Universal Discrete Denoising: Known Channel

Tsachy Weissman, *Member, IEEE*, Erik Ordentlich, *Member, IEEE*, Gadiel Seroussi, *Fellow, IEEE*, Sergio Verdú, *Fellow, IEEE*, and Marcelo J. Weinberger, *Senior Member, IEEE*

***Abstract*—A discrete denoising algorithm estimates the input sequence to a discrete memoryless channel (DMC) based on the observation of the entire output sequence. For the case in which the DMC is known and the quality of the reconstruction is evaluated with a given single-letter fidelity criterion, we propose a discrete denoising algorithm that does not assume knowledge of statistical properties of the input sequence. Yet, the algorithm is universal in the sense of asymptotically performing as well as the optimum denoiser that knows the input sequence distribution, which is only assumed to be stationary. Moreover, the algorithm is universal also in a semi-stochastic setting, in which the input is an individual sequence, and the randomness is due solely to the channel noise. The proposed denoising algorithm is practical, requiring a linear number of register-level operations and sublinear working storage size relative to the input data length.**

***Index Terms*—Context models, denoising, discrete filtering, discrete memoryless channels (DMCs), individual sequences, noisy channels, universal algorithms.**

> "*If the source already has a certain redundancy and no attempt is made to eliminate it... a sizable fraction of the letters can be received incorrectly and still reconstructed by the context.*" Claude Shannon, 1948

I. Introduction

CONSIDER the problem of estimating a signal $\{X_t\}_{t\in T}$ from a noisy version $\{Z_t\}_{t\in T}$, which has been corrupted by a memoryless channel. The estimation is assumed to depend on the *entire* signal $\{Z_t\}_{t\in T}$. In the Shannon paradigm [59] redundancy is added to the noiseless signal in order to protect it from the channel noise, and a decoder that knows the codebook and the channel statistics can recover the noiseless signal with arbitrary reliability, provided that the coding scheme respects the fundamental limits of information theory. In contrast, situations abound where no channel coding is performed and the recovery of the noiseless signal can only be accomplished with a certain distortion. This problem, for various types of index sets T, input–output alphabets, and channels, arises naturally in a wide range of applications spanning fields such as statistics, engineering, computer science, image processing, astronomy, biology, cryptography, and information theory.

The continuous case, where the input and output alphabets are the real line (or other Euclidean spaces), has received significant attention for over half a century. From the linear filters of Wiener [75], [6] and Kalman [34], to Donoho and Johnstone's nonlinear denoisers [20], [22], the amount of work and literature in between is far too extensive even to be given a representative sample of references. In fact, the practice of denoising, as influenced by the theory, at least for the problem of one-dimensionally indexed data corrupted by additive Gaussian white noise, is believed by some to have reached a point where substantial improvement in performance is unlikely for most applications of interest [8].

Less developed are the theory and practice of denoising for the case where the alphabet of the noiseless, as well as that of the noise-corrupted signal, are finite. The problem arises in a variety of situations ranging from typing and/or spelling correction [1], [15], [40] to hidden Markov model (HMM) state estimation (cf. [24] for the many applications); from DNA sequence analysis and processing [60], [63], [64] to enhancement of facsimile and other binary images; from blind equalization problems to joint source–channel decoding when a discrete source is sent uncompressed (or suboptimally compressed) through a noisy channel [9], [45] (and references therein).

A commonly analyzed denoising setting is one in which the underlying noiseless signal and noisy channel are assumed to be stochastic with known distributions. It is assumed that the goal of a denoising algorithm is to minimize the expected distortion of its output with respect to the unobserved noiseless signal, where the distortion is measured by a single-letter loss function. In such a Bayesian setting, the joint distribution of the noiseless and noisy signals can be obtained. The latter, in turn, gives rise to the posterior distribution of the noiseless signal conditioned on the noisy observation signal, which determines an optimal denoiser achieving the above minimization. Thus, though it may not always be practical to explicitly obtain this posterior distribution, in principle it is computable from the statistical descriptions of the input and the channel.

Certain instances of the discrete denoising problem have been extensively studied, particularly in the context of state

Manuscript received February 10, 2003; revised September 28, 2004. The material in this paper was presented in part at the 2002 IEEE Information Theory Workshop, Bangalore, India, at the 2003 IEEE International Symposium on Information Theory, Yokohama, Japan, and at the 2003 IEEE International Conference on Image Processing, Barcelona, Catalonia, Spain. The work of T. Weissman was supported in part by the National Science Foundation under Grant CCR-0312839. Part of this work was performed while T. Weissman was with Hewlett-Packard Laboratories and S. Verdú was a Hewlett-Packard/Mathematical Sciences Research Institute (MSRI) Visiting Research Professor.

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

E. Ordentlich, G. Seroussi, and M. J. Weinberger are with the Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA (e-mail: eord@hpl.hp.com; seroussi@hpl.hp.com; marcelo@hpl.hp.com).

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: verdu@princeton.edu).

Communicated by A. B. Nobel, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2004.839518

estimation for hidden Markov processes (cf. [24] and the many references therein). Indeed, for the case where the states evolve according to a known Markov process and the channel (from state to observation) is known, the optimum Bayesian scheme can be implemented with reasonable complexity via forward–backward dynamic programming [13], [4]. It should be mentioned, however, that even for the simplest among cases where the underlying signal has memory, namely, the case of a binary-symmetric Markov chain observed through a binary-symmetric channel (BSC), the bit-error rate of the optimal denoiser is not explicitly known for all values of the transition probability and the channel error rate; only the asymptotic behavior of the bit-error rate, as the transition probabilities become small [35], [61], and conditions for the optimality of "singlet decoding" (cf. [19], [47]), are known.

In this work, we address a *universal* version of the discrete denoising problem in which there is uncertainty about the distribution of the underlying noiseless signal, so that the posterior distribution on which the optimal Bayesian denoiser is based is not available. We are thus interested in denoisers which operate independently of the noiseless signal distribution and consider the following basic questions.

1) *Theoretical.* How well can a distribution-independent denoiser perform? Can it attain, universally, the performance of the best distribution-dependent denoiser?
2) *Algorithmic.* If we can answer the previous question in the affirmative, can we find a practically implementable universal denoiser? What is its complexity?

To study these questions, we restrict our attention to the case of finite alphabets and a known discrete memory channel (DMC) whose transition probability matrix has full rank.[1] In this case, the distribution of the channel output uniquely determines the distribution of the input.

The main contribution of this work is a discrete denoising algorithm performing favorably from both the theoretical and the algorithmic viewpoints. Specifically, we propose and analyze an algorithm with the following properties.

1) The algorithm is asymptotically optimal in
   a) *The semi-stochastic setting*. In this setting, we make no assumption on a probabilistic or any other type of mechanism that may be generating the underlying noiseless signal and assume it to be an "individual sequence" unknown to the denoiser. The randomness in this setting is due solely to the channel noise. We show that for every underlying individual sequence, our denoising algorithm is guaranteed to attain the performance of the best finite-order sliding-window denoiser, tuned to the noiseless sequence and the observed noisy sequence. Competing with finite-order sliding-window denoisers is similar to the setting introduced in universal lossless coding by Ziv and Lempel (LZ) [79].
   b) *The stochastic setting*. We show that our denoising algorithm asymptotically attains the performance of the optimal distribution-dependent scheme, for any stationary source that may be generating the underlying signal. This property follows easily from the result in the semi-stochastic setting.
2) The algorithm is practical. Implementation of the denoiser requires a linear number of register-level operations, and working storage complexity which is sublinear in the data size. *Register-level operations* are arithmetic and logic operations, address computations, and memory references, on operands of size $O(\log n)$ bits, where n is the input size. *Working storage* refers to the memory required by the algorithm for its internal data structures, book-keeping, etc.

The proposed universal denoising algorithm is the first discrete denoiser to provably attain the distribution-dependent optimum performance in either the semi-stochastic or stochastic universal settings described above (and in more detail later in Sections V and VI).

For concreteness and simplicity of the exposition, we assume one-dimensionally indexed data, though all our results can be readily extended to the multidimensional case. In fact, Section VIII presents experimental results for two-dimensional image denoising, and the multidimensional formalism is discussed in more detail in [46]. For the sake of clarity, most of the presentation is given for the case where the channel input and output alphabets are identical. In Section IV-C, it is indicated how our algorithm and results carry over to the general case where this condition might not hold.

The proposed denoising algorithm makes two passes over the noisy observation sequence. For a fixed k, counts of the occurrences of all the strings of length $2k+1$ appearing along the noisy observation sequence are accumulated in the first pass. The actual denoising is done in the second pass where, at each location along the noisy sequence, an easily implementable metric computation is carried out (based on the known channel matrix, the loss function, and the counts acquired in the previous pass) to determine what the denoised value of the symbol at that location should be. A judicious choice of k (as a function of the sequence length) yields a denoiser with the claimed properties.

Our work is conceptually connected with some of the previous literature on universal denoising and related problems. Most closely connected to our stochastic and semi-stochastic settings are the *empirical Bayes* and *compound decision* methods, respectively, from the statistics literature [51], [31], [53]–[55], [57], [58] (cf. [77] for a more comprehensive list of references). Most of the work on the compound decision problem has focused on competing with a "symbol-by-symbol" denoiser, and can be viewed as a particularization of our semi-stochastic setting to the case $k = 0$. Under the rubric of k-extended compound decision problems, competition with higher order sliding-window denoisers was also considered to a limited extent in that framework [2], [3], [66], [67]. After concretely describing the problem and stating our main results for the semi-stochastic setting in Section V, we elaborate on the connections between the problems and on how our results can be regarded as contributions to the compound decision problem.

[1]Here and throughout, by "full rank" we mean "full row-rank."

One basic element common to our denoiser and the approach in compound decision theory is the step of estimating the empirical distribution of the noiseless signal $\{X_t\}$ from the observed statistics of the noisy signal $\{Z_t\}$. This step has also been at the heart of an algorithm proposed in [1] for learning the empirical distribution of m-blocks in $\{X_t\}$ by solving a system of linear equations involving the empirical distribution of m-blocks in $\{Z_t\}$, similar to the k-extended compound decision problem. In contrast to this blockwise approach, our denoiser proceeds in symbol-by-symbol fashion using conditional marginal distributions of $\{Z_t\}$. The resulting estimate of the distribution of m-blocks in the noiseless signal is used in [1] to build a finite-state tree model [68] of $\{X_t\}$, whose accuracy is analyzed when $\{X_t\}$ is assumed to be distributed according to an arbitrary Markov source of order bounded by $m-1$. Empirical results are provided in [1] for text correction using dynamic programming to compute the likeliest state sequence given $\{Z_t\}$ under the assumption that $\{X_t\}$ is distributed according to the learned finite-state tree model.

The recovery of noise-corrupted signals generated by unknown deterministic dynamical systems is considered in [36]–[38] in a continuous alphabet setting. Conditions are found under which signals generated by classes of unknown discrete-time dynamical systems can be recovered in an asymptotically lossless fashion after being corrupted by additive memoryless noise known only to have zero-mean and bounded support.

Compression-based approaches for denoising have been proposed in a variety of universal settings, involving both continuous and discrete signals, in several papers including [11], [12], [21], [42]–[44], [50], [63], [64], and references therein. The intuition motivating the compression-based approach is that the noise constitutes that part of the noisy signal which is hardest to compress. Thus, by lossily compressing the noisy signal and appropriately tuning the fidelity level of the compressor to match the noise level, it may be expected that the part of the noisy signal that will be lost will mainly consist of the noise, so that the reconstructed signal will, in effect, be a denoised version of the observation signal. A variant of the compression-based approach to denoising is formalized and analyzed in [21] as one of the applications of the "Kolmogorov sampler." At the heart of this denoising algorithm is an optimal (in the rate-distortion sense) *universal* lossy compression of the noisy signal. The resulting denoiser is shown in [21] (cf. also [74]) to achieve an asymptotic distortion that is bounded strictly away from (but is within a factor of two of) the distortion achieved by the optimum distribution-dependent denoiser in certain instances of the stochastic setting (see 1a) described above. In addition, the complexity of the "Kolmogorov sampler" is at least that of optimal universal lossy compression for which no known computationally efficient algorithms exist (cf. [5, Sec. VI]). Compression-based schemes for denoising under source uncertainty have been also considered from a somewhat different perspective in [16], [69]. These papers are concerned with rate-distortion coding of noisy sources and characterize tradeoffs that are universally attainable between denoising performance and the rate constraint. An essential difference between the above mentioned compression-based approach and the setting of [16], [69] is that while in the former the rate constraint is imposed as a tool to facilitate denoising, in the latter this constraint is assumed real and the goal is to obtain a (Shannon-theoretic) characterization of optimum performance subject to this constraint.

The remainder of the paper is organized as follows. Section II presents our notation and conventions. In Section III, we show the form of the optimal nonuniversal denoiser in a stochastic Bayesian setting. In Section IV, we describe the proposed denoiser, analyze its complexity, and explain, at an intuitive level, why it can indeed be expected to attain universal optimality. Although the form of the denoiser is motivated in the Bayesian setting of Section III, the most powerful optimality results for the universal denoiser are those pertaining to a semi-stochastic setting where the input to the channel is an unknown individual sequence. This analysis is presented in Section V. The results for the fully stochastic setting, where the noiseless sequence is assumed emitted by a probabilistic source, then follow using standard techniques in Section VI. In Section VII, we discuss some theoretical and practical aspects of the choice of context model size for the denoiser. In Section VIII, we report the results of experiments in which our algorithm was employed on simulated data, English text, and images. We also briefly discuss some additional practical aspects of the implementation, as well as possible theoretical and practical extensions.

II. Notation and Conventions

Throughout we assume that the components of the noiseless signal, as well as those of the noisy observation sequence and the reconstruction sequence, take their values in an M-letter alphabet $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$. We will sometimes use elements of $\mathcal{A}$ as indices to M-vectors and $M \times M$ matrices, in which cases we identify a symbol with its index in the alphabet. The simplex of M-dimensional column probability vectors will be denoted by $\mathcal{M}$.

As stated in the Introduction, we assume a given channel whose transition probability matrix $\mathbf{\Pi} = \{\Pi(i,j)\}_{i,j\in\mathcal{A}}$ is known to the denoiser. Here, $\Pi(i,j)$ denotes the probability of output symbol j when the input is i. Moreover, we extend this notation to subsets $J \subseteq \mathcal{A}$, by denoting

$$\Pi(i,J) = \sum_{j\in J} \Pi(i,j).$$

We also assume a given loss function (fidelity criterion) $\Lambda : \mathcal{A}^2 \to [0,\infty)$, represented by the matrix $\mathbf{\Lambda} = \{\Lambda(i,j)\}_{i,j\in\mathcal{A}}$, where $\Lambda(i,j)$ denotes the loss incurred by estimating the symbol i with the symbol j. The maximum single-letter loss will be denoted $\Lambda_{\max} = \max_{i,j\in\mathcal{A}} \Lambda(i,j)$. We let π_i denote the ith column of $\mathbf{\Pi}$, and λ_j denote the jth column of $\mathbf{\Lambda}$. Hence, we have

$$\mathbf{\Pi} = [\pi_1 \mid \cdots \mid \pi_M], \qquad \mathbf{\Lambda} = [\lambda_1 \mid \cdots \mid \lambda_M].$$

Note that the columns of the channel transition probability matrix need not be probability vectors (though all the rows are).

For a vector or matrix $\mathbf{\Gamma}$, $\mathbf{\Gamma}^T$ will denote transposition and, for an invertible matrix, $\mathbf{\Gamma}^{-T}$ will denote the transpose of $\mathbf{\Gamma}^{-1}$.

The ith component of a vector $\mathbf{u}$ will be denoted by u_i or, when indexing some vector-valued expressions, $\mathbf{u}[i]$ (in the case of a probability vector $\mathbf{P}$, the ith component is denoted $\mathbf{P}(i)$). For M-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$, $\mathbf{u} \odot \mathbf{v}$ will denote the vector obtained from componentwise multiplication, i.e.,

$$(\mathbf{u} \odot \mathbf{v})[i] = u_i v_i.$$

In terms of order of operations, $\odot$ will have the usual multiplicative precedence over addition and subtraction. Notice that for column M-vectors $\mathbf{u}, \mathbf{v}$, and $\mathbf{w}$, we have

$$(\mathbf{u} \odot \mathbf{v})^T \cdot \mathbf{w} = \mathbf{u}^T \cdot (\mathbf{v} \odot \mathbf{w}). \tag{1}$$

The L_p norm of any vector $\mathbf{v}$ will be denoted by $\|\mathbf{v}\|_p$. Similarly, following standard conventions (cf., e.g., [28]), $\|\mathbf{A}\|_p$ will denote the L_p matrix norm of $\mathbf{A}$ defined by

$$\|\mathbf{A}\|_p = \sup_{\|\mathbf{v}\|_p = 1} \|\mathbf{A}\mathbf{v}\|_p$$

with $\mathbf{v}$ denoting a column vector. The notation $|\cdot|$ will be used to denote both absolute value and cardinality, according to whether the argument is real- or set-valued.

We let $\mathcal{A}^\infty$ denote the set of one-sided infinite sequences with $\mathcal{A}$-valued components, i.e., $\mathbf{a} \in \mathcal{A}^\infty$ is of the form $\mathbf{a} = (a_1, a_2, \ldots), a_i \in \mathcal{A}, i \geq 1$. For $\mathbf{a} \in \mathcal{A}^\infty$, let $a^n = (a_1, \ldots, a_n)$ and $a_i^j = (a_i, \ldots, a_j)$. More generally, we will allow the indices of vector components to be negative as well, so, for example, $u_{-k}^k = (u_{-k}, \ldots, u_0, \ldots u_k)$. For positive integers k_1, k_2 and strings $s_i \in \mathcal{A}^{k_i}$, we let $s_1 s_2$ denote the string of length $k_1 + k_2$ formed by concatenation.

For $\mathbf{P} \in \mathcal{M}$, let

$$U(\mathbf{P}) = \min_{\hat{x} \in \mathcal{A}} \sum_{a \in \mathcal{A}} \Lambda(a, \hat{x}) \mathbf{P}(a) = \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{P} \tag{2}$$

denote the *Bayes envelope* (cf., e.g., [30], [56], [41]) associated with the distribution $\mathbf{P}$ and the loss function Λ. Following [30], it will be convenient to extend the definition of $U(\cdot)$ to cases in which the argument is *any* M-vector $\mathbf{v}$, not necessarily in the simplex $\mathcal{M}$. We denote the minimizing symbol $\hat{x}$ in (2), namely, the *Bayes response* to $\mathbf{v}$, by $\hat{x}(\mathbf{v})$, i.e.,

$$\hat{x}(\mathbf{v}) = \arg\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{v} \tag{3}$$

where throughout $\arg\min_{\hat{x} \in \mathcal{A}}$ ($\arg\max_{\hat{x} \in \mathcal{A}}$) denotes the minimizing (maximizing) argument, resolving ties by taking the letter in the alphabet with the lowest index.

An *n-block denoiser* is a mapping $\hat{X}^n : \mathcal{A}^n \to \mathcal{A}^n$. We let $L_{\hat{X}^n}(x^n, z^n)$ denote the normalized cumulative loss, as measured by $\mathbf{\Lambda}$, of the denoiser $\hat{X}^n$ when the observed sequence is $z^n \in \mathcal{A}^n$ and the underlying noiseless one is $x^n \in \mathcal{A}^n$, i.e.,

$$L_{\hat{X}^n}(x^n, z^n) = \frac{1}{n} \sum_{i=1}^n \Lambda(x_i, \hat{X}^n(z^n)[i]). \tag{4}$$

Additional notation will be introduced in the sequel where needed.

