UNIVERSAL DENOISING OF CONTINUOUSAMPLITUDE SIGNALS WITH APPLICATIONSTO IMAGES

Kamakshi Sivaramakrishnan Tsachy Weissman
Department of Electrical Engineering Stanford University Stanford, California 94305–9505

ABSTRACT
We consider the problem of image denoising wherein the statistical characterization of the noise corruption mechanism is known. We make no assumptions on the nature or statistics of the underlying noise-free signal. A denoiser is proposed which, although ignorant of the statistical properties of the noise-free image, does essentially as well as a scheme with full knowledge of the statistics. The solution is presented as a sequence of schemes that progressively consider larger neighborhoods (contexts) around a pixel being denoised, to achieve optimum performance under a user-defined distortion measure.

Index Terms—Universal Denoising, kernel density estimation, Quantization, Sliding Window Denoiser, Denoisability, Continuous Memoryless Channels.

I. INTRODUCTION
The digitization of most forms of visual representation has made the problem of image denoising and restoration a particularly relevant one in today’s context. This problem, in spite of being a well researched topic over the years has suffered from the rather specific nature of its theoretical formulation in the past. There are abundant heuristic techniques and smoothing-based approaches which give good performance when adapted to a particular image/data in question. The wavelet domain denoising based on soft-thresholding and Gaussian scale mixtures [1] provides an elegant solution to the problem for the specific case of an additive white gaussian noise corruption source under a mean-squared error metric. It is of both theoretical and practical interest to investigate the problem of denoising with no assumptions on the nature of the corruption source or the statistics of the ‘clean’ source. Recent developments in [2] provide an extension of the DUDE framework based on lossless compression techniques to obviate some of the computational issues presented by applying the DUDE of [3]. It, however, leaves room for performance gains which are affected by the procedure for estimation of the distributions of the noisy symbols. This procedure is based on quantization of the space of the output noisy symbols combined with prediction-based techniques.

In this paper, we propose an asymptotically universal optimal scheme that has no knowledge of the statistics or type of the ‘clean’ image, and is computationally viable, outperforming some of the state-of-the art schemes. This scheme also obviates the computational intractableness of the DUDE-based approach in [4].

II. PROBLEM SETTING AND NOTATION
Let \( x = (x_1, x_2, \ldots) \) be the noise-free source signal with components taking values in \([a, b]\) \( \subset \mathbb{R} \) and \( Y = (Y_1, Y_2, \ldots), Y_i \in \mathbb{R} \) be the corresponding noisy observations, also referred to as the output of the channel (corruption source). The channel considered here is memoryless, specified by a family of distribution functions \( \mathcal{C} = \{ F_{Y|X} \}_{x \in [a, b]} \) with the assumption of the associated family of measures \( \mathcal{Y} = \{ \mu_x \}_{x \in [a, b]} \) being tight i.e. \( \sup_{x \in [a, b]} \mu_x([-T, T]^c) \rightarrow 0 \) as \( T \rightarrow \infty \). We also assume that \( F_{Y|X} \) is absolutely continuous \( \forall x \in [a, b] \), letting \( f_{Y|X} \) denote the associated density w.r.t Lebesgue measure. An n-block denoiser is a Borel-measurable mapping taking \( \mathbb{R}^n \) into \([a, b]^n\). We assume a bounded loss function \( \Lambda : [a, b]^2 \rightarrow [0, \infty) \) and denote the normalized cumulative loss of n-block denoiser \( \hat{X}_n \) by

\[
L_{\hat{X}_n}(x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, \hat{X}_n(y^n)[i]) \tag{1}
\]

where \( \hat{X}_n(y^n)[i] \) denotes the i-th component of \( \hat{X}_n(y^n) \), the denoised sequence having observed the noisy sequence \( y^n \). We also define the minimum symbol-by-symbol loss of \( x^n \) as

\[
D(x^n) = \min_{g} \left[ \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, g(Y_i)) \right] \tag{2}
\]

where the minimum is over all measurable maps \( g : \mathbb{R} \rightarrow [a, b] \). It has been shown in [5] that the minimizer of equation (2) is given by

\[
g_{\text{opt}}[F_{x^n}][y] = \arg \min_{\hat{x} \in [a, b]} \int_{[a, b]} \Lambda(x, \hat{x}) d[F_{x^n} \otimes \mathcal{C}]_{X|Y}(x) \tag{3}
\]
where, \([F_x^n \otimes C]_{X=x/y}\) denotes the conditional distribution of the underlying clean symbol value being \(x\) given the noisy symbol is \(Y = y\), with \(F_x^n\) being the input empirical distribution of \(x\). This also defines the optimum denoising candidate for the symbol-symbol case, \(F_x^n\) being unknown, the crux of the denoising process lies in estimation of this input empirical distribution. We have presented a one-dimensional formulation of the problem for sake of notational simplicity but the setting and results extend very easily without loss of generality to higher dimensions.

