Universal Denoising of Discrete-time Continuous-Amplitude Signals

Kamakshi Sivaramakrishnan Tsachy Weissman
Department of Electrical Engineering
Stanford University
Stanford, California 94305–9505
Email: {ksivaram, tsachy}@stanford.edu

***Abstract*— We consider the problem of reconstructing a discrete-time continuous-amplitude signal corrupted by a known memoryless channel with a general output alphabet. We develop a sequence of denoisers that asymptotically achieve optimum performance in a semi-stochastic setting of an unknown individual noiseless signal, where the quality of reconstruction is measured with respect to a general given loss function satisfying mild conditions. We also extend this to the fully stochastic setting and show that our denoiser is asymptotically optimal for any stationary noiseless source. We conclude with some experimental validations of the proposed theory.**

I. Introduction

Consider the problem of estimating the clean signal $\{X_t\}_{t\in\mathbb{T}}$, $X_t \in [a,b] \subset \mathbb{R}$, from its noisy observations $\{Z_t\}_{t\in\mathbb{T}}$, $Z_t \in \mathbb{R}$, where $\{Z_t\}$ is the output of a memoryless channel whose input is $\{X_t\}$. This problem finds applications in areas ranging from engineering, cryptography, astronomy to bioinformatics. There is significant literature on particular instantiations of this problem, for example, where the noise corruption (channel) is additive in nature and has a specific form of the distribution function, most notably Gaussian (cf. [1], [2] and references therein). Recently, universal denoising for discrete signals and channels was considered in [3]. The results of [3], and the denoising scheme DUDE suggested therein, although attractive theoretically, are restricted in their practicality to problems with small alphabets. This is a result of computational issues involved with collecting higher-order statistics from the noisy data, mapping an estimated channel input distribution to an estimated channel output distribution, and statistical issues having to do with count statistics that are too sparse to be reliable for even moderately large alphabets sizes. This leaves open challenges in the application of DUDE to problems like gray-scale image denoising. The problem was further extended to the discrete-valued input and general output alphabet setting in [4]. This approach proposes quantization of the output alphabet space and proceeds on an approach similar to that in [3], showing that there is no essential loss of optimality in quantizing the channel output before denoising. In spite of its theoretical elegance, this approach faces similar issues as the scheme of [3], limiting its scope of applications to small channel input alphabets. More recently, a modified DUDE, using ideas from lossless image compression, was presented in [5]. As discussed in that work, in spite of circumventing some of the computational issues mentioned above, the approach leaves room for improvement in the denoising performance.

Recent developments in universal denoising have also been reported in [2]. Their approach is based on local smoothing methods that make assumptions on the underlying structure of the data which are more relevant in image denoising due to inherent redundancy in natural images. The consistency results showed the convergence of the denoising rule to the conditional expected value of the clean symbol given the noisy neighborhood sans the noisy symbol being denoised. There is potential to improve this result by incorporating the information from the noisy pixel that is being denoised too, an approach at the heart of the denoisers we present below. We establish the universal optimality of the suggested denoisers in a generality that applies to arbitrarily distributed noiseless signals, arbitrary memoryless channels, and arbitrary loss functions (with some benign regularity conditions).

The remainder of the paper is organized as follows. In section II, we discuss the problem setup and notations. This is followed by a description of some technical results that are key to the construction of the denoisers and their performance analysis in Section III. Section IV details the construction and performance guarantees for our suggested universal "symbol by symbol" denoiser. Section V is devoted to extending the ideas to the construction of a sliding window context-aided denoiser and detailing a few of its theoretical performance guarantees implying its universal asymptotic optimality. Section VI briefly mentions some promising preliminary experimental results. Proofs and associated details for the theorems and lemmas, as well as additional theoretical and experimental results, are given in [6].

