Universal Denoising of Discrete-time Continuous-Amplitude Signals

Kamakshi Sivaramakrishnan † and Tsachy Weissman §, Member, IEEE

Abstract

We consider the problem of reconstructing a discrete-time signal with continuous valued components corrupted by a known memoryless channel. When performance is measured using a per-symbol loss function satisfying mild regularity conditions, we develop a sequence of denoisers that, although independent of the distribution of the underlying ‘clean’ signal, is universally optimal. This sequence is universal in the sense of performing as well as any sliding window denoising scheme which may be optimized for the underlying clean signal. Our results are initially developed in a “semi-stochastic” setting, where the noiseless signal is an individual unknown sequence, and the only source of randomness is due to the channel noise. It is subsequently shown that in the fully stochastic setting where the noiseless sequence is a stationary stochastic process, our schemes universally attain optimum performance. The proposed schemes draw from nonparametric density estimation techniques and are implementable with low complexity. We demonstrate efficacy of the proposed schemes in denoising gray-scale images in the conventional additive white Gaussian noise setting, with additional promising results for less conventional noise distributions.

Index Terms

Universal Denoising, kernel density estimation, Quantization, Sliding Window Denoiser, Denoisability, Memory-less Channels, semi-stochastic setting.

I. INTRODUCTION

Consider the problem of estimating a clean signal \( \{X_t\}_{t \in \mathbb{T}}, X_t \in [a, b] \subset \mathbb{R} \), from its noisy observations \( \{Z_t\}_{t \in \mathbb{T}}, Z_t \in \mathbb{R} \), where \( \{Z_t\} \) is the output of a memoryless channel whose input is \( \{X_t\} \). This problem finds applications in areas ranging from engineering, cryptography, to bioinformatics and beyond. There is significant literature on particular instantiations of this problem, most notably for the case where signal and noise components are real-valued and the noise is additive, most commonly Gaussian (cf. [6] and references therein). Solutions to this problem in [6] are based on wavelet-based soft thresholding and have various asymptotic optimality properties under a minimax criterion. The scope of wavelet-based thresholding in [6] has been extended beyond the additive white Gaussian case.
in [10], [1] where optimality is again established in an asymptotic minimax sense. The soft-thresholding scheme proposed in [1] is among the few denoisers found in the literature [10], [16] that are designed for the case of a non-Gaussian corruption mechanism. Even in this case, restrictions to additive noise and symmetry assumptions on the noise distribution are made in order to provide asymptotic performance guarantees. Compression based approaches pioneered in [19] and [7], as discussed in [26], are provably sub-optimal and suffer from non-practicality of implementation of optimal lossy compression schemes. The wavelet-based Bayesian estimation approach in [20], has demonstrated significant improvement in image denoising. However, despite much recent progress, the problem of universal denoising for discrete-time continuous-amplitude data is still a largely open problem of both theoretical and practical value. The problem is particularly relevant in new emerging areas as microarray imaging [25], array-based comparative genomic hybridization (array-CGH) [14] and medical imaging [24], [12], [17], where parametric noise models that are currently used often fail to capture the true nature of the noise.

Recently, universal denoising for discrete signals and channels was considered in [26]. The results of [26], and the denoising scheme DUDE suggested therein, although attractive theoretically, are restricted in their practicality to problems with small alphabets. This is a result of computational issues involved with collecting higher-order statistics from the noisy data, mapping an estimated channel output distribution to an estimated channel input distribution, and issues having to do with count statistics being too sparse to be reliable for even moderately large alphabet sizes. This leaves open challenges in the application of DUDE to problems like gray-scale image denoising. More recently, a modified DUDE, using ideas from lossless compression, was presented in [18]. As discussed in that work, in spite of circumventing some of the computational issues mentioned above, the approach leaves room for improvement in the denoising performance. The problem was further extended to the discrete-valued input and general output alphabet setting in [3]. This approach proposes quantization of the output alphabet space and proceeds on an approach similar to that in [26], showing that there is no essential loss of optimality in quantizing the channel output before denoising. In spite of its theoretical elegance, this approach faces similar issues as the scheme of [26], limiting its scope of applications to small channel input alphabets. The authors in [3] while conjecturing the need for mild restrictions on the channel, suggest an extension of the proposed scheme to the case where both the input and output alphabet space is continuous-valued and general. The present work proposes an extension of a two-stage DUDE-like approach in [26], [3] to the case of denoising for general alphabet spaces. A natural extension would have been to quantize both the input and the output space and apply a similar count-statistic based two-pass approach. The vast literature on nonparametric density estimation (cf. [5] and references therein), however, points to the opportunity of extracting more reliable statistics from the observed data, hence leading to a ‘better performing’ (measured under a specified loss function) denoiser. We, however, maintain the $2k + 1$-context length sliding window denoising approach in [3], [26] and show asymptotic optimality of our schemes with increasing context lengths.

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

The remainder of the paper is organized as follows. In section II, we discuss the problem setup and notations. This is followed by a description of the technical results that are key to the construction of the denoisers in section III. We establish asymptotic optimality of the proposed denoiser in the semi-stochastic setting in Section IV and provide concrete bounds on the difference between the performance of the proposed denoiser and that of the best symbol-by-symbol denoiser chosen by a genie with access to the clean signal. Section V details the extension of the proposed denoiser to one that competes with sliding window schemes of arbitrary order. Section VI discusses the implication of the performance guarantees in the semi-stochastic setting to the fully stochastic setting where the clean signal is a stationary stochastic process, rather than an individual sequence. A slightly modified version of the proposed denoiser is shown in section VII to reduce to the scheme of [3] when the clean signal has a finite alphabet. The proposed denoiser can, hence, be seen as a natural extension of the Universal Denoiser in [3] to the current setting of denoising continuous valued symbols corrupted by a continuous memoryless channel. In section VIII we present some preliminary experimental results of applying the proposed schemes to denoising of gray-scale images. We conclude in section IX with some propositions for future research directions. Throughout this paper, we maintain the flow by stating the Theorems and Lemmas corresponding to the optimality results in the main body of the paper and relegate most of the proofs to the appendices.

II. Problem Setting and Notations

Let \( x = (x_1, x_2, \cdots) \) be an individual (deterministic) noise-free source signal with components taking values in \([a, b] \subset \mathbb{R}\) and \( Y = (Y_1, Y_2, \cdots) \), \( Y_i \in \mathbb{R} \) be the corresponding noisy observations, also referred to as the output of the channel (corruption source). This setting, where both the underlying clean sequence and the noisy sequence are continuous valued, is the continuous-amplitude analog of the semi-stochastic setting discussed in [3]. The channel is 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 \( x \). Also, we denote the probability measure on \( \mathbb{R} \) corresponding to \( F_{Y|x} \) by \( \mu_x \). We assume,

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 \( x \), 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 [26] 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,

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

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

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

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

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).

C7a. The family of conditional densities, $C$, have uniformly bounded second order universal derivatives, i.e., $\exists a B_C$ s.t. $0 < B_C < \infty$ and $D_2^* (f_{Y|x}) < B_C, \forall x \in [a, b]$, where

$$D_2^* (f_{Y|x}) = \lim \inf_{h \to 0} \int \left| (f_{Y|x} * \phi_h)^{(2)} \right| dy$$

$\phi_h(x) = \frac{1}{h} \phi \left( \frac{x}{h} \right), \phi \in C^\infty, C^\infty$ is a set of functions that have infinitely many continuous derivatives with compact support and $f^{(s)}$ denotes the $s$-th 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. Note that we are not imposing the differentiability of the conditional densities of the channel themselves. We are, instead, proposing a milder constraint that the smoothed version of the channel conditional densities is “differentiable enough”. This condition is trivially satisfied if we have a family of conditional densities that have a uniformly absolutely continuous derivative.
C7b. An alternative to the previous condition on the family of conditional densities of the channel is, \( \lim_{|t| \to 0} \Omega_C(t) = 0 \), where

\[
\Omega_C(t) = \sup_{x \in [a,b]} \omega_x(t)
\]  

and

\[
\omega_x(t) = \int |f_{Y|x}(y - t) - f_{Y|x}(y)| \, dy
\]

From the fact [27] that, for any \( f \in L_1(\mathbb{R}) \), the corresponding, \( L_1 \)-modulus of continuity,

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

and

\[\|\omega\|_{\infty} \leq 2\|f\|_1 < \infty\]

it follows that the global \( L_1 \)-modulus of continuity, \( \Omega_C(t) \), is well-defined for all \( t \) and families of conditional densities, \( C \). 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.

The above, are rather benign conditions obeyed by most channels arising in practice, an example of this being the most commonly addressed channel, viz., the Additive White Gaussian Noise Channel (AWGN). It is easy to verify that even the multiplicative (non-additive) Gaussian channel with a finite variance and mean satisfies these requirements. In this case, the channel input (underlying clean signal) affects the variance of the channel. The fact that the underlying clean signal takes only bounded values implies that the tightness condition, \( C2 \), is satisfied. In fact, any additive noise channel with distribution functions that are absolutely continuous and the corresponding densities (of finite mean and variance) satisfying conditions \( C4-7 \) will satisfy the above requirements.

An \( n \)-block denoiser is a measurable mapping taking \( \mathbb{R}^n \) into \([a, b]^n\). We assume a loss function \( \Lambda : [a, b]^2 \to [0, \infty) \) and denote the normalized cumulative loss of an \( n \)-block denoiser \( \hat{X}^n \), when the underlying sequence is \( x^n \) and the observed sequence is \( y^n \), by

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

where \( \hat{X}^n(y^n)[i] \) denotes the \( i \)-th component of \( \hat{X}^n(y^n) \). In addition to the constraints on the channel, we impose some conditions on the permissible loss functions, \( \Lambda \). We assume the loss function, \( \Lambda \),

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

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

\[
||\Lambda||_L = \sup_{0 \leq \Delta < (b-a)} \frac{\lambda(\Delta)}{\Delta}
\]  

(11)

where,

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

(12)
\[ \lambda(\Delta) = \sup_{x \in [a,b]} \lambda(\Delta, x) \]  

(13)

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.

It can be easily verified that the commonly used loss functions of \( L_2, L_1 \) norms satisfy the aforementioned condition.

Let \( F[a,b] \) denote the set of all probability distribution functions with support contained in the interval \([a,b]\). For \( F \in F[a,b] \), we let

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

(14)

denote its ‘Bayes envelope’ (our assumptions on the loss function will imply existence of the minimum). In other words, \( U(F) \) 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

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

(15)

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 \( x^n \). \( D_0(x^n) \), which is the expected per-symbol loss of the optimal symbol-by-symbol rule for the individual sequence \( x^n \), will be our benchmark for assessing the performance of the universal symbol-by-symbol denoiser that we construct in the next section. The same benchmark was used also in [3]. This is slightly different than the benchmark used in [26], which corresponded to a genie that can choose the best symbol-by-symbol rule with knowledge not only of the individual sequence \( x^n \), but also of the noisy sequence realization \( Y^n \). The latter is irrelevant for our current setting where each of the components of \( Y^n \) will take on a different value, with probability one. For \( x^n \in [a,b]^n \), define

\[ F_{x^n}(x) = \frac{|\{1 \leq i \leq n : x_i \leq x\}|}{n}, \]  

(16)

i.e., the CDF associated with the empirical distribution of \( x^n \). Note that \( D_0(x^n) \) can be expressed as

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

(17)

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

\[ E_x \Lambda(x, g(Y)) = \int \Lambda(x, g(y)) f_{Y|x}(y) dy \]  

(18)
For $F \in \mathcal{F}[a,b]$, let $F \otimes C$ and $E_{F \otimes C}$ denote, respectively, probability and expectation when the channel input $X \sim F$ and $Y$ is the channel output. So that,

$$
E_{F \otimes C}(X, g(Y)) = \int_{[a,b]} E_x(\Lambda(x, g(Y))dF(x)
$$

$$
= \int_{[a,b]} \int_{\mathbb{R}} \Lambda(x, g(y))f_{Y|x}(y)dy dF(x)
$$

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

$$
\min_g E_{F \otimes C}(X, g(Y)) = E_{F \otimes C}(U([F \otimes C]_{X|Y}))
$$

with $U$ denoting the Bayes envelope as defined above. Letting $g_{\text{opt}}[F]$ denote the achiever of the minimum in (20), we note that is given by the Bayes response to $[F \otimes C]_{X|Y}$, namely,

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

$$
= \arg \min_{\tilde{x} \in [a,b]} \int_{[a,b]} \Lambda(x, \tilde{x})f_{Y|x}(y)dF(x)
$$

In Lemma 9, we will establish the concavity of $U(F)$, and minimizing this bounded (by our assumption of bounded $\Lambda$) concave function over a closed compact interval, $[a, b]$, guarantees the existence of the minimizer, $g_{\text{opt}}$. Note that from (17), (18) and (19) we have

$$
D_0(x^n) = \min_g E_{F_x \otimes C}(X, g(Y))
$$

where $F_x$ was defined in (16) and the minimum is attained by $g_{\text{opt}}[F_x^n]$. Thus, only a “genie” with access to the empirical distribution of the noiseless sequence could employ $g_{\text{opt}}[F_x^n]$.

III. CONSTRUCTION OF UNIVERSAL “SYMBOL-BY-SYMBOL” DE NOISER AND PRELIMINARIES

$F_x$ and, hence, $g_{\text{opt}}[F_x^n]$ are not known to an observer of the noisy sequence. The first step towards constructing an estimate of $g_{\text{opt}}[F_x^n]$ is to estimate the input empirical distribution from the observable noisy sequence, $Y^n$, and knowledge of the channel, $C$. 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 function we are interested in estimating is

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

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 [5], [4], the details of which are discussed in Subsection III-A below. Once we have
an estimate \( f_n^y = f_n^y[Y^n] \) for this function, we use it to estimate the input empirical distribution by

\[
\hat{F}_{x^n}[Y^n] = \arg \min_{F \in \mathcal{F}_n^{[a,b]}} d \left( f_n^y, \int f_{Y|x} dF(x) \right) 
\]

(24)

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 \( [F \otimes C]_Y \) 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

\[
F(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[x \leq x_i]}
\]

(25)

for some \( n \)-tuple, \( x^n = (x_1, x_2, \cdots, x_n) \), with \([a, b] \)-valued components. The norm, \( d \), in (24) is defined as

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

(26)

The channel, \( C \), 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, \( f_n^y \), which is constructed only from the noisy sequence, \( Y^n \), is oblivious to the set of achievable marginal densities and hence could lie outside this set. 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 (24). The uniqueness of the minimizer in (24) follows from the fact that the objective function being minimized is a norm-function and hence convex, coupled with the linear independence assumption of the channel, C4. The assumption, C4, implies a one-to-one correspondence between channel input and channel output distributions (i.e., “invertibility” of the channel). Additionally, the search for the minimizer is conducted on a convex set of distribution functions, \( \mathcal{F}_n^{[a,b]} \), resulting in uniquely achieving the minimizer or in other words, the candidate input empirical distribution estimate.

A two-stage quantization of both, the support of the underlying clean symbol, \([a, b]\), and the levels of the estimate of its empirical distribution function, \( \hat{F}_{x^n} \), is carried out to give the corresponding quantized probability mass function that has mass points only at the quantized symbols.

Q1. The quantization of the interval \([a, b]\) is depicted in Fig. 1. For a given quantization step size, \( \Delta \), the quantized symbols, \( a_i \) in the interval \([a, b]\) are constructed in the following manner.

For \( \Delta > 0 \), \( N(\Delta) = \frac{b-a}{\Delta} \), if \( m = \lfloor \frac{b-a}{\Delta} \rfloor \), consider a family of vectors,

\[
\mathcal{F}^{\Delta} = \{ P^{\Delta} : P^{\Delta} = (P(a_0), P(a_1), \cdots, P(a_{N(\Delta)})) \}
\]

\[
\mathcal{A}^{\Delta} = \{ a_i = a + i \Delta, i = 0, \cdots, N(\Delta) \}
\]

s.t.

\[
\sum_{i=1}^{N(\Delta)} P(a_i) = 1
\]

else, define the family of vectors as \( \mathcal{F}^{\Delta} = \{ P^{\Delta} : P^{\Delta} = (P(a_0), P(a_1), \cdots, P(a_{N(\Delta)-1}), P(a_{N(\Delta)})) \} \), \( \mathcal{A}^{\Delta} = \{ a_i = a + i \Delta, i = 0, \cdots, N(\Delta) - 1 \}, a_{N(\Delta)} = b, \sum_{i=1}^{N(\Delta)} P(a_i) = 1. \)

As indicated in Fig. 1, the probability mass function, \( P^{\Delta} \), that we propose is constructed by allocating the
mass of the distribution function, $F$, in any quantization interval (of length $\Delta$) to the higher end point in that interval. More precisely,

$$P^\Delta(a_i) = F(a_i) - F(a_{i-1})$$  \hspace{1cm} (27)

where $a_i$’s as defined above and note that

$$P^\Delta(B) = \sum_{a_i \in B} P(a_i)$$

with any $B \in \mathcal{B}^{[a,b]}$, $\mathcal{B}^{[a,b]}$ is the Borel sigma-algebra generated by open sets in $[a,b]$.

