Discrete Denoising With Shifts

Taesup Moon, *Member, IEEE*, and Tsachy Weissman, *Senior Member, IEEE*

Abstract—We introduce S-DUDE, a new algorithm for denoising discrete memoryless channel (DMC)-corrupted data. The algorithm, which generalizes the recently introduced DUDE (Discrete Universal DEnoiser), aims to compete with a genie that has access, in addition to the noisy data, also to the underlying clean data, and that can choose to switch, up to m times, between sliding-window denoisers in a way that minimizes the overall loss. When the underlying data form an individual sequence, we show that the S-DUDE performs essentially as well as this genie, provided that m is sublinear in the size of the data. When the clean data are emitted by a piecewise stationary process, we show that the S-DUDE achieves the optimum distribution-dependent performance, provided that the same sublinearity condition is imposed on the number of switches. To further substantiate the universal optimality of the S-DUDE, we show that when the number of switches is allowed to grow linearly with the size of the data, *any* (sequence of) scheme(s) fails to compete in the above sense. Using dynamic programming, we derive an efficient implementation of the S-DUDE, which has complexity (time and memory) growing linearly with the data size and the number of switches m. Preliminary experimental results are presented, suggesting that S-DUDE has the capacity to improve on the performance attained by the original DUDE in applications where the nature of the data abruptly changes in time (or space), as is often the case in practice.

Index Terms—Competitive analysis, discrete denoising, discrete memoryless channel (DMC), dynamic programming, forward-backward recursions, individual sequence, piecewise stationary processes, switching experts, universal algorithms.

I. INTRODUCTION

DISCRETE denoising is the problem of reconstructing the components of a finite-alphabet sequence based on the *entire* observation of its discrete memoryless channel (DMC)-corrupted version. The quality of the reconstruction is evaluated via a user-specified (single-letter) loss function. *Universal* discrete denoising, in which no statistical or other properties are known *a priori* about the underlying clean data and the goal is to attain optimum performance, was considered and solved in [1] with the assumption of known DMC. The main problem

T. Moon was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. He is now with Yahoo! Labs, Sunnyvale, CA 95054 USA (e-mail: tsmoon@ymail.com).

T. Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA and with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: tsachy@stanford.edu).

setting there is the "semi-stochastic" one, in which the underlying signal is assumed to be an "individual sequence," and the randomness is due solely to the channel noise. In this setting, it is unreasonable to expect to attain the best performance among all the denoisers in the world, since for every given sequence, there exists a denoiser that recovers all the sequence components perfectly. Thus, [1] limits the comparison class, also known as (a.k.a.) expert class, and uses the competitive analysis approach. Specifically, it is shown that regardless of what the underlying individual sequence may be, the Discrete Universal DEnoiser (DUDE) essentially attains the performance of the best sliding-window denoiser that would be chosen by a genie with access to the underlying clean sequence, in addition to the observed noisy sequence. This semi-stochastic setting result is shown in [1] to imply the stochastic setting result, i.e., that for any underlying stationary signal, the DUDE attains the optimal distribution-dependent performance.

The setting of an arbitrary individual sequence, combined with competitive analysis, has been very popular in many other research areas, especially for problems of sequential decision making. Examples include universal compression [2], universal prediction [3], universal filtering [4], [5], repeated game playing [6]–[8], universal portfolios [9], online learning [10], [11], zero-delay coding [12], [13], and much more. A comprehensive account of this line of research can be found in [14]. The beauty of this approach is the fact that it leads to the construction of schemes that perform, on every individual sequence, essentially as well as the best in a class of experts, which is the performance of a genie that had hindsight on the entire sequence before selecting its actions. Moreover, if the expert class is judiciously chosen, the relative sense of such a performance results can, in many cases, imply optimum performances in absolute senses as well.

One extension to this approach is competition with an expert class and a genie that has the freedom to form a *compound* action, which breaks the sequence into a certain (limited) number of segments, applies different experts in each segment, and achieves an even better performance overall. Note that the optimal segmentation of the sequence and the choice of the best expert in each segment is also determined by hindsight. Clearly, competing with the best compound action is more challenging, since the number of possible compound actions is exponential in the sequence length n, and the brute-force vanilla implementation of the ordinary universal scheme requires prohibitive complexity. However, clever schemes with linear complexity that successfully track the best segments and experts have been devised in many different areas, such as online learning, universal prediction [15], [16], universal compression [17], [18], online linear regression [19], universal portfolios [20], and zero-delay lossy source coding [21].

In this paper, we expand the idea of compound actions and apply it to the discrete denoising problem. The motivation of this expansion is natural: the characteristics of the underlying data in the denoising problem often tend to be time- or space-varying. In this case, determining the best segmentation and the best expert for each segment requires complete knowledge of both clean and noisy sequences. Therefore, whereas the challenge in sequential decision-making problems is to track the shift of the best expert based on the past, true observation, the challenge in the denoising problem is to learn the shift based on the entire, but noisy, observation. We extend DUDE to meet this challenge and provide results that parallel and strengthen those of [1].

Specifically, we introduce the S-DUDE and show first that, for every underlying noiseless sequence, it attains the performance of the best compound finite-order sliding-window denoiser (concretely defined later), both in expectation and in a high probability sense. We develop our scheme in the semi-stochastic setting as in [1]. The toolbox for the construction and analysis of our scheme draws on ideas developed in [4]. We circumvent the difficulty of not knowing the exact true loss by using an observable unbiased estimate of it. This kind of an estimate has proved to be very useful in [4] and [22] to devise schemes for filtering and for denoising with dynamic contexts. Building on this semi-stochastic setting result, we also establish a stochastic setting result, which can be thought of as a generalization and strengthening of the stochastic setting results of [1], from the world of stationary processes to that of piecewise stationary processes.

Our stochastic setting has connections to other areas, such as change-point detection problems in statistics [23], [24] and switching linear dynamical systems in machine learning and signal processing [25], [26]. Both of these lines of research share a common approach with S-DUDE, in that they try to learn the change of the underlying time-varying parameter or state of stochastic models, based on noisy observations of the parameter or state. One difference is that, whereas our goal is the noncausal estimation, i.e., denoising of the general underlying piecewise stationary process, the change-point detection problems mainly focus on sequentially detecting the time point where the change of model happened. Another difference is in that the switching linear dynamical systems focus on a special class of underlying processes, the linear dynamical system. In addition, they deal with continuous-valued signals, whereas our focus is the discrete case, with finite-alphabet signals.

As we explain in detail, the S-DUDE can be practically implemented using a two-pass algorithm with complexity (both space and time) linear in the sequence length and the number of switches. We also present initial experimental results that demonstrate the S-DUDE's potential to outperform the DUDE on both simulated and real data.

The remainder of the paper is organized as follows. Section II provides the notation, preliminaries, and motivation for the paper; in Section III, we present our scheme and establish its strong universality properties via an analysis of its performance in the semi-stochastic setting. Section IV establishes the universality of our scheme in a fully stochastic setting, where the underlying noiseless sequence is emitted by a piecewise stationary process. Algorithmic aspects and complexity of the

actual implementation of the scheme is considered in Section V, and some experimental results are displayed in Section VI. In Section VII, we conclude with a summary of our findings and some possible future research directions.

II. NOTATION, PRELIMINARIES, AND MOTIVATION

A. Notation

We use a combination of notation from [1] and [4]. Let $\mathcal{X}$, $\mathcal{Z}$, $\hat{\mathcal{X}}$ denote, respectively, the alphabet of the clean, noisy, and reconstructed sources, which are assumed finite.[1] As in [1] and [4], the noisy sequence is a DMC-corrupted version of the clean one, where the channel matrix $\mathbf{\Pi} = \{\Pi(x,z)\}_{x \in \mathcal{X}, z \in \mathcal{Z}}$, $\Pi(x,z)$ denoting the probability of a noisy symbol z when the underlying clean symbol is x, is assumed to be *known and fixed* throughout the paper, and *of full row rank*. The zth column of $\mathbf{\Pi}$ will be denoted as π_z. Upper case letters will denote random variables as usual; lower case letters will denote either individual deterministic quantities or specific realizations of random variables.

Without loss of generality, the elements of any finite set $\mathcal{V}$ will be identified with $\{0, 1, \ldots, |\mathcal{V}| - 1\}$. We let $\mathcal{V}^\infty$ denote the set of one-sided infinite sequences with $\mathcal{V}$-valued components, i.e., $\boldsymbol{v} \in \mathcal{V}^\infty$ is of the form $\boldsymbol{v} = (v_1, v_2, \ldots), v_i \in \mathcal{V}, i \geq 1$. For $\boldsymbol{v} \in \mathcal{V}^\infty$, let $v^n = (v_1, \ldots, v_n)$ and $v_m^n = (v_m, \ldots, v_n)$. Furthermore, we let $v^{n \backslash t}$ denote the sequence $v^{t-1} v_{t+1}^n$. $\mathbb{R}^{|\mathcal{V}|}$ is a space of $|\mathcal{V}|$-dimensional column vectors with real-valued components indexed by the elements of $\mathcal{V}$. The ath component of $q \in \mathbb{R}^{|\mathcal{V}|}$ will be denoted by either q_a or $q[a]$. Subscripting a vector or a matrix by "max" will represent the difference between the maximum and minimum of all its components. Thus, for example, if Γ is a $|\mathcal{Z}| \times |\mathcal{X}|$ matrix, then $\Gamma_{\max}$ stands for $\max_{z \in \mathcal{Z}}{}_{x \in \mathcal{X}} \Gamma(z,x) - \min_{z \in \mathcal{Z}, x \in \mathcal{X}} \Gamma(z,x)$ (in particular, if the components of Γ are nonnegative and $\Gamma(z,x) = 0$ for some z and x, then $\Gamma_{\max} = \max_{z \in \mathcal{Z}, z \in \mathcal{X}} \Gamma(z,x)$.) In addition, $\mathbf{1}_{\{\cdot\}}$ denotes an indicator of the event inside $\{\cdot\}$.

Generally, let the finite sets $\mathcal{Y}$, $\mathcal{A}$ be, respectively, a source alphabet and an action space. For a general loss function $l : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, a *Bayes response* for $\zeta \in \mathbb{R}^{|\mathcal{Y}|}$ under the loss function l is given as

$$b_l(\zeta) = \arg\min_{a \in \mathcal{A}} \zeta^T \cdot \mathcal{L}_a \tag{1}$$

where $\mathcal{L}_a$ denotes the column of the matrix of the loss function l corresponding to the ath action, and ties are resolved lexicographically. The corresponding *Bayes envelope* is denoted as

$$U_l(\zeta) = \min_{a \in \mathcal{A}} \zeta^T \cdot \mathcal{L}_a. \tag{2}$$

Note that when ζ is a probability, namely, it has nonnegative components summing to one, $U_l(\zeta)$ is the minimum achievable expected loss (as measured under the loss function l) in guessing the value of $Y \in \mathcal{Y}$ which is distributed according to ζ, with some action in $\mathcal{A}$. The associated optimal action for the guess is $b_l(\zeta)$.

---

[1] To avoid trivialities, we assume the cardinalities of $\mathcal{X}$, $\mathcal{Z}$, $\hat{\mathcal{X}}$ are all greater than 1.

An *n-block denoiser* is a collection of n mappings $\hat{\boldsymbol{X}}^n = \{\hat{X}_t\}_{1 \le t \le n}$, where $\hat{X}_t : \mathcal{Z}^n \to \hat{\mathcal{X}}$. We assume a given loss function $\Lambda : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$, where the maximum single-letter loss is denoted by $\Lambda_{\max}$ and the minimum value of the loss is 0, and $\lambda_{\hat{x}}$ denotes the $\hat{x}$th column of the loss matrix. The normalized cumulative loss of the denoiser $\hat{\boldsymbol{X}}^n$ on the individual sequence pair (x^n, z^n) is represented as

$$L_{\hat{\boldsymbol{X}}^n}(x^n, z^n) = \frac{1}{n} \sum_{t=1}^{n} \Lambda(x_t, \hat{X}_t(z^n)).$$

In words, $L_{\hat{\boldsymbol{X}}^n}(x^n, z^n)$ is the normalized (per-symbol) loss, as measured under the loss function Λ, when using the denoiser $\hat{\boldsymbol{X}}^n$ and when the observed noisy sequence is z^n while the underlying clean one is x^n. The notation $L_{\hat{\boldsymbol{X}}^n}$ is extended for $1 \le i \le j \le n$

$$L_{\hat{\boldsymbol{X}}^n}(x_i^j, z^n) = \frac{1}{j - i + 1} \sum_{t=i}^{j} \Lambda(x_t, \hat{X}_t(z^n))$$

denoting the normalized (per-symbol) loss between (and including) locations i and j.

Now, consider the set $\mathcal{S} = \{s : \mathcal{Z} \to \hat{\mathcal{X}}\}$, which is the (finite) set of mappings that take $\mathcal{Z}$ into $\hat{\mathcal{X}}$. We refer to elements of $\mathcal{S}$ as "single-symbol denoisers," since each $s \in \mathcal{S}$ can be thought of as a rule for estimating $X \in \mathcal{X}$ on the basis of $Z \in \mathcal{Z}$. Now, for any $s \in \mathcal{S}$, an unbiased estimator for $\Lambda(x, s(Z))$ (based on Z only), where x is a deterministic symbol and Z is the output of the DMC when the input is x, can be obtained as in [4]. First, pick a function $h : \mathcal{Z} \to \mathbb{R}^{|\mathcal{X}|}$ with the property that, for $a, b \in \mathcal{X}$

$$E_a h_b(Z) = \sum_{z \in \mathcal{Z}} \Pi(a, z) h_b(z)$$
$$= \delta(a, b) \triangleq \left\{ \begin{array}{ll} 1, & \text{if } a = b \\ 0, & \text{otherwise} \end{array} \right\} \quad (3)$$

where E_a denotes expectation over the channel output Z when the underlying channel input is a, and $h_b(z)$ denotes the bth component of $h(z)$. Let H denote the $|\mathcal{Z}| \times |\mathcal{X}|$ matrix whose zth row is $h^T(z)$, i.e., $H(z, b) = h_b(z)$. To see that our assumption of a channel matrix with full row rank guarantees the existence of such an h, note that (3) can equivalently be stated in matrix form as

$$\Pi H = I \quad (4)$$

where I is the $|\mathcal{X}| \times |\mathcal{X}|$ identity matrix. Thus, e.g., any H of the form $H = \Gamma^T(\Pi\Gamma^T)^{-1}$, for any Γ such that $\Pi\Gamma^T$ is invertible, satisfies (4). In particular, $\Gamma = \Pi$ is a valid choice ($\Pi\Pi^T$ is invertible, since Π is of full row rank) corresponding to the Moore–Penrose generalized inverse [27]. Now, for any $s \in \mathcal{S}$, let $\rho(s) \in \mathbb{R}^{|\mathcal{X}|}$ denote the column vector with xth component

$$\rho_x(s) = \sum_z \Pi(x, z) \Lambda(x, s(z))$$
$$= E_x \Lambda(x, s(Z)). \quad (5)$$

In words, $\rho_x(s)$ is the expected loss using the single-symbol denoiser s, while the underlying symbol is x. Considering $\mathcal{S}$

as an action space alphabet, we define a loss function $\ell : \mathcal{Z} \times \mathcal{S} \to \mathbb{R}$ as

$$\ell(z, s) = h(z)^T \cdot \rho(s). \quad (6)$$

We observe from (3) and (5) that $\ell(Z, s)$ is an unbiased estimate of $\Lambda(x, s(Z))$ since

$$E_x \ell(Z, s) = E_x h(Z)^T \cdot \rho(s) = \sum_{x'} E_x h_{x'}(Z) \rho_{x'}(s)$$
$$= \sum_{x'} \delta(x, x') \rho_{x'}(s) = \rho_x(s)$$
$$= E_x \Lambda(x, s(Z)), \quad \forall x \in \mathcal{X}.$$

Note that $\ell(Z, s)$ only depends on the noisy symbol Z and the single-symbol denoiser s, thus it is observable by the denoiser. As mentioned in the Introduction, this unbiased estimate $\ell(Z, s)$ is the key tool in overcoming the difficulty of not knowing the true loss in devising our new denoiser.

