Beyond thresholding: analysis and improvements for deterministic parameter estimation

Beyond Thresholding: Analysis and Improvements for Deterministic Parameter Estimation

Baris I. Erkmen, *Student Member, IEEE* and Vivek K. Goyal, *Senior Member, IEEE*

Abstract

Hard-threshold estimators are popular in signal processing applications. We provide a detailed study of using hard-threshold estimators for estimating an unknown deterministic signal when additive white Gaussian noise corrupts observations. The analysis, depending heavily on Cramér-Rao bounds, motivates piecewise-linear estimation as a simple improvement to hard thresholding. We compare the performance of two piecewise-linear estimators to a hard-threshold estimator. When either piecewise-linear estimator is optimized for the decay rate of the basis coefficients, its performance is better than the best possible with hard thresholding.

Index Terms

biased estimation, Cramér-Rao bound, hard thresholding, semisoft thresholding, wavelet shrinkage

EDICS Categories: SSP-PARE, SSP-REST, SSP-PERF

This work was supported in part by the Texas Instruments Leadership University Consortium Program.

B. I. Erkmen (email: erkmen@mit.edu) and V. K. Goyal (email: vgoyal@mit.edu) are with the Department of Electrical Engineering and Computer Science and the Research Laboratory of Electronics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

I. INTRODUCTION

Removing noise from signals ("denoising") is a problem central to many engineering disciplines. As summarized nicely by Moulin and Liu [1], most methods fall into at least one of three categories: Bayesian techniques, which assume a probabilistic prior for the unknown signal and minimize an error measure given the observations; minimax techniques, which are designed for good worst-case performance over some broad class of signals; and techniques based on the minimum description length (MDL) principle. In the electrical engineering literature, the Bayesian and MDL approaches are more common than the minimax approach familiar to statisticians.

In many fields—especially image processing and geophysics—the use of the wavelet domain is prominent, regardless of which of the three approaches is undertaken. For Bayesian techniques, the wavelet domain is convenient because it allows low-complexity diagonal [2]–[4] or nearly-diagonal [5] estimators to be used with little loss in performance. Minimax estimation performance is intimately connected with nonlinear approximation [6], so the approximation power of wavelet bases for many classes of signals make them appropriate [7]–[9]. Finally, the success of wavelet-based compression makes wavelet representations suitable for generation of regularization terms in MDL [10], [11]. All three approaches, under appropriate conditions, justify simple hard threshold or soft threshold (shrinkage) estimators.

In this paper we study the classical problem of estimating a signal in the presence of additive white Gaussian noise (AWGN) with the goal of minimizing mean-squared error (MSE). Rather than applying the Bayesian formulation, we cast this as the estimation of a non-random parameter vector. This allows us to explain the performance of hard threshold estimators through bias-variance trade-off and the Cramér-Rao bound (CRB). The (biased) CRB provides more insight than the standard "oracle" bound of [12, Ch. 10]. Shaping the bias and resulting MSE inspires the analysis and optimization of two alternatives to hard thresholding. Both are piecewise-linear functions, and one has been proposed previously as the "semisoft shrinkage" estimator [13]. Our focus is not on the invention of simple estimators, but rather on the fact that the performance of such estimators can be better understood with the analysis presented herein. Furthermore, we show that the degrees of freedom in these estimators can be optimized, given the decay rate of coefficients, to achieve lower average estimation error than that incurred by hard thresholding.

The paper is organized as follows. In Section II we review the definition of bias and estimation error bounds for both unbiased and biased estimators. We then analyze hard-threshold estimators—notably explaining their performance using Cramér-Rao Bounds (CRBs)—in Section III. Inspired by this way of understanding hard-threshold estimators, we analyze two alternative estimators in Section IV. In particular, we show that these estimators can be optimized for the decay rate of the unknown deterministic parameter vector, resulting in uniform improvement over hard thresholding. We also discuss the limiting cases that relate the alternative estimators to hard-threshold estimators. Finally, Section V provides a discussion of the key results that emerge from our analysis.

II. Background on Estimation Error Bounds

A. Estimation in White Gaussian Noise

In this paper we consider non-random signal estimation when the observation is the signal plus white Gaussian noise. In particular, assume that the observed signal is expanded in some basis of our choice, and the $N \in \mathbb{Z}^{+}$ basis coefficients of interest are stacked in a $N \times 1$ vector $\boldsymbol{y} \in \mathbb{R}^{N}$, such that

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^{N}$ and $\boldsymbol{w} \in \mathbb{R}^{N}$ are $N \times 1$ column vectors representing the corresponding signal and noise basis coefficients respectively, when expanded in the same basis.[1] Here, $\boldsymbol{x}$ is deterministic yet unknown, and $\boldsymbol{w}$ is a random, zero-mean ($E[\boldsymbol{w}] = 0$) white Gaussian noise vector with correlation matrix $E[\boldsymbol{w}\boldsymbol{w}^{T}] = \sigma_w^2 I_N$, where I_N is the $N \times N$ identity matrix. Thus $\boldsymbol{y}$ is a Gaussian random vector with mean $\boldsymbol{x}$ and covariance matrix $\sigma_w^2 I_N$; i.e. the probability density function of $\boldsymbol{y}$ is

$$p(\boldsymbol{y};\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{x}, \sigma_w^2 I_N) = \frac{1}{(2\pi\sigma_w^2)^{N/2}} e^{-\|\boldsymbol{y}-\boldsymbol{x}\|^2/(2\sigma_w^2)}. \tag{2}$$

Before embarking on our analysis let us review and establish some notation. An estimator for $\boldsymbol{x}$, denoted by $\widehat{\boldsymbol{x}}(\boldsymbol{y})$, is a deterministic function of $\boldsymbol{y}$ that maps an observation vector in $\mathbb{R}^{N}$ into the parameter space $\mathbb{R}^{N}$.[2] Given an estimator, we define the error as

$$\boldsymbol{e}(\boldsymbol{y}) \equiv \widehat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}, \tag{3}$$

which is a $N \times 1$ random vector. Its mean value is termed the bias of the estimator and is denoted with

$$\boldsymbol{b}(\boldsymbol{x}) \equiv E[\widehat{\boldsymbol{x}}(\boldsymbol{y})] - \boldsymbol{x}. \tag{4}$$

Note that the bias of an estimator is in general a function of the parameters that are being estimated (i.e. $\boldsymbol{x}$), so in general it is not trivial to arbitrarily modify or eliminate the bias of an estimator.