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, $\hat{P}^\Delta_{x^n}$

$$\hat{P}^\Delta_{x^n}(a_i) = \hat{F}_{x^n}(a_i) - \hat{F}_{x^n}(a_{i-1})$$  \hspace{1cm} (28)

where, $a_i \in A^\Delta$.

Q2. The quantization of the values $\hat{P}_{x^n}$ is carried out using a uniform quantizer, $Q_\delta$

$$\tilde{\hat{P}}^{\delta,\Delta}_{x^n} = Q_\delta(\hat{P}_{x^n})$$  \hspace{1cm} (29)

where, $\delta$ denotes the quantization step-size on the interval $[0,1]$.

This is primarily motivated by tractability of the proof of the asymptotic optimality results. But, it can also be argued that any practical implementation of this proposed denoiser only has a finite precision representation of the underlying clean symbol and the distribution function values itself. Analysis of the asymptotic optimality results also lends itself nicely to viewing the distribution of the underlying clean symbol, $\hat{F}_{x^n}$, as the asymptotic limit attained by its quantized, finite precision representation, $\hat{P}^{\delta,\Delta}_{x^n}$. This is formalized in section III-C where we discuss the precise convergence notion of $\hat{P}^\Delta_{x^n}$ to the un-quantized probability measure.
The minimizer of the Bayes envelope in (21) is then constructed from the quantized probability mass function,

\[ P_{x^n}^{\delta, \Delta}, \text{as } g_{\text{opt}} \left[ P_{x^n}^{\delta, \Delta} \right], \]

where \( g_{\text{opt}} \) for the quantized clean symbol is,

\[ g_{\text{opt}}[P](y) = \arg \min_{\hat{x} \in A^\Delta} \sum_{a \in A^\Delta} \Lambda(a, \hat{x}) \cdot f_{Y|x=a}(y) \cdot P(X = a) \tag{30} \]

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

\[ \hat{X}_{n}^{\delta, \Delta}[y^n](i) = g_{\text{opt}}[\hat{P}_{x^n}^{\delta, \Delta}(y^n)](y_i), \quad 1 \leq i \leq n \tag{31} \]

where, \( g_{\text{opt}} \) is given in (30). We now proceed to discuss in detail the construction and consistency results of the estimate, \( f_{n}^{Y}, \hat{F}_{x^n} \) and its quantized version, \( \hat{P}_{x^n}^{\delta, \Delta} \).

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

We now obtain an estimator \( f_{n}^{Y} \), for the function in (23) which depends on \( x^n \) and therefore unknown to the denoiser. Given the memoryless nature of the channel, the sequence of output symbols, \( Y_1, Y_2, \cdots, Y_n \) are independent random variables taking values in \( \mathbb{R} \), having conditional densities, \( f_{Y|x_1}, f_{Y|x_2}, \cdots, f_{Y|x_n} \) respectively. A density estimate is a sequence \( f_{1}, f_{2}, \cdots, f_{n} \), where for each \( n \), \( f_{n}[Y^n](y) = f_{n}(y; Y_{1}, \cdots, Y_{n}) \) is a real-valued Borel measurable function of its arguments, and for fixed \( n \), \( f_{n} \) is a density estimate on \( \mathbb{R} \). The kernel density estimate is given by

\[ f_{n}(y) = \frac{1}{nh^n} \sum_{i=1}^{n} K \left( \frac{y - Y_i}{h} \right) \tag{32} \]

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

\[ J_n = \int \left| f_{n}(y) - \frac{1}{n} \sum_{i=0}^{n} f_{Y|x_i}(y) \right| dy \tag{33} \]

The choice of \( L_1 \) distance as elaborated by the authors in [5] is motivated by its invariance under monotone transformations of the coordinate axes and the fact that it is always well-defined. Before proceeding to discuss convergence results for \( J_n \), we present definitions of certain types of kernel functions, \( K \), that are the backbone to kernel density estimation techniques, [4].

**Definition 1:** The class of kernels, \( \mathcal{K} \) s.t. \( \forall K \in \mathcal{K}, \) we have

\[ \int K = 1 \]

and \( K \) is symmetric about 0 are called **class 0 kernels**.

**Definition 2:** A class \( s \) kernel is a class 0 kernel for which

\[ \int |x|^s |K(x)|dx < \infty \]
and
\[ \int x^i K(x) dx = 0 \]
for all \( i = 1, \cdots, s - 1 \). Most class 0 kernels are in fact class 2 kernels, the only additional condition being that \( \int |x|^2 K(x) < \infty \). However, nonnegative class 0 kernels cannot possibly of class \( s \geq 3 \).

**Theorem 1:** Let \( K \) be a nonnegative Borel measurable function on \( \mathbb{R} \) with \( \int K = 1 \) of class \( s = 2 \). Let \( f^n_Y \) be the kernel estimate in (32) and \( J_n \), the corresponding error as defined in (33). Consider

1) \( J_n \to 0 \) in probability as \( n \to \infty \), for some sequence \( x = (x_1, x_2, \cdots) \)
2) \( J_n \to 0 \) in probability as \( n \to \infty \), for all sequences \( x = (x_1, x_2, \cdots) \)
3) \( J_n \to 0 \) almost surely as \( n \to \infty \), for all sequences \( x = (x_1, x_2, \cdots) \)
4) For all \( \epsilon > 0 \), there exist \( r, n_0 > 0 \) such that \( P(J_n \geq \epsilon) \leq e^{-rn}, n \geq n_0 \), for all sequences \( x \).
5) \( \lim_{n \to \infty} h = 0, \lim_{n \to \infty} nh = \infty \)

Then, \( 5 \Rightarrow 4 \Rightarrow 3 \Rightarrow 2 \Rightarrow 1 \).

The following lemma is key to the proof of Theorem 1.

**Lemma 1:** For any family of channel probability density functions, \( \{f_{Y|x} \mid x \in [a,b] \} \) on \( \mathbb{R} \), satisfying assumptions C1-C7, and any non-negative, integrable function \( K \), with \( \int K(x) dx = 1 \), condition 4) in Theorem 1 holds whenever
\[
\lim_{n \to \infty} h_n = 0 \quad \text{and} \quad \lim_{n \to \infty} nh_d = \infty
\]  

**Proof:** [Proof of Theorem 1]

The implication that \( 5 \Rightarrow 4 \) is proved in Lemma 1. Since clearly, \( 4 \Rightarrow 3 \Rightarrow 2 \Rightarrow 1 \), the proof of Theorem 1 is complete.

**B. Channel Inversion**

The estimate in (24) projects the kernel estimate of \( \frac{1}{n} \sum_{i=1}^n f_{Y|x_i}(y) \) to an estimate of the empirical distribution, \( F_{x^n} \). This projection is such that it best approximates (in the \( L_1 \) sense), the kernel density estimate with a member in the set of achievable channel output distributions. From the construction of \( f^n_Y \) in (32), it is clear that \( f^n_Y \) is a bona fide density on \( \mathbb{R} \). Additionally, from the construction of \( F_{x^n} \) in (24), we see that for every \( F \in F^{[n,b]}_n \), \( [F \otimes C]_Y \) is also a valid density in \( \mathbb{R} \). Finally, from the definition of the norm, \( d \), in (26), it is true that for \( f^n_Y \) and \( [F \otimes C]_Y \) being bona fide densities on \( \mathbb{R} \), \( 0 \leq d(f^n_Y, [F \otimes C]_Y) \leq 2, \forall, n \). These facts, together with the convexity
Lemma 2: \( d(f^n_Y, [\hat{F}^n_x \otimes C]_Y) \to 0 \) a.s.

Proof: By definition,

\[ 0 \leq d(f^n_Y, [\hat{F}^n_x \otimes C]_Y) \leq d(f^n_Y, [F^n_x \otimes C]_Y), \forall n \]

Taking limit \( n \to \infty \) in the inequality of (35), we get

\[ 0 \leq \lim_{n \to \infty} d(f^n_Y, [\hat{F}^n_x \otimes C]_Y) \leq \lim_{n \to \infty} d(f^n_Y, [F^n_x \otimes C]_Y) = 0 \text{ a.s.} \]

where the second part of the inequality in (35) follows from Theorem 1.

Lemma 3: \( d([F^n_x \otimes C]_Y, [\hat{F}^n_x \otimes C]_Y) \to 0 \) a.s.

Proof: 

\[ 0 \leq d([F^n_x \otimes C]_Y, [\hat{F}^n_x \otimes C]_Y) \leq d([F^n_x \otimes C]_Y, f^n_Y) + d(f^n_Y, [\hat{F}^n_x \otimes C]_Y) \]

We have already seen \( d([F^n_x \otimes C]_Y, f^n_Y) \to a.s \) and by Lemma 2,

\[ d(f^n_Y, [\hat{F}^n_x \otimes C]_Y) \to 0 \text{ a.s.} \]

Whence,

\[ d([F^n_x \otimes C]_Y, [\hat{F}^n_x \otimes C]_Y) \to 0 \text{ a.s.} \]

Definition 3 (Levy metric): The Levy distance \( \lambda(F, G) \) between any two distributions \( F \) and \( G \) is defined as

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

Theorem 2: For the estimator, \( \hat{F}^n_x \) defined in equation (24) we have \( \lambda(F^n_x, \hat{F}^n_x) \to 0 \text{ a.s.} \)

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

In this section, we discuss the convergence notion of \( \hat{P}^n_x \) to the law corresponding to the un-quantized distribution function \( \hat{F}^n_x \).

Definition 4 (\( \beta \) metric): For any two laws \( P \) and \( Q \) on \( S \), \( f : S \to \mathbb{R} \) let \( \int f d(P - Q) := \int f dP - \int f dQ \), for bounded \( \int f dP \) and \( \int f dQ \), the Prohorov metric is defined as

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

where

\[ \| f \|_{BL} = \| f \|_L + \| f \|_\infty \]

(35)
and
\[
\| f \|_L := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)}, \quad \| f \|_\infty = \sup_x |f(x)|
\]

Equipped with this definition, we now state the following theorem,

**Theorem 3:**
\[
\lim_{\Delta \to \infty} \beta \left( \hat{P}_{x^n}, \hat{P}_{x^n}^\Delta \right) = 0
\]

where, \( \hat{P}_{x^n} \) denotes the law associated with the distribution function \( \hat{F}_{x^n} \).

**Proof:** Follows directly from Lemma 4.

**Lemma 4:** For any \( F \in \mathcal{F}^{[a, b]} \), there exists, a family \( \{ P^\Delta \}_{A^\Delta} \), such that
\[
\lim_{\Delta \to 0} \beta (P, P^\Delta) = 0
\]

where, \( P \) is the law associated with distribution functions in the family \( \mathcal{F}^{[a, b]} \). Particularly, the \( P^\Delta \) that satisfies (38) has the form,
\[
P^\Delta(a_i) = F(a_i) - F(a_{i-1})
\]

where \( a_i \in A^\Delta \) and \( A^\Delta \) is the finite alphabet approximation of \([a, b]\) discussed earlier.

As a result of Theorem 3, we are able to show that uniform quantization is a continuous mapping (under the beta metric) of the empirical distribution of the underlying clean symbol to a probability mass function that corresponds to its quantized representation. In other words, we are asymptotically able to reconstruct the true empirical distribution of the clean signal with vanishing quantization steps sizes of its support.

**IV. PERFORMANCE GUARANTEES**

The main result of this section is Theorem 5 which establishes the strong universal asymptotic optimality of our proposed symbol-by-symbol denoiser in (31) with respect to the class of symbol-by-symbol schemes. The predominant technical result leading to Theorem 5 is Theorem 4. We continue to restrict ourselves to the semi-stochastic setting where the underlying clean sequence is an unknown, but deterministic sequence, \( x \). The benchmark performance for the clean sequence is, the minimum possible symbol-by-symbol loss, \( D_0(x^n) \), defined in section II.

The main result of this section is Theorem 5, where we show that the proposed denoiser, \( g_{\text{opt}} \left[ \tilde{P}_{x^n}^{\delta, \Delta} \right] \), asymptotically achieves the performance of the best possible symbol-by-symbol denoiser. This is achieved by bounding the deviation of the cumulative loss incurred by \( g_{\text{opt}} \left[ \tilde{P}_{x^n}^{\delta, \Delta} \right] \) from the minimum possible symbol-by-symbol loss in Theorem 4 for any block length, \( n \), of the underlying clean sequence. Hence we show that, \( g_{\text{opt}} \left[ \tilde{P}_{x^n}^{\delta, \Delta} \right] \) performs essentially as well as the best possible symbol-by-symbol denoiser.

In preparation of Theorem 4 we let \( \mathcal{F}_{\delta, \Delta} \) denote the set of scalars with components in \([0,1]\) that are integer multiples of \( \delta \). Note that \( \tilde{P}_{x^n}^{\delta, \Delta} \in \mathcal{F}_{\delta, \Delta} \). Also, let \( G_{\delta, \Delta} = \{ g_{\text{opt}}[P] \} \) denote the set of all possible denoisers that can be constructed from the members of the set \( \mathcal{F}_{\delta, \Delta} \) using (30).
Theorem 4: For all $\epsilon > 0$, $\rho \in (0, 1)$, $\delta > 0$, $\Delta > 0$ and $x^n \in [a, b]^n$ let,

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

then, we have

$$\Pr(|L_{\hat{X}n, \delta, \Delta}(x^n, Y^n) - D_0(x^n)| > 3\epsilon + 5\delta \Lambda_{\text{max}} + 4\delta \Delta \Lambda_{\text{max}} + 4\lambda(\Delta)(1 + \delta \Delta))$$

$$\leq |G_{\delta, \Delta}| \left[2e^{-G(\epsilon + \delta \Lambda_{\text{max}}, \Lambda_{\text{max}})n} + e^{-(1 - \rho)\frac{n \gamma^2}{2}} + e^{-(1 - \rho)\frac{n \gamma^2}{2}}\right]$$ \hspace{1cm} (40)

for all $n > n_0(C, \rho, K, h_n, \gamma)$

where, $G(\epsilon, B) = \frac{2\epsilon^2}{B^2}$ and $D_0(x^n)$ is the symbol-by-symbol minimum loss of $x^n$ defined in (15).

A. Intuition behind the Proof of Theorem 4

The benchmark for assessing the performance of the proposed denoiser is the minimum possible symbol-by-symbol cumulative loss, $D_0(x^n)$. It has been shown in (22), that this is the minimum over all measurable mappings, $g : \mathbb{R} \rightarrow [a, b]$, of the expected loss under the marginal density induced by the true distribution of the underlying clean sequence. This has been further shown in (20) to be equal to the expected value of the Bayes envelope under the true conditional empirical distribution of the underlying clean signal given the noisy observation. This true conditional empirical distribution of the underlying clean signal is the quantity that is unknown to us. However, if we have an estimate of this conditional empirical distribution that is in some sense “close” to the true conditional empirical distribution and asymptotically is essentially “it”, we are on the right track. Since, this is clearly derived as a function of the marginal distribution of the underlying clean signal, all that is needed is, “closeness” of the estimate of the marginal distribution of the underlying clean signal to the true marginal distribution. The almost sure convergence of the marginal density at the output of the memoryless channel gives us, through the mapping in (24), an estimate of the input empirical distribution that weakly converges, as shown in Theorem 2, to the true empirical distribution of the underlying clean signal. This, then subsequently lends itself to the convergence of the expected loss under the corresponding induced densities at the output of the memoryless channel. From (20) and (22), the fact that we have well-behaved (satisfying conditions C1-C7) channel conditional densities, $\{f_{Y|x} \}_{x \in [a, b]}$, and loss function, $\Lambda$ (satisfying conditions L1-L2), we can bound the deviation of the expected value of $U \left([-F \otimes C]_{X|Y}\right)$ under the two corresponding induced densities.

The goal, eventually, is to bound the deviation of the cumulative loss, $L_{\hat{X}_{\text{suniv}}}$, incurred by our proposed denoiser from $D_0(x^n)$ as a function of the block length, $n$. This is done by using Lemmas 5, 6 which formalize the deviation bounds of the expected loss under densities induced by weakly converging distributions. Finally, Lemma 7 is used to bound the deviation of the empirical expected loss from the true expected loss. The combination of these results is used to bound the deviation of $L_{\hat{X}_{n, \delta, \Delta}}$ from $D_0(x^n)$ in the proof of Theorem 4 that follows.
B. Proof of Theorem 4

Using the definition of the Lipschitz norm of the loss function, $\Lambda$, and the channel continuity function, $\delta_\Delta$, we bound the deviation of the expected value of the loss function under two marginal densities induced at the output of the memoryless channel by the corresponding empirical distributions of the underlying clean signal at the input of the memoryless channel.

