A Context Quantization Approach to Universal Denoising

Kamakshi Sivaramakrishnan, *Member, IEEE*, and Tsachy Weissman, *Senior Member, IEEE*

***Abstract*—We revisit the problem of denoising a discrete-time, continuous-amplitude signal corrupted by a known memoryless channel. By modifying our earlier approach to the problem, we obtain a scheme that is much more tractable than the original one and at the same time retains the universal optimality properties. The universality refers to the fact that the proposed denoiser asymptotically (with increasing block length of the data) achieves the performance of an optimum denoiser that has full knowledge of the distribution of a source generating the underlying clean sequence; the only restriction being that the distribution is stationary. The optimality, in a sense we will make precise, of the denoiser also holds in the case where the underlying clean sequence is unknown and deterministic and the only source of randomness is in the noise. The schemes involve a simple preprocessing step of quantizing the noisy symbols to generate quantized *contexts*. The quantized context value corresponding to each sequence component is then used to partition the unquantized symbols into subsequences. A universal symbol-by-symbol denoiser (for unquantized sequences) is then separately employed on each of the subsequences. We identify a rate at which the context length and quantization resolution should be increased so that the resulting scheme is universal. The proposed family of schemes is computationally attractive with an upper bound on complexity which is independent of the context length and the quantization resolution. Initial experimentation seems to indicate that these schemes are not only superior from a computational viewpoint, but also achieve better denoising in practice.**

***Index Terms*—Denoisability, kernel density estimation, linear programming, memoryless channel, context, quantized context, semi-stochastic setting, sliding window denoiser, symbol-by-symbol denoiser, universal denoising.**

I. Introduction

CONSIDER the problem of estimating the discrete-time clean signal $\{X_t\}_{t\in\mathbb{T}}, X_t \in [a,b] \subset \mathbb{R}$, from its noisy observations $\{Y_t\}_{t\in\mathbb{T}}, Y_t \in \mathbb{R}$, where $\{Y_t\}$ is the output of a known, memoryless channel whose input is $\{X_t\}$. A specific instance of this problem where the noise is additive in nature and Normal (Gaussian) distributed is very well studied in the literature. Solutions to this specific class of denoising problems date as far back as Wiener filtering in [1] where the underlying clean signal-generating source is assumed to be known. A plethora of schemes have since been proposed on various adaptations of this fundamental problem where, assumptions on knowledge of the clean signal were relaxed. Notable works in this regard are the wavelet shrinkage techniques for denoising proposed in [2]. The approach in [2] is based on two facts, the first one being—Gaussian white noise retains its properties of Gaussianity and whiteness under any orthogonal transformation. The second, certain *types* of signals have a sparse representation in the wavelet domain with most of the energy compacted into a small number of high magnitude coefficients. This framework has been used in subsequent works [3], [4] leading to significant improvements in the denoising performance. Recently, there have been attempts [5]–[8] at extending the regularized least squares estimation to denoising in the case where the noise continues to be additive but the statistics are unknown. An important limitation of these schemes is that performance guarantees hold almost exclusively for the case of additive and Gaussian statistics of the noise. An alternative perspective of compression-based denoising has been proposed in [9] and [10] among others. This approach is motivated by the fact that the hardest part to compress in a noisy signal is essentially the noise. Hence, a lossy compression scheme whose distortion level is tuned to match the noise level is set to meet the objective of eliminating noise. However, the algorithmic and performance disadvantages of such schemes have been discussed in [11]. To reiterate the performance disadvantages, it has been shown in [12] that compression-based schemes for universal denoising fall short of the optimal distribution-dependent denoiser. This motivates the work presented in this paper of universal denoising schemes for continuous valued signals that are of manageable complexity. The development of denoising techniques for noise corrupted images in many practical signal processing applications has resulted in a class of schemes that focus on a finite sequence length setting. Among the state-of-the-art approaches are the application of Gaussian scale mixture (GSM) models to wavelet coefficients [13], non local means (NLM) filtering [14], [15], block-matching and 3-D filtering [16], adaptive image filtering [30], [17], shape adaptive DCT [18], SVD based learning approaches of [19] and intelligent local smoothing based denoising in [20]. The NLM filtering is a contextual scheme that constructs a globally smoothed estimate of the clean signal by weighting all the noisy observations. It was shown that these global smoothed estimates are good proxies for the optimal denoiser in the case of minimum mean-square loss based reconstruction, viz., the Bayes estimate

Manuscript received February 20, 2008; accepted November 16, 2008. First published December 31, 2008; current version published May 15, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pierre Vandergheynst. Work supported in part by National Science Foundation through Grants CCR-0311633 and the NSF CAREER.

K. Sivaramakrishnan was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. She is now with Admob Inc., San Mateo, CA 94401 USA (e-mail: kamakshis@gmail.com).

T. Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA, and also with the Technion—Israel Institute of Technology, Haifa, 32000, Israel (e-mail: tsachy@stanford.edu).

Digital Object Identifier 10.1109/TSP.2008.2011847

of the underlying clean signal. These techniques lend themselves to good results for the particular case where the noise is additive and Gaussian distributed. There are, however, scientific and practical implications to the study and development of generalized universal denoising schemes that do not make any assumptions on the clean signal statistics. There have been a spate of recent results in this direction triggered by the universal denoising scheme for discrete valued data (also known as DUDE) proposed in [11]. The universality and asymptotic optimality of the DUDE has parallels in the literature in the form of Empirical Bayesian approach to statistical decision problems in [21]. As pointed in [11], the DUDE has a similar spirit of estimating the input empirical distribution as in the compound decision approach in [21] and [22]. It is also important to point out that the *Bayes envelope* functional defined in [21] (and similarly in [11]) is done so for the symbol-by-symbol case. Both the DUDE in [11] and [23] have extended the Bayes envelope functional to the case where the loss is measured over joint distributions of arbitrary order on entire sequences. Similar to the flavor of the DUDE like schemes (as in the current work and [23]), the empirical Bayes approach in [22] can be viewed as a compound decision problem that competes with a genie with access to the unknown finite number of hypotheses (prior distributions). The work in [21] also has a discussion similar to that in [23] regarding convergence of the estimated prior distribution to the true prior of the decision variable. As already clarified in [11], the DUDE like schemes are nonsequential versions of the compound decision problems in [22] in the symbol-by-symbol setting. A detailed discussion of the similarities between the symbol-by-symbol approach of the DUDE and the compound decision problems in [22] is found in [11] (and references therein). The current work serves to enhance the scope of DUDE (and hence, Empirical Bayes approach) beyond the symbol-by-symbol setting to continuous valued data with performance guarantees that are stronger (than the work in [21]) due to the ability of the proposed algorithm to learn higher order joint distributions.

Recently, universal denoising for continuous valued signals and channels was considered in [23]–[27]. These approaches were motivated by the DUDE framework in [11] and [28] and nonparametric techniques in density estimation [29]. This framework consists of a "two-pass" approach in which the first pass involves accruing the statistics of the noisy sequence and using knowledge of the channel to then estimate statistics of the underlying clean signal itself. The second pass is one where, having learned the statistics, denoising is carried out to minimize the loss under a user-specified loss function. A similar two-pass approach was presented in [30] wherein image denoising was considered for the specific case of a Ricean noise model. The approach was based on principles of Empirical Bayesian estimation techniques with an interesting analogue of the first step to that in the DUDE-like schemes. A nonparametric Markov random field (MRF) described the prior distribution of the underlying clean signal and the approach used the expectation-maximization algorithm to estimate the same. The approach demonstrated compelling results in the case of MRI image denoising as a specific assumption of the noise model was made. Hence, universality of this scheme was established under specific assumptions on the noise model. Universal optimality of the denoisers in [23] and [24], however, was established in a generality that applies to arbitrarily distributed clean signals, and *arbitrary but known memoryless channels (noise models)* and loss functions (with some benign regularity conditions). The theoretical optimality results were also translated to some encouraging, practically implementable schemes discussed in [23] and [25]. However, this family of denoisers suffers from computational issues that could render the scheme unattractive for real-time applications. The denoiser in [23] learns higher order statistics by considering groups of symbols in the noisy sequence. For a chosen member in a group, the remaining symbols form a context. In other words, a context is a collection of symbols from the neighborhood of a certain size/shape associated with every symbol in a sequence. The sparsity of occurrences of these contexts in large alphabet size or the continuous valued problems necessitates borrowing from the knowledge of "similar" contexts. The denoiser in [23] uses a natural context aggregation (or borrowing) from similar contexts by using nonparametric density estimation techniques to learn the statistics. The order of the denoiser is characterized and given by the order of the statistics of the noisy sequence it learns. The full benefits of the approach in [23] are achieved only with increasing order of the denoiser which come at the expense of computational intractabilities and related statistical sparsity issues. It did, however, provide a paradigm to address denoising of real valued signals from the view point of universal asymptotic optimality. There has been work in extending the application of the DUDE to address the computational intractability and its application to denoising (large alphabet size) gray scale images in [27]. The denoising performance in [27] leaves room for improvement in comparison to many of the state of the art schemes in [13], [16]–[18], [30]. In addition, the nature of the work in [27] is more experimental in nature that does not discuss theoretical guarantees and analysis of the proposed scheme. Much of the work in this paper is motivated by the need for an extended framework that handles continuous valued and large alphabet size signals and like the DUDE in [11], [28] provides concrete performance guarantees.

The symbol-by-symbol denoiser (which makes decisions using only the marginal statistics) discussed in [23], however, is computationally attractive. One possible approach to reduce the computational burden could be to quantize the continuous valued noisy symbols (as proposed in [28] for the discrete-input general output setting) and then apply higher order denoisers. This would, however, lead to suboptimal performance since the optimum scheme denoises based on the unquantized values. In this paper, we propose a middle ground solution using quantization only for the contexts of the symbols in the noisy sequence. We quantize the contexts of every noisy symbol resulting in classes of quantized contexts and apply the symbol-by-symbol denoiser of [23] to the (unquantized) subsequences associated with each class. This approach emulates the higher order functionality of the sliding window denoisers through the quantized contexts while still maintaining the low complexity of the symbol-by-symbol denoiser of [23] as it is applied within each context subsequence. The complexity of the denoisers obtained in this way is not only polynomial in the data size n for a fixed

context length k and quantization resolution M but in fact is bounded with a proportionality constant that does not depend on k and M. This is in stark contrast to the complexity of the scheme in [23] which, although polynomial in n, is exponential in k and is, consequently, formidable for even moderate values of k. The natural question is whether the new denoiser we propose, beyond its superiority over that of [23] from a computational standpoint, preserves the asymptotic optimality and universality properties of the denoiser in [23]. In this paper, we answer this question in the affirmative. We address the answer to this question by considering two scenarios, the first being the semi-stochastic setting, where the underlying clean sequence is an unknown, deterministic sequence. In this setting, we first fix the number of quantization levels M of the context and the context length k. We then demonstrate the asymptotic optimality of the proposed denoisers w.r.t. a sequence of denoisers that also make their decisions based on the quantized noisy contexts, but applying an optimal symbol-by-symbol denoiser to the middle symbol that uses the true statistics of the underlying clean sequence. In the second step, we *appropriately* grow the context length k and the number of quantization levels M with the data size. This then enables us to establish the fact that, the proposed sequence of denoisers asymptotically (for n large) achieves the performance of any sliding window scheme of arbitrary order, on the noisy sequence that has the luxury of knowledge of the true distribution of the underlying clean sequence.

The remainder of the paper is organized as follows. In Section II, we discuss the problem setup and notations. This is followed by Section III with a brief discussion of the denoiser in [23] and some key technical results therein. Section IV details the construction of the proposed denoiser and some of its performance guarantees for the semi-stochastic setting. We then proceed to analyze the scheme in the fully stochastic setting where the underlying clean sequence is now a stochastic process rather than an individual sequence. To extend the optimality results to the fully stochastic setting, we characterize the necessary conditions of the quantizer in Section V. Conditions on the quantizer are imposed in the form of growth rates in the quantization resolution with increasing block lengths of the noisy sequence. Another condition on the quantizer is that the sequence of partitions generated is asymptotically fine. This is formalized and defined in Section V but it goes to say that (with increasing block lengths) the quantizer must "faithfully" learn the distribution of the noisy sequence from its quantized counterparts. Section VI discusses some promising and competitive experimental results of the application of the proposed denoiser in the case of additive white Gaussian noise (AWGN) channels. Proofs and associated details for the theorems and lemmas are given in the Appendixes.