III. Nonuniversal Discrete Denoising

In this section, we consider a stochastic Bayesian setting in which the distribution of the input is known and the optimality criterion is the expected loss between the noiseless and denoised sequences. The optimal denoiser for this setting is readily seen to output the symbol that minimizes the expected loss, given the available noisy observations. In other words, the output is the Bayes response (3) to the posterior distribution of the noiseless symbol given the noisy sequence. The optimal nonuniversal denoiser that we obtain in this section will play a role in the motivation of the structure of the universal denoiser.

Denoting by $\mathbf{P}_{X_t|z^n}$ the column M-vector whose αth component is $\Pr(X_t = \alpha | Z^n = z^n)$ according to the given input and channel distributions, the optimal nonuniversal denoiser for the symbol in position t is the Bayes response to $\mathbf{P}_{X_t|z^n}$, namely

$$\hat{X}^{\mathsf{opt}}(z^n)[t] = \arg\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{P}_{X_t|z^n} = \arg\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{P}_{X_t, z^n} \tag{5}$$

where the M-vector $\mathbf{P}_{X_t, z^n}$ is equal to the M-vector $\mathbf{P}_{X_t|z^n}$ multiplied by the scalar $\Pr(Z^n = z^n)$. Thus, the two vectors can be used interchangeably where scaling is immaterial, as in (5).

The conditional marginals $\mathbf{P}_{X_t|z^n}$ of the input given the observed process can be computed from the distribution of the input process and the channel transition probability matrix $\mathbf{\Pi}$. Alternatively, when $\mathbf{\Pi}$ is invertible we can derive an equivalent denoiser which uses only the conditional marginals of the output given the other outputs. The structure of the universal denoiser will be motivated by this alternative.

For any sequence u_m^n and integer t such that $m \leq t \leq n$, denote by $u_m^{n \backslash t}$ the sequence $u_m^{t-1} u_{t+1}^n$. As usual, we omit the subscript m when $m = 1$. By the memorylessness of the channel, we have that for every t

$$\begin{aligned} &\Pr(X_t = x_t, Z_t = z_t, Z^{n \backslash t} = z^{n \backslash t}) \\ &\quad = \Pr(X_t = x_t, Z^{n \backslash t} = z^{n \backslash t}) \, \Pi(x_t, z_t) \end{aligned} \tag{6}$$

or, in vector notation

$$\mathbf{P}_{X_t, z^n} = \pi_{z_t} \odot \mathbf{P}_{X_t, z^{n \backslash t}}. \tag{7}$$

Here, $\mathbf{P}_{X_t, z^{n \backslash t}}$ denotes the M-vector whose αth component is $\Pr(X_t = \alpha, Z^{n \backslash t} = z^{n \backslash t})$. Furthermore, marginalizing (6) with respect to X_t and iterating over all possible $z_t \in \mathcal{A}$ we obtain

$$\mathbf{P}_{Z_t, z^{n \backslash t}} = \mathbf{\Pi}^T \mathbf{P}_{X_t, z^{n \backslash t}}. \tag{8}$$

Putting together (7) and (8) yields

$$\mathbf{P}_{X_t, z^n} = \pi_{z_t} \odot \left[\mathbf{\Pi}^{-T} \mathbf{P}_{Z_t, z^{n \backslash t}} \right] \tag{9}$$

which together with (5) and (1) leads to the following nonuniversal estimator that minimizes the expected loss:

$$\hat{X}^{\mathsf{opt}}(z^n)[t] = \arg\min_{\hat{x} \in \mathcal{A}} \left[\mathbf{P}_{Z_t, z^{n \backslash t}} \right]^T \mathbf{\Pi}^{-1} \left[\lambda_{\hat{x}} \odot \pi_{z_t} \right]. \tag{10}$$

Note that the foregoing derivation of the nonuniversal denoiser has used the memoryless nature of the channel but has not assumed any statistical properties on the input process.

IV. The Discrete Universal DEnoiser (DUDE)

In this section, we present our Discrete Universal DEnoiser (DUDE). We describe the algorithm and assess its complexity in Section IV-A before we proceed to provide intuition on its optimality in Section IV-B. For the sake of clarity, we concentrate on the case of a square channel matrix $\mathbf{\Pi}$ (equal channel input and output alphabets), which is invertible. The more general case, in which $\mathbf{\Pi}$ is nonsquare, is treated in Section IV-C, assuming the matrix rows are linearly independent. In Section IV-D, we particularize the algorithm to several channels of interest, and conclude with Section IV-E emphasizing that the core of the DUDE can be viewed as an estimation of a conditional distribution.

A. The Algorithm: Description and Implementation

For $2k < n$, $a^n \in \mathcal{A}^n$, $b^k \in \mathcal{A}^k$, $c^k \in \mathcal{A}^k$, let $\mathbf{m}(a^n, b^k, c^k)$ denote the M-dimensional column vector whose βth component ($\beta \in \mathcal{A}$) is equal to

$$\mathbf{m}(a^n, b^k, c^k)[\beta] = \left|\left\{i : k+1 \le i \le n-k, a_{i-k}^{i+k} = b^k \beta c^k\right\}\right| \tag{11}$$

namely, the number of appearances of the string $b^k \beta c^k$ along the sequence a^n. For such an appearance, we say that β occurs in *left context* b^k, *right context* c^k, and *double-sided context* (b^k, c^k). The normalized (unit sum) version of the vector $\mathbf{m}(a^n, b^k, c^k)$ gives the empirical conditional distribution of a single letter given that the double-sided context is (b^k, c^k)

For a given noisy sequence z^n, the output of the algorithm at location i will be defined as a fixed function of z_i and of the vector of counts $\mathbf{m}(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})$, where the context length k may depend on n. Specifically, for a sequence $a^n \in \mathcal{A}^n$, a context length k, a double-sided context $(b^k, c^k) \in \mathcal{A}^{2k}$, and a symbol $\alpha \in \mathcal{A}$, we define the function

$$g_{a^n}^k(b^k, \alpha, c^k) = \arg\min_{\hat{x} \in \mathcal{A}} \mathbf{m}^T(a^n, b^k, c^k) \mathbf{\Pi}^{-1} \left[\lambda_{\hat{x}} \odot \pi_\alpha\right]. \tag{12}$$

For arbitrary $n > 2k$, let $\hat{X}^{n,k}$ denote the n-block denoiser given by

$$\hat{X}^{n,k}(z^n)[i] = g_{z^n}^k(z_{i-k}^{i-1}, z_i, z_{i+1}^{i+k}), \qquad k+1 \le i \le n-k. \tag{13}$$

Notice the similarity between the right-hand sides of (10) and (13). The value of $\hat{X}^{n,k}(z^n)[i]$ for $i \le k$ and $i > n-k$ will be (asymptotically) inconsequential in subsequent developments but, for concreteness, can be assumed to be identically given by an arbitrary fixed symbol.[2] Finally, for each n, our asymptotic analysis of the DUDE algorithm will focus on the n-block denoiser $\hat{X}_{\mathsf{univ}}^n$ defined as

$$\hat{X}_{\mathsf{univ}}^n = \hat{X}^{n,k_n} \tag{14}$$

where, for asymptotic optimality k_n is any unboundedly increasing function of n such that[3]

$$k_n M^{2k_n} = o(n/\log n). \tag{15}$$

A valid choice of k_n is given, for example, by $k_n = \lceil c \log_M n \rceil$ with $c < \frac{1}{2}$. Notice that this freedom in the choice of k_n is similar to the situation arising in universal prediction of individual sequences, where any growth rate for the order of a Markov predictor slower than some threshold guarantees universality [25]. The choice of a logarithmic growth rate (the fastest in the allowable range) would be similar to the choice implicit in the LZ predictor. The tradeoffs involved in this choice will become clearer in the sequel.

A natural implementation of the DUDE algorithm for a given k makes two passes through the observations z^n. The empirical counts $\mathbf{m}(z^n, u_{-k}^{-1}, u_1^k)[u_0]$, for the various strings u_{-k}^k appearing along the sequence z^n, are accumulated and stored in the first pass while the actual application of $g_{z^n}^k(\cdot)$, as determined by the accumulated empirical counts via (12), is performed in the second pass. We analyze the computational complexity of the following embodiment of the algorithm.

- **Preprocessing.** Before the data is read, the inverse of the channel transition probability matrix is computed in addition to $[\lambda_{\hat{x}} \odot \pi_\alpha]$ for all $(\hat{x}, \alpha) \in \mathcal{A}^2$. This takes $O(M^3)$ arithmetic operations and requires $O(M^3)$ storage.
- **Computation of counts.** The computation of the empirical counts can be organized efficiently in various ways. One possibility is to regard the double-sided context (b^k, c^k) of an input symbol z_i as a *state* of a finite-state automaton with M^{2k} states. As the denoiser transitions from location i to location $i+1$, the state following (b^k, c^k) can assume M^2 possible values of the form $(b_2^k z_i, c_2^k z_{i+k+1})$. Associated with each state (b^k, c^k) is an M-vector of counts, which, at time i, contains $\mathbf{m}(z^{i+k}, b^k, c^k)$. Each automaton transition requires a constant number of register-level operations: incrementing one of the components in one of the count vectors, and retrieving a pointer to the next state. Thus, the number of operations required in the first pass of the DUDE is linear in n. The storage requirements for this pass are, in the worst case, $O(M^{2k+1})$. Using an alternative lexicon, the finite automaton can also be described as a trellis with the same set of M^{2k} states, with the input sequence representing a path through the trellis. In many applications such as text correction, only a small subset of states are actually visited, and the implementation can allocate their storage dynamically as new states occur, resulting in significant storage savings. It can be shown that the data structure can be dynamically grown with $O(N_{k,n})$ arithmetic operations, where we denote $N_{k,n} = \min\{n, M^{2k}\}$.

  The described finite state automaton lends itself to a representation with the additional properties of a tree data structure, similar to the *tree model* representations used in source coding (cf., e.g., [68]). This representation is convenient when the function $g_{z^n}^k(\cdot)$ is to be computed for multiple values of k, since internal nodes of the tree correspond to different possible double-sided context lengths. In this case, the information stored at the leaves is sufficient to infer the counts corresponding to the internal nodes.

[2] In practice, a more judicious choice for the boundary symbols is the corresponding estimate obtained with the longest possible context that fits within the data sequence.

[3] As will be discussed in Section V, the condition (15) can be slightly relaxed depending on the type of universality that is required.

- **Precomputations for the second pass.** The unnormalized input probability vectors $\mathbf{m}^T(z^n, b^k, c^k)\mathbf{\Pi}^{-1}$ are computed for each double-sided context (b^k, c^k) actually encountered in the sequence. Since there are $N_{k,n}$ double-sided contexts in the worst case, and each computation takes $O(M^2)$ arithmetic operations, the computational complexity and the space required to store the computations are both $O(N_{k,n})$. The algorithm then proceeds to precompute the values of $g^k_{z^n}(b^k, \alpha, c^k)$ according to (12), for each state (b^k, c^k) and alphabet symbol α. There are at most $MN_{k,n}$ such combinations, each requiring $O(M^2)$ operations, for a total of $O(N_{k,n})$ operations requiring $O(N_{k,n})$ storage.
- **Denoising.** The algorithm scans the sequence z^n a second time. At each sequence location, the context $(z_{i-k}^{i-1}, z_{i+1}^{i+k})$ and input symbol z_i are observed, and used to address the table of precomputed values of $g^k_{z^n}(\cdot)$ from the previous step. The automaton transitions are followed as in the first pass, yielding, again, running time linear in n.

Adding up the contributions of the various steps, the overall running time complexity of the algorithm, measured in register-level operations, is $O(n + N_{k,n}) = O(n)$. The working storage complexity is $O(N_{k,n})$, which is sublinear when $k = k_n$ satisfies (15). This choice of growth rate for k will also allow us to prove the asymptotic optimality of the DUDE. This estimate does not take into account memory that might be required to store the input sequence between the two passes. In many practical applications, the sequence is stored in secondary memory (e.g., hard disk), and read twice by the algorithm. Notice that the computation does not require more than $2k + 1$ symbols from the input sequence at any one time. In applications where there is no distinction between fast working memory and secondary storage, the storage complexity becomes linear in n.

The linear time complexity of the DUDE implementation just described stems from the fact that the data is scanned sequentially, and that in the transition from one symbol to the next, a constant number of "new" symbols is introduced to the context. This will not be the case in multidimensional implementations, however, where the number of new symbols introduced in a context transition will generally be of the form $O(K^\eta)$, where K is the total number of symbols in the context, and $0 < \eta \leq 1$. Since the multidimensional case still requires $K = K_n \to \infty$ with $K_n = O(\log n)$ for asymptotic optimality as $n \to \infty$, the running time of the denoiser will be super-linear, but no worse than $O(n^{1+\epsilon})$ for any $\epsilon > 0$. This upper bound holds for the DUDE also in the one-dimensional case under the more stringent computational model where we count bit operations, rather than register-level ones. Notice also that the fact that a sequential scanning is not essential for the DUDE's function makes the algorithm highly parallelizable.

B. Intuition Behind the Optimality of the DUDE

The n-block denoiser $\hat{X}^{n,k}$, as is evident from (13), employs a denoising function $g^k_{z^n} : \mathcal{A}^{2k+1} \to \mathcal{A}$, which depends on z^n but is the same for all locations $k + 1 \leq i \leq n - k$. This property characterizes kth order sliding-window schemes, which will be formally defined in Section V. Our main goal in the present subsection is to heuristically argue why this particular sliding-window scheme can be expected to do well and, in fact, essentially as well as the best scheme of its type, regardless of the characteristics of the underlying noise-free sequence. Moreover, we will argue that with $k = k_n$ and $k_n \to \infty$ at an appropriate rate, this property guarantees asymptotic optimality in a strong sense.

For $a^n, b^n \in \mathcal{A}^n$, $c^{2k+1} \in \mathcal{A}^{2k+1}$, let $\mathbf{q}(a^n, b^n, c^{2k+1})$ denote the M-dimensional column vector whose αth component, $\alpha \in \mathcal{A}$, is

$$\mathbf{q}(a^n, b^n, c^{2k+1})[\alpha] = \left|\left\{i : k+1 \leq i \leq n-k, a_{i-k}^{i+k} = c^{2k+1}, b_i = \alpha\right\}\right| \quad (16)$$

namely, the number of appearances of the string c^{2k+1} along the sequence a^n when the letter in the sequence b^n corresponding to the center of c^{2k+1} is α. Observe that

$$\sum_{\alpha \in \mathcal{A}} \mathbf{q}(a^n, b^n, c^{2k+1})[\alpha] = \mathbf{m}(a^n, c_1^k, c_{k+2}^{2k+1})[c_{k+1}]. \quad (17)$$

Now, by summing over all possible $(2k + 1)$-tuples, the cumulative loss incurred by a sliding-window denoiser that uses a denoising function f on x_{k+1}^{n-k} takes the form

$$\sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \lambda_{f(u_{-k}^k)}^T \mathbf{q}(z^n, x^n, u_{-k}^k)\,. \quad (18)$$

Therefore, a genie getting to select the best kth order sliding-window scheme (in the sense of minimizing the overall loss) based on knowledge of both x^n and z^n would choose (in analogy with (5) in the nonuniversal setting) the function f given by the Bayes response to $\mathbf{q}$, namely

$$f(u_{-k}^k) = \arg\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}(z^n, x^n, u_{-k}^k) = \hat{x}(\mathbf{q}(z^n, x^n, u_{-k}^k)). \quad (19)$$

Thus, if we are ambitious enough to strive for attaining the performance of the genie, a plausible approach would be via a denoiser $\hat{X}^n$ given, for $k + 1 \leq i \leq n - k$, by

$$\hat{X}^n(z^n)[i] = \hat{x}(\hat{\mathbf{q}}(z^n, z_{i-k}^{i+k})) \quad (20)$$

where, for $u_{-k}^k \in \mathcal{A}^{2k+1}$, $\hat{\mathbf{q}}(z^n, u_{-k}^k)$ would be some estimate, based on z^n alone, for the unobserved $\mathbf{q}(z^n, x^n, u_{-k}^k)$. Indeed, comparing (20) with (19), it is natural to expect, by continuity arguments, that the normalized loss of the denoiser in (20) be "close" to that of the genie whenever $\hat{\mathbf{q}}(z^n, u_{-k}^k)$ is "close" to $\mathbf{q}(z^n, x^n, u_{-k}^k)$ for all u_{-k}^k. Note that the denoiser in (13) is exactly of the form (20) if we choose

$$\hat{\mathbf{q}}(z^n, u_{-k}^k) = \pi_{u_0} \odot \left[\mathbf{\Pi}^{-T} \mathbf{m}(z^n, u_{-k}^{-1}, u_1^k)\right]. \quad (21)$$

It thus remains to be argued why the right-hand side of (21) can be expected to be close to $\mathbf{q}(z^n, x^n, u_{-k}^k)$. To this end, take a double-sided context (u_{-k}^{-1}, u_1^k) and a symbol $a \in \mathcal{A}$, and consider the number of locations for which z_i, $k+1 \leq i \leq n-k$, appears in context (u_{-k}^{-1}, u_1^k) and the noiseless symbol is a, i.e.,

$$\sum_{b \in \mathcal{A}} \mathbf{q}(z^n, x^n, u_{-k}^{-1} b u_1^k)[a].$$

It seems plausible to expect that the fraction of locations for which the noise-corrupted symbol is u_0 be approximately $\Pi(a,u_0)$, i.e.,

$$\mathbf{q}(z^n,x^n,u_{-k}^k)[a]\approx\Pi(a,u_0)\cdot\sum_{b\in\mathcal{A}}\mathbf{q}(z^n,x^n,u_{-k}^{-1}bu_1^k)[a] \tag{22}$$

or, in vector notation

$$\mathbf{q}(z^n,x^n,u_{-k}^k)\approx\pi_{u_0}\odot\sum_{b\in\mathcal{A}}\mathbf{q}(z^n,x^n,u_{-k}^{-1}bu_1^k). \tag{23}$$

This will indeed be shown in Section V to be the case, in the sense that the magnitude of the difference between the two sides of (22) divided by n vanishes with high probability and in expectation as n tends to infinity. Observe that (23) can be *formally* obtained from (7) by replacing the vector $\mathbf{P}_{X_t,z^n}$ with the vector $\mathbf{q}(z^n,x^n,u_{-k}^k)$ and the equality with the $\approx$ sign. Thus, a derivation *formally identical* to the one leading, in the nonuniversal setting, from (6) to (9), will now lead us, by (17), to the desired conclusion that $\hat{\mathbf{q}}$ in (21) is close to $\mathbf{q}$. Notice that this observation, together with the correspondence between (19) and (5), establishes a complete analogy between the two settings in case the observed data in Section III is limited to a $(2k+1)$-tuple. Specifically, summing over the components of the vectors on both sides of (23), using (17), and iterating over all possible $u_0\in\mathcal{A}$, we obtain

$$\mathbf{m}(z^n,u_{-k}^{-1},u_1^k)\approx\mathbf{\Pi}^T\left[\sum_{b\in\mathcal{A}}\mathbf{q}(z^n,x^n,u_{-k}^{-1}bu_1^k)\right] \tag{24}$$

(which corresponds to (8)). Putting together (23) and (24) yields

$$\mathbf{q}(z^n,x^n,u_{-k}^k)\approx\pi_{u_0}\odot\left[\mathbf{\Pi}^{-T}\mathbf{m}(z^n,u_{-k}^{-1},u_1^k)\right] \tag{25}$$

as claimed (the precise statement of this conclusion will be given in the proof of Theorem 2, Section V).

