To Appear in: Proc. ICIP-07, San Antonio, TX, Oct 2007.

OPTIMAL DENOISING IN REDUNDANT BASES

Martin Raphan

Courant Inst. of Mathematical Sciences
New York University

Eero P. Simoncelli

Howard Hughes Medical Institute,
Center for Neural Science, and
Courant Inst. of Mathematical Sciences
New York University

Image denoising methods are often based on estimators chosen to minimize mean squared error (MSE) within the subbands of a multi-scale decomposition. But this does not guarantee optimal MSE performance in the image domain, unless the decomposition is orthonormal. We prove that despite this suboptimality, the expected image-domain MSE resulting from a representation that is made redundant through spatial replication of basis functions (e.g., cycle-spinning) is less than or equal to that resulting from the original non-redundant representation. We also develop an extension of Stein's unbiased risk estimator (SURE) that allows minimization of the image-domain MSE for estimators that operate on subbands of a redundant decomposition. We implement an example, jointly optimizing the parameters of scalar estimators applied to each subband of an overcomplete representation, and demonstrate substantial MSE improvement over the suboptimal application of SURE within individual subbands.

Index Terms— denoising, Bayes least squares, SURE, overcomplete, redundant, translation invariance, cycle spinning

1. INTRODUCTION

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) [1] may be used to select the estimator that minimizes the mean squared error (MSE) [2]. The most popular transforms are multi-scale decompositions, and within this family, empirical evidence indicates that redundant representations are more effective than orthonormal representations [3, 4, 5]. This fact is somewhat mysterious since the estimators are usually optimized for MSE in the transform domain, which, for an overcomplete basis, is not the same as the MSE in the image domain.

In this paper we extend the SURE methodology to 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 cycle-spinning or elimination of decimation is guaranteed to be no worse in MSE (and is in practice typically better) 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.

2. STEIN'S LEMMA FOR OVERCOMPLETE BASES

Given a noisy image Y, we wish to compute an estimate of the form

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

by selecting $g \in \mathcal{G}$ that minimizes the expected squared error:

$$g_{\text{opt}} = \arg\min_{g\in\mathcal{G}} E\left\{|X - (Y + g(Y))|^2\right\}$$

where X is the original (clean) image [1]. Stein's Lemma [1] implies that, for additive Gaussian noise, the MSE may be rewritten without reference to X:

$$g_{\text{opt}} = \arg\min_{g\in\mathcal{G}} E\left\{|g(Y)|^2 + 2\sigma^2(\nabla \cdot g)(Y)\right\}. \quad (1)$$

Given a single vector-valued sample Y (e.g., an image), g_{opt} can be approximated by minimizing the expression in braces, which is (up to an additive constant) Stein's unbiased risk estimator (SURE) [1]. This result can be generalized to non-Gaussian noise, as well as a variety of non-additive corruption processes [6].

It is common to apply estimators to a linearly transformed version of the image, in which the statistical properties are simplified. Stein's Lemma is readily extended to this situation. Suppose we have a family of estimators $\{u + g_u(u) :$

[1] From a frequentist perspective, X is fixed but unknown, and the expectation is taken over Y. One may also consider both X and Y as random, taking the expected value over both.

$g_u \in \mathcal{G}_U\}$ which act on $U = WY$, a transformed version of the image Y. Here W can be a complete or overcomplete linear transformation (an m by n matrix, $m \geq n$, where n is the dimension of image space), that has a left inverse $W^\dagger$. The estimate is computed by transforming with W, applying g_u, and inverse transforming with $W^\dagger$:

$$\begin{aligned} \hat{x}(Y) &= W^\dagger(WY + g_u(WY)) \\ &= Y + W^\dagger g_u(WY). \end{aligned} \tag{2}$$

To optimize this for MSE, we replace $g(Y)$ by $W^\dagger g_u(WY)$ in Eq. (1), and after a bit of calculus obtain:

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

where $\mathrm{tr}(\cdot)$ indicates the trace of a matrix. As before, the expression in braces is an unbiased estimate of risk and can be optimized even over a single sample of Y. For simplicity, in what follows we will assume that the transform is a tight frame, for which $W^\dagger = W^T$. This includes orthonormal, cycle-spun and undecimated wavelet transforms, as well as other overcomplete transforms such as the steerable pyramid [3].

