Since

$$D(\widehat{Y}^n\|X^n) = \log\frac{1}{P_{Y^n}(C_n)} + \int_{C_n}\log\frac{dP_{Y^n}}{dP_{X^n}}(y)dP_{\widehat{Y}^n}(y)$$
$$\le \log\frac{N_n(F_n)}{N_n(C_n)} - \log P_{Y^n}(F_n) + n(r_1+\varepsilon)$$
$$\le nr - \log P_{Y^n}(F_n)$$

we have

$$D_u(\widehat{\boldsymbol{Y}}\|\boldsymbol{X}) \le r. \tag{93}$$

Since $\varepsilon > 0$ is arbitrary, it follows from Theorem 2, (92) and (93) that

$$R_e^*(D,r|\boldsymbol{X}) = \inf_{\boldsymbol{Y}:D_u(\boldsymbol{Y}\|\boldsymbol{X})\le r} R(D|\boldsymbol{Y}) \le g(r) - r$$

and that (90) holds. We now suppose that $r > g(r_1)$. Then

$$R_e^*(D,r|\boldsymbol{X}) \le R_e^*(D,g(r_1)|\boldsymbol{X})$$
$$\le [g(g(r_1)) - g(r_1)]^+ = [g(r_1) - g(r_1)]^+ = 0$$

yielding (78).

Finally, if $g(r,\cdot)$ is continuous, then (79) follows from (78). □

Example 2 (Stationary Ergodic Source): Let $\boldsymbol{X} = \{X^n\}$ be a source corresponding to a stationary ergodic process $\{X_n\}$. We denote by $\mathcal{D}$ the set of all $\mathcal{X}$-valued random variables Y for which there exists $a \in \mathcal{X}$ such that $E[d(Y,a)] < \infty$, and by $\mathcal{E}$ the set of all stationary ergodic sources $\boldsymbol{Y} = \{Y^n\}$ such that $Y_1 \in \mathcal{D}$. It is known (cf. [3], [10]) that

$$\widetilde{R}_{fm}(D|\boldsymbol{Y}) = \widetilde{R}_{va}(D|\boldsymbol{Y}) = \lim_{n\to\infty}\frac{1}{n}R(D|Y^n) \tag{94}$$

if $\boldsymbol{Y} = \{Y^n\} \in \mathcal{E}$. Combining Theorem 2, Theorem 4 and (94), we have

$$R_e^*(D,r|\boldsymbol{X}) \le \inf_{\boldsymbol{Y}=\{Y^n\}\in\mathcal{E}:D_u(\boldsymbol{Y}\|\boldsymbol{X})\le r}\lim_{n\to\infty}\frac{1}{n}R(D_1|Y^n).$$

$0 < D_1 < D$. This inequality may be useful to show the direct part of the coding theorem for the rate $R_e^*(D,r|\boldsymbol{X})$.

REFERENCES

[1] I. Csiszár and Körner, *Information Theory: Coding Theorems for Discrete Memoryless System*. London, U.K.: Academic, 1981.

[2] T. S. Han, "An information-spectrum approach to source coding theorems with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. 43, no. 4, pp. 1145–1164, Jul. 1997.

[3] ——, *Information-Spectrum Methods in Information Theory*. Tokyo, Japan: Baifukan, 1998.

[4] ——, "The reliability functions of the general source with fixed-length coding," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 2117–2132, Sep. 2000.

[5] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf.. Theory*, vol. 39, no. 3, pp. 752–772, May 1993.

[6] S. Ihara and M. Kubo, "Error exponent for coding of memoryless Gaussian sources with a fidelity criterion," *IEICE Trans.*, vol. E83-A, pp. 1891–1897, Oct. 2000.

[7] K. Iriyama, "Probability of error for the fixed-length source coding of general sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1537–1543, May 2001.

[8] K. Iriyama and S. Ihara, "The error exponent and minimum achievable rates for the fixed-length coding of general sources," *IEICE Trans. Fundament.*, vol. E84-A, pp. 2466–2473, Oct. 2001.

[9] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 197–199, Mar. 1974.

[10] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 63–86, May 1996.

Universal Denoising for the Finite-Input General-Output Channel

Amir Dembo and Tsachy Weissman, *Member, IEEE*

***Abstract*—We consider the problem of reconstructing a finite-alphabet signal corrupted by a known memoryless channel with a general output alphabet. The goodness of the reconstruction is measured by a given loss function. We (constructively) establish the existence of a universal (sequence of) denoiser(s) attaining asymptotically the optimum distribution-dependent performance for any stationary source that may be generating the noiseless signal. We show, in fact, that there is a whole family of denoiser sequences with this property. These schemes are shown to be universal also in a semistochastic setting, where the only randomness assumed is that associated with the channel noise. The scheme is practical, requiring $O(n^{1+\varepsilon})$ operations (for any $\varepsilon > 0$) and working storage size sublinear in the input data length. This extends recent work that presented a discrete universal denoiser for recovering a discrete source corrupted by a discrete memoryless channel (DMC).**

***Index Terms*—Denoising, discrete universal denoising, filtering, individual sequences, memoryless channels, noisy channels, quantization, sliding-window schemes, universal algorithms.**

I. INTRODUCTION

The goal of a denoising algorithm is to recover a signal from its noise-corrupted observations. Perfect recovery is seldom possible and performance is measured under a given fidelity criterion. The problem lies at the heart of a wide range of scenarios spanning such fields as statistics, engineering, and bioinformatics (cf. [10] for a broader discussion of the problem and for a sample from the many references). For discrete signals corrupted by discrete memoryless channels it was recently shown in [10] that this task can be optimally and practically performed with no knowledge of statistical (or any other) properties of the signal.

Our interest in the present work is in the case where the components of the underlying noise-free signal are still finite valued, yet their noisy observations take values in a general alphabet. For concreteness, we will let the alphabet of the noise-free signal be $\mathcal{A} = \{0,\ldots,M-1\}$, $M < \infty$, and assume the channel output alphabet is $\mathbb{R}$. The channel in this case, which we denote by $\mathcal{C}$, is assumed memoryless and is given by the set $\{f_a\}_{a\in\mathcal{A}}$, f_a denoting the density with respect to (w.r.t.) Lebesgue measure, assumed for concreteness to exist, associated with the channel output distribution for an input symbol a. All the derivations and results in this work carry over to the more general case where the channel output distribution for some inputs may not have a density w.r.t. Lebesgue measure by considering the density w.r.t. a different dominating measure (which will always exist). Similarly, the results carry over to output alphabets other than the real line. In particular, the schemes and results that will be presented are applicable for the case

Manuscript received May 14, 2004; revised November 7, 2004. The work of T. Weissman was supported in part by the National Science Foundation under Grants DMS-0072331 and CCR-0312839. Part of this work was performed while T. Weissman was visiting the Statistics Department, Stanford University, Stanford, CA. The material in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004.

A. Dembo is with the Departments of Mathematics and Statisitcs, Stanford University, Stanford, CA 94305 USA (e-mail: amir@stanford.edu).

T. Weissman is with the Electrical Engineering Department, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.edu).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2005.844104

0018-9448/$20.00 © 2005 IEEE

of a finite channel output alphabet as well. Letting μ denote Lebesgue measure, our assumption on the channel is as follows.

Assumption 1: The set of densities $\{f_a\}_{a\in\mathcal{A}}$ is a set of linearly independent functions in $L_1(\mu)$.

Note that this parallels the assumption that the channel matrix is of full row rank in the finite alphabet setting of [10]. This assumption is equivalent to the requirement that the channel output distribution (in the single input–output case) uniquely determine its input distribution, clearly a necessary condition if universal denoising is to be feasible.

For a fixed k, the Discrete Universal DEnoiser (DUDE) of [10] was based on accumulating counts of occurrences of strings of length $2k+1$ appearing along the noisy observation signal (first pass), and then (second pass) employing a scheme tailored for the output statistics associated with each "double-sided context" of order k (i.e., the conditional distribution of the middle component conditioned on the remaining components of the $2k+1$ tuple). The efficiency of the scheme was largely due to the fact that for small enough k, the limited number of possible noisy $2k+1$ tuples guaranteed that most of these appear enough times for decisions based on the empirical distribution within the associated double-sided context to be reliable. This breaks down completely in our setting (assuming channel output densities), where all appearing $2k+1$ tuples will be distinct, with probability one.

Our approach in this work is to use a processed version of the noisy signal to estimate the distribution of a $2k+1$-tuple in the underlying noise-free signal, and then to operate "Bayesianly," employing a sliding-window scheme which estimates the clean input symbol at each location based on the $2k+1$ noisy components around it, assuming the estimated input distribution.

To demonstrate the simplicity and efficiency of this approach, we begin by considering an asymptotically optimal scheme based on scalar quantization of the noisy observations. The idea is to estimate the $2k+1$th-order distribution of the input based on the quantized observation signal in a first pass. We show that this can be done efficiently, with any scalar quantizer having the property that the channel matrix from clean to quantized observation is invertible. This, in turn, will be shown to lead to performance bounds guaranteeing the asymptotic optimality of the associated denoising scheme (which employs the said sliding-window denoiser assuming the estimate of the input distribution from the first pass). Note that the first pass coincides with the first pass of the DUDE of [10], when employed on the *quantized* noisy signal. It is somewhat surprising that this quantization does not prevent our schemes from attaining the optimum performance.

After setting up some notation in Section II, we shall turn in Section III to motivating the algorithm, presenting it concretely, and establishing performance bounds and its asymptotic optimality for the semistochastic setting (where the underlying noise-free signal is an individual sequence, and the only randomness is due to the channel noise). Section IV will show how the performance guarantees from the semistochastic setting of Section III directly imply asymptotic optimality of the denoising schemes when the noise-free signal is a stationary stochastic process. In Section V, we show that estimating the distribution of the noise-free signal based on scalar quantization of the channel output can be viewed as a special case of a more general approach which uses a set of basis functions for scalar (symbol-by-symbol) processing of the observations before using the processed signal for the said estimation. We show that denoisers belonging to this more general family are asymptotically optimal, and point to a plausible guideline for the choice of a particular denoiser from this family per a given channel. In Section VI, we show that for the case where the effective channel output alphabet is equal to the clean source alphabet, the denoiser of the previous sections coincides with the discrete denoiser of [10] and, hence, can be seen as its natural extension to the present setting. Section VII details the explicit form of the denoiser for a couple of special cases, Section VIII presents a few preliminary experimental results, and in Section IX we conclude, mentioning a few directions for related future research.

II. Basic Setup and Notation

We assume that $\mathcal{A}$ is a finite alphabet of size M where both the clean and reconstruction signal components take values. For any finite set $\mathcal{B}$, $\mathcal{M}(\mathcal{B})$ will denote the simplex of probabilities on $\mathcal{B}$. More concretely, $v\in\mathcal{M}(\mathcal{B})$ will be regarded as a $|\mathcal{B}|$-dimensional column vector with components corresponding to some arbitrary ordering of the elements of $\mathcal{B}$.