II. Problem Setting and Notations

Let $\mathbf{x} = (x_1, x_2, \ldots)$ be an individual (or deterministic) noise-free source signal (sequence) with components taking values in $[a,b] \subset \mathbb{R}$ and $\mathbf{Y} = (Y_1, Y_2, \ldots), Y_i \in \mathbb{R}$ be the corresponding noisy observations, also referred to as the output of the channel (corruption source). This setting, where both the underlying clean sequence and the noisy sequence are real-valued, is the continuous-alphabet analog of the semi-stochastic setting discussed in [28]. The channel 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. Memoryless here means that the components Y_i in the noisy sequence $\mathbf{Y}$ are independent with each Y_i distributed according to $F_{Y\,|\,x_i}$. Also, we denote the probability measure on $\mathbb{R}$ corresponding to $F_{Y\,|\,x}$ by μ_x. Let $x^n = (x_1, \ldots, x_n)$ and $Y^n = (Y_1, \ldots, Y_n)$ denote n-tuples. An *n-block* denoiser, $\hat{X}^n(y^n)$, 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 define the normalized cumulative loss of the n-block denoiser 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 the underlying sequence is x^n and the observed noisy sequence is y^n. $\hat{X}^n(y^n)[i]$ denotes the ith component of $\hat{X}^n(y^n)$. We impose continuity and boundedness conditions on both the channel $\mathcal{C}$ and the loss function Λ. The distribution functions $F_{Y\,|\,x}$ are also assumed to be absolutely continuous for all $x \in [a,b]$ w.r.t. the Lebesgue measure and $\{f_{Y\,|\,x}\}$ denotes the corresponding densities. We defer to Appendix A for a detailed discussion and specifications of these conditions but briefly motivate them here. These are benign conditions ensuring the invertibility and its smoothness in a sense that (makes it possible to estimate the empirical distribution of the underlying clean sequence). Let $\mathcal{F}^{[a,b]}$ denote 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 on the loss function detailed in Appendix A imply existence of the minimum). In other words, $\mathcal{U}(F)$ denotes the minimum achievable expected loss when guessing the value of $X \sim F$. Define the symbol-by-symbol minimum loss of x^n by

$$D_0(x^n) = \min_g E\left[\frac{1}{n}\sum_{i=1}^{n} \Lambda(x_i, g(Y_i))\right] \tag{3}$$

where the minimum is over all measurable maps $g : \mathbb{R} \to [a,b]$. $D_0(x^n)$ denotes the minimum expected loss in denoising the sequence x^n, using a time-invariant symbol-by-symbol rule. This can be attained by a "genie" with access to the clean sequence x^n. We use $D_0(x^n)$ as a benchmark for assessing the performance of a universal symbol-by-symbol denoiser that we construct in the next section. 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. Note that $D_0(x^n)$ can be expressed as

$$D_0(x^n) = \min_g \int_{[a,b]} E_x \Lambda(x, g(Y)) dF_{X^n}(x) \tag{5}$$

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, probability and expectation when the channel input $X \sim F$ and Y is the channel output. So that

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

Letting $[F \otimes \mathcal{C}]_{X\,|\,y}$ denote the conditional distribution of X given $Y = y$ under $F \otimes \mathcal{C}$, we have

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

where $\mathcal{U}$ denotes the Bayes envelope as defined in (2). Letting $g_{\mathrm{opt}}[F]$ denote the achiever of the minimum in (8) (which exists under our assumptions on the loss function), we note that it is given by the Bayes response to $[F \otimes \mathcal{C}]_{X\,|\,y}$, namely,

$$\begin{aligned} g_{\mathrm{opt}}[F](y) &= \arg\min_{\hat{x}\in[a,b]} \int_{[a,b]} \Lambda(x,\hat{x})d[F\otimes\mathcal{C}]_{X\,|\,y}(x) \\ &= \arg\min_{\hat{x}\in[a,b]} \int_{[a,b]} \Lambda(x,\hat{x})f_{Y\,|\,x}(y)dF(x). \end{aligned} \tag{9}$$

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

$$D_0(x^n) = \min_g E_{F_{x^n}\otimes\mathcal{C}}\Lambda(X,g(Y)) \tag{10}$$

where F_{x^n} was defined in (4) and the minimum is attained by $g_{\mathrm{opt}}[F_{x^n}]$. Thus, only a "genie" with access to the empirical distribution of the noiseless sequence could employ $g_{\mathrm{opt}}[F_{x^n}]$.

III. A Competitive Symbol-by-Symbol Denoiser and Its Extension

In this section, we discuss the construction of a *universal* "symbol-by-symbol" denoiser and establish its asymptotic optimality by comparing its performance to that of a "genie-aided expert." This expert, with access to the *true* empirical distribution of the underlying clean sequence, chooses the optimal mapping between each component Y_i of the noisy sequence and an estimate in the interval $[a,b]$. The *universality* of the denoiser is in the fact that, for any reasonably well-behaved noise distribution and loss function, it attains the performance of the said genie of the underlying clean sequence. We begin by conceptually constructing the symbol-by-symbol denoiser, following it up with a concrete discussion of the implementation steps. We also briefly present some performance guarantees of the proposed denoiser, referring the reader to [23] for more details and related proofs. Finally, we discuss the extension of the symbol-by-symbol denoiser to higher order sliding window schemes which achieve better performance. This is made possible by learning and exploiting the knowledge of higher order statistics of the underlying clean sequence. The greater performance benefits of this extension come accompanied with higher computational requirements, the details of which are also discussed.

A. Construction of Universal "Symbol-by-Symbol" Denoiser and Preliminaries

F_{x^n} and, hence, $g_{\mathrm{opt}}[F_{x^n}]$ are not known to an observer of the noisy sequence. The first step toward constructing an estimate of $g_{\mathrm{opt}}[F_{x^n}]$ is to estimate the input empirical distribution F_{x^n} from the observable noisy sequence Y^n and knowledge of the channel $\mathcal{C}$. We approach this problem by first estimating a function that tracks the evolution of the "average" density function according to which the noisy symbols are distributed. For an input sequence x^n, given the memoryless nature of the channel, the output symbols will be independent with respective distributions $\{F_{Y\,|\,x_1},\ldots,F_{Y\,|\,x_n}\}$ and have the corresponding density functions $\{f_{Y\,|\,x_1},\ldots,f_{Y\,|\,x_n}\}$. The first function we are interested in estimating is

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

which can be thought of as the marginal density of the noisy symbols in the semi-stochastic setting where x^n is the unknown deterministic sequence. The estimation of this function is done by exploiting the vast literature on density estimation techniques [29], [31]. 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 $\mathbb{R}$ having conditional densities $f_{Y\,|\,x_1}, f_{Y\,|\,x_2}, \ldots, f_{Y\,|\,x_n}$, respectively. A density estimate $f_Y^n(y) = f(y; Y_1, \ldots, Y_n)$ is a real-valued Borel measurable function of its arguments and for fixed n, f_Y^n is a density on $\mathbb{R}$. The kernel density estimate is given by

$$f_Y^n(y) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{y-Y_i}{h}\right) \tag{12}$$

where $h = h_n$ is a sequence of positive numbers, such that $\lim_{n\to\infty} nh_n = \infty$, and K is a Borel measurable function satisfying $K \geq 0, \int K = 1$. Once we have an estimate f_Y^n of the output marginal density we use it to estimate the input empirical distribution by

$$\hat{F}_{x^n} = \hat{F}_{x^n}[Y^n] = \arg\min_{F\in\mathcal{F}_n^{[a,b]}} d\left(f_Y^n, \underbrace{\int f_{Y\,|\,x}dF(x)}_{[F\otimes\mathcal{C}]_Y}\right) \tag{13}$$

where $\mathcal{F}_n^{[a,b]} \subseteq \mathcal{F}^{[a,b]}$ denotes the set of empirical distributions induced by n-tuples with $[a,b]$-valued components and $[F \otimes \mathcal{C}]_Y$ denotes the marginal density induced at the output of the channel by an input distribution F. That is, every member $F(x)$ of $\mathcal{F}_n^{[a,b]}$ is of the form

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(x_i\leq x)} \tag{14}$$

for some n-tuple, $x^n = (x_1, x_2, \ldots, x_n)$, with $[a,b]$-valued components. The norm, d, in (13) is the L_1 norm defined as

$$d(f,g) = \int |f(y) - g(y)|\, dy. \tag{15}$$

The channel $\mathcal{C}$ induces a set of "feasible" densities of the output noisy symbol corresponding to the family of empirical distributions of the underlying clean sequence at the input of the channel. The density estimate f_Y^n, which is constructed only from the noisy sequence Y^n, is oblivious to this set of feasible marginal densities and hence could lie outside it. It is thus natural to estimate the unobserved F_{x^n} by the member of $\mathcal{F}_n^{[a,b]}$ leading to a channel output distribution closest to the estimated one f_Y^n. This is exactly the estimate in (13). The uniqueness of the minimizer in (13) follows from the invertibility of the channel and the fact that the objective function being minimized is a norm-function and hence convex.

A two-stage quantization of both the support of the underlying clean symbol $[a,b]$ and the values of the estimate of its empirical distribution function $\hat{F}_{x^n}$ is carried out to give the estimated probability mass function that has mass points only at the quantized symbols. Applying this quantization of the support of the underlying clean symbol to the estimate $\hat{F}_{x^n}$, we construct now the corresponding probability mass function $\hat{P}_{x^n}^{\Delta}$

$$\hat{P}_{x^n}^{\Delta}(a_i) = \hat{F}_{x^n}(a_i) - \hat{F}_{x^n}(a_{i-1}) \tag{16}$$

where $a_i \in \mathcal{A}^{\Delta}$. The values of $\hat{P}_{x^n}^{\Delta}$ are quantized using a uniform quantizer, Q_δ, to give its corresponding quantized representation $\hat{P}_{x^n}^{\delta,\Delta}$. Specifically

$$\hat{P}_{x^n}^{\delta,\Delta}(a_i) = Q_\delta\left(\hat{P}_{x^n}^{\Delta}\right) \tag{17}$$

where $Q_\delta : [0,1] \to \{\delta k, k = 0, 1, \ldots, \lfloor 1/\delta \rfloor\}$ and δ is the quantization resolution for representing the distribution function values. The minimizer of the Bayes envelope in (9) is then constructed from the quantized probability mass function $\hat{P}_{x^n}^{\delta,\Delta}$ as $g_{\text{opt}}[\hat{P}_{x^n}^{\delta,\Delta}]$, where in this case g_{opt} assumes the form

$$g_{\text{opt}}[P](y) = \arg\min_{\hat{x}\in[a,b]} \sum_{a\in\mathcal{A}^{\Delta}} \Lambda(a,\hat{x}) f_{Y\,|\,x=a}(y) \cdot P(X=a). \tag{18}$$

$\mathcal{A}^{\Delta}$ is finite alphabet approximation of $[a,b]$ corresponding to the quantization step size of Δ. Note that we have extended the definition of g_{opt} to accommodate the case when P is not a valid probability, i.e., $\hat{P}_{x^n}^{\delta,\Delta}$ (it does not necessarily sum up to 1). Equipped with $\hat{P}_{x^n}^{\delta,\Delta}$, the n-block symbol-by-symbol denoiser is naturally defined by

$$\tilde{X}^{n,\delta,\Delta}(y^n)[i] = g_{\text{opt}}\left[\hat{P}_{x^n}^{\delta,\Delta}\right](y_i), \quad 1 \le i \le n \tag{19}$$

where g_{opt} is given in (18). Finally, the suggested symbol-by-symbol denoiser is

$$\hat{X}_{\text{ss univ}}^{n} = \tilde{X}^{n,\delta,\Delta} \tag{20}$$

for $\delta = \delta_n = \log n$ and $\Delta = \Delta_n = \log n$.

B. Implementation of the Symbol-by-Symbol Denoiser

The implementation of the denoiser in the previous section involves a discretization of the density estimation and the channel inversion steps. The discretized version of the kernel density estimate $f_Y^n(y)$ in (12) is evaluated at a set of discrete points $\{y_1, \ldots, y_N\}$. This gives an N-dimensional vector of the distribution function $p_Y^n(y)$. The "channel inversion" in (13) is also discretized using the estimate $p_Y^n(y)$.

1) Fast Kernel Density Estimation: The Kernel density estimation in (12) for a given kernel function K, although simple in construction, for a brute-force computation, is faced with computational burden on the order of $O(Nn)$ corresponding to n data points and N points $\{y_1, \ldots, y_N\}$ at which $p_Y^n(y)$ is evaluated. The computational complexity can be greatly reduced by using FFT based methods [32]. Recently, there has been extensive work on the use of fast Gauss transform-based techniques based on [33] for reduction of computational complexity. These techniques reduce the complexity from $O(Nn)$ to $O(N+n)$. This scheme uses a clever factorization of the terms in the evaluation of the kernel density estimate defined in (12) for the particular case of Gaussian kernel function K. This reduces the complexity of computation for (12) from $O(Nn)$ to $O(N+n)$ where N is the number of discrete points at which f_Y^n are calculated and n is the length of the noisy sequence. This is based on fast Gauss transform proposed in [33]. However, the constant factor in $O(n+N)$ grows exponentially with increasing context length k, which makes the algorithm impractical in higher dimensions. The work in [34] proposed an improved fast Gauss transform where the constant factor is reduced to asymptotically polynomial order. This is based on a preprocessing step which assigns the N source points into K clusters using the farthest-point clustering algorithm in [35]. The cardinal factor in nonparametric density estimation procedures is the choice of the *optimal* bandwidth h in (12). There has been some recent work in [36] on using dual-tree methods to derive fast methods for optimal bandwidth choice that continues to maintain the complexity of this step at $O(N+n)$. For $N = O(n)$, this reduces to $O(n)$.