3. POINT ESTIMATORS ON SUBBANDS

Suppose now that g_u consists of functions g_i that operate pointwise on (i.e., on each element of) U. The unbiased risk estimator becomes

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

where n_{ii} are the diagonal elements of WW^T (the squared norms of the basis functions). Typically, the transform coefficients are partitioned into subbands $\{\mathcal{S}_k; k = 1, 2, \ldots K\}$, corresponding to shifted versions of the same basis function, all of which are are assumed to have the same marginal statistical properties. In this case, the same estimator will be applied to all coefficients within a subband, and the unbiased risk estimator becomes

$$|W^T g(U)|^2 + 2\sigma^2 \sum_k n_k \sum_{i \in \mathcal{S}_k} g_k'(U_i) \tag{3}$$

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 $\{g_k\}_{k=1}^K$ so as to minimize the MSE in the image domain.

4. REDUNDANCY IMPROVES PERFORMANCE

Equation (3) allows us to explain why the performance of marginal denoising in orthonormal wavelet bases can be improved by adding redundancy to the transform through cycle spinning or elimination of decimation [4, 5] For didactic purposes, we will consider cycle spinning. For W an orthonormal wavelet decomposition, the unbiased estimate of the risk given in Eq. (3) may be written

$$\sum_k \sum_{i \in \mathcal{S}_k} g_k(U_i)^2 + 2\sigma^2 \sum_k n_k \sum_{i \in \mathcal{S}_k} g_k'(U_i). \tag{4}$$

The n_k are all identically one in this case. 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 contain N times as many coefficients, relative to to the orthonormal representation, each reduced by factor of $\sqrt{N}$. As such, the coefficients in each band will have the same marginal statistics[2], when rescaled by a factor of $\sqrt{N}$. We can thus rewrite Eq. (4), the unbiased estimator of risk for the orthonormal transform, in terms of the *cycle-spun* coefficients, U_i^c:

$$\sum_k \frac{1}{N} \sum_{i \in \mathcal{S}_k} g_k(\sqrt{N} U_i^c)^2 + 2\sigma^2 \sum_k \frac{n_k}{N} \sum_{i \in \mathcal{S}_k} g_k'(\sqrt{N} U_i^c). \tag{5}$$

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 assumption that the redundant coefficients in a band have the same marginal statistics as the original orthonormal coefficients implies that

$$h_k(u) = \frac{1}{\sqrt{N}} g_k(\sqrt{N} u)$$

is the marginal function that should be applied to the coefficients in the cycle-spun representation. Equation (5) may thus be rewritten as:

$$\sum_k \sum_{i \in \mathcal{S}_k} h_k(U_i^c)^2 + 2\sigma^2 \sum_k n_k^c \sum_{i \in \mathcal{S}_k} h_k'(U_i^c) \tag{6}$$

where $n_k^c = n_k/N$, since the norms of the cycle-spun basis vectors are a factor of $\sqrt{N}$ less than those of the original orthonormal basis.

Now we wish to compare this with the unbiased estimate of the risk in denoising the cycle spun transform, which by Eq. (3) is

$$|(W^c)^T h(U^c)|^2 + 2\sigma^2 \sum_k n_k^c \sum_{i \in \mathcal{S}_k} h_k'(U_i^c). \tag{7}$$

If W^c is the overcomplete cycle-spun transformation matrix, then $(W^c)^T$ is a projection operator. This means that

$$|(W^c)^T u|^2 \leq |u|^2$$

[2] We have taken the view that coefficients of the original, noiseless subband are drawn from a random distribution. If, instead, we adopt a frequentist view that the image is fixed but unknown, we must assume that the histogram of coefficient values for an orthonormal subband is a good approximation to the histogram for the corresponding subband of the cycle-spun representation.