For $\xi \in \mathbb{R}^{|\mathcal{Z}|}$, let $B_H(\xi, \cdot) \in \mathcal{S}$ be defined by

$$B_H(\xi, z) = \arg \min_{\hat{x}} \xi^T \cdot H \cdot [\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_z] \quad (7)$$

where, for vectors v_1 and v_2 of equal dimensions, $v_1 \odot v_2$ denotes the vector obtained by component-wise multiplication. Note that, similarly as in [4, eqs. (88), (89)]

$$B_H(\xi, \cdot) = \arg \min_{s \in \mathcal{S}} \sum_z \xi^T \cdot H \cdot [\boldsymbol{\lambda}_{s(z)} \odot \boldsymbol{\pi}_z]$$
$$= \arg \min_{s \in \mathcal{S}} \xi^T \cdot H \cdot \rho(s)$$
$$= \arg \min_{s \in \mathcal{S}} \sum_z \xi_z \cdot [h^T(z) \cdot \rho(s)]$$
$$= \arg \min_{s \in \mathcal{S}} \sum_z \xi_z \cdot \ell(z, s) = b_\ell(\xi). \quad (8)$$

Thus, $B_H(\xi, \cdot)$ is a Bayes response for ξ under the loss function ℓ defined in (6).

B. Preliminaries

In this subsection, we summarize the results from [1] and motivate the approach underlying the construction of our new class of denoisers. Analogously as in [4], the n-block denoiser $\hat{\boldsymbol{X}}^n = \{\hat{X}_t\}_{1 \le t \le n}$ can be associated with $\boldsymbol{F}^n = \{F_t\}_{1 \le t \le n}$, where $F_t : \mathcal{Z}^{n \backslash t} \to \mathcal{S}$ is defined as follows: $F_t(z^{n \backslash t}, \cdot)$ is the single-symbol denoiser in $\mathcal{S}$ satisfying

$$\hat{X}_t(z^n) = F_t(z^{n \backslash t}, z_t), \quad \forall z_t. \quad (9)$$

Therefore, we can adopt the view that at each time t, an n-block denoiser is choosing a single-symbol denoiser based on all the noisy sequence components but z_t, and applying that single-symbol denoiser on z_t to yield the tth reconstruction $\hat{x}_t$. Conversely, any sequence of mappings into single-symbol denoisers $\boldsymbol{F}^n$ defines a denoiser $\hat{\boldsymbol{X}}^n$, again via (9). We will adhere to this viewpoint in what follows.

One special class of widely used n-block denoisers is that of kth-order "sliding-window" denoisers, which we denote by $\hat{\boldsymbol{X}}^{n,\mathcal{S}_k}$. Such denoisers are of the form

$$\hat{X}_t^{s_k}(z^n) = s_k(z_{t-k}^{t+k}), \quad t = k+1, \ldots, n-k \quad (10)$$

where s_k is an element of $\mathcal{S}_k = \{s_k : \mathcal{Z}^{2k+1} \rightarrow \hat{\mathcal{X}}\}$, the (finite) set of mappings from $\mathcal{Z}^{2k+1}$ into $\hat{\mathcal{X}}$.[2] We also refer to $s_k \in \mathcal{S}_k$ as a "kth-order denoiser." Note that $\mathcal{S}_0 = \mathcal{S}$. From the definition (10), it follows that

$$\hat{X}_i^{s_k}(z^n) = \hat{X}_j^{s_k}(z^n) \text{ whenever } z_{i+k}^{i+k} = z_{j+k}^{j+k}. \quad (11)$$

Following the association in (9), we can adopt the alternative view that the kth-order sliding-window denoiser chooses a single-symbol denoiser $s_k(z_{t-k}^{t-1}, z_{t+1}^{t+k}, \cdot) \in \mathcal{S}$ at time t on the basis of the context, and $\hat{X}_t^{s_k}(z^n) = s_k(z_{t-k}^{t-1}, z_{t+1}^{t+k}, z_t)$.

We denote $\boldsymbol{c}_t \triangleq (z_{t-k}^{t-1}, z_{t+1}^{t+k})$ as a (two-sided) context for z_t, and define the set of all possible kth-order contexts

$$\boldsymbol{C}_k \triangleq \{(u_{-k}^{-1}, u_1^k) : (u_{-k}^{-1}, u_1^k) \in \mathcal{Z}^{2k}\}.$$

Then, for given z^n and for each $\boldsymbol{c} \in \boldsymbol{C}_k$, we define

$$\begin{aligned}\mathcal{T}(\boldsymbol{c}) &\triangleq \{t : \boldsymbol{c}_t = \boldsymbol{c}, \quad k+1 \leq t \leq n-k\} \\ &= \{t : (z_{t-k}^{t-1}, z_{t+1}^{t+k}) = \boldsymbol{c}, \quad k+1 \leq t \leq n-k\}\end{aligned}$$

the set of indices where the context equals $\boldsymbol{c}$. Now, an equivalent interpretation for (11) is that for each $\boldsymbol{c} \in \boldsymbol{C}_k$, the kth-order sliding-window denoiser employs a time-invariant single-symbol denoiser, $s_k(\boldsymbol{c}, \cdot)$, at all points $t \in \mathcal{T}(\boldsymbol{c})$. In other words, the sequence z^n is partitioned into the subsequences associated with the various contexts, and on each such subsequence a time-invariant single-symbol scheme is employed.

In [1], for integers $k \geq 0$ and $n > 2k$, the kth-order minimum loss of (x^n, z^n) is defined by

$$\begin{aligned}D_k(x^n, z^n) &\triangleq \min_{\hat{\boldsymbol{X}}^n \in \hat{\boldsymbol{X}}^{n,\mathcal{S}_k}} L_{\hat{\boldsymbol{X}}^n}(x_{k+1}^{n-k}, z^n) \\ &= \min_{s_k \in \mathcal{S}_k} \frac{1}{n-2k} \sum_{t=k+1}^{n-k} \Lambda(x_t, s_k(\boldsymbol{c}_t, z_t)). \quad (12)\end{aligned}$$

The identity of the element $s_k \in \mathcal{S}_k$ that achieves (12) depends not only on z^n, but also on x^n, since (12) can be expressed as

$$\frac{1}{n-2k} \sum_{\boldsymbol{c} \in \boldsymbol{C}_k} \left[\min_{s \in \mathcal{S}} \sum_{\tau \in \mathcal{T}(\boldsymbol{c})} \Lambda(x_\tau, s(z_\tau)) \right]$$

and at each time t, the best kth-order sliding-window denoiser that achieves (12) will employ the single-symbol denoiser

$$\arg \min_{s \in \mathcal{S}} \sum_{\tau \in \mathcal{T}(\boldsymbol{c}_t)} \Lambda(x_\tau, s(z_\tau)) \quad (13)$$

[2]The value of $\hat{X}_t^{s_k}(z^n)$ for $t \leq k$ and $t > n-k$ is defined, for concreteness and simplicity, as an arbitrary fixed symbol in $\hat{\mathcal{X}}$.

which is determined from the joint empirical distribution of pairs $\{(x_\tau, z_\tau) : \tau \in \mathcal{T}(\boldsymbol{c}_t)\}$.

It was shown in [1] that, despite the lack of knowledge of x^n, $D_k(x^n, Z^n)$ is achievable in a sense made precise below, in the limit of growing n, by a scheme that only has access to Z^n. This scheme is dubbed in [1] as the Discrete Universal DEnoiser (DUDE), $\hat{\boldsymbol{X}}_{\text{univ}}^{n,k} = \{\hat{X}_{\text{univ},t}^k\}_{1 \leq t \leq n}$. The algorithm is defined by

$$\hat{X}_{\text{univ},t}^k(z^n) = B_H(\boldsymbol{m}(z^n, z_{t-k}^{t-1}, z_{t+1}^{t+k}), z_t) \quad (14)$$

where $\boldsymbol{m}(z^n, \boldsymbol{c})$ is the vector of counts of the appearances of the various symbols within the context $\boldsymbol{c}$ along the sequence z^n. That is, for all $\beta \in \mathcal{Z}$, $\boldsymbol{m}(z^n, \tilde{z}_{-k}^{-1}, \tilde{z}_1^k)$ is the $|\mathcal{Z}|$-dimensional column vector whose βth component is

$$\boldsymbol{m}(z^n, \tilde{z}_{-k}^{-1}, \tilde{z}_1^k)[\beta] = \left|\{t : k+1 \leq t \leq n-k, z_{t-k}^{t+k} = \tilde{z}_{-k}^{-1}\beta\tilde{z}_1^k\}\right|$$

namely, the number of appearances of $\tilde{z}_{-k}^{-1}\beta\tilde{z}_1^k$ along the sequence z^n.

The main result of [1] is the following theorem, pertaining to the semi-stochastic setting of an individual sequence $\boldsymbol{x} = (x_1, x_2, \ldots)$ corrupted by a DMC that yields the stochastic noisy sequence $\boldsymbol{Z} = (Z_1, Z_2, \ldots)$.

Theorem 1 ([1, Theorem 1]): Take $k = k_n$ satisfying $k_n|\mathcal{Z}|^{2k_n} = o(n/\log n)$. Then, for all $\boldsymbol{x} \in \mathcal{X}^\infty$, the sequence of denoisers $\{\hat{\boldsymbol{X}}_{\text{univ}}^{n,k_n}\}$ defined in (14) satisfies:
a)

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

b)

$$E\left[L_{\hat{\boldsymbol{X}}_{\text{univ}}^{n,k_n}}(x^n, Z^n) - D_{k_n}(x^n, Z^n) \right] = O\left(\sqrt{\frac{k_n|\mathcal{Z}|^{2k_n}}{n}} \right).$$

Theorem 1 was further shown in [1] to imply the universality of the DUDE in the fully stochastic setting where the underlying sequence is emitted by a stationary source (and the goal is to attain the performance of the optimal distribution-dependent denoiser).

From (14), it is apparent that the DUDE ends up employing a kth-order sliding-window denoiser (where the sliding window scheme the DUDE chooses depends on z^n). Moreover, (8) implies that, at each time t, DUDE is merely employing the single-symbol denoiser $B_H(\boldsymbol{m}(z^n, z_{t-k}^{t-1}, z_{t+1}^{t+k}), \cdot) \in \mathcal{S}$, which can be obtained by finding the Bayes response $b_\ell(\boldsymbol{m}(z^n, z_{t-k}^{t-1}, z_{t+1}^{t+k}))$ or, equivalently, the mapping in $\mathcal{S}$ given by

$$\arg \min_{s \in \mathcal{S}} \sum_{\tau \in \mathcal{T}(\boldsymbol{c}_t)} \ell(z_\tau, s) \quad (15)$$

where $\ell(z, s)$ is the loss function defined in (6). By comparing (13) with (15), and from Theorem 1, we observe that working with the estimated loss $\ell(z_\tau, s)$ in lieu of the genie-aided $\Lambda(x_\tau, s(z_\tau))$ allows us to essentially achieve the genie-aided performance in (12).

C. Motivation

Our motivation for this paper is based on the observation that the kth-order sliding-window denoisers ignore the time-varying nature of the underlying sequence x^n. That is, as discussed above, for time instances with the same contexts, the single-symbol denoiser employed along the associated subsequence is time invariant. In other words, for each t, only the empirical distribution of the sequence $\{(x_\tau, z_\tau) : \tau \in \mathcal{T}(c_t)\}$ matters, but its order of composition, i.e., its time-varying nature, is not considered. It is clear, however, that when the characteristics of the underlying clean sequence x^n are changing, the (normalized) cumulative loss that is achieved by sliding-window denoisers that can shift from one rule to another along the sequence may be strictly lower (better) than (12). We now devise and analyze our new scheme that achieves this more ambitious target performance.

III. THE SHIFTING DENOISER (S-DUDE)

In this section, we derive our new class of denoisers and analyze their performance. In Section III-A, we begin with the simplest case, competing with shifting symbol-by-symbol denoisers, or, in other words, shifting zeroth-order denoisers. The argument is generalized to shifting kth-order denoisers in Section III-B, and the framework and results include Section III-A as a special case. We will use the notation $\mathcal{S}_0$, instead of $\mathcal{S}$, for consistency in denoting the class of single-symbol denoisers. Throughout this section, we assume the semi-stochastic setting.

A. Switching Between Symbol-by-Symbol (Zeroth-Order) Denoisers

Consider an n-tuple of single-symbol denoisers $\boldsymbol{S} = \{s_1, \ldots, s_n\} \in \mathcal{S}_0^n$. Then, as mentioned in Section II-B, for such $\boldsymbol{S}$, we can define the associated n-block denoiser $\hat{\boldsymbol{X}}^{n,\boldsymbol{S}} = \{\hat{X}_t^{\boldsymbol{S}}(z^n)\}_{1 \le t \le n}$ as

$$\hat{X}_t^{\boldsymbol{S}}(z^n) = s_t(z_t). \tag{16}$$

Note that in this case, the single-symbol denoiser applied at each time may depend on the time t (but not on $z^{n \setminus t}$, as would be the case for a general denoiser). We also denote the estimated normalized cumulative loss as

$$\tilde{L}_{\boldsymbol{S}}(z^n) \triangleq \frac{1}{n} \sum_{t=1}^{n} \ell(z_t, s_t) \tag{17}$$

whose property is given in the following lemma, which parallels [4, Theorem 4].

Lemma 1: Fix $\epsilon > 0$. For fixed $\boldsymbol{S} \in \mathcal{S}_0^n$, and all $x^n \in \mathcal{X}^n$

$$P\left(L_{\hat{X}^n, S}(x^n, Z^n) - \tilde{L}_{\boldsymbol{S}}(Z^n) > \epsilon\right)$$
$$\le \exp\left(-n \frac{2\epsilon^2}{L_{\max}^2}\right) \quad \text{and} \tag{18}$$
$$P\left(\tilde{L}_{\boldsymbol{S}}(Z^n) - L_{\hat{X}^n, S}(x^n, Z^n) > \epsilon\right)$$
$$\le \exp\left(-n \frac{2\epsilon^2}{L_{\max}^2}\right) \tag{19}$$

where $L_{\max} = \Lambda_{\max} + \ell_{\max}$.

In words, the lemma shows that for every $\boldsymbol{S} \in \mathcal{S}_0^n$, the estimated loss $\tilde{L}_{\boldsymbol{S}}(Z^n)$ is concentrated around the true loss $L_{\hat{X}^n, S}(x^n, Z^n)$ with high probability, as n becomes large, regardless of the underlying sequence x^n.

Proof of Lemma 1: See the Appendix, Part A.   □

Now, let the integer $0 \le m \le \lfloor \frac{n}{2} \rfloor$ denote the maximum number of shifts allowed along the sequence. Then, define the set $\mathcal{S}_{0,m}^n \subseteq \mathcal{S}_0^n$ as

$$\mathcal{S}_{0,m}^n = \left\{\boldsymbol{S} \in \mathcal{S}_0^n : \sum_{t=2}^{n} \mathbf{1}_{\{s_{t-1} \neq s_t\}} \le m\right\} \tag{20}$$

namely, $\mathcal{S}_{0,m}^n$ is the set of n-tuples of single-symbol denoisers with at most m shifts from one mapping to another.[3] Analogously to (12), for the class of n-block denoisers $\hat{\boldsymbol{X}}^{n,\boldsymbol{S}}$ with $\boldsymbol{S} \in \mathcal{S}_{0,m}^n$, we define

$$D_{0,m}(x^n, z^n) \triangleq \min_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} L_{\hat{X}^n, S}(x^n, z^n)$$
$$= \min_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} \frac{1}{n} \sum_{t=1}^{n} \Lambda(x_t, s_t(z_t)) \tag{21}$$

which is the minimum normalized cumulative loss that can be achieved for (x^n, z^n) by the sequence of n single-symbol denoisers that allow at most m shifts. Our goal in this section is to build a universal scheme that only has access to Z^n, but still essentially achieves $D_{0,m}(x^n, Z^n)$.

As hinted by the DUDE, we build our universal scheme by working with the estimated loss. That is, define

$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{S}}(z^n) \triangleq \arg \min_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} \tilde{L}_{\boldsymbol{S}}(z^n) \tag{22}$$

and our $(0, m)$-Shifting DUDE (S-DUDE), $\hat{\boldsymbol{X}}_{\text{univ}}^{n,0,m}$, is defined as $\hat{\boldsymbol{X}}^{n,\hat{\boldsymbol{S}}}$. It is clear that, by definition, $L_{\hat{X}^n, \hat{S}}(x^n, z^n) \ge D_{0,m}(x^n, z^n)$ for all x^n and z^n, but we can also show that, with high probability, $L_{\hat{X}^n, \hat{S}}(x^n, Z^n)$ does not exceed $D_{0,m}(x^n, Z^n)$ by much, as stated in the following theorem.