2) Channel Inversion Using Linear Programming Techniques: In solving the channel inversion problem in (13), we are looking for a vector in the probability simplex, $\mathcal{F}^{\Delta} = \{P : \sum_{i=1}^{N(\Delta)} P(a_i), a_i \in \mathcal{A}^{\Delta}\}$, for our candidate distribution function, $\hat{P}_{x^n}^{\delta,\Delta}$. The discretized version of (13) is given by

$$\hat{P}_{x^n}^{\delta,\Delta} = \arg\min_{p\in\mathcal{F}^{\Delta}} \sum_{i=1}^{N} \left| p_Y^n(y_i) - \sum_{j=1}^{N(\Delta)} f_{Y\,|\,x=x_j}(y_i) Q_\delta(p(x_j)) \right|. \tag{21}$$

The objective function, being an L_1-norm, is clearly a convex function and the candidate minimizer also resides in the convex subspace, viz., the probability simplex $\mathcal{F}^{\Delta}$. This can be easily solved using well-studied linear programming algorithms in the broader area of convex optimization techniques. The two-pronged quantization discussed in the previous section can be naturally built into the optimization problem in (21) by

TABLE I
ALGORITHM I: SYMBOL-BY-SYMBOL DENOISER

input : Noisy Sequence y^n, Channel $\mathcal{C}$

output: Denoised Sequence, $\hat{x}^n$

FIRST PASS

Density Estimation Step

input : Noisy Sequence, y^n

output: Density Estimate, f_Y^n

Determine the optimal bandwidth from any one of the techniques discussed in [32], e.g., Cross-validation

Use techniques discussed in [36] for *fast* evaluation of (12)

Channel Inversion Step

input : $\hat{f}_Y^n$, Quantization resolutions, δ,Δ

output: $\hat{P}_{x^n}^{\delta,\Delta}$

Construct an LP (Linear Program) as in (22) and use `linprog` (in MATLAB) or any complex program solver to solve it. Alternatively, use log-barrier methods discussed in [38] to solve for the estimate, $\hat{F}_{x^n}$

Use the quantization mapping in (16) to map $\hat{F}_{x^n}$ to $\hat{P}_{x^n}^{\Delta}$

Then use a uniform quantizer with resolution δ to get $\hat{P}_{x^n}^{\delta,\Delta} \leftarrow Q_\delta\left(\hat{P}_{x^n}^{\Delta}\right)$

SECOND PASS

input : Noisy Sequence, y^n, Channel $\mathcal{C}$, estimate of input distribution $\hat{F}_{x^n}$

output: Denoised Sequence, $\hat{x}^n$

Use equation (18), (19) to denoise at every location, i

for $i \leftarrow 1$ **to** n **do**

  $\hat{x}_i \leftarrow g_{\text{opt}}[\hat{P}_{x^n}^{\delta,\Delta}](y_i)$

end

searching in $\mathcal{F}^{\delta,\Delta} = \{Q_\delta(P) : P \in \mathcal{F}^{\Delta}\}$, the set of N (Δ)-tuples with components in [0, 1] that are integer multiples of δ with point masses on the set $\mathcal{A}^{\Delta}$. The formulation would then be

$$\begin{aligned}\hat{P}_{x^n}^{\delta,\Delta} = &\arg\min_{p\in\mathcal{F}^{\delta,\Delta}} \sum_{i=1}^{N} \varepsilon_i \\ \text{s.t.}\quad & p_Y^n(y_i) - \sum_{j=1}^{N(\Delta)} f_{Y\,|\,x=x_j}(y_i)p(x_j) \le \varepsilon_i \\ & \sum_{j=1}^{N(\Delta)} f_{Y\,|\,x=x_j}(y_i)p(x_j) - p_Y^n(y_i) \le \varepsilon_i \\ & \qquad\qquad \forall i \in \{1,\ldots,N\}. \quad (22)\end{aligned}$$

The computational complexity of solving this problem using the popular interior point methods [37] is $O((N + N(\Delta))^3) = O((N + (1/\Delta))^3) = O((N + \log n)^3)$. This again, for $N = O(n)$, reduces to $O((n + \log n)^3) = O(n^3)$. This channel inversion is the heart of the denoiser in (18) and its simple formulation makes the scheme particularly elegant and attractive. The estimate of the empirical distribution in (21) is then plugged into (18) to finally give an estimate of the underlying clean symbol according to (19). The denoiser is described as Algorithm 1 (see Table I).

C. Analysis

In this section, we state the main result of [23] in the semi-stochastic setting, which establishes universal asymptotic optimality of the universal symbol-by-symbol denoiser in (20) for any unknown individual underlying clean sequence, $\mathbf{x}$.

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

$$\lim_{n\to\infty}\left[L_{\hat{X}^n_{\text{ss univ}}}(x^n, Y^n) - D_0(x^n)\right] = 0 \quad \text{a.s.} \qquad (23)$$

This theorem is a consequence of a concentration inequality (refer to Appendix B for the statement and [23] for further details) that bounds the deviation of the cumulative incurred by the symbol-by-symbol denoiser, $L_{\tilde{X}_{\text{ss univ}}}(x^n, Y^n)$ from $D_0(x^n)$.

TABLE II
ALGORITHM 2: FURTHER SUBSEQUENCING OF THE SUBSEQUENCE $x^{m/n}$

input : Noisy Subsequence $y^{m,n}$

output: Subsubsequences, $\{y^{m,n,l}\}_{l=1}^{2k+1}$

initialization: $\tilde{y}^{m,n} = y^{m,n}$;

for $l \leftarrow 1$ **to** $2k+1$ **do**

  $y^{m,n,l} \leftarrow \{\tilde{y}_j^{m,n} : d\left(\mathcal{M}(\tilde{y}_j^{m,n}), \mathcal{M}(\tilde{y}_1^{m,n})\right) \geq 2k+1\}$;

  $\tilde{y}^{m,n} \leftarrow \tilde{y}^{m,n} \setminus \tilde{y}^{m,n,l}$;

  Apply Algorithm 1 for each $\{y^{m,n,l}\}$

end

D. Extension to the $2k+1$-Window Length Denoiser and Associated Computational Complexity

The symbol-by-symbol denoiser discussed in the previous section aims at attaining the performance of a denoiser that bases its decisions on marginal, first-order statistics of the underlying clean sequence. It was shown in [11] and [23] that greater performance benefits are achieved by using a denoiser that makes decisions using higher order, joint statistics of the underlying clean sequence. This is based on the fact that the minimum possible loss as a function of the order k of the joint statistics is a monotonically (and, usually strictly) decreasing function. The extension of the symbol-by-symbol denoiser to the kth-order sliding window setting, involving the idea of subsequencing, has been discussed in detail in [23]. The subsequencing involves breaking up the sequence into subsequences of symbols that are $2k$ apart and then applying the symbol-by-symbol denoiser in Section III-A to $2k+1$-tuple super symbols (of a symbol and its kth order contexts). This extension entails learning higher order statistics which include the kth-order density of the noisy sequence Y^n, the kth-order channel inversion and the kth-order denoiser construction. These are kth-order equivalents of (12), (13), and (18), respectively. Precise rates of growth of the order k of the denoiser as a function of the block length of the underlying clean sequence are derived in [23]. We have only briefly touched upon the kth-order denoiser here, primarily to motivate the problem of increased computational burden and details of the construction are available in [23]. The complexity of the kernel density estimation increases with k as $O(Nk^{\gamma})$, where $\gamma > 1$ is a function of the accuracy in estimating the kth-order equivalent of the noisy statistics in (12) using IGFT [39], Dual-tree methods [40]. The channel inverse, like in the symbol-by-symbol case can be solved using linear programming techniques. The increase of dimensionality in the channel inversion problem results in searching on a $2k+1$-dimensional space which manifests itself in a computational complexity of $O(n^{6k})$ [37]. The exponential complexity with increasing context lengths (block lengths of data) renders the scheme with formidable computational burden. This leads us to discuss a modified kth-order sliding window scheme in the following section that achieves similar performance benefits while maintaining the computational advantages of the symbol-by-symbol scheme in Section III-B.

IV. UNIVERSAL DENOISING WITH QUANTIZED CONTEXTS

The kth-order sliding window denoiser discussed in Section III-D is in principle an elegant approach to the problem of universal denoising of continuous-amplitude data. This denoiser systematically approaches the problem using nonparametric techniques to learn a quantized version of the *a posteriori* distribution of the underlying clean symbol given its observed noisy context, $[\tilde{P}_{x^n}^{\delta,\Delta} \otimes \mathcal{C}]_{X \mid y_{-k}^k}$, as induced by the channel $\mathcal{C}$. Motivated, primarily, by the need to reduce the computational burden of the scheme in Section III-D, we propose and study the modification shown in Fig. 1. In words, we are proposing the following:

Fix a window size k and the number of quantization levels M

- Fix an M-level vector quantizer Q_M for the $2k$-length contexts, (y_{-k}^{-1}, y_1^k), that maps them to one of M possible $2k$-tuples, $(\hat{y}_{-k}^{-1}, \hat{y}_1^k)$. For a given M, let $y^M = \{y_1, \ldots, y_m, \ldots, y_M\}$ denote the set of $M, 2k$-tuples in which the quantized contexts, $(\hat{y}_{-k}^{-1}, \hat{y}_1^k)$ take values, i.e., $(\hat{y}_{-k}^{-1}, \hat{y}_1^k) \in y^M$. Hence, the quantizer is a mapping, $Q_M : \mathbb{R}^{2k} \rightarrow y^M$. Let n_m be the number of $2k$-tuples at quantization level, m, i.e., $n_m = |\{k+1 \leq j \leq n-k : (\hat{y}_{j-k}^{j-1}, \hat{y}_{j+1}^{j+k}) = y_m\}|$.
- For each m, collect all the *unquantized* middle symbols that have quantized contexts y_m to form the subsequence $y^{m,n} = \{y_j : (\hat{y}_{j-k}^{j-1}, \hat{y}_{j+1}^{j+k}) = y_m\}$, where $(\hat{y}_{j-k}^{j-1}, \hat{y}_{j+1}^{j+k}) = Q_M(y_{j-k}^{j-1}, y_{j+1}^{j+k})$. The subsequence $\{y^{n,m}\}$ is further split into $2k+1$ subsequences depending on the distance between their corresponding contexts. Let

$$\mathcal{M}\left(y_j^{m,n}\right) = \{y_i : y_i = y_j^{m,n}\}$$

and

$$d(y_i, y_j) = |i-j|.$$

Further subsequencing followed by denoising is carried out using Algorithm 2 (see Table II). This corresponds to a bank of M symbol-by-symbol denoisers as shown in Fig. 1.

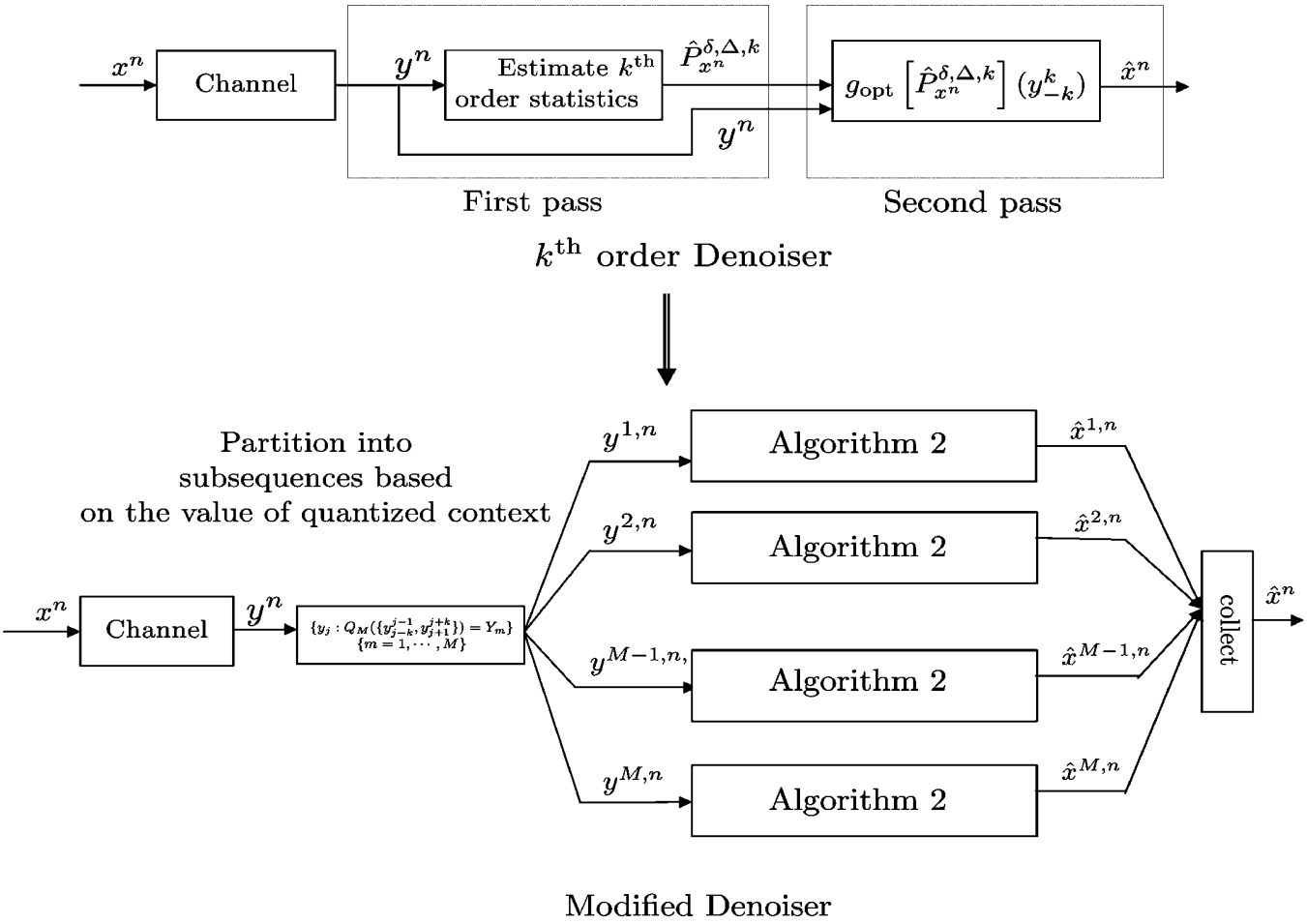

Fig. 1. Proposed modification to the denoiser of [23].

- Collect the denoised estimates $\hat{x}^{m,n}$ and combine them in appropriate order from all the quantization levels, $m \in \{1, \ldots, M\}$ to produce the denoised sequence $\hat{x}^n$.

The details of the construction of this modified Denoiser is discussed in the following subsections.

