Assumptions H1–H3 and H4' are clearly satisfied. Therefore, since the theoretical trajectory matches the theoretical trajectory predicted by (3) for $r = 0.8182$ and $\mu = 0.001$. To determine the steady-state distribution of the parameter estimate errors $\hat{W}_k = r - W_k$, 1 million iterations of the algorithm were run, and the last 500,000 were used to compute the simulated error densities. The plots show both the simulated and theoretical densities. In Fig. 6, $\alpha = -0.8$, and in Fig. 7, $\alpha = 0$, whereas $\alpha = 0.5$ in Fig. 8.

V. CONCLUSIONS

Asymptotic results were derived for the MLMS and other momentum algorithms. The analysis was based on the results of [5]. The effect of the momentum factor on convergence was studied, and expressions for the asymptotic distribution of the parameter estimates were derived.

REFERENCES


OcCam Filters for Stochastic Sources with Application to Digital Images

Balas Natarajan, Konstantinos Konstantinides, and Cormac Herley

Abstract—An OcCam filter employs lossy data compression to separate signal from noise. Previously, it was shown that OcCam filters can filter random noise from deterministic signals. Here, we show that OcCam filters can also separate two stochastic sources, depending on their relative compressibility. We also compare the performance of OcCam filters and wavelet-based denoising on digital images.

I. INTRODUCTION

A practical problem in real signal processing systems is the treatment of noise-corrupted signals. A commonly used noise-removal approach is the Wiener filter, which is a linear filter that weights the spectrum of the signal by an amount that depends on the noise strength at a given frequency. Recently, there have been a number of alternative nonlinear approaches to noise removal, representative of which is the soft-thresholding approach introduced by Carlson et al. [3] and furthered by Donoho and others [5], [6]. In this approach, the authors are with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA (e-mail: balas@hpl.hp.com).

Publisher Item Identifier S 1053-587X(98)02010-8.
the noisy signal is wavelet transformed, the wavelet coefficients are passed through a thresholding nonlinearity, and then, the inverse wavelet transform is applied to the thresholded coefficients. The idea of the approach is that when the wavelet transform is applied to the noisy signal, much of the signal energy is localized at low frequencies, whereas the noise is dispersed across the spectrum. In spirit, this approach is similar to Wiener filtering in that frequencies at which the signal-to-noise ratio (SNR) is low are de-emphasized relative to those where it is high. For soft thresholding, this is done subtractively rather than multiplicatively and in some wavelet rather than Fourier space. Thus, it resembles an older nonlinear technique known as spectral subtraction [2]. Examples of Wiener filtering and soft thresholding are shown in Fig. 1. Observe that the noise amplitude is suppressed most in regions where the signal strength is low: generally the high-frequency regions.

A general framework that allows analysis of both linear and nonlinear noise removal approaches is that of Occam filters, which were introduced by Natarajan [8]. A novelty of the approach is that it is not tied to a frequency domain interpretation of separating signal and noise, but rather exploits the difference in compressibility that it is not tied to a frequency domain interpretation of separating signal and noise, but rather exploits the difference in compressibility between the signal and the noise. Let \( f \) be a signal source that is corrupted by a noise source \( \nu \) in that we observe discrete samples \( f_1 + \nu_1, f_2 + \nu_2, \ldots \). Suppose we feed these noisy samples to a lossy coding scheme, whose allowed loss can be controlled in the mean-squared error sense. Set the allowed error to be the power of the noise, compress the samples, and examine the decompressed signal \( g \). In particular, we compare the decompressed signal \( g \) and the noise-free signal \( f \). Will \( g \) be closer to \( f \) than the noisy signal? That is, is it reasonable to expect that the coding error and the noise will cancel rather than add? Previously [8], it was shown that the noise and the coding error cancel if the the noise \( \nu \) is random, and the signal \( f \) is deterministic, continuous, and finitely supported. The extent of the cancelation depends on the effective of the coding and the sampling rate, with the cancelation tending to completion as increasing numbers of samples of the signal are taken over its finite support. The cancelation is not contingent on the linearity of the coding process and therefore offers a technique for constructing nonlinear filters for random noise that is more general than the wavelet-based schemes. In this correspondence, using rate-distortion theoretic techniques, we extend this result to stochastic signals by showing that the coding error and the noise cancel, with the extent of the cancelation depending on the relative compressibility of the noise and the signal. We also specialize our rather general main result to the case of normally distributed noise and derive a simple inequality linking the SNR's of the output signal, the SNR of the input signal, and the coding rate of the compression method. Finally, we present some experimental results to exemplify our theoretical results. We construct an Occam filter for images based on thresholded singular value decomposition (SVD). While SVD thresholding is well established as a filtering technique in the literature, the threshold value is usually selected by ad hoc means. Our contribution here is to select the threshold value simply in accordance with the theory. In experiments on still images corrupted with synthetic Gaussian noise, we find the SVD-based Occam filter outperforms the wavelet denoising technique of Donoho and others. We also compare the SNR's achieved in the experiments with the SNR's that are predicted similarly to select the threshold value simply in accordance with the theory. In experiments on still images corrupted with synthetic Gaussian noise, we find the SVD-based Occam filter outperforms the wavelet denoising technique of Donoho and others. We also compare the SNR's achieved in the experiments with the SNR's that are predicted by our theory and find that the analytic bounds are asymptotically valid although overly conservative.

