Universal Discrete Denoising: Known Channel

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi
Sergio Verdú, Marcelo Weinberger
Information Theory Research
HP Laboratories Palo Alto
HPL-2003-29
February 10th , 2003*

context models, denoising, discrete filtering, discrete memoryless channels, individual sequences, noisy channels, universal algorithms

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 and ergodic. 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 sub-linear working storage size relative to the input data length.

* Internal Accession Date Only
Approved for External Publication

Universal Discrete Denoising: Known Channel *

Tsachy Weissman Erik Ordentlich Gadiel Seroussi Sergio Verdú
Marcelo Weinberger

February 10, 2003

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 and ergodic. 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 sub-linear working storage size relative to the input data length.

Key words and phrases: Context models, Denoising, Discrete filtering, Discrete Memoryless Channels, 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

1 Introduction

Consider the problem of recovering 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 recovery is assumed to start once the *entire* signal $\{Z_t\}_{t\in T}$ is available. 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.

*Part of this work was performed while S. Verdú was a Hewlett-Packard/Mathematical Sciences Research Institute (MSRI) visiting research professor; he is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: verdu@princeton.edu). The other authors are with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA (e-mail: tsachyw@hpl.hp.com; eord@hpl.hp.com; seroussi@hpl.hp.com; marcelo@hpl.hp.com). T. Weissman is also with the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.edu).

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 [57, 3] and Kalman [27], to Donoho and Johnstone's nonlinear denoisers [14, 15], 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 [5].

Considerably 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 [30, 10] to Hidden Markov Model (HMM) state estimation (cf. [18] for the many applications); from DNA sequence analysis and processing [45, 49, 48] to enhancement of facsimile and other binary images; from blind equalization problems to joint source-channel decoding when a discrete source is sent unencoded through a noisy channel [8, 21]. Here, 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 (measured by a single-letter loss function), as evaluated under the posterior distribution of the noiseless signal given its noisy observation. When the distribution of the noiseless signal and the channel are known, 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, on which the optimal denoiser is based. Thus, though it may not always be practical to explicitly obtain this posterior distribution, in principle it is available.

Certain instances of the discrete denoising problem have been extensively studied, particularly in the context of state estimation for hidden Markov processes (cf. [18] 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 above optimum Bayesian scheme can be implemented with reasonable complexity via forward-backward dynamic programming [8, 1]. It should be mentioned, however, that even for the simplest among cases where the underlying signal has memory, namely the case of a binary 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, [28, 46] and conditions for the optimality of "singlet decoding" (cf. [13]), are known.

The literature on the *universal* discrete denoising setting is even sparser. In this setting, there is uncertainty regarding the distribution of the underlying noiseless signal and/or regarding the channel, so that the posterior distribution on which the optimal denoiser is based is not available. One recent line of attack to this problem is the compression-based approach, encompassing Natarajan's "Occam filters"

[32, 33, 34], Yu et al.'s "compresstimation" [7, 26], Donoho's "Kolmogorov sampler" [16], and Tabus-Rissanen-Astola's "normalized maximum likelihood" models [48, 49, 38]. 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. Unfortunately, the compression-based approach to denoising, as formalized and analyzed concretely in [16], appears to suffer from the following two drawbacks. The first one is algorithmic: Its faithful implementation essentially boils down to the implementation of a universal lossy compression scheme. Performing the latter optimally (in the rate distortion sense) and with manageable complexity[1] is one of the notoriously hard open problems in contemporary information theory (cf. [2, Section VI]). The second drawback, perhaps more fundamental than the first one, is the fact, established in [16], that optimal universal lossy compression of the noisy signal fails to achieve the optimal distribution dependent denoising performance with stationary ergodic input signals, for the two concrete settings of the BSC and the additive Gaussian white noise channel. The fact that compression-based schemes for universal denoising fall short of the optimal distribution dependent performance was consolidated from a somewhat different perspective by Dembo and Weissman [11, 52], who consider universal rate distortion coding of noisy sources and characterize tradeoffs between the attainable denoising performance and the rate constraint.

In principle, a denoising scheme that fails to attain the distribution-dependent optimum performance for all possible stationary ergodic sources (such as the compression-based scheme of [16]), is not necessarily suboptimal in the universal setting. Since in this setting the posterior distribution cannot be obtained, it may seem plausible that universally attaining optimum distribution-dependent performance is an unreachable goal. This supposition, in fact, seems to be implicit in the existent literature on universal denoising. In the recent [16], for example, the optimality of the proposed universal "Kolmogorov sampler" was not ruled out. Therefore, the following basic questions arise:

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, is the universal denoiser practical? What is its complexity?

To study these questions, we restrict attention to the case of finite alphabets and a known DMC

[1]Some of the mentioned compression-based schemes are implemented with practical, sub-optimal, lossy compression schemes and have no performance guarantees, as the performance analysis of [16] applies to optimal compression (in the rate-distortion sense) of the noisy signal.

whose transition probability matrix has full rank.[2] In this case, the distribution of the channel output uniquely determines the distribution of the input.

As discussed above, no discrete denoiser available in the literature universally attains the distribution-dependent optimum performance. The main contribution of this work is a discrete denoising algorithm performing favorably from both the theoretical and the algorithmic viewpoints. Specifically, we shall propose and analyze an algorithm that is:

1. 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 our denoising algorithm is guaranteed to attain the performance of the best finite-order sliding-window denoiser in an almost sure sense, for every underlying individual sequence. Here, competing with finite-order sliding-window denoisers is akin to the setting introduced in the universal lossless coding literature by Ziv and Lempel [60].

   (b) *The stochastic setting.* We show that our denoising algorithm asymptotically attains the performance of the optimal distribution-dependent scheme, for any stationary ergodic source that may be generating the underlying signal. This property follows easily from the result in the semi-stochastic setting.

2. Practical. Implementation of the denoiser requires a linear number of register-level operations, and working storage complexity which is sub-linear 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.

For concreteness and simplicity of the exposition, we assume one-dimensionally indexed data, though all our results can be readily extended to the multi-dimensional case. In fact, Section 7 presents experimental results for two-dimensional image denoising, and the multi-dimensional formalism is discussed in more detail in [35]. 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 3-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

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

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 context statistics 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.

We remark that in the statistics literature, the semi-stochastic setting dates nearly half a century back to the so-called *compound decision* problem [25, 39, 40, 42, 43, 50], which can be viewed as the particular case of our denoising setting in which the denoiser is constrained to be context-independent, corresponding to $k = 0$ in the above discussion.

The remainder of the paper is organized as follows. Section 2 presents our notation and conventions. In Section 3, we introduce the denoising algorithm, analyze its complexity, motivate its structure, and detail its explicit form for a few special cases. Section 4 is devoted to assessing the performance of the proposed algorithm in the semi-stochastic setting. The fully stochastic setting, where the underlying noiseless sequence is assumed generated by a stationary ergodic source, is considered in Section 5. In Section 6 we discuss some theoretical and practical aspects of the choice of context model size for the denoiser. In Section 7, we report the results of a few experiments where 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 enhancements. Finally, Section 8 discusses some related directions for future work.

2 Notation and Conventions

We present some definitions and notation that are used throughout the paper. Additional, "local" notation is introduced where needed.

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 a 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 an 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 $\mathbf{\Lambda}_{\max} = \max_{i,j\in\mathcal{A}} \Lambda(i,j)$.

We let $\boldsymbol{\pi}_i$ denote the i-th column of $\mathbf{\Pi}$, and $\boldsymbol{\lambda}_j$ denote the j-th column of $\mathbf{\Lambda}$. Hence, we have,

$$\mathbf{\Pi} = [\boldsymbol{\pi}_1 \mid \cdots \mid \boldsymbol{\pi}_M], \qquad \mathbf{\Lambda} = [\boldsymbol{\lambda}_1 \mid \cdots \mid \boldsymbol{\lambda}_M].$$

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 i-th component of a vector $\mathbf{u}$ will be denoted by $\mathbf{u}_i$, or $\mathbf{u}[i]$ (the latter notation used for indexing vector-valued expressions). For M-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$, $\mathbf{u} \odot \mathbf{v}$ will denote the vector obtained from componentwise multiplication, i.e., $(\mathbf{u} \odot \mathbf{v})[i] = \mathbf{u}[i]\mathbf{v}[i]$. In terms of order of operations, $\odot$ will have the usual multiplicative precedence over addition and subtraction. The L_p norm of any vector $\mathbf{v}$ will be denoted by $\|\mathbf{v}\|_p$. Similarly, following standard conventions (cf., e.g., [22]), $\|\mathbf{A}\|_p$ 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 $2k < n$, $\mathbf{a} \in \mathcal{A}^n$, $\mathbf{b} \in \mathcal{A}^k$, $\mathbf{c} \in \mathcal{A}^k$ let $\mathbf{m}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ denote the M-dimensional column vector whose β-th component, $\beta \in \mathcal{A}$, is equal to

$$\begin{aligned} \mathbf{m}(\mathbf{a}, \mathbf{b}, \mathbf{c})[\beta] &= \left| \left\{ i : k+1 \leq i \leq n-k, a_{i-k}^{i-1} = \mathbf{b}, a_i = \beta, a_{i+1}^{i+k} = \mathbf{c} \right\} \right| \\ &= \sum_{i \in \{k+1, \ldots, n-k\} : a_i = \beta} \mathbf{1}_{\{a_{i-k}^{i-1} = \mathbf{b}\}} \mathbf{1}_{\{a_{i+1}^{i+k} = \mathbf{c}\}} \end{aligned} \tag{1}$$

where throughout $\mathbf{1}_{\{\cdot\}}$ will denote the indicator function. The component $\mathbf{m}(\mathbf{a}, \mathbf{b}, \mathbf{c})[\beta]$ is the number of appearances of the string $\mathbf{b}\beta\mathbf{c}$ along the sequence $\mathbf{a}$. For such an appearance, we say that β occurs in *left context* $\mathbf{b}$, *right context* $\mathbf{c}$, and *double-sided context* $(\mathbf{b}, \mathbf{c})$. The normalized (unit sum) version of the vector $\mathbf{m}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ gives the empirical conditional distribution of a single letter given that the double-sided context is $(\mathbf{b}, \mathbf{c})$. For $\mathbf{a}, \mathbf{b} \in \mathcal{A}^n$, $\mathbf{c} \in \mathbf{A}^{2k+1}$ let $\mathbf{q}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ denote the M-dimensional column vector whose α-th component, $\alpha \in \mathcal{A}$, is

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

namely, the number of appearances of the string $\mathbf{c}$ along the sequence $\mathbf{a}$ when the letter in the sequence

$\mathbf{b}$ corresponding to the center of $\mathbf{c}$ is α. Note that, for every $\mathbf{d} \in \mathcal{A}^n$,

$$\sum_{\alpha \in \mathcal{A}} \mathbf{q}(\mathbf{a}, \mathbf{d}, \mathbf{b}\beta\mathbf{c})[\alpha] = \mathbf{m}(\mathbf{a}, \mathbf{b}, \mathbf{c})[\beta].$$

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

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

denote the *Bayes envelope* (cf., e.g., [24, 41, 31]) associated with the distribution $\mathbf{P}$ and the loss measure Λ. Following [24], 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 (3), namely the *Bayes response* to $\mathbf{v}$, by $\hat{x}(\mathbf{v})$, i.e.,

$$\hat{x}(\mathbf{v}) = \arg\min_{\hat{x} \in \mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \mathbf{v}, \tag{4}$$

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 Λ, 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.,

$$L_{\hat{X}^n}(x^n, z^n) = \frac{1}{n} \sum_{i=1}^{n} \Lambda(x_i, \hat{X}^n(z^n)[i]). \tag{5}$$

3 The Discrete Universal DEnoiser (DUDE)

In this section we present our Discrete Universal DEnoiser (DUDE). We describe the algorithm and assess its complexity in subsection 3-A before we proceed to motivate the form of this algorithm in subsection 3-B. For the sake of clarity, we concentrate on the case of a square channel matrix $\boldsymbol{\Pi}$ (equal channel input and output alphabets), which is invertible. The more general case, in which $\boldsymbol{\Pi}$ is non-square, is treated in subsection 3-C, assuming the matrix rows are linearly independent. Then, in subsection 3-D, we particularize the algorithm to several channels of interest.

3-A The Algorithm: Description and Implementation

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 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 $\mathbf{a} \in \mathcal{A}^n$, a context length k, two contexts $\mathbf{b} \in \mathcal{A}^k$ and $\mathbf{c} \in \mathcal{A}^k$, and a symbol $\alpha \in \mathcal{A}$, we define the function

$$g_{\mathbf{a}}^k(\mathbf{b}, \alpha, \mathbf{c}) = \arg\min_{\hat{x} \in \mathcal{A}} \mathbf{m}^T(\mathbf{a}, \mathbf{b}, \mathbf{c}) \boldsymbol{\Pi}^{-1} \left[\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_\alpha \right]. \tag{6}$$

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

$$\hat{X}^{n,k}(z^n)[i] = g^k_{z^n}(z_{i-k}^{i-1}, z_i, z_{i+1}^{i+k}), \quad k+1 \le i \le n-k. \tag{7}$$

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.[3] Finally, for each n, our asymptotic analysis of the DUDE algorithm will focus on the n-block denoiser $\hat{X}^n_{\mathsf{univ}}$ defined as

$$\hat{X}^n_{\mathsf{univ}} = \hat{X}^{n,k_n}, \tag{8}$$

where, for asymptotic optimality, k_n is any unboundedly increasing function of n such that[4]

$$k_n M^{2k_n} = o(n/\log n). \tag{9}$$

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 akin 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 [19]. 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 trade-offs 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 $g^k_{z^n}(\cdot)$, as determined by the accumulated empirical counts via (6), 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 $[\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_\alpha]$ 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 two-sided context $(\mathbf{b}, \mathbf{c})$ 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 $(\mathbf{b}, \mathbf{c})$ can assume M^2 possible values of the form $(\mathbf{b}' z_i, \mathbf{c}' z_{i+k+1})$, where $\mathbf{b}'$ and $\mathbf{c}'$ are the suffixes of length $k-1$ of $\mathbf{b}$ and $\mathbf{c}$, respectively. Associated with each state $(\mathbf{b}, \mathbf{c})$ is an M-vector of counts, which, at time i, contains $\mathbf{m}(z^{i+k}, \mathbf{b}, \mathbf{c})$. Each

[3] 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.

