

# Universal Denoising of Discrete-time Continuous-Amplitude Signals

Kamakshi Sivaramakrishnan <sup>†</sup> and Tsachy Weissman <sup>†,§</sup>, *Senior Member, IEEE*

## Abstract

We consider the problem of reconstructing a discrete-time signal (sequence) 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’ sequence, is universally optimal in the limit of large sequence length. This sequence of denoisers 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 unknown individual 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 practically implementable. 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, Memoryless Channels, semi-stochastic setting, discrete denoising.

## I. INTRODUCTION

Consider the problem of estimating a clean discrete-time signal (sequence)  $\{X_t\}_{t \in \mathbb{T}}$ ,  $X_t \in [a, b] \subset \mathbb{R}$ , based on its noisy observations  $\{Z_t\}_{t \in \mathbb{T}}$ ,  $Z_t \in \mathbb{R}$ , where  $\{Z_t\}$  is the output of a corruption mechanism, a memoryless channel. This problem finds applications in areas ranging from engineering, cryptography and statistics, 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. [9])

Work supported in part by National Science Foundation through grants CCR-0311633 and the NSF CAREER.

<sup>†</sup> The research was partially supported by NASA New Horizons program through grant No. 399840Q.

The material in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Seattle, WA, July 2006

<sup>†</sup> Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mails: {ksivaram, tsachy}@stanford.edu)

<sup>§</sup> Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, Israel.

and references therein). Solutions to this problem in [9] are based on wavelet-based soft thresholding and have various asymptotic optimality properties under a minimax criterion. The scope of wavelet-based thresholding in [9] has been extended beyond the additive white Gaussian case in [13], [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 [13], [21] 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. For the case of a random vector  $Y = X + Z$ , where  $X$  is independent of  $Z$  (with known distribution). The Minimum Mean Squared Estimate (MMSE) of  $X$  is well-known to be given by  $\hat{X} = \psi(Y) = E\{X|Y\}$ . It was shown in [27] that, for  $Z \sim \mathcal{N}(\mu, \Sigma)$ ,  $\psi(\cdot)$  satisfies  $\psi(Y) = \frac{(Y - \mu) - \nabla_y \ln f_Y(Y)}{f_Y}$ , where  $f_Y(y)$  is the marginal density of  $Y$ , which can be learned from the noisy samples  $Y^n = \{Y_1, \dots, Y_n\}$  of  $Y$ . Using techniques for nonparametric density estimation in [7], an estimate of  $f_Y(y)$ ,  $\hat{f}_Y(y)$ , can be computed, the (appropriate) gradient of which leads to the following estimate:

$$\hat{\psi}(Y) = \frac{(Y - \mu) - \nabla_y \ln \hat{f}_Y(Y)}{\hat{f}_Y} \quad (1)$$

The authors in [27] also discuss expressions for  $\hat{\psi}(Y)$  for a certain class of non-Gaussian noise distributions with the corruption mechanism continuing to be additive. This leaves room for universal denoising schemes for continuous valued data for a general class of noise distributions where the corruption mechanism is also arbitrary. Compression based approaches pioneered in (cf., e. g., [25] and [10]), as discussed in [36], are provably sub-optimal and suffer from non-practicality of implementation of optimal lossy compression schemes. The wavelet-based Bayesian estimation approach in [26], 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 [35], array-based comparative genomic hybridization (array-CGH) [19] and medical imaging [34], [17], [22], 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 [36]. The results of [36], and the denoising scheme DUDE proposed 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 joint distributions from the noisy data.
- mapping an estimated channel output distribution to an estimated channel input distribution.
- 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 [24]. 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 [5]. This approach proposes quantization of the output alphabet space and proceeds on an a

similar line to that in [36], showing that there is no essential loss of optimality in quantizing the channel output before denoising (insofar as learning the statistics of the underlying data is concerned). In spite of its theoretical elegance, this approach faces similar issues as the scheme of [36], limiting its scope of applications to small channel input alphabets. The authors of [5], 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 the two-stage DUDE-like approach in [36], [5] to the case of denoising for general alphabets. 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. [7] and references therein), however, points to the opportunity of extracting more reliable statistics from the observed data, that would lead to better denoising (as measured under a specified loss function). We do, however, maintain the sliding window approach of [5], [36] and show asymptotic universal optimality of our schemes with increasing context lengths in the limit of large sequence lengths.

Recent developments in universal denoising in the particular context of images have also been reported in [4]. 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 particular noisy symbol in question. 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. In section IV, we establish universality of a family of denoisers that we develop for the semi-stochastic setting, in which the clean data is an individual sequence and provide bounds on the difference between the performance of this proposed family of denoisers and that of the best ‘symbol-by-symbol’ denoiser chosen by a genie with full knowledge of the distribution (or probability law) of the clean data. Section V details an extension of this proposed family of denoisers to a genie that can select the best sliding window scheme, of any order, with knowledge of the underlying clean data. Section VI discusses the implication of the performance guarantees in the semi-stochastic setting to the fully stochastic setting where the clean data is generated by a stationary stochastic process, rather than an individual sequence. A slightly modified version of the proposed denoiser is shown to reduce to the scheme of [5] when the underlying clean data have finite alphabet size. The proposed family of denoisers can, hence, be seen as a natural extension of those in [5] to the current setting of denoising continuous valued symbols corrupted by a continuous memoryless channel where the clean data components may take values in a continuum. In section VII, we present some preliminary experimental results of applying the proposed schemes to denoising of gray-scale images. We conclude in section VIII with a summary of 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 relegating most of the proofs to the appendices.

## II. PROBLEM SETTING AND NOTATIONS

Let  $\mathbf{x} = (x_1, x_2, \dots)$  be an individual (deterministic) noise-free source signal <sup>1</sup> with components taking values in  $[a, b] \subset \mathbb{R}$  and  $\mathbf{Y} = (Y_1, Y_2, \dots)$ ,  $Y_i \in \mathbb{R}$  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 [5]. 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 make the following assumptions about the channel,

- C1. A memoryless channel, which is to say that the components of  $\mathbf{Y}$  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,  $\mathcal{C}$ , is uniformly tight in the sense

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

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

- C3. The distribution functions  $F_{Y|x}$  are absolutely continuous for all  $x \in [a, b]$  w.r.t the Lebesgue measure and  $\{f_{Y|x}\}$  denotes the corresponding densities. This assumption is not crucial for the validity of our approach but is made for concreteness in the construction of our schemes and the development of their performance guarantees.
- C4. The conditional densities of the channel form a set of linearly independent functions. This is equivalent to the ‘invertibility’ condition of [36] which ensures that, to any distribution on the input to the channel there corresponds a unique channel output.
- C5. The 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 is continuous. The precise analytical expression describing this condition is discussed in Appendix I.
- C6. The expected loss, for reasonably well-behaved loss functions (conditions L1-L2 listed subsequently in this section), induced by two output distributions that are close (under the metric discussed in section III) is continuous. Again, the analytical expression describing this condition is in the Appendix I.

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

<sup>1</sup>throughout the paper we will be using the terms ‘signal’ and ‘sequence’ interchangeably

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 (C7 discussed in Appendix I) 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  $\Lambda : [a, b]^2 \rightarrow [0, \infty)$  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

$$L_{\hat{X}^n}(x^n, y^n) = \frac{1}{n} \sum_{i=1}^n \Lambda(x_i, \hat{X}^n(y^n)[i]) \quad (2)$$

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.,  $\Lambda_{\max} < \infty$  where  $\Lambda_{\max} = \sup_{x, \hat{x} \in [a, b]} \Lambda(x, \hat{x})$

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

$$\|\Lambda\|_L = \sup_{0 < \Delta < (b-a)} \frac{\lambda(\Delta)}{\Delta} \quad (3)$$

where,

$$\lambda(\Delta, x) = \sup_{y \in [a, b]} \sup_{x' : |x-x'| < \Delta} |\Lambda(x, y) - \Lambda(x', y)| \quad (4)$$

and

$$\lambda(\Delta) = \sup_{x \in [a, b]} \lambda(\Delta, x) \quad (5)$$

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  $\mathcal{F}^{[a, b]}$  denote the set of all probability distribution functions with support contained in the interval  $[a, b]$ . For  $F \in \mathcal{F}^{[a, b]}$ , we let

$$\mathcal{U}(F) = \min_{\hat{x} \in [a, b]} \int_{x \in [a, b]} \Lambda(x, \hat{x}) dF(x) \quad (6)$$

denote its ‘Bayes envelope’ (our assumptions on the loss function will imply existence of the minimum). In other words,  $\mathcal{U}(F)$  denotes the minimum achievable expected loss when guessing the value of  $X \sim F$ . Define the symbol-by-symbol minimum loss of  $x^n$  by

$$D_0(x^n) = \min_g E \left[ \frac{1}{n} \sum_{i=1}^n \Lambda(x_i, g(Y_i)) \right] \quad (7)$$

where the minimum is over all measurable maps  $g : \mathbb{R} \rightarrow [a, b]$ .  $D_0(x^n)$  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 [5]. This is slightly different than the benchmark used in [36], 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

$$F_{x^n}(x) = \frac{|\{1 \leq i \leq n : x_i \leq x\}|}{n}, \quad (8)$$

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

$$D_0(x^n) = \min_g \int_{[a,b]} E_x \Lambda(x, g(Y)) dF_{X^n}(x) \quad (9)$$

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

$$E_x \Lambda(x, g(Y)) = \int \Lambda(x, g(y)) f_{Y|x}(y) dy \quad (10)$$

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

$$\begin{aligned} E_{F \otimes \mathcal{C}} \Lambda(X, g(Y)) &= \int_{[a,b]} E_x \Lambda(x, g(Y)) dF(x) \\ &= \int_{[a,b]} \left[ \int_{\mathbb{R}} \Lambda(x, g(y)) f_{Y|x}(y) dy \right] dF(x) \end{aligned} \quad (11)$$

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

$$\min_g E_{F \otimes \mathcal{C}} \Lambda(X, g(Y)) = E_{F \otimes \mathcal{C}} \mathcal{U}([F \otimes \mathcal{C}]_{X|Y}) \quad (12)$$

with  $\mathcal{U}$  denoting the Bayes envelope as defined above. Letting  $g_{\text{opt}}[F]$  denote the achiever of the minimum in (12), we note that is given by the Bayes response to  $[F \otimes \mathcal{C}]_{X|y}$ , namely,

$$\begin{aligned} g_{\text{opt}}[F](y) &= \arg \min_{\hat{x} \in [a,b]} \int_{[a,b]} \Lambda(x, \hat{x}) d[F \otimes \mathcal{C}]_{X|y}(x) \\ &= \arg \min_{\hat{x} \in [a,b]} \int_{[a,b]} \Lambda(x, \hat{x}) f_{Y|x}(y) dF(x) \end{aligned} \quad (13)$$

In Lemma 12, we will establish the concavity of  $\mathcal{U}(F)$ , 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 (9), (10) and (11) we have

$$D_0(x^n) = \min_g E_{F_{x^n} \otimes \mathcal{C}} \Lambda(X, g(Y)) \quad (14)$$

where  $F_{x^n}$  was defined in (8) 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’ DENOISER AND PRELIMINARIES

$F_{x^n}$  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,  $\mathcal{C}$ . We approach this problem by first estimating a function that tracks the evolution of the ‘average’ density function according to which the noisy symbols are distributed. For an input sequence  $x^n$ , given the memoryless nature of the channel, the output symbols will be independent with respective distributions,  $\{F_{Y|x_1}, \dots, F_{Y|x_n}\}$  and have the corresponding density functions,  $\{f_{Y|x_1}, \dots, f_{Y|x_n}\}$ . The function we are interested in estimating is

$$f_Y^n(y) = \frac{1}{n} \sum_{i=1}^n f_{Y|x_i}(y) \quad (15)$$

which can be thought of as the marginal density,  $f_Y^n$ , 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 [7], [6], the details of which are discussed in Subsection III-A below. Once we have an estimate  $f_Y^n = f_Y^n[Y^n]$  for this function, we use it to estimate the input empirical distribution by

$$\hat{F}_{x^n}[Y^n] = \arg \min_{F \in \mathcal{F}_n^{[a,b]}} d \left( f_Y^n, \underbrace{\int f_{Y|x} dF(x)}_{[F \otimes \mathcal{C}]_Y} \right) \quad (16)$$

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

$$F(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(x \leq x_i)} \quad (17)$$

for some  $n$ -tuple,  $x^n = (x_1, x_2, \dots, x_n)$ , with  $[a, b]$ -valued components. The norm,  $d$ , in (16) is defined as

$$d(f, g) = \int |f(y) - g(y)| dy \quad (18)$$

The channel,  $\mathcal{C}$ , induces a set of ‘feasible’ densities of the output noisy symbol corresponding to the family of empirical distributions of the underlying clean sequence at the input of the channel. The density estimate,  $f_Y^n$ , which is constructed only from the noisy sequence,  $Y^n$ , is oblivious to 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 (16). The uniqueness of the minimizer in (16) 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 below. For a given quantization step size,  $\Delta$ , the

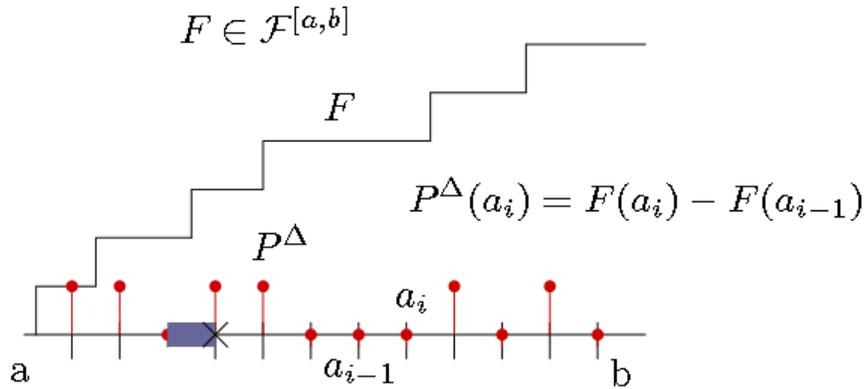


Fig. 1. Quantization of the support of a distribution function,  $F \in \mathcal{F}^{[a,b]}$

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,

$$\mathcal{F}^\Delta = \{P^\Delta : P^\Delta = (P(a_0), P(a_1), \dots, P(a_{N(\Delta)}))\}$$

$$\mathcal{A}^\Delta = \{a_i = a + i\Delta, i = 0, \dots, N(\Delta)\}$$

$$\text{s.t. } \sum_{i=1}^{N(\Delta)} P(a_i) = 1$$

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

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,

$$P^\Delta(a_i) = F(a_i) - F(a_{i-1}) \quad (19)$$

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

$$\hat{P}_{x^n}^\Delta(a_i) = \hat{F}_{x^n}(a_i) - \hat{F}_{x^n}(a_{i-1}) \quad (20)$$

where,  $a_i \in \mathcal{A}^\Delta$ .

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

$$\hat{P}_{x^n}^{\delta,\Delta} = Q_\delta(\hat{P}_{x^n}^\Delta) \quad (21)$$

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,  $\hat{P}_{x^n}^{\delta,\Delta}$ . This is formalized in section III-C where we discuss the precise convergence notion of  $\hat{P}_{x^n}^\Delta$  to the un-quantized probability measure.

The minimizer of the Bayes envelope in (13) is then constructed from the quantized probability mass function,  $\hat{P}_{x^n}^{\delta,\Delta}$ , as  $g_{\text{opt}}[\hat{P}_{x^n}^{\delta,\Delta}]$ , where  $g_{\text{opt}}$  for the quantized clean symbol is,

$$g_{\text{opt}}[P](y) = \arg \min_{\hat{x} \in \mathcal{A}^\Delta} \sum_{a \in \mathcal{A}^\Delta} \Lambda(a, \hat{x}) \cdot f_{Y|X=a}(y) \cdot P(X = a) \quad (22)$$

$\mathcal{A}^\Delta$  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.,  $\hat{P}_{x^n}^{\delta,\Delta}$  (it does not sum up to 1). Equipped with  $\hat{P}_{x^n}^{\delta,\Delta}$ , the candidate for the  $n$ -block symbol-by-symbol denoiser is now given by

$$\tilde{X}^{n,\delta,\Delta}[y^n](i) = g_{\text{opt}}[\hat{P}_{x^n}^{\delta,\Delta}[y^n]](y_i), \quad 1 \leq i \leq n \quad (23)$$

where,  $g_{\text{opt}}$  is given in (22). We now proceed to discuss in detail the construction and consistency results of the estimate,  $f_Y^n$ ,  $\hat{F}_{x^n}$  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  $f_Y^n$ , for the function in (15) which depends on  $x^n$  and therefore unknown to the denoiser. Given the memoryless nature of the channel, the sequence of output symbols,  $Y_1, Y_2, \dots, Y_n$  are independent random variables taking values in  $\mathbb{R}$ , having conditional densities,  $f_{Y|x_1}, f_{Y|x_2}, \dots, f_{Y|x_n}$  respectively. A density estimate is a sequence  $f^1, f^2, \dots, f^n$ , where for each  $n$ ,  $f_Y^n(y) = f^n(y; Y_1, \dots, Y_n)$  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

$$f_Y^n(y) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{y - Y_i}{h}\right) \quad (24)$$

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

$$J_n = \int \left| f_Y^n(y) - \frac{1}{n} \sum_{i=0}^n f_{Y|x_i}(y) \right| dy \quad (25)$$

The choice of  $L_1$  distance as elaborated by the authors in [7] 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, [6].