A. Notations

The application of the density estimator in (12) at each of the M branches of the *modified denoiser* gives an estimate of the quantity $f_{Y_0 \mid \hat{y}_{-k}^{-1}, \hat{y}_1^k}(y)$. Thus, for a quantization level m, (12) becomes

$$f_{Y_0 \mid y_m}^n(y) = \frac{1}{n_m h} \sum_{j:\left(\hat{Y}_{j-k}^{j-1}, \hat{Y}_1^{j+k}\right)=y_m} K\left(\frac{y - Y_j}{h}\right), \quad m \in 1, \ldots, M \tag{24}$$

where Y_j are such that $(\hat{Y}_{j-k}^{j-1}, \hat{Y}_{j+1}^{j+k}) = y_m$, the quantized $2k$-tuple at level m. In addition, application of (13) to $f_{Y_0 \mid y_m}^n(y)$ now gives

$$\hat{F}_{x_0 \mid y_m}[Y^{m,n}] = \arg \min_{F \in \mathcal{F}[a,b]} d\left(f_{Y_0 \mid y_m}^n, \underbrace{\int f_{Y \mid x} dF(x)}_{[F \otimes \mathcal{C}]_Y}\right) \tag{25}$$

which is an estimate of $F_{x_0 \mid y_m}$, the *true* empirical distribution of the underlying clean symbol given the observed noisy context quantized to level m. Note here a key fact about the conditional distribution $F_{x_0 \mid y_m}$: it is now a random object induced by the M-level quantization of the noisy symbols. Finally, the two-stage quantization for step-sizes δ and Δ, corresponding to the support of the underlying clean symbol $[a, b]$ and distribution function levels, respectively, gives $\hat{P}_{x_0 \mid y_m}^{\delta,\Delta}$. For a given M, k, δ, Δ the denoiser is now given by the sequence

$$\tilde{\mathbf{X}}^{n,M,\delta,\Delta,k} = \{\tilde{X}^{n,m,\delta,\Delta}\}_{1 \le m \le M} \tag{26}$$

where

$$\tilde{\mathbf{X}}^{n,m,\delta,\Delta,k} = \{\tilde{X}^{n,m,l,\delta,\Delta}\}_{1 \le l \le 2k+1} \tag{27}$$

and

$$\tilde{X}^{n,m,l,\delta,\Delta}\left[y^{m,n,l}\right](j) = g_{\text{opt}}\left[\hat{P}_{x_0 \mid y_m}^{\delta,\Delta}[y^{m,n,l}]\right]\left(y_j^{m,n,l}\right)$$

$y^{m,n,l} = \{y_j^{m,n,l}\}, j = 1, \ldots, \lfloor (n_m - 2k - l - i/2k + 1) \rfloor$, $l = 1 \ldots, 2k+1$ and g_{opt} is given by (18). The cumulative loss at each of the quantization levels, m, is given by

$$L_{\tilde{\mathbf{X}}^{n,m,\delta,\Delta,k}}(x^n, Y^n) = \frac{1}{n_m} \sum_{i=1}^{n_m} \Lambda\left(x_i^{m,n}, \tilde{X}^{n,m,\delta,\Delta}[y^{m,n}](i)\right) \tag{28}$$

where $x^{m,n} = \{x_j^{m,n}\}, j = 1, \ldots, n_m$ is the underlying clean sequence corresponding to $y^{m,n}$. It is important to note that n_m is itself a random variable. In other words, $L_{\tilde{X}^{n,m,\delta,\Delta}}$ (a random variable) is nothing but the cumulative loss incurred by applying the symbol-by-symbol denoiser in Section III within each class of the subsequence, $\{x^{n,m}\}_{m=1,\ldots,M}$ induced by the quantizer Q.

B. Computational Complexity of the Proposed Denoiser

We now demonstrate the edge of the proposed denoiser, computationally, in comparison with the denoiser of [23] discussed in Section III-D. We briefly list the steps in the implementation of the proposed denoiser and then discuss the computational complexity of the scheme. For a fixed context length, k, and number of quantization levels M.

1) Quantization of the $2k$-Tuple Contexts: The quantization of the $2k$-tuples of the noisy sequence can be viewed as a clustering problem with number of clusters being equal to M. With a plethora of clustering techniques available in the literature, we focus on two types of schemes, viz., k-means and k-cluster partitioning relocation based-clustering. An approximate k-cluster algorithm was proposed in [41] that has a complexity of $O(n \log k), k$ being the number of clusters. In our context, this is equal to $O(n \log M)$.

2) Estimation of $\hat{P}^{\delta,\Delta}_{x_0 \mid y_m}$: As is evident from the discussion in the previous section, we solve a symbol-by-symbol problem at each quantization level which reduces the computational burden dramatically from exponential to polynomial order. This gives an upper bound on the computational complexity of $O(n + (n + M \log n)^3) = O(n^3)$, which is the complexity of the symbol-by-symbol scheme.

Furthermore, we would like to highlight the fact that the context quantization technique proposed here mandates only a symbol-by-symbol (scalar valued) density estimation step in (24). Hence, there is no loss of benefits in the computational complexity of the density estimation step by the (added) complexity of the preprocessing step (discussed in Section III-D, which again would involve only scalar valued clustering algorithms).

In the following section, we proceed to demonstrate the asymptotic universal optimality of the proposed scheme by comparing its performance to a sliding window scheme (whose order increases as a function of the block length) while maintaining the computational complexity of the symbol-by-symbol scheme in Section III.

C. Analysis

We begin by recalling the kth-order sliding window loss as defined in [23], using unquantized contexts, i.e., the $2k+1$ noisy tuples, Y_{-k}^k,

$$D_k(x^n) = \min_{g:\mathbb{R}^{2k+1} \to [a,b]} E\left[\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda\left(x_i, g\left(Y_{i-k}^{i+k}\right)\right)\right]. \tag{29}$$

As has been shown in [23], $D_k(x^n)$ can be expressed as

$$D_k(x^n) = \min_g E_{F_{x^n}^k \otimes \mathcal{C}} \Lambda\left(X, g(Y_{-k}^k)\right) \tag{30}$$

where $F_{x^n}^k$ is the kth-order empirical distribution of the underlying clean sequence and $F_{x^n}^k \otimes \mathcal{C}$ stands for the joint distribution of (X_{-k}^k, Y_{-k}^k) when $X_{-k}^k \sim F_{x^n}^k$ and Y_{-k}^k is the output of the channel $\mathcal{C}$ whose input is X_{-k}^k.

In a similar vein, for a fixed quantizer $Q_M : \mathbb{R}^{2k} \to y^M$, let $D_k^M(x^n)$ be the minimum possible kth-order sliding window loss corresponding to an underlying clean sequence x^n and noisy sequence Y^n with (M-level) quantization of noisy contexts. Specifically,

$$D_k^M(x^n) = \min_g E\left[\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(x_i, g(\hat{Y}_{i-k}^{i-1}, \hat{Y}_{i+1}^{i+k}, Y_i))\right] \tag{31}$$

where $g : y^M \times \mathbb{R} \to \mathbb{R}$,

$$\left(\hat{Y}_{i-k}^{i-1}, \hat{Y}_{i+1}^{i+k}\right) = Q_M\left(\left(Y_{i-k}^{i-1}, Y_{i+1}^{i+k}\right)\right). \tag{32}$$

Note that $D_k^M(x^n)$ depends on the particular M-level quantizer, Q_M used though we suppress this dependence in the notation. Similarly as in (30), the minimum possible kth-order sliding window loss using quantized contexts $D_k^M(x^n)$ can also be expressed as

$$\begin{aligned} D_k^M(x^n) &= \min_g E_{P(\hat{Y}_{-k}^{-1}, \hat{Y}_1^k)} E_{F_{x_0 \mid \hat{Y}_{-k}^{-1}, \hat{Y}_1^k} \otimes \mathcal{C}} \Lambda \\ &\quad \times \left(X_0, g\left(\hat{Y}_{-k}^{-1}, \hat{Y}_1^k, Y_0\right)\right) \qquad (33) \\ &= \min_g E_{P(y_m)} E_{F_{x_0 \mid y_m} \otimes \mathcal{C}} \Lambda(X_0, g(y_m, Y_0)) \qquad (34) \end{aligned}$$

where $P(y_m)$ is the distribution induced by the M-level quantization, Q_M, of the $2k$-tuples (and the underlying x^n). The following theorem, for a given number of quantization levels, M of the noisy contexts bounds the difference between the cumulative loss incurred by the proposed sequence of denoisers and $D_k^M(x^n)$ defined in (31).

Theorem 2: For every sequence, x^n, channel, $\mathcal{C}$, loss function, Λ satisfying conditions C1-7 and L1-2 respectively, $\epsilon > 0, \varepsilon > 0, \delta > 0, \Delta > 0, 1 \le k \le \lfloor (n/2) \rfloor, \exists \zeta = \zeta(\mathcal{C}, \Lambda, \Delta)$ which satisfies

$$\lim_{\Delta \to 0} \zeta(\mathcal{C}, \Lambda, \Delta) = 0$$

s.t.

$$\begin{aligned} &P\left(L_{\tilde{\mathbf{X}}^{n,M,\delta,\Delta,k}}(x^n, Y^n) - D_k^M(x^n)\right. \\ &\quad \left. > 3\epsilon + 5\delta\Lambda_{\max} + 4\zeta + \varepsilon\right) \\ &\quad \le \phi\left(\delta, \Delta, k, \epsilon, \varepsilon, \mathcal{C}, \Lambda, M, n\right) \end{aligned}$$

where

$$\begin{aligned} &\phi(\delta, \Delta, k, \epsilon, \varepsilon, \mathcal{C}, \Lambda, M, n) \\ &\quad = Mn\alpha\left(\delta, \Delta, k, \epsilon, \mathcal{C}, \Lambda, n\right) + Mne^{-\frac{2\varepsilon^2(n-2k)}{\Lambda_{\max}^2(2k+1)}} \end{aligned} \tag{35}$$

and α is given in

$$\begin{aligned}\alpha(\delta,\Delta,k,\epsilon,\mathcal{C},K,\Lambda,n) &= \left[1+\frac{1}{\delta}\right]^{\frac{1}{\Delta}}(2k+1)\left[e^{-\frac{2(\epsilon+\delta\Lambda_{\max})^2(n-2k)}{(2k+1)\Lambda_{\max}^2}} + e^{-\frac{(n-2k)(1-\rho(\epsilon,K))\gamma^2}{2(2k+1)}}\right] \\ &\quad + (2k+1)e^{-\frac{(n-2k)(1-\rho(\epsilon,K))\gamma^2}{2(2k+1)}} \end{aligned} \tag{36}$$

with $\gamma > 0$ and $\gamma = \gamma(\epsilon, \Lambda, \mathcal{C})$.

For a given number of quantization levels, M, growth rates of $k = k_n, \delta = \delta_n, \Delta = \Delta_n$ define

$$\hat{X}^{n,M}_{\text{qc univ}} = \tilde{\mathbf{X}}^{n,M,\delta,\Delta,k} \tag{37}$$

For sufficiently slow growth rates of $M = M_n$, s.t.

$$\sum_{n=1}^{\infty} \phi(\delta_n, \Delta_n, k_n, \epsilon, \mathcal{C}, \Lambda, M_n, n) < \infty \tag{38}$$

application of the Borel–Cantelli Lemma gives the following main result in the semi-stochastic setting.

Theorem 3: For all $\mathbf{x} \in [a,b]^{\infty}$

$$\lim_{n\to\infty}\left[L_{\hat{X}^{n,M_n}_{\text{qc univ}}}(x^n, Y^n) - D^{M_n}_{k_n}(x^n)\right] = 0 \quad \text{a.s.} \tag{39}$$

For example, $M_n = \log n$ along with the growth rates of $k_n = \log n, \delta_n, \Delta_n = (1/\log n)$ satisfy the condition in (35). Also, note that in particular, for a fixed number of quantization levels, $M > 0$ application of Theorem 3 gives

$$\lim_{n\to\infty}\left[L_{\hat{X}^{n,M}_{\text{qc univ}}}(x^n, Y^n) - D^{M}_{k_n}(x^n)\right] = 0 \quad \text{a.s.} \tag{40}$$

V. Stochastic Setting

The discussion so far has focused on the semi-stochastic setting wherein the choice of the quantization scheme was arbitrary. The constraint to guarantee optimality being on the growth rate of quantization resolution with the block length, n, of the noisy sequence. That optimality in the semi-stochastic setting was characterized by comparing with a genie-aided scheme that also uses the same quantization (in both the number of levels and the quantizer itself) as the denoiser. This did not mandate any conditions on the quantization scheme, it could very well be rather a poor choice and yet not affect the performance guarantees at all. Our results also imply optimality for the stochastic setting when the underlying clean signal is now a stationary process, $\mathbf{X}$, with distribution $F_{\mathbf{X}}$. In the stochastic setting, however, the quantizer needs to be able to learn the distribution of the underlying clean sequence and hence arbitrary choice of the quantizer does not suffice any more. This now mandates additional constraints on the quantization scheme itself. It necessitates a quantization scheme that is able to "faithfully" learn the distribution of the noisy sequence and is made precise here in this section. In other words, the quantizer should be such that it generates an "asymptotically fine" sequence of partitions. Asymptotically fine goes to say that the quantizer resolution diminishes with increasing block lengths in a manner such that it generates a sequence of partitions that are also nested. More precisely, define

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

where the expectation is assuming X^n are the first n symbols of the source with distribution $F_{\mathbf{X}}$ and the limit is guaranteed to exist by sub-additivity. Assuming $M_n \to \infty$ at a rate for which Theorem 3 holds and a quantization scheme Q_n that for any n and context length k partitions the space $\mathbb{R}^{2k}$ in a symbol-by-symbol fashion, in such a way that the resulting partition, $\mathcal{P}_n = \mathcal{P}_{n,1} \times \cdots \times \mathcal{P}_{n,2k}, \{\mathcal{P}_{n,i}\}_{i=1}^{2k}$ being the partition in $\mathbb{R}$ corresponding to symbol-by-symbol quantization of each element in the $2k$-tuple, satisfies the following conditions:
Let $\mathcal{P}_{n,i} = \{A_{n,i,j}, j = 1, \ldots, m_{n,i}\}$ be sequences of finite partitions:

- M1) the sequences of partition $\mathcal{P}_{n,i}$ is nested, that is, any cell $\mathcal{P}_{l+1,i}$ is a subset of a cell of $\mathcal{P}_{l,i}, l = 1, 2, \ldots$ for each $i = 1, \ldots, 2k$;
- M2) the sequences of partition $\mathcal{P}_{n,i}$ is asymptotically fine, i.e., if

$$\operatorname{diam}(A) = \sup_{x,y\in A} \|x - y\|$$

denotes the diameter of a set, then for each sphere S centered at the origin

$$\lim_{l\to\infty} \max_{j: A_{l,i,j}\cap S \neq \phi} \operatorname{diam}(A_{l,i,j}) = 0 \quad \text{for each } i. \tag{42}$$

Throughout the remainder of this section, $D^M_k(X^n)$ stands for the denoisability using the symbol-by-symbol denoiser on the quantized kth order contexts and a quantizer that satisfies the above conditions of asymptotical fineness. With the assumptions on the quantization, we have Lemma 1.