Finally, to perform asymptotically as well as *any* given sliding-window denoiser, we need $k=k_n\to\infty$. At first sight, it may seem that the higher the value of k the better the performance since this is certainly the case for the target genie-aided denoiser. However, not only is there a computational disincentive to make k too large, but there is also a performance penalty for letting k_n grow too fast with n: the context counts become too sparse and noisy. In particular, the error term in the above approximations will be seen to increase rapidly with k, necessitating a limit on the growth rate of k_n.

Note that the above argumentation involved no assumption on a probabilistic (or any other type of) mechanism that may have generated the noise-free signal. The approximate relationship (22) on which the whole line of reasoning hinges was argued solely on the basis of the randomness in the channel. We now argue heuristically how the semi-stochastic optimality of the DUDE just sketched translates into optimality in the fully stochastic setting. Let $\mathbf{X}=\{X_t\}_{t\in\mathbb{Z}}$ be a stationary ergodic process with $\mathcal{A}$-valued components, and let $\mathbf{Z}$ denote the output of the memoryless channel $\mathbf{\Pi}$ whose input is $\mathbf{X}$. The semi-stochastic optimality of the DUDE implies that it performs no worse, asymptotically, than the nonuniversal sliding-window denoiser obtained by replacing $\mathbf{P}_{Z_t,z^n\setminus t}$ in (10) with $\mathbf{P}_{Z_t,z_{t-k_\eta}^{(t+k_n)\setminus t}}$, on any individual sequence x^n, and hence in any conceivable probabilistic sense. That the DUDE performs at least as well as (10) then follows by a martingale argument (plus the ergodic theorem for the almost sure case) showing that the asymptotic performance of these nonuniversal sliding-window denoisers converges to that of (10). This argument is made precise in Section VI. We emphasize that our method for establishing the fully stochastic optimality of the DUDE does *not* rely on showing (nor does it imply) that $(1/n)\mathbf{m}(z^n,u_{-k_n}^{-1},u_1^{k_n})$ converges to $\mathbf{P}_{Z_t,u_{-k_n}^{k_n}\setminus 0}$ in any sense whatsoever.[4] Instead, not unlike the proofs of analogous results in universal compression and prediction, we establish the fully stochastic optimality of the DUDE by leveraging its stronger semi-stochastic optimality.

C. Nonsquare Channel Transition Probability Matrix

It is easy to generalize the DUDE to the case where the channel transition probability matrix is nonsquare, as long as its rows are linearly independent. The input and output alphabets are now denoted by $\mathcal{A}$ and $\mathcal{B}$, respectively, with $|\mathcal{A}|=M$ and $|\mathcal{B}|=M'$. Note that the channel transition probability matrix $\mathbf{\Pi}$ is $M\times M'$ where $M\le M'$. The loss matrix is still $M\times M$ since we assume the reconstruction alphabet to equal the noiseless source alphabet $\mathcal{A}$.[5] A common channel encompassed by this generalization is the erasure channel.

In order to generalize the DUDE to this setting, it suffices to replace (12) by

$$g_{a^n}^k(b^k,\alpha,c^k)=\arg\min_{\hat{x}\in\mathcal{A}}\mathbf{m}^T(a^n,b^k,c^k)\mathbf{\Pi}^T\left(\mathbf{\Pi}\,\mathbf{\Pi}^T\right)^{-1}[\lambda_{\hat{x}}\odot\pi_\alpha]. \tag{26}$$

To motivate (26), write

$$\sum_{b\in\mathcal{A}}\mathbf{q}(z^n,x^n,u_{-k}^{-1}bu_1^k)=(\mathbf{\Pi}\mathbf{\Pi}^T)^{-1}\mathbf{\Pi}\mathbf{\Pi}^T\sum_{b\in\mathcal{A}}\mathbf{q}(z^n,x^n,u_{-k}^{-1}bu_1^k)\approx(\mathbf{\Pi}\mathbf{\Pi}^T)^{-1}\mathbf{\Pi}\,\mathbf{m}(z^n,u_{-k}^{-1},u_1^k) \tag{27}$$

where (27) follows from (24). Substituting the rightmost side of (27) in lieu of $\left[\mathbf{\Pi}^{-T}\mathbf{m}(z^n,u_{-k}^{-1},u_1^k)\right]$ in (21) yields (26).

The above derivation can be readily extended by replacing the Moore–Penrose generalized inverse (cf., e.g., [39]) $\mathbf{\Pi}^T(\mathbf{\Pi}\,\mathbf{\Pi}^T)^{-1}$ appearing in (26) with any other generalized inverse of the form $\mathbf{\Gamma}^T(\mathbf{\Pi}\,\mathbf{\Gamma}^T)^{-1}$, where $\mathbf{\Gamma}$ is any $M\times M'$ matrix for which $\mathbf{\Pi}\,\mathbf{\Gamma}^T$ is invertible. While any generalized inverse of this form will give rise to an asymptotically optimal DUDE, some choices may be more effective than others in terms of convergence rates. For expository convenience, subsequent sections will assume $\mathcal{B}=\mathcal{A}$, though all the results we present can be seen to carry over to the case $|\mathcal{B}|>|\mathcal{A}|$ for full row rank $\mathbf{\Pi}$ and the DUDE defined through (26).

[4]On the other hand, stochastic optimality of the DUDE among all kth order sliding-window denoisers *can* be established directly from the fact that $(1/n)\mathbf{m}(z^n,u_{-k}^{-1},u_1^k)$ converges to $\mathbf{P}_{Z_t,u_{-k}^k\setminus 0}$ for fixed k (by ergodicity).

[5]The derivation extends to a general reconstruction alphabet in a straightforward way.

D. A Closer Look at Special Cases

We now derive the explicit form of the denoiser for a few cases of special interest. Hamming loss is assumed (with equal loss for any errors in the nonbinary case) in all the examples below.

- *BSC:* For a BSC with error probability δ, $\delta < 1/2$

$$\mathbf{\Pi} = \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix}, \ \mathbf{\Pi}^{-1} = \frac{1}{1-2\delta}\begin{pmatrix} 1-\delta & -\delta \\ -\delta & 1-\delta \end{pmatrix}.$$

Substituting the value of $\mathbf{\Pi}^{-1}$ into (12) yields, following simple algebraic manipulations:

$$g^k_{z^n}(u^{-1}_{-k}, u_0, u^k_1) = \begin{cases} u_0, & \text{if } \frac{\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[u_0]}{\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[u_0] + \mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[\bar{u}_0]} \geq 2\delta(1-\delta) \\ \bar{u}_0, & \text{otherwise} \end{cases} \tag{28}$$

where $\bar{u}_0$ denotes the binary complement of u_0. In words, for each bit u_0 in the noisy sequence, the DUDE counts how many bits occurring within the same double-sided context are equal to u_0. If the fraction of such occurrences among the total number of occurrences of this double-sided context is below $2\delta(1-\delta)$ then u_0 is deemed to be an error introduced by the BSC.

To gain some intuition regarding the form that the DUDE assumes in this case, consider the situation of an independent and identically distributed (i.i.d.) Bernoulli (θ) process corrupted by the BSC with crossover probability δ $(\theta, \delta < 1/2)$. It is easy to see that the optimal (distribution-dependent) scheme for this case leaves the ones in the noisy signal untouched whenever $\delta \leq \theta$, and flips all ones into zeros otherwise. Notice that the condition $\delta \leq \theta$ for leaving the signal untouched can be written in the equivalent form $\theta(1-\delta) + (1-\theta)\delta \geq 2\delta(1-\delta)$, and that the noisy signal is Bernoulli $(\theta(1-\delta) + (1-\theta)\delta)$. Now, since the frequency of ones in the noisy signal is an efficient estimate for $\theta(1-\delta) + (1-\theta)\delta$, a scheme which compares the frequency of ones in the noisy signal to the threshold $2\delta(1-\delta)$, flipping the ones only if the threshold is exceeded, will be asymptotically optimal in this i.i.d. example. Comparing this now with (28), it can be seen that this is precisely the kind of scheme that the DUDE is independently employing within each of the occurring double-sided k-contexts.

Another point we mention in this context is that the DUDE, as well as the optimal distribution-dependent scheme, may be making as few as zero flips (corresponding to the case, for the i.i.d. example above, of $\delta < \theta$) and as many as $\approx 2\delta(1-\delta)n$ flips (for $\theta \approx \delta$). This is in contrast to the compression-based scheme of [21] which makes at most $n\delta$ flips.

- *M-ary Symmetric Channel:* Generalizing the previous example, we consider the channel

$$\Pi(i,j) = \begin{cases} 1-\delta, & \text{if } i = j \\ \frac{\delta}{M-1}, & \text{otherwise} \end{cases}$$

for which the matrix is easily seen to be invertible for $\delta \neq (M-1)/M$, and the inverse takes the form

$$[\mathbf{\Pi}^{-1}](i,j) = \begin{cases} \alpha, & \text{if } i = j \\ \beta, & \text{otherwise} \end{cases}$$

where $\beta/\alpha = \frac{-\delta}{M-1-\delta}$, and $\alpha > 0$ or $\alpha < 0$ according to whether $\delta < (M-1)/M$ or $\delta > (M-1)/M$. Substituting into (12) yields, after straightforward manipulations, for $\delta < (M-1)/M$

$$g^k_{z^n}(u^k_{-k}) = \begin{cases} u_0, & \text{if } \varsigma \mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[u_0] \\ & \quad -\mu \mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[x^*] \\ & \quad \geq \xi(u^{-1}_{-k}, u^k_1) \\ x^*, & \text{otherwise} \end{cases} \tag{29}$$

where

$$\xi(u^{-1}_{-k}, u^k_1) = \sum_{a \in \mathcal{A}} \mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[a]$$

$$x^* = \arg\max_{\hat{x} \neq u_0} \mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[\hat{x}]$$

$$\varsigma = \frac{(M-1)^2(1-\delta)}{\delta[(1-\delta)M - 1]}$$

and $\mu = (M-1)/[(1-\delta)M - 1]$.

- *The Z Channel:* The channel probability matrix, and its inverse, for this case are given by

$$\mathbf{\Pi} = \begin{pmatrix} 1-\delta & \delta \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{\Pi}^{-1} = \begin{pmatrix} \frac{1}{1-\delta} & \frac{-\delta}{1-\delta} \\ 0 & 1 \end{pmatrix}.$$

Since only locations i where $z_i = 1$ may need correction, we are only interested in the evaluation of $g^k_{z^n}$ at $(u^{-1}_{-k}, 1, u^k_1)$. Equation (12) takes the form

$$g^k_{z^n}(u^{-1}_{-k}, 1, u^k_1) = \begin{cases} 0, & \text{if } \frac{1-\delta}{2\delta} < \frac{\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[0]}{\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[1]} \\ 1, & \text{otherwise.} \end{cases} \tag{30}$$

- *The Erasure Channel:* Consider the case where $\{1, \ldots, M\}$ is the alphabet of the noiseless signal, which is corrupted by an erasure channel with erasure probability δ. Thus, the channel output alphabet is $\{1, \ldots, M, e\}$ and the $M \times (M+1)$ channel matrix is of the form

$$\mathbf{\Pi} = \left[\begin{array}{c|c} (1-\delta)\mathbf{I}_M & \begin{matrix} \delta \\ \vdots \\ \delta \end{matrix} \end{array}\right] \tag{31}$$

where $\mathbf{I}_M$ denotes the $M \times M$ identity matrix. This setting falls within the purview of the DUDE derived in Section IV-C, (26). Evaluating the explicit form of this denoiser shows (after straightforward manipulations) that, as one may have expected, it corrects every erasure with the most frequent symbol for its context. In particular, the denoiser does not depend on the channel parameter δ.

E. Estimation of Distributions

To conclude this section, we emphasize that the core of the DUDE can be viewed as the estimation of the statistics of the

noiseless signal components, given the observed noisy data. Indeed, the DUDE bases its estimates of the noiseless symbols on the vector

$$\pi_{z_i} \odot \left[\mathbf{\Pi}^{-T}\mathbf{m}(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})\right].$$

As was heuristically argued in Section IV-B, and will be made precise in the next section, this vector is, up to normalization, an efficient estimate of the conditional distribution of the noiseless symbol X_i given the noisy data z_{i-k}^{i+k}. There are problems [45] outside denoising where the computation of these quantities plays a fundamental role.

We also mention that having an efficient way of estimating the empirical distribution of x_i given z_{i-k}^{i+k} (as well as the empirical joint distribution of x_i and z_{i-k}^{i+k}) allows us to efficiently estimate the empirical distribution of any function of these variables, and, in turn, any quantity corresponding to an expectation under such an empirical distribution.

As one example, consider the following explicit estimate, obtained with this approach, of the empirical joint distribution of x_i and the DUDE's output $\hat{x}_i = \hat{X}^{n,k}(z^n)[i]$ (defined in (13)). For $x, \hat{x} \in \mathcal{A}$, the estimate takes the form of (32) at the bottom of the page. It is not hard to show that, in various quantitative senses $\hat{Q}_{z^n}$ is an efficient estimate of the joint empirical distribution of $(x_i, \hat{x}_i)$ (cf., e.g., [27] for a more general result). In particular, the expected value of $\Lambda(X, \hat{X})$ with respect to $\hat{Q}_{z^n}$ gives an efficient estimate of the overall normalized cumulative loss of the DUDE. This estimate, as well as the estimated empirical conditional distribution of $\hat{x}_i$ given x_i, play important roles in algorithms for channel decoding with redundant information sources [45].

V. The Semi-Stochastic Setting

In this section, we assess the strong asymptotic optimality of the DUDE, as defined in Section IV-A. To this end, we define a *semi-stochastic setting*, in which $\mathbf{x}$ is an *individual sequence* and its noise-corrupted version, a random variable $\mathbf{Z}$, is the output of the memoryless channel $\mathbf{\Pi}$, whose input is $\mathbf{x}$. This setting is assumed throughout this section. We shall use $\mathbf{z}$ to denote an individual sequence, or a specific sample value of $\mathbf{Z}$. Though we suppress this dependence in the notation for readability, probabilities of events (as well as associated expectations) relate to the underlying individual sequence. Thus, we shall write, for example, $\Pr(Z^n = z^n)$ to denote the probability that the channel output is z^n, when the input sequence was the individual sequence x^n. Note that in this case we have the explicit relation

$$\Pr(Z^n = z^n) = \prod_{i=1}^{n} \Pi(x_i, z_i).$$

A setting involving a noise-corrupted individual sequence was introduced into information theory by Ziv in his work [78] on rate-distortion coding of individual sequences. More recently, problems of prediction [70], [73], as well as of limited-delay coding [71] of noise-corrupted individual sequences were also considered. As mentioned in Section I and as we elaborate on below, denoising in the semi-stochastic setting is also closely related to the classical compound decision problem [31], [54], [55], [57], [58].

A. Main Results: Statement and Discussion

To state our results in the semi-stochastic setting, we define a class of n-block denoisers, characterized by sliding windows of length $2k+1$. Specifically, a kth-order sliding-window denoiser $\hat{X}^n$ is characterized by the property that for all $z^n \in \mathcal{A}^n$

$$\hat{X}^n(z^n)[i] = \hat{X}^n(z^n)[j], \qquad \text{whenever} \quad z_{i-k}^{i+k} = z_{j-k}^{j+k}.$$

Thus, for each sequence z^n, the denoiser defines a mapping

$$f_{z^n} : \mathcal{A}^{2k+1} \to \mathcal{A}$$

so that

$$\hat{X}^n(z^n)[i] = f_{z^n}(z_{i-k}^{i+k}), \qquad i = k+1, \ldots, n-k.$$

We let $\mathcal{S}_k$ denote the class of kth-order sliding-window denoisers. Note that these are two-pass denoisers in the sense that the (fixed) sliding-window function is allowed to depend on the entire observed sequence. For an n-block denoiser $\hat{X}^n$, we now extend the scope of the notation $L_{\hat{X}^n}$ by defining, for $1 \le l \le m \le n$

$$L_{\hat{X}^n}(x_l^m, z^n) = \frac{1}{m-l+1}\sum_{i=l}^{m} \Lambda(x_i, \hat{X}^n(z^n)[i])$$

namely, the normalized cumulative loss incurred between (and including) locations l and m. Note that for $\hat{X}^n \in \mathcal{S}_k$ with an associated collection of mappings $\{f_{z^n}\}$ we have

$$\begin{aligned}
(n-2k)&L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n) \\
&= \sum_{i=k+1}^{n-k} \Lambda(x_i, \hat{X}^n(z^n)[i]) \\
&= \sum_{i=k+1}^{n-k} \Lambda(x_i, f_{z^n}(z_{i-k}^{i+k})) \\
&= \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \sum_{a \in \mathcal{A}} \mathbf{q}(z^n, x^n, u_{-k}^k)[a]\Lambda(a, f_{z^n}(u_{-k}^k)) \\
&= \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \lambda_{f_{z^n}(u_{-k}^k)}^T \mathbf{q}(z^n, x^n, u_{-k}^k) \qquad (33)
\end{aligned}$$

where the statistics $\mathbf{q}$ are defined in (16). Note also that the DUDE, $\hat{X}_{\mathsf{univ}}^n$, is a member of $\mathcal{S}_{k_n}$, with each mapping f_{z^n} given

$$\hat{Q}_{z^n}(x, \hat{x}) = \frac{1}{n-2k} \sum_{u_{-k}^k : g_{z^n}^k(u_{-k}^{-1}, u_0, u_1^k) = \hat{x}} \left(\pi_{u_0} \odot \left[\mathbf{\Pi}^{-T}\mathbf{m}(z^n, u_{-k}^{-1}, u_1^k)\right]\right)[x]. \qquad (32)$$

by $g^k_{z^n}$ for $k = k_n$ (see (14)). Here, k_n is any unboundedly increasing function of n with certain limitations on the growth rate, which are required for universality (recall (15)).

For an individual noiseless sequence $\mathbf{x} \in \mathcal{A}^\infty$, noisy observation sequence $\mathbf{z} \in \mathcal{A}^\infty$, and integers $k \geq 0$ and $n > 2k$, we define the kth-order minimum loss of (x^n, z^n) by

$$D_k(x^n, z^n) \triangleq \min_{\hat{X}^n \in \mathcal{S}_k} L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n) = \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(x_i, f(z_{i-k}^{i+k})). \quad (34)$$

The minimum loss $D_k(x^n, z^n)$ is the benchmark against which we will assess the performance of denoisers in the class $\mathcal{S}_k$ (we ignore any loss contributed by the boundaries, as $k = o(n)$ in the cases of interest). The minimizing argument in (34) depends on both x^n and z^n. It follows, *a fortiori*, that the definition of the class of kth-order sliding-window denoisers could have been restricted to only those denoisers for which the mapping f_{z^n} is the same for all sequences z^n ("*one-pass*" denoisers). This restricted class would still contain at least one denoiser achieving $D_k(x^n, z^n)$. As noted, the DUDE is a member of $\mathcal{S}_{k_n}$, yet note that it does *not* belong to the restricted class of k_nth-order sliding-window one-pass denoisers.

By (33), the kth-order minimum loss takes the form

$$D_k(x^n, z^n) = \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}(z^n, x^n, u_{-k}^k) = \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} U(\mathbf{q}(z^n, x^n, u_{-k}^k)). \quad (35)$$

Our main result, Theorem 1, states that for any input sequence $\mathbf{x}$, the DUDE, as defined in (14), performs essentially as well as the best sliding-window denoiser with the same window length.