[1] Throughout this paper we are going to assume N is finite, although it can be arbitrarily large.

[2] The observation space and parameter space need not be identical in general.

The performance of an estimator is often assessed by its $N \times N$, positive-definite error correlation matrix

$$\Lambda_e(\boldsymbol{x}) \equiv E[\boldsymbol{e}(\boldsymbol{y})\boldsymbol{e}^T(\boldsymbol{y})]. \tag{5}$$

Note that the ℓ_2-norm of the error vector can be obtained from the error correlation matrix by taking its trace; i.e.

$$\text{mse}(\boldsymbol{x}) \equiv E\left[\|\widehat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}\|^2\right] = \text{tr}\left(\Lambda_e\right). \tag{6}$$

B. Unbiased Estimators

An estimator is *unbiased* if it satisfies $\boldsymbol{b}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$. It is well known in estimation theory [14, Ch. 2] that the Cramér-Rao bound yields a global performance bound for all unbiased estimators as

$$\Lambda_e - I_{\boldsymbol{y}}^{-1}(\boldsymbol{x}) \geq 0, \tag{7}$$

where '$\geq$' indicates that the matrix on the left hand side is positive semi-definite. Here $I_{\boldsymbol{y}}(\boldsymbol{x})$ is the $N \times N$ Fisher Information matrix with elements

$$[I_{\boldsymbol{y}}(\boldsymbol{x})]_{n,m} = -E\left[\tfrac{\partial^2}{\partial x_n \partial x_m} \ln p(\boldsymbol{y};\boldsymbol{x})\right], \tag{8}$$

for $n, m = 1, \ldots, N$. The Cramér-Rao bound is always a lower bound (in the positive semi-definite sense), however it may not be possible to satisfy it with equality. In particular, the left hand side of (7) is equal to 0 if and only if the efficient estimator

$$\widehat{\boldsymbol{x}}(\boldsymbol{y}) = \boldsymbol{x} + I_{\boldsymbol{y}}^{-1}(\boldsymbol{x})\Big(\nabla_{\boldsymbol{x}} \ln\, p(\boldsymbol{y};\boldsymbol{x})\Big)^T, \tag{9}$$

where $\nabla_{\boldsymbol{x}} \equiv [\partial/\partial x_1, \, \ldots, \, \partial/\partial x_N]$, exists; i.e. the right hand-side of (9) must be independent of $\boldsymbol{x}$. This is indeed the case for the AWGN problem studied in this paper. In particular, we find that

$$I_{\boldsymbol{y}}^{-1}(\boldsymbol{x}) = \sigma_w^2 I_N, \tag{10}$$

and $\Lambda_e = I_{\boldsymbol{y}}^{-1}(\boldsymbol{x})$ is satisfied when the maximum-likelihood estimator is used, i.e. when

$$\widehat{\boldsymbol{x}}(\boldsymbol{y}) = \boldsymbol{y}. \tag{11}$$

Equations (10) and (11) imply that the ℓ_2-norm of the error for any unbiased estimator is no less than $N\sigma_w^2$ and this minimum is achieved by the maximum-likelihood estimator, which is not only a linear estimator, but also the trivial identity function. Furthermore, when the maximum-likelihood estimator is employed, the estimation errors for each element in $\boldsymbol{x}$ are statistically independent.

For unbiased estimators, the CRB on the error covariance matrix is *independent* of the basis representation, so in particular a decomposition into a wavelet basis has no advantage over any other basis [14, Ch. 2]. This picture changes significantly however, when we turn our attention to biased estimators.

C. Biased Estimators

The set of all estimators that satisfy $\boldsymbol{b}(\boldsymbol{x}) \neq 0$ for some $\boldsymbol{x}$ constitute the class of biased estimators. The Cramér-Rao bound for biased estimators is given by [14, Ch. 2]

$$\Lambda_e - \boldsymbol{b}(\boldsymbol{x})\boldsymbol{b}^T(\boldsymbol{x}) \\ - \Big(I_N + \nabla_{\boldsymbol{x}}\boldsymbol{b}(\boldsymbol{x})\Big) I_{\boldsymbol{y}}^{-1}(\boldsymbol{x}) \Big(I_N + \nabla_{\boldsymbol{x}}\boldsymbol{b}(\boldsymbol{x})\Big)^T \geq 0, \tag{12}$$

where '$\geq$', once again, denotes the positive semi-definiteness of the matrix on the left hand side. Unfortunately, the Cramér-Rao bound in the biased case is less useful than it is in the unbiased case. In particular, (12) is interpreted as a lower bound for all estimators with bias $\boldsymbol{b}(\boldsymbol{x})$, but it is not guaranteed that multiple estimators can satisfy a given bias function $\boldsymbol{b}(\boldsymbol{x})$ (unless $\boldsymbol{b}(\boldsymbol{x})$ is trivially a constant). Furthermore, in general the bound cannot be satisfied with equality. However, as we shall see shortly, the Cramér-Rao bound can provide valuable insight into the performance of particular estimators.

It is worth emphasizing that the Cramér-Rao bound in (12) depends on only three quantities: The Fisher Information matrix, the bias, and the gradient of the bias in the parameter space. The Fisher Information matrix is a function of the probability density function of the observed data and therefore is fixed for a given problem setup; for example the Fisher Information matrix for the AWGN problem analyzed in this paper is $\sigma_w^{-2} I_N$. Thus the Cramér-Rao bound can be thought as a function of only the latter two quantities.