Let the measurable $Q:\mathbb{R}\to\mathcal{A}$ be a quantizer and $\mathbf{\Pi}$ denote the $M\times M$ channel matrix associated with the channel induced by quantizing the output of the original channel with the quantizer Q

$$\Pi(i,j)=\int_{y:Q(y)=j} f_i(y)dy. \tag{1}$$

Assumption 1 guarantees the existence of a quantizer for which $\mathbf{\Pi}$ is invertible, and we assume Q has this property. We mention in passing that, for "natural" channels, a simple and natural quantization will be evident. For example, for the binary-input additive white Gaussian noise (BIAWGN) channel, a natural quantizer will map the real line into two values according to whether the argument is positive or negative, in which case the resulting channel in (1) will be a binar-symmetric channel (BSC) (details for this example will be given in Section VII).

With slight abuse of notation we shall let $\Pi^{-1}_{\max}$ denote $\max_{i,j}\Pi^{-1}(i,j)$, $\Pi^{-1}(i,j)$ denoting the (i,j)th component of the inverse of the channel matrix.

We will denote the source signal by[1] $\boldsymbol{x}=(x_1,x_2,\ldots)$ and its noisy observation process by $\boldsymbol{Y}=(Y_1,Y_2,\ldots)$. $\boldsymbol{Z}=(Z_1,Z_2,\ldots)$, where $Z_i=Q(Y_i)$, will denote the quantized channel output. An *n-block denoiser* is a measurable mapping taking $\mathbb{R}^n$ into $\mathcal{A}^n$. We assume a given loss function $\Lambda:\mathcal{A}^2\to[0,\infty)$ and denote the normalized cumulative loss of an n-block denoiser $\hat{X}^n$ by

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

where $\hat{X}^n(y^n)[i]$ denotes the ith component of $\hat{X}^n(y^n)$. We denote $\Lambda_{\max}=\max\{\max_{i,j}\Lambda(i,j),1\}$ and let $\boldsymbol{\lambda}_{\hat{x}}$ denote the $\hat{x}$th column of Λ. We let $\mathcal{M}$ denote the set of M-dimensional column simplex probability vectors (corresponding to distributions on $\mathcal{A}$). For $\boldsymbol{P}\in\mathcal{M}$, we let

$$U(\boldsymbol{P})=\min_{\hat{x}\in\mathcal{A}}\sum_{a\in\mathcal{A}}\Lambda(a,\hat{x})\boldsymbol{P}(a)=\min_{\hat{x}\in\mathcal{A}}\boldsymbol{\lambda}_{\hat{x}}^T\boldsymbol{P} \tag{3}$$

denote its "Bayes envelope" [4], [8], [6]. In words, $U(\boldsymbol{P})$ is the minimum achievable expected loss in guessing the value of a variable distributed according to $\boldsymbol{P}$, as measured by the loss function Λ. Finally, $\|v\|_p$ will denote the p-norm of any vector v with real-valued components.

III. Semistochastic Setting

We start by assuming the semistochastic setting in which $\boldsymbol{x}=(x_1,x_2,\ldots)$ is an individual sequence corrupted by the channel $\mathcal{C}=\{f_a\}_{a\in\mathcal{A}}$, its noise corrupted version being the stochastic process

[1]Lower case letters will be used to denote "individual" sequences or symbols in the semistochastic setting where the noiseless source is assumed deterministic. They will also be used to denote specific realization values that the associated random quantities may assume. Upper case letters will denote random quantities.

$\boldsymbol{Y} = (Y_1, Y_2, \ldots)$. Define the kth-order sliding-window minimum loss of x^n by

$$D_k(x^n) = \min_g E\left[\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda\left(x_i, g\left(Y_{i-k}^{i+k}\right)\right)\right] \quad (4)$$

where the minimum is over all measurable maps $g : \mathbb{R}^{2k+1} \to \mathcal{A}$. Define further the sliding-window denoisability of $\boldsymbol{x}$ by

$$D(\boldsymbol{x}) = \lim_{k\to\infty}\limsup_{n\to\infty} D_k(x^n) \quad (5)$$

the limit existing for all $\boldsymbol{x}$ by monotonicity. For $x^n \in \mathcal{A}^n$ and $u_{-k}^k \in \mathcal{A}^{2k+1}$ define

$$\boldsymbol{r}\left[x^n, u_{-k}^k\right] = \left|\left\{k+1 \le i \le n-k : x_{i-k}^{i+k} = u_{-k}^k\right\}\right|. \quad (6)$$

Note that $D_k(x^n)$ can be expressed as

$$D_k(x^n) = \min_g \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \frac{\boldsymbol{r}\left[x^n, u_{-k}^k\right]}{n-2k} E_{u_{-k}^k}\Lambda\left(u_0, g\left(Y_{-k}^k\right)\right) \quad (7)$$

where $E_{u_{-k}^k}$ denotes expectation when the underlying clean symbols are u_{-k}^k (the expectation being, as in (4), over the channel noise) so that

$$E_{u_{-k}^k}\Lambda\left(u_0, g\left(Y_{-k}^k\right)\right) = \int_{\mathbb{R}^{2k+1}}\Lambda\left(u_0, g\left(y_{-k}^k\right)\right)\left[\prod_{i=-k}^{k} f_{u_i}(y_i)\right] dy_{-k}\ldots dy_k. \quad (8)$$

For a probability distribution P over $\mathcal{A}^{2k+1}$ let further $P \otimes \mathcal{C}$ and $E_{P\otimes\mathcal{C}}$ denote, respectively, probability and expectation when the channel input $U_{-k}^k \sim P$ and Y_{-k}^k is the channel output of the channel $\mathcal{C}$ whose input is U_{-k}^k. So that, e.g.,

$$\begin{aligned} E_{P\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y_{-k}^k\right)\right) &= \sum_{u_{-k}^k} P\left(u_{-k}^k\right) E_{u_{-k}^k}\Lambda\left(u_0, g\left(Y_{-k}^k\right)\right) \\ &= \sum_{u_{-k}^k} P\left(u_{-k}^k\right) \cdot \left[\int_{\mathbb{R}^{2k+1}}\Lambda\left(u_0, g\left(y_{-k}^k\right)\right) \cdot \left[\prod_{i=-k}^{k} f_{u_i}(y_i)\right] dy_{-k}\ldots dy_k\right]. \end{aligned} \quad (9)$$

Let $P \otimes \mathcal{C}\left(y_{-k}^k | U_0 = a\right)$ denote the density associated with the distribution of Y_{-k}^k conditioned on $U_0 = a$, as induced by the distribution $P \otimes \mathcal{C}$, namely,

$$\begin{aligned} P \otimes \mathcal{C}\left(y_{-k}^k | U_0 = a\right) &= \sum_{u_{-k}^{-1}, u_1^k} P \otimes \mathcal{C}\left(y_{-k}^k | U_{-k}^{-1} = u_{-k}^{-1}, U_1^k = u_1^k, U_0 = a\right) \\ &\quad \cdot P\left(U_{-k}^{-1} = u_{-k}^{-1}, U_1^k = u_1^k | U_0 = a\right) \quad (10) \\ &= \sum_{u_{-k}^k \in \mathcal{A}^{2k+1} : u_0 = a} P \otimes \mathcal{C}\left(y_{-k}^k | U_{-k}^k = u_{-k}^k\right) \\ &\quad \cdot P\left(U_{-k}^{-1} = u_{-k}^{-1}, U_1^k = u_1^k | U_0 = a\right) \quad (11) \\ &= \sum_{u_{-k}^k \in \mathcal{A}^{2k+1} : u_0 = a}\left[\prod_{i=-k}^{k} f_{u_i}(y_i)\right] \\ &\quad \cdot P\left(U_{-k}^{-1} = u_{-k}^{-1}, U_1^k = u_1^k | U_0 = a\right). \quad (12) \end{aligned}$$

Similarly

$$P \otimes \mathcal{C}\left(U_0 = a | y_{-k}^k\right) = \frac{1}{K\left(y_{-k}^k\right)} \cdot \sum_{u_{-k}^k \in \mathcal{A}^{2k+1} : u_0 = a}\left[\prod_{i=-k}^{k} f_{u_i}(y_i)\right] P\left(U_{-k}^k = u_{-k}^k\right) \quad (13)$$

$K(y_{-k}^k)$ being the normalization term

$$K\left(y_{-k}^k\right) = \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}}\left[\prod_{i=-k}^{k} f_{u_i}(y_i)\right] P\left(U_{-k}^k = u_{-k}^k\right). \quad (14)$$

Letting $[P \otimes \mathcal{C}]_{U_0|y_{-k}^k}$ denote the vector in $\mathcal{M}$ whose ath component is $P \otimes \mathcal{C}(U_0 = a | y_{-k}^k)$, it is then clear that

$$\min_g E_{P\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y_{-k}^k\right)\right) = E_{P\otimes\mathcal{C}} U\left([P \otimes \mathcal{C}]_{U_0|Y_{-k}^k}\right) \quad (15)$$

(U denoting the Bayes envelope defined in (3)), where the minimum is attained by the Bayes response to $[P \otimes \mathcal{C}]_{U_0|y_{-k}^k}$, namely

$$\begin{aligned} g_{\text{opt}}[P]\left(y_{-k}^k\right) &= \arg\min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T [P \otimes \mathcal{C}]_{U_0|y_{-k}^k} \\ &= \arg\min_{\hat{x}\in\mathcal{A}} \sum_{a\in\mathcal{A}} \Lambda(a, \hat{x}) \\ &\quad \cdot \left\{\sum_{u_{-k}^k \in \mathcal{A}^{2k+1} : u_0 = a}\left[\prod_{i=-k}^{k} f_{u_i}(y_i)\right] \cdot P\left(U_{-k}^k = u_{-k}^k\right)\right\} \end{aligned} \quad (16)$$

where the second line follows by substitution of the expression from (13).

For $x^n \in \mathcal{A}^n$, define now the distribution $P_{x^n}^k$ over $\mathcal{A}^{2k+1}$ by

$$P_{x^n}^k\left(u_{-k}^k\right) = \frac{\boldsymbol{r}\left[x^n, u_{-k}^k\right]}{n-2k}$$

i.e., $P_{x^n}^k$ is the empirical distribution of $2k+1$-tuples, as induced by x^n. It is then clear from (7) and (9) that

$$D_k(x^n) = \min_g E_{P_{x^n}^k \otimes \mathcal{C}}\Lambda\left(U_0, g\left(Y_{-k}^k\right)\right) \quad (17)$$

which is attained by $g_{\text{opt}}\left[P_{x^n}^k\right]$.