Theorem 2: Let $\hat{\boldsymbol{X}}_{\text{univ}}^{n,0,m}$ be defined as $\hat{\boldsymbol{X}}^{n,\hat{\boldsymbol{S}}}$, where $\hat{\boldsymbol{S}}$ is given in (22). Then, for all $\epsilon > 0$ and $x^n \in \mathcal{X}^n$

$$P\left(L_{\hat{X}_{\text{univ}}^{n,0,m}}(x^n, Z^n) - D_{0,m}(x^n, Z^n) > \epsilon\right)$$
$$\le 2 \exp\left(-n\left[\frac{\epsilon^2}{2L_{\max}^2} - 2\left\{h\left(\frac{m}{n}\right) + \frac{(m+1)\ln N}{n}\right\}\right]\right)$$

where $h(x) = -x \ln x - (1-x) \ln(1-x)$ for $0 \le x \le 1$, and $N = |\mathcal{S}| = |\hat{\mathcal{X}}|^{|\mathcal{Z}|}$. In particular, the right-hand side of the inequality is exponentially small, provided $m = o(n)$.

Remark: It is reasonable to expect this theorem to hold, given Lemma 1. That is, since, for fixed $\boldsymbol{S} \in \mathcal{S}_{0,m}^n$, $\tilde{L}_{\boldsymbol{S}}(Z^n)$ is concentrated on $L_{\hat{X}^n, S}(x^n, Z^n)$, it is plausible that the $\hat{\boldsymbol{S}}$ that achieves $\min_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} \tilde{L}_{\boldsymbol{S}}(Z^n)$ will have a loss $L_{\hat{X}^n, \hat{S}}(x^n, Z^n)$ close to $\min_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} L_{\hat{X}^n, S}(x^n, Z^n)$, i.e., $D_{0,m}(x^n, Z^n)$.

Proof of Theorem 2: See the Appendix, Part B.   □

---

[3]Note that, when $m = 0$, $\mathcal{S}_{0,0}^n$ is the set of constant n-tuples consisting of the same single-symbol denoiser.

B. Switching Between kth-Order Denoisers

Now, we extend the result from Section III-A to the case of shifting between kth-order denoisers. The argument parallels that of Section III-A. Let $\{s_{k,t}\}_{t=k+1}^{n-k}$ be an arbitrary sequence of the kth-order denoiser mappings, i.e., $s_{k,t} \in \mathcal{S}_k$ for $k+1 \leq t \leq n-k$. Now, for given z^n, define an $(n-2k)$-tuple of (kth-order denoiser induced) single-symbol denoisers

$$\boldsymbol{S}_k(z^n) \triangleq \{s_{k,t}(\boldsymbol{c}_t, \cdot)\}_{t=k+1}^{n-k} \in \mathcal{S}_0^{n-2k} \qquad (23)$$

where, to recall, $\boldsymbol{c}_t = (z_{t-k}^{t-1}, z_{t+1}^{t+k})$, and $s_{k,t}(\boldsymbol{c}_t, \cdot)$ is the single-symbol denoiser induced from $s_{k,t} \in \mathcal{S}_k$ and $\boldsymbol{c}_t$. For brevity of notation, we will suppress the dependence on z^n in $\boldsymbol{S}_k(z^n)$ and denote it as $\boldsymbol{S}_k$. Then, as in (16), we define the associated n-block denoiser $\hat{\boldsymbol{X}}^{n,\boldsymbol{S}_k} = \{\hat{X}_t^{\boldsymbol{S}_k}(z^n)\}_{1 \leq t \leq n}$ as[4]

$$\hat{X}_t^{\boldsymbol{S}_k}(z^n) = s_{k,t}(\boldsymbol{c}_t, z_t). \qquad (24)$$

In addition, extending (17), the estimated normalized cumulative loss is given as

$$\tilde{L}_{\boldsymbol{S}_k}(z^n) = \frac{1}{n-2k} \sum_{t=k+1}^{n-k} \ell(z_t, s_{k,t}(\boldsymbol{c}_t, \cdot)). \qquad (25)$$

Then, we have the following lemma, which parallels Lemma 1.

Lemma 2: Fix $\epsilon > 0$. For any fixed sequence $\{s_{k,t}\}_{t=k+1}^{n-k}$, and all $x^n \in \mathcal{X}^n$

$$\Pr\left(L_{\hat{\boldsymbol{X}}^n, s_k}(x_{k+1}^{n-k}, Z^n) - \tilde{L}_{\boldsymbol{S}_k}(Z^n) > \epsilon \right)$$
$$\leq (k+1) \exp\left(-\frac{2(n-2k)\epsilon^2}{(k+1)L_{\max}^2} \right) \quad \text{and} \quad (26)$$

$$\Pr\left(\tilde{L}_{\boldsymbol{S}_k}(Z^n) - L_{\hat{\boldsymbol{X}}^n, s_k}(x_{k+1}^{n-k}, Z^n) > \epsilon \right)$$
$$\leq (k+1) \exp\left(-\frac{2(n-2k)\epsilon^2}{(k+1)L_{\max}^2} \right) \qquad (27)$$

where $L_{\max} = \Lambda_{\max} + \ell_{\max}$.

Remark: Note that when $k = 0$, this lemma coincides with Lemma 1. The proof of this lemma combines Lemma 1 and the de-interleaving argument in the proof of [1, Theorem 2]. Namely, we de-interleave Z^n into $(k+1)$ subsequences consisting of symbols separated by blocks of k symbols, and exploit the conditional independence of symbols in each subsequence, given all symbols not in that subsequence, to use Lemma 1.

Proof of Lemma 2: See the Appendix, Part C. □

Now, for an integer $0 \leq m \leq \lfloor \frac{n-2k}{2} \rfloor$ and given z^n, let $n(\boldsymbol{c}) \triangleq |\mathcal{T}(\boldsymbol{c})|$, and $m(\boldsymbol{c}) \triangleq \min\{n(\boldsymbol{c}), m\}$ for $\boldsymbol{c} \in \boldsymbol{C}_k$. Then, analogously as in (20), we define

$$\mathcal{S}_{k,m}^n(z^n) = \Big\{ \boldsymbol{S}_k(z^n) \in \mathcal{S}_0^{n-2k} :$$
$$\{s_{k,\tau}(\boldsymbol{c}, \cdot)\}_{\tau \in \mathcal{T}(\boldsymbol{c})} \in \mathcal{S}_{0,m(\boldsymbol{c})}^{n(\boldsymbol{c})} \text{ for all } \boldsymbol{c} \in \bar{\boldsymbol{C}}_k \Big\}. \quad (28)$$

[4]Again, the value of $\hat{X}_t^{\boldsymbol{S}_k}(z^n)$ for $t \leq k$ and $t > n-k$ can be defined as an arbitrary fixed symbol, since it will be inconsequential in subsequent development.

In words, $\mathcal{S}_{k,m}^n(z^n)$ is the set of $(n-2k)$-tuples of (kth-order denoiser induced) single-symbol denoisers that allow at most $m(\boldsymbol{c})$ shifts within the subsequence $\{t : t \in \mathcal{T}(\boldsymbol{c})\}$ for each context $\boldsymbol{c} \in \boldsymbol{C}_k$.[5] Again, for brevity, the dependence on z^n in $\mathcal{S}_{k,m}^n(z^n)$ is suppressed, and we write simply $\mathcal{S}_{k,m}^n$. It is worth noting that $\mathcal{S}_{k,m}^n$ is a larger class than the class of kth-order "sliding-window" denoisers that are allowed to shift at most m times. The reason is that in $\mathcal{S}_{k,m}^n$, the shift within each subsequence associated with each context can occur at any time, regardless of the shifts in other subsequences, whereas in the latter class, the shifts in each subsequence occur together with other shifts in other subsequences.

For integers $k \geq 0$ and $n > 2k$, we now define, for the class of n-block denoisers $\hat{\boldsymbol{X}}^{n,\boldsymbol{S}}$ with $\boldsymbol{S} \in \mathcal{S}_{k,m}^n$

$$D_{k,m}(x^n, z^n) \triangleq \min_{\boldsymbol{S} \in \mathcal{S}_{k,m}^n} L_{\hat{\boldsymbol{X}}^n, s}(x_{k+1}^{n-k}, z^n)$$
$$= \min_{\boldsymbol{S} \in \mathcal{S}_{k,m}^n} \frac{1}{n-2k} \sum_{t=k+1}^{n-k} \Lambda(x_t, s_{k,t}(\boldsymbol{c}_t, z_t)) \qquad (29)$$

the minimum normalized cumulative loss of (x^n, z^n) that can be achieved by the sequence of kth-order denoisers that allow at most m shifts within each context. Now, to build a legitimate (non-genie-aided) universal scheme achieving (29) on the basis of Z^n only, we define

$$\hat{\boldsymbol{S}}_{k,m} = \arg \min_{\boldsymbol{S} \in \mathcal{S}_{k,m}^n} \tilde{L}_{\boldsymbol{S}}(z^n) \qquad (30)$$

and the (k,m)-S-DUDE, $\hat{\boldsymbol{X}}_{\text{univ}}^{n,k,m}$, is defined as $\hat{\boldsymbol{X}}^{n,\hat{\boldsymbol{S}}_{k,m}}$. Note that when $m = 0$, $\hat{\boldsymbol{X}}^{n,\hat{\boldsymbol{S}}_{k,m}}$ coincides with the DUDE in [1]. The following theorem generalizes Theorem 2 to the case of general $k \geq 0$.

Theorem 3: Let $\hat{\boldsymbol{X}}_{\text{univ}}^{n,k,m}$ be given by $\hat{\boldsymbol{X}}^{n,\hat{\boldsymbol{S}}_{k,m}}$, where $\hat{\boldsymbol{S}}_{k,m}$ is defined in (30). Then, for all $\epsilon > 0$ and $x^n \in \mathcal{X}^n$

$$\Pr\left(L_{\hat{\boldsymbol{X}}_{\text{univ}}^{n,k,m}}(x_{k+1}^{n-k}, Z^n) - D_{k,m}(x^n, Z^n) > \epsilon \right) \qquad (31)$$

$$\leq 2(k+1) \exp\left(-(n-2k) \cdot \left[\frac{\epsilon^2}{2(k+1)L_{\max}^2} \right. \right.$$
$$\left. \left. - 2|\mathcal{Z}|^{2k} \cdot \left\{ h\left(\frac{m}{n-2k}\right) + \frac{(m+1)\ln N}{n-2k} \right\} \right] \right) \quad (32)$$

where $h(x) = -x \ln x - (1-x) \ln(1-x)$ for $0 \leq x \leq 1$, and $N = |\mathcal{S}| = |\hat{\mathcal{X}}|^{|\mathcal{Z}|}$.

Remark: Note that when $k = 0$, this theorem coincides with Theorem 2. Similarly to the way Theorem 2 was plausible given Lemma 1, Theorem 3 can be expected given Lemma 2, since $\hat{\boldsymbol{S}}_{k,m}$ achieves $\min_{\boldsymbol{S} \in \mathcal{S}_{k,m}^n} \tilde{L}_{\boldsymbol{S}}(Z^n)$, and we expect $L_{\hat{\boldsymbol{X}}^n, \hat{s}_{k,m}}(x_{k+1}^{n-k}, Z^n)$ to be close to $D_{k,m}(x^n, Z^n)$ from the concentration of $\tilde{L}_{\boldsymbol{S}}(Z^n)$ to $L_{\hat{\boldsymbol{X}}^n, s}(x_{k+1}^{n-k}, Z^n)$ for all $\boldsymbol{S} \in \mathcal{S}_{k,m}^n$.

[5]When $m = 0$, $\mathcal{S}_{k,0}^n(z^n)$ becomes the set of n-block kth-order "sliding-window" denoisers.

Proof of Theorem 3: See the Appendix, Part D.   □

From Theorem 3, we now easily obtain one of the main results of the paper, which extends Theorem 1 from the case $m = 0$ to the case of general $0 \leq m \leq \lfloor \frac{n-2k}{2} \rfloor$. That is, the following theorem asserts that, for every underlying sequence $\boldsymbol{x} \in \mathcal{X}^\infty$, our (k, m)-S-DUDE performs essentially as well as the best shifting kth-order denoiser that allows at most m shifts within each context, both in high probability and expectation sense, provided a growth condition on k and m is satisfied.

Theorem 4: Suppose $k = k_n$ and $m = m_n$ are such that the right-hand side of (32) is summable in n. Then, for all $\boldsymbol{x} \in \mathcal{X}^\infty$, the sequence of denoisers $\{\hat{\boldsymbol{X}}_{\text{univ}}^{n,k,m}\}$ satisfies

a)

$$\lim_{n \to \infty} \left[L_{\hat{\boldsymbol{X}}_{\text{univ}}^{n,k,m}}(x^n, Z^n) - D_{k,m}(x^n, Z^n) \right] = 0 \quad \text{a.s.} \quad (33)$$

b) For any $\delta > 0$

$$E\left[L_{\hat{\boldsymbol{X}}_{\text{univ}}^{n,k,m}}(x^n, Z^n) - D_{k,m}(x^n, Z^n) \right]$$
$$= O\left(\sqrt{k_n |\mathcal{Z}|^{2k_n} \left(\frac{m_n}{n} \right)^{1-\delta}} \right). \quad (34)$$

Remark: It will be seen in Claim 1 below that the stipulation in the theorem implies $\lim_{n \to \infty} k_n |\mathcal{Z}|^{2k_n} \left(\frac{m_n}{n} \right)^{1-\delta} = 0$, which, when combined with (34), implies that the expected difference on the left-hand side of (34) vanishes with increasing n. That in itself, however, can easily be deduced from (33) and bounded convergence. The more significant value of (34) is in providing a rate of convergence result for the "redundancy" in the S-DUDE's performance, as a function of both k and m. In particular, note that for any $\eta > 0$, $O(n^{-1/2+\eta})$ is achievable provided $k_n = c \log n$ and $m_n = n^\xi$, for small enough positive constants c, ξ.

In what follows, we specify the maximal growth rates for $k = k_n$ and $m = m_n$ under which the summability condition stipulated in Theorem 4 holds.

Claim 1:
a) Maximal growth rate for k: The summability condition in Theorem 4 is satisfied provided $k_n = c_1 \log n$ with $c_1 < \frac{1}{2 \log |\mathcal{Z}|}$ and m_n grows at any subpolynomial rate. On the other hand, the condition is not satisfied for $k_n = c_1 \log n$ with any $c_1 \geq \frac{1}{2 \log |\mathcal{Z}|}$, even when m is fixed (not growing with n).
b) Maximal growth rate for m: The summability condition in Theorem 4 is satisfied for any sublinear growth rate of m_n, provided k_n is taken to increase sufficiently slowly that $k_n |\mathcal{Z}|^{2k_n} = o((n/m_n)^{1-\delta})$ for some $\delta > 0$. On the other hand, the condition is not satisfied whenever m_n grows linearly with n, even when k is fixed.

Proof of Claim 1: See Appendix, Part E.   □

Proof of Theorem 4: See the Appendix, Part F.   □

C. A "Strong Converse"

In Claim 1, we have shown the necessity of $m_n = o(n)$ for the condition required in Theorem 4 to hold. However, we can prove the necessity of $m_n = o(n)$ in a much stronger sense, described in the following theorem.

Theorem 5: Suppose that $\mathcal{X} = \hat{\mathcal{X}}$, that $\Lambda(x, \hat{x}) \geq 0$ for all $x, \hat{x}$ with equality if and only if $x = \hat{x}$, and that $\Pi(x, z) > 0$ for all x, z. If $m = m_n = \Theta(n)$, then for any sequence of denoisers $\{\hat{\boldsymbol{X}}^n\}$, there exists $\boldsymbol{x}^\infty \in \mathcal{X}^\infty$ such that

$$\limsup_{n \to \infty} E\left[L_{\hat{\boldsymbol{X}}^n}(x^n, Z^n) - D_{0,m}(x^n, Z^n) \right] > 0. \quad (35)$$

Remark: The theorem establishes the fact that when $m_n = o(n)$ does not hold, namely, when $m_n = \Theta(n)$, not only does the almost-sure convergence in Theorem 4 not hold but, in fact, even the much weaker convergence in expectation would fail. Further, it shows that this would be the case for *any* sequence of denoisers, not necessarily the S-DUDE. In addition, (35) features $D_{0,m}(x^n, Z^n)$, pertaining to competition with a genie that shifts among single-symbol denoisers so, *a fortiori*, it implies that for any fixed $k > 0$ or k that grows with n

$$\limsup_{n \to \infty} E\left[L_{\hat{\boldsymbol{X}}^n}(x^n, Z^n) - D_{k,m}(x^n, Z^n) \right] > 0 \quad (36)$$

also holds since, by definition, $D_{0,m}(x^n, z^n) \geq D_{k,m}(x^n, z^n)$ for all x^n, z^n and $k \geq 0$. Therefore, the theorem asserts that for *any* sequence of denoisers to compete with $D_{k,m}(x^n, Z^n)$, even in expectation sense, $m_n = o(n)$ is necessary. Finally, we mention that the conditions stipulated in the statement of the theorem regarding the loss function and the channel can be considerably relaxed without compromising the validity of the theorem. These conditions are made to allow for the simple proof that we give in the Appendix, Part G.