Theorem 1: For all $\mathbf{x} \in \mathcal{A}^\infty$, the sequence of denoisers $\{\hat{X}^n_{\mathsf{univ}}\}$ defined in (14) satisfies

a)
$$\lim_{n\to\infty} [L_{\hat{X}^n_{\mathsf{univ}}}(x^n, Z^n) - D_{k_n}(x^n, Z^n)] = 0 \text{ a.s.}$$

provided that $k_n M^{2k_n} = o(n/\log n)$.

b) $E\left[L_{\hat{X}^n_{\mathsf{univ}}}(x^n, Z^n) - D_{k_n}(x^n, Z^n)\right] = O\left(\sqrt{\frac{k_n M^{2k_n}}{n}}\right)$.

Remark: Part b) of the theorem states convergence in expectation provided that $k_n M^{2k_n} = o(n)$, a condition slightly less stringent than the one required in Part a). This convergence, however, may be seen as less relevant to the semi-stochastic setting than the almost sure convergence of Part a), since an expectation criterion is more naturally targeted to situations in which repeated experiments can be carried out. The result is, in any case, in line with the fully stochastic setting assumed in Section VI. We include it here as it does not require a probabilistic assumption on $\mathbf{x}$, and its proof uses similar tools as that of Part a).

Theorem 2 is the key result underlying the proof of Theorem 1. To state this theorem, we further define the function[6] [32, Theorem 1]

$$\varphi(p) \triangleq \begin{cases} \frac{1}{1-2p}\log\frac{1-p}{p}, & 0 \leq p < 1/2, \\ \frac{1}{2p(1-p)}, & 1/2 \leq p \leq 1. \end{cases} \quad (36)$$

The function $\varphi(p)$ is continuous and $\min_{0\leq p\leq 1}\varphi(p) = \varphi(1/2) = 2$ (see [32, Theorem 1]).

Theorem 2: Let

$$F_{\mathbf{\Pi}} \triangleq \sum_{a\in\mathcal{A}} \left[\min_{A\subseteq\mathcal{A}} \varphi(\Pi(a, A))\right]^{-1}$$

$$C_{\mathbf{\Lambda},\mathbf{\Pi}} \triangleq \Lambda_{\max}\left(1 + \|\mathbf{\Pi}^{-1}\|_\infty\right)$$

and

$$V_{\mathbf{\Pi}} \triangleq \left[\sum_{a\in\mathcal{A}} \left(\sum_{b\in\mathcal{A}} \sqrt{\Pi(a,b)(1-\Pi(a,b))}\right)^2\right]^{\frac{1}{2}}.$$

Then, for any $k \geq 0$, $n > 2k$, $x^n \in \mathcal{A}^n$, and $\varepsilon > 0$, the denoiser $\hat{X}^{n,k}$ defined in (13) satisfies

$$\Pr\left(L_{\hat{X}^{n,k}}(x_{k+1}^{n-k}, Z^n) - D_k(x^n, Z^n) > \varepsilon\right) \leq K_1(k+1)M^{2k+1}\exp\left(-\frac{(n-2k)\varepsilon^2}{4(k+1)M^{2k}F_{\mathbf{\Pi}}C_{\mathbf{\Lambda},\mathbf{\Pi}}^2}\right) \quad (37)$$

$$E\left[L_{\hat{X}^{n,k}}(x_{k+1}^{n-k}, Z^n) - D_k(x^n, Z^n)\right] \leq \sqrt{\frac{2}{\pi}} C_{\mathbf{\Lambda},\mathbf{\Pi}} V_{\mathbf{\Pi}} M^k \sqrt{\frac{k+1}{n-2k}} + C_{\mathbf{\Lambda},\mathbf{\Pi}} M^{2k+2}\frac{k+1}{n-2k} \quad (38)$$

where K_1 depends only on the channel.

In words: Regardless of the underlying noiseless individual sequence, the event that the normalized cumulative loss of the denoiser $\hat{X}^{n,k}$ will exceed that of the best kth-order sliding-window denoiser by $\varepsilon > 0$ is exponentially unlikely in the sequence length. In addition, the expected excess loss vanishes at a rate $O(1/\sqrt{n})$ for fixed k. The factor $V_{\mathbf{\Pi}}$ in the right-hand side of (38) tells us that the bound on the expected excess loss becomes smaller for "skewed" channels. For example, for the BSC with transition probability δ, $V_{\mathbf{\Pi}} = \sqrt{8\delta(1-\delta)}$. The factor $F_{\mathbf{\Pi}}$, which also tends to zero as the channel becomes less "noisy" (as $\varphi(0) = \varphi(1) = \infty$), captures the analogous dependency on $\mathbf{\Pi}$ in the exponent of (37). In any case, $V_{\mathbf{\Pi}} \leq \sqrt{M(M-1)}$ by the Cauchy–Schwarz inequality, whereas $F_{\mathbf{\Pi}} \leq M/2$. The factor $C_{\mathbf{\Lambda},\mathbf{\Pi}}$, on the other hand, tends to infinity as the channel matrix "approaches" a nonfull-rank matrix, reflecting the fact that universal denoising becomes increasingly difficult in this regime. The proof of Theorem 2 is deferred to Section V-B.

[6]Throughout this section, and in the statement of Theorem 2 in particular, we assume the following conventions concerning ∞, as shorthand for more formal but straightforward limit and continuity arguments: For any $c > 0$, $c/0 = \infty$, $c/\infty = 0$, $c\infty = \infty$, $\log(\infty) = \infty$, and $e^{-\infty} = 0$. Furthermore, $\log(\cdot)$ denotes the natural logarithm throughout.

Remark: It can be shown that the minimum over A appearing in the definition of $F_{\mathbf{\Pi}}$ is achieved at

$$A^* = \arg\max_{A \subseteq \mathcal{A}} \min\{\Pi(a, A), 1 - \Pi(a, A)\}.$$

Proof of Theorem 1: Fix throughout the proof $\mathbf{x} \in \mathcal{A}^\infty$. To prove Part a), choose $\varepsilon > 0$, and, for each n, use (37) with $k = k_n$. It is easy to see that for $k_n M^{2k_n} = o(n/\log n)$, the right-hand side of (37) is summable. Thus, by the Borel–Cantelli lemma

$$L_{\hat{X}^{n,k_n}}(x_{k_n+1}^{n-k_n}, Z^n) - D_{k_n}(x^n, Z^n) \leq \varepsilon \text{ eventually almost surely.} \quad (39)$$

Now, for any n-block denoiser $\hat{X}^n$ and $k \geq 0$ [7]

$$L_{\hat{X}^n}(x^n, z^n) = \frac{1}{n}\sum_{i=1}^{n} \Lambda(x_i, \hat{X}^n(z^n)[i]) \leq \frac{2k\Lambda_{\max}}{n} + L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n). \quad (40)$$

In particular, (40) holds for the sequence of denoisers $\{\hat{X}^n_{\mathsf{univ}}\}$. Taking limit suprema in (39), using (40) with $k = k_n$, and noticing that k_n/n vanishes, we obtain, for any $\varepsilon > 0$

$$\limsup_{n\to\infty}[L_{\hat{X}^n_{\mathsf{univ}}}(x^n, Z^n) - D_{k_n}(x^n, Z^n)] \leq \varepsilon \quad \text{a.s.}$$

Since ε is arbitrary, the proof of Part a) is complete by noticing that $\hat{X}^n_{\mathsf{univ}} \in \mathcal{S}_{k_n}$, and therefore, for all pairs of sequences $\mathbf{x}, \mathbf{z}$ and all n

$$L_{\hat{X}^n_{\mathsf{univ}}}(x^n, z^n) - \frac{n - 2k_n}{n} D_{k_n}(x^n, z^n) \geq 0$$

implying, in turn

$$\liminf_{n\to\infty}[L_{\hat{X}^n_{\mathsf{univ}}}(x^n, z^n) - D_{k_n}(x^n, z^n)] \geq 0.$$

Part b) follows directly from using (38) in Theorem 2 with $k = k_n$ and (40). □

It should be noticed that, in the semi-stochastic setting, it is possible to define a notion of "denoisability" of an individual sequence, analogous to the finite-state (FS) compressibility of [79], the FS predictability of [25], and, in particular, the conditional FS predictability of [73]. To this end, we define the sliding-window minimum loss of $(\mathbf{x}, \mathbf{z})$ by

$$D(\mathbf{x}, \mathbf{z}) = \lim_{k\to\infty} D_k(\mathbf{x}, \mathbf{z}) \quad (41)$$

where

$$D_k(\mathbf{x}, \mathbf{z}) = \limsup_{n\to\infty} D_k(x^n, z^n). \quad (42)$$

Note that $D_k(\mathbf{x}, \mathbf{z})$ is nonincreasing with k so that $D(\mathbf{x}, \mathbf{z})$ is well defined. The corresponding random variable $D(\mathbf{x}, \mathbf{Z})$ in principle depends on the realization of the channel noise. However, it turns out to be degenerate.

[7]Here and throughout, equalities or inequalities between random variables can be understood to hold, when not explicitly mentioned, for *all* possible realizations.

Claim 1: For any $\mathbf{x} \in \mathcal{A}^\infty$, there exists a deterministic real number $D(\mathbf{x})$ (which depends on $\mathbf{\Pi}$) such that

$$D(\mathbf{x}, \mathbf{Z}) = D(\mathbf{x}) \quad \text{a.s.} \quad (43)$$

Remark: We refer to $D(\mathbf{x})$ as the *denoisability* of $\mathbf{x}$. Intuitively, (43) is to be regarded as a law of large numbers, as $D_k(\mathbf{x}, \mathbf{z})$ depends on $\mathbf{x}$ and $\mathbf{z}$ only through the joint $(2k+1)$th-order empirical statistics of the two sequences, which for each given noise-free $(2k+1)$-tuple will converge to deterministic (channel-dependent) values. The technical proof is best handled by direct use of Kolmogorov's 0–1 law (cf., e.g., [23]).

Proof of Claim 1: For fixed $\mathbf{x} \in \mathcal{A}^\infty$ and k, $D_k(\mathbf{x}, \mathbf{z})$ is, by definition, invariant to changes in a finite number of coordinates of $\mathbf{z}$. Thus, by Kolmogorov's 0–1 law, there exists a deterministic constant $D_k(\mathbf{x})$ such that $D_k(\mathbf{x}, \mathbf{Z}) = D_k(\mathbf{x})$ a.s. Letting $D(\mathbf{x}) = \lim_{k\to\infty} D_k(\mathbf{x})$ completes the proof. □

The following result, which is a corollary to Theorem 1, establishes the asymptotic optimality of the DUDE in the semi-stochastic setting.

Corollary 1: The sequence of denoisers $\{\hat{X}^n_{\mathsf{univ}}\}$ satisfies

$$\limsup_{n\to\infty} L_{\hat{X}^n_{\mathsf{univ}}}(x^n, Z^n) \leq D(\mathbf{x}) \quad \text{a.s. } \forall \mathbf{x} \in \mathcal{A}^\infty \quad (44)$$

provided that $\lim_{n\to\infty} k_n = \infty$ and $k_n M^{2k_n} = o(n/\log n)$.

Proof: For fixed k and n large enough to guarantee $k_n \geq k$, we have

$$(n - 2k_n) D_{k_n}(x^n, Z^n) \leq (n - 2k) D_k(x^n, Z^n).$$

It follows that

$$\limsup_{n\to\infty} D_{k_n}(x^n, Z^n) \leq \limsup_{n\to\infty}\left[\frac{n-2k}{n-2k_n} D_k(x^n, Z^n)\right] = D_k(\mathbf{x}, \mathbf{Z}) \quad (45)$$

implying, by the arbitrariness of k, that

$$\limsup_{n\to\infty} D_{k_n}(x^n, Z^n) \leq D(\mathbf{x}, \mathbf{Z}). \quad (46)$$

The proof is completed by combining Theorem 1, Part a), with (46), and invoking Claim 1. □

To end this subsection, we discuss the form of our denoiser and its semi-stochastic performance guarantees (Theorems 1 and 2) in relation to the algorithms and results pertaining to the compound decision problem. Whereas in the classical hypothesis testing problem one out of m distributions is selected upon observation of a realization, Robbins [51] proposed the "compound decision problem" (also known as the "compound Bayes decision problem"), where n such independent hypothesis tests are to be solved simultaneously. In our terminology this is nothing but the problem of denoising an individual sequence with m-valued components corrupted by a memoryless channel, where the m distributions are given by the rows of $\mathbf{\Pi}$. Instead of assuming a Bayesian setting in which a prior distribution is available, the goal Robbins set was to compete

with the "time-invariant" scalar (also known as "symbol-by-symbol" [19], [47]) decision rule that minimizes its expected loss, as selected by a "genie" that knows the n true hypotheses. Robbins' setting has since been further developed in various directions such as accommodating sequentiality and refining convergence rates [31], [52]–[55], [57], [58], [65]. The proposed schemes obtain estimates of the unknown prior probabilities of the m hypotheses from the noisy observations, and then use the Bayesian "symbol-by-symbol" scheme replacing the unknown priors with those estimated probabilities. It can be verified that the nonsequential solution to the compound decision problem (given in its most general form in [53]), when specialized to a DMC of equal input and output alphabets, coincides with the DUDE in the special case $k = 0$, i.e., the "contextless" DUDE. The compound decision literature shows that as $n \to \infty$, the proposed schemes indeed achieve the performance of the best "symbol-by-symbol" decision rule aided by a genie that is allowed to observe the true hypotheses and select the rule with minimum expected loss (expectation with respect to the randomness in the noise). In contrast, our semi-stochastic result is stronger in the sense that the genie is allowed to see not only the clean signal but also the channel realization, thus selecting the rule minimizing the actual (rather than the expected) loss.

Competition with classes of schemes other than the symbol-by-symbol denoisers has received far less attention in the literature pertaining to the compound decision problem. The idea of using standards more stringent than attaining the performance of the best symbol-by-symbol scheme was suggested by Johns [33]. Whereas the rationale underlying Robbins' compound decision problem (and its solution) is that the ordering of the n hypotheses plays no role and therefore any arbitrary shuffling of them leads to the same solution, Johns [33] realizes the importance of context and considers sliding-window denoisers of a given order k. Under the label of "the extended compound decision problem" subsequent work [2], [3], [66], [67] reduced the problem back to one of competing with symbol-by-symbol schemes by regarding the noisy observations in the sliding window as one observation from a kth-order "super-symbol." This approach does not entirely reduce the problem to its classical setting since the noisy super-symbols are statistically dependent. The dependence is, however, rather weak (complete independence between components farther than k symbols apart, a property referred to as k-dependence [2]), allowing to extend the performance guarantees from the original problem. Although this reduction to the original symbol-by-symbol approach is conceptually straightforward, it gives rise to computationally complex schemes. While the computation performed by the DUDE at every location in the denoising pass is of a "single-letter" nature regardless of the value of k, the computation employed at each location by a scheme induced from the "super-symbol" approach is exponentially complex in k. More concretely, for a fixed k, n, and M, the running time of the latter is M^{2k+1} times greater than that of the DUDE. This difference is of considerable practical significance for even moderate values of k and M.

It can be shown via a derivation similar to that in [17, Sec. 6] that, after appropriate simplification, the (nonsequential version of the) denoising rule induced from the "super-symbol" approach would coincide with the kth-order DUDE when the input and output alphabets coincide.[8] More generally, however, when the input and output alphabets are not equal, the schemes significantly differ: while the DUDE (recall Section IV-C for its form in this case) retains its "single-letter" nature, the "super-symbol" approach yields schemes requiring computations that are exponentially complex in k for the estimation of each signal component.

Finally, the issue of whether and how to increase the window order with the sequence length has not been addressed in the literature on the extended compound decision problem, which deals with a regime of a fixed k and $n \to \infty$. In contrast, the increase of k with n, along with the characterization of the allowable rate for this increase, are key to the performance guarantees of the DUDE in both the semi-stochastic setting (attaining the denoisability of any individual sequence) and the stochastic setting of the next section (where optimum performance is shown to be attained for every stationary source).

B. Proof of Theorem 2

To prove Theorem 2, we first present three lemmas. The first two lemmas formalize the continuity intuitively claimed in Section IV-B and establish inequalities that are valid for any pair of sequences x^n, z^n. The third lemma, on the other hand, is probabilistic, and formalizes the approximate relation (22).

Lemma 1: Fix $k \geq 0$ and some collection of M^{2k+1} M-vectors $\{\mathbf{v}(u_{-k}^k)\}$ indexed by $u_{-k}^k \in \mathcal{A}^{2k+1}$. Construct a kth-order sliding-window denoiser $\hat{X}^n$ with sliding-block function given by the Bayes responses to $\{\mathbf{v}(u_{-k}^k)\}$

$$f(u_{-k}^k) = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{v}(u_{-k}^k) = \hat{x}(\mathbf{v}(u_{-k}^k))$$
$$\hat{X}^n(z^n)[i] = f(z_{i-k}^{i+k}).$$

Then, for all $x^n, z^n \in \mathcal{A}^n$

$$0 \leq L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n) - D_k(x^n, z^n) \leq \frac{\Lambda_{\max}}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \left\| \mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{v}(u_{-k}^k) \right\|_1 . \quad (47)$$

Proof: The first inequality in (47) follows trivially from the fact that $\hat{X}^n \in \mathcal{S}_k$. To derive the second inequality, notice that by (35), (33), and the definition of the denoiser $\hat{X}^n$, we have (48)–(49) at the top of the following page, where, for simplicity, we have dropped the arguments of $\mathbf{v}$ and $\mathbf{q}$ when used for indexing columns of $\mathbf{\Lambda}$, and (48) holds since, for any pair of M-vectors $\mathbf{v}$ and $\mathbf{w}$, we have

$$[\lambda_{\hat{x}(\mathbf{v})} - \lambda_{\hat{x}(\mathbf{w})}]^T \mathbf{v} \leq 0. \qquad \square$$

The continuity property established in Lemma 1 is, in fact, typical of finite matrix games [30, eq. (14)]. In particular, the proposed denoiser is clearly of the form covered by the lemma, with

$$\mathbf{v}(u_{-k}^k) = \pi_{u_0} \odot \left[\mathbf{\Pi}^{-T} \mathbf{m}(z^n, u_{-k}^{-1}, u_1^k) \right]. \quad (50)$$

[8] Consequently, our stronger performance guarantees apply also to that approach.

$$L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n) - D_k(x^n, z^n) = \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} [\lambda_{\hat{x}(\mathbf{v})}^T - \lambda_{\hat{x}(\mathbf{q})}^T]\, \mathbf{q}(z^n, x^n, u_{-k}^k)$$

$$\leq \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} [\lambda_{\hat{x}(\mathbf{v})}^T - \lambda_{\hat{x}(\mathbf{q})}^T][\mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{v}(u_{-k}^k)] \tag{48}$$

$$\leq \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \left\|\lambda_{\hat{x}(\mathbf{v})} - \lambda_{\hat{x}(\mathbf{q})}\right\|_\infty \cdot \left\|\mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{v}(u_{-k}^k)\right\|_1$$

$$\leq \frac{\Lambda_{\max}}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \left\|\mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{v}(u_{-k}^k)\right\|_1 \tag{49}$$

For this case, the upper bound (47) can be further upper-bounded as follows.

Lemma 2: For all $x^n, z^n \in \mathcal{A}^n$, and $u_{-k}^{-1}, u_1^k \in \mathcal{A}^k$

$$\sum_{u_0 \in \mathcal{A}} \left\|\mathbf{q}(z^n, x^n, u_{-k}^k) - \pi_{u_0} \odot \left[\mathbf{\Pi}^{-T} \mathbf{m}(z^n, u_{-k}^{-1}, u_1^k)\right]\right\|_1 \leq \left(1 + \|\mathbf{\Pi}^{-1}\|_\infty\right) \sum_{u_0 \in \mathcal{A}} \left\|\mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{q}'(z^n, x^n, u_{-k}^k)\right\|_1 \tag{51}$$

where

$$\mathbf{q}'(z^n, x^n, u_{-k}^k) \triangleq \pi_{u_0} \odot \sum_{b \in \mathcal{A}} \mathbf{q}(z^n, x^n, u_{-k}^{-1} b u_1^k). \tag{52}$$

Lemma 2 is proved in Appendix I.