**Lemma 5:** For any $F, \hat{F} \in \mathcal{F}[a,b]$, measurable $g : \mathbb{R} \to [a,b]$ and a bounded Lipschitz loss function with $E_{f,\Lambda} \Lambda(u, g(Y)) < \infty, \forall u$,

$$\left| E_{F \otimes C} \Lambda(U_0, g(Y)) - E_{\hat{F} \otimes C} \Lambda(U_0, g(Y)) \right| \leq (\| \Lambda \|_L + \Lambda_{\max} \| \delta \|_L + (b-a) \| \Lambda \|_L + \Lambda_{\max}) \beta(P, \hat{P})$$

(41)

where $P$ and $\hat{P}$ are the laws associated with $F$ and $\hat{F}$, $\beta(P, \hat{P})$ is the $\beta$ metric between the corresponding laws.

Similarly, we bound the deviation of the expected loss function under the marginal density induced by any empirical distribution at the input of the memoryless channel from that of the expected loss under the marginal density induced by the corresponding probability mass function (under the mapping discussed in section III-C), in the following Lemma

**Lemma 6:** For any $\Delta > 0, F \in \mathcal{F}[a,b]$ with the associated law $P, P^\Delta \in \mathcal{F}^\Delta$, measurable $g : \mathbb{R} \to [a,b]$ and a continuous bounded loss function with $E_{f,\Lambda} \Lambda(u, g(Y)) < \infty, \forall u$,

$$\left| E_{P^\Delta \otimes C} \Lambda(U_0, g(Y)) - E_{F \otimes C} \Lambda(U_0, g(Y)) \right| \leq \delta_\Delta \Lambda_{\max} + \lambda(\Delta) (1 + \delta_\Delta)$$

where $\lambda(\Delta)$ is the global modulus of continuity of the loss function $\Lambda$ as defined in equation (12) and $\delta_\Delta$ is as defined in (6)

**Lemma 7:** For every $n \geq 1, x^n \in [a,b]^n$, measurable $g : \mathbb{R} \to [a,b]$, and $\epsilon > 0$,

$$Pr \left( \left| \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, g(Y_i)) - E_{F_{x^n} \otimes C} \Lambda(U, g(Y)) \right| > \epsilon \right) \leq 2 \exp(-G(\epsilon, \Lambda_{\max})n)$$

(42)

**Proof:** By linearity of expectation, $\frac{1}{n} \sum_{i=1}^{n} E\Lambda(x_i, g(Y_i)) = E_{F_{x^n} \otimes C} \Lambda(U, g(Y))$. Thus, the expression inside the absolute value brackets in (42) is a sum of zero mean random variables, bounded in magnitude by $\Lambda_{\max}$. Furthermore, $\Lambda(x_i, g(Y_i))$ and $\Lambda(x_j, g(Y_j))$ are independent whenever $i \neq j$. This allows the use of Hoeffding inequality as in [3] leading to (42).
Proof: [Proof of Theorem 4] We fix \( n \geq 1, x^n \in [a, b]^n \),
\[
\left| E_{\hat{P}}^{x^n, [Y^n] \otimes C} \Lambda(U, g(Y)) - E_{F_x^n \otimes C} \Lambda(U, g(Y)) \right| \leq \left| E_{\hat{P}}^{x^n, [Y^n] \otimes C} \Lambda(U, g(Y)) - E_{F_x^n} [Y^n] \otimes C \Lambda(U, g(Y))' \right| + \left| E_{F_x^n} [Y^n] \otimes C \Lambda(U, g(Y)) - E_{F_x^n} \otimes C \Lambda(U, g(Y)) \right|
\]
Hence,
\[
Pr \left( \sup_{g : \mathcal{G} \to [a, b]} \left| E_{\hat{P}}^{x^n, [Y^n] \otimes C} \Lambda(U, g(Y)) - E_{F_x^n \otimes C} \Lambda(U, g(Y)) \right| > \epsilon + \delta \Lambda_{\max} + \delta \Delta \Lambda_{\max} + \lambda(1 + \delta) \right) \leq Pr \left( \left| E_{\hat{P}}^{x^n, [Y^n] \otimes C} \Lambda(U, g(Y)) - E_{F_x^n \otimes C} \Lambda(U, g(Y)) \right| > \epsilon \right) + Pr \left( \left| E_{F_x^n} [Y^n] \otimes C \Lambda(U, g(Y)) - E_{\hat{P}}^{x^n, [Y^n] \otimes C} \Lambda(U, g(Y)) \right| > \delta \Lambda_{\max} + \delta \Delta \Lambda_{\max} + \lambda(1 + \delta) \right)
\]
Now,
\[
Pr \left( \left| E_{\hat{P}}^{x^n, [Y^n] \otimes C} \Lambda(U, g(Y)) - E_{F_x^n \otimes C} \Lambda(U, g(Y)) \right| > \epsilon \right) \leq Pr \left( \left\| C \right\| L + \Lambda_{\max} \left\| \Lambda \right\| L + \left( b - a \right) \left\| \Lambda \right\| L + \Lambda_{\max} \right) \beta \left( P_{x^n}, \hat{P}_{x^n} \right) > \epsilon
\]
\[
\leq Pr \left( \left\| C \right\| L + \Lambda_{\max} \left\| \Lambda \right\| L + \left( b - a \right) \left\| \Lambda \right\| L + \Lambda_{\max} \right) d \left( F_{x^n} \otimes C, \hat{F}_{x^n} \otimes C \right) > \epsilon \right) \leq e^{-(1-\rho)^{\Delta^2 / 2}}, \quad \text{for all } n > n_0(C, \rho, K, h_n, \gamma)
\]
where \( \mathcal{C} \) is the family of channel densities \( \{ f_Y | x \} \). The inequality in (45) is due to Lemma 5, while the first inequality in (46) is by application of Theorem 2 and the second inequality is by due to Lemma 3 and Theorem 1. Finally, application of Lemma 6 to (44) yields
\[
Pr \left( \sup_{g : \mathcal{G} \to [a, b]} \left| E_{\hat{P}}^{x^n, [Y^n] \otimes C} \Lambda(U, g(Y)) - E_{F_x^n \otimes C} \Lambda(U, g(Y)) \right| > \epsilon + \delta \Lambda_{\max} + \delta \Delta \Lambda_{\max} + \lambda(1 + \delta) \right) \leq e^{-(1-\rho)^{\Delta^2 / 2}}, \quad \text{for all } n > n_0(C, \rho, K, h_n, \gamma)
\]
Combining (47) with Lemma 7 gives
\[
Pr \left( \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, g(Y_i)) - E_{\hat{P}}^{x^n, [Y^n] \otimes C} \Lambda(U, g(Y)) \right) > 2 \epsilon + 2 \delta \Lambda_{\max} + \delta \Delta \Lambda_{\max} + \lambda(1 + \delta) \right) \leq 2e^{-G(\epsilon + \delta \Lambda_{\max}, \Lambda_{\max})^n + e^{-(1-\rho)^{\Delta^2 / 2}}, \quad \text{for all } n > n_0(C, \rho, K, h_n, \gamma)
\]
By the union bound, (48) guarantees that for any class \( \mathcal{G} \)
\[
Pr \left( \max_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, g(Y_i)) - E_{\hat{P}}^{x^n, [Y^n] \otimes C} \Lambda(U, g(Y)) \right| > 2 \epsilon + 2 \delta \Lambda_{\max} + \delta \Delta \Lambda_{\max} + \lambda(1 + \delta) \right) \leq |\mathcal{G}| \left[ 2e^{-G(\epsilon + \delta \Lambda_{\max}, \Lambda_{\max})^n + e^{-(1-\rho)^{\Delta^2 / 2}}\right]
\]
Consequently,

\[
Pr\left(\left| L_{\tilde{X}_{n,\delta,\Delta}}(x^n, Y^n) - \min_{g \in \tilde{G}_{\delta,\Delta}} E_{\tilde{\rho}_{x^n,\delta,\Delta}} \otimes C \Lambda(U, g(Y)) \right| > 2\epsilon + 2\delta\Lambda_{\text{max}} + C\Delta\Lambda_{\text{max}} + \lambda(\Delta)(1 + \delta\Delta)\right) \\
= Pr\left(\left| \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, g(Y_i)) - E_{\tilde{\rho}_{x^n,\delta,\Delta}} \otimes C \Lambda(U, g(Y)) \right| > 2\epsilon + 2\delta\Lambda_{\text{max}} + C\Delta\Lambda_{\text{max}} + \lambda(\Delta)(1 + \delta\Delta)\right) \\
\leq Pr\left(\max_{g \in \tilde{G}_{\delta,\Delta}} \left| \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, g(Y_i)) - E_{\tilde{\rho}_{x^n,\delta,\Delta}} \otimes C \Lambda(U, g(Y)) \right| > 2\epsilon + 2\delta\Lambda_{\text{max}} + C\Delta\Lambda_{\text{max}} + \lambda(\Delta)(1 + \delta\Delta)\right) \\
\leq |\tilde{G}_{\delta,\Delta}| \left[ 2e^{-G(\epsilon + \delta\Lambda_{\text{max}}, \Lambda_{\text{max}})n} + e^{-(1-\rho)\frac{n\lambda^2}{2}} \right] \tag{50}
\]

where the first equality follows from the definition of \( \tilde{X}_{n,\delta,\Delta} \) and the fact that for any \( \tilde{P} \in \tilde{F}_{\delta,\Delta} \),

\[
\min_{g \in \tilde{G}_{\delta,\Delta}} E_{\tilde{P}} \otimes C \Lambda(U, g(Y)) = E_{\tilde{P}} \otimes C \Lambda(U, g_{\text{opt}}(P)(Y))
\]

The first inequality follows by the fact that \( \tilde{P}_{x^n,\delta,\Delta} \in \tilde{F}_{\delta,\Delta} \) and therefore \( g_{\text{opt}}(\tilde{P}_{x^n,\delta,\Delta}) \in \tilde{G}_{\delta,\Delta} \) and finally the last inequality follows from (49). It also follows, from (47), that

\[
Pr\left(\left| \min_{g \in \tilde{G}_{\delta,\Delta}} E_{\tilde{\rho}_{x^n,\delta,\Delta}} \otimes C \Lambda(U, g(Y)) - \min_{g \in \tilde{G}_{\delta,\Delta}} E_{\tilde{\rho}_{x^n,\delta,\Delta}} \otimes C \Lambda(U, g(Y)) \right| > \epsilon + \delta\Lambda_{\text{max}} + \delta\Lambda_{\text{max}} + \lambda(\Delta)(1 + \delta\Delta) \right) \leq e^{-(1-\rho)\frac{n\lambda^2}{2}} \tag{51}
\]

Combining (50) and (51) gives

\[
Pr\left(\left| L_{\tilde{X}_{n,\delta,\Delta}}(x^n, Y^n) - \min_{g \in \tilde{G}_{\delta,\Delta}} E_{\tilde{\rho}_{x^n,\delta,\Delta}} \otimes C \Lambda(U, g(Y)) \right| > 3\epsilon + 3\delta\Lambda_{\text{max}} + 2\delta\Lambda_{\text{max}} + 2\lambda(\Delta)(1 + \delta\Delta) \right) \\
\leq |\tilde{G}_{\delta,\Delta}| \left[ 2e^{-G(\epsilon + \delta\Lambda_{\text{max}}, \Lambda_{\text{max}})n} + e^{-(1-\rho)\frac{n\lambda^2}{2}} \right] + e^{-(1-\rho)\frac{n\lambda^2}{2}} \tag{52}
\]
On the other hand, letting \( \hat{F}^{|G, \Delta}_{x^n} \) denote the element in \( \mathcal{F}_{|G, \Delta} \) closest (under the Prohorov metric of the corresponding measures) to \( F_{x^n} \),

\[
\left| D_0(x^n) - \min_{g \in \mathcal{G}_{|G, \Delta}} E_{F_{x^n} \otimes C} \Lambda(U, g(Y)) \right| = \min_{F \in \mathcal{F}_{|G, \Delta}} E_{\hat{F}^{|G, \Delta}_{x^n} \otimes C} \Lambda(U, g_{\text{opt}}(F)(Y)) - \min_{g \in \mathcal{G}_{|G, \Delta}} E_{F_{x^n} \otimes C} \Lambda(U, g(Y)) \leq \min_{F \in \mathcal{F}_{|G, \Delta}} E_{\hat{F}^{|G, \Delta}_{x^n} \otimes C} \Lambda(U, g(Y)) - \min_{g \in \mathcal{G}_{|G, \Delta}} E_{F_{x^n} \otimes C} \Lambda(U, g(Y)) + \Lambda_{\text{max}} \delta + \delta \Lambda_{\text{max}} + \lambda(\Delta)(1 + \delta \Delta) = \min_{g \in \mathcal{G}_{|G, \Delta}} E_{\hat{F}^{|G, \Delta}_{x^n} \otimes C} \Lambda(U, g(Y)) - \min_{g \in \mathcal{G}_{|G, \Delta}} E_{F_{x^n} \otimes C} \Lambda(U, g(Y)) + \Lambda_{\text{max}} \delta + \delta \Lambda_{\text{max}} + \lambda(\Delta)(1 + \delta \Delta) \leq 2 (\Lambda_{\text{max}} \delta + \delta \Lambda_{\text{max}} + \lambda(\Delta)(1 + \delta \Delta))
\]

(53)

(54)

(55)

(56)

(57)

where (54) and (57) follow from Lemma 6, and (55) follows from the fact that the achiever of the minimum in the first term of (54) is \( \hat{F}^{|G, \Delta}_{x^n} \) which, by definition, is a member of \( \mathcal{F}_{|G, \Delta} \). Finally, combining (51) with (57) gives

\[
\Pr \left( |L_{X_{n, \Delta}}(x^n, Y^n) - D_0(x^n)| > 3\varepsilon + 5\delta \Lambda_{\text{max}} + 4\delta \Lambda_{\text{max}} + 4\lambda(\Delta)(1 + \delta \Delta) \right) \leq |\mathcal{G}_{|G, \Delta}| \left[ e^{-G(\varepsilon + \delta \Lambda_{\text{max}})n} + e^{-(1 - \rho) \frac{n \varepsilon^2}{2}} \right] + e^{-(1 - \rho) \frac{n \delta^2}{2}}
\]

(58)

for all \( n > n_0(C, \rho, K, h_n, \gamma) \).

Note that the tightness condition on the probability measures associated with the family of the conditional densities of the channel, \( C \), guarantees that \( n_0(C, \rho, K, h_n, \gamma) < \infty \) \( \forall \rho \in (0, 1) \) and \( \gamma > 0 \). From the definition of \( \mathcal{G}_{|G, \Delta} \), it is clear that \( |\mathcal{G}_{|G, \Delta}| \leq \left[ \frac{1}{\delta} + 1 \right]^{\Delta} \). Let,

\[
\alpha(\varepsilon, \delta, \Delta, \rho, \gamma) = \left[ \frac{1}{\delta} + 1 \right]^{\Delta} e^{-nG(\varepsilon + \delta \Lambda_{\text{max}}) \lambda_{\text{max}}} + e^{-(1 - \rho) \frac{n \varepsilon^2}{2}}
\]

(59)

Take now, \( \delta = \delta_n, \Delta = \Delta_n \) such that \( \delta_n \downarrow 0, \Delta_n \downarrow 0 \) for all \( \varepsilon > 0 \) and \( \sum_{n=1}^{\infty} \alpha(\varepsilon, \delta_n, \Delta_n, \rho, \gamma) < \infty \). For example, \( \delta_n, \Delta_n = \frac{1}{\log n} \) would satisfy the above requirements of summability and growth for any \( \varepsilon > 0 \). With the growth rates that satisfy the summability condition for \( \alpha(\varepsilon, \delta_n, \Delta_n, \rho, \gamma) \) let,

\[
\hat{X}^n_{\text{ssuniv}} = \hat{X}^n_{X_{n, \Delta}}
\]

(60)

where, the subscript ‘ssuniv’ stands for symbol-by-symbol universal denoiser. A direct consequence of Theorem 4 and the Borel-Cantelli Lemma gives us the following main theorem that establishes universal asymptotic optimality
of our proposed symbol-by-symbol denoiser for any unknown individual underlying clean sequence, $x$.

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

$$\lim_{n \to \infty} \left[ L_{\hat{X}_{\text{univ}}} (x^n, Y^n) - D_0(x^n) \right] = 0 \quad a.s. \quad (61)$$

**V. CONSTRUCTION OF THE UNIVERSAL DENOISER AND ITS PERFORMANCE GUARANTEES**

In this section, we propose an extension of the symbol-by-symbol denoiser discussed in section III to a $2k + 1$-length sliding window denoising scheme. The performance guarantees made in the symbol-by-symbol case also hold in the proposed extension. We discuss concrete bounds for the performance deviation of the proposed scheme from the best possible $2k + 1$-length sliding window scheme for a finite block length data, as chosen by a “genie” that knows the clean sequence $x^n$, and also discuss asymptotic results. The main result in this section is Theorem 7, which establishes the strong universal asymptotic optimality of our proposed sliding window denoiser, the length of which grows asymptotically with the block length, $n$, according to certain prescriptions discussed in Theorem 7 with respect to the class of sliding window denoisers of the same order. Analogous to Theorem 4 in the symbol-by-symbol case, Theorem 6 leads to the main universal asymptotic sliding window result in Theorem 7.