IV. THE STOCHASTIC SETTING

In [1], the semi-stochastic setting result, [1, Theorem 1], was shown to imply the result for the stochastic setting as well. That is, when the underlying data form a stationary process, [1, Sec. VI] shows that the DUDE attains optimum distribution-dependent performance. Analogously, we can now use the results from the semi-stochastic setting of the previous section to generalize the results of [1, Sec. VI] and show that our S-DUDE attains optimum distribution-dependent performance when the underlying data form a piecewise stationary process. We first define the precise notion of the class of piecewise stationary processes in Section IV-A, and discuss the richness of this class in Section IV-B. Section IV-C gives the main result of this section: the stochastic setting optimality of the S-DUDE.

A. Definition of the Class of Processes $\mathcal{P}\{m_n\}$

Let $P_{\boldsymbol{X}}^{(1)}, \ldots, P_{\boldsymbol{X}}^{(M)}$ be a finite collection of M probability distributions of stationary processes, with components taking the values in $\mathcal{X}$. Let $\boldsymbol{A}$ be a process with components taking the values in $\{1, \ldots, M\}$. Then, a piecewise stationary process $\boldsymbol{X}$ is generated by shifting between the M processes in a way specified by the "switching process" $\boldsymbol{A}$, as we now describe.

First, denote $r(A^n)$ as the number of shifts that have occurred along the n-tuple A^n, i.e.,

$$r(A^n) \triangleq \sum_{j=1}^{n-1} \mathbf{1}_{\{A_j \neq A_{j+1}\}}.$$

Thus, there are $r(A^n) + 1$ "blocks" in A^n, where each block is a tuple of constant values that are different from the values of adjacent blocks. Now, for each $1 \leq i \leq r(A^n) + 1$, we define

$$\tau_i(A^n) \triangleq \begin{cases} \inf\{t : \sum_{j=1}^{t} \mathbf{1}_{\{A_j \neq A_{j+1}\}} = i\}, & \text{if } 1 \leq i \leq r(A^n) \\ n, & \text{if } i = r(A^n) + 1 \end{cases}$$

as the last time instance of the ith block in A^n. In addition, define $\tau_0(A^n) \triangleq 0$. Clearly, $r(A^n)$ and $\tau_i(A^n)$ depend on A^n and, thus, are random variables. However, for brevity, we suppress the dependence on A^n when there is no confusion, and write simply r and τ_i, respectively.

Using these definitions, and by denoting P_{A^n} as the nth-order marginal distribution of $\boldsymbol{A}$, we define a piecewise stationary process $\boldsymbol{X}$ by characterizing its nth-order marginal distribution P_{X^n} as

$$P_{X^n}(X^n = x^n) = \sum_{a^n} P_{A^n}(a^n) P(X^n = x^n | A^n = a^n)$$
$$= \sum_{a^n} P_{A^n}(a^n) \prod_{i=1}^{r+1} P_{\boldsymbol{X}}^{(a_{\tau_i})}(x_{\tau_{i-1}+1}^{\tau_i}) \quad (37)$$

for each n. The corresponding distribution of the process $\boldsymbol{X}$ is denoted as $P_{\boldsymbol{X}}$.[6] In words, $\boldsymbol{X}$ is constructed by following one of the M probability distributions in each block, switching from one to another depending on $\boldsymbol{A}$. Furthermore, conditioned on the realization of $\boldsymbol{A}$, each stationary block is independent of other blocks, even if the distribution of distinct blocks is the same. This property of conditional independence is reasonable for modeling many types of data arising in practice, since we can think of the M distributions as different "modes"; if the process returns to the same mode, it is reasonable to model the new block as a new independent realization of that same distribution. In other words, the "mode" may represent the kind of "texture" in a certain region of the data, but two different regions with the same "texture" should have independent realizations from the texture-generating source. Our notion of a piecewise stationary process almost coincides with that developed in [28]. The main difference is that we allow an arbitrary distribution for the process $\boldsymbol{A}$.

Now, we define $\mathcal{P}\{m_n\}$ to be the class of all process distributions that can be constructed as in (37) for some M, some collection $P_{\boldsymbol{X}}^{(1)}, \ldots, P_{\boldsymbol{X}}^{(M)}$ of stationary processes, and some switching process $\boldsymbol{A}$ whose number of shifts satisfies

$$r(A^n) \leq m_n \quad \text{a.s.} \quad \forall n. \quad (38)$$

In words, a process $\boldsymbol{X}$ belongs to[7] $\mathcal{P}\{m_n\}$ if and only if it can be formed by switching between a finite collection of independent processes in which the number of switches by time n does not exceed m_n.

---

[6] $\{P_{X^n}\}_{n \geq 1}$ is readily verified to be a consistent family of distributions and, thus, by Kolmogorov's extension theorem, uniquely defines the distribution of the process $\boldsymbol{X}$.

[7] The phrase "the process $\boldsymbol{X}$ belongs to $\mathcal{P}\{m_n\}$" is shorthand for "the distribution of the process $\boldsymbol{X}$, $P_{\boldsymbol{X}}$, belongs to $\mathcal{P}\{m_n\}$."

B. Richness of $\mathcal{P}\{m_n\}$

In this subsection, we examine how rich the class $\mathcal{P}\{m_n\}$ is, in terms of the growth rate m_n and the existence of denoising schemes that are universal with respect to $\mathcal{P}\{m_n\}$. First, given any distribution on a noiseless n-tuple, P_{X^n}, we define

$$\mathbb{D}(P_{X^n}, \boldsymbol{\Pi}) \triangleq \min_{\hat{\boldsymbol{X}}^n \in \mathcal{D}_n} EL_{\hat{\boldsymbol{X}}^n}(X^n, Z^n) \quad (39)$$

where $\mathcal{D}_n$ is the class of all n-block denoisers. The expectation on the right-hand side of (39) assumes that X^n is generated from P_{X^n} and that Z^n is the output of the DMC, $\boldsymbol{\Pi}$, whose input is X^n. Thus, $\mathbb{D}(P_{X^n}, \boldsymbol{\Pi})$ is the optimum denoising performance (in the sense of expected per-symbol loss) attainable when the source distribution P_{X^n} is known.

What happens when the source distribution is unknown? Theorem 3 of [1] established the fact that[8]

$$\lim_{n \to \infty} \left[EL_{\hat{\boldsymbol{X}}_{\mathrm{DUDE}}^{n,k_n}}(X^n, Z^n) - \mathbb{D}(P_{X^n}, \boldsymbol{\Pi}) \right] = 0 \quad (40)$$

for all stationary $P_{\boldsymbol{X}}$. Note that our newly defined class of processes, $\mathcal{P}\{m_n\}$, is simply the class of all stationary processes if one takes the sequence m_n to be $m_n \equiv 0$ for all n. Thus, assuming $m_n \equiv 0$, (40) is equivalent to

$$\lim_{n \to \infty} \left[EL_{\hat{\boldsymbol{X}}_{\mathrm{DUDE}}^{n,k_n}}(X^n, Z^n) - \mathbb{D}(P_{X^n}, \boldsymbol{\Pi}) \right] = 0 \quad (41)$$

for all $P_{\boldsymbol{X}} \in \mathcal{P}\{m_n\}$. At the other extreme, when $m_n = n$, $\mathcal{P}\{m_n\}$ consists of all possible (not necessarily stationary) processes. We can observe this equivalence by having $M = |\mathcal{X}|$ processes each be a constant process at a different symbol in $\mathcal{X}$, and creating any process by switching to the appropriate symbol. In this case, not only does (41) not hold for the DUDE, but clearly (41) cannot hold under any sequence of denoisers. In other words, $\mathcal{P}\{m_n\}$ is far too rich to allow for the existence of schemes that are universal with respect to it.

It is obvious then that $\mathcal{P}\{m_n\}$ is significantly richer than the family of stationary processes whenever m_n grows with n. It is of interest then to identify the maximal growth rate of m_n that allows for the existence of schemes that are universal with respect to $\mathcal{P}\{m_n\}$, and to find such a universal scheme. In what follows, we offer a complete answer to these questions. Specifically, we show that if the growth rate of m_n allows for the existence of any scheme which is universal with respect to $\mathcal{P}\{m_n\}$, the S-DUDE is universal, as well.

C. Universality of S-DUDE

Here, we state our stochastic setting result, which establishes the universality of (k, m)-S-DUDE with respect to the class $\mathcal{P}\{m_n\}$.

Theorem 6: Let $k = k_n$ and $m = m_n$ satisfy the growth rate condition stipulated in Theorem 4, in addition to $\lim_{n \to \infty} k_n = \infty$. Then, the sequence of denoisers $\{\hat{\boldsymbol{X}}_{\mathrm{univ}}^{n,k,m}\}$ defined in Section III satisfy

$$\lim_{n \to \infty} \left[EL_{\hat{\boldsymbol{X}}_{\mathrm{univ}}^{n,k,m}}(X^n, Z^n) - \mathbb{D}(P_{X^n}, \boldsymbol{\Pi}) \right] = 0 \quad (42)$$

for all $P_{\boldsymbol{X}} \in \mathcal{P}\{m_n\}$.

---

[8] When $P_{\boldsymbol{X}}$ is stationary, the limit $\lim_{n \to \infty} \mathbb{D}(P_{X^n}, \boldsymbol{\Pi}) \triangleq \mathbb{D}(P_{\boldsymbol{X}}, \boldsymbol{\Pi})$ was shown to exist in [1]. Thus, (40) was equivalently stated as $\lim_{n \to \infty} EL_{\hat{\boldsymbol{X}}_{\mathrm{DUDE}}^n} = \mathbb{D}(P_{\boldsymbol{X}}, \boldsymbol{\Pi})$ in [1, Theorem 3].

Remark 1: Recall that, as noted in Claim 1, $m_n = o(n)$ together with appropriately slowly growing $k = k_n$ is sufficient to guarantee the growth rate condition stipulated in Theorem 4. Hence, by Theorem 6, $m = o(n)$ and the sufficiently slowly growing $k = k_n$ suffices for (42) to hold. Therefore, Theorem 6 implies the existence of schemes that are universal with respect to $\mathcal{P}\{m_n\}$ whenever m_n increases sublinearly in n. Since, as discussed in Section IV-B, no universal scheme exists for $\mathcal{P}\{m_n\}$ when m_n is linear in n, we conclude that the sublinearity of m_n is the necessary and sufficient condition for a universal scheme to exist with respect to $\mathcal{P}\{m_n\}$. Moreover, Theorem 6 establishes the strong sense of optimality of the S-DUDE, as it shows that whenever $\mathcal{P}\{m_n\}$ is universally "competable," the S-DUDE does the job. This fact is somewhat analogous to the situation in [28], where the optimality of the universal lossless coding scheme presented therein for piecewise stationary sources was established under the condition that $m = o(n)$.

Remark 2: A pointwise result

$$\lim_{n \to \infty} \left[L_{\hat{\mathbf{X}}^{n,k,m}_{\text{univ}}}(X^n, Z^n) - \mathbb{D}(P_{X^n}, \mathbf{\Pi}) \right] = 0 \quad \text{a.s.}$$

for all $P_{\mathbf{X}} \in \mathcal{P}\{m_n\}$, which is analogous to [1, Theorem 4], can also be derived. However, we omit such a result here since the details required for stating it rigorously would be convoluted, and its added value over the strong point-wise result we have already established in the semi-stochastic setting would be little.

Proof of Theorem 6: See the Appendix, Part H. □

V. ALGORITHM AND COMPLEXITY

A. An Efficient Implementation of S-DUDE

In the preceding two sections, we gave strong asymptotic performance guarantees for the new class of schemes, the S-DUDE. However, the question regarding the practical implementation of (30), i.e., obtaining

$$\hat{\mathbf{S}}_{k,m} = \arg \min_{\mathbf{S} \in \mathcal{S}^n_{k,m}} \tilde{L}_{\mathbf{S}}(z^n)$$

for fixed k, m, and n remains and, at first glance, may seem to be a difficult combinatorial optimization problem. In this section, we devise an efficient two-pass algorithm, which yields (30) and performs denoising with linear complexity in the sequence length n. A recursion similar to that in the first pass of the algorithm we present appears also in the study of tracking the best expert in online learning [15], [16].

From the definition of $\mathcal{S}^n_{k,m}$, (28), we can see that obtaining (30) is equivalent to obtaining the best combination of single-symbol denoisers with at most $m(\mathbf{c})$ shifts that minimizes the cumulative estimated loss along $\{t : t \in \mathcal{T}(\mathbf{c})\}$, for each $\mathbf{c} \in \mathbf{C}_k$. Thus, our problem breaks down to $|\mathbf{C}_k|$ independent problems, each being a problem of competing with the best combination of single-symbol schemes allowing m switches.

To describe an algorithm that implements this parallelization efficiently, we first define variables. For (k, m)-S-DUDE, let $I = m + 1, J = N$, where $N = |\mathcal{S}| = |\hat{\mathcal{X}}|^{|\mathcal{Z}|}$. Then, a matrix $LM_t \in \mathbb{R}^{I \times J}$ is defined for $k + 1 \leq t \leq n - k$, where $LM_t(i, j)$ for $1 \leq i \leq I$ and $1 \leq j \leq J$ represents the minimum (un-normalized) cumulative estimated loss of the sequence of single-symbol denoisers along the time index $\{\tau : \tau \leq t, \mathbf{c}_\tau = \mathbf{c}_t\}$, allowing at most $(i - 1)$ shifts between single-symbol denoisers and applying $s_t = j$. Moreover, a vector $IM_t \in \mathbb{R}^I$ is defined for the same range of t, where $IM_t(i)$, for $1 \leq i \leq I$, is the symbol-by-symbol denoiser that attains the minimum value of the ith row of LM_t, i.e., $\arg \min_{1 \leq j \leq J} LM_t(i, j)$. A time pointer $T \in \mathbb{R}^D$, where $D = |\mathbf{C}_k| = |\mathcal{Z}|^{2k}$, is defined to store the closest time index that has the same context as current time, during the first and second pass. That is,

$$T(\mathbf{c}_t) \triangleq \begin{cases} \max\{\tau : \tau < t, \mathbf{c}_\tau = \mathbf{c}_t\}, & \text{when first pass} \\ \min\{\tau : \tau > t, \mathbf{c}_\tau = \mathbf{c}_t\}, & \text{when second pass}. \end{cases} \quad (43)$$

We also define $r \in \mathbb{R}^D$ and $q \in \mathbb{R}^D$ as variables for storing the pointer enabling our scheme to follow the best combination of single-symbol denoisers during the second pass. Thus, the total memory size required is $O(mNn + |\mathcal{Z}|^{2k}) = O(mNn)$ (assuming that k satisfies the growth rate stipulated in the previous sections, which implies $|\mathcal{Z}|^{2k} = o(n)$).

Our two-pass algorithm has ingredients from both the DUDE and from the forward–backward recursions of hidden Markov models [29] and, in fact, the algorithm becomes equivalent to DUDE when $m = 0$. The first pass of the algorithm runs forward from $t = k + 1$ to $t = n - k$, and updates the elements of LM_t recursively. The recursions have a natural dynamic programming structure. For $2 \leq i \leq I, 1 \leq j \leq J$, $LM_t(i, j)$ is determined by

$$LM_t(i, j) = \ell(z_t, j) + \min \left\{ LM_{T(\mathbf{c}_t)}(i, j), LM_{T(\mathbf{c}_t)}(i-1, IM_{T(\mathbf{c}_t)}(i-1)) \right\} \quad (44)$$

that is, adding the current loss to the best cumulative loss up to $T(\mathbf{c}_t)$ along $\{\tau : \tau < t, \mathbf{c}_\tau = \mathbf{c}_t\}$. When $i = 1$, the second term in the minimum of (44) is not defined, and $LM_t(i, j)$ just becomes $\ell(z_t, j) + LM_{T(\mathbf{c}_t)}(i, j)$. The validity of (44) can be verified by observing that there are two possible cases in achieving $LM_t(i, j)$: either the $(i - 1)$th shift to the single-symbol denoiser j occurred before t, or it occurred at time t. We can see that the first term in the minimum of (44) corresponds to the former case; the second term corresponds to the latter. Obviously, the minimum of these two (where ties may be resolved arbitrarily), leads to the value of $LM_t(i, j)$ as in (44). After updating all LM_t's during the first pass, the second pass runs backwards from $t = n - k$ to $t = k + 1$, and extracts $\hat{\mathbf{S}}_{k,m}$ from $\{LM_t\}_{t=k+1}^{n-k}$ by following the best shifting between single-symbol denoisers. The actual denoising (i.e., assembling the reconstruction sequence $\hat{X}^n$) is also performed in that pass. The pointers $r(\mathbf{c}_t)$ and $q(\mathbf{c}_t)$ are updated recursively, and they track the best shifting point and combination of single-symbol denoisers, respectively, for each of the subsequences associated with the various contexts. A succinct description of the algorithm is provided in Algorithm 1. The time complexity of the algorithm is readily seen to be $O(mNn)$ as well.