As hinted by Lemmas 1 and 2, a key step in the proof of Theorem 2 will be to show that, with high probability, the vector $\mathbf{q}'(Z^n, x^n, u_{-k}^k)$ is close to $\mathbf{q}(Z^n, x^n, u_{-k}^k)$. As discussed in Section IV-B, this step is indeed plausible (see (23), where the right-hand side is precisely $\mathbf{q}'(z^n, x^n, u_{-k}^k)$). However, there are two apparent obstacles to making the intuition given in (23) precise. One is that the number of symbols in z^n which occur in double-sided context (u_{-k}^{-1}, u_1^k), and such that the corresponding noiseless symbol is a, is itself a random variable. The other is that these symbols are in general dependent random variables, since their contexts might also consist of symbols with the same property. In the technique that follows, we surmount these difficulties by first deinterleaving z^n into subsequences, and then conditioning the contribution of each subsequence to the right-hand side of (51) on all symbols not in the subsequence. The symbols in each subsequence are just far enough apart for the conditioning to determine each symbol's context, thereby fixing the cardinality and positions of those symbols in the subsequence which occur in double-sided context (u_{-k}^{-1}, u_1^k), and such that the corresponding noiseless symbol is a. Additionally, since the channel is memoryless, the symbols in a subsequence are conditionally independent. Thus, the conditioning permits a conventional analysis, and the final result is obtained by extracting the worst case conditional behavior. To implement this analysis, we first break the statistics $\mathbf{q}(z^n, x^n, u_{-k}^k)$ into partial counts, each corresponding to occurrences of u_0 at indices i such that $i \equiv \ell \bmod (k+1)$, $\ell = 0, 1, \ldots, k$. There are thus k intervening symbols between any two symbols contributing to a given partial count, which is the smallest gap that induces fixed contexts after conditioning on all noncontributing symbols.

Specifically, for $a^n, b^n \in \mathcal{A}^n$, $c^{2k+1} \in \mathcal{A}^{2k+1}$, let $\mathbf{q}_\ell(a^n, b^n, c^{2k+1})[\alpha]$ denote the M-dimensional column vector whose αth component ($\alpha \in \mathcal{A}$) is

$$\mathbf{q}_\ell(a^n, b^n, c^{2k+1})[\alpha] = \left|\left\{i : i \in \mathcal{I}_\ell, a_{i-k}^{i+k} = c^{2k+1}, b_i = \alpha\right\}\right| \tag{53}$$

where

$$\mathcal{I}_\ell \triangleq \{i : k+1 \leq i \leq n-k, i \equiv \ell \bmod (k+1)\}.$$

The cardinality n_ℓ of the index set $\mathcal{I}_\ell$ is $\lfloor (n-\ell-k)/(k+1) \rfloor$. By definition

$$\mathbf{q}(a^n, b^n, c^{2k+1}) = \sum_{\ell=0}^{k} \mathbf{q}_\ell(a^n, b^n, c^{2k+1}).$$

Similarly, we define, as in (52)

$$\mathbf{q}'_\ell(z^n, x^n, u_{-k}^k) \triangleq \pi_{u_0} \odot \sum_{b \in \mathcal{A}} \mathbf{q}_\ell(z^n, x^n, u_{-k}^{-1} b u_1^k).$$

In the sequel, for simplicity, our notation will occasionally omit the first two arguments of the vectors $\mathbf{q}$, $\mathbf{q}'$, $\mathbf{q}_\ell$, and $\mathbf{q}'_\ell$, as these arguments will always be z^n and x^n, respectively. By the triangle inequality, we can further upper-bound the bound in Lemma 2 to obtain

$$\sum_{u_0 \in \mathcal{A}} \left\|\mathbf{q}(u_{-k}^k) - \pi_{u_0} \odot \left[\mathbf{\Pi}^{-T} \mathbf{m}(z^n, u_{-k}^{-1}, u_1^k)\right]\right\|_1 \leq \left(1 + \|\mathbf{\Pi}^{-1}\|_\infty\right) \sum_{\ell=0}^{k} \sum_{u_0 \in \mathcal{A}} \left\|\mathbf{\Delta}_\ell(u_{-k}^k)\right\|_1 \tag{54}$$

where

$$\mathbf{\Delta}_\ell(u_{-k}^k) \triangleq \mathbf{q}_\ell(u_{-k}^k) - \mathbf{q}'_\ell(u_{-k}^k).$$

We will bound each sum $\sum_{u_0} \left\|\mathbf{\Delta}_\ell(u_{-k}^k)\right\|_1$ in probability and expectation, conditioned on the collection of random variables $Z(\ell)$ given by

$$Z(\ell) \triangleq \{Z_i : 1 \leq i \leq n, i \notin \mathcal{I}_\ell\}.$$

We denote by $z(\ell) \in \mathcal{A}^{n-n_\ell}$ a particular realization of $Z(\ell)$. Now, for each ℓ, let

$$n_\ell(u_{-k}^{-1}, u_1^k, a) \triangleq \sum_{b \in \mathcal{A}} \mathbf{q}_\ell(u_{-k}^{-1} b u_1^k)[a]$$

denote the number of times z_i, $i \in \mathcal{I}_\ell$, occurrs in double-sided context (u_{-k}^{-1}, u_1^k), when $x_i = a$. Notice that (for the fixed x^n) *conditioned on* $Z(\ell) = z(\ell)$, $n_\ell(u_{-k}^{-1}, u_1^k, a)$ is *deterministic*, as it depends only on x^n and $z(\ell)$.

Lemma 3: Let

$$F_{\mathbf{\Pi},a} \triangleq \min_{A \subseteq \mathcal{A}} \varphi(\Pi(a, A))$$

where the function $\varphi(\cdot)$ is given in (36), and let

$$V_{\mathbf{\Pi},a} \triangleq \sum_{b \in \mathcal{A}} \sqrt{\Pi(a,b)(1-\Pi(a,b))}.$$

Then, for all $x^n \in \mathcal{A}^n, z(\ell) \in \mathcal{A}^{n-n_\ell}, u_{-k}^{-1}, u_1^k \in \mathcal{A}^k, a \in \mathcal{A}$, and $\varepsilon > 0$, we have

$$\Pr\left(\sum_{u_0 \in \mathcal{A}} \left| \mathbf{\Delta}_\ell(u_{-k}^k)[a] \right| > n_\ell(u_{-k}^{-1}, u_1^k, a)\varepsilon \;\middle|\; Z(\ell) = z(\ell) \right) \leq (2^M - 2) e^{-n_\ell(u_{-k}^{-1}, u_1^k, a) F_{\mathbf{\Pi},a} \frac{\varepsilon^2}{4}} \quad (55)$$

and

$$E\left[\sum_{u_0 \in \mathcal{A}} \left| \mathbf{\Delta}_\ell(u_{-k}^k)[a] \right| \;\middle|\; Z(\ell) = z(\ell) \right] \leq \sqrt{\frac{2}{\pi}} V_{\mathbf{\Pi},a} \sqrt{n_\ell(u_{-k}^{-1}, u_1^k, a)} + M. \quad (56)$$

Remark: Notice that $Z_{\ell+k+1}, Z_{\ell+2(k+1)}, \ldots, Z_{\ell+n_\ell(k+1)}$, are the only random variables in the lemma that have not been fixed.

We will obtain the bound (55) of Lemma 3 by applying the following proposition, where

$$d(p_1 \| p_2) = p_1 \log \frac{p_1}{p_2} + (1-p_1) \log \frac{1-p_1}{1-p_2}$$

will denote the binary divergence, which we take to be ∞ if $p_1 > 1$.

Proposition 1: Let P be a probability distribution on the set $\{1, \ldots, M\}$ and $\mathbf{P} = [P(1), \ldots, P(M)]$. Let $X_1, X_2, \ldots, X_m$ be i.i.d. random variables distributed according to P, and let $\hat{\mathbf{P}}$ denote the probability vector

$$\hat{\mathbf{P}} = \frac{1}{m} \left[\sum_i \mathbf{1}_{\{X_i = 1\}}, \ldots, \sum_i \mathbf{1}_{\{X_i = M\}} \right]$$

corresponding to the empirical distribution $\hat{P}$. Then, for all $\varepsilon > 0$

$$\Pr(\|\mathbf{P} - \hat{\mathbf{P}}\|_1 \geq \epsilon) \leq \sum_{A \subseteq \mathcal{A}} e^{-m d(P(A) + \frac{\epsilon}{2} \,\|\, P(A))} \quad (57)$$

$$\leq (2^M - 2) e^{-m(\min_{A \subseteq \mathcal{A}} \varphi(P(A))) \frac{\epsilon^2}{4}} \quad (58)$$

where the function $\varphi(\cdot)$ is given in (36).

Proof: By the well-known equality

$$\|\mathbf{P} - \hat{\mathbf{P}}\|_1 = 2 \max_{A \subseteq \mathcal{A}} (P(A) - \hat{P}(A)) \quad (59)$$

a union bound implies that

$$\Pr\left(\|\mathbf{P} - \hat{\mathbf{P}}\|_1 \geq \epsilon \right) \leq \sum_{A \subseteq \mathcal{A}} \Pr\left(P(A) - \hat{P}(A) \geq \frac{\epsilon}{2} \right). \quad (60)$$

Applying the standard Chernoff bounding technique to each term in the summation yields (57), as in [32, Theorem 1, eq. (2.1)]. The terms in the summation of (57) corresponding to $A = \emptyset$ and $A = \mathcal{A}$ are clearly 0. Equation (58) then follows from [32, Theorem 1, eq. (2.2)] by lower-bounding the smallest exponent. □

Proof of Lemma 3: For all i such that $x_i = a$, and for each $u_0 \in \mathcal{A}$, we have $\Pr(Z_i = u_0) = \Pi(a, u_0)$. Thus, by definition, conditioned on $Z(\ell) = z(\ell)$, $\mathbf{q}_\ell(Z^n, x^n, u_{-k}^k)[a]$ is the sum of the $n_\ell(u_{-k}^{-1}, u_1^k, a)$ i.i.d. Bernoulli-$\Pi(a, u_0)$ random variables $\mathbf{1}_{\{Z_i = u_0\}}$, where i belongs to the index set

$$\mathcal{I}_\ell(u_{-k}^{-1}, u_1^k, a) \triangleq \{ i \in \mathcal{I}_\ell : z_{i-k}^{i-1} = u_{-k}^{-1}, z_{i+1}^{i+k} = u_1^k, x_i = a \}$$

which is completely determined by x^n and $z(\ell)$. Moreover, by (52)

$$\mathbf{q}'_\ell(Z^n, x^n, u_{-k}^k)[a] = \Pi(a, u_0) n_\ell(u_{-k}^{-1}, u_1^k, a).$$

Therefore, after normalization by $n_\ell(u_{-k}^{-1}, u_1^k, a)$, the sum in the left-hand sides of (55) and (56) is the L_1-distance between the distribution $\Pi(a, u_0)$ on u_0, and the corresponding empirical distribution $\mathbf{q}_\ell(u_{-k}^k)[a] / n_\ell(u_{-k}^{-1}, u_1^k, a)$. The upper bound (55) then follows from Proposition 1 with $P = \Pi(a, \cdot)$ and $m = n_\ell(u_{-k}^{-1}, u_1^k, a)$.

As for the bound on the expectation, notice that each term

$$E\left[\left| \mathbf{\Delta}_\ell(u_{-k}^k)[a] \right| \;\middle|\; Z(\ell) = z(\ell) \right]$$

is the expected magnitude of the difference between the number $S_{m,p}$ of successes in m Bernoulli trials with success probability p and its average mp, with $p = \min\{\Pi(a, u_0), (1 - \Pi(a, u_0))\}$ and $m = n_\ell(u_{-k}^{-1}, u_1^k, a)$. In particular, for $0 \leq p \leq 1/2$ [26, Ch. IX, Problem 35]

$$E|S_{m,p} - mp| = E|S_{m,q} - mq| = 2\nu q \binom{m}{\nu} p^\nu q^{m-\nu} \quad (61)$$

where $q = 1 - p$ and $\nu = \lfloor mp \rfloor + 1$. For a given positive integer ν, it is easy to see that the value of p that maximizes the right-hand side of (61) is $p' = \nu/(m+1)$. Thus, applying Stirling's formula to $\binom{m+1}{\nu}$ we obtain, after straightforward algebraic manipulations

$$E|S_{m,p} - mp| \leq \sqrt{\frac{2(m+1)p'(1-p')}{\pi}}.$$

Clearly

$$p'(1-p') \leq p''(1-p'')$$

where $p'' = \min\{(mp+1)/(m+1), 1/2\}$. Moreover, $p'' \geq p$ and $(m+1)p'' \leq mp + 1$, so that

$$E|S_{m,p} - mp| \leq \sqrt{\frac{2(mp+1)q}{\pi}}.$$

The proof is complete by observing that $\sqrt{mp+1} \le \sqrt{mp}+1$, applying the resulting upper bound to each u_0, and then summing over u_0. □

Regarding Lemma 3, note the following.

a) It can be shown that the smallest exponent of the terms in the summation of (57) coincides with that given by Sanov's theorem and hence is the best possible.[9] A stronger version of Lemma 3, based on this optimal rate, could have been derived. The use of this rate in obtaining a closed-form bound on the probability of Theorem 2, however, appears to require the weaker version (55).

b) The constant $F_{\mathbf{\Pi},a}$ in the exponent of (55) is lower-bounded by 2, and indeed replacing $F_{\mathbf{\Pi},a}$ by 2 coincides simply with the application of the more widely used weaker form of Hoeffding's inequality [32, Theorem 1, eq. (2.3)] to each probability on the right-hand side of (60) in the proof of Proposition 1. Such a bound, however, would not reflect the intuitively appealing fact that less "noisy" channels result in larger denoising exponents.

Proof of Theorem 2: Using Lemma 1 with $\{\mathbf{v}(u_{-k}^{k})\}$ given by (50), and (54), we obtain, for any $\varepsilon > 0$

$$\begin{aligned} P &\triangleq \Pr\left(L_{\hat{X}^{n,k}}(x_{k+1}^{n-k}, Z^n) - D_k(x^n, Z^n) > \varepsilon\right) \\ &\le \Pr\left(\sum_{\ell=0}^{k} \sum_{\mathbf{u}\in\mathcal{A}^{2k+1}} \sum_{a\in\mathcal{A}} |\mathbf{\Delta}_\ell(\mathbf{u})[a]| > \frac{(n-2k)\varepsilon}{C_{\mathbf{\Lambda},\mathbf{\Pi}}}\right) \\ &\le \sum_{\ell=0}^{k} \Pr\left(\sum_{\mathbf{u}\in\mathcal{A}^{2k+1}} \sum_{a\in\mathcal{A}} |\mathbf{\Delta}_\ell(\mathbf{u})[a]| > \frac{(n-2k)\gamma_\ell\varepsilon}{C_{\mathbf{\Lambda},\mathbf{\Pi}}}\right) \end{aligned} \tag{62}$$

where $\{\gamma_\ell\}$ is a set of nonnegative constants (to be specified later) satisfying $\sum_{\ell=0}^{k}\gamma_\ell = 1$, and the last inequality follows from the union bound. To further upper-bound each probability in the rightmost side of (62) via Lemma 3, we condition the events on the random variables $Z(\ell)$, to obtain (63) at the bottom of the page. Letting P_ℓ denote the conditional probability in the right-hand side of (63), the union bound yields the second equation at the bottom of the page, where, again, conditioned on $Z(\ell)$, $\{\beta_{a,\mathbf{u}}\} \triangleq \{\beta_{a,\mathbf{u}_L,\mathbf{u}_R}\}$ is a set of nonnegative constants (to be specified later) satisfying $\sum_{\mathbf{u}_L,\mathbf{u}_R,a}\beta_{a,\mathbf{u}} = 1$. We can now apply (55) in Lemma 3, which yields (64) at the bottom of the page. Now, choose

$$\beta_{a,\mathbf{u}} = \frac{\sqrt{n_\ell(\mathbf{u}_L,\mathbf{u}_R,a)/F_{\mathbf{\Pi},a}}}{\sum_{\mathbf{u}_L,\mathbf{u}_R\in\mathcal{A}^k}\sum_{a\in\mathcal{A}}\sqrt{n_\ell(\mathbf{u}_L,\mathbf{u}_R,a)/F_{\mathbf{\Pi},a}}}$$

so that

$$\begin{aligned} \frac{n_\ell(\mathbf{u}_L,\mathbf{u}_R,a)}{F_{\mathbf{\Pi},a}\beta_{a,\mathbf{u}}^2} &= \left(\sum_{\mathbf{u}_L,\mathbf{u}_R\in\mathcal{A}^k}\sum_{a\in\mathcal{A}}\sqrt{n_\ell(\mathbf{u}_L,\mathbf{u}_R,a)/F_{\mathbf{\Pi},a}}\right)^2 \\ &\le n_\ell M^{2k}\sum_{a\in\mathcal{A}} F_{\mathbf{\Pi},a}^{-1} = M^{2k} n_\ell F_{\mathbf{\Pi}} \end{aligned}$$

where we used the Cauchy–Schwarz inequality and the fact that $\sum_{\mathbf{u}_L,\mathbf{u}_R}\sum_a n_\ell(\mathbf{u}_L,\mathbf{u}_R,a) = n_\ell$. With this choice, which equalizes the exponents in (64), (63) and (64) yield the fourth equation at the bottom of the page. We complete the proof of the bound (37) by choosing

$$\gamma_\ell = \frac{\sqrt{n_\ell}}{\sum_j \sqrt{n_j}}$$

applying similarly the Cauchy–Schwarz inequality, and using the fact that $\sum_{\ell=0}^{k} n_\ell = n-2k$.