*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  $i = 1, \dots, s-1$ . 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 be 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  $f_Y^n$  be the kernel estimate in (24) and  $J_n$ , the corresponding error as defined in (25). Consider

- 1)  $J_n \rightarrow 0$  in probability as  $n \rightarrow \infty$ , for some sequence  $\mathbf{x} = (x_1, x_2, \dots)$
- 2)  $J_n \rightarrow 0$  in probability as  $n \rightarrow \infty$ , for all sequences  $\mathbf{x} = (x_1, x_2, \dots)$
- 3)  $J_n \rightarrow 0$  almost surely as  $n \rightarrow \infty$ , for all sequences  $\mathbf{x} = (x_1, x_2, \dots)$
- 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  $\mathbf{x}$ .
- 5)  $\lim_{n \rightarrow \infty} h = 0$ ,  $\lim_{n \rightarrow \infty} nh = \infty$

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,  $\{f_{Y|x}\}_{x \in [a,b]}$  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

$$\lim_{n \rightarrow \infty} h_n = 0 \text{ and } \lim_{n \rightarrow \infty} nh^d = \infty \quad (26)$$

*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.  $\blacksquare$

### B. Channel Inversion

The mapping in (16) projects the kernel density 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  $f_Y^n$  in (24), it is clear that  $f_Y^n$  is a bona fide density on  $\mathbb{R}$ . Additionally, from the construction of  $\hat{F}_{x^n}$  in (16), we see that for every  $F \in \mathcal{F}_n^{[a,b]}$ ,  $[F \otimes \mathcal{C}]_Y$  is also a valid density in  $\mathbb{R}$ . Finally, from the definition of the norm,  $d$ , in (18), it is true that for  $f_Y^n$  and  $[F \otimes \mathcal{C}]_Y$  being bona fide densities on  $\mathbb{R}$ ,  $0 \leq d(f_Y^n, [F \otimes \mathcal{C}]_Y) \leq 2$ ,  $\forall n$ . These facts, together with the convexity of  $\mathcal{F}_n^{[a,b]}$  show that the estimator in (16) is well defined. With the Levy metric defined as:

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

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

we have:

*Theorem 2:* For the estimator,  $\hat{F}_{x^n}$  defined in equation (16) we have  $\lambda(F_{x^n}, \hat{F}_{x^n}) \rightarrow 0$  a.s. for all  $x \in [a, b]^\infty$ . The proof of Theorem 2 is discussed in detail in the Appendix III.

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

In this section, we discuss the convergence notion of  $\hat{P}_{x^n}^\Delta$  to the law corresponding to the un-quantized distribution function  $\hat{F}_{x^n}$ .

*Definition 4 ( $\beta$  metric):* For any two laws  $P$  and  $Q$  on  $S$ ,  $f : S \rightarrow \mathbb{R}$  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

$$\|f\|_{BL} = \|f\|_L + \|f\|_\infty \quad (27)$$

and

$$\|f\|_L := \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x - y|}, \quad \|f\|_\infty = \sup_x |f(x)| \quad (28)$$

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

*Theorem 3:*

$$\lim_{\Delta \rightarrow \infty} \beta(\hat{P}_{x^n}, \hat{P}_{x^n}^\Delta) = 0 \quad (29)$$

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

*Proof:* Follows directly from Lemma 2. ■

*Lemma 2:* For any  $F \in \mathcal{F}^{[a,b]}$ ,

$$\lim_{\Delta \rightarrow 0} \beta(P, P^\Delta) = 0 \quad (30)$$

where  $P$  is the law associated with distribution functions in the family  $\mathcal{F}^{[a,b]}$ . Particularly, the  $F$  and  $P^\Delta$  that satisfies (30) is defined by,

$$P^\Delta(a_i) = F(a_i) - F(a_{i-1}) \quad (31)$$

where  $a_i \in \mathcal{A}^\Delta$  and  $\mathcal{A}^\Delta$  is the finite alphabet approximation of  $[a, b]$  discussed earlier.

In words, any empirical distribution of the underlying clean sequence is approximated arbitrarily well with a PMF on the quantized set of points when the quantization is fine enough.

Next we discuss the mechanics of the construction of the denoiser, which has the density estimation and the channel inversion steps as its core.

#### D. Implementation of the symbol-by-symbol denoiser

The implementation of the denoiser in the previous section involves a discretization of the density estimation and the channel inversion steps. The discretized version of the kernel density estimate,  $f_Y^n(y)$ , in (24) is evaluated at a set of discrete points,  $\{y_1, \dots, y_N\}$ . This gives an  $N$ -dimensional vector of the distribution function,  $p_Y^n(y)$ . The ‘‘channel inversion’’ in (16) is also discretized using the estimate,  $p_Y^n(y)$ .

1) *Fast kernel density estimation:* The Kernel density estimation in (24) for a given kernel function,  $K$ , although simple in construction, is faced with a significant computational burden for a brute-force computation of  $O(Nn)$  corresponding to  $n$  data points and  $N$  points  $\{y_1, \dots, y_N\}$  at which  $p_Y^n(y)$  is evaluated. The computational complexity can be greatly reduced by using FFT based methods [31]. Recently, there has been extensive work on the use of fast gauss transform-based techniques [16] for reduction of computational complexity. These techniques reduce the complexity from  $O(Nn)$  to  $O(N+n)$ . The cardinal factor in nonparametric density estimation procedures is the choice of the *optimal* bandwidth,  $h$ , in (24). There has been some recent work in [14] on using dual-tree methods to derive fast methods for optimal bandwidth choice that continues to maintain the complexity of this step at  $O(N+n)$ . For  $N = O(n)$ , this reduces to  $O(n)$ .

2) *Channel inversion using linear programming techniques*: In solving the channel inversion problem in (16), we are looking for a vector in the probability simplex,  $\mathcal{F}^\Delta = \{P : \sum_{i=1}^{N(\Delta)} P(a_i), a_i \in \mathcal{A}^\Delta\}$ , for our candidate distribution function,  $\hat{P}_{x^n}^{\delta, \Delta}$ . The discretized version of (16) is given by,

$$\hat{P}_{x^n}^{\delta, \Delta} = \arg \min_{p \in \mathcal{F}^\Delta} \sum_{i=1}^N \left| p_Y^n(y_i) - \sum_{j=1}^{N(\Delta)} f_{Y|x=x_j}(y_i) Q_\delta(p(x_j)) \right| \quad (32)$$

The objective function, being an  $L_1$ -norm, is clearly a convex function (of the input distribution,  $p(\cdot)$ ) and the candidate minimizer also resides in the convex subspace, viz., the probability simplex  $\mathcal{F}^\Delta$ . This can be easily solved using well-studied linear programming algorithms in the broader area of convex optimization techniques.

The particular reformulation of the problem solved is of the form

$$\begin{aligned} \hat{P}_{x^n}^{\delta, \Delta} &= \arg \min_{p \in \mathcal{F}^\Delta} \sum_{i=1}^N \varepsilon_i \\ \text{s.t.} \quad & p_Y^n(y_i) - \sum_{j=1}^{N(\Delta)} f_{Y|x=x_j}(y_i) Q_\delta(p(x_j)) \leq \varepsilon_i \\ & \sum_{j=1}^{N(\Delta)} f_{Y|x=x_j}(y_i) Q_\delta(p(x_j)) - p_Y^n(y_i) \leq \varepsilon_i \quad \forall i \in \{1, \dots, N\} \end{aligned} \quad (33)$$

The computational complexity of solving this problem using the popular interior point methods [2] is  $O((N + N(\Delta))^3) = O((N + \frac{1}{\Delta})^3) = O((N + \log n)^3)$ . This again, for  $N = O(n)$ , reduces to  $O((n + \log n)^3) = O(n^3)$ .

The two-pronged quantization discussed in the previous section can be naturally built into the optimization problem in (32) by searching in

$$\mathcal{F}^{\delta, \Delta} = \{Q_\delta(P) : P \in \mathcal{F}^\Delta\} \quad (34)$$

the set of  $N(\Delta)$ -tuples with components in  $[0,1]$  that are integer multiples of  $\frac{1}{\delta}$  with point masses on the set  $\mathcal{A}^\Delta$ .

The formulation would then be

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

This channel inversion is at the heart of the denoiser in (22) and its simple formulation makes the scheme particularly elegant and practically implementable. The estimate of the empirical distribution in (32) is then plugged into (22) to finally give an estimate of the underlying clean symbol according to (23). The denoiser is described as Algorithm 1 below.

#### IV. PERFORMANCE GUARANTEES FOR THE SYMBOL BY SYMBOL DENOISER

The main result of this section is Theorem 5 below, which establishes the universal asymptotic optimality of our proposed symbol-by-symbol denoiser in (23) with respect to the class of symbol-by-symbol schemes. The

<p><b>input</b> : Noisy sequence <math>y^n</math>, channel <math>\mathcal{C}</math></p> <p><b>output</b>: Denoised sequence, <math>\hat{x}^n</math></p> <p><b>1 FIRST PASS</b></p> <p><b>2 Density estimation step</b></p> <p><b>input</b> : Noisy sequence, <math>y^n</math></p> <p><b>output</b>: Density estimate, <math>f_Y^n</math></p> <p>3 Determine the optimal bandwidth from any one of the techniques discussed in [31], e.g., cross-validation</p> <p>4 Use techniques discussed in [14] for <i>fast</i> evaluation of (24)</p> <p><b>5 Channel inversion step</b></p> <p><b>input</b> : <math>f_Y^n</math>, Quantization resolutions, <math>\delta, \Delta</math></p> <p><b>output</b>: <math>\hat{P}_{x^n}^{\delta, \Delta}</math></p> <p>6 Construct an LP (Linear Program) as in (33) and use <code>linprog</code> (in MATLAB) or any complex program solver to solve it. Alternatively, use log-barrier methods discussed in [3] to solve for the estimate, <math>\hat{F}_{x^n}</math></p> <p>7 Use the quantization mapping in (20) to map <math>\hat{F}_{x^n}</math> to <math>\hat{P}_{x^n}^{\Delta}</math></p> <p>8 Then use a uniform quantizer with resolution <math>\delta</math> to get <math>\hat{P}_{x^n}^{\delta, \Delta} \leftarrow Q_{\delta} \left( \hat{P}_{x^n}^{\Delta} \right)</math></p> <p><b>9 SECOND PASS</b></p> <p><b>input</b> : Noisy sequence, <math>y^n</math>, channel <math>\mathcal{C}</math>, estimate of input distribution <math>\hat{P}_{x^n}^{\delta, \Delta}</math></p> <p><b>output</b>: Denoised Sequence, <math>\hat{x}^n</math></p> <p>10 Use equation (22), (23) to denoise at every location, <math>i</math></p> <p>11 <b>for</b> <math>i \leftarrow 1</math> <b>to</b> <math>n</math> <b>do</b></p> <p>12     <math>\hat{x}_i \leftarrow g_{\text{opt}} \left[ \hat{P}_{x^n}^{\delta, \Delta} \right] (y_i)</math></p> <p>13 <b>end</b></p>
---

**Algorithm 1:** Symbol-by-symbol denoiser in Section III

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  $\mathbf{x}$ . The benchmark performance for the clean sequence is the minimum possible symbol-by-symbol loss,  $D_0(x^n)$ , defined in Section II. Theorem 5 shows that our proposed denoiser,  $g_{\text{opt}} \left[ \hat{P}_{x^n}^{\delta, \Delta} \right]$ , asymptotically (as the number of observations increases) achieves that benchmark performance. This is achieved by bounding the deviation of the cumulative loss incurred by  $g_{\text{opt}} \left[ \hat{P}_{x^n}^{\delta, \Delta} \right]$  from the minimum possible symbol-by-symbol loss in Theorem 4 for any block length,  $n$ . Hence we show that,  $g_{\text{opt}} \left[ \hat{P}_{x^n}^{\delta, \Delta} \right]$  performs essentially as well as the best possible symbol-by-symbol denoiser,  $D_0(x^n)$ .

In preparation for Theorem 4 let  $\mathcal{F}^{\delta, \Delta}$ , defined in (34), denote the set of probabilities with components in  $[0, 1]$  that are integer multiples of  $\delta$  (defined under Q2. in section III). Note that  $\hat{P}_{x^n}^{\delta, \Delta} \in \mathcal{F}^{\delta, \Delta}$ , where  $\hat{P}_{x^n}^{\delta, \Delta}$  was defined in (21). Also, let  $\mathcal{G}_{\delta, \Delta} = \{g_{\text{opt}}[P]\}_{P \in \mathcal{F}^{\delta, \Delta}}$  denote the set of all possible denoisers that can be constructed from the

members of the set  $\mathcal{F}^{\delta, \Delta}$  using (22). Define  $G(\epsilon, B) = \frac{2\epsilon^2}{B^2}$ ,

$$\alpha_n(\epsilon, \delta, \Delta, \rho, \gamma) = \left[ \frac{1}{\delta} + 1 \right]^\Delta \left[ 2e^{-G(\epsilon + \delta\Lambda_{\max}, \Lambda_{\max})n} + e^{-(1-\rho)\frac{n\gamma^2}{2}} \right] + e^{-(1-\rho)\frac{n\gamma^2}{2}} \quad (35)$$

$$\nu(\epsilon, \delta, \Delta, \Lambda, \mathcal{C}) = 3\epsilon + 5\delta\Lambda_{\max} + 4\xi_\Delta\Lambda_{\max} + 4\lambda(\Delta)(1 + \xi_\Delta) \quad (36)$$

$$1 - \frac{\rho(\epsilon, \delta)}{2} = \left( 1 - \frac{6\epsilon}{\delta} \right)^2 \quad (37)$$

where

$$\xi_\Delta = \sup_{x \in [a, b]} \sup_{\substack{\hat{x} \in [a, b] \\ |x - \hat{x}| \leq \Delta}} \int |f_{Y|x}(y) - f_{Y|\hat{x}}(y)| dy \quad (38)$$

and  $\lambda(\Delta)$  is the moduli of continuity defined in (5). The Lipschitz norm,  $\|\Xi\|_L$  of  $\xi_\Delta$  is given by

$$\|\Xi\|_L = \sup_{0 < \Delta < (b-a)} \frac{\xi_\Delta}{\Delta} \quad (39)$$

$D_0(x^n)$  is the symbol-by-symbol minimum loss of  $x^n$  defined in (7).

*Theorem 4:* For all  $\epsilon > 0$ ,  $\delta > 0$ ,  $\rho = \rho(\epsilon, \delta)$ ,  $\Delta > 0$  and  $x^n \in [a, b]^n$  let,

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

then, we have

$$\Pr(|L_{\tilde{X}^{n, \delta, \Delta}}(x^n, Y^n) - D_0(x^n)| > \nu(\epsilon, \delta, \Delta, \Lambda, \mathcal{C})) \leq \alpha_n(\epsilon, \delta, \Delta, \rho, \gamma) \quad \forall n \text{ s.t. } nh_n > n_0(\mathcal{C}, \rho, \delta, K) \quad (40)$$

where,  $\|\Xi\|_L$  is defined in (39) and the form of  $n_0$  in (112). Note that the tightness condition on the probability measures associated with the family of the conditional densities of the channel,  $\mathcal{C}$ , guarantees that  $n_0(\mathcal{C}, \rho, \delta, K) < \infty$ ,  $\forall \rho \in (0, 1)$ . Theorem 4 formalizes the fact that the probability of deviation of the cumulative symbol-by-symbol loss,  $L_{\tilde{X}^{n, \delta, \Delta}}(x^n, Y^n)$  from the minimum possible loss,  $D_0(x^n)$  is exponentially small with the block length  $n$ .

#### *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 (14), 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 (12) 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 derived as a function of the marginal empirical 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 empirical distribution. The almost sure

convergence of the marginal density at the output of the memoryless channel gives us, through the mapping in (16), 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 (12) and (14), 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  $\mathcal{U}([F \otimes \mathcal{C}]_{X|Y})$  under the two corresponding induced densities.

The goal, eventually, is to bound the deviation of the cumulative loss,  $L_{\hat{X}^n, \delta, \Delta}$ , incurred by the proposed denoiser in (23) 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. These Lemmas are analogous (in spirit) to the corresponding ones, i.e., Lemmas 1, 2, 3 (for context length,  $k = 0$ ) in the discrete-input, general valued output setting in [5]. There are, however, subtle differences in the bounds and the requirements on the channel, loss functions (C1-7, L1-2) that make it possible in this continuous valued setting. 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. Take now,  $\delta = \delta_n, \Delta = \Delta_n$  such that  $\delta_n \downarrow 0, \Delta_n \downarrow 0$  for all  $\epsilon > 0$  and

$$\sum_{n=1}^{\infty} \alpha_n(\epsilon, \delta_n, \Delta_n, \rho, \gamma) < \infty \quad (41)$$

For example,  $\delta_n, \Delta_n = \frac{1}{\log n}$  would satisfy the above requirements of summability and growth for any  $\epsilon > 0$ . With the growth rates that satisfy the summability condition in (41) for  $\alpha_n(\epsilon, \delta_n, \Delta_n, \rho, \gamma)$  let,

$$\hat{X}_{\text{ssuniv}}^n = \tilde{X}^{n, \delta_n, \Delta_n} \quad (42)$$

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,  $\mathbf{x}$ .

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

$$\lim_{n \rightarrow \infty} \left[ L_{\hat{X}_{\text{ssuniv}}^n}(x^n, Y^n) - D_0(x^n) \right] = 0 \quad a.s. \quad (43)$$

## 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 previous sections to a  $2k+1$ -length sliding window denoising scheme, one that competes with sliding window schemes. The performance guarantees made in the symbol-by-symbol case also hold in the proposed extension. The first result of this section is presented in Theorem 6, which assess the performance of our proposed scheme by showing that it does well

relative to that of the best sliding window scheme of order  $2k + 1$ , as would be chosen by a “genie” that knows the underlying clean sequence  $x^n$ . The main result of this section is Theorem 7, which establishes the strong universality of our proposed sliding window denoiser, showing that it does essentially as well as any sliding window scheme, of any order, as the length of the data increases, regardless of what the underlying clean sequence may be. Theorem 7 will be shown to be a direct consequence of Theorem 6, analogously as Theorem 5 of the previous section followed from Theorem 4.

