Twice-universal denoising

Erik Ordentlich, Krishnamurthy Viswanathan, Marcelo J. Weinberger

HP Laboratories
HPL-2011-205

Keyword(s):
Universal denoising; universal data compression; loss estimation

Abstract:
We propose a sequence of universal denoisers motivated by the goal of extending the notion of twice-universality from universal data compression theory to the sliding window denoising setting. Given a sequence length n and a denoiser, the k-th order regret of the latter is the maximum excess expected denoising loss relative to sliding window denoisers with window length $2k+1$, where, for a given clean sequence, the expectation is over all channel realizations and the maximum is over all clean sequences of length n. We define the twice-universality penalty of a denoiser as its excess k-th order regret when compared to the k-th order regret of the DUDE with parameter k, and we are interested in denoisers with a small penalty for all k simultaneously. We consider a class of denoisers that apply one of a number of constituent denoisers based on minimizing an estimated denoising loss and establish a formal relationship between errors in the estimated denoising loss and the twice-universality penalty of the resulting denoiser. Given a sequence of window parameters k_n, increasing in n sufficiently fast, we use this approach to construct and analyze a specific sequence of denoisers that achieves a much smaller twice--universality penalty for $k < k_n$ than the sequence of DUDEs with parameter k_n.

External Posting Date: October 22, 2011 [Fulltext] Approved for External Publication
Internal Posting Date: October 22, 2011 [Fulltext]

Twice–universal denoising*

Erik Ordentlich, Krishnamurthy Viswanathan, Marcelo J. Weinberger
Hewlett Packard Labs, Palo Alto, CA 94304.

October 13, 2011

Abstract

We propose a sequence of universal denoisers motivated by the goal of extending the notion of twice–universality from universal data compression theory to the sliding window denoising setting. Given a sequence length n and a denoiser, the k-th order regret of the latter is the maximum excess expected denoising loss relative to sliding window denoisers with window length $2k+1$, where, for a given clean sequence, the expectation is over all channel realizations and the maximum is over all clean sequences of length n. We define the twice–universality penalty of a denoiser as its excess k-th order regret when compared to the k-th order regret of the DUDE with parameter k, and we are interested in denoisers with a small penalty for all k simultaneously. We consider a class of denoisers that apply one of a number of constituent denoisers based on minimizing an estimated denoising loss and establish a formal relationship between errors in the estimated denoising loss and the twice–universality penalty of the resulting denoiser. Given a sequence of window parameters k_n, increasing in n sufficiently fast, we use this approach to construct and analyze a specific sequence of denoisers that achieves a much smaller twice–universality penalty for $k < k_n$ than the sequence of DUDEs with parameter k_n.

Keywords: Universal denoising, universal data compression, loss estimation.

1 Introduction

The problem of denoising is one of signal reproduction based on noisy observations, with the quality of the reproduction being measured by a fidelity criterion. In one version of this problem, a clean *discrete* sequence x^n of length n is passed through a known discrete memoryless channel (DMC) to obtain a noisy sequence z^n, and the goal of the denoiser is to produce a reconstruction $\hat{x}^n$ whose quality is measured by a single-letter loss function. This problem is studied in [1], where a denoising algorithm, DUDE, is proposed. The DUDE algorithm takes as input a non-negative integer parameter k, computes the number of occurrences of all $2k+1$-tuples of symbols in z^n, and bases its reconstruction on these counts. It is shown in [1] that DUDE with parameter k is *universal* in the sense that, for any input sequence x^n, the difference between its loss and that of the best k-th

*This work was presented, in part, at the IEEE Intl. Symp. on Inform. Theory (ISIT), in Seoul, Korea (2009) and in Austin, Texas (2010).

order *sliding window* denoiser for the pair (x^n, z^n) (the difference being the k-th order "regret") vanishes in the limit of large n (and fixed k), both in expectation and with high probability (where the randomness comes from the known DMC). A k-th order sliding window denoiser is one whose decision at time i depends only on the window z_{i-k}^{i+k}.

The k-th order regret bound of the DUDE [1] (see also (4) below) continues to vanish for sufficiently slowly increasing sequences $k = k_n \to \infty$. It follows that DUDE with such parameters k_n (as a sequence of denoisers) competes successfully with any sequence of k'_n-th order sliding window denoisers for sufficiently slowly growing order $k'_n \leq k_n$ (including any fixed k). This property is akin to that of the Lempel–Ziv algorithm in data compression [2], which is able to compress every sequence essentially as well as any compressor with a fixed (or a slowly growing) number of states. However, this view of universality does not address the issue of how fast the range of competing window lengths should grow, or what is the additional cost, in terms of the k-th order regret, incurred over an algorithm optimized for a specific k or specific sequence k'_n, for dealing with a faster growing range of window lengths.

These are important considerations, since, ideally, we would like a universal denoiser to compete with a family of denoisers as large as possible (in other words, we would like k_n to grow at the fastest possible rate) while at the same time pay as little as possible for this additional universality. Following the above data compression analogy, such desiderata are akin to those formulated in the context of twice–universal data compression [4, 5]. For example, a particularly desirable denoiser would be one that for *all* sequences k_n, incurs a k_n-th order regret that is only negligibly more than that incurred by the DUDE of order k_n, that is, the DUDE designed specifically to compete with sliding window denoisers of order k_n. The envisioned denoiser would thus be independent of *any* sequence k_n, unlike the DUDE. One might think that the sequence of DUDEs with parameter k_n for some rapidly increasing k_n would come close to having such an ideal property. This, however, is not the case since the best that can be said about the k'_n-th order regret of such a sequence of DUDEs, when $k'_n \leq k_n$, is that it equals the k_n-th order regret, which is roughly exponential in k_n (see (4) below), thus potentially far exceeding the k'_n-order regret of a k'_n-th order DUDE. Moreover, for $k'_n > k_n$, the k'_n-th order regret of the k_n-th order DUDE is not even guaranteed to vanish, even if the k'_n-order regret of the k'_n-th order DUDE did vanish. This weakness is the result of a naive policy for dealing with a growing range of window lengths, namely one which is *data independent*. In the data compression setting, this policy corresponds to a Markov compressor which increases the Markov order sufficiently slowly in a data independent manner, as in the early days of universal data compression.

Defining the twice–universality penalty of a denoiser as its excess k-th order regret when compared to the k-th order regret of the DUDE with parameter k, we are interested in a denoiser with a small penalty for all k *simultaneously*. For such a denoiser, the k-th order regret would not only be vanishing, but would be close to the best possible. In this paper, we present a denoiser that is a step towards such an ideal denoiser in terms of these aspects of universality. The proposed algorithm, dubbed TU-DUDE, is based on loss estimation results and a variant of DUDE, and is also defined by a growing sequence k_n (though, as noted below, there is a natural "best" such sequence). The

TU-DUDE is shown to have several advantages over the DUDE. First, for k_n increasing with n sufficiently fast, a much smaller k-th order regret upper bound can be proved for k, $0 \leq k \leq k_n$, than that of the DUDE of order k_n. Moreover, for the "most ambitious" choice of k_n, denoted κ_n (to be specified in Section 3), the k-th order regret of the TU-DUDE is essentially the same as the regret of DUDE with parameter k for every k roughly in the range $\kappa_n/2 < k \leq \kappa_n$, making the penalty for competing with a *range* of window lengths rather than with a specific window length k, negligible. In contrast, as noted above, in order to be universal for a range $0 \leq k \leq k_n$, DUDE (with parameter k_n) incurs for any k in the range, under the best available bounds, a regret corresponding to the maximum order in the range, k_n, which is exponential in k_n. Second, the TU-DUDE is also shown to compete against a slightly larger family of sliding window denoisers (with growing order) than any sequence of DUDEs for any choice of parameters k_n. The above properties can be seen as a step toward twice–universality in denoising, similar to the counterpart concept in data compression, thus motivating the name TU-DUDE, for twice–universal DUDE. As discussed in Section 6.3, it should be noted, however, that the TU-DUDE falls short of being twice–universal in the data compression sense.

The TU-DUDE is based on a DUDE-like universal denoiser dubbed the D-DUDE (for reasons explained below), whose regret satisfies the same bounds from [1] as that of the DUDE with the corresponding parameter. The TU-DUDE takes the approach first proposed in [7], namely to select the value of k that minimizes an *estimate* of the loss of D-DUDE with parameter k for $0 \leq k \leq k_n$ and then to apply D-DUDE with the selected parameter. The idea is that if the expected error in estimating the loss is small for all clean sequences x^n, then the loss incurred for the selected k does not deviate much from the loss incurred for the best k for the underlying clean sequence. The D-DUDE for the best k, in turn, clearly has a smaller regret than the D-DUDE for any k. We prove this result in greater generality by showing that the existence of a loss estimator whose expected error is small for a sequence of denoisers, with regret satisfying the same bounds as the DUDE, provides a way to construct a twice–universal denoiser with a small excess regret. Finally, we also show that for a restricted subset of clean sequences, dubbed *non-pathological* sequences, the excess loss incurred by the TU-DUDE (the twice–universality penalty) may be much smaller than the bounds we can prove without any restriction on the clean sequences. We highlight scenarios for which the fraction of clean sequences that are non-pathological tends to one.

The rest of the paper is organized as follows. Basic definitions and notations are presented in Section 2. The notion of twice–universality is defined in Section 3, where we also formally state our main results. The next two sections present results and tools necessary for constructing the TU-DUDE and for proving our main theorem. A lemma establishing the connection between good loss estimation and twice universality is proved in Section 4. The specific loss estimator that we consider is presented, along with its properties, in Section 5. The TU-DUDE is then formally described in Section 6. The proof of our main theorem is also presented in this section. Finally, in Section 7, we show that stronger results are possible in the sense that if one defined the regret based on the worst-case expected excess loss over a subset of non-pathological clean sequences, a much smaller twice–universality penalty can be achieved.

2 Notation and Preliminaries

The notation we employ is similar to the one in [1]. We first define the notation we use to refer to vectors, matrices and sequences. For any matrix A, a_i will denote its i_{th} column, and for a vector $\mathbf{u}$ its i_{th} component will be denoted by u_i or $\mathbf{u}[i]$. Often, the indices may belong to any discrete set of appropriate size. For two vectors $\mathbf{u}$ and $\mathbf{v}$ of the same dimension, $\mathbf{u} \odot \mathbf{v}$ will denote the vector obtained from componentwise multiplication. For any vector or matrix A, A^T will denote transposition, for an invertible matrix A^{-T} will denote the transpose of its inverse A^{-1}, and $||A||_\infty$ will denote the largest absolute value of any entry in the matrix or vector.

For any set $\mathcal{A}$, let $\mathcal{A}^\infty$ denote the set of one-sided infinite sequences with $\mathcal{A}$-valued components, *i.e.,* $\mathbf{a} \in \mathcal{A}^\infty$ is of the form $\mathbf{a} = (a_1, a_2, \ldots)$, $a_i \in \mathcal{A}$, $i \geq 1$. For $\mathbf{a} \in \mathcal{A}^\infty$, let $a^n = (a_1, a_2, \ldots, a_n)$ and $a_i^j = (a_i, a_{i+1}, \ldots, a_j)$. More generally, we will permit the indices to be negative as well, for example, $u_{-k}^k = (u_{-k}, \ldots, u_0, \ldots, u_k)$. For positive integers k_1, k_2, and strings $s_i \in \mathcal{A}^{k_i}$, $i = 1, 2$, let $s_1 s_2$ denote the string formed by the concatenation of s_1 and s_2. Sometimes we will also refer to the i-th component of a sequence $\mathbf{a}$ by $\mathbf{a}[i]$.