To prove the bound (38), we use again Lemma 1 with $\{\mathbf{v}(u_{-k}^{k})\}$ given by (50), and (54), to obtain the equation at the top of the following page. By (56) in Lemma 3, we can further upper-bound the expectation to obtain the second set of expressions at the top of the following page, where the second and fourth lines follow from the Cauchy–Schwarz inequality, completing the proof of (38). □

[9] Moreover, note that the bound in (57) is preferable since it avoids the factor resulting from the use of the method of types in Sanov's theorem, which is *polynomial* in m; cf., e.g., [14, Theorem 12.4.1].

$$P \le \sum_{\ell=0}^{k}\sum_{z(\ell)\in\mathcal{A}^{n-n_\ell}} \Pr(Z(\ell)=z(\ell))\Pr\left(\sum_{\mathbf{u}\in\mathcal{A}^{2k+1}}\sum_{a\in\mathcal{A}}|\mathbf{\Delta}_\ell(\mathbf{u})[a]| > \frac{(n-2k)\gamma_\ell\varepsilon}{C_{\mathbf{\Lambda},\mathbf{\Pi}}}\,\middle|\, Z(\ell)=z(\ell)\right). \tag{63}$$

$$P_\ell \le \sum_{\mathbf{u}_L,\mathbf{u}_R\in\mathcal{A}^k}\sum_{a\in\mathcal{A}} \Pr\left(\sum_{u_0\in\mathcal{A}}|\mathbf{\Delta}_\ell(\mathbf{u}_L u_0 \mathbf{u}_R)[a]| > \frac{(n-2k)\gamma_\ell\beta_{a,\mathbf{u}}\varepsilon}{C_{\mathbf{\Lambda},\mathbf{\Pi}}}\,\middle|\, Z(\ell)=z(\ell)\right)$$

$$P_\ell \le (2^M-2)\sum_{\mathbf{u}_L,\mathbf{u}_R\in\mathcal{A}^k}\sum_{a\in\mathcal{A}}\exp\left(-\frac{F_{\mathbf{\Pi},a}(n-2k)^2\gamma_\ell^2\beta_{a,\mathbf{u}}^2}{4n_\ell(\mathbf{u}_L,\mathbf{u}_R,a)}\cdot\frac{\varepsilon^2}{C_{\mathbf{\Lambda},\mathbf{\Pi}}^2}\right). \tag{64}$$

$$P \le (2^M-2)M^{2k+1}\sum_{\ell=0}^{k}\exp\left(-\frac{(n-2k)^2\gamma_\ell^2}{4M^{2k}n_\ell F_{\mathbf{\Pi}}}\cdot\frac{\varepsilon^2}{C_{\mathbf{\Lambda},\mathbf{\Pi}}^2}\right)\sum_{z(\ell)\in\mathcal{A}^{n-n_\ell}}\Pr(z(\ell)).$$

$$E \triangleq E\left[(n-2k)\left(L_{\hat{X}^{n,k}}(x_{k+1}^{n-k}, Z^n) - D_k(x^n, Z^n)\right)\right] \le C_{\mathbf{\Lambda},\mathbf{\Pi}} \sum_{\ell=0}^{k} \sum_{\mathbf{u}\in\mathcal{A}^{2k+1}} \sum_{a\in\mathcal{A}} E\left[\,|\mathbf{\Delta}_\ell(\mathbf{u})[a]|\,\right]$$

$$= C_{\mathbf{\Lambda},\mathbf{\Pi}} \sum_{\ell=0}^{k} \sum_{\mathbf{u}_L,\mathbf{u}_R\in\mathcal{A}^k} \sum_{a\in\mathcal{A}} \sum_{z(\ell)\in\mathcal{A}^{n-n_\ell}} \Pr(z(\ell))\, E\left[\sum_{u_0\in\mathcal{A}} |\,\mathbf{\Delta}_\ell(\mathbf{u}_L u_0 \mathbf{u}_R)[a]\,| \,\middle|\, Z(\ell) = z(\ell)\right].$$

$$\begin{aligned} E &\le C_{\mathbf{\Lambda},\mathbf{\Pi}} \sum_{\ell=0}^{k} \sum_{z(\ell)\in\mathcal{A}^{n-n_\ell}} \Pr(z(\ell)) \sum_{\mathbf{u}_L,\mathbf{u}_R\in\mathcal{A}^k} \sum_{a\in\mathcal{A}} \left(\sqrt{\frac{2}{\pi}} V_{\mathbf{\Pi},a} \sqrt{n_\ell(\mathbf{u}_L,\mathbf{u}_R,a)} + M\right) \\ &\le C_{\mathbf{\Lambda},\mathbf{\Pi}} \sum_{\ell=0}^{k} \sum_{z(\ell)\in\mathcal{A}^{n-n_\ell}} \Pr(z(\ell)) \left(\sqrt{\frac{2}{\pi}} \sqrt{M^{2k} V_{\mathbf{\Pi}}^2 n_\ell} + M^{2k+2}\right) \\ &= C_{\mathbf{\Lambda},\mathbf{\Pi}} \sum_{\ell=0}^{k} \left(\sqrt{\frac{2}{\pi}} V_{\mathbf{\Pi}} M^k \sqrt{n_\ell} + M^{2k+2}\right) \\ &\le \sqrt{\frac{2}{\pi}} C_{\mathbf{\Lambda},\mathbf{\Pi}} V_{\mathbf{\Pi}} M^k \sqrt{(k+1)(n-2k)} + C_{\mathbf{\Lambda},\mathbf{\Pi}}(k+1) M^{2k+2} \end{aligned}$$

VI. The Stochastic Setting

Consider the fully stochastic analogue of the setting of Section V where the underlying noiseless signal is a stochastic process rather than an individual sequence. Specifically, we assume that $\mathbf{Z}$ is the output of the memoryless, invertible, channel $\mathbf{\Pi}$ whose input is the double-sided stationary process $\mathbf{X}$. Letting $\mathbf{P}_{X^n}, \mathbf{P}_{\mathbf{X}}$ denote, respectively, the distributions of $X^n, \mathbf{X}$, and $\mathcal{D}_n$ denote the class of all n-block denoisers, define

$$\mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) = \min_{\hat{X}^n\in\mathcal{D}_n} E\left[L_{\hat{X}^n}(X^n, Z^n)\right] \tag{65}$$

the expectation on the right-hand side assuming that X^n was generated according to $\mathbf{P}_{X^n}$ (and Z^n is the output of the DMC $\mathbf{\Pi}$ whose input is X^n). By stationarity, for all $m, n \ge 0$

$$(m+n)\mathbb{D}(\mathbf{P}_{X^{m+n}}, \mathbf{\Pi}) \le m\mathbb{D}(\mathbf{P}_{X^m}, \mathbf{\Pi}) + n\mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}). \tag{66}$$

Thus, by the Subadditivity Lemma (cf., e.g., [18, Lemma 6.1.11])

$$\lim_{n\to\infty} \mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) = \inf_{n\ge1} \mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) \triangleq \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}). \tag{67}$$

By definition, $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$ is the (distribution-dependent) optimal asymptotic denoising performance attainable when the noiseless signal is emitted by the source $\mathbf{P}_{\mathbf{X}}$ and corrupted by the channel $\mathbf{\Pi}$. The main goal of this section is to establish the fact that the DUDE asymptotically attains $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$ no matter what stationary source has emitted $\mathbf{X}$. Note that in the definition leading to $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$ we minimize over *all* denoising schemes, not necessarily sliding-block schemes of the type considered in Section V. This is in accord with analogous situations in universal compression [79], prediction [41], and noisy prediction [72], where in the individual-sequence setting the set of schemes in the comparison class is limited in some computational sense. In the fully stochastic setting, on the other hand, such a limitation takes the form of a restriction on the class of allowable sources (cf. discussion on the duality between the viewpoints in [41]).

Let $\mathbf{P}_{X_0|z_i^j} \in \mathcal{M}$ denote the M-dimensional probability vector whose ath component is[10] $P(X_0 = a | Z_i^j = z_i^j)$.

Claim 2: $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) = E\left[U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})\right]$.

The claim results from its well-known counterpart for a finite set of noisy observations, and from the following lemma.

Lemma 4:

1) For $k, l \ge 0$, $E\left[U(\mathbf{P}_{X_0|Z_{-k}^{l}})\right]$ is decreasing in both k and l.
2) For any two unboundedly increasing sequences of positive integers $\{k_n\}, \{l_n\}$
$$\lim_{n\to\infty} E\left[U(\mathbf{P}_{X_0|Z_{-k_n}^{l_n}})\right] = E\left[U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})\right].$$

Lemma 4 and Claim 2 parallel similar results in sequential decision theory [41] (e.g., in the data compression case, the limiting values of the block and conditional entropies coincide, defining the entropy rate). Their proofs are also standard, but are given in Appendix II for completeness.

The first main result of this section, Theorem 3, follows now from the properties shown for the semi-stochastic setting and the above claims.

Theorem 3: The sequence of denoisers $\{\hat{X}_{\mathsf{univ}}^n\}$ in (14) with $\lim_{n\to\infty} k_n = \infty$ satisfies

$$\lim_{n\to\infty} E\left[L_{\hat{X}_{\mathsf{univ}}^n}(X^n, Z^n)\right] = \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$$

provided that $\sqrt{k_n} M^{k_n} = o(\sqrt{n})$.

Proof: By Claim 2

$$\begin{aligned} &E L_{\hat{X}_{\mathsf{univ}}^n}(X^n, Z^n) - \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \\ &\quad = E\left[U(\mathbf{P}_{X_0|Z_{-k_n}^{k_n}}) - U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})\right] \\ &\qquad + E\left[L_{\hat{X}_{\mathsf{univ}}^n}(X^n, Z^n) - U(\mathbf{P}_{X_0|Z_{-k_n}^{k_n}})\right]. \end{aligned} \tag{68}$$

[10]The definition is rigorously extended to cases with $i = -\infty$ and/or $j = \infty$, by assuming $\mathbf{P}_{X_0|z_i^j}$ to be a regular version of the conditional distribution (cf., e.g., [23]) of X_0 given Z_i^j, evaluated at z_i^j.

The first expectation on the right-hand side of (68) vanishes in the limit by Lemma 4, whereas for the second expectation we notice that, for any $k \geq 0$

$$E\Big[U(\mathbf{P}_{X_0|Z_{-k}^k})\Big] = \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} E\left[\Lambda(X_0, f(Z_{-k}^k))\right] \quad (69)$$

$$= \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} E\left[\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda(X_i, f(Z_{i-k}^{i+k}))\right] \quad (70)$$

$$\geq E\left[\min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda(X_i, f(Z_{i-k}^{i+k}))\right] \quad (71)$$

$$= E\left[D_k(X^n, Z^n)\right] \quad (72)$$

where (70) follows by stationarity. Thus, the second expectation in the right-hand side of (68) vanishes in the limit by Theorem 1, Part b). □

Regarding Theorem 3, note the following.

a) Equation (68) provides insight into the convergence of $E\left[L_{\hat{X}^n_{\text{univ}}}(X^n, Z^n)\right]$ to $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$. The vanishing rate of the first expectation in the right-hand side depends on the underlying process, and there is no upper bound on this rate which holds uniformly for all stationary $\mathbf{X}$. In contrast, the second expectation is uniformly bounded by Theorem 1, Part b). A slower growing rate for k_n yields a faster vanishing rate for the second expectation but the price, of course, is a slower vanishing rate for the first one.

b) Inequality (71) parallels the well-known property that the conditional entropy of order k is an upper bound on the expectation of the corresponding empirical entropy.

Theorem 3 guarantees the asymptotic expected performance of the DUDE for any stationary noiseless process. Our goal in the remainder of this section is to establish the sample path optimality of the DUDE, namely, the fact that its actual (rather than expected) asymptotic performance is optimal, universally for all stationary and *ergodic* sources. Our first step toward this end is to show that when $\mathbf{P}_{\mathbf{X}}$ is also ergodic, the sliding-window minimum loss (cf. the definition (41)) of the emitted source sequence coincides, with probability one, with $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$. This result parallels [79, Theorem 4], where it is shown that the finite-state compressibility of a sequence drawn from a stationary ergodic source coincides with the entropy of the source with probability one.

Claim 3: If $\mathbf{P}_{\mathbf{X}}$ is stationary and ergodic then, with probability one

$$D(\mathbf{X}, \mathbf{Z}) = \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}).$$

Proof: Recall the definition of $D_k(\mathbf{X}, \mathbf{Z})$ in (42), and notice that assuming stationarity and ergodicity of $\mathbf{P}_{\mathbf{X}}$, for each k and each map f taking $\mathcal{A}^{2k+1}$ into $\mathcal{A}$

$$\lim_{n\to\infty}\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda(X_i, f(Z_{i-k}^{i+k})) = E\left[\Lambda(X_0, f(Z_{-k}^k))\right] \quad \text{a.s.} \quad (73)$$

Since the set of all maps taking $\mathcal{A}^{2k+1}$ into $\mathcal{A}$ is finite, (73) implies

$$D_k(\mathbf{X}, \mathbf{Z}) = \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} E\left[\Lambda(X_0, f(Z_{-k}^k))\right] \quad \text{a.s.}$$

The proof is completed by letting $k \to \infty$ in (69) and invoking Lemma 4 and Claim 2. □

Our main result on the sample path behavior of the DUDE now follows from Claim 3 and properties established in the semi-stochastic setting.

Theorem 4: The sequence of denoisers $\{\hat{X}^n_{\text{univ}}\}$ with $\lim_{n\to\infty} k_n = \infty$ satisfies

$$\limsup_{n\to\infty} L_{\hat{X}^n_{\text{univ}}}(X^n, Z^n) = \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \quad \text{a.s.}$$

for every stationary and ergodic $\mathbf{P}_{\mathbf{X}}$, provided that $k_n M^{2k_n} = o(n/\log n)$.

Proof: Corollary 1 (in Section V) holds for all sequences $\mathbf{x}$ and, *a fortiori*, almost surely. It thus follows by invoking Claims 1 and 3 that

$$\limsup_{n\to\infty} L_{\hat{X}^n_{\text{univ}}}(X^n, Z^n) \leq \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \quad \text{a.s.} \quad (74)$$

The reverse inequality follows from Fatou's lemma and the fact (obvious from the definition of $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi})$) that

$$\limsup_{n\to\infty} E\left[L_{\hat{X}^n_{\text{univ}}}(X^n, Z^n)\right] \geq \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}). \quad \square$$

Remark: For stationary and ergodic $\mathbf{P}_{\mathbf{X}}$ and for the range of values of k_n covered by Theorem 4, the convergence in expectation of Theorem 3 could be directly derived from Theorem 4 via Fatou's lemma. In fact, this approach can be employed without the additional ergodicity requirement, by first conditioning on the ergodic mode and then applying Theorem 4 separately on each mode.

VII. Context-Length Selection

A. The "Best" k

The optimality results shown in the preceding sections provide asymptotic guidance on the choice of the context length for universal denoising. However, these results refer to a sequence of problems, shedding little light on how the order k ought to be selected for a specific sequence z^n. In particular, notice that even though Theorem 2 provides *nonasymptotic* information about how the denoiser $\hat{X}^{n,k}$ compares with the best kth-order sliding-window denoiser, it does not address the issue of comparing different choices of k.

The problem of choosing k is, in many aspects, similar to that of double-universality in universal data compression (see, e.g., [41]). In the data compression case, once a specific algorithm is shown to be universal for a given model structure parameterized by a value k (e.g., a mixture probability assignment for some Markovian order k), the question arises as to what value of k minimizes the code length assigned by that specific algorithm to an individual sequence. Notice that a comparison class is used for analyzing universality for a given k, but once a universal algorithm is selected, the criterion for choosing k is independent of the comparison class. The encoder can, for example,

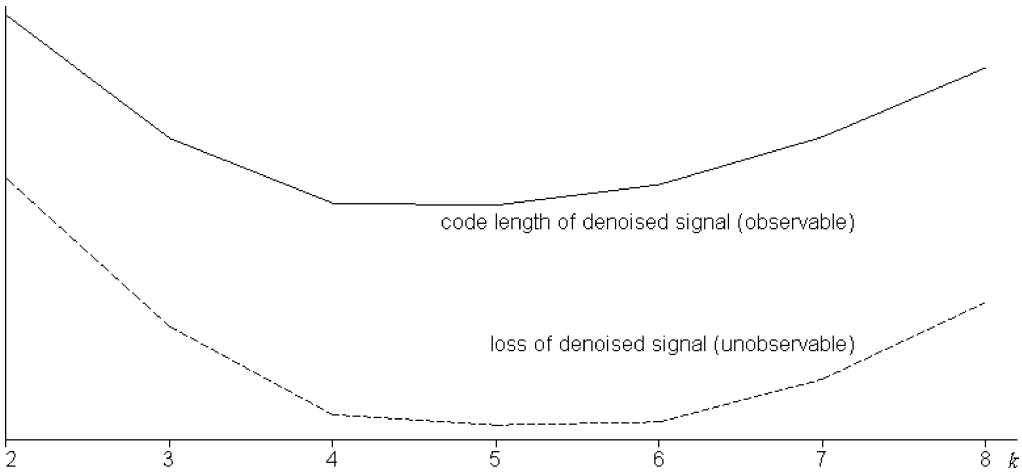

Fig. 1. Code length and loss of denoised signal as a function of k.

search for the optimal k, and describe its value to the decoder. The key difference in the denoising problem, however, is that the sequence x^n on which the optimal k depends is not observable. Yet, the data compression analogy suggests the following formalization as a possible criterion for choosing k.

For a given pair (x^n, z^n), let

$$k^*(x^n, z^n) \triangleq \arg\min_k L_{\hat{X}^{n,k}}(x^n, z^n).$$

In words: $k^*(x^n, z^n)$ is the optimal order for a denoiser having the same form as the DUDE. This best value of k is, of course, unavailable as it depends on x^n. Now, define the function $k_n : \mathcal{A}^n \to \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ given by (75) at the bottom of the page. The order $k_n(z^n)$ provides a possible benchmark for choosing k as a function of z^n (as opposed to the order k_n in previous sections, which depends just on n and was selected based on considerations of asymptotic optimality). This choice aims at minimizing, in the worst case of x^n, the expected *excess* loss over the loss we would have achieved with the optimal order $k^*(x^n, z^n)$ (namely, the "regret"). With $k_{\max} = \max_{z^n \in \mathcal{A}^n} k_n(z^n)$, $\hat{X}^{n,k(\cdot)}$ is a $k_{\max}$th-order sliding-window denoiser. Notice that (75) would provide a sensible order selection criterion only if this worst case regret vanishes asymptotically. Beyond this open problem, $k_n(z^n)$ seems difficult to compute even in the simplest cases, and in the next subsection we consider heuristics for selecting k, or more generally, an appropriately sized context model, in practice.

B. Heuristics for Choice of Context Size

As mentioned, choosing "the best" k seems to present some theoretical and practical difficulties. Ideally, we would like to be able to choose a value of k that approaches the DUDE's best denoising performance for the given input data sequence, and such that its determination from observable quantities is computationally feasible. Fortunately, it was observed in experiments where the original noiseless sequence x^n was available as a reference, that the value of k that minimizes the loss $L_{\hat{X}^{n,k}}(x^n, z^n)$ is consistently close to the value that makes $\hat{X}^{n,k}(z^n)$ most compressible. The intuition behind this heuristic is similar to the motivation for the compression-based schemes for denoising: the better the denoising performance is, the more structure of the noiseless signal is unveiled, and thus the higher its compressibility. Compressibility of $\hat{X}^{n,k}(z^n)$ can be estimated from observable data by using a practical implementation of a universal lossless compression scheme. Fig. 1 shows (suitably scaled) typical plots of compressed code length and loss of the denoised signal as a function of k, corresponding to one of the data sets reported on in Section VIII. All data sets mentioned in Section VIII actually exhibit a similar behavior. A formalization of the link between compressibility and the best k for denoising is an open problem of theoretical and practical interest.

The preceding discussion also applies to more general context models, in which the context length depends not only on z^n, but may vary from location to location, similar to the tree models customary in data compression (see, e.g., [68], [76]). Moreover, the context length need not be equal on the left and on the right (see [80] for a formal definition). As mentioned in Section IV, the internal data structure of the DUDE can be readily designed to support these models. Choosing an appropriately sized context model is important in all applications, but essential in applications with large alphabets (e.g., continuous tone images), as is evident from the error terms in Theorem 2 in Section V. Similar issues of *model cost* [49] have been addressed in related areas of lossless image compression (see, for instance, [10]), and significant knowledge and experience have been generated, which can be brought to bear on the discrete denoising problem.