A. Description of the Algorithm

$P_{x^n}^k$, and, therefore, also $g_{\text{opt}}[P_{x^n}^k]$, are unknown to an observer of the noisy sequence. Our approach is to obtain an estimate $\hat{P}_{x^n}^k$ of $P_{x^n}^k$ from the quantized versions of the noisy data $Z_i = Q(Y_i)$, and then, by a plug-in approach, employ the sliding-window scheme $g_{\text{opt}}[\hat{P}_{x^n}^k]$. To this end, we start by defining $\hat{P}_{x^n}^k[z^n]$ by

$$\hat{P}_{x^n}^k[z^n]\left(u_{-k}^k\right) = \frac{1}{n-2k}\sum_{v_{-k}^k \in \mathcal{A}^{2k+1}} \boldsymbol{r}\left[z^n, v_{-k}^k\right]\prod_{j=-k}^{k}\Pi^{-1}(v_j, u_j) \quad (18)$$

where $\boldsymbol{r}[z^n, \cdot]$ are the count statistics associated with the quantized noisy signal (recall definition in (6)), as obtained in a first pass through the noisy data (cf. Fig. 1). Note that the expression on the right-hand side of (18) involves only the inversion of the $M \times M$ channel matrix (from x_i to $Z_i = Q(Y_i)$), and a summation over M^{2k+1} terms. As we show later (Lemma 1), $\hat{P}_{x^n}^k[Z^n]$ is an asymptotically efficient estimate of the unobserved $P_{x^n}^k$. Intuition for why this may be true can be gained by verifying that

$$E\left[\hat{P}_{x^n}^k[Z^n]\left(u_{-k}^k\right)\right] = P_{x^n}^k\left(u_{-k}^k\right)$$

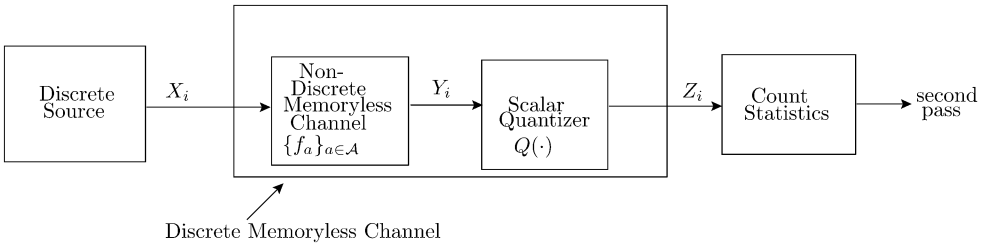

Fig. 1. First pass.

for each u_{-k}^{k}, i.e., $\hat{P}_{x^n}^{k}[Z^n]\left(u_{-k}^{k}\right)$ is an unbiased estimate of $P_{x^n}^{k}\left(u_{-k}^{k}\right)$.

For a fixed data size n, there are two ways in which the estimate $\hat{P}_{x^n}^{k}[Z^n]$ can be enhanced. The first is by ensuring that the estimated probabilities are all nonnegative. The second is by quantizing their values to some finite precision so as to facilitate their storage and retrieval. We thus define the modified estimate $\tilde{P}_{x^n}^{k,\delta}$ by

$$\tilde{P}_{x^n}^{k,\delta}\left(u_{-k}^{k}\right) = Q_\delta\left(\hat{P}_{x^n}^{k}[z^n]\left(u_{-k}^{k}\right)\right) \tag{19}$$

where Q_δ denotes quantization to the nearest nonnegative integer multiple of δ which is ≤ 1. Equipped with $\tilde{P}_{x^n}^{k,\delta}[Z^n]$, the natural candidate for a denoiser is

$$\tilde{X}^{n,k,\delta}[y^n](i) = g_{\text{opt}}\left[\tilde{P}_{x^n}^{k,\delta}[z^n]\right]\left(y_{i-k}^{i+k}\right), \qquad k+1 \leq i \leq n-k \tag{20}$$

where $g_{\text{opt}}[\cdot](\cdot)$ is given explicitly in[2] (16) and z^n is the scalar quantization of y^n, i.e., $z_i = Q(y_i)$. The value of $\tilde{X}^{n,k,\delta}[y^n](i)$ for i's outside the range $k+1 \leq i \leq n-k$, as in [10], will be asymptotically inconsequential. A rough analysis of the time complexity associated with the implementation of $\tilde{X}^{n,k,\delta}$ is as follows.

1. Acquisition of $\boldsymbol{r}\left[z^n, v_{-k}^{k}\right]$ for the different contexts: as in the DUDE of [10], linear.
2. Computation of $\hat{P}_{x^n}^{k}[z^n]\left(u_{-k}^{k}\right)$ for all u_{-k}^{k}:
$$O(M^{2k+1}M^{2k+1}(2k+1))$$
multiplications and summation operations.
3. Computation of $\tilde{P}_{x^n}^{k,\delta}[z^n]\left(u_{-k}^{k}\right)$ for all u_{-k}^{k}: $O(M^{2k+1})$ operations.
4. Application of $\tilde{X}^{n,k,\delta}(y^n)[i]$ for all i's:
$$O(nM^2M^{2k+1}(2k+1))$$
operations.

Summing up, we get number of operations which is

$$O(nM^{2k+\log k}) = O(n^{1+\varepsilon}), \qquad \text{for } k = k(n) = \frac{\varepsilon \log n}{3\log M}.$$

Storage complexity, beyond the cost of storing the noisy data (bounding generously) is $O(\log[1/\delta^{M^{2k+1}}])$, which will be made sublinear by our choice of δ (to follow).

B. Analysis

Let

$$\begin{aligned}\alpha(\varepsilon,k,\delta) = \left[\frac{1}{\delta}+1\right]^{M^{2k+1}} & (2M^{2k+1}+1) \\ & \cdot A\left(k, \varepsilon+\delta\Lambda_{\max}, \min\left\{\Lambda_{\max}, \Pi_{\max}^{-(2k+1)}\right\}\right) \\ & \cdot \exp\left[-nG\left(k, \varepsilon/(2\Lambda_{\max}), \max\left\{\Lambda_{\max}, \Pi_{\max}^{-(2k+1)}\right\}\right)\right]\end{aligned} \tag{21}$$

[2]Though $g_{\text{opt}}[P]$ was derived assuming P was a probability, $\tilde{P}_{x^n}^{k,\delta}[z^n]$ may not be a probability since after quantization its components may not sum to unity. We extend the definition of $g_{\text{opt}}[P]$ to the right-hand side of (16) to accommodate this case.

where $A(k,\varepsilon,B) = (2k+1)\exp\left(\frac{2\varepsilon^2}{B^2}\right)$ and $G(k,\varepsilon,B) = \frac{2\varepsilon^2}{(2k+1)B^2}$. Take now $k = k(n)$ and $\delta = \delta(n)$ such that $k(n) \to \infty$ and $\delta(n) \downarrow 0$ while

$$\sum_n \alpha(\varepsilon, k(n), \delta(n)) < \infty, \qquad \text{for all } \varepsilon > 0$$

and satisfying the growth order required in the above complexity analysis. For example, it is readily verified that any unboundedly increasing $k(n)$ with $k(n)/\log n \to 0$ and, say, $\delta(n) = 1/\log n$ jointly satisfy these requirements. Letting now

$$\tilde{X}_{\text{univ}}^{n} = \tilde{X}^{n,k(n),\delta(n)} \tag{22}$$

our main result is the following.

Theorem 1: For all $\boldsymbol{x} \in \mathcal{A}^\infty$

$$\lim_{n\to\infty}\left[L_{\tilde{X}_{\text{univ}}^{n}}(x^n, Y^n) - D_{k(n)}(x^n)\right] = 0 \text{ a.s.} \tag{23}$$

An immediate corollary is the following:

Corollary 1: For all $\boldsymbol{x} \in \mathcal{A}^\infty$

$$\limsup_{n\to\infty} L_{\tilde{X}_{\text{univ}}^{n}}(x^n, Y^n) \leq D(\boldsymbol{x}) \text{ a.s.} \tag{24}$$

Proof of Corollary 1 (Assuming Theorem 1): For any fixed k clearly, $\limsup_{n\to\infty} D_{k(n)}(x^n) \leq D_k(\boldsymbol{x})$ and, therefore, also $\limsup_{n\to\infty} D_{k(n)}(x^n) \leq D(\boldsymbol{x})$which, combined with (23), implies (24). □

Theorem 1 is a direct consequence of the following (combined with the Borel–Cantelli lemma).

Theorem 2: For all $n \geq 1, k, \varepsilon > 0, \delta > 0$ and $x^n \in \mathcal{A}^n$

$$\Pr\left(|L_{\tilde{X}^n,k,\delta}(x^n, Y^n) - D_k(x^n)| > 4\varepsilon + 6\delta\Lambda_{\max}\right) \leq \alpha(\varepsilon,k,\delta). \tag{25}$$

Three lemmas employed in the proof of Theorem 2 are as follows.

Lemma 1: For all $n \geq 1$, $x^n \in \mathcal{A}^n$, and $\delta, \varepsilon > 0$

$$\begin{aligned}\Pr\left(\left\|\tilde{P}_{x^n}^{k,\delta}[Z^n] - P_{x^n}^{k}\right\|_\infty > \varepsilon+\delta\right) & \leq \Pr\left(\left\|\hat{P}_{x^n}^{k}[Z^n] - P_{x^n}^{k}\right\|_\infty > \varepsilon\right) \\ & \leq M^{2k+1}A\left(k, \varepsilon, \left(\Pi_{\max}^{-1}\right)^{2k+1}\right) \\ & \quad \cdot \exp\left(-G\left(k, \varepsilon, \left(\Pi_{\max}^{-1}\right)^{2k+1}\right)n\right).\end{aligned} \tag{26}$$

Lemma 2: For every $n \geq 1$, $x^n \in \mathcal{A}^n$, measurable $g : \mathbb{R}^{2k+1} \to \mathcal{A}$, and $\varepsilon > 0$

$$\begin{aligned}\Pr\Bigg(\Bigg|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda\left(x_i, g\left(Y_{i-k}^{i+k}\right)\right) & \\ - E_{P_{x^n}^{k}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y_{-k}^{k}\right)\right)\Bigg| > \varepsilon\Bigg) & \\ \leq A\left(k,\varepsilon,M,\Lambda_{\max}\right)\exp\left(-G\left(k,\varepsilon,\Lambda_{\max}\right)n\right). &\end{aligned} \tag{27}$$

Lemma 3: For $P, \tilde{P} \in \mathcal{M}(\mathcal{A}^{2k+1})$ and $g : \mathbb{R}^{2k+1} \to \mathcal{A}$

$$\left| E_{P\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y_{-k}^{k}\right)\right) - E_{\tilde{P}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y_{-k}^{k}\right)\right)\right| \le \Lambda_{\max}\|P - \tilde{P}\|_1. \quad (28)$$

Proofs of these three lemmas are given in Appendix A, while the proof of Theorem 2 is relegated to Appendix B.