**A. Extension to the $2k + 1$-window length Denoiser**

The scheme is pictorially depicted in Fig. 2 below. The necessity for independence of the symbols in the density estimation procedure discussed in section III-A coupled with the memoryless nature of the channel is the motivation

![Fig. 2. Schematic representation of the $2k + 1$-length sliding window denoiser](image-url)
for partitioning the problem into subsequences that are processed similarly, but separately. A $2k + 1$-tuple super-symbol is formed by jumping a length of $2k + 1$ to achieve the independence condition between the successive symbols in this case. Note that there are $2k + 1$ such subsequences and each subsequence, $i$ (counting in the order of symbols in the sequence), has $\binom{[\frac{n-2k-i-1}{2k+1}] - 1}{2k+1}$ $2k + 1$-tuple super symbols. We label the subsequences as $x^{n_i}$, for $1 \leq i \leq 2k + 1$. For a fixed $n$, each subsequence $x^{n_i}$ has the following super symbols,

$$x^{n_i} = \{x^{2k+i}, x^{2k+1+i}, \cdots, x^{2k+i+2k}\}$$

This facilitates the extension of the ideas from the symbols of the symbol-by-symbol denoiser to the super-symbol of the $2k + 1$ sliding window denoiser. Some definitions are in order before we set to investigate the optimality results of the scheme. As in the symbol-by-symbol scheme, let $f^n_{Y,i}$ denote the $k$th order density estimate of the noisy sequence of symbols and is computed exactly as in (32) except $y, Y_i \in \mathbb{R}^{2k+1}$. Denote $F^{[a,b],k}$ to be the set of all probability distribution functions with support contained in the hypercube $[a, b]^{2k+1}$. Let $D_k(x^n)$ denote the $k$th-order sliding window minimum loss and is defined as

$$D_k(x^n) = \min_g E_{\mathcal{F}^{[a,b],k}} \Lambda(X, g(Y^{k}_{n,k}))$$

Note the similar definition of symbol-by-symbol denoisability in (15). As before, $D_k(x^n)$ can be expressed as

$$D_k(x^n) = \min_g E_{\mathcal{F}^{[a,b],k}} \Lambda(X, g(Y^{k}_{n,k}))$$

where $F^{[a,b],k}_{x^n}$ is the $k$th order empirical distribution of the source. Define further the sliding window denoisability of the individual sequence $x = (x_1, x_2, x_3, \cdots)$ by

$$D(x) = \lim_{k \to \infty} \limsup_{n \to \infty} \min D_k(x^n)$$

where, the limit exists by monotonicity.

Extending the definition of $k$th-order minimum loss to a subsequence, $x^{n_i}$ as

$$D_k(x^{n_i}) = \min_g E_{\mathcal{F}^{[a,b],k}_{x^{n_i}}} \Lambda(X, g(Y^{k}_{n,k}))$$

The mapping to the corresponding $k$th order input empirical distribution is given by

$$\hat{F}^{[a,b],k}_{x^n}[Y^n] = \arg \min_{F \in \mathcal{F}^{[a,b],k}} d(F_{Y^n}, \int f^{Y, k}_{y^{n}, i} \prod_{i=-k}^{k} f_{Y_i|Y_i} dF^{k}_{x^{n_i}} | F \otimes \mathcal{C}^{[a,b],k}_{Y})$$

where $\mathcal{F}^{[a,b],k}_{n, i} \subseteq \mathcal{F}^{[a,b],k}$ denotes the set of $k$th order $(1 \leq k \leq \lfloor \frac{n}{2k+1} \rfloor)$ empirical distributions induced by $n$-tuples with $[a, b]^{2k+1}$-valued components. $\hat{F}^{[a,b],k}_{x^n}[n, i]$ denotes the $k$-th order estimate of the input empirical distribution of the source analogously defined as in the symbol-by-symbol case. The $2k + 1$-length sliding window denoiser for each of the subsequences, $i$, is given by

$$\hat{X}^{n_i, \Delta, k}[n, j] = \arg \min_{\tilde{F}^{[a,b],k}_{x^n_i}[n]} \left( \tilde{F}^{[a,b],k}_{x^n_i}[n, j^{i+k}] \right)_{j \in \{k + i, 3k + 1 + i, \cdots, \left\lfloor \frac{n - 2k - i - 1}{2k + 1} \right\rfloor \}}$$
where the $k^{th}$ order equivalent of the denoiser in (30) is given by
\[
g_{opt}[P] (y_{-k}^k) = \arg \min_{\hat{x} \in A} \Lambda(\cdot, \hat{x})^T [P \otimes C]_{U} | y_{-k}^k
\]
\[
= \arg \min_{\hat{x} \in A} \sum_{\alpha \in A} \Lambda(\alpha, \hat{x}) \cdot \left\{ \sum_{u \in F^k \in A^{2k+1}: u_0 = a} \prod_{i=1}^k f_{y_i | x = u_i} (y_i) P (U_{-k}^k = u_{-k}^k) \right\} \tag{68}
\]

Let, $F^k_{\delta, \Delta}$ denote the set of $2k + 1$- dimensional vectors with components in $[0, 1]$ that are integers multiples of $\delta$. Note that, $F^k_{x, |z^n|} \in F^k_{\delta, \Delta}$ for all $z^n$. Finally, let $G^k_{\delta, \Delta} = \{g_{opt}[P]\}_{P \in F^k_{\delta, \Delta}}$ and
\[
\tilde{X}^{n, \delta, \Delta, k} = \{\tilde{X}^{n_i, \delta, \Delta, k}\}_{1 \leq i \leq 2k+1}
\]
be our candidate for the $n$-block $2k + 1$-length sliding window denoiser. It is the sequence of $2k + 1$ denoisers that operate individually on each of the subsequences. The cumulative loss incurred by this sequence of denoisers is defined as
\[
L_{\tilde{X}^{n, \delta, \Delta, k}} = \frac{1}{2k + 1} \sum_{i=1}^{2k+1} L_{\tilde{X}^{n_i, \delta, \Delta, k}} \tag{70}
\]
where, $L_{\tilde{X}^{n_i, \delta, \Delta, k}}$ is the cumulative loss incurred by the proposed denoiser for the $i^{th}$- subsequence. The following Lemma illustrates a rather intuitive fact, the average minimum $k^{th}$ order sliding window loss incurred by operating on each of the subsequences is at most the minimum $k^{th}$ order sliding window loss for the entire sequence.

**Lemma 8:** For all $n \geq 1$, $k \leq \lfloor \frac{n}{2} \rfloor$,
\[
\frac{1}{2k + 2} D_k(x^n) \leq D_k(x^n) \tag{71}
\]

**B. Performance Guarantees**

In this section we discuss Theorem 7 wherein we demonstrate that, provided certain growth constraints on the context length $k$, quantization step sizes $\delta, \Delta$ and width of the Kernel density estimate, $h$ are satisfied the cumulative loss, $L_{\tilde{X}^{n, \delta, \Delta, k}}$, incurred by the proposed denoiser asymptotically approaches the sliding window denoisability. The exact handle on the growth constraints are discussed at the end of this section. They are a fall out of sufficient conditions to prove Theorem 7 from exponential deviation bounds of the cumulative loss, $L_{\tilde{X}^{n, \delta, \Delta, k}}$ from $D_k$ for any underlying clean sequence $x^n$.

Let
\[
\alpha (\epsilon, k, \delta, \Delta, \rho, \gamma_k) = \left[ \frac{1}{\delta} + 1 \right]^{\Delta^{2k+1}} \cdot \left[ A (k, \epsilon + \delta \Lambda_{\text{max}}, \Lambda_{\text{max}}) \exp \left( -(n + 1) G (k, \epsilon + \delta \Lambda_{\text{max}}, \Lambda_{\text{max}}) \right) + \right.
\]
\[
A \left( k, \sqrt{1 - \rho}, \frac{2}{\gamma_k} \right) \exp \left( -(n + 1) G \left( k, \sqrt{1 - \rho}, \frac{2}{\gamma_k} \right) \right) \left. \right] + e^{-\left(\rho \frac{(n-2k)\gamma_k^2}{2}\right)}
\]
where,

\[ A(k, \epsilon, B) = (2k + 1) \exp \left( \frac{2\epsilon^2}{B^2} \right) \]  \hspace{1cm} (72)

\[ G(k, \epsilon, B) = \frac{2\epsilon^2}{(2k + 1)B^2} \]  \hspace{1cm} (73)

We now state the \( k \)-th-order equivalent of Theorem 4 which bounds the deviation of the cumulative loss incurred by the proposed \( 2k + 1 \)-length sliding window denoiser from the minimum possible \( k \)-th-length sliding window loss, \( D_k(x^n) \). Note here, \( x \in [a, b]^{2k+1} \) and \( Y \in [a, b]^{2k+1} \) (\( 2k+1 \)-tuple supersymbols) is the output of the continuous valued memoryless channel.

**Theorem 6:** For all \( n \geq 1, \epsilon > 0, \Delta > 0, 1 \leq k \leq \left\lfloor \frac{n}{2} \right\rfloor \) and \( x^n \in [a, b]^n \)

\[ Pr(\left| L_{X_n, \delta, \Delta, k}(x^n, Y^n) - D_k(x^n) \right| > 3\epsilon + 5\delta\Lambda_{\max} + 4\delta\Lambda_{\max} + 4\lambda(\Delta)(1 + \delta_{\Delta})) \leq \alpha(k, \epsilon, \Delta, \rho, \gamma_k) \]

for all \( n > n_0(C, k, K, h_n, \gamma_k) \)

**Proof:**

\[
L_{X_n, \delta, \Delta, k}(x^n, Y^n) - D_k(x^n) = L_{X_n, \delta, \Delta, k}(x^n, Y^n) - \frac{1}{2k + 1} \sum_{i=1}^{2k+1} D_k(x^{n_i}) + \frac{1}{2k + 1} \sum_{i=1}^{2k+1} D_k(x^{n_i}) - D_k(x^n) \] \hspace{1cm} (74)

From Lemma 8, we have

\[
L_{X_n, \delta, \Delta, k}(x^n, Y^n) - D_k(x^n) \leq L_{X_n, \delta, \Delta, k}(x^n, Y^n) - \frac{1}{2k + 1} \sum_{i=1}^{2k+1} D_k(x^{n_i}) 
= \frac{1}{2k + 1} \sum_{i=1}^{2k+1} L_{X_{n_i}, \delta, \Delta, k}(x^{n_i}, Y^{n_i}) - \frac{1}{2k + 1} \sum_{i=1}^{2k+1} D_k(x^{n_i}) 
\leq \frac{1}{2k + 1} \sum_{i=1}^{2k+1} \left| L_{X_{n_i}, \delta, \Delta, k}(x^{n_i}, Y^{n_i}) - D_k(x^{n_i}) \right| \] \hspace{1cm} (75)

Hence,

\[
Pr \left( L_{X_n, \delta, \Delta, k}(x^n, Y^n) - D_k(x^n) > 3\epsilon + 5\delta\Lambda_{\max} + 4\delta\Lambda_{\max} + 4\lambda(\Delta)(1 + \delta_{\Delta}) \right) \leq Pr \left( \sum_{i=1}^{2k+1} \left| L_{X_{n_i}, \delta, \Delta, k}(x^{n_i}, Y^{n_i}) - D_k(x^{n_i}) \right| > 3\epsilon + 5\delta\Lambda_{\max} + 4\delta\Lambda_{\max} + 4\lambda(\Delta)(1 + \delta_{\Delta}) \right) \leq \sum_{i=1}^{2k+1} Pr \left( \left| L_{X_{n_i}, \delta, \Delta, k}(x^{n_i}, Y^{n_i}) - D_k(x^{n_i}) \right| > 3\epsilon + 5\delta\Lambda_{\max} + 4\delta\Lambda_{\max} + 4\lambda(\Delta)(1 + \delta_{\Delta}) \right) \leq (2k + 1) G_{\delta_{\Delta}}^k \left[ e^{-G(\epsilon + \delta\Lambda_{\max}) (\frac{n - 2k}{20k + 1})} + e^{-\left(1 - \frac{n - 2k}{20k + 1}\right) (\frac{n - 2k}{20k + 1})^2} \right] + e^{-\left(1 - \frac{n - 2k}{20k + 1}\right) (\frac{n - 2k}{20k + 1})^2} \] \hspace{1cm} (76)

This is true by applying Theorem 12 to the \( 2k + 1 \) subsequences of independent supersymbols with at most \( \frac{n - 2k}{20k + 1} \) supersymbols in each of them. Also, the cardinality of the set of all possible proposed \( 2k + 1 \)-length sliding window denoisers is bounded by the cardinality of the set of all possible quantized \( k \)-th-order probability mass functions,
\[ P_{x^n}^{\delta,\Delta, k}, \text{i.e.,} \ |G_{x^n}^{k}| \leq \left[ \frac{1}{\theta} + 1 \right]^{\Delta^2 k + 1}. \]

Take now, \( k = k_n, \delta = \delta_n \) and \( \Delta = \Delta_n \) such that \( k_n \to \infty, \delta_n \downarrow 0, \Delta_n \downarrow 0, \sum_{n=1}^{\infty} \alpha (k_n, \epsilon, \delta_n, \Delta_n, \rho, \gamma_k) < \infty \) and \( n_0 (C, \rho, K, h_n, \gamma_k) < \infty \). With growth rates that satisfy these conditions let,

\[ \hat{X}_{univ} = \tilde{X}_{n, \delta_n, \Delta_n, k_n} \quad (77) \]

For example, it can be verified that unbounded increasing \( k_n = \log (\log(n)) \), \( h_n = \frac{1}{\log(n)} \), \( \delta_n k_n \to 0, \left( \delta_n, \Delta_n = \frac{1}{\log(n)} \right) \) satisfies the requirements for a family, \( C \), that has \( \delta^{2k_n+1} \Delta_n \to 0 \) and loss functions that have \( \lambda (\Delta_n) \delta^{2k_n+1} \Delta_n \to 0. \)

We now have the following result as a direct consequence of Theorem 4 and the Borel-Cantelli Lemma.

**Theorem 7:** For all \( x \in [a, b]^{\infty} \)

\[ \lim_{n \to \infty} \left[ L_{\hat{X}_{univ}}^{\hat{X}_{univ}} (x^n, Y^n) - D_k(x^n) \right] = 0 \quad a.s. \quad (78) \]

In fact, we can go a step further and show that the \( \lim \sup \) of the cumulative loss incurred by the proposed denoiser is bounded by the sliding window denoisiability. Specifically,

**Corollary 1:** For all \( x \in [a, b]^{\infty} \)

\[ \lim_{n \to \infty} \sup \left[ L_{\hat{X}_{univ}}^{\hat{X}_{univ}} (x^n, Y^n) - D(x^n) \right] \leq 0 \quad a.s. \quad (79) \]

which is an immediate corollary of Theorem 7.

**Proof:** [Proof of Corollary 1] The proof of the above corollary is the same as Corollary 1 in [3].

**VI. Stochastic Setting**

Our results also imply optimality for the stochastic setting when the source (clean signal) is now a stationary stochastic process with distribution \( F_X \). For the distribution pair \( (F_X, C) \), defining the denoisiability, \( \mathbb{D}(F_X, C) \), now as

\[ \mathbb{D}(F_X, C) = \lim_{n \to \infty} \min_{X^n} E L_{\hat{X}_n} (X^n, Y^n) \quad (80) \]

where, the expectation is assuming \( X^n \) are the first \( n \) symbols emitted by a source with distribution \( F_X \) and \( Y^n \) is, as defined before, the \( n \)-tuple of output noisy symbols from the channel \( C \) that corrupts \( X^n \). This is achieved by a “genie” that has access to the true distribution, \( F_X \), of the underlying clean signal, \( X \). It has been shown in [26], [3] that the limit in (80) exists and hence the denoisiability, \( \mathbb{D}(F_X, C) \), is well-defined for every stationary \( F_X \).

We now state the main result for the stochastic setting wherein, we establish that, for any stationary underlying clean sequence \( X \sim F \), the expected cumulative loss incurred by our proposed scheme asymptotically achieves the denoisiability, \( \mathbb{D}(F_X, C) \).
Theorem 8: For all stationary \( X \)

\[
\lim_{n \to \infty} EL_{X_{univ}}^{\hat{X}(n)}(X^n, Y^n) = D(F_X, C) \quad (81)
\]

If \( X \) is also ergodic then

\[
\lim_{n \to \infty} L_{X_{univ}}^{\hat{X}(n)}(X^n, Y^n) = D(F_X, C) \quad a.s. \quad (82)
\]

The proof of this theorem is essentially the same as Theorem 3 in [3] except for some subtle differences in our setting due to the continuous input and output alphabets. We, however, do provide the proof of the above statement for completeness, simultaneously illustrating the differences in our setting. The following Claim is necessary for the proof of Theorem 8.

Claim 1:

\[
\lim_{k \to \infty} \min_{g} E \Lambda \left( X_0, g \left( Y_{\hat{k}}^k \right) \right) = D(F_X, C)
\]

The claim results from the following lemma.