Lemma 1: For all stationary $\mathbf{X}$, and any $n, 1 \le k \le \lfloor (n/2) \rfloor$

$$\lim_{M\to\infty} ED^M_k(X^n) = ED_k(X^n) \tag{43}$$

where $D^M_k(X^n)$ is as defined in (57). This lemma characterizes the asymptotic optimality of the finite quantization resolution denoisability, D^M_k. In particularly it states, for any stationary source $\mathbf{X}$, generating the underlying sequence X^n, the expected value of D^M_k achieves the quantity of interest ED_k (denoisability corresponding to infinite resolution) in the limit, with increasing levels of the quantization resolution, M. In addition, this lemma is instrumental in proving the Theorem 4 which cements the optimality of the proposed denoiser in the stochastic setting.

We are now ready for the final result which proves the optimality of the proposed denoiser for the fully stochastic setting. The following theorem establishes the fact that proposed denoiser achieves the optimal distribution dependent performance for any stationary source generating the underlying clean sequence.

Theorem 4: For all stationary $\mathbf{X}$,

$$\lim_{n\to\infty} EL_{\hat{X}^{n,M_n}_{\text{qc univ}}}(X^n, Y^n) = \mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \tag{44}$$

Fig. 2. Top-left: Original image; top-right: Noisy image (corrupted by an AWGN, $\sigma = 15$); bottom-left: Denoised image using the proposed scheme ($2k + 1 = 3$) RMSE $= 6.8$; bottom right: Denoised image using the scheme in [16]RMSE $= 6.36$.

VI. Experimental Results

Results of applying the proposed scheme to a natural test images, are shown in Figs. 2 and 3. The images are corrupted by an AWGN source with $\sigma = 15$. In addition to the significant computational advantages, we are able to achieve better denoising performance of the denoiser by considering higher order contexts, which are computationally far more tractable with this modified denoiser. The results in this section are obtained from further heuristic amends of the fundamental ideas and scheme presented in this paper. The heuristically motivated enhancements are in each of the steps in the denoiser, viz., the quantization scheme, density estimation and the channel inversion steps in the symbol-by-symbol denoiser. The quantization of the kth order contexts is implemented using an off-the-shelf clustering scheme, the k-means or k-medoid clustering algorithm [42], [43]. These quantization approaches are primarily chosen for ease of implementation. There are greater performance benefits to be gained in using a quantization scheme that uses knowledge of the noise variance in determining "similar" context vectors. This type of clustering deserves further investigation and study which we propose as future directions in en-

hancing the performance of the denoiser proposed here. The crucial step in the kernel density estimation is the choice of the smoothing parameter, h which is chosen here as a "rule-of-thumb" [32] because of its computational speed. The choice of h governed by more sophisticated approaches like cross-validation techniques yield better estimates of multimodal distributions. This is an avenue for further exploration in the implementation details of the proposed scheme. The channel inversion is another important step in the process of construction of the denoiser which is solved as a linear program here. For the AWGN case, the matrix that defines the channel is low rank and we are looking for the "best" candidate of the input distribution among all the possibilities that are consistent with the constraints of the inversion problem. The results using this scheme are compared with those from the kth order denoiser proposed in [24]. The results here are presented primarily as a supporting experimental verification of the proposed scheme and proof-of-concept. Hence, we have limited the presentation to two images and the comparison with other state-of-the-art schemes in literature. An experimental exposition of this work with a thorough study of the avenues for further research suggested above is the focus of our current study and is in preparation. In particular, we are studying application of the proposed technique in the reconstruction of undersampled medical resonance images (MRI) [44].

VII. Conclusion

We have presented a scheme for denoising real-valued signals that is asymptotically optimal and universal. A salient feature of the proposed scheme is its computational tractability along with the vital asymptotic universal optimality. This optimality is guaranteed under a large class of well-behaved, user specified, loss functions and channels (noise distributions). The technique presented in this paper draws from the symbol-by-symbol scheme in [23] and the "DUDE framework" in [11]. The sliding window kth-order denoiser in [11] can be viewed as a symbol-by-symbol denoiser that operates on each of the possible context classes ($2k$-tuples of possible contexts induced by the order k). The inapplicability of this scheme to the current setting lies in the fundamental nature of the type of real-valued signals considered here and the resulting irrelevance of a count statistic based scheme. In response to a weighted context aggregation suggested in [11] to enhance the performance of the DUDE and apply it in current settings of real-valued signals, the technique in [23] provides a natural context aggregation mechanism. The exponential dependence of the complexity to the context length limits the practical applications of the denoiser in [23]. In this paper, we have proposed a scheme that is a via-media solution of applying the symbol-by-symbol denoiser (of very low complexity) in [23] to classes of quantized contexts. In addition to the computational benefits, the proposed scheme also addresses the problem of sparse statistics that affects the quality of higher order statistics that are accrued by the sliding window denoiser in [23]. This technique thus aggregates similar contexts to be able to reliably learn the conditional distribution of a noisy symbol given its context in a computationally efficient manner. This then lends itself to learning posterior distribution of the clean symbol given its noisy context which is the quantity the denoiser is after. We also simultaneously prove the optimality of the proposed scheme in the stochastic setting by characterizing the nature of permissible quantizers deployed on the contexts. In the fully stochastic setting the proposed denoiser asymptotically (with increasing block lengths of the noisy sequence) attains the performance of the optimal distribution-dependent denoiser for any stationary ergodic source that generated the clean sequence. Finally, the proposed scheme is an elegant approach to denoising real-valued signals in a tractable manner while maintaining the theoretical guarantees of [23].

The experimental results in this paper compete with many state-of-the-art schemes in the literature [13], [14] that are specifically catered for the particular case of additive Gaussian noise. The scheme presented in this paper however can address more general noise mechanisms and distributions, yet being competitive to schemes that are more specific in their design. Inspired by the performance results in the case of images, we propose to research the application of these ideas to denoising video sequences, audio clips and explore the performance boundaries.

Appendix A
Conditions on the Channel and Loss Function

C1) A memoryless channel, which is to say that the components of y are independent with $Y_i \sim F_{Y\,|\,x_i}$.

C2) The family of measures, $\{\mu_x\}_{x\in[a,b]}$, associated with the channel, $\mathcal{C}$, to be uniformly tight in the sense

$$\sup_{x\in[a,b]} \mu_x([-T,T]^c) \to 0 \quad \text{as} \quad T \to \infty.$$

This condition will be needed to guarantee that one can consistently track the evolution of the marginal density of the noisy symbols at the output of the memoryless channel, regardless of the underlying $\mathbf{x}$, using nonparametric Kernel density estimation techniques.

C3) The distribution functions $F_{Y\,|\,x}$ are assumed to be absolutely continuous for all $x \in [a,b]$ w.r.t. the Lebesgue measure and $\{f_{Y\,|\,x}\}$ denotes the corresponding densities.

C4) The conditional densities of the channel form a set of linearly independent functions. This is equivalent to the "invertibility" condition of [11] which ensures that, to any distribution of the input to the channel there corresponds a unique channel output.

C5) The channel satisfies the uniform Lipschitz continuity condition,

$$\sup_{y\in\mathbb{R}} \|f_{Y\,|\,x}(y)\|_{BL} < \infty \tag{45}$$

where

$$\|f_{Y\,|\,x}(y)\|_{BL} = \|f_{Y\,|\,x}(y)\|_{L} + \|f_{Y\,|\,x}(y)\|_{\infty} \tag{46}$$

$$\|f_{Y\,|\,x}(y)\|_{L} = \sup_{\substack{x\neq z \\ x,z\in[a,b]}} \frac{|f_{Y\,|\,x}(y) - f_{Y\,|\,z}(y)|}{|x-z|} < \infty, \quad \forall y \in \mathbb{R} \tag{47}$$

Fig. 3. Top-left: Original image; top-right: Noisy image (corrupted by an AWGN, $\sigma = 15$); bottom: Denoised image using the proposed scheme, RMSE $= 5.099$.

$$\|f_{Y\,|\,x}(y)\|_\infty = \sup_{x\in[a,b]} f_{Y\,|\,x}(y). \tag{48}$$

This condition guarantees a continuous mapping, w.r.t. a metric that will be detailed in Section III-A, from the space of channel input distributions to the corresponding channel output distributions.

C6) The conditional densities, additionally, satisfy the following Lipschitz continuity condition:

$$\| \delta \|_L = \sup_{0<\Delta<(b-a)} \frac{\delta_\Delta}{\Delta} < \infty \tag{49}$$

where

$$\delta_\Delta = \sup_{x\in[a,b]} \sup_{\substack{\hat{x}\in[a,b] \\ |x-\hat{x}|\le\Delta}} \int |f_{Y\,|\,x}(y) - f_{Y\,|\,\hat{x}}(y)|\, dy. \tag{50}$$

This condition ensures, for reasonably well-behaved loss functions (conditions L1–L2 listed subsequently in this section), continuity in the expected loss induced by two output distributions that are close together (under the metric discussed in Section III-A).

C7) a. The family of conditional densities, $\mathcal{C}$, have uniformly bounded second order universal derivatives, i.e., $\exists$ a $\mathcal{B}_\mathcal{C}$

s.t. $0 < \mathcal{B}_{\mathcal{C}} < \infty$ and $D_2^*(f_{Y\,|\,x}) < \mathcal{B}_{\mathcal{C}}, \forall x \in [a,b]$, where

$$D_2^*(f_{Y\,|\,x}) = \liminf_{h \downarrow 0} \int \left| (f_{Y\,|\,x} * \phi_h)^{(2)} \right| \, dy \tag{51}$$

$\phi_h(x) = (1/h)\phi(x/h), \phi \in C^\infty, C^\infty$ is a set of functions that have infinitely many continuous derivatives with compact support and $f^{(s)}$ denotes the sth derivative of f. This is a mild technical condition that enables the proof of the convergence of marginal density estimates at the output of the memoryless channel to the true marginal density. This condition is trivially satisfied if we have a family of conditional densities that have a uniformly absolutely continuous derivative.

C8) b. An alternative to the previous condition on the family of conditional densities of the channel is, $\lim_{|t| \to 0} \Omega_{\mathcal{C}}(t) = 0$, where

$$\Omega_{\mathcal{C}}(t) = \sup_{x \in [a,b]} \omega_x(t) \tag{52}$$

and

$$\omega_x(t) = \int |f_{Y\,|\,x}(y-t) - f_{Y\,|\,x}(y)| \, dy. \tag{53}$$

From the fact [45] that, for any $f \in L_1(\mathbb{R})$, the corresponding, L_1-modulus of continuity

$$\omega(t) = \int |f(x-t) - f(x)| \, dx \to 0, \quad \text{as } |t| \to 0$$

and

$$\|\omega\|_\infty \leq 2\|f\|_1 < \infty$$

it follows that the global L_1-modulus of continuity $\Omega_{\mathcal{C}}(t)$ is well-defined for all t and families of conditional densities $\mathcal{C}$. In other words, this condition demands uniform convergence of the L_1-moduli of continuity of the individual members comprising the family of conditional densities.

In addition to the constraints on the channel, conditions are imposed on the permissible loss functions, Λ. We assume the loss function Λ:

L1) to be bounded, i.e., $\Lambda_{\max} < \infty$, where $\Lambda_{\max} = \sup_{x,\hat{x} \in [a,b]} \Lambda(x,\hat{x})$;

L2) to be a bounded Lipschitz function. More formally, we require the Lipschitz norm, $\|\Lambda\|_L < \infty$. The Lipschitz norm of the loss function, is defined as

$$\|\Lambda\|_L = \sup_{0 < \Delta < (b-a)} \frac{\lambda(\Delta)}{\Delta} \tag{54}$$

where

$$\lambda(\Delta, x) = \sup_{y \in [a,b]} \sup_{x' : |x - x'| < \Delta} |\Lambda(x,y) - \Lambda(x',y)| \tag{55}$$

and

$$\lambda(\Delta) = \sup_{x \in [a,b]} \lambda(\Delta, x). \tag{56}$$

In words, this condition necessitates continuity of the mapping that takes the estimates of the underlying symbol to the corresponding loss incurred. We require that estimates of the underlying clean symbol that are close together have corresponding loss values that are also close to each other.

Appendix B
Proof of Theorem 2

Before we address the Proof of Theorem 2, we state the analogous theorem, Theorem 5 below, for the symbol-by-symbol denoiser defined in (20) and refer the reader to [23] for details on the proof. This theorem, for a given sequence x^n, gives a bound on the deviation of the cumulative loss incurred by the denoiser in (19) from the minimum symbol-by-symbol loss, $D_0(x^n)$. In addition to providing a performance bound on the deviation, an important consequence of Theorem 5 is Theorem 6 which states an analogous result for the proposed denoiser using quantized contexts.