We now define the parameters associated with the universal denoising problem, namely, the channel transition probabilities, the loss function, and relevant classes of denoisers. Let the sequences x^n, $z^n \in \mathcal{A}^n$, respectively denote the noiseless input to and the noisy output from a discrete memoryless channel (DMC) whose input and output alphabet are both $\mathcal{A}$. Let $M = |\mathcal{A}|$ denote the size of the alphabet and $\mathcal{M}$ the simplex of M-dimensional probability vectors. Let the matrix $\mathbf{\Pi} = \{\mathbf{\Pi}(i,j)\}_{i,j \in \mathcal{A}}$, whose components are indexed by members of $\mathcal{A}$, denote the *transition probability matrix* of the channel, where $\mathbf{\Pi}(i,j)$ is the probability that the output symbol is j when the input symbol is i. Also, for $j \in \mathcal{A}$, π_j denotes the j-th column of $\mathbf{\Pi}$. We are interested in channels whose transition matrix $\mathbf{\Pi}$ is invertible. For technical reasons stemming from our proof technique (cf. Lemma 8 and its proof), we also assume throughout that all entries of $\mathbf{\Pi}$ are strictly positive. We believe, however, that this assumption can be relaxed considerably while preserving our results.

Upon observing a noisy sequence $z^n \in \mathcal{A}^n$, a denoiser outputs a reconstruction sequence $\{\hat{x}_t\}_{t=1}^n \in \mathcal{A}^n$. The *loss matrix* $\mathbf{\Lambda} = \{\Lambda(i,j)\}_{i,j \in \mathcal{A}}$ represents the loss function associated with the denoising problem, namely, $\Lambda(i,j) \geq 0$ denotes the loss incurred by a denoiser that outputs $\hat{x} = j$ when the channel input $x = i$. Also, for $i \in \mathcal{A}$, λ_i denotes the i-th column of $\mathbf{\Lambda}$.

An *n-block denoiser* is a mapping $\hat{X}^n : \mathcal{A}^n \to \mathcal{A}^n$. For any $z^n \in \mathcal{A}^n$, let $\hat{X}^n(z^n)[i]$ denote the i-th term of the sequence $\hat{X}^n(z^n)$. For a noiseless input sequence x^n and the observed output sequence z^n, the *normalized cumulative loss* $L_{\hat{X}^n}(x^n, z^n)$ of the denoiser $\hat{X}^n$ is

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

Let $\mathcal{D}_n$ denote the class of all n-block denoisers. A *k-th order sliding window denoiser* $\hat{X}^n$ is a denoiser with the property that for all $z^n \in \mathcal{A}^n$, if $z_{i-k}^{i+k} = z_{j-k}^{j+k}$ then

$$\hat{X}^n(z^n)[i] = \hat{X}^n(z^n)[j].$$

Thus, the denoiser defines a mapping,

$$f : \mathcal{A}^{2k+1} \to \mathcal{A}$$

so that for all $z^n \in \mathcal{A}^n$

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

Let $\mathcal{S}_k$ denote the class of k-th order sliding window denoisers. In the sequel, we define the best loss obtainable for a given pair of noiseless and noisy sequences with a k-th order sliding window denoiser. For convenience, we will modify the definition of normalized cumulative loss to accomodate noiseless and noisy sequences of differing lengths.

For an individual noiseless sequence $x^n \in \mathcal{A}^n$ and a noisy sequence $z^n \in \mathcal{A}^n$, $k \geq 0$ and $n > 2k$, $D_k(x^n, z^n)$, the *k-th order minimum loss* of (x^n, z^n) is defined to be

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

the least loss incurred by any k-th order denoiser on the pair (x^n, z^n). For a given channel $\mathbf{\Pi}$ and a noiseless sequence x^n define

$$\hat{D}_k(x^n) \stackrel{\text{def}}{=} E[D_k(x^n, Z^n)] \tag{1}$$

the expected k-th order minimum loss incurred when each random noisy sequence Z^n produced when x^n is input to the channel is denoised by the best k-th order denoiser for the pair (x^n, Z^n). For any n-block denoiser $\hat{X}^n$ we define its *k-th order regret* to be

$$\hat{R}_k\Big(\hat{X}^n\Big) \stackrel{\text{def}}{=} \max_{x^n \in \mathcal{A}^n} \Big\{E\Big[L_{\hat{X}^n}\Big(x_{k+1}^{n-k}, Z^n\Big)\Big] - \hat{D}_k(x^n)\Big\}. \tag{2}$$

Note that since $D_k(x^n, z^n)$ is non-increasing in k for any fixed x^n and z^n, $\hat{D}_k(x^n)$ is non-increasing in k for all x^n so that, for all $\hat{X}^n$, $\hat{R}_k\Big(\hat{X}^n\Big)$ is non-decreasing in k. Given a non-decreasing sequence $\{k_n\}$, a sequence of denoisers $\{\hat{X}^n\}$ is *universal* for the classes $\mathcal{S}_{k_n}$ if the k_n-th order regret of $\hat{X}^n$

$$\hat{R}_{k_n}\Big(\hat{X}^n\Big) = o(1).$$

The Discrete Universal Denoiser (DUDE) was proposed in [1], and is described below. The vector $\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k)$ is defined as

$$\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k)[c_0] \stackrel{\text{def}}{=} \Big|\Big\{i : k+1 \leq i \leq n-k, z_{i-k}^{i+k} = c_{-k}^k\Big\}\Big|$$

for $c_0 \in \mathcal{A}$. Then, the DUDE with parameter k denoises according to

$$\hat{X}_{\text{DUDE}}^{n,k}(z^n)[i] = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T\Big((\mathbf{\Pi}^{-T}\mathbf{m}(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})) \odot \pi_{z_i}\Big), \tag{3}$$

where ties in the minimization are broken according to an arbitrary, but fixed rule. For all $a \in \mathcal{A}$, the a-th component of $((\mathbf{\Pi}^{-T}\mathbf{m}(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})) \odot \pi_{z_i})$ is a good estimate for the number of indices

j in x^n that take the value a when the j-th noisy symbol equals z_i and the context of the j-th noisy symbol equals $\left(z_{i-k}^{i-1}, z_{i+1}^{i+k}\right)$. The DUDE denoises by selecting that reconstruction symbol that minimizes the loss assuming this estimate is exact. Note that for the DUDE with parameter k, $\hat{X}_{\mathrm{DUDE}}^{n,k} \notin \mathcal{S}_k$ (since the effective $2k+1$ window-wise denoising mapping depends on z^n). It was shown in [1] that for all k, all sufficiently large n and all clean sequences x^n, $\hat{X}_{\mathrm{DUDE}}^{n,k}$ satisfies, for some constant[1] C independent of k, n, and x^n,

$$\hat{R}_k\left(\hat{X}_{\mathrm{DUDE}}^{n,k}\right) \le C\sqrt{\frac{M^{2k}(k+1)}{n}}. \tag{4}$$

Therefore, $\hat{X}_{\mathrm{DUDE}}^{n,k_n}$ is universal for $\mathcal{S}_{k_n}$ if $k_n M^{2k_n} = o(n)$. In [6], it was shown that the regret in (4) is close to the best possible. Specifically for most $(\mathbf{\Pi}, \Lambda)$, and all k and any sequence $\{\hat{X}^n \in \mathcal{D}_n\}$ of denoisers, as n tends to infinity

$$\hat{R}_k\left(\hat{X}^n\right) \ge \frac{c^k}{\sqrt{n}}$$

where c is a positive function of $(\mathbf{\Pi}, \Lambda)$. For certain $(\mathbf{\Pi}, \Lambda)$ pairs, c equals M, and in those cases the regret of the DUDE with parameter k is optimal up to a factor of $\sqrt{k}$. This suggests a definition of twice–universality that is based on the regret of the DUDE, rather than a more elusive optimal regret.

3 Twice–universality: definition and main results

In this paper, we explore the notion of twice–universal denoisers, namely denoisers whose regret with respect to all $\mathcal{S}_k, k \le k_n$, for a given sequence k_n, is not just vanishing but close to the best possible. We shall say that a sequence of denoisers $\hat{X}^n$ is "twice–universal" with penalty $\epsilon(k,n)$ if its regret satisfies

$$\hat{R}_k\left(\hat{X}^n\right) \le C\sqrt{\frac{M^{2k}(k+1)}{n}} + \epsilon(k,n) \tag{5}$$

for all sufficiently large n, and all k *simultaneously*. We shall sometimes refer to $\epsilon(k,n)$ satisfying (5) as the *twice–universality penalty* of $\hat{X}^n$. This definition is DUDE-specific in the sense that the penalty $\epsilon(k,n)$ is the excess loss incurred by $\hat{X}^n$ beyond the regret bound (4) achieved by DUDE with parameter k.

Note that for a sequence k_n, a sequence of DUDEs with parameter k_n satisfies

$$\hat{R}_k\left(\hat{X}_{\mathrm{DUDE}}^{n,k_n}\right) \le C\sqrt{\frac{M^{2k_n}(k_n+1)}{n}}, \tag{6}$$

for all $k \le k_n$ and n sufficiently large, since for all $\hat{X}^n$, $\hat{R}_k\left(\hat{X}^n\right)$ is non-decreasing in k. Thus, the

[1] Several constants pertaining to different results throughout the paper will be denoted by C. These are not necessarily the same constant.

DUDE with parameter k_n is provably twice–universal with penalty

$$\begin{aligned}\epsilon_{\mathrm{D},k_n}(n,k) &= C\sqrt{\frac{M^{2k_n}(k_n+1)}{n}} - C\sqrt{\frac{M^{2k}(k+1)}{n}} \\ &\approx C\sqrt{\frac{M^{2k_n}(k_n+1)}{n}}\end{aligned}$$

for $k \leq k_n$, where the latter approximation holds for k sufficiently smaller than k_n. The use of the term "provably" is due to the fact that we have computed $\epsilon_{\mathrm{D},k_n}(n,k)$ from an upper bound on $\hat{R}_k\left(\hat{X}^{n,k_n}_{\mathrm{DUDE}}\right)$, not from an exact characterization. Thus, while the actual penalty could be smaller than $\epsilon_{\mathrm{D},k_n}(n,k)$, no such reduced penalty has been proved. Note also that, based on known bounds, the best we can say about the twice–universality penalty of the DUDE with parameter k_n for $k > k_n$ is that it is $||\Lambda||_\infty$. We investigate whether it is possible to achieve a smaller penalty than $\epsilon_{\mathrm{D},k_n}(n,k)$.

Given a sequence k_n, the main result of this paper is a denoiser dubbed the TU-DUDE and denoted as $\hat{X}^{n,k_n}_{\text{TU-DUDE}}$ and the following theorem concerning its twice–universality penalty.

Theorem 1. *For all channels such that the transition matrix* $\mathbf{\Pi}$ *has only non-zero entries, given a sequence* k_n, $\hat{X}^{n,k_n}_{TU\text{-}DUDE}$ *is twice–universal with penalty* $\epsilon_{TU,k_n}(k,n) = \tilde{C}((k_n+1)^{5/4}/n^{1/4})$ *for a constant* $\tilde{C}$ *and* $k \leq k_n$.

The construction of the TU-DUDE with parameter k_n and the proof of Theorem 1 are presented in Section 6. Here, we make the following observations concerning the theorem.