Lemma 9:

- For \( k, l \geq 0 \), \( E \mathcal{U} \left( F_{X_0 | Y_{\hat{l}}^l} \right) \) is decreasing in both \( k \) and \( l \).
- For any two unboundedly increasing sequences of positive integers \( \{k_n\}, \{l_n\} \),

\[
\lim_{n \to \infty} E \mathcal{U} \left( F_{X_0 | Y_{\hat{l}_n}^{l_n}} \right) = E \mathcal{U} \left( F_{X_0 | Y_{\infty}} \right) \quad (83)
\]

Proof: A direct consequence of the definition of the Bayes envelope \( \mathcal{U}(\cdot) \) is a concave function. Specifically, for two distribution functions \( F \) and \( G \) defined on \([a, b]\), and \( \alpha \in [0, 1] \),

\[
\mathcal{U}(\alpha F + (1 - \alpha)G) = \min_{\hat{x} \in [a, b]} \int_{x \in [a, b]} \Lambda(x, \hat{x})d(\alpha F + (1 - \alpha)G)(x)
\]

\[
= \alpha \min_{\hat{x} \in [a, b]} \int_{x \in [a, b]} \left[ \Lambda(x, \hat{x})dF(x) + (1 - \alpha)\Lambda(x, \hat{x})dG(x) \right] +
\]

\[
\geq \alpha \min_{\hat{x} \in [a, b]} \int_{x \in [a, b]} \Lambda(x, \hat{x})dF(x) + (1 - \alpha) \min_{\hat{x} \in [a, b]} \int_{x \in [a, b]} \Lambda(x, \hat{x})dG(x)
\]

\[
= \alpha \mathcal{U}(F) + (1 - \alpha)\mathcal{U}(G)
\]

where the first equality follows from the fact that the mapping, \( F \mapsto Ff, Ff = \int f dF \), for a bona fide distribution.
function, is linear. Next, to show that \( E\mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{t+k}^l} \right) \) decreases with \( l \), observe that

\[
E\mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{t+k}^l} \right) = \int_{g_{t+k}^l} \mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{t+k}^l} \right) dF_{Y_{t+k}^l}
\]

\[
= \int_{g_{t+k}^l} \mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{t+k}^l} \right) dF_{Y_{t+k}^l}
\]

\[
\leq \int_{g_{t+k}^l} \mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{t+k}^l} \right) dF_{Y_{t+k}^l}
\]

\[
= \int_{g_{t+k}^l} \mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{t+k}^l} \right) dF_{Y_{t+k}^l}
\]

\[
= \int_{g_{t+k}^l} \mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{t+k}^l} \right) dF_{Y_{t+k}^l}
\]

\[
= \int_{g_{t+k}^l} \mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{t+k}^l} \right) dF_{Y_{t+k}^l}
\]

\[
= \int_{g_{t+k}^l} \mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{t+k}^l} \right) dF_{Y_{t+k}^l}
\]

where, the first inequality follows from the fact that \( \mathcal{U} \) is a concave functional mapping. The definition of \( [F \otimes \mathcal{C}]_{X|Y} \) is bona fide from the assumption that the family of conditional measures, \( \mathcal{C} \), is absolutely continuous. Finally, application of Fubini’s theorem permits the change of order of integration to achieve the final inequality. The fact that \( F_{X|Y_{t+k}} \rightarrow F_{X|Y_{\infty}} \) a.s., implying \( F_{X|Y_{t+k}} \rightarrow F_{X|Y_{\infty}} \). Using the convergence of random measures [15, Theorem 16.16], we have

\[
F_{X|Y_{t+k}} \rightarrow F_{X|Y_{\infty}}, \forall f \in C_K^+,\text{ the class of continuous positive valued functions with compact support. Here, the notation } Ff = \int f dF \text{ for any measurable } f \text{ and bona fide probability distribution function, } F. \text{ In section IV, we have imposed the condition of continuity of the loss function, } \Lambda, \text{ and since the input alphabet space is restricted to a closed compact interval } [a, b], \text{ we satisfy the condition, } \Lambda \in C_K^+. \text{ Hence, we have, } F_{X|Y_{t+k}} \Lambda (\cdot, \hat{x}) \rightarrow F_{X|Y_{\infty}} \Lambda (\cdot, \hat{x}), \forall, \hat{x}. \text{ Since } \Lambda (\cdot, \hat{x}) : [a, b] \times [a, b] \rightarrow \mathbb{R}^+ \text{ is a continuous mapping, in } \hat{x}, \min_{x \in [a, b]} \int \Lambda (x, \hat{x}) dF(x) \text{ is also a continuous mapping. Using the fact that } \Lambda \text{ is a bounded mapping and the continuous mapping theorem [9], } \mathcal{U} \left( F_{X|Y_{t+k}} \right) \rightarrow \mathcal{U} \left( F_{X|Y_{\infty}} \right) \text{ and } E\mathcal{U} \left( F_{X|Y_{t+k}} \right) \rightarrow E\mathcal{U} \left( F_{X|Y_{\infty}} \right). \]

The proof for Claim 1 is very similar to that of Claim 2 in [26] but we, nevertheless, discuss it here for the sake of completeness.
Proof: [Proof of Claim 1]

\[ \mathbb{D}(F_{X^n}, C) = \min_{X^n \in \mathcal{D}_n} EL_{X^n}(X^n, Y^n) = \frac{1}{n} \sum_{i=1}^{n} \min_{\hat{X} \in \mathbb{R}^{n-[a,b]}} E\Lambda(X_i, \hat{X}(Y^n)) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}^{n}} \min_{\hat{x} \in [a,b]} E[\Lambda(X_i, \hat{x})|Y^n = y^n] dF_{Y^n} \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}^{n}} \mathcal{U}(F_{X_i}|Y^n = y^n) dF_{Y^n} \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \mathcal{U}(F_{X_i}|Y^n = y^n) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{U}(F_{X_i}|Z_{1-i}) \] (85)

where the last equality follows by stationarity. Since by Lemma 9, \( \mathcal{U}(F_{X_i}|Y^n = y^n) \geq \mathcal{U}(F_{X_0}|Y^n = y^n) \), it follows from (85) that \( \mathbb{D}(F_{X^n}, C) \geq \mathcal{U}(F_{X_0}|Y^n = y^n) \) for all \( n \) and, therefore, \( \mathbb{D}(F_X, C) \geq \mathcal{U}(F_{X_0}|Y_{-k}) \). On the other hand, for any \( k, 0 \leq k \leq n \), Lemma 9 and (85) yield the upper bound

\[ \mathbb{D}(F_X, C) \leq \frac{1}{n} \left[ 2k \mathcal{U}(F_{X_0}) + \sum_{i=1}^{n-k} \mathcal{U}(F_{X_0}|Y^n_{1-i}) \right] \] (86)

\[ \leq \frac{1}{n} \left[ 2k \mathcal{U}(F_{X_0}) + \sum_{i=1}^{n-k} \mathcal{U}(F_{X_0}|Y^k_{1-k}) \right] \] (87)

\[ = \frac{1}{n} \left[ 2k \mathcal{U}(F_{X_0}) + (n - 2k) \mathcal{U}(F_{X_0}|Y^k_{1-k}) \right] \] (88)

Considering the limit as \( n \to \infty \) of both ends of the above chain yields \( \mathbb{D}(F_X, C) \leq \mathcal{U}(F_{X_0}|Y_{-k}) \). Letting now \( k \to \infty \) and invoking Lemma 9 implies \( \mathbb{D}(F_X, C) \leq \mathcal{U}(F_{X_0}|Y_{-k}) \).

\[ \blacksquare \]

Proof: [Proof of Theorem 8]

By definition of \( \mathbb{D}(F_X, C) \) clearly

\[ \lim_{n \to \infty} \inf_{X_n \in \mathcal{D}_n} EL_{X_n}(X^n, Y^n) \geq \mathbb{D}(F_X, C) \]

On the other hand, from (63), for any \( k \)

\[ ED_k(X^n) = E \min_{g} E_{F^n_{X \oplus C}} \Lambda(X, g(Y^k X_{-k})) \]

\[ \leq \min_{g} E \left[ E_{F^n_{X \oplus C}} \Lambda(X, g(Y^k X_{-k})) \right] \]

\[ = \min_{g} E\Lambda(X, g(Y^k X_{-k})) \] (89)

where, the right side \( X^k_{-k} \) is emitted from the (unique) double-sided extension of the source \( F_X \). Using the result from equation (89), we get

\[ \lim_{n \to \infty} \sup ED_{k_n}(X^n) \leq \mathbb{D}(F_X, C) \] (90)

implying, by Theorem 7 and bounded convergence, that

\[ \lim_{n \to \infty} \sup ED_{k_n}(X^n) \leq \mathbb{D}(F_X, C) \]

implying, by Theorem 7 and bounded convergence, that

\[ \lim_{n \to \infty} \sup EL_{X_n}(X^n, Y^n) \leq \mathbb{D}(F_X, C) \] (91)
and proving (81). To prove (82) assume stationary ergodic X. We have established the continuity of \( E_{F \otimes \mathcal{C}}A(\mathcal{U}_0, g(Y)) \) w.r.t \( F \in \mathcal{F}^{[a, b]} \) in Lemma 5 and it is easily extendible to \( \min_{g} E_{F \otimes \mathcal{C}}A(\mathcal{U}_0, g(Y)) \). By the ergodic theorem and continuity of \( \min_{g} E_{F \otimes \mathcal{C}}A(\mathcal{U}_0, g(Y)) \) in \( F \in \mathcal{F}^{[a, b]} \), it follows from the representation in (63) that
\[
D_k (X) = \lim_{n \to \infty} D_k (X^n) = \min_{g} EA (X_0, g (Y_{-k}^k)) \quad a.s. \tag{92}
\]
and by Claim 1,
\[
D(X) = D (F_X, C) \quad a.s. \tag{93}
\]
Thus, the fact that \( \limsup_{n \to \infty} D_{k_n} (x) \), \( \forall x \in [a, b]^\infty \) (recall proof of Corollary 1), combined with Theorem 7, implies
\[
\limsup_{n \to \infty} L_{X_n} (X^n, Y^n) \leq D (F_X, C) \quad a.s. \tag{94}
\]
On the other hand, by Fatou’s lemma and definition of \( D (F_X, C) \)
\[
E \left[ \limsup_{n \to \infty} L_{X_n} (X^n, Y^n) \right] \geq \limsup_{n \to \infty} EL_{X_n} (X^n, Y^n) \geq D (F_X, C) \tag{95}
\]
The combination of (94) and (95) completes the proof of (82) \( \blacksquare \)

VII. COMPARISON TO THE DENOISER IN [3]

In this section, we compare the proposed denoiser to the DUDE-like scheme in [3] for the finite input and continuous-real valued output alphabet case. We will proceed to show that, by a minor modification to the proposed denoiser, we essentially achieve the denoiser in [3]. This is illustrated by comparing the first pass of the DUDE-like denoiser with a modified version of the proposed scheme through the schematic representation in Fig. 3. Each output alphabet is uniformly quantized to the same number of levels, \( M \), as the input (for \( Y \in \mathbb{R} \), the end-intervals are greater than quantization step size). We label the set of quantization intervals at the output as \( \mathcal{O} = \{O_1, \ldots, O_M\} \) and let the quantization step size be \( \alpha \). Corresponding to the channel output, \( Y^n \), let \( Z^n \) be the corresponding quantized version. Also, let \( \mathcal{A} \) denote the \( M \)-level finite alphabet set at the input.

As a result of the quantization, we propose mapping the \( k \)-th order kernel density estimate at the output, \( f_{Y}^{n,k} \), to the corresponding probability mass function, \( \hat{Q}_{z^n}^{k} \), with mass at the quantized output alphabets in the following manner,
\[
\hat{Q}_{z^n}^{k} [y^n] (v^n_{-k}) = \int_{v^n_{-k} \in \mathcal{O}_{2k+1}} f_{Y}^{n,k} (y^n_{-k}) dy^n_{-k} \tag{96}
\]
where, \( v^n_{-k} \) is the corresponding \( 2k + 1 \)-tuple of the quantized levels. The channel conditional densities also get correspondingly mapped to an \( M \times M \) channel matrix that is formed using,
\[
\Pi(i, j) = \int_{y:Q_\alpha (y) = j} f_{Y|x=i} (y) dy \tag{97}
\]
where \( Q_\alpha (\cdot) \) denotes a uniform quantizer with a quantization step size \( \alpha \).

We compare \( \hat{Q}_{z^n}^{k} [y^n] (v^n_{-k}) \) to \( \hat{P}_{z^n}^{k} (v^n_{-k}) \), the \( k \)-th order distribution of the quantized output symbols, using the notation in [3].
\[
\hat{P}_{z^n}^{k} (v^n_{-k}) = \frac{r [z^n, v^n_{-k}]}{n - 2k} \tag{98}
\]
The density estimate, $f_{Y}^{n,k}$, we consider is the cubic histogram estimate. The histogram estimate is defined by

$$f_{Y}^{n,k}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \in A_{nj}} \lambda(A_{nj}), \quad y \in A_{nj}, y \in \mathbb{R}^{2k+1}$$ (99)

where, $\mathcal{P}_n = \{A_{nj}, j = 1, 2, \cdots\}, n \geq 1$ is a sequence of partitions and $A_{nj}$’s are Borel sets with finite nonzero Lebesgue measure. The sequence of partitions is rich enough such that the class of Borel sets ($\mathcal{B}^{[a,b]}$) is equal to

$$\bigcap_{n=1}^{\infty} \sigma \left( \bigcup_{m=n}^{\infty} \mathcal{P}_m \right)$$ (100)

where $\sigma$ is the usual notation of the $\sigma$-algebra generated by a class of sets. In particular, the cubic histogram estimate is constructed when we consider sets $A_{nj}$ of the form, $\prod_{i=1}^{2k+1} [a_i k_i h, a_i (k_i + 1)h)$, $k_i$’s are integers, $h$ is a smoothing factor as for the kernel density estimate in (99) and $a_i$’s are positive constants s.t. $a_i k_i h \in [a, b]$, $\forall h, k_i$. The following result similar to that in Theorem 1, for $J_n$ defined in equation (33), is true for histogram density estimates.

**Theorem 9:** Assume that the sequence of partitions $\mathcal{P}_n$ satisfies (100). Consider

1) $J_n \to 0$ in probability as $n \to \infty$, for all sequences $x^n$
2) $J_n \to 0$ almost surely as $n \to \infty$, for all sequences $x^n$
3) $J_n \to 0$ exponentially as $n \to \infty$, for all sequences $x^n$
4) For all $A \in \mathcal{B}$ with $0 < \lambda(A) < \infty$, and all $\varepsilon > 0$ there exists $n_0$ such that for all $n \geq n_0$, we can find
\[ A_n \in \sigma(P_n) \text{ with } \lambda(A \Delta A_n) < \varepsilon \text{ and} \]

\[
\sup_{M > 0 \text{ all sets } C \text{ of finite Lebesgue measure}} \lim_{n \to \infty} \lambda \left( \bigcup_{j, \lambda(A_n \cap C) \leq \frac{\varepsilon}{M}} A_{nj} \cap C \right) = 0 \tag{101}
\]

It is then true that \(4 \Rightarrow 3 \Rightarrow 2 \Rightarrow 1\).

For the proof of this theorem, please refer to [5] with the added condition of tightness imposed on the family of measures associated with the channel, \(C\).