Because there is no global lower bound on the performance of biased estimators, a general treatment is not possible. Hence, the utility of the biased Cramér-Rao bound is best demonstrated via an example. Due to its popularity in wavelet-based estimation techniques, we shall first focus our attention on hard thresholding estimators.

III. Analysis of Hard Thresholding

A hard thresholding estimator acts on each basis coefficient observation $\{y_n\}_{n=1}^N$ independently and estimates the true value of each signal basis coefficient according to

$$\hat{x}_n^{(\text{ht})}(y_n) = \begin{cases} 0, & \text{if } |y_n| < T; \\ y_n, & \text{if } |y_n| \geq T, \end{cases} \tag{13}$$

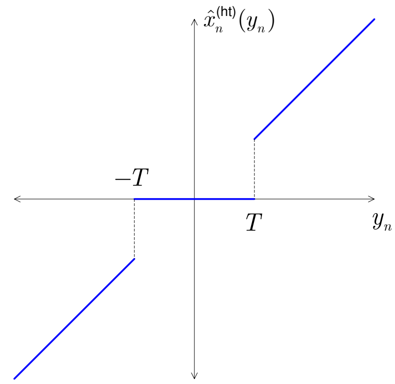
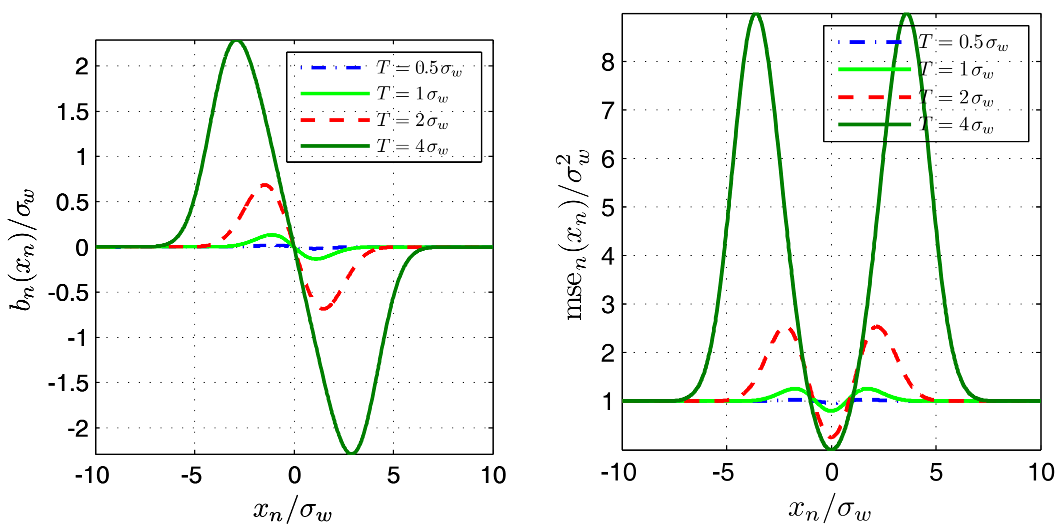


Fig. 1. The input/output relation of a hard thresholding estimator.



(a) Bias **(b)** Mean-square error

Fig. 2. The bias and mean-square error of a hard thresholding estimator as a function of the true value of the parameter. The mean-square error is normalized to the variance of the noise, whereas the bias and the parameter values are normalized to the standard deviation of the noise, such that all axis variables are dimensionless.

where $T > 0$ is called the *threshold*, as shown in Figure 1. Note that, because the estimator acts on each coefficient separately and because the noise on each coefficient is independent in the AWGN estimation problem, we can restrict our analysis to the scalar estimation case with no loss of generality.

The hard thresholding estimator differs from the maximum-likelihood estimator in (11) only when the observation y_n has absolute value less than the threshold T. In this case the hard thresholding estimator estimates the true value of the underlying signal as 0, and this squelching action introduces a bias given by[3]

$$b_n^{(\mathrm{ht})}(x_n) = -\int_{-T}^{T} y\, p_w(y - x_n)\, \mathrm{d}y, \tag{14}$$

where $p_w(y) = \exp(-y^2/2\sigma_w^2)/\sqrt{2\pi\sigma_w^2}$. This bias is plotted in Figure 2(a) for various threshold values. Note that the bias is anti-symmetric; i.e. $b_n^{(\mathrm{ht})}(x_n) = -b_n^{(\mathrm{ht})}(-x_n)$.

The mean-square error is found to be

$$\begin{aligned}\mathrm{mse}_n^{(\mathrm{ht})}(x_n) = \sigma_w^2 - 2x_n b_n^{(\mathrm{ht})}(x_n) \\ - \int_{-T}^{T} y^2\, p_w(y - x_n)\, \mathrm{d}y.\end{aligned} \tag{15}$$

It is worth identifying the terms contributing to this expression. The first term on the right hand side of (15) is the mean-square error obtained when a maximum-likelihood estimator is used. The middle term is always positive and therefore always increases the mean-square error above the minimum mean-square error achievable with an unbiased estimator. On the other hand, the rightmost term is negative, hence it reduces the mean-square error. Therefore, the overall performance of a hard thresholding estimator, relative to the maximum-likelihood estimator, is determined by which of the latter two terms has greater magnitude. Figure 2(b) plots the mean-square error normalized to the variance of the noise, against the true value of the parameter x_n normalized to the standard-deviation of the noise. The figure shows that the performance of a hard thresholding estimator depends on the value of x_n. Recalling that the maximum-likelihood estimator has mean-square error equal to σ_w^2 independent of x_n, we observe from the plots that when x_n is within approximately one standard-deviation of 0, the mean-square error of the hard thresholding estimator is less than that of the maximum-likelihood estimator. On the other hand, if $x_n \gg T$, then the mean-square error approaches that of the maximum-likelihood estimator, because the probability that the noise will push the observation into the regime of thresholding becomes very small. However, in the intermediate regime of x_n, when the true value of the signal is on the same order as the threshold, the mean-square error of the thresholding estimator is worse than the maximum-likelihood estimator, because the noise can push the observation to either side of the threshold, leading to significant errors in the estimate.