A. *Extension to competition with  $2k + 1$ -order sliding window denoisers*

The scheme we propose is pictorially depicted in Fig. 2 below. The necessity for independence of the symbols in



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

the density estimation procedure discussed in section III-A coupled with the memoryless nature of the channel is the motivation 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 super-symbols. Note that there are  $2k + 1$  such subsequences and each subsequence,  $i$  (counting in the order of symbols in the sequence), consists of  $\lceil \frac{n-2k-i-1}{2k+1} \rceil$ ,  $2k + 1$ -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,

$$x^{n_i} = \left\{ x_i^{2k+i}, x_{2k+1+i}^{4k+1+i}, \dots, x_{\left(\lceil \frac{n-2k-i-1}{2k+1} \rceil - 1\right)(2k+1)+i}^{(2k+1)+i+2k} \right\}$$

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  $f_Y^{n,k}$  denote the  $k^{\text{th}}$  order density estimate of the noisy sequence of symbols and is computed exactly as in (24) except  $y, Y_i \in \mathbb{R}^{2k+1}$ . Denote  $\mathcal{F}^{[a,b],k}$  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  $k^{\text{th}}$ -order sliding window minimum loss and is defined as

$$D_k(x^n) = \min_g E \left[ \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(x_i, g(Y_{i-k}^{i+k})) \right] \quad (44)$$

Note the similar definition of symbol-by-symbol denoisability in (7). As before,  $D_k(x^n)$  can be expressed as

$$D_k(x^n) = \min_g E_{F_{x^n}^k \otimes \mathcal{C}} \Lambda(X, g(Y_{-k}^k)) \quad (45)$$

where  $F_{x^n}^k$  is the  $k^{\text{th}}$  order empirical distribution of the source. Define further the sliding window denoisability of the individual sequence  $\mathbf{x} = (x_1, x_2, x_3, \dots)$  by

$$D(\mathbf{x}) = \lim_{k \rightarrow \infty} \limsup_{n \rightarrow \infty} D_k(x^n) \quad (46)$$

where the limit exists by monotonicity. In words,  $D(\mathbf{x})$  is the loss of a genie who knows the underlying clean sequence and can choose to denoise with the best sliding window scheme, of arbitrary order. Extending the definition of  $k^{\text{th}}$ -order minimum loss to a subsequence,  $x^{n_i}$  as

$$D_k(x^{n_i}) = \min_g E_{F_{x^{n_i}}^k \otimes \mathcal{C}} \Lambda(X, g(Y_{-k}^k)) \quad (47)$$

The mapping to the corresponding  $k^{\text{th}}$  order input empirical distribution is given by

$$\hat{F}_{x^n}^k[Y^n] = \arg \min_{F \in \mathcal{F}_n^{[a,b],k}} d \left( f_Y^{n,k}, \underbrace{\int \prod_{i=-k}^k f_{Y|x_i} dF(x_{-k}^k)}_{[F \otimes \mathcal{C}]_Y^k} \right) \quad (48)$$

where  $\mathcal{F}_n^{[a,b],k} \subseteq \mathcal{F}^{[a,b],k}$  denotes the set of  $k^{\text{th}}$  order ( $1 \leq k \leq \lfloor \frac{n}{2} \rfloor$ ) empirical distributions induced by  $n$ -tuples with  $[a, b]^{2k+1}$ -valued components.  $\hat{F}_{x^n}^{\delta, \Delta, k}$  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

$$\tilde{X}^{n_i, \delta, \Delta, k}[y^n](j) = g_{\text{opt}} \left[ \hat{F}_{x^{n_i}}^{\delta, \Delta, k}[y^{n_i}] \right] \left( y_{j-k}^{j+k} \right), \quad j \in \left\{ k+i, 3k+1+i, \dots, \lceil \frac{n-2k-i-1}{2k+1} \rceil \right\} \quad (49)$$

where the  $k^{\text{th}}$  order equivalent of the denoiser in (22) is given by

$$\begin{aligned} g_{\text{opt}}[P](y_{-k}^k) &= \arg \min_{\hat{x} \in \mathcal{A}} \Lambda(\cdot, \hat{x})^T [P \otimes \mathcal{C}]_{U|y_{-k}^k} \\ &= \arg \min_{\hat{x} \in \mathcal{A}} \sum_{a \in \mathcal{A}} \Lambda(a, \hat{x}) \cdot \left\{ \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}: u_0=a} \left[ \prod_{i=-k}^k f_{Y|X=u_i}(y_i) P(U_{-k}^k = u_{-k}^k) \right] \right\} \end{aligned} \quad (50)$$

Let,  $\mathcal{F}_{\delta, \Delta}^k$  denote the set of  $2k+1$ - dimensional vectors with components in  $[0,1]$  that are integers multiples of  $\delta$ . Note that,  $\hat{P}_{x^{n_i}}^{\delta, \Delta}[z^{n_i}] \in \mathcal{F}_{\delta, \Delta}^k$  for all  $z^n$ . Finally, let  $\mathcal{G}_{\delta, \Delta}^k = \{g_{\text{opt}}[P]\}_{P \in \mathcal{F}_{\delta, \Delta}^k}$  and

$$\tilde{X}^{n, \delta, \Delta, k} = \{\tilde{X}^{n_i, \delta, \Delta, k}\}_{1 \leq i \leq 2k+1} \quad (51)$$

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

$$L_{\tilde{X}^{n, \delta, \Delta, k}} = \frac{1}{2k+1} \sum_{i=1}^{2k+1} L_{\tilde{X}^{n_i, \delta, \Delta, k}} \quad (52)$$

where,  $L_{\tilde{X}^{n_i, \delta, \Delta, k}}$  is the cumulative loss incurred by the proposed denoiser for the  $i^{\text{th}}$ - subsequence. The following Lemma illustrates a rather intuitive fact, the average minimum  $k^{\text{th}}$  order sliding window loss incurred by operating on each of the subsequences is at most the minimum  $k^{\text{th}}$  order sliding window loss for the entire sequence.

*Lemma 3:* For all  $n \geq 1$ ,  $k \leq \lfloor \frac{n}{2} \rfloor$ ,

$$\frac{1}{2k+2} D_k(x^{n_i}) \leq D_k(x^n) \quad (53)$$

### B. Performance guarantees

In this section we present 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 growth constraints are specified at the end of this section. They are dictated by an exponential bound on the deviation between the cumulative loss,  $L_{\tilde{X}^{n, \delta, k, \Delta}}$  and  $D_k$  which we now develop.

Let

$$\begin{aligned} \alpha_n(\epsilon, k, \delta, \Delta, \rho, \gamma) &= \\ & \left[ \frac{1}{\delta} + 1 \right]^{\Delta^{2k+1}} \cdot \left[ A(k, \epsilon + \delta \Lambda_{\max}, \Lambda_{\max}) \exp(-(n+1)G(k, \epsilon + \delta \Lambda_{\max}, \Lambda_{\max})) + \right. \\ & \left. A\left(k, \sqrt{1-\rho}, \frac{2}{\gamma}\right) \exp\left(-(n+1)G\left(k, \sqrt{1-\rho}, \frac{2}{\gamma}\right)\right) \right] + e^{-(1-\rho)\frac{(n-2k)\gamma^2}{2(2k+1)}} \end{aligned}$$

where,

$$A(k, \epsilon, B) = (2k + 1) \exp\left(\frac{2\epsilon^2}{B^2}\right) \quad (54)$$

$$G(k, \epsilon, B) = \frac{2\epsilon^2}{(2k + 1)B^2} \quad (55)$$

and

$$\nu(\epsilon, \delta, \Delta, \Lambda, \mathcal{C}, k) = 3\epsilon + 5\delta\Lambda_{\max} + 4\xi_{\Delta}^{2k+1}\Lambda_{\max} + 4\lambda(\Delta)(1 + \xi_{\Delta}^{2k+1}) \quad (56)$$

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

*Theorem 6:* For all  $n \geq 1$ ,  $\epsilon > 0$ ,  $\delta > 0$ ,  $\rho = \rho(\epsilon, \delta)$  defined in (37),  $\Delta > 0$ ,  $1 \leq k \leq \lfloor \frac{n}{2} \rfloor$  and  $x^n \in [a, b]^n$

$$\Pr(L_{\tilde{X}^{n, \delta, \Delta, k}}(x^n, Y^n) - D_k(x^n) > \nu(\epsilon, \delta, \Delta, \Lambda, \mathcal{C}, k)) \leq \alpha_n(\epsilon, k, \delta, \Delta, \rho, \gamma_k) \quad \forall n \text{ s.t } nh_n^k > n_k(\mathcal{C}, \rho, \delta, K) \quad (57)$$

where,

$$\gamma_k = \frac{\epsilon}{(\|\Lambda\|_L + \Lambda_{\max} \|\Xi\|_L^k + (b-a) \|\Lambda\|_L \|\Xi\|_L^k + \Lambda_{\max})} \quad (58)$$

$\|\Xi\|_L^k$  (the  $k^{\text{th}}$  order equivalent of  $\|\Xi\|_L$  in (39)) and  $n_k(\mathcal{C}, \rho, \delta, K)$  are defined in (159) and (110) respectively.

Take now,  $k = k_n$ ,  $\delta = \delta_n$  and  $\Delta = \Delta_n$  such that  $k_n \rightarrow \infty$ ,  $\delta_n \downarrow 0$ ,  $\Delta_n \downarrow 0$ ,

$$\sum_{n=1}^{\infty} \alpha_n(\epsilon, k_n, \delta_n, \Delta_n, \rho, \gamma_{k_n}) < \infty$$

and  $n_k(\mathcal{C}, \rho, \delta, K) < \infty$ . With growth rates that satisfy these conditions let,

$$\hat{X}_{\text{univ}}^n = \tilde{X}^{n, \delta_n, \Delta_n, k_n} \quad (59)$$

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

Particularly for additive Gaussian noise channels of finite variance, squared and absolute loss functions with the aforementioned growth rates of  $k_n$ ,  $\Delta_n$ ,  $\delta_n$  satisfy the conditions of  $\lambda(\Delta_n) \delta_{\Delta_n}^{2k_n+1} \rightarrow 0$  and  $\delta_{\Delta_n}^{2k_n+1} \rightarrow 0$ .

We now have the following result as a direct consequence of Theorem 6 and the Borel-Cantelli Lemma.

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

$$\lim_{n \rightarrow \infty} \left[ L_{\hat{X}_{\text{univ}}^n}(x^n, Y^n) - D_{k_n}(x^n) \right] = 0 \quad a.s. \quad (60)$$

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 denoisability. Specifically,

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

$$\limsup_{n \rightarrow \infty} \left[ L_{\hat{X}_{\text{univ}}^n}(x^n, Y^n) - D(\mathbf{x}) \right] \leq 0 \quad a.s. \quad (61)$$

which is a corollary of Theorem 7, proved similarly as corollary 1 in [5].

### C. Computation complexity of the proposed denoiser

Let us summarize the computational complexity of the proposed denoisers: the “symbol-by-symbol” and the  $k^{\text{th}}$  order extensions. For the symbol by-symbol denoiser, we have already covered the analysis in Sections III-D.1, III-D.2. For  $X_{\text{univ}}^n$  defined in (59), we have:

a) *Symbol-by-symbol scheme:*

- 1) Fast Kernel Density Estimation,  $O(n)$

Using the techniques of fast kernel density estimation in [29], [28], [23], [14] it was shown that the complexity can be reduced from  $O(n^2)$  to  $O(n)$ .

- 2) Channel Inversion,  $O(n^3)$

The polynomial complexity of the simplex approach in linear programming problems is discussed in detail in [2].

b)  *$k^{\text{th}}$  order sliding window scheme:*

- 1) Fast Kernel Density Estimation,  $O(n)$

As before, the complexity of the denoiser continues to be linear in the length of the data,  $n$  and the context length,  $k$ , i.e.,  $O(nk^\gamma)$   $\gamma > 0$  [14].

- 2) Channel Inversion,  $O(n^{6k})$

From the fact that the dimensionality of the contexts is length  $2k$ , the channel inversion now increases in complexity exponentially and is given by  $O(n^{6k})$ . Thus, our schemes are practical for small values of  $k$ , but become unrealistic to implement as  $k$  grows.

This lead to our follow up work in [33] that uses quantized contexts in conjunction with the (low complexity) symbol-by-symbol denoiser that asymptotically (with increasing levels of quantization of the contexts) achieves the performance of the sequence of denoisers proposed here.

## VI. UNIVERSALITY IN THE STOCHASTIC SETTING

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

$$\mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) = \lim_{n \rightarrow \infty} \min_{\hat{X}_n} EL_{\hat{X}_n}(X^n, Y^n), \quad (62)$$

where the expectation is assuming  $X^n$  are the first  $n$  symbols emitted by a source with distribution  $F_{\mathbf{X}}$  and  $Y^n$  is, as before, the  $n$ -tuple of output noisy symbols from the channel  $\mathcal{C}$  that corrupts  $X^n$ . This is achieved by a “genie”

that has access to the true distribution,  $F_{\mathbf{X}}$ , of the underlying clean signal,  $\mathbf{X}$ . It has been shown in [36], [5] that the limit in (62) exists and hence the denoisability,  $\mathbb{D}(F_{\mathbf{X}}, \mathcal{C})$ , is well-defined for every stationary  $F_{\mathbf{X}}$ .

We now state the main result for the stochastic setting wherein we establish that for any stationary underlying clean sequence  $\mathbf{X} \sim F_{\mathbf{X}}$ , the expected cumulative loss incurred by our proposed scheme asymptotically achieves the denoisability,  $\mathbb{D}(F_{\mathbf{X}}, \mathcal{C})$ .

*Theorem 8:* For all stationary  $\mathbf{X}$

$$\lim_{n \rightarrow \infty} EL_{\hat{X}_{\text{univ}}^n}(X^n, Y^n) = \mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \quad (63)$$

If  $\mathbf{X}$  is also ergodic then

$$\limsup_{n \rightarrow \infty} L_{\hat{X}_{\text{univ}}^n}(X^n, Y^n) = \mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \text{ a.s.} \quad (64)$$

Given the results established for the semi-stochastic setting, the proof is analogous to that of Theorem 3 in [5] 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 and for accommodating these differences in Appendix VIII.

We conclude this section by comparing the proposed sequence of denoisers to the DUDE-like schemes in [5] for the case of finite input (or underlying clean data) and continuous valued output (noisy data). By a minor modification, the proposed denoiser collapses to that in [5] when, as in the setting of [5], the channel input alphabet is finite. 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. The theoretical details of the equivalence of the modification shown in Fig. 3 below to the denoiser in [5] are elaborated in Appendix IX.

## VII. 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. In addition, we consider a highly nonlinear, non-conventional noise distribution: a locally varying Rayleigh noise whose variance is a function of the gray level of the underlying clean image. The first pass of the denoiser is performed using a Fast Kernel Density Estimation approach proposed in [15] and a channel inversion procedure. This channel inversion is performed using a convex optimization linear programming technique that maps the output  $k^{\text{th}}$ -order density estimate to the corresponding input  $k^{\text{th}}$ -order input empirical distribution in accordance with (48). The experimental results presented in this section have been obtained by implementing the scheme of the previous sections, with no heuristic modifications that are likely to boost the performance. The practical implementation aspects are discussed in greater detail and depth in [32], [33].

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  $\mathbb{R} \rightarrow \mathcal{A} = \{0, \dots, 255\}$

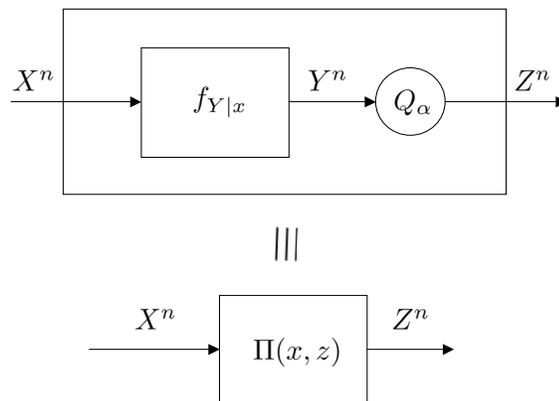


Fig. 3. Modification to our proposed scheme that is equivalent to that in [5]

and reduces to that in (50). 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 [9] and Bayesian Least Squares Gaussian Scaled Mixture (BLS-GSM) denoiser in [26]. Increasing context lengths,  $k$ , translates to accruing increasing  $k^{\text{th}}$ -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 uncon-

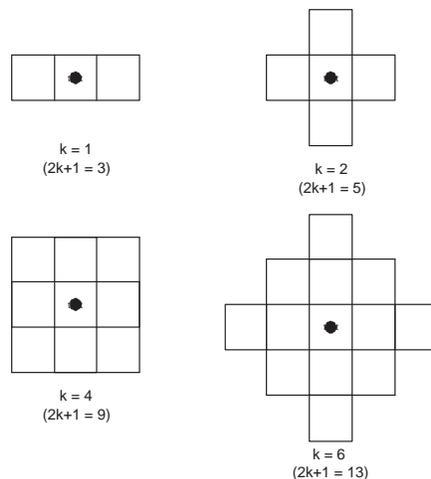


Fig. 4. 2D Contexts for context length,  $k$

ventional 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,  $f(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}$ ) 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 [32]. 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 [15]. 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 [26] to validate its efficacy.

## VIII. CONCLUSION AND FUTURE DIRECTIONS

We have presented a family of schemes for denoising continuous amplitude signals that is universally optimal. A salient feature of our setting and results is the wide generality of channels and loss functions for which they apply. The techniques presented in this paper draw from the “DUDE framework” in [36]. A weighted ‘context aggregation’ was suggested in [36] 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 [5] 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 [5] 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 [5]. 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 [18] 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

### CONDITIONS ON THE CHANNEL

In addition to conditions C1-C4 in section II, the following conditions on the channel (noise distribution) round up the necessary assumptions for the performance guarantees made in this work.