IV. Stochastic Setting

Assume the source signal is now a stationary stochastic process $\boldsymbol{X} = (X_1, X_2, \ldots)$, whose distribution is $\boldsymbol{P}_{\boldsymbol{X}}$. Let

$$\mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}) = \lim_{n\to\infty} \min_{\hat{X}^n} E L_{\hat{X}^n}(X^n, Y^n) \quad (29)$$

where the expectation on the right side is assuming X^n are the first n symbols emitted by the source $\boldsymbol{P}_{\boldsymbol{X}}$ and that Y^n are the noisy output from the channel $\mathcal{C}$. The minimum is taken over all n-block denoisers and the limit is guaranteed to exist by subadditivity.

Theorem 3: For all stationary $\boldsymbol{X}$

$$\lim_{n\to\infty} E L_{\tilde{X}^n_{\text{univ}}}(X^n, Y^n) = \mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}). \quad (30)$$

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

$$\limsup_{n\to\infty} L_{\tilde{X}^n_{\text{univ}}}(X^n, Y^n) = \mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}) \text{ a.s.} \quad (31)$$

Proof: By the definition of $\mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C})$ clearly

$$\liminf_{n\to\infty} E L_{\tilde{X}^n_{\text{univ}}}(X^n, Y^n) \ge \mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}).$$

On the other hand, by (17), for any k

$$\begin{aligned} E D_k(X^n) &= E \min_g E_{P^k_{X^n}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y_{-k}^{k}\right)\right) \\ &\le \min_g E\left[E_{P^k_{X^n}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y_{-k}^{k}\right)\right)\right] \\ &= \min_g E\Lambda\left(X_0, g\left(Y_{-k}^{k}\right)\right) \end{aligned} \quad (32)$$

where in the right side X_{-k}^{k} is emitted from the (unique) double-sided extension of the source $\boldsymbol{P}_{\boldsymbol{X}}$. Using a standard martingale argument as in [10, Sec. 5] one can show that

$$\lim_{k\to\infty} \min_g E\Lambda\left(X_0, g\left(Y_{-k}^{k}\right)\right) = \mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}). \quad (33)$$

It thus follows from (32) that

$$\limsup_{n\to\infty} E D_{k(n)}(X^n) \le \mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}) \quad (34)$$

implying, by Theorem 1 and bounded convergence, that

$$\limsup_{n\to\infty} E L_{\tilde{X}^n_{\text{univ}}}(X^n, Y^n) \le \mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}) \quad (35)$$

and proving (30). To prove (31) assume stationary ergodic $\boldsymbol{X}$. By the ergodic theorem and continuity of $\min_g E_{P\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y_{-k}^{k}\right)\right)$ in $P \in \mathcal{M}(\mathcal{A}^{2k+1})$, it follows from the representation in (17) that

$$D_k(\boldsymbol{X}) = \lim_{n\to\infty} D_k(X^n) = \min_g E\Lambda\left(X_0, g\left(Y_{-k}^{k}\right)\right) \text{ a.s.} \quad (36)$$

and, by (33), that

$$D(\boldsymbol{X}) = \mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}) \text{ a.s.} \quad (37)$$

Thus, the fact that $\limsup_{n\to\infty} D_{k(n)}(x^n) \le D(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{A}^\infty$ (recall proof of Corollary 1), combined with Theorem 1, implies

$$\limsup_{n\to\infty} L_{\tilde{X}^n_{\text{univ}}}(X^n, Y^n) \le \mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}) \text{ a.s.} \quad (38)$$

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

$$\begin{aligned} E\left[\limsup_{n\to\infty} L_{\tilde{X}^n_{\text{univ}}}(X^n, Y^n)\right] &\ge \limsup_{n\to\infty} E L_{\tilde{X}^n_{\text{univ}}}(X^n, Y^n) \\ &\ge \mathbb{D}(\boldsymbol{P}_{\boldsymbol{X}}, \mathcal{C}). \end{aligned} \quad (39)$$

The combination of (38) and (39) implies (31). □

V. A Generalized Scheme

We now generalize the scheme of the previous sections. Our idea is to pass the $(2k+1)$th-order empirical distribution of a noisy tuple through a transformation which maps distributions on $2k+1$-tuples at the channel output into the distribution of the channel input $2k+1$-tuple that gives rise to it. Assumption 1 guarantees that to every output distribution on $2k+1$-tuples there exists a unique channel input distribution. Therefore, the said transformation is unique on the set of output distributions that are *bona fide* in the sense of governing a $2k+1$ noisy tuple at the channel output. One would like, however, to extrapolate this transformation to be able to take *any* distribution of a $2k+1$ noisy tuple into a channel input distribution whose associated output distribution is "close" to the given one. The reason is that one can only hope for the $2k+1$th-order empirical distribution observed at the channel output to be "close" to the true distribution, but cannot expect it to be *bona fide* in the above sense. As we show later, this extrapolation is not unique. The scheme of the previous sections can be seen as a particular choice for this extrapolation from a larger family of schemes which we now develop.

Let $\Gamma = \{\gamma_a\}_{a\in\mathcal{A}}$ be a collection of measurable functions with the property that $A = \Gamma\mathcal{C}^T$ is invertible, namely, the $M \times M$ matrix whose (i,j)th entry is given by

$$A(i,j) = \int \gamma_i(y) f_j(y) dy \quad (40)$$

is nonsingular. For any $P_X \in \mathcal{M}(\mathcal{A})$ we then have

$$\begin{aligned} P_X &= A^{-1}AP_X && (41)\\ &= A^{-1}\Gamma\mathcal{C}^T P_X && (42)\\ &= A^{-1}\Gamma P_Y && (43) \end{aligned}$$

where P_Y is the distribution of the output distribution (thought of here as an infinite-dimensional column vector) when the channel input distribution is P_X. ΓP_Y is the M-dimensional column vector whose ath component is $E_{P_Y}\gamma_a(Y) = \int \gamma_a(y) dP_Y(y)$. It follows that if P_Y is the output distribution associated with the input distribution P_X then, for $u \in \mathcal{A}$

$$P_X(u) = \sum_{v\in\mathcal{A}} A^{-1}(u,v) \int \gamma_v(y) dP_Y(y). \quad (44)$$

Similarly, it can be seen that if $P_{Y_{-k}^{k}}$ is the channel output distribution of a $2k+1$-tuple when the input distribution is $P_{X_{-k}^{k}}$ then, for $u_{-k}^{k} \in \mathcal{A}^{2k+1}$

$$P_{X_{-k}^{k}}\left(u_{-k}^{k}\right) = \sum_{v_{-k}^{k}\in\mathcal{A}^{2k+1}} \prod_{j=-k}^{k} A^{-1}(u_j, v_j) \times \int \prod_{j=-k}^{k} \gamma_{v_j}(y_j) dP_{Y_{-k}^{k}}\left(y_{-k}^{k}\right). \quad (45)$$

Our approach to the estimation of $P^k_{x^n}$ is to substitute the empirical distribution of a $2k+1$-tuple at the channel output for $P_{Y_{-k}^{k}}$ on the

right-hand side of (45). The integral in (45) is then, up to normalization, given by

$$\alpha\left[y^n, v_{-k}^k\right] = \sum_{i=k+1}^{n-k} \prod_{j=-k}^{k} \gamma_{v_j}(y_{i+j}) \tag{46}$$

leading to our estimate of $P_{x^n}^k$ as

$$\bar{P}_{x^n}^k[y^n]\left(u_{-k}^k\right) = \frac{1}{n-2k} \sum_{v_{-k}^k \in \mathcal{A}^{2k+1}} \alpha\left[y^n, v_{-k}^k\right] \prod_{j=-k}^{k} A^{-1}(u_j, v_j). \tag{47}$$

We note that $\hat{P}_{x^n}^k[z^n]$ is obtained as a special case of $\bar{P}_{x^n}^k[y^n]$ under a particular Γ.

Example 1: Letting, for $a \in \mathcal{A}$

$$\gamma_a(y) = \begin{cases} 1, & \text{if } Q(y) = a \\ 0, & \text{otherwise} \end{cases} \tag{48}$$

the matrix A defined in (40) coincides with Π^T (for the matrix Π defined in (1)) and for this case

$$\alpha\left[y^n, v_{-k}^k\right] = \boldsymbol{r}\left[z^n, v_{-k}^k\right], \qquad \text{for all } v_{-k}^k \in \mathcal{A}^{2k+1}.$$

Substituting into (47) and comparing with (18) shows that in this case $\bar{P}_{x^n}^k[y^n]$ becomes $\hat{P}_{x^n}^k[z^n]$.

Indeed, extending (19) to

$$\tilde{P}_{x^n}^{k,\delta}\left(u_{-k}^k\right) = Q_\delta\left(\bar{P}_{x^n}^k[y^n]\left(u_{-k}^k\right)\right) \tag{49}$$

and defining $\hat{X}^{n,k,\delta}$ as in (20), Theorem 2 and, therefore, also Theorem 1 and Corollary 1, remain intact for this generalized scheme. This is because the following extension of Lemma 1 holds.

Claim 1: For all $n \geq 1$, $x^n \in \mathcal{A}^n$, and $\varepsilon > 0$

$$\begin{aligned} &\Pr\left(\left\|\bar{P}_{x^n}^k[Y^n] - P_{x^n}^k\right\|_\infty > \varepsilon\right) \\ &\quad \leq M^{2k+1} A\left(k, \varepsilon, \left(\Gamma_{\max} A_{\max}^{-1}\right)^{2k+1}\right) \\ &\qquad \cdot \exp\left(-G\left(k, \varepsilon, \left(\Gamma_{\max} A_{\max}^{-1}\right)^{2k+1}\right) n\right) \end{aligned} \tag{50}$$

where

$$\Gamma_{\max} = \max_a \sup_y \gamma_a(y) \quad \text{and} \quad A_{\max}^{-1} = \max_{i,j} A^{-1}(i,j). \tag{51}$$

The proof of Claim 1, which is very similar to that of Lemma 1, can be found in Appendix C. Equipped with Claim 1 it is easy to check that all the bounds and performance guarantees of previous sections remain intact for the scheme defined via (49), with the constant $\Pi_{\max}^{-1}$ replaced by $\Gamma_{\max} A_{\max}^{-1}$.

Example 2: Letting, for $a \in \mathcal{A}$

$$\gamma_a(y) = f_a(y) \tag{52}$$

results in

$$A(i,j) = \int f_i(y) f_j(y) dy \tag{53}$$

which is nonsingular under our standing Assumption 1. In fact, for this case, $A^{-1}\Gamma P_Y$ can be thought of as the Moore–Penrose inverse (cf., e.g., [5]) of the channel "matrix." $\bar{P}_{x^n}^k[y^n]$ in this case can be thought of as the Moore–Penrose inverse of the channel matrix associated with the $2k+1$th-order super-symbols (the $2k+1$-fold tensor product of the original channel matrix) applied to the $2k+1$th-order empirical distribution of the noisy signal.