[4] As will be discussed in Section 4, the condition (9) can be slightly relaxed depending on the type of universality that is required.

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. The number of operations required to dynamically grow the data structure is $O(M^{2k+2})$.

The described finite state automaton lends itself to a representation with the additional properties of a tree data structure, akin to the *tree model* representations used in source coding (cf., e.g.,[51]). This representation is convenient when the function $g^k_{z^n}(\cdot)$ 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.

- **Pre-computations for the second pass.** The unnormalized input probability vectors $\mathbf{m}^T(z^n, \mathbf{b}, \mathbf{c})\mathbf{\Pi}^{-1}$ are computed for each two-sided context $(\mathbf{b}, \mathbf{c})$ actually encountered in the sequence. Since there are M^{2k} two-sided contexts in the worst case, and each computation takes $O(M^2)$ arithmetic operations, the computational complexity is $O(M^{2k+2})$ and the space required to store the computations is $O(M^{2k+1})$. The algorithm then proceeds to pre-compute the values of $g^k_{z^n}(\mathbf{b}, \alpha, \mathbf{c})$ according to (6), for each state $(\mathbf{b}, \mathbf{c})$ and alphabet symbol α. There are at most M^{2k+1} such combinations, each requiring $O(M^2)$ operations, for a total of $O(M^{2k+3})$ operations requiring $O(M^{2k+1})$ storage.

- **Denoising.** The algorithm scans the sequence z^n a second time. At each sequence location, the context $(\mathbf{b}, \mathbf{c})$ and input symbol z_i are observed, and used to address the table of pre-computed 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 $O(n+M^{2k+3})$. For $k = k_n$ satisfying the requirement (9), the dominant term in the time complexity is $O(n)$. The working storage complexity is $O(M^{2k+1}) = o(n)$. This 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 multi-dimensional 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 multi-dimensional case still requires $K = K_n \to \infty$ with $K_n = o(n)$ 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. By partitioning the input data into ρ equally sized portions, and assigning each portion to a processor (for both passes), ρ processors can run the DUDE in time $O(\tau/\rho + M^{2k}\log\rho)$, where τ is the sequential, single-processor running time. The $O(M^{2k}\log\rho)$ term stems from the need to merge the statistics gathered by the ρ processors at the end of the first pass, and re-distributing the merged statistics for use in the second pass. The inter-processor communication and joint memory access requirements of the parallelized DUDE are fairly weak.

3-B Derivation of the Denoiser

To motivate the form of the DUDE, consider first the case in which we only have two jointly distributed $\mathcal{A}$-valued random variables X and Z, and that our goal is to estimate X based on Z, minimizing the expected loss as measured by the loss function Λ. Letting $\mathbf{P}_{X|z}$ denote the vector in $\mathcal{M}$ whose a-th component is $P(X=a|Z=z)$, optimum performance for this problem is readily seen to be given by

$$E\left[U(\mathbf{P}_{X|Z})\right], \tag{10}$$

where the Bayes envelope $U(\cdot)$ is defined in (3) and the expectation is with respect to Z. The minimum loss in (10) is attained by the estimator

$$\hat{X}(z) = \arg\min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \mathbf{P}_{X|z}, \tag{11}$$

namely the Bayes response to $\mathbf{P}_{X|z}$. Further suppose now that Z is the output of the channel $\boldsymbol{\Pi}$ whose input is X. Letting $\mathbf{P}_X, \mathbf{P}_Z \in \mathcal{M}$ denote the respective distributions of X and Z, we have $\mathbf{P}_Z = \boldsymbol{\Pi}^T\mathbf{P}_X$ so that, in terms of $\mathbf{P}_Z$, $\mathbf{P}_{X|z}$ is given by

$$\mathbf{P}_{X|z}(a) = \frac{P(X=a,Z=z)}{\mathbf{P}_Z(z)} = \frac{P(Z=z|X=a)\mathbf{P}_X(a)}{\mathbf{P}_Z(z)} = \frac{\Pi(a,z)[\boldsymbol{\Pi}^{-T}\mathbf{P}_Z](a)}{\mathbf{P}_Z(z)},$$

or, in vector notation,

$$\mathbf{P}_{X|z} = \frac{1}{\mathbf{P}_Z(z)}\ \boldsymbol{\pi}_z \odot [\boldsymbol{\Pi}^{-T}\mathbf{P}_Z]. \tag{12}$$

Consequently, the optimal estimator in (11) for this case assumes the form

$$\hat{X}(z) = \arg\min_{\hat{x}\in\mathcal{A}} \mathbf{P}_Z^T \boldsymbol{\Pi}^{-1} \left[\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_z\right]. \tag{13}$$

Although, in general, an optimum estimate of X requires knowledge of its prior distribution, the invertibility of the channel probability matrix has allowed us to express the optimal estimator solely in terms of the channel output distribution $\mathbf{P}_Z$ and the inverse channel matrix.

Let now X_1, X_2 be jointly distributed $\mathcal{A}$-valued random variables and let Z_1, Z_2 denote the respective outputs of the memoryless channel $\boldsymbol{\Pi}$ fed with X_1, X_2. Suppose that we would like to estimate X_1 based on observing Z_1, Z_2. Letting $\mathbf{P}_{X_1|z_1,z_2}$ denote the analogue of $\mathbf{P}_{X|z}$ for the distribution of X_1 conditioned on $Z_1 = z_1, Z_2 = z_2$, it is clear that the minimum-mean-loss estimator of X_1 based on Z_1, Z_2 is, similarly to (11), the Bayes response

$$\hat{X}_1(z_1, z_2) = \arg\min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \mathbf{P}_{X_1|z_1,z_2}. \tag{14}$$

Note that the memorylessness of the channel implies

$$\mathbf{P}_{Z_1|z_2} = \boldsymbol{\Pi}^T \mathbf{P}_{X_1|z_2}. \tag{15}$$

Consequently, for $x_1 \in \mathcal{A}, z_1 \in \mathcal{A}, z_2 \in \mathcal{A}$,

$$\begin{aligned}
\mathbf{P}_{X_1|z_1,z_2}(x_1) &= \frac{P(X_1 = x_1, Z_1 = z_1, Z_2 = z_2)}{P(Z_1 = z_1, Z_2 = z_2)} \\
&= \frac{P(Z_1 = z_1|X_1 = x_1, Z_2 = z_2)P(X_1 = x_1|Z_2 = z_2)P(Z_2 = z_2)}{P(Z_1 = z_1|Z_2 = z_2)P(Z_2 = z_2)} \\
&= \frac{P(Z_1 = z_1|X_1 = x_1)P(X_1 = x_1|Z_2 = z_2)}{P(Z_1 = z_1|Z_2 = z_2)} \qquad (16) \\
&= \frac{P(Z_1 = z_1|X_1 = x_1)\left[\boldsymbol{\Pi}^{-T}\mathbf{P}_{Z_1|z_2}\right](x_1)}{P(Z_1 = z_1|Z_2 = z_2)} \qquad (17) \\
&= \frac{\pi(x_1, z_1)\left[\boldsymbol{\Pi}^{-T}\mathbf{P}_{Z_1|z_2}\right](x_1)}{P(Z_1 = z_1|Z_2 = z_2)}, \qquad (18)
\end{aligned}$$

where (16) follows from the conditional independence of Z_1 and Z_2 given X_1, and (17) follows from (15). In vector notation, (18) assumes the form

$$\mathbf{P}_{X_1|z_1,z_2} = \frac{1}{\mathbf{P}_{Z_1|z_2}(z_1)}\ \boldsymbol{\pi}_{z_1} \odot \left[\boldsymbol{\Pi}^{-T}\mathbf{P}_{Z_1|z_2}\right]. \tag{19}$$

Substituting (19) into (14), the optimal estimator for X_1, based on Z_1, Z_2, becomes

$$\hat{X}_1(z_1, z_2) = \arg\min_{\hat{x}\in\mathcal{A}} \left[\mathbf{P}_{Z_1|z_2}\right]^T \boldsymbol{\Pi}^{-1} \left[\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_{z_1}\right]. \tag{20}$$

Two key features of the estimator in (20) to be noted are:

1. It is given solely in terms of the conditional distribution $\mathbf{P}_{Z_1|z_2}$ associated with the channel output random variables.

2. Though the inverse problem it solves is now two-dimensional (i.e., the size of the "data set" is 2), its application involves inversion of the channel matrix $\mathbf{\Pi}$ corresponding to just one input-output pair.

As can be expected, this strategy is not limited to a data set of size 2. Indeed, let now T be an arbitrary index set and assume $X(T) = \{X_t\}_{t\in T}$ to be any stochastic process (or random field) with components taking values in $\mathcal{A}$. Suppose that $Z(T)$ is the noisy version of $X(T)$ corrupted by the memoryless channel $\mathbf{\Pi}$. For $t \in T$, and $\mathbf{e} \in \mathcal{A}^{T\setminus t}$ (where $T \setminus t$ denotes the set $T \setminus \{t\}$), consider the M-dimensional column probability vectors on $\mathcal{A}$ with components:

$$\begin{aligned}
\mathbf{P}_{X_t|\mathbf{e}}(a) &= P(X_t = a|Z(T \setminus t) = \mathbf{e}),\\
\mathbf{P}_{Z_t|\mathbf{e}}(a) &= P(Z_t = a|Z(T \setminus t) = \mathbf{e}),\\
\mathbf{P}_{Z_t,\mathbf{e}}(a) &= P(Z_t = a, Z(T \setminus t) = \mathbf{e}).
\end{aligned} \tag{21}$$

The analogue of (15) for this case, again, due to the memorylessness of the channel, is the following: For any $t \in T$ and $z(T \setminus t) \in \mathcal{A}^{T\setminus t}$,

$$\mathbf{P}_{Z_t|z(T\setminus t)} = \mathbf{\Pi}^T \mathbf{P}_{X_t|z(T\setminus t)}. \tag{22}$$

By (22), a chain of equalities completely analogous to that leading to (18) yields, for $x_t \in \mathcal{A}$ and $z(T) \in \mathcal{A}^T$,

$$P(X_t = x_t|Z(T) = z(T)) = \frac{\pi(x_t, z_t)\left[\mathbf{\Pi}^{-T}\mathbf{P}_{Z_t|z(T\setminus t)}\right](x_t)}{P(Z_t = z_t|Z(T \setminus t) = z(T \setminus t))}, \tag{23}$$

which, in vector notation, assumes the form

$$\mathbf{P}_{X_t|z(T)} = \frac{1}{\mathbf{P}_{Z_t|z(T\setminus t)}(z_t)}\ \boldsymbol{\pi}_{z_t} \odot \left[\mathbf{\Pi}^{-T}\mathbf{P}_{Z_t|z(T\setminus t)}\right]. \tag{24}$$

Consequently, proceeding as in (20), the optimal estimator $\hat{X}^{\mathsf{opt}}(\cdot)[t]$ for the value of X_t based on observing $Z(T)$, in the sense of minimizing the expected loss, is

$$\begin{aligned}
\hat{X}^{\mathsf{opt}}(z(T))[t] &= \arg\min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \left[\frac{1}{\mathbf{P}_{Z_t|z(T\setminus t)}(z_t)}\boldsymbol{\pi}_{z_t} \odot \left[\mathbf{\Pi}^{-T}\mathbf{P}_{Z_t|z(T\setminus t)}\right]\right]\\
&= \arg\min_{\hat{x}\in\mathcal{A}} \left[\mathbf{P}_{Z_t|z(T\setminus t)}\right]^T \mathbf{\Pi}^{-1}\left[\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_{z_t}\right].
\end{aligned} \tag{25}$$

Again, we see that the estimator is given solely in terms of the distribution of the channel output process $Z(T)$, and involves the inversion of the channel matrix $\mathbf{\Pi}$ corresponding to *one* input-output pair, regardless of the size of the index set T.

As it depends on the unknown input distributions, the vector $\mathbf{P}_{Z_t|z(T\setminus t)}$ is not available in the universal setting. Our approach consists in estimating the empirical conditional output distributions from the observed data and use them in (25) in lieu of $\mathbf{P}_{Z_t|z(T\setminus t)}$.