C5. The channel satisfies the uniform Lipschitz continuity condition,

$$\sup_{y \in \mathbb{R}} \|f_{Y|x}(y)\|_{BL} < \infty \quad (65)$$

where

$$\|f_{Y|x}(y)\|_{BL} = \|f_{Y|x}(y)\|_L + \|f_{Y|x}(y)\|_\infty \quad (66)$$

$$\|f_{Y|x}(y)\|_L = \sup_{\substack{x \neq z \\ x, z \in [a, b]}} \frac{|f_{Y|x}(y) - f_{Y|z}(y)|}{|x - z|} < \infty, \forall y \in \mathbb{R} \quad (67)$$

$$\|f_{Y|x}(y)\|_\infty = \sup_{x \in [a, b]} f_{Y|x}(y) \quad (68)$$

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

$$\|\Xi\|_L = \sup_{0 < \Delta < (b-a)} \frac{\xi_\Delta}{\Delta} < \infty \quad (69)$$

where,  $\xi_\Delta$  is defined in (38).

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

(refer [6] for further details)

$$D_2^*(f_{Y|x}) = \liminf_{h \downarrow 0} \int \left| (f_{Y|x} * \phi_h)^{(2)} \right| dy \quad (70)$$

$\phi_h(x) = \frac{1}{h} \phi\left(\frac{x}{h}\right)$ ,  $\phi \in C^\infty$ ,  $C^\infty$  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,  $\lim_{|t| \rightarrow 0} \Omega_{\mathcal{C}}(t) = 0$ , where

$$\Omega_{\mathcal{C}}(t) = \sup_{x \in [a, b]} \omega_x(t) \quad (71)$$

and

$$\omega_x(t) = \int |f_{Y|x}(y-t) - f_{Y|x}(y)| dy \quad (72)$$

From the fact [37] that, for any  $f \in L_1(\mathbb{R})$ , the corresponding,  $L_1$ -modulus of continuity,

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

and

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

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

## APPENDIX II PROOF OF LEMMA 1

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

*Theorem 9: Every kernel  $K$  with  $\int K = 1$ ,  $K \geq 0$  is an approximate identity, i.e for  $\lim_{n \rightarrow \infty} h_n = 0$  and every  $f_i \in L_1$ , s.t.  $D_2^*(f_i) < \infty$  are uniformly bounded we have*

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

An alternate formulation of the approximation identity is the following,

*Theorem 10:* Every kernel  $K$  with  $\int K = 1, K \geq 0$  is an approximate identity, i.e for  $\lim_{n \rightarrow \infty} h_n = 0$  and every  $f_i \in L_1$ , s.t.  $\lim_{|t| \rightarrow 0} \Omega_C(t) = 0$

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

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

$$\begin{aligned} L(x) &= (-1)^s \int_x^\infty \frac{(y-x)^{s-1}}{(s-1)!} K(y) dy & (x > 0) \\ L(-x) &= (-1)^s L(x) & (x < 0) \end{aligned}$$

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,

$$\int |L| \leq \frac{1}{s!} \int |x|^s |K(x)| dx \quad (73)$$

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

$$\int |L| = \frac{1}{s!} \int |x|^s |K(x)| dx \quad (74)$$

Finally,

$$\begin{aligned} \int L &= \int \frac{x^s}{s!} K(x) dx \\ &= \begin{cases} 0 & : s \text{ odd} \\ 0 & : s \text{ even, and the order of } K \text{ is } > s \end{cases} \end{aligned} \quad (75)$$

*Proof:* [Proof of Theorem 9]

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  $K$ ,

$$\begin{aligned}
& \left( \frac{1}{n} \sum_{i=1}^n f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^n f_i \right) \\
&= \frac{1}{n} \int \left( \sum_{i=1}^n f_i(x+y) - \sum_{i=1}^n f_i(x) \right) K_{h_n}(y) dy \quad (\text{recall that } \int K = 1) \\
&= \frac{1}{n} \sum_{i=1}^n \left[ \sum_{j=1}^{s-1} 0 + \int \int_x^{x+y} \frac{(x+y-u)^{s-1}}{(s-1)!} f_i^{(s)}(u) du K_{h_n}(y) dy \right] \\
&= \frac{1}{n} \sum_{i=1}^n \left[ \int_x^\infty f_i^{(s)}(u) \int_{u-x}^\infty \frac{(x+y-u)^{s-1}}{(s-1)!} K_{h_n}(y) dy du \right. \\
&\quad \left. - \int_{-\infty}^x f_i^{(s)}(u) \int_{-\infty}^{u-x} \frac{(x+y-u)^{s-1}}{(s-1)!} K_{h_n}(y) dy du \right] \\
&= \frac{1}{n} \sum_{i=1}^n \left[ \int_x^\infty f_i^{(s)}(u) (-1)^s (L)_{h_n}(u-x) du \right. \\
&\quad \left. - \int_{-\infty}^x f_i^{(s)}(u) (-1) (-1)^s (-1)^s (L)_{h_n}(x-u) du \right] \\
&= \frac{1}{n} \sum_{i=1}^n \left[ \int_{-\infty}^\infty f_i^{(s)}(u) (L)_{h_n}(x-u) du \right] \\
&= \frac{1}{n} \sum_{i=1}^n h^s f_i^{(s)} * L_{h_n}
\end{aligned} \tag{76}$$

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 [30],

$$\begin{aligned}
\int \left| \left( \frac{1}{n} \sum_{i=1}^n f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^n f_i \right) \right| &= \int \left| \frac{1}{n} \sum_{i=1}^n h_n^s f_i^{(s)} * L_{h_n} \right| \\
&\leq \frac{h_n^s}{n} \int \left| \sum_{i=1}^n f_i^{(s)} \right| \int |L| \\
&\leq \frac{h_n^s}{n} \left( \sum_{i=1}^n \int |f_i^{(s)}| \right) \int |L|
\end{aligned} \tag{77}$$

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 (77) simplifies to

$$\int \left| \left( \frac{1}{n} \sum_{i=1}^n f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^n f_i \right) \right| \leq h_n^s M \int |L| \tag{78}$$

Since,

$$\int |L| \leq \frac{1}{s!} \int |x|^s |K(x)| dx = B_K < \infty \tag{79}$$

for  $K$  being an order  $s$  kernel, inequality in equation (78) becomes

$$\int \left| \left( \frac{1}{n} \sum_{i=1}^n f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^n f_i \right) \right| \leq h_n^s M B_K \tag{80}$$

Taking limit  $n \rightarrow \infty$  on either sides, we get

$$0 \leq \lim_{n \rightarrow \infty} \int \left| \left( \frac{1}{n} \sum_{i=1}^n f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^n f_i \right) \right| \leq \lim_{n \rightarrow \infty} h_n^s MB_K = 0 \quad (81)$$

This can be extended to the general  $f_i$ 's using the universal derivative defined earlier. As a reminder,

$$D_s^*(f_i) \triangleq \liminf_{h \downarrow 0} \int |(f_i * \phi_h)^{(s)}| \quad (82)$$

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

$$K(x) = Ce^{-\frac{1}{1-x^2}}, \quad |x| \leq 1 \quad (83)$$

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.,  $D_2^*(f_i) < \mathcal{B}_C < \infty, \forall i \in \mathbb{N}$ , it can then be shown that,

$$\begin{aligned} \int \left| \left( \frac{1}{n} \sum_{i=1}^n f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^n f_i \right) \right| &\leq \frac{1}{n} \sum_{i=1}^n \int |f_i * K_{h_n} - f_i| \\ &\leq \frac{1}{n} \sum_{i=1}^n h_n^s D_s^*(f_i) \int |L| \\ &\leq \frac{1}{n} \sum_{i=1}^n h_n^s \mathcal{B}_C \int |L| \\ &= h_n^s \mathcal{B}_C \int |L| \end{aligned} \quad (84)$$

Taking limits on both sides we get,

$$\lim_{n \rightarrow \infty} \int \left| \left( \frac{1}{n} \sum_{i=1}^n f_i \right) * K_{h_n} - \left( \frac{1}{n} \sum_{i=1}^n f_i \right) \right| = 0 \quad (85)$$

■

*Proof:* [Proof of Theorem 10]

$$f_i(x) = f_i(x) \int K_h(t) dt = \int f_i(x) K_h(t) dt, \quad \forall i \quad (86)$$

Therefore,

$$\begin{aligned} \left| \left( \frac{1}{n} \sum_{i=1}^n f_i * K_h \right) (x) - \frac{1}{n} \sum_{i=1}^n f_i(x) \right| &= \left| \int \left[ \frac{1}{n} \sum_{i=1}^n f_i(x-t) - \frac{1}{n} \sum_{i=1}^n f_i(x) \right] K_h(t) dt \right| \\ &\leq \int \left| \frac{1}{n} \sum_{i=1}^n f_i(x-t) - \frac{1}{n} \sum_{i=1}^n f_i(x) \right| |K_h(t)|^{\frac{1}{p}} |K_h(t)|^{\frac{1}{p'}} dt \end{aligned} \quad (87)$$

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

$$\begin{aligned}
& \int \left| \left( \frac{1}{n} \sum_{i=1}^n f_i * K_h \right) (x) - \frac{1}{n} \sum_{i=1}^n f_i(x) \right|^p dx \\
& \leq \int \left[ \int \left| \frac{1}{n} \sum_{i=1}^n f_i(x-t) - \frac{1}{n} \sum_{i=1}^n f_i(x) \right|^p |K_h(t)| dt \right] \left[ \int |K_h(t)| dt \right]^{\frac{p}{p'}} dx \\
& = \|K\|_1^{\frac{p}{p'}} \int \left[ \int \left| \frac{1}{n} \sum_{i=1}^n f_i(x-t) - \frac{1}{n} \sum_{i=1}^n f_i(x) \right|^p |K_h(t)| dt \right] dx \\
& \leq \|K\|_1^{\frac{p}{p'}} \int \left[ \frac{1}{n} \sum_{i=1}^n \int |f_i(x-t) - f_i(x)|^p |K_h(t)| dt \right] dx \tag{88}
\end{aligned}$$

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

$$\begin{aligned}
\left\| \left( \frac{1}{n} \sum_{i=1}^n f_i \right) * K_h - \frac{1}{n} \sum_{i=1}^n f_i \right\|_p^p & \leq \|K\|_1^{\frac{p}{p'}} \int |K_h(t)| \frac{1}{n} \sum_{i=1}^n \omega_i(t) dt \\
& \leq \|K\|_1^{\frac{p}{p'}} \int |K_h(t)| \Omega(t) dt \tag{89}
\end{aligned}$$

For  $\delta > 0$ ,

$$I_h = \int |K_h(t)| \Omega(t) dt = \int_{|t| < \delta} + \int_{|t| \geq \delta} = A_{h,\delta} + B_{h,\delta} \tag{90}$$

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

$$A_{h,\delta} \leq \eta \int_{|t| < \delta} |K_h(t)| dt \leq \eta \|K\|_1, \quad \forall h > 0 \tag{91}$$

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  $I_h \rightarrow 0$  as  $h \rightarrow 0$  and the theorem follows.  $\blacksquare$

Another lemma necessary for the proof of Lemma 1 is the following.

*Lemma 4:* (A Multinomial distribution inequality)

Let  $N_1, \dots, N_k$  be a multinomial random vector with parameters  $n, p_1, \dots, p_k$ . Then

$$P \left( \sum_{i=1}^k \left| \frac{N_i}{n} - p_i \right| \geq \epsilon \right) \leq 2^{k+1} e^{-\frac{n\epsilon^2}{2}} \tag{92}$$

**Proof**

By Scheffe's theorem,

$$\sum_{i=1}^k \left| \frac{N_i}{n} - p_i \right| = 2 \sup_A \left| \frac{N(\mathbb{A})}{n} - P(\mathbb{A}) \right| \tag{93}$$

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

$$P \left( \sup_{\mathbb{A}} \left| \frac{N(\mathbb{A})}{n} - P(\mathbb{A}) \right| \geq \frac{\epsilon}{2} \right) \leq 2^k 2e^{-2n(\frac{\epsilon}{2})^2} \tag{94}$$

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

$$g_h(x) = E(f^n(x)) = \frac{1}{nh^d} \sum_{i=1}^n \int K\left(\frac{x-y}{h}\right) f_i(y) dy \quad (95)$$

*Proof:* [Proof of Lemma 1]

Let  $g_h$  be defined as in (95). By Theorem 1, it is enough to show that  $\int |f^n(x) - g_h(x)| dx \rightarrow 0$  exponentially. Let  $\mu_n$  be the empirical probability measure for  $X_1, X_2, \dots, X_n$  and note that

$$f^n(x) = \frac{1}{h^d} \int K\left(\frac{x-y}{h}\right) \mu_n(dy) \quad (96)$$

$$(97)$$

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

$$K^*(x) = \sum_{i=1}^N a_i I_{A_i}(x) \quad (98)$$

satisfies:  $|K^*| \leq M$ ,  $K^* = 0$  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

$$\begin{aligned} \int |f^n(x) - g_h(x)| dx &\leq \int |f^n(x) - f^{n*}(x)| dx + \int |f^{n*}(x) - g_h^*(x)| dx + \int |g_h^*(x) - g_h(x)| dx \\ &\leq \int \frac{1}{h^d} \int \left| K^*\left(\frac{x-y}{h}\right) - K\left(\frac{x-y}{h}\right) \right| \mu_n(dy) dx \\ &\quad + \int \frac{1}{nh^d} \sum_{i=1}^n \int \left| K^*\left(\frac{x-y}{h}\right) - K\left(\frac{x-y}{h}\right) \right| f_i(y) dy dx \\ &\quad + \int |f^{n*}(x) - g_h^*(x)| dx \\ &\leq 2\epsilon + \int |f^{n*}(x) - g_h^*(x)| dx \end{aligned}$$

by a double change of integral. But, if  $\mu$  is the probability measure for  $f$ ,

$$\begin{aligned} \int |f^{n*}(x) - g_h^*(x)| dx &\leq \sum_{i=1}^N |a_i| \int \left| \frac{1}{nh^d} \sum_{j=1}^n \int_{x-hA_i} f_j(y) dy - \frac{1}{h^d} \int_{x-hA_i} \mu_n(dy) \right| dx \\ &\leq \frac{1}{h^d} \sum_{i=1}^N |a_i| \int \left| \frac{1}{n} \sum_{j=1}^n \mu_j(x-hA_i) - \mu_n(x-hA_i) \right| dx \quad (99) \end{aligned}$$

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

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

Choose an  $\mathbb{A}$ , 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 II. Let

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

and

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

Define

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

Clearly, for any  $n$

$$\begin{aligned} \int \left| \frac{1}{n} \sum_{j=1}^n \mu_j(x - h\mathbb{A}) - \mu_n(x - h\mathbb{A}) \right| dx &\leq \int \sum_{\substack{B \in \Pi \\ B \subseteq x - h\mathbb{A}}} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| dx \\ &\quad + \int \left( \frac{1}{n} \sum_{j=1}^n \mu_j + \mu_n \right) (C_x^*) \end{aligned} \quad (100)$$

The last term in (100) equals

$$2\lambda(h(\mathbb{A} - \mathbb{A}^*)) = 2h^d \lambda(\mathbb{A} - \mathbb{A}^*) \quad (101)$$

$$= 2h^d \left( \prod_{i=1}^d a_i - \prod_{i=1}^d \left( a_i - \frac{2}{N} \right) \right) \quad (102)$$

where  $\lambda$  is the Lebesgue measure. Now, putting (102), (100) and (99) together, we get

$$\begin{aligned} &\int |f^n(x) - g_h(x)| dx \leq 2\epsilon + \int |f^{n*}(x) - g_h^*(x)| \\ &\leq 2\epsilon + \sum_{i=1}^N |a_i| \frac{1}{h^d} \int \sum_{\substack{B \in \Pi \\ B \subseteq x - hA_i}} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| dx + \sum_{i=1}^N |a_i| \frac{2}{h^d} h^d \lambda(A_i - A_i^*) \\ &\leq 2\epsilon + \frac{1}{h^d} \sum_{i=1}^N |a_i| \sum_{B \in \Pi} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| \int_{B \subseteq x - hA_i} dx + \sum_{i=1}^N |a_i| \frac{2}{h^d} h^d \lambda(A_i - A_i^*) \\ &\leq 2\epsilon + \frac{1}{h^d} \sum_{i=1}^N |a_i| \sum_{B \in \Pi} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| h^d \lambda(A_i) + \sum_{i=1}^N |a_i| \frac{2}{h^d} h^d \lambda(A_i - A_i^*) \\ &\leq 2\epsilon + \left( \sum_{i=1}^N |a_i| \lambda(A_i) \right) \sum_{B \in \Pi} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| + 2 \sum_{i=1}^N |a_i| \lambda(A_i - A_i^*) \end{aligned} \quad (103)$$

The third term on the right hand side can be made smaller than  $\epsilon$  by choosing  $N$  large enough ( $A_i^* \rightarrow A_i, \forall i$  as  $N \rightarrow \infty$ ). 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

$$\begin{aligned} \int |f^n(x) - g_h(x)| dx &\leq 3\epsilon + (1 + \epsilon) \sum_{B \in \Pi} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| \\ &\leq 5\epsilon + \sum_{B \in \Pi} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| \end{aligned} \quad (104)$$

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

$$\begin{aligned} \sum_{B \in \Pi} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| &\leq \sum_{B \in \Pi_r} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| + \frac{1}{n} \sum_{j=1}^n \mu_j(T) + \mu_n(T) \\ &\leq \sum_{B \in \Pi_r} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| + 2 \frac{1}{n} \sum_{j=1}^n \mu_j(T) + \left| \frac{1}{n} \sum_{j=1}^n \mu_j(T) - \mu_n(T) \right| \\ &\leq \sum_{B \in \Pi_r^*} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| + 2 \frac{1}{n} \sum_{j=1}^n \mu_j(T) \\ &\leq \sum_{B \in \Pi_r^*} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| + 2 \sup_{i \in \mathcal{I}} \mu_i(T) \end{aligned} \quad (105)$$

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

$$\int |f^n(x) - g_h(x)| dx \leq 6\epsilon + \sum_{B \in \Pi_r^*} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| \quad (106)$$

