the course of this work.

Finally, I would like to thank my family and friends for their warm support. I would particularly like to thank my father for his encouragement throughout the course of this research.

Contents

List of Figures

List of Tables

Chapter 1

Contrast Dependent Receptive Field Size

In this chapter, we will begin by describing a nonlinear phenomenon observed in the primary visual cortex, that of increasing spatial integration area with decreasing contrast. We will give a possible physiological function of this behavior in terms of optimal filtering of stochastic processes. Namely, we will show that in certain cases of spatial stochastic processes, with a certain type of random corruption, the linear filter which should be applied to the corrupted measurements in order to optimally estimate the uncorrupted process has the same shape for all contrasts, with the width decreasing as contrast increases. We will then discuss the possibility of this variable receptive field size being achieved through positive feedback that attenuates at high contrast, and we will illustrate two mechanisms of cortical connectivity that could allow for this type of feedback. Finally, we point out that the the optimal linear filter can be written without reference to the statistics of the signal to be estimated, but rather entirely in terms of the statistics of

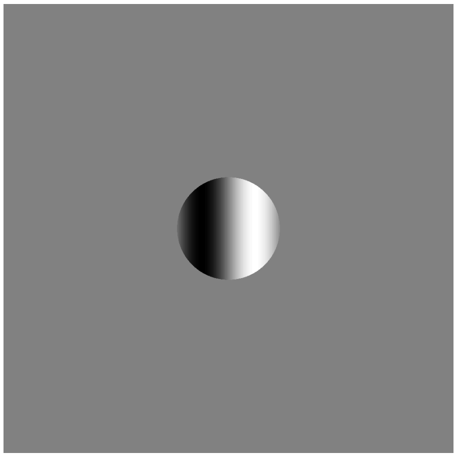


Fig. 1.1: Stimulus used to test receptive field size. The stimulus in the central region is a sinusoidal drifting grating with temporal frequency, $f = 5Hz$ and spatial frequency $k = 0.6$ cycle/deg.

the corrupted observation, a theme which we will elaborate on in further chapters.

1.1 Contrast Dependent Receptive Field Size

In the work of Sceniak, et al [1], experiments showed a dependency of spatial summation area on contrast. In one experiment, for example, a cell in the primary visual cortex (V1) was stimulated by a circular aperture with a drifting sinusoidal grating inside, as shown in Fig. 1.1. This sinusoid was chosen to have spatial and temporal frequencies, as well as orientation, which were known, from separate experiments, to optimally stimulate the cell. Then, for various values of contrast and aperture radius, the first harmonic amplitude (F1) of the cell's firing rate was measured and recorded. When the F1 response was plotted as a function of the aperture radius, as depicted in the cartoon of Fig. 1.2, it was noted that for a given cell, the aperture radius that gave the peak F1 response at low contrast

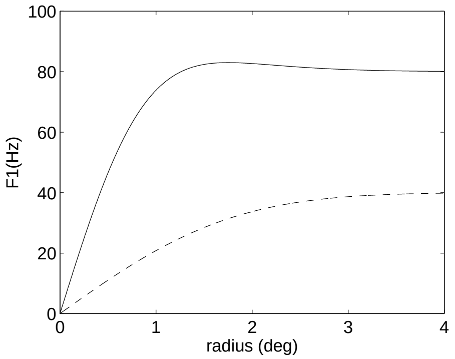


Fig. 1.2: Cartoon depicting F1 response as a function of aperture radius. Dashed curve is at low contrast, solid curve is at high contrast.

was larger than the aperture that gave the peak F1 response at high contrast. To give a quantitative measure of this phenomenon, the F1 response as a function of aperture radius was fit by a Difference of Gaussians (DoG) model:

$$R(s) = K_e \int_{-\frac{s}{2}}^{\frac{s}{2}} e^{-(2y/a)^2} dy - K_i \int_{-\frac{s}{2}}^{\frac{s}{2}} e^{-(2y/b)^2} dy \tag{1.1}$$

where R denotes the F1 response, s denotes the aperture radius, a the width of the central excitatory Gaussian, b the width of the wider inhibitory Gaussian, and $K_{e,i}$ are the weights of the excitatory and inhibitory Gaussians, respectively. It was found that as contrast increased, the area of spatial integration as measured by the width, a, of the center Gaussian, decreased. This is a nonlinear phenomenon, and so cannot be explained by a linear model of the visual pathway.

1.2 Why Contrast Dependent Receptive Field Size?

Before discussing mechanisms for achieving this nonlinear behavior, we first describe a possible physiological function of this behavior in terms of optimal filtering. Imagine that we have a two dimensional stationary process X on a discrete lattice, and that we observe N, a corrupted version of X. We will assume that N_{ij}, the elements of N, are independent given X, and that the statistics of each element N_{ij} are dependent only on X_{ij}, with all elements having the conditional density $P(N_{ij}|X_{ij})$. We can express this mathematically as:

$$P(N|X) = \prod_{i,j} P(N_{ij}|X_{ij}). \tag{1.2}$$

For example one might consider a positive process X, that is used to generate Poisson variables N_{ij}, which are independent and have rate X_{ij}, respectively. A continuous version of the Poisson case is discussed in [2].

Suppose that we wish to find the linear operation on N which gives the estimate of X that minimizes variance of the difference between X and its estimate. We begin by writing the two dimensional variables X, and N as vectors by concatenating the columns. Then an arbitrary linear operator may be written in terms of some matrix H. Our optimal estimator will then be

$$\hat{X}(N) = HN \tag{1.3}$$

where H is chosen to minimize

$$E\{|X-HN-(E\{X\}-HE\{N\})|^2\} = E\{|(X-E\{X\})-H(N-E\{N\})|^2\} \quad (1.4)$$

Using the orthogonality principle, this can be solved by insisting that

$$E\Big\{\Big(X-E\{X\}-H(N-E\{N\})\Big)\Big(N-E\{N\}\Big)^T\Big\} = 0 \quad (1.5)$$

or equivalently

$$E\{(X-E\{X\})(N-E\{N\})^T\} = HE\{(N-E\{N\})(N-E\{N\})^T\} \quad (1.6)$$

This may also be written as

$$C_{XN} = HC_{NN} \quad (1.7)$$

where C_{XN} is the matrix of covariances between elements of X and N, and C_{NN} is the covariance matrix for the vector N. Eq. (1.7) is a standard result on optimal linear estimation [3], but is rather general. We will now see what this solution looks like for the particular corruption model we are assuming.

If we make the assumption that

$$E\{N|X\} = X \quad (1.8)$$

which holds, for example, in the Poisson case, and which implies that

$$E\{N\} = E\{X\}, \quad (1.9)$$

then

$$\begin{aligned}
C_{XN} &= E\{(X - E\{X\})(N - E\{N\})^T\} \\
&= E\{(X - E\{X\})(N - E\{X\})^T\} \\
&= E\{E\{(X - E\{X\})(N - E\{X\})^T \Big| X\}\} \\
&= E\{(X - E\{X\})(E\{N|X\} - E\{X\})^T\} \\
&= E\{(X - E\{X\})(X - E\{X\})^T\} \\
&= C_{XX} \qquad (1.10)
\end{aligned}$$

where C_{XX} is the covariance matrix of the vector X. Next we have that,

$$\begin{aligned}
C_{NN} &= E\{(N - E\{N\})(N - E\{N\})^T\} \\
&= E\{(N - E\{X\})(N - E\{X\})^T\} \qquad (1.11)
\end{aligned}$$

We note that, since for $i \neq j$ N_i and N_j are independent given X

$$\begin{aligned}
E\{N_i N_j\} &= E\{E\{N_i N_j|X\}\} \\
&= E\{E\{N_i|X\}E\{N_j|X\}\} \\
&= E\{X_i X_j\} \qquad (1.12)
\end{aligned}$$

Also we have the trivial identity

$$E\{N_i^2\} = E\{X_i^2\} + (\sigma_n^2 - \sigma_x^2) \qquad (1.13)$$

where σ_x^2 and σ_n^2 denote the variances of the elements of X and N. This allows us to rewrite Eq. (1.11) as

$$\begin{aligned} C_{NN} &= E\{(X - E\{X\})(X - E\{X\})^T\} + (\sigma_n^2 - \sigma_x^2)I \\ &= C_{XX} + (\sigma_n^2 - \sigma_x^2)I \\ &= C_{XX} + \sigma^2 I \qquad (1.14) \end{aligned}$$

where σ^2 is the difference between the variance of N and X. In the case of additive noise corruption, σ^2 will be the variance of the additive noise, but this solution works for any type of corruption which fits our model, including Poisson.

Putting our identities for C_{XN} and C_{NN} into Eq. (1.7) gives

$$C_{XX} = H(C_{XX} + \sigma^2 I) \qquad (1.15)$$

or

$$H = C_{XX}(C_{XX} + \sigma^2 I)^{-1} \qquad (1.16)$$

In the case of additive Gaussian noise, this is the well known Weiner filter [3]. We now see that the optimal linear solution for any independent componentwise corruption process which satisfies $E\{N|X\} = X$ will have an optimal linear estimator of the same form.

Since C_{XX} is a covariance matrix, and since we assumed that X was stationary, it can be diagonalized by the matrix Q that represents the two dimensional Fourier transform, so that

$$Q^* H Q = S_X(S_X + \sigma^2 I)^{-1} \qquad (1.17)$$

where S_X is a diagonal matrix with the Power Spectral Density (PSD) of X, Φ

along its diagonal. Therefore, in the Fourier representation, H is diagonal, which means that the optimal linear operation is a shift invariant filter. We also see that the Fourier transform of the optimal filter for estimating X from N is given by:

$$\hat{h}_{opt}(\omega) = \frac{\Phi(\omega)}{\sigma^2 + \Phi(\omega)} \tag{1.18}$$

where σ^2 is a constant that depends on the second order statistics of the N_i's and the X_i's.

In our problem we will view the true firing rates of the retinal ganglion cells as the stationary process X and the number of spikes generated by the retinal ganglion cells as being a stochastic process with the property described in Eq. (1.2). We will further assume that the firing rates scale with contrast of the stimulus. This will be the case, for example, if the retinal firing rates are well modeled using linear operations and rectifications, since both these operations posses the property that if the input is scaled by a positive value, the output is scaled by the same value. Also, we will ignore the fact that a true optimal filter would also have a causal, temporal part, which filters in time. According to [2], ignoring temporal filtering does not qualitatively change the problem.

Since the firing rate X, scales with c, its PSD must scale with c^2. Including the dependence on c explicitly in Eq. (1.18) gives

$$\hat{h}_{opt} = \frac{c^2\Phi_0}{\sigma^2 + c^2\Phi_0} \tag{1.19}$$

where Φ_0 is the PSD of the stimulus at unit contrast.

If Φ_0 takes the the particular form

$$\Phi_0 = \frac{k_0}{1 + |a\omega|^\alpha} \tag{1.20}$$

for constants k_0, α and a, then the optimal filter will have Fourier transform

$$\begin{aligned}
\hat{h}_{opt}(\omega) &= \frac{c^2 k_0}{\sigma^2(1 + |a\omega|^\alpha) + c^2 k_0} \\
&= \frac{c^2}{(\sigma^2 + c^2 k_0)}(\frac{k_0}{(1 + |a\omega/((1 + \frac{c^2 k_0}{\sigma^2})^{1/\alpha})|^\alpha}) \\
&= \frac{c^2}{(\sigma^2 + c^2 k_0)}\Phi_0(\frac{\omega}{\beta}) \qquad (1.21)
\end{aligned}$$

where

$$\beta = (1 + \frac{c^2 k_0}{\sigma^2})^{\frac{1}{\alpha}} \tag{1.22}$$

Using the scaling properties of the Fourier transform and the fact that we are dealing with a two dimensional space, we get an optimal filter with spatial profile

$$\begin{aligned}
h_{opt}(x) &= \frac{c^2 \beta^2}{(\sigma^2 + c^2 k_0)}\phi_0(\beta x) \\
&= \frac{c^2}{\sigma^2}(1 + \frac{c^2 k_0}{\sigma^2})^{\frac{2}{\alpha} - 1}\phi_0(\beta x) \qquad (1.23)
\end{aligned}$$

This shows that the optimal filter for any contrast is just a scaled version of some fixed filter, which gets narrower as contrast increases. We note that a commonly assumed value of α for natural images is 2, which further simplifies the form of this equation and gives

$$h_{opt}(x) = \frac{c^2}{\sigma^2}\phi_0(\beta x) \tag{1.24}$$

Since β increases with increasing contrast, the effective width of the optimal filter, proportional to $\frac{1}{\beta}$, decreases with increasing contrast. The choice of optimal filter is a tradeoff between two opposing limitations. A wider filter would have low variances, since it is averaging the responses of many neurons, but such a filter would also have more bias, because these cells will have different firing rates. At higher firing rates, there is less of a need to average over many cells to decrease the variance. Therefore, the optimal filter will have smaller field size so that it may reduce bias. At lower firing rates, however, it is the variance problem which dominates, and so we use a larger receptive field, trading off increased bias for a reduction in variance.

1.3 Increased Receptive Field Size through Positive Feedback

Physiological studies show that spatial summation areas of V1 cells at low contrast are wider than can be accounted for by the feedforward connectivity that these cells receive from the LGN [4]. Therefore, we are looking for a mechanism that will expand these receptive field sizes at low contrast (as opposed to shrinking them at high contrast, for example). In this section we will discuss how this might be achieved through a positive feedback mechanism. We will begin the discussion of the proposed mechanism for contrast dependent receptive field size by discussing the intuition which leads to the conclusion that positive feedback does indeed increase receptive field size. We then consider a heuristic example which makes this more quantitatively concrete.

Consider the cartoon shown in Fig. 1.3. In this figure we have a number of

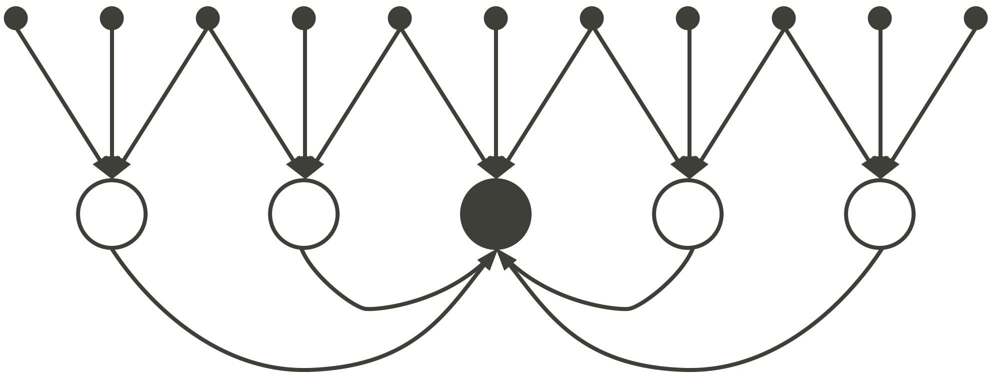


Fig. 1.3: Positive feedback loop

excitatory cells with excitatory lateral connections. Indicated is the feedforward receptive fields of these cells, each one receiving input from three points of stimulus space. Also indicated is the fact that through the lateral excitatory connections, the cells "talk" to a reference cell, colored black. This means that even though a position in stimulus space may not be in the feedforward receptive field of the reference neuron, it may be in the receptive field of another neuron which in turn "talks" to the reference neuron. So, although the position in stimulus space is not in the feedforward receptive field, it is in the receptive field of the reference neuron when it is part of the network.

For a more quantitative heuristic example of such a system, consider the feedback loop shown in Fig. 1.4. Here, we have a positive feedback loop with feedforward system having spatial frequency response function (the Fourier transform of its impulse response function) G and a feedback system having frequency response function K. A basic result in control systems theory tells us that the overall system will have frequency response function given by

$$H = \frac{G}{1 - KG} \tag{1.25}$$

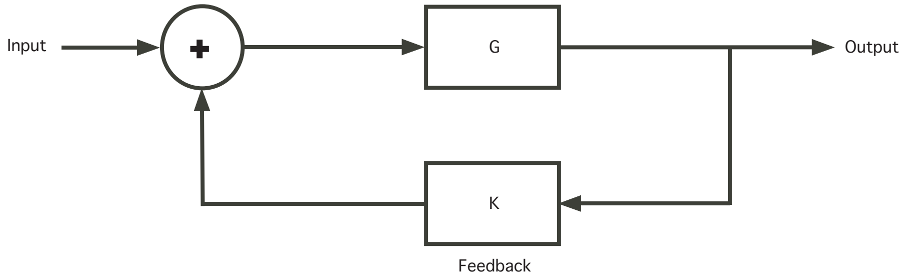


Fig. 1.4: Positive feedback loop

We now examine what happens if the feedforward connectivity has a form similar to that in Eq. (1.20)

$$G(\omega) = \frac{1}{1 + |b\omega|^{\alpha}} \tag{1.26}$$

This happens, for example, in one dimension, if the feedforward connectivity has a Laplacian impulse response function with width b:

$$g(x) = \frac{1}{2b} e^{-|x/b|} \tag{1.27}$$

which has Fourier transform

$$\frac{1}{1 + b^2\omega^2} \tag{1.28}$$

corresponding to $\alpha = 2$. We will also assume that K is just multiplication by a constant $0 < k < 1$. In this situation the effective frequency response function of the entire system is

$$\begin{aligned} H &= \frac{1}{1 + |b\omega|^{\alpha} - k} \\ &= \frac{1}{(1-k)} \frac{1}{|\frac{b}{(1-k)^{\frac{1}{\alpha}}}\omega|^{\alpha} + 1} \end{aligned} \tag{1.29}$$

which is proportional to the Fourier transform of the feedforward impulse response function stretched by a factor $(1-k)^{-\frac{1}{\alpha}}$. Since $0 < k < 1$, this factor will be greater than one. Thus, the positive feedback yields a larger effective receptive field size.

In particular, if we allow k to depend on c

$$k = k(c)$$

and try to make the frequency response in Eq. (1.29) proportional to the ideal response in Eq. (1.21), we will pick $k(c)$ to satisfy

$$\frac{b^\alpha}{1-k(c)} = \frac{a^\alpha}{1+\frac{c^2 k_0}{\sigma^2}} \tag{1.30}$$

This gives

$$k(c) = 1 - \frac{(1+\frac{c^2 k_0}{\sigma^2})b^\alpha}{a^\alpha} \tag{1.31}$$

Since this function decreases with increases in c, we either have positive feedback ($k(c) > 0$) whose magnitude decreases with increasing c, or negative feedback ($k(c) < 0$) whose magnitude increases with increasing c. As mentioned earlier, for physiological reasons, we restrict ourselves to positive feedback values. To do this, we can require that if $0 \leq c \leq c_{max}$ then

$$(1+\frac{c_{max}^2 k_0}{\sigma^2}) \leq (\frac{a}{b})^\alpha \tag{1.32}$$

or equivalently

$$b \leq a(1+\frac{c_{max}^2 k_0}{\sigma^2})^{-\frac{1}{\alpha}} \tag{1.33}$$

Thus we see that positive feedback that decreases at high contrast would allow for an effective receptive field predicted by optimal filtering theory.

1.4 Mechanisms for Positive Feedback

In this section we discuss two possible neural mechanisms for achieving positive feedback whose gain naturally decreases with increasing contrast. The first mechanism involves lateral excitatory cortical connections, which are effectively weakened at high contrast through synaptic depression. The second mechanism is based on the model of [5], in which excitatory cells receive antiphase inhibition as in the model of [6], and in which there is feedback to the LGN that is negative on balance. We will show that the antiphase inhibition can turn this negative feedback to the LGN into a positive feedback loop which disappears at high contrast.

1.4.1 Synaptic Depression

The first mechanism is rather straightforward. The positive feedback comes from lateral excitatory connections between excitatory cells. Although this has been called "lateral" the effects are much the same as "feedback". At low contrast this mechanism will increase the size of the effective receptive field from that of the feedforward connectivity. In this model, the mechanism which reduces this effect at high contrast is that of synaptic depression. Whenever a cell provides synaptic input to another cell, there will be a depletion of synaptic resources, lessening the chance that this presynaptic cell will successfully transmit again. These resources recover, and the likelihood of transmission increases correspondingly, until the cell sends another synaptic transmission. Thus, this synaptic depression will be activated at high firing rates, and will decrease the efficacy of synaptic transmission. In order to quantify this, we used a model described in [7], in which cells having higher firing rates have a lower probability of synaptic transmission. Because cells will tend to have higher firing rates at higher contrast, the effects of synaptic depression will be more severe. The result is that the strength of connectivity between the cells is diminished at high contrast, reducing the strength of the positive feedback, and thus shrinking the receptive field.

1.4.2 Antiphase Inhibition and Feedback to LGN

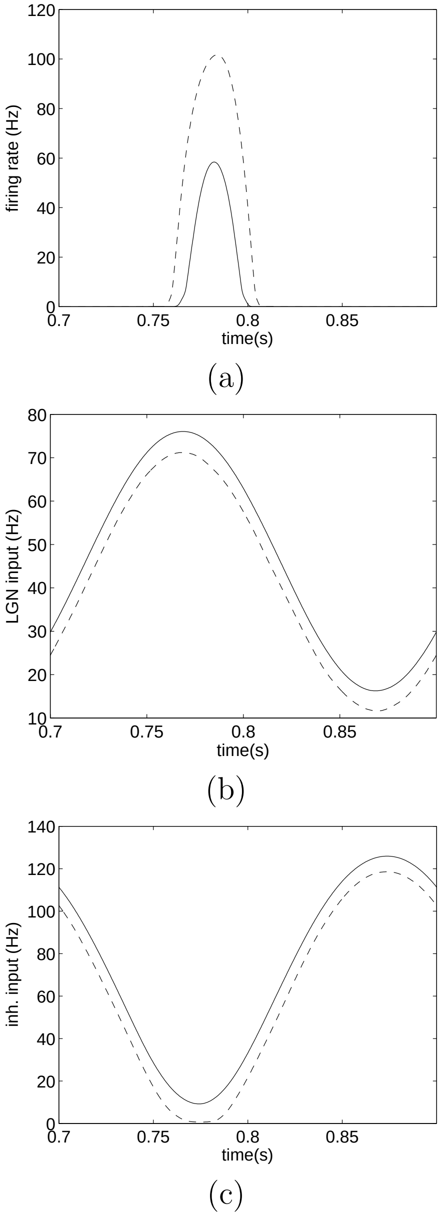
The second mechanism is based on the model of [5]. In this model, excitatory cells receive strong inhibition from inhibitory cortical cells with opposite phase, or antiphase, preference (as in the model of [6]). Also, this model includes feedback from the excitatory cortical cells to the LGN which is negative on balance. The antiphase nature of the inhibitory connectivity means that when the excitatory input from the LGN input to cortical cells is in its sinusoidal "upswing" the inhibitory input will be in its "downswing". Because of this, the strength of the inhibition may be made quite strong without the cell shutting off. This strong inhibition was introduced in the models of [5] and [6] to achieve contrast invariant orientation tuning width.

In Fig. (1.5), a plot of one model cell's cortical response, its excitatory geniculate input, and the inhibitory cortical input to the cell (in this case proportional to the antiphase inhibitory cell's firing rate) are reproduced for a single contrast with and without feedback to the LGN in place. As can be seen, although the feedback to the LGN is negative, and causes the LGN response to drop, the overall effect on the cortical firing rate is a positive one, raising the cortical firing rate. The reason for this can also seen in the concurrent drop in inhibitory input, which also results from the drop of LGN input to the inhibitory cortical cells. Since the inhibition is given so much weight, the cell operates at a point where a drop in LGN input which leads to a corresponding drop in inhibitory input leads to an overall increase in the cortical cell firing rate. Viewed in linear terms, this changes the negative feedback loop to a positive one.

As can be seen from the figure, it is only the behavior in the "downswing" of the inhibitory input which has any effect. In the "upswing" the cortical cell is just turned off. At very high contrast, the inhibitory input during the downswing is clipped at zero, and cannot be made to decrease any further. Because of this, the circuit no longer possesses positive feedback and the receptive field size will be that of the feedforward connectivity.

1.5 Computational Models

As mentioned above, in the expanding aperture experiments of [1], the orientation of the stimulus was fixed, and the only parameter which was altered was the aperture radius. Thus, the experiments did not depend on the two dimensional structure of the stimulus. Furthermore, our theoretical development does not rely on the two dimensional structure of the stimulus or cortical connectivity, and would be equally valid in a model with one dimensional stimulus and cortical connectivity, as long as the model has the correct feedback structure. Hence, for computational and conceptual simplicity, we will study two one-dimensional models with the feedback mechanisms we have described. The first model will have lateral excitation and synaptic depression. The second model will have feedback



(a)

(b)

(c)

Fig. 1.5: responses of the cortical cell at the center of the circular window (Fig. 1.1) (- is without feedback - - is with) (a) cortical excitatory cell (b) excitatory geniculate input (c) inhibitory input

to the LGN and antiphase inhibition of excitatory cortical cells.

The structure of the models, as well as the modeling of cellular behavior will be based on the model developed in [5]. Although we do not discuss it here, we have implemented a fully two dimensional version of the model with LGN feedback and antiphase inhibition, and this model was able to produce contrast dependent receptive field size similar to the one dimensional model, while maintaining the two-dimensional characteristics studied in [5, 6], namely, contrast independent orientation preference and sensitivity to orientation discontinuity.

1.5.1 Elements Common to both Models

The basic structure of both models is a one dimensional lattice of retinal ganglion cells (RGC) which spatially pool and temporally process the one dimensional stimulus, followed by a lattice of LGN cells which receive input from these RGCs and which in turn excite lattices of both excitatory and inhibitory cortical cells (See Fig. 1.6 for a cartoon summarizing the general structure). In this model, stimulus space is represented as a one dimensional lattice of points. Model RGCs are located on a lattice subsampled from the stimulus space lattice, with both an ON and OFF RGC located at each position on the sublattice. For this reason, cells at different positions may be indexed by referencing their associated position in stimulus space, x. The ON and OFF cells are excited or inhibited, respectively, by a light stimulus presented in the center of the receptive field.