Evidently, there is a lot of freedom in the choice of Γ. The only requirement for the asymptotic optimality of the induced sequence of schemes to be guaranteed is that the resulting matrix A (i.e., (40)) be nonsingular. The dependence of the bounds on Γ seems to imply that a good choice should strive to minimize $\Gamma_{\max} A_{\max}^{-1}$.

VI. Collapse to the DUDE for the Invertible Discrete Memoryless Channel (DMC)

Consider the more brute force scheme which uses $\hat{P}_{x^n}^k[z^n]$ instead of its processed version $\tilde{P}_{x^n}^{k,\delta}[z^n]$, i.e.,

$$\hat{X}^{n,k}[y^n](i) = g_{\text{opt}}\left[\hat{P}_{x^n}^k[z^n]\right]\left(y_{i-k}^{i+k}\right), \qquad k+1 \leq i \leq n-k. \tag{54}$$

Consider further the special case where the effective channel output alphabet is[3] $\mathcal{A}$, namely, the setting of [10]. We shall now show that, for this case, letting Q be the identity map (so that $\boldsymbol{Y} = \boldsymbol{Z}$), the scheme in (20) coincides with that of [10].

For this case, using the notation of [10], we get (55) and (56) at the bottom of the page, where (55) was established in [10] (cf., in particular, (9) therein), the $\propto$ notation indicating equality up to normalization of the vector whose ath component is given, and the equality in (56) follows from writing out the relationship

$$P_{z_{-k}^{-1}, Z_0, z_1^k}(b) = \sum_{u_{-k}^k} \left[\prod_{i \in \{-k,\ldots k\}\setminus 0} \Pi(u_i, z_i)\right] \Pi(u_0, b) P\left(U_{-k}^k = u_{-k}^k\right).$$

Evidently, for $k+1 \leq l \leq n-k$, we get (57) at the bottom of the following page, where $\boldsymbol{m}\left(z^n, z_{l-k}^{l-1}, z_{l+1}^{l+k}\right)$ is the notation from [10] for the noisy context statistics. Equality (a) follows from (56). For

[3]Note that, formally, this is outside the setting of output distributions given by densities but, as indicated, the schemes and results presented carry over to this case in an obvious way.

$$\sum_{u_{-k}^k \in \mathcal{A}^{2k+1}: u_0 = a} \left[\prod_{i=-k}^{k} \Pi(u_i, z_i)\right] P\left(U_{-k}^k = u_{-k}^k\right)$$

$$\propto \left[\boldsymbol{\pi}_{z_0} \odot \Pi^{-T} P_{z_{-k}^{-1}, Z_0, z_1^k}\right](a) \tag{55}$$

$$= \left[\boldsymbol{\pi}_{z_0} \odot \Pi^{-T} \begin{pmatrix} \sum_{u_{-k}^k} \left[\prod_{i \in \{-k,\ldots k\}\setminus 0} \Pi(u_i, z_i)\right] \Pi(u_0, 0) P\left(U_{-k}^k = u_{-k}^k\right) \\ \vdots \\ \sum_{u_{-k}^k} \left[\prod_{i \in \{-k,\ldots k\}\setminus 0} \Pi(u_i, z_i)\right] \Pi(u_0, M-1) P\left(U_{-k}^k = u_{-k}^k\right) \end{pmatrix}\right](a) \tag{56}$$

(b), we have used tensor product notation where $\Pi^{\otimes(2k+1)}$ denotes a $(2k+1)$-fold tensor power of the matrix Π, which is the channel matrix associated with symbols from the super-alphabet of $2k+1$-tuples under lexicographic ordering. The notation $\boldsymbol{r}[z^n, \cdot]$ denotes the M^{2k+1}-dimensional vector with components $\boldsymbol{r}[z^n, u_{-k}^k]$ (ordering the u_{-k}^k's lexicographically). From the expression on the righthand side of (57) it is clear that $\hat{X}^{n,k}$ is nothing but the kth-order DUDE of [10].

VII. Explicit Form of Denoiser for Binary Input Additive White Noise

A. BIAWGN Channel With Hamming Distortion

For this case $\mathcal{A} = \{-1, 1\}$ and f_a is the density of an $N(a, \sigma^2)$. The natural quantizer to employ here would seem to be

$$Q(y) = \begin{cases} 1, & \text{for } y \geq 0 \\ -1, & \text{otherwise} \end{cases} \tag{58}$$

so the induced channel from clean bit to quantized observation is a BSC$(G(1/\sigma))$, with $G(t)$ denoting the probability that a standard Gaussian be larger than t. For this case we get

$$\hat{P}_{x^n}^k[z^n]\left(u_{-k}^k\right) = \frac{1}{n-2k} \cdot \sum_{v_{-k}^k \in \mathcal{A}^{2k+1}} \boldsymbol{r}\left[z^n, v_{-k}^k\right] \left(\frac{\eta}{\eta - 1}\right)^{d_H\left(v_{-k}^k, u_{-k}^k\right)} \tag{59}$$

where d_H denotes Hamming distance and $\eta = \eta(\sigma) = G(1/\sigma)$. For this case

$$g_{\text{opt}}\left[\tilde{P}_{x^n}^{k,\delta}[z^n]\right]\left(y_{-k}^k\right) \tag{60}$$

$$= \arg\min_{\hat{x} \in \mathcal{A}} \sum_{a \in \mathcal{A}} \Lambda(a, \hat{x}) \cdot \left\{ \sum_{u_{-k}^k \in \mathcal{A}^{2k+1} : u_0 = a} \left[\prod_{i=-k}^{k} f_{u_i}(y_i)\right] \tilde{P}_{x^n}^{k,\delta}[z^n]\left(u_{-k}^k\right) \right\} \tag{61}$$

$$= \arg\min_{\hat{x} \in \mathcal{A}} \sum_{a \in \mathcal{A}} \Lambda(a, \hat{x}) \cdot \left\{ \sum_{u_{-k}^k \in \mathcal{A}^{2k+1} : u_0 = a} Q_\delta\left(\hat{P}_{x^n}^k[z^n]\left(u_{-k}^k\right)\right) \cdot \exp\left(-\frac{1}{2\sigma^2}\left\|y_{-k}^k - u_{-k}^k\right\|_2^2\right) \right\}. \tag{62}$$

$Q_\delta\left(\hat{P}_{x^n}^k[z^n]\left(u_{-k}^k\right)\right)$, for each u_{-k}^k, is calculated according to the right-hand side of (59). Note, however, that this computation need be performed once (after the first pass through the (quantized) data), when $\{\boldsymbol{r}[z^n, v_{-k}^k]\}_{v_{-k}^k}$ is available. Then, for the actual denoising, the $Q_\delta(\hat{P}_{x^n}^k[z^n](u_{-k}^k))$ in the summation in (62) are given constants.

B. Binary-Input Additive White pth-Power Noise With Hamming Distortion

The above example immediately generalizes to the case $\mathcal{A} = \{-1, 1\}$ and

$$f_a(x) = c(\alpha) e^{-\alpha |x-a|^p}, \qquad x \in \mathbb{R} \tag{63}$$

where $p > 0$, $\alpha > 0$, and $c(\alpha)$ is the normalization constant. Employing the quantizer in (58), the induced channel from clean bit to quantized observation is a BSC(δ), where

$$\eta = \eta(\alpha) = \int_1^\infty c(\alpha) e^{-\alpha x^p}.$$

In this case

$$g_{\text{opt}}\left[\tilde{P}_{x^n}^{k,\delta}[z^n]\right]\left(y_{-k}^k\right) = \arg\min_{\hat{x} \in \mathcal{A}} \sum_{a \in \mathcal{A}} \Lambda(a, \hat{x}) \cdot \left\{ \sum_{u_{-k}^k \in \mathcal{A}^{2k+1} : u_0 = a} Q_\delta\left(\hat{P}_{x^n}^k[z^n]\left(u_{-k}^k\right)\right) \cdot \exp\left(-\alpha \left\|y_{-k}^k - u_{-k}^k\right\|_p^p\right) \right\}. \tag{64}$$

$$\begin{aligned}
\hat{X}^{n,k}[y^n](l) &= g_{\text{opt}}\left[\hat{P}_{x^n}^k[z^n]\right]\left(y_{l-k}^{l+k}\right) \\
&= \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_{u_i}(y_{l+i})\right] \hat{P}_{x^n}^k[z^n]\left(u_{-k}^k\right) \right\} \\
&= \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} \Pi(u_i, z_{l+i})\right] \left[\sum_{v_{-k}^k \in \mathcal{A}^{2k+1}} \boldsymbol{r}\left[z^n, v_{-k}^k\right] \prod_{j=-k}^{k} \Pi^{-1}(v_j, u_j)\right] \right\} \\
&\overset{(a)}{=} \arg\min_{\hat{x} \in \mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \left[\boldsymbol{\pi}_{z_l} \odot \Pi^{-T} \begin{pmatrix} \sum_{u_{-k}^k} \left[\prod_{i \in \{-k, \ldots k\} \setminus 0} \Pi(u_i, z_{l+i})\right] \Pi(u_0, 0) \left[\sum_{v_{-k}^k} \boldsymbol{r}\left[z^n, v_{-k}^k\right] \prod_{j=-k}^{k} \Pi^{-1}(v_j, u_j)\right] \\ \vdots \\ \sum_{u_{-k}^k} \left[\prod_{i \in \{-k, \ldots k\} \setminus 0} \Pi(u_i, z_{l+i})\right] \Pi(u_0, M-1) \left[\sum_{v_{-k}^k} \boldsymbol{r}\left[z^n, v_{-k}^k\right] \prod_{j=-k}^{k} \Pi^{-1}(v_j, u_j)\right] \end{pmatrix} \right] \\
&\overset{(b)}{=} \arg\min_{\hat{x} \in \mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \left[\boldsymbol{\pi}_{z_l} \odot \Pi^{-T} \begin{pmatrix} \left[\left[\Pi^{\otimes(2k+1)}\right]^T \left[(\Pi^{-1})^{\otimes(2k+1)}\right]^T \boldsymbol{r}[z^n, \cdot]\right]\left(z_{l-k}^{l-1}, 0, z_{l+1}^{l+k}\right) \\ \vdots \\ \left[\left[\Pi^{\otimes(2k+1)}\right]^T \left[(\Pi^{-1})^{\otimes(2k+1)}\right]^T \boldsymbol{r}[z^n, \cdot]\right]\left(z_{l-k}^{l-1}, M-1, z_{l+1}^{l+k}\right) \end{pmatrix} \right] \\
&= \arg\min_{\hat{x} \in \mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \left[\boldsymbol{\pi}_{z_l} \odot \Pi^{-T} \begin{pmatrix} \boldsymbol{r}[z^n, \cdot]\left(z_{l-k}^{l-1}, 0, z_{l+1}^{l+k}\right) \\ \vdots \\ \boldsymbol{r}[z^n, \cdot]\left(z_{l-k}^{l-1}, M-1, z_{l+1}^{l+k}\right) \end{pmatrix} \right] \\
&= \arg\min_{\hat{x} \in \mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \left[\boldsymbol{\pi}_{z_l} \odot \Pi^{-T} \boldsymbol{m}\left(z^n, z_{l-k}^{l-1}, z_{l+1}^{l+k}\right) \right]
\end{aligned} \tag{57}$$