The results in this paper appeared in preliminary form in [9].

II. PRELIMINARIES

We now establish our notational conventions. A stochastic source produces an infinite sequence of observations of a random variable.
encoding the sample $f_i$, the coder is given access to the infinite history of
the signal in that it is allowed to see $f_{-1}, f_{-2}, \ldots$. The coder
outputs a sequence of codes that can be decoded by a decoder that
produces a sequence of reals $g_1, g_2, \ldots$, constituting the reconstructed
signal $g$. The decoder is also given access to the infinite past of coded
symbols but must decode each symbol before seeing the next. The
coder may choose to encode several samples $f_i, f_{i+1}, \ldots, f_{i+k}$ as
a single symbol, but $k$ must be finite. The parameter $k$ can be viewed as
the delay of the coder. The loss of the coder is the expected deviation
between $f$ and the reconstructed signal $g$ in a predetermined metric.
Throughout this correspondence, we use the mean-square measure
\[ \|f - g\|^2 = E\{ (f_i - g_i)^2 \}. \] (1)
We use $f \cdot g$ to denote the expectation of the product of $f$ and $g$,
i.e., the inner product
\[ f \cdot g = E\{ f \cdot g_i \}. \] (2)

We say a coder/decoder pair is admissible if the error $g - f$ and the
reconstructed signal $g$ are uncorrelated, i.e., $(f - g) \cdot g = 0$. The notion
of admissibility is to capture the idea that the decoder extracts all the
information about $f$ from the stream of code symbols. Henceforth, we
will restrict our discussion to admissible pairs of coders and decoders.

By definition, for an admissible coder
\[ \|f\| = \|f - g\| + \|g\|. \] (3)
The rate of a coder $c$ when applied to a source $f$ at distortion $D$
is the average number of bits per input sample in the coded stream.
The operational rate-distortion function of a coder, which is denoted
by $R_c(f, D)$, is the rate of a coder for source $f$ as a function of the
distortion $D$. The rate-distortion function for a source $f$, which is
denoted by $R(f, D)$, is the pointwise minimum of the rates of all possible
coders for $f$ as a function of the distortion. For a detailed
discussion of rate-distortion theory, see [1]. We say two sources $f$
and $g$ are independent if a) $f$ and $g$ are uncorrelated, i.e., $f \cdot g = 0$,
and b) $R(f + g, D) \geq R(f, D)$ and $R(f + g, D) \geq R(g, D)$.

III. Theoretical Results

We now state a general theorem on the convergence of Occam
filters. Intuitively speaking, the theorem implies that if a noisy signal
is coded with the loss equal to the noise power, then the noise
and the coding loss tend to cancel, with the extent of the coding
loss depending on the incompressibility of the noise relative to the
signal. More precisely, the theorem bounds the residual noise in the
reconstructed signal as a function of the ratio of the coding rate
achieved and the incompressibility of the noise. The numerator in
the estimate is the rate at which the noisy signal is coded, i.e., the
effectiveness of the coder, whereas the denominator is the derivative
of the rate-distortion function of the noise, i.e., a measure of the
incompressibility of the noise. We tacitly assume that the rate-
distortion function of the noise is differentiable. In a corollary to
the theorem, we address the special case of Gaussian noise, wherein
the incompressibility of the noise can be explicitly estimated.