III. CONSTRUCTION OF UNIVERSAL DENOISER

Given the memoryless nature of the channel, the sequence of output symbols, \(Y_1, Y_2, \ldots, Y_n\) are independent random variables taking values in \(\mathbb{R}\) and have conditional densities, \(f_{Y_2|Y_1}, f_{Y_3|Y_2}, \ldots, f_{Y_n|Y_{n-1}}\), respectively. In order to estimate the input empirical distribution of the clean signal, we begin by estimating the probability density function of the noisy output symbols. We map this density estimate of the output noisy symbols to a corresponding channel input distribution, which gives an estimate for the empirical distribution of the clean signal.

III-A. Density Estimation

A density estimate is a sequence \(f^1, f^2, \ldots, f^n\), where for each \(n\), \(f^n(y) = f^n(y; Y_1, \ldots, Y_n)\) is a real-valued Borel measurable functions of its arguments, and for fixed \(n\), \(f^n\) is a density estimate on \(\mathbb{R}\). The kernel estimate is defined by

\[
f^n_k(y) = \frac{1}{nh^d} \sum_{i=1}^n K \left( \frac{y - Y_i}{h} \right)
\]

where \(h = h_n\) is a sequence of positive numbers and \(K\) is a Borel measurable function satisfying \(K \geq 0, \int K = 1\).

It has been shown in [5] that such a kernel density estimate when applied to the noisy observations \(Y_1, \ldots, Y_n\), regardless of the noiseless symbol \(x^n\), yields

\[
\int \left| f^n_k(y) - \int f_{Y|x} dF_x^n(x) \right| dy = o(n^{-1/2}) \quad \text{a.s.}
\]

where \([F_x^n \otimes C]_Y\) denotes the marginal distribution of the observed noisy symbols. This can be mapped to a corresponding estimate of the empirical distribution of the input sequence using a mapping which we specify next.

III-B. Channel Inverse

Consider the following mapping,

\[
\hat{F}_x^n[Y^n] = \arg \min_{F \in C_{x^n}} d \left( f^n_k, \int f_{Y|x} dF(x) \right)
\]

where \(F_{x^n}^{[a,b]}\) denotes the set of empirical distributions induced by \(n\)-tuples with \([a, b]\)-valued components. The definition for the norm, \(d\), is given by

\[
d(f, g) = \int |f(y) - g(y)| dy
\]

The estimate in equation (6) can be shown to satisfy

\[
\hat{F}_x^n \to 0 \quad \text{a.s.}
\]

where \(\lambda\) is the Levy metric defined as

\[
\lambda(F, G) = \inf\{\varepsilon > 0 : F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon \text{ for all } x\}
\]

For practical computational purposes, this estimate \(\hat{F}_x^n\) is approximated using its quantized version as

\[
P_{x^n}^{\delta}(a_i) = \hat{F}_x^n(a_i) - \hat{F}_x^n(a_{i-1})
\]

where \(a_i\) are s.t. \(A = \{a_i = a + i\Delta, i = 0, \ldots, N(\Delta)\}, N(\Delta) = \frac{(b-a)}{\Delta} + 1. \Delta\) is the quantization step-size of the interval \([a, b]\). We also have

\[
P_{x^n}^{\delta}(\delta) = Q_{\delta} \left( \hat{F}_x^n \right)
\]

where \(Q_{\delta}\) denotes quantization to the nearest non-negative integer multiple of \(\delta\) where \(\delta > 0\). The denoiser then reduces to

\[
\hat{X}^{n,\delta,\Delta} \approx \{y^n\} = g_{\text{opt}}[P_{x^n}^{\delta,\Delta}[y^n]](y_i), \quad 1 \leq i \leq n
\]

with

\[
g_{\text{opt}}[P_{x^n}^{\delta,\Delta}](y) = \arg \min_{x \in \mathcal{A}} \Lambda^T_{\hat{x}} [P_{x^n}^{\delta,\Delta} \otimes C]_{X=x/y}
\]

\[
\Lambda_{\hat{x}} = \{\Lambda (x, x) | x \in \mathcal{A}\} \text{ is a column vector in } \mathbb{R}^{\mid \mathcal{A}\mid}
\]

\[
\Delta_{\hat{x}} = \{\Delta (x, y) | x \in \mathcal{A}\}
\]

\[
\Lambda_{\hat{x}} = \{\Lambda (x, x) | x \in \mathcal{A}\}
\]

where \([F_x^n \otimes C]_{X=x/y}\) is a column vector of size \(\mid \mathcal{A}\mid\) denoting conditional distribution of the clean symbol, \(X = x\) given \(Y = y\).