Algorithm 1: The (k, m)-Shifting Discrete Denoising Algorithm

---

Require: $LM_t(i, j) \in \mathbb{R}^{I \times J}$, $IM_t(i) \in \mathbb{R}^I$, $k + 1 \leq t \leq n - k, 1 \leq i \leq I, 1 \leq j \leq J$,

$$T \in \mathbb{R}^D,\ r \in \mathbb{R}^D, q \in \mathbb{R}^D,\ L \in \mathbb{R}$$

Ensure: $\hat{\boldsymbol{S}}_k = \{s_{k,t}(\boldsymbol{c}_t, \cdot)\}_{t=k+1}^{n-2k}$ in (30) and the denoised output $\{\hat{x}_t\}_{t=k+1}^{n-k}$

  $\tau(\boldsymbol{c}) \Leftarrow \phi$ for all $\boldsymbol{c} \in \boldsymbol{C}_k$

  for $t = k + 1$ to $n - k$ **do**

    if $T(\boldsymbol{c}_t) = \phi$ **then**

      $LM_t(i, j) \Leftarrow \ell(z_t, j)$ for $1 \leq i \leq I$, $1 \leq j \leq J$

      $IM_t(i) \Leftarrow \arg\min_{1 \leq j \leq J} LM_t(i, j)$ for $1 \leq i \leq I$

    else see $LM_{T(\boldsymbol{c}_t)}^*(i, j)$ defined at the bottom of the page.

      $LM_t(i, j) \Leftarrow LM_{T(\boldsymbol{c}_t)}^*(i, j) + \ell(z_t, j)$, for $1 \leq i \leq I$, $1 \leq j \leq J$

      $IM_t(i) \Leftarrow \arg\min_{1 \leq j \leq J} LM_t(i, j)$, for $1 \leq i \leq I$

    end if

    $T(\boldsymbol{c}_t) \Leftarrow t$

  end for

  $T(\boldsymbol{c}) \Leftarrow \phi$ for all $\boldsymbol{c} \in \bar{\boldsymbol{C}}_k$

  for $t = n - k$ to $k + 1$ **do**

    if $T(\boldsymbol{c}_t) = \phi$ **then**

      $r(\boldsymbol{c}_t) \Leftarrow I, q(\boldsymbol{c}_t) \Leftarrow IM_t(r(\boldsymbol{c}_t))$

    else

      $L \Leftarrow LM_{T(\boldsymbol{c}_t)}(r(\boldsymbol{c}_t),\ q(\boldsymbol{c}_t)) - \ell(z_t, q(\boldsymbol{c}_t))$

      if $L < LM_t(r(\boldsymbol{c}_t), q(\boldsymbol{c}_t))$ **then**

        $r(\boldsymbol{c}_t) \Leftarrow r(\boldsymbol{c}_t) - 1, q(\boldsymbol{c}_t) \Leftarrow IM_t(r(\boldsymbol{c}_t))$

      end if

    end if

    $T(\boldsymbol{c}_t) \Leftarrow t, s_{k,t}(\boldsymbol{c}_t, \cdot) \Leftarrow q(\boldsymbol{c}_t)$

    $\hat{x}_t \Leftarrow s_{k,t}(\boldsymbol{c}_t, z_t)$

  end for

B. Extending the S-DUDE to Multidimensional Data

As noted, our algorithm is essentially separately employing the same algorithm to compete with the best shifting single-symbol denoisers, on each subsequence associated with each context. The overall algorithm is the result of parallelizing the operations of the schemes for the different subsequences, which allows for a more efficient implementation than if these schemes were to be run completely independently of one another. This characteristic of running the same algorithm in parallel along each subsequence enables us to extend S-DUDE to the case of multidimensional data: run the same algorithm along each subsequence associated with each (this time multidimensional) context. It should be noted, however, that the extension of the S-DUDE to the multidimensional case is not as straightforward as the extension of the DUDE was, since, whereas the DUDE's output is independent of the ordering of the data within each context, this ordering may be very significant in its effect on the output and, hence, the performance of S-DUDE. Therefore, the choice of a scheme for scanning the data and capturing its local spatial stationarity, e.g., Peano–Hilbert scanning [30], is an important ingredient in extending S-DUDE to the denoising of multidimensional data. Findings from the recent study on universal scanning reported in [31], [32] can be brought to bear on such an extension.

VI. Experimentation

In this section, we report some preliminary experimental results obtained by applying S-DUDE to several kinds of noise-corrupted data.

A. Image Denoising

In this subsection, we report some experimental results of denoising a binary image under the Hamming loss function. The first and most simplistic experiment is with the 400×400 black-and-white binary image shown in Fig. 1. The first figure is the clean underlying image. The image is raster scanned and passed through a binary-symmetric channel (BSC) with crossover probability $\delta = 0.1$, to obtain the noisy image in Fig. 1(b). Note that in this case, there are only four symbol-by-symbol denoisers, namely, $\mathcal{S} = \{0, 1, z, \bar{z}\}$, representing always-say-0, always-say-1, say-what-you-see, and flip-what-you-see, respectively. Fig. 1(c) is the DUDE output with $k = 0$, Fig. 1(d) is the DUDE output with $k = 4$, and Fig. 1(e) is the output of our S-DUDE with $k = 0, m = 1$. The DUDE with $k = 0$ competes with the best time-invariant symbol-by-symbol denoiser which, in this case, is the say-what-you-see denoiser, since the empirical distribution of the clean image is $(0.5, 0.5)$ and $\delta = 0.1$. Given that the DUDE with $k = 0$ turns out to be also the say-what-you-see denoiser, the DUDE output with $k = 0$ is the same as the noisy image, and so no denoising is performed. The performance of DUDE with larger $k = 4$ shows a decent performance with the

---

$$LM_{T(\boldsymbol{c}_t)}^*(i, j) \Leftarrow \begin{cases} LM_{T(\boldsymbol{c}_t)}(i, j), & \text{for } i = 1, \quad 1 \leq j \leq J \\ \min\left\{ LM_{T(\boldsymbol{c}_t)}(i, j), LM_{T(\boldsymbol{c}_t)}(i - 1, IM_{T(\boldsymbol{c}_t)}(i - 1)) \right\}, & \text{for } 2 \leq i \leq I, \quad 1 \leq j \leq J \end{cases}$$

(a) Clean   (b) Noisy

(c) DUDE $k = 0$ (d) DUDE $k = 4$
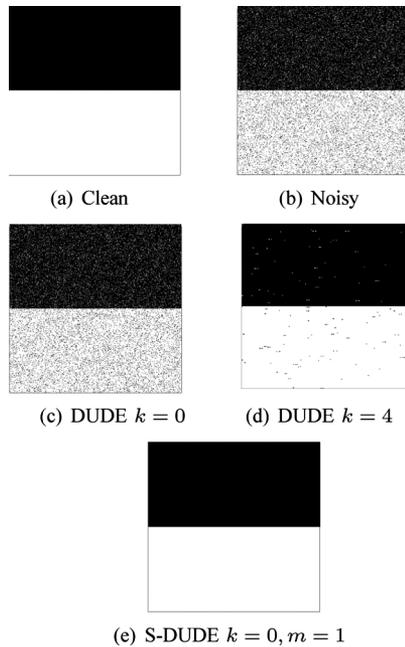
(e) S-DUDE $k = 0, m = 1$

Fig. 1. 400×400 binary images.

bit-error rate (BER) 0.0019, but still shows errors. However, it is clear that, for this image, the best compound action of the symbol-by-symbol denoisers is always-say-0 for the first half and then a shift to always-say-1 for the remainder. We can see that our $(0, 1)$-S-DUDE successfully captures this shift from the noisy observations, and results in *perfect* denoising with zero bit errors.

Now, we move on to a more realistic example. Fig. 2(a), a concatenation of a half-toned Einstein image (300×300) and scanned Shannon's 1948 paper (300×300), is the clean image. We raster-scan the image and pass it through a BSC with crossover probability $\delta = 0.1$, to obtain the noisy image Fig. 2(b), which we raster-scan and employ the S-DUDE on the resulting one-dimensional sequence. Since the two concatenated images are of a very different nature, we expect our S-DUDE to perform better than the DUDE, because it is designed to adapt to the possibility of employing different schemes in different regions of the data. The plot in Fig. 2(c) shows the performance of our (k, m)-S-DUDE with various values of k and m. The horizontal axis reflects k, and the vertical axis represents the ratio of BER to $\delta = 0.1$. Each curve represents the BER of schemes with different $m = 0, 1, 2, 3$. Note that $m = 0$ corresponds to the DUDE. We can see that S-DUDE with $m > 0$ mostly dominates the DUDE, with an additional BER reduction of $\sim 11\%$, including when $k = 6$, the best k value for the DUDE. The bottom three figures show the denoised images with $(k, m) = (4, 0), (4, 2), (6, 1)$, achieving BERs of $\delta \times (0.744, 0.6630, 0.4991)$, respectively. Thus, in this example, $(4, 2)$-S-DUDE achieves an additional BER reduction of 11% over the DUDE with $k = 4$, and the overall best performance is achieved by $(6, 1)$-S-DUDE. Given the nature of the image, which is a concatenation of two completely different types of images, each reasonably uniform in texture, it is not surprising to find that the S-DUDE with $m = 1$ performs the best.
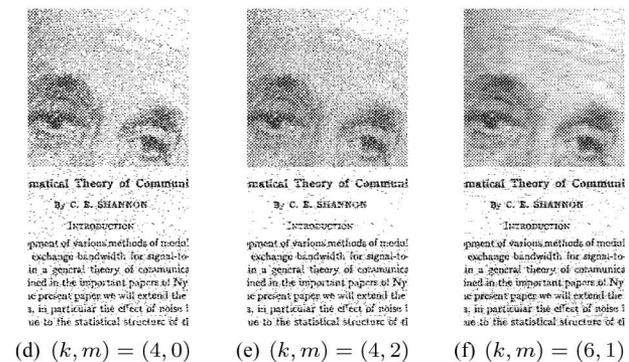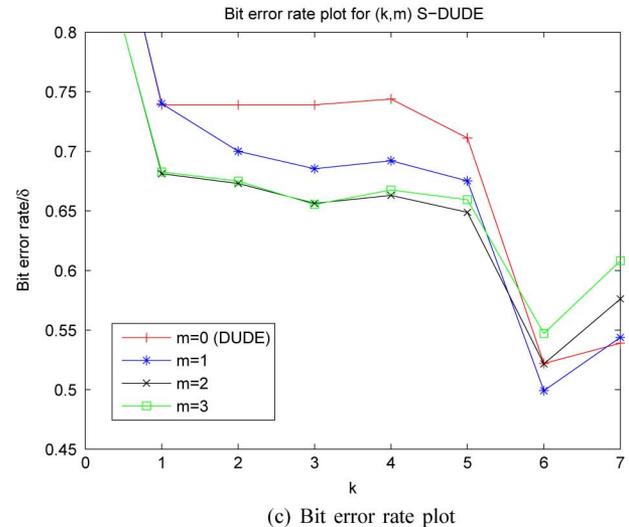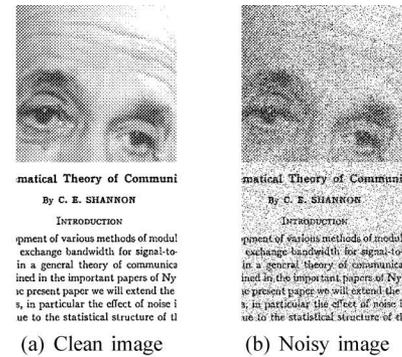

(a) Clean image   (b) Noisy image

(c) Bit error rate plot

(d) $(k, m) = (4, 0)$ (e) $(k, m) = (4, 2)$ (f) $(k, m) = (6, 1)$

Fig. 2. Clean and noisy images, the BER plot for (k, m)-S-DUDE, and three denoised outputs for $(k, m) = (4, 0), (4, 2), (6, 1)$, respectively.

B. State Estimation for a Switching Binary Hidden Markov Process

Here, we give a stochastic setting experiment. A switching binary hidden Markov process in this example is defined as a binary symmetric Markov chain observed through a BSC, where the transition probabilities of the Markov chain switches over time. The goal of a denoiser here is to estimate the underlying Markov chain based on the noisy output.

In our example, we construct a simple switching binary hidden Markov process of length $n = 10^6$, in which the transition probability of the underlying binary symmetric Markov source switches from $p = 0.01$ to $p = 0.2$ at the midpoint of the sequence, and the crossover probability of the BSC is $\delta = 0.1$. Then, we estimate the state of the underlying Markov
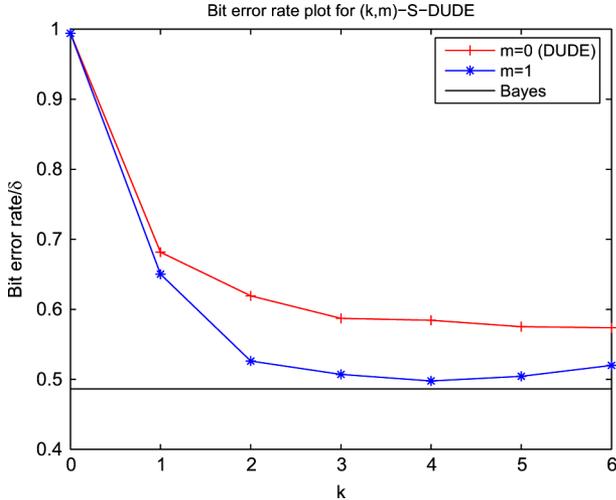
Fig. 3. BER for switching binary hidden Markov process ($\delta = 0.1$, $n = 10^6$). The switch of the underlying binary Markov chain occurs when $t = 5 \times 10^5$, from the transition probability $p = 0.01$ to $p = 0.2$.

chain based on the BSC output. The goodness of the estimation is again measured by the Hamming loss, i.e., the fraction of errors made. Slightly better than the optimal Bayesian distribution-dependent performance for this case can be obtained by employing the forward–backward recursion scheme, incorporating the varying transition probabilities with the help of a genie that knows the exact location of the change in the process distribution. Fig. 3 plots the BER of (k, m)-S-DUDE with various k and m, compared to the genie-aided Bayes optimal BER. The horizontal axis represents k, and the two curves refer to $m = 0$ (DUDE) and $m = 1$. The vertical axis is the ratio of BER to $\delta = 0.1$.

We can observe that the optimal Bayesian BER is (lower-bounded by) $0.4865 \times \delta$. The best performance of the DUDE was achieved when $k = 6$ with a BER of $0.5738 \times \delta$, which is far above (18% more than) the optimal BER. It is clear that, despite the size of the data, the DUDE fails to converge to the optimum, as it is confined to be employing the same sliding-window scheme throughout the whole data. However, we can see that the $(4, 1)$-S-DUDE achieves a BER of $.4979 \times \delta$, which is within 2.3% of the optimal BER. This example shows that our S-DUDE is competent in attaining the optimum performance for a class richer than that of the stationary processes. Specifically, it attains the optimum performance for piecewise stationary processes, on which the DUDE generally fails.

VII. CONCLUSION AND SOME FUTURE DIRECTIONS

Inspired by the DUDE algorithm, we have developed a generalization that accommodates switching between sliding-window rules. We have shown a strong semi-stochastic setting result for our new scheme in competing with shifting kth-order denoisers. This result implies a stochastic setting result as well, asserting that the S-DUDE asymptotically attains the optimal distribution-dependent performance for the case in which the underlying data is piecewise stationary. We also described an efficient low-complexity implementation of the algorithm, and presented some simple experiments that demonstrate the potential benefits of employing S-DUDE in practice.

There are several future research directions related to this work. The S-DUDE can be thought of as a generalization of the DUDE, with the introduction of a new component captured by the nonnegative integer parameter m. Many previous extensions of the DUDE, such as the settings of channel with memory [33], channel uncertainty [34], applications to channel decoding [35], discrete-input continuous-output data [36], denoising of analog data [37], and decoding in the Wyner–Ziv problem [38], may stand to benefit from a revision that would incorporate the viewpoint of switching between time-invariant schemes. Particularly, extending S-DUDE to the case where the data are analog as in [37] will be nontrivial and interesting from both a theoretical and a practical viewpoint. In addition, as mentioned in Section V, an extension of the S-DUDE to the case of multidimensional data is not as straightforward as the extension of the DUDE was. Such an extension should prove interesting and practically important. Finally, it would be useful to devise guidelines for the choice of k and m. The heuristics suggested in [1, Sec. VII], would still be useful in the choice of k, and techniques in [22], [39] may also be helpful in choosing k and m. Further practical guidelines could be developed via additional experimentation.