**Theorem:** Let $f$ and $\nu$ be the signal and noise, respectively, where
$f$ and $\nu$ are independent. Let $g$ be the signal obtained by coding
and decoding $f + \nu$ using an admissible coder $c$, whose distortion is $\|\nu\|$. The
residual noise in the reconstructed signal is bounded as
\[ \frac{\|f - g\|}{\|f\|} \leq (2 + \sqrt{2}) \sqrt{-R'(\nu, \|\nu\|)} \] (4)
where $R'(\nu, \|\nu\|)$ is the left derivative of the rate-distortion function
$R(f, D)$ with respect to the distortion $D$ evaluated at $D = \|\nu\|$.

**Proof:** Express the reconstructed signal $g$ as the sum of two
uncorrelated signals $\hat{f}$ and $\hat{\nu}$ representing the distorted versions
of the signal $f$ and the noise $\nu$ captured in $g$. Specifically, $\hat{f}$ is the
projection of $g$ on $f$, and $\hat{\nu}$ is the portion of $g$ that is orthogonal to $f$
\[ \hat{f} = \frac{f \cdot g}{\|f\|^2} f, \quad \hat{\nu} = g - \hat{f}. \] (5)

Since the distortion in the coder is $\|\nu\|$, we have
\[ \|\nu\| = \|(f + \nu) - g\| \]
\[ = \|(f + \nu) - (\hat{f} + \hat{\nu})\| \]
\[ = \|(f - \hat{f}) + (\nu - \hat{\nu})\| \]
\[ = \|f - \hat{f}\| + \|\nu - \hat{\nu}\| \] (6)
where the last equality follows from the fact that $\nu$ and $\hat{\nu}$ are
orthogonal to $f$ and $\hat{f}$ by definition. If we set
\[ \zeta = \|f - \hat{f}\| \] (10)
in the above equation, we get
\[ \|\nu - \hat{\nu}\| = \|\nu\| - \zeta. \] (11)

Now
\[ \|f + \nu\| = \|(f + \nu) - g\| + \|g\| \]
\[ = \|f + \nu - g\| + \|g\| \]
\[ = \|\nu\| + \|g\| \] (14)
where (13) follows by virtue of the admissibility assumption,
and (14) follows by (13) by the fact that the loss in the coder is $\|\nu\|$. From
the orthogonality of $f$ and $\nu$, the left-hand side of (14) can be
written as $\|f\| + \|\nu\|$, yielding that
\[ \|f\| = \|g\| = \|\hat{f}\| + \|\hat{\nu}\| \] (15)
and hence
\[ \|\hat{\nu}\| = \|f\| - \|\hat{f}\|. \] (16)

Since $f$ and $\hat{f}$ are colinear, it follows that
\[ \frac{\|f\| - \|\hat{f}\|}{\|f\|} \leq 2 \sqrt{\frac{\|f - \hat{f}\|}{\|f\|}} = 2 \sqrt{\frac{\zeta}{\|f\|}}. \] (17)
Combining (16) and (17) we get
\[ \frac{\|\hat{\nu}\|}{\|f\|} \leq 2 \sqrt{\frac{\zeta}{\|f\|}}. \] (18)

We can now estimate $\|f - g\|/\|f\|$ as
\[ \frac{\|f - g\|}{\|f\|} = \frac{\|f - (f + \nu)\|}{\|f\|} = \|f - \hat{f}\| + \|\hat{\nu}\| \] (19)
Substituting (10) and (18) in the above, we have
\[ \frac{\|f - g\|}{\|f\|} \leq \frac{\zeta}{\|f\|} + 2 \sqrt{\frac{\zeta}{\|f\|}}. \] (20)