The condition 4) in Theorem 9 translates to \(\lim_{n \to \infty} h = 0, \lim_{n \to \infty} nh^d = \infty\). It can be shown as in [5] that they are necessary sufficient conditions for that specified in 4) in Theorem 9. By choosing the smoothing factor, \(h\) to be a decreasing sequence of numbers that are all integers fractions of the quantization step size \(\alpha\), such that \(nh^d \to \infty\) is also simultaneously satisfied, we get the mapping in equation (96) to reduce that to equation (98) for the subsequences described in section V. This is because we split the sequence \(x^n\) into \(2k + 1\) subsequences whose \(2k + 1\)-length super symbols are independent so that we can apply Theorem 9. Now,

\[
\hat{Q}_{z_k}^k(v_{-k}) = \int_{y_{-k}^k \in O^{2k+1}} f_{y, -k}^k(y_{-k}) dy_{-k} \tag{102}
\]

\[
= \int_{y_{-k}^k \in O^{2k+1}} \frac{1}{\lambda(A_{n+1})} \sum_{j=0}^{\lfloor \frac{n-2k-1}{2k+1} \rfloor} 1_{\{y_{j(2k+1)}^k \in A_{n+1}\}} \tag{103}
\]

\[
= \frac{1}{\lfloor \frac{n-2k-1}{2k+1} \rfloor} r \left[ z^{n_k}, v_{-k}^k \right] \tag{104}
\]

If we mapped the finite input-continuous output channel, \(C\), to \(\Pi\), the mapping in equation (66) would then reduce to,

\[
\hat{Q}_{z_k}^k = \arg \min_{P \in \mathcal{F}^{A,k}} \sum_{v_{-k}^k} \hat{Q}_{z_k}^k(v_{-k}) - \sum_{u_{-k}^k \in A^{2k+1}} \Pi(u_j, v_j) P(u_{-k}) \tag{105}
\]

where, \(\mathcal{F}^{A,k}\) denote the space of all possible \(k\)-th-order distributions on \(A\). If we lift the constraints of the minimizer being a bona fide element of \(\mathcal{F}^{A,k}\), we get the following candidate for the minimizer in (105)

\[
\hat{Q}_{z_k}^k[u_{-k}^k] = \frac{1}{\lfloor \frac{n-2k-1}{2k+1} \rfloor} \sum_{u_{-k}^k} r \left[ z^{n_k}, v_{-k}^k \right] \prod_{j=-k}^{k} \Pi^{-1}(v_j, u_j) \tag{106}
\]

which is exactly the same as \(\hat{P}_{z_k}^k[z^{n_k}] (u_{-k}^k)\) using equation (18) in [3], also given below.

\[
\hat{P}_{z_k}^k[u_{-k}^k] = \frac{1}{\lfloor \frac{n-2k-1}{2k+1} \rfloor} \sum_{u_{-k}^k} r \left[ z^{n_k}, v_{-k}^k \right] \prod_{j=-k}^{k} \Pi^{-1}(v_j, u_j) \tag{107}
\]

Now, using the construction of the discrete denoiser in equation (68), for \(\hat{Q}_{x_k}\), we get

\[
g_{\text{opt}}[\hat{Q}_{x_k}] (y_{-k}^k) = \arg \min_{\hat{\beta} \in \mathcal{A}} \Lambda(\hat{\beta}, \hat{\beta})^T [\hat{Q}_{x_k} \otimes C]_{U|y_{-k}^k} \]

\[
= \arg \min_{\hat{\beta} \in \mathcal{A}} \sum_{\hat{a} \in \mathcal{A}} \Lambda(\hat{a}, \hat{\beta}) \cdot \left\{ \sum_{u_{-k}^k \in A^{2k+1}: u_{0_0}=a} \prod_{j=-k}^{k} f_{y|z=x}(y_j) \hat{Q}_{x_k} (U_{-k}^k = u_{-k}^k) \right\} \tag{108}
\]
which is exactly the same as $g_{\text{opt}}[P] (y_k^k)$ in equation (16) in [3]. Hence, the proposed denoiser with histogram density estimate of the output symbols and quantization gives us the same candidate as that of [3] applied to the $2k + 1$ subsequences of the output sequence $Y^n$.

VIII. Experimental Results

In this section, we discuss experimental results of applying the proposed scheme to denoising 256-level gray scale images. We demonstrate efficacy of the scheme with results of its application to cases of additive and multiplicative Gaussian noise cases. In addition, we also consider a highly nonlinear, non-conventional noise distribution, a locally varying Rayleigh distributed noise whose variance is a function of the gray level of the underlying clean image. The first pass of the denoiser which involves collection of the statistics of the underlying clean sequence is performed using a Fast Kernel Density Estimation approach proposed in [11] and a channel inversion procedure. This channel inversion is performed using a convex optimization linear programming technique that maps the output $k$th-order density estimate to the corresponding input $k$th-order input empirical distribution in accordance with (66). The experiment results presented in this section have been done so with little or no heuristical modification to the theoretical discussions earlier in this paper. The practical implementation aspects of the scheme are discussed in greater details and depth in [22], [23].

The first example we consider is, denoising of the boats image that is corrupted by an additive white noise channel (AWGN) with, $\sigma = 20$. The loss function, $\Lambda$, to be minimized in this case is the squared error between the true clean image and our denoised estimate. The denoiser in this case is a mapping from $\mathbb{R} \to A = \{0, \ldots, 255\}$ and reduces to that in (68). Results of the proposed denoising scheme are shown in the Fig. 5 below with context length, $k$, ranging from 1 to 6. The context (for $k > 1$) around any location, $i$, in the block of noisy data are 2D neighborhoods. The 2D contexts for various values of $k$ are shown in Fig. 4 below. As is evident from both, the reported Root Mean Squared Error (RMSE) figures and the perceptual quality, we are able to achieve improved denoising performance with increasing context lengths. Finally, we compare the results of the proposed scheme to that achieved by wavelet-based thresholding scheme [6] and Bayesian Least Squares Gaussian Scaled Mixture (BLS-GSM) denoiser in [20]. Increasing context lengths, $k$, translates to accruing increasing $k$th-order statistics from the finite block length data. This is the classic trade-off between increasing context lengths and reliability of the associated higher order statistics is seen in Fig. 6 where we see only marginal gains in the RMSE between, $k = 4$ and $k = 6$. The results for the AWGN case are primarily aimed at demonstrating the practicality of the proposed scheme fully acknowledging the performance lead of schemes like the BLS-GSM that are particularly catered to the problem of denoising in the case of AWGN channels. The benefits of the proposed approach are in fact highlighted in unconventional cases like nonlinear noise channels which will be discussed next.

Another example of the application of the proposed scheme is in denoising an image corrupted with an unconventional distribution as discussed earlier in this section. More specifically, we simulate the noisy image by using a gray-level dependent Rayleigh distribution (with probability density function, $f(x) = \frac{x}{2B^2} e^{-\frac{x^2}{2B^2}}$) whose variance parameter, $B$, is chosen as a function of clean image’s gray level at that location. In this particular example, we
generate a matrix of 256x256 Rayleigh distributed random variables whose parameters $B$ are chosen according to the following rule, $B(i, j) = I(i, j) + 35/256$, where $I(i, j)$ is the true value of the clean image at location $(i, j)$. We will discuss the denoising performance only in the symbol-by-symbol case in this setting in favor of succinctness to convey the point of efficacy of the proposed scheme. More detailed results and discussions on this problem setting can be found in [22]. We compare, in Fig. 7, the empirical distribution estimate, $\hat{F}_{x^n}$, of the underlying clean image with the histogram generated from access to the “true” clean image. We also compare these results to the smoothed histogram estimate of the true clean image that was produced using the Kernel Density estimation approach in [11]. From a visual inspection of the figure, it is evident that we are able to reasonably recover the true marginal empirical distribution of the underlying clean image and correspondingly the estimate of the true image.

Finally, we present the results of denoising the boats image that is corrupted by a multiplicative Gaussian noise with a distribution, $\mathcal{N}(1.0.2)$ in Fig. 8. The noise in this case literally multiplies this case literally multiplies the original clean image to corrupt it and as such, the effects are relatively more catastrophic. We compare, qualitatively, the results from the proposed denoiser with that of [20] to validate its efficacy.

IX. CONCLUSION AND FUTURE DIRECTIONS

We have presented a scheme for denoising continuous amplitude signals that is asymptotically optimal and universal. A salient feature of the proposed scheme is that the asymptotic universal optimality is guaranteed under a large class of reasonably well-behaved, user specified, loss functions. The technique presented in this paper draws from the “DUDE framework” in [26]. A weighted ‘context aggregation’ was suggested in [26] as an approach to enhance the performance of the DUDE in the first pass of the statistics collection. The proposed technique provides a natural context aggregation mechanism whereby neighboring contexts in addition to the observed are weighted by the kernel in the density estimation step. The denoiser proposed in [3] was shown to be asymptotically universal.
and extended the domain of applicability of DUDE-like schemes to cases where the noise is continuous valued. This approach, even though elegant theoretically, suffers from some of the same issues as the DUDE in terms of sparseness of statistics for large alphabet sizes. Our technique addresses this problem for the problem setting considered in [3] by natural context aggregation induced by the kernel density estimation. In the setting where the underlying clean signal is discrete-valued, taking values in a finite alphabet space, a slight modification of our scheme has been shown to reduce to the scheme in [3]. We also simultaneously provide a framework to address the case of continuous valued alphabets, where there is need to learn distribution functions instead of individual mass points as in the discrete-valued case. Finally, the proposed scheme is practical and tractable in its computational requirements as demonstrated by the experimental results.

The experimental results in this paper seem promising enough to motivate further exploration of practical aspects of the proposed scheme. This is an interesting future direction that is currently under investigation. Additional directions of research include studying the applicability of recursive density estimation techniques discussed in [13] in designing recursive denoisers as an alternative to the scheme presented in this paper. This would be particularly useful in multidimensional data applications like denoising noise corrupted video. It could also be of theoretical interest to understand the implications of a recursive structure to the denoiser and its associated optimality results.

**Appendix I**

**Proof of Lemma 1**

A theorem necessary for the proof of Lemma 1 is as follows

**Theorem 10:** Every Kernel $K$ with $\int K = 1, K \geq 0$ is an approximate identity, i.e for $\lim_{n \to \infty} h_n = 0$ and every $f_i \in L_1$, s.t. $D^2 \left( f_i \right) < \infty$ are uniformly bounded we have

$$\lim_{n \to \infty} \int \left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) \right| = 0$$

An alternate formulation of the approximation identity is the following,

**Theorem 11:** Every Kernel $K$ with $\int K = 1, K \geq 0$ is an approximate identity, i.e for $\lim_{n \to \infty} h_n = 0$ and every $f_i \in L_1$, s.t. $\lim_{|t| \to 0} \Omega_C(t) = 0$

$$\lim_{n \to \infty} \int \left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) \right| = 0$$

A definition regarding the notion of an associated kernel, $L$, with the kernel, $K$ that is necessary for the subsequent proof is,

**Definition 5:** The function $L$ defined by

$$L(x) = (-1)^s \int_{x}^{\infty} \frac{(y - x)^{s-1}}{(s-1)!} K(y) dy \quad (x > 0)$$

$$L(-x) = (-1)^s L(x) \quad (x < 0)$$
is the kernel associated with kernel $K$. The function $L$ is sometimes said to have a parameter $s$ since it figures in the definition of $L$. When $K$ is symmetric, $L$ is symmetric.

Furthermore,

$$
\int |L| \leq \frac{1}{s!} \int |x|^s |K(x)| dx
$$

for all nonnegative integers $s$. For $s = 0$, we define $L = K$. For $K \geq 0$, we have the equality

$$
\int |L| = \frac{1}{s!} \int |x|^s |K(x)| dx
$$

Finally,

$$
\int L = \int \frac{x^s}{s!} K(x) dx
= \left\{ \begin{array}{ll}
0 & : s \text{ odd} \\
0 & : s \text{ even, and the order of } K \text{ is } > s
\end{array} \right.
$$

**Proof:** [Proof of Theorem 10]

Let us start with the case that $f_i$ has $s-1$ absolutely continuous derivatives. Then, by Taylor’s series expansion with remainder,

$$
f_i(x + y) - f_i(x) = \sum_{j=1}^{s-1} \frac{y^j}{j!} f_i^{(j)}(x) + \int_x^{x+y} \frac{(x+y-u)^{s-1}}{(s-1)!} f_i^{(s)}(u) du
$$

so that, for class $s$ kernels $K$,

$$
\left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) = \frac{1}{n} \int \left( \sum_{i=1}^{n} f_i(x + y) - \sum_{i=1}^{n} f_i(x) \right) K_{h_n}(y) dy
$$

(recall that $\int K = 1$)

$$
= \frac{1}{n} \sum_{i=1}^{n} \left[ \sum_{j=1}^{s-1} \int_x^{x+y} \frac{(x+y-u)^{s-1}}{(s-1)!} f_i^{(s)}(u) du \right] K_{h_n}(y) dy
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \left[ \int_{-\infty}^{x} f_i^{(s)}(u) \int_{u-x}^{\infty} \frac{(x+y-u)^{s-1}}{(s-1)!} K_{h_n}(y) dy du - \int_{-\infty}^{x} f_i^{(s)}(u) \int_{-\infty}^{u-x} \frac{(x+y-u)^{s-1}}{(s-1)!} K_{h_n}(y) dy du \right]
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \left[ \int_{-\infty}^{x} f_i^{(s)}(u) \int_{-\infty}^{u-x} \frac{(-1)^s (L)_{h_n}(u-x) du}{(s-1)!} \right]
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \left[ \int_{-\infty}^{x} f_i^{(s)}(u) (L)_{h_n}(x-u) du \right]
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} h^s f_i^{(s)} * L_{h_n}
$$

(112)
where \((L)_{h_n}\) is the kernel associated with \(K_{h_n}\) and \(L\) is the kernel associated with \(K\). Therefore, by Young’s inequality,

\[
\int \left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) \right| |L| \leq \frac{h_n^s}{n} \sum_{i=1}^{n} \left| f_i^{(s)} \right| |L| \leq \frac{h_n^s}{n} \left( \sum_{i=1}^{n} \left| f_i^{(s)} \right| \right) |L| \tag{113}
\]

Since \(f_i\)’s have \((s - 1)\) absolutely continuous derivatives, \(\int |f_i^{(s)}| < \infty\), and further if \(\int |f_i^{(s)}| < M < \infty\), \(\forall i\) (uniformly bounded) the inequality in (113) simplifies to

\[
\int \left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) \right| \leq h_n^s M \int |L| \tag{114}
\]

Since,

\[
\int |L| \leq \frac{1}{s!} \int |x|^s |K(x)| dx = B_K < \infty \tag{115}
\]

for \(K\) being an order \(s\) kernel, inequality in equation (114) becomes

\[
\int \left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) \right| \leq h_n^s M B_K \tag{116}
\]

Taking limit \(n \to \infty\) on either sides, we get

\[
0 \leq \lim_{n \to \infty} \int \left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) \right| \leq \lim_{n \to \infty} h_n^s M B_K = 0 \tag{117}
\]

This can be extended to the general \(f_i\)’s using the universal derivative defined earlier. As a reminder,

\[
D^s_n(f_i) \triangleq \lim_{h \to 0} \inf \int \left| (f_i * \phi_h)^{(s)} \right| \tag{118}
\]

where, \(\phi\) is a mollifier.

Mollifiers are class 0 kernels, nonnegative and zero outside \([-1, 1]\). They also have infinitely many continuous derivatives and is called a mollifier because of its exceptional smoothing properties. An example of a mollifier is

\[
K(x) = Ce^{-\frac{1}{1-x^2}}, \quad |x| \leq 1 \tag{119}
\]

For a class \(s\) kernel, \(K\), and a family of density functions \(\{f_i\}_{i \in \mathbb{N}}\) with associated universal derivatives that are uniformly bounded, i.e., \(D^s_2(f_i) < B_C < \infty, \forall i \in \mathbb{N}\), it can then be shown that,

\[
\int \left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) \right| \leq \frac{1}{n} \sum_{i=1}^{n} \left| f_i * K_{h_n} - f_i \right| \leq \frac{1}{n} \sum_{i=1}^{n} h_n^s D^s_n(f_i) \int |L| \leq \frac{1}{n} \sum_{i=1}^{n} h_n^s B_C \int |L| = h_n^s B_C \int |L| \tag{120}
\]
Taking limits on both sides we get,
\[
\lim_{n \to \infty} \int \left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{\Omega} - \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) \right| = 0
\] (121)

**Proof:** [Proof of Theorem 11]

\[
f_i(x) = f_i(x) \int K_{h}(t)dt = \int f_i(x)K_{h}(t)dt, \quad \forall i
\] (122)

Therefore,
\[
\left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i * K_{h} \right)(x) - \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right| = \int \left| \left[ \frac{1}{n} \sum_{i=1}^{n} f_i(x - t) - \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right] K_{h}(t)\right| dt
\]
\[
\leq \int \left| \frac{1}{n} \sum_{i=1}^{n} f_i(x - t) - \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right| |K_{h}(t)|^{\frac{1}{p}} |K_{h}(t)|^{\frac{1}{p'}} dt
\] (123)

where \( \frac{1}{p} + \frac{1}{p'} = 1, \left( \frac{1}{p'} = 0 \right. \text{if } p \left. = 1 \right) \). Applying Holder’s inequality with exponents \( p \) and \( p' \), and then raising both sides to the \( p^{th} \) power and integrating with respect to \( x \), we obtain

\[
\int \left| \left( \frac{1}{n} \sum_{i=1}^{n} f_i * K_{h} \right)(x) - \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right|^p dx
\]
\[
\leq \int \left[ \int \left| \frac{1}{n} \sum_{i=1}^{n} f_i(x - t) - \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right|^p |K_{h}(t)| dt \right]^{\frac{1}{p}} |K_{h}(t)|^{\frac{1}{p'}} dx
\]
\[
= \|K\|_1^{\frac{p}{p'}} \int \left[ \int \left| \frac{1}{n} \sum_{i=1}^{n} f_i(x - t) - \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right|^p |K_{h}(t)| dt \right] dx
\]
\[
\leq \|K\|_1^{\frac{p}{p'}} \int \left[ \frac{1}{n} \sum_{i=1}^{n} \int |f_i(x - t) - f_i(x)|^p |K_{h}(t)| dt \right] dx
\] (124)