For simplicity, we demonstrate this approach in the case where T is the one-dimensional index set $\{1, 2, \ldots, n\}$. Specifically, let now $\mathbf{X} = \{X_n\}_{n\geq 1}$ be a stationary ergodic process taking values in $\mathcal{A}^\infty$ and let $\mathbf{Z}$ denote the output of the memoryless channel $\mathbf{\Pi}$ whose input is $\mathbf{X}$. To estimate $\mathbf{P}_{Z_t|z(T\setminus t)} = \mathbf{P}_{Z_t|(z_1^{t-1} z_{t+1}^n)}$ from the data, consider the problem of estimating from the data the conditional probability vector $\mathbf{P}_{Z_t|(z_{t-k}^{t-1} z_{t+1}^{t+k})}$, for some $k \geq 0$. Two conflicting goals compete in the choice of k to produce a good estimate of $\mathbf{P}_{Z_t|(z_1^{t-1} z_{t+1}^n)}$. On one hand, we would like to choose a large value of k in order to approach the probabilities conditioned on the entire "punctured" sequence $z_1^{t-1} z_{t+1}^n$. On the other hand, k cannot be too large for otherwise our estimates would be based on sparse data. This trade-off is customary in sequential prediction and compression problems. Let further $g_{\mathsf{opt}}^k : \mathcal{A}^{2k+1} \to \mathcal{A}$ denote the minimum-mean-loss estimator for X_{k+1} based on Z^{2k+1} which, by (25), is given by

$$g_{\mathsf{opt}}^k(z^{2k+1}) = \arg\min_{\hat{x}\in\mathcal{A}} \left[\mathbf{P}_{Z_{k+1},(z_1^k z_{k+2}^{2k+1})}\right]^T \mathbf{\Pi}^{-1} \left[\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_{z_{k+1}}\right] \tag{26}$$

where we replaced $\mathbf{P}_{Z_{k+1}|(z_1^k z_{k+2}^{2k+1})}$ by $\mathbf{P}_{Z_{k+1},(z_1^k z_{k+2}^{2k+1})}$ (see definition (21)) using the fact that the normalization constant is independent of the optimization variable $\hat{x}$ in (25). Note that for $n > 2k$, by stationarity, g_{opt}^k minimizes

$$E\left[\sum_{i=k+1}^{n-k} \Lambda(X_i, g(Z_{i-k}^{i+k}))\right] \tag{27}$$

over all $g : \mathcal{A}^{2k+1} \to \mathcal{A}$.

The mapping g_{opt}^k, as defined in (26), depends on the distribution of the $2k+1$-tuple Z^{2k+1}. When this distribution is not known, it can be estimated from the data. Indeed, the ergodicity of $\mathbf{X}$, and hence of $\mathbf{Z}$, implies that for all $k \geq 0$ and any string $u_1^{2k+1} \in \mathcal{A}^{2k+1}$,

$$\frac{1}{n}\mathbf{m}(Z^n, u_1^k, u_{k+2}^{2k+1})[u_{k+1}] \stackrel{n\to\infty}{\longrightarrow} P(Z^{2k+1} = u_1^{2k+1}) \quad a.s., \tag{28}$$

or, in vector notation, that

$$\frac{1}{n}\mathbf{m}(Z^n, u_1^k, u_{k+2}^{2k+1}) \stackrel{n\to\infty}{\longrightarrow} \mathbf{P}_{Z_{k+1},(u_1^k u_{k+2}^{2k+1})} \quad a.s. \tag{29}$$

This relationship motivates using the left-hand side of (29) to estimate the right-hand side.

Note that while $\mathbf{\Pi}^{-T}\mathbf{P}_{Z_t|z(T\setminus t)}$ is the conditional input distribution, $\mathbf{\Pi}^{-T}\mathbf{m}(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})$ need not be a distribution, as not only it is unnormalized but it may have negative valued components.

For fixed $k \geq 0$ and $u^{2k+1} \in \mathcal{A}^{2k+1}$, comparing (6) with (26) and keeping (29) in mind, it may be hoped that, for large n, $g_{Z^n}^k(u_1^k, u_{k+1}, u_{k+2}^{2k+1}) \approx g_{\mathsf{opt}}^k(u^{2k+1})$ with high probability, or at least that the performance of $g_{Z^n}^k$ in the sense of (27) is close to that of g_{opt}^k. This, as we show in Section 4,

turns out to be the case in a remarkably strong sense, not only for stationary and ergodic $\mathbf{X}$, but in an individual-sequence setting as well.

3-C Non-Square Channel Transition Probability Matrix

It is easy to generalize the previous derivation of the DUDE to the case where the channel transition probability matrix is non-square, 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 (6) by

$$g^k_{\mathbf{a}}(\mathbf{b}, \alpha, \mathbf{c}) = \arg\min_{\hat{x} \in \mathcal{A}} \mathbf{m}^T(\mathbf{a}, \mathbf{b}, \mathbf{c}) \mathbf{\Pi}^T \left(\mathbf{\Pi}\mathbf{\Pi}^T\right)^{-1} \left[\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_\alpha\right]. \tag{30}$$

To motivate this form we follow all the steps in subsection 3-B verbatim, except that from (22) we now write

$$\begin{aligned} \mathbf{P}_{X_t|z(T\backslash t)} &= (\mathbf{\Pi}\mathbf{\Pi}^T)^{-1}\mathbf{\Pi}\mathbf{\Pi}^T \mathbf{P}_{X_t|z(T\backslash t)} \\ &= (\mathbf{\Pi}\mathbf{\Pi}^T)^{-1}\mathbf{\Pi}\mathbf{P}_{Z_t|z(T\backslash t)}. \end{aligned} \tag{31}$$

Substituting the right-most side of (31) in lieu of $\left[\mathbf{\Pi}^{-T}\mathbf{P}_{Z_t|z(T\backslash t)}\right]$ in (23)-(24), we obtain

$$\hat{X}^{\mathsf{opt}}(z(T))[t] = \arg\min_{\hat{x} \in \mathcal{A}} \left[\mathbf{P}_{Z_t|z(T\backslash t)}\right]^T \mathbf{\Pi}^T (\mathbf{\Pi}\mathbf{\Pi}^T)^{-1} \left[\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_{z_t}\right]. \tag{32}$$

The above derivation can be readily extended by replacing the Moore-Penrose generalized inverse (cf., e.g., [29]) $\mathbf{\Pi}^T(\mathbf{\Pi}\mathbf{\Pi}^T)^{-1}$ appearing in (30) and (32) 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 the generalized inverse exists. 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 rank $\mathbf{\Pi}$ and the DUDE defined through (30).

3-D A Closer Look at Special Cases

To conclude this section, we 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 non-binary case) in all the examples below.

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

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

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

Substituting the value of $\mathbf{\Pi}^{-1}$ into (6) yields, following simple arithmetic,

$$g^k_{z^n}(u^{-1}_{-k}, u_0, u^k_1) = \begin{cases} u_0 & \text{if } \frac{\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[u_0]}{\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[u_0] + \mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[\bar{u}_0]} \geq 2\delta(1-\delta) \\ \bar{u}_0 & \text{otherwise,} \end{cases} \tag{33}$$

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 left and right k-contexts are equal to u_0 among the total number of occurrences of this double-sided context. If the fraction of such occurrences 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 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. Since the noisy signal is Bernoulli($\theta(1-\delta) + (1-\theta)\delta$), we can express the above condition for leaving the signal untouched as $\theta(1-\delta) + (1-\theta)\delta \geq 2\delta(1-\delta)$. Now, since the frequency of appearances 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 (33), it can be seen that this is precisely the kind of scheme that the DUDE is independently employing within each of the double-sided k-contexts occurring.

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 (sub-optimal) compression-based scheme of [16] which, by definition, 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 (6) yields, for $\delta < (M-1)/M$,

$$g^k_{z^n}(u^k_{-k}) = \arg\min_{\hat{x}\in\mathcal{A}} \left[\delta\xi(u^{-1}_{-k}, u^k_1) - (M-1)\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[\hat{x}]\right] \cdot \left(\delta + [(1-\delta)M - 1]\,\mathbf{1}_{\{\hat{x}=u_0\}}\right) \tag{34}$$

where

$$\xi(u^{-1}_{-k}, u^k_1) = \sum_{a\in\mathcal{A}} \mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[a]\,. \tag{35}$$

Letting

$$x^* = \arg\max_{\hat{x}\neq u_0} \mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[\hat{x}]\,, \tag{36}$$

the decision rule in (34) assumes the simpler form

$$g^k_{z^n}(u^k_{-k}) = \begin{cases} u_0 & \text{if } \varsigma\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[u_0] - \mu\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[x^*] \geq \xi(u^{-1}_{-k}, u^k_1) \\ x^* & \text{otherwise,} \end{cases} \tag{37}$$

where

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

and

$$\mu = \frac{M-1}{(1-\delta)M-1}\,.$$

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

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

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 (6) takes the form

$$g^k_{z^n}(u^{-1}_{-k}, 1, u^k_1) = \begin{cases} 0 & \text{if } \frac{1-\delta}{2\delta} < \frac{\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[0]}{\mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)[1]} \\ 1 & \text{otherwise.} \end{cases} \tag{38}$$

- *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

$$\mathbf{\Pi} = \left[\begin{array}{c|c} & \delta \\ (1-\delta)\mathbf{I}_M & \vdots \\ & \delta \end{array}\right] \tag{39}$$

where $\mathbf{I}_M$ denotes the $M \times M$ identity matrix. This setting falls within the purview of the DUDE derived in subsection 3-C, equation (30) (or (32)), which we now explicitly obtain. We have

$$\mathbf{\Pi}\,\mathbf{\Pi}^T = (1-\delta)^2\mathbf{I}_M + \delta^2\mathbf{U}_M, \tag{40}$$

where $\mathbf{U}_M$ denotes the $M \times M$ matrix all whose entries equal 1. The inverse of the matrix in (40) is readily verified to be given by

$$\left(\mathbf{\Pi}\,\mathbf{\Pi}^T\right)^{-1} = a\mathbf{I}_M + b\mathbf{U}_M, \tag{41}$$

where $a = 1/(1-\delta)^2$ and $b = -a\delta^2/((1-\delta)^2 + \delta^2 M)$. Thus

$$\mathbf{\Pi}^T\left(\mathbf{\Pi}\,\mathbf{\Pi}^T\right)^{-1} = \left[\begin{array}{c} (1-\delta)[a\mathbf{I}_M + b\mathbf{U}_M] \\ \hline \delta(Mb+a)[1\cdots 1] \end{array}\right] \tag{42}$$

and, consequently,

$$\frac{1}{\delta}\mathbf{\Pi}^T\left(\mathbf{\Pi}\,\mathbf{\Pi}^T\right)^{-1}(\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_e) = \left[\begin{array}{c} (1-\delta)(M-1)b \\ \vdots \\ \vdots \\ \vdots \\ (1-\delta)(M-1)b \\ \delta(M-1)(bM+a) \end{array}\right] - \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ (1-\delta)a \\ 0 \\ \vdots \\ 0 \\ 0 \end{array}\right] + \left[\begin{array}{c} 1 \\ 1 \\ \vdots \\ \vdots \\ \vdots \\ 1 \\ 0 \end{array}\right](1-\delta)a, \tag{43}$$

where the non-zero term in the second vector on the right side of (43) is the $\hat{x}$-th component. Since the first and third vectors do not depend on $\hat{x}$, we obtain

$$g^k_{z^n}(u^{-1}_{-k}eu^k_1) = \arg\min_{\hat{x}\in\mathcal{A}} \mathbf{m}^T(z^n, u^{-1}_{-k}, u^k_1)\mathbf{\Pi}^T\left(\mathbf{\Pi}\,\mathbf{\Pi}^T\right)^{-1}(\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_e) = \arg\max_{\hat{x}\in\mathcal{A}} \mathbf{m}(z^n, u^{-1}_{-k}, u^k_1)(\hat{x}) \tag{44}$$

and, of course, $g^k_{z^n}(u^k_{-k}) = u_0$ for $u_0 \neq e$, since only locations i for which $z_i = e$ need be corrected. As one may have expected, we correct every erasure with the most frequent symbol for its context. Note that this denoiser does not depend on the channel parameter δ.

4 Universal Optimality: The Semi-Stochastic Setting

In this section, we assess the strong asymptotic optimality of the denoiser introduced in Subsection 3-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 first introduced into information theory by Ziv in his work [59] on rate distortion coding of individual sequences. More recently, problems of prediction [53, 56], as well as of limited-delay coding [54] of noise-corrupted individual sequences were also considered. As mentioned in Section 1, the semi-stochastic setting is also related to the classical compound decision problem [25, 39, 40, 42, 43, 50], which can be viewed as the particular case of our denoising setting with $k = 0$.

4-A Statement of the Main Result

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 *k-th order sliding-window denoiser* $\hat{X}^n$ is characterized by the property that for all $z^n \in \mathcal{A}^n$,

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

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

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

so that

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

We let $\mathcal{S}_k$ denote the class of k-th order sliding-window denoisers. 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

$$\begin{aligned}(n-2k)L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n) &= \sum_{i=k+1}^{n-k} \Lambda(x_i, \hat{X}^n(z^n)[i]) \\ &= \sum_{i=k+1}^{n-k} \Lambda(x_i, f_{z^n}(z_{i-k}^{i+k}))\end{aligned}$$

$$= \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \sum_{a \in \mathcal{A}} \mathbf{q}(z^n, x^n, u_{-k}^k)[a] \Lambda(a, f_{z^n}(u_{-k}^k))$$
$$= \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \boldsymbol{\lambda}_{f_{z^n}(u_{-k}^k)}^T \mathbf{q}(z^n, x^n, u_{-k}^k) \tag{45}$$

where the statistics $\mathbf{q}$ are defined in (2). Note also that the DUDE, $\hat{X}_{\mathsf{univ}}^n$, is a member of $\mathcal{S}_{k_n}$, with the mappings $\{f_{z^n}\}$ given by $g_{z^n}^k(z_{i-k}^{i-1}, z_i, z_{i+1}^{i+k})$ for $k = k_n$ (see Equation (8)). Here, k_n is any unboundedly increasing function of n with certain limitations on the growth rate, which are required for universality (recall (9)).