VIII. Experimental Results and Practical Considerations

In this section, we report on experimental results obtained by applying the DUDE to a variety of noise-corrupted data sets.

$$k_n(\cdot) \triangleq \arg\min_{\kappa(\cdot)} \max_{x^n \in \mathcal{A}^n} E\left[L_{\hat{X}^{n,\kappa(Z^n)}}(x^n, Z^n) - L_{\hat{X}^{n,k^*(x^n,Z^n)}}(x^n, Z^n)\right]. \quad (75)$$

TABLE I
BIT-ERROR RATE OF DENOISED SEQUENCES EMITTED BY A MARKOV SOURCE THROUGH A BSC

	$\delta = 0.01$		$\delta = 0.10$		$\delta = 0.20$	
p	DUDE	Bayes	DUDE	Bayes	DUDE	Bayes
0.01	$0.072\,\delta$ [3]	$0.072\,\delta$	$0.066\,\delta$ [5]	$0.057\,\delta$	$0.127\,\delta$ [6]	$0.082\,\delta$
0.05	$0.422\,\delta$ [3]	$0.420\,\delta$	$0.301\,\delta$ [5]	$0.297\,\delta$	$0.375\,\delta$ [5]	$0.358\,\delta$
0.10	$1.021\,\delta$ [8]	$1.002\,\delta$	$0.560\,\delta$ [3]	$0.557\,\delta$	$0.602\,\delta$ [4]	$0.593\,\delta$
0.15	$1.017\,\delta$ [8]	$1.005\,\delta$	$0.755\,\delta$ [5]	$0.752\,\delta$	$0.766\,\delta$ [4]	$0.765\,\delta$
0.20	$0.999\,\delta$ [8]	$0.994\,\delta$	$0.923\,\delta$ [3]	$0.923\,\delta$	$0.882\,\delta$ [4]	$0.881\,\delta$

A. Binary-Symmetric Markov Source Corrupted by a BSC

We implemented the DUDE for the BSC, as derived in Section IV-D. A first-order symmetric binary Markov source was simulated and corrupted by a simulated BSC for five values of the transition probability p associated with the Markov source, $\{0.01, 0.05, 0.1, 0.15, 0.2\}$, and for three values of the crossover probability δ associated with the BSC, $\{0.01, 0.1, 0.2\}$. In each case, only one realization of the source was generated (with sequence length $n = 10^6$).

Table I shows the bit-error rate (expressed as a multiple of δ) of the denoised signal obtained when employing the DUDE for the 15 combinations of the pair (p, δ). The number in square brackets is the value of k employed, which was obtained using the compressibility heuristic described in Section VII-B. For each combination, we also show the residual error rate of the optimal Bayesian distribution-dependent scheme tailored for the specific corresponding value of the pair (p, δ), as implemented by the forward–backward recursions [13], [4]. Note that at times the Bayes solution is shown to give worse bit-error rate than the crossover probability of the channel, because in this experiment no averaging is performed with respect to either input or channel realization.

We observe that in the majority of the cases shown in the table, the DUDE approaches optimum performance within a rather small margin. The somewhat less negligible gaps between the performance of the DUDE and that of the optimal scheme are observed in the first line of the table, corresponding to $p = 0.01$. A qualitative explanation for this performance may be that in this case the process is less mixing or more "slowly varying," so in order to approach the performance of the optimal scheme (which bases its denoising decisions for each location on the whole noisy signal) to within a certain margin, a sliding-window denoiser of higher order is needed. However, the sequence length in the experiments is probably not sufficient to get close enough to the optimum for these larger values of k.

B. Text Denoising

We employed the DUDE on a corrupted version of *Don Quixote de La Mancha* (in English translation), by Miguel de Cervantes Saavedra (1547–1616). The text of this novel[11] consists of approximately $2.3 \cdot 10^6$ characters. It was artificially corrupted by flipping each letter, independently, with probability 0.05, equiprobably into one of its nearest neighbors in the QWERTY keyboard. The resulting number of errors in the corrupted text came out to 89 087. The DUDE, employed with $k = 2$, reduced the number of errors to 50 250, which is approximately a 44% error-correction rate. Following are two segments from the corrupted text, with the corresponding DUDE output.

1) *Noisy Text (21 errors):*

```
"Whar giants?" said Sancho Panza. "Those
thou seest theee," snswered yis master,
"with the long arms, and spne have tgem
ndarly two leagues long." "Look, ylur wor-
ship," sair Sancho; "what we see there zre
not gianrs but windmills, and what seem
to be their arms are the sails that turned
by the wind make rhe millstpne go." "Kt is
easy to see," replied Don Quixote, "that
thou art not used to this business of ad-
ventures; fhose are giantz; and if thou
arf wfraod, away with thee out of this and
betake thysepf to prayer while I engage
them in fierce and unequal combat."
```

DUDE output (7 errors):

```
"What giants?" said Sancho Panza. "Those
thou seest there," answered his master,
"with the long arms, and spne have them
nearly two leagues long." "Look, your wor-
ship," said Sancho; "what we see there are
not giants but windmills, and what seem
to be their arms are the sails that turned
by the wind make the millstone go." "It is
easy to see," replied Don Quixote, "that
thou art not used to this business of ad-
ventures; fhose are giantz; and if thou
arf wfraod, away with thee out of this and
betake thyself to prayer while I engage
them in fierce and unequal combat."
```

2) *Noisy Text (4 errors):*

```
… in the service of such a masger ws Dpn
Qhixote …
```

DUDE output (0 errors):

```
… in the service of such a master as Don
Quixote …
```

C. Image Denoising

The binary implementation of the DUDE was used to denoise binary images corrupted by BSCs of various parameter values. In this setting, the input to the denoiser is a sequence $z^{m \times n}$, with components $z_\ell \in \{0, 1\}$, where $\ell = (i, j)$, $1 \le i \le m$, $1 \le j \le n$. We define two-dimensional context patterns as follows. Let $(0, 0), (-1, 0), (1, 0), \ldots$, be an ordering of $\mathbb{Z}^2$ by increasing L_2 norm, with ties broken first by increasing L_∞ norm, then by increasing value of j, and finally by increasing value of i. Denote by Δ_t, $t \ge 0$, the tth integer pair in the order. For an integer $K \ge 0$, the Kth order context for z_ℓ consists of the symbols with coordinates $\ell + \Delta_1, \ell + \Delta_2, \ldots, \ell + \Delta_K$ (with appropriate provisions for image boundaries).

[11] Available online from the Project Gutenberg website at http://promo.net/pg/.

For the image experiments, an attempt was made to estimate the BSC parameter δ, rather than assume it known. It was found that given K, a good estimate of δ is given by[12]

$$\hat{\delta} = \min_{\mathbf{c}} \frac{\min\{\mathbf{m}(z^{m\times n}, \mathbf{c})[0], \mathbf{m}(z^{m\times n}, \mathbf{c})[1]\}}{\mathbf{m}(z^{m\times n}, \mathbf{c})[0] + \mathbf{m}(z^{m\times n}, \mathbf{c})[1]}$$

the minimum taken over contexts $\mathbf{c} \in \mathcal{A}^K$ that occur in $z^{m\times n}$ with frequency surpassing a given threshold (to avoid "diluted" contexts). The intuition behind this heuristic is that if the image is denoisable, then some significant context must exhibit skewed statistics, where the least probable symbol has a low count, thus "exposing" the outcomes of the BSC. Notice that this estimate of δ can be computed after running the first pass of the DUDE, and used during the second pass.

The compressibility heuristic of Section VII-B was used to determine the context order K. The steps of empirically estimating δ and K might need to be iterated, as the estimate of one depends on the estimate of the other. In practice, however, it was observed that very few, if any, iterations are needed if one starts from a reasonable guess of the channel parameter. The best K is estimated given this guess, and from it a more accurate estimate of δ is obtained. In the majority of cases, no further iterations were needed.

We now present denoising results for two images. The first image is the first page from a scanned copy of a famous paper [59]. The results are shown in the upper portion of Table II, which lists the normalized bit-error rate of the denoised image, relative to the original one. The table also shows results of denoising the same image with a 3×3 median filter [29], and a morphological filter [62] available under MATLAB. The results for the morphological filter are for the best ordering of the morphological open and close operations based on a 2×2 structural element, which was found to give the best performance. The results in the table show that the DUDE significantly outperforms the reference filters. Fig. 2 shows corresponding portions of the noiseless, noisy, and DUDE-denoised images, respectively, for the experiment with $\delta = 0.05$ (the whole image is not shown due to space constraints and to allow easy comparison of the three versions).

The second image reported on is a halftoned portrait of a famous physicist. While it is arguable whether denoising of halftone images is a common application, these images provide good test cases for a denoiser, which has to distinguish between the random noise and the "texture" of the half-tone pattern. The numerical results are shown in the lower part of Table II, which shows that the DUDE is able to achieve significant denoising of the half-tone. In contrast, the more traditional (median and morphological filtering) algorithms fail, and, in fact, significantly increase the bit-error rate as well as the perceived distortion. Portions of the noiseless, noisy, and DUDE-denoised half-tone images for the experiment with $\delta = 0.02$ are shown in Fig. 3. The experiments on halftones serve to showcase the universality of the DUDE: the same algorithm that performed well on the scanned text of the first example, also performs well for the halftoned photograph, a very different type of image.

[12]The vector-valued function $\mathbf{m}(\cdot)$ now takes two arguments, as $\mathbf{c}$ represents the whole context, which was represented by b^k, c^k in the one-dimensional case.

TABLE II
BIT-ERROR RATES OF DENOISED BINARY IMAGES

		Channel parameter δ			
Image	Scheme	0.01	0.02	0.05	0.10
Shannon 1800×2160	DUDE	$0.096\,\delta$ K=11	$0.090\,\delta$ K=12	$0.082\,\delta$ K=12	$0.091\,\delta$ K=12
	median	$0.483\,\delta$	$0.285\,\delta$	$0.164\,\delta$	$0.141\,\delta$
	morpho.	$0.270\,\delta$	$0.195\,\delta$	$0.162\,\delta$	$0.161\,\delta$
Einstein 896×1160	DUDE	$0.350\,\delta$ K=18	$0.375\,\delta$ K=14†	$0.362\,\delta$ K=12†	$0.391\,\delta$ K=12†
	median	$15.60\,\delta$	$7.90\,\delta$	$3.28\,\delta$	$1.80\,\delta$
	morpho.	$14.90\,\delta$	$7.55\,\delta$	$3.26\,\delta$	$1.93\,\delta$

One-dimensional contexts of size K, consisting of $K/2$ samples to the left, and $K/2$ to the right of the denoised sample, were used in these cases to obtain the best results. While a two-dimensional context scheme obtains bit-error rates that are not far from those reported, the visual quality of the denoised halftone was superior with the one-dimensional contexts.

D. Other Practical Considerations and Improvements

We briefly mention a few other possible avenues for improvement of the DUDE's performance in practical settings, in addition to those discussed in conjunction with the experimental results. Given the diversity of applications of the algorithm, we expect that additional structure, specific to each application, could be exploited to improve performance.

- **Context aggregation.** The crux of the DUDE algorithm is the estimation of the empirical statistics of the noiseless sequence x^n from those of the noisy z^n. If the context of a particular symbol has been contaminated by one or more errors, the count corresponding to the symbol will be credited to the "wrong" context, and, conversely, the statistics used for the correction of the symbol will be partially based on counts of the "wrong" contexts. Thus, the statistics of contexts that are close, in the sense of a higher probability of confusion due to the channel, get intermixed. This suggests a strategy of making decisions based on the counts obtained not only from the observed context, but also from neighboring contexts. To that end, the first pass of the denoiser proceeds as usual; for the second pass, the counts of similar contexts are aggregated, weighing them by the similarity of the context to the observed one. The aggregation of statistics can occur before the second pass, and its complexity is independent of the data size. The context aggregation mechanism is different from, and complementary to, the techniques of context tree pruning from the data compression literature [68], [76] mentioned in Section VII-B. In particular, context aggregation need not reduce the size of the context model. The idea of estimating a symbol at a given location based on counts associated not only with the context observed at that location, but also with other contexts that are only similar (but not identical) to it, has been used in [38] (cf. also [36], [37]) for recovering a noise-corrupted chaotic signal.
- **Nonstationary data.** While the algorithm presented in this work is well suited for stationary sources (or for individual sequences having a low sliding-window minimum

top-right: original

bottom-left: noisy, δ=0.05

bottom-right: denoised, k=12 (2D)

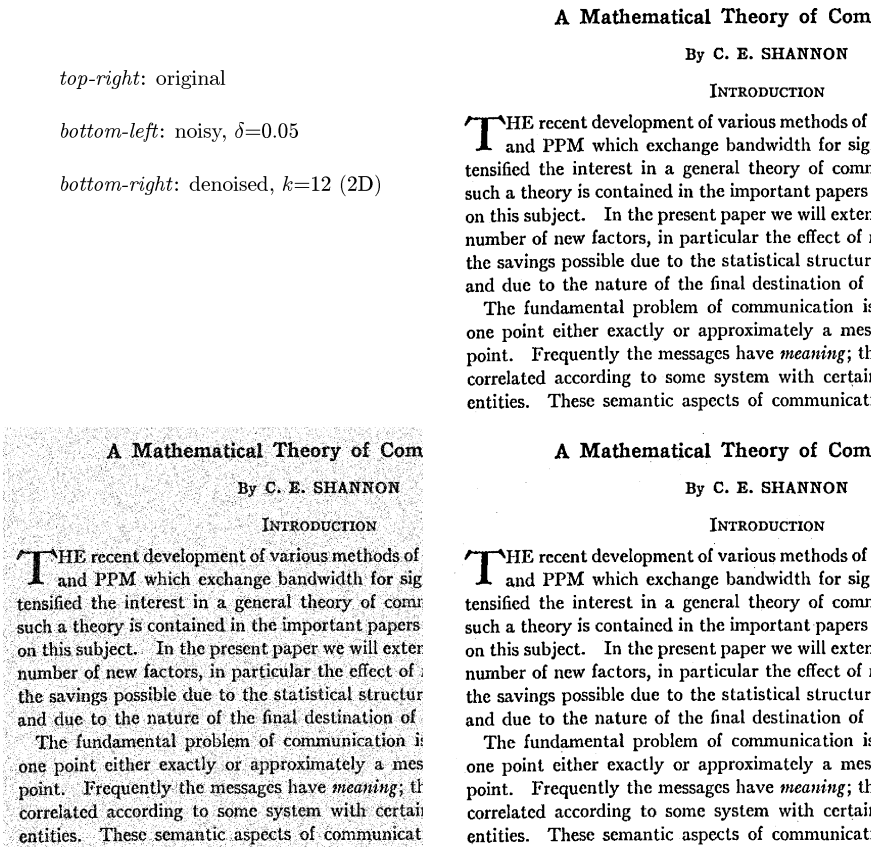

Fig. 2. Denoising of a scanned text image.

top-right: original

bottom-left: noisy, δ=0.02

bottom-right: denoised, k=14 (1D)

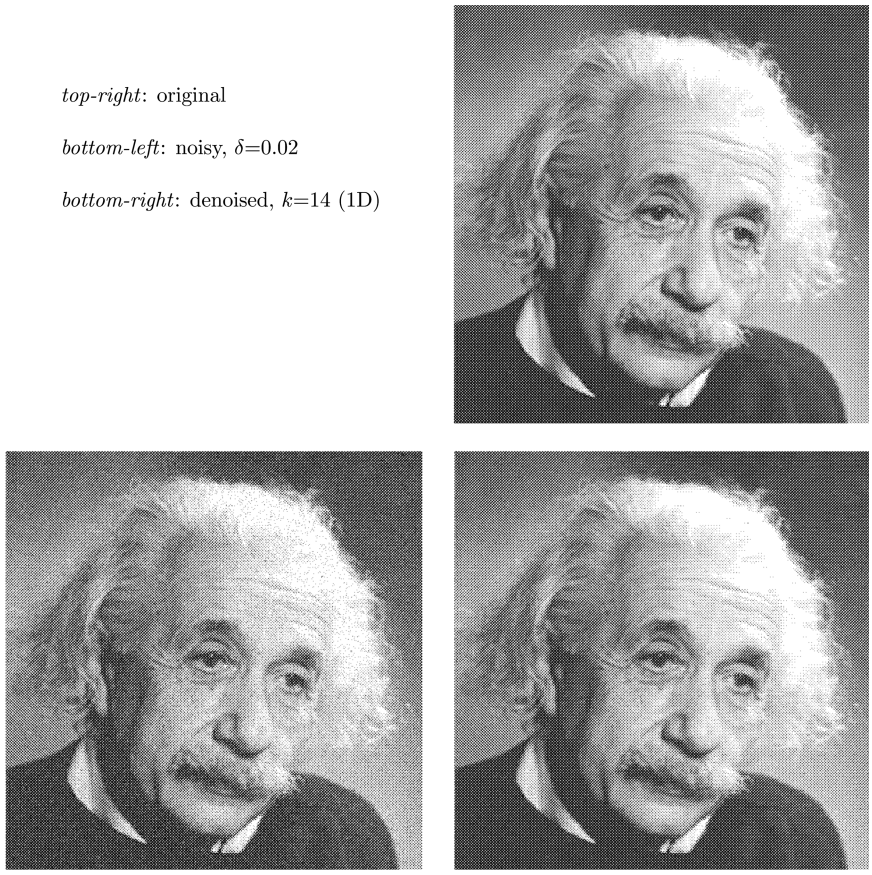

Fig. 3. Denoising of a binary halftone image.

loss), it lends itself to nonstationarity. For example, when the data may be assumed piecewise stationary (e.g., images and various types of audio signals), the counting of the appearances of the strings can include a "forgetting factor" to discount the contribution of strings according to their distance from the relevant location. To that end, judicious segmentation of the input data sequence depending on the expected dynamics of the data statistics can be helpful.

Related theoretical and practical directions that have been pursued since the submission of this work include causal denoising (filtering) [48], the case of channel uncertainty [27], [81], the case of a general (not necessarily discrete) channel output alphabet [17], the case of channel memory [82], loss estimation for efficient pruning of bi-directional context trees [80], and applications of the DUDE to joint source channel decoding of an unknown source [45].