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

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

$$\begin{aligned} P \left( \int |f^n - g_h| > \delta \right) &\leq P \left( \sum_{B \in \Pi_r^*} \left| \frac{1}{n} \sum_{j=1}^n \mu_j(B) - \mu_n(B) \right| > \delta - 6\epsilon \right) \\ &\leq 2^{2 + (2 + \frac{2rN}{h})^d} e^{-\frac{1}{2}n(\delta - 6\epsilon)^2} \end{aligned} \quad (107)$$

$$\leq e^{-(1-\rho)\frac{n\delta^2}{2}}, n \geq n_0(\rho, \delta, K, \Upsilon, h) \quad (108)$$

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

$$n > \frac{16 + 4^{d+1}}{\rho\delta^2} \quad (109)$$

$$nh^d > n_0^d(\mathcal{C}, \rho, \delta, K, d) = \frac{42^d(2r(\mathcal{C}, K)N)^d}{\rho\delta^2} \quad (110)$$

if we pick

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

For the symbol-by-symbol case,  $d = 1$  and (110) becomes

$$n > \frac{16 + 4^{d+1}}{\rho\delta^2} \quad (111)$$

$$nh^d > n_0(\mathcal{C}, \rho, \delta, K) = \frac{16r(\mathcal{C}, K)N}{\rho\delta^2} \quad (112)$$

■

### APPENDIX III

#### 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

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

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  $\hat{F}_{x^n}$ . From [11, Theorem 11.7.1],  $\rho(P_n, Q_n) \rightarrow 0 \Rightarrow \beta(P_n, Q_n)$  then by definition of the  $\beta$ -metric, we have

$$\lim_{n \rightarrow \infty} \left| \int f d(P_n - Q_n) \right| = 0 \quad \forall \|f\|_{BL} \leq 1 \quad (113)$$

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

$$\lim_{n \rightarrow \infty} \left| \int f(x, y) d(P_n - Q_n) \right| = 0 \quad \forall y \text{ and } f \in \mathcal{S}^{[a,b] \times \mathbb{R}} \quad (114)$$

where  $\mathcal{S}_M^{[a,b] \times \mathbb{R}} := \{f : [a, b] \times \mathbb{R} \rightarrow \mathbb{R}, \|f(y)\|_{BL} < M \forall y\}$  for some  $M < \infty$  and

$$\|f(y)\|_L := \sup_{x \neq z} \frac{|f(x, y) - f(z, y)|}{|x - z|} \quad (115)$$

$$\|f(y)\|_\infty := \sup_x f(y, x) \quad (116)$$

$$\|f(y)\|_{BL} := \|f(y)\|_L + \|f(y)\|_\infty \quad (117)$$

Hence, for a channel with conditional densities,  $\{f_{Y|x}\}_{x \in [a,b]} \in \mathcal{S}_M^{[a,b] \times \mathbb{R}}$ , we have

$$\left| \int f_{Y|x} dF_{x^n} - \int f_{Y|x} d\hat{F}_{x^n} \right| \rightarrow 0 \quad \forall y \in \mathbb{R} \quad (118)$$

and by dominated convergence theorem,

$$\int \left| \int f_{Y|x} dF_{x^n} - \int f_{Y|x} d\hat{F}_{x^n} \right| dy \rightarrow 0 \quad (119)$$

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

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

$$f_{Y^n}(y) = [F_{x^n} \otimes \mathcal{C}]_Y = \int f_{Y|x} dF_{x^n}(x) \quad (120)$$

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 [11, 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,  $([\mathcal{F}^{[a,b]} \otimes \mathcal{C}], d)$ , we can apply the continuous mapping theorem [30] to get continuity in the inverse mapping too. This gives the desired result that as  $d([F_{x^n} \otimes \mathcal{C}]_Y, [\hat{F}_{x^n} \otimes \mathcal{C}]_Y) \rightarrow 0$ , we have  $\beta(P_n, Q_n) \rightarrow 0$  and  $\rho(P_n, Q_n) \rightarrow 0$ . Finally using the fact [11],  $\lambda \leq \rho$ ,  $\lambda(F_{x^n}, \hat{F}_{x^n}) \rightarrow 0$ . ■

#### APPENDIX IV PROOF OF LEMMA 2

*Proof:*

Consider  $f \in \mathcal{C}_b([a, b])$ , where  $\mathcal{C}_b$  denotes the set of all continuous bounded functions,  $f : [a, b] \rightarrow \mathbb{R}$ . For any  $F \in \mathcal{F}^{[a,b]}$  and  $P^\Delta$  that is constructed using (31)

$$\begin{aligned} & \left| \int f dF(x) - \int f P^\Delta(dx) \right| \\ &= \left| \int f (dF(x) - P^\Delta(dx)) \right| \\ &= \left| \int f dF(x) - \sum_{i=1}^N f(a_i) P(a_i) \right| \\ &\leq \left| \sum_{i=0}^{N-1} \int_{a_i}^{a_{i+1}} (f(a_i) + \omega_f(\Delta)) dF(x) - \sum_{i=1}^N f(a_i) P(a_i) \right| \\ &= \left| \sum_{i=0}^{N-1} (f(a_i) + \omega_f(\Delta)) P(a_i) - \sum_{i=1}^N f(a_i) P(a_i) \right| \\ &= \left| \omega_f(\Delta) \sum_{i=1}^N P(a_i) \right| \\ &= \omega_f(\Delta) \end{aligned} \quad (121)$$

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,

$$\lim_{\Delta \rightarrow 0} |P^\Delta f - P f| = \left| \lim_{\Delta \rightarrow 0} \int f (dF(x) - P^\Delta(dx)) \right| \quad (122)$$

$$= \lim_{\Delta \rightarrow 0} \omega_f(\Delta) \quad (123)$$

$$= 0, \quad \forall f \in \mathcal{C}_b([a, b]) \quad (124)$$

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

APPENDIX V  
PROOF OF THEOREM 4

Using the definition of the Lipschitz norm of the loss function,  $\Lambda$ , and the channel continuity function,  $\xi_\Delta$ , 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  $g : \mathbb{R} \rightarrow [a, b]$  and a bounded Lipschitz loss function with  $E_{f_{Y|u}} \Lambda(u, g(Y)) < \infty, \forall u$ ,

$$\begin{aligned} & |E_{F \otimes C} \Lambda(U_0, g(Y)) - E_{\hat{F} \otimes C} \Lambda(U_0, g(Y))| \\ & \leq (\|\Lambda\|_L + \Lambda_{\max} \|\Xi\|_L + (b-a) \|\Lambda\|_L \|\Xi\|_L + \Lambda_{\max}) \beta(P, \hat{P}) \end{aligned} \quad (125)$$

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  $g : \mathbb{R} \rightarrow [a, b]$  and a continuous bounded loss function with  $E_{f_{Y|u}} \Lambda(u, g(Y)) < \infty, \forall u$ ,

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

where  $\lambda(\Delta)$  is the global modulus of continuity of the loss function  $\Lambda$  as defined in equation (4) and  $\xi_\Delta$  is as defined in (38).

The proofs for Lemmas 5 and 6 are discussed in the following section, Appendix VI

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

$$Pr \left( \left| \frac{1}{n} \sum_{i=1}^n \Lambda(x_i, g(Y_i)) - E_{F_{x^n} \otimes C} \Lambda(U, g(Y)) \right| > \epsilon \right) \leq 2 \exp(-G(\epsilon, \Lambda_{\max})n) \quad (126)$$

*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 (126) 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 [8] as in [5] leading to (126). ■

In preparation of the proof of Theorem 4, we need also the following two Lemmas

*Lemma 8:*  $d(f_Y^n, [\hat{F}_{x^n} \otimes \mathcal{C}]_Y) \rightarrow 0$  a.s.

*Proof:* By definition,

$$0 \leq d(f_Y^n, [\hat{F}_{x^n} \otimes \mathcal{C}]_Y) \leq d(f_Y^n, [F_{x^n} \otimes \mathcal{C}]_Y), \forall n$$

Taking limit  $n \rightarrow \infty$  in the inequality of (127), we get

$$0 \leq \lim_{n \rightarrow \infty} d(f_Y^n, [\hat{F}_{x^n} \otimes \mathcal{C}]_Y) \leq \lim_{n \rightarrow \infty} d(f_Y^n, [F_{x^n} \otimes \mathcal{C}]_Y) = 0 \text{ a.s.}$$

where the second part of the inequality in (127) follows from Theorem 1. ■

*Lemma 9:*  $d([F_{x^n} \otimes \mathcal{C}]_Y, [\hat{F}_{x^n} \otimes \mathcal{C}]_Y) \rightarrow 0$  a.s.

*Proof:*

$$0 \leq d([F_{x^n} \otimes \mathcal{C}]_Y, [\hat{F}_{x^n} \otimes \mathcal{C}]_Y) \leq d([F_{x^n} \otimes \mathcal{C}]_Y, f_Y^n) + d(f_Y^n, [\hat{F}_{x^n} \otimes \mathcal{C}]_Y)$$

We have already seen  $d([F_{x^n} \otimes \mathcal{C}]_Y, f_Y^n) \rightarrow 0$  a.s and by Lemma 8,

$$d(f_Y^n, [\hat{F}_{x^n} \otimes \mathcal{C}]_Y) \rightarrow 0 \text{ a.s.}$$

Whence,

$$d([F_{x^n} \otimes \mathcal{C}]_Y, [\hat{F}_{x^n} \otimes \mathcal{C}]_Y) \rightarrow 0 \text{ a.s.} \quad \blacksquare$$

We are now ready for the proof of Theorem 4, *Proof:* [Proof of Theorem 4] We fix  $n \geq 1$ ,  $x^n \in [a, b]^n$ ,

$$\begin{aligned} & \left| E_{\hat{P}_{x^n}^{\delta, \Delta} [Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) - E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| \leq \\ & \left| E_{\hat{P}_{x^n}^{\delta, \Delta} [Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) - E_{\hat{F}_{x^n} [Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| + \\ & \left| E_{\hat{F}_{x^n} [Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) - E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| \end{aligned} \quad (127)$$

Hence,

$$\begin{aligned} & Pr \left( \sup_{g: \mathbb{R} \rightarrow [a, b]} \left| E_{\hat{P}_{x^n}^{\delta, \Delta} [Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) - E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > \epsilon + \delta \Lambda_{\max} + \xi_{\Delta} \Lambda_{\max} + \right. \\ & \left. \lambda(\Delta)(1 + \xi_{\Delta}) \right) \leq Pr \left( \left| E_{\hat{F}_{x^n} [Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) - E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > \epsilon \right) + \end{aligned} \quad (128)$$

$$Pr \left( \left| E_{\hat{F}_{x^n} [Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) - E_{\hat{P}_{x^n}^{\delta, \Delta} [Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > \delta \Lambda_{\max} + \xi_{\Delta} \Lambda_{\max} + \lambda(\Delta)(1 + \xi_{\Delta}) \right) \quad (129)$$

Now,

$$\begin{aligned}
& Pr \left( \left| E_{\hat{F}_{x^n}[Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) - E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > \epsilon \right) \leq \\
& Pr \left( (\| \Lambda \|_L + \Lambda_{\max} \| \Xi \|_L + (b-a) \| \Lambda \|_L \| \Xi \|_L + \Lambda_{\max}) \beta \left( P_{x^n}, \hat{P}_{x^n} \right) > \epsilon \right) \quad (130) \\
& \leq Pr \left( (\| \Lambda \|_L + \Lambda_{\max} \| \Xi \|_L + (b-a) \| \Lambda \|_L \| \Xi \|_L + \Lambda_{\max}) d \left( F_{x^n} \otimes \mathcal{C}, \hat{F}_{x^n} \otimes \mathcal{C} \right) > \epsilon \right) \\
& \leq e^{-(1-\rho) \frac{n\gamma^2}{2}}, \\
& \text{for all } nh_n > n_0(\mathcal{C}, \rho, \delta, K) \quad (131)
\end{aligned}$$

where  $\mathcal{C}$  is the family of channel densities  $\{f_{Y|x}\}$ . The inequality in (130) is due to Lemma 5, while the first inequality in (131) is by application of Theorem 2 and the second inequality is due to Lemma 9 and Theorem 1. Finally, application of Lemma 6 to (129) yields

$$\begin{aligned}
& Pr \left( \sup_{g: \mathbb{R} \rightarrow [a, b]} \left| E_{\hat{P}_{x^n}^{\delta, \Delta}[Y^n] \otimes \mathcal{C}} \Lambda(U, g(Y)) - E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > \epsilon + \delta \Lambda_{\max} + \xi_{\Delta} \Lambda_{\max} + \right. \\
& \left. \lambda(\Delta)(1 + \xi_{\Delta}) \right) \leq e^{-(1-\rho) \frac{n\gamma^2}{2}}, \text{ for all } n > n_0(\mathcal{C}, \rho, \delta, K) \quad (132)
\end{aligned}$$

Combining (132) with Lemma 7 gives

$$\begin{aligned}
& Pr \left( \left| \frac{1}{n} \sum_{i=1}^n \Lambda(x_i, g(Y_i)) - E_{\hat{P}_{x^n}^{\delta, \Delta} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > 2\epsilon + 2\delta \Lambda_{\max} + \xi_{\Delta} \Lambda_{\max} + \lambda(\Delta)(1 + \xi_{\Delta}) \right) \\
& \leq 2e^{-G(\epsilon + \delta \Lambda_{\max}, \Lambda_{\max})n} + e^{-(1-\rho) \frac{n\gamma^2}{2}}, \text{ for all } nh_n > n_0(\mathcal{C}, \rho, \delta, K) \quad (133)
\end{aligned}$$

By the union bound, (133) guarantees that for any class  $\mathcal{G}$

$$\begin{aligned}
& Pr \left( \max_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \Lambda(x_i, g(Y_i)) - E_{\hat{P}_{x^n}^{\delta, \Delta} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > 2\epsilon + 2\delta \Lambda_{\max} + C_{\Delta} \Lambda_{\max} \right. \\
& \left. + \lambda(\Delta)(1 + \xi_{\Delta}) \right) \leq |\mathcal{G}| \left[ 2e^{-G(\epsilon + \delta \Lambda_{\max}, \Lambda_{\max})n} + e^{-(1-\rho) \frac{n\gamma^2}{2}} \right] \quad (134)
\end{aligned}$$

Consequently,

$$\begin{aligned}
& Pr \left( \left| L_{\tilde{X}^{n, \delta, \Delta}}(x^n, Y^n) - \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{\hat{P}_{x^n}^{\delta, \Delta} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > 2\epsilon + 2\delta \Lambda_{\max} + C_{\Delta} \Lambda_{\max} \right. \\
& \left. + \lambda(\Delta)(1 + \xi_{\Delta}) \right) = Pr \left( \left| \frac{1}{n} \sum_{i=1}^n \Lambda(x_i, g_{opt}[\hat{P}_{x^n}^{\delta, \Delta}[Y^n]](Y_i)) - E_{\hat{P}_{x^n}^{\delta, \Delta} \otimes \mathcal{C}} \Lambda(U, g_{opt}[\hat{P}_{x^n}^{\delta, \Delta}[Y^n]](Y)) \right| \right. \\
& \left. > 2\epsilon + 2\delta \Lambda_{\max} + C_{\Delta} \Lambda_{\max} + \lambda(\Delta)(1 + \xi_{\Delta}) \right) \\
& \leq Pr \left( \max_{g \in \mathcal{G}_{\delta, \Delta}} \left| \frac{1}{n} \sum_{i=1}^n \Lambda(x_i, g(Y_i)) - E_{\hat{P}_{x^n}^{\delta, \Delta} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > 2\epsilon + 2\delta \Lambda_{\max} + C_{\Delta} \Lambda_{\max} \right. \\
& \left. + \lambda(\Delta)(1 + \xi_{\Delta}) \right) \leq |\mathcal{G}_{\delta, \Delta}| \left[ 2e^{-G(\epsilon + \delta \Lambda_{\max}, \Lambda_{\max})n} + e^{-(1-\rho) \frac{n\gamma^2}{2}} \right] \quad (135)
\end{aligned}$$

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

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

The first inequality follows by the fact that  $\hat{P}_{x^n}^{\delta, \Delta}[Y^n] \in \mathcal{F}_{\delta, \Delta}$  and therefore  $g_{opt}[\hat{P}_{x^n}^{\delta, \Delta}[Y^n]] \in \mathcal{G}_{\delta, \Delta}$ , and finally the last inequality follows from (134). It also follows, from (132), that

$$Pr \left( \left| \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{\hat{P}_{x^n}^{\delta, \Delta} \otimes \mathcal{C}} \Lambda(U, g(Y)) - \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > \epsilon + \delta \Lambda_{\max} + \xi_{\Delta} \Lambda_{\max} + \lambda(\Delta)(1 + \xi_{\Delta}) \right) \leq e^{-(1-\rho) \frac{n\gamma^2}{2}} \quad (136)$$

Combining (135) and (136) gives

$$Pr \left( \left| L_{\tilde{X}^n, \delta, \Delta}(x^n, Y^n) - \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| > 3\epsilon + 3\delta \Lambda_{\max} + 2\xi_{\Delta} \Lambda_{\max} + 2\lambda(\Delta)(1 + \xi_{\Delta}) \right) \leq |\mathcal{G}_{\delta, \Delta}| \left[ 2e^{-G(\epsilon + \delta \Lambda_{\max}, \Lambda_{\max})n} + e^{-(1-\rho) \frac{n\gamma^2}{2}} \right] + e^{-(1-\rho) \frac{n\gamma^2}{2}} \quad (137)$$