In comparing the twice–universal penalty $\epsilon_{\mathrm{TU},k_n}(k,n)$ from the theorem to $\epsilon_{\mathrm{D},k_n}(n,k)$, we see that it is smaller roughly when $k_n > \log n/(4\log M)$. In particular, consider the choice $k_n = \kappa_n$ for k_n, where

$$\kappa_n = \max\left\{k \in \mathbb{Z}^+ : C\sqrt{\frac{M^{2k}(k+1)}{n}} \leq ||\Lambda||_\infty\right\} \approx \frac{\log n}{2\log M} \tag{7}$$

which is the largest k_n for which the DUDE with parameter k_n may be doing any *provable* denoising for a given n. It is thus a natural sequence to consider in light of the definition (5) of a twice–universal denoiser with a given penalty: for $k > \kappa_n$ the k-th order DUDE regret component of the overall regret fails to vanish and is thus of limited interest to track, while for $k < \kappa_n$ the k-th order DUDE may be carrying out some denoising and its performance may thus be worthwhile tracking. Note also that the k-th order regret $\hat{R}_k\left(\hat{X}^{n,\kappa_n}_{\text{TU-DUDE}}\right)$, which, based on the theorem, is upper bounded by the sum of the k-th order regret of the DUDE with parameter k and the stated twice–universality penalty, vanishes for k up to almost κ_n, while for $k \geq \kappa_n$ this may no longer be the case. In this sense, κ_n constitutes a "most ambitious" sequence for TU-DUDE, as choosing $k_n > \kappa_n$ does not increase the range of k over which the k-th order regret of TU-DUDE provably vanishes.

It is not hard to see that for this natural sequence κ_n, the penalty term $\epsilon_{\mathrm{TU},\kappa_n}(k,n)$ of the TU-DUDE $\hat{X}^{n,\kappa_n}_{\text{TU-DUDE}}$ is negligible relative to the k-th order DUDE regret bound (4) for k roughly in the range $\kappa_n/2 < k \leq \kappa_n$. No other previously proposed denoising approach is twice–universal with a smaller penalty (or even a vanishing one) for this range of k. In the case of the DUDE, for

example, not only is $\epsilon_{\mathrm{D},\kappa_n}(n,k)$ not negligible relative to the k-th order DUDE regret bound, but, more severely, it does not even vanish for most of the range $0 \le k \le \kappa_n$ (the exception being at k extremely close to κ_n). It follows that the k-th order regret of $\hat{X}^{n,\kappa_n}_{\mathrm{DUDE}}$ is bounded away from that of $\hat{X}^{n,k}_{\mathrm{DUDE}}$, the DUDE with parameter k, and, as a result, its k-th order regret $\hat{R}_k\Big(\hat{X}^{n,\kappa_n}_{\mathrm{DUDE}}\Big)$ may, in turn, not vanish for the same range of k.

The following new *universality* result also follows as a corollary of Theorem 1. Let

$$\mathcal{K}_n \stackrel{\mathrm{def}}{=} \left\{\{k_n\} : \sqrt{\frac{M^{2k_n}(k_n+1)}{n}} \to 0\right\} \tag{8}$$

denote the set of sequences $\{k_n\}$ for which the right-hand side of (6) vanishes.

Corollary 2. *Under the assumptions of Theorem 1, the TU-DUDE with parameter κ_n is universal with respect to all sliding window denoisers of order k_n for all sequences $\{k_n\} \in \mathcal{K}_n$, in the sense that $\lim_{n\to\infty} \hat{R}_{k_n}\Big(\hat{X}^{n,\kappa_n}_{TU}\Big) = 0$ for all $\{k_n\} \in \mathcal{K}_n$.*

Note that, on the other hand, no sequence of parameters $\{k'_n\}$ is known for which the universality of DUDE, with respect to the class of sliding window denoisers in the corollary, can be shown to follow from known results. In particular, if the sequence $\{k'_n\} \notin \mathcal{K}_n$ then the resulting DUDE would not be universal with respect to even 0-th order sliding windows (corresponding to the trivial sequence $k_n = 0$ which clearly belongs to $\mathcal{K}_n$), since the best we can say about the regret in this case, namely $\hat{R}_0\Big(\hat{X}^{n,k'_n}_{\mathrm{D}}\Big)$, is that it is bounded above by $C\sqrt{\frac{M^{2k'_n}(k'_n+1)}{n}}$ which does not tend to zero by virtue of $\{k'_n\} \notin \mathcal{K}_n$. Suppose, on the other hand, that the sequence $\{k'_n\} \in \mathcal{K}_n$. It is not hard to see that in this case, for any such $\{k'_n\}$, one can find another sequence $\{k_n\} \in \mathcal{K}_n$ which grows faster than $\{k'_n\}$. The DUDE with parameters k'_n, however, may not be effective against sliding window denoisers with the faster growing window sizes k_n, and indeed the best we can say about the regret via known bounds is that it is $||\Lambda||_\infty$.

Proof of Corollary 2. For every sequence $\{k_n\} \in \mathcal{K}_n$, the overall k_n-th order regret bound of the TU-DUDE with parameter κ_n, comprised of the DUDE regret bound and the twice–universality penalty, satisfies

$$\lim_{n\to\infty} \hat{R}_{k_n}\Big(\hat{X}^{n,\kappa_n}_{\mathrm{TU}}\Big) \le \lim_{n\to\infty} \left[C\sqrt{\frac{M^{2k_n}(k_n+1)}{n}} + \tilde{C}\left(\frac{(\kappa_n+1)^{5/4}}{n^{1/4}}\right)\right] \tag{9}$$

$$= 0, \tag{10}$$

where (9) follows from Theorem 1 and the fact that for all $\{k_n\} \in \mathcal{K}_n$, $k_n \le \kappa_n$ for n sufficiently large, and (10) follows from the definition of $\mathcal{K}_n$, and the fact that $\kappa_n = O(\log n)$. □

4 Loss estimation and twice–universal denoising

The TU-DUDE denoises a sequence by estimating the loss that would be incurred by candidate DUDE-like denoisers and then selecting the one with the lowest estimated loss. It is described in

detail in Section 6. In this section, we derive a more general result that will also be used in the analysis of TU-DUDE. We show that if there exists a "good" estimator for the loss incurred by $\hat{X}^{n,k}_{\mathrm{DUDE}}$, or an alternative denoiser with similar performance guarantees, for all $k < k_n$, then one can construct a "good" twice–universal denoiser, namely one that is twice–universal with a small penalty.

A loss estimator for a denoiser $\hat{X}^n$ is a mapping

$$\hat{L}_{\hat{X}^n} : \mathcal{A}^n \to \mathbb{R}$$

that, given a noisy sequence z^n, estimates the loss $L_{\hat{X}^n}(x^n, z^n)$ incurred by $\hat{X}^n$ to be $\hat{L}_{\hat{X}^n}(z^n)$. Consider a sequence $\{k_n\}$ of integers. For each n, let $\{\hat{X}^{n,k}\}$ denote a set of denoisers indexed by k in the range $0 \le k \le k_n$, whose regret satisfies the upper bound in (6), *i.e.*,

$$\hat{R}_k\Big(\hat{X}^{n,k}\Big) \le C\sqrt{\frac{M^{2k}(k+1)}{n}},\ k \le k_n \tag{11}$$

for all n sufficiently large. Let $\{\hat{L}_{\hat{X}^{n,k}}\}$, $k \le k_n$, be loss estimators for the denoisers $\{\hat{X}^{n.k}\}$. Given n and k_n, the denoiser $\hat{X}^n_U$, in turn, evaluates the estimated loss of each of the denoisers $\{\hat{X}^{n,k}\}$ for $k \le k_n$ using the loss estimators $\{\hat{L}_{\hat{X}^{n,k}}\}$ and denoises using the denoiser with the minimum estimated loss. Formally, $\hat{X}^n_U$ is given by

$$\hat{X}^n_U(z^n)[i] = \hat{X}^{n,\hat{k}^*_n}(z^n)[i] \tag{12}$$

where

$$\hat{k}^*_n = \arg\min_{k \le k_n} \hat{L}_{\hat{X}^{n,k}}(z^n). \tag{13}$$

Lemma 3. *Let $\{k_n\}$ be a sequence of integers, and for all n, let $\{\hat{X}^{n,k}\}$ denote a set of denoisers indexed by $k \le k_n$ satisfying* (11) *for n sufficiently large. If for all $k \le k_n$ and all x^n*

$$E\Big[\Big|L_{\hat{X}^{n,k}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}}(Z^n)\Big|\Big] \le \alpha(k_n, n) \tag{14}$$

for all n, then $\hat{X}^n_U$ is twice–universal with penalty $2(k_n+1)\alpha(k_n, n)$.

Proof. Observe that for all k, all n sufficiently large, and all x^n

$$\begin{aligned} E(L_{\hat{X}^n_U}(x^n, Z^n) - D_k(x^n, Z^n)) &= E(L_{\hat{X}^{n,k}}(x^n, Z^n) - D_k(x^n, Z^n)) \\ &\quad + E(L_{\hat{X}^n_U}(x^n, Z^n) - L_{\hat{X}^{n,k}}(x^n, Z^n)) \\ &\le C\sqrt{\frac{M^{2k}(k+1)}{n}} + E\Big(L_{\hat{X}^n_U}(x^n, Z^n) - L_{\hat{X}^{n,k}}(x^n, Z^n)\Big) \end{aligned} \tag{15}$$

where (15) follows from (11) and the definition of regret. In analogy to $\hat{k}^*_n$ defined in (13), let

$$k^*_n = \arg\min_{k \le k_n} L_{\hat{X}^{n,k}}(x^n, z^n), \tag{16}$$

be the actual loss-minimizing parameter for $\{\hat{X}^{n,k}\}$. This parameter is a function of both the clean and noisy sequences, unlike its counterpart $\hat{k}_n^*$, which is a function only of the noisy sequence. We then have

$$\begin{aligned}
E\Big(L_{\hat{X}_U^n}(x^n, Z^n) - L_{\hat{X}^{n,k}}(x^n, Z^n)\Big) &= E\big(L_{\hat{X}^{n,\hat{k}_n^*}}(x^n, Z^n) - L_{\hat{X}^{n,k_n^*}}(x^n, Z^n)\big) \\
&\quad + E\big(L_{\hat{X}^{n,k_n^*}}(x^n, Z^n) - L_{\hat{X}^{n,k}}(x^n, Z^n)\big) && (17) \\
&\le E\big(L_{\hat{X}^{n,\hat{k}_n^*}}(x^n, Z^n) - L_{\hat{X}^{n,k_n^*}}(x^n, Z^n)\big) && (18) \\
&= E\Big(L_{\hat{X}^{n,\hat{k}_n^*}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,\hat{k}_n^*}}(Z^n)\Big) \\
&\quad + E(\hat{L}_{\hat{X}^{n,\hat{k}_n^*}}(Z^n) - L_{\hat{X}^{n,k_n^*}}(x^n, Z^n)) \\
&\le E\Big(L_{\hat{X}^{n,\hat{k}_n^*}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,\hat{k}_n^*}}(Z^n)\Big) \\
&\quad + E\Big(\hat{L}_{\hat{X}^{n,k_n^*}}(Z^n) - L_{\hat{X}^{n,k_n^*}}(x^n, Z^n)\Big) && (19) \\
&\le 2\sum_{k\le k_n} E\Big(\Big|L_{\hat{X}^{n,k}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}}(Z^n)\Big|\Big) && (20) \\
&\le 2(k_n+1)\alpha(k_n, n) && (21)
\end{aligned}$$

where (17) follows from (12), (18) follows from (16), (19) follows from (13), (20) follows by taking absolute values and summing over all k in each expectation, not just the two in the previous step, and (21) follows from (14). Substituting (21) in (15), we obtain the lemma. □

5 A loss estimator and its properties

The previous section suggests that one way to obtain a denoiser that is twice–universal with a small penalty is through a good estimator for the losses of a collection of constituent denoisers. In this section, we study the properties of one such estimator, first proposed in [7]. The estimate of the loss incurred by *any* denoiser $\hat{X}^n$ proposed in [7] is given by

$$\hat{L}_{\hat{X}^n}(z^n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathcal{A}} \mathbf{\Pi}^{-T}(x, z_i)\sum_{z\in\mathcal{A}}\Lambda(x, \hat{x}_i(z))\mathbf{\Pi}(x, z) \tag{22}$$

where we use $\hat{x}_i(z)$ to abbreviate $\hat{X}^n(z_1^{i-1}\cdot z\cdot z_{i+1}^n)[i]$. A derivation of this estimator is provided in [8]. One way to view this expression is to observe that $\sum_{z\in\mathcal{A}}\Lambda(x, \hat{x}_i(z))\mathbf{\Pi}(x, z)$ is the expected loss of denoising the i-th symbol when the clean symbol is x, while $\mathbf{\Pi}^{-T}(x, z_i)$ is a weighting on the different values of x that results in the overall estimate being unbiased. Indeed, the expected value of $\mathbf{\Pi}^{-T}(x, Z)$ is 1 precisely for $x = x_i$ and 0 otherwise, and thus $\mathbf{\Pi}^{-T}(x, z_i)$ can be loosely interpreted as an instantaneous (for each index i) estimate of this indicator function/vector.