The first stage of processing in the model is a linear spatial pooling mechanism for the RGCs, which represents their receptive field properties. This spatially pooled data is given by the convolution of the stimulus $S(x, t)$ with the receptive

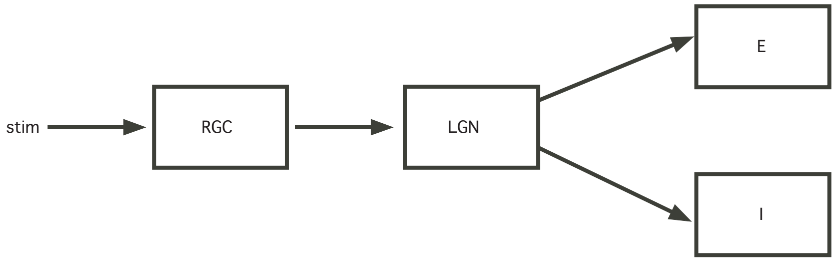


Fig. 1.6: General structure common to both models

field H:

$$S_{pool}(x,t) = H(x) \star S(x,t) \tag{1.34}$$

where $\star$ is the spatial convolution operator for fixed t, and where the receptive field, H, is modeled by a DoG. The spatial convolution is computed by multiplying the fast Fourier transforms (FFT) of the data and the receptive field and computing the inverse FFT. The convolution used is therefore the circular convolution, which can be interpreted as implying a periodic structure on the stimulus space or on the neuronal sublattice, or as implying long range connectivity between cells at the edges. However, in all of the experiments we will discuss, it is the response of the center cell that is measured, and since the connections between the center cell and cells at the edge are weak in this model, the measured response will not be much altered by boundary effects.

After the spatial pooling, the pooled stimulus is downsampled, and this signal, $s(x,t)$, is given on the retinal sublattice, with x referring to the cells at position x. Next, $s(x,t)$ is used as the input for the temporal model used in [8]. This model is described by four temporal equations:

$$\begin{aligned} x_{lo}(t) &= \int_0^\infty \frac{s(t-t')}{N_L!}(\frac{t'}{T_L})^{N_L-1}e^{-t'/T_L}dt' && (1.35) \\ T_S(t)\dot{x}_{hi}(t) &= -x_{hi}(t) + T_S(t)\dot{x}_{lo}(t) + (1-H_S)x_{lo}(t) && (1.36) \\ T_S(t) &= \frac{T_0}{1+\frac{c(t)}{c_{\frac{1}{2}}}} && (1.37) \\ T_C\dot{c}(t) &= |x_{hi}(t)| - c(t) && (1.38) \end{aligned}$$

The spatial indexing in these equations is omitted, since the calculation for each spatial index is carried out separably for all time. The calculation in Eq. (1.35) is carried out by writing $x_{lo}(t)$ as the solution of the system of first order differential equations

$$x_i + T_L\frac{dx_i}{dt} = x_{i-1}\text{for i=1,...}N_L \qquad (1.39)$$

where x_0 is the stimulus, $s(t)$, and the output, $x_{lo}(t)$, is identified as x_{N_L}. In [5] a simpler model for RGC firing rates is used, which does not have a temporal component. However, this model does not account for saturation effects that are found experimentally, and so multiplication by a nonlinear function of contrast is introduced. The retinal model which we use gives more saturation then the model of [5] without making explicit use of the contrast stimulus parameter. It is unlikely, however, that differences in the models of individual RGC cells would have much impact on phenomena like receptive field sizes, which are more heavily determined by interactions between cortical and LGN cells.

The output of each model retinal cell is then put through a nonlinearity of the form

$$r_{out} = \frac{r_{max}x_{hi}}{r_{\frac{1}{2}} + |x_{hi}|} \qquad (1.40)$$

where, again, the spatial and temporal indexes are omitted, since the operation is performed on each value of $x_{hi}(x, y, t)$ separately. The purpose of this nonlinearity is to introduce more saturation into the LGN response, and the parameters r_{max} and $r_{\frac{1}{2}}$ are chosen so that the F1 component of r_{out} in response to a drifting grating at various contrasts matches that of the model of [5].

Next, the firing rate of the retinal ganglion cell, $r^R(x, t)$ is computed from

$$r^R(x,t) = [r_0 \pm r_{out}(x,t)]^+ \tag{1.41}$$

where the "+" and "-" in the brackets are for ON and OFF cells, respectively, and where the $[\cdot]^+$ operation is rectification.

The next stage in the model is a set of LGN ON and OFF cells, with one of each located at every position on the retinal sublattice. The model LGN cell at location x receives input from the corresponding RGC with the same location x and same ON/OFF label as the LGN cell, at rate $\nu_e^{R\to L}(x, t)$, which is proportional to the RGC's firing rate. There will also be a background firing rate input ν_e^0, which is stimulus independent. In the model which contains feedback from the cortex to LGN there will also be input from the cortex at with excitatory rate $\nu_e^{C\to L}(x, t)$ and inhibitory rate $\nu_i^{C\to L}(x, t)$. As in [5], the model used for cells in the LGN and cortex was a conductance based rate model, in which a quasi-stationary assumption on the conductance allows for the modeling of the evolution of the conductance and conversion of the conductance to instantaneous firing rate via a nonlinearity. If a model cell receives excitatory input at rates ν_e^k and inhibitory input at rates ν_i^k, and if g_e and g_i are the excitatory and inhibitory conductances, respectively, of the cell, the conductances evolve according to

$$\begin{aligned}\frac{\partial g_e}{\partial t} &= -\frac{g_e}{\tau_e} + \sum_k \gamma_e^k \nu_e^k \\ \frac{\partial g_i}{\partial t} &= -\frac{g_i}{\tau_i} + \sum \gamma_i^k \nu_i^k \qquad (1.42)\end{aligned}$$

for appropriate time constants τ_e and τ_i and connectivity constants γ_i^k and γ_i^k. The firing rate is then modeled as an instantaneous function of the conductances

$$r = \frac{1}{\tau_{ref} - \tau f(V)} \qquad (1.43)$$

where

$$\tau = \tau_m/(1 + g_e + g_i) \qquad (1.44)$$

for appropriate time constants τ_m and τ_{ref} and where

$$V = (g_e E_e + g_i E_i)/(1 + g_e + g_i). \qquad (1.45)$$

Here E_e and E_i are, respectively, the excitatory and inhibitory equilibrium potentials measured relative to rest, and normalized by the firing threshold potential. We use a nonlinearity of the form

$$f(v) = \frac{c}{\sigma \exp(-\frac{(v-v_{th})^2}{2\sigma^2}) + (v - v_{th})\mathrm{erfc}(-\frac{v-v_{th}}{\sigma})} \qquad (1.46)$$

which is equal to

$$\frac{1}{E([V - v_{th}]^+)} \qquad (1.47)$$

where V is $N(v, \sigma^2)$. This functional form is based on an alteration of the neuronal model of [7], which models the firing rate as the expectation of a Gaussian random variable which has mean equal to a voltage variable. The parameters v_{th}, σ and c are chosen to fit the function

$$\begin{cases} -\ln[1-\frac{1}{v}], & v > 1 \\ 0, & v \leq 1 \end{cases}$$

which is the function used in [5], based on an integrate and fire neuron. The nonlinearity we use is therefore nothing more than a smoothed version of that for the integrate and fire neuron. This smoother nonlinearity allows for a smoother onset of firing, which in turn allows for more stable behavior of the network, and also allows for low firing rates in the model.

The next stage in the model is the layer 6 cortical cells, which are also located on the retinal sublattice. At every location on the sublattice there is a single model excitatory cell and a single model inhibitory cell for each phase preference ϕ, ($\phi = 0°, 180°$). The rate of excitatory input from the LGN to cortical cells at position x with phase preference ϕ is related to the one dimensional Gabor function

$$\mathcal{G}(x,\phi) = \exp(-[\frac{x^2}{{\sigma_w}^2}]) \times \cos[2\pi k'(x) + \phi] \tag{1.48}$$

where $\sigma_w = 0.9°$ and $k' = 0.6$ cyc/deg. The excitatory input rate from the LGN is given by

$$\begin{aligned} \nu_e^{L\to C}(x,t,\phi) &= \sum_i([\mathcal{G}(x-x_i,\phi)]^+ \times r_{ON}^L(x_i,t-t_d) \\ &- [\mathcal{G}(x-x_i,\phi)]^- \times r_{OFF}^L(x_i,t-t_d)) \end{aligned} \tag{1.49}$$

here "-" means the negative part, and "+" as before is the positive part. The delay time t_d simply models synaptic delay. Excitatory cells also receive inhibitory input from the inhibitory cortical cells at a rate $\nu_i^{C\to C}$, which will be modeled differently for the two models, and which will be described for each model in later subsections. In the model which uses the mechanism of synaptic depression, there will also be lateral excitatory input at rate $\nu_e^{C\to C}$.

The inhibitory cell receives excitatory input from the LGN, using the same type of connectivity as for the input to the excitatory cells. The evolution of the excitatory and inhibitory firing rates in terms of their input rates is given by the same equations which govern the evolution of the geniculate firing rates, Eq. (1.42)-Eq. (1.46).

1.5.2 Synaptic Depression

In the model which used the mechanism of synaptic depression, the model also includes excitatory to excitatory connectivity which gives heavier weights to cells which are located nearby, and which gives equal weights to cells of both phase preference. Also, the firing rates of the inhibitory cells are pooled across phase preference to provide inhibitory input to the excitatory cells that has no phase preference (see Fig. 1.7 for a cartoon). Although pooling lateral excitatory and feedforward inhibitory input across phase preferences may not be necessary, we want to illustrate that this connectivity does not require the antiphase connectivity of [6], since there are models which do not use this connectivity [9].

The rate of lateral excitation to a cell at position x, $\nu_e^{C\to C}(x,t)$ is determined

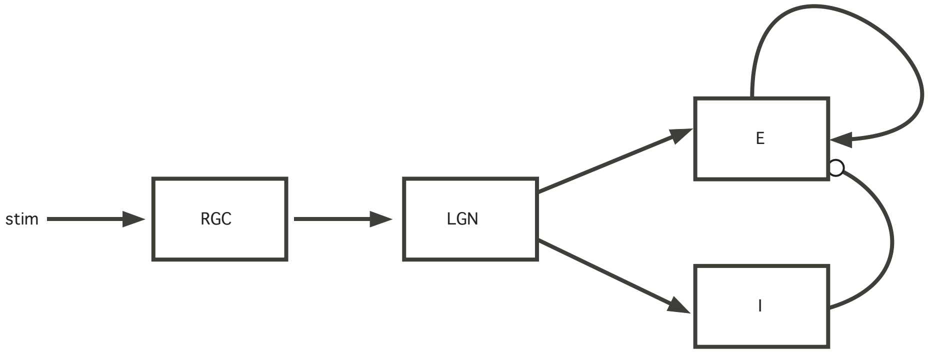


Fig. 1.7: Structure of model using synaptic depression mechanism

by

$$\nu_e^{C\to C}(x,t) = \sum_{\phi} C_{e\to e}(x) \star \Big(r_e^C(x,t,\phi) p(x,t,\phi) \Big) \tag{1.50}$$

where $r_e^C(x,t,\phi)$ is the firing rate at time t of the excitatory cell at location x with phase preference ϕ, p represents the probability of synaptic transmission for this cell, to be discussed shortly, and where $C_{e\to e}$ represents the lateral excitatory connectivity (In our model we use a Gaussian with standard deviation of one degree). Similarly, the excitatory cells receive inhibitory input, $\nu_i^{C\to C}(x,t)$, from the inhibitory cells, at rate

$$\nu_i^{C\to C}(x,t) = \sum_{\phi} C_{i\to e}(x) \star_x r_i^C(x,t,\phi) \tag{1.51}$$

where $r_i^C(x,t,\phi)$ is the firing rate at time t of the inhibitory cell at location x with phase preference ϕ. We do not spatially pool inhibitory input, so $C_{i\to e}$ is just a delta function in space.

As mentioned earlier, the mechanism for attenuating the excitatory input is to include the effects of synaptic depression in this connectivity. In order to model this synaptic depression we use the simplified model cited in [7]

$$\frac{dp}{dt} = \frac{u - p}{\tau_R} - upr_e^C \tag{1.52}$$

where p is the probability of synaptic transmission r_e^C is the presynaptic firing rate of an excitatory cell and u is a utilization parameter and where we have left out indexing for simplicity. In our model, the excitatory firing rates are multiplied by these probabilities in order to determine the lateral excitatory input. In order to see how this affects the input we solve for the steady state solution of this equation, with constant firing rate f

$$p_\infty = \frac{u}{1 + \tau_R u f} \tag{1.53}$$

As we can see, the probability decreases with the presynaptic firing rate, as expected. Thus, at higher contrasts the higher firing rates cause the lateral excitatory input to have less of an effect.

1.5.3 Antiphase Inhibition and Feedback to LGN

In this model the LGN stage not only receives spatially pooled input from the retina, but also receives feedback input from the excitatory cortical cells which is pooled spatially and across phase preference. The excitatory cortical neurons provide excitatory feedback through interneurons to the LGN cell at position x at a rate $\nu_e^{C \to L}(x, t)$. At the same time, the excitatory cortical cells stimulate perigeniculate neurons, which in turn provide inhibitory input to the LGN cell at position x at a rate $\nu_i^{C\to L}(x,t)$. These signals are given by:

$$\nu_{e/i}^{C\to L}(x,t) = \sum_{m=1}^{2}\sum_{n} P_{e/i}(x-x_n) \times r_e^C(x_n, t-t_{e/i}^d, \phi_m) \tag{1.54}$$

where $r_e^C(x,t,\phi)$ is the firing rate of an excitatory cortical cell with preferred spatial phase ϕ, $\phi_1 = 0$, $\phi_2 = \pi$, and $t_{e/i}^d$ is the net delay for excitation/inhibition, set to mimic transmission through interneurons or perigeniculate neurons. $P_{e/i}(x)$ is a Gaussian function describing the synaptic footprint of excitatory/inhibitory cortical feedback:

$$P_{e/i}(x) = \exp\Big(-[x^2/\sigma_{W_{e/i}}^2]\Big) \tag{1.55}$$

For excitatory cortical feedback, the characteristic width of the two Gaussians for the synaptic footprint (Eq. 1.55) that we use are $\sigma_{W_e} = 0.2$ and $\sigma_{W_i} = 0.32$.

Also, in this model, the inhibitory input to the excitatory neurons is antiphase, with inhibitory neurons of one phase preference inhibiting excitatory neurons of opposite preference. The excitatory cell at position x with phase preference ϕ receives inhibitory input from inhibitory neurons at a rate $\nu_i^{C\to C}(x,t,\phi)$, where

$$\nu_i^{C\to C}(x,t,\phi) = r_i^C(x,t,\phi+\pi) \tag{1.56}$$

As described, this strong antiphase inhibition has the effect of turning the negative feedback to the LGN into a positive feedback loop (see Fig. 1.8 for a cartoon of this connectivity).

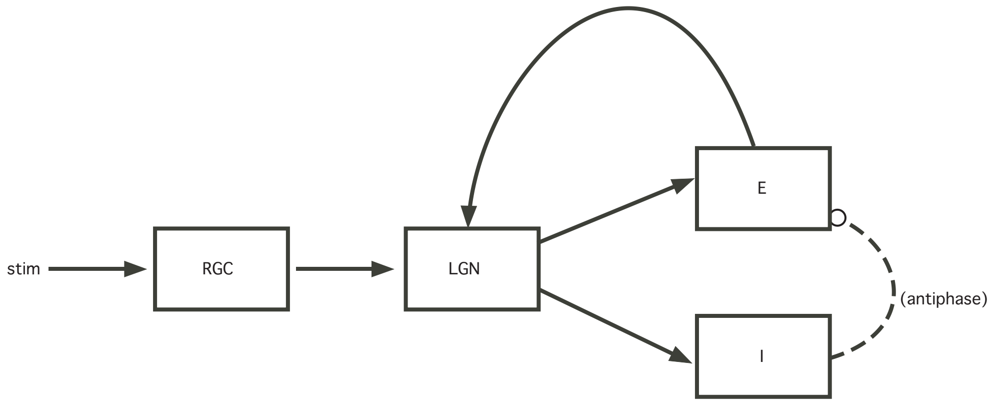


Fig. 1.8: Structure of model using feedback to LGN mechanism

1.6 Simulations

In this section we will describe the results of simulations performed with the described models. The simulation for each model was done using a drifting one dimensional sinusoid which was inside of an interval $(-s, s)$, and the F1 response of the excitatory cortical cell located at the center of the aperture was recorded. This experiment was performed for various values of the contrast of the sinusoid. As we have discussed, for real cortical cells, the aperture radius which gives the peak or saturating response, tends to decrease with an increase in contrast. This can be measured quantitatively, by fitting the F1 response as a function of s by a DoG model.

1.6.1 Results for synaptic depression

In Fig. 1.9 , F1 as a function of aperture size is plotted for stimuli at various contrasts. Curves having higher firing rates corresponding to higher contrast. As can be seen in this figure, the effective receptive field size at low contrast is larger

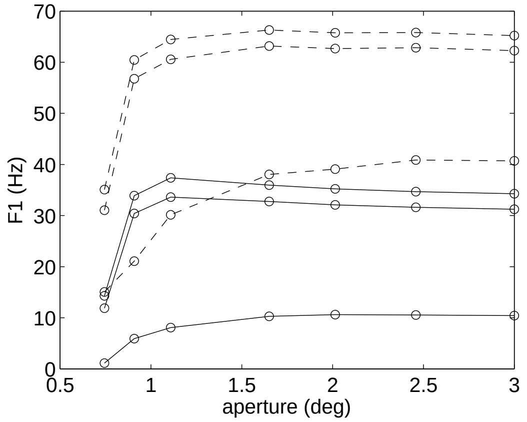


Fig. 1.9: Responses of the cortical cell at the center of the circular window (Fig. 1.1). - is without lateral excitation and – is with lateral excitation and suppression

than at high contrast for the neural circuit which makes use of the mechanism of lateral excitation with synaptic depression. Curves for two high contrast values are shown, to show that the curve at high contrast is not peaking at lower radius because its response is saturating. For comparison purposes, 1.9 also shows the result when lateral excitatory connectivity is removed. As can be seen, the difference in receptive field sizes is not as pronounced. It should be noted that removing lateral excitation significantly alters the operating point of the circuit, and so this comparison is not necessarily a fair one. However, it does serve to illustrate that it is the feedback mechanism which is causing this phenomenon.

Figure 1.10 shows a weighted least squares fit using a difference of Gaussians (DoG) fit as in [1], and the ratio of high contrast to low contrast excitatory width with and without feedback. This further illustrates that the positive feedback is giving us a larger receptive field size at low contrast. While the increased receptive field size at low contrast is not as pronounced as that using the model with feedback

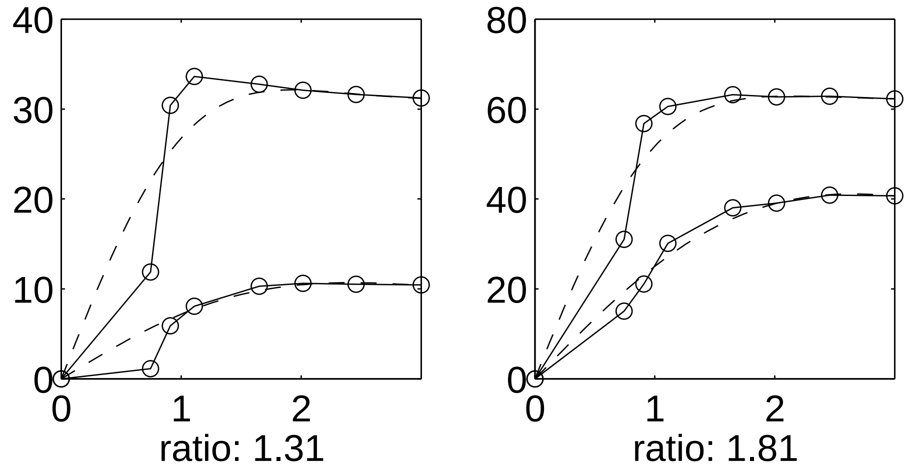


Fig. 1.10: Fit of responses with DoG; left is without lateral excitation, right is with lateral excitation and synaptic depression

to the LGN, which we will discuss shortly, this model does have the benefit of using a more physiologically plausible mechanism, which does not rely on antiphase inhibition.

1.6.2 Results for model with feedback to LGN

Figure 1.11, shows the plot of F1 response as a function of aperture size for the model that uses feedback to the LGN and antiphase inhibition. As can be seen, this mechanism gives a pronounced increase in effective receptive field size at low contrast. Figure 1.12 shows the DoG fit for this model, with the ratio of high contrast to low contrast excitatory width with and without feedback. Note that, because the response does not asymptote within the region tested, the ratio for the case with feedback is larger that it should be. Also note that in this case, responses with and without feedback may be more fairly compared since they have firing rates of the same order of magnitude. The results serve to illustrate that we

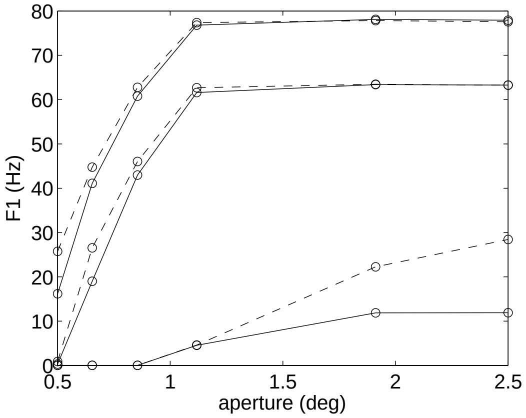


Fig. 1.11: responses of the cortical cell at the center of the circular window(Fig. 1.1) - is without feedback to LGN in place – is with

are getting an increase in receptive field size by using the mechanism of positive feedback.

1.7 Discussion

In this chapter, we have shown that optimal filtering theory provides a heuristic framework for understanding the physiological purpose behind an increase in receptive field at low contrast. We have also shown that this effect may be achieved either through positive feedback which decreases in strength with increasing contrast, or through negative feedback which increases in strength with increasing contrast. For physiological reasons, we have concentrated on the former mechanism. We have introduced two possible neural mechanisms for achieving positive feedback which attenuates at high contrast. The first is through lateral excitation which is weakened by synaptic depression. The second is through feedback to the LGN which is negative on balance, coupled with antiphase inhibition. Finally we

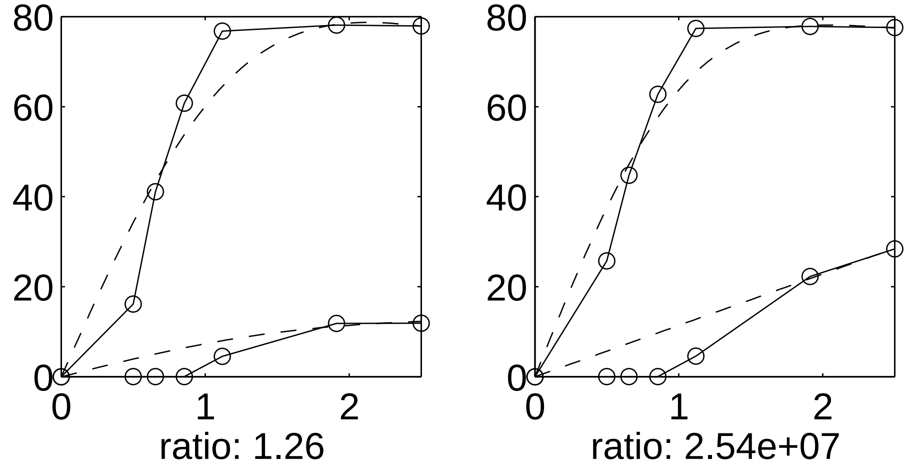


Fig. 1.12: Fit of responses with DoG left is without feedback to LGN right is with feedback

have performed experiments with models based on these connectivities and shown that these models do exhibit contrast dependent receptive field sizes.

As noted, the model which uses feedback to the LGN with antiphase inhibition provides a more pronounced increase in receptive field size at low contrast that the model which utilizes the mechanism of lateral excitation. Furthermore it should be noted that the mechanism which involved lateral excitation is naturally prone to instability because of this excitation, and so parameters had to be carefully chosen to provide stable circuit behavior. It might be feasible to stabilize this model using inhibition which receives excitatory input from the excitatory neurons. However, this would then introduce a negative feedback loop, and so the connectivities must be carefully chosen in order to stabilize the model while preserving overall lateral excitation.

Despite these problems, the mechanism of lateral excitation provides a more physiologically reasonable mechanism for positive feedback, since it is local to the cortex and does not involve antiphase connectivity. In order to see which if either of these mechanisms are responsible for the increase in receptive field size at low contrast, it would seemingly be necessary to perform experiments in which feedback to the LGN is removed. However, this cannot be done by freezing cortical cells, since it is their response which we need to measure.

1.8 Prelude to Prior Free

As a prelude to the next chapter, we recall some facts about the optimal linear filter for the model of corruption discussed in Section 1.2. Recall that in this model, the components of the underlying signal, X, were corrupted by independent and identical corruption processes, and that we made the assumption that

$$E\{N|X\} = X$$

We also assumed in Section 1.2 that X was stationary, an assumption which we shall drop in this section. Proceeding along the same lines as in this section, we can show that in this case as well the optimal linear estimator

$$\widehat{X}(N) = HN$$

is determined by

$$C_{XN} = HC_{NN}$$

As in that section, we may also show that

$$C_{XN} = C_{XX}$$

However, we must now slightly alter the equation for C_{NN} to be

$$C_{NN} = C_{XX} + D_\sigma \tag{1.57}$$

where D_σ is a diagonal matrix whose i^{th} diagonal element is equal to

$$(D_\sigma)_{ii} = Var(N_i) - Var(X_i) \tag{1.58}$$

In the stationary case, all diagonal elements are the same, so D_σ reduces to multiple of the identity.