Now
\[ \zeta = \|f - \hat{f}\| \leq \|f\| + \|\hat{f}\| \]
\[ \leq \|f\| + \|g\| \]
\[ = 2\|f\| \] (23)
where the last inequality is obtained via (15). It follows that
\[ \frac{\zeta}{\|f\|} \leq \sqrt{\frac{\zeta}{\|f\|}}. \] (24)
Substituting the above in (20), we get
\[
\frac{\|f - g\|}{\|f\|} \leq (2 + \sqrt{2}) \sqrt{\frac{\zeta}{\|f\|}}.
\] (25)
It remains to estimate \(\zeta\). Now
\[
R_s(f + \nu, \|\nu\|) \geq R(f + \hat{\nu}, 0)
\]
\[
= R(\nu + (\hat{\nu} - \nu) + \tilde{f}, 0)
\]
\[
\geq R(\nu + \tilde{f}, \|\nu - \tilde{f}\|)
\]
\[
\geq R(\nu, \|\nu - \tilde{f}\|)
\]
\[
= R(\nu, \|\nu - \tilde{\nu}\|) - \zeta.
\] (30)
In the above, (28) follows from (27) by the definition of the rate-distortion function, and (29) follows from (28) since \(\tilde{f}\) is a scalar multiple of \(f\) and, hence, is independent of \(\nu\). Inequality (30) follows from (29) by virtue of (11). Using a Taylor expansion for the rate-distortion function of \(\nu\) and invoking the convexity of the rate-distortion function
\[
R(\nu, \|\nu - \tilde{\nu}\|) \geq R(\nu, \|\nu\|) - \zeta R'(\nu, \|\nu\|).
\] (31)
Noting that the \(R(\nu, \|\nu\|) = 0\) and combining (30) and (31), we have
\[
\zeta \leq -\frac{R_s(f + \nu, \|\nu\|)}{R'(\nu, \|\nu\|)}.
\] (32)
Substituting the above in (25), we have the theorem.

**Corollary 1:** If the noise variable is normally distributed with zero mean, then the residual noise in the reconstructed signal is given by
\[
\frac{\|g - f\|}{\|f\|} \leq 4.02 \sqrt{R_s(f + \nu, \|\nu\|)} \sqrt{\frac{\|\nu\|}{\|f\|}} \] (33)
Proof: Suppose that the noise variable \(\nu\) is normally distributed with zero mean and variance \(\|\nu\|\). The rate-distortion function \(R(\nu, D)\) is given by [1]
\[
R(\nu, D) = \frac{1}{2 \ln(2)} \ln \frac{\|\nu\|}{D} \quad D \leq \|\nu\| \] (34)
in bits per sample. Taking the derivative and evaluating it at \(D = \|\nu\|\), we get
\[
R'(\nu, \|\nu\|) = -\frac{1}{2 \ln(2)} \|\nu\|.
\] (35)
Substituting the above in the statement of the theorem and rounding up the constants, the corollary follows.

Define the SNR of the noisy signal to be
\[
\text{SNR}_\text{noisy} = 10 \log_{10} \left( \frac{\|f\|}{\|\nu\|} \right) \text{ dB} \] (36)
as measured in decibels. Similarly, the SNR of the reconstructed signal is given by
\[
\text{SNR}_\text{out} = 10 \log_{10} \left( \frac{\|f\|}{\|g - f\|} \right) \text{ dB}. \] (37)
By taking logarithms on both sides of (33), we get the following.

**Corollary 2:** If the noise variable is normally distributed with zero mean, then the SNR of the reconstructed signal in decibels satisfies
\[
\text{SNR}_\text{out} \geq \frac{1}{2} \text{SNR}_\text{noisy} - 10 \log_{10}(4.02) - 5 \log_{10} R_s(\nu, \|\nu\|).
\] (38)
From (38), if \(R_s(\nu, \|\nu\|) < 1\), that is, the filtered signal can be compressed with less than one bit per symbol, then \(\text{SNR}_\text{out}\) is expected to improve.

**IV. EXPERIMENTAL RESULTS**

We now turn our attention to the practical problem of filtering digital images corrupted with random noise. In particular, we build an Occam filter for images using a coder consisting of two independent phases: a quantization phase followed by a lossless coding phase. For the lossless coding phase, we used JPEG-LS, which is a new JPEG lossless standard (ISO CD 14495) [10], [11]. Since the decoder simply inverts the lossless coding, the lossless coding phase is not critical to the strength of the experiment, and any other lossless coding scheme could be used.

The quantization phase is based on the SVD. Given an \(m \times n\) matrix \(B\), the SVD of \(B\) is given by \(B = U_n \Sigma_n V_n^T\), where \(U_n\) and \(V_n\) are orthogonal matrices, and \(\Sigma_n = \text{diag}(\beta_1, \beta_2, \ldots, \beta_n)\) is a diagonal matrix. The diagonal elements of \(\Sigma_n\) can be arranged in a descending order and are called the singular values of \(B\).