5.1 Unbiasedness

It was stated in [7] that the estimate (22) is unbiased for all denoisers and clean sequences. An outline of the proof was also provided there. We present a formal proof of this fact for completeness

here. The proof argument is similar to the one used for filtering, a causal version of the denoising problem, in [9]. A similar result for denosing was proved in [8]. In fact, we will require (and prove) the stronger property that the estimate of the loss incurred by a denoiser on the i-th symbol is conditionally unbiased given the other noisy symbols. Let

$$\tilde{\Lambda}_{i,\hat{X}^n}(z^n) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{A}} \mathbf{\Pi}^{-T}(x, z_i) \sum_{z\in\mathcal{A}} \Lambda(x, \hat{x}_i(z))\mathbf{\Pi}(x, z)$$

denote the estimate of the loss incurred on the i-th symbol. Then

$$\hat{L}_{\hat{X}^n}(z^n) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\Lambda}_{i,\hat{X}^n}(z^n).$$

Lemma 4. *[7] For all x^n, all denoisers $\hat{X}^n$, and all i, $1 \le i \le n$, z_1^{i-1}, z_{i+1}^n*

$$E\Big[\tilde{\Lambda}_{i,\hat{X}^n}(Z^n)\,\Big|Z_1^{i-1} = z_1^{i-1}, Z_{i+1}^n = z_{i+1}^n\Big] = E\Big[\Lambda\Big(x_i, \hat{X}^n(Z^n)[i]\Big)\,\Big|Z_1^{i-1} = z_1^{i-1}, Z_{i+1}^n = z_{i+1}^n\Big] \quad (23)$$

and therefore

$$E\Big[\hat{L}_{\hat{X}^n}(z^n)\Big] = E\big[L_{\hat{X}^n}(x^n, Z^n)\big]. \quad (24)$$

Proof. Observe that

$$\begin{aligned}
&E\Big[\tilde{\Lambda}_{i,\hat{X}^n}(Z^n)\,\Big|Z_1^{i-1} = z_1^{i-1}, Z_{i+1}^n = z_{i+1}^n\Big] \\
&= \sum_{x\in\mathcal{A}} E\big[\mathbf{\Pi}^{-T}(x, Z_i)\,\big|Z_1^{i-1} = z_1^{i-1}, Z_{i+1}^n = z_{i+1}^n\big]\left(\sum_{z\in\mathcal{A}} \Lambda\Big(x, \hat{X}^n(z_1^{i-1}\cdot z\cdot z_{i+1}^n)[i]\Big)\mathbf{\Pi}(x, z)\right) \\
&= \sum_{x\in\mathcal{A}} E\big[\mathbf{\Pi}^{-T}(x, Z_i)\big]\left(\sum_{z\in\mathcal{A}} \Lambda\Big(x, \hat{X}^n(z_1^{i-1}\cdot z\cdot z_{i+1}^n)[i]\Big)\mathbf{\Pi}(x, z)\right) \quad (25) \\
&= \sum_{x\in\mathcal{A}}\sum_{z_i\in\mathcal{A}} \mathbf{\Pi}(x_i, z_i)\mathbf{\Pi}^{-T}(x, z_i)\left(\sum_{z\in\mathcal{A}} \Lambda\Big(x, \hat{X}^n(z_1^{i-1}\cdot z\cdot z_{i+1}^n)[i]\Big)\mathbf{\Pi}(x, z)\right) \\
&= \sum_{x\in\mathcal{A}} 1(x = x_i)\left(\sum_{z\in\mathcal{A}} \Lambda\Big(x, \hat{X}^n(z_1^{i-1}\cdot z\cdot z_{i+1}^n)[i]\Big)\mathbf{\Pi}(x, z)\right) \quad (26) \\
&= \sum_{z\in\mathcal{A}} \Lambda\Big(x_i, \hat{X}^n(z_1^{i-1}\cdot z\cdot z_{i+1}^n)[i]\Big)\mathbf{\Pi}(x_i, z) \\
&= E\Big[\Lambda\Big(x_i, \hat{X}^n(Z^n)[i]\Big)\,\Big|Z_1^{i-1} = z_1^{i-1}, Z_{i+1}^n = z_{i+1}^n\Big]
\end{aligned}$$

where $1(\cdot)$ is the indicator function, (25) holds as Z_i is independent of Z_1^{i-1} and Z_{i+1}^n, and (26) is true since

$$\sum_{z_i\in\mathcal{A}} \mathbf{\Pi}(x_i, z_i)\mathbf{\Pi}^{-T}(x, z_i) = (\mathbf{\Pi}\mathbf{\Pi}^{-1})(x_i, x),$$

the (x_i, x)-th entry of the matrix $\mathbf{\Pi}\mathbf{\Pi}^{-1}$, which is one when $x_i = x$, and 0 otherwise. Summing over i and taking expectation on both sides of (23) over all values of Z_1^{i-1} and Z_{i+1}^n gives (24). □

5.2 Concentration

Following Lemma 3, we seek to bound the expected absolute error $E\Big[\Big|L_{\hat{X}^{n,k}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}}(Z^n)\Big|\Big]$ in the loss estimate. Observe that for any bounded random variable Z with $|Z| \leq M$, and any $c > 0$

$$cP(|Z| \ \geq c) \leq E[|Z|] \leq MP(|Z| \geq c) + c.$$

Therefore, one way to establish if an estimator is satisfactory or otherwise for our purposes is to derive concentration bounds (upper or lower) on the expected absolute loss estimation error, *i.e.,* the probability that it deviates significantly from 0. Such concentration bounds have been derived for the class of sliding window denoisers in [8]. It is shown in [8] that for all $\hat{X}^n \in \mathcal{S}_k$, $\mathbf{\Pi}$, $\mathbf{\Lambda}$, x^n, and $\tau > 0$

$$P\Big(\Big|L_{\hat{X}^n}(x^n, Z^n) - \hat{L}_{\hat{X}^n}(Z^n)\Big| \geq \tau\Big) \leq (k+1)e^{\frac{-2(n-2k)\tau^2}{(k+1)||\mathbf{\Lambda}||_\infty^2(1+M||\mathbf{\Pi}^{-1}||_\infty)^2}}. \tag{27}$$

This bound implies that as long as $k = o(n/\log n)$, the estimated loss of a given k-th order sliding window denoiser concentrates around the true loss. While such strong concentration results, where the probability that the estimate deviates significantly from the true loss decreases exponentially in n for fixed k, are possible for sliding window denoisers, we show in Section 5.3 that similar strong results are not possible for the DUDE. It turns out, however, that the method for showing that exponential concentration is not possible provides us the insight that leads to weaker concentration bounds for 0-th order DUDE. We then build on the concentration bounds for 0-th order DUDE to derive bounds on the expected absolute loss estimation error for a variant of $\hat{X}^{n,k}_{\text{DUDE}}$, for $k > 0$, that will then be suitable for the twice–universal denoising paradigm of Section 4.

5.3 Non-exponential concentration for the DUDE of 0-th order

To show that exponential concentration is not possible in general, we consider the special case of the binary symmetric channel (BSC) and Hamming loss. Let $\mathcal{A} = \{0, 1\}$. Let $\mathbf{\Pi}$ correspond to the BSC with crossover probability $\delta < \frac{1}{2}$, and $\mathbf{\Lambda}$ correspond to the Hamming loss. Note that $\hat{X}^{n,0}_{\text{DUDE}}$ denotes DUDE of 0-th order. To prove our result, we will require the DeMoivre-Laplace theorem, which approximates the binomial distribution close to the mean using the normal distribution. We state it below as a lemma. Let

$$b_{n,k}(p) \stackrel{\text{def}}{=} \binom{n}{\lfloor np \rfloor + k} p^{\lfloor np \rfloor + k}(1-p)^{n - \lfloor np \rfloor - k} \tag{28}$$

denote the probability that a binomial random variable with parameters n and p takes the value $\lfloor np \rfloor + k$. Let

$$f(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{29}$$

denote the probability density function (pdf) of the standard Normal distribution.

Lemma 5. *([10], VII.3, Theorem 1) Let $\{K_n\}$ be a sequence such that*

$$K_n = o(n^{2/3}).$$

Then, for all p, $0 < p < 1$, all $\epsilon > 0$, and all n sufficiently large,

$$(1-\epsilon) \leq \frac{\sqrt{np(1-p)}b_{n,k}(p)}{f\left(\frac{k}{\sqrt{np(1-p)}}\right)} \leq (1+\epsilon)$$

for all k, $|k| < K_n$.

Theorem 6. *There exists a clean sequence x^n and constants K and τ_0 such that*

$$P\Big(\Big|L_{\hat{X}^{n,0}_{DUDE}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,0}_{DUDE}}(Z^n)\Big| \geq \tau_0\Big) \geq \frac{K}{\sqrt{n}}.$$

Proof. For any $z^n \in \{0,1\}^n$, let $T(z^n)$ denote the *type* of z^n, *i.e.,* the number of 1s in z^n. Then, the DUDE of zero order is given by [1]

$$\hat{X}^{n,0}_{\rm DUDE}(z^n)[i] = \begin{cases} 0, & T(z^n) \leq n2\delta(1-\delta) \\ z_i, & n2\delta(1-\delta) < T(z^n) \leq n(1-2\delta(1-\delta)) \\ 1, & T(z^n) > n(1-2\delta(1-\delta)). \end{cases} \tag{30}$$

Consider z^n such that $T(z^n) = \lfloor n2\delta(1-\delta) \rfloor$. Then, by (30),

$$L_{\hat{X}^{n,0}_{\rm DUDE}}(x^n, z^n) = \frac{1}{n}\sum_{i=1}^{n} \Lambda(x_i, 0) = \frac{T(x^n)}{n}. \tag{31}$$

Observe that for the same z^n, if $z_i = 0$, then the second case in (30) holds for $z_1^{i-1} \cdot 1 \cdot z_{i+1}^n$, and therefore $\hat{X}^{n,0}_{\rm DUDE}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i] = z, \quad z = 0, 1$. Consequently, for all $x \in \{0,1\}$,

$$\sum_{z \in \{0,1\}} \Lambda(x, \hat{X}^{n,0}_{\rm DUDE}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i])\mathbf{\Pi}(x, z) = \delta.$$

Hence,

$$\sum_{x \in \{0,1\}} \mathbf{\Pi}^{-T}(x, z_i) \sum_{z \in \{0,1\}} \Lambda(x, \hat{X}^{n,0}_{\rm DUDE}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i])\mathbf{\Pi}(x, z) = \delta \sum_{x \in \{0,1\}} \mathbf{\Pi}^{-T}(x, 0) = \delta. \tag{32}$$

If $z_i = 1$ instead, then, with a similar argument, $\hat{X}^{n,0}_{\rm DUDE}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i] = 0, \quad z = 0, 1$. Therefore, for all $x \in \{0,1\}$,

$$\sum_{z \in \{0,1\}} \Lambda(x, \hat{X}^{n,0}_{\rm DUDE}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i])\mathbf{\Pi}(x, z) = \Lambda(x, 0).$$