We can therefore write the optimal linear operator in a form very similar to that used in Section 1.2

$$H = C_{XX}(C_{XX} + D_\sigma)^{-1} \tag{1.59}$$

Even if X is not stationary, it still may happen that D_σ is a multiple of the identity. In that case there will be a unitary matrix, Q which diagonalizes C_{XX}

$$Q^* C_{XX} Q = S_X \tag{1.60}$$

where S_X is a diagonal matrix, so that

$$Q^* H Q = S_X(S_X + \sigma I)^{-1} \tag{1.61}$$

This is just as in Section 1.2, except that the unitary operator need not be the Fourier transform. Note that this form of the optimal linear solution makes it clear that the optimal linear solution is a contraction towards the origin. We will return to this theme in Chapter 2.

A possible criticism of using Eq. (1.59) as a basis for explaining physiological behavior is its reliance on knowledge of the statistics of X. Considering that the system only observes corrupted measurements, N, it does not seem possible for such a system to "learn" the statistics of X. However, in this situation, it is easy to see that this problem may be easily overcome by writing the optimal filter in its equivalent form

$$H = (C_{NN} - D_{\sigma})(C_{NN})^{-1} \tag{1.62}$$

In this equation, the only possible involvement of the statistics of X is through the matrix D_{σ}. For many corruption processes, however, D_{σ} does not depend on X. For example, if the corruption process is additive noise with known variance σ^2, then D_{σ} will be $\sigma^2 I$. In the case of corruption by Poisson noise (where each element of N is Poisson with rate equal to the corresponding element of X), it is easily shown that, since $E\{N_i\} = E\{X_i\}$

$$\begin{aligned} & var\{N_i\} - var\{X_i\} \\ = \;& E\{N_i^2\} - E\{X_i^2\} \\ = \;& E\{N_i\} \end{aligned} \tag{1.63}$$

In such cases, it is therefore possible to write the difference between the variance of the observed process and that of the underlying process without reference to the statistics of the underlying process. In this way, we are able to write the optimal linear operator without referring to the statistics of the underlying process, but, rather, entirely in terms of the statistics of the observed process. In the next chapter, we shall show that this phenomenon occurs for many other types of corruption processes, and discuss ways of using this to obtain optimal estimators without needing to use the statistics of the underlying signals, which are unavailable to the system which is trying to do the estimating.

In our particular case, we are interested in finding the covariance of the observed data, C_{NN}. We may do this by either using an ensemble of images to estimate the covariance between any two elements of the observed vector, as in [2]. Alternatively, if the observed vector is assumed to be wide sense stationary (which would be true if the underlying signal was wide sense stationary), then the covariance between any two elements will depend only on the displacement between them. Thus, we may calculate the covariance for a given displacement by averaging the product between all pairs of elements of the observed vector with that displacement.

Chapter 2

Prior-Free Estimation

In the previous chapter we discussed the variance minimizing linear estimator which estimates an underlying signal from a corrupted or observed version of that signal. We pointed out that under certain circumstances, if we know the observation process, it is possible to write this optimal solution without reference to the statistics of the underlying signal, but, rather, entirely in terms of the statistics of the observed signal. In this chapter we will generalize these results significantly. We will begin by showing that in many cases the optimal estimator (not necessarily linear) can be formulated without reference to the statistics of the underlying signal. We will also show that for a wide range of observation processes, if we have a family of estimators (which does not necessarily contain the optimal estimator) we can pick the optimal member of the family using the statistics of the underlying signal. This generalizes the result of the last chapter, which was for a particular set of corruption processes and for a particular family of estimators, namely linear ones.

2.1 BLS Estimators

While it was natural, in the physiological setting of the previous chapter, to minimize the variance of the error, in this chapter we will concentrate on minimizing the more commonly used Mean Squared Error (MSE) of the estimate for which the results of the previous chapter would be very similar. The estimator which minimizes the MSE is known as the Bayesian Least Squares (BLS) estimator because, as we shall see, it is often formulated using Bayes' rule. This estimator is also commonly referred to as the regression function. Bayesian methods are widely used throughout engineering for estimating quantities from corrupted measurements, with the BLS estimators being particularly widespread. These estimators are commonly derived by assuming explicit knowledge of the observation process (expressed as the conditional density of the observation given the underlying signal), and the distribution of the underlying signal, known as the prior distribution. Despite its appeal, this approach is often criticized for this reliance on knowledge of the prior distribution, since the true prior is usually not known, and in many cases one does not have data drawn from this distribution with which to approximate it. In this case, it must be learned from the same observed measurements that are available in the estimation problem. In general, learning the prior distribution from the observed data presents a difficult, if not impossible task, even when the observation process is known. In the commonly used ”Empirical Bayesian” approach [10], one assumes a parametric family of prior densities and then chooses the parametric prior which, together with the known corruption model, gives the best fit to the observed data data. This prior is then used to derive a Bayes estimator that may be applied to the data. If the true prior is not a member of the assumed parametric family, however, such estimators can perform quite poorly.

An estimator may also be obtained in a *supervised* setting, in which one is provided with many pairs containing a corrupted observation along with the true value of the quantity to be estimated. In this case, selecting an estimator is a classic regression problem: find a function that best maps the observations to the correct values, in a least squares sense. Given a large enough number of training samples, this function will approach the BLS estimate, and should perform well on new samples drawn from the same distribution as the training samples. In many real-world situations, however, one does not have access to such training data.

In this chapter, we examine the BLS estimation problem in a setting that lies between the two cases described above. Specifically, we assume the observation process (but not the prior) is known, and we assume *unsupervised* training data, consisting only of corrupted observations (without the correct values). We show that for many observation processes, the BLS estimator may be written directly in terms of the density of the corrupted observations. The precise form of the estimator depends on the observation process, and the examples we derive here provide an illustration of the diversity of such formulations. We also show a dual formulation, in which the BLS estimator may be obtained by minimizing an expression for the mean squared error that is written only in terms of the observation density. There are a few special cases of the first formulation in the Empirical Bayes literature [11], and of the second formulation in another branch of the statistical literature concerned with improvement of estimators [12, 13, 14]. Our work serves to unify these methods within a linear algebraic framework, and to generalize them to a wider range of cases. We develop practical examples of nonparametric approximations for several different observation processes, demonstrating empirically that the resulting estimators converge to the Bayes least squares estimator as the amount of observed data increases. We also develop a parametric family of estimators for use in the additive Gaussian case, and examine their empirical convergence properties. We expect such BLS estimators, constructed from corrupted observations without explicit knowledge of, assumptions about, or samples from the prior, to prove useful in a variety real-world estimation problems faced by machine or biological systems that must learn from examples.

2.2 Common Formulations of BLS Estimators

Suppose we make an observation, Y, that depends on a hidden variable X, where X and Y may be scalars or vectors. Given this observation, the BLS estimate of X is simply the conditional expectation of the posterior density, $E\{X|Y=\mathbf{y}\}$. If the prior distribution on X is P_X, and the likelihood function is $P_{Y|X}$ then this can be written using Bayes' rule as

$$\begin{aligned} E\{X|Y=\mathbf{y}\} &= \int \mathbf{x}\, P_{X|Y}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \\ &= \int \mathbf{x}\, P_{Y|X}(\mathbf{y}|\mathbf{x})\, P_X(\mathbf{x})\, d\mathbf{x} \Big/ P_Y(\mathbf{y}) \ , \end{aligned} \tag{2.1}$$

where the denominator is the distribution of the observed data:

$$P_Y(\mathbf{y}) = \int P_{Y|X}(\mathbf{y}|\mathbf{x})\, P_X(\mathbf{x})\, d\mathbf{x} \ . \tag{2.2}$$

If we know P_X and $P_{Y|X}$, we can calculate this explicitly.

Alternatively, if we do not know P_X or $P_{Y|X}$, but are given independent identically distributed (i.i.d.) samples (X_n, Y_n) drawn from the joint distribution of

(X, Y), then we can solve for the estimator $f(\mathbf{y}) = E\{X|Y = \mathbf{y}\}$ nonparametrically, or we could choose a parametric family of estimators $\{f_\theta\}$, and choose θ to minimize the empirical squared error:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} |f_\theta(Y_n) - X_n|^2. \tag{2.3}$$

However, in many situations, one does not have access to P_X, or to samples drawn from P_X.

2.3 Prior-free reformulation of the BLS estimator

In many cases, the BLS estimate may be written without explicit reference to the prior distribution. We begin by noting that in Eq. (2.1), the prior appears only in the numerator

$$N(\mathbf{y}) = \int P_{Y|X}(\mathbf{y}|\mathbf{x})\, \mathbf{x}\, P_X(\mathbf{x})\, d\mathbf{x}. \tag{2.4}$$

This equation may be viewed as a composition of linear transformations of the function $P_X(\mathbf{x})$

$$N(\mathbf{y}) = (\mathbf{A} \circ \mathbf{X})\{P_X\}(\mathbf{y}),$$

where

$$\mathbf{X}\{f\}(\mathbf{x}) = \mathbf{x} f(\mathbf{x}),$$

and the operator $\mathbf{A}$ computes an inner product with the likelihood function

$$\mathbf{A}\{f\}(\mathbf{y}) = \int P_{Y|X}(\mathbf{y}|\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x}.$$

Similarly, Eq. (2.2) may be viewed as the linear transformation $\mathbf{A}$ applied to $P_X(\mathbf{x})$.

If the linear transformation $\mathbf{A}$ is 1-1, and we restrict P_Y to lie in the range of $\mathbf{A}$, then we can then write the numerator as a linear transformation on P_Y alone, without explicit reference to P_X:

$$\begin{aligned} N(\mathbf{y}) &= (\mathbf{A} \circ \mathbf{X} \circ \mathbf{A}^{-1})\{P_Y\}(\mathbf{y}) \\ &= \mathbf{L}\{P_Y\}(\mathbf{y}). \end{aligned} \tag{2.5}$$

In the definition of the operator $\mathbf{L}$, $\mathbf{A}^{-1}$ effectively inverts the observation process, recovering P_X from P_Y. We will refer to the observation process where $\mathbf{A}$ is one to one as invertible observation processes. This allows us to write the BLS estimator as

$$E\{X|Y = \mathbf{y}\} = \frac{\mathbf{L}\{P_Y\}(\mathbf{y})}{P_Y(\mathbf{y})}. \tag{2.6}$$

In the discrete case, $P_Y(\mathbf{y})$ and $N(\mathbf{y})$ are each vectors, $\mathbf{A}$ is a matrix containing $\mathbf{P}_{Y|X}$, $\mathbf{X}$ is a diagonal matrix containing values of $\mathbf{x}$, and $\circ$ is matrix multiplication. $\mathbf{L}$ will then be a matrix. As we will see, Eq. (2.6) serves to generalize and unify some specific results which appear in the Empirical Bayes literature [15, 16, 11].

Note that if we wished to calculate $E\{X^n|Y\}$, then Eq. (2.5) would be replaced by $(\mathbf{A} \circ \mathbf{X}^n \circ \mathbf{A}^{-1})\{P_Y\} = (\mathbf{A} \circ \mathbf{X} \circ \mathbf{A}^{-1})^n\{P_Y\} = \mathbf{L}^n\{P_Y\}$. By linearity of the conditional expectation, we may extend this to any polynomial function (and thus to any function that can be approximated with a polynomial):

$$E\left\{f(X)|Y = \mathbf{y}\right\} = \frac{f(\mathbf{L})\{P_Y\}(\mathbf{y})}{P_Y(\mathbf{y})}. \tag{2.7}$$

with

$$f(x) = \sum_{k=-N}^{M} c_k x^k \tag{2.8}$$

Thus, our linear algebraic framework allows us to find the linear operator for the estimator of a function of the hidden variable by taking the same function of the linear operator. In particular this implies that

$$E\left\{f(X)\right\} = E\left\{\frac{f(\mathbf{L})\{P_Y\}(Y)}{P_Y(Y)}\right\}. \tag{2.9}$$

which allows us to obtain the statistics of X from those of Y. This shows that, while practically unfeasible, it is theoretically possible to calculate all the moments of X using corresponding statistics of Y. For invertible observation processes this will uniquely determine the prior density, P_X. However, as we shall see, for some noninvertible observation processes, we may still be able to define an appropriate operator, L, that allows for prior free estimators. In this situation, we will still be able to compute the moments of X in terms of statistics of Y, but these moments will not uniquely define the distribution P_X.

More generally, we may wish to find

$$E\left\{\frac{\mathbf{M}\{P_X\}(X)}{P_X(X)}|Y = y\right\} \tag{2.10}$$

for some linear operator $\mathbf{M}$. Of course, a special case of a linear operator is multiplication by a function of X, which we have just covered. Another case of interest is if X is in turn a corrupted observation of another hidden variable Z, where $\mathbf{M}$ is the linear operator associated with the prior free estimator of Z in terms of X. In this case we may write

$$\begin{aligned} E\{Z|Y=y\} &= E\{E\{Z|X\}|Y=y\} \\ &= E\left\{\frac{\mathbf{M}\{P_X\}(X)}{P_X(X)}|Y=y\right\} \end{aligned} \tag{2.11}$$

as desired. If we try to calculate the linear operator for estimating Z from Y directly using a calculation similar to Eq. (2.5), we would get

$$E\{Z|Y=y\} = \frac{(\mathbf{A}\circ\mathbf{M}\circ\mathbf{A}^{-1})\{P_Y\}(y)}{P_Y(y)} \tag{2.12}$$

which involves knowing the operator $\mathbf{A}$. Instead, we show that we may write this estimator without having to do this inversion.

To calculate Eq. (2.10), we may write

$$\begin{aligned} &\int \frac{\mathbf{M}\{P_X\}(x)}{P_X(x)} P_X(x) P_{Y|X}(y|x)dx \\ =& \int \mathbf{M}\{P_X\}(x) P_{Y|X}(y|x)dx \\ =& \int P_X(x)\mathbf{M}^*_x\{P_{Y|X}(y|x)\}dx \\ =& \int P_X(x)P_{Y|X}(y|x)\frac{\mathbf{M}^*_x\{P_{Y|X}(y|x)\}}{P_{Y|X}(y|x)}dx \end{aligned} \tag{2.13}$$

where $\mathbf{M}^*_x$ is the dual operator of $\mathbf{M}$, with the subscript reminding us that this operates on $P_{Y|X}(y|x)$ as a function of x. We therefore have that

$$E\left\{\frac{\mathbf{M}\{P_X\}(X)}{P_X(X)}\Big|Y=y\right\} = E\left\{\frac{\mathbf{M}^*_x\{P_{Y|X}\}(y|X)}{P_{Y|X}(y|X)}\Big|Y=y\right\} \tag{2.14}$$

Since we are assuming that $P_{Y|X}$ is known, this reduces to the previously discussed situation of find the conditional mean of a function of X given $Y = y$.

If instead of finding the BLS estimator of X given Y we may wish to find the estimator of X using Z where

$$r(Z) = Y \tag{2.15}$$

for an invertible, differentiable, function r. Using known properties of change of variables for densities we have

$$\begin{aligned} E\{X|Z = \mathbf{z}\} &= E\{X|Y = r(\mathbf{z})\} \\ &= \frac{\mathbf{L}\{P_Y\}(r(\mathbf{z}))}{P_Y(r(\mathbf{z}))} \\ &= \frac{J(\mathbf{z})\mathbf{L}\{\frac{P_Z(r^{-1}(\mathbf{y}))}{J(r^{-1}(\mathbf{y}))}\}(\mathbf{y})\Big|_{\mathbf{y}=r(\mathbf{z}))}}{P_Z(\mathbf{z})} \end{aligned} \tag{2.16}$$

where $J(\mathbf{z})$ is the Jacobian of the transformation.

Obviously, our derivation of the operator is valid even if $\mathbf{A}$ is not invertible for all observed densities, as long as there is a unique prior which gives rise to the particular observed density we are dealing with. In such a situation, the prior is sometimes said to be *i*dentifiable in the observed density. See Chapter 2 of [11] for a further discussion of identifiability. In some situations, however, $\mathbf{A}$ may not be one to one and the prior density may not be identifiable, and so it is impossible to uniquely define its inverse. In some of these noninvertible cases, it is still possible to define an operator $\mathbf{L}$ which will give us a prior free estimator, while in other cases we must place restrictions on the set of allowable priors in order to define $\mathbf{L}$ for this restricted family. Examination of the definition of $\mathbf{L}$ in the case where $\mathbf{A}$ is invertible shows that, if we want to define $\mathbf{L}$ when $\mathbf{A}$ is not invertible we must insist that for any two prior densities $P_1(\mathbf{x}), P_2(\mathbf{x})$ such that

$$\mathbf{A}\left\{P_1\right\}(\mathbf{y}) = \mathbf{A}\left\{P_2\right\}(\mathbf{y}) \tag{2.17}$$

we must have

$$(\mathbf{A} \circ \mathbf{X})\left\{P_1\right\}(\mathbf{y}) = (\mathbf{A} \circ \mathbf{X})\left\{P_2\right\}(\mathbf{y}) \tag{2.18}$$

If this is already the case for all priors, then we may still define the operator $\mathbf{L}$ without restricting the set of possible priors. If it is not the case, then we must restrict the prior distribution so that we can define $\mathbf{L}$ for use with this restricted family of priors. One way of doing this is to restrict the set of priors so that $\mathbf{A}$ is one to one. However, this may be unnecessarily restrictive. For example, we may restrict our prior to lie in the subspace

$$\mathcal{P} = \left((\mathcal{N}(\mathbf{A} \circ \mathbf{X})^c) \cap \mathcal{N}(\mathbf{A})\right)^{\perp} \tag{2.19}$$

where $\mathcal{N}$ denotes the nullspace of an operator, ^{c} denotes set complement and $^{\perp}$ denotes the orthogonal complement of a subspace. If we make this restriction, then if Eq. (2.17) holds for two priors in this subspace then we have

$$\mathbf{A}\left\{f\right\} = 0 \tag{2.20}$$

with

$$f = P_1 - P_2 \in \mathcal{P} \tag{2.21}$$

since $\mathcal{P}$ is a subspace. Since f is in $\mathcal{N}(\mathbf{A})$, yet must be orthogonal to $((\mathcal{N}(\mathbf{A} \circ \mathbf{X})^c) \cap \mathcal{N}(\mathbf{A}))$ we get that $f \in \mathcal{N}(\mathbf{A} \circ \mathbf{X})$, as desired.

Even in cases where $\mathbf{A}$ is invertible its inverse may not be well-behaved. For example, in the case of additive Gaussian noise, $\mathbf{A}^{-1}$ is a deconvolution operation which is inherently unstable for high frequencies. The usefulness of Eq. (2.6) comes from the fact that in many cases, the composite operation $\mathbf{L}$ may be written explicitly, even when the inverse operation is poorly defined or unstable. In section 2.4, we develop examples of operators $\mathbf{L}$ for a variety of observation processes.

2.4 Example estimators

We will now discuss some specific cases and how to find appropriate operators. As mentioned earlier, in general, it can be difficult to obtain the operator $\mathbf{L}$ directly from the definition in Eq. (2.5), because inversion of the operator $\mathbf{A}$ could be unstable or undefined. In those cases where the operator $\mathbf{L}$ exists, however, it is not necessary to actually invert $\mathbf{A}$ to find it. Instead, a solution may often be obtained by noting that the definition implies that

$$\mathbf{L} \circ \mathbf{A} = \mathbf{A} \circ \mathbf{X},$$

or, equivalently

$$\mathbf{L}\{P_{Y|X}(\mathbf{y}|\mathbf{x})\} = \mathbf{x} P_{Y|X}(\mathbf{y}|\mathbf{x}).$$

This is an eigenfunction equation: for each value of $\mathbf{x}$, the conditional density $P_{Y|X}(\mathbf{y}|\mathbf{x})$ must be an eigenfunction (eigenvector, for discrete variables) of operator $\mathbf{L}$, with associated eigenvalue $\mathbf{x}$. We may therefore try to find such an operator by inspection of $P_{Y|X}$.

2.4.1 Additive noise

Consider a standard example, in which the variable of interest is corrupted by independent additive noise: $Y = X + W$. The conditional density is

$$P_{Y|X}(\mathbf{y}|\mathbf{x}) = P_W(\mathbf{y} - \mathbf{x}).$$

We wish to find an operator which when applied to this conditional density (viewed as a function of $\mathbf{y}$) will give

$$\mathbf{L}\{P_W(\mathbf{y} - \mathbf{x})\} = \mathbf{x}\, P_W(\mathbf{y} - \mathbf{x}). \tag{2.22}$$

Subtracting $\mathbf{y}\, P_W(\mathbf{y} - \mathbf{x})$ from both sides gives

$$\mathbf{M}\,\{P_W(\mathbf{y} - \mathbf{x})\} = -(\mathbf{y} - \mathbf{x})\, P_W(\mathbf{y} - \mathbf{x}), \tag{2.23}$$

where

$$\mathbf{M}\,\{f\}\,(\mathbf{y}) = \mathbf{L}\{f\}(\mathbf{y}) - \mathbf{y}\, f(\mathbf{y})$$

is a linear shift-invariant operator (acting in $\mathbf{y}$).

Taking Fourier transforms and using the convolution and differentiation properties gives:

$$\begin{aligned}\widehat{\mathbf{M}}(\omega)\widehat{P_W}(\omega) &= -\widehat{(\mathbf{y}P_W)}(\omega) \\ &= -i\nabla_\omega \widehat{P_W}(\omega), \qquad (2.24)\end{aligned}$$

so that

$$\begin{aligned}
\widehat{\mathbf{M}}(\omega) &= -i\frac{\nabla_\omega \widehat{P_W}(\omega)}{\widehat{P_W}(\omega)} \\
&= -i\nabla_\omega \ln\left(\widehat{P_W}(\omega)\right). \qquad (2.25)
\end{aligned}$$

This gives us the linear operator

$$\mathbf{L}\{f\}(\mathbf{y}) = \mathbf{y}\, f(\mathbf{y}) - \mathcal{F}^{-1}\left\{i\nabla_\omega \ln\left(\widehat{P_W}(\omega)\right)\ \widehat{f}(\omega)\right\}(\mathbf{y}), \qquad (2.26)$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. Note that throughout this discussion X and W played symmetric roles. Thus, in cases with known prior density and unknown additive noise density, one can formulate the estimator entirely in terms of the prior.

If the additive noise is such that the corruption process is not invertible, i.e. if the Fourier transform of P_W is bandlimited, the proof of Eq. (2.26) shows that this equation is still valid as long as we define

$$\nabla_\omega \ln\left(\widehat{P_W}(\omega)\right) = 0 \qquad (2.27)$$

whenever $\widehat{P_W}(\omega) = 0$. In this case we will have an observation process which is not one to one. To see this, consider a density P_0 which has finite moments of all order and is bandlimited, which therefore has a Fourier transform which is infinitely differentiable and has compact support. (To see that such a density exists, we can start with any function in the Fourier domain which is infinitely differentiable with compact support. Since the Fourier transform converts multiplication into convolution, convolving this function with itself gives an infinitely differentiable function with compact support, whose inverse transform is positive. We may then normalize this inverse transform to get P_0.) If we then consider

$$P_1(x) = P_0(x)(1 + \cos(w_h x)) \tag{2.28}$$

a simple argument using the Fourier modulation theorem tells us that, for w_h high enough, the Fourier transforms of P_0 and P_1 will be identical over the support of the Fourier transform of P_W. Therefore, both these prior densities will give rise to the same observed density, P_Y. We therefore have a situation where the observation process is not invertible, yet we still have an operator $\mathbf{L}$ which gives us a prior free formulation of our estimator without having to restrict our set of allowable priors. Again, this is because any information lost in the observation process is not required in calculating the BLS estimator. We may also use Eq. (2.9) to calculate all the moments of the prior density, but as we mentioned in discussing this equation, this will not uniquely determine the prior. Any two priors which have the same observed density (e.g. P_0 and P_1 above), will have Fourier transforms which are the same for low frequencies. Since the moments of a density can be calculated by taking the derivatives of the Fourier transform at 0 frequency [3], any two such densities will have equal moments of all orders.

Gaussian case. The most commonly used additive noise model is Gaussian:

$$P_W(\mathbf{x}) = \frac{1}{(2\pi|\Lambda|)^{n/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Lambda^{-1}(\mathbf{x}-\mu)} \tag{2.29}$$

with covariance matrix Λ, and mean vector μ. In this case, the Fourier transform of the density is:

$$\widehat{P_W}(\omega) = e^{-i\omega\cdot\mu-\frac{1}{2}\omega^T\Lambda\omega} \tag{2.30}$$

which, upon substitution into Eq. (2.25) yields:

$$\widehat{M}(\omega) = [i\Lambda\omega - \mu]\,\widehat{P_Y}(\omega) \tag{2.31}$$

Finally, computing the inverse Fourier transform and substituting into Eq. (2.26) yields

$$\begin{aligned} E\{X|Y\} &= \mathbf{y} - \mu + \frac{\Lambda\nabla_{\mathbf{y}}P_Y(\mathbf{y})}{P_Y(\mathbf{y}))} \\ &= \mathbf{y} - \mu + \Lambda\nabla_{\mathbf{y}}ln(P_Y(\mathbf{y})) \end{aligned} \tag{2.32}$$

This formulation for the case of additive Gaussian noise was described by [16]. It implies that, assuming one has sufficient data to compute an approximation of the gradient of the log of the observation density P_Y, one can compute Bayesian least squares estimates without knowing the prior P_X. In Chapter 3 we will describe in detail a particular method for doing this.