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 *k-th order minimum loss* of (x^n, z^n) by

$$\begin{aligned} D_k(x^n, z^n) &\stackrel{\triangle}{=} \min_{\hat{X}^n \in \mathcal{S}_k} L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n) \\ &= \min_{f:\mathcal{A}^{2k+1} \to \mathcal{A}} \left[\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(x_i, f(z_{i-k}^{i+k})) \right]. \end{aligned} \tag{46}$$

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 (46) depends on both x^n and z^n. It follows, a fortiori, that the definition of the class of k-th 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_n-th order sliding window one-pass denoisers.

By (45), the k-th order minimum loss takes the form

$$\begin{aligned} D_k(x^n, z^n) &= \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \min_{\hat{x} \in \mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \mathbf{q}(z^n, x^n, u_{-k}^k) \\ &= \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} U(\mathbf{q}(z^n, x^n, u_{-k}^k)). \end{aligned} \tag{47}$$

Our main result, Theorem 1, states that for any input sequence $\mathbf{x}$, the DUDE, as defined in (8), 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 $\{\hat{X}_{\mathsf{univ}}^n\}$ defined in (8) with $\lim_{n\to\infty} k_n = \infty$ satisfies*

(a) $\lim_{n\to\infty} [L_{\hat{X}_{\mathsf{univ}}^n}(x^n, Z^n) - D_{k_n}(x^n, Z^n)] = 0$ *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 5. 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).

The following theorem is the key result underlying the proof of Theorem 1. Throughout this section, and in the statement of Theorem 2 below 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. To state Theorem 2, we further define the function

$$\varphi(p) \stackrel{\triangle}{=} \frac{1}{1-2p} \log \frac{1-p}{p}, \quad 0 \le p < 1/2. \tag{48}$$

We extend the definition (48), again by continuity, to $\varphi(1/2) = 2$.

Theorem 2 *Let*

$$F_{\mathbf{\Pi}} \stackrel{\triangle}{=} \sum_{a \in \mathcal{A}} \left[\varphi(\max_{A \subseteq \mathcal{A}} \min(\Pi(a, A), \Pi(a, A^c)))\right]^{-1}, \quad C_{\mathbf{\Lambda},\mathbf{\Pi}} \stackrel{\triangle}{=} \mathbf{\Lambda}_{\max}\left(1 + \|\mathbf{\Pi}^{-1}\|_\infty\right),$$

and

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

Then, for any $k \ge 0$, $n > 2k$, $x^n \in \mathcal{A}^n$, and $\varepsilon > 0$, the denoiser $\hat{X}^{n,k}$ defined in (7) satisfies

$$\Pr\left(L_{\hat{X}^{n,k}}(x_{k+1}^{n-k}, Z^n) - D_k(x^n, Z^n) > \varepsilon\right) \le K_1(k+1)M^{2k+1} \exp\left(-\frac{(n-2k)\varepsilon^2}{4(k+1)M^{2k}F_{\mathbf{\Pi}}C^2_{\mathbf{\Lambda},\mathbf{\Pi}}}\right) \tag{49}$$

$$E\left[L_{\hat{X}^{n,k}}(x_{k+1}^{n-k}, Z^n) - D_k(x^n, Z^n)\right] \le \sqrt{\frac{2}{\pi}} C_{\mathbf{\Lambda},\mathbf{\Pi}} V_{\mathbf{\Pi}} M^k \sqrt{\frac{k+1}{n-2k}} + C_{\mathbf{\Lambda},\mathbf{\Pi}} M^{2k+2} \frac{k+1}{n-2k} \tag{50}$$

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 k-th 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 (50) tells us that the bound

on the expected excess loss becomes smaller for "skewed" channels. For example, for the BSC, $V_{\mathbf{\Pi}} = \sqrt{8\mathrm{Var}(\mathbf{\Pi})}$, where $\mathrm{Var}(\mathbf{\Pi})$ denotes the variance of the channel conditional distributions. The factor $F_{\mathbf{\Pi}}$, which also tends to zero as the channel becomes less "noisy", captures the analogous dependency on $\mathbf{\Pi}$ in the exponent of (49). Notice that $V_{\mathbf{\Pi}} \leq \sqrt{M(M-1)}$ 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 non-full-rank matrix, reflecting the fact that universal denoising becomes increasingly difficult in this regime. The proof of Theorem 2 is deferred to Subsection 4-C.

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 (49) 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 (49) is summable. Thus, by the Borel-Cantelli Lemma

$$L_{\hat{X}^{n,k_n}}(x_{k_n+1}^{n-k_n}, Z^n) - D_{k_n}(x^n, Z^n) \leq \varepsilon \quad \text{eventually almost surely.} \tag{51}$$

Now, for any n-block denoiser $\hat{X}^n$ and $k \geq 0$,[6]

$$L_{\hat{X}^n}(x^n, Z^n) = \frac{1}{n}\sum_{i=1}^{n} \Lambda(x_i, \hat{X}^n(Z^n)[i]) \leq \frac{2k\mathbf{\Lambda}_{\max}}{n} + L_{\hat{X}^n}(x_{k+1}^{n-k}, Z^n)\,. \tag{52}$$

In particular, (52) holds for the sequence of denoisers $\{\hat{X}^n_{\mathsf{univ}}\}$. Taking limit suprema in (51), using (52) with $k = k_n$, and noticing that k_n/n vanishes, we obtain, for any $\varepsilon > 0$,

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

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 Equation (50) in Theorem 2 with $k = k_n$, (52), and the fact that for the allowable range of k_n, $k_n/n = O((\log n)/n)$. □

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 [60], the FS predictability of [19], and, in particular, the conditional FS predictability of [56]. To this end, we define the *sliding-window minimum loss* of $(\mathbf{x}, \mathbf{z})$ by

$$D(\mathbf{x}, \mathbf{z}) = \lim_{k\to\infty} D_k(\mathbf{x}, \mathbf{z}) \tag{53}$$

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

where

$$D_k(\mathbf{x},\mathbf{z}) = \limsup_{n\to\infty} D_k(x^n, z^n). \tag{54}$$

Note that $D_k(\mathbf{x},\mathbf{z})$ is non-increasing 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, thanks to the memorylessness of the channel, it is in fact degenerate:

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

$$D(\mathbf{x},\mathbf{Z}) = D(\mathbf{x}) \quad a.s. \tag{55}$$

Remark: We refer to $D(\mathbf{x})$ as the *denoisability* of $\mathbf{x}$. Intuitively, Equation (55) 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 k-th order empirical statistics of the two sequences, which for each given input k-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., [17]).

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*

$$\limsup_{n\to\infty} L_{\hat{X}^n_{\mathsf{univ}}}(x^n, Z^n) \le D(\mathbf{x}) \quad a.s. \quad \forall \mathbf{x}\in\mathcal{A}^\infty \tag{56}$$

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 $k_n \ge k$, we have

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

It follows that

$$\limsup_{n\to\infty} D_{k_n}(x^n, Z^n) \le \limsup_{n\to\infty} \left[\frac{n-2k}{n-2k_n} D_k(x^n, Z^n)\right] = D_k(\mathbf{x},\mathbf{Z}) \tag{57}$$

implying, by the arbitrariness of k, that

$$\limsup_{n\to\infty} D_{k_n}(x^n, Z^n) \le D(\mathbf{x},\mathbf{Z}). \tag{58}$$

The proof is completed by combining Theorem 1, Part (a), with (58), and invoking Claim 1. □

4-B Intuition and Idea behind the Proof of Theorem 2

It may seem striking that a denoiser that was derived via heuristics from the fully stochastic setting (Subsection 3-B) performs so well in the semi-stochastic setting. Our goal in this subsection is to provide intuition as to why this is the case, while outlining the main idea behind the proof of Theorem 2 (which is deferred to the next subsection).

First, observe that, for $n > 2k$, the function f attaining $D_k(x^n, z^n)$ in (47) is given by the Bayes response (cf. Section 2)

$$f(u_{-k}^k) = \arg\min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \mathbf{q}(z^n, x^n, u_{-k}^k) = \hat{x}(\mathbf{q}(z^n, x^n, u_{-k}^k)). \tag{59}$$

Unfortunately, this mapping depends on x^n and, hence, cannot be implemented by a denoiser observing solely z^n. Thus, if our goal is to construct a denoiser approximately attaining $D_k(x^n, z^n)$, a plausible approach would be a denoiser $\hat{X}^n$ given, for $k+1 \le i \le n-k$, by

$$\hat{X}^n(z^n)[i] = \hat{x}(\hat{\mathbf{q}}(z^n, z_{i-k}^{i+k})) \tag{60}$$

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 (60) with (59), it is natural to expect, by continuity arguments, that the normalized loss of the denoiser in (60) be "close" to attaining $D_k(x^n, z^n)$ whenever $\hat{\mathbf{q}}(z^n, u_{-k}^k)$ is "close" to $\mathbf{q}(z^n, x^n, u_{-k}^k)$ for any u_{-k}^k. This intuition will be made precise in Lemma 1 below. Note that our denoiser in (7) is exactly of the form (60) if we choose

$$\hat{\mathbf{q}}(z^n, u_{-k}^k) = \boldsymbol{\pi}_{u_0} \odot \left[\boldsymbol{\Pi}^{-T} \mathbf{m}(z^n, u_{-k}^{-1}, u_1^k)\right]. \tag{61}$$

It thus remains to be argued that, for the semi-stochastic setting, the right-hand side of (61) is an efficient estimate of $\mathbf{q}(Z^n, x^n, u_{-k}^k)$. To get a feel for why this should be the case, take two contexts $u_{-k}^{-1}, u_1^k \in \mathcal{A}^k$, a symbol $a \in \mathcal{A}$, and consider the number of locations for which z_i, $k+1 \le i \le n-k$, appears in left context u_{-k}^{-1}, right context u_1^k, and the noiseless symbol is a, i.e., $\sum_{b\in\mathcal{A}} \mathbf{q}(Z^n, x^n, u_{-k}^{-1} b u_1^k)[a]$. Furthermore, 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.,

$$\Pi(a, u_0) \cdot \left[\sum_{b\in\mathcal{A}} \mathbf{q}(Z^n, x^n, u_{-k}^{-1} b u_1^k)[a]\right] \approx \mathbf{q}(Z^n, x^n, u_{-k}^k)[a], \tag{62}$$

no matter what the individual sequence x^n may be. This can indeed be shown to be the case (Lemma 3) in the strong sense that, regardless of the underlying individual sequence, the normalized magnitude of the difference between the two sides of (62) exceeds any $\varepsilon > 0$ with probability that decays exponentially with n (for fixed k), and vanishes with n in expectation. Summing now (62) over $a \in \mathcal{A}$ gives

$$\sum_{b\in\mathcal{A}} \boldsymbol{\pi}_{u_0}^T \mathbf{q}(Z^n, x^n, u_{-k}^{-1} b u_1^k) \approx \mathbf{m}(Z^n, u_{-k}^{-1}, u_1^k)[u_0], \tag{63}$$

or, in vector notation, iterating (63) over the possible $u_0 \in \mathcal{A}$,

$$\sum_{b\in\mathcal{A}} \mathbf{\Pi}^T \mathbf{q}(Z^n, x^n, u_{-k}^{-1} b u_1^k) \approx \mathbf{m}(Z^n, u_{-k}^{-1}, u_1^k), \tag{64}$$

implying, in turn,

$$\sum_{b\in\mathcal{A}} \mathbf{q}(Z^n, x^n, u_{-k}^{-1} b u_1^k) \approx \mathbf{\Pi}^{-T} \mathbf{m}(Z^n, u_{-k}^{-1}, u_1^k). \tag{65}$$

Combining (65) with (62) (written in vector notation) leads to

$$\boldsymbol{\pi}_{u_0} \odot \left[\mathbf{\Pi}^{-T} \mathbf{m}(Z^n, u_{-k}^{-1}, u_1^k)\right] \approx \mathbf{q}(Z^n, x^n, u_{-k}^k), \tag{66}$$

which is the desired conclusion (the precise statement of this conclusion is given in the proof of Theorem 2).

4-C Proof of Theorem 2

To prove Theorem 2 we first present three lemmas. The first two lemmas establish inequalities that are valid for any pair of sequences x^n, z^n, whereas the third one is probabilistic.