Hence,

$$\begin{aligned} \sum_{x \in \{0,1\}} \mathbf{\Pi}^{-T}(x, z_i) \sum_{z \in \{0,1\}} \Lambda(x, \hat{X}^{n,0}_{\rm DUDE}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i]))\mathbf{\Pi}(x, z) &= \sum_{x \in \{0,1\}} \mathbf{\Pi}^{-T}(x, 1)\Lambda(x, 0) \\ &= \mathbf{\Pi}^{-T}(1, 1) = \frac{1-\delta}{1-2\delta}. \end{aligned} \tag{33}$$

Combining (22), (32), and (33), we obtain that if $T(z^n) = \lfloor n2\delta(1-\delta) \rfloor$, and $\delta < \frac{1}{2}$, then

$$\begin{aligned}
\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(z^n) &= \left(1 - \frac{T(z^n)}{n}\right)\delta + \frac{T(z^n)}{n}\frac{1-\delta}{1-2\delta} \\
&\geq (1 - 2\delta(1-\delta))\delta + \frac{2\delta(1-\delta)^2}{1-2\delta} - \frac{1-\delta}{n(1-2\delta)} \\
&= \delta + 2\delta(1-\delta)\left(\frac{1-\delta}{1-2\delta} - \delta\right) - \frac{1-\delta}{n(1-2\delta)} \\
&= \delta + 2\delta(1-\delta) + 2\delta(1-\delta)\left(\frac{2\delta^2}{1-2\delta}\right) - \frac{1-\delta}{n(1-2\delta)} \\
&> \delta + 2\delta(1-\delta)
\end{aligned}$$

for all δ, $1/2 > \delta > 0$ and all sufficiently large n. From (31), we obtain that if $T(z^n) = \lfloor n2\delta(1-\delta) \rfloor$

$$\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(z^n) - L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n, z^n) > \left(\delta - \frac{T(x^n)}{n}\right) + 2\delta(1-\delta).$$

On the other hand, if $T(x^n) = \lfloor n\delta \rfloor$, then

$$\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(z^n) - L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n, z^n) > 2\delta(1-\delta).$$

But when $T(x^n) = \lfloor n\delta \rfloor$, we will show that for some constant K and all sufficiently large n

$$P(T(Z^n) = \lfloor n2\delta(1-\delta) \rfloor) \geq \frac{K}{\sqrt{n}}$$

which will prove the theorem. Observe that $T(Z^n) = \lfloor n2\delta(1-\delta) \rfloor$ when i, $0 \leq i \leq \lfloor n\delta \rfloor$, of the bits in x^n that are 1 remain unflipped by the channel and $\lfloor n2\delta(1-\delta) \rfloor - i$ of the bits that are 0 in x^n are flipped. Therefore, using the notation in (28)

$$\begin{aligned}
&P(T(Z^n) = \lfloor n2\delta(1-\delta) \rfloor) \\
&= \sum_{i=0}^{\lfloor n\delta \rfloor} \binom{\lfloor n\delta \rfloor}{i} \delta^{\lfloor n\delta \rfloor - i}(1-\delta)^i \binom{n - \lfloor n\delta \rfloor}{\lfloor n2\delta(1-\delta) \rfloor - i} \delta^{\lfloor n2\delta(1-\delta) \rfloor - i}(1-\delta)^{n - \lfloor n\delta \rfloor - \lfloor n2\delta(1-\delta) \rfloor + i} \\
&= \sum_{i=-\lfloor \lfloor n\delta \rfloor (1-\delta) \rfloor}^{\lfloor n\delta \rfloor - \lfloor \lfloor n\delta \rfloor (1-\delta) \rfloor} b_{\lfloor n\delta \rfloor, i}(1-\delta)\, b_{n - \lfloor n\delta \rfloor, \Delta - i}(\delta) \\
&\geq \sum_{i=-\lfloor \sqrt{n} \rfloor}^{\lfloor \sqrt{n} \rfloor} b_{\lfloor n\delta \rfloor, i}(1-\delta) b_{n - \lfloor n\delta \rfloor, \Delta - i}(\delta) \qquad (34)
\end{aligned}$$

where

$$\Delta = \lfloor n2\delta(1-\delta) \rfloor - \lfloor n\delta - \lfloor n\delta \rfloor \delta \rfloor - \lfloor \lfloor n\delta \rfloor (1-\delta) \rfloor$$

and the inequality holds for sufficiently large n. Observe that $|\Delta| \leq 3$. Therefore, for $-\lfloor \sqrt{n} \rfloor \leq i \leq \lfloor \sqrt{n} \rfloor$,

$$\frac{\Delta - i}{(n - \lfloor n\delta \rfloor)^{\frac{2}{3}}} \to 0 \text{ and also } \frac{i}{n\delta^{\frac{2}{3}}} \to 0.$$

Thus, applying Lemma 5 to each of the terms in the summation in (34), we obtain that for any $\epsilon > 0$ and all sufficiently large n

$$P(T(Z^n) = \lfloor n2\delta(1-\delta) \rfloor) \geq \sum_{i=-\lfloor\sqrt{n}\rfloor}^{\lfloor\sqrt{n}\rfloor} (1-\epsilon)^2 \frac{f\left(\frac{i}{\sqrt{\lfloor n\delta \rfloor (1-\delta)\delta}}\right)}{\sqrt{\lfloor n\delta \rfloor (1-\delta)\delta}} \frac{f\left(\frac{\Delta - i}{\sqrt{(n-\lfloor n\delta \rfloor)\delta(1-\delta)}}\right)}{\sqrt{(n-\lfloor n\delta \rfloor)\delta(1-\delta)}} \tag{35}$$

$$\geq \sum_{i=-\lfloor\sqrt{n}\rfloor}^{\lfloor\sqrt{n}\rfloor} (1-\epsilon)^2 \frac{C}{\sqrt{n}} \cdot \frac{C}{\sqrt{n}} \tag{36}$$

$$\geq \frac{2(1-\epsilon)^2 C^2}{\sqrt{n}}$$

for some constant C, where (36) follows from (29) and the observation that for $-\lfloor\sqrt{n}\rfloor \leq i \leq \lfloor\sqrt{n}\rfloor$, the absolute values of both arguments of the function f in (35) are bounded from above by a constant. □

While Theorem 6 shows that an exponential concentration bound as in (27) is not possible for the DUDE with the estimator $\hat{L}$, the proof suggests a way to bound the error in the loss estimate to obtain suitable non-exponential concentration bounds. We first provide a brief outline of the argument by continuing to analyze the case of a BSC and Hamming loss. Notice that when $T(z^n) < \lfloor n2\delta(1-\delta) \rfloor$, then $\hat{X}^{n,0}_{\mathrm{DUDE}}$ always returns 0. Therefore, the true loss is

$$L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n, z^n) = \frac{T(x^n)}{n}.$$

Also, for all $x \in \{0,1\}$,

$$\sum_{z \in \{0,1\}} \Lambda(x, \hat{X}^{n,0}_{\mathrm{DUDE}}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i]) \mathbf{\Pi}(x,z) = \Lambda(x,0)$$

and therefore

$$\sum_{x \in \{0,1\}} \mathbf{\Pi}^{-T}(x, z_i) \sum_{z \in \{0,1\}} \Lambda(x, \hat{X}^{n,0}_{\mathrm{DUDE}}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i])) \mathbf{\Pi}(x,z) = \sum_{x \in \{0,1\}} \mathbf{\Pi}^{-T}(x, z_i) \Lambda(x,0) = \mathbf{\Pi}^{-T}(1, z_i).$$

Hence, the loss estimate is

$$\begin{aligned}
\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(z^n) &= \frac{T(z^n)}{n} \mathbf{\Pi}^{-T}(1,1) + \left(1 - \frac{T(z^n)}{n}\right) \mathbf{\Pi}^{-T}(1,0) \\
&= \frac{T(z^n)}{n} \frac{1-\delta}{1-2\delta} - \left(1 - \frac{T(z^n)}{n}\right) \frac{\delta}{1-2\delta} \\
&= \frac{T(z^n)}{n(1-2\delta)} - \frac{\delta}{1-2\delta}.
\end{aligned}$$

Since

$$E(T(Z^n)) = T(x^n)(1-\delta) + (n - T(x^n))\delta = n\delta + T(x^n)(1-2\delta),$$

it follows from standard concentration results that $\frac{T(Z^n)}{n(1-2\delta)} - \frac{\delta}{1-2\delta}$ concentrates around the true loss $T(x^n)/n$. Thus, when $T(z^n) < \lfloor n2\delta(1-\delta) \rfloor$ the loss estimate is likely to be close to the true

loss. More generally, we observe that the set of noisy sequences z^n can be partitioned into subsets based on the type of z^n. If the type does not equal $\lfloor n2\delta(1-\delta)\rfloor$ or $\lfloor n(1-2\delta(1-\delta))\rfloor$, the decision boundaries where $\hat{X}^{n,0}_{\mathrm{DUDE}}$ changes its denoising rule, then the loss estimate concentrates around the true loss. This observation suggests the following approach to bounding the loss-estimation error. The cases when the type of z^n corresponds to a decision boundary, and when it does not, can be separated. The loss-estimation error in the former case can be bounded from above with the probability of observing such a z^n, whereas the error in the latter case is small for the reasons described above. This approach can be generalized to derive an upper bound on the expected error in the loss estimate for arbitrary $\mathbf{\Pi}$ and $\mathbf{\Lambda}$. We first derive this upper bound for $k=0$ and then use that result to obtain a more general result for $k>0$. To derive the bound for $k=0$, we first identify the z^ns that result in poor estimates and bound the probability of their occurrence.

When $k=0$, the notation $\mathbf{m}(\cdot)$ defined in Section 2 can be simplified as follows. Let $\mathbf{m}(z^n)$ be the vector whose c_0-th component, $c_0 \in \mathcal{A}$, is

$$\mathbf{m}(z^n)[c_0] = |\{i : 1 \le i \le n, z_i = c_0\}|$$

the number of occurrences of c_0 in z^n. Define $g : \mathcal{A}^n \times \mathcal{A} \to \mathcal{A}$ to be

$$g(z^n, z') \stackrel{\mathrm{def}}{=} \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \left((\mathbf{\Pi}^{-T}\mathbf{m}(z^n)) \odot \pi_{z'} \right) \tag{37}$$

where ties are broken based on some fixed ordering of the elements of $\mathcal{A}$, so that

$$\hat{X}^{n,0}_{\mathrm{DUDE}}(z^n)[i] = g(z^n, z_i).$$

A sequence z^n is said to be $\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}$-*continuous* if $\forall i, z'$

$$g\big(z_1^{i-1} \cdot z' \cdot z_{i+1}^n, z'\big) = g\big(z_1^{i-1} \cdot z_i \cdot z_{i+1}^n, z'\big). \tag{38}$$

It is said to be $\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}$-*discontinuous* otherwise. We will now bound the probability of observing a $\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}$-*discontinuous* Z^n. To do so we require the following version of the Berry-Esseen theorem (see, *e.g.,* [11]), stated as a lemma.