Theorem 5: For all $\epsilon > 0, \delta > 0, \Delta > 0$ and $x^n \in [a,b]^n$, channel, $\mathcal{C}$, loss Λ satisfying the regularity conditions in Appendix A (C1–C7), Kernel function K (in (12)) there exist positive constants $\psi = \psi(\mathcal{C}, \Lambda, \Delta, \epsilon, \delta), \chi = \chi(\Lambda, \mathcal{C}, \Delta)$ and $\zeta = \zeta(\mathcal{C}, \Lambda, \Delta)$ which satisfies

$$\lim_{\Delta \to 0} \zeta(\mathcal{C}, \Lambda, \Delta) = 0$$

s.t.

$$P\left(|L_{\tilde{X}^n,\delta,\Delta}(x^n, Y^n) - D_0(x^n)| > 3\epsilon + 5\delta\Lambda_{\max} + 4\zeta\right) \leq |\mathcal{G}_{\delta,\Delta}| e^{-n\psi} + e^{-n\chi}, \quad \forall\, n > n_0(\mathcal{C}, K, \{h\}, \epsilon, \Delta, \Lambda)$$

$\mathcal{G}_{\delta,\Delta} = \{g_{\text{opt}}[P]\}_{P \in \mathcal{F}^{\delta,\Delta}}$ denotes the set of all possible denoisers that can be constructed from the members of the set $\mathcal{F}^{\delta,\Delta}$ using (18). It is also true that $|\mathcal{G}_{\delta,\Delta}| \leq [1 + (1/\delta)]^{(1/\Delta)}$. The precise functional forms of ζ, ψ and χ can be found in Theorem 4 in [23]. An analogous result is derived for competing with the kth-order sliding window denoiser in [23].

A consequence of Theorem 5, the Borel–Cantelli Lemma and the definition of the universal symbol-by-symbol denoiser in (20) gives us Theorem 1.

Now, in preparation of Theorem 6 we introduce D_k^m as

$$D_k^m(x^n) = D_0(x^{m,n}). \tag{57}$$

In words, it is the minimum sliding window loss for the mth subsequence $x^{m,n}$. The subsequence $x^{m,n}$ is now random and induced by the (M-level) vector quantization of the kth-order contexts in the noisy sequence.

Theorem 6:

$$P\left(|L_{\tilde{X}^n,m,\delta,\Delta,k}(x^n, Y^n) - D_k^m(x^n)| > 3\epsilon + 5\delta\Lambda_{\max} + 4\zeta | n_m\right) \leq \alpha(\delta, \Delta, k, \epsilon, \mathcal{C}, K, \Lambda, n_m) \quad \forall n_m > n_0^{k,m}(\mathcal{C}, K, \{h\}, \epsilon, \Delta, \Lambda, k) \tag{58}$$

where $\alpha(\delta, \Delta, k, \epsilon, \mathcal{C}, K, \Lambda, n_m)$ is as defined in (36).

Proof of Theorem 6: Using Lemma 8 in [24]

$$L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^n,Y^n)-D_k^m(x^n) \le \frac{1}{2k+1}\sum_{l=1}^{2k+1}\left[L_{\tilde{X}^{n,m,l,\delta,\Delta,k}}(x^{n,m},Y^{n,m})-D_k^{m,l}(x^{n,m})\right] \tag{59}$$

where

$$D_k^{m,l}(x^{n,m}) = D_0(x^{n,m,l})$$

and $L_{\tilde{X}^{n,m,l,\delta,\Delta,k}}(x^{n,m},Y^{n,m})$ are defined for the subsubsequences $x^{n,m,l}$ and are equivalent to (57) and (28), respectively. This then translates to

$$\begin{aligned}&\Pr\left(L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^n,Y^n)-D_k^m(x^n) > 3\epsilon+5\delta\Lambda_{\max}+4\zeta|n_m\right)\\ &\le \Pr\left(\frac{1}{2k+1}\sum_{l=1}^{2k+1}L_{\tilde{X}^{n,m,,l,\delta,\Delta,k}}(x^{m,n},Y^{m,n}) - D_k^{m,l}(x^{m,n}) > 3\epsilon+5\delta\Lambda_{\max}+4\zeta|n_m\right)\\ &\le \sum_{l=1}^{2k+1}\Pr\left(L_{\tilde{X}^{n,m,l,\delta,\Delta,k}}(x^{m,n},Y^{m,n}) - D_k^{m,l}(x^{m,n}) > 3\epsilon+5\delta\Lambda_{\max}+4\zeta|n_m\right).\end{aligned}$$

Theorem 7 below dictates the bound for the deviation in each of the above l terms and is a direct consequence of Theorem 5.

Theorem 7:

$$\begin{aligned}&\Pr\left(\left|L_{\tilde{X}^{n,m,l,\delta,\Delta,k}}(x^{m,n},Y^{m,n}) - D_k^{m,l}(x^{m,n})\right| > 3\epsilon+5\delta\Lambda_{\max}+4\zeta|n_m\right)\\ &\le \left[1+\frac{1}{\delta}\right]^{\frac{1}{\Delta}}\left[e^{-\frac{2(\epsilon+\delta\Lambda_{\max})^2 n_m}{\Lambda_{\max}^2}}+e^{-\frac{n_m(1-\rho(\epsilon,K))\gamma^2}{2}}\right] + e^{-\frac{n_m(1-\rho(\epsilon,K))\gamma^2}{2}}\\ &\qquad \forall\, n_m > n_0^{k,m}(\mathcal{C},K,\{h\},\epsilon,\Delta,\Lambda,k)\end{aligned}$$

where $\rho(\epsilon,K)\in(0,1)$ is a function of the Kernel function K and the deviation bound ϵ given by

$$\rho(\epsilon,K)=\frac{\epsilon}{6\nu}.$$

For any $\epsilon>0, 0<6\nu<\epsilon$ describes the smoothness of the kernel function, K, i.e., for given $\epsilon>0$, find finite constants $M,L,N,a_1,\ldots,a_N$ and disjoint finite rectangles $A_1,\ldots,A_N$ in $\mathbb{R}^d$ such that the function

$$K^*(x)=\sum_{i=1}^{N}a_i I_{A_i}(x). \tag{60}$$

satisfies: $|K^*|\le M, K^*=0$ outside $[-L,L]^d$, and $\int|K(x)-K^*(x)|\,dx<\nu$.

$$\begin{aligned}&\zeta(\mathcal{C},\Lambda,\Delta)\\ &\quad=\delta_\Delta\Lambda_{\max}+4\lambda(\Delta)(1+\delta_\Delta)\end{aligned}$$

and

$$\begin{aligned}&\gamma=\gamma(\epsilon,\Lambda,\mathcal{C})\\ &\quad=\frac{\epsilon}{||\Lambda||_L+\Lambda_{\max}||\delta||_L+(b-a)||\delta||_L||\Lambda||_L+\Lambda_{\max}}.\end{aligned}$$

$||\delta||_L, ||\Lambda||_L$ are defined in (49) and (54).

This then leads to

$$\begin{aligned}&\Pr\left(L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^n,Y^n)-D_k^m(x^n) > 3\epsilon+5\delta\Lambda_{\max}+4\zeta|n_m\right)\\ &\le \left[1+\frac{1}{\delta}\right]^{\frac{1}{\Delta}}(2k+1)\left[e^{-\frac{2(\epsilon+\delta\Lambda_{\max})^2 n_m}{\Lambda_{\max}^2}} + e^{-\frac{n_m(1-\rho(\epsilon,K))\gamma^2}{2}}\right]+(2k+1)e^{-\frac{n_m(1-\rho(\epsilon,K))\gamma^2}{2}}\\ &\qquad \forall\, n_m > n_0^{k,m}(\mathcal{C},K,\{h\},\epsilon,\Delta,\Lambda,k).\end{aligned}$$

$n_0^{k,m}$ has the same functional form as the n_0 in Theorem 5 but, also depends on the quantization level m through the induced distribution of the quantized noisy symbols and context length, k through the conditional distribution, $F_{x_0\,|\,y_m}$, induced by the kth-order contexts. This is similar to Theorem 4 in [23] in that it bounds the deviation of the cumulative loss from the minimum loss, D_k^m. A key point to note here is that the above theorem is now true at each quantization level, m and $n_0^{k,m}$, the sample size at which the bound is valid varies with each quantization level, m. $n_0^{k,m}$ is also a function of the length k of the contexts, the quantization of which induces the context classes.

Hence

$$\begin{aligned}&P\left(|L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^n,Y^n)-D_k^m(x^n)| > 3\epsilon+5\delta\Lambda_{\max}+4\zeta|n_m\right)\\ &\le \alpha(\delta,\Delta,k,\epsilon,\mathcal{C},K,\Lambda,n_m)\\ &\qquad \forall\, n_m > n_0^{k,m}(\mathcal{C},K,\{h\},\epsilon,\Delta,\Lambda,k)\end{aligned} \tag{61}$$

where $\alpha(\delta,\Delta,k,\epsilon,\mathcal{C},K,\Lambda,n_m)$ is as defined in (36). ■

The following lemma formalizes the fact that, for any sequence, by performing *optimally* within every quantized context, wherein we allow the denoiser chosen to be different for every quantization level, we will be doing at least as well as the scheme which fixes one denoiser for all the quantization levels. In relation to this, we define

$$\begin{aligned}P(y_m)&=P\left(\hat{Y}_{-k}^{-1},\hat{Y}_1^k\right)\\ &=\frac{1}{n-2k}\sum_{i=k+1}^{n-k}P\left(\hat{Y}_{i-k}^{i-1},\hat{Y}_{i+1}^{i+k}=y_m\right).\end{aligned} \tag{62}$$

Lemma 2: For any sequence, x^n and $M>0$

$$D_k^M(x^n)\ge E_{P(y_m)}D_k^m(x^n). \tag{63}$$

Proof: By definition of D_k^M in (34)

$$
\begin{aligned}
&D_k^M(x^n)\\
&\quad= \min_{g:y^M\times\mathbb{R}\to[a,b]} E_{P(\hat{Y}_{-k}^{-1},\hat{Y}_1^k)} E_{F_{x_0|(\hat{Y}_{-k}^{-1},\hat{Y}_1^k)}\otimes\mathcal{C}\Lambda}\\
&\qquad\times\left(X_0, g\left(\hat{Y}_{-k}^{-1},\hat{Y}_1^k, Y_0\right)\right) && (64)\\
&\quad= \min_{g:y^M\times\mathbb{R}\to[a,b]} \sum_{m=1}^{M} E_{F_{x_0|(\hat{Y}_{-k}^{-1},\hat{Y}_1^k)=y_m}\otimes\mathcal{C}\Lambda}\\
&\qquad\times\left(X_0, g\left(y_m, Y_0\right)\right) P(y_m) && (65)\\
&\quad\geq \sum_{m=1}^{M} \min_{g_m:\mathbb{R}\to[a,b]} E_{F_{x_0|y_m}\otimes\mathcal{C}\Lambda}\left(X_0, g_m\left(Y_0\right)\right) P(y_m) && (66)\\
&\quad= \sum_{m=1}^{M} P(y_m) \min_{g_m:\mathbb{R}\to[a,b]} E_{F_{x_0|y_m}\otimes\mathcal{C}\Lambda}\left(X_0, g_m\left(Y_0\right)\right) && (67)\\
&\quad= E_{P(y_m)} \min_{g_m:\mathbb{R}\to[a,b]} E_{F_{x_0|y_m}\otimes\mathcal{C}\Lambda}(X_0, g_m(Y_0)) && (68)\\
&\quad= E D_k^m\left(x_0^{n_m}; y_m\right). && (69)
\end{aligned}
$$

The inequality in (66) follows from the fact that $\{g_1(\cdot) = g(y_1,\cdot) : \mathbb{R}\to[a,b]\}\times\{g_2(\cdot)=g(y_2,\cdot):\mathbb{R}\to[a,b]\}\cdots\times\{g_M(\cdot)=g(y_M,\cdot):\mathbb{R}\to[a,b]\}\supset\{g : y^M\times\mathbb{R}\to[a,b]\}$ and hence minimizing the mapping g for each y_m lower bounds the one where g remains the same at each quantization level m.