Changing the order of integration in the last expression (which is justified since the integrand is nonnegative), we obtain

\[
\left\| \left( \frac{1}{n} \sum_{i=1}^{n} f_i \right) * K_{h} - \frac{1}{n} \sum_{i=1}^{n} f_i \right\|_p \leq \|K\|_1^{\frac{p}{p'}} \int |K_{h}(t)| \frac{1}{n} \sum_{i=1}^{n} \omega_i(t) dt
\]
\[
\leq \|K\|_1^{\frac{p}{p'}} \int |K_{h}(t)| \Omega(t) dt
\] (125)

For \( \delta > 0 \),

\[
I_h = \int |K_{h}(t)| \Omega(t) dt = \int_{|t| < \delta} + \int_{|t| \geq \delta} = A_{h,\delta} + B_{h,\delta}
\] (126)

Since, we have \( \Omega(t) \to 0 \) as \( |t| \to 0 \), for \( \eta > 0 \), we can choose \( \delta \) so small that \( \Omega(t) < \eta \) if \( |t| < \delta \). Then

\[
A_{h,\delta} \leq \eta \int_{|t| < \delta} |K_{h}(t)| dt \leq \eta \|K\|_1, \quad \forall h > 0
\] (127)

Also, \( \Omega \) is a bounded function by Minkowski’s inequality [note that \( \|\Omega\|_{\infty} \leq \sup_{i \in \mathbb{N}} \|\omega_i\|_{\infty} \leq \sup_{i \in \mathbb{N}} (2\|f_i\|_p)^p \), which for \( p = 1 \), becomes \( \|\Omega\|_{\infty} \leq 2 \)], so that \( B_{h,\delta} \) is less than a constant multiple of \( \int_{|t| \geq \delta} |K_{h}(t)| dt \), which tends to zero with \( h \). This proves that \( I_h \to 0 \) as \( h \to 0 \) and the theorem follows. ■
Another lemma necessary for the proof of Lemma 1 is the following

**Lemma 10:** (A Multinomial distribution inequality)

Let \( N_1, \cdots, N_k \) be a multinomial random vector with parameters \( n, p_1, \cdots, p_k \). Then

\[
P \left( \sum_{i=1}^{k} \left| \frac{N_i}{n} - p_i \right| \geq \epsilon \right) \leq 2^{k+1} e^{-n\epsilon^2} \tag{128}
\]

**Proof**

By Scheffe’s theorem,

\[
\sum_{i=1}^{k} \left| \frac{N_i}{n} - p_i \right| = 2 \sup_{A} \left| \frac{N(A)}{n} - P(A) \right| \leq 2^{k} 2e^{-2n} \tag{129}
\]

where, \( A = \{ \text{all } 2^k \text{ possible sets of integers from } 1, \cdots, k \} \) and \( N(A) \) is the cardinality of \( A \). By Bonferroni’s inequality and Hoeffding’s inequality,

\[
P \left( \sup_{k} \left| \frac{N(A)}{n} - P(A) \right| \geq \frac{\epsilon}{2} \right) \leq 2^{k} 2e^{-2n} \tag{130}
\]

The expected value of \( f^n(x) \) is denoted by,

\[
g_h(x) = E(f^n(x)) = \frac{1}{nh^n} \sum_{i=1}^{n} \int K \left( \frac{x - y}{h} \right) f_i(y) dy \tag{131}
\]

**Proof:** [Proof of Lemma 1]

Let \( g_h \) be defined as in (131). By Theorem 1, it is enough to show that \( \int |f^n(x) - g_h(x)| dx \to 0 \) exponentially. Let \( \mu_n \) be the empirical probability measure for \( X_1, X_2, \cdots, X_n \) and note that

\[
f^n(x) = \frac{1}{h^n} \int K \left( \frac{x - y}{h} \right) \mu_n(dy) \tag{132}
\]

\[
\mu_n(dy) = \frac{1}{h^n} \sum_{i=1}^{n} 1_{A_i}(y) \tag{133}
\]

For given \( \epsilon > 0 \), find finite constants \( M, L, N, a_1, \cdots, a_N \) and disjoint finite rectangles \( A_1, \cdots, A_N \) in \( \mathbb{R}^d \) such that the function

\[
K^*(x) = \sum_{i=1}^{N} a_i 1_{A_i}(x) \tag{134}
\]

satisfies: \( |K^*| \leq M, K^* = 0 \text{ outside } [-L, L]^d \), and \( \int |K(x) - K^*(x)| dx < \epsilon \). Define \( g_h^* \) and \( f^n^* \) as \( g_h \) and \( f^n \) with \( K^* \) instead of \( K \). Then

\[
\int |f^n(x) - g_h(x)| dx \leq \int |f^n(x) - f^n^*(x)| dx + \int |f^n^*(x) - g_h^*(x)| dx + \int |g_h^*(x) - g_h(x)| dx
\]

\[
\leq \int \frac{1}{h^n} \int \left| K^* \left( \frac{x - y}{h} \right) K \left( \frac{x - y}{h} \right) \right| \mu_n(dy) dx + \frac{1}{nh^n} \sum_{i=1}^{n} \int \left| K^* \left( \frac{x - y}{h} \right) K \left( \frac{x - y}{h} \right) \right| f_i(y) dy dx
\]

\[
+ \int |f^n^*(x) - g_h^*(x)| dx
\]

\[
\leq 2\epsilon + \int |f^n^*(x) - g_h^*(x)| dx
\]
by a double change of integral. But, if $\mu$ is the probability measure for $f$,

$$
\int |f^n(x) - g^n(x)| dx \leq \sum_{i=1}^{N} |a_i| \int \left| \frac{1}{nh^d} \sum_{j=1}^{n} f_j(y) - \frac{1}{h^d} \int_{x-hA_i} \mu_n(dy) \right| dx
$$

$$
\leq \frac{1}{h^d} \sum_{i=1}^{N} |a_i| \int \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(x-hA_i) - \mu_n(x-hA_i) \right| dx
$$

(135)

Lemma 1 follows if we can show that for all finite rectangles $A$ of $\mathbb{R}^d$

$$
\frac{1}{h^d} \sum_{i=1}^{N} \int \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(x-hA_i) - \mu_n(x-hA_i) \right| dx \to 0 \text{ exponentially as } n \to \infty
$$

Choose an $A$, and let $\epsilon > 0$ be arbitrary. Consider the partition of $\mathbb{R}^d$ into sets $B$ that are $d$-fold products of intervals of the form $\left[ \frac{(i-1)h}{N}, \frac{ih}{N} \right)$, where $i$ is an integer, and $N$ is a new constant to be chosen later. Call the partition $\Pi$. Let

$$
A = \prod_{i=1}^{d} [x_i, x_i + a_i), \min_i a_i \geq \frac{2}{N}
$$

and

$$
A^* = \prod_{i=1}^{d} \left[ x_i + \frac{1}{N}, x_i + a_i - \frac{1}{N} \right)
$$

Define

$$
C_x = \left( x - hA - \bigcup_{B \subseteq x-hA} B \right) \subseteq x + h(A - A^*) = C_x^*
$$

Clearly, for any $n$

$$
\int \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(x-hA) - \mu_n(x-hA) \right| dx \leq \int \sum_{B \subseteq x-hA} \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \right| dx
$$

$$
+ \int \left( \frac{1}{n} \sum_{j=1}^{n} \mu_j + \mu_n \right) (C_x^*)
$$

(136)

The last term in (136) equals

$$
2\lambda(h(A - A^*)) = 2h^d \lambda(A - A^*)
$$

$$
= 2h^d \left( \prod_{i=1}^{d} a_i - \prod_{i=1}^{d} \left( a_i - \frac{2}{N} \right) \right)
$$

(137)
where \( \lambda \) is the Lebesgue measure. Now, putting (138), (136) and (135) together, we get

\[
\int |f^n(x) - g_h(x)| \, dx \leq 2\epsilon + \int |f^{n*}(x) - g^*_h(x)| \\
\leq 2\epsilon + \frac{1}{h^d} \sum_{i=1}^{N} |a_i| \int_{B \subseteq x - hA_i} \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \, dx + \sum_{i=1}^{N} |a_i| \frac{2}{h^d} h^d \lambda(A_i - A_i^*)
\]

\[
\leq 2\epsilon + \frac{1}{h^d} \sum_{i=1}^{N} |a_i| \sum_{B \subseteq \Pi} \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \int_{B \subseteq x - hA_i} \, dx + \sum_{i=1}^{N} |a_i| \frac{2}{h^d} h^d \lambda(A_i - A_i^*)
\]

\[
\leq 2\epsilon + \left( \sum_{i=1}^{N} |a_i| \lambda(A_i) \right) \sum_{B \subseteq \Pi} \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) + 2 \sum_{i=1}^{N} |a_i| \lambda(A_i - A_i^*)
\]

(139)

The third term on the right hand side can be made smaller than \( \epsilon \) by choosing \( N \) large enough \( (A_i^* - A_i, \forall i as \ N \to \infty) \). The coefficient of the first term on the right hand side is equal to \( \int |K^*| \leq 1 + \epsilon \). Thus, we have shown that for every \( \epsilon > 0 \), we can find \( N \) large enough such that

\[
\int |f^n(x) - g_h(x)| \, dx \leq 3\epsilon + (1 + \epsilon) \sum_{B \subseteq \Pi} \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \\
\leq 5\epsilon + \sum_{B \subseteq \Pi} \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B)
\]

(140)

We are almost in a position to use the Multinomial inequality were it not for the fact that the partition \( \Pi \) is infinite. Thus, it is necessary to "cut-off" the tails of the distribution. Consider a finite partition, \( \Pi_r \), consisting of sets of \( \Pi \) that has a non-empty intersection with \( [-r, r]^d \) where \( r > 0 \) is to be picked later. Let \( \Pi_r^* \) be \( \Pi_r \cup [-r, r]^d \). The cardinality of \( \Pi_r^* \) is at most

\[
\left( \frac{2rN}{h} + 2 \right)^d = O(n)
\]

To take care of the tails we argue as follows: let \( T \) stand for the tail set, i.e., the complement of \( [-r, r]^d \), then

\[
\sum_{B \subseteq \Pi} \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \right| \leq \sum_{B \subseteq \Pi_r} \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \right| + \frac{1}{n} \sum_{j=1}^{n} \mu_j(T) + \mu_n(T)
\]

\[
\leq \sum_{B \subseteq \Pi_r} \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \right| + 2\frac{1}{n} \sum_{j=1}^{n} \mu_j(T) + \frac{1}{n} \sum_{j=1}^{n} \mu_j(T) - \mu_n(T)
\]

\[
\leq \sum_{B \subseteq \Pi_r} \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \right| + 2 \sup_{i \in I} \mu_i(T)
\]

(141)
Now, \(2 \sup_{i \in I} \mu_i(T)\) can be made smaller than \(\epsilon\) by choice of \(r\). This gives,

\[
\int |f^n(x) - g_h(x)| \, dx \leq 6\epsilon + \sum_{B \pi_{r^*}} \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \right| \tag{142}
\]

where \(r\) depends on \(\epsilon, \Upsilon,\) and \(N\) depends on \(\epsilon, K\).

By Lemma 1, for \(\delta > 6\epsilon\) and \(\rho \in (0, 1)\),

\[
P \left( \int |f^n - g_h| > \delta \right) \leq P \left( \sum_{B \pi_{r^*}} \left| \frac{1}{n} \sum_{j=1}^{n} \mu_j(B) - \mu_n(B) \right| > \delta - 6\epsilon \right) \tag{143}
\]

\[
\leq e^{-(1-\rho) \frac{n^d}{\rho \delta^2}}, n \geq n_0(\rho, \delta, K, \Upsilon, h) \tag{144}
\]

This concludes that the proof \(5 \Rightarrow 4\) for nonnegative \(K\). Note that the inequality can be forced for all \(n, h\) with

\[
n > \frac{16 + 4^{d+1}}{\rho \delta^2} \tag{145}
\]

\[
nh^d > \frac{42^d (2rN)^d}{\rho \delta^2} \tag{146}
\]

if we pick

\[
\epsilon = \frac{\delta}{6} \left(1 - \sqrt{1 - \frac{\rho}{2}}\right)
\]

\[
\blacksquare
\]

APPENDIX II

PROOF OF THEOREM 2

Definition 6 (Prohorov metric): For any two laws \(P\) and \(Q\) on the set \([a, b] \subset \mathbb{R}\), the Prohorov metric, \(\rho\) is defined as

\[
\rho(P, Q) := \inf \{\epsilon > 0 : P(B^\epsilon) \leq P(B^\epsilon) + \epsilon, B \in \mathcal{B}^{[a, b]}\}
\]

where \(B^\epsilon = \{\tilde{x} : |x - \tilde{x}| < \epsilon, x \in B\}\).

Proof: [Proof of Theorem 2] Let \(P_n\) and \(Q_n\) denote the laws associated with the distribution functions, \(F_{x^n}\) and \(\tilde{F}_{x^n}\). From [8, Theorem 11.7.1], \(\rho(P_n, Q_n) \to 0 \Rightarrow \beta(P_n, Q_n)\) then by definition of the \(\beta\)-metric, we have

\[
\lim_{n \to \infty} \int f \, d(P_n - Q_n) = 0 \quad \forall \|f\|_{BL} \leq 1 \tag{147}
\]

By a mere scaling, the above statement is also true for a uniformly bounded Lipschitz class of functions, \(S_M^{[a, b]} = \{f : \|f\|_{BL} < M, f : [a, b] \to \mathbb{R}\}\) for some \(M < \infty\). It is also true that

\[
\lim_{n \to \infty} \int f(x, y) d(P_n - Q_n) = 0 \quad \forall y \text{ and } f \in S^{[a, b] \times \mathbb{R}} \tag{148}
\]
where $S_M^{[a,b] × R} := \{ f : [a,b] × R \to R, \| f(y) \|_{BL} < M \forall y \}$ for some $M < \infty$ and

\begin{align}
\| f(y) \|_L &:= \sup_{x \neq z} \frac{|f(x, y) - f(z, y)|}{|x - z|} \\
\| f(y) \|_\infty &:= \sup_x f(y, x) \\
\| f(y) \|_{BL} &:= \| f(y) \|_L + \| f(y) \|_\infty
\end{align}

(149) (150) (151)

Hence, for a channel with conditional densities, $\{ f_{Y|x} \}_{x \in [a,b]} \in S_M^{[a,b] × R}$, we have

\[ \left| \int f_{Y|x} dF_{x^n} - \int f_{Y|x} d\hat{F}_{x^n} \right| \to 0 \quad \forall y \in R \]

(152)

and by dominated convergence theorem,

\[ \int \left| \int f_{Y|x} dF_{x^n} - \int f_{Y|x} d\hat{F}_{x^n} \right| dy \to 0 \]

(153)

and hence, $d \left( [F_{x^n} \otimes C]_Y, [\hat{F}_{x^n} \otimes C]_Y \right) \to 0$.

Hence, the mapping of input empirical distributions to output densities induced by the channel,

\[ f_{Y^n}(y) = [F_{x^n} \otimes C]_Y = \int f_{Y|x} dF_{x^n}(x) \]

(154)

is continuous with respect to the $\beta$ metric on the input distributions and the total variation metric on the output densities. We also have the fact that $(\mathcal{F}^{[a,b]}, \beta)$ is a compact [8, Theorem 11.5.4, Corollary 11.5.5] metric space. Since, we have a continuous 1-1 (bijection) mapping between the compact metric space of input distributions with the $\beta$ metric, $(\mathcal{F}^{[a,b]}, \beta)$, and the space of output densities, with the total variation metric, $( [\mathcal{F}^{[a,b]} \otimes C], d)$, we can apply the continuous mapping theorem [21] to get continuity in the inverse mapping too. This gives the desired result that as $d([F_{x^n} \otimes C]_Y, [\hat{F}_{x^n} \otimes C]_Y) \to 0$, we have $\beta(P_n, Q_n) \to 0$ and $\rho(P_n, Q_n) \to 0$. Finally using the fact [8], $\lambda \leq \rho$, $\lambda \left( F_{x^n}, \hat{F}_{x^n} \right) \to 0$.

\section*{Appendix III}

\section*{Proof of Lemma 4}

\textit{Proof:}

Consider $f \in C_b([a,b])$, where $C_b$ denotes the set of all continuous bounded functions, $f : [a,b] \to R$. For any
\( F \in \mathcal{F}^{[a,b]} \) and \( P^\Delta \) that is constructed using (39)

\[
\left| \int f dF(x) - \int f P^\Delta(dx) \right| \\
= \left| \int f \left( dF(x) - P^\Delta(dx) \right) \right| \\
= \left| \int dF(x) - \sum_{i=1}^{N} f(a_i) P(a_i) \right| \\
\leq \sum_{i=0}^{N-1} \int_{a_i}^{a_{i+1}} (f(a_i) + \omega_f(\Delta)) dF(x) - \sum_{i=1}^{N} f(a_i) P(a_i) \\
= \sum_{i=0}^{N-1} (f(a_i) + \omega_f(\Delta)) P(a_i) - \sum_{i=1}^{N} f(a_i) P(a_i) \\
= \omega_f(\Delta) \sum_{i=1}^{N} P(a_i) \\
= \omega_f(\Delta) \\
\tag{155}
\]

where \( \omega_f(\Delta) = \max_{y \in [a,b]} |f(y + \Delta) - f(y)| \) and \( N \) is the number of quantization levels as defined previously. Hence,

\[
\lim_{\Delta \to 0} |P^\Delta f - Pf| = \left| \lim_{\Delta \to 0} \int f \left( dF(x) - P^\Delta(dx) \right) \right| \\
= \lim_{\Delta \to 0} \omega_f(\Delta) \\
= 0, \quad \forall f \in C_b([a,b]) \\
\tag{156}
\]

This implies weak convergence of \( P^\Delta \Rightarrow P \). Hence, the statement of the theorem follows from the Prohorov metric that metrizes weak convergence.