Lemma 1 *Fix $k \geq 0$, $z^n \in \mathcal{A}^n$, 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 k-th order sliding window denoiser $\hat{X}^n$ with sliding-block function given by the Bayes responses to $\{\mathbf{v}(u_{-k}^k)\}$:*

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

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

$$0 \leq L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n) - D_k(x^n, z^n) \leq \frac{\mathbf{\Lambda}_{\max}}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \left\| \mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{v}(u_{-k}^k) \right\|_1 . \tag{67}$$

Proof: The left inequality in (67) follows trivially from the fact that $\hat{X}^n$ is a k-th order sliding-window denoiser. To derive the second inequality, notice that by (47), (45), and the definition of the denoiser $\hat{X}^n$, we have

$$\begin{aligned}
L_{\hat{X}^n}(x_{k+1}^{n-k}, z^n) - D_k(x^n, z^n) &= \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} [\boldsymbol{\lambda}_{\hat{x}(\mathbf{v})}^T - \boldsymbol{\lambda}_{\hat{x}(\mathbf{q})}^T]\, \mathbf{q}(z^n, x^n, u_{-k}^k) \\
&\leq \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} [\boldsymbol{\lambda}_{\hat{x}(\mathbf{v})}^T - \boldsymbol{\lambda}_{\hat{x}(\mathbf{q})}^T][\mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{v}(u_{-k}^k)] \qquad (68) \\
&\leq \frac{1}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \left\| \boldsymbol{\lambda}_{\hat{x}(\mathbf{v})} - \boldsymbol{\lambda}_{\hat{x}(\mathbf{q})} \right\|_\infty \cdot \left\| \mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{v}(u_{-k}^k) \right\|_1 \\
&\leq \frac{\mathbf{\Lambda}_{\max}}{n-2k} \sum_{u_{-k}^k \in \mathcal{A}^{2k+1}} \left\| \mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{v}(u_{-k}^k) \right\|_1 \qquad (69)
\end{aligned}$$

where, for simplicity, we have dropped the arguments of $\mathbf{v}$ and $\mathbf{q}$ when used for indexing columns of $\mathbf{\Lambda}$, and (68) holds since, for any pair of M-vectors $\mathbf{v}$ and $\mathbf{w}$, we have

$$[\boldsymbol{\lambda}_{\hat{x}(\mathbf{v})} - \boldsymbol{\lambda}_{\hat{x}(\mathbf{w})}]^T \mathbf{v} \leq 0.$$

□

The continuity property established in Lemma 1 is, in fact, typical of finite matrix games [24, Equation (14)]. In particular, the proposed denoiser is clearly of the form covered by the lemma, with

$$\mathbf{v}(u_{-k}^k) = \boldsymbol{\pi}_{u_0} \odot \left[\mathbf{\Pi}^{-T} \mathbf{m}(z^n, u_{-k}^{-1}, u_1^k) \right]. \tag{70}$$

For this case, the upper bound (67) can be further upper-bounded as follows.

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

$$\sum_{u_0 \in \mathcal{A}} \left\| \mathbf{q}(z^n, x^n, u_{-k}^k) - \boldsymbol{\pi}_{u_0} \odot \left[\mathbf{\Pi}^{-T} \mathbf{m}(z^n, u_{-k}^{-1}, u_1^k) \right] \right\|_1 \leq$$
$$\left(1 + \|\mathbf{\Pi}^{-1}\|_\infty\right) \sum_{u_0 \in \mathcal{A}} \left\| \mathbf{q}(z^n, x^n, u_{-k}^k) - \mathbf{q}'(z^n, x^n, u_{-k}^k) \right\|_1 \tag{71}$$

where

$$\mathbf{q}'(z^n, x^n, u_{-k}^k) \stackrel{\triangle}{=} \boldsymbol{\pi}_{u_0} \odot \sum_{b \in \mathcal{A}} \mathbf{q}(z^n, x^n, u_{-k}^{-1} b u_1^k). \tag{72}$$

Lemma 2 is proved in Appendix A.

As is 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 a good estimate of $\mathbf{q}(Z^n, x^n, u_{-k}^k)$. As discussed in Subsection 4-B, this step is indeed plausible (see (62), where the left-hand side is precisely $\mathbf{q}'(Z^n, x^n, u_{-k}^k)[a]$). However, there are two apparent obstacles to making the intuition given in (62) precise. One is that the number of symbols in z^n which occurred with left and right contexts u_{-k}^{-1} and 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 (71) 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 occurred with left and right contexts u_{-k}^{-1} and 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 time indices i such that $i \equiv \ell \bmod (k+1)$, $\ell = 0, 1, \cdots, k$. 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 non-contributing symbols.

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

$$\mathbf{q}_\ell(\mathbf{a}, \mathbf{b}, \mathbf{c})[a] = \left| \left\{ i : i \in \mathcal{I}_\ell, a_{i-k}^{i+k} = \mathbf{c}, b_i = a \right\} \right|, \tag{73}$$

where

$$\mathcal{I}_\ell \stackrel{\triangle}{=} \{ i : k+1 \le i \le n-k, i \equiv \ell \bmod (k+1) \}.$$

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

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

Similarly, we define, as in (72),

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

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

$$\sum_{u_0 \in \mathcal{A}} \left\| \mathbf{q}(u_{-k}^k) - \boldsymbol{\pi}_{u_0} \odot \left[\boldsymbol{\Pi}^{-T} \mathbf{m}(z^n, u_{-k}^{-1}, u_1^k) \right] \right\|_1 \le \left(1 + \|\boldsymbol{\Pi}^{-1}\|_\infty\right) \sum_{\ell=0}^{k} \sum_{u_0 \in \mathcal{A}} \left\| \boldsymbol{\Delta}_\ell(u_{-k}^k) \right\|_1 \tag{74}$$

where

$$\boldsymbol{\Delta}_\ell(u_{-k}^k) \stackrel{\triangle}{=} \mathbf{q}_\ell(u_{-k}^k) - \mathbf{q'}_\ell(u_{-k}^k).$$

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

$$Z(\ell) \stackrel{\triangle}{=} \{ Z_i : 1 \le i \le n, i \notin \mathcal{I}_\ell \}.$$

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

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

denote the number of times z_i, $i \in \mathcal{I}_\ell$, occurs with left context u_{-k}^{-1} and right context u_1^k, when $x_i = a$. Notice that given x^n and *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*

$$F_{\mathbf{\Pi},a} \stackrel{\triangle}{=} \varphi\left(\max_{A\subseteq\mathcal{A}} \min(\Pi(a,A),\Pi(a,A^c))\right)$$

where the function $\varphi(\cdot)$ is given in (48), and let

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

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

$$\Pr\left(\sum_{u_0\in\mathcal{A}} \left|\mathbf{\Delta}_\ell(u_{-k}^k)[a]\right| > n_\ell(u_{-k}^{-1},u_1^k,a)\varepsilon \;\middle|\; Z(\ell)=z(\ell)\right) \le (2^M-2)e^{-n_\ell(u_{-k}^{-1},u_1^k,a)F_{\mathbf{\Pi},a}\varepsilon^2/4} \tag{75}$$

and

$$E\left[\sum_{u_0\in\mathcal{A}} \left|\mathbf{\Delta}_\ell(u_{-k}^k)[a]\right| \;\middle|\; Z(\ell)=z(\ell)\right] \le \sqrt{\frac{2}{\pi}} V_{\mathbf{\Pi},a}\sqrt{n_\ell(u_{-k}^{-1},u_1^k,a)} + M. \tag{76}$$

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 (75) of Lemma 3 by applying the following result of [36], where

$$D_B(p_1\|p_2) = p_1\log(p_1/p_2) + (1-p_1)\log((1-p_1)/(1-p_2))$$

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 $\{1,\ldots,d\}$ and $\mathbf{P} = [P(1),\ldots,P(d)]$. Let $X_1, X_2, ..., X_m$ be i.i.d. random variables distributed according to P, and let $\hat{\mathbf{P}}$ denote the probability vector corresponding to the empirical distribution, $\hat{\mathbf{P}} = (1/m)[\sum_i \mathbf{1}_{\{X_i=1\}},\ldots,\sum_i \mathbf{1}_{\{X_i=d\}}]$. Then, for all $\varepsilon > 0$,*

$$\begin{aligned} \Pr(\|\mathbf{P}-\hat{\mathbf{P}}\|_1 \ge \epsilon) &\le (2^d-2)e^{-mD_B(p^*+\frac{\epsilon}{2}\,\|\,p^*)} &(77)\\ &\le (2^d-2)e^{-m\varphi(p^*)\frac{\epsilon^2}{4}}, &(78)\end{aligned}$$

where

$$p^* = \max_{A\subseteq\{1,\ldots,d\}} \min\left(P(A),P(A^c)\right)$$

and the function $\varphi(\cdot)$ is given in (48).

Proof sketch: The event $\{\|\mathbf{P}-\hat{\mathbf{P}}\|_1 \ge \varepsilon\}$ is equivalent to the union of events $\{\mathbf{s}^T[\mathbf{P}-\hat{\mathbf{P}}] \ge \varepsilon\}$ where $\mathbf{s}$ ranges over the 2^d-2 non-constant vectors in $\{-1,1\}^d$. Applying the union bound and then the Chernoff bounding technique to each sub-event yields, after bounding by the worst case, the first inequality. For

$p \le 1/2$, elementary calculus shows that $\inf_{0\le\epsilon\le 1-p} D_B(p+\epsilon\|p)/\varepsilon^2 = \varphi(p)$, yielding (78). See [36] for the details. □

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 $\mathcal{I}_\ell(u_{-k}^{-1}, u_1^k, a) \stackrel{\triangle}{=} \{i \in \mathcal{I}_\ell : z_{i-k}^{i-1} = u_{-k}^{-1}, z_{i+1}^{i+k} = u_1^k, x_i = a\}$, which is *completely determined* by x^n and $z(\ell)$. Moreover, by (72),

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

Therefore, after normalization by $n_\ell(u_{-k}^{-1}, u_1^k, a)$, the sum in the left-hand sides of (75) and (76) 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 (75) 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 $0 \le p \le 1/2$ [20, Chapter IX, Problem 35],

$$E|S_{m,p} - mp| = E|S_{m,q} - mq| = 2\nu q \binom{m}{\nu} p^\nu q^{m-\nu} \tag{79}$$

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 (79) is $p' = \nu/(m+1)$. Thus, applying Stirling's formula to $\binom{m+1}{\nu}$ we obtain, after straightforward algebraic manipulations,

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

Clearly, $p'(1-p') \le p''(1-p'')$, where $p'' = \min((mp+1)/(m+1), 1/2)$. Moreover, $p'' \ge p$ and $(m+1)p'' \le mp + 1$, so that

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

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. □

Discussion:

(a) It is shown in [36] that the exponent in (77) coincides with that given by Sanov's Theorem and hence is the best possible.[7] A stronger version of Lemma 3, based on this optimal rate, could have been derived. The integration of this rate into the proof of Theorem 2, however, appears to require the weaker version (75).

[7] Moreover, note that the bound in (77) 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., [9, Theorem 12.4.1].

(b) The constant $F_{\mathbf{\Pi},a}$ in the exponent of (75) is bounded from below by 2, and indeed replacing $F_{\mathbf{\Pi},a}$ by 2 coincides simply with the application of Pinsker's inequality [9, Lemma 12.6.1] to $D_B(p^* + \epsilon/2 \| p^*)$ in (77). Such a bound, however, would not reflect the intuitively appealing fact that less "noisy" channels result in larger exponents.

Proof of Theorem 2: Using Lemma 1 with $\{\mathbf{v}(u_{-k}^k)\}$ given by (70), and (74), we obtain, for any $\varepsilon > 0$,

$$\begin{aligned} P &\stackrel{\triangle}{=} \Pr\left(L_{\hat{X}^{n,k}}(x_{k+1}^{n-k}, Z^n) - D_k(x^n, Z^n) > \varepsilon\right) \\ &\leq \Pr\left(\sum_{\ell=0}^{k} \sum_{\mathbf{u}\in\mathcal{A}^{2k+1}} \sum_{a\in\mathcal{A}} |\mathbf{\Delta}_\ell(\mathbf{u})[a]| > \frac{(n-2k)\varepsilon}{C_{\mathbf{\Lambda},\mathbf{\Pi}}}\right) \\ &\leq \sum_{\ell=0}^{k} \Pr\left(\sum_{\mathbf{u}\in\mathcal{A}^{2k+1}} \sum_{a\in\mathcal{A}} |\mathbf{\Delta}_\ell(\mathbf{u})[a]| > \frac{(n-2k)\gamma_\ell\varepsilon}{C_{\mathbf{\Lambda},\mathbf{\Pi}}}\right) \end{aligned} \tag{80}$$

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 right-most side of (80) via Lemma 3, we condition the events on the random variables $Z(\ell)$, to obtain

$$P \leq \sum_{\ell=0}^{k} \sum_{z(\ell)\in\mathcal{A}^{n-n_\ell}} \Pr(Z(\ell) = z(\ell)) \Pr\left(\sum_{\mathbf{u}\in\mathcal{A}^{2k+1}} \sum_{a\in\mathcal{A}} |\mathbf{\Delta}_\ell(\mathbf{u})[a]| > \frac{(n-2k)\gamma_\ell\varepsilon}{C_{\mathbf{\Lambda},\mathbf{\Pi}}} \middle| Z(\ell) = z(\ell)\right). \tag{81}$$

Letting P_ℓ denote the conditional probability in the right-hand side of (81), the union bound yields

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

where, again, conditioned on $Z(\ell)$, $\{\beta_{a,\mathbf{u}}\} \stackrel{\triangle}{=} \{\beta_{a,\mathbf{u}_L,\mathbf{u}_R}\}$ is a set of non-negative constants (to be specified later) satisfying $\sum_{\mathbf{u}_L,\mathbf{u}_R,a} \beta_{a,\mathbf{u}} = 1$. We can now apply Equation (75) in Lemma 3, which yields

$$P_\ell \leq (2^M - 2) \sum_{\mathbf{u}_L,\mathbf{u}_R\in\mathcal{A}^k} \sum_{a\in\mathcal{A}} \exp\left(-\frac{F_{\mathbf{\Pi},a}(n-2k)^2\gamma_\ell^2\beta_{a,\mathbf{u}}^2}{4n_\ell(\mathbf{u}_L,\mathbf{u}_R,a)} \cdot \frac{\varepsilon^2}{C_{\mathbf{\Lambda},\mathbf{\Pi}}^2}\right). \tag{82}$$

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

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

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 (82), equations (81) and (82) yield