The main steps in our SVD-based quantizer are as follows [7]:

a) Divide an \(M \times N\) matrix \(B\) into nonoverlapping blocks of size \(8 \times 8\) pixels.

b) Perform SVD on each block. Set to zero the singular values that are smaller than a threshold \(\epsilon\). Using the quantized singular values, compute the inverse SVD of each block to obtain the quantized (filtered) image block.

By selecting the threshold \(\epsilon\), the loss of the coder can be controlled simply in the squared-error sense, owing to the orthogonality of the transforms involved. Traditional SVD-based techniques use ad hoc schemes to select the threshold \(\epsilon\) to distinguish between significant and nonsignificant singular values. Our contribution here is to use the principle of Occam filtering in selecting the threshold \(\epsilon\) so that the loss of the coder is equal to the power of the noise in the squared-error sense.

In our experiments, we took three images of \(512 \times 512\) pixels: **Barbara**, **Couple**, and **Lena**. For each image, we added Gaussian noise of zero mean and variance 25, 50, and 100 to obtain three noisy images. For each noisy image, we applied the SVD quantizer with loss equal to the noise variance and then declared the quantized image to be the filtered image. Table I shows the mean squared error in the filtered image as compared with the noise-free image for each of the choices of image and noise. Table I also shows the mean square error obtained by filtering each image with the wavelet denoising technique of Donoho and Johnstone [5], using length 18 coiflets [4] for a depth-three wavelet tree. It is clear that the SVD-based Occam technique gives consistently better results. However, we note that the SVD is potentially more expensive to compute than the wavelet transform, although this is mitigated to some extent by our use of block-based SVD’s.

To relate our theoretical results with our experimental results, we compare the predictions of Corollary 2 with the actual SNR’s achieved in the experiments. Fig. 2 shows a plot of achieved output SNR (SNR_out) on the vertical axis versus the predicted bound on the horizontal axis, i.e., the right-hand side of the inequality (38) for each of the three images. Note that \(R_s(f + \nu, \|\nu\|)\) is the bit rate
TABLE I

<table>
<thead>
<tr>
<th>Equation</th>
<th># Mults.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y(n - L) = \begin{bmatrix} 0 \ Y(n - L - 1) \end{bmatrix} + x(n - L)\hat{W}(n - L) + \epsilon^{(n)}(n - L - 1)\hat{R}(n)$</td>
<td>$2L$</td>
</tr>
<tr>
<td>$\hat{W}(n - L + 1) = \hat{W}(n - L) + x(n - 2L)E^{(n)}(n - L - 1)$</td>
<td>$L$</td>
</tr>
<tr>
<td>$\epsilon^{(n)}(n - L) = \mu^{(n)}(n - L){ d(n - L) - yL(n - L) }$</td>
<td>$1$</td>
</tr>
<tr>
<td>$E^{(n)}(n - L) = \begin{bmatrix} \epsilon^{(n)}(n - L) \ E^{(n)}(n - L - 1) \end{bmatrix}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\hat{R}(n + 1) = \hat{R}(n) + x(n - L)X(n) - x(n - 2L)X(n - L)$</td>
<td>$2L$</td>
</tr>
<tr>
<td>Total ($\mu(n)$ arbitrary): $5L + 1$</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Predicted versus computed output SNR for SVD-based Occam filtering on the Lena, Barbara, and Couple images.

(bits per sample) at the output of our lossless coder. The input SNR ($SNR_{in}$) ranged roughly between 20 and 30 dB, corresponding to the noise variance in the range 25–100, as in Table I. The predicted SNR is much too conservative in that it is about 20 dB smaller than the achieved value. However, it is significant that the achieved SNR is linear in the predicted value for each of the three images. This points to the validity of Corollary 2 within a scalar constant.

V. CONCLUSION

An Occam filter employs lossy data compression to separate signal from noise. We proved that Occam filters can be used to separate two stochastic sources with the effectiveness of the separation depending on their relative compressibility. To illustrate the practicability of Occam filters, we constructed a simple Occam filter based on the singular value decomposition and applied it to digital images corrupted with Gaussian noise. We observed that the SVD-based Occam filter outperformed the wavelet-based denoising method of Donoho and others on several sample images.

ACKNOWLEDGMENT

The authors wish to thank T. Berger, P. Chou, R. Gray, and G. Chang for comments and discussions.

REFERENCES