II. Problem Setting and Notations

Let $\mathbf{x} = (x_1, x_2, \cdots)$ be the individual noise-free source signal with components taking values in $[a,b] \subset \mathbb{R}$ and $\mathbf{Y} = (Y_1, Y_2, \cdots)$, $Y_i \in \mathbb{R}$ 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 $\mathcal{C} = \{F_{Y|x}\}_{x\in[a,b]}$, where $F_{Y|x}$ denotes the distribution of the channel output symbol when the input symbol is x. We assume the associated family of measures $\Upsilon = \{\mu_x\}_{x\in[a,b]}$ to be tight in the sense that $\sup_{x\in[a,b]} \mu_x([-T,T]^c) \to 0$ as $T \to \infty$.

1-4244-0504-1/06/$20.00 ©2006 IEEE

An *n-block denoiser* is a measurable mapping taking $\mathbb{R}^n$ into $[a,b]^n$. We assume a loss function $\Lambda : [a,b]^2 \to [0,\infty)$ and denote the normalized cumulative loss of an 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 $\hat{X}^n(y^n)$. We denote $\Lambda_{\max} = \sup_{x,y\in[a,b]} \Lambda(x,y)$, and assume $\Lambda_{\max} < \infty$. Denote $\mathcal{F}^{[a,b]}$ to be the set of all probability distribution functions with support contained in the interval $[a,b]$. For $F \in \mathcal{F}^{[a,b]}$, we let

$$\mathcal{U}(F) = \min_{\hat{x}\in[a,b]} \int_{x\in[a,b]} \Lambda(x,\hat{x}) dF(x) \tag{2}$$

denote its 'Bayes envelope' (our assumptions below on the loss function will imply existence of the minimum). Define the symbol-by-symbol minimum loss of x^n by

$$D(x^n) = \min_g E\left[\frac{1}{n}\sum_{i=1}^{n} \Lambda(x_i, g(Y_i))\right] \tag{3}$$

where the minimum is over all measurable maps $g : \mathbb{R} \to [a,b]$. For $x^n \in [a,b]^n$ define

$$F_{x^n}(x) = \frac{|\{1 \le i \le n : x_i \le x\}|}{n}, \tag{4}$$

i.e., the CDF associated with the empirical distribution of x^n. For simplicity we also assume henceforth 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. Note that $D(x^n)$ can be expressed as

$$D(x^n) = \min_g \int_{[a,b]} E_x \Lambda(x, g(Y)) dF_{X^n}(x) \tag{5}$$

where E_x denotes expectation when the underlying clean symbol is x, the expectation being over the channel noise

$$E_x \Lambda(x, g(Y)) = \int \Lambda(x, g(y)) f_{Y|x}(y) dy \tag{6}$$

For $F \in \mathcal{F}^{[a,b]}$, let $F \otimes \mathcal{C}$ and $E_{F\otimes\mathcal{C}}$ denote, respectively, distribution and expectation when the channel input $X \sim F$ and Y is the channel output. So that,

$$\begin{aligned} E_{F\otimes\mathcal{C}}\Lambda(X, g(Y)) &= \int_{[a,b]} E_x \Lambda(x, g(Y)) dF(x) \qquad (7)\\ &= \int_a^b \left[\int_{\mathbb{R}} \Lambda(x, g(y)) f_{Y|x}(y) dy\right] dF(x) \end{aligned}$$

Letting $[F \otimes \mathcal{C}]_{X|y}$ denote the conditional distribution of X given $Y = y$ under $F \otimes \mathcal{C}$ (which can be obtained explicitly given F and $\mathcal{C}$), we have

$$\min_g E_{F\otimes\mathcal{C}}\Lambda(X, g(Y)) = E_{F\otimes\mathcal{C}}\mathcal{U}\left([F\otimes\mathcal{C}]_{X|Y}\right) \tag{8}$$

with $\mathcal{U}$ denoting the Bayes envelope as defined above, and where the minimum is attained by the Bayes response to $[F \otimes \mathcal{C}]_{X|y}$, namely,

$$g_{opt}[F](y) = \arg\min_{\hat{x}\in[a,b]} \int_{[a,b]} \Lambda(x,\hat{x}) d[F\otimes\mathcal{C}]_{X|y}(x) \tag{9}$$