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

We complete the proof of the bound (49) 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 (50), we use again Lemma 1 with $\{\mathbf{v}(u_{-k}^k)\}$ given by (70), and (74), to obtain

$$\begin{aligned}
E &\stackrel{\triangle}{=} E\left[(n-2k)\left(L_{\hat{X}^{n,k}}(x_{k+1}^{n-k},Z^n)-D_k(x^n,Z^n)\right)\right] \le C_{\mathbf{\Lambda},\mathbf{\Pi}}\sum_{\ell=0}^{k}\sum_{\mathbf{u}\in\mathcal{A}^{2k+1}}\sum_{a\in\mathcal{A}}E\left[\,|\mathbf{\Delta}_\ell(\mathbf{u})[a]|\,\right]\\
&= C_{\mathbf{\Lambda},\mathbf{\Pi}}\sum_{\ell=0}^{k}\sum_{\mathbf{u}_L,\mathbf{u}_R\in\mathcal{A}^k}\sum_{a\in\mathcal{A}}\sum_{z(\ell)\in\mathcal{A}^{n-n_\ell}}\Pr(z(\ell))\,E\left[\sum_{u_0\in\mathcal{A}}|\,\mathbf{\Delta}_\ell(\mathbf{u}_Lu_0\mathbf{u}_R)[a]\,|\,\middle|\,Z(\ell)=z(\ell)\right].
\end{aligned}$$

By (76) in Lemma 3, we can further upper-bound the expectation to obtain

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

where the second and fourth lines follow from the Cauchy-Schwarz inequality, completing the proof of (50). □

5 Universal Optimality: The Stochastic Setting

Consider the fully stochastic analogue of the setting of Section 4 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 ergodic $\mathbf{X}$. 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

$$\mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) = \min_{\hat{X}^n \in \mathcal{D}_n} E\left[L_{\hat{X}^n}(X^n, Z^n)\right], \tag{83}$$

the expectation on the right-hand side assuming $X^n \sim \mathbf{P}_{X^n}$. By stationarity, for all $m, n \geq 0$,

$$(m+n)\mathbb{D}(\mathbf{P}_{X^{m+n}}, \mathbf{\Pi}) \leq m\mathbb{D}(\mathbf{P}_{X^m}, \mathbf{\Pi}) + n\mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}). \tag{84}$$

Thus, by the Sub-additivity Lemma (cf., e.g., [12, Lemma 6.1.11]),

$$\lim_{n\to\infty} \mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) = \inf_{n\geq 1} \mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) \stackrel{\triangle}{=} \mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}). \tag{85}$$

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 ergodic 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 4. This is in accord with analogous situations in universal compression [60], prediction [31], and noisy prediction [55], where in the individual-sequence setting the class 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 of the class of allowable sources (cf. discussion on the duality between the viewpoints in [31]).

For integers i, j, let now $\mathbf{P}_{X_0|z_i^j} \in \mathcal{M}$ denote the M-dimensional probability vector whose a-th component is[8] $P(X_0 = a | Z_i^j = z_i^j)$. In Subsection 3-B, we used the fact that for a finite set of random variables, the best denoising performance is given by the conditional Bayes envelope (cf., Equation (10)). This property is now stated for a random process in Claim 2 below.

Claim 2 $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) = EU(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})$.

The claim results from the following lemma.

Lemma 4 *1. For $k, l \geq 0$, $EU(\mathbf{P}_{X_0|Z_{-k}^{l}})$ is decreasing in both k and l.*

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

$$\lim_{n\to\infty} EU(\mathbf{P}_{X_0|Z_{-k_n}^{l_n}}) = EU(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}}).$$

[8]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., [17]) of X_0 given Z_i^j, evaluated at z_i^j.

Lemma 4 and Claim 2 parallel similar results in sequential decision theory [31] (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 B for completeness.

Next, we show that with probability one, the sliding window minimum loss (cf. the definition (53)) for an individual sequence drawn from a stationary ergodic source, coincides with $\mathbb{D}(\mathbf{P_X}, \mathbf{\Pi})$. This result parallels [60, Theorem 4], where it is shown that the finite-state compressibility of an individual sequence drawn from a stationary ergodic source coincides with the entropy of the source with probability one.

Claim 3 *$D(\mathbf{X}, \mathbf{Z}) = \mathbb{D}(\mathbf{P_X}, \mathbf{\Pi})$ a.s.*

Proof: Recall the definition of $D_k(\mathbf{X}, \mathbf{Z})$ in (54), and notice that by stationarity and ergodicity, for each k and each map f taking $\mathcal{A}^{2k+1}$ into $\mathcal{A}$,

$$\lim_{n\to\infty} \left[\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(X_i, f(Z_{i-k}^{i+k})) \right] = E\Lambda(X_0, f(Z_{-k}^{k})) \quad a.s. \tag{86}$$

Since the set of all maps taking $\mathcal{A}^{2k+1}$ into $\mathcal{A}$ is finite, it follows that (86) implies

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

Since

$$\min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} E\Lambda(X_0, f(Z_{-k}^{k})) = EU(\mathbf{P}_{X_0|Z_{-k}^{k}}),$$

the proof is completed by letting $k \to \infty$ and invoking Lemma 4 and Claim 2. □

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

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

(a) $\limsup_{n\to\infty} L_{\hat{X}_{\mathsf{univ}}^n}(X^n, Z^n) \le \mathbb{D}(\mathbf{P_X}, \mathbf{\Pi})$ *a.s., provided that* $k_n M^{2k_n} = o(n/\log n)$.

(b) $\lim_{n\to\infty} EL_{\hat{X}_{\mathsf{univ}}^n}(X^n, Z^n) = \mathbb{D}(\mathbf{P_X}, \mathbf{\Pi})$, *provided that* $\sqrt{k_n} M^{k_n} = o(\sqrt{n})$.

Proof: To derive Part (a), notice that Corollary 1 (in Section 4) holds for all sequences $\mathbf{x}$ and, *a fortiori*, almost surely. Thus, the result follows by invoking Claims 1 and 3. As for Part (b), by Claim 2, we have

$$EL_{\hat{X}_{\mathsf{univ}}^n}(X^n, Z^n) - \mathbb{D}(\mathbf{P_X}, \mathbf{\Pi}) = E\left[U(\mathbf{P}_{X_0|Z_{-k_n}^{k_n}}) - U(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}}) \right] + E\left[L_{\hat{X}_{\mathsf{univ}}^n}(X^n, Z^n) - U(\mathbf{P}_{X_0|Z_{-k_n}^{k_n}}) \right]. \tag{87}$$

The first expectation in the right-hand side of (87) vanishes in the limit by Lemma 4, whereas for the second expectation we notice that, for any $k \geq 0$,

$$\begin{aligned} EU(\mathbf{P}_{X_0|Z_{-k}^k}) &= \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} E\Lambda(X_0, f(Z_{-k}^k)) && (88)\\ &= \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} E\left[\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda(X_i, f(Z_{i-k}^{i+k}))\right] \\ &\geq E\left[\min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\Lambda(X_i, f(Z_{i-k}^{i+k}))\right] && (89)\\ &= ED_k(X^n, Z^n) && (90) \end{aligned}$$

where the second equality follows by stationarity. Thus, the second expectation in the right-hand side of (87) vanishes in the limit by Theorem 1, Part (b). □

Remarks:

(a) Equation (87) provides insight into the convergence of $EL_{\hat{X}_{\mathsf{univ}}^n}(X^n, Z^n)$ 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 ergodic $\mathbf{X}$. 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) The inequality (90) parallels the well-known property that the conditional entropy of order k is an upper-bound on the expectation of the corresponding empirical entropy.

(c) For the range of values of k_n covered by Part (a) of the theorem, the convergence in expectation could be easily derived from the almost sure convergence by use of Fatou's Lemma.

6 Context-length selection

6-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 k ought to be selected for a specific sequence z^n. In particular, notice that even though Theorem 2 provides *non-asymptotic* information about how the denoiser $\hat{X}^{n,k}$ compares with the best k-th 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, akin to that of double-universality in universal data compression (see, e.g., [31]). 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, search for the optimal k, and transmit its value to the decoder. The key difference with the denoising setting, however, is that the optimal denoiser depends on the unobserved sequence x^n. 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

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

In words, $k^*(x^n, z^n)$ is the order of the best 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

$$k_n(\cdot) \stackrel{\triangle}{=} \arg\min_{\kappa(\cdot)} \max_{x^n \in \mathcal{A}^n} E\left[L_{\hat{X}^{n,\kappa(Z^n)}}(x^n, Z^n) - L_{\hat{X}^{n,k^*(x^n,Z^n)}}(x^n, Z^n)\right]\,. \tag{91}$$

The order $k_n(z^n)$ provides a 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 one we would have achieved had we been able to access the best order $k^*(x^n, z^n)$ (namely, the "regret"). Notice that 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. Unfortunately, $k_n(z^n)$ may be difficult to compute, and in the next subsection we consider heuristics for selecting k, or more generally, an appropriately sized context model, in practice.

6-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 distortion of the denoised sequence $\hat{X}^{n,k}(z^n)$ relative to the original x^n, is consistently close to the value that makes $\hat{X}^{n,k}(z^n)$ most compressible. 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. Figure 1 shows (suitably scaled) typical plots of compressed code length and distortion of the denoised signal as a function of k, corresponding to one of the data sets reported on in Section 7. All data sets mentioned in Section 7 actually exhibit

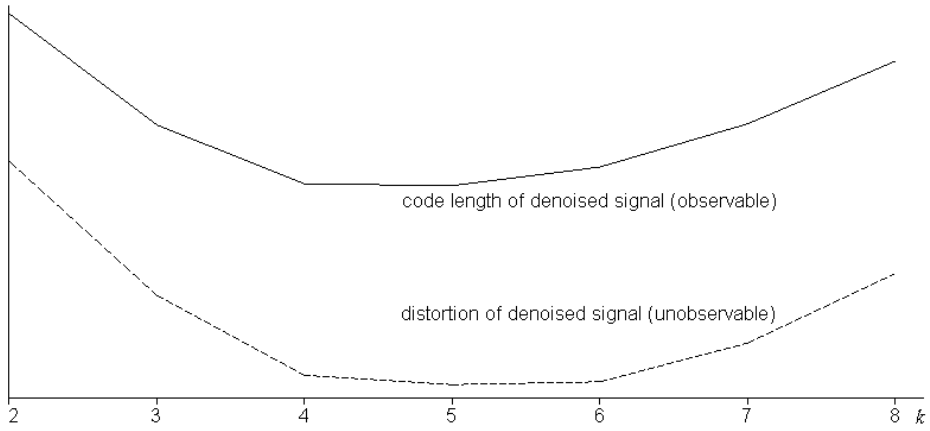

Figure 1: Code length and distortion of denoised signal as a function of k

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 above 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., [51, 58]). Moreover, the context length need not be equal on the left and on the right. As mentioned in Section 3, 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 4. Similar issues of *model cost* [37] have been addressed in related areas of lossless image compression (see, for instance, [6]), and significant knowledge and experience have been generated, which can be brought to bear on the discrete denoising problem. Finally, we mention that if a general tree model is used for the count statistics, the logarithm of its size (number of states) can be used in lieu of k for the compressibility heuristics mentioned above.

7 Experimental Results and Practical Considerations

In this section, we report on experimental results obtained by implementation of the DUDE and its application to a few noise-corrupted data sets.

7-A Binary Symmetric Markov Source Corrupted by a BSC

We implemented the DUDE for the BSC, as derived in subsection 3-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

	$\delta = 0.01$		$\delta = 0.10$		$\delta = 0.20$	
p	DUDE	Bayes	DUDE	Bayes	DUDE	Bayes
0.01	0.000723 [3]	0.000721	0.006648 [5]	0.005746	0.025301 [6]	0.016447
0.05	0.004223 [3]	0.004203	0.030084 [5]	0.029725	0.074936 [5]	0.071511
0.10	0.010213 [8]	0.010020	0.055976 [3]	0.055741	0.120420 [4]	0.118661
0.15	0.010169 [8]	0.010050	0.075474 [5]	0.075234	0.153182 [4]	0.152903
0.20	0.009994 [8]	0.009940	0.092304 [3]	0.092304	0.176354 [4]	0.176135

Table 1: Denoising a BSC-corrupted Markov source

crossover probability δ associated with the BSC, $\{0.01, 0.1, 0.2\}$ (one realization for each). A moderate sequence length of $n = 10^6$ was used.

Table 1 shows the bit error rate of the denoised signal obtained when employing the DUDE for the fifteen 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 6-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 [8, 1].

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 DUDE's performance 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 for a close enough convergence to the optimum for these larger values of k.