APPENDIX

A. Proof of Lemma 1

We first establish the fact that for all $x^n \in \mathcal{X}^n$, and for fixed $\boldsymbol{S} \in \mathcal{S}_0^n$

$$\left\{ n \left[L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(x^n, Z^n) - \tilde{L}_{\boldsymbol{S}}(Z^n) \right] \right\}_{n \geq 1}$$

is a $\{Z_n\}$-martingale. This is not hard to see by the following argument:

$$E\left(n[L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(x^n, Z^n) - \tilde{L}_{\boldsymbol{S}}(Z^n)] \big| Z^{n-1} \right)$$
$$= E\left(\sum_{t=1}^n \Lambda(x_t, s_t(Z_t)) - \sum_{t=1}^n \ell(Z_t, s_t) \big| Z^{n-1} \right)$$
$$= (n-1)[L_{\hat{\boldsymbol{X}}^{n-1}, \boldsymbol{s}}(x^{n-1}, Z^{n-1}) - \tilde{L}_{\boldsymbol{S}}(Z^{n-1})]$$
$$\quad + E\left(\Lambda(x_n, s_n(Z_n)) - \ell(Z_n, s_n) \big| Z^{n-1} \right)$$
$$= (n-1)[L_{\hat{\boldsymbol{X}}^{n-1}, \boldsymbol{s}}(x^{n-1}, Z^{n-1}) - \tilde{L}_{\boldsymbol{S}}(Z^{n-1})] \quad (45)$$

where (45) follows from the fact that Z_n is independent of Z^{n-1}, and $E_{x_n} \Lambda(x_n, s_n(Z_n)) = E_{x_n} \ell(Z_n, s_n)$. Therefore, $L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(x^n, Z^n) - \tilde{L}_{\boldsymbol{S}}(Z^n)$ is a normalized sum of bounded martingale differences; therefore the inequalities (18) and (19) follow directly from the Hoeffding–Azuma inequality [14, Lemma A.7]. $\qquad \square$

B. Proof of Theorem 2

First, recall (22) that the notation $\hat{\boldsymbol{S}}$ stands for $\hat{\boldsymbol{S}}(Z^n)$ which depends on the noisy sequence Z^n. Then, consider the following chain of inequalities:[9]

$$P\left(L_{\hat{\boldsymbol{X}}^n, \hat{\boldsymbol{s}}}(x^n, Z^n) - D_{0, m}(x^n, Z^n) > \epsilon \right)$$
$$= P\left(\max_{\boldsymbol{S} \in \mathcal{S}_{0, m}^n} \left\{ L_{\hat{\boldsymbol{X}}^n, \hat{\boldsymbol{s}}}(x^n, Z^n) - L_{\hat{\boldsymbol{X}}^n, \boldsymbol{s}}(x^n, Z^n) \right\} > \epsilon \right)$$

[9]All inequalities between random variables in this proof are in almost sure sense.

$$\Pr\Big(L_{\hat{\boldsymbol{X}}^n, \boldsymbol{s}_k}(x_{k+1}^{n-k}, Z^n) - \tilde{L}_{\boldsymbol{S}_k}(Z^n) > \epsilon\Big)$$

$$= \Pr\Bigg(\sum_{d=0}^{k} \sum_{\tau \in \mathcal{I}_d} \Big\{\Lambda\big(x_\tau, s_{k,\tau}(\boldsymbol{C}_\tau, Z_\tau)\big) - \ell\big(Z_\tau, s_{k,\tau}(\boldsymbol{C}_\tau, \cdot)\big)\Big\} > (n - 2k)\epsilon\Bigg)$$

$$\leq \sum_{d=0}^{k} \Pr\Bigg(\sum_{\tau \in \mathcal{I}_d} \Big\{\Lambda\big(x_\tau, s_{k,\tau}(\boldsymbol{C}_\tau, Z_\tau)\big) - \ell\big(Z_\tau, s_{k,\tau}(\boldsymbol{C}_\tau, \cdot)\big)\Big\} > (n - 2k)\gamma_d \epsilon\Bigg) \qquad (54)$$

$$\leq \sum_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} P\Big(L_{\hat{\boldsymbol{X}}^n, \hat{\boldsymbol{s}}}(x^n, Z^n) - L_{\hat{\boldsymbol{X}}^n, \boldsymbol{s}}(x^n, Z^n) > \epsilon\Big) \qquad (46)$$

$$\leq \underbrace{\sum_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} P\Big(L_{\hat{\boldsymbol{X}}^n, \hat{\boldsymbol{s}}}(x^n, Z^n) - \tilde{L}_{\hat{\boldsymbol{S}}}(Z^n) > \epsilon/2\Big)}_{(i)}$$

$$+ \underbrace{\sum_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} P\Big(\tilde{L}_{\hat{\boldsymbol{S}}}(Z^n) - L_{\hat{\boldsymbol{X}}^n, \boldsymbol{s}}(x^n, Z^n) > \epsilon/2\Big)}_{(ii)} \qquad (47)$$

where (46) follows from the union bound, and (47) follows from adding and subtracting $\tilde{L}_{\hat{\boldsymbol{S}}}(Z^n)$, and the union bound. For term (i) in (47),

$$(i) \leq \sum_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} P\Big(\max_{\boldsymbol{S}' \in \mathcal{S}_{0,m}^n} \{L_{\hat{\boldsymbol{X}}^n, \boldsymbol{s}'}(x^n, Z^n) - \tilde{L}_{\boldsymbol{S}'}(Z^n)\} > \frac{\epsilon}{2}\Big) \qquad (48)$$

$$\leq \sum_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} \sum_{\boldsymbol{S}' \in \mathcal{S}_{0,m}^n} \exp\Big(-n\frac{\epsilon^2}{2L_{\max}^2}\Big) \qquad (49)$$

where (48) follows from

$$L_{\hat{\boldsymbol{X}}^n, \hat{\boldsymbol{s}}}(x^n, Z^n) - \tilde{L}_{\hat{\boldsymbol{S}}}(Z^n)$$
$$\leq \max_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} \{L_{\hat{\boldsymbol{X}}^n, \boldsymbol{s}}(x^n, Z^n) - \tilde{L}_{\boldsymbol{S}}(Z^n)\}$$

and (49) follows from the union bound and (18). Similarly, for term (ii) in (47)

$$(ii) \leq \sum_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} P\Big(\tilde{L}_{\boldsymbol{S}}(Z^n) - L_{\hat{\boldsymbol{X}}^n, \boldsymbol{s}}(x^n, Z^n) > \frac{\epsilon}{2}\Big) \qquad (50)$$

$$\leq \sum_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} \exp\Big(-n\frac{\epsilon^2}{2L_{\max}^2}\Big) \qquad (51)$$

where (50) follows from $\tilde{L}_{\hat{\boldsymbol{S}}}(Z^n) \leq \tilde{L}_{\boldsymbol{S}}(Z^n)$, and (51) follows from (19). Therefore, continuing (47), we obtain

$$(47) \leq 2 \sum_{\boldsymbol{S} \in \mathcal{S}_{0,m}^n} \sum_{\boldsymbol{S}' \in \mathcal{S}_{0,m}^n} \exp\Big(-n\frac{\epsilon^2}{2L_{\max}^2}\Big)$$

$$= 2\Bigg[\sum_{i=0}^{m} \binom{n-1}{i} N(N-1)^i\Bigg]^2 \exp\Big(-n\frac{\epsilon^2}{2L_{\max}^2}\Big) \qquad (52)$$

$$\leq 2 \exp\Bigg(-n\Big[\frac{\epsilon^2}{2L_{\max}^2} - 2h\Big(\frac{m}{n}\Big) - \frac{2(m+1)\ln N}{n}\Big]\Bigg) \qquad (53)$$

where (52) follows from $N = |\mathcal{S}| = |\mathcal{Z}|^{|\mathcal{X}|}$ and $|\mathcal{S}_{0,m}^n| = \sum_{i=0}^{m} \binom{n-1}{i} N(N-1)^i$, and (53) follows from the assumption $m \leq \lfloor\frac{n}{2}\rfloor$ and $|\mathcal{S}_{0,m}^n| \leq N^{m+1} \exp\big(nh(\frac{m}{n})\big)$. Hence, the theorem is proved. $\square$

C. Proof of Lemma 2

We will prove (26) since the proof of (27) is essentially identical. As in [1], define

$$\mathcal{I}_d \triangleq \{t : k+1 \leq t \leq n-k, t \equiv d \bmod (k+1)\},$$

whose cardinality is denoted $n_d = \lfloor(n-d-k)/(k+1)\rfloor$. Then, by denoting $\boldsymbol{C}_t = (Z_{t-k}^{t-1}, Z_{t+1}^{t+k})$, we start the chain of inequalities, given in (54) at the top of the page, where (54) follows from the union bound, and $\{\gamma_d\}$ is a set of nonnegative constants (to be specified later) satisfying $\sum_d \gamma_d = 1$. In the sequel, for simplicity, we will denote $\Lambda\big(x_\tau, s_{k,\tau}(\boldsymbol{C}_\tau, Z_\tau)\big)$ and $\ell\big(Z_\tau, s_{k,\tau}(\boldsymbol{C}_\tau, \cdot)\big)$ in (46) as Λ_τ and ℓ_τ, respectively. Now, the collection of random variables $Z(d)$ is defined to be

$$Z(d) \triangleq \{Z_t : 1 \leq t \leq n, t \notin \mathcal{I}_d\}$$

and $z(d) \in \mathcal{Z}^{n-n_d}$ denotes a particular realization of $Z(d)$. Then, by conditioning, we have

$$(54) \leq \sum_{d=0}^{k} \sum_{z(d) \in \mathcal{Z}^{n-n_d}} \Pr(Z(d) = z(d))$$

$$\times \Pr\Bigg(\sum_{\tau \in \mathcal{I}_d} \{\Lambda_\tau - \ell_\tau\} > (n-2k)\gamma_d\epsilon \Bigg| Z(d) = z(d)\Bigg) \qquad (55)$$

and let P_d denote the conditional probability of (55). Now, since $\{Z_\tau\}_{\tau \in \mathcal{I}_d}$ are all conditionally independent given $Z(d) = z(d)$, the summation in P_d beomes

$$\sum_{\tau \in \mathcal{I}_d} \Big\{\Lambda\big(x_\tau, s_{k,\tau}(\boldsymbol{c}_\tau, Z_\tau)\big) - \ell\big(Z_\tau, s_{k,\tau}(\boldsymbol{c}_\tau, \cdot)\big)\Big\}$$

given $\{Z_\tau\}_{\tau \in \mathcal{I}_d}$, which is the sum of the differences of the true and estimated losses of the symbol-by-symbol denoisers $s_{k,\tau}(\boldsymbol{c}_\tau, \cdot)$ over $\tau \in \mathcal{I}_d$. Thus, we can apply (18), and obtain

$$P_d = \Pr\Bigg(\sum_{\tau \in \mathcal{I}_d} \{\Lambda_\tau - \ell_\tau\} > n_d$$
$$\cdot \frac{(n-2k)\gamma_d\epsilon}{n_d} \Bigg| Z(d) = z(d)\Bigg)$$

$$\leq \exp\Big(-\frac{2(n-2k)^2\gamma_d^2\epsilon^2}{L_{\max}^2 n_d}\Big). \qquad (56)$$

$$\Pr\Big(L_{\hat{\boldsymbol{X}}^n,\hat{s}_{k,m}}(x_{k+1}^{n-k},Z^n) - D_{k,m}(x^n,Z^n) > \epsilon\Big)$$

$$= P\Big(\max_{\boldsymbol{S}\in\mathcal{S}_{k,m}^n}\{L_{\hat{\boldsymbol{X}}^n,\hat{s}_{k,m}}(x_{k+1}^{n-k},Z^n) - L_{\hat{\boldsymbol{X}}^n,\boldsymbol{s}}(x_{k+1}^{n-k},Z^n)\} > \epsilon\Big)$$

$$\leq \sum_{\boldsymbol{S}\in\mathcal{S}_{k,m}^n} P\Big(L_{\hat{\boldsymbol{X}}^n,\hat{s}_{k,m}}(x_{k+1}^{n-k},Z^n) - L_{\hat{\boldsymbol{X}}^n,\boldsymbol{s}}(x_{k+1}^{n-k},Z^n) > \epsilon\Big) \tag{58}$$

$$\leq \sum_{\boldsymbol{S}\in\mathcal{S}_{k,m}^n}\Big\{P\Big(L_{\hat{\boldsymbol{X}}^n,\hat{s}_{k,m}}(x_{k+1}^{n-k},Z^n) - \tilde{L}_{\hat{\boldsymbol{S}}_{k,m}}(Z^n) > \frac{\epsilon}{2}\Big) + P\Big(\tilde{L}_{\hat{\boldsymbol{S}}_{k,m}}(Z^n) - L_{\hat{\boldsymbol{X}}^n,\boldsymbol{s}}(x_{k+1}^{n-k},Z^n) > \frac{\epsilon}{2}\Big)\Big\} \tag{59}$$

$$\leq 2(k+1)\sum_{\boldsymbol{S}\in\mathcal{S}_{k,m}^n}\sum_{\boldsymbol{S}'\in\mathcal{S}_{k,m}^n}\exp\Big(-\frac{(n-2k)\epsilon^2}{2(k+1)L_{\max}^2}\Big) \tag{60}$$

$$= 2(k+1)\Big[\sum_{i=0}^{m(\boldsymbol{c})}\binom{n(\boldsymbol{c})-1}{i}N(N-1)^i\Big]^{2|\boldsymbol{C}_k|}\exp\Big(-\frac{(n-2k)\epsilon^2}{2(k+1)L_{\max}^2}\Big) \tag{61}$$

Following [1], we choose $\gamma_d = \frac{\sqrt{n_d}}{\sum_j \sqrt{n_j}}$, and from the Cauchy–Schwarz inequality and $\sum_d n_d = n-2k$, we arrive at

$$\frac{n_d}{\gamma_d^2} \leq (k+1)\sum_{d=0}^k n_d = (k+1)(n-2k)$$

and hence

$$P_d \leq \exp\Big(-\frac{2(n-2k)\epsilon^2}{(k+1)L_{\max}^2}\Big). \tag{57}$$

Therefore, plugging (57) into (55), we finally have

$$(55) \leq (k+1)\exp\Big(-\frac{2(n-2k)\epsilon^2}{(k+1)L_{\max}^2}\Big),$$

which proves the lemma. □

D. Proof of Theorem 3

The proof resembles that of Theorem 2. Consider (58)–(61) at the top of the page, where (58)–(59) follow similarly as in (46)–(47); (60) follows from arguments similar to (48), (50), and Lemma 2 (which plays the role that Lemma 1 played there); and (61) follows from $|\mathcal{S}_{k,m}^n| = [(\sum_{i=0}^{m(\boldsymbol{c})}\binom{n(\boldsymbol{c})-1}{i}N(N-1)^i]^{|\boldsymbol{C}_k|}$. Now, for all $\boldsymbol{c}\in\boldsymbol{C}_k$, clearly

$$\sum_{i=0}^{m(\boldsymbol{c})}\binom{n(\boldsymbol{c})-1}{i}N(N-1)^i$$

$$\leq \sum_{i=0}^m \binom{n-2k-1}{i}N(N-1)^i \tag{62}$$

$$\leq N^{m+1}\exp\Big((n-2k)h\Big(\frac{m}{n-2k}\Big)\Big) \tag{63}$$

where (62) follows from $m(\boldsymbol{c}) \leq m$ and $n(\boldsymbol{c}) \leq n-2k$, and (63) follows similarly as the reason for (53). Therefore, together with $|\boldsymbol{C}_k| = |\mathcal{Z}|^{2k}$, we have

$$(61) \leq 2(k+1)\exp\Big(-(n-2k)\cdot\Big[\frac{\epsilon^2}{2(k+1)L_{\max}^2}$$

$$-2|\mathcal{Z}|^{2k}\cdot\Big\{h\Big(\frac{m}{n-2k}\Big)+\frac{(m+1)\ln N}{n-2k}\Big\}\Big]\Big) \tag{64}$$

which proves the theorem. □

E. Proof of Claim 1

For part a), to show the necessity first, suppose $c_1 \geq \frac{1}{2\log|\mathcal{Z}|}$. Then, from $|\mathcal{Z}|^{2k} = n^{2k\log|\mathcal{Z}|/\log n}$, we have

$$2|\mathcal{Z}|^{2k}\cdot\Big\{h\Big(\frac{m}{n-2k}\Big)+\frac{(m+1)\ln N}{n-2k}\Big\}$$

$$= \Omega\Big(n^{2k\log|\mathcal{Z}|/\log n}\cdot\frac{m}{n}\log\frac{n}{m}\Big)$$

which will grow to infinity as n grows, even when m is fixed. Therefore, the right-hand side of (32) is not summable. On the other hand, $k = c_1\log n$ with $c_1 < \frac{1}{2\log|\mathcal{Z}|}$ is readily verified to suffice for the summability, provided that $m = m_n$ grows at any subpolynomial rate, i.e., grows more slowly than n^α for any $\alpha > 0$ (e.g., $c_2\log n$).