Note that from (5), (6) and (7) we have

$$D(x^n) = \min_g E_{F_{x^n}\otimes\mathcal{C}}\Lambda(X, g(Y)) \tag{10}$$

where F_{x^n} was defined in equation (4) and the minimum is attained by $g_{opt}\left[F_{x^n}\right]$

III. Towards Construction of Denoiser

F_{x^n} and, hence, $g_{opt}[F_{x^n}]$ are not known to an observer of the noisy sequence, Y^n. The first order of business is to estimate the input empirical distribution from the observable noisy sequence and knowledge of the channel. We approach this problem by first estimating a function that tracks the evolution of the "average" density function according to which the output symbols are distributed. This, for the case of a sequence of a finite length amounts to estimating the density function according to which the sequence of output noisy symbols are distributed. For an input sequence x^n, given the memoryless nature of the channel, the output symbols will be distributed as $\{F_{Y|x_1}, \cdots, F_{Y|x_n}\}$ and have the corresponding density functions, $\{f_{Y|x_1}, \cdots, f_{Y|x_n}\}$. The function we are interested in estimating is

$$\frac{1}{n}\sum_{i=1}^{n} f_{Y|x_i}(y). \tag{11}$$

Once we have an estimate $f_Y^n = f_Y^n[Y^n]$ for this function, we use it to estimate the input empirical distribution by

$$\hat{F}_{x^n}[Y^n] = \arg\min_{F\in\mathcal{F}_n^{[a,b]}} d\left(f_Y^n, \underbrace{\int f_{Y|x} dF(x)}_{[F\otimes\mathcal{C}]_Y}\right) \tag{12}$$

where $[F\otimes\mathcal{C}]_Y$ denotes the marginal of the output symbol Y under the distribution $F\otimes\mathcal{C}$ described earlier. $\mathcal{F}_n^{[a,b]} \subseteq \mathcal{F}^{[a,b]}$ denotes the set of empirical distributions induced by n-tuples with $[a,b]$-valued components where every member, $F(x)$, of $\mathcal{F}_n^{[a,b]}$ is of the form

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(x\le x_i)} \tag{13}$$

where, $x^n = \{x_1, x_2, \cdots, x_n\}$, is the $[a,b]$-valued n-tuple. The definition for the norm, d, is

$$d(f,g) = \int |f(y) - g(y)|\, dy \tag{14}$$

A. Density Estimation for independent and non identically distributed random variables

Towards our first order of business, which is to estimate $\hat{F}_{x^n}$, we estimate the function in (11). Given the memoryless nature of the channel, the sequence of output symbols, $Y_1, Y_2, \cdots, Y_n$ are independent random variables taking values in $\mathbb{R}$ and have conditional densities, $f_{Y|x_1}, f_{Y|x_2}, \cdots, f_{Y|x_n}$ respectively. A density estimate is a sequence $f^1, f^2, \cdots, f^n$, where for each n, $f_Y^n(y) = f^n(y; Y_1, \cdots, Y_n)$ is a real-valued Borel measurable function of its arguments, and for fixed n, f_Y^n is a density estimate on $\mathbb{R}$. The *kernel estimate* is given by

$$f_Y^n(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right) \tag{15}$$

where $h = h_n$ is a sequence of positive numbers and K is a Borel measurable function satisfying $K \geq 0$, $\int K = 1$. The L_1 distance, J_n, is defined as

$$J_n = \int \left| f_Y^n(y) - \frac{1}{n} \sum_{i=0}^{n} f_{Y|x_i}(y) \right| dy \tag{16}$$

A result very similar to that in [7] is elaborated here, viz., for the kernel estimate, all types of convergence of J_n to 0 are equivalent. The choice of L_1 distance, as elaborated in [7], is motivated by its invariance under monotone transformations of the coordinate axes and the fact that it is always well-defined.