Laplacian case. When the noise, W, is drawn from a Laplacian distribution, we have

$$P_W(x) = \frac{1}{2\alpha}e^{-|x/\alpha|} \tag{2.33}$$

The Fourier transform of the noise density will be

$$\widehat{P_W}(\omega) = \frac{1}{1+(\alpha\omega)^2} \tag{2.34}$$

which gives

$$\widehat{M}(\omega) = 2i\alpha^2\omega\widehat{P_W}(\omega) \tag{2.35}$$

This gives us the BLS estimator

$$E\{X|Y\} = y + \frac{2\alpha^2 P'_W(x) \star P_Y(y)}{P_Y(y)} \tag{2.36}$$

where $\star$ denotes convolution and

$$P'_W(x) = -(\frac{1}{\alpha})\text{sgn}(x)\, P_W(x) \tag{2.37}$$

with

$$sgn(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases} \tag{2.38}$$

This formulation of the BLS estimator involves a convolutional operator, as compared to the differential operator found in the Gaussian case. There are a variety of noise densities (for example, the family of generalized Gaussian distributions) for which the operator will be a convolution with some known kernel, K, that depends on the form of the noise. In such instances, the kernel may be used directly to approximate the convolutional operator from observations $\{Y_i\}$:

$$K \star P_Y(\mathbf{y}) \approx \frac{1}{N}\sum_{i=1}^{N} K(\mathbf{y} - Y_i) \tag{2.39}$$

It is interesting to note that kernel estimators such as those in Eq. (2.39), with positive kernels which integrate to one, are commonly used as density estimators.

While such density estimators are generally biased [17], in our situation this approximation is unbiased and converges to the desired convolution $K \star P_Y$ as the amount of data (N) increases, since

$$E\{\frac{1}{N}\sum_{i=1}^{N} K(\mathbf{y} - Y_i)\} = \int K(\mathbf{y} - \tilde{\mathbf{y}}) P_Y(\tilde{\mathbf{y}}) d\tilde{\mathbf{y}} \tag{2.40}$$

The denominator of Eq. (2.36) may be approximated by any of the myriad density estimators (see [17] for a review and further references).

2.4.2 Mixture of Uniform

The next case we will discuss is a mixture of uniform densities

$$P_{Y|X}(y|x) = \begin{cases} \frac{1}{2x}, & |y| \leq x \\ 0, & |y| > x \end{cases} \tag{2.41}$$

where $x \geq 0$. We note that

$$\begin{aligned} \int_{|y|}^{\infty} P_{Y|X}(\tilde{y}|x) d\tilde{y} &= \begin{cases} \frac{1}{2x}(x - |y|), & |y| \leq x \\ 0, & |y| > x \end{cases} \\ &= (x - |y|) P_{Y|X}(y|x) \end{aligned} \tag{2.42}$$

so that the operator we want in this case is

$$L\{f\}(y) = \int_{|y|}^{\infty} f(\tilde{y}) d\tilde{y} + |y| f(y) \tag{2.43}$$

giving

$$\begin{aligned} E\{X|Y=y\} &= |y| + \frac{\int_{|y|}^{\infty} P_Y(\tilde{y})d\tilde{y}}{P_Y(y)} \\ &= |y| + \frac{Pr\{Y > |y|\}}{P_Y(y)} \\ &= |y| + \frac{1 - Pr\{Y \leq |y|\}}{P_Y(y)} \end{aligned} \tag{2.44}$$

2.4.3 Multiplicative Lognormal Noise

The next case we will discuss is that of multiplicative lognormal noise, where

$$Y = Xe^W \tag{2.45}$$

where W is independent Gaussian noise with $N(0, \sigma^2)$ distribution. In this case, taking logarithms gives

$$\ln(Y) = \ln(X) + W \tag{2.46}$$

which is an additive Gaussian noise model. Thus, using the prior free operator for additive Gaussian noise, we have, with $Z = \ln(Y)$

$$E\{\ln(X)|Z=z\} = \frac{(z + \sigma^2 D_z)\{P_Z\}(z)}{P_Z(z)} \tag{2.47}$$

where D_z represents the derivative operator with respect to z. However, we wish to find $E\{X|Y\}$ so we need to use the change of variables formulas we have derived. Firstly, since $X = e^{\ln(X)}$, we have

$$E\{X|Z=z\} = \frac{e^{(z+\sigma^2 D_z)}\{P_Z\}(z)}{P_Z(z)} \tag{2.48}$$

By the Baker-Campbell-Hausdorff formula [18] we have that

$$\begin{aligned} e^{(z+\sigma^2 D_z)}\{f\}(z) &= e^{z+\frac{1}{2}\sigma^2}(e^{\sigma^2 D_z}\{f\}(z)) \\ &= e^{z+\frac{1}{2}\sigma^2} f(z+\sigma^2) \end{aligned} \tag{2.49}$$

so that

$$E\{X|Z=z\} = \frac{e^{z+\frac{1}{2}\sigma^2} P_Z(z+\sigma^2)}{P_Z(z)} \tag{2.50}$$

If we wish to check this formula, we may see directly that, since

$$P_{Z|\ln(X)}(z|\ln(x)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\ln(x))^2}{2\sigma^2}} \tag{2.51}$$

$$\begin{aligned} e^{z+\frac{1}{2}\sigma^2} P_{Z|\ln(X)}(z+\sigma^2|\ln(x)) &= e^{z+\frac{1}{2}\sigma^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\ln(x)+\sigma^2)^2}{2\sigma^2}} \\ &= e^{\ln(x)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\ln(x))^2}{2\sigma^2}} \\ &= x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\ln(x))^2}{2\sigma^2}} \end{aligned} \tag{2.52}$$

Next, using the fact that $\ln(Y) = Z$, we have by the change of variables formula

$$P_Y(y) = \frac{P_Z(\ln(y))}{y} \tag{2.53}$$

so that

$$E\{X|Y=y\} = \frac{e^{\frac{3}{2}\sigma^2} P_Y(e^{\sigma^2} y)}{P_Y(y)} y \tag{2.54}$$

This formula may also be shown by directly using the change of variables formula to find $P_{Y|X}$ and applying the linear operator

$$\mathbf{L}\{f\}(y) = e^{\frac{3}{2}\sigma^2}\, y\, f(e^{\sigma^2} y) \tag{2.55}$$

to show that

$$\mathbf{L}\{P_{Y|X}\}(y|x) = xP_{Y|X}(y|x)$$

2.4.4 Power of Fixed Density

An interesting family of observation processes are those for which

$$\widehat{P}_{Y|X}(\omega) = [\widehat{P_W}(\omega)]^X \tag{2.56}$$

for some density P_W. This occurs, for example, when X takes on integer values, and Y is a sum of X i.i.d. random variables with distribution P_W:

$$Y = \sum_{n=1}^{X} W_n$$

Three other special cases of this are of particular interest. The first occurs when X is a positive variable and Y is a Poisson random variable with rate X. In this case Eq. (2.56) will hold for

$$\widehat{P_W}(\omega) = e^{(e^{-i\omega}-1)} \tag{2.57}$$

The second example arises when X is a positive random variable and Y is a zero mean Gaussian with variance X, a case known as the Gaussian Scale Mixture

(GSM) [19]. In this case Eq. (2.56) will hold for

$$\widehat{P_W}(\omega) = e^{-\frac{1}{2}\omega^2} \tag{2.58}$$

The third special case is when X is a random positive value, W is an independent variable drawn from an α-stable distribution with Fourier transform

$$\widehat{P_W}(\omega) = e^{-|\omega|^\alpha} \tag{2.59}$$

and

$$Y = X^{\frac{1}{\alpha}} W \tag{2.60}$$

Generally, if P_W is an infinitely divisible distribution and X is an arbitrary positive real number, then the right side of Eq.(2.56) will be the Fourier transform of some density, which can be used as the observation process.

All of these cases can be written in prior-free form. Taking the derivative of Eq. (2.56) gives

$$\begin{aligned}
\widehat{P_{Y|X}}'(\omega) &= \widehat{P_W}'(\omega)\, x\, \widehat{P_W}(\omega)^{x-1} \\
&= \frac{\widehat{P_W}'(\omega)}{\widehat{P_W}(\omega)}\, x\, \widehat{P_W}(\omega)^{x} \\
&= \frac{d}{d\omega} ln(\widehat{P_W}(\omega))\, x\, \widehat{P_{Y|X}}(\omega) \qquad (2.61)
\end{aligned}$$

Rearranging this equality gives

$$\begin{aligned} x\,\widehat{P_{Y|X}}(\omega) &= \frac{1}{\frac{d}{d\omega}ln(\widehat{P_W}(\omega))}\widehat{P_{Y|X}}'(\omega) \\ &= \frac{1}{i\frac{d}{d\omega}ln(\widehat{P_W}(\omega))}\widehat{yP_{Y|X}}(\omega) \end{aligned} \tag{2.62}$$

Thus, the linear operation first multiplies P_Y by y and then applies the linear shift-invariant transform:

$$\widehat{M}(\omega) = \frac{1}{i\frac{d}{d\omega}ln(\widehat{P_W}(\omega))} \tag{2.63}$$

In the cause of Poisson with random rates this will be

$$\widehat{M}(\omega) = e^{i\omega} \tag{2.64}$$

Substituting this into Eq. (2.62), taking the inverse Fourier transform and substituting into Eq. (2.1), gives

$$E\{X|Y = n\} = \frac{(n+1)P_Y(n+1)}{P_Y(n)} \tag{2.65}$$

a fact which can be verified by direct calculation.

In the case of the GSM the operator will be

$$\widehat{M}(\omega) = \frac{-1}{i\omega} \tag{2.66}$$

which gives

$$E\{X|Y = y\} = \frac{-(H(y) - \frac{1}{2}) \star (yP_Y(y))}{P_Y(y)} \tag{2.67}$$

where H is the Heavyside function. Since $yP_Y(y)$ is odd this is equal to

$$\begin{aligned} E\{X|Y=y\} &= \frac{-(H(y))\star(yP_Y(y))}{P_Y(y)} \\ &= \frac{-\int_{-\infty}^{y}\tilde{y}P_Y(\tilde{y})d\tilde{y}}{P_Y(y)} \\ &= \frac{-E_Y\{Y;Y<y\}}{P_Y(y)} \end{aligned} \tag{2.68}$$

where the numerator is now the mean of the density to the left of y and may be approximated in an unbiased way by the sum of data less than y divided by the total number of data points.

2.4.5 Exponential Families

Another important case in which the linear operator may be solved for explicitly is for exponential families of distributions.

Discrete Exponential The first case we discuss is discrete exponential families of the form

$$Pr\{Y=n|X=x\} = h(x)g(n)x^n \tag{2.69}$$

where h is chosen so that summing over n gives one. This case includes the Poisson case discussed in the previous section among others (see [11]). In this case it may be shown that

$$E\{X|Y=n\} = \frac{g(n)P_Y(n+1)}{g(n+1)P_Y(n)} \tag{2.70}$$

Also we note from Eq. (2.7) that

$$E\{\frac{1}{X}|Y=n\} = \frac{g(n)P_Y(n-1)}{g(n-1)P_Y(n)} \tag{2.71}$$

Continuous Exponential The next case we discuss is continuous families of the form

$$P_{Y|X}(y|x) = h(x)g(y)e^{T(y)x} \tag{2.72}$$

where we assume that T is differentiable, a case which includes the GSM discussed in the previous section. In this case

$$\begin{aligned} E\{X|Y=y\} &= \frac{g(y)\frac{d}{dy}\{\frac{P_Y(y)}{g(y)}\}}{T'(y)P_Y(y)} \\ &= \frac{1}{T'(y)}\frac{d}{dy}ln(\frac{P_Y(y)}{g(y)}) \end{aligned} \tag{2.73}$$

Also

$$E\{\frac{1}{X}|Y=y\} = \frac{g(y)\int_{-\infty}^{y}\frac{T'(\tilde{y})}{g(\tilde{y})}P_Y(\tilde{y})d\tilde{y}}{P_Y(y)} \tag{2.74}$$

Our prior-free estimator methodology is quite general, and can often be applied to more complicated observation processes. In order to give some sense of the diversity of forms that can arise, Table 2.1 provides additional examples. In this table, functions written with hats or in terms of ω represent multiplication in the Fourier Domain, and n replaces y for discrete distributions. References for the specific cases that we have found in the statistics literature are provided in table.

2.4.6 Noninvertible Observation Processes

In all the cases we have discussed so far, we have been able to define an operator **L** regardless of whether the observation process was invertible. In some cases, however, the noninvertiblity of an observation process will prevent us from writing the optimal estimator entirely in terms of P_Y, unless we put some restriction on the prior P_X. In this section we will discuss such an example, and illustrate the kind of restrictions that will be put on the prior.

Suppose we randomly choose a coin with probability of heads $X, 0 \leq X \leq 1$, the density of X being P_X. We then perform a binomial trial of flipping the chosen coin n times and observing the number of heads, so that

$$Pr\{Y = k|X = x\} = \binom{n}{k} x^k (1-x)^{n-k} \tag{2.75}$$

which gives the observed probability

$$Pr(Y = k) = \binom{n}{k} \int_0^1 P_X(x) x^k (1-x)^{n-k} dx \tag{2.76}$$

Here, we have gone from the infinite set of prior densities, P_X, on the interval $[0, 1]$ to the finite set of observed probabilities $\{Pr(Y = k)\}_{k=0}^n$ on the set $\{0, ..., n\}$, so a simple dimensionality argument tells us that this process is not invertible. We will now show that this prevents us from writing the BLS estimator entirely in terms of P_Y unless further restriction is placed on P_X.

For each value of k, $x^k(1-x)^{n-k}$ is a polynomial of degree n, and $Pr(Y = k)$ gives us the dot product of P_X with these polynomials. In order to obtain the BLS estimator of X, however, we need to know the numerator in Eq. (2.4), which in our case is

$$N(k) = \binom{n}{k} \int P_X(x) x^{k+1} (1-x)^{n-k} dx \tag{2.77}$$

which is the dot product of P_X with a polynomial of degree $n + 1$, for $k = 0, ..., n$. In order to be able to write the estimator entirely in terms of $Pr(Y = k)$, then, we need to be able to write the dot product of P_X with a polynomial of degree $n+1$, in terms of the dot products of P_X with polynomials of degree n. This is obviously impossible to do for general P_X. However, since the set of polynomials $\{x^k(1-x)^{n-k}\}_{k=0}^n$ forms a linearly independent set (look at the lowest order terms), knowing $Pr(Y = k)$ for all k allows us to find the dot product of P_X with any polynomial of degree less than or equal to n. Therefore, restricting P_X to be orthogonal to the $(n+1)^{st}$ degree polynomial in a Gram-Schmidt orthogonal basis of polynomials on the interval, or equivalently, requiring the expected value of this polynomial applied to X to be zero, allows us to write the numerator in terms of P_Y alone. This does not make the observation process invertible, but, rather, requires that two priors in the restricted set of priors with the same observed density also have the same BLS estimator. This is less restrictive than requiring the observation process to be invertible, which could be accomplished by requiring P_X to lie in the space of polynomials of degree less than or equal to n, or equivalently, requiring moments of order higher than n to be zero.

This behavior is tied to the parametrization used. If, for example, we choose $X \in [0, \infty)$ with density P_X and then perform the Bernoulli experiment with probability of heads $\frac{X}{X+1}$ then

$$\begin{aligned} Pr(Y = k) &= \binom{n}{k} \int P_X(x) \frac{x}{x+1}^k \left(\frac{1}{x+1}\right)^{n-k} dx \\ &= \binom{n}{k} \int P_X(x) x^k \left(\frac{1}{x+1}\right)^n dx \end{aligned} \tag{2.78}$$

In order to obtain the BLS estimator of X we need to know

$$N(k) = \binom{n}{k} \int P_X(x) x^{k+1} \left(\frac{1}{x+1}\right)^n dx \tag{2.79}$$

Now it is easy to see that for $k < n$

$$N(k) = \frac{\binom{n}{k}}{\binom{n}{k+1}} Pr(Y = k + 1) \tag{2.80}$$

Knowing $Pr(Y = k)$ gives the dot product of P_X, using weighting function $(\frac{1}{x+1})^n$, with all polynomials up to degree n. However, in order to know $N(n)$, we need to know the dot product of P_X with x^{n+1}. Therefore, in general we cannot solve for $N(n)$. Again, we can get around this by limiting the prior distribution in an appropriate way.

If instead of performing a Bernoulli experiment with the randomly weighted coin, we performed a geometric experiment with probability of heads X, observing the number of tosses before we get a tail, then

$$Pr(Y = k) = \int P_X(x)x^k(1 - x)dx, \;\; k = 0, 1, 2, ... \tag{2.81}$$

a special case of the discrete exponential family in Eq. (2.69). Here the number of possible observations is countably infinite, and gives the dot product of P_X with all polynomials (we already know the integral of P_X is one). In this case we want

$$N(k) = \int P_X(x)x^{k+1}(1 - x)dx \tag{2.82}$$

which is easily seen to be Pr(Y=k+1). Therefore our estimator will be

$$E\{X|Y = k\}\frac{Pr(Y = k + 1)}{Pr(Y = k)} \tag{2.83}$$

which agrees with our formula in Table 2.1.

2.5 Prior-free reformulation of the mean squared error

In some cases, developing a stable nonparametric approximation of the ratio in Eq. (2.6) may be difficult. However, the linear operator formulation of the BLS estimator also leads to a dual expression for the mean squared error that does not depend explicitly on the prior, and this may be used to select an optimal estimator from a parametric family of estimators. Specifically, for any estimator $f_\theta(Y)$ parameterized by θ, the mean squared error may be expanded as

$$\begin{aligned} E\left\{|f_\theta(Y) - X|^2\right\} &= E\left\{|f_\theta(Y)|^2 - 2X \cdot f_\theta(Y)\right\} + E\left\{|X|^2\right\} \\ &= E\left\{|f_\theta(Y)|^2 - 2E\{X|Y\} \cdot f_\theta(Y)\right\} + E\left\{|X|^2\right\} \end{aligned} \tag{2.84}$$

Since $E\{|X|^2\}$ does not depend on f_θ, it is irrelevant for optimizing θ. Using the prior-free formulation of the previous sections, the second component of the expectation may be written as

$$\begin{aligned} E\left\{f_\theta(Y)E(X|Y)\right\} &= E\left\{f_\theta(Y)\frac{\mathbf{L}\{P_Y\}(Y)}{P_Y(Y)}\right\} \\ &= \int f_\theta(\mathbf{y})\frac{\mathbf{L}\{P_Y\}(\mathbf{y})}{P_Y(\mathbf{y})}P_Y(\mathbf{y})d\mathbf{y} \\ &= \int f_\theta(\mathbf{y})\ \mathbf{L}\{P_Y\}(\mathbf{y})d\mathbf{y} \\ &= \int \mathbf{L}^*\{f_\theta\}(\mathbf{y})P_Y(\mathbf{y})d\mathbf{y} \\ &= E\left\{\mathbf{L}^*\{f_\theta\}(Y)\right\}, \end{aligned}$$

where $\mathbf{L}^*$ is the dual operator of $\mathbf{L}$ (in the discrete case, $\mathbf{L}^*$ is the matrix transpose of $\mathbf{L}$). Combining all of the above, we have:

$$E\left\{|f_\theta(Y) - X|^2\right\} = E\left\{|f_\theta(Y)|^2 - 2\mathbf{L}^*\{f_\theta\}(Y)\right\} + \text{const.} \tag{2.85}$$

where the constant, $E\{|X|^2\}$, does not depend on θ.

Some special examples of this type of formulation have appeared in the literature in the context of improving estimators[12, 13, 14] (see also Table 2.1). Our approach serves to unify and generalize them, and to tie them to the prior free formulation of the BLS estimator. It should be noted that these papers work in a framework where X is not random, but rather fixed and unknown. However, such formulas may be easily derived from our context by fixing the prior to be degenerate at this fixed and unknown value, in which case all expectations become expectations conditioned on X. Conversely, if such a formulation is written in the case of a fixed and unknown X, all expectations will be written in terms of conditional densities. It is then easy to convert formulas written in terms of conditional densities into our framework by taking expectations with respect to X.

The way these formulations of the MSE are most often used is in the context of improving estimators. For example, if X is a fixed but unknown vector of dimension $d > 2$ and Y is $N(X, \mathbf{I}_d)$, we may show as in [12] that

$$Y - \frac{d-2}{|Y|^2}Y \tag{2.86}$$

has lower MSE than the MSE we would get by using Y to estimate X.

In this case

$$\mathbf{L}^* = y - \nabla\cdot \tag{2.87}$$

It turns out that it is easier to represent the estimator as

$$f(\mathbf{y}) = \mathbf{y} + g(\mathbf{y}). \tag{2.88}$$

Substituting into Eq. (2.85) gives

$$E\{|f(Y) - X|^2\} = E\{|g(Y)|^2 + 2\nabla \cdot g(Y)\} + \text{const.}, \tag{2.89}$$

where the constant does not depend on g. Examining the case where $g = \mathbf{0}$ (or equivalently $f(\mathbf{y}) = \mathbf{y}$) shows that the constant must be equal to d so that

$$MSE = E\{|g(Y)|^2 + 2\nabla \cdot g(Y)\} + d \tag{2.90}$$

Letting

$$g(\mathbf{y}) = -\frac{d-2}{|\mathbf{y}|^2}\mathbf{y} \tag{2.91}$$

gives

$$\nabla \cdot g(\mathbf{y}) = -\frac{(d-2)^2}{|\mathbf{y}|^2} \tag{2.92}$$

so that

$$|g(\mathbf{y})|^2 + 2\nabla \cdot g(Y) = -\frac{(d-2)^2}{|\mathbf{y}|^2} < 0 \tag{2.93}$$

which tells us that the MSE using this value of g gives us a MSE smaller than d, which is the value of the MSE in using Y to estimate X. This fact, known as Stein's Paradox [20, 21], says that for all fixed X, shrinking the observed vector, Y, towards zero gives a better estimate of X than using Y. In our context we may extend this statement to random X for all possible prior densities on X.

We may gain some insight to this paradox by comparing this estimator to the

Weiner estimator. Recall that this estimator may be defined as the linear estimator which minimizes the MSE. In our case, this implies that the associated function g in Eq. (2.88) may be written as

$$g(\mathbf{y}) = G\mathbf{y} \tag{2.94}$$

for some matrix G. Some tedious but straightforward calculation then gives that Eq. (2.89) is minimized by choosing

$$G = -R_{YY}^{-1} \tag{2.95}$$

where

$$R_{YY} = E\{Y^T Y\} \tag{2.96}$$

is the correlation matrix of Y. The associated estimator is then

$$(I - R_{YY}^{-1})\mathbf{y} = (R_{YY} - I)R_{YY}^{-1}\mathbf{y} \tag{2.97}$$

This is the same result derived at the end of Chapter 1, except that, since we are minimizing MSE, this result uses the autocorrelation matrix, R_{YY} instead of the covariance matrix, C_{YY}. By diagonalizing R_{YY} in Eq. (2.97), it is easy to see that the optimal linear estimator is a shrinkage estimator as well. For example, restricting ourselves to one dimension, this estimator may be written as ay with

$$a = \frac{E\{Y^2\} - 1}{E\{Y^2\}} \tag{2.98}$$

so that $a < 1$. Thus, we see that for linear operators, shrinkage ($a < 1$) gives improved performance over the identity ($a = 1$), regardless of the true value of the unknown mean X. For the linear estimator, this may be explained heuristically by noticing that this estimator moves an observed y by a distance

$$y - ay = (1 - a)y \tag{2.99}$$

which is proportional to y, This in turn implies that larger values move by greater distances than smaller values, so that while all values shrink towards the origin, they also can have an overall contraction towards another value as well. The particular choice of a made by the Weiner estimator allows an overall contraction towards the unknown mean X.

For the general corruption process, we may try to show that one estimator, f_{θ_1}, is better than another, f_{θ_2}, for all possible priors (or more particularly for all fixed values of X) by showing that

$$|f_{\theta_1}(\mathbf{y})|^2 - 2\mathbf{L}^*\{f_{\theta_1}\}(\mathbf{y}) < |f_{\theta_2}(\mathbf{y})|^2 - 2\mathbf{L}^*\{f_{\theta_2}\}(\mathbf{y}) \tag{2.100}$$

for all values of $\mathbf{y}$. Combining this with Eq. (2.85) shows that the MSE for f_{θ_1} is smaller than that for f_{θ_2}.