$$P^k_{x^n}\left(u^k_{-k}\right) - \hat{P}^k_{x^n}[Z^n]\left(u^k_{-k}\right) \tag{A1}$$

$$= \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left[\mathbf{1}_{\left\{x^{i+k}_{i-k}=u^k_{-k}\right\}} - \sum_{v^k_{-k}\in\mathcal{A}^{2k+1}} \mathbf{1}_{\left\{Z^{i+k}_{i-k}=v^k_{-k}\right\}} \prod_{j=-k}^{k} \Pi^{-1}(v_j,u_j)\right] \tag{A2}$$

$$= \frac{1}{n-2k}\sum_{m=0}^{2k}\sum_{\substack{i\in\{k+1,\ldots,n-k\},\\ \lceil (i-m)/(2k+1)\rceil=(i-m)/(2k+1)}}\left[\mathbf{1}_{\left\{x^{i+k}_{i-k}=u^k_{-k}\right\}} - \sum_{v^k_{-k}\in\mathcal{A}^{2k+1}} \mathbf{1}_{\left\{Z^{i+k}_{i-k}=v^k_{-k}\right\}} \prod_{j=-k}^{k} \Pi^{-1}(v_j,u_j)\right]. \tag{A3}$$

VIII. Sample of Experimental Results

In the following, we present one representative sample of initial experimental results. A comprehensive study of the performance of the proposed schemes, as well as good choices of the basis filters $\{\gamma_a\}_{a\in\mathcal{A}}$ for various data sets will be conducted separately.

A first-order symmetric binary Markov chain of transition probability 0.2 from one state to the other was simulated, and corrupted by a BIAWGN channel of unit variance. The scheme $\tilde{X}^{n,k}$ detailed in Section VII-A was used, on a sequence length of 10^5, for $k=1,2,3$. The following table encapsulates the results. The first line shows the error rates when employing a genie-aided scheme, which is informed of the true empirical distribution of order $2k+1$ of the clean source realization and operates optimally based on it. The second line shows the results of employing the optimal distribution-dependent filter basing its decisions on a noisy observation window of length $2k+1$ around each location. The third line contains the error rate of the optimum distribution-dependent scheme, which bases its decision for each location on all the observations (implemented via the backward–forward recursions). The fourth line details the performance of $\tilde{X}^{n,k}$.

$n=10^5$, $\theta=0.2$	k=1	k=2	k=3
Genie-aided	0.1148	0.1096	0.1082
Source-dependent	0.1148	0.1096	0.1083
Optimum	0.1081	0.1081	0.1081
$\tilde{X}^{n,k}$	0.1148	0.1094	0.1089

Similar experiments, with different noise-free process distributions and channel characteristics (not necessarily Gaussian) were conducted, with similar[4] results.

IX. Conclusion

We have presented a family of (conceptually and algorithmically) simple and universally optimal denoising algorithms for estimating the components of a finite-alphabet signal corrupted by a general memoryless channel. This extends the recent work [10] which presented universal denoisers for the case where the corrupting channel is a DMC. It was shown that our schemes, when specialized to the case of equal channel input and output symbols and an invertible DMC, coincide with the scheme in [10] and can thus be regarded a natural generalization of it.

Though the emphasis of our work was on the case where the channel output alphabet is the entire real line, our schemes apply just as well to the case of any finite channel output alphabet larger than the input alphabet (provided the channel matrix is of full row rank). Interestingly, they do not coincide in this case with the scheme suggested in [10, IV-C]. Indeed, one interesting direction for future work is a comparison of the performance of our denoisers to that of [10, Subsec. 3-C] for the case of a finite channel output alphabet larger than the input alphabet. We believe that our schemes will compare favorably in the sense that, for fixed k and n (i.e., before the asymptotics "kick in"), the excess loss relative to the class of kth-order sliding-window schemes (i.e., the "redundancy") will depend on the cardinality of the channel input alphabet for our schemes, while being dependent on the size of the channel output alphabet for the scheme of [10, Subsec. 3-C].

Additional directions of interest include the sequential denoising problem (filtering), and the case of channel uncertainty. For the filtering, a modified version of our scheme, where both collection of the counts and the actual symbol estimation would be performed concurrently in one pass, can be shown to work. Results in this direction for the discrete output case can be found in [7]. Regarding channel uncertainty, at first glance, our scheme may seem to accommodate uncertainty, since one quantizer can satisfy the requirement for the invertibility of the induced DMC (in (1)) simultaneously for a variety of different channels. The second pass, however, is channel dependent, since the estimate of the empirical distribution of the noiseless sequence (recall (18)) depends on the induced DMC, which depends on the original channel. It can be shown [2], [3] that, under channel uncertainty, attaining optimum noiseless-signal-dependent performance (in both the senses of Section III and of Section IV) is, in general, not feasible. Variations on the scheme developed in this work, to accommodate channel uncertainty, can be derived and shown optimal w.r.t. a minimax criterion, analogously to what was done for the DMC in [2], [3].

Finally, a natural next step is the problem where the input alphabet is also arbitrary. For sufficiently "well-behaved" channels, a slight variation on our scheme can be shown to work. The general case is under investigation.

Appendix A
Proof of Lemmas

A. Proof of Lemma 1

The first inequality is immediate from the definition of $\tilde{P}^{k,\delta}_{x^n}$. Turning to the second inequality, we fix $n\geq 1$, $x^n\in\mathcal{A}^n$, and $u^k_{-k}\in\mathcal{A}^{2k+1}$. Writing out the definitions for $P^k_{x^n}\left(u^k_{-k}\right)$ and $\hat{P}^k_{x^n}[Z^n]\left(u^k_{-k}\right)$ and interchanging order of summation we obtain (A1)–(A3) at the top of the page. It is easy to verify (cf., e.g., [9, Lemma 4]) that, for each i, the bracketed term is a zero-mean random variable. Furthermore, note that each of the inner sums in (A3) is a sum of at most $\frac{n-2k}{2k+1}$ *independent* random variables, bounded in magnitude by $\left(\Pi^{-1}_{\max}\right)^{2k+1}$. The reason for this independence is that each summand in the inner sum depends on the sequence Z^n only through Z^{i+k}_{i-k}, and Z^{i+k}_{i-k} is independent of $Z^{i'+k}_{i'-k}$ whenever $|i-i'|>2k$. Applying Hoeffding's inequality (cf., e.g., [1, Theorem 8.1]) gives for each $0\leq m\leq 2k$ and $\varepsilon>0$, (A4) at the top of the following page. Combining (A4) with (A3) and applying the union bound

$$\Pr\left(P^k_{x^n}\left(u^k_{-k}\right) - \hat{P}^k_{x^n}[Z^n]\left(u^k_{-k}\right) > \varepsilon\right) \leq (2k+1)\exp\left(-\frac{2\varepsilon^2(n-2k)}{(2k+1)\left(\Pi^{-1}_{\max}\right)^{4k+2}}\right). \tag{A5}$$

[4]Similar in the sense that the error rate of $\tilde{X}^{n,k}$ with $k=2$ and $k=3$ was already within a fraction of 1% from the optimum.

$$\Pr\left(\frac{1}{n-2k}\sum_{\substack{i\in\{k+1,\ldots,n-k\},\\ \lceil (i-m)/(2k+1)\rceil=(i-m)/(2k+1)}}\left[\mathbf{1}_{\left\{x_{i-k}^{i+k}=u_{-k}^{k}\right\}}-\sum_{v_{-k}^{k}\in\mathcal{A}^{2k+1}}\mathbf{1}_{\left\{Z_{i-k}^{i+k}=v_{-k}^{k}\right\}}\prod_{j=-k}^{k}\Pi^{-1}(v_j,u_j)\right]>\varepsilon\right)$$
$$\leq \exp\left(-\frac{2\varepsilon^2(n-2k)}{(2k+1)\left(\Pi_{\max}^{-1}\right)^{4k+2}}\right). \quad \text{(A4)}$$

Since a similar argument would give the same bound on

$$\Pr\left(P_{x^n}^k\left(u_{-k}^k\right)-\hat{P}_{x^n}^k[Z^n]\left(u_{-k}^k\right)<-\varepsilon\right)$$

we obtain

$$\Pr\left(\left|P_{x^n}^k\left(u_{-k}^k\right)-\hat{P}_{x^n}^k[Z^n]\left(u_{-k}^k\right)\right|>\varepsilon\right)$$
$$\leq 2(2k+1)\exp\left(-\frac{2\varepsilon^2(n-2k)}{(2k+1)\left(\Pi_{\max}^{-1}\right)^{4k+2}}\right) \quad \text{(A6)}$$

which, by another application of the union bound, gives (26). □

B. Proof of Lemma 2

By linearity of expectation,

$$\frac{1}{n-2k}\sum_{i=k+1}^{n-k}E\Lambda\left(x_i,g\left(Y_{i-k}^{i+k}\right)\right)=E_{P_{x^n}^k\otimes\mathcal{C}}\Lambda\left(U_0,g\left(Y_{-k}^k\right)\right).$$

Thus, the expression inside the absolute value brackets in (27) is a sum of zero-mean random variables, bounded in magnitude by $\Lambda_{\max}$. Furthermore,

$$\Lambda\left(x_i,g\left(Y_{i-k}^{i+k}\right)\right) \quad \text{and} \quad \Lambda\left(x_j,g\left(Y_{j-k}^{j+k}\right)\right)$$

are independent whenever $|i-j|>2k$. This allows the same decomposition of the sum as in the proof of Lemma 1 into $2k+1$ sums of independent, bounded, zero-mean variables, which, by a similar derivation, leads to (27). □