7-B Denoising of Don Quixote

We employed the DUDE on a corrupted version of the book *Don Quixote of La Mancha* (English translation), by Miguel de Cervantes Saavedra (1547-1616). The text, available online from the Project Gutenberg web-site at `http://promo.net/pg/`, 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 89087. The DUDE, employed with $k = 2$, reduced the number of errors to 50250, 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 worship," 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 adventures; 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 worship," 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 adventures; 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 ...
   ```

7-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 order of 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 t-th integer pair in the order. For an integer $K \ge 0$, the K-th 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). The sequence of context patterns used in the one-dimensional case can be seen to follow the same rules, except that the context size K is restricted to even values, i.e., $K = 2k$.

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 the channel parameter δ is given by[9]

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

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 6-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 [44], available in the publications data base of the IEEE Information Theory Society. The results are shown in the upper portion of Table 2, 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 [23], and a morphological filter [47] 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. Figure 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).

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

		Channel parameter δ			
Image	Scheme	0.01	0.02	0.05	0.10
Shannon	DUDE	0.00096	0.0018	0.0041	0.0091
1800×2160		K=11	K=12	K=12	K=12
	median	0.00483	0.0057	0.0082	0.0141
	morpho.	0.00270	0.0039	0.0081	0.0161
Einstein	DUDE	0.0035	0.0075	0.0181	0.0391
896×1160		K=18	K=14†	K=12†	K=12†
	median	0.156	0.158	0.164	0.180
	morpho.	0.149	0.151	0.163	0.193

Table 2: Denoising results for binary images

The second image reported on is a half-toned portrait of a famous physicist. While it is arguable whether denoising of half-tone 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 2, which shows that the DUDE is able to achieve significant denoising of the half-tone. In contrast, the more traditional algorithms fail, and, in fact, significantly amplify the distortion. Portions of the noiseless, noisy, and DUDE-denoised half-tone images for the experiment with $\delta = 0.02$ are shown in Figure 3. The experiments on half-tones 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 half-toned photograph, a very different type of image.

7-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

†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.

top-right: original

bottom-left: noisy, δ=0.05

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

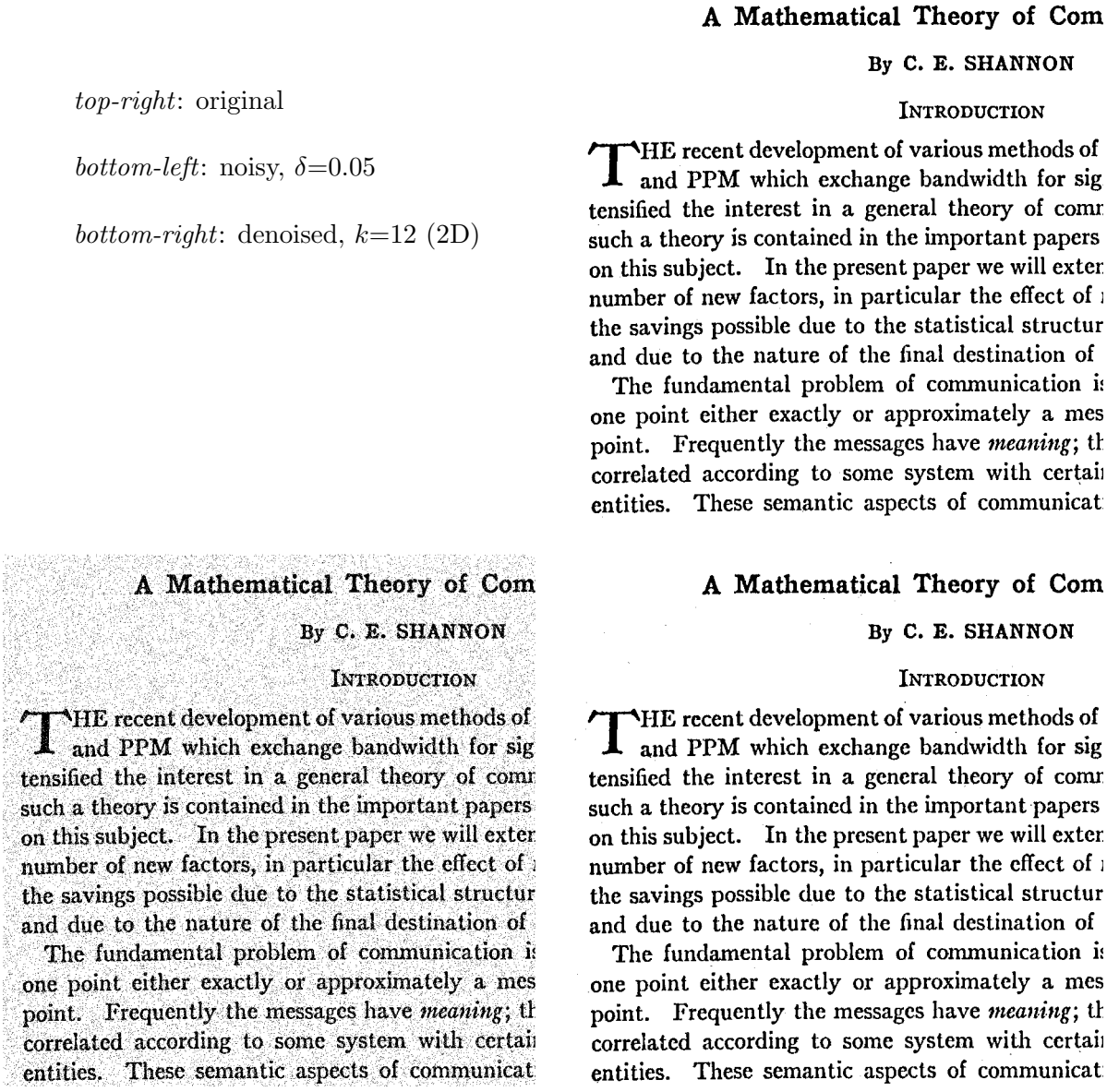

Figure 2: Denoising of a scanned text image

top-right: original

bottom-left: noisy, δ=0.02

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

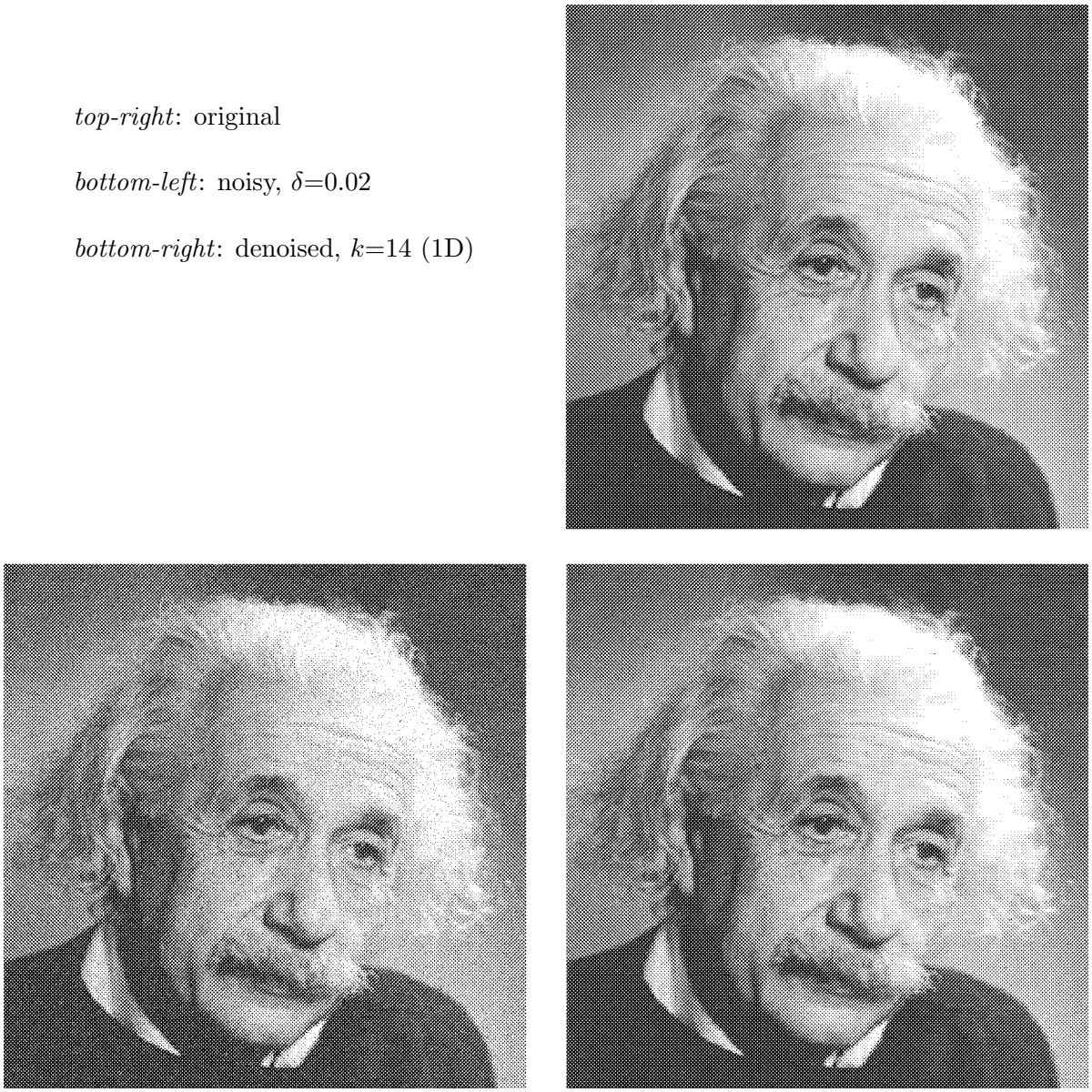

Figure 3: Denoising of a binary halftone image

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 [51, 58] mentioned in Section 6-B. In particular, context aggregation need not reduce the size of the context model.

- **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 loss), it lends itself to non-stationarity. 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.

8 Related Directions

We have presented an asymptotically optimal, universal, low-complexity denoiser for the finite-alphabet case, where the noiseless data is corrupted with a known channel that is invertible in the sense that its associated matrix has full rank.

We next outline a few directions for related research that are currently under investigation.

The setting emphasized in this work, where little (or nothing) is known about the noiseless signal yet the noisy channel is known, arises naturally in many applications. In many other cases, however, there is also uncertainty in the characteristics of the noisy channel. The results attainable in the latter setting are of a basically different nature due to the fact that knowledge of the channel output distribution may not uniquely determine the joint input-output distribution when there is channel uncertainty. When the optimal distribution-dependent denoisers corresponding to the possible input-output distributions consistent with the observations do not coincide, it will not be possible to attain the distribution-dependent optimal performance. For a simple illustration of this fact, consider the case of a BSC with crossover probability δ only known to lie in $[0, 1/2]$. There is, of course, no way to distinguish between, say, the all-zero input signal with $\delta = 1/4$ and the i.i.d. Bernoulli(1/4) signal with a noise-free channel. Since the optimal denoisers corresponding to the two possibilities are completely different, there exists no scheme which will universally attain optimal distribution-dependent performance for all $\delta \in [0, 1/2]$ and all stationary ergodic input processes. Thus, in general, in the case of channel uncertainty, if one's goal is, say, to attain optimal distribution-dependent performance for the stochastic setting, the class of

allowable sources must be more limited than the class of all stationary ergodic processes. On the other hand, we saw in Section 7 that accurate estimation of the channel parameters is feasible in practical applications, the reason likely to be that real-life data sources are more limited and structured than, say, the class of stationary ergodic sources. A related setting is the case of a channel whose associated matrix is not of full rank. The similarity with the case of channel uncertainty is in that, here too, the channel output distribution does not uniquely determine its input distribution. A fuller characterization of these settings is, therefore, an open problem of both theoretical and practical interest.

The approach underlying the construction of the DUDE can be applied to the problem of causal and delay-constrained denoising, also referred to in the HMM literature, respectively, as *filtering* and *smoothing*. Specifically, define a *delay-d denoiser* to be a sequence of functions $\hat{\mathbf{X}} = \{\hat{X}[t]\}_{t \geq 1}$, where $\hat{X}[t] : \mathcal{A}^{t+d} \rightarrow \mathcal{A}$. For each point in time, t, the delay-d denoiser outputs a reconstruction for X_t based on observing Z^{t+d}, namely, $\hat{X}(Z^{t+d})[t]$. For positive integers n, k, d, and $\mathbf{a} \in \mathcal{A}^n$, $\mathbf{b} \in \mathcal{A}^k$, $\mathbf{c} \in \mathcal{A}^d$, let now $\mathbf{m}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ and $g_{\mathbf{a}}^{k,d}(\mathbf{b}, \alpha, \mathbf{c})$ denote the obvious extensions of the definitions in (1) and (6) to accommodate the possibility $k \neq d$. Consider now the delay-d denoiser $\hat{\mathbf{X}}^k$ given, for $t > k$, by

$$\hat{X}^k(z^{t+d})[t] = g_{z^{t+d}}^{k,d}(z_{t-k}^{t-1}, z_t, z_{t+1}^{t+d}). \tag{92}$$

It is then plausible to expect that letting $\hat{\mathbf{X}}$ be the delay-d denoiser in (92) with $k = k_t$, where $\{k_t\}$ is a sequence of integers increasing at a slow enough rate, will result in a scheme with optimality and practicality properties analogous to those established in this work for the DUDE. The scheme in this case is sequential, both acquisition of the statistics and the actual denoising being performed in one pass.

An extension of the DUDE approach to the continuous case is also plausible. Though the technicalities would be somewhat more involved and the context-based approach to estimating the finite-order distributions would have to be replaced by an appropriate density estimator, it seems that the essential feature of having to invert (or de-convolute) the output distribution of just one channel output would remain. It would be of interest to see whether this approach could give rise to a practical scheme and to compare its performance with that of Donoho and Johnstone's wavelet thresholding techniques [15].

Additional interesting cases related to the setting of this work, but not covered by it, include: Channels with memory; channels with deletions, insertions and transpositions; and rate-constrained denoising.

Appendix

A 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 $a \in \mathcal{A}$,

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

where the last equality follows from the fact that, by the definition (72), the only dependence of $\mathbf{q}'(b)[a']$ on b is due to the factor $\boldsymbol{\Pi}(a', b)$, and from the identity $\sum_b \boldsymbol{\Pi}(a', b)\boldsymbol{\Pi}^{-1}(b, a) = \mathbf{1}_{a=a'}$. Thus,

$$\begin{aligned}
\left\|\mathbf{q}(u_0) - \boldsymbol{\pi}_{u_0} \odot (\boldsymbol{\Pi}^{-T}\mathbf{m})\right\|_1 &= \sum_{a \in \mathcal{A}} \left| \mathbf{q}(u_0)[a] - \mathbf{q}'(u_0)[a] - \boldsymbol{\Pi}(a, u_0) \sum_{a', b \in \mathcal{A}} \boldsymbol{\Pi}^{-1}(b, a) \left[\mathbf{q}(b)[a'] - \mathbf{q}'(b)[a']\right] \right| \\
&\leq \left\|\mathbf{q}(u_0) - \mathbf{q}'(u_0)\right\|_1 + \sum_{a, b \in \mathcal{A}} \boldsymbol{\Pi}(a, u_0) \left|\boldsymbol{\Pi}^{-1}(b, a)\right| \left\|\mathbf{q}(b) - \mathbf{q}'(b)\right\|_1 . \qquad \text{(A.2)}
\end{aligned}$$

Summing (A.2) over u_0 yields

$$\begin{aligned}
\sum_{u_0 \in \mathcal{A}} \left\|\mathbf{q}(u_0) - \boldsymbol{\pi}_{u_0} \odot (\boldsymbol{\Pi}^{-T}\mathbf{m})\right\|_1 &\leq \sum_{u_0 \in \mathcal{A}} \left\|\mathbf{q}(u_0) - \mathbf{q}'(u_0)\right\|_1 + \sum_{b \in \mathcal{A}} \left[\left\|\mathbf{q}(b) - \mathbf{q}'(b)\right\|_1 \sum_{a \in \mathcal{A}} \left|\boldsymbol{\Pi}^{-1}(b, a)\right| \right] \\
&\leq (1 + \left\|\boldsymbol{\Pi}^{-1}\right\|_\infty) \sum_{u_0 \in \mathcal{A}} \left\|\mathbf{q}(u_0) - \mathbf{q}'(u_0)\right\|_1 \qquad \text{(A.3)}
\end{aligned}$$

where (A.3) follows from the definition (Section 2):

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

□

B Proof of Claim 2

Proof of Lemma 4: First, we recall that, as a direct consequence of its definition, the Bayes envelope $U(\cdot)$ is a concave function. Specifically, for two M-vectors $\mathbf{u}$ and $\mathbf{v}$, and $\alpha \in [0,1]$,