Theorem 1: Let K be a nonnegative Borel measurable function on $\mathbb{R}$ with $\int K = 1$ of class $2 \leq s \leq 3$ (refer to [8] for class definitions). Consider

1) $J_n \to 0$ in probability as $n \to \infty$, for some sequence x^n
2) $J_n \to 0$ in probability as $n \to \infty$, for all sequences x^n
3) $J_n \to 0$ almost surely as $n \to \infty$, for all sequences x^n
4) For all $\epsilon > 0$, there exist r, $n_0 > 0$ such that $P(J_n \geq \epsilon) \leq e^{-rn}$, $n \geq n_0$, all sequences x^n.
5) $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh = \infty$

It is then true that 5 $\Rightarrow$ 4 $\Rightarrow$ 3 $\Rightarrow$ 2 $\Rightarrow$ 1.

B. Channel Inversion

The mapping in equation (12) projects the kernel estimate at the output of the channel to an estimate of the input empirical distribution. This projection is such that it best approximates (in the L_1 sense), the kernel density estimate with a member in the set of achievable output distributions.

For the mapping defined in equation (12),

Lemma 1: As $J_n \to 0$, $d([F_{x^n} \otimes \mathcal{C}]_Y, \left[\hat{F}_{x^n} \otimes \mathcal{C}\right]_Y) \to 0$ a.s.

Definition 1 (Levy metric): The Levy distance $\lambda(F, G)$ between any two distributions F and G is defined as

$$\lambda(F,G) = \inf\{\varepsilon > 0 : F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon \text{ for all } x\} \lim_{\Delta\to 0} \inf_{\varepsilon > 0} \{\varepsilon : P^{\Delta}(\mathbb{I}) \leq P(\mathbb{I}^{\varepsilon}) + \varepsilon, \mathbb{I} \in [a,b]\} = 0 \tag{21}$$

Definition 2 (Prohorov metric): For any two laws P and Q on S, $f : S \to \mathbb{R}$ let $\int f d(P - Q) := \int f dP - \int f dQ$, for bounded $\int f dP$ and $\int f dQ$, the Prohorov metric is defined as

$$\beta(P,Q) := \sup\left\{ \left| \int f d(P-Q) \right| : \| f \|_{BL} \leq 1 \right\}$$

where

$$\| f \|_{BL} = \| f \|_L + \| f \|_\infty \tag{17}$$

and

$$\| f \|_L =: = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}, \qquad \| f \|_\infty = \sup_x |f(x)| \tag{18}$$

Lemma 2: For the channel, $\mathcal{C}$, define

$$\varepsilon_\Delta(y) = \sup_{x \in [a,b]} \sup_{\substack{\hat{x} \in [a,b] \\ |x - \hat{x}| \leq \Delta}} \left| f_{Y|x}(y) - f_{Y|\hat{x}}(y) \right| \tag{19}$$

and suppose that the channel satisfies the following two conditions

1) $\lim_{\Delta\to 0} \varepsilon_\Delta(y) = 0$, $\forall\ y$
2) The set of densities $\{f_{Y|x}\}_{x\in[a,b]}$ is a set of linearly independent functions in $L_1(\mu)$

If, for distributions F and G, $d([F \otimes \mathcal{C}]_Y, [G \otimes \mathcal{C}]_Y) \to 0$ then, $\lambda(F, G) \to 0$.

Thus for a channel, $\mathcal{C}$,

1) whose associated measures are both, tight and absolutely continuous
2) that satisfies the continuity conditions in Lemma 2

and mapping defined in equation (12) we have, $\lambda\left(F_{x^n}, \hat{F}_{x^n}\right) \to 0$

C. Distribution-independent Approximation of the Estimate of the Input empirical distribution

We develop a distribution-independent approximation of the input empirical distribution, $\hat{F}_{x^n}[Y^n]$. We begin by defining some new quantities. For $\Delta > 0$, if $\frac{b-a}{\Delta} \in \mathbb{Z}^+$, consider a family of vectors, $\mathcal{F}^\Delta = \{P^\Delta\colon P^\Delta = \big(P(a_0), P(a_1), \cdots, P(a_{N(\Delta)})\big)\}, N(\Delta) = \frac{(b-a)}{\Delta}, \mathcal{A} = \{a_i = a + i\Delta, i = 0, \cdots, N(\Delta)\}, \sum_{i=1}^{N(\Delta)} P(a_i) = 1$ else, define the family of vectors as $\mathcal{F}^\Delta = \{P^\Delta\colon P^\Delta = \big(P(a_0), P(a_1), \cdots, P(a_{N(\Delta)-1}),\ P(a_{N(\Delta)})\big)\}, N(\Delta) - 1 = \lfloor \frac{(b-a)}{\Delta} \rfloor, \mathcal{A} = \{a_i = a + i\Delta, i = 0, \cdots, N(\Delta)\}, a_{N(\Delta)} = b, \sum_{i=1}^{N(\Delta)} P(a_i) = 1$. Further, defining P as the probability measure associated with a distribution function F, i.e.,