Instead of just trying to use these prior free formulations of the MSE to show that one estimator is better than another for all priors, we may use these formulas in a data adaptive way to try and choose the best estimator out of a family of estimators, as we did for the family of linear estimators. In this approach, we note that

$$\arg\min_{\theta} E\left\{|f_\theta(Y) - X|^2\right\} = \arg\min_{\theta} E\left\{|f_\theta(Y)|^2 - 2\mathbf{L}^*\{f_\theta\}(Y)\right\}. \tag{2.101}$$

where the expectation on the right is over the observation variable, Y. In practice, we can solve for an optimal θ by minimizing the sample mean of this quantity:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \left\{ |f_\theta(Y_n)|^2 - 2\mathbf{L}^*\{f_\theta\}(Y_n) \right\}. \tag{2.102}$$

where $\{Y_n\}$ is a set of observed data. We can than apply the estimator $f_{\hat{\theta}}$ to the same data that we used to pick $\hat{\theta}$. Again this does not require any knowledge of, or samples drawn from, the prior P_X. This is the unsupervised counterpart of Eq. (2.3). In the supervised situation, we train on observed data for which we are given the corresponding hidden values, and then apply the trained estimator to new samples for which we do not have the corresponding hidden values. In our situation, however, we are never given the underlying values, and so it does not make sense to set aside any particular data as training data. Instead we train on the same data to which we want to apply the estimator. In some situations, we may wish to use cross validation (CV) or some other resampling method, but there is no reason to set aside any particular data values in this process, all data has equal status. This implies that as we apply our estimator to more observations, the estimator should be able to improve by using this new data to further train the estimator. This cannot be said of the supervised situation, in which case all training takes place on the training data. Another implication is that if the prior distribution of the underlying values changes over time, our estimator should be able to adapt to these changes using the new observations. Again, this cannot be said of the supervised situation, where, if the distribution of the underlying values changes from that during the training phase, the estimator will not be able to adapt without retraining on new supervised data drawn from the new distribution. We also note that it is quite easy to extend our methodology to a semi-supervised setting, in which we are given some pairs $\{(X_i, Y_i)\}_{i\in\mathcal{S}}$ of clean samples together with their corrupted observations, and some samples $\{Y_i\}_{i\in\mathcal{U}}$ for which we only have corrupted observations. Since we know that both

$$|f_\theta(Y)|^2 - 2\mathbf{L}\{f_\theta\}(Y) \tag{2.103}$$

and

$$|f_\theta(Y) - X| = |f_\theta(Y)|^2 - 2X \cdot f_\theta(Y) + |X|^2 \tag{2.104}$$

are both unbiased estimator of MSE, we can try and minimize

$$\sum_{i\in\mathcal{S}\cup\mathcal{U}} |f_\theta(Y_i)|^2 - 2\sum_{i\in\mathcal{U}} \mathbf{L}\{f_\theta\}(Y_i) - 2\sum_{i\in\mathcal{S}} X_i \cdot f_\theta(Y_i) \tag{2.105}$$

In the unsupervised one dimensional additive Gaussian case, letting

$$g_\theta(y) = f_\theta(y) - y$$

we get

$$\hat{\theta} = \arg\min_\theta \frac{1}{N}\sum_{n=1}^{N} \left\{ |g_\theta(Y_n)|^2 + 2\sigma^2 g'_\theta(Y_n) \right\}. \tag{2.106}$$

SUREShrink [22] uses this equation to find the optimal threshold for denoising data corrupted by additive Gaussian noise. In [23, 24] something similar is done for additive Gaussian noise and a linear family of estimators. As far as we are aware, [24] is the only work which suggests this approach for non-Gaussian noise.

Recently [25], an expression similar to Eq. (2.106), with

$$g_\theta = \frac{d}{dy} \ln P_\theta(y) \tag{2.107}$$

was used as a criterion for choosing a density estimate from a family P_θ in cases where the normalizing constant, or partition function, is difficult to obtain. The prior-free approach we are discussing provides a justification and interpretation for this procedure: the optimal density is the one which, when converted into an estimator using the formula in Table 2.1 for the additive Gaussian case, gives the best MSE. We might have tried, as in the commonly used Empirical Bayes procedure [10, 26], to pick this density using a more widely used criterion such as Maximum Likelihood (ML), and then convert the density to an estimator. The problem is that the quantity we are eventually trying to optimize is the MSE, and the ML does not minimize this criteria (it actually minimizes the Kullback-Leibler divergence between the true density and the parametric density [20]). Instead, we are now able to directly choose the estimator which optimizes the criteria we really wish to optimize, namely the MSE. This density estimation procedure may be extended to any of the linear operators in Table 2.1. We can pick the parametric density P_θ which minimizes Eq. (2.85) with

$$f_\theta(y) = \frac{\mathbf{L}\{P_\theta\}(y)}{P_\theta} \tag{2.108}$$

being the estimator associated with the density $P_\theta(y)$, and where $\mathbf{L}$ is the appropriate operator form Table 2.1.

When using this formulation in the context of density estimation, it is more natural to parametrize the observed density which may then be converted into the corresponding estimator using the appropriate linear operator. In the estimation framework, however, it is more natural to parametrize the estimator. This leads to the general question of whether, given the BLS estimator

$$\hat{X}(y) = \frac{\mathbf{L}\{P_Y\}(y)}{P_Y(y)}$$

it is possible to recover the density $P_Y(y)$. While it is difficult to formulate a general solution to this problem, in the continuous and discrete exponential cases, we may come up with a general expression of the density in terms of the estimator.

For example, in the discrete exponential case (see Table 2.1), recall that we have

$$\hat{X}(n) = \frac{g(n)P_Y(n+1)}{g(n+1)P_Y(n)} \tag{2.109}$$

so that

$$\ln P_Y(n+1) - \ln P_Y(n) = \ln \hat{X}(n) + \ln g(n+1) - \ln g(n) \tag{2.110}$$

Equivalently, we write this as

$$\Delta \ln P_Y = \ln \hat{X}(n) + \Delta \ln g \tag{2.111}$$

where Δ represents the difference operator. The solution to this equation is

$$\ln P_Y(n) = \sum_{k=0}^{n-1} \ln \hat{X}(k) + \ln g(n) + const. \tag{2.112}$$

where the constant is such that $P_Y(n)$ sums to one. It is interesting to note that in this situation

$$\ln P_{Y|X}(n|x) = \ln h(x) + \ln g(n) + n \ln x \tag{2.113}$$

so that

$$\Delta_n \ln P_{Y|X} = \Delta \ln g + \ln x \tag{2.114}$$

where the subscript on the Δ_n is to remind us that we are taking the difference with respect to n. We therefore have that

$$\ln P_Y(n) = \sum_{k=0}^{n-1} \left\{ \Delta_n \ln P_{Y|X} \right\} (k|\widehat{X}(k)) + const. \tag{2.115}$$

which is the desired expression for P_Y in terms of $\hat{X}(y)$.

Similarly, for the continuous exponential case (see Table 2.1), we have

$$\begin{aligned} \widehat{X}(y) &= \frac{g(y)}{T'(y)P_Y(y)} \frac{d}{dy} \left(\frac{P_y(y)}{g(y)} \right) \\ &= \frac{1}{T'(y)} \frac{d}{dy} \ln \left(\frac{P_y(y)}{g(y)} \right) \end{aligned} \tag{2.116}$$

so that

$$\frac{d}{dy} \ln P_Y(y) = T'(y)\widehat{X}(y) + \frac{d}{dy} \ln g(y) \tag{2.117}$$

The solution to this equation is

$$\ln P_Y(y) = \int T'(\tilde{y})\, \widehat{X}(\tilde{y})\, d\tilde{y} + \ln g(y) + const. \tag{2.118}$$

where, again, the constant is such that $P_Y(y)$ integrates to one. In this situation

$$\ln P_{Y|X}(y|x) = \ln h(x) + \ln g(y) + T(y)x \tag{2.119}$$

so that

$$\frac{d}{dy} \ln P_{Y|X}(y|x) = \frac{d}{dy} \ln g(y) + T'(y)x \tag{2.120}$$

We therefore have that

$$\ln P_Y(y) = \int \left\{ \frac{d}{dy} \ln P_{Y|X} \right\} \left(\tilde{y} \,\middle|\, \widehat{X}(\tilde{y}) \right) d\tilde{y} + const. \tag{2.121}$$

We have shown how to go from a prior free expression for the BLS estimator as might appear in [11, 15, 16, 24] to a prior free expression for the MSE, as might appear in [12, 13, 14, 24]. For completeness, we show that it is possible to go from the prior free formulation of MSE to the prior free formulation of the BLS estimator. To see this suppose that for a particular observation process, and for every estimator f, we have

$$E\left\{ |f(Y) - X|^2 \right\} = E\left\{ |f(Y)|^2 - 2h_f(Y) \right\} + c. \tag{2.122}$$

for a constant c which does not depend on f, and a function h_f which depends on f, but does not depend on the statistics of X. Then, expanding the left side of this equation, we see that we must have

$$E\left\{ X \cdot f(Y) \right\} = E\left\{ h_f(Y) \right\}. \tag{2.123}$$

It is then obvious that if this holds for all f, then h_f must be linear in f,

$$h_f(\mathbf{y}) = M\{f\}(\mathbf{y}) \tag{2.124}$$

so that

$$\begin{aligned}
E\{E\{X|Y\} \cdot f(Y)\} &= E\{X \cdot f(Y)\} \\
&= E\{M\{f\}(Y)\} \\
&= \int M\{f\}(\mathbf{y})P_Y(\mathbf{y})d\mathbf{y} \\
&= \int f(\mathbf{y})M^*\{P_Y\}(\mathbf{y})d\mathbf{y} \\
&= \int f(\mathbf{y})\frac{M^*\{P_Y\}(\mathbf{y})}{P_Y(\mathbf{y})}P_Y(\mathbf{y})d\mathbf{y} \\
&= E\left\{f(Y)\frac{M^*\{P_Y\}(Y)}{P_Y(Y)}\right\} \qquad (2.125)
\end{aligned}$$

Since this is true for arbitrary f we have that

$$E\{X|Y = \mathbf{y}\} = \frac{M^*\{P_Y\}(\mathbf{y})}{P_Y(\mathbf{y})} \tag{2.126}$$

as desired.

We will illustrate another method of proof which provides insight. Since the MSE may be written as

$$E\left\{|f(Y) - X|^2\right\} = E\left\{|f(Y)|^2 - 2M\{f\}(Y)\right\} + c. \tag{2.127}$$

this functional is minimized at

$$f_0(\mathbf{y}) = E\left\{X \middle| Y = \mathbf{y}\right\}$$

This means that for an arbitrary function f_1, the functional

$$\begin{aligned} J(\epsilon) &= E\left\{|f_0(Y) + \epsilon f_1(Y) - X|^2\right\} \\ &= E\left\{|f_0(Y) + \epsilon f_1(Y)|^2 - 2M\{f_0 + \epsilon f_1\}(Y)\right\} + c. \end{aligned} \tag{2.128}$$

is minimized at $\epsilon = 0$. Setting the derivative of this functional with respect to ϵ equal to zero at $\epsilon = 0$ gives

$$E\left\{f_0(Y) \cdot f_1(Y)\right\} = E\left\{M\{f_1\}(Y)\right\} \tag{2.129}$$

As before we show that

$$E\left\{M\{f_1\}(Y)\right\} = E\left\{f_1(Y)\frac{M^*\{P_Y\}(Y)}{P_Y(Y)}\right\} \tag{2.130}$$

so that we have

$$E\left\{f_0(Y) \cdot f_1(Y)\right\} = E\left\{f_1(Y)\frac{M^*\{P_Y\}(Y)}{P_Y(Y)}\right\} \tag{2.131}$$

for arbitrary f_1. This gives us

$$f_0(\mathbf{y}) = \frac{M^*\{P_Y\}(\mathbf{y})}{P_Y(\mathbf{y})} \tag{2.132}$$

as desired.

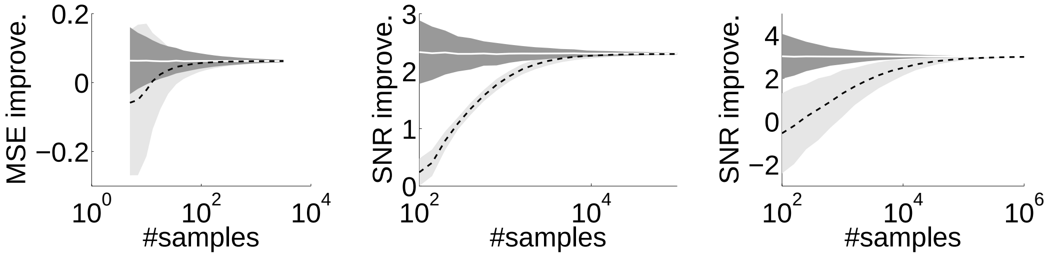
2.6 Simulations

2.6.1 Prior Free BLS Estimators

Since each of the prior-free BLS estimators discussed above relies on approximating values from the observed data, the behavior of such estimators should approach the BLS estimator as the number of data samples grows. In Fig. 2.1, we examine the behavior of three non-parametric prior-free estimators based on Eq. (2.6). The first case corresponds to data drawn independently from a binary source, which are observed through a process in which bits are switched with probability $\frac{1}{4}$. The estimator does not know the binary distribution of the source (which was a "fair coin" for our simulation), but does know the bit-switching probability. For this estimator we use the observations to approximate P_Y using a simple histogram, and then use the matrix version of the linear operator in Eq. (2.5) to construct the estimator. We then apply the constructed estimator to the same observed data to estimate the uncorrupted value associated with each observation. We measure the behavior of the estimator, $\hat{X}$, using the the empirical MSE,

$$\frac{1}{N}\sum_{k=1}^{N}(\hat{X}_i - X_i)^2 \tag{2.133}$$

where $\{X_i\}$ are the underlying values and $\{\hat{X}_i\}$ are the corresponding estimates based on the observations. We characterize the behavior of this estimator as a function of the number of data points, N, by running many Monte Carlo simulations for each N, constructing the estimator using the N observations, applying the constructed estimator to these observations and recording the empirical MSE. Figure 2.1 indicates the mean improvement in empirical MSE (measured by the



(a) (b) (c)

Fig. 2.1: Empirical convergence of prior-free estimator to optimal BLS solution, as a function number of observed samples of Y. For each number of observations, each estimator is simulated many times. Black dashed lines show the improvement of the prior-free estimator, averaged over simulations, relative to the ML estimator. White line shows the mean improvement using the conventional BLS solution, $E\{X|Y = \mathbf{y}\}$, assuming the prior density is known. Gray regions denote ± one standard deviation. (**a**) Binary noise (10,000 simulations for each number of observations); (**b**) additive Gaussian noise (1,000 simulations); (**c**) Poisson noise (1,000 simulations).

increase in empirical MSE compared with using the ML estimator, which, in this case, is the identity function) over the Monte Carlo simulations, the mean improvement using the conventional BLS estimation function, $E\{X|Y = \mathbf{y}\}$ assuming the prior density is known, and the standard deviations of the improvements taken over our simulations. Note that the variance in the BLS estimator for small numbers of data points comes from the fact that we are using empirical MSE.

Figure 2.1**b** shows similar results for additive Gaussian noise, with the empirical MSE being replaced by the empirical Signal to Noise Ratio (SNR), which is defined as

$$SNR(dB) = 20\log_{10}\left(\frac{\sum_{k=1}^{N} X_k^2}{\sum_{k=1}^{N}(\hat{X}_k - X_k)^2}\right) \qquad (2.134)$$

Signal density is a generalized Gaussian with exponent 0.5, and the noisy SNR is 4.8 dB. In this case, we compute Eq. (2.6) using a more sophisticated approximation method, as described in [27], which we will also describe in Chapter 3. We fit a local exponential model similar to that used in [28] to the data in bins, with binwidth adaptively selected so that the product of the number of points in the bin and the squared binwidth is constant. This binwidth selection procedure, analogous to adaptive binning procedures developed in the density estimation literature [17], provides a reasonable tradeoff between bias and variance, and converges to the correct answer for any well-behaved density [27]. Note that in this case, convergence is substantially slower than for the binary case, as might be expected given that we are dealing with a continuous density rather than a single scalar probability. But the variance of the estimates is very low.

Figure 2.1**c** shows the case of estimating a randomly varying rate parameter that governs an inhomogeneous Poisson process. The prior on the rate (unknown to the estimator) is exponential. The observed values Y are the (integer) values drawn from the Poisson process. In this case the histogram of observed data was used to obtain a naive approximation of $P_Y(n)$, the appropriate operator from Table 2.1 was used to convert this into an estimator, and this estimator was then applied to the observed data. It should be noted that improved performance for this estimator is expected if we were to use a more sophisticated approximation of the ratio of densities.

2.6.2 Parametric examples

In this section we discuss the empirical behavior of the parametric approach applied to the additive Gaussian case. Recall from Eq. (2.87) that in this case we have

$$\mathbf{L}^* = y - \sigma^2 \frac{d}{dy}.$$

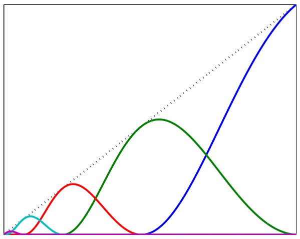


Fig. 2.2: Example bump functions, used for linear parameterization of estimators in Figs. 2.3(a) and 2.3(b).

We also have, from Eq. (2.89)

$$E\{|f(Y)-X|^2\} = E\{|g(Y)|^2 + \sigma^2 g'(Y)\} + \text{const},$$

with

$$f(y) = y + g(y).$$

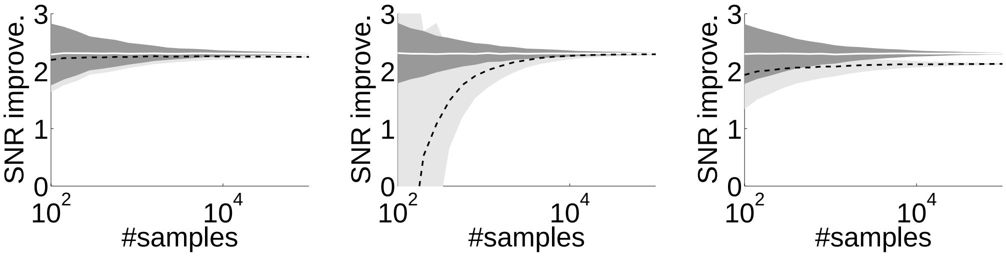
where the constant does not depend on g. Therefore, if we have a parametric family $\{g_\theta\}$ of such g, and are given data $\{Y_n\}$ we can try and pick θ to minimize

$$\frac{1}{N}\sum_{n=1}^{N}\{|g_\theta(Y_n)|^2 + \sigma^2 g'_\theta(Y_n)\}. \tag{2.135}$$

This expression, known as Stein's unbiased risk estimator (SURE) [12], favors estimators g_θ that have small amplitude, and highly negative derivatives at the data values. This is intuitively sensible: the resulting estimators will "shrink" the data toward regions of high probability.

As an example, we parametrize g as a linear combination of nonlinear "bump" functions

$$g_\theta(y) = \sum_k \theta_k g_k(y) \tag{2.136}$$



(a) (b) (c)

Fig. 2.3: Empirical convergence of parametric prior-free method to optimal BLS solution, as a function number of data observations, for three different parameterized estimators. (**a**)3 bump; (**b**)15 bumps; (**c**) Soft thresholding. All cases use a generalized Gaussian prior (exponent 0.5), and additive Gaussian noise. Noisy SNR is 4.8 dB.

where the functions g_k are of the form

$$g_k(y) = y \ \cos^2\left(\frac{1}{\alpha}\text{sgn}(y)\log_2\left(|y|/\sigma + 1\right) - \frac{k\pi}{2}\right),$$

as illustrated in Fig. 2.2. Recently, linear parameterizations have been used in conjunction with Eq. (2.135) for image denoising in the wavelet domain [23].

We can substitute Eq. (2.136) into Eq. (2.135) and pick coefficients $\{\theta_k\}$ to minimize this criteria, which is a quadratic function of the coefficients. We then apply the resulting estimator the observations and measure the empirical SNR. For our simulations, we used a generalized Gaussian prior, with exponent 0.5. The noisy SNR was 4.8 dB. Figure 2.3 shows the empirical behavior of these "SURE-bump" estimators when using three bumps (Fig. 2.3**a**) and fifteen bumps (Fig. 2.3**b**), illustrating the bias-variance tradeoff inherent in the fixed parameterization. Three bumps behaves fairly well for small amounts of data, though the asymptotic behavior for large amounts of data is biased and thus falls short of ideal. Fifteen bumps asymptotes correctly but has very large variance for small amounts of data (over-fitting). A more sophisticated method might use cross validation or some other resampling method to appropriately set the number of bumps to try and minimize both these effects. For comparison purposes, we have included the behavior of SUREShrink [29], in which Eq. (2.6.2) is used to choose an optimal threshold, θ, for the function

$$f_\theta(y) = \mathrm{sgn}(y)(|y| - \theta)^+.$$

As can be seen, SURE thresholding shows significant asymptotic bias although the variance behavior is nearly ideal.

2.7 Discussion

We have reformulated the Bayes least squares estimation problem for a setting in which one knows the observation process, and has access to many observations. We do not assume the prior density is known, nor do we assume access to samples from the prior. Our formulation thus acts as a bridge between a conventional Bayesian setting in which one derives the optimal estimator from known prior and likelihood functions, and a data-oriented regression setting in which one learns the optimal estimator from samples of the prior paired with corrupted observations of those samples. In many cases, the prior-free estimator can be written explicitly, and we have shown a number of examples to illustrate the diversity of estimators that can arise under different observation processes. For three simple cases, we developed implementations and demonstrated that these converge to optimal BLS estimators as the amount of data grows. We also have derived a prior-free formulation of the MSE, which allows selection of an estimator from a parametric family. We have shown simulations for a linear family of estimators in the additive Gaussian case.

These simulations serve to demonstrate the potential of this approach, which holds particular appeal for real-world systems (machine or biological) that must learn the priors from environmental observations. Both methods can be enhanced by using data-adaptive parameterizations or fitting procedures in order to properly trade off bias and variance (as we will see, for example in Chapter 3, see also [27]). Included in this would be resampling techniques such as CV, which would allow appropriate choice of parameters for the different methods in a data adaptive way. Such methods will naturally be more computationally intensive. It is of particular interest to develop incremental implementations, which would update the estimator based on incoming observations. This would further enhance the applicability of this approach for systems that must learn to do optimal estimation from corrupted observations.

Obs. process	**Obs. density:** $P_{Y\|X}(\mathbf{y}\|\mathbf{x})$	**Numerator:** $N(\mathbf{y}) = \mathbf{L}\{P_Y\}(\mathbf{y})$
Discrete	$\mathbf{A}$	$(\mathbf{A} \circ X \circ \mathbf{A}^{-1})P_Y(\mathbf{y})$
Gen. Add.	$P_W(y-x)$	$\mathbf{y}P_Y$ $-\mathcal{F}^{-1}\left\{i\nabla_\omega \ln\left(\widehat{P_W}(\omega)\right)\widehat{P_Y}(\omega)\right\}$
Add. Gaussian [16]/[12]*	$\frac{\exp -\frac{1}{2}(\mathbf{y}-\mathbf{x}-\mu)^T\Lambda^{-1}(\mathbf{y}-\mathbf{x}-\mu)}{\sqrt{\|2\pi\Lambda\|}}$	$(\mathbf{y}-\mu)P_Y(\mathbf{y}) + \Lambda\nabla_\mathbf{y}P_Y(\mathbf{y})$
Add. Poisson	$\sum \frac{\lambda^k e^{-\lambda}}{k!}\delta(y-x-ks)$	$yP_Y(y) - \lambda s P_Y(y-s)$
Add. Laplacian	$\frac{1}{2\alpha}e^{-\|(y-x)/\alpha\|}$	$yP_Y(y) + 2\alpha^2\{P'_W \star P_Y\}(y)$
Add. Cauchy	$\frac{1}{\pi}(\frac{\alpha}{(\alpha(y-x))^2+1})$	$yP_Y(y) - \{\frac{1}{2\pi\alpha y} \star P_Y\}(y)$
Add. Uniform	$\begin{cases} \frac{1}{2a}, & \|y-x\| \le a \\ 0, & \|y-x\| > a \end{cases}$	$yP_Y(y) + a\sum \mathrm{sgn}(k)P_Y(y-ak)$ $-\frac{1}{2}\int P_Y(\tilde{y})\mathrm{sgn}(y-\tilde{y})d\tilde{y}$
Add. Random # of Components	$P_W(y-x)$, where: $W \sim \sum_{k=0}^{K} W_k$, W_k i.i.d. (P_c), $K \sim Poiss(\lambda)$	$yP_Y(y) - \lambda\{(yP_c) \star P_Y\}(y)$
Discr. Exp. [11]/[14]*	$h(x)g(n)x^n$	$\frac{g(n)}{g(n+1)}P_Y(n+1)$
Discr. Exp. (inv.) [14]*	$h(x)g(n)x^{-n}$	$\frac{g(n)}{g(n-1)}P_Y(n-1)$
Cont. Exp. [11]/[13]*	$h(x)g(y)e^{T(y)x}$	$\frac{g(y)}{T'(y)}\frac{d}{dy}\{\frac{P_Y(y)}{g(y)}\}$
Cont. Exp. (inv) [13]*	$h(x)g(y)e^{T(y)/x}$	$g(y)\int_{-\infty}^{y}\frac{T'(\tilde{y})}{g(\tilde{y})}P_Y(\tilde{y})d\tilde{y}$
Poisson [15]/[14]*	$\frac{x^n e^{-x}}{n!}$	$(n+1)P_Y(n+1)$
GSM	$\frac{1}{\sqrt{2\pi x}}e^{-\frac{y^2}{2x}}$	$-E_Y\{Y; Y < y\}$
Laplacian Scale Mixture	$\frac{1}{x}e^{-\frac{y}{x}}; x, y > 0$	$P_Y\{Y > y\}$

Table 2.1: Prior-free estimation formulas. Functions written with hats or in terms of ω represent multiplication in the Fourier Domain. n replaces y for discrete distributions. Bracketed numbers are references for operators $\mathbf{L}$, with * denoting references for the parametric (dual) operator, $\mathbf{L}^*$.

Chapter 3

Nonparametric Denoising of Additive Gaussian Noise

In Chapter 2, we illustrated two approaches for prior-free estimation. In the first method, we were able to write the BLS estimator using a linear operator derived from the known corruption process

$$E\{X|Y = \mathbf{y}\} = \frac{\mathbf{L}\{P_Y\}(\mathbf{y})}{P_Y(\mathbf{y})} \tag{3.1}$$

In this approach, we would like to find a nonparametric estimator of this quantity which is based on samples $\{Y_i\}$ drawn from P_Y, which we would then apply to those samples to estimate the corresponding hidden values. We would also like this estimator to adapt to the data in a way such that, as the amount of data increases, the estimator comes closer to the ideal BLS estimator. In this section, we will illustrate such an estimator for the particular case of additive Gaussian noise.