$$\begin{aligned} U(\alpha\mathbf{u}+(1-\alpha)\mathbf{v}) &= \min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T[\alpha\mathbf{u}+(1-\alpha)\mathbf{v}] \\ &\geq \alpha \min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T\mathbf{u} + (1-\alpha)\min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T\mathbf{v} = \alpha U(\mathbf{u}) + (1-\alpha)U(\mathbf{v})\,. \end{aligned} \tag{A.4}$$

Next, to show that $EU(\mathbf{P}_{X_0|Z_{-k}^l})$ decreases with l, observe that

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

where the inequality follows by concavity. The fact that $EU(\mathbf{P}_{X_0|Z_{-k}^l})$ decreases with k is established similarly, concluding the proof of the first item. For the second item note that, by martingale convergence (cf., in particular, [4, 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 $U(\mathbf{P}) \leq \boldsymbol{\Lambda}_{\max}$ for all $\mathbf{P}\in\mathcal{M}$,

$$EU(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}}) = E\lim_{n\to\infty} U(\mathbf{P}_{X_0|Z_{-k_n}^{l_n}}) = \lim_{n\to\infty} EU(\mathbf{P}_{X_0|Z_{-k_n}^{l_n}}),$$

the second equality following by bounded convergence. □

Proof of Claim 2: We have

$$\begin{aligned} \mathbb{D}(\mathbf{P}_{X^n},\boldsymbol{\Pi}) &= \min_{\hat{X}^n\in\mathcal{D}_n} EL_{\hat{X}^n}(X^n,Z^n) = \frac{1}{n}\sum_{i=1}^n \min_{\hat{X}:\mathcal{A}^n\to\mathcal{A}} E\Lambda(X_i,\hat{X}(Z^n)) \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{z^n\in\mathcal{A}^n} P(Z^n=z^n)\min_{\hat{x}\in\mathcal{A}} E[\Lambda(X_i,\hat{x})|Z^n=z^n] \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{z^n\in\mathcal{A}^n} P(Z^n=z^n)U(\mathbf{P}_{X_i|Z^n=z^n}) \\ &= \frac{1}{n}\sum_{i=1}^n EU(\mathbf{P}_{X_i|Z^n}) = \frac{1}{n}\sum_{i=1}^n EU(\mathbf{P}_{X_0|Z_{1-i}^{n-i}}), \end{aligned} \tag{A.6}$$

where the last equality follows by stationarity. Since, by Lemma 4, $EU(\mathbf{P}_{X_0|Z_{1-i}^{n-i}}) \geq EU(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})$, it follows from (A.6) that $\mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) \geq EU(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})$ for all n and, therefore, $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \geq EU(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})$. On the other hand, for any k, $0 \leq k \leq n$, Lemma 4 and (A.6) yield the upper bound

$$\begin{aligned}
\mathbb{D}(\mathbf{P}_{X^n}, \mathbf{\Pi}) &\leq \frac{1}{n}\left[2kU(\mathbf{P}_{X_0}) + \sum_{i=k+1}^{n-k} EU(\mathbf{P}_{X_0|Z_{1-i}^{n-i}})\right] \\
&\leq \frac{1}{n}\left[2kU(\mathbf{P}_{X_0}) + \sum_{i=k+1}^{n-k} EU(\mathbf{P}_{X_0|Z_{-k}^{k}})\right] \\
&= \frac{1}{n}\left[2kU(\mathbf{P}_{X_0}) + (n-2k)EU(\mathbf{P}_{X_0|Z_{-k}^{k}})\right]. \qquad \text{(A.7)}
\end{aligned}$$

Considering the limit as $n \to \infty$ of both ends of the above chain yields $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \leq EU(\mathbf{P}_{X_0|Z_{-k}^{k}})$. Letting now $k \to \infty$ and invoking Lemma 4 implies $\mathbb{D}(\mathbf{P}_{\mathbf{X}}, \mathbf{\Pi}) \leq EU(\mathbf{P}_{X_0|Z_{-\infty}^{\infty}})$. □

References

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

[2] T. Berger and J. D. Gibson. Lossy source coding. *IEEE Trans. Inform. Theory*, 44(6):2693–2723, October 1998.

[3] H. W. Bode and C. E. Shannon. A simplified derivation of linear least-squares smoothing and prediction theory. *Proceedings IRE*, 38:417–425, 1950.

[4] L. Breiman. *Probability*. SIAM, Philadelphia, 1992.

[5] A. Bruce, D. L. Donoho, and H. Y. Gao. Wavelet analysis. *IEEE Spectrum*, pages 26–35, October 1996.

[6] B. Carpentieri, M.J. Weinberger, and G. Seroussi. Lossless compression of continuous-tone images. *Proceedings of the IEEE*, 88:1797–1809, 2000.

[7] G. Chang, B. Yu, and M. Vetterli. Bridging compression to wavelet thresholding as a denoising method. *Proc. Conf. Inf. Sciences and Systems*, 1997.

[8] R. W. Chang and J. C. Hancock. On receiver structures for channels having memory. *IEEE Trans. Inform. Theory*, 12:463–468, October 1966.

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

[10] F. J. Damerau and E. Mays. An examination of undetected typing errors. *Information Processing and Management: an International Journal*, 25(6):659–664, 1989.

[11] A. Dembo and T. Weissman. The minimax distortion redundancy in noisy source coding. *Technical Report, Dept. of Statistics, Stanford University*, (2002-22), June 2002. Also submitted to IEEE Trans. Inform. Theory.

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

[13] J.L. Devore. A note on the observation of a Markov source through a noisy channel. *IEEE Trans. Inform. Theory*, 20:762–764, November 1974.

[14] D. L. Donoho. De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*, 41(3):613–627, May 1995.

[15] D. L. Donoho and I. M. Johnstone. Wavelet shrinkage: Asymptopia? *J. R. Statist. Soc.*, 57(2):301–369, 1995.

[16] D.L. Donoho. The Kolmogorov sampler. January 2002. (manuscript available at: `http://www-stat.stanford.edu/donoho/`).

[17] R. Durret. *Probability: Theory and Examples*. Duxbury Press, Belmont, California, 1991.

[18] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, 48(6):1518–1569, June 2002.

[19] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. Inform. Theory*, 38:1258–1270, July 1992.

[20] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, 1968.

[21] M. Gastpar, B. Rimoldi, and M. Vetterli. To code or not to code. *Proc. IEEE Int. Symp. Info. Theory*, page 236, June 2000. Submitted.

[22] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Ed. Johns Hopkins, third edition, 1996.

[23] R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Addison Wesley, New York, 1992.

[24] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, III:97–139, 1957. Princeton, NJ.

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

[26] R. Jörnstein and B. Yu. Adaptive quantization. *Technical Report, Department of Statistics*, 2001. UC Berkeley.

[27] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

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

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

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

[31] N. Merhav and M. Feder. Universal prediction. *IEEE Trans. Inform. Theory*, 44(6):2124–2147, October 1998.

[32] B. Natarajan. Filtering random noise via data compression. *Data Compression Conference, DCC '93*, pages 60–69, 1993.

[33] B. Natarajan. Filtering random noise from deterministic signals via data compression. *IEEE Trans. Signal Proc.*, 43(11):2595–2605, November 1995.

[34] B. Natarajan, K. Konstantinides, and C. Herley. Occam filters for stochastic sources with application to digital images. *IEEE Trans. Signal Proc.*, 46:1434–1438, November 1998.

[35] E. Ordentlich, G. Seroussi, S. Verdú, M.J. Weinberger, and T. Weissman. A universal discrete image denoiser and its application to binary images. *Submitted to IEEE International Conference on Image Processing*, 2003.

[36] E. Ordentlich and M. J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. Technical report, Hewlett-Packard Laboratories, 2003.

[37] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1998.

[38] J. Rissanen. MDL denoising. *IEEE Trans. Inform. Theory*, 46:2537–2543, November 2000.

[39] J. Van Ryzin. The sequential compound decision problem with $m \times n$ finite loss matrix. *Ann. Math. Statist.*, 37:954–975, 1966.

[40] E. Samuel. Asymptotic solutions of the sequential compound decision problem. *Ann. Math. Statist.*, pages 1079–1095, 1963.

[41] E. Samuel. An empirical Bayes approach to the testing of certain parametric hypotheses. *Ann. Math. Statist.*, 34(4):1370–1385, 1963.

[42] E. Samuel. Convergence of the losses of certain decision rules for the sequential compound decision problem. *Ann. Math. Statist.*, pages 1606–1621, 1964.

[43] E. Samuel. On simple rules for the compound decision problem. *J. Royal Stat. Society*, series B-27:238–244, 1965.

[44] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, (27):379–423, 623–656, 1948.

[45] J. W. Shavlik. Case-based reasoning with noisy case boundaries: An application in molecular biology. Technical Report CS-TR-1990-988, 1990.

[46] L. Shue, B.D.O. Anderson, and F. De Bruyne. Asymptotic smoothing error for hidden Markov models. *IEEE Trans. on Signal Processing*, 12(48), 2000.

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

[48] I. Tabus, J. Rissanen, and J. Astola. Classification and feature gene selection using the normalized maximum likelihood model for discrete regression. Submitted April 2002 to Signal Processing, Special issue on genomic signal processing.

[49] I. Tabus, J. Rissanen, and J. Astola. Normalized maximum likelihood models for boolean regression with application to prediction and classification in genomics. *Computational and Statistical Approaches to Genomics*, March 2002. (W. Zhang and I. Shmulevich, eds.), Kluwer Acadmic Publishers, Boston.

[50] S. B. Vardeman. Admissible solutions of k-extended finite state set and the sequence compound decision problems. *J. Multiv. Anal.*, 10:426–441, 1980.

[51] M. J. Weinberger, J. J. Rissanen, and M. Feder. A universal finite-memory source. *IEEE Trans. Inform. Theory*, 41:643–652, May 1995.

[52] T. Weissman. Universally attainable error-exponents for rate-constrained denoising of noisy sources. *HP Laboratories Technical Report*, HPL-2002-214, August 2002. Also submitted to IEEE Trans. Inform. Theory.

[53] T. Weissman and N. Merhav. Universal prediction of individual binary sequences in the presence of noise. *IEEE Trans. Inform. Theory*, 47(6):2151–2173, September 2001.

[54] T. Weissman and N. Merhav. Finite-delay lossy coding and filtering of individual sequences corrupted by noise. *IEEE Trans. Inform. Theory*, 48(3):721–733, March 2002.

[55] T. Weissman and N. Merhav. Universal prediction of random binary sequences in a noisy environment. *Ann. App. Prob.*, 2003. To Appear.

[56] 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. Inform. Theory*, 47(5):1849–1866, July 2001.

[57] N. Wiener. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. John Wiley & Sons, New York, N.Y., 1949.

[58] F. M. J. Willems, Y. M. Shtarkov, and T.J. Tjalkens. The context-tree weighting method: Basic properties. *IEEE Trans. Inform. Theory*, 41:653–664, May 1995.

[59] J. Ziv. Distortion-rate theory for individual sequences. *IEEE Trans. Inform. Theory*, 26(2):137–143, March 1980.

[60] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, 24(5):530–536, September 1978.