We can utilize the Cramér-Rao bound for biased estimators to gain further insight into hard thresholding. Figure 3 compares the normalized mean-square error of a hard thresholding estimator to the unbiased and biased Cramér-Rao bounds, as well as the “optimal oracle” lower bound obtained in [12, Ch. 10]

[3]Because the explicit expressions of the bias and mean-square error are cumbersome, we shall defer them to Appendix A and state here instead concise integral expressions for these quantities.

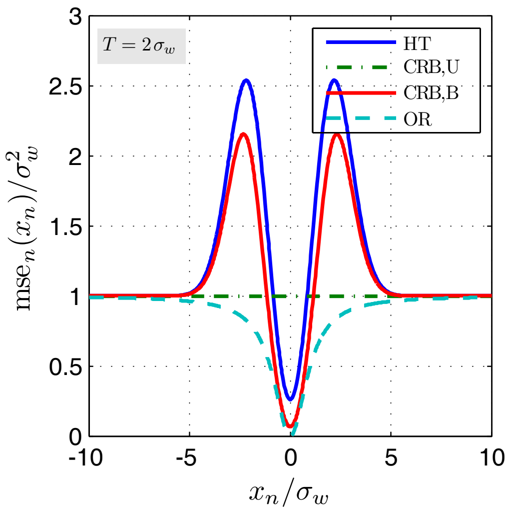


Fig. 3. The mean-square error of a hard thresholding estimator with $T = 2\sigma_w$, together with the unbiased Cramér-Rao bound (CRB,U), the biased Cramér-Rao bound (CRB,B) for all estimators with bias given by $b_n^{(\mathrm{ht})}(x_n)$, and the "optimal oracle" bound (OR) derived in [12, Ch. 10] (see Appendix B). Because the hard thresholding estimator is biased, its mean-square error can go lower than the unbiased Cramér-Rao bound.

and reproduced in Appendix B for convenience. Notice from the figure that the oracle bound is a very weak lower bound which does not capture the oscillatory behavior of the hard thresholding mean-square error. On the other hand, the Cramér-Rao bound—a lower bound for all estimators with bias given by (14)—follows the same oscillatory trend as the hard thresholding mean-square error. In the scalar AWGN case, the biased Cramér-Rao bound simplifies to

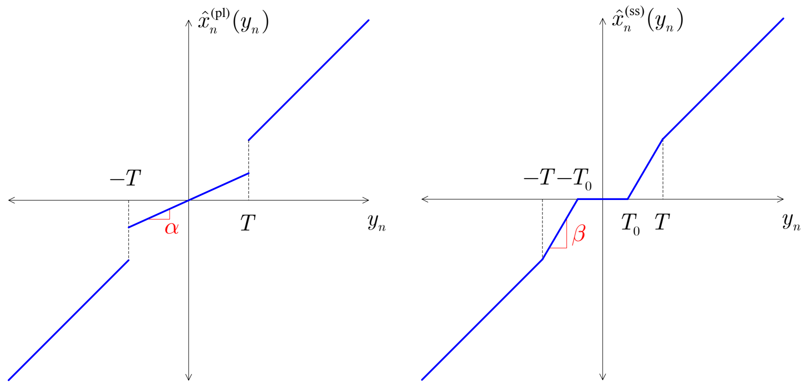
$$\mathrm{mse}_n(x_n) \geq b_n^2(x_n) + \sigma_w^2 \left(1 + \frac{\partial}{\partial x_n} b_n(x_n)\right)^2, \tag{16}$$

which depends *only* on the bias and its derivative with respect to x_n. Hence, the oscillatory behavior of the hard thresholding mean-square error is primarily a consequence of its *bias*. From Figure 2(a) we can verify that the improvement in mean-square error for $x_n \approx 0$, is due to $\partial b_n^{(\mathrm{ht})}(x_n)/\partial x_n < 0$, whereas both $|b_n(x_n)| \gg 0$ and $\partial b_n^{(\mathrm{ht})}(x_n)/\partial x_n > 0$ contribute to the peak in the mean-square error.

We can further exploit the bias dependence of the mean-square error to improve the performance of the estimator on a sequence of coefficients with a given decay rate. We develop this in the next section.

IV. Alternatives to Hard Thresholding

As evidenced in hard thresholding estimators, the bias plays a significant role in determining the mean-square error behavior of an estimator. This dependence has been exploited in previous work to improve



(a) Piecewise-linear **(b)** Semisoft shrinkage

Fig. 4. The input/output relations for two alternative piecewise-linear estimators.

the mean-square error performance—with respect to the unbiased Cramér-Rao bound—over the entire parameter space [15]. Here, we shall instead aim at improving the average error performance of the hard thresholding estimator when applied to multiple basis coefficients with known decay rate.

In this section we consider two piecewise-linear estimators that generalize the hard thresholding estimator. The first is given by

$$\hat{x}_n^{(\mathrm{pl})}(y_n) = \begin{cases} \alpha y_n, & \text{if } |y_n| < T \\ y_n, & \text{if } |y_n| \geq T \end{cases}, \tag{17}$$

where $\alpha \in [0,1]$ and $n = 1,\ldots,N$. From the input/output relation of the piecewise-linear estimator shown in Figure 4(a), it is clear that $\alpha = 0$ corresponds to hard thresholding, whereas $\alpha = 1$ yields the maximum-likelihood estimator. Thus, the slope of the line segment over $y_n \in [-T, T]$ is a degree of freedom in the piecewise-linear estimator that encompasses both the maximum-likelihood estimator and the hard thresholding estimator as special instances.