■

Proof of Theorem 2

The cumulative loss incurred by the sequence of denoisers, $L_{\tilde{\mathbf{X}}^{n,M,\delta,\Delta,k}}$ is given by

$$
\begin{aligned}
&L_{\tilde{\mathbf{X}}^{n,M,\delta,\Delta,k}}(x^n,Y^n)\\
&\quad= \sum_{m=1}^{M}\frac{n_m}{n-2k} L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^n,Y^n) && (70)\\
&\quad= \sum_{m=1}^{M}\frac{n_m}{n-2k}\frac{1}{n_m}\sum_{i=1}^{n_m}\Lambda\left(x_i^{m,n}, \tilde{X}^{n,m,\delta,\Delta,k}[y^{m,n}](i)\right) && (71)\\
&\quad= \frac{1}{n-2k}\sum_{m=1}^{M}\sum_{i=1}^{n_m}\Lambda\left(x_i^{m,n}, \tilde{X}^{n,m,\delta,\Delta,k}[y^{m,n}](i)\right) && (72)\\
&\quad= \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda\left(x_i, \hat{\mathbf{X}}^{n,M,\delta,\Delta,k}[y^n](i)\right) && (73)
\end{aligned}
$$

where the last equality follows from the definition of the sequence of denoisers in (26). In words, the cumulative loss incurred by the sequence of denoisers $\hat{\mathbf{X}}^{n,M,\delta,\Delta,k}$ is a weighted average of the cumulative losses incurred by the individual symbol-by-symbol denoisers on each of the M subsequences. Using this fact, we can proceed with the proof.

$$
\begin{aligned}
&(n-2k)L_{\tilde{\mathbf{X}}^{n,M,\delta,\Delta,k}}(x^n,Y^n) - (n-2k)D_k^M(x^n)\\
&\quad= \sum_{m=1}^{M} n_m L_{\tilde{X}^{n,m,\delta,\Delta,k}}\left(x^{m,n},Y^{m,n}\right) - (n-2k)D_k^M(x^n)\\
&\quad\leq \sum_{m=1}^{M} n_m L_{\tilde{X}^{n,m,\delta,\Delta,k}}\left(x^{m,n},Y^{m,n}\right) - \sum_{m=1}^{M} n_m D_k^m(x^n)\\
&\qquad+\sum_{m=1}^{M} n_m D_k^m(x^n) - (n-2k)E_{P(y_m)}D_k^m(x^n).
\end{aligned}
$$

Now

$$
\begin{aligned}
&\left|(n-2k)L_{\tilde{\mathbf{X}}^{n,M,\delta,\Delta,k}}(x^n,Y^n) - (n-2k)D_k^M(x^n)\right|\\
&\quad\leq \sum_{m=1}^{M} n_m\left|L_{\tilde{X}^{n,m,\delta,\Delta,k}}\left(x^{m,n},Y^{m,n}\right) - D_k^m(x^n)\right|\\
&\qquad+\left|\sum_{m=1}^{M} n_m D_k^m(x^n) - (n-2k)E_{P(y_m)}D_k^m(x^n)\right|
\end{aligned}
$$

and

$$
\begin{aligned}
&\Pr\left(\left|(n-2k)L_{\tilde{\mathbf{X}}^{n,M,\delta,\Delta,k}}(x^n,Y^n) - (n-2k)D_k^M(x^n)\right|\right.\\
&\quad\left.> (n-2k)(3\epsilon+5\delta\Lambda_{\max}+4\zeta+\varepsilon)\right)\\
&\quad\leq \Pr\left(\sum_{m=1}^{M} n_m\left|L_{\tilde{X}^{n,m,\delta,\Delta,k}}\left(x^{m,n},Y^{m,n}\right) - D_k^m(x^n)\right|\right.\\
&\qquad\left.> (n-2k)(3\epsilon+5\delta\Lambda_{\max}+4\zeta)\right)\\
&\qquad+\Pr\left(\left|\sum_{m=1}^{M} n_m D_k^m(x^n) - (n-2k)E_{P(y_m)}D_k^m(x^n)\right|\right.\\
&\qquad\left.> (n-2k)\varepsilon\right).
\end{aligned}
$$

Using a simple union bound

$$
\begin{aligned}
&\Pr\left(\left|(n-2k)L_{\tilde{\mathbf{X}}^{n,M,\delta,\Delta,k}}(x^n,Y^n) - (n-2k)D_k^M(x^n)\right|\right.\\
&\quad\left.> (n-2k)(3\epsilon+5\delta\Lambda_{\max}+4\zeta+\varepsilon)\right)\\
&\quad\leq \sum_{m=1}^{M}\Pr\left(\left|L_{\tilde{X}^{n,m,\delta,\Delta,k}}\left(x^{m,n},Y^{m,n}\right) - D_k^m(x^n)\right|\right.\\
&\qquad\left.> \frac{n-2k}{n_m}(3\epsilon+5\delta\Lambda_{\max}+4\zeta)\right)\\
&\qquad+\sum_{m=1}^{M}\Pr\left(\left|n_m D_k^m(x^n) - (n-2k)E_{P(y_m)}D_k^m(x^n)\right|\right.\\
&\quad\left.> (n-2k)\varepsilon\right). && (74)
\end{aligned}
$$

Now, for each m, it is true that

$$
\begin{aligned}
&\Pr\left(\left|L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^{m,n},Y^{m,n}) - D_k^m(x^n)\right|\right.\\
&\quad\left.> \frac{n-2k}{n_m}(3\epsilon+5\delta\Lambda_{\max}+4\zeta)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{\nu=1}^{n} \Pr\left(|L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^{m,n}, Y^{m,n}) - D_k^m(x^n)|\right. \\
&> (3\epsilon + 5\delta\Lambda_{\max} + 4\zeta)|n_m = \nu)\, P(\nu) \\
&\le n \max_{1\le n_m\le n} \Pr\left(|L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^{m,n}, Y^{m,n}) - D_k^m(x^n)|\right. \\
&> \left.\left.\frac{n-2k}{n_m}(3\epsilon + 5\delta\Lambda_{\max} + 4\zeta)\,|n_m\right)
\end{aligned}
$$

where n_m, the number of points at quantization level m is a random variable and can be anywhere between 1 and n.

Further

$$
\begin{aligned}
&P\left(|L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^{m,n}, Y^{m,n}) - D_k^m(x^n)|\right. \\
&> \left.\frac{n-2k}{n_m}(3\epsilon + 5\delta\Lambda_{\max} + 4\zeta)\right| n_m\Big) \\
&\le \left[1+\frac{1}{\delta}\right]^{\frac{1}{\Delta}} (2k+1)\left[e^{-\frac{2(\epsilon+\delta\Lambda_{\max})^2(n-2k)^2}{(2k+1)\Lambda_{\max}^2 n_m}}\right. \\
&\quad \left. + e^{-\frac{(n-2k)^2(1-\rho(\epsilon,K))\gamma^2}{2n_m}}\right] + (2k+1)e^{-\frac{(n-2k)^2(1-\rho(\epsilon,K))\gamma^2}{2n_m}} \\
&\qquad\qquad \forall n_m > n_0^{k,m}(\mathcal{C}, K, \{h\}, \epsilon, \Delta, \Lambda) \quad (75)
\end{aligned}
$$

where the last inequality follows from Theorem 6. Now, the RHS of (75) is maximized by $n_m = n - 2k$

$$
\begin{aligned}
&\max_{1\le n_m\le n} \Pr\left(|L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^{m,n}, Y^{m,n}) - D_k^m(x^n)|\right. \\
&> \left.\frac{(n-2k)}{n_m}(3\epsilon + 5\delta\Lambda_{\max} + 4\zeta)\right| n_m\Big) \\
&\le \left[1+\frac{1}{\delta}\right]^{\frac{1}{\Delta}} (2k+1)\left[e^{-\frac{2(\epsilon+\delta\Lambda_{\max})^2(n-2k)}{(2k+1)\Lambda_{\max}^2}}\right. \\
&\quad \left. + e^{-\frac{(n-2k)(1-\rho(\epsilon,K))\gamma^2}{2}}\right] + (2k+1)e^{-\frac{(n-2k)(1-\rho(\epsilon,K))\gamma^2}{2}}.
\end{aligned}
$$

Now, for each m, it is true that

$$
\begin{aligned}
&\Pr\left(|L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^{m,n}, Y^{m,n}) - D_k^m(x^n)|\right. \\
&> \left.\frac{n-2k}{n_m}(3\epsilon + 5\delta\Lambda_{\max} + 4\zeta)\right) \\
&= \sum_{\nu=1}^{n} \Pr\left(|L_{\tilde{X}^{n,m,\delta,\Delta,k}}(x^{m,n}, Y^{m,n}) - D_k^m(x^n)|\right. \\
&> (3\epsilon + 5\delta\Lambda_{\max} + 4\zeta)|n_m = \nu)\, P(\nu) \\
&\le n\left[\left(1+\frac{1}{\delta}\right)^{\frac{1}{\Delta}} (2k+1)\left(e^{-\frac{2(\epsilon+\delta\Lambda_{\max})^2(n-2k)}{(2k+1)\Lambda_{\max}^2}}\right.\right. \\
&\quad \left.\left. + e^{-\frac{(n-2k)(1-\rho(\epsilon,K))\gamma^2}{2}}\right) + (2k+1)e^{-\frac{(n-2k)(1-\rho(\epsilon,K))\gamma^2}{2}}\right] \\
&= n\alpha(\delta, \Delta, k, \epsilon, \mathcal{C}, K, \Lambda, n). \quad (76)
\end{aligned}
$$

Also,

$$
\begin{aligned}
&\left|\sum_{m=1}^{M} n_m D_k^m(x^n) - (n-2k)E_{P(y_m)} D_k^m(x^n)\right| \\
&\qquad = \sum_{m=1}^{M} D_k^m(x^n)|n_m - (n-2k)P(y_m)|. \quad (77)
\end{aligned}
$$

We can bound the term on the RHS in (77) using the following arguments, for every m,

$$
\begin{aligned}
n_m - (n-2k)P(y_m) = &\sum_{i=k+1}^{n-k} \mathbf{1}_{\{\hat{Y}_{i-k}^{i-1}, \hat{Y}_{i+1}^{i+k}\}=y_m} \\
&- (n-2k)P(y_m) \quad (78)
\end{aligned}
$$

where, $\mathbf{1}_{\{.\}}$ is the indicator random variable. It is clear that

$$
E\left[\mathbf{1}_{\{\hat{Y}_{i-k}^{i-1}, \hat{Y}_{i+1}^{i+k}\}s=y_m}\right] = P(y_m). \quad (79)
$$

Hence

$$
E\left[\sum_{i=k+1}^{n-k} \mathbf{1}_{\{\hat{Y}_{i-k}^{i-1}, \hat{Y}_{i+1}^{i+k}\}=y_m}\right] = (n-2k)P(y_m). \quad (80)
$$

Thus, the term within the bracket in (77) is a sum of $\lfloor (n_m - 2k/2k+1) \rfloor$ random variables that are bounded in [0, 1].

$$
\begin{aligned}
&\Pr\left(\sum_{m=1}^{M} D_k^m(x^n)|n_m - (n-2k)P(y_m)| > (n-2k)\varepsilon\right) \\
&\le \Pr\left(\sum_{m=1}^{M} |n_m - (n-2k)P(y_m)| > (n-2k)\frac{\varepsilon}{\Lambda_{\max}}\right) \quad (81) \\
&\le \sum_{m=1}^{M} \Pr\left(|n_m - (n-2k)P(y_m)| > (n-2k)\frac{\varepsilon}{\Lambda_{\max}}\right). \quad (82)
\end{aligned}
$$

The inequality in (81) follows from the fact that $D_k^m(x^n) < \Lambda_{\max}, \forall\, m$ coupled with the union bound. Again, using similar arguments as before

$$
\begin{aligned}
&\Pr\left(|n_m - (n-2k)P(y_m)| > (n-2k)\frac{\varepsilon}{\Lambda_{\max}}\right) \\
&= \sum_{\nu=1}^{n} \Pr\left(|n_m - (n-2k)P(y_m)|\right. \\
&\quad \left.> \frac{(n-2k)\varepsilon}{\Lambda_{\max}}\right| n_m = \nu\Big) P(\nu) \quad (83) \\
&\le n \max_{1\le n_m\le n} \Pr\left(|n_m - (n-2k)P(y_m)|\right. \\
&\quad \left.> \frac{(n-2k)\varepsilon}{\Lambda_{\max}}\right| n_m\Big). \quad (84)
\end{aligned}
$$

Using Hoeffding's inequality [28]

$$\begin{aligned}&\Pr\left(|n_m-(n-2k)P(y_m)|>\frac{(n-2k)\varepsilon}{\Lambda_{\max}}\middle| n_m\right)\\&\quad=\Pr\left(|n_m-(n-2k)P(y_m)|\right.\\&\qquad\left.>(n_m-2k)\frac{(n-2k)\varepsilon}{\Lambda_{\max}(n_m-2k)}\middle| n_m\right)\\&\quad\le e^{-\frac{2\varepsilon^2(n-2k)^2(n_m-2k)}{\Lambda_{\max}^2(2k+1)(n_m-2k)^2}}\\&\quad=e^{-\frac{2\varepsilon^2(n-2k)^2}{\Lambda_{\max}^2(2k+1)(n_m-2k)}}.\end{aligned}\tag{85}$$

Similarly, the RHS is maximized when $n_m=n$

$$\begin{aligned}&\Pr\left(|n_m-(n-2k)P(y_m)|>\frac{(n-2k)\varepsilon}{\Lambda_{\max}}\middle| n_m\right)\\&\qquad\le e^{-\frac{2\varepsilon^2(n-2k)}{\Lambda_{\max}^2(2k+1)}}.\end{aligned}\tag{86}$$

Substituting in (84)

$$\begin{aligned}&\Pr\left(|n_m-(n-2k)P(y_m)|>(n-2k)\frac{\varepsilon}{\Lambda_{\max}}\right)\\&\qquad\le ne^{-\frac{2\varepsilon^2(n-2k)}{\Lambda_{\max}^2(2k+1)}}.\end{aligned}\tag{87}$$

Combining (74), (76), (82) and (87) we finally get

$$\begin{aligned}&\Pr\left(\left|L_{\tilde{\mathbf{X}}^{n,M,\delta,\Delta,k}}(x^n,Y^n)-D_k^M(x^n)\right|\right.\\&\quad\left.>3\epsilon+5\delta\Lambda_{\max}+4\zeta+\varepsilon\right)\\&\quad\le Mn\alpha\left(\delta,\Delta,k,\epsilon,\mathcal{C},\Lambda,n\right)+Mne^{-\frac{2\varepsilon^2(n-2k)}{\Lambda_{\max}^2(2k+1)}}.\end{aligned}$$