Lemma 7. *([11], XVI.5, Theorem 1) Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of independent real-valued random variables such that for all i, $1 \le i \le n$,*

$$E[Y_i] = 0,\ E\big[Y_i^2\big] = \sigma_i^2,\ E\big[|Y_i|^3\big] = \rho_i.$$

Let

$$s_n^2 = \sum_{i=1}^n \sigma_i^2,\ r_n = \sum_{i=1}^n \rho_i$$

and let F_n be the cumulative distribution function (cdf) of the normalized sum $(Y_1+Y_2+\ldots+Y_n)/s_n$. Then, for all x and n,

$$|F_n(x) - \Phi(x)| \le 6\frac{r_n}{s_n^3}$$

where $\Phi(x)$ is the standard normal cdf, namely,

$$\Phi(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du.$$

Lemma 8. *For all channels such that the transition matrix* $\mathbf{\Pi}$ *has only non-zero entries, and all* x^n*,*

$$P\Big(Z^n \text{ is } \hat{L}_{\hat{X}^{n,0}_{DUDE}}\text{-discontinuous}\Big) \le \frac{C}{\sqrt{n}}$$

where the constant C *depends on* $\mathbf{\Pi}, \Lambda$ *and* M*.*

Proof. Let z^n be $\hat{L}_{\hat{X}^{n,0}_{\text{DUDE}}}$*-discontinuous.* Then there exist i, z' such that

$$\hat{x}_1 \stackrel{\text{def}}{=} g\big(z_1^{i-1} \cdot z' \cdot z_{i+1}^n, z'\big) \neq g\big(z_1^{i-1} \cdot z_i \cdot z_{i+1}^n, z'\big) \stackrel{\text{def}}{=} \hat{x}_2. \tag{39}$$

Noting that $\mathbf{a}^T((A^T\mathbf{b}) \odot \mathbf{c}) = (\mathbf{a} \odot \mathbf{c})^T(A^T\mathbf{b}) = (A(\mathbf{a} \odot \mathbf{c}))^T\mathbf{b}$ for all vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and matrix A, it follows from (37) and (39) that

$$\big(\mathbf{\Pi}^{-1}(\lambda_{\hat{x}_1} \odot \pi_{z'})\big)^T \mathbf{m}\big(z_1^{i-1} \cdot z' \cdot z_{i+1}^n\big) \le \big(\mathbf{\Pi}^{-1}(\lambda_{\hat{x}_2} \odot \pi_{z'})\big)^T \mathbf{m}\big(z_1^{i-1} \cdot z' \cdot z_{i+1}^n\big) \tag{40}$$

and

$$\big(\mathbf{\Pi}^{-1}(\lambda_{\hat{x}_1} \odot \pi_{z'})\big)^T \mathbf{m}\big(z_1^{i-1} \cdot z_i \cdot z_{i+1}^n\big) \ge \big(\mathbf{\Pi}^{-1}(\lambda_{\hat{x}_2} \odot \pi_{z'})\big)^T \mathbf{m}\big(z_1^{i-1} \cdot z_i \cdot z_{i+1}^n\big) \tag{41}$$

with the equality holding in at most one of (40) and (41) (since equality in both would imply that ties were broken differently for different values of $\mathbf{m}$ in (37)). For $z \in \mathcal{A}$, let $\mathbf{e}_z$ denote the M dimensional unit column vector, *i.e.,* $\mathbf{e}_z[z] = 1$ and $\mathbf{e}_z[a] = 0$ for all $a \neq z$. Then

$$\mathbf{m}\big(z_1^{i-1} \cdot z' \cdot z_{i+1}^n\big) = \mathbf{m}\big(z_1^{i-1} \cdot z_i \cdot z_{i+1}^n\big) + \mathbf{e}_{z'} - \mathbf{e}_{z_i}.$$

Substituting in (40), and using (41), we obtain that if z^n is $\hat{L}_{\hat{X}^{n,0}_{\text{DUDE}}}$*-discontinuous* then, for some i, z',

$$0 \le \big(\mathbf{\Pi}^{-1}((\lambda_{\hat{x}_1} - \lambda_{\hat{x}_2}) \odot \pi_{z'})\big)^T \mathbf{m}(z^n) \le \big(\mathbf{\Pi}^{-1}((\lambda_{\hat{x}_1} - \lambda_{\hat{x}_2}) \odot \pi_{z'})\big)^T (\mathbf{e}_{z_i} - \mathbf{e}_{z'})$$

with one of the inequalities being necessarily strict. Hence, if z^n is $\hat{L}_{\hat{X}^{n,0}_{\text{DUDE}}}$*-discontinuous* then, for some a, b, z',

$$0 \le \big(\mathbf{\Pi}^{-1}((\lambda_a - \lambda_b) \odot \pi_{z'})\big)^T \mathbf{m}(z^n) \le \max_{z'' \in \mathcal{A}} \big(\mathbf{\Pi}^{-1}((\lambda_a - \lambda_b) \odot \pi_{z'})\big)^T (\mathbf{e}_{z''} - \mathbf{e}_{z'}), \tag{42}$$

with at least one of the inequalities being strict, so that

$$0 < \max_{z'' \in \mathcal{A}} \big(\mathbf{\Pi}^{-1}((\lambda_a - \lambda_b) \odot \pi_{z'})\big)^T (\mathbf{e}_{z''} - \mathbf{e}_{z'}). \tag{43}$$

Let

$$\alpha(a, b, z') \stackrel{\text{def}}{=} \mathbf{\Pi}^{-1}((\lambda_a - \lambda_b) \odot \pi_{z'}),$$

and let

$$\Delta^*(a, b, z') \stackrel{\text{def}}{=} \max_{z'' \in \mathcal{A}} \alpha(a, b, z')^T (\mathbf{e}_{z''} - \mathbf{e}_{z'}) > 0.$$

For $i = 1, \ldots, n$, we define independent random variables Y_i to be

$$Y_i = \alpha(a, b, z')[Z_i]$$

so that for all $z \in \mathcal{A}$,

$$P(Y_i = \alpha(a, b, z')[z]) = \mathbf{\Pi}(x_i, z). \tag{44}$$

Then

$$\left(\mathbf{\Pi}^{-1}((\lambda_a - \lambda_b) \odot \pi_{z'})\right)^T \mathbf{m}(z^n) = \alpha(a, b, z')^T \mathbf{m}(z^n) = \sum_{i=1}^{n} Y_i,$$

and (42) takes the form

$$0 \leq \sum_{i=1}^{n} Y_i \leq \Delta^*(a, b, z'). \tag{45}$$

Let $\sigma_i^2 = E\big[(Y_i - E[Y_i])^2\big]$, $s_n^2 = \sum_{i=1}^n \sigma_i^2$, $\rho_i = E\big[|Y_i - E[Y_i]|^3\big]$, $r_n = \sum_{i=1}^n \rho_i$, and let F_n denote the cdf of $s_n^{-1} \sum_{i=1}^n (Y_i - E[Y_i])$. Observe that (43) implies that $\alpha(a, b, z')$ is not a constant vector and recall the assumption that all the entries of $\mathbf{\Pi}$ are non-zero. Thus, by (44), it follows that, for all $i = 1, \ldots, n$, Y_i takes at least two different values with non-zero probability and therefore

$$\sigma_i^2 > 0. \tag{46}$$

It follows from Lemma 7 that for all x and n

$$|F_n(x) - \Phi(x)| \leq 6 \frac{r_n}{s_n^3}$$

where $\Phi(x)$ is the normal cdf. Since $|Y_i|$ is bounded, there exists a constant c_1 such that

$$r_n = \sum_{i=1}^{n} E\big[|Y_i - E[Y_i]|^3\big] \leq c_1 n$$

and, by (46), there exists a constant c_2 such that

$$s_n^2 = \sum_{i=1}^{n} E\big[|Y_i - E[Y_i]|^2\big] \geq c_2 n. \tag{47}$$

Therefore, for all x and n, there exists a constant $c_3 > 0$ such that

$$|F_n(x) - \Phi(x)| \leq \frac{c_3}{\sqrt{n}}. \tag{48}$$

Note that for all $\delta > 0$ and x

$$\begin{aligned} F_n(x + \delta) - F_n(x) &\leq |F_n(x + \delta) - \Phi(x + \delta)| + \Phi(x + \delta) - \Phi(x) + |F_n(x) - \Phi(x)| \\ &\leq \frac{2c_3}{\sqrt{n}} + \sup_x \left(\Phi(x + \delta) - \Phi(x)\right) \\ &\leq \frac{2c_3}{\sqrt{n}} + \frac{\delta}{\sqrt{2\pi}} \end{aligned} \tag{49}$$

where the second inequality follows from (48) and the third from mean value theorem. From (45), it follows that for all x^n, and all $\epsilon > 0$

$$
\begin{aligned}
&P\Big(Z^n \text{ is } \hat{L}_{\hat{X}^{n,0}_{\text{DUDE}}}\text{-}\textit{discontinuous}\Big) \\
&\leq \sum_{a,b,z' \in \mathcal{A}:\Delta^*(a,b,z')>0} P\left(0 \leq \sum_{i=1}^{n} Y_i \leq \Delta^*(a,b,z')\right) && (50)\\
&\leq \sum_{a,b,z' \in \mathcal{A}:\Delta^*(a,b,z')>0} \left[F_n\left(s_n^{-1}\left(\Delta^*(a,b,z') - \sum_{i=1}^{n} E[Y_i]\right)\right) - F_n\left(-s_n^{-1}\left(\epsilon + \sum_{i=1}^{n} E[Y_i]\right)\right)\right] && (51)\\
&\leq \sum_{a,b,z' \in \mathcal{A}:\Delta^*(a,b,z')>0} \left(\frac{2c_3}{\sqrt{n}} + \frac{s_n^{-1}(\Delta^*(a,b,z') + \epsilon)}{\sqrt{2\pi}}\right) && (52)\\
&\leq M^3 \frac{c_4}{\sqrt{n}} && (53)
\end{aligned}
$$

for some constant $c_4 > 0$, where (50) follows from the union bound, (51) from the definition of F_n, (52) from the application of (49), and (53) from (47). □

Now that we have bounded the probability of observing a $\hat{L}_{\hat{X}^{n,0}_{\text{DUDE}}}$-*discontinuous* sequence, we use it to bound the error in the loss estimate when $k = 0$. We do so by showing that when the observed Z^n is $\hat{L}_{\hat{X}^{n,0}_{\text{DUDE}}}$-*continuous*, the error in the loss estimate can be bounded from above. We will use this result to prove a similar result for a variant of $\hat{X}^{n,k}_{\text{DUDE}}$ for $k > 0$.

Lemma 9. *For all channels such that the transition matrix* $\mathbf{\Pi}$ *has only non-zero entries, and all* x^n*,*

$$E\Big[(L_{\hat{X}^{n,0}_{DUDE}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,0}_{DUDE}}(Z^n))^2\Big] \leq \frac{C}{\sqrt{n}} \tag{54}$$

for a constant C *independent of* n*.*

Proof. For all x^n, z^n,

$$\Big|L_{\hat{X}^{n,0}_{\text{DUDE}}}(x^n, z^n)\Big| \leq ||\mathbf{\Lambda}||_\infty,$$

and, by (22),

$$
\begin{aligned}
\Big|\hat{L}_{\hat{X}^{n,0}_{\text{DUDE}}}(z^n)\Big| &\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathcal{A}}\big|\mathbf{\Pi}^{-T}(x,z_i)\big|\sum_{z\in\mathcal{A}}\Lambda(x,\hat{x}_i(z))\mathbf{\Pi}(x,z) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathcal{A}}\big|\mathbf{\Pi}^{-T}(x,z_i)\big|\,||\mathbf{\Lambda}||_\infty\sum_{z\in\mathcal{A}}\mathbf{\Pi}(x,z) \\
&\leq M||\mathbf{\Lambda}||_\infty||\mathbf{\Pi}^{-1}||_\infty.
\end{aligned}
$$

Therefore, for all x^n, z^n,

$$\Big(L_{\hat{X}^{n,0}_{\text{DUDE}}}(x^n, z^n) - \hat{L}_{\hat{X}^{n,0}_{\text{DUDE}}}(z^n)\Big)^2 \leq ||\mathbf{\Lambda}||_\infty^2\big(1 + M||\mathbf{\Pi}^{-1}||_\infty\big)^2. \tag{55}$$