The bias of this estimator is given by

$$b_n^{(\mathrm{pl})}(x_n) = (1-\alpha) b_n^{(\mathrm{ht})}(x_n), \tag{18}$$

and because $0 \leq (1-\alpha) \leq 1$, the bias and its derivative have smaller magnitude in comparison to hard thresholding, as can be verified from Figure 5(a). The mean-square error is given by

$$\begin{aligned} \mathrm{mse}_n^{(\mathrm{pl})}(x_n) = \sigma_w^2 &- (1-\alpha) 2 x_n b_n^{(\mathrm{ht})}(x_n) \\ &- (1-\alpha^2) \int_{-T}^{T} y^2\, p_w(y - x_n)\, \mathrm{d}y, \end{aligned} \tag{19}$$

and is plotted together with the mean-square error from hard thresholding in Figure 5(b). It is clear from the plot that the piecewise-linear estimator has better worst-case mean-square error than hard thresholding, but this comes at the price of worse best-case mean-square error.

The second estimator we consider here is the “semisoft shrinkage” estimator [13] which replaces the discontinuities in the hard thresholding estimator with a linear segment connecting the left and right limit points, i.e.,

$$\hat{x}_n^{(\mathrm{ss})}(y_n) = \begin{cases} 0, & \text{if } |y_n| < T_0; \\ \beta(y_n - \mathrm{sgn}(y_n)T_0) & \text{if } T_0 \le |y| < T; \\ y_n, & \text{if } |y_n| \ge T, \end{cases} \tag{20}$$

where $\beta \equiv T/(T - T_0) > 1$ denotes the slope of the line segment shown in Figure 4(b), and

$$\mathrm{sgn}(y) = \begin{cases} 1, & y > 0 \\ 0, & y = 0 \\ -1, & y < 0 \end{cases} \tag{21}$$

is the signum function. Note that the shrinkage estimator reduces to hard thresholding when $T = T_0$, but it does not otherwise encompass the piecewise-linear estimator in (17).

The bias and mean-square error expressions for the shrinkage estimator are less tractable than the previous case. Nonetheless, its bias can be expressed as

$$\begin{aligned} b_n^{(\mathrm{ss})}(x_n) = b_n^{(\mathrm{ht})}(x_n) + \beta \int_{T_0}^{T} (y - T_0) \\ \times \big(p_w(y - x_n) - p_w(y + x_n)\big)\,\mathrm{d}y, \end{aligned} \tag{22}$$

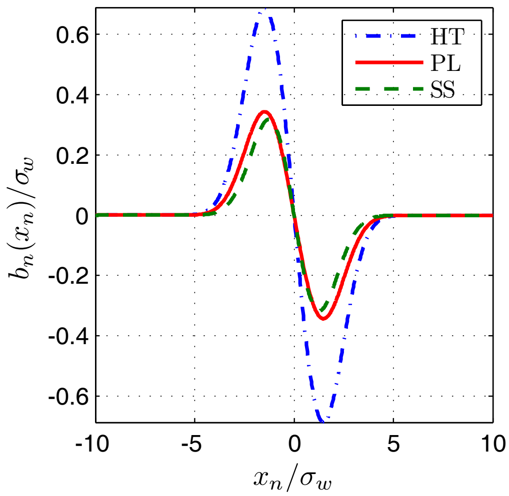
and is plotted in Figure 5(a) together with those obtained from the estimators introduced thus far. The mean-square error of this shrinkage estimator takes on the form

$$\mathrm{mse}_n^{(\mathrm{ss})}(x_n) = \mathrm{mse}_n^{(\mathrm{ht})}(x_n) + f(x_n) + f(-x_n) \tag{23}$$

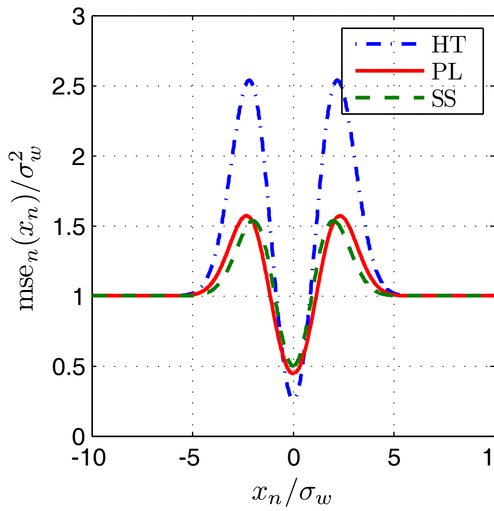
where

$$f(x) \equiv \int_{T_0}^{T} \big(\beta^2 (y - T_0)^2 - 2x\beta(y - T_0)\big) p_w(y - x)\,\mathrm{d}y. \tag{24}$$

This mean-square error is compared to that of the previous estimators in Figure 5(b). It is seen that the shrinkage estimator has similar error to that obtained from (17), but the peak of the oscillation is slightly skewed.



(a) Bias



(b) Mean-square error

Fig. 5. The bias and mean-square error of the two piecewise-linear estimators (PL, SS) and the hard thresholding (HT) estimator are plotted as a function of the true value of the signal coefficient. All axis variables are normalized to be dimensionless. $T = 2\sigma_w$ for all of the plots, $\alpha = 0.5$ is the slope of the piecewise-linear estimator, and $T_0 = 0.5T$ for the shrinkage estimator.