Appendix I
Proof of Lemma 2

Throughout the proof, we will simplify our notation by omitting the first two arguments in the vectors $\mathbf{q}(z^n, x^n, u_{-k}^{-1} b u_1^k)$ and $\mathbf{q}'(z^n, x^n, u_{-k}^{-1} b u_1^k)$, as these arguments will always be z^n and x^n, respectively, and we will replace the third argument, in which u_{-k}^{-1} and u_1^k are fixed, by its central symbol, $b \in \mathcal{A}$. Similarly, we will omit all the arguments in the vector $\mathbf{m}(z^n, u_{-k}^{-1}, u_1^k)$. Since, for all $b \in \mathcal{A}$, we have by definition

$$\mathbf{m}[b] = \sum_{a' \in \mathcal{A}} \mathbf{q}(b)[a']$$

it follows that, for all $a \in \mathcal{A}$

$$\begin{aligned}
&[\pi_{u_0} \odot (\Pi^{-T}\mathbf{m})][a] \\
&= \Pi(a, u_0) \sum_{a', b \in \mathcal{A}} \Pi^{-T}(a, b)\mathbf{q}(b)[a'] \\
&= \Pi(a, u_0) \sum_{a', b \in \mathcal{A}} \Pi^{-1}(b, a)[\mathbf{q}(b)[a'] - \mathbf{q}'(b)[a'] + \mathbf{q}'(b)[a']] \\
&= \mathbf{q}'(u_0)[a] + \Pi(a, u_0) \sum_{a', b \in \mathcal{A}} \Pi^{-1}(b, a)[\mathbf{q}(b)[a'] - \mathbf{q}'(b)[a']] \qquad \text{(A1)}
\end{aligned}$$

where the last equality follows from the fact that, by the definition (52), the only dependence of $\mathbf{q}'(b)[a']$ on b is due to the factor $\Pi(a', b)$, and from the identity $\sum_b \Pi(a', b)\Pi^{-1}(b, a) = \mathbf{1}_{a=a'}$. Thus, we have (A2) at the bottom of the page. Summing (A2) over u_0 yields

$$\begin{aligned}
&\sum_{u_0 \in \mathcal{A}} \left\| \mathbf{q}(u_0) - \pi_{u_0} \odot (\mathbf{\Pi}^{-T}\mathbf{m}) \right\|_1 \\
&\quad \le \sum_{u_0 \in \mathcal{A}} \|\mathbf{q}(u_0) - \mathbf{q}'(u_0)\|_1 \\
&\qquad + \sum_{b \in \mathcal{A}} \left[\|\mathbf{q}(b) - \mathbf{q}'(b)\|_1 \sum_{a \in \mathcal{A}} \left| \Pi^{-1}(b, a) \right| \right] \\
&\quad \le (1 + \left\| \mathbf{\Pi}^{-1} \right\|_\infty) \sum_{u_0 \in \mathcal{A}} \|\mathbf{q}(u_0) - \mathbf{q}'(u_0)\|_1 \qquad \text{(A3)}
\end{aligned}$$

where (A3) follows from the definition (Section II)

$$\left\| \mathbf{\Pi}^{-1} \right\|_\infty = \max_{b \in \mathcal{A}} \sum_{a \in \mathcal{A}} \left| \Pi^{-1}(b, a) \right|. \qquad \square$$

Appendix II
Proof of Claim 2

A. Proof of Lemma 4

Recall first that the Bayes envelope $U(\cdot)$ is a concave function. Specifically, for two M-vectors $\mathbf{u}$ and $\mathbf{v}$ and $\alpha \in [0, 1]$

$$\begin{aligned}
U(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) &= \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T [\alpha\mathbf{u} + (1-\alpha)\mathbf{v}] \\
&\ge \alpha \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{u} + (1-\alpha) \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{v} \\
&= \alpha U(\mathbf{u}) + (1-\alpha) U(\mathbf{v}). \qquad \text{(A4)}
\end{aligned}$$

Next, to show that $E\left[U(\mathbf{P}_{X_0|Z_{-k}^l})\right]$ decreases with l, observe equation (A5) at the bottom of the page, where the inequality follows by concavity. The fact that $E\left[U(\mathbf{P}_{X_0|Z_{-k}^l})\right]$ decreases with k is established similarly, concluding the proof of the first item. For the second item note that, by martingale convergence

$$\begin{aligned}
\left\| \mathbf{q}(u_0) - \pi_{u_0} \odot (\mathbf{\Pi}^{-T}\mathbf{m}) \right\|_1 &= \sum_{a \in \mathcal{A}} \left| \mathbf{q}(u_0)[a] - \mathbf{q}'(u_0)[a] - \Pi(a, u_0) \sum_{a', b \in \mathcal{A}} \Pi^{-1}(b, a) \left[\mathbf{q}(b)[a'] - \mathbf{q}'(b)[a'] \right] \right| \\
&\le \|\mathbf{q}(u_0) - \mathbf{q}'(u_0)\|_1 + \sum_{a, b \in \mathcal{A}} \Pi(a, u_0) \left| \Pi^{-1}(b, a) \right| \|\mathbf{q}(b) - \mathbf{q}'(b)\|_1 . \qquad \text{(A2)}
\end{aligned}$$

$$\begin{aligned}
E\left[U(\mathbf{P}_{X_0|Z_{-k}^{l+1}})\right] &= \sum_{z_{-k}^{l+1} \in \mathcal{A}^{k+l+2}} U(\mathbf{P}_{X_0|Z_{-k}^{l+1} = z_{-k}^{l+1}}) P(Z_{-k}^{l+1} = z_{-k}^{l+1}) \\
&= \sum_{z_{-k}^{l} \in \mathcal{A}^{k+l+1}} \left[\sum_{z_{l+1} \in \mathcal{A}} U(\mathbf{P}_{X_0|Z_{-k}^{l} = z_{-k}^{l}, Z_{l+1} = z_{l+1}}) P(Z_{l+1} = z_{l+1} | Z_{-k}^{l} = z_{-k}^{l}) \right] P(Z_{-k}^{l} = z_{-k}^{l}) \\
&\le \sum_{z_{-k}^{l} \in \mathcal{A}^{k+l+1}} U\left(\sum_{z_{l+1} \in \mathcal{A}} \mathbf{P}_{X_0|Z_{-k}^{l} = z_{-k}^{l}, Z_{l+1} = z_{l+1}} P(Z_{l+1} = z_{l+1} | Z_{-k}^{l} = z_{-k}^{l}) \right) P(Z_{-k}^{l} = z_{-k}^{l}) \\
&= \sum_{z_{-k}^{l} \in \mathcal{A}^{k+l+1}} U(\mathbf{P}_{X_0|Z_{-k}^{l} = z_{-k}^{l}}) P(Z_{-k}^{l} = z_{-k}^{l}) = E\left[U(P_{X_0|Z_{-k}^{l}})\right] \qquad \text{(A.5)}
\end{aligned}$$

(cf., in particular, [7, Theorem 5.21]), $\mathbf{P}_{X_0|Z_{-k_n}^{l_n}} \to \mathbf{P}_{X_0|Z_{-\infty}^{\infty}}$ a.s., implying, by the (easily verified) continuity of $U(\cdot)$ that $U(\mathbf{P}_{X_0|Z_{-k_n}^{l_n}}) \to U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})$ a.s. Consequently, since $U(\mathbf{P}) \le \Lambda_{\max}$ for all $\mathbf{P} \in \mathcal{M}$

$$\begin{aligned} E\left[U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})\right] &= E\left[\lim_{n\to\infty} U(\mathbf{P}_{X_0|Z_{-k_n}^{l_n}})\right] \\ &= \lim_{n\to\infty} E\left[U(\mathbf{P}_{X_0|Z_{-k_n}^{l_n}})\right] \end{aligned}$$

the second equality following by bounded convergence. □

Proof of Claim 2

We have

$$\begin{aligned} &\mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) \\ &= \min_{\hat{X}^n \in \mathcal{D}_n} E\left[L_{\hat{X}^n}(X^n, Z^n)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \min_{\hat{X}:\mathcal{A}^n \to \mathcal{A}} E\left[\Lambda(X_i, \hat{X}(Z^n)[i])\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \sum_{z^n \in \mathcal{A}^n} P(Z^n = z^n) \min_{\hat{x}\in\mathcal{A}} E\left[\Lambda(X_i, \hat{x}) | Z^n = z^n\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \sum_{z^n \in \mathcal{A}^n} P(Z^n = z^n) U(\mathbf{P}_{X_i|Z^n=z^n}) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[U(\mathbf{P}_{X_i|Z^n})\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[U(\mathbf{P}_{X_0|Z_{1-i}^{n-i}})\right] \quad \text{(A6)} \end{aligned}$$

where the last equality follows by stationarity. Since, by Lemma 4

$$E\left[U(\mathbf{P}_{X_0|Z_{1-i}^{n-i}})\right] \ge E\left[U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})\right]$$

it follows from (A6) that

$$\mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) \ge E\left[U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})\right]$$

for all n and, therefore,

$$\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \ge E\left[U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})\right].$$

On the other hand, for any k, $0 \le k \le n$, Lemma 4 and (A6) yield the upper bound

$$\begin{aligned} \mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) &\le \frac{1}{n}\left[2kU(\mathbf{P}_{X_0}) + \sum_{i=k+1}^{n-k} E[U(\mathbf{P}_{X_0|Z_{1-i}^{n-i}})]\right] \\ &\le \frac{1}{n}\left[2kU(\mathbf{P}_{X_0}) + \sum_{i=k+1}^{n-k} E[U(\mathbf{P}_{X_0|Z_{-k}^{k}})]\right] \\ &= \frac{1}{n}\Big[2kU(\mathbf{P}_{X_0}) + (n-2k)E[U(\mathbf{P}_{X_0|Z_{-k}^{k}})]\Big]. \quad \text{(A7)} \end{aligned}$$

Considering the limit as $n \to \infty$ of both ends of the (A7) yields

$$\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \le E\left[U(\mathbf{P}_{X_0|Z_{-k}^{k}})\right].$$

Letting now $k \to \infty$ and invoking Lemma 4 implies

$$\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \le E\left[U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})\right]. \qquad \square$$

REFERENCES

[1] D. Angluin and M. Csürös, "Learning Markov chains with variable memory length from noisy output," in *Proc. 10th Annu. Conf. Computational Learning Theory (COLT 1997)*, Nashville, TN, Jul. 1997.

[2] R. J. Ballard, "Extended rules for the sequence compound decision problem with $m \times n$ component," Ph.D. dissertation, Michigan State Univ., East Lansing, 1974.

[3] R. J. Ballard, D. C. Gilliland, and J. Hannan, "$O(N^{-1/2})$ convergence to k-extended Bayes risk in the sequence compound decision problem with $m \times n$ component," Statistics and Probability, Michigan State Univ., East Lansing, RM-333, 1974.

[4] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[5] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2693–2723, Oct. 1998.

[6] H. W. Bode and C. E. Shannon, "A simplified derivation of linear least-squares smoothing and prediction theory," *Proc. IRE*, vol. 38, pp. 417–425, 1950.

[7] L. Breiman, *Probability*. Philadelphia, PA: SIAM, 1992.

[8] A. Bruce, D. L. Donoho, and H. Y. Gao, "Wavelet analysis," *IEEE Spectrum*, vol. 33, no. 10, pp. 26–35, Oct. 1996.

[9] G. Caire, S. Shamai (Shitz), and S. Verdú, "Almost-noiseless joint source-channel coding-decoding of sources with memory," in *Proc. 5th Int. ITG Conf. Source and Channel Coding (SCC)*, Jan 14–16, 2004, pp. 295–303.

[10] B. Carpentieri, M. J. Weinberger, and G. Seroussi, "Lossless compression of continuous-tone images," *Proc. IEEE*, vol. 88, no. 11, pp. 1797–1809, Nov. 2000.

[11] G. Chang, B. Yu, and M. Vetterli, "Bridging compression to wavelet thresholding as a denoising method," in *Proc. Conf. Information Sciences and Systems*, vol. 2, Mar. 1997, pp. 568–573.

[12] ——, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000.

[13] R. W. Chang and J. C. Hancock, "On receiver structures for channels having memory," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 4, pp. 463–468, Oct. 1966.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[15] F. J. Damerau and E. Mays, "An examination of undetected typing errors," *Inf. Process. and Management: An Int. J.*, vol. 25, no. 6, pp. 659–664, 1989.

[16] A. Dembo and T. Weissman, "The minimax distortion redundancy in noisy source coding," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 3020–3030, Nov. 2003.

[17] ——, "Universal denoising for the finite-input-general-output channel," *IEEE Trans. Inf. Theory*. Available [Online] at http://www.stanford.edu/~tsachy/interest.htm, to be published.

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

[19] J. L. Devore, "A note on the observation of a Markov source through a noisy channel," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 6, pp. 762–764, Nov. 1974.

[20] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[21] ——, "The Kolmogorov sampler," preprint, Jan. 2002. Available [Online] at http://playfair.stanford.edu/~donoho/Reports/.

[22] D. L. Donoho, I. M. Johnstone, G. Keryacharian, and D. Picard, "Wavelet shrinkage: Asymptopia?," *J. Roy. Statist. Soc.*, vol. 57, no. 2, pp. 301–369, 1995.

[23] R. Durret, *Probability: Theory and Examples*. Belmont, CA: Duxbury, 1991.

[24] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1518–1569, Jun. 2002.

[25] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inf. Theory*, vol. 38, no. 4, pp. 1258–1270, Jul. 1992.

[26] W. Feller, *An Introduction to Probability Theory and its Applications*. New York: Wiley, 1968.

[27] G. Gemelos, S. Sigurjònsson, and T. Weissman, "Universal discrete denoising under channel uncertainty," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 199.

[28] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[29] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Reading, MA: Addison Wesley, 1992.

[30] J. Hannan, "Approximation to Bayes risk in repeated play," in *Contributions to the Theory of Games*. Princeton, NJ: Princeton Univ. Press, 1957, vol. III, pp. 97–139.

[31] J. Hannan and H. Robbins, "Asymptotic solutions of the compound decision problem for two completely specified distributions," *Ann. Math. Statist.*, vol. 26, pp. 37–51, 1955.

[32] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Amer. Statist. Assoc.*, vol. 58, pp. 13–30, 1963.

[33] M. V. Johns, Jr., "Two-action compound decision problems," in *Proc. 5th Berkeley Symp. Mathematical and Statistical Probability*, Berkeley, CA, 1967, pp. 463–478.

[34] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME—J. Basic Eng.*, ser. D, vol. 82, pp. 35–45, 1960.

[35] R. Khasminskii and O. Zeitouni, "Asymptotic filtering for finite state Markov chains," *Stochastic Processes and Their Applications*, vol. 63, pp. 1–10, 1996.

[36] S. P. Lalley, "Beneath the noise, chaos," *Ann. Statist.*, vol. 27, pp. 461–479, 1999.

[37] ——, "Removing the noise from chaos plus noise," in *Nonlinear Dynamics and Statistics*, A. I. Mees, Ed. Basel, Switzerland: Birkhauser, 1999, pp. 233–244.

[38] S. P. Lalley and A. B. Nobel, "Denoising deterministic time series," preprint, 2002. Available [Online] at http://www.stat.unc.edu/faculty/nobel/links/denoise.pdf.

[39] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*. Orlando, FL: Academic, 1985.

[40] E. Mays, F. J. Damerau, and R. L. Mercer, "Context based spelling correction," in *Proc. IBM Natural Language ITL*, Paris, France, 1990, pp. 517–522.

[41] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.

[42] B. Natarajan, "Filtering random noise via data compression," in *Proc. Data Compression Conf. (DCC '93)*, Snowbird, UT, Mar. 1993, pp. 60–69.

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

[44] B. Natarajan, K. Konstantinides, and C. Herley, "Occam filters for stochastic sources with application to digital images," *IEEE Trans. Signal Process.*, vol. 46, no. 11, pp. 1434–1438, Nov. 1998.

[45] E. Ordentlich, G. Seroussi, S. Verdú, K. Viswanathan, M. J. Weinberger, and T. Weissman, "Channel decoding of systematically encoded unknown redundant sources," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 165.

[46] E. Ordentlich, G. Seroussi, S. Verdú, M. J. Weinberger, and T. Weissman, "A universal discrete image denoiser and its application to binary images," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Barcelona, Catalonia, Spain, Sep. 2003, pp. 117–120.

[47] E. Ordentlich and T. Weissman, "On the optimality of symbol by symbol filtering and denoising," *IEEE Trans. Inf. Theory*. Available: [Online] at http://www.stanford.edu/~tsachy/interest.htm, to be published.

[48] E. Ordentlich, T. Weissman, M. J. Weinberger, A. Somekh-Baruch, and N. Merhav, "Discrete universal filtering through incremental parsing," in *Proc. 2004 Data Compression Conf. (DCC'04)*, Snowbird, UT, Mar. 2004, pp. 352–361.

[49] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific, 1989.

[50] ——, "MDL denoising," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2537–2543, Nov. 2000.

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

[52] H. Robbins and E. Samuel, "Testing statistical hypotheses; the compound approach," in *Recent Developments in Information and Decision Processes*. New York: Macmillan, 1962, pp. 63–70.

[53] J. Van Ryzin, "The compound decision problem with $m \times n$ finite loss matrix," *Ann. Math. Statist.*, vol. 37, pp. 412–424, 1966.

[54] ——, "The sequential compound decision problem with $m \times n$ finite loss matrix," *Ann. Math. Statist.*, vol. 37, pp. 954–975, 1966.

[55] E. Samuel, "Asymptotic solutions of the sequential compound decision problem," *Ann. Math. Statist.*, vol. 34, no. 3, pp. 1079–1095, 1963.

[56] ——, "An empirical Bayes approach to the testing of certain parametric hypotheses," *Ann. Math. Statist.*, vol. 34, no. 4, pp. 1370–1385, 1963.

[57] ——, "Convergence of the losses of certain decision rules for the sequential compound decision problem," *Ann. Math. Statist.*, vol. 35, no. 4, pp. 1606–1621, 1964.

[58] ——, "On simple rules for the compound decision problem," *J. Roy. Statist. Soc.*, ser. B, vol. 27, pp. 238–244, 1965.

[59] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[60] J. W. Shavlik, "Finding genes by case-based reasoning in the presence of noisy case boundaries," in *Proc. DARPA Cased-Based Reasoning Workshop*, vol. 1, 1991, pp. 327–338.

[61] L. Shue, B. D. O. Anderson, and F. De Bruyne, "Asymptotic smoothing error for hidden Markov models," *IEEE Trans. Signal Process.*, vol. 48, no. 12, pp. 3289–3302, Dec. 2000.

[62] P. Soille, *Morphological Image Analysis: Principles and Applications*. Berlin, Germany: Springer-Verlag, 1999.

[63] I. Tabus, J. Rissanen, and J. Astola, "Normalized maximum likelihood models for boolean regression with application to prediction and classification in genomics," in *Computational and Statistical Approaches to Genomics*, W. Zhang and I. Shmulevich, Eds. Boston, MA: Kluwer Acadmic, 2002.

[64] ——, "Classification and feature gene selection using the normalized maximum likelihood model for discrete regression," *Signal Process., Special Issue: Genomic Signal Processing*, vol. 83, no. 4, pp. 713–727, Apr. 2003.

[65] S. B. Vardeman, "Admissible solutions of finite state sequence compound decision problems," *Ann. Statist.*, vol. 6, no. 3, pp. 673–679, 1978.

[66] ——, "Admissible solutions of k-extended finite state set and sequence compound decision problems," *J. Multiv. Anal.*, vol. 10, pp. 426–441, 1980.

[67] ——, "Approximation to minimum k-extended Bayes risk in sequences of finite state decision problems and games," *Bull. Inst. Math. Academia Sinica*, vol. 10, no. 1, pp. 35–52, Mar. 1982.

[68] M. J. Weinberger, J. J. Rissanen, and M. Feder, "A universal finite-memory source," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 643–652, May 1995.

[69] T. Weissman, "Universally attainable error-exponents for rate-distortion coding of noisy sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1229–1246, Jun. 2004.

[70] T. Weissman and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2151–2173, Sep. 2001.

[71] ——, "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.

[72] ——, "Universal prediction of random binary sequences in a noisy environment," *Ann. Appl. Probab.*, vol. 14, no. 1, pp. 54–89, Feb. 2004.

[73] T. Weissman, N. Merhav, and A. Baruch, "Twofold universal prediction schemes for achieving the finite-state predictability of a noisy individual binary sequence," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1849–1866, Jul. 2001.

[74] T. Weissman and E. Ordentlich, "The empirical distribution of rate-constrained codes," *IEEE Trans. Inf. Theory*. Available [Online] at http://www.stanford.edu/~tsachy/interest.htm, to be published.

[75] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. New York: Wiley, 1949.

[76] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.

[77] C. H. Zhang, "Compound decision theory and empirical Bayes methods," *Ann. Statist.*, vol. 31, no. 2, pp. 379–390, 2003.

[78] J. Ziv, "Distortion-rate theory for individual sequences," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 2, pp. 137–143, Mar. 1980.

[79] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 5, pp. 530–536, Sep. 1978.

[80] E. Ordentlich, M.J. Weinberger, and T. Weissman, "Efficient pruning of bi-directional context trees with applications to universal denoising and compression," in *Proc. 2004 Information Theory Workshop*, San Antonio, TX, Oct. 2004.

[81] G. Gemelos, S. Sigurjønsson, and T. Weissman, "Universal minimax binary image denoising under channel uncertainty," in *Proc. IEEE Int. Conf. Image Processing*, Singapore, Sep. 2004, pp. 997–1000.

[82] R. Zhang and T. Weissman, "On discrete denoising for the burst noise channel," in *Proc. 42nd Annu. Allerton Conf. Communication, Control, and Computing*, vol. 1, Monticello, IL, Sep. 2004.