$$P(A) = \int_{A \in \mathcal{B}^{[a,b]}} dF(x) \tag{20}$$

where $\mathcal{B}^{[a,b]}$ is the Borel sigma-algebra generated by open sets in $[a, b]$, we state the following theorem,

Theorem 2: For any $F \in \mathcal{F}^{[a,b]}$, $\exists\ P^\Delta \in \mathcal{F}^\Delta$ s.t.

where $\mathbb{I}$ is any closed interval in $[a,b]$, $\mathbb{I}^\varepsilon = \{\tilde{x} : |x-\tilde{x}| < \varepsilon, x \in \mathbb{I}\}$ and P is the probability measure associated with the distribution function, F. Particularly, the P^Δ that satisfies (21) has the form,

$$P^\Delta(a_i) = F(a_i) - F(a_{i-1}) \tag{22}$$

where a_i's are as defined above.

IV. Analysis

Definition 3: For a bounded continuous Lipschitz loss function Λ with

$$\lambda(\Delta, y) = \sup_x \sup_{x':|x-x'|<\Delta} |\Lambda(x,y) - \Lambda(x',y)| \tag{23}$$

$$\lambda(\Delta) = \sup_y \lambda(\Delta, y) \tag{24}$$

let

$$\| \Lambda \|_L = \sup_{0<\Delta<(b-a)} \frac{\lambda(\Delta)}{\Delta} \tag{25}$$

Definition 4: For a channel which satisfies the continuity condition $\lim_{\Delta\to 0} \delta_\Delta = 0$ where

$$\delta_\Delta = \sup_{x\in[a,b]} \sup_{\substack{\hat{x}\in[a,b] \\ |x-\hat{x}|\le\Delta}} \int \underbrace{\left|f_{Y|x}(y) - f_{Y|\hat{x}}(y)\right|}_{\varepsilon_\Delta(y)} dy \tag{26}$$

let

$$\| \delta \|_L = \sup_{0<\Delta<(b-a)} \frac{\delta_\Delta}{\Delta} \tag{27}$$

Lemma 3: For any $F, \hat{F} \in \mathcal{F}^{[a,b]}$, $U \sim F$, a channel $\mathcal{C}$ s.t. $\| \delta \|_L < \infty$ and a bounded Lipschitz loss function

$$\begin{aligned}&\left|E_{F\otimes\mathcal{C}}\Lambda(U, g(Y)) - E_{\hat{F}\otimes\mathcal{C}}\Lambda(U, g(Y))\right| \\ &\le (\| \Lambda \|_L + \Lambda_{\max} \| \delta \|_L + (b-a) \| \Lambda \|_L \| \delta \|_L + \\ &\qquad \Lambda_{\max})\, \beta\left(P, \hat{P}\right)\end{aligned}$$

where P and $\hat{P}$ are the laws associated with F and $\hat{F}$.

Lemma 4: For any $\Delta > 0$, $F \in \mathcal{F}^{[a,b]}$, $U \sim F$ with the associated measure P, $P^\Delta \in \mathcal{F}^\Delta$ and a continuous bounded loss function

$$\begin{aligned}&|E_{P^\Delta\otimes\mathcal{C}}\Lambda(U, g(Y)) - E_{F\otimes\mathcal{C}}\Lambda(U, g(Y))| \le \\ &\qquad \delta_\Delta \Lambda_{\max} + \lambda(\Delta)\,(1+\delta_\Delta)\end{aligned}$$

where $\lambda(\Delta)$ is the global modulus of continuity of the loss function Λ as defined in equation (23) and δ_Δ is as defined in equation (26).