To demonstrate that piecewise-linear estimation improves performance over hard thresholding, we return to the vector-valued AWGN estimation problem and compare the mean-square error per symbol obtained with a hard thresholding estimator and with the two piecewise-linear estimators introduced above. We assume that there are N basis coefficients to be estimated from the same number of observations. Furthermore, because both estimators have symmetric mean-square error as a function of the true value of the parameter, we shall assume with no loss of generality that $x_n \geq 0$ for all $n = 1, \ldots, N$. Finally, we assume that the true values of the coefficients—when sorted—have a decay rate governed by a generalized Gaussian function, i.e. we assume the coefficient sequence is given by

$$x_n = \kappa(p) e^{-[\lambda(n-1)]^p} \tag{25}$$

for $n = 1 \ldots, N$, where $p > 0$ is the decay rate, $\kappa(p) > 0$ is a scaling factor such that the energy of the sequence is equal for all values of p, and the closest integer to $1/\lambda$ is approximately the e^{-1} attenuation point of the coefficients. The signal-to-noise ratio of such a sequence is defined as the ratio of the total signal energy to the total noise energy, i.e.

$$\mathrm{SNR} = \frac{\sum_{n=1}^{N} |x_n|^2}{N\sigma_w^2}. \tag{26}$$

Recall that the estimators act on each coefficient separately and the noise on each coefficient is independent. Therefore, for both estimators, the total mean-square error is equal to the sum of the mean-square error from each coefficient, i.e.

$$E[\|\widehat{\boldsymbol{x}}^{(\mathrm{m})}(\boldsymbol{y})-\boldsymbol{x}\|^2]=\sum_{n=1}^{N}E\Big[\big(\hat{x}_n^{(\mathrm{m})}(y_n)-x_n\big)^2\Big], \tag{27}$$

where method m is ht for hard thresholding, pl for piecewise linear, or ss for semisoft shrinkage. Furthermore, the average mean-square error per symbol is defined as

$$E\big[\|\widehat{\boldsymbol{x}}^{(\mathrm{m})}(\boldsymbol{y})-\boldsymbol{x}\|^2\big]/N. \tag{28}$$

Figure 6 plots the average mean-square error obtained from the three estimators as a function of the decay rate of coefficients when $N=101$, $\lambda=0.04$ and all the degrees of freedom for the estimators are optimized; i.e. the optimization is carried out over T in (13), α and T in (17), and T_0 and T in (20).[4]

The plot shows that the average mean-square error improves for all values of the decay rate when either of the piecewise-linear estimators is utilized in place of the hard thresholding estimator. Let us consider the limiting cases. The histogram for $p=75$ shows that for fast decay rates the coefficients are clustered into two groups: a significant number of the coefficients are very close to 0, while the remaining coefficients are grouped at a large value determined by the SNR of the signal. Consequently, very few coefficients are in the intermediate region. Such a histogram is ideal for hard thresholding because the optimum threshold aligns the region with significant mean-square error in the gap between the two sets of coefficients. Thus, the error incurred per large coefficient becomes σ_w^2 (the maximum-likelihood estimator limit), whereas coefficients that are approximately 0 yield error that is a fraction of σ_w^2. In this regime, the optimal slope α in (17) approaches 0, while the optimal threshold equals that of hard thresholding. Hence the mean-square error from the two estimators converge for $p \gg 1$. On the other hand, the smoother transition at the thresholding boundaries in the semisoft shrinkage estimator reduces the error incurred from the few coefficients that fall within the intermediate region, without significant impact on the errors incurred from the two main clusters of coefficients. Hence, for $p \gg 1$ the shrinkage estimator slightly outperforms the other two estimators.

In the opposite limiting case, the slow decay rate implies that the coefficients will be more spread out over the parameter space, as evidenced by the histogram for $p=1$. Consequently, the error incurred from the coefficients with intermediate values becomes prohibitively large when either the hard thresholding or

[4]All coefficient sequences (each with different p) were normalized to the same energy (to attain identical SNR in all sequences), where the normalization constant was chosen such that the largest coefficient over all p was $10\sigma_w$. Then, the optimization was performed numerically (using the analytical expressions given in Appendix A), given the constraints $\alpha \in [0,1]$ and $T \geq T_0 \geq 0$.

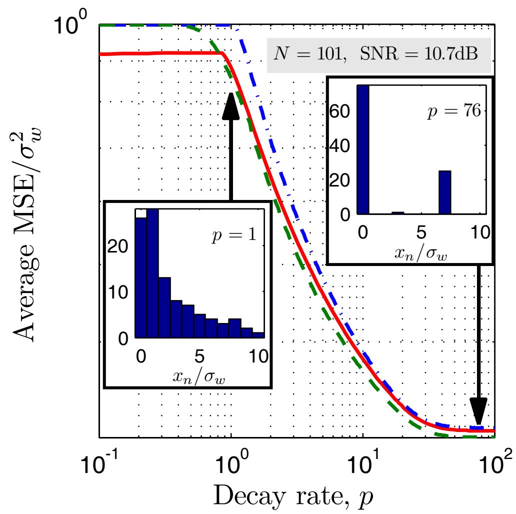
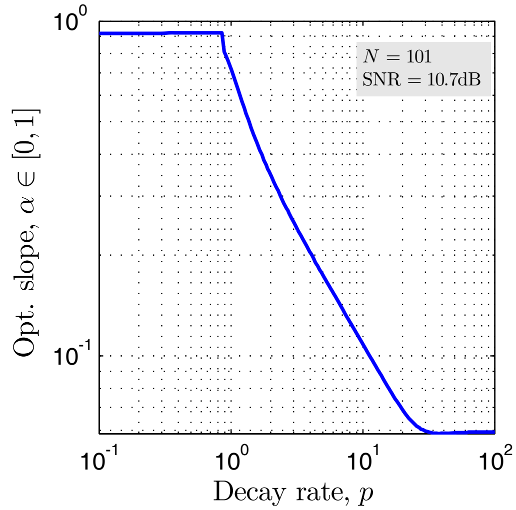