\section*{Appendix IV}

\textbf{Proof of Lemmas 5 and 6}

We need the following proposition for the proof of Lemma 5

\textit{Proposition 1:} \( A(x) = \int \Lambda(x, g(y)) f_{Y|x}(y) dy \) is a bounded Lipschitz function for any measurable \( g : \mathbb{R} \to [a, b] \).

\textit{Proof:} Let \( \Delta = |x - x'| \),

\[
A(x) - A(x') = \int \Lambda(x, g(y)) f_{Y|x}(y) dy - \int \Lambda(x', g(y)) f_{Y|x'}(y) dy \\
\leq \int (\Lambda(x', g(y)) + \Lambda(\Delta, x)) f_{Y|x}(y) dy - \int (\Lambda(x', g(y)) f_{Y|x'}(y) dy \\
\leq \int (\Lambda(x', g(y)) + \Lambda(\Delta, x)) (f_{Y|x'}(y) + \varepsilon(\Delta(y)) dy - \int (\Lambda(x', g(y)) f_{Y|x'}(y) dy \\
\leq \Lambda(\Delta, x) + \Lambda_{\text{max}} \delta + \Lambda(\Delta, x) \delta_{\Delta}
\]

\[
\boxed{
\text{APPENDIX IV}
\text{PROOF OF LEMMAS 5 AND 6}
\}
\]

We need the following proposition for the proof of Lemma 5

\textit{Proposition 1:} \( A(x) = \int \Lambda(x, g(y)) f_{Y|x}(y) dy \) is a bounded Lipschitz function for any measurable \( g : \mathbb{R} \to [a, b] \).

\textit{Proof:} Let \( \Delta = |x - x'| \),

\[
A(x) - A(x') = \int \Lambda(x, g(y)) f_{Y|x}(y) dy - \int \Lambda(x', g(y)) f_{Y|x'}(y) dy \\
\leq \int (\Lambda(x', g(y)) + \Lambda(\Delta, x)) f_{Y|x}(y) dy - \int (\Lambda(x', g(y)) f_{Y|x'}(y) dy \\
\leq \int (\Lambda(x', g(y)) + \Lambda(\Delta, x)) (f_{Y|x'}(y) + \varepsilon(\Delta(y)) dy - \int (\Lambda(x', g(y)) f_{Y|x'}(y) dy \\
\leq \Lambda(\Delta, x) + \Lambda_{\text{max}} \delta + \Lambda(\Delta, x) \delta_{\Delta}
\]

\[
\boxed{
\text{APPENDIX IV}
\text{PROOF OF LEMMAS 5 AND 6}
\}
\]
Also,
\[
A(x) - A(x') = \int \Lambda(x, g(y)) f_{Y|x}(y) dy - \int \Lambda(x', g(y)) f_{Y|x'}(y) dy
\geq \int \left( \Lambda(x', g(y)) - \lambda(\Delta, x) \right) f_{Y|x}(y) dy - \int \left( \Lambda(x', g(y)) \right) f_{Y|x'}(y) dy
\geq \int \left( \Lambda(x', g(y)) - \lambda(\Delta, x) \right) \left( f_{Y|x'}(y) - \varepsilon(\Delta) dy - \int \left( \Lambda(x', g(y)) \right) f_{Y|x'}(y) dy
\geq -\lambda(\Delta, x) - \lambda_{\max} \Delta + \lambda(\Delta, x) \delta_{\Delta}
\geq -\lambda(\Delta, x) - \lambda_{\max} \delta_{\Delta} + \lambda(\Delta, x) \delta_{\Delta}
\]
Hence, \( |A(x) - A(x')| \leq \lambda(\Delta) + \lambda_{\max} \delta_{\Delta} + \lambda(\Delta) \).

The assumption of Lipschitz continuity (condition, C6) of the channel guarantees \( \lim_{\Delta \to 0} \delta_{\Delta} = 0 \). With this and the fact that \( \lim_{\Delta \to 0} \lambda(\Delta) = 0 \), we have \( \lim_{\Delta \to 0} |A(x) - A(x')| = 0 \)

Moreover,
\[
\| A \|_L = \sup_{0 < \Delta < (b-a)} \| \Lambda \|_L + \lambda_{\max} \| \delta \|_L + (b-a) \| \Lambda \|_L \| \delta \|_L \leq 0 \quad \Delta \to 0 \tag{159}
\]

Hence,
\[
\| A \|_{BL} = \| A \|_L + \| A \|_{\infty} \leq \| A \|_L + \lambda_{\max} \| \delta \|_L + (b-a) \| \Lambda \|_L \| \delta \|_L + \lambda_{\max} \tag{160}
\]

**Proof:** [Proof of Lemma 5]

\[
\| F_{\otimes C} \Lambda(U_0, g(Y)) - E_{\tilde{F}_{\otimes C}} \Lambda(U_0, g(Y)) \|
\leq \| F_{\otimes C} \Lambda(U_0, g(Y)) \|
\leq \| A \|_{BL} \beta \left( P, \hat{P} \right)
\leq \left( \| A \|_L + \lambda_{\max} \| \delta \|_L + (b-a) \| \Lambda \|_L \| \delta \|_L + \lambda_{\max} \right) \beta \left( P, \hat{P} \right)
\leq \left( \| A \|_L + \lambda_{\max} \| \delta \|_L + (b-a) \| \Lambda \|_L \| \delta \|_L + \lambda_{\max} \right) \beta \left( P, \hat{P} \right)
\leq 0 \quad \beta \left( P, \hat{P} \right) \to 0 \quad \text{we have} \quad \| E_{\tilde{F}_{\otimes C}} \Lambda(U_0, g(Y)) - E_{\tilde{F}_{\otimes C}} \Lambda(U_0, g(Y)) \| \to 0.
\]
Proof: [Proof of Lemma 6]

\[ |E_{P^\Delta \otimes C \Lambda}(U_0, g(Y)) - E_{F^\otimes C \Lambda}(U_0, g(Y))| = \]

\[ \sum_{i=1}^{N(\Delta)} \int_{a_{i-1}}^{a_i} dF(u') \left( \int \Lambda(u', g(y)) f_{Y|X=u'}(y)dy \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y)dy \right) \]

\[ = \sum_{i=1}^{N(\Delta)} \int dy \left( \int_{a_{i-1}}^{a_i} f_{Y|X=u'}(y)dF(u') \Lambda(u', g(y)) \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y)dy \right) \]

Equality in (162) is due to application of Fubini’s theorem. Hence,

\[ |E_{P^\Delta \otimes C \Lambda}(U_0, g(Y)) - E_{F^\otimes C \Lambda}(U_0, g(Y))| \leq \sum_{i=1}^{N(\Delta)} \int dy \left( \int_{a_{i-1}}^{a_i} f_{Y|X=u'}(y)dF(u') (\Lambda(a_i, g(y)) + \lambda(\Delta)) \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y)dy \right) \]

\[ = \sum_{i=1}^{N(\Delta)} \int dy (\Lambda(a_i, g(y)) + \lambda(\Delta)) \left( \int_{a_{i-1}}^{a_i} f_{Y|X=u'}(y)dF(u') \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y)dy \right) \]

\[ < \sum_{i=1}^{N(\Delta)} \int dy (\Lambda(a_i, g(y)) + \lambda(\Delta)) \left( f_{Y|X=a_i}(y) + \varepsilon(y) \right) \left( \int_{a_{i-1}}^{a_i} dF(u') \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y)dy \right) \]

\[ < \sum_{i=1}^{N(\Delta)} \left( \int_{a_{i-1}}^{a_i} dF(u') \right) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y)dy \right) \]

\[ + \int \varepsilon(y)\Lambda(a_i, g(y)) dy + \lambda(\Delta) \int f_{Y|X=a_i}(y)dy + \lambda(\Delta) \int \varepsilon(y)dy - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y)dy \right) \]
\[
\left| \sum_{i=1}^{N(\Delta)} \left( \int_{a_{i-1}}^{a_i} dF(u') \right) \left[ \int \Lambda (a_i, g(y)) f_{Y|X=a_i}(y) dy \right. \\
\left. + \int \varepsilon(y) \Lambda (a_i, g(y)) dy + \lambda(\Delta) \int f_{Y|X=a_i}(y) dy + \lambda(\Delta) \int \varepsilon(y) dy - \sum_{i=1}^{N(\Delta)} P^A (a_i) \left( \int \Lambda (a_i, g(y)) f_{Y|X=a_i}(y) dy \right) \right] \right| \\
= \left| \sum_{i=1}^{N(\Delta)} \left( \int_{a_{i-1}}^{a_i} dF(u') \right) \left[ \int \varepsilon(y) \Lambda (a_i, g(y)) dy + \lambda(\Delta) + \lambda(\Delta) \delta_\Delta \right] \right| \\
= \left| \sum_{i=1}^{N(\Delta)} (F(a_i) - F(a_{i-1})) \left[ \int \varepsilon(y) \Lambda (a_i, g(y)) dy + \lambda(\Delta) + \lambda(\Delta) \delta_\Delta \right] \right| \\
\leq \int \sum_{i=1}^{N(\Delta)} \varepsilon(y) \Lambda (a_i, g(y)) P^A (a_i) dy + (\lambda(\Delta) + \lambda(\Delta) \delta_\Delta) \\
\leq \delta_\Delta \Lambda_{\text{max}} + (\lambda(\Delta) + \lambda(\Delta) \delta_\Delta) \\
= \delta_\Delta \Lambda_{\text{max}} + \lambda(\Delta) (1 + \delta_\Delta)
\]

Hence,

\[
\lim_{\Delta \to 0} \left| E_{P^A \otimes C} \Lambda(U_0, g(Y)) - E_{F^\otimes C} \Lambda(U_0, g(Y)) \right| = 0
\]
APPENDIX V
PROOF OF LEMMA 8 AND K\textsuperscript{th}-ORDER EXTENSION OF THEOREM 4

Proof: [Proof of Lemma 8]

\[ D_k (x^n) = \min_g \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} \Lambda \left( X_0, g \left( Y^k_i \right) \right) \]

\[ = \min_g \int \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} \Lambda \left( x, g \left( y_{i-k}^i \right) \right) \prod_{l=i-k}^{i+k} f_{Y|X=x_l} (y_l) dy_l \]

\[ = \min_g \frac{1}{n - 2k} \sum_{i=k+1}^{2k+1} \int \frac{1}{n - 2k} \sum_{j=0}^{\left[ \frac{n-2k-i-1}{2k+1} \right]-1} \Lambda \left( x, g \left( y_{j(2k+1)+i}^j \right) \right) \prod_{l=j(2k+1)+i}^{j(2k+1)+i+2k} f_{Y|X=x_l} (y_l) dy_l \]

\[ \geq \frac{1}{2k+1} \sum_{i=1}^{2k+1} \min_g \int \frac{1}{n - 2k} \sum_{j=0}^{\left[ \frac{n-2k-i-1}{2k+1} \right]-1} \Lambda \left( x, g_i \left( y_{j(2k+1)+i}^j \right) \right) \prod_{l=j(2k+1)+i}^{j(2k+1)+i+2k} f_{Y|X=x_l} (y_l) dy_l \]

\[ \geq \frac{1}{2k+1} \sum_{i=1}^{2k+1} D_k (x^n) \]

Proposition 1, Lemmas 5 and 6 are extendible to their k\textsuperscript{th}-order equivalents with the proofs carrying over directly from the symbol-by-symbol case. We hence merely state the Lemmas for the k\textsuperscript{th}-order case and proofs are left out in this discussion.

Proposition 2: \( A(x) = \int \Lambda \left( x, g \left( y^k_i \right) \right) \prod_{i=k}^{k} f_{Y|X,x_i} (y_i) dy^k_i \) is a bounded Lipschitz function for any measurable \( g : [a, b]^{2k+1} \rightarrow \mathbb{R} \).

Lemma 11: For any \( F, \bar{F} \in \mathcal{F}^{[a, b], k} \), measurable \( g : \mathbb{R}^{2k+1} \rightarrow [a, b] \) and a bounded Lipschitz loss function with
\[ E_{f_Y|u}(u, g(Y_k^k)) < \infty, \forall u, \]
\[ |E_{F \otimes C}(U_0, g(Y_k^k)) - E_{F \otimes C}(U_0, g(Y_k^k))| \]
\[ \leq \left( \| \Lambda \|_L + \Lambda_{\text{max}} \| \delta \|_L^k + (b - a) \| \Lambda \|_L \| \delta \|_L^k + \Lambda_{\text{max}} \right) \beta \left( P, \hat{P} \right) \]

where \( P \) and \( \hat{P} \) are the laws associated with \( F \) and \( \hat{F} \) and \( \beta \) is the usual \( \beta \)-metric

\[ \| \delta \|_L^k \text{ is the } k^{th} \text{ order Lipschitz norm of the channel.} \]
\[ \| \delta \|_L^k \equiv \sup_{0 < \Delta < (b-a)} \frac{\delta^{2k+1}}{\Delta} \quad (172) \]

and \( \delta_{\Delta} \) is as defined in (6).

**Lemma 12:** For any \( \Delta > 0, F \in \mathcal{F}^{[a,b],k} \) with the associated measure \( P, P^{\Delta,k} \in \mathcal{F}^{\Delta,k} \), measurable \( g : \mathbb{R}^{2k+1} \rightarrow [a, b] \) and a continuous bounded loss function with \( E_{f_Y|u}(u, g(Y_k^k)) < \infty, \forall u, \)

\[ |E_{P^{\Delta,k} \otimes C}(U_0, g(Y_k^k)) - E_{F \otimes C}(U_0, g(Y_k^k))| \leq \delta_{\Delta}^{2k+1} \Lambda_{\text{max}} + \lambda(\Delta) \left( 1 + \delta_{\Delta}^{2k+1} \right) \]

These are then used to bound the deviation of the cumulative loss incurred by the proposed denoiser for each of the \( 2k+1 \) subsequences from the minimum possible \( k^{th} \)-order sliding window loss for that subsequence. We now, state the \( k^{th} \)-order equivalent of Theorem 4 for each subsequence.

**Theorem 12:** For all \( m \geq 1, k \geq 1, \epsilon > 0, \rho \in (0, 1), \delta > 0, \Delta > 0 \), and \( x^m \in [a,b]^{(2k+1)m} \)

\[ Pr\left( |L_{x^m, \delta, \Delta}(x^m, Y^m) - D_0(x^m)| > 3\epsilon + 5\delta \Lambda_{\text{max}} + 4\delta_{\Delta}^{2k+1} \Lambda_{\text{max}} + 4\lambda(\Delta)(1 + \delta_{\Delta}^{2k+1}) \right) \leq |G_{\delta,\Delta}^k| \left[ e^{-G(\epsilon+\delta \Lambda_{\text{max}} \Lambda_{\text{max}})m} + e^{(-1-\rho)\frac{m\gamma_k^2}{2}} \right] + e^{(-1-\rho)\frac{m\gamma_k^2}{2}} \quad (173) \]

for all \( m > m_0(C, k, h_n, \gamma_k) \)

where,
\[ \gamma_k = \frac{\epsilon}{\left( \| \Lambda \|_L + \Lambda_{\text{max}} \| \delta \|_L^k + (b - a) \| \Lambda \|_L \| \delta \|_L^k + \Lambda_{\text{max}} \right)} \]

and \( G, G_{\delta,\Delta}^k \) are as defined in Theorem 6.

**Proof:** The proof of this theorem carries over directly from the proof of Theorem 4 using Proposition 2, Lemmas 11, 12 and 7.
REFERENCES

Fig. 5. Row 1- left: Original image, right: Noisy image, $\sigma = 20$; Denoised Images using, Row 2- left: $k = 1$ right: $k = 2$; Row 3- left: $k = 4$, right: $k = 6$; Row 4- left: the scheme in [6], right: the scheme in [20]
Fig. 6. Comparison of RMSE of the denoised image for various context lengths, $k$

![Graph showing RMSE comparison](image)

RMSE = 11.2839

Fig. 7. Row 1- left: Original image, right: Noisy image; Denoised images using Row 2- left: symbol-symbol scheme, right: Comparison of Distribution estimates for the symbol-by-symbol denoiser

RMSE$_{\text{noise}} = 38.802$

RMSE = 11.2839
Fig. 8. Row 1- left: Original image, right: Noisy image; Denoised images using Row 2- left: proposed scheme, right: BLS-GSM [20]