Lemma 5: For every $n \ge 1$, $x^n \in [a,b]^n$, measurable $g : \mathbb{R} \to [a,b]$, and $\varepsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \Lambda(x_i, g(Y_i)) - E_{F_{x^n}\otimes\mathcal{C}}\Lambda(U, g(Y))\right| > \epsilon\right) \tag{28}$$

$$\le A(\epsilon, \Lambda_{\max}) \exp(-G(\epsilon, \Lambda_{\max}) n) \tag{29}$$

where,

$$A(\epsilon, B) = \exp\left(\frac{2\epsilon^2}{B^2}\right), \qquad G(\epsilon, B) = \frac{2\epsilon^2}{B^2} \tag{30}$$

Let $\gamma = (\| \Lambda \|_L + \Lambda_{\max} \| \delta \|_L + (b-a) \| \Lambda \|_L \| \delta \|_L + \Lambda_{\max})$

Theorem 3: For all $\epsilon > 0$, $\rho \in (0,1)$, $\delta > 0$, $\Delta > 0$ and $x^n \in [a,b]^n$

$$\begin{aligned}&P\left(|L_{\tilde{X}^{n,\delta,\Delta}}(x^n, Y^n) - D(x^n)| > \right. \\ &\qquad \left. 3\epsilon + 5\delta\Lambda_{\max} + 4\delta_\Delta\Lambda_{\max} + 4\lambda(\Delta)(1+\delta_\Delta)\right) \\ &\le |\mathcal{G}_{\delta,\Delta}| \left[A(\epsilon + \delta\Lambda_{\max}, \Lambda_{\max}) e^{-G(\epsilon+\delta\Lambda_{\max},\Lambda_{\max})n} + e^{-(1-\rho)\frac{n\gamma^2}{2}}\right] \\ &\qquad + e^{-(1-\rho)\frac{n\gamma^2}{2}}, \qquad \text{for all } n > n_0\left(\mathcal{C}, n, K, \{h\}, \frac{\epsilon}{\gamma}\right)\end{aligned}$$

where

$$\tilde{X}^{n,\delta,\Delta}[y^n](i) = g_{opt}[\tilde{P}_{x^n}^{\delta,\Delta}[y^n]](y_i), \qquad 1 \le i \le n \tag{31}$$

$\tilde{P}_{x^n}^{\delta,\Delta}$ is the quantized version of $\hat{P}_{x^n}^\Delta$, the closest member in $\mathcal{F}^\Delta$ to $\hat{F}_{x^n}$ in the sense defined in section III C and

$$\tilde{P}_{x^n}^{\delta,\Delta} = Q_\delta(\hat{P}_{x^n}^\Delta) \tag{32}$$

Let $\mathcal{F}_{\delta,\Delta}$ denote the set of scalars with components in [0,1] that are integer multiples of δ. Note that $\tilde{P}_{x^n}^{\delta,\Delta}[y^n] \in \mathcal{F}_{\delta,\Delta}$ for all y^n. Also, let $\mathcal{G}_{\delta,\Delta} = \{g_{opt}[P]\}_{P\in\mathcal{F}_{\delta,\Delta}}$ (extending the definition of g_{opt} in (9) to quantized distributions).
Take now, $\delta = \delta_n, \Delta = \Delta_n$ such that $\delta_n \downarrow 0, \Delta_n \downarrow 0$ for all $\epsilon > 0$ and $\sum_{n=1}^\infty (1/\delta_n)^{\Delta_n} A(\epsilon + \delta_n\Lambda_{\max}, \Lambda_{\max}) e^{-G(\epsilon+\delta_n\Lambda_{\max},\Lambda_{\max})n} < \infty$. Let

$$\hat{X}_{\text{univ}}^n = \tilde{X}^{n,\delta,\Delta} \tag{33}$$

A direct consequence of Theorem 3 and the Borel-Cantelli Lemma gives us the main theorem.