For part b), to show the necessity, suppose $m = \Theta(n)$. Then, $h\Big(\frac{m}{n-2k}\Big)+\frac{(m+1)\ln N}{n-2k} = \Theta(1)$, and, thus, for sufficiently small ϵ,

$$\frac{\epsilon^2}{2(k+1)L_{\max}^2}-2|\mathcal{Z}|^{2k}\cdot\Big\{h\Big(\frac{m}{n-2k}\Big)+\frac{(m+1)\log N}{n-2k}\Big\} < 0$$

even for k fixed. Therefore, the right-hand side of (32) is not summable. Hence, $m = o(n)$ is necessary for the summability. For sufficiency, suppose $m = m_n$ is any rate, such that $\lim_{n\to\infty}\frac{m_n}{n} = 0$. Then

$$\frac{\epsilon^2}{2(k+1)L_{\max}^2} - 2|\mathcal{Z}|^{2k}$$

$$\cdot\Big\{h\Big(\frac{m}{n-2k}\Big)+\frac{(m+1)\log N}{n-2k}\Big\}$$

$$= \frac{1}{k}\Big\{\frac{\epsilon^2}{2(1+\frac{1}{k}L_{\max}^2)}-2k|\mathcal{Z}|^{2k}\cdot O\Big(\Big(\frac{m_n}{n}\Big)^{1-\delta}\Big)\Big\} \tag{65}$$

for some $\delta > 0$. Thus, if k grows sufficiently slowly that

$$k|\mathcal{Z}|^{2k} = o\Big(\Big(\frac{n}{m_n}\Big)^{1-\delta}\Big)$$

then (65) becomes positive for sufficiently large n, and the right-hand side of (32) becomes summable. □

$$
EL_{\hat{\boldsymbol{X}}_{\mathrm{univ}}^{n,k,m}}(X^n, Z^n) - \mathbb{D}(P_{X^n}, \boldsymbol{\Pi})
$$
$$
= \underbrace{EL_{\hat{\boldsymbol{X}}_{\mathrm{univ}}^{n,k,m}}(X^n, Z^n) - E\left[\min_{\boldsymbol{S} \in \mathcal{S}_{k,m}^n} E\left[L_{\hat{\boldsymbol{X}}^n, \boldsymbol{s}}(X^n, Z^n)\Big| A^n\right]\right]}_{(\mathrm{i})} + \underbrace{E\left[\min_{\boldsymbol{S} \in \mathcal{S}_{k,m}^n} E\left[L_{\hat{\boldsymbol{X}}^n, \boldsymbol{s}}(X^n, Z^n)\Big| A^n\right]\right] - \mathbb{D}(P_{X^n}, \boldsymbol{\Pi})}_{(\mathrm{ii})}. \quad (74)
$$

F. Proof of Theorem 4

First, denote the random variable

$$
A_{k,m}^n \triangleq L_{\hat{\boldsymbol{X}}_{\mathrm{univ}}^{n,k,m}}(x_{k+1}^{n-k}, Z^n) - D_{k,m}(x^n, Z^n).
$$

Then, for part a), we have

$$
L_{\hat{\boldsymbol{X}}_{\mathrm{univ}}^{n,k,m}}(x^n, Z^n) - D_{k,m}(x^n, Z^n) \leq \frac{2k\Lambda_{\max}}{n} + A_{k,m}^n \quad \text{a.s.}
$$

Since the maximal rate for k is $c_1 \log n$ as specified in Claim 1, $\lim_{n \to \infty} \frac{2k_n \Lambda_{\max}}{n} = 0$. Furthermore, from the summability condition on k and m, Theorem 3, and the Borel–Cantelli lemma, we get $\lim_{n \to \infty} A_{k_n, m_n}^n = 0$ with probability 1, which proves part a). To prove part b), note that, for any $\epsilon > 0$

$$
E\left[L_{\hat{\boldsymbol{X}}_{\mathrm{univ}}^{n,k,m}}(x^n, Z^n) - D_{k,m}(x^n, Z^n)\right]
$$
$$
\leq \frac{2k\Lambda_{\max}}{n} + E(A_{k,m}^n)
$$
$$
= \frac{2k\Lambda_{\max}}{n} + E(A_{k,m}^n | A_{k,m}^n \leq \epsilon)\Pr(A_{k,m}^n \leq \epsilon)
$$
$$
\quad + E(A_{k,m}^n | A_{k,m}^n > \epsilon)\Pr(A_{k,m}^n > \epsilon)
$$
$$
\leq \frac{2k\Lambda_{\max}}{n} + \epsilon + \Lambda_{\max} \cdot \Pr(A_{k,m}^n > \epsilon)
$$
$$
\leq \frac{2k\Lambda_{\max}}{n} + \epsilon + \Lambda_{\max}
$$
$$
\quad \cdot (\text{right-hand side of } (32)). \quad (66)
$$

From the proof of Claim 1, the condition of Theorem 4 requires $k = k_n$ and $m = m_n$ to satisfy

$$
\lim_{n \to \infty} k_n |\mathcal{Z}|^{2k_n} \left(\frac{m_n}{n}\right)^{1-\delta} = 0
$$

for some $\delta > 0$. Therefore, if we set

$$
\epsilon^2 = \Theta\left(k_n |\mathcal{Z}|^{2k_n} \left(\frac{m_n}{n}\right)^{1-\delta}\right)
$$

with sufficiently large constant then, from (65), we can see that the right-hand side of (32) will decay almost exponentially, which is much faster than $\epsilon = \Theta\left(\sqrt{k_n |\mathcal{Z}|^{2k_n}(\frac{m_n}{n})^{1-\delta}}\right)$. Hence, from (66), we conclude that

$$
E(A_{k,m}^n) = O\left(\sqrt{k_n |\mathcal{Z}|^{2k_n}(\frac{m_n}{n})^{1-\delta}}\right)
$$

which results in part b). $\qquad \square$

G. Proof of Theorem 5

The fact that $m = \Theta(n)$ implies the existence of $\alpha > 0$, such that $m \geq n\alpha$ for all sufficiently large n. Let $\boldsymbol{X}$ be the process formed by concatenating independent and identically distributed (i.i.d.) blocks of length $\lceil 1/\alpha \rceil$, each block consisting of the same repeated symbol chosen uniformly from $\mathcal{X}$. The first observation to note is that, for all n large enough that $m \geq n\alpha$

$$
D_{0,m}(X^n, Z^n) = 0 \quad \text{a.s.} \quad (67)
$$

This is because, by construction, X^n is, with probability 1, piecewise constant with constancy subblocks of length, at least, $\lceil 1/\alpha \rceil$. Thus, a genie with access to X^n can choose a sequence of symbol-by-symbol schemes (in fact, ignoring the noisy sequence), with less than $n\alpha$ (and, therefore, less than m) switches, that perfectly recover X^n (and, therefore, by our assumption on the loss function, suffers zero loss). On the other hand, the assumptions on the loss function and the channel imply that, for the process $\boldsymbol{X}$ just constructed

$$
\limsup_{n \to \infty} \min_{\hat{\boldsymbol{X}}^n} EL_{\hat{\boldsymbol{X}}^n}(X^n, Z^n) > 0 \quad (68)
$$

since even the Bayes-optimal scheme for this process incurs a positive loss, with a positive probability, on each $\lceil 1/\alpha \rceil$ super-symbol. Thus, we get

$$
E\left\{\limsup_{n \to \infty} E\left[L_{\hat{\boldsymbol{X}}^n}(X^n, Z^n) - D_{0,m}(X^n, Z^n) | X^n\right]\right\} \quad (69)
$$
$$
\geq \limsup_{n \to \infty} E\left[L_{\hat{\boldsymbol{X}}^n}(X^n, Z^n) - D_{0,m}(X^n, Z^n)\right] \quad (70)
$$
$$
= \limsup_{n \to \infty} EL_{\hat{\boldsymbol{X}}^n}(X^n, Z^n) \quad (71)
$$
$$
\geq \limsup_{n \to \infty} \min_{\hat{\boldsymbol{X}}^n} EL_{\hat{\boldsymbol{X}}^n}(X^n, Z^n) > 0 \quad (72)
$$

where (70) follows from Fatou's lemma[10] and $L_{\hat{\boldsymbol{X}}^n}(X^n, Z^n) - D_{0,m}(X^n, Z^n) \leq \Lambda_{\max}$; (71) follows from (67); and (72) follows from (68). In particular, there must be one particular individual sequence $\boldsymbol{x} \in \mathcal{X}^\infty$ for which the expression inside the curled brackets of (69) is positive, i.e.,

$$
\limsup_{n \to \infty} E\left[L_{\hat{\boldsymbol{X}}^n}(X^n, Z^n) - D_{0,m}(X^n, Z^n) | X^n = x^n\right] > 0, \quad (73)
$$

which is equivalent to (35). $\qquad \square$

H. Proof of Theorem 6

First, by adding and subtracting the same terms, we obtain (74) at the top of the page. We will consider term (i) and term (ii) separately. For term (i), see (75)–(77) at the top of the following page, where (75) follows from upper-bounding and omitting the losses for time instances $t \leq k$ and $t > n - k$ in the first and second terms of (i), respectively; (76) follows from exchanging

---

[10]Fatou's lemma [40, Sec. 1.3]: If $X_n \geq 0$, then $\liminf_{n \to \infty} EX_n \geq E(\liminf_{n \to \infty} X_n)$. We apply Fatou's lemma to $\Lambda_{\max} - L_{\hat{\boldsymbol{X}}^n}(X^n, Z^n) - D_{0,m}(X^n, Z^n)$ to get (70).

$$(i) = EL_{\hat{\boldsymbol{X}}^{n,k,m}_{\text{univ}}}(X^n, Z^n) - E\Big[\min_{\boldsymbol{S} \in \mathcal{S}^n_{k,m}} E\big[L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(X^n, Z^n)\big|A^n\big]\Big]$$

$$\leq \frac{2k\Lambda_{\max}}{n} + \frac{n - 2k}{n} \cdot \Big[EL_{\hat{\boldsymbol{X}}^{n,k,m}_{\text{univ}}}(X^{n-k}_{k+1}, Z^n) - E\Big[\min_{\boldsymbol{S} \in \mathcal{S}^n_{k,m}} E\big[L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(X^{n-k}_{k+1}, Z^n)\big|A^n\big]\Big]\Big] \tag{75}$$

$$\leq \frac{2k\Lambda_{\max}}{n} + \frac{n - 2k}{n} \cdot E\Big[L_{\hat{\boldsymbol{X}}^{n,k,m}_{\text{univ}}}(X^{n-k}_{k+1}, Z^n) - \min_{\boldsymbol{S} \in \mathcal{S}^n_{k,m}} L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(X^{n-k}_{k+1}, Z^n)\Big] \tag{76}$$

$$\leq \frac{2k\Lambda_{\max}}{n} + E\Big[L_{\hat{\boldsymbol{X}}^{n,k,m}_{\text{univ}}}(X^{n-k}_{k+1}, Z^n) - D_{k,m}(X^n, Z^n)\Big], \tag{77}$$

the minimum with the expectation and the definition of iterated expectation, i.e.,

$$E\Big[\min_{\boldsymbol{S} \in \mathcal{S}^n_{k,m}} E\big[L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(X^{n-k}_{k+1}, Z^n)\big|A^n\big]\Big]$$
$$\geq E\Big[E\big[\min_{\boldsymbol{S} \in \mathcal{S}^n_{k,m}} L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(X^{n-k}_{k+1}, Z^n)\big|A^n\big]\Big]$$
$$= E\Big[\min_{\boldsymbol{S} \in \mathcal{S}^n_{k,m}} L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(X^{n-k}_{k+1}, Z^n)\Big]$$

and (77) follows from the definition (29) and $\frac{n-2k}{n} \leq 1$.

For term (ii), we bound the first term in (ii) as

$$E\Big[\min_{\boldsymbol{S} \in \mathcal{S}^n_{k,m}} E\big[L_{\hat{\boldsymbol{X}}^n, \boldsymbol{S}}(X^{n-k}_{k+1}, Z^n)\big|A^n\big]\Big]$$
$$\leq \frac{2k(m+1)\Lambda_{\max}}{n}$$
$$+ \frac{1}{n}E\Big[\min_{\boldsymbol{S} \in \mathcal{S}^n_{k,m}} E\Big[\sum_{i=1}^{r+1}\sum_{j=\tau_{i-1}+k+1}^{\tau_i - k}$$

$$\times \Lambda(X_j, s_{k,j}(Z^{j+k}_{j-k}))\Big|A^n\Big]\Big] \tag{78}$$

by upper-bounding the losses with $\Lambda_{\max}$ on the boundary of the shifting points. Now, let $\boldsymbol{P}_{X_j|Z^l_i, A^n} \in \mathbb{R}^{|\mathcal{X}|}$ denote the $|\mathcal{X}|$-dimensional probability vector whose xth component is $\Pr(X_j = x|Z^l_i, A^n)$. Then, we can bound the second term in (78) by the chain of inequalities given in (79)–(84) at the bottom of the page, where (80) follows from the stationarity of the distribution in each block as well as the fact that the combination of the best kth-order sliding-window denoiser for each block is in $\mathcal{S}^n_{k,m}$ and achieves the minimum in (79); (81) follows from conditioning; (82) follows from the definition (2); (83) follows from the stationarity of the distribution in each ith block; and (84) follows from adding more nonnegative terms.