On the other hand, letting  $\hat{P}_{x^n}^{\delta, \Delta}$  denote the element in  $\mathcal{F}_{\delta, \Delta}$  closest (under the Prohorov metric of the corresponding measures) to  $F_{x^n}$ ,

$$\begin{aligned} & \left| D_0(x^n) - \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| \\ &= \left| \min_{F \in \mathcal{F}_n^{[a, b]}} E_{F \otimes \mathcal{C}} \Lambda(U, g_{opt}[F](Y)) - \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| \end{aligned} \quad (138)$$

$$\leq \left| \min_{F \in \mathcal{F}_n^{[a, b]}} E_{\hat{F}_{x^n}^{\delta, \Delta} \otimes \mathcal{C}} \Lambda(U, g_{opt}[F](Y)) - \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| + \Lambda_{\max} \delta + \xi_{\Delta} \Lambda_{\max} + \lambda(\Delta)(1 + \xi_{\Delta}) \quad (139)$$

$$= \left| \min_{P \in \mathcal{F}^{\delta, \Delta}} E_{\hat{P}_{x^n}^{\delta, \Delta} \otimes \mathcal{C}} \Lambda(U, g_{opt}[P](Y)) - \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| + \Lambda_{\max} \delta + \xi_{\Delta} \Lambda_{\max} + \lambda(\Delta)(1 + \xi_{\Delta}) \quad (140)$$

$$= \left| \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{\hat{F}_{x^n}^{\delta, \Delta} \otimes \mathcal{C}} \Lambda(U, g(Y)) - \min_{g \in \mathcal{G}_{\delta, \Delta}} E_{F_{x^n} \otimes \mathcal{C}} \Lambda(U, g(Y)) \right| + \Lambda_{\max} \delta + \xi_{\Delta} \Lambda_{\max} + \lambda(\Delta)(1 + \xi_{\Delta}) \quad (141)$$

$$\leq 2(\Lambda_{\max} \delta + \xi_{\Delta} \Lambda_{\max} + \lambda(\Delta)(1 + \xi_{\Delta})) \quad (142)$$

where (139) and (142) follow from Lemma 6, and (140) follows from the fact that the achiever of the minimum in the first term of (139) is  $F_{x^n}^{\delta, \Delta}$  which, by definition, is a member of  $\mathcal{F}_{\delta, \Delta}$ . Finally, combining (136) with (142) gives

$$\begin{aligned} Pr \left( |L_{\tilde{X}^n, \delta, \Delta}(x^n, Y^n) - D_0(x^n)| > 3\epsilon + 5\delta \Lambda_{\max} + 4\xi_{\Delta} \Lambda_{\max} + 4\lambda(\Delta)(1 + \xi_{\Delta}) \right) \\ \leq |\mathcal{G}_{\delta, \Delta}| \left[ e^{-G(\epsilon + \delta \Lambda_{\max}, \Lambda_{\max})n} + e^{-(1-\rho) \frac{n\gamma^2}{2}} \right] + e^{-(1-\rho) \frac{n\gamma^2}{2}} \end{aligned} \quad (143)$$

for all  $nh_n > n_0(\mathcal{C}, \rho, \delta, K)$

From the definition of  $\mathcal{G}_{\delta,\Delta}$ , it is clear that  $|\mathcal{G}_{\delta,\Delta}| \leq \left[\frac{1}{\delta} + 1\right]^\Delta$ . Hence,

$$\begin{aligned} Pr(|L_{\bar{X}^n, \delta, \Delta}(x^n, Y^n) - D_0(x^n)| > 3\epsilon + 5\delta\Lambda_{\max} + 4\xi_\Delta\Lambda_{\max} + 4\lambda(\Delta)(1 + \xi_\Delta)) \\ \leq \left[1 + \frac{1}{\delta}\right]^\Delta \left[ e^{-G(\epsilon + \delta\Lambda_{\max}, \Lambda_{\max})n} + e^{-(1-\rho)\frac{n\gamma^2}{2}} \right] + e^{-(1-\rho)\frac{n\gamma^2}{2}} \end{aligned} \quad (144)$$

for all  $nh_n > n_0(\mathcal{C}, \rho, \delta, K)$

■

## APPENDIX VI

### PROOF OF LEMMAS 5 AND 6

We need the following proposition for the proof of Lemma 5

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

*Proof:* Let  $\Delta = |x - x'|$ ,

$$\begin{aligned} A(x) - A(x') &= \int \Lambda(x, g(y)) f_{Y|x}(y) dy - \int \Lambda(x', g(y)) f_{Y|x'}(y) dy \\ &\leq \int (\Lambda(x', g(y)) + \lambda(\Delta, x)) f_{Y|x}(y) dy - \int (\Lambda(x', g(y))) f_{Y|x'}(y) dy \\ &\leq \int (\Lambda(x', g(y)) + \lambda(\Delta, x)) (f_{Y|x'}(y) + \varepsilon_\Delta(y)) dy - \int (\Lambda(x', g(y))) f_{Y|x'}(y) dy \\ &\leq \lambda(\Delta, x) + \Lambda_{\max}\xi_\Delta + \lambda(\Delta, x)\xi_\Delta \end{aligned}$$

Also,

$$\begin{aligned} A(x) - A(x') &= \int \Lambda(x, g(y)) f_{Y|x}(y) dy - \int \Lambda(x', g(y)) f_{Y|x'}(y) dy \\ &\geq \int (\Lambda(x', g(y)) - \lambda(\Delta, x)) f_{Y|x}(y) dy - \int (\Lambda(x', g(y))) f_{Y|x'}(y) dy \\ &\geq \int (\Lambda(x', g(y)) - \lambda(\Delta, x)) (f_{Y|x'}(y) - \varepsilon_\Delta(y)) dy - \int (\Lambda(x', g(y))) f_{Y|x'}(y) dy \\ &\geq -\lambda(\Delta, x) - \Lambda_{\max}\xi_\Delta + \lambda(\Delta, x)\xi_\Delta \\ &\geq -\lambda(\Delta, x) - \Lambda_{\max}\xi_\Delta - \lambda(\Delta, x)\xi_\Delta \end{aligned}$$

Hence,  $|A(x) - A(x')| \leq \lambda(\Delta) + \Lambda_{\max}\xi_\Delta + \lambda(\Delta)$ .

The assumption of Lipschitz continuity (condition, C6) of the channel guarantees  $\lim_{\Delta \rightarrow 0} \xi_\Delta = 0$ . With this and the fact that  $\lim_{\Delta \rightarrow 0} \lambda(\Delta) = 0$ , we have

$$\lim_{\substack{|x-x'| < \Delta \\ \Delta \rightarrow 0}} |A(x) - A(x')| = 0$$

■

Moreover,

$$\begin{aligned}
\| A \|_L &= \sup_{0 < \Delta < (b-a)} \sup_{\substack{x \neq x' \\ |x-x'|=\Delta}} \frac{|A(x) - A(x')|}{|x - x'|} \\
&\leq \sup_{0 < \Delta < (b-a)} \frac{\lambda(\Delta) + \Lambda_{\max} \xi_\Delta + \lambda(\Delta) \xi_\Delta}{\Delta} \\
&\leq \| \Lambda \|_L + \Lambda_{\max} \| \Xi \|_L + (b-a) \| \Lambda \|_L \| \Xi \|_L
\end{aligned} \tag{145}$$

Hence,

$$\begin{aligned}
\| A \|_{BL} &= \| A \|_L + \| A \|_\infty \\
&\leq \| \Lambda \|_L + \Lambda_{\max} \| \Xi \|_L + (b-a) \| \Lambda \|_L \| \Xi \|_L + \Lambda_{\max}
\end{aligned} \tag{146}$$

*Proof:* [Proof of Lemma 5]

$$\begin{aligned}
&|E_{F \otimes C} \Lambda(U_0, g(Y)) - E_{\hat{F} \otimes C} \Lambda(U_0, g(Y))| \\
&= \left| \int dF(x) \left( \int \Lambda(x, g(y)) f_{Y|x}(y) dy \right) \right. \\
&\quad \left. - \int d\hat{F}(x) \left( \int \Lambda(x, g(y)) f_{Y|x}(y) dy \right) \right| \\
&= \left| \int dF(x) A(x) - \int d\hat{F}(x) A(x) \right| \\
&= \left| \int A(x) d(F - \hat{F})(x) \right| \\
&\leq \| A \|_{BL} \beta(P, \hat{P}) \\
&\leq (\| \Lambda \|_L + \Lambda_{\max} \| \Xi \|_L + (b-a) \| \Lambda \|_L \| \Xi \|_L + \Lambda_{\max}) \beta(P, \hat{P})
\end{aligned} \tag{147}$$

where, (147) follows from the fact that  $A(x)$  is a bounded Lipschitz function as shown in Proposition 1. Hence, as  $\beta(P, \hat{P}) \rightarrow 0$  we have  $|E_{F \otimes C} \Lambda(U_0, g(Y)) - E_{\hat{F} \otimes C} \Lambda(U_0, g(Y))| \rightarrow 0$ . ■

*Proof:* [Proof of Lemma 6]

$$\begin{aligned}
&|E_{P^\Delta \otimes C} \Lambda(U_0, g(Y)) - E_{F \otimes C} \Lambda(U_0, g(Y))| \\
&= \left| \sum_{i=1}^{N(\Delta)} \int_{a_{i-1}}^{a_i} dF(u') \left( \int \Lambda(u', g(y)) f_{Y|X=u'}(y) dy \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y) dy \right) \right| \\
&= \left| \sum_{i=1}^{N(\Delta)} \int dy \left( \int_{a_{i-1}}^{a_i} f_{Y|X=u'}(y) dF(u') \Lambda(u', g(y)) \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y) dy \right) \right|
\end{aligned} \tag{148}$$

Equality in (148) is due to application of Fubini's theorem. Hence,

$$\begin{aligned}
& |E_{P^\Delta \otimes C} \Lambda(U_0, g(Y)) - E_{F \otimes C} \Lambda(U_0, g(Y))| \\
& < \left| \sum_{i=1}^{N(\Delta)} \int dy \left( \int_{a_{i-1}}^{a_i} f_{Y|X=u'}(y) dF(u') (\Lambda(a_i, g(y)) + \lambda(\Delta)) \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y) dy \right) \right| \\
& = \left| \sum_{i=1}^{N(\Delta)} \int dy (\Lambda(a_i, g(y)) + \lambda(\Delta)) \left( \int_{a_{i-1}}^{a_i} f_{Y|X=u'}(y) dF(u') \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y) dy \right) \right| \tag{149}
\end{aligned}$$

$$\begin{aligned}
& < \left| \sum_{i=1}^{N(\Delta)} \int dy (\Lambda(a_i, g(y)) + \lambda(\Delta)) (f_{Y|X=a_i}(y) + \varepsilon(y)) \left( \int_{a_{i-1}}^{a_i} dF(u') \right) - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y) dy \right) \right| \\
& < \left| \sum_{i=1}^{N(\Delta)} \left( \int_{a_{i-1}}^{a_i} dF(u') \right) \left[ \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y) dy + \int \varepsilon(y) \Lambda(a_i, g(y)) dy + \lambda(\Delta) \int f_{Y|X=a_i}(y) dy \right. \right. \\
& \quad \left. \left. + \lambda(\Delta) \int \varepsilon(y) dy - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y) dy \right) \right] \right| \tag{150}
\end{aligned}$$

$$\begin{aligned}
& < \left| \sum_{i=1}^{N(\Delta)} \left( \int_{a_{i-1}}^{a_i} dF(u') \right) \left[ \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y) dy \right. \right. \\
& \quad \left. \left. + \int \varepsilon(y) \Lambda(a_i, g(y)) dy + \lambda(\Delta) \int f_{Y|X=a_i}(y) dy + \lambda(\Delta) \int \varepsilon(y) dy \right. \right. \\
& \quad \left. \left. - \sum_{i=1}^{N(\Delta)} P^\Delta(a_i) \left( \int \Lambda(a_i, g(y)) f_{Y|X=a_i}(y) dy \right) \right] \right| \tag{151}
\end{aligned}$$

$$\begin{aligned}
& = \left| \sum_{i=1}^{N(\Delta)} \left( \int_{a_{i-1}}^{a_i} dF(u') \right) \left[ \int \varepsilon(y) \Lambda(a_i, g(y)) dy + \lambda(\Delta) + \lambda(\Delta) \xi_\Delta \right] \right| \\
& = \left| \sum_{i=1}^{N(\Delta)} (F(a_i) - F(a_{i-1})) \left[ \int \varepsilon(y) \Lambda(a_i, g(y)) dy + \lambda(\Delta) + \lambda(\Delta) \xi_\Delta \right] \right| \\
& \leq \int \sum_{i=1}^{N(\Delta)} \varepsilon(y) \Lambda(a_i, g(y)) P^\Delta(u_i) dy + (\lambda(\Delta) + \lambda(\Delta) \xi_\Delta) \\
& \leq \xi_\Delta \Lambda_{\max} + (\lambda(\Delta) + \lambda(\Delta) \xi_\Delta) \\
& = \xi_\Delta \Lambda_{\max} + \lambda(\Delta) (1 + \xi_\Delta)
\end{aligned}$$

Hence,

$$\lim_{\Delta \rightarrow 0} |E_{P^\Delta \otimes C} \Lambda(U_0, g(Y)) - E_{F \otimes C} \Lambda(U_0, g(Y))| = 0 \tag{152}$$

## APPENDIX VII

## PROOF OF THEOREM 6

In preparation of Theorem 6 we start by presenting the proof of Lemma 3 and Theorem 11 of Lemma 3]

*Proof:* [Proof

$$\begin{aligned} D_k(x^n) &= \min_g E \left[ \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(X_0, g(Y_{-k}^k)) \right] \\ &= \min_g \int \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(x_i, g(y_{i-k}^{i+k})) \prod_{l=i-k}^{i+k} f_{Y|X=x_l}(y_l) dy_l \end{aligned} \quad (153)$$

$$= \min_g \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \int \Lambda(x_i, g(y_{i-k}^{i+k})) \prod_{l=i-k}^{i+k} f_{Y|X=x_l}(y_l) dy_l \quad (154)$$

$$= \min_g \frac{1}{2k+1} \sum_{i=1}^{2k+1} \int \frac{1}{\frac{n-2k}{2k+1}} \sum_{j=0}^{\lceil \frac{n-2k-i-1}{2k+1} \rceil - 1} \Lambda(x_{j(2k+1)+k+1}, \quad (155)$$

$$g\left(y_{j(2k+1)+i}^{j(2k+1)+i+2k}\right) \prod_{l=j(2k+1)+i}^{j(2k+1)+i+2k} f_{Y|X=x_l}(y_l) dy_l$$

$$\geq \min_g \frac{1}{2k+1} \sum_{i=1}^{2k+1} \int \frac{1}{\lceil \frac{n-2k}{2k+1} \rceil} \sum_{j=0}^{\lceil \frac{n-2k-i-1}{2k+1} \rceil - 1} \Lambda(x_{j(2k+1)+k+1}, \quad (156)$$

$$g\left(y_{j(2k+1)+i}^{j(2k+1)+i+2k}\right) \prod_{l=j(2k+1)+i}^{j(2k+1)+i+2k} f_{Y|X=x_l}(y_l) dy_l$$

$$\geq \frac{1}{2k+1} \sum_{i=1}^{2k+1} \min_{g_i} \int \frac{1}{\lceil \frac{n-2k}{2k+1} \rceil} \sum_{j=0}^{\lceil \frac{n-2k-i-1}{2k+1} \rceil (2k+1)+k+1+i} \Lambda(x_i, \quad (157)$$

$$g_i\left(y_{j(2k+1)+i}^{j(2k+1)+i+2k}\right) \prod_{l=j(2k+1)+i}^{j(2k+1)+i+2k} f_{Y|X=x_l}(y_l) dy_l$$

$$= \frac{1}{2k+1} \sum_{i=1}^{2k+1} D_k(x^{n_i}) \quad (158)$$

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

*Proposition 2:*  $A(x) = \int \Lambda(x, g(y_{-k}^k)) \prod_{i=-k}^k f_{Y|x_i}(y_i) dy_{-k}^k$  is a bounded Lipschitz function for any measurable  $g : [a, b]^{2k+1} \rightarrow \mathbb{R}$ .

*Lemma 10:* For any  $F, \hat{F} \in \mathcal{F}^{[a,b],k}$ , measurable  $g : \mathbb{R}^{2k+1} \rightarrow [a, b]$  and a bounded Lipschitz loss function with  $E_{f_{Y|u}} \Lambda(u, g(Y_{-k}^k)) < \infty, \forall u$ ,

$$\begin{aligned} & |E_{F \otimes C} \Lambda(U_0, g(Y_{-k}^k)) - E_{\hat{F} \otimes C} \Lambda(U_0, g(Y_{-k}^k))| \\ & \leq (\| \Lambda \|_L + \Lambda_{\max} \| \Xi \|_L^k + (b-a) \| \Lambda \|_L \| \Xi \|_L^k + \Lambda_{\max}) \beta(P, \hat{P}) \end{aligned}$$

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

$\| \Xi \|_L^k$  is the  $k^{\text{th}}$  order Lipschitz norm of the channel.

$$\| \Xi \|_L^k = \sup_{0 < \Delta < (b-a)} \frac{\xi_{\Delta}^{2k+1}}{\Delta} \quad (159)$$

and  $\xi_{\Delta}$  is as defined in (38).

*Lemma 11:* 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  $E_{f_{Y|u}} \Lambda(u, g(Y_{-k}^k)) < \infty, \forall u$ ,

$$|E_{P^{\Delta,k} \otimes C} \Lambda(U_0, g(Y_{-k}^k)) - E_{F \otimes C} \Lambda(U_0, g(Y_{-k}^k))| \leq \xi_{\Delta}^{2k+1} \Lambda_{\max} + \lambda(\Delta) (1 + \xi_{\Delta}^{2k+1})$$

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  $k^{\text{th}}$ -order sliding window loss for that subsequence. We now, state the  $k^{\text{th}}$ -order equivalent of Theorem 4 for each subsequence.