Theorem 4: For all $\mathbf{x} \in \mathbb{R}^\infty$,

$$\lim_{n\to\infty}\left[L_{\hat{X}_{\text{univ}}^n}(x^n, Y^n) - D(x^n)\right] = 0 \qquad a.s. \tag{34}$$

V. Extension to $2k+1$-context length denoiser

In this section, we propose an extension of the symbol-by-symbol denoiser discussed in the earlier sections to the $2k+1$-length sliding window denoising scheme. The scheme is pictorially depicted in the figure below. The $2k+1$-tuple super-

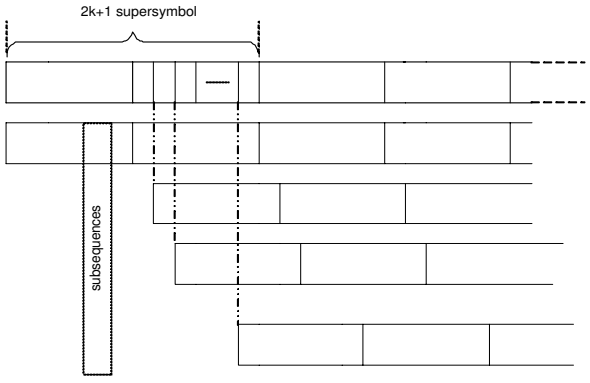

Fig. 1. Schematic representation of the $2k+1$-window denoiser

symbol is formed by jumping a length of $2k+1$ to achieve the independence condition between the super-symbols. This facilitates the extension of the ideas from the symbols of the symbol-by-symbol denoiser to the super-symbol of the $2k+1$ sliding window denoiser. As seen in Fig. 1 , this subsequencing gives rise to $2k+1$ subsequences (of supersymbols of length $2k+1$). As in the symbol-by-symbol scheme, let $\tilde{P}_{x^n}^{\delta,\Delta,k}$ denote the estimate of the $2k+1$-th order input empirical distribution of the source and the denoiser is defined as

$$\tilde{X}^{n,\delta,\Delta,k}[y^n](i) = g_{opt}[\tilde{P}_{x^n}^{\delta,\Delta,k}[y^n]](y), \qquad k+1 \le i \le n-k \tag{35}$$

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

$$D_k(x^n) = \min_g E\left[\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda(x_i, g(Y_{i-k}^{i+k}))\right] \tag{36}$$

Note, this is the $2k+1$ analog of $D(x^n)$ defined in equation (3). As before, $D_k(x^n)$ can be expressed as

$$D_k(x^n) = \min_g E_{F_{x^n}^k \otimes \mathcal{C}}\Lambda(X, g(Y_{-k}^k)) \tag{37}$$

where $F_{x^n}^k$ is the $2k+1$-th order empirical distribution of the source. Again, let $k = k_n \to \infty, \delta = \delta_n \downarrow 0, \Delta = \Delta_n \downarrow 0$ s.t. we have summability in the $2k+1$-th order analog of the inequality in (29) over n, [6], [4] and denote

$$\hat{X}_{\text{univ}}^n = \tilde{X}^{n,\delta,\Delta,k} \tag{38}$$

Theorem 5: For all $\mathbf{x} \in \mathbb{R}^\infty$, k

$$\limsup_{n\to\infty}\left[L_{\hat{X}_{\text{univ}}^n}(x^n, Y^n) - D_k(x^n)\right] \le 0 \qquad a.s. \tag{39}$$

Our results also imply optimality for the stochastic setting when the source (clean signal) is now a stationary process, $\mathbf{X}$, with distribution $F_{\mathbf{X}}$. Defining $\mathbb{D}(F_{\mathbf{X}}, \mathcal{C})$ as

$$\mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) = \lim_{n\to\infty}\min_{\hat{X}^n} EL_{\hat{X}^n}(X^n, Y^n) \tag{40}$$

where, the expectation is assuming X^n are the first n symbols of the source with distribution $F_{\mathbf{X}}$ and Y^n is as defined before.