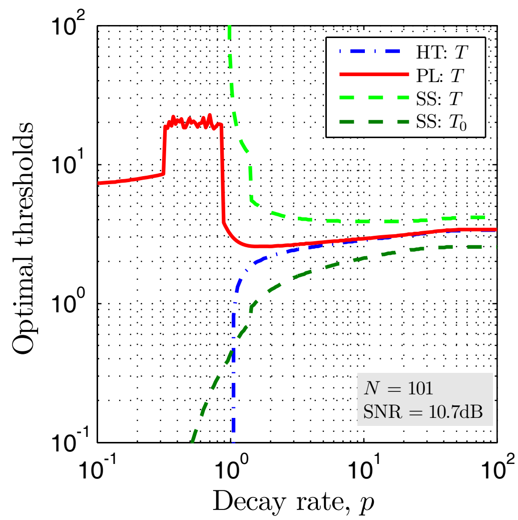
Fig. 6. Mean-square error per symbol of the three estimators analyzed herein. The dash-dotted (blue) curve is the mean-square error of hard thresholding estimators, the solid (red) curve is that of piecewise-linear estimators of form (17), and the dashed (green) curve is the mean-square error of semisoft shrinkage estimators. Histograms refer to signal coefficients with decay rate $p = 1$ and $p = 75$ respectively. Optimization over threshold values and slope is performed numerically from analytic expressions (see footnote 4 for details). Parameter values are $N = 101$, $\lambda = 0.04$ and SNR=10.7dB.

the shrinkage estimator is utilized. Hence, the optimal threshold parameters for both of these estimators, when $p \ll 1$, leads to the maximum-likelihood estimator (i.e. $T = 0$ for hard thresholding and $T_0 = 0$ for shrinkage). On the other hand, the optimal values for the piecewise-linear estimator in (17) turns out to be a large threshold value combined with a slope slightly smaller than unity. Therefore, an estimator that is linear over a significant portion of the coefficients, but with a more conservative slope than the maximum-likelihood estimator, reduces the average mean-square error in comparison to the that obtained from the maximum-likelihood estimator. This result is not entirely unexpected, as it has been shown previously that a (biased) linear estimator with slope less than unity performs better than the maximum-likelihood estimator over the entire parameter space [15]. In the case considered herein, the slow decay rate implies that the coefficients will be spread out over the parameter space, thus the estimator that minimizes the average mean-square error per coefficient must perform well over a large subspace of the parameter space, and this is consistent with the optimization criterion considered in [15].

The optimal values of the degrees of freedom for all three estimators are plotted in Figure 7 as a function of the decay rate p.



(a) Slope



(b) Thresholds

Fig. 7. The optimum slope of the piecewise-linear estimator and the optimal threshold values of all estimators are plotted as a function of the decay rate of the signal coefficients. Values are determined numerically from analytical expressions given in Appendix A. 'HT' denotes the hard thresholding estimator, 'PL' is the piecewise-linear estimator in (17), and 'SS' denotes the semisoft shrinkage estimator. Parameter values are $N = 101$, $\lambda = 0.04$ and SNR=10.7dB.

V. Discussion

In this paper we have provided an estimation theoretic study of non-random signal estimation in order to deepen our understanding of some of the common results encountered in wavelet-based estimation techniques. We focused on the problem of estimating the basis expansion coefficients of a signal when the coefficients are corrupted with additive white Gaussian noise.

The main results developed in this paper can be summarized as follows. The mean-square error lower bound that applies to unbiased estimators indicates that linear estimation is optimal, and furthermore, the basis choice for decomposing the signal does not affect optimal performance. This is, of course, an expected result. Optimal processing of Gaussian random vectors is always linear. Furthermore, if one obtains an optimal solution in one basis, a non-singular transformation of the coordinate-system does not affect the minimum achievable mean-square error, because the optimal estimator in the new basis will simply invert the transformation and apply the former estimator to achieve the same minimum.

The results for biased estimators however differ notably from those for unbiased estimators. Our analysis of hard thresholding as an example of a biased estimator shows that biased estimators are not constrained by the unbiased version of the Cramér-Rao bound. Furthermore, the extension of this bound for biased estimators does not yield achievable lower bounds on performance. Hence, optimality arguments for biased estimators are inevitably more heuristic. Our analysis for hard thresholding demonstrated that basis representation indeed does affect performance when such an estimator is used, as the basis coefficients must be well separated into values that are very close to zero (approximately within one standard-deviation of the noise) and values that are large (significantly larger than the noise standard deviation and the threshold value) to obtain mean-square error smaller than the error of a maximum-likelihood estimator. In other words, the decay rate of the sorted basis coefficients must be fast. Therefore, for the class of signals that have fast-decaying wavelet coefficients, wavelet-basis decompositions in conjunction with hard thresholding will be effective in denoising the observed signal.

Nevertheless, the Cramér-Rao bound for biased estimators provides additional information on how the bias of an estimator affects mean-square error performance. In particular, through our analysis of this bound for the hard thresholding estimator we motivated piecewise-linear estimators and subsequently demonstrated that they achieve smaller average mean-square error when the decay rate of the coefficients are governed by a generalized Gaussian distribution.

In summary, although prior literature provides abundant analysis on the reduction of mean-square error by utilizing wavelet basis expansions and thresholding estimators, additional insight can be obtained by connecting the recent advances in wavelet-based techniques with more traditional estimation theoretic analysis. In this paper we have provided such a connection through non-random signal estimation theory for the AWGN problem, and we have used our analysis to demonstrate that piecewise-linear estimators improve the average mean-square error attained via hard thresholding.