For the second term in (ii), we first define

$$n_i(A^n) \triangleq \tau_i(A^n) - \tau_{i-1}(A^n)$$

as the length of the ith block, for $1 \leq i \leq r(A^n)+1$. Obviously, $n_i(A^n)$ also depends on A^n, and, thus, is a random variable,

$$\frac{1}{n}E\Big[\min_{\boldsymbol{S} \in \mathcal{S}^n_{k,m}} E\Big[\sum_{i=1}^{r+1}\sum_{j=\tau_{i-1}+k+1}^{\tau_i - k} \Lambda(X_j, s_{k,j}(Z^{j+k}_{j-k}))\Big|A^n\Big]\Big] \tag{79}$$

$$= \frac{1}{n}E\Big[\sum_{i=1}^{r+1}\sum_{j=\tau_{i-1}+k+1}^{\tau_i - k} \min_{s_k \in \mathcal{S}_k} E\Big[\Lambda(X_j, s_k(Z^{j+k}_{j-k}))\Big|A^n\Big]\Big] \tag{80}$$

$$= \frac{1}{n}E\Big[\sum_{i=1}^{r+1}\sum_{j=\tau_{i-1}+k+1}^{\tau_i - k} \sum_{z^k_{-k} \in \mathcal{Z}^{2k+1}} P(Z^{j+k}_{j-k} = z^k_{-k}|A^n) \times \min_{\hat{x} \in \hat{\mathcal{X}}} E\Big[\Lambda(X_j, \hat{x})|Z^{j+k}_{j-k} = z^k_{-k}, A^n\Big]\Big] \tag{81}$$

$$= \frac{1}{n}E\Big[\sum_{i=1}^{r+1}\sum_{j=\tau_{i-1}+k+1}^{\tau_i - k} \sum_{z^k_{-k} \in \mathcal{Z}^{2k+1}} P(Z^{j+k}_{j-k} = z^k_{-k}|A^n) \times U_\Lambda(\boldsymbol{P}_{X_j|Z^{j+k}_{j-k}=z^k_{-k}, A^n})\Big] \tag{82}$$

$$= \frac{1}{n}E\Big[\sum_{i=1}^{r+1}\sum_{j=\tau_{i-1}+k+1}^{\tau_i - k} E\big[U_\Lambda(\boldsymbol{P}_{X_j|Z^{j+k}_{j-k}, A^n})\big|A^n\big]\Big]$$

$$= \frac{1}{n}E\Big[\sum_{i=1}^{r+1}\sum_{j=\tau_{i-1}+k+1}^{\tau_i - k} E\big[U_\Lambda(\boldsymbol{P}^{(A_{\tau_i})}_{X_0|Z^k_{-k}})\big|A^n\big]\Big] \tag{83}$$

$$\leq \frac{1}{n}E\Big[\sum_{i=1}^{r+1}\sum_{j=\tau_{i-1}+1}^{\tau_i} E\big[U_\Lambda(\boldsymbol{P}^{(A_{\tau_i})}_{X_0|Z^k_{-k}})\big|A^n\big]\Big] \tag{84}$$

but we again suppress A^n for brevity and denote it as n_i. Then, similar to the first term in (ii), we obtain

$$
\mathbb{D}(P_{X^n}, \mathbf{\Pi})
$$
$$
= \min_{\hat{\boldsymbol{X}}^n \in \mathcal{D}_n} EL_{\hat{X}^n}(X^n, Z^n)
$$
$$
= \frac{1}{n} \min_{\hat{\boldsymbol{X}}^n \in \mathcal{D}_n} E\Big[E\Big[\sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} \Lambda(X_j, \hat{X}_j(Z^n))\Big|A^n\Big]\Big]
$$
$$
\geq \frac{1}{n} E\Big[\sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} \min_{\hat{X}:\mathcal{Z}^n \to \hat{\mathcal{X}}} E\Big[\Lambda(X_j, \hat{X}(Z^n))\Big|A^n\Big]\Big]
$$
$$(85)$$
$$
= \frac{1}{n} E\Big[\sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} \min_{\hat{X}:\mathcal{Z}^{n_i} \to \hat{\mathcal{X}}} E\Big[\Lambda(X_j, \hat{X}(Z_{\tau_{i-1}+1}^{\tau_i}))\Big|A^n\Big]\Big]
$$
$$(86)$$
$$
= \frac{1}{n} E\Big[\sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} E\big[U_\Lambda(\boldsymbol{P}_{X_j|Z_{\tau_{i-1}+1}^{\tau_i}, A^n})\big|A^n\big]\Big]
$$
$$
= \frac{1}{n} E\Big[\sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} E\big[U_\Lambda(\boldsymbol{P}_{X_0|Z_{\tau_{i-1}-j+1}^{\tau_i-j}}^{(A_{\tau_i})})\big|A^n\big]\Big]
$$
$$(87)$$
$$
\geq \frac{1}{n} E\Big[\sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} E\big[U_\Lambda(\boldsymbol{P}_{X_0|Z_{-\infty}^\infty}^{(A_{\tau_i})})\big|A^n\big]\Big]
$$
$$(88)$$

where (85) follows from exchanging the minimum with expectation; (86) follows from the conditional independence between different blocks, given A^n; (87) follows from the stationarity of the distribution in each block, and (88) follows from [1, Lemma 4 (1)].[11] Therefore, from (78), (84), and (88), we obtain

$$
\text{(ii)} = \min_{\boldsymbol{S} \in \mathcal{S}_{k,m}^n} EL_{\hat{X}^n, s}(X^n, Z^n) - \mathbb{D}(P_{X^n}, \mathbf{\Pi})
$$
$$
\leq \frac{2k(m+1)\Lambda_{\max}}{n}
$$
$$
+ \frac{1}{n} E\Big[\sum_{i=1}^{r+1} \sum_{j=\tau_{i-1}+1}^{\tau_i} E\Big[U_\Lambda\Big(\boldsymbol{P}_{X_0|Z_{-k}^k}^{(A_{\tau_i})}\Big)\Big|A^n\Big]
$$
$$
- E\Big[U_\Lambda\Big(\boldsymbol{P}_{X_0|Z_{-\infty}^\infty}^{(A_{\tau_i})}\Big)\Big|A^n\Big]\Big]
$$
$$
= \frac{2k(m+1)\Lambda_{\max}}{n}
$$
$$
+ E\Big[\sum_{i=1}^{r+1} \frac{n_i}{n} \cdot \Big\{E\Big[U_\Lambda\Big(\boldsymbol{P}_{X_0|Z_{-k}^k}^{(A_{\tau_i})}\Big)\Big|A^n\Big]
$$
$$
- E\Big[U_\Lambda\Big(\boldsymbol{P}_{X_0|Z_{-\infty}^\infty}^{(A_{\tau_i})}\Big)\Big|A^n\Big]\Big\}\Big].
$$
$$(89)$$

Now, observe that, regardless of A^n, the sequence of numbers $\{\frac{n_i}{n}\}_{i=1}^{r+1}$ forms a probability distribution, since $\sum_{i=1}^{r+1} \frac{n_i}{n} = 1$ and $\frac{n_i}{n} \geq 0$ for all i, with probability 1. Then, based on the fact

[11]For $k, l > 0$, $E[U(\boldsymbol{P}_{X_0|Z_{-k}^l}]$ is decreasing in both k and l.

that the average is less than the maximum, we obtain the further upper bound

$$
(89) \leq \frac{2k(m+1)\Lambda_{\max}}{n}
$$
$$
+ E\Big[\max_{i \in \{1,\dots,M\}} \Big\{E[U_\Lambda(\boldsymbol{P}_{X_0|Z_{-k}^k}^{(i)})]
$$
$$
- E[U_\Lambda(\boldsymbol{P}_{X_0|Z_{-\infty}^\infty}^{(i)})]\Big\}\Big].
$$
$$(90)$$

The remaining argument to prove the theorem is to show that the upper bounds (77) and (90) converge to 0 as n tends to infinity. First, from the given condition on $k = k_n$ and $m = m_n$, the maximal allowable growth rate for k is $k = c_1 \log n$, which leads to $\lim_{n \to \infty} \frac{2k\Lambda_{\max}}{n} = 0$. In addition, the condition requires $m = o(n)$, and k to be sufficiently slow, such that $k|\mathcal{Z}|^{2k} = o((\frac{n}{m})^{1-\delta})$, which implies $k = o(\frac{n}{m})$. Therefore

$$
\lim_{n \to \infty} \frac{2k(m+1)\Lambda_{\max}}{n} = 0.
$$

Furthermore, from conditioning on X^n, bounded convergence theorem, and part b) of Theorem 4, we obtain

$$
\lim_{n \to \infty} E[L_{\hat{X}_{\mathrm{univ}}^{n,k,m}}(X_{k+1}^{n-k}, Z^n) - D_{k,m}(X^n, Z^n)] = 0.
$$

Thus, we have

$$
\limsup_{n \to \infty} \Big[EL_{\hat{X}_{\mathrm{univ}}^{n,k,m}}(X^n, Z^n) - \mathbb{D}(P_{X^n}, \mathbf{\Pi})\Big]
$$
$$
\leq \limsup_{n \to \infty} E\Big[\max_{i \in \{1,\dots,M\}} \Big\{E\Big[U_\Lambda\Big(\boldsymbol{P}_{X_0|Z_{-k}^k}^{(i)}\Big)\Big]
$$
$$
- E\Big[U_\Lambda\Big(\boldsymbol{P}_{X_0|Z_{-\infty}^\infty}^{(i)}\Big)\Big]\Big\}\Big]
$$
$$
\leq E\Big[\limsup_{n \to \infty} \max_{i \in \{1,\dots,M\}} \Big\{E\Big[U_\Lambda\Big(\boldsymbol{P}_{X_0|Z_{-k}^k}^{(i)}\Big)\Big]
$$
$$
- E\Big[U_\Lambda\Big(\boldsymbol{P}_{X_0|Z_{-\infty}^\infty}^{(i)}\Big)\Big]\Big\}\Big]
$$
$$(91)$$
$$
= 0
$$
$$(92)$$

where (91) follows from the Fatou's lemma, and (92) follows from [1, Lemma 4 (2)], and M being finite. Since it is clear that $\liminf_{n \to \infty}[EL_{\hat{X}_{\mathrm{univ}}^{n,k,m}}(X^n, Z^n) - \mathbb{D}(P_{X^n}, \mathbf{\Pi})] \geq 0$ by definition of $\mathbb{D}(P_{X^n}, \mathbf{\Pi})$, the theorem is proved. $\square$

Remark: As in [1, Theorem 3], the convergence rate in (42) may depend on $P_{\boldsymbol{X}}$, and there is no vanishing upper bound on this rate that holds for all $P_{\boldsymbol{X}} \in \mathcal{P}\{m_n\}$. However, we can glean some insight into the convergence rate from (i) and (ii): whereas the term (i) is uniformly upper-bounded for all $P_{\boldsymbol{X}} \in \mathcal{P}\{m_n\}$,[12] the rate at which term (ii) vanishes depends on $P_{\boldsymbol{X}}$. In general, we observe that the slower the rate of increase of $k = k_n$, the faster the convergence in (i), but the convergence

[12]Recall part b) of Theorem 4, where a uniform bound (uniform in the underlying individual sequence) on $E\Big[L_{\hat{X}_{\mathrm{univ}}^{n,k,m}}(x^n, Z^n) - D_{k,m}(x^n, Z^n)\Big]$ was provided in the semi-stochastic setting. Clearly, in the stochastic setting the same bound holds on $E\Big[L_{\hat{X}_{\mathrm{univ}}^{n,k,m}}(X^n, Z^n) - D_{k,m}(X^n, Z^n)\Big]$, regardless of the distribution of X^n.

in (ii) is slower. With respect to the rate of increase of m_n, the slower it is, the faster the convergence in (i), but whether or not the convergence in (ii) is accelerated by a slower rate of increase of m_n may depend on the underlying process distribution $P_{\mathbf{X}}$.

ACKNOWLEDGMENT

T. Moon wishes to thank Prof. Manfred Warmuth for introducing him to a substantial amount of literature on expert tracking problems in online learning.

REFERENCES

[1] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 5–28, Jan. 2005.

[2] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 5, pp. 5530–536, Sep. 1978.

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

[4] T. Weissman, E. Ordentlich, M. Weinberger, A. Somekh-Baruch, and N. Merhav, "Universal filtering via prediction," *IEEE Trans. Inf. Theory*, vol. 53, no. 4, pp. 1253–1264, Apr. 2007.

[5] T. Weissman, "How to filter an individual sequence with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3831–3841, Aug. 2008.

[6] D. Blackwell, "Controlled random walks," in *Proc. Int. Congr. Mathematicians*, Amsterdam, The Netherlands, 1956, vol. 3, pp. 336–338.

[7] D. Blackwell, "An analog of the minimax theorem for vector payoffs," *Pacific J. Math*, vol. 6, pp. 1–8, 1956.

[8] J. Hannan, "Approximation to Bayes risk in repeated play," *Contributions to the Theory of Games*, vol. III, pp. 97–139, 1957.

[9] E. Ordentlich and T. Cover, "The cost of achieving the best portfolio in hindsight," *Math. Oper. Res.*, vol. 23, no. 4, pp. 960–982, 1998.

[10] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Inf. Comput.*, vol. 108, no. 2, pp. 212–261, 1994.

[11] V. Vovk, "Aggregating strategies," in *Proc. 3rd Annu. Workshop on Computational Learning Theory*, Rochester, NY, 1990, pp. 371–382.

[12] A. György, T. Linder, and G. Lugosi, "Efficient algorithms and minimax bounds for zero-delay lossy source coding," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2337–2347, Aug. 2004.

[13] S. Matloub and T. Weissman, "Universal zero-delay joint source-channel coding," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5240–5250, Dec. 2006.

[14] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[15] M. Herbster and M. K. Warmuth, "Tracking the best expert," *J. Mach. Learn.*, vol. 32, no. 2, pp. 151–178, 1998.

[16] O. Bousquet and M. K. Warmuth, "Tracking a small set of experts by mixing past posteriors," *J. Mach. Learn. Res.*, vol. 3, pp. 363–396, 2002.

[17] G. I. Shamir and N. Merhav, "Low-complexity sequential lossless coding for piecewise-stationary memoryless sources," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1498–1519, Jul. 1999.

[18] F. M. J. Willems, "Coding for binary independent piecewise identically distributed source," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2210–2217, Sep. 1996.

[19] S. S. Kozat and A. C. Singer, "Universal switching linear least squares prediction," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 189–204, Jan. 2007.

[20] Y. Singer, "Switching portfolios," *Int. J. Neural Syst.*, pp. 488–495, 1998.

[21] A. György, T. Linder, and G. Lugosi, "Tracking the best quantizer," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1604–1625, Apr. 2008.

[22] E. Ordentlich, M. Weinberger, and T. Weissman, "Multi-directional context sets with applications to universal denoising and compression," in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 1270–1274.

[23] D. Siegmund, "Confidence sets in change-point problems," *Int. Statist. Rev.*, vol. 56, pp. 31–48, 1989.

[24] D. Siegmund and E. S. Venkatraman, "Using the generalized likelihood ratio statistic for sequential detection of a change-point," *Ann. Statist.*, vol. 23, pp. 255–271, 1995.

[25] S. M. Oh, J. M. Rehg, and F. Dellaert, "Parameterized duration modeling for switching linear dynamic systems," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, New York, Oct. 2006, vol. 2, pp. 1694–1700.

[26] B. Mesot and D. Barber, "Switching linear dynamical systems for noise robust speech recognition," *IEEE Trans. Audio, Speech and Language Process.*, vol. 15, pp. 1850–1858, Aug. 2007.

[27] P. Lancaster and M. Tismenetsky, "The theory of matrices," *Academic*, vol. 51, no. 1, pp. 5–28, 1985.

[28] G. I. Shamir and D. J. Costello, Jr, "On the redundancy of universal lossless coding for general piecewise stationary sources," *Commun. Inf. Syst.*, vol. 1, no. 3, pp. 305–332, 2001.

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

[30] A. Lempel and J. Ziv, "Compression of two-dimensional data," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 1, pp. 2–8, Jan. 1986.

[31] A. Cohen, N. Merhav, and T. Weissman, "Scanning and sequential decision making for multi-dimensional data—Part I: The noiseless case," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3001–3020, Sep. 2007.

[32] A. Cohen, T. Weissman, and N. Merhav, "Scanning and sequential decision making for multidimensional data—Part II: The noisy case," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5609–5631, Dec. 2008.

[33] R. Zhang and T. Weissman, "Discrete denoising for channels with memory," *Commun. Inf. Syst.*, vol. 5, no. 2, pp. 257–288, 2005.

[34] G. M. Gemelos, S. Sigurjónsson, and T. Weissman, "Universal minimax discrete denoising under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3476–3497, Aug. 2006.

[35] E. Ordentlich, G. Seroussi, S. Verdú, and K. Viswanathan, "Universal algorithms for channel decoding of uncompressed sources," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2243–2262, May 2008.

[36] A. Dembo and T. Weissman, "Universal denoising for the finite-input-general-output channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1507–1517, Apr. 2005.

[37] K. Sivaramakrishnan and T. Weissman, "Universal denoising of discrete-time continuous-amplitude signals," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5632–5660, Dec. 2008.

[38] S. Jalali, S. Verdú, and T. Weissman, "A universal Wyner-Ziv scheme for discrete sources," in *Proc. IEEE Int. Symp. Information Theory*, Nice, France, Jun. 2007, pp. 1951–1955.

[39] J. Yu and S. Verdú, "Schemes for bidirectional modeling of discrete stationary sources," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4789–4807, Nov. 2006.

[40] R. Durrett, *Probability: Theory and Examples*, 3 ed. Pacific Grove, CA: Duxbury, 2005.

Taesup Moon (S'04–M'08) received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 2004 and 2008, respectively.

He joined Yahoo! Labs, Sunnyvale, CA, as a research scientist in 2008. His research interests are in information theory, statistical signal processing, machine learning, and information retrieval.

Dr. Moon was awarded the Samsung Scholarship and a fellowship from the Korean Foundation of the Advanced Studies.

Tsachy Weissman (S'99–M'02–SM'07) received the B.Sc. and Ph.D. degrees in electrical engineering from the Technion–Israel Institute of Technology, Haifa, Israel, in 1997 and 2001, respectively.

He has held postdoctoral appointments with the Statistics Department at Stanford University, Stanford, CA, and with Hewlett-Packard Laboratories, Palo, CA. Currently he is with the departments of Electrical Engineering at Stanford University and at the Technion. His research interests span information theory and its applications, and statistical signal processing. His papers thus far have focused mostly on data compression, communications, prediction, denoising, and learning. He is also inventor or coinventor of several patents in these areas and has been involved in a number of high-tech companies as a researcher or member of the technical board.

Prof Weissman recently was awarded the NSF CAREER award and a Horev fellowship for leaders in Science and Technology. He is a Robert N. Noyce Faculty Scholar of the School of Engineering at Stanford, and a recipient of the 2006 IEEE joint IT/COM societies best paper award.