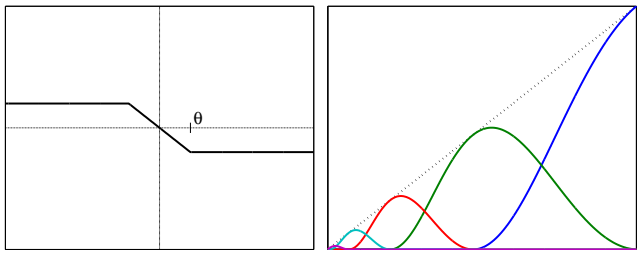

Fig. 1. Two families of pointwise estimator functions, $g_\theta(y)$. Left: soft threshold. Right: linear basis of "bump" functions.

Orthonormal wavelet		Undecimated wavelet			
thresh	bumps	thresh		bumps	
		subband	im	subband	im
23.3	23.5	24.2	24.3	24.1	24.5

Table 1. Comparison of various denoising methods, expressed as PSNR, applied to the "Barbara" image. In the undecimated cases, we subdivide into two cases: one with the estimator independently optimized to minimize the MSE of each subband, and the other with the estimators jointly optimized to minimize MSE in the image domain. Noisy PSNR is 15.2 dB (σ=44.4).

for any vector u, which in turn implies that

$$\sum_k \sum_{i \in \mathcal{S}_k} h_k(U_i^c)^2 \geq |(W^c)^T h(U^c)|^2 \tag{8}$$

where h is the function that applies h_k to each of the bands $\mathcal{S}_k$. Comparing Eq. (6) and Eq. (7), we see that the MSE estimate for the orthonormal case is always greater than or equal to that for the cycle-spun case. The result may be extended to undecimated wavelets, in which the number of coefficients in each band will be multiplied by a different factor.

5. SIMULATIONS

Equation (3) 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 will discuss two families of estimators, illustrated in Fig. 1. The first consists of soft thresholding functions:

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

The second is constructed from a basis of "bump" functions:

$$g_\theta(y) = \sum_k \theta_k b_k(y), \tag{9}$$

where

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

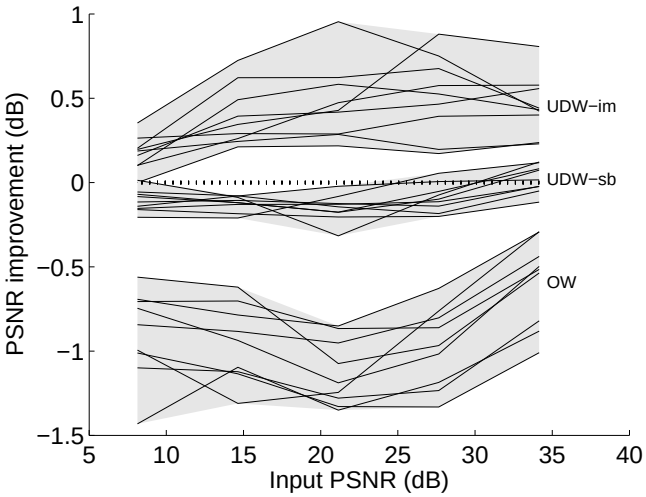

Fig. 2. Comparison of denoising results for three estimators. Each group of lines (indicated by gray regions) shows results for one estimator. Each line within a group indicates improvement in PSNR (dB) of the denoised image relative to SUREshrink with undecimated wavelets (optimized within subbands[4]), as a function of input PSNR, for one of eight images. Bottom group: SUREbumps with orthonormal wavelets; middle group: SUREbumps with undecimated wavelets, optimized within subbands; top group: SUREbumps, with undecimated wavelets, optimized for image-domain MSE.