For all x^n,

$$
\begin{aligned}
&E\Big[(L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n,Z^n)-\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(Z^n))^2\Big]\\
&\quad=E\Big[(L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n,Z^n)-\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(Z^n))^2 1\Big(Z^n \text{ is } \hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}\text{-}\textit{continuous}\Big)\Big]\\
&\qquad+E\Big[(L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n,Z^n)-\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(Z^n))^2 1\Big(Z^n \text{ is } \hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}\text{-}\textit{discontinuous}\Big)\Big]\\
&\quad\le E\Big[(L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n,Z^n)-\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(Z^n))^2 1\Big(Z^n \text{ is } \hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}\text{-}\textit{continuous}\Big)\Big]\\
&\qquad+||\mathbf{\Lambda}||_\infty^2\big(1+M||\mathbf{\Pi}^{-1}||_\infty\big)^2 P\Big(Z^n \text{ is } \hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}\text{-}\textit{discontinuous}\Big)\\
&\quad\le E\Big[(L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n,Z^n)-\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(Z^n))^2 1\Big(Z^n \text{ is } \hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}\text{-}\textit{continuous}\Big)\Big]\\
&\qquad+||\mathbf{\Lambda}||_\infty^2\big(1+M||\mathbf{\Pi}^{-1}||_\infty\big)^2\frac{C}{\sqrt{n}} \qquad (56)
\end{aligned}
$$

where the first inequality follows from (55) and the second from Lemma 8. Recall that if z^n is $\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}$-*continuous* then for all i, $1\le i\le n$, and all $z'\in\mathcal{A}$,

$$
g\big(z_1^{i-1}\cdot z'\cdot z_{i+1}^n,z'\big)=g\big(z_1^{i-1}\cdot z_i\cdot z_{i+1}^n,z'\big). \qquad (57)
$$

Therefore, if z^n is $\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}$-*continuous* then

$$
\begin{aligned}
\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(z^n)&=\frac{1}{n}\sum_{i=1}^n\sum_{x\in\mathcal{A}}\mathbf{\Pi}^{-T}(x,z_i)\sum_{z\in\mathcal{A}}\Lambda(x,g(z_1^{1-i}\cdot z\cdot z_{i+1}^n,z))\mathbf{\Pi}(x,z)\\
&=\frac{1}{n}\sum_{i=1}^n\sum_{x\in\mathcal{A}}\mathbf{\Pi}^{-T}(x,z_i)\sum_{z\in\mathcal{A}}\Lambda(x,g(z^n,z))\mathbf{\Pi}(x,z) \qquad (58)\\
&=\frac{1}{n}\sum_{z'\in\mathcal{A}}\mathbf{m}(z^n)[z']\sum_{x\in\mathcal{A}}\mathbf{\Pi}^{-T}(x,z')\sum_{z\in\mathcal{A}}\Lambda(x,g(z^n,z))\mathbf{\Pi}(x,z)\\
&=\frac{1}{n}\sum_{x\in\mathcal{A}}\left(\sum_{z'\in\mathcal{A}}\mathbf{m}(z^n)[z']\mathbf{\Pi}^{-T}(x,z')\right)\sum_{z\in\mathcal{A}}\Lambda(x,g(z^n,z))\mathbf{\Pi}(x,z)\\
&=\frac{1}{n}\sum_{x\in\mathcal{A}}\sum_{z\in\mathcal{A}}\hat{\mathbf{q}}(z^n)(x,z)\Lambda(x,g(z^n,z))
\end{aligned}
$$

where (58) follows from (57), and where

$$
\hat{\mathbf{q}}(z^n)(x,z)=\left(\sum_{z'\in\mathcal{A}}\mathbf{m}(z^n)[z']\mathbf{\Pi}^{-T}(x,z')\right)\mathbf{\Pi}(x,z).
$$

The $M\times M$ matrix $\mathbf{q}(x^n,z^n)$ is defined as

$$
\mathbf{q}(x^n,z^n)(a,b)\stackrel{\text{def}}{=}\left|\{i:1\le i\le n,x_i=a,z_i=b\}\right|.
$$

Then, again by (57), the true loss is

$$
L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n,z^n)=\frac{1}{n}\sum_{i=1}^n\Lambda(x_i,g(z^n,z_i))=\frac{1}{n}\sum_{x\in\mathcal{A}}\sum_{z\in\mathcal{A}}\mathbf{q}(x^n,z^n)(x,z)\Lambda(x,g(z^n,z)).
$$

Therefore, if z^n is $\hat{L}_{\hat{X}^{n,0}_{\rm DUDE}}$-*continuous* then

$$\begin{aligned}\left|L_{\hat{X}^{n,0}_{\rm DUDE}}(x^n,z^n)-\hat{L}_{\hat{X}^{n,0}_{\rm DUDE}}(z^n)\right| &= \frac{1}{n}\left|\sum_{x\in\mathcal{A}}\sum_{z\in\mathcal{A}}(\mathbf{q}(x^n,z^n)(x,z)-\hat{\mathbf{q}}(z^n)(x,z))\Lambda(x,g(z^n,z))\right| \\ &\le \frac{1}{n}\sum_{x\in\mathcal{A}}\sum_{z\in\mathcal{A}}|\mathbf{q}(x^n,z^n)(x,z)-\hat{\mathbf{q}}(z^n)(x,z)|\,||\mathbf{\Lambda}||_\infty.\end{aligned}$$

Using the fact that for all random variables X_i, $1\le i\le M$,

$$E\left[\left(\sum_{i=1}^M X_i\right)^2\right]\le M\sum_{i=1}^M E\big[X_i^2\big]$$

we obtain that

$$\begin{aligned}&E\left[\Big(L_{\hat{X}^{n,0}_{\rm DUDE}}(x^n,Z^n)-\hat{L}_{\hat{X}^{n,0}_{\rm DUDE}}(Z^n)\Big)^2 1\Big(Z^n \text{ is } \hat{L}_{\hat{X}^{n,0}_{\rm DUDE}}\text{-}continuous\Big)\right] \\ &\le \frac{1}{n^2}||\mathbf{\Lambda}||_\infty^2 E\left[\sum_{x\in\mathcal{A}}\sum_{z\in\mathcal{A}}|\mathbf{q}(x^n,Z^n)(x,z)-\hat{\mathbf{q}}(Z^n)(x,z)|\right]^2 \\ &\le \frac{1}{n^2}||\mathbf{\Lambda}||_\infty^2 M^2\sum_{x\in\mathcal{A}}\sum_{z\in\mathcal{A}}E[\mathbf{q}(x^n,Z^n)(x,z)-\hat{\mathbf{q}}(Z^n)(x,z)]^2. \qquad (59)\end{aligned}$$

Now, notice that one can write

$$\mathbf{q}(x^n,z^n)(x,z)-\hat{\mathbf{q}}(z^n)(x,z)=\sum_{i=1}^n\xi_i$$

where

$$\xi_i=1(x_i=x,z_i=z)-\left(\sum_{z'\in\mathcal{A}}1\big(z_i=z'\big)\mathbf{\Pi}^{-T}(x,z')\right)\mathbf{\Pi}(x,z).$$

Observe that

$$\begin{aligned}E[\xi_i] &= \sum_{z''}\mathbf{\Pi}(x_i,z'')1\big(x_i=x,z''=z\big)-\sum_{z''}\mathbf{\Pi}(x_i,z'')\left(\sum_{z'\in\mathcal{A}}1\big(z''=z'\big)\mathbf{\Pi}^{-T}(x,z')\right)\mathbf{\Pi}(x,z) \\ &= \mathbf{\Pi}(x,z)1(x_i=x)-\left(\sum_{z'\in\mathcal{A}}\mathbf{\Pi}(x_i,z')\mathbf{\Pi}^{-T}(x,z')\right)\mathbf{\Pi}(x,z) \\ &= \mathbf{\Pi}(x,z)1(x_i=x)-1(x_i=x)\mathbf{\Pi}(x,z) \\ &= 0.\end{aligned}$$

Further, the ξ_i are independent and bounded. Therefore, there exists a constant $c_1>0$ such that

$$E[\mathbf{q}(x^n,z^n)(x,z)-\hat{\mathbf{q}}(z^n)(x,z)]^2=\sum_{i=1}^n E\big[\xi_i^2\big]\le c_1 n.$$

Substituting in (59) we obtain that

$$E\left[\Big(L_{\hat{X}^{n,0}_{\rm DUDE}}(x^n,Z^n)-\hat{L}_{\hat{X}^{n,0}_{\rm DUDE}}(Z^n)\Big)^2 1\Big(Z^n \text{ is } \hat{L}_{\hat{X}^{n,0}_{\rm DUDE}}\text{-}continuous\Big)\right]\le\frac{c_1||\mathbf{\Lambda}||_\infty^2 M^4}{n}. \qquad (60)$$

Substituting this in (56), we obtain the lemma. □

6 TU-DUDE construction and proof

In this section, we present the construction of the TU-DUDE $\hat{X}^{n,k_n}_{\text{TU-DUDE}}$ and prove Theorem 1 establishing that it is twice–universal with a penalty $\epsilon_{\text{TU},k_n}(k,n) = \tilde{C}((k_n+1)^{5/4}/n^{1/4})$. We conclude the section with a comparison of TU-DUDE to twice universal data compression schemes and some comments about the TU-DUDE construction.

6.1 Construction

The TU-DUDE is based on the D-DUDE, a deinterleaved version of the DUDE algorithm.[2] For $j = 0, \ldots, k$, define $\tilde{\mathbf{m}}_j(z^n, c_{-k}^{-1}, c_1^k)$ as

$$\tilde{\mathbf{m}}_j\Big(z^n, c_{-k}^{-1}, c_1^k\Big)[c_0] = \Big|\Big\{i : k+1 \le i \le n-k, i = j \bmod (k+1), z_{i-k}^{i+k} = c_{-k}^k\Big\}\Big|$$

for $c_0 \in \mathcal{A}$. The D-DUDE with parameter k denoises according to

$$\hat{X}^{n,k}_{\text{D-DUDE}}(z^n)[i] = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T\Big(\Big(\mathbf{\Pi}^{-T}\tilde{\mathbf{m}}_{j(i)}\Big(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}\Big)\Big) \odot \pi_{z_i}\Big), \tag{61}$$

where $j(i) = i \bmod (k+1)$. Thus, the D-DUDE denoises the i-th symbol using only symbol occurrences for which the index coincides with i modulo $k+1$. The D-DUDE satisfies all of the performance guarantees proved in [1] for the DUDE, including (4); in fact, the proofs in [1] actually involve a deinterleaving step. Thus, we have the following lemma.

Lemma 10. *For all k and sufficiently large n*

$$\hat{R}_k\Big(\hat{X}^{n,k}_{D\text{-}DUDE}\Big) \le C\sqrt{\frac{M^{2k}(k+1)}{n}}. \tag{62}$$

Following the paradigm of Section 4, given a sequence k_n, the TU-DUDE evaluates the estimated loss of the D-DUDE for all parameter values $k \le k_n$ and denoises using the minimizing value. Formally, the TU-DUDE is defined as

$$\hat{X}^{n,k_n}_{\text{TU-DUDE}}(z^n)[i] = \hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}(z^n)[i] \tag{63}$$

where

$$\hat{k}_n^* = \arg\min_{k\le k_n} \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(z^n), \tag{64}$$

with $\hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(z^n)$ corresponding to (22).

[2] As explained in more detail in Section 6.3, we rely on D-DUDE as opposed to DUDE for technical reasons related to the proof of the main theorem.