C. Proof of Lemma 3

By (9)

$$\left|E_{P\otimes\mathcal{C}}\Lambda(U_0,g(Y_{-k}^k))-E_{\tilde{P}\otimes\mathcal{C}}\Lambda(U_0,g(Y_{-k}^k))\right|$$
$$\leq\sum_{u_{-k}^k}\left|P\left(u_{-k}^k\right)-\tilde{P}\left(u_{-k}^k\right)\right|$$
$$\cdot\left[\int_{\mathbb{R}^{2k+1}}\Lambda(u_0,g(y_{-k}^k))\left[\prod_{i=-k}^{k}f_{u_i}(y_i)\right]dy_{-k}\ldots dy_k\right]$$
$$\leq\Lambda_{\max}\sum_{u_{-k}^k}\left|P\left(u_{-k}^k\right)-\tilde{P}\left(u_{-k}^k\right)\right|. \quad \text{(A7)}$$

□

Appendix B
Proof of Theorem 2

We let $\mathcal{F}_\delta^k$ denote the set of $\mathcal{A}^{2k+1}$-dimensional vectors with components in $[0,1]$ that are integer multiples of δ. Note that $\tilde{P}_{x^n}^{k,\delta}[z^n]\in\mathcal{F}_\delta^k$ for all z^n. Just as the notation $g_{\text{opt}}[P]$ (given by (16)) was extended to accommodate P's that are not *bona fide* simplex members, so we agree to extend other notation. Thus, for example, $E_{P\otimes\mathcal{C}}\Lambda\left(U_0,g\left(Y_{-k}^k\right)\right)$ should be understood as the right-hand side of (9) even when P is not a probability. It is readily verified that Lemma 3 continues to hold for general P, $\tilde{P}$ that are not necessarily probabilities. Finally, let

$$\mathcal{G}_\delta^k=\{g_{\text{opt}}[P]\}_{P\in\mathcal{F}_\delta^k}.$$

Now for the proof: We fix $n\geq 1$, $x^n\in\mathcal{A}^n$, and note that

$$\Pr\left(\sup_{g:\mathbb{R}^{2k+1}\to\mathcal{A}}\left|E_{\tilde{P}_{x^n}^{k,\delta}[Z^n]\otimes\mathcal{C}}\Lambda\left(U_0,g\left(Y_{-k}^k\right)\right)\right.\right.$$
$$\left.\left.-E_{P_{x^n}^k\otimes\mathcal{C}}\Lambda\left(U_0,g\left(Y_{-k}^k\right)\right)\right|>\varepsilon+\delta\Lambda_{\max}\right) \quad \text{(A8)}$$
$$\leq\Pr\left(\left\|\tilde{P}_{x^n}^{k,\delta}[Z^n]-P_{x^n}^k\right\|>\frac{\varepsilon}{\Lambda_{\max}}+\delta\right)$$
$$\leq M^{2k+1}A\left(k,\varepsilon/\Lambda_{\max},\left(\Pi_{\max}^{-1}\right)^{2k+1}\right)$$
$$\cdot\,e^{-G\left(k,\varepsilon/\Lambda_{\max},(\Pi_{\max}^{-1})^{2k+1}\right)n} \quad \text{(A9)}$$

where[5] the first inequality in (A9) follows from Lemma 3 and the second one from Lemma 1. Combining (A9) with Lemma 2 gives

$$\Pr\left(\left|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda\left(x_i,g\left(Y_{i-k}^{i+k}\right)\right)\right.\right.$$
$$\left.\left.-E_{\tilde{P}_{x^n}^{k,\delta}[Z^n]\otimes\mathcal{C}}\Lambda\left(U_0,g\left(Y_{-k}^k\right)\right)\right|>2\varepsilon+2\delta\Lambda_{\max}\right)$$
$$\leq A\left(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max}\right)e^{-G(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max})n}$$
$$+M^{2k+1}A\left(k,\varepsilon/(2\Lambda_{\max}),\left(\Pi_{\max}^{-1}\right)^{2k+1}\right)$$
$$\cdot\,e^{-G\left(k,\varepsilon/(2\Lambda_{\max}),(\Pi_{\max}^{-1})^{2k+1}\right)n}. \quad \text{(A10)}$$

By the union bound, (A10) guarantees that for any class $\mathcal{G}$

$$\Pr\left(\max_{g\in\mathcal{G}}\left|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda\left(x_i,g\left(Y_{i-k}^{i+k}\right)\right)\right.\right.$$
$$\left.\left.-E_{\tilde{P}_{x^n}^{k,\delta}[Z^n]\otimes\mathcal{C}}\Lambda(U_0,g(Y_{-k}^k))\right|>2\varepsilon+2\delta\Lambda_{\max}\right)$$
$$\leq|\mathcal{G}|\left[A\left(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max}\right)e^{-G(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max})n}\right.$$
$$+M^{2k+1}A\left(k,\varepsilon/(2\Lambda_{\max}),\left(\Pi_{\max}^{-1}\right)^{2k+1}\right)$$
$$\left.\cdot\,e^{-G\left(k,\varepsilon/(2\Lambda_{\max}),(\Pi_{\max}^{-1})^{2k+1}\right)n}\right]. \quad \text{(A11)}$$

Consequently

$$\Pr\left(\left|L_{\tilde{X}^n,k,\delta}(x^n,Y^n)-\min_{g\in\mathcal{G}_\delta^k}E_{\tilde{P}_{x^n}^{k,\delta}[Z^n]\otimes\mathcal{C}}\Lambda\left(U_0,g\left(Y_{-k}^k\right)\right)\right|\right.$$
$$\left.>2\varepsilon+2\delta\Lambda_{\max}\right)$$
$$=\Pr\left(\left|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda\left(x_i,g_{\text{opt}}\left[\tilde{P}_{x^n}^{k,\delta}[Z^n]\right]\left(Y_{i-k}^{i+k}\right)\right)\right.\right.$$
$$\left.-E_{\tilde{P}_{x^n}^{k,\delta}[Z^n]\otimes\mathcal{C}}\Lambda(U_0,g_{\text{opt}}\left[\tilde{P}_{x^n}^{k,\delta}[Z^n]\right]\left(Y_{-k}^k\right))\right|$$
$$\left.>2\varepsilon+2\delta\Lambda_{\max}\right) \quad \text{(A12)}$$

[5]The supremum inside the probability in (A8) is over measurable functions. Although this is an uncountable set, the associated event is measurable as $\hat{P}_{x^n}^k[Z^n]$ is finitely valued.

$$P^k_{x^n}\left(u^k_{-k}\right) - \bar{P}^k_{x^n}[Y^n]\left(u^k_{-k}\right) \tag{A24}$$

$$= \frac{1}{n-2k}\sum_{m=0}^{2k}\sum_{\substack{i\in\{k+1,\ldots,n-k\}\\ \lceil (i-m)/(2k+1)\rceil=(i-m)/(2k+1)}}\left[\mathbf{1}_{\left\{x^{i+k}_{i-k}=u^k_{-k}\right\}} - \sum_{v^k_{-k}\in\mathcal{A}^{2k+1}}\prod_{j=-k}^{k}\gamma_{v_j}(Y_{i+j})A^{-1}(u_j,v_j)\right]. \tag{A25}$$

$$\le \Pr\left(\max_{g\in\mathcal{G}^k_\delta}\left|\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda\left(x_i, g\left(Y^{i+k}_{i-k}\right)\right) - E_{\tilde{P}^{k,\delta}_{x^n}[Z^n]\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right)\right| > 2\varepsilon + 2\delta\Lambda_{\max}\right) \tag{A13}$$

$$\le |\mathcal{G}^k_\delta|\left[A\left(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max}\right)e^{-G(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max})n} + M^{2k+1}A\left(k,\varepsilon/(2\Lambda_{\max}),(\Pi^{-1}_{\max})^{2k+1}\right)\cdot e^{-G\left(k,\varepsilon/(2\Lambda_{\max}),(\Pi^{-1}_{\max})^{2k+1}\right)n}\right], \tag{A14}$$

where (A12) follows from the definition of $\tilde{X}^{n,k,\delta}$ and the fact that for any $P\in\mathcal{F}^k_\delta$

$$\min_{g\in\mathcal{G}^k_\delta} E_{P\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right) = E_{P\otimes\mathcal{C}}\Lambda\left(U_0, g_{\text{opt}}[P]\left(Y^k_{-k}\right)\right)$$

(A13) follows by the fact that $\tilde{P}^{k,\delta}_{x^n}[Z^n]\in\mathcal{F}^k_\delta$ and, therefore, $g_{\text{opt}}[\tilde{P}^{k,\delta}_{x^n}[Z^n]]\in\mathcal{G}^k_\delta$, and (A14) follows from (A11). It also follows, from (A9), that

$$\Pr\left(\left|\min_{g\in\mathcal{G}^k_\delta} E_{\tilde{P}^{k,\delta}_{x^n}[Z^n]\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right) - \min_{g\in\mathcal{G}^k_\delta} E_{P^k_{x^n}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right)\right| > \varepsilon+\delta\Lambda_{\max}\right) \le M^{2k+1}A\left(k,\varepsilon/\Lambda_{\max},\left(\Pi^{-1}_{\max}\right)^{2k+1}\right)\cdot e^{-G\left(k,\varepsilon/\Lambda_{\max},(\Pi^{-1}_{\max})^{2k+1}\right)n}. \tag{A15}$$

Combining (A15) and (A14) gives

$$\Pr\left(\left|L_{\tilde{X}^n,k,\delta}(x^n,Y^n) - \min_{g\in\mathcal{G}^k_\delta} E_{P^k_{x^n}\otimes\mathcal{C}}\Lambda(U_0, g(Y^k_{-k}))\right| > 4\varepsilon+4\delta\Lambda_{\max}\right) \le \left|\mathcal{G}^k_\delta\right|\left[A\left(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max}\right)e^{-G(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max})n} + M^{2k+1}A\left(k,\varepsilon/(2\Lambda_{\max}),\left(\Pi^{-1}_{\max}\right)^{2k+1}\right)\cdot e^{-G\left(k,\varepsilon/(2\Lambda_{\max}),(\Pi^{-1}_{\max})^{2k+1}\right)n}\right] + M^{2k+1}A\left(k,\varepsilon/\Lambda_{\max},\left(\Pi^{-1}_{\max}\right)^{2k+1}\right)\cdot e^{-G\left(k,\varepsilon/\Lambda_{\max},(\Pi^{-1}_{\max})^{2k+1}\right)n}. \tag{A16}$$