III-C. Extension to \(2k + 1\)-window length denoiser

The ideas presented in the preceding sections can be extended to the case where every symbol is associated with a corresponding neighborhood of symbols, called the ‘context’. A context of length \(2k\) associated with a symbol \(y_i\) of the sequence \(y^n\), \(k + 1 \leq i \leq n - k\) is defined as

\[
\mathcal{N}_i = \{y_j | i - k \leq j \leq i + k, j \neq i\}
\]

The context, \(\mathcal{N}_i\), together with the symbol, \(y_i\), form a \(2k + 1\)-tuple ‘super-symbol’. The consistency results in the previous sections extend to the case of these super-symbols.

The \(2k + 1\)-sliding window denoiser then operates by sliding through the super-symbols denoising the center-symbol by estimating the statistics of order \(2k + 1\). An explicit description of the denoiser can be found in [5]. Let \(\hat{F}_{x^n}^{\delta,\Delta,k}\) denote the \(2k + 1\)-th order estimate of the input empirical distribution of the \(x\) and the \(2k + 1\) sliding window denoiser is defined as

\[
\hat{X}^{n,\delta,\Delta,k} \approx \{y^n\} = g_{\text{opt}}[\hat{F}_{x^n}^{\delta,\Delta,k}[y^n]](y_i), \quad k + 1 \leq i \leq n - k
\]
IV. ASYMPTOTIC OPTIMALITY

Let $D_k(x^n)$ denote the $2k + 1$-th order sliding window minimum loss defined as

$$D_k(x^n) = \min_g \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(x_i, g(Y_{i+k}^{i-k}))$$  (14)

where the minimum is over all measurable maps $g : \mathbb{R}^{2k+1} \rightarrow [a, b]$. Letting $\hat{X}_n^{\alpha} = \hat{X}^{n,\Delta,k}$.

Theorem 1: For all $x \in \mathbb{R}^{\infty,k}$

$$\limsup_{n \to \infty} \left[ L_{\hat{X}_n^{\alpha}}(x^n, Y^n) - D_k(x^n) \right] \leq 0 \quad a.s.$$  (15)

The above theorem establishes the ability of the proposed scheme to compete with the best possible gene-aided scheme that has full knowledge of the distribution of the clean image and the channel asymptotically.

V. APPLICATION TO CONTINUOUS TONE IMAGE DENOISING

The kernel density estimation in equation (4) is a computationally intensive procedure, and fast algorithms for its computation have been presented in [6]. The implementation uses “kd-trees”, a hierarchical representation for point sets which caches sufficient statistics about point locations in order to achieve potential speedups in computation. We use the software provided at [7] to implement the kernel density estimation on the noisy observations. The mapping in equation (6) can be framed as an $L_1$ norm problem solved using the simplex method in linear programming. The $2k + 1$ super-symbol is formed by considering a $\lfloor \sqrt{2k} + 1 \rfloor$-length square region of pixels around the pixel being denoised. For a user-defined loss function, $\Lambda$, the universal denoiser is constructed using equation (12).

VI. RESULTS

Results of applying the proposed scheme to a natural test image are shown in Figs 1 and 2 and presented in Table I. We illustrate the application of the proposed scheme to a non-additive noise corruption mechanism. The noise is introduced by corrupting every gray-level by a parameter that is picked from a Rayleigh distribution whose parameter is dependent on the pixel/gray level. Rayleigh-distributed speckle noise is a commonly observed phenomena in SAR imagery. As can be seen from the figure, the symbol-symbol scheme that relies only on marginal distributions leaves remnant artifacts in denoising the background curtain and even in reproducing the contrast of the original image. The context $'2k + 1'$ indicates knowledge of one adjoining pixel while higher context lengths consider knowledge of the 2-D neighborhood of the pixel being denoised. Table I shows successive improvements of the scheme with increasing lengths of the context, $k$, in terms of the loss metric, viz., the RMSE. The impact of the improved loss metric on the visual quality is evident in a truer reproduction of

![Fig. 1. Row 1- left: Original image, right: Noisy image; Denoised images using Row 2- left: symbol-symbol scheme, right: '2k+1'=2; Row 3- left: 2k+1 = 3, right: [1]; Row 4- left: 2k+1 = 3 followed by adaptive Wiener filtering, right: the $L_1$ metric and the proposed scheme; Row 5- left: Error image for 2k+1=3, right: Error image for [1]](image-url)
We have presented a sequence of universally optimal denoisers for estimating the components of a real valued sequence corrupted by a known memoryless channel. This provides a conclusive extension to recent works [3], [4] extending the framework to handle the denoising problem in much greater generality. The current work addresses the limitations of the DUDE in [3] in terms of intractability of the computations of the context-aided distributions. The present formulation of the problem also provides the advantage of not quantizing the output alphabet space, as in [4], which results in more efficient estimation of the distributions of the output alphabets and the corresponding clean sequence given the noisy observations. Finally, the proposed scheme is computationally tractable, and appears to have potential to do well on various types of real data and noise sources.

VIII. REFERENCES