Theorem 6: For all stationary $\mathbf{X}$

$$\lim_{n\to\infty} EL_{\hat{X}_{\text{univ}}}(X^n, Y^n) = \mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \tag{41}$$

VI. Experimental Results

Results of applying the proposed scheme to a natural test image, shown in Fig. 2, and presented in the table below. The image is corrupted by an AWGN source with $\sigma = 20$. As can be seen from the figure, arguably the denoised image reproduces the contrast of the original 'clean' image better than that on the lower-right corner. The context '$2k+1$'$= 2$ indicates knowledge of one adjoining pixel from the left while higher context lengths considers knowledge of the 2-D neighborhood of the noisy pixel being denoised. The table below shows successive improvements of the scheme with increasing lengths of the context, k. It essentially attains with $2k+1$ as low as 5, the performance of the scheme in [1] which is specifically tuned for AWGN with squared-loss metric. For less standard (but no less realistic) noise models and loss metrics, the proposed technique outperforms many of the state-of-the-art denoising schemes, as is discussed in [9].

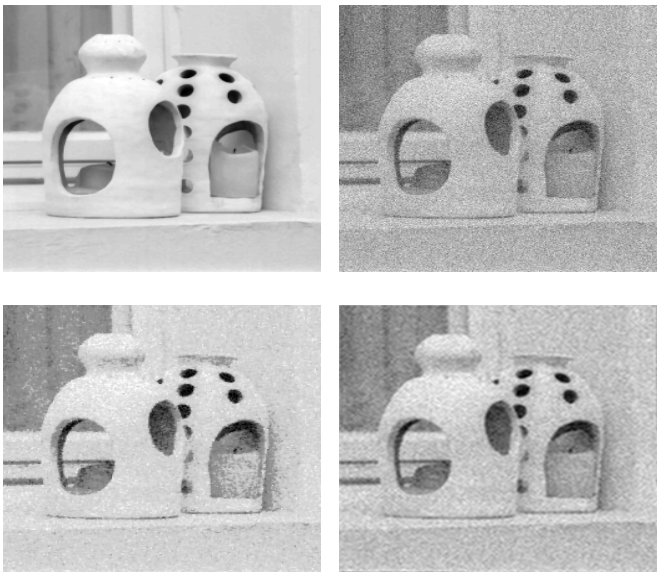

Fig. 2. Top-left: Original image, top-right: Noisy image (corrupted by an AWGN, $\sigma = 20$), bottom-left: Denoised image using the proposed scheme ($2k+1 = 5$), bottom right: Denoised image using wavelet-based soft-thresholding [1]

scheme	wavelet [1]	sym-sym	'2k+1' =2	2k+1 = 3	2k+1 = 5
RMSE	9.5359	15.4143	13.0226	10.7619	9.6207

TABLE I

Root Mean squared error (RMSE) in denoising

References

[1] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transcations on Information Theory*, vol. 41, pp. 613–627, May. 1995.

[2] A. Buades, B. Coll, and J. M. Morel, "A review of image de-noising algorithms with a new one," *Multiscale Modeling and Simulation*, vol. 4, pp. 490–530, Jul. 2005.

[3] T. Weissman, E. Orendtlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Transcations on Information Theory*, vol. 51, pp. 1229 – 1246, Jan. 2005.

[4] A. Dembo and T. Weissman, "Universal denoising for the finite input general output channel," *IEEE Transcations on Information Theory*, vol. 51, pp. 1507 – 1517, Apr. 2005.

[5] G. Motta, E. Ordentlich, I. Ramirez, G. Seroussi, and M. J. Weinberger, "The DUDE framework for continuous tone image denoising," in *2005 IEEE International Conference on Image Processing*, Sept. 2005.

[6] K. Sivaramakrishnan and T. Weissman, "Universal denoising of discrete-time continuous amplitude signals," in preparation.

[7] L. Devroye and L. Gyorfi, *Nonparametric Denisty Estimation, the L_1 view*, Wiley Series in Probability and Mathematical Statistics, New York, NY, 1985.

[8] L. Devroye, *A Course in Density Estimation*, Birkhauser, Boston, MA, 1987.

[9] K. Sivaramakrishnan and T. Weissman, "Universal denoising of continuous valued signals with applications to images," to appear in proceedings of *2006 International Conference on Image Processing*.