Appendix A

Analytical Expressions for Bias and Mean-Square Error in Hard Thresholding

Let us define

$$x_S = (x_n + T)/(\sqrt{2}\,\sigma_w), \tag{29}$$

$$x_D = (x_n - T)/(\sqrt{2}\,\sigma_w). \tag{30}$$

Evaluating the bias for the hard thresholding estimator from (14) gives

$$\begin{aligned} b_n^{(\mathrm{ht})}(x_n)/\sigma_w = (2\pi)^{-1/2} &\left(e^{-x_D^2} - e^{-x_S^2}\right) \\ &- (x_n/\sigma_w)\big(Q(x_D) - Q(x_S)\big), \end{aligned} \tag{31}$$

where

$$Q(x) \equiv \pi^{-1/2} \int_x^\infty e^{-t^2} \mathrm{d}t. \tag{32}$$

The mean-square error, on the other hand, is obtained from (15) as

$$\begin{aligned} \mathrm{mse}_n^{(\mathrm{ht})}(x_n)/\sigma_w^2 = 1 &+ (x_n/\sigma_w)^2\big(Q(x_D) - Q(x_S)\big) \\ &+ \frac{1}{2}\Big(\mathrm{sgn}(x_D)\Gamma_{\mathrm{inc}}\big(x_D^2, 3/2\big) - \mathrm{sgn}(x_S)\Gamma_{\mathrm{inc}}\big(x_S^2, 3/2\big)\Big), \end{aligned} \tag{33}$$

where $\mathrm{sgn}(x)$ is defined in (21), and

$$\Gamma_{\mathrm{inc}}\big(x, 3/2\big) \equiv 2/\sqrt{\pi} \int_0^x t^{1/2} e^{-t} \mathrm{d}t \tag{34}$$

is the Gamma distribution of order $3/2$.

To provide analytical expressions for the shrinkage estimator we must define two new dimensionless variables,

$$\xi_S = (x_n + T_0)/(\sqrt{2}\,\sigma_w), \tag{35}$$

$$\xi_D = (x_n - T_0)/(\sqrt{2}\,\sigma_w). \tag{36}$$

Now the bias is given by

$$\begin{aligned} b_n^{\text{(ss)}}(x_n) &= \beta b_n^{\text{(ht)}}(x_n;T_0) - (\beta-1) b_n^{\text{(ht)}}(x_n;T) \\ &\quad - \beta T_0 \big(Q(x_D) - Q(\xi_D) + Q(x_S) - Q(\xi_S)\big), \end{aligned} \tag{37}$$

where the second argument in $b_n^{\text{(ht)}}$ indicates the threshold value. The mean-square error expression (23) depends on (33) and the function

$$\begin{aligned} & f(x)/\beta^2\sigma_w^2 = \\ & \Big(x_D(1-2T_0/T)e^{-x_D^2} - \big(\xi_S - \sqrt{2}T_0 x/(\sigma_w T)\big)e^{-\xi_D^2}\Big)/\sqrt{\pi} \\ & + \Big(1 - 2\xi_D\big(\xi_S - \sqrt{2}T_0 x/(\sigma_w T)\big)\Big)\big(Q(x_D) - Q(\xi_D)\big). \end{aligned} \tag{38}$$

Appendix B
"Optimal Oracle" Bound

We simply reproduce the derivation in [12, Ch. 10]. Consider a scalar estimator of the form

$$\hat{x}(y) = ay, \tag{39}$$

where $a \in \mathbb{R}$ is deterministic, for the scalar AWGN estimation problem. Then

$$E\big[(\hat{x}(y) - x)^2\big] = a^2\sigma_w^2 + (1-a)^2 x^2. \tag{40}$$

Differentiating this expression with respect to a and setting it to 0, we find the value that minimizes the mean-square error as

$$a_{\text{opt}} = \frac{x^2}{x^2 + \sigma_w^2}, \tag{41}$$

and the minimum mean-square error is

$$\min_{a\in\mathbb{R}} E[(ay - x)^2] = \frac{\sigma_w^2 x^2}{x^2 + \sigma_w^2}. \tag{42}$$

Equation (42) is the oracle bound used in [12, Ch. 10] and plotted in Figure 3. Because a_{opt} depends on x (which is unknown), such an estimator is not feasible. Hence this mean-square error is not achievable by any feasible estimator of form given in (39).

REFERENCES

[1] P. Moulin and J. Liu, “Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors,” *IEEE Trans. Inform. Theory*, vol. 45, no. 3, pp. 909–919, Apr. 1999.

[2] E. P. Simoncelli and E. H. Adelson, “Noise removal via Bayesian wavelet coring,” in *Proc. IEEE Int. Conf. Image Process.*, vol. I, Lausanne, Switzerland, Sep. 1996, pp. 379–382.

[3] P. Moulin and J. Liu, “Adaptive Bayesian wavelet shrinkage,” *J. Amer. Stat. Assoc.*, vol. 92, no. 440, pp. 1413–1421, Dec. 1997.

[4] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, “Image denoising using scale mixtures of Gaussians in the wavelet domain,” *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[5] L. Şendur and I. W. Selesnick, “Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency,” *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2744–2756, Nov. 2002.

[6] D. L. Donoho, “Unconditional bases are optimal bases for data compression and for statistical estimation,” *Appl. Comput. Harm. Anal.*, vol. 1, no. 1, pp. 100–115, Dec. 1993.

[7] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, “Image coding using wavelet transform,” *IEEE Trans. Image Process.*, vol. 1, no. 2, pp. 205–220, Apr. 1992.

[8] R. A. DeVore, B. Jawerth, and B. J. Lucier, “Image compression through wavelet transform coding,” *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 719–746, Mar. 1992.

[9] R. A. DeVore, “Nonlinear approximation,” *Acta Numerica*, pp. 51–150, 1998.

[10] S. G. Chang, B. Yu, and M. Vetterli, “Adaptive wavelet thresholding for image denoising and compression,” *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000.

[11] M. Hansen and B. Yu, “Wavelet thresholding via MDL for natural images,” *IEEE Trans. Inform. Theory*, vol. 46, no. 5, pp. 1778–1788, Aug. 2000.

[12] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Academic Press, 1999.

[13] H.-Y. Gao and A. G. Bruce, “WaveShrink and semisoft shrinkage,” StatSci Division of MathSoft, Inc., Seattle, WA, Research Report 39, Sep. 1995.

[14] H. L. Van Trees, *Detection, Estimation and Modulation Theory, Part I*. New York, NY: Wiley, 2001.

[15] Y. C. Eldar, “Uniformly improving the Crameér-Rao bound and maximum-likelihood estimation,” *IEEE Trans. Signal Process.*, vol. 54, no. 8, pp. 2943–2956, Aug. 2006.