3.1 Prior-free formulation of BLS estimator for Additive Gaussian Noise

In the case of additive Gaussian noise with zero mean and covariance matrix Λ, we have from Chapter 2 [16] that

$$E\{X|Y=\mathbf{y}\} = \mathbf{y} + \frac{\Lambda\nabla_{\mathbf{y}}P_Y(\mathbf{y})}{P_Y(\mathbf{y}))} \quad (3.2)$$

$$= \mathbf{y} + \Lambda\nabla_{\mathbf{y}}ln(P_Y(\mathbf{y})), \quad (3.3)$$

where Λ is the covariance matrix of the noise. In the Chapter 2, we derived this as a special case of the general additive noise case. Since, in this chapter, we we will be focusing on the additive Gaussian case, we will re-derive this result more directly, using a more straightforward proof. First, we write the observation equation for additive Gaussian noise contamination:

$$P_{Y|X}(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Lambda|^{1/2}}e^{-\frac{1}{2}(\mathbf{y}-\mathbf{x})^T\Lambda^{-1}(\mathbf{y}-\mathbf{x})} \quad (3.4)$$

Next, note that this expression implies that

$$\nabla_{\mathbf{y}}P_{Y|X}(\mathbf{y}|\mathbf{x}) = \Lambda^{-1}P_{Y|X}(\mathbf{y}|\mathbf{x})(\mathbf{x}-\mathbf{y})\ . \quad (3.5)$$

Taking the gradient of

$$P_Y(\mathbf{y}) = \int P_{Y|X}(\mathbf{y}|\mathbf{x})P_X(\mathbf{x})d\mathbf{x} \quad (3.6)$$

with respect to $\mathbf{y}$, dividing by $P_Y(\mathbf{y})$, and substituting Eq. (3.5) yields:

$$\begin{aligned}
\frac{\nabla_{\mathbf{y}} P_Y(\mathbf{y})}{P_Y(\mathbf{y})} &= \frac{\int P_X(\mathbf{x})\, \nabla_{\mathbf{y}} P_{Y|X}(\mathbf{y}|\mathbf{x})\, d\mathbf{x}}{P_Y(\mathbf{y})} \\
&= \frac{\Lambda^{-1} \int P_X(\mathbf{x})\, P_{Y|X}(\mathbf{y}|\mathbf{x})\, (\mathbf{x}-\mathbf{y})\, d\mathbf{x}}{P_Y(\mathbf{y})} \\
&= \Lambda^{-1} \int P_{X|Y}(\mathbf{x}|\mathbf{y})\, (\mathbf{x}-\mathbf{y})\, d\mathbf{x} \\
&= \Lambda^{-1} \left[E\{X|Y=\mathbf{y}\} - \mathbf{y} \right] . \qquad (3.7)
\end{aligned}$$

Finally, rearranging the terms gives Eq. (3.2). In what follows, we will restrict ourselves to discussing the case of scalar data.

3.2 Learning the estimator function from data

The prior-free formulation of the BLS estimator shifts the problem from one of approximating the prior, a distribution from which samples are generally not available, to one of approximating the noisy distribution, from which samples are available. But simple histograms will not suffice for this approximation, because Eq.(3.2) requires us to find the logarithmic derivative of the distribution.

3.2.1 Approximating local logarithmic derivative

A natural solution for this problem is to approximate the logarithmic derivative of the density at the observation $Y_k = y$ as being constant over some interval (x_0, x_1) containing y. This will be a good approximation if the density is approximately exponential in the interval:

$$P_Y(y) \approx ce^{-ay}, \qquad x_0 < y < x_1 \qquad (3.8)$$

where a is the estimate of the logarithmic derivative in the interval (x_0, x_1). Note that, in our problem, it is the a's which are to be used for the estimator, while the c's are irrelevant. For this reason, we look instead at the conditional density of y given that y is in the interval (x_0, x_1)

$$\begin{aligned} P_{Y|Y\in(x_0,x_1)}(y) &= \frac{e^{-ay}}{\int_{x_0}^{x_1} e^{-ay}dy} I_{(x_0,x_1)} \\ &= \frac{\frac{a}{2}e^{-a(y-\bar{x})}}{\sinh(\frac{a}{2}\Delta x)} I_{(x_0,x_1)} \end{aligned} \tag{3.9}$$

where $I_{(x_0,x_1)}$ denotes the indicator function of (x_0, x_1), $\bar{x} = \frac{x_0+x_1}{2}$ and $\Delta x = x_1 - x_0$. Comparing this with Eq. (3.8), we see that the conditional density is also an exponential function of y over the interval (x_0, x_1), with the same exponent a, but is normalized so that c no longer appears, and so that it integrates to one over the interval. If we then have observations Y_n drawn from $P_Y(y)$, and keep only data which fall in (x_0, x_1), these data will have distribution $P_{Y|Y\in(x_0,x_1)}(y)$, so we can use this to estimate the parameter a.

One very popular estimator used for such type of problems is the Maximum Likelihood (ML) estimator. Assuming that Eq.(3.9) is a good approximation of the conditional density on (x_0, x_1), this estimator can be written

$$\begin{aligned} \hat{a} &= \arg\max_a \sum_{\{n:Y_n\in(x_0,x_1)\}} \ln(P_{Y|Y\in(x_0,x_1)}(Y_n)) \\ &= \arg\max_a \{\ln(a) - a(\bar{Y} - \bar{x}) - \ln(\sinh(\frac{a}{2}\Delta x))\} \end{aligned} \tag{3.10}$$

where

$$\bar{Y} \overset{def}{=} \frac{1}{\#\{Y_n \in (x_0, x_1)\}} \sum_{Y_n\in(x_0,x_1)} Y_n \tag{3.11}$$

is the average of the data that fall into (x_0, x_1). Setting the derivative of Eq. (3.10) with respect to a equal to zero yields

$$\frac{1}{\hat{a}} - (\bar{Y} - \bar{x}) - \coth(\frac{\hat{a}}{2}\Delta x)\frac{\Delta x}{2} = 0 \tag{3.12}$$

or

$$\frac{1}{\frac{\hat{a}\Delta x}{2}} - \coth(\frac{\hat{a}\Delta x}{2}) = \frac{\bar{Y} - \bar{x}}{\frac{\Delta x}{2}} \tag{3.13}$$

Solving this for $\hat{a}$ gives

$$\hat{a} = \frac{2}{\Delta x} f^{-1}(\frac{\bar{Y} - \bar{x}}{\frac{\Delta x}{2}}) \tag{3.14}$$

where

$$f(y) = \frac{1}{y} - \coth(y) \tag{3.15}$$

This local exponential approximation is similar to that used in [28] except that, since we are approximating the local *conditional* density, c disappears from the equation for $\hat{a}$. This has the benefit that we only need to invert a scalar function of one variable, f, to calculate the estimate at all points, instead of inverting a two dimensional vector function of two variables, as in the other method.

Obviously, it is $\bar{Y}$, the local mean, which requires the most calculation, but, since most of this calculation comes from adding up the value of data which fall in the interval, this may be done in an iterative way, subtracting or adding from a running sum. This method is efficient enough that it may be calculated at each data point, instead of on a grid with interpolation.

3.2.2 Choice of binwidth

In order to calculate Eq. (3.14) for a particular y, it is necessary to choose the interval (x_0, x_1), or, equivalently, to choose the binwidth $h = x_1 - x_0$. In order to say what we mean by an optimal binwidth, we must choose a measure of how "good" an estimate is. Once again, we will use the MSE of the estimate, which may be separated into a variance term and a bias term

$$\begin{aligned} E\{(\hat{a}-a)^2\} &= E\{((\hat{a}-E\{\hat{a}\})+(E\{\hat{a}\}-a))^2\} \\ &= E\{(\hat{a}-E\{\hat{a}\})^2\}+(E\{\hat{a}\}-a)^2 \\ &= Var\{\hat{a}\}+(E\{\hat{a}\}-a))^2 \qquad (3.16) \end{aligned}$$

where $\hat{a}$ is the data-dependent estimate of the true value a. The first term is the variance of the estimator, $\hat{a}$ and will decrease as the binwidth of the interval is increased, since more data will fall into the interval, giving a more reliable estimate. The second term is the squared bias, which will conversely increase as the interval is increased, since the exponential fit of the density over the interval will in general become worse, which means that the estimate $\hat{a}$ will not give a good estimate of the true value of the logarithmic derivative, a. This is known as the bias-variance tradeoff.

In order to choose an optimal binwidth, we must analyze how Eq. (3.16) behaves as a function of the binwidth, h. For large amounts of data, we expect h to be small, and so we may use small h approximations for the bias and variance. In general, the variance in estimating the parameter, a, for the interval (x_0, x_1) will depend inversely on the amount of data which falls in the interval. If there are N total data points, we will approximate the number falling in the interval (x_0, x_1)

as

$$n \approx P_Y(y)Nh \tag{3.17}$$

Hence, we will assume that

$$Var\{\hat{a}\} \approx \frac{C}{P_Y(y)Nh} \tag{3.18}$$

for an appropriate constant, C.

On the other hand, the squared bias will generally depend only on how well the exponential fits the true density over the interval. As $h \to 0$ the bias for the interval will decrease to zero. For small h we will therefore assume that

$$(E\{\hat{a}\} - a)^2 \approx Dh^m \tag{3.19}$$

where $D = D(P_Y, y)$ depends only on the shape of P_Y in the interval, but not on the actual value $P_Y(y)$ (see [28]). In what follows, we will assume that the density is smooth enough that we may ignore the dependence of D on shape, and treat D as constant for all values of y. Since, in our case, P_Y comes from convolving P_X with a Gaussian, P_Y will be at least as smooth as P_X, and will become smoother as the noise variance increases. Therefore, this approximation will become better as the amount of noise increases.

Assuming that the approximation of the true logarithmic derivative of the density by a constant is of first order in h leads to the result that the squared bias will be of order h^2, which gives $m = 2$. This may be justified by the use of Taylor series when h is very small.

Putting everything together than yields the approximation

$$E\{(\hat{a}-a)^2\} \approx \frac{C}{P_Y(y)Nh} + Dh^m \tag{3.20}$$

Setting the derivative of this equation with respect to h equal to zero yields

$$Dmh^{m+1} - \frac{C}{P_Y(y)N} = 0 \tag{3.21}$$

or

$$h = (\frac{C}{DmP_Y(y)N})^{\frac{1}{m+1}} \tag{3.22}$$

which verifies our assumption that $h \to 0$ as the amount of data increases. Substituting this into Eq. (3.20) gives

$$\left(\frac{(DmC^m)^{\frac{1}{m+1}}}{(P_Y(y)^{\frac{m}{m+1}})} + D^{\frac{1}{m+1}} (\frac{C}{mP_Y(y)})^{\frac{m}{m+1}} \right) \frac{1}{N^{\frac{m}{m+1}}} \tag{3.23}$$

which also shows that both the squared bias and variance, and hence the MSE, go to zero as $N \to \infty$. Using Eq. (3.17) to approximate P_Y in Eq. (3.22) gives

$$h \approx (\frac{Ch}{Dmn})^{\frac{1}{m+1}} \tag{3.24}$$

Rearranging this equation gives

$$nh^m = \frac{C}{Dm} \tag{3.25}$$

which says that the optimal binwidth is chosen such that the product of the number of points which fall in the interval times some power of the binwidth of the interval is constant.

3.2.3 Choice of power

To determine the binwidth, it is necessary to determine the constant m. If $m = 0$, then n, the number of data points in the neighborhood, will be constant for all data points, a method known as k nearest neighbors (KNN). In the limit as $m \to \infty$, the binwidth will be fixed at a constant value for all data points. As discussed, a first order assumption of the fit will lead to $m = 2$, in which case there will be an interplay between the binwidth and number of points in the interval.

In this section we compare the empirical behavior of these three methods of binwidth selection to see how they behave for two different distributions. To put all three methods on the same footing, the constant product for each is chosen so that the average binwidth across data points is the same for all three methods. Thus, we are looking at how well the three methods allocate this average binwidth.

The first density we examine is the Cauchy distribution.

$$P_Y(y) \propto \frac{1}{1 + 0.5y^2} \tag{3.26}$$

so that

$$\frac{d}{dy} \ln(P_Y(y)) = \frac{y}{1 + 0.5y^2} \tag{3.27}$$

Figure 3.1 shows the behavior of the estimate of the logarithmic derivative for the three different methods of binwidth selection for a sample of $9,000$ points drawn from the Cauchy distribution. As can be seen, the KNN method ($m = 0$) has a systematic bias in the tails, the fixed binwidth ($m \to \infty$) method has larger variance in the tails, while the $m = 2$ method has reduced the bias seen in the KNN

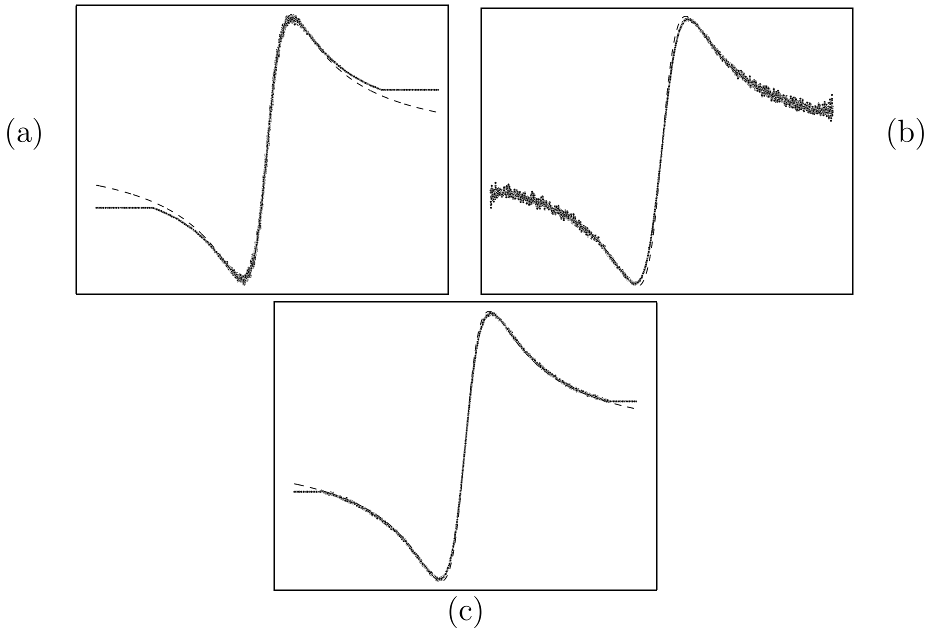


Fig. 3.1: Estimate of logarithmic derivative of Cauchy (dashed line is actual value) (**a**) using KNN ($m = 0$); (**b**)using fixed binwidth ($m = \infty$; (**c**) using $m = 2$ to select binwidth

method without introducing the variance present in the fixed binwidth method.

The second density we will look at will be the Laplacian distribution.

$$P_Y(y) \propto e^{-|x|} \tag{3.28}$$

which gives

$$\frac{d}{dy}\ln(P_Y(y)) = \text{sgn}(x) \tag{3.29}$$

Figure 3.2 shows the behavior of the estimate of the logarithmic derivative for the three different methods of binwidth selection on $9,000$ points drawn from the Laplacian distribution. Notice that in this case, since the logarithmic derivative *is* constant away from the origin, there won't be any bias problem. As can be seen in this case, the KNN method has more of a variance problem near the origin, the

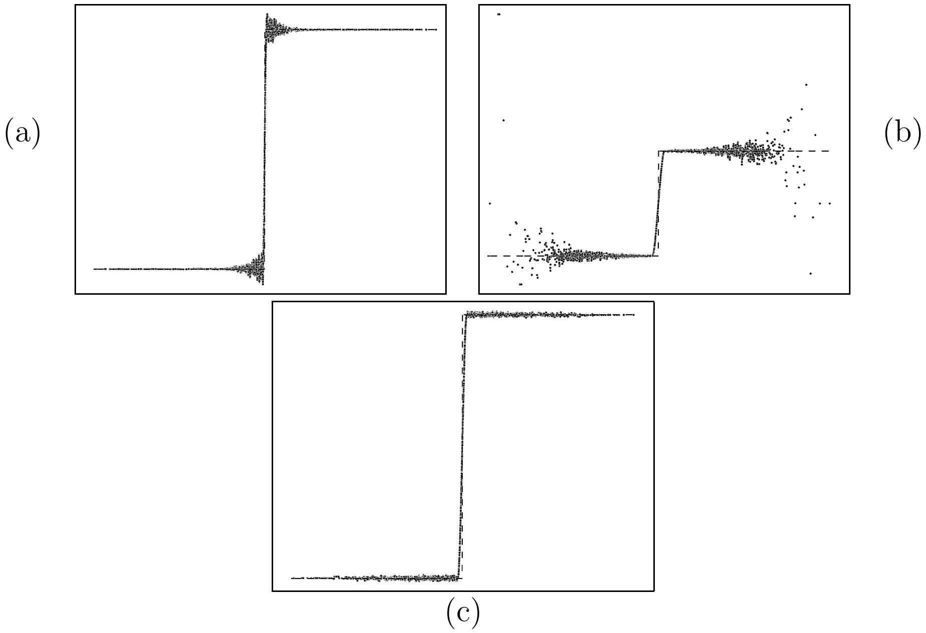


Fig. 3.2: Estimate of logarithmic derivative of Laplacian (dashed line is actual value) (**a**) using KNN ($m = 0$); (**b**)using fixed binwidth ($m = \infty$; (**c**) using $m = 2$ to select binwidth

fixed binwidth method has larger variance in the tails, while the $m = 2$ method has reduced the variance near the origin without introducing variance in the tails. Based on this analysis, in what follows we will restrict ourselves to using the $m = 2$ method.

The next question is how to choose what the average binwidth should be. Equivalently, we are trying to determine what the constant value of the product in Eq. (3.25) should be. In the examples that follow, we will choose the constant so that the average binwidth across the data is proportional to the $\sigma_Y N^{-\frac{1}{m+1}}$, where σ_Y is the standard deviation of the observed data Y. The dependence on σ_Y stems from the intuition that if the data are multiplied by some constant the density will simply be stretched out by that factor, and so the binwidth should get proportionally wider to include the same data and exponential fit. The behavior as a function of N is read off Eq. (3.22).

Now that we have a method of binwidth selection, $\bar{Y}$,$\bar{x}$ and Δx, can all be calculated, then Eq. (3.14) applied to obtain the estimate of the logarithmic derivative, which is then used in Eq. (3.2) to obtain the BLS estimator.

3.3 Approach to ideal BLS estimator with increase in data

Since each binwidth shrinks and the amount of data in each bin increases with increasing amounts of data, our BLS estimator will approach the ideal BLS estimator as the amount of data increases. In Fig. 3.3, (which also appears as a sub-figure of Fig. 2.1 in Chapter 2) we illustrate this behavior. For this figure, the density of the prior signal is a generalized Gaussian distribution (GGD)

$$P_X(\mathbf{x}) \propto e^{-|\mathbf{x}/\mathbf{s}|^p} \; . \tag{3.30}$$

with $s = 1$, and exponent $p = 0.5$. As described in Chapter 2, we characterize the behavior of this estimator as a function of the number of data points, N, by running many Monte Carlo simulations for each N and indicating the mean improvement in empirical SNR (as measured by increase in empirical SNR compared to the ML estimator, which is the identity function), the mean improvement using the

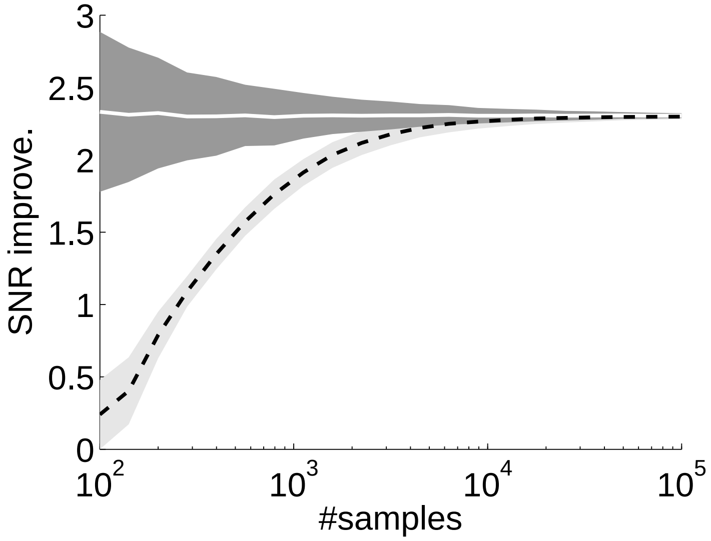


Fig. 3.3: Empirical convergence of prior-free estimator to optimal BLS solution, as a function number of observed samples of Y. For each number of observations, each estimator is simulated many times. Black dashed lines show the improvement of the prior-free estimator, averaged over simulations, relative to the ML estimator. White line shows the mean improvement using the conventional BLS solution, $E\{X|Y=\mathbf{y}\}$, assuming the prior density is known. Gray regions denote ± one standard deviation.

conventional BLS estimation function,

$$\begin{aligned} E\{X|Y=y\} &= \frac{\int xP_X(x)P_{Y|X}(y|x)dx}{\int P_X(x)P_{Y|X}(y|x)dx} \\ &= \frac{\int xP_X(x)e^{\frac{-(y-x)^2}{2\sigma^2}}dx}{\int P_X(x)e^{\frac{-(y-x)^2}{2\sigma^2}}dx}, \end{aligned} \tag{3.31}$$

and the standard deviations of the improvements taken over our simulations.

As can be seen, our estimator does approach the behavior of the ideal BLS estimator as the amount of data increases. It does this without making any assumption about the prior density of the data, instead adapting to the data it does observe. As can also be seen, the variance of this estimator is quite low, for even moderate amounts of data.

3.4 Comparison with Empirical Bayes

As we have discussed, our prior free estimator will adapt to the observed data, and, given enough data, will give behavior that is near ideal, regardless of the form of the prior distribution. If, instead, we were to assume a particular parametric form for the prior distribution, as in the commonly used Empirical Bayes method[10], and the true prior did not fall into this parametric family, then the behavior of this estimator would likely be compromised. Thus, our estimator gives a potential advantage over methods which use parametric forms for estimators, since it makes no assumptions about the prior distribution. In exchange, it may require more data than a parametric method. In this section, we will compare the empirical behavior of our estimator with that of a parametric estimator under conditions where the assumptions of the parametric estimator are valid and under conditions where these assumptions are false.

For our simulations, the Empirical Bayes estimator, based on [26], assumes a GGD form for the prior, as in Eq. (3.30). The parameters, p and s, are fit to the noisy observation by maximizing the likelihood of the noisy data, and the estimator is computed by numerical integration of

$$\hat{X}_{GGD}(y) = \frac{\int x e^{-|x/s|^p} e^{\frac{-(y-x)^2}{2\sigma^2}} dx}{\int e^{-|x/s|^p} e^{\frac{-(y-x)^2}{2\sigma^2}} dx} \tag{3.32}$$

and this estimator is then applied to the noisy observations.

3.4.1 Prior Distributions

Since the eventual application we have in mind is in image processing, we picked $9,000$ data points in our simulation, a reasonable number for such applications. The priors we will deal with are shown in Fig. 3.4. The first is the Laplacian prior (a special case of the GGD), the second is a Laplacian prior with shifted mean, the third is a bimodal Laplacian

$$P_X(x) \propto \frac{1}{2}e^{-|x-m|} + \frac{1}{2}e^{-|x+m|} \tag{3.33}$$

and the fourth is an asymmetric GGD:

$$P_X(x) \propto \begin{cases} e^{-|\frac{x}{s_1}|^{p_1}}, & x \leq 0 \\ e^{-|\frac{x}{s_2}|^{p_2}}, & x > 0 \end{cases} \tag{3.34}$$

where the constants are chosen such that the distribution still has zero mean. Thus, the first distribution fits the model assumed by the Empirical Bayes method, whereas the last three break it in some simple ways.