On the other hand, letting $P^k_{x^n}[\delta]$ denote the element in $\mathcal{F}^k_\delta$ closest (under L_∞ norm) to $P^k_{x^n}$

$$\left|D_k(x^n) - \min_{g\in\mathcal{G}^k_\delta} E_{P^k_{x^n}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right)\right| \tag{A17}$$

$$= \left|\min_{P\in\mathcal{M}(\mathcal{A}^{2k+1})} E_{P^k_{x^n}\otimes\mathcal{C}}\Lambda\left(U_0, g_{\text{opt}}[P]\left(Y^k_{-k}\right)\right) - \min_{g\in\mathcal{G}^k_\delta} E_{P^k_{x^n}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right)\right| \tag{A18}$$

$$\le \left|\min_{P\in\mathcal{M}(\mathcal{A}^{2k+1})} E_{P^k_{x^n}[\delta]\otimes\mathcal{C}}\Lambda\left(U_0, g_{\text{opt}}[P]\left(Y^k_{-k}\right)\right) - \min_{g\in\mathcal{G}^k_\delta} E_{P^k_{x^n}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right)\right| + \Lambda_{\max}\delta \tag{A19}$$

$$= \left|\min_{P\in\mathcal{F}^k_\delta} E_{P^k_{x^n}[\delta]\otimes\mathcal{C}}\Lambda\left(U_0, g_{\text{opt}}[P]\left(Y^k_{-k}\right)\right) - \min_{g\in\mathcal{G}^k_\delta} E_{P^k_{x^n}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right)\right| + \Lambda_{\max}\delta \tag{A20}$$

$$= \left|\min_{g\in\mathcal{G}^k_\delta} E_{P^k_{x^n}[\delta]\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right) - \min_{g\in\mathcal{G}^k_\delta} E_{P^k_{x^n}\otimes\mathcal{C}}\Lambda\left(U_0, g\left(Y^k_{-k}\right)\right)\right| + \Lambda_{\max}\delta \tag{A21}$$

$$\le \Lambda_{\max}\delta + \Lambda_{\max}\delta = 2\Lambda_{\max}\delta, \tag{A22}$$

where (A19) and (A22) follow from Lemma 3, and (A20) follows since the achiever of the minimum in the first term of (A19) is $P^k_{x^n}[\delta]$, which, by definition, is a member of $\mathcal{F}^k_\delta$. Finally, combining (A16) with (A22) gives

$$\Pr\left(|L_{\tilde{X}^n,k,\delta}(x^n,Y^n) - D_k(x^n)| > 4\varepsilon+6\delta\Lambda_{\max}\right) \le \left|\mathcal{G}^k_\delta\right|\left[A\left(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max}\right)e^{-G(k,\varepsilon+\delta\Lambda_{\max},\Lambda_{\max})n} + M^{2k+1}A\left(k,\varepsilon/(2\Lambda_{\max}),(\Pi^{-1}_{\max})^{2k+1}\right)\cdot e^{-G\left(k,\varepsilon/(2\Lambda_{\max}),(\Pi^{-1}_{\max})^{2k+1}\right)n}\right] + M^{2k+1}A\left(k,\varepsilon/\Lambda_{\max},(\Pi^{-1}_{\max})^{2k+1}\right)\cdot e^{-G\left(k,\varepsilon/\Lambda_{\max},(\Pi^{-1}_{\max})^{2k+1}\right)n}. \tag{A23}$$

The fact that both $A(k,\varepsilon,B)$ and $G(k,\varepsilon,B)$ are increasing in ε and decreasing in B, and the fact that $\left|\mathcal{G}^k_\delta\right| \le \left[\frac{1}{\delta}+1\right]^{M^{2k+1}}$ imply that the right-hand side of (A23) is upper-bounded by $\alpha(\varepsilon,k,\delta)$, as defined in (21). □

Appendix C
Proof of Claim 1

The proof is similar to that of Lemma 1. By the definitions of $P^k_{x^n}\left(u^k_{-k}\right)$ and $\bar{P}^k_{x^n}[Y^n]\left(u^k_{-k}\right)$ one can show (A24) and (A25) at the top of the page. It can now be verified that, for each i, the bracketed term is a zero-mean random variable and that the inner sums in (A25) are sums of *independent* random variables, bounded in magnitude by $\left(\Gamma_{\max}\Pi^{-1}_{\max}\right)^{2k+1}$. The proof is completed by applying Hoeffding's inequality on each of the inner sums. □

ACKNOWLEDGMENT

The authors are grateful to Taesup Moon for his help with the experimental results. They are also grateful to Erik Ordentlich and Sergio Verdú for helpful discussions.

REFERENCES

[1] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.
[2] G. Gemelos, S. Sigurjònsson, and T. Weissman, "Universal minimax binary image denoising under channel uncertainty," in *Proc. 11th Int. Conf. Image Processing, ICIP 2004*, Singapore, Oct. 24–27, pp. 997–1000.
[3] ——, "Universal minimax discrete denoising under channel uncertainty," in *Proc. Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 199.
[4] J. Hannan, "Approximation to Bayes risk in repeated play," in *Contributions to the Theory of Games, Ann. Math. Study*. Princeton, NJ: Princeton Univ. Press, 1957, vol. III, pp. 97–139.
[5] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*. Orlando, FL: Academic, 1985.
[6] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.
[7] E. Ordentlich, T. Weissman, M. J. Weinberger, A. Somekh-Baruch, and N. Merhav, "Discrete universal filtering through incremental parsing," in *Proc. 2004 Data Compression Conference (DCC'04)*, Snowbird, UT, Mar. 2004, pp. 352–361.
[8] E. Samuel, "An empirical Bayes approach to the testing of certain parametric hypotheses," *Ann. Math. Statist.*, vol. 34, no. 4, pp. 1370–1385, 1963.
[9] T. Weissman and N. Merhav, "Finite-delay lossy coding and filtering of individual sequences corrupted by noise," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 721–733, Mar. 2002.
[10] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 2–28, Jan. 2005.

Error Exponents for Hypothesis Testing of the General Source

Kiminori Iriyama

***Abstract*—In this correspondence, we consider the simple hypothesis testing problems for general sources in the sence of Han and Verdú. Recently Han established a compact formula for the supremum of achievable exponents for the second-kind of error probability under the asymptotic constraint of the form $\mu_n \sim e^{-nr}$ $(n \to \infty)$ on the first-kind of error probability μ_n, where r is a given positive number. We investigate the same hypothesis testing problems studied by Han. The aim of the correspondence is to give a new expression for the supremum of achievable error exponents. Our formula is expressed in terms of the divergences and given in quite different forms from Han's expression.**

***Index Terms*—Abstract alphabet, divergence, error exponent, general source, hypothesis testing, information spectrum, large deviation.**

I. INTRODUCTION

In this correspondence, we consider the simple hypothesis testing problems for general sources in the sence of Han and Verdú [5]. We investigate the same problems studied by Han [3]. Let $\boldsymbol{X} = \{X^n\}_{n=1}^{\infty}$ and $\overline{\boldsymbol{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ be general sources taking values in the same source alphabets $\{\mathcal{X}_n\}_{n=1}^{\infty}$. Han [3] has defined the general hypothesis tesiting problem with $\boldsymbol{X}$ as the null hypothesis and $\overline{\boldsymbol{X}}$ as the alternative hypothesis. One of the basic problems is to determine the supremum $B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ of achievable exponents for the second-kind of error probability λ_n under the asymptotic constraint of the form

$$\mu_n \sim e^{-nr}, \qquad n \to \infty \tag{1}$$

on the first-kind of error probability μ_n, where $r > 0$ is a prescribed arbitrary constant. The quantity $B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ is called the supremum of r-achievable error exponents. Another basic problem is to determine the infimum $B_e^*(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ of achievable exponents for the second-kind of correct probability $1 - \lambda_n$ under the asymptotic constraint (1).

Han [3] proposed these problems and established general formulas (see Theorems 1 and 3) for $B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ and $B_e^*(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$. It may be of interest to know that we can give different formulas for these quantities. The main aim of the correspondence is to derive new formulas for $B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ and $B_e^*(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$, which are expressed in terms of divergence and given in quite different forms from the Han's expressions. Our results are stated in Theorem 2 and Theorem 4.

The quantities $B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ and $B_e^*(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ have been studied for stationary memoryless sources (SMSs), Markov sources and stationary Gaussian sources (see [3] and references therein), where $\boldsymbol{X} = \{X^n\}$ is said to be a SMS if $\{X_i\}_{i=1}^{\infty}$ is a stationary memoryless process and $X^n = (X_1, \ldots, X_n)$. If $\boldsymbol{X}$ and $\overline{\boldsymbol{X}}$ are SMSs, then it is known (cf. [3], [4], [6], [11]) that $B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ is given by

$$B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}}) = \inf_{Y: D(Y\|X_1) < r} D(Y\|\overline{X}_1) \tag{2}$$

where $D(Y\|X)$ denotes the divergence of Y with respect to X. In the case of Markov sources (cf. [3], [12]) and stationary Gaussian sources ([1], [7]) it is also shown that $B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ is expressed in terms of the divergence. It should be emphasized that our formula (Theorem 2) for $B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ seems to be a natural extension of the formula (2) for the SMS.

Han [3] has shown that his results can be applied to obtain the formulas of $B_e(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ and $B_e^*(r|\boldsymbol{X}\|\overline{\boldsymbol{X}})$ of SMSs and Markov sources. Our theorems can also be applied to SMSs and Markov sources (see Examples 1 and 2).

The main theorems are stated in Section II, and the proofs are given in Section III. We apply our main theorems to some special cases in Section IV.

II. MAIN RESULTS

Let $\boldsymbol{X} = \{X^n\}_{n=1}^{\infty}$, $\overline{\boldsymbol{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ and $\boldsymbol{Y} = \{Y^n\}_{n=1}^{\infty}$ be general sources, where X^n, $\overline{X}^n$ and Y^n are random variables taking values in $\mathcal{X}_n$ and $(\mathcal{X}_n, \mathcal{B}_n)$ is a measurable space. Note that the random sequence $\boldsymbol{X} = \{X^n\}$ is not necessarily stationary nor consistent. We denote by P_X the probability distribution of a random variable X.

Our results are expressed in terms of the following divergences. The divergence $D(Y^n\|X^n) \equiv D(P_{Y^n}\|P_{X^n})$ of Y^n with respect to X^n is defined by

$$D(Y^n\|X^n) = \int_{\mathcal{X}_n} \log \frac{dP_{Y^n}}{dP_{X^n}}(x) dP_{Y^n}(x)$$

if P_{Y^n} is absolutely continuous with respect to P_{X^n}, otherwise $D(Y^n\|X^n) = \infty$. We define $D_u(\boldsymbol{Y}\|\boldsymbol{X})$ and $D_l(\boldsymbol{Y}\|\boldsymbol{X})$ by

$$D_u(\boldsymbol{Y}\|\boldsymbol{X}) = \limsup_{n\to\infty} \frac{1}{n} D(Y^n\|X^n)$$

$$D_l(\boldsymbol{Y}\|\boldsymbol{X}) = \liminf_{n\to\infty} \frac{1}{n} D(Y^n\|X^n).$$

Manuscript received April 30, 2002; revised December 7, 2004.

The author is at 1-3 Mukaiyama-chou, Ichinomiya-shi, Aichi 491-0869, Japan.

Communicated by P. Narayan, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2004.842774

0018-9448/$20.00 © 2005 IEEE