We used Eq. (4) to optimize the selection of soft-thresholds for orthonormal wavelet subbands, a method known as SUREshrink [7]. We used the same equation to optimize estimators constructed from the bumps basis, a method which we will refer to as SUREbumps (a similar method, using a different basis, was used with orthonormal wavelets in [8]). As can be seen in Table 1, SUREbumps gives some improvement over SUREshrink in an orthonormal basis. Next, we used Eq. (4) to optimize parameters for the soft-threshold (as in [4]) and the bumps in an undecimated wavelet transform. The estimator for each subband was chosen to minimize the MSE for that subband, producing a suboptimal result in the image domain (since the transform is overcomplete). As expected from the proof of section 4, this gives improvement for both methods. But whereas SUREbumps is the superior method for denoising on an orthonormal wavelet decomposition, SUREshrink is superior when optimized on subbands of the redundant basis. Finally, we used Eq. (3) to optimize parameters for both methods in the image domain. This produces improvement in both methods, but the improvement for SUREbumps is more substantial, and it now surpasses the thresholding results. We note that while optimizing Eq. (3) for bumps in an overcomplete basis is a relatively simple least squares problem, optimizing for the thresholds is a nonconvex optimization problem, and so our solution may represent a local minimum. As such, it might be possible to improve the result for optimizing thresholding in the image-domain in Table 1.

Figure 2 illustrates the performance of these methods over a wide range of noise levels and for a number of images. The graph shows the improvement in PSNR of three SUREbumps estimators (applied to orthonormal wavelets, undecimated wavelets optimized within subbands, and undecimated wavelets optimized in the image domain) relative to the SUREshrink estimator on the undecimated wavelet optimized within subbands. We did not include comparisons to thresholding optimized in the image domain because of the uncertainty in finding the globally optimimum solution, but our experiments indicate that SUREbumps generally outperforms SUREshrink when applied in an orthonormal wavelet basis. As can be seen in the figure, using SUREbumps on an undecimated wavelet improves its performance substantially, compared to the orthonormal wavelet case. This performance generally falls slighly short of the behavior of SUREshrink optimized within subbands of the undecimated wavelet. However, when optimized for image domain MSE, the behavior of SUREbumps on undecimated wavelets consistently and significantly outperforms SUREshrink on undecimated wavelets.

6. DISCUSSION

We have generalized Stein's Lemma to examine overcomplete representations of a signal, and used this generalization 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 substantial improvement over the suboptimal application of SURE in each the subbands.

The results illustrate the importance of distinguishing between the method of denoising (e.g., thresholding or bumps), the decomposition to which it is applied (e.g., orthonormal 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 always 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 orthonormal basis, this benefit may be lost when applying the method to a redundant basis.

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 [3], complex wavelets [9], curvelets [10]). Finally, the image-domain SURE methodology that we have developed is relevant for any estimator that is applied to a transformed version of the data. We are currently pursuing the optimization of more complex estimators that operate on clusters of coefficients, [11, 12].

7. REFERENCES

[1] C M. Stein, "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, November 1981.

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

[3] E P Simoncelli, W T Freeman, E H Adelson, and D J Heeger, "Shiftable multi-scale transforms," *IEEE Trans Information Theory*, vol. 38, no. 2, pp. 587–607, March 1992.

[4] R R Coifman and D L Donoho, "Translation-invariant denoising," in *Wavelets and statistics*, A Antoniadis and G Oppenheim, Eds. Springer-Verlag lecture notes, San Diego, 1995.

[5] 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., chapter 18, pp. 291–308. Springer-Verlag, New York, 1999.

[6] 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. May 2007, vol. 19, MIT Press.

[7] D Donoho and I Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J American Stat Assoc*, vol. 90, no. 432, December 1995.

[8] 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, October 2006, pp. 1457–1460.

[9] N Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 234–253, May 2001.

[10] 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, 2000, pp. 105– V120, Vanderbilt Univ. Press.

[11] L Şendur and I W Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Sig. Proc.*, vol. 50, no. 11, pp. 2744–2756, November 2002.

[12] 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, no. 11, pp. 1338–1351, November 2003.