3.4.2 Results

In these cases, since the prior is known the optimal solution may be calculated directly numerically integrating Eq. (3.31). Figure 3.5 shows the estimators, also known as coring functions, obtained for the prior-free and GGD methods from the observed data, as compared with the optimal solution calculated by numerical integration of Eq. (3.31). Table 3.4.2 shows the empirical SNR obtained from applying these methods to the observed data, for the priors discussed, as simulated

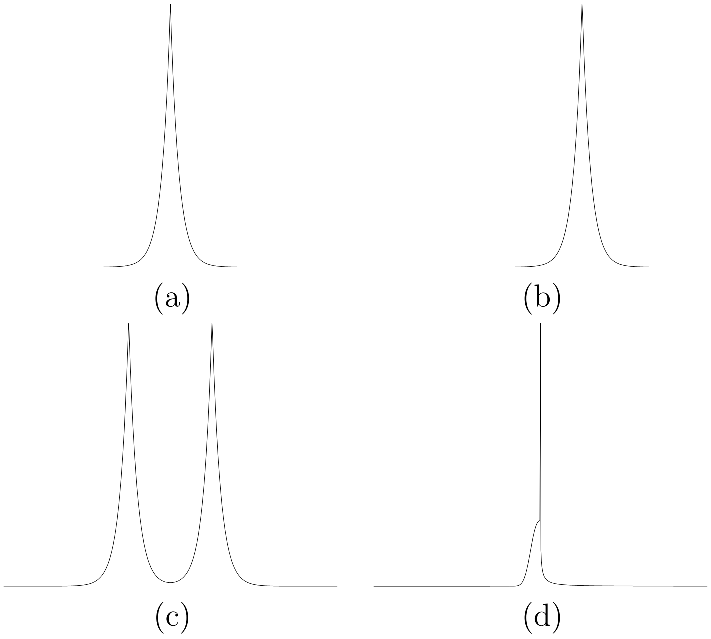


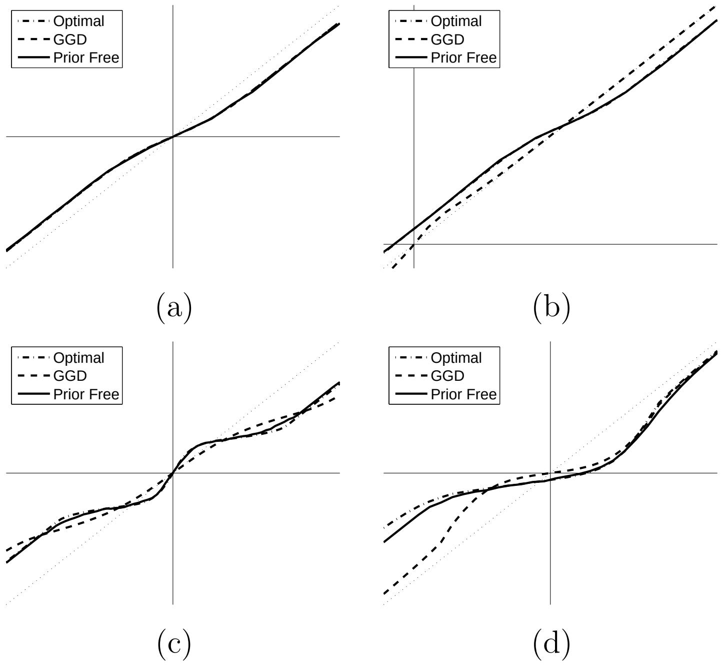
Fig. 3.4: Other Priors: (**a**) Laplacian (**b**) shifted Laplacian (**c**) bimodal Laplacian (**d**) asymmetric GGD

for various values of noise power.

As is to be expected, in the case where the prior actually fits the assumptions of the GGD model, then the GGD method will outperform the prior-free method, though, it should be noted, not by very much. In the cases where the assumption on the prior is broken in some simple ways, however, the performance of the GGD method degrades considerably while that of the the prior-free method remains surprisingly close to ideal.

3.5 Image denoising example

In this section we describe a specific example of this prior-free approach as applied to image denoising. The development of multi-scale (wavelet) representa-



(a) (b)

(c) (d)

Fig. 3.5: Coring Functions for: (**a**) Laplacian (**b**) shifted Laplacian (**c**) bimodal Laplacian (**d**) asymmetric GGD prior distributions. In all figures, the dotted line denotes the identity function for reference.

Prior	Noise	Denoised SNR		
Distn.	SNR	Opt.	GGD	Prior-free
Lapl.	1.800	4.226	4.225	4.218
	4.800	6.298	6.297	6.291
	7.800	8.667	8.667	8.666
	10.800	11.301	11.301	11.299
Shifted	1.800	4.219	2.049	4.209
	4.800	6.273	4.920	6.268
	7.800	8.655	7.762	8.651
	10.800	11.285	10.735	11.284
Bimodal	1.800	4.572	4.375	4.547
	4.800	7.491	6.767	7.468
	7.800	10.927	9.262	10.885
	10.800	13.651	11.776	13.603
Asym.	1.800	7.102	6.398	7.055
	4.800	8.944	8.170	8.915
	7.800	10.787	10.044	10.767
	10.800	12.811	12.143	12.791

Table 3.1: Simulated denoising results.

tions has led to substantial improvements in many signal processing applications, especially denoising. Typically, the signal (or image) is decomposed into frequency bands at multiple scales, each of which is independently denoised by applying a pointwise nonlinear shrinkage function that suppresses low-amplitude values. The concept was developed originally in the television engineering literature (where it is known as "coring"[30, 31, e.g.]), and specific shrinkage functions have been derived under a variety of formulations, including minimax optimality under a smoothness condition [32, 33, 34], and Bayesian estimation with non-Gaussian priors [26, 35, 36, 37, 38, 39, 40, 41, e.g.]. Note that, although such methods denoise each coefficient separately, a process which will not generally be optimal unless the coefficients are independent (which is impossible for redundant transformations. for example), such marginal denoising methods have proven effective. This must be because the statistics of the coefficients, while not independent, are sufficiently "close" to independent for this method to give improvement.

As in [26, 38, 42], we begin by decomposing the noisy image using a steerable pyramid. This is a redundant, invertible linear transform that separates the image content into oriented octave-bandwidth frequency subbands. We apply our prior free estimator to each subband separately, using the noisy data in a subband to construct an estimator for that subband. We then apply the subband estimator to the noisy coefficients in the subband in order to estimate the values of the original, noise-free subband. After the coefficients of each subband have been processed, the inverse pyramid transform is applied in order to reconstruct the denoised image.

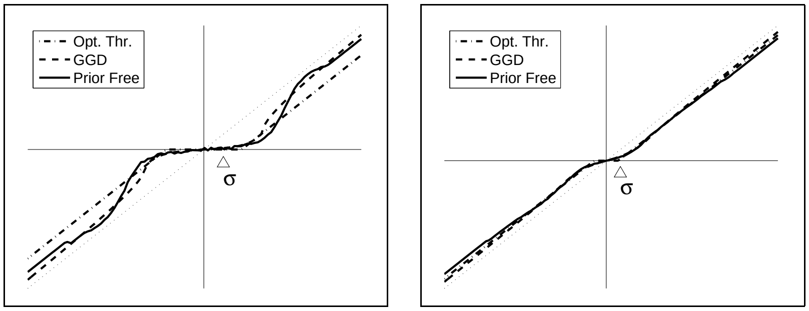


Fig. 3.6: Example estimators (coring functions) for the two subbands: Prior-free Bayesian estimator (solid), BLS estimator for a GGD (dashed), and optimal soft threshold (dash-dotted). Dotted line indicates the identity function. Noise standard deviation σ is also indicated.

3.5.1 Results

We have applied our prior-free Bayesian estimator to several images contaminated with simulated Gaussian noise. For all examples, the noise variance was assumed to be known. The results were compared with two other methods of denoising. The first method [26], described in the last section, uses ML to fit the parameters of a GGD prior, Eq. (3.30), to the noisy data in the subband. This is justified by the fact that the GGD is a parametric form which is known to provide good fits for the marginal densities of coefficients in image subbands [26, 38, 39]. We then use use this parametric prior to find the associated estimator by numerical integration of Eq. (3.32).

The second estimator is a “soft threshold” function[32]:

$$\hat{x}(Y) = \begin{cases} Y - t, & t \leq Y \\ 0, & -t < Y < t \\ Y + t, & Y \leq -t \ . \end{cases} \tag{3.35}$$

We make use of the clean, original data to find a soft threshold for each subband that minimizes the empirical mean squared error in that subband. Thus, the performance of this method should not be interpreted as resulting from a feasible denoising algorithm, but rather as an upper bound on thresholding approaches to denoising. Two example estimators are shown in Fig. 3.6.

Figure 3.7 shows a sample of an image denoised using these three methods. Table 3.5.1 shows denoising results for some sample images under several noise conditions. As can be seen, the prior-free approach compares favorably to the other two, despite the fact that it makes weaker assumptions about the prior than does the generalized Gaussian, and doesn't have access to the clean data, as does the optimum thresholding. Figure 3.8 shows a histogram of SNR improvement of the prior-free algorithm over optimal thresholding and generalized Gaussian approaches for nine images at four different noise levels. As we can see, our prior free method compares favorably with the parametric method, which was based on detailed empirical knowledge of the statistics of image coefficients.

3.6 Discussion

We've developed a modified formulation of the Bayes least squares estimator in the case of additive Gaussian noise. Unlike the traditional form, this estimator is written in terms of the distribution of the noisy measurement data, and is thus more natural for situations in which the prior must be learned from the data. We've shown that as the amount of data is increased, the prior free estimator will tend to give performance that is near ideal. We've also shown that breaking the assumptions of parametric models of the prior leads to a drastic reduction in the

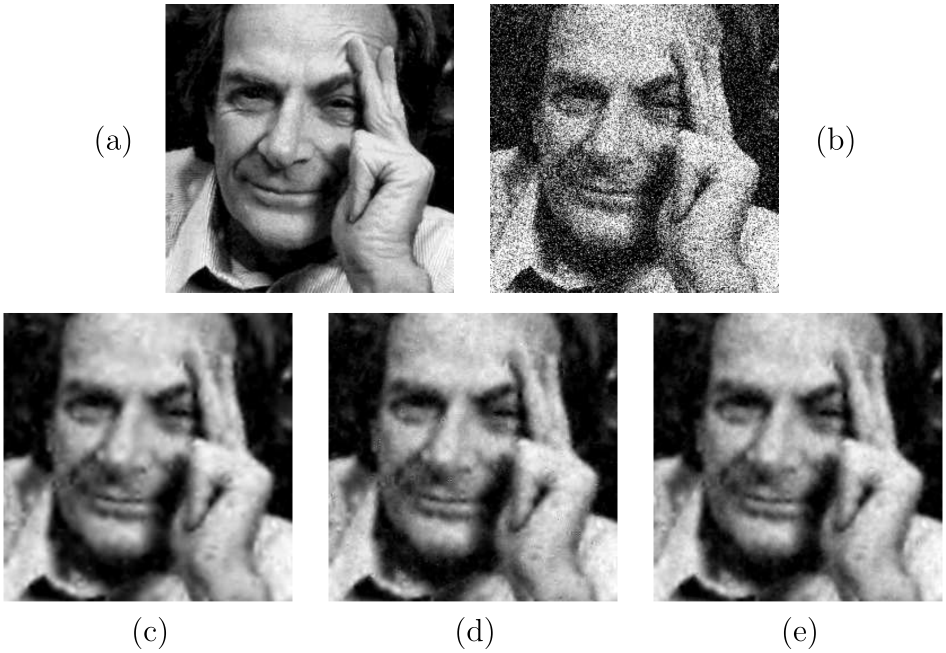


Fig. 3.7: Denoising results for the “Feynman” image. (**a**) original; (**b**) noisy image (SNR = 1.8 dB); (**c**) using optimal thresholding (SNR = 14.11 dB) (**d**) using generalized Gaussian (SNR = 13.86 dB) (**e**) using prior-free denoising (SNR = 13.95 dB)

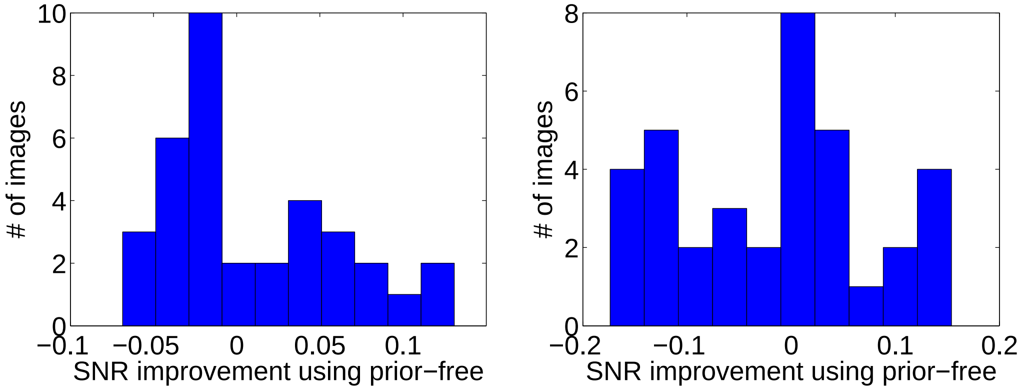


Fig. 3.8: Improvement in SNR for prior-free approach compared with the GGD estimator (left) and optimal thresholding (right). Histograms summarize data for 9 images at 4 noise levels.

Image	Noise	Denoised SNR		
	SNR	Opt. Thr.	GGD	Prior-free
crowd	1.8000	12.3873	12.1682	12.2547
	4.8000	13.9415	13.7585	13.7996
	7.8000	15.6572	15.4715	15.5225
	10.8000	17.4312	17.2917	17.3145
feynman	1.8000	14.1194	13.8432	13.9457
	4.8000	15.2441	15.1393	15.1612
	7.8000	16.5077	16.4731	16.4417
	10.8000	17.7889	17.8045	17.7658
boats	1.8000	12.5215	12.3593	12.4807
	4.8000	13.9687	13.8955	13.9661
	7.8000	15.5719	15.5383	15.6021
	10.8000	17.2541	17.2601	17.3484
einstein	1.8000	10.9483	10.8459	10.7773
	4.8000	12.3319	12.2796	12.2191
	7.8000	13.7506	13.7469	13.6893
	10.8000	15.3258	15.3685	15.3277
lena	1.8000	13.5310	13.2506	13.3814
	4.8000	15.0002	14.8227	14.8809
	7.8000	16.4823	16.3755	16.4229
	10.8000	18.1261	18.0736	18.1233
bench	1.8000	8.0068	8.0795	8.0389
	4.8000	9.3484	9.4211	9.3905
	7.8000	10.9993	11.0393	11.0213
	10.8000	12.9475	12.9762	12.9735
brick	1.8000	7.7446	7.6724	7.6425
	4.8000	9.2697	9.2819	9.2668
	7.8000	10.8920	11.0206	11.0063
	10.8000	12.7608	12.9270	12.9148
bridge	1.8000	9.4794	9.4669	9.4705
	4.8000	10.7470	10.7599	10.7403
	7.8000	12.1666	12.1983	12.1842
	10.8000	13.8568	13.9175	13.9012

Table 3.2: Simulated denoising results.

performance of methods based on such assumptions, while the prior-free method is able to deal with such changes. Finally, we've demonstrated the feasibility of this methodology by applying it to the problem of image denoising, demonstrating that it performs as well or better than estimators based on marginal prior models found in the literature, which are based on empirical studies of the marginal statistics of clean image subbands. Therefore, in situations where the prior distribution of the clean data is unknown, our method can be used, with some confidence that not too much is lost by not studying the empirical statistics of clean data, which may not even be possible in some situations.

It must be pointed out that the prior-free method is restricted in that it requires a lot of data to be feasible. Also, in cases where an accurate model of the prior is available, methods that make use of this explicit model will give some improvement, although we have seen some situations where this is not by much. If nothing is known about the prior, and there is a lot of data, then the prior-free method should give improvement over an ad-hoc assumption about the prior.

In order to obtain image denoising results which are competitive with the state of the art, it is necessary to jointly denoise vectors of coefficients, instead of one coefficient at a time [42]. While it is true that Eq. (3.2) holds for vectors as well as scalars, finding neighborhoods of vectors to use in estimating the logarithmic gradient at a point becomes much more difficult. For higher dimensions the vectors will tend to be further and further apart (the "curse" of dimensionality), so great care must be taken in choosing the shape of the large neighborhoods required to include sufficient number of data points.

Chapter 4

Optimal Denoising in Redundant Bases

As discussed in Chapter 3, image denoising has undergone dramatic improvement over the past decade, due to both the development of linear decompositions that simplify the statistical characteristics of the signal, and to new estimators that are optimized for those characteristics. A standard methodology proceeds by linearly transforming the image, operating on the transform coefficients with pointwise nonlinear functions, and then applying the inverse linear transformation. If the pointwise nonlinearity is chosen from a parametric family, Stein's unbiased risk estimator (SURE) [12], introduced in Chapter 2, may be used to select the estimator that minimizes the MSE [22]. The most popular transforms are multi-scale decompositions, and within this family, empirical evidence indicates that redundant, or overcomplete, representations are more effective than orthogonal representations [43, 44, 38]. This fact is somewhat mysterious since the estimators are usually chosen to minimize MSE within each subband of the the transform domain, which, for an overcomplete basis, is not the same as the choosing the functions to minimize MSE in the image domain.

In this chapter we extend the SURE methodology to approximate the image-domain MSE that results from denoising in an overcomplete basis. We use this to prove that application of a given denoising function to a basis made overcomplete through the method known as cycle-spinning or through elimination of decimation is guaranteed to be no worse in MSE than applying the same function in an orthonormal basis. We also use this extension of SURE to optimize two example pointwise estimators, operating on undecimated wavelet subbands, to minimize MSE in the image domain. We show through simulations that this can result in significant performance improvements.

4.1 Stein's Lemma for overcomplete bases

Recall from Chapter 2 that if X is a random vector which is corrupted by Additive White Gaussian Noise (AWGN), Y is the observed, noisy vector and our estimator is of the form

$$\hat{X}(Y) = Y + g(Y)$$

for g in some family, $\mathcal{G}$, of vector functions, then the MSE, otherwise known as risk, may be written in a prior free way using Stein's Lemma:

$$E\left\{|X - (Y + g(Y))|^2\right\} = E\left\{|g(Y)|^2 + 2\sigma^2(\nabla \cdot g)(Y)\right\} + \sigma^2 d \qquad (4.1)$$

where d is the dimension of X. Noting that σ^2 is a constant, we can write the optimal estimator as

$$g_{\text{opt}} = \arg\min_{g \in \mathcal{G}} E\left\{|g(Y)|^2 + 2\sigma^2(\nabla \cdot g)(Y)\right\}. \tag{4.2}$$

We now think of a clean image as a single vector-valued sample, X, which is corrupted by AWGN, to form the noisy image Y, which is also a single vector valued sample. Given Y, g_{opt} can be approximated by minimizing the expression in braces, which is (up to an additive constant) SURE [12].

It is common to apply estimators to a linearly transformed version of the image, in which the statistical properties are simplified, and in which the form of the estimators are simpler. Stein's Lemma is readily extended to this situation. Suppose we have a family of estimators $\{u + g_u(u) : g_u \in \mathcal{G}_U\}$ which act on $U = WY$, a transformed version of the image Y. Here W is an m by n matrix representing a linear transformation which can be complete $(m = n)$ or overcomplete$(m > n)$, and which has a left inverse $W^\dagger$. The estimate is computed by taking the transform of the sample vector using W, applying g_u, and taking the inverse transform using $W^\dagger$:

$$\begin{aligned} \hat{X}(Y) &= W^\dagger(WY + g_u(WY)) \\ &= Y + W^\dagger g_u(WY). \end{aligned} \tag{4.3}$$

To optimize this for MSE, we replace $g(Y)$ by $W^\dagger g_u(WY)$ in Eq. (4.2), and after a bit of calculus obtain:

$$g_{u,\text{opt}} = \arg\min_{g_u \in \mathcal{G}_U} E\left\{ |W^{\dagger} g_u(U)|^2 + 2\sigma^2 \text{tr}\left(W W^{\dagger} \frac{\partial g_u}{\partial u}(U) \right) \right\} \tag{4.4}$$

where tr(·) indicates the trace of the matrix. As before, the expression in braces is an unbiased estimate of MSE, and can be optimized for the single sample of Y. For simplicity, in what follows we will assume that the transform is a tight frame, defined as one for which $W^{\dagger} = W^T$.

4.2 Point Estimators

Suppose now that g_u operates by applying the scalar function g_i to the i^{th} element of U (the transformed version of Y). The unbiased risk estimator then becomes

$$|W^T g_u(U)|^2 + 2\sigma^2 \sum_i n_{ii} g_i'(U_i)$$

where

$$n_{ij} = \left(W W^T \right)_{ij}$$

so that n_{ii} are the diagonal elements of (WW^T) (the squared norms of the basis functions). Often, the transform coefficients are separated into subbands $\{\mathcal{S}_k; k = 1, 2, \ldots K\}$, where each subband contains coefficients which are calculated by taking the dot product of the image with shifted versions of the same basis function. If the image has shift invariant statistics, the coefficients in a particular subband will have the same marginal statistical properties. In this case, the same estimator

g_i will be applied to all coefficients within a band, $\mathcal{S}_i$. The unbiased risk estimator now becomes

$$|W^T g_u(U)|^2 + 2\sigma^2 \sum_k n_k \sum_{i \in \mathcal{S}_k} g_k'(U_i) \tag{4.5}$$

where n_k is the common value of n_{ii} for $i \in \mathcal{S}_k$. For a single transformed image $U = WY$, this expression provides a criterion for choosing $\{g_k\}_{k=1}^K$ from an appropriate family so as to minimize the MSE in the image domain.

4.3 Optimal nonlinearity

It is of theoretical interest to find an equation for the optimal nonlinearities that one can use for denoising in an overcomplete basis. Suppose that we denoise by applying g_k to the k^{th} band. Then, from Eq.(4.5), the unbiased estimator of risk will be

$$\sum_{k,l} \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l}} g_k(U_i) n_{ij} g_l(U_j) + 2\sigma^2 \sum_k n_k \sum_{i \in \mathcal{S}_k} g_k'(U_i) \tag{4.6}$$

If the MSE is minimized by the pointwise nonlinearities $\{g_{k0}\}$, then

$$\begin{aligned} E\Big\{ \sum_{k,l} \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l}} &(g_{k0}(U_i) + \epsilon_k g_{k1}(U_i)) n_{ij} (g_{l0}(U_j) + \epsilon_l g_{l1}(U_j)) \\ &+ 2\sigma^2 \sum_k n_k \sum_{i \in \mathcal{S}_k} (g_{k0}'(U_i) + \epsilon_k g_{k1}'(U_i)) \Big\} \end{aligned} \tag{4.7}$$

will be minimized at $\vec{\epsilon} = \mathbf{0}$. Setting the gradient with respect to $\vec{\epsilon}$ equal to zero at $\vec{\epsilon} = \mathbf{0}$ gives

$$E\Big\{ \sum_{l} \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l}} g_{k1}(U_i) n_{ij} g_{l0}(U_j) + \sigma^2 n_k \sum_{i \in \mathcal{S}_k} g'_{k1}(U_i) \Big\} = 0 \tag{4.8}$$

for every $1 \leq k \leq K$, and for arbitrary g_{k1}.

Writing the expectations out explicitly gives

$$\begin{aligned} \int\int \sum_{l} \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l}} g_{k1}(u) n_{ij} g_{l0}(v) P_{U_i,U_j}(u,v) du dv \\ = -\sigma^2 n_k \int \sum_{i \in \mathcal{S}_k} g'_{k1}(u) P_{U_i}(u) du \\ = \sigma^2 n_k \int \sum_{i \in \mathcal{S}_k} g_{k1}(u) P'_{U_i}(u) du \end{aligned} \tag{4.9}$$

where the last step uses integration by parts. We therefore have that

$$\begin{aligned} \int g_{k1}(u) \left(\sum_{l} \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l}} n_{ij} \int g_{l0}(v) P_{U_i,U_j}(u,v) dv \right) du \\ = \int g_{k1}(u) \left(\sigma^2 n_k \sum_{i \in \mathcal{S}_k} P'_{U_i}(u) \right) du \end{aligned} \tag{4.10}$$

Since this is true for arbitrary g_{k1} we get

$$\sum_{l} \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l}} n_{ij} \int g_{l0}(v) P_{U_i,U_j}(u,v) dv = \sigma^2 n_k \sum_{i \in \mathcal{S}_k} P'_{U_i}(u) \tag{4.11}$$

for all k and u. Separating out the term where $i = j$ in the sum we have

$$\begin{aligned}
& \sum_{i \in \mathcal{S}_k} n_k g_{k0}(u) P_{U_i}(u) + \sum_l \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l \\ i \neq j}} n_{ij} \int g_{l0}(v) P_{U_i,U_j}(u,v) dv \\
= \; & \sigma^2 n_k \sum_{i \in \mathcal{S}_k} P'_{U_i}(u) \qquad (4.12)
\end{aligned}$$

Since the marginal statistics in a band are all the same we have

$$\begin{aligned}
& g_{k0}(u) N_k n_k P_k(u) + \sum_l \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l \\ i \neq j}} n_{ij} \int g_{l0}(v) P_{U_i,U_j}(u,v) dv \\
= \; & \sigma^2 n_k N_k P'_k(u) \qquad (4.13)
\end{aligned}$$

where N_k is the number of elements in the k^{th} band, and P_k denotes the marginal distribution of coefficients in that band. This can be rewritten

$$\begin{aligned}
& g_{k0}(u) + \frac{1}{N_k n_k} \sum_l \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l \\ i \neq j}} n_{ij} \int g_{l0}(v) \frac{P_{U_i,U_j}(u,v)}{P_k(u)} dv \\
= \; & g_{k0}(u) + \frac{1}{N_k n_k} \sum_n \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l \\ i \neq j}} n_{ij} E\{g_{l0}(U_j) | U_i = u\} \\
= \; & \sigma^2 \frac{P'_k(u)}{P_k(u)} \qquad (4.14)
\end{aligned}$$

We can also write this as a fixed point equation

$$\begin{aligned}
& g_{k0}(u) \\
= \; & \sigma^2 \frac{P'_k(u)}{P_k(u)} - \frac{1}{N_k n_k} \sum_l \sum_{\substack{i \in \mathcal{S}_k \\ j \in \mathcal{S}_l \\ i \neq j}} n_{ij} E\{g_{l0}(U_j) | U_i = u\} \qquad (4.15)
\end{aligned}$$

Thus, in general, we see that the optimal nonlinearity will depend on the pairwise statistics of coefficients. Since we are dealing with AWGN, this dependency can only be introduced through the redundancy of the transform or through dependencies introduced by the statistics of the underlying signal, X. If the basis is a complete one, i.e. if $n_{ij} = \delta_{ij}$, then this equation reduces to

$$g_{k0}(u) = \sigma^2 \frac{P_k'(u)}{P_k(u)} \tag{4.16}$$

which is same result as would be obtained by assuming that the coefficients in a subband are iid(since MSE in the transform domain is the same as MSE in the image domain in this case), as in Eq. (3.2). Recall however, that we made no assumption about the statistics of the underlying vector signal, X. We see that even if the coefficients have dependencies introduced by the statistics of X, the optimal marginal denoiser for an complete decomposition is not affected by these dependencies, and, instead, only depends on the marginal statistics of the coefficients.