*Theorem 11:* For all  $m \geq 1$ ,  $k \geq 1$ ,  $\epsilon > 0$ ,  $\rho \in (0, 1)$ ,  $\delta > 0$ ,  $\Delta > 0$ , and  $x^m \in [a, b]^{(2k+1)m}$

$$\begin{aligned} Pr(|L_{\bar{X}^m, \delta, \Delta, k}(x^m, Y^m) - D_k(x^m)| > 3\epsilon + 5\delta\Lambda_{\max} + 4\xi_{\Delta}^{2k+1}\Lambda_{\max} + 4\lambda(\Delta)(1 + \xi_{\Delta}^{2k+1})) \\ \leq |\mathcal{G}_{\delta, \Delta}^k| \left[ e^{-G(\epsilon + \delta\Lambda_{\max}, \Lambda_{\max})m} + e^{-(1-\rho)\frac{m\gamma_k^2}{2}} \right] + e^{-(1-\rho)\frac{m\gamma_k^2}{2}} \end{aligned} \quad (160)$$

for all  $mh_m^k > m_k(C, \rho, \delta, K)$

where,

$$\gamma_k = \frac{\epsilon}{(\| \Lambda \|_L + \Lambda_{\max} \| \Xi \|_L^k + (b-a) \| \Lambda \|_L \| \Xi \|_L^k + \Lambda_{\max})}$$

and  $G, \mathcal{G}_{\delta, \Delta}^k$  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 10, 11 and 7. ■

*Proof:* [Proof of Theorem 6]

$$L_{\tilde{X}^{n,\delta,\Delta,k}}(x^n, Y^n) - D_k(x^n) =$$

$$L_{\tilde{X}^{n,\delta,\Delta,k}}(x^n, Y^n) - \frac{1}{2k+1} \sum_{i=1}^{2k+1} D_k(x^{n_i}) + \frac{1}{2k+1} \sum_{i=1}^{2k+1} D_k(x^{n_i}) - D_k(x^n) \quad (161)$$

From Lemma 3, we have

$$\begin{aligned} L_{\tilde{X}^{n,\delta,\Delta,k}}(x^n, Y^n) - D_k(x^n) &\leq L_{\tilde{X}^{n,\delta,\Delta,k}}(x^n, Y^n) - \frac{1}{2k+1} \sum_{i=1}^{2k+1} D_k(x^{n_i}) \\ &= \frac{1}{2k+1} \sum_{i=1}^{2k+1} L_{\tilde{X}^{n_i,\delta,\Delta,k}}(x^{n_i}, Y^{n_i}) - \frac{1}{2k+1} \sum_{i=1}^{2k+1} D_k(x^{n_i}) \\ &\leq \frac{1}{2k+1} \sum_{i=1}^{2k+1} [|L_{\tilde{X}^{n_i,\delta,\Delta,k}}(x^{n_i}, Y^{n_i}) - D_k(x^{n_i})|] \end{aligned} \quad (162)$$

Hence,

$$\begin{aligned} &Pr(L_{\tilde{X}^{n,\delta,\Delta,k}}(x^n, Y^n) - D_k(x^n) > 3\epsilon + 5\delta\Lambda_{\max} + 4\xi_{\Delta}^{2k+1}\Lambda_{\max} + 4\lambda(\Delta)(1 + \xi_{\Delta}^{2k+1})) \\ &\leq Pr\left(\frac{1}{2k+1} \sum_{i=1}^{2k+1} |L_{\tilde{X}^{n_i,\delta,\Delta,k}}(x^{n_i}, Y^{n_i}) - D_k(x^{n_i})| > 3\epsilon + 5\delta\Lambda_{\max} + 4\xi_{\Delta}^{2k+1}\Lambda_{\max} + 4\lambda(\Delta)(1 + \xi_{\Delta}^{2k+1})\right) \\ &\leq \sum_{i=1}^{2k+1} Pr(|L_{\tilde{X}^{n_i,\delta,\Delta,k}}(x^{n_i}, Y^{n_i}) - D_k(x^{n_i})| > 3\epsilon + 5\delta\Lambda_{\max} + 4\xi_{\Delta}^{2k+1}\Lambda_{\max} + 4\lambda(\Delta)(1 + \xi_{\Delta}^{2k+1})) \\ &\leq (2k+1)|\mathcal{G}_{\delta,\Delta}^k| \left[ e^{-G(\epsilon+\delta\Lambda_{\max},\Lambda_{\max})\frac{(n-2k)}{2k+1}} + e^{-(1-\rho)\frac{(n-2k)\gamma_k^2}{2(2k+1)}} \right] + e^{-(1-\rho)\frac{(n-2k)\gamma_k^2}{2(2k+1)}} \end{aligned}$$

This is true by applying Theorem 11 to the  $2k+1$  subsequences of independent supersymbols with at most  $\frac{n-2k}{2k+1}$  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  $k^{\text{th}}$ -order probability mass functions,  $\hat{P}_{x^n}^{\delta,\Delta,k}$ , i.e.,  $|\mathcal{G}_{\delta,\Delta}^k| \leq \lceil \frac{1}{\delta} + 1 \rceil^{\Delta^{2k+1}}$ . ■

## APPENDIX VIII

### PROOF OF THEOREM 8

The following claim is necessary for the proof of Theorem 8.

*Claim 1:*

$$\lim_{k \rightarrow \infty} \min_g E\Lambda(X_0, g(Y_{-k}^k)) = \mathbb{D}(F_{\mathbf{X}}, \mathcal{C})$$

The claim results from the following lemma.

*Lemma 12:* • For  $k, l \geq 0$ ,  $EU(F_{X_0|Y_{-k}^l})$  is decreasing in both  $k$  and  $l$ .

- For any two unboundedly increasing sequences of positive integers  $\{k_n\}, \{l_n\}$ ,

$$\lim_{n \rightarrow \infty} EU \left( F_{X_0|Y_{-k_n}^{l_n}} \right) = EU \left( F_{X_0|Y_{-\infty}^{\infty}} \right) \quad (163)$$

Equipped with Lemma 12, the proof for Claim 1 is very similar to that of Claim 2 in [36] but we, nevertheless, present here for completeness.

#### A. Proof of Lemma 12

*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  $\alpha \in [0, 1]$ ,

$$\begin{aligned} \mathcal{U}(\alpha F + (1 - \alpha)G) &= \min_{\hat{x} \in [a, b]} \int_{x \in [a, b]} \Lambda(x, \hat{x}) d(\alpha F + (1 - \alpha)G)(x) \\ &= \alpha \min_{\hat{x} \in [a, b]} \int_{x \in [a, b]} [\Lambda(x, \hat{x}) dF(x) + (1 - \alpha) \Lambda(x, \hat{x}) dG(x)] \\ &\geq \alpha \min_{\hat{x} \in [a, b]} \int_{x \in [a, b]} \Lambda(x, \hat{x}) dF(x) + \\ &\quad (1 - \alpha) \min_{\hat{x} \in [a, b]} \int_{x \in [a, b]} \Lambda(x, \hat{x}) dG(x) \\ &= \alpha \mathcal{U}(F) + (1 - \alpha) \mathcal{U}(G) \end{aligned}$$

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  $EU \left( [F \otimes \mathcal{C}]_{X|Y_{-k}^l} \right)$  decreases with  $l$ , observe that

$$\begin{aligned}
EU \left( [F \otimes \mathcal{C}]_{X|Y_{-k}^{l+1}} \right) &= \int_{y_{-k}^{l+k+2}} \mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{-k}^{l+1}} \right) dF_{Y_{-k}^{l+1}} \\
&= \int_{y_{-k}^l} \left[ \int_{y_{l+1}} \mathcal{U} \left( [F \otimes \mathcal{C}]_{X|Y_{-k}^l, Y_{l+1}} \right) dF_{Y_{l+1}|Y_{-k}^l} \right] dF_{Y_{-k}^l} \\
&\leq \int_{y_{-k}^l} \mathcal{U} \left[ \int_{y_{l+1}} \left( [F \otimes \mathcal{C}]_{X|Y_{-k}^l, Y_{l+1}} \right) dF_{Y_{l+1}|Y_{-k}^l} \right] dF_{Y_{-k}^l} \\
&= \int_{y_{-k}^l} \mathcal{U} \left[ \int_{y_{l+1}} \left( \int_a^x \frac{f_{Y_{-k}^{l+1}|X=\alpha} dF_X(\alpha)}{f_{Y_{-k}^{l+1}}} \right) dF_{Y_{l+1}|Y_{-k}^l} \right] dF_{Y_{-k}^l} \\
&= \int_{y_{-k}^l} \mathcal{U} \left[ \int_{y_{l+1}} \left( \int_a^x \frac{f_{Y_{-k}^{l+1}|X=\alpha} dF_X(\alpha)}{f_{Y_{l+1}|Y_{-k}^l} f_{Y_{-k}^l}} \right) dF_{Y_{l+1}|Y_{-k}^l} \right] dF_{Y_{-k}^l} \\
&= \int_{y_{-k}^l} \mathcal{U} \left[ \int_{y_{l+1}} \left( \int_a^x \frac{f_{Y_{-k}^{l+1}|X=\alpha} dF_X(\alpha)}{f_{Y_{-k}^l}} \right) dy_{l+1} \right] dF_{Y_{-k}^l} \\
&= \int_{y_{-k}^l} \mathcal{U} \left[ \int_a^x \left( \int_{y_{l+1}} \frac{f_{Y_{-k}^{l+1}|X=\alpha} dF_X(\alpha)}{f_{Y_{-k}^l}} \right) dy_{l+1} \right] dF_{Y_{-k}^l} \\
&= \int_{y_{-k}^l} \mathcal{U} \left[ \int_a^x \left( \frac{f_{Y_{-k}^l|X=\alpha} dF_X(\alpha)}{f_{Y_{-k}^l}} \right) \right] dF_{Y_{-k}^l} \\
&= \int_{y_{-k}^l} \mathcal{U} [F \otimes \mathcal{C}]_{X|Y_{-k}^l} dF_{Y_{-k}^l} \\
&= EU \left( [F \otimes \mathcal{C}]_{X|Y_{-k}^l} \right)
\end{aligned} \tag{164}$$

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  $EU \left( [F \otimes \mathcal{C}]_{X|Y_{-k}^{l+1}} \right)$  decreases with  $k$  is established similarly, concluding the proof of the first item. For the second item, similar to the proof of Lemma 4 in [36], by the martingale convergence theorem, we have,  $F_{X|Y_{-k_n}^{l_n}} \rightarrow F_{X|Y_{-\infty}^{\infty}}$  a.s., implying  $F_{X|Y_{-k_n}^{l_n}} \xrightarrow{d} F_{X|Y_{-\infty}^{\infty}}$ . Using the convergence of random measures [20, Theorem 16.16], we have  $F_{X|Y_{-k_n}^{l_n}} f \xrightarrow{d} F_{X|Y_{-\infty}^{\infty}} f, \forall f \in C_K^+$ , the class of continuous positive valued functions with compact support. Here, the notation  $Ff = \int f dF$  for any measurable  $f$  and bona fide probability distribution function,  $F$ . In section IV, we have imposed the condition of continuity of the loss function,  $\Lambda$ , and since the input alphabet space is restricted to a closed compact interval  $[a, b]$ , we satisfy the condition,  $\Lambda \in C_K^+$ . Hence, we have,  $F_{X|Y_{-k_n}^{l_n}} \Lambda(\cdot, \hat{x}) \xrightarrow{d} F_{X|Y_{-\infty}^{\infty}} \Lambda(\cdot, \hat{x}), \forall \hat{x}$ . Since  $\Lambda(\cdot, \hat{x}) : [a, b] \times [a, b] \rightarrow \mathbb{R}^+$  is a continuous mapping, in  $\hat{x}$ ,  $\min_{\hat{x} \in [a, b]} \int \Lambda(x, \hat{x}) dF(x)$  is also a continuous mapping. Using the fact that  $\Lambda$  is a bounded mapping and the continuous mapping theorem [12],  $\mathcal{U} \left( F_{X|Y_{-k_n}^{l_n}} \right) \xrightarrow{d} \mathcal{U} \left( F_{X|Y_{-\infty}^{\infty}} \right)$  and  $EU \left( F_{X|Y_{-k_n}^{l_n}} \right) \rightarrow EU \left( F_{X|Y_{-\infty}^{\infty}} \right)$ . ■

### B. Proof of Claim 1

*Proof:* [Proof of Claim 1]

$$\begin{aligned}
\mathbb{D}(F_{X^n}, \mathcal{C}) &= \min_{\hat{X}^n \in \mathcal{D}_n} EL_{\hat{X}^n}(X^n, Y^n) = \frac{1}{n} \sum_{i=1}^n \min_{\hat{X}: \mathbb{R}^n \rightarrow [a, b]} E\Lambda(X_i, \hat{X}(Y^n)) \\
&= \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^n} \min_{\hat{x} \in [a, b]} E[\Lambda(X_i, \hat{x}) | Y^n = y^n] dF_{Y^n} \\
&= \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^n} \mathcal{U}(F_{X_i | Y^n = y^n}) dF_{Y^n} \\
&= \frac{1}{n} \sum_{i=1}^n EU(F_{X_i | Y^n = y^n}) = \frac{1}{n} \sum_{i=1}^n EU(F_{X_0 | Z_{1-i}^{n-i}}) \tag{165}
\end{aligned}$$

where the last equality follows by stationarity. Since by Lemma 12,  $EU(F_{X_0 | Y_{1-i}^{n-i}}) \geq EU(F_{X_0 | Y_{-\infty}^{\infty}})$ , it follows from (165) that  $\mathbb{D}(F_{X^n}, \mathcal{C}) \geq EU(F_{X_0 | Y_{-\infty}^{\infty}})$  for all  $n$  and, therefore,  $\mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \geq EU(F_{X_0 | Y_{-\infty}^{\infty}})$ . On the other hand, for any  $k$ ,  $0 \leq k \leq n$ , Lemma 12 and (165) yield the upper bound

$$\mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \leq \frac{1}{n} \left[ 2k\mathcal{U}(F_{X_0}) + \sum_{i=k+1}^{n-k} EU(F_{X_0 | Y_{1-i}^{n-i}}) \right] \tag{166}$$

$$\leq \frac{1}{n} \left[ 2k\mathcal{U}(F_{X_0}) + \sum_{i=k+1}^{n-k} EU(F_{X_0 | Y_{-k}^k}) \right] \tag{167}$$

$$= \frac{1}{n} \left[ 2k\mathcal{U}(F_{X_0}) + (n-2k) EU(F_{X_0 | Y_{-k}^k}) \right] \tag{168}$$

Considering the limit as  $n \rightarrow \infty$  of both ends of the above chain yields  $\mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \leq EU(F_{X_0 | Y_{-k}^k})$ . Letting now  $k \rightarrow \infty$  and invoking Lemma 12 implies  $\mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \leq EU(F_{X_0 | Y_{-\infty}^{\infty}})$ . ■

### C. Proof of Theorem 8

*Proof:* By definition of  $\mathbb{D}(F_{\mathbf{X}}, \mathcal{C})$  clearly

$$\liminf_{n \rightarrow \infty} EL_{\tilde{X}_{\text{univ}}^n}(X^n, Y^n) \geq \mathbb{D}(F_{\mathbf{X}}, \mathcal{C})$$

On the other hand, from (45), for any  $k$

$$\begin{aligned}
ED_k(X^n) &= E \min_g E_{F_{x^n}^k \otimes \mathcal{C}} \Lambda(X, g(Y_{-k}^k)) \\
&\leq \min_g E \left[ E_{F_{x^n}^k \otimes \mathcal{C}} \Lambda(X, g(Y_{-k}^k)) \right] \\
&= \min_g E\Lambda(X, g(Y_{-k}^k)) \tag{169}
\end{aligned}$$

where, the right side  $X_{-k}^k$  is emitted from the (unique) double-sided extension of the source  $F_{\mathbf{X}}$ . Using the result from equation (169), we get

$$\limsup_{n \rightarrow \infty} ED_{k_n}(X^n) \leq \mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \tag{170}$$

implying, by Theorem 7 and bounded convergence, that

$$\limsup_{n \rightarrow \infty} EL_{\tilde{X}_{\text{univ}}} (X^n, Y^n) \leq \mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \quad (171)$$

and proving (63). To prove (64) assume stationary ergodic  $\mathbf{X}$ . We have established the continuity of  $E_{F \otimes \mathcal{C}} \Lambda(U_0, g(Y))$  w.r.t  $F \in \mathcal{F}^{[a,b]}$  in Lemma 5 and it is easily extendible to  $\min_g E_{F \otimes \mathcal{C}} \Lambda(U_0, g(Y))$ . By the ergodic theorem and continuity of  $\min_g E_{F \otimes \mathcal{C}} \Lambda(U_0, g(Y))$  in  $F \in \mathcal{F}^{[a,b]}$ , it follows from the representation in (45) that

$$D_k(\mathbf{X}) = \lim_{n \rightarrow \infty} D_k(X^n) = \min_g E \Lambda(X_0, g(Y_{-k}^k)) \quad a.s. \quad (172)$$

and by Claim 1,

$$D(\mathbf{X}) = \mathbb{D}(F_{\mathbf{X}}, \mathcal{C}) \quad a.s. \quad (173)$$

Thus, the fact that  $\limsup_{n \rightarrow \infty} D_{k_n}(\mathbf{x}), \forall \mathbf{x} \in [a, b]^\infty$  (recall proof of Corollary 1), combined with Theorem 7, implies

$$\limsup_{n \rightarrow \infty} L_{\tilde{X}_{\text{univ}}} (X^n, Y^n) \leq \mathbb{D}(F_X, \mathcal{C}) \quad a.s. \quad (174)$$

On the other hand, by Fatou's lemma and definition of  $\mathbb{D}(F_X, \mathcal{C})$

$$E \left[ \limsup_{n \rightarrow \infty} L_{\tilde{X}_{\text{univ}}} (X^n, Y^n) \right] \geq \limsup_{n \rightarrow \infty} EL_{\tilde{X}_{\text{univ}}} (X^n, Y^n) \geq \mathbb{D}(F_X, \mathcal{C}) \quad (175)$$

The combination of (174) and (175) completes the proof of (64)  $\blacksquare$

## APPENDIX IX

### COMPARISON TO THE DENOISER IN [5]

Referring to 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  $\mathcal{O} = \{O_1, \dots, O_M\}$  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  $k^{\text{th}}$ -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,

$$\hat{Q}_{z^n}^k [y^n] (v_{-k}^k) = \int_{y_{-k}^k \in \mathcal{O}^{2k+1}} f_Y^{n,k}(y_{-k}^k) dy_{-k}^k \quad (176)$$

where,  $v_{-k}^k$  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,

$$\Pi(i, j) = \int_{y: Q_\alpha(y)=j} f_{Y|x=i}(y) dy \quad (177)$$

where  $Q_\alpha(\cdot)$  denotes a uniform quantizer with a quantization step size  $\alpha$ .