Appendix C
Proof of Lemma 1

Lemma 3: For all stationary $\mathbf{X}$, and any $n, 1\le k\le\lfloor(n/2)\rfloor$

$$\lim_{M\to\infty}ED_k^M(X^n)=ED_k(X^n).$$

Proof: For a fixed context length, k, let Σ_M denote the σ-field generated by $\{Q_M(Y_{-k}^{-1},Y_1^k),Y_0\}$, i.e.,

$$\Sigma_M=\sigma\left(Q_M\left(Y_{-k}^{-1},Y_1^k\right),Y_0\right)$$

and Σ_∞, the σ-field generated by Y_{-k}^k, $\Sigma_\infty=\sigma(Y_{-k}^k)$. From the constraints M1, M2 on the quantizer, Q_M, it is clear that $\mathbf{\Sigma}=(\Sigma_M)$ forms a discrete filtration. Also, note that

$$\Sigma_\infty=\bigvee_{M\in\mathbb{Z}^+}\Sigma_M$$

and $Y_{-k}^k\in\Sigma_\infty$. From our assumption on the loss function, it is also clear that, $\Lambda(X_0,\cdot)\in L^1$. Now, using the uniform integrability and convergence result of Martingales in L^1 [46], we get

$$\begin{aligned}&E\left(\Lambda\left(X_0,g\left(Q_M\left(Y_{-k}^{-1},Y_1^k\right),Y_0\right)\right)|\Sigma_M\right)\\&\qquad\xrightarrow[\text{a.s.}]{M\to\infty}E\left(\Lambda\left(X_0,g\left(Y_{-k}^k\right)\right)|\Sigma_\infty\right).\end{aligned}$$

Further application of the bounded convergence theorem gives

$$\begin{aligned}&E\left(\Lambda\left(X_0,g\left(Q_M\left(Y_{-k}^{-1},Y_1^k\right),Y_0\right)\right)\right)\\&\qquad\xrightarrow{M\to\infty}E\left(\Lambda\left(X_0,g\left(Y_{-k}^k\right)\right)\right).\end{aligned}$$

Further, the continuity conditions on Λ [23] guarantee the following:

$$\begin{aligned}&ED_k^M(X^n)\\&\quad=\min_g E\left(\Lambda\left(X_0,g\left(Q_M\left(Y_{-k}^{-1},Y_1^k\right),Y_0\right)\right)\right)\\&\quad\xrightarrow{M\to\infty}E\left(\Lambda\left(X_0,g\left(Y_{-k}^k\right)\right)\right)=ED_k(X^n).\end{aligned}$$

■

Appendix D
Proof of Theorem 4

Proof: By the definition of $\mathbb{D}(F_{\mathbf{X}},\mathcal{C})$ and the sub-additivity lemma [47]

$$\begin{aligned}&\liminf_{n\to\infty}EL_{\hat{X}_{\text{qc univ}}^{n,M_n}}(X^n,Y^n)\\&\quad\ge\inf_{n\ge1}EL_{\hat{X}_{\text{qc univ}}^{n,M_n}}(X^n,Y^n)\\&\quad\ge\inf_{n\ge1}\min_{\hat{X}^n}EL_{\hat{X}^n}(X^n,Y^n)=\mathbb{D}(F_{\mathbf{X}},\mathcal{C}).\end{aligned}\tag{88}$$

On the other hand, from Theorem 3 and dominated convergence theorem

$$\begin{aligned}\limsup_{n\to\infty}EL_{\hat{X}_{\text{qc univ}}^{n,M_n}}(X^n,Y^n)&=\limsup_{n\to\infty}ED_{k_n}^{M_n}(X^n)\\&\le\lim_{M\to\infty}\limsup_{n\to\infty}ED_{k_n}^M(X^n).\end{aligned}\tag{89}$$

Using Lemma 1, it is also true that, for any fixed k

$$\begin{aligned}&\lim_{M\to\infty}\limsup_{n\to\infty}ED_{k_n}^M(X^n)\\&\quad\le\lim_{M\to\infty}\limsup_{n\to\infty}ED_k^M(X^n)\le\limsup_{n\to\infty}\lim_{M\to\infty}ED_k^M(X^n)\\&\quad=\limsup_{n\to\infty}ED_k(X^n)\end{aligned}\tag{90}$$

which, by arbitrariness of k, implies

$$\lim_{M\to\infty}\limsup_{n\to\infty}ED_{k_n}^M(X^n)\le\lim_{k\to\infty}\limsup_{n\to\infty}ED_k(X^n).\tag{91}$$

Further, as shown in Claim 1 in [23],

$$\lim_{k\to\infty}\min_g E\Lambda\left(X_0,g\left(Y_{-k}^k\right)\right)=\mathbb{D}(F_{\mathbf{X}},\mathcal{C}).\tag{92}$$

For any k

$$\begin{aligned} ED_k(X^n) &= E \min_g E_{F_{x^n}^k \otimes \mathcal{C}} \Lambda\left(X, g\left(Y_{-k}^k\right)\right) \\ &\le \min_g E\left[E_{F_{x^n}^k \otimes \mathcal{C}} \Lambda\left(X, g(Y_{-k}^k)\right)\right] \\ &= \min_g E\Lambda\left(X_0, g\left(Y_{-k}^k\right)\right) \end{aligned} \tag{93}$$

where the right side X_{-k}^k is emitted from the (unique) double-sided extension of the source $F_{\mathbf{X}}$. Using the result from (92) and (93), we get

$$\lim_{k\to\infty} \limsup_{n\to\infty} ED_k(X^n) \le \mathbb{D}(F_{\mathbf{X}}, \mathcal{C}). \tag{94}$$

Combining (89), (91), and (94), we get

$$\limsup_{n\to\infty} EL_{\hat{X}_{\text{qc univ}}^{n, M_n}}(X^n, Y^n) \le \mathbb{D}(F_{\mathbf{X}}, \mathcal{C}). \tag{95}$$

Finally, (88) and (95) together give us the desired result. ■

REFERENCES

[1] N. Wiener, *Density Estimation for Statistics and Data Analysis*. Cambridge, MA: Technology Press of MIT, 1949.

[2] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, pp. 613–627, May 1995.

[3] R. R. Coifman and D. L. Donoho, *Translation-Invariant De-Noising, Lecture Notes in Statistics: Wavelets and Statistics*. New York: Springer-Verlag, 1995.

[4] J.-L. Starck, E. J. Candés, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, pp. 670–684, Jun. 2002.

[5] A. Beck and Y. C. Eldar, "Regularization in regression with bounded noise: A Chebyshev center approach," *SIAM J. Matrix Anal. Appl.*, vol. 29, no. 2, pp. 606–625, Nov. 2007.

[6] A. Beck, Y. C. Eldar, and A. Ben-Tal, "Minimax mean-squared error estimation of multichannel signals," *SIAM J. Matrix Anal. Appl.*, vol. 29, no. 3, pp. 712–730, Nov. 2007.

[7] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 349–366, Feb. 2007.

[8] Y. Hel-Or and D. Shaked, "Slicing the transform—A discriminative approach for wavelet denoising," Hewlett-Packard Labs, Tech. Rep. HPL-2006-103R1, 2006.

[9] B. Natarajan, "Filtering random Boise from deterministic signals via data compression," *IEEE Trans. Signal Process.*, vol. 43, no. 11, pp. 2595–2605, Nov. 1995.

[10] D. L. Donoho, "Kolmogorov sampler," Stanford Univ., Stanford, CA, Tech. Rep., 2002.

[11] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 1229–1246, Jan. 2005.

[12] A. Dembo and T. Weissman, "The minimax distortion redundancy in noisy source coding," *IEEE Trans. Inf. Theory*, vol. 49, pp. 3020–3030, Nov. 2003.

[13] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, pp. 1338–1351, Nov. 2003.

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

[15] M. Mahmoudi and G. Sapiro, "Fast image and video denoising via nonlocal means of similar neighborhoods," *IEEE Signal Process. Lett.*, vol. 12, no. 12, pp. 839–842, Dec. 2005.

[16] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, pp. 2080–2095, Aug. 2007.

[17] S. P. Awate and R. T. Whitaker, "Unsupervised, information-theoretic, adaptive image filtering for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 364–376, 2006.

[18] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images," *IEEE Trans. Image Process.*, vol. 16, pp. 1395–1411, May 2007.

[19] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[20] C. Kervrann and J. Boulanger, "Local adaptivity to variable smoothness for exemplar-based image regularization and representation," *Int. J. Comput. Vis.*, vol. 79, no. 1, pp. 45–69, Aug. 2008.

[21] H. Robbins, "The empirical Bayes approach to statistical decision problems," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 1–20, Mar. 1964.

[22] H. Robbins, "Asymptotically subminimax solutions of compound decision problems," in *Proc. 2nd Berkeley Symp. Mathematical Statistical Probability*, Berkeley, CA, 1951, pp. 131–148.

[23] K. Sivaramakrishnan and T. Weissman, "Universal denoising of discrete-time continuous-amplitude signals," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5632–5660, Dec. 2008.

[24] K. Sivaramakrishnan and T. Weissman, "Universal denoising of discrete-time continuous-amplitude signals," presented at the 2006 IEEE Int. Symp. Information Theory, Seattle, WA, Jul. 2006.

[25] K. Sivaramakrishnan and T. Weissman, "Universal denoising of continuous valued signals with applications to images," presented at the 2006 IEEE Int. Conf. Image Processing, Atlanta, GA, Sep. 2006.

[26] K. Sivaramakrishnan and T. Weissman, "A context quantization approach to universal denoising," presented at the 2007 IEEE Int. Symp. Information Theory, Nice, France, Jul. 2007.

[27] G. Motta, E. Ordentlich, I. Ramirez, G. Seroussi, and M. J. Weinberger, "The DUDE framework for continuous tone image denoising," presented at the 2005 IEEE Int. Conf. Image Processing, Genoa, Italy, Sep. 2005.

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

[29] L. Devroye and L. Gyorfi, *Nonparametric Denisty Estimation, the L_1 View*, ser. Probability and StatisticsMathematical. New York: Wiley, 1985.

[30] S. P. Awate and R. T. Whitaker, "Feature-preserving MRI denoising: A nonparametric empirical Bayes approach," *IEEE Trans. Med. Imag.*, vol. 26, no. 9, pp. 1242–1255, Sep. 2007.

[31] L. Devroye, *A Course in Density Estimation*. Boston, MA: Birkhäuser, 1987.

[32] B. W. Silverman, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. London, U.K.: Chapman & Hall, 1986.

[33] L. Greengard and J. Strain, "The fast gauss transform," *SIAM J. Sci. Statist. Comput.*, vol. 12, no. 1, pp. 79–94, 1991.

[34] C. Yang, R. Duraiswami, N. A. Gumerov, and L. Davis, "Improved fast gauss transform and efficient kernel density estimation," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 464–471.

[35] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theor. Comput. Sci.*, vol. 38, pp. 293–306, 1985.

[36] A. G. Gray and A. W. Moore, "Very fast multivariate kernel density estimation via computational geometry," in *Proc. Joint Statistical Meet.*, San Francisco, CA, Aug. 2003.

[37] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. New York: Athena Scientific, 1997.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge Univ. Press, 2004.

[39] V. C. Raykar, C. Yang, R. Duraiswami, and N. Gumerov, "Fast computation of sums of Gaussians in high dimensions," Univ. of Maryland, College Park, MD, Tech. Rep., 2006.

[40] A. G. Gray and A. W. Moore, "Rapid evaluation of multiple density models," in *Proc. 9th Int. Workshop Artificial Intelligence Statistics*, Key West, FL, Jan. 2003.

[41] T. Feder and D. H. Greene, "Optimal algorithms for approximate clustering," in *Proc. Symp. Theory Computing*, 1988, pp. 434–444.

[42] R. M. Gray, "Vector quantization," *IEEE ASSP Mag.*, vol. 1, no. 2, pp. 4–29, Apr. 1984.

[43] J. Abonyi, B. Balasko, and B. Feil, Fuzzy Clustering Toolbox [Online]. Available: http://www.fmt.vein.hu/softcomp/fclusttoolbox

[44] K. Sivaramakrishnan, M. Lustig, and T. Weissman, "Application of the context quantized universal denoising techniques to medical resonance images (MRI)," in preparation.

[45] R. Wheeden and A. Zymund, *Measure and Integral*. New York: Marcel Dekker, 1977.

[46] R. Durrett, *Probability: Theory and Examples*, ser. Duxbury Advanced Series, 3rd ed. Pacific Grove, CA: Brooks/Cole, 2005.

[47] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1998.

Kamakshi Sivaramakrishnan (S'00–M'08) received the M.S. degree from Boston University, Boston, MA, in 2000 and the Ph.D. degree from Stanford University, Stanford, CA, in 2008, both in electrical engineering.

From 2003 to 2005, she was a principal contributor to New Horizons, a space mission to Pluto, launched by NASA in January 2006. Between June and November 2007, she was as a Postdoctoral Research Fellow at Hewlett-Packard Laboratories. Since November 2007, she has been an Applied Research Scientist with Admob, Inc., San Mateo, CA. Her research interests range from information theory and its applications to, more recently, machine learning techniques in computational advertisement. She has published in the fields of signal processing and information theory; her current work has expanded to complexity of algorithms. She has several patents in the area of machine learning and data mining, primarily, for their applications in mobile advertising.

Tsachy Weissman (S'99–M'02–SM'07) received the B.Sc. and Ph.D. degrees in electrical engineering from the Technion—Israel Institute of Technology, Haifa.

He has held postdoctoral appointments with the Statistics Department at Stanford University, Stanford, CA, and with Hewlett-Packard Laboratories. Currently, he is with the Department of Electrical Engineering at both Stanford University and the Technion. His research interests span information theory and its applications and statistical signal processing. His papers thus far have focused mostly on data compression, communications, prediction, denoising, and learning. He is also inventor or co-inventor of several patents in these areas and involved in a number of high-tech companies as a researcher or member of the technical board.

Dr. Weissman is a Robert N. Noyce Faculty Scholar of the School of Engineering at Stanford University and a recipient of the 2006 IEEE joint IT/COM societies Best Paper Award. His recent prizes include the NSF CAREER award and a Horev fellowship for leaders in Science and Technology.