4.4 Redundancy improves performance

Eq. (4.5) allows us to explain the empirically observed fact [44, 38] that the performance of marginal denoising in orthogonal wavelet bases can be improved by adding redundancy to the transform through cycle spinning or elimination of decimation. We begin by describing these methods of adding redundancy.

In an orthogonal wavelet decomposition W, subbands are calculated by taking the dot product of an image with shifts of a basis vector and then subsampling, or decimating, the subbands so that the transform of the image has the same dimension as the image, and so that the transform is unitary:

$$W^T W = W W^T = I \tag{4.17}$$

One method of introducing redundancy to the transform is to remove the decimation, which gives an undecimated wavelet transform represented by the matrix W^{ud}. Each subband of this transform of the image will be the same size as the original image, which is therefore redundant by the subsampling factor. If we normalize the basis vectors for each subband by the square root of the redundancy of that subband, then the undecimated transform will continue to be a tight frame

$$(W^{ud})^T W^{ud} = I \tag{4.18}$$

The dimension of transform domain will be equal to the product of the dimension of the image domain and the number of subbands.

A cycle-spun decomposition [44], W^c, on the other hand, acts by taking wavelet decompositions of all possible shifts of the image, and dividing by the square root of the dimension of the image domain. To reconstruct, the coefficients are again divided by the square root of the dimension of the image domain, the inverse wavelet transforms are computed for each shift, and the corresponding images are shifted back and added. Note that applying the transform and then inverse transforming is equivalent to taking the wavelet transform of all shifts, inverse transforming these, shifting back and then averaging, which does indeed give back the original image. The dimension of the transform domain will be the square of the dimension of the original image domain, so that this transform is very overcomplete. Note that the transform, besides being overcomplete, will also have repetition of some coefficients. Instead of shifting the image and calculating the wavelet transform, we may equivalently use a redundant basis comprised of all possible shifts of the wavelet basis vectors. We can therefore view the transform, W^c, as a single overcomplete transform. Because each wavelet transform is unitary and we have appropriately normalized the coefficients, the cycle-spun transform will also be a tight frame

$$(W^c)^T W^c = I \tag{4.19}$$

For didactic purposes we will show that using a cycle spun decomposition can give MSE no worse than using the associated orthogonal wavelet decomposition. The result for the undecimated wavelet is very similar.

For W an orthogonal wavelet decomposition, the risk may be written using the unbiased estimate given in Eq. (4.5)

$$E\left\{\sum_k \sum_{i \in \mathcal{S}_k} g_k(U_i)^2 + 2\sigma^2 \sum_k n_k \sum_{i \in \mathcal{S}_k} g_k'(U_i)\right\}. \tag{4.20}$$

The n_k are all identically one in this case, but we leave them in nonetheless. Since both terms are summed over k, each g_k can be independently optimized over the data from the corresponding subband, $\mathcal{S}_k$.

Cycle spinning corresponds to replicating each basis function at N translated positions. Each subband will therefore contain N times as many coefficients, and their magnitudes will be reduced by factor of $\sqrt{N}$, in order to insure that the overcomplete transform is still unitary. As such, the coefficients in each band of the overcomplete transform will have the same marginal statistics as those in the corresponding band of the complete transform, when rescaled by a factor of $\sqrt{N}$.

An unbiased estimate of the MSE for the *complete* decomposition in Eq. (4.20), may therefore be re-written instead in terms of the *cycle-spun* coefficients as

$$E\left\{\sum_k \frac{1}{N}\sum_{i\in\mathcal{S}_k} g_k(\sqrt{N}U_i^c)^2 + 2\sigma^2\sum_k \frac{n_k}{N}\sum_{i\in\mathcal{S}_k} g_k'(\sqrt{N}U_i^c)\right\} \tag{4.21}$$

where the c superscript denotes the cycle-spun coefficients. Defining $h_k(u) = \frac{1}{\sqrt{N}}g_k(\sqrt{N}u)$, and noting that the norms of the cycle-spun basis vectors are a factor of $\sqrt{N}$ less than those of the original orthogonal basis, we can write Eq. (4.21) as the expected value of

$$\sum_k\sum_{i\in\mathcal{S}_k} h_k(U_i^c)^2 + 2\sigma^2\sum_k n_k^c\sum_{i\in\mathcal{S}_k} h_k'(U_i^c) \tag{4.22}$$

where $n_k^c = n_k/N$. Note that if we are using g_k as the marginal function to denoise the coefficients in the wavelet representation, the scaling of the coefficients and the fact that new coefficients in a band have the same marginal statistics, implies that h_k is the marginal function we would apply to the coefficients in the cycle-spun representation.

Last, if W^c is the overcomplete cycle-spun transformation matrix, then $(W^c)^T$ is a projection operator and $|(W^c)^T u|^2 \leq |u|^2$ for any vector u. This implies that

$$\sum_k\sum_{i\in\mathcal{S}_k} h_k(U_i^c)^2 \geq |(W^c)^T h(U^c)|^2 \tag{4.23}$$

where h is the function that applies h_k to each of the bands $\mathcal{S}_k$. Putting this all together, the MSE estimate for the orthogonal case, given by Eq. (4.22), is greater

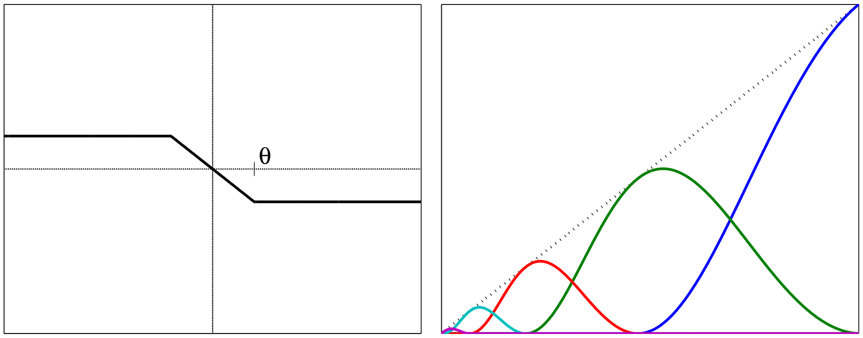


Fig. 4.1: Two families of pointwise estimator functions, $g_\theta(y)$. Left: soft threshold. Right: linear basis of “bump” functions.

than or equal to

$$|(W^c)^T h(U^c)|^2 + 2\sigma^2 \sum_k n_k^c \sum_{i \in \mathcal{S}_k} h_k'(U_i^c). \tag{4.24}$$

Comparing this to Eq. (4.5), we see that it is just the unbiased estimator of the error in using the cycle-spun decomposition to denoise, thus concluding the proof. The result may be extended to undecimated wavelets, in which the number of coefficients in each band will be multiplied by a different factor.

4.5 Simulations

Equation (4.5) may be used to jointly optimize a set of estimators, g_k, to be applied to the subbands $\mathcal{S}_k$. In this section we demonstrate this for two families of estimators, illustrated in Fig. 4.1. The first consists of soft thresholding functions,

Orthonormal wavelet		Undecimated wavelet			
SUREShrink	SUREBumps	SUREShrink		SUREBumps	
		subband	im	subband	im
23.3	23.5	24.2	24.3	24.1	24.5

Table 4.1: Comparison of various denoising methods, expressed as PSNR, applied to the "Barbara" image. In the undecimated cases, we subdivide into cases where the estimator for each subband was optimized separately, and those where the estimators are jointly optimized to minimize MSE in the image domain. Noisy PSNR is 15.2 dB (σ=44.4).

first introduced in Chapter 2. For these functions

$$g_\theta(y) = \begin{cases} -y, & |y| \leq \theta \\ -\text{sgn}(y)\theta, & |y| > \theta \end{cases}$$

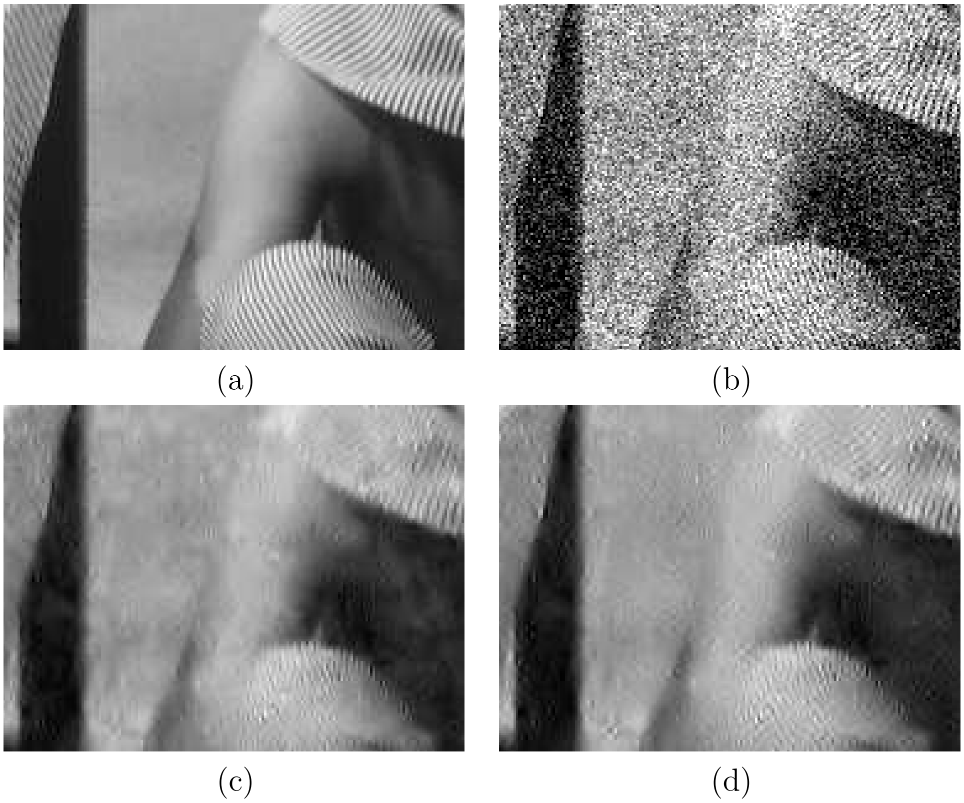
The second is constructed from the "bump" basis of Chapter 2:

$$g_\theta(y) = \sum_k \theta_k b_k(y), \tag{4.25}$$

where

$$b_k(y) = y \ \cos^2\left(\frac{1}{\alpha}\text{sgn}(y)\log_2\left(|y|/\sigma + 1\right) - \frac{k\pi}{2}\right).$$

We use Eq. (4.20) to optimize the selection of thresholds for orthogonal wavelet subbands, a method known as SUREShrink [29]. We use the same equation to optimize subband estimators constructed from the bumps basis, a method which we will refer to as SUREBumps [24] (a similar method, using a different basis, was used with orthogonal wavelets in [23]). As can be seen in Table 4.1, SUREBumps gives some improvement over SUREShrink. Note that in this table, instead of


(a)

(b)

(c)

(d)

Fig. 4.2: Denoising results corresponding to Table 4.1. (**a**) original (cropped); (**b**) noisy; (**c**) Thresholding, optimized for each subband of an undecimated wavelet; (**f**) Bumps, jointly optimized to minimize MSE in the image domain. All PSNRs are listed in Table4.1

using SNR to measure performance, we use PSNR, defined as

$$SNR(dB) = 20\log_{10}(\frac{255^2}{\sum_{k=1}^{N}(\hat{X}_k - X_k)^2}) \tag{4.26}$$

Comparing this with Eq. (2.134), we see that, for a given image, PSNR differs from SNR by a constant. Therefore differences between two PSNRs are equal to the differences between the corresponding SNRs. Next, we use Eq. (4.20) to optimize parameters for the soft-threshold (as in [44]) and the bumps in the bands of an undecimated wavelet transform. Note that the estimator for each subband is chosen to minimize the MSE for that subband, which is suboptimal in the image domain since the transform is overcomplete. This gives improvement for both methods, as expected from the proof of section 4.4. But whereas SUREBumps seemed the superior method for denoising on an orthonormal wavelet decomposition, SUREShrink appears to be superior when applied in the redundant basis. However, if we now use Eq. (4.5) to optimize in the image domain, we see that SUREBumps in the image domain shows significant improvement over SUREBumps optimized for subbands, surpassing the result with thresholding. The PSNR improvements are consistent with visual appearance, as can be seen in example images shown in Fig. 4.2.

We note that while optimizing Eq. (4.5) for bumps in an overcomplete basis is a relatively simple least squares problem, optimizing for the thresholds is a nonconvex optimization problem, and attempts to solve it may get stuck in local minima. As such, it might be possible to improve the result for optimizing thresholding in the image-domain in Table 4.1.

We have examined the behavior of SUREBumps with orthonormal wavelets and undecimated wavelets, over a wide range of noise levels and for a number of images. We did not include thresholding in these comparisons because of the difficulty in optimizing for thresholding in the image domain, and because our experiments indicate that SUREBumps consistently outperforms SUREShrink in an orthonormal wavelet basis. Figure 4.3 shows the improvement in PSNR relative to SUREShrink on the undecimated wavelet, optimized on subbands. As mentioned above (but not shown in the figure), SUREBumps generally outperforms thresholding on an orthonormal wavelet. Using SUREBumps on an undecimated wavelet improves its performance, but as can be seen, this performance generally falls short of the behavior of SUREShrink optimized within subbands of the undecimated wavelet.

However, if we now optimize for image domain MSE, the behavior of SUREBumps on undecimated wavelets significantly outperforms SUREShrink on undecimated wavelets.

4.6 Discussion

We have generalized Stein's Lemma to examine overcomplete representations of the signal, and used this to prove that the expected MSE for marginal denoising in a representation that is made redundant through spatial replication of basis functions (e.g. cycle-spinning, undecimated wavelets) is never larger than in the original non-redundant representation. We have used this extended SURE to design estimators that are applied to subbands of an overcomplete representation, but that are optimized for MSE in the image domain. We have shown simulations demonstrating that optimization of the estimator in the image domain leads to substantial improvement over the suboptimal application of SURE in each the subbands.

The results demonstrate the importance of distinguishing between the method of denoising (e.g., thresholding or bumps), the decomposition to which it is applied (e.g., orthogonal vs. redundant), and the domain in which it is optimized (subbands vs. image). If we were to compare, say, SUREBumps on an orthonormal wavelet and SUREShrink on an undecimated wavelet, we might come to the erroneous conclusion that thresholding is superior to bumps, when in fact the advantage is entirely derived from the overcompleteness of the basis. In addition, while one method of marginal denoising may be superior to another on an orthogonal basis, this benefit may be lost when applying the method to a redundant basis. In future studies, we plan to elaborate on the interaction of the choice of basis with the complexity of denoising method.

The denoising results shown here are meant to illustrate the use of Stein's lemma in the overcomplete case. The methodology is simple, and one can imagine many improvements. In the case of bumps, we have chosen a fixed number of bumps for all bands in all simulations. This could be improved by adapting the dimensionality of the basis both to the noise level and to amount of data in each band. It is also likely that improvement could come from use of an oriented basis (e.g., steerable pyramid [45], complex wavelets [46], curvelets [47]). Finally, the image-domain SURE methodology that we have developed applies to any sort of denoiser that is applied to a transformed version of the data. This can be used to optimize more complex estimators that operate on clusters of coefficients, [48, 42, 23].

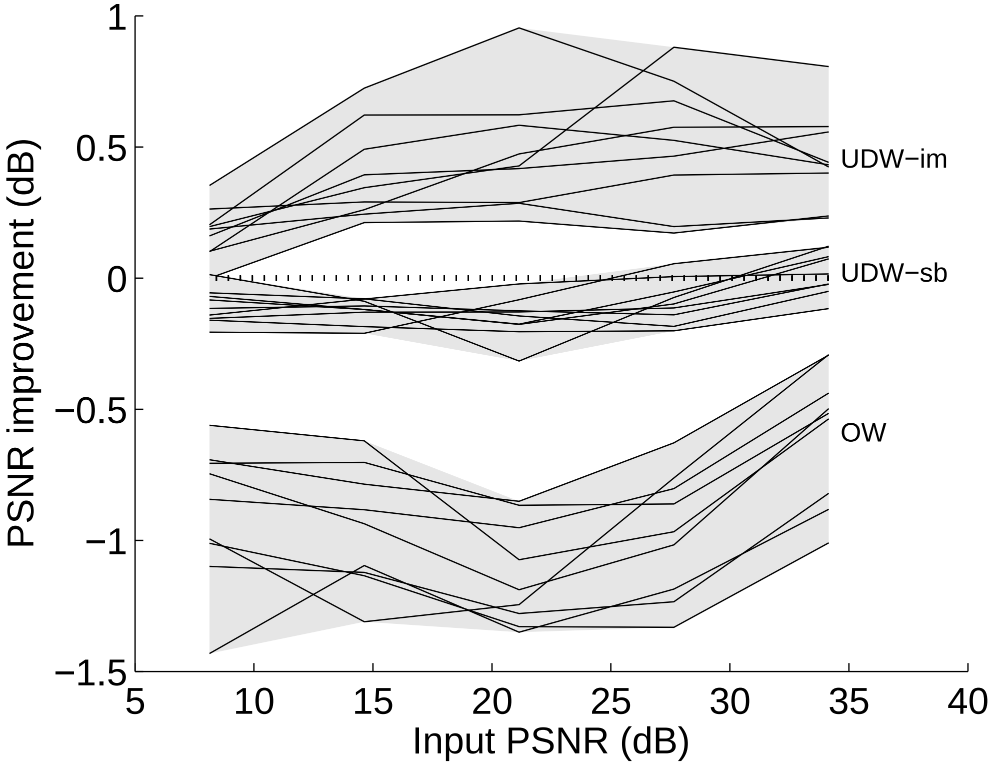


Fig. 4.3: Comparison of denoising results for three estimators. Each group of lines (indicated by gray regions) shows results for one estimator. Each line within a group indicates improvement in PSNR (dB) of the denoised image relative to SUREShrink with undecimated wavelets (optimized within subbands), as a function of input PSNR, for one of eight images. Bottom group: SUREBumps with orthogonal wavelets; middle group: SUREBumps with undecimated wavelets, optimized within subbands; top group: SUREBumps, with undecimated wavelets, optimized for image-domain MSE.

Bibliography

[1] M. P. Sceniak, D. L. Ringach, M. J. Hawken, and R. Shapley, "Contrast's effect on spatial summation by macaque v1 neurons," *Nature Neuroscience*, vol. 2, pp. 733–739, 1999.

[2] C. S. Peskin, D. Tranchina, and D. M. Hull, "How to see in the dark: Photon noise in vision and nuclear medicine," *Annals of the New York Academy of Sciences*, pp. 48–72, 1984.

[3] A. Papoulis, *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, 4th ed., 2001.

[4] A. Angelucci, J. B. Levitt, E. J. S. Walton, J.-M. Hupe, J. Bullier, and J. S. Lund, "Circuits for local and global signal integration in primary visual cortex," *The Journal of Neuroscience*, vol. 22, pp. 8633–8646, 2002.

[5] F. Hayot and D. Tranchina, "Modeling corticofugal feedback and the sensitivity of lateral geniculate neurons to orientation discontinuity," *Visual Neuroscience*, vol. 18, pp. 865–877, 2001.

[6] T. W. Troyer, A. E. Krukowski, N. J. Priebe, and K. D. Miller, "Contrast-invariant orientation tuning in cat visual cortex: Thalamcortical input tun-

ing and correlation-based intracortical connectivity," *The Journal of Neuroscience*, vol. 18, pp. 5908–5927, 1998.

[7] M. Carandini, D. J. Heeger, and W. Senn, "A synaptic explanation of suppression in visual cortex," *The Journal of Neuroscience*, vol. 22, pp. 10053–10065, 2002.

[8] J. D. Victor, "The dynamics of the cat retinal x cell centre," *J. Physiol.*, vol. 386, pp. 219–246, 1987.

[9] D. McLaughlin, R. Shapley, M. Shelley, and D. J. Wielaard, "A neuronal network model of macaque primary visual cortex (V1): Orientation selectivity and dynamics in the input layer 4Cα," *PNAS*, vol. 97, pp. 8087–8092, 2000.

[10] G. Casella, "An introduction to empirical Bayes data analysis," *Amer. Statist.*, vol. 39, pp. 83–87, 1985.

[11] J. S. Maritz and T. Lwin, *Empirical Bayes Methods.* Chapman & Hall, 2nd ed., 1989.

[12] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, vol. 9, pp. 1135–1151, November 1981.

[13] J. Berger, "Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters," *The Annals of Staistics*, vol. 8, pp. 545–571, 1980.

[14] J. T. Hwang, "Improving upon standard estimators in discrete exponential families with applications to poisson and negative binomial cases," *The Annals of Staistics*, vol. 10, pp. 857–867, 1982.

[15] H. Robbins, "An empirical bayes approach to statistics," *Proc. Third Berkley Symposium on Mathematcal Statistics*, vol. 1, pp. 157–163, 1956.

[16] K. Miyasawa, "An empirical bayes estimator of the mean of a normal population," *Bull. Inst. Internat. Statist.*, vol. 38, pp. 181–188, 1961.

[17] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley, 1992.

[18] R. M. Wilcox, "Exponential operators and parameter differentiation in quantum physics," *Journal of Mathematical Physics*, vol. 8, pp. 962–982, 1967.

[19] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. Royal Stat. Soc.*, vol. 36, pp. 99–102, 1974.

[20] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference.* Springer, 2004.

[21] L. Wasserman, *All of Nonparametric Statistics.* Springer, 2006.

[22] D. Donoho, "Denoising by soft-thresholding," *IEEE Trans. Info. Theory*, vol. 43, pp. 613–627, 1995.

[23] F. Luisier, T. Blu, and M. Unser, "SURE-based wavelet thresholding integrating inter-scale dependencies," in *Proc IEEE Int'l Conf on Image Proc*, (Atlanta GA, USA), pp. 1457–1460, October 2006.

[24] M. Raphan and E. P. Simoncelli, "Learning to be Bayesian without supervision," in *Adv. Neural Information Processing Systems (NIPS*06)* (B. Schölkopf, J. Platt, and T. Hofmann, eds.), vol. 19, MIT Press, May 2007.

[25] A. Hyvarinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, pp. 695–709, 2005.

[26] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *Proc 3rd IEEE Int'l Conf on Image Proc*, vol. I, (Lausanne), pp. 379–382, IEEE Sig Proc Society, September 16-19 1996.

[27] M. Raphan and E. P. Simoncelli, "Empirical Bayes least squares estimation without an explicit prior." manuscript in preparation, 2006.

[28] C. R. Loader, "Local likelihood density estimation," *Annals of Statistics*, vol. 24, no. 4, pp. 1602–1618, 1996.

[29] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J American Stat Assoc*, vol. 90, December 1995.

[30] J. P. Rossi, "Digital techniques for reducing television noise," *JSMPTE*, vol. 87, pp. 134–140, 1978.

[31] B. E. Bayer and P. G. Powell, "A method for the digital enhancement of unsharp, grainy photographic images," *Adv in Computer Vision and Im Proc*, vol. 2, pp. 31–88, 1986.

[32] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[33] A. Chambolle, R. A. DeVore, and B. J. L. N. Lee, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Proc.*, vol. 7, pp. 319–335, March 1998.

[34] D. Leporini and J. C. Pesquet, "Multiscale regularization in Besov spaces," in *31st Asilomar Conf on Signals, Systems and Computers*, (Pacific Grove, CA), November 1998.

[35] H. A. Chipman, E. D. Kolaczyk, and R. M. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J American Statistical Assoc*, vol. 92, no. 440, pp. 1413–1421, 1997.

[36] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J R Stat Soc B*, vol. 60, pp. 725–749, 1998.

[37] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *Journal of the American Statistical Association*, vol. 93, pp. 173–179, 1998.

[38] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models* (P. Müller and B. Vidakovic, eds.), ch. 18, pp. 291–308, New York: Springer-Verlag, 1999.

[39] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using a generalized Gaussian and complexity priors," *IEEE Trans. Info. Theory*, vol. 45, pp. 909–919, 1999.

[40] Hyvarinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computation*, vol. 11, no. 7, pp. 1739–1768, 1999.

[41] J. Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Proc.*, vol. 11, pp. 670–684, June 2002.

[42] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli, “Image denoising using a scale mixture of Gaussians in the wavelet domain,” *IEEE Trans Image Processing*, vol. 12, pp. 1338–1351, November 2003.

[43] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, “Shiftable multi-scale transforms,” *IEEE Trans Information Theory*, vol. 38, pp. 587–607, March 1992.

[44] R. R. Coifman and D. L. Donoho, “Translation-invariant de-noising,” in *Wavelets and statistics* (A. Antoniadis and G. Oppenheim, eds.), San Diego: Springer-Verlag lecture notes, 1995.

[45] E. Simoncelli and H. Farid, “Steerable wedge filters for local orientation analysis,” *IEEE Trans Image Proc*, vol. 5, pp. 1377–1382, September 1996.

[46] N. Kingsbury, “Complex wavelets for shift invariant analysis and filtering of signals,” *Applied and Computational Harmonic Analysis*, vol. 10, pp. 234–253, May 2001.

[47] E. J. Candès and D. L. Donoho, “Curvelets - a surprisingly effective nonadaptive representation for objects with edges,” in *Curves and Surfaces* (C. Rabut, A. Cohen, and L. L. Schumaker, eds.), (Nashville, TN), pp. 105– V120, Vanderbilt Univ. Press, 2000.

[48] L. Şendur and I. W. Selesnick, “Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency,” *IEEE Trans. Sig. Proc.*, vol. 50, pp. 2744–2756, November 2002.