We compare  $\hat{Q}_{z^n}^k [y^n] (v_{-k}^k)$  to  $\hat{P}_{z^n}^k (v_{-k}^k)$ , the  $k$ -th order distribution of the quantized output symbols, using the notation in [5].

$$\hat{P}_{z^n}^k (v_{-k}^k) = \frac{\mathbf{r} [z^n, v_{-k}^k]}{n - 2k} \quad (178)$$

The density estimate,  $f_Y^{n,k}$ , we consider is the cubic histogram estimate. The histogram estimate is defined by

$$f_Y^{n,k}(y) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbf{1}_{[Y_i \in A_{n,j}]}}{\lambda(A_{n,j})}, \quad y \in A_{n,j}, y \in \mathbb{R}^{2k+1} \quad (179)$$

where,  $\mathcal{P}_n = \{A_{n,j}, j = 1, 2, \dots\}$ ,  $n \geq 1$  is a sequence of partitions and  $A_{n,j}$ '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

$$\bigcap_{n=1}^{\infty} \sigma \left( \bigcup_{m=n}^{\infty} \mathcal{P}_m \right) \quad (180)$$

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  $A_{n,j}$  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 (179) 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 (25), holds for histogram density estimates.

*Theorem 12:* Assume that the sequence of partitions  $\mathcal{P}_n$  satisfies (180). Consider

- 1)  $J_n \rightarrow 0$  in probability as  $n \rightarrow \infty$ , for all sequences  $x^n$
- 2)  $J_n \rightarrow 0$  almost surely as  $n \rightarrow \infty$ , for all sequences  $x^n$
- 3)  $J_n \rightarrow 0$  exponentially as  $n \rightarrow \infty$ , 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  $A_n \in \sigma(\mathcal{P}_n)$  with  $\lambda(A \Delta A_n) < \varepsilon$  and

$$\sup_{M > 0 \text{ all sets } C \text{ of finite Lebesgue measure}} \limsup_{n \rightarrow \infty} \lambda \left( \bigcup_{j: \lambda(A_{n,j} \cap C) \leq \frac{M}{n}} A_{n,j} \cap C \right) = 0 \quad (181)$$

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

For the proof of this theorem, refer to [7] with the added condition of tightness imposed on the family of measures associated with the channel,  $\mathcal{C}$ .

The condition 4) in Theorem 12 translates to  $\lim_{n \rightarrow \infty} h = 0$ ,  $\lim_{n \rightarrow \infty} nh^d = \infty$ . It can be shown as in [7] that they are necessary sufficient conditions for that specified in 4) in Theorem 12. 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  $nh^d \rightarrow \infty$  is also simultaneously satisfied, we get the mapping in equation (176) to reduce to that in equation (178) 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 12. Now,

$$\hat{Q}_{z^{n_i}}^k(v_{-k}^k) = \int_{y_{-k}^k \in \mathcal{O}^{2k+1}} f_Y^{n_i, k}(y_{-k}^k) dy_{-k}^k \quad (182)$$

$$= \int_{y_{-k}^k \in \mathcal{O}^{2k+1}} \frac{1}{\lceil \frac{n-2k-i-1}{2k+1} \rceil} \sum_{j=0}^{\lceil \frac{n-2k-i-1}{2k+1} \rceil} \frac{\mathbf{1}_{[Y_{j(2k+1)+i}^{j(2k+1)+i+2k} \in A_{n_i l}]}}{\lambda(A_{n_i l})} \quad (183)$$

$$= \frac{1}{\lceil \frac{n-2k-i-1}{2k+1} \rceil} \mathbf{r}[z^{n_i}, v_{-k}^k] \quad (184)$$

If we mapped the finite input-continuous output channel,  $\mathcal{C}$ , to  $\Pi$ , the mapping in equation (48) would then reduce to,

$$\hat{Q}_{x^{n_i}}^k = \arg \min_{P \in \mathcal{F}^{\mathcal{A}, k}} \sum_{v_{-k}^k} \left| \hat{Q}_{z^{n_i}}^k(v_{-k}^k) - \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \prod_{j=-k}^k \Pi(u_j, v_j) P(u_{-k}^k) \right| \quad (185)$$

where,  $\mathcal{F}^{\mathcal{A}, k}$  denote the space of all possible  $k^{\text{th}}$ -order distributions on  $\mathcal{A}$ . If we lift the constraints of the minimizer being a bona fide element of  $\mathcal{F}^{\mathcal{A}, k}$ , we get the following candidate for the minimizer in (185)

$$\hat{Q}_{x^{n_i}}^k[u_{-k}^k] = \frac{1}{\lceil \frac{n-2k-i-1}{2k+1} \rceil} \sum_{v_{-k}^k} \mathbf{r}[z^{n_i}, v_{-k}^k] \prod_{j=-k}^k \Pi^{-1}(v_j, u_j) \quad (186)$$

which is exactly the same as  $\hat{P}_{x^{n_i}}[z^{n_i}](u_{-k}^k)$  using equation (18) in [5], also given below.

$$\hat{P}_{x^{n_i}}^k[u_{-k}^k] = \frac{1}{\lceil \frac{n-2k-i-1}{2k+1} \rceil} \sum_{v_{-k}^k} \mathbf{r}[z^{n_i}, v_{-k}^k] \prod_{j=-k}^k \Pi^{-1}(v_j, u_j) \quad (187)$$

Now, using the construction of the discrete denoiser in equation (50), for  $\hat{Q}_{x^{n_i}}$ , we get

$$\begin{aligned} g_{\text{opt}}[\hat{Q}_{x^{n_i}}](y_{-k}^k) &= \arg \min_{\hat{x} \in \mathcal{A}} \Lambda(\cdot, \hat{x})^T [\hat{Q}_{x^{n_i}} \otimes \mathcal{C}]_{U|y_{-k}^k} \\ &= \arg \min_{\hat{x} \in \mathcal{A}} \sum_{\hat{a} \in \mathcal{A}} \Lambda(\hat{a}, \hat{x}) \cdot \left\{ \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}: u_0 = \hat{a}} \left[ \prod_{j=-k}^k f_{Y|x=u_j}(y_j) \hat{Q}_{x^{n_i}}(U_{-k}^k = u_{-k}^k) \right] \right\} \end{aligned} \quad (188)$$

which is exactly the same as  $g_{\text{opt}}[P](y_{-k}^k)$  in equation (16) in [5]. Hence, the proposed denoiser with histogram density estimate of the output symbols and quantization gives us the same denoising rule as that of [5] applied to the  $2k + 1$  subsequences of the output sequence  $Y^n$ .

## REFERENCES

- [1] R. Averkamp and C. Hourdré, "Wavelet thresholding for non-necessarily Gaussian noise:idealism," *The Annals of Statistics*, vol. 31, no. 1, pp. 110–151, 2003.
- [2] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. New York, NY: Athena Scientific, 1997.
- [3] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [4] A. Buades, B. Coll, and J. M. Morel, "A review of image de-noising algorithms with a new one," *Multiscale Modeling and Simulation*, vol. 4, pp. 490–530, Jul. 2005.
- [5] A. Dembo and T. Weissman, "Universal denoising for the finite input general output channel," *Information Theory, IEEE Transactions on*, vol. 51, pp. 1507 – 1517, Apr. 2005.
- [6] L. Devroye, *A Course in Density Estimation*. Boston, MA: Birkhauser, 1987.

- [7] L. Devroye and L. Györfi, *Nonparametric Density Estimation, the  $L_1$  View*. New York, NY: Wiley Series in Probability and Mathematical Statistics, 1985.
- [8] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.
- [9] D. L. Donoho, “De-noising by soft-thresholding,” *Information Theory, IEEE Transactions on*, vol. 41, pp. 613–627, May. 1995.
- [10] —, “Kolmogorov sampler,” Stanford University, Tech. Rep., 2002.
- [11] R. M. Dudley, *Real Analysis and Probability*. New York, NY: Cambridge studies in advanced mathematics, 2002.
- [12] R. Durrett, *Probability: Theory and Examples*, 3rd ed., ser. Duxbury Advanced Series. Thomson Books/Cole, 2005.
- [13] H. Y. Gao, “Wavelet estimation of spectral densities in time series analysis,” Ph.D. dissertation, University of California, Berkeley, 1993.
- [14] A. G. Gray and A. W. Moore, “Rapid evaluation of Multiple Density Models,” in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Jan. 2003.
- [15] —, “Very fast multivariate kernel density estimation via computational geometry,” in *Proceedings of the Joint Statistical Meeting*, Aug. 2003.
- [16] L. Greengard and J. Strain, “The fast Gauss transform,” *SIAM Journal on Scientific and Statistical Computing*, vol. 12, no. 1, pp. 79–94, 1991.
- [17] H. Gudbjartsson and S. Patz, “The rician distribution of noisy MRI data,” *Magn. Reson. Med*, vol. 34, pp. 910–914, 1995.
- [18] L. Györfi and E. Masry, “The  $L_1$  and  $L_2$  strong consistency of recursive kernel density estimation from dependent samples,” *Information Theory, IEEE Transactions on*, vol. 36, pp. 531–539, 1990.
- [19] L. Hsu, S. G. Self, D. Grove, T. Randolph, K. Wang, J. J. Delrow, L. Loo, and P. Porter, “Denoising array-based comparative genomic hybridization data using wavelets,” *Biostat*, vol. 6, pp. 211–226, 2005.
- [20] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed., ser. Probability and Its Applications. Springer, 2001.
- [21] E. D. Kolaczyk, “Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds,” *Statistica Sinica*, vol. 9, pp. 119–136, 1999.
- [22] E. McVeigh, R. Henkelman, and M. Bronskill, “Noise and filtration in magnetic resonance imaging,” *Med. Phys.*, vol. 3, pp. 604–618, 1985.
- [23] A. W. Moore, “An introductory tutorial on kd-trees,” University of Cambridge, Tech. Rep. 209, 1991.
- [24] G. Motta, E. Ordentlich, I. Ramirez, G. Seroussi, and M. J. Weinberger, “The dude framework for continuous tone image denoising,” in *Image Processing, 2005 IEEE International Conference on*, Sept. 2005.
- [25] B. Natarajan, “Filtering random noise from deterministic signals via data compression,” *Signal Processing, IEEE Transactions on*, vol. 43, no. 11, pp. 2595–2605, 1995.
- [26] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli, “Image denoising using scale mixtures of Gaussians in the wavelet domain,” *Image Processing, IEEE Transactions on*, vol. 12, pp. 1338–1351, Nov. 2003.
- [27] M. Raphan and E. P. Simoncelli, “Learning to be bayesian without supervision,” in *Neural Information Processing Systems*, Dec. 2006, pp. 1145–1152.
- [28] V. C. Raykar and R. Duraiswami, “Very fast optimal bandwidth selection for univariate kernel density estimation,” Department of computer science, University of Maryland, College Park, Tech. Rep. CS-TR-4774, 2005.
- [29] —, “Fast optimal bandwidth selection for kernel density estimation,” in *Proceedings of the sixth SIAM International Conference on Data Mining*, J. Ghosh, D. Lambert, D. Skillicorn, and J. Srivastava, Eds., 2006, pp. 524–528.
- [30] W. Rudin, *Principles of Mathematical Analysis*. New York, NY: McGraw Hill Book Company, 1976.
- [31] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.
- [32] K. Sivaramakrishnan and T. Weissman, “Universal denoising of continuous valued signals with applications to images,” in *Image Processing, 2006 IEEE International Conference on*, Atlanta, Sept. 2006.
- [33] —, “A context quantization approach to universal denoising,” 2008, submitted for publication in *IEEE Transactions on Signal Processing*.
- [34] R. Wagner, S. Smith, J. Sandrik, and H. Lopez, “Statistics of speckle in ultrasound b-scans,” *IEEE Trans. on Sonics and Ultrasonics*, vol. 30, no. 3, pp. 156–163, May 1983.
- [35] X. H. Wang, R. S. H. Istepanian, and Y. H. Song, “Microarray image enhancement by denoising using stationary wavelet transform,” *Neuro Biology, IEEE Transactions on*, vol. 14, pp. 184–189, Dec. 2003.

- [36] T. Weissman, E. Ordentlich, G. Seroussi, and M. W. S. Verdu, "Universal discrete denoising: Known channel," *Information Theory, IEEE Transactions on*, vol. 51, pp. 1229 – 1246, Jan. 2005.
- [37] R. Wheeden and A. Zymund, *Measure and Integral*. New York, NY: Marcel Dekker, 1977.



RMSE = 14.7354



RMSE = 13.0945



RMSE = 11.2899



RMSE = 11.2610



RMSE = 11.1782



RMSE = 7.842

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 [9], right: the scheme in [26]

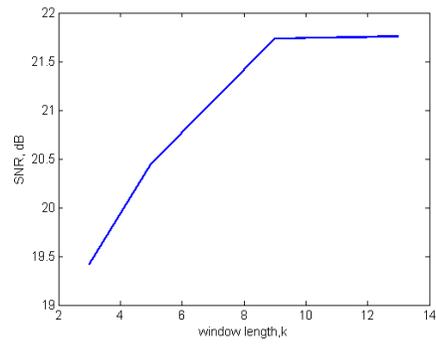


Fig. 6. Comparison of RMSE of the denoised image for various context lengths,  $k$

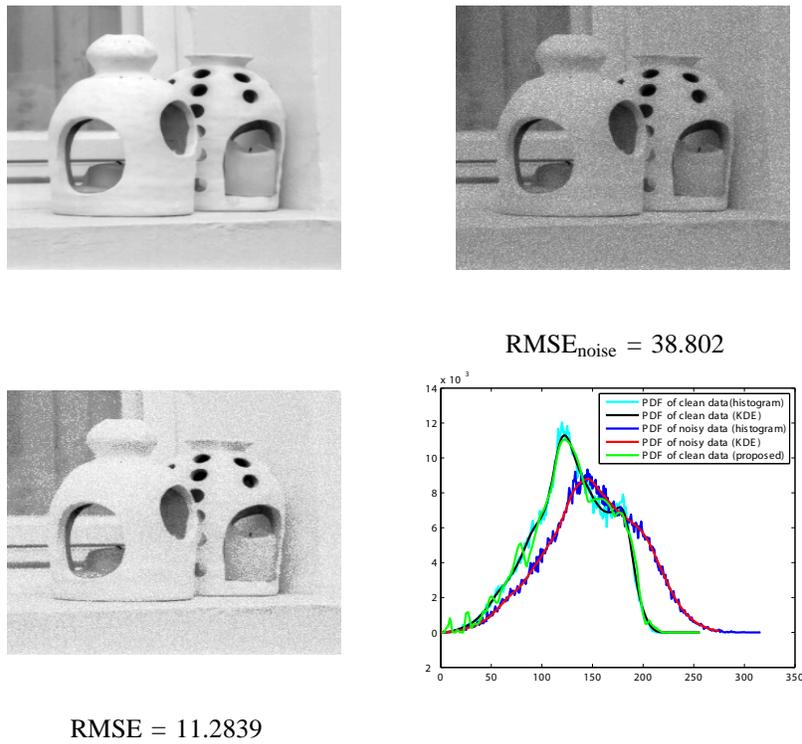


Fig. 7. Row 1- left: Original image, right: Noisy image; Denoised images using Row 2- left: symbol-symbol scheme, right: Comparison of Distribution estimates for the symbol-by-symbol denoiser

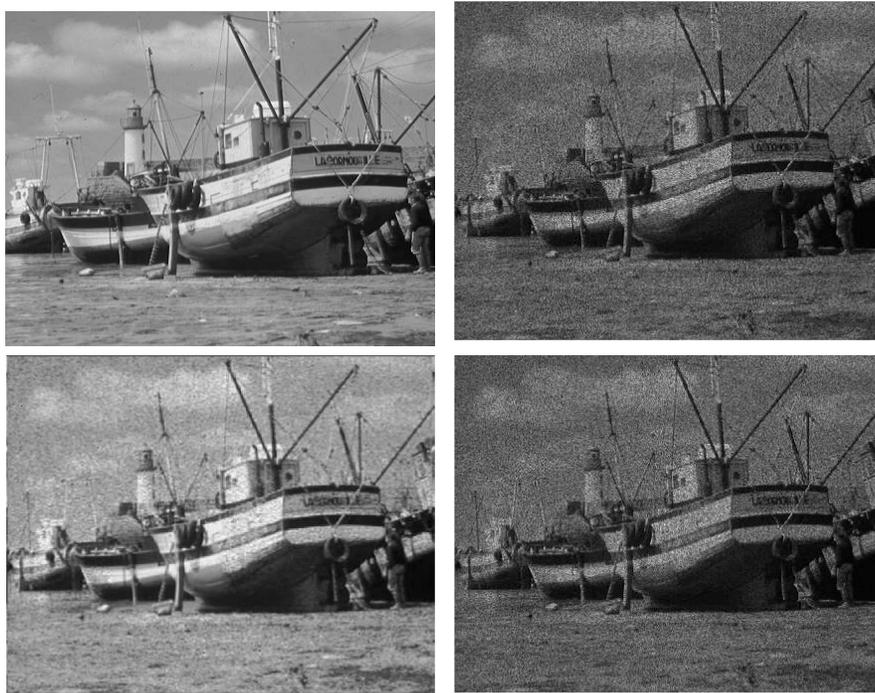


Fig. 8. Row 1- left: Original image, right: Noisy image; Denoised images using Row 2- left: proposed scheme, right: BLS-GSM [26]