6.2 Proof of Theorem 1

Since $\hat{X}^{n,k}_{\text{D-DUDE}}$ satisfies (62), then by Lemma 3 it suffices to show that for all $k \le k_n$ and all x^n

$$E\Big[\Big|L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(Z^n)\Big|\Big] \le \frac{\tilde{C}((k_n+1)/n)^{1/4}}{2} \tag{65}$$

for a constant $\tilde{C}$. For $j = 0, \ldots, k$, let

$$\Delta_{k,j}(x^n, z^n) = \sum_{i: i = j \bmod (k+1)} \Big[\Lambda(x_i, \hat{X}^{n,k}_{\text{D-DUDE}}(z^n)[i]) - \hat{\Lambda}_i(z^n)\Big] \tag{66}$$

where

$$\hat{\Lambda}_i(z^n) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{A}} \mathbf{\Pi}^{-T}(x, z_i) \sum_{z\in\mathcal{A}} \Lambda(x, \hat{X}^{n,k}_{\text{D-DUDE}}(z_1^{i-1}, z, z_{i+1}^n)[i])\mathbf{\Pi}(x, z). \tag{67}$$

Note the identity

$$L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, z^n) - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(z^n) = \frac{1}{n}\sum_{j=0}^{k} \Delta_{k,j}(x^n, z^n). \tag{68}$$

It follows that

$$E\Big[\Big|L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(Z^n)\Big|\Big] \le \frac{1}{n}\sum_{j=0}^{k} E(|\Delta_{k,j}(x^n, Z^n)|). \tag{69}$$

Let $\mathcal{S}_j \stackrel{\triangle}{=} \{i : i \ne j \bmod (k+1)\}$. By conditioning on $Z^{\mathcal{S}_j} \stackrel{\triangle}{=} \{Z_i : i \in \mathcal{S}_j\}$, it follows that

$$E[|\Delta_{k,j}(x^n, Z^n)|] = E\big[E\big[|\Delta_{k,j}(x^n, Z^n)|\,\big|Z^{\mathcal{S}_j}\big]\big] \le \max_{z^{\mathcal{S}_j}} E\big[|\Delta_{k,j}(x^n, Z^n)|\,\big|Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\big] \tag{70}$$

where $z^{\mathcal{S}_j}$ is a sequence over $\mathcal{A}$ indexed by elements of $\mathcal{S}_j$. Notice that for any index $i \in \mathcal{S}_j^c$, $\{i-k, i-k+1, \ldots, i-1\} \subset \mathcal{S}_j$ and $\{i+1, i+2, \ldots, i+k\} \subset \mathcal{S}_j$. Define

$$\tilde{\mathcal{S}}_{j, c_{-k}^{-1}, c_1^k}(z^{\mathcal{S}_j}) \stackrel{\text{def}}{=} \{i \in \mathcal{S}_j^c : z_{i-k}^{i-1} = c_{-k}^{-1}, z_{i+1}^{i+k} = c_1^k\} \tag{71}$$

and let

$$\tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(x^n, z^n) = \sum_{i\in\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})} \Big[\Lambda(x_i, \hat{X}^{n,k}_{\text{D-DUDE}}(z^n)[i]) - \hat{\Lambda}_i(z^n)\Big] \tag{72}$$

so that

$$\Delta_{k,j}(x^n, z^n) = \sum_{(c_{-k}^{-1}, c_1^k)\in\mathcal{A}^{2k}} \tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(x^n, z^n). \tag{73}$$

It follows from Lemma 4, that for all j and $(c_{-k}^{-1}, c_1^k) \in \mathcal{A}^{2k}$

$$E\Big[\tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(x^n, Z^n)|Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\Big] = 0$$

and for all j, the random variables $\{\tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(x, Z^n) : (c_{-k}^{-1}, c_1^k) \in \mathcal{A}^{2k}\}$ are conditionally independent given $Z^{\mathcal{S}_j}$. Moreover, each such random variable is conditionally distributed like the difference

between the actual and estimated loss for a zero-th order DUDE operating on the subsequence of noisy symbols with indices in $\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})$. Therefore, it follows from Lemma 9 (in unnormalized form) that

$$\sigma^2_{k,j,c_{-k}^{-1},c_1^k} \stackrel{\text{def}}{=} E\left[\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n, Z^n)\big|Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\right]$$

satisfies

$$\sigma^2_{k,j,c_{-k}^{-1},c_1^k} \le b_1 \left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})\right|^{3/2}$$

for some constant b_1. The conditional independence of the $\tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(x^n, Z^n)$ then implies (from (73)) that the conditional variance of $\Delta_{k,j}(x^n, Z^n)$, denoted by

$$\sigma^2_{k,j} \stackrel{\text{def}}{=} E\big[\Delta^2_{k,j}(x^n, Z^n)\big|Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\big] \tag{74}$$

satisfies

$$\sigma^2_{k,j} = \sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} \sigma^2_{k,j,c_{-k}^{-1},c_1^k} \le b_1 \sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} \left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})\right|^{3/2}. \tag{75}$$

Noting that

$$\sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} \left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})\right| = \left|\mathcal{S}_j^c\right| \le \frac{n}{k+1} + 1,$$

it follows that

$$\begin{aligned}
&\sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} \left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})\right|^{3/2} \\
&\qquad \le \left(\max_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} \left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})\right|\right)^{1/2} \sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} \left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})\right| \\
&\qquad \le \left(\frac{n}{k+1}+1\right)^{3/2}
\end{aligned}$$

so that, from (75),

$$\sigma^2_{k,j} \le b_2 \left(\frac{n}{k+1}\right)^{3/2} \tag{76}$$

for some constant b_2 independent of k, n and j. Together with Jensen's inequality, (76) implies that

$$E\big[|\Delta_{k,j}(x^n, Z^n)|\,\big|Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\big] \le (\sigma^2_{k,j})^{1/2} \tag{77}$$

$$\le (b_2)^{1/2}\left(\frac{n}{k+1}\right)^{3/4} \tag{78}$$

which, together with (69) and (70), implies

$$\begin{aligned}
E\Big[\Big|L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(Z^n)\Big|\Big] &\le (b_2)^{1/2}\frac{k+1}{n}\left(\frac{n}{k+1}\right)^{3/4} = b_3\left(\frac{k+1}{n}\right)^{1/4} \\
&\le b_3\left(\frac{k_n+1}{n}\right)^{1/4}
\end{aligned} \tag{79}$$

for some constant b_3. □

6.3 Discussion

Twice–universal denoising versus twice–universal compression. In lossless data compression, a universal code for a class of compressors has the property that its redundancy with respect to the class vanishes with increasing length. A twice–universal code for a hierarchy of compressor classes has the property that its redundancy with respect to any given compressor equals the best possible redundancy with respect to the smallest class in the hierarchy containing this compressor, plus a "penalty" for twice–universality that is negligible relative to the main redundancy term. In particular, for the hierarchy of Markov compressors which parallels the family of sliding window denoisers, the main redundancy term is essentially $0.5K(\log n)/n$, where K is the number of free parameters corresponding to the given compressor, and the penalty is $c(K)/n$, where the numerator is a function of K. While in this setting the redundancy bound is required to hold for *any* K, it should be noticed that universality is only interesting when the redundancy vanishes, which implicitly limits K to $o(n/(\log n))$.

In the denoising setting, the corresponding property for the proposed TU-DUDE is formulated using the regret of DUDE with respect to sliding window denoisers, which is larger than the best possible regret, as established in [6]. In addition, in the case of the TU-DUDE with the the above parameter κ_n, for example, we cannot claim the penalty to be negligible for any *fixed* k, but only for values of k in the upper half of the range $0 \le k \le \kappa_n$. The relative weakness of these claims are best understood by examining the workings of the simplest twice–universal source codes, which basically search for the best model size, encode this value, and then universally encode the data with a code designed for this optimal model size.[3] A similar approach in the denoising setting is not possible, since the loss resulting from a given choice of k cannot be computed, as the clean sequence x^n is not observable. The TU-DUDE overcomes this problem by applying the value of k that minimizes an *estimated* loss over the range $0 \le k \le \kappa_n$. As the search space increases, the effect of the estimation error becomes more noticeable, and therefore the penalty grows with κ_n, unlike the data compression case.

Why D-DUDE and not DUDE? The technique underlying the proof of Theorem 1 does not directly apply to a denoiser based on the original DUDE with a context parameter selected using the loss estimator. The difficulty is that in a DUDE-based denoiser, the random variables $\tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(x^n, Z^n)$ for different contexts may no longer be conditionally independent given $Z^{\mathcal{S}_j}$, thereby greatly complicating the analysis of the variance of their sum. The D-DUDE, on the other hand, induces such a conditional independence. Whether or not replacing D-DUDE with the original DUDE in $\hat{X}^{n,k_n}_{\text{TU-DUDE}}$ continues to yield the twice–universality properties of Theorem 1 is thus an open question.

[3] Though leading to twice–universality, this approach is known to be sub–optimal, since a mixture over all classes will produce a shorter code length [4].

7 Non-pathological sequences

In this section, we explore the extent to which it is possible to claim a smaller twice–universality penalty than in Theorem 1 if the set of underlying clean sequences x^n is restricted to be "benign." To this end, we show that the loss estimator (22) applied to the D-DUDE concentrates more closely around the true loss for certain "non-pathological" clean sequences. Specifically, the source of much of the twice–universality penalty of Theorem 1 can be attributed to the probability that the accumulated D-DUDE counts $\tilde{\mathbf{m}}_j$ are such that the decision problem (3) has two or more (near) optimal solutions (i.e., lies at a decision boundary). Non-pathological clean sequences are those for which this probability decays sufficiently rapidly. We formalize this notion and show that for the corresponding non-pathological clean sequences x^n, the TU-DUDE with parameter k_n attains an improved twice–universality penalty of roughly

$$\epsilon_{\text{TU-NP},k_n}(k,n) = \mathcal{O}\left(\sqrt{\frac{(k_n+1)^3 \log n}{n}}\right)$$

for $k \le k_n$, which is less than the one proved in Theorem 1.

Let $\tilde{\mathbf{m}}_j(z^n, c_{-k}^{-1}, c_1^k)$ be as defined in Section 6 with respect to the D-DUDE. Let $\mathcal{M}$ denote the set of all M-dimensional vectors with non-negative components. For $a, b, c \in \mathcal{A}$, let

$$\mathcal{M}^*(a,b,c) \stackrel{\text{def}}{=} \{\mathbf{m} \in \mathcal{M} : (\lambda_a - \lambda_b)^T \big((\mathbf{\Pi}^{-T}\mathbf{m}) \odot \pi_c\big) = 0\}$$

denote the M-dimensional semi-hyperplane that contains all the $\mathbf{m}$'s that might fall on a decision boundary of the decision rule (37) underlying DUDE and D-DUDE, involving reconstruction symbols a, b, and noisy symbol c. Let

$$\mathcal{M}^* \stackrel{\text{def}}{=} \bigcup_{\substack{a,b,c \in \mathcal{A}: \\ a \neq b}} \mathcal{M}^*(a,b,c).$$

We term a clean sequence "non-pathological" if the expected values of the counts $\tilde{\mathbf{m}}_j$, when significant, are bounded away from $\mathcal{M}^*$ for all j. Formally, x^n is said to be (α, γ, k_n)-non-pathological, if, for all $k \le k_n$, all j, $0 \le j \le k$, and all c_{-k}^{-1}, c_1^k, if $E\big[||\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)||_1\big] \ge \frac{\alpha \log n}{2(1+\gamma)}$, then, for all $\mathbf{m}^* \in \mathcal{M}^*$,

$$\left|\left|E\Big[\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)\Big] - \mathbf{m}^*\right|\right|_1 > \gamma \left|\left|E\Big[\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)\Big]\right|\right|_1 + 2 + \frac{\alpha \log n}{2}. \tag{80}$$

Letting $\mathcal{N}_n$ denote the set of (α, γ, k_n)-non-pathological sequences[4] we shall define the (α, γ, k_n)-non-pathological k-th order regret of the TU-DUDE of order k_n to be

$$\hat{R}_{NP,k}\Big(\hat{X}^{n,k_n}_{\text{TU-DUDE}}\Big) \stackrel{\text{def}}{=} \max_{x^n \in \mathcal{N}_n} \Big\{E\Big[L_{\hat{X}^{n,k_n}_{\text{TU-DUDE}}}\Big(x_{k+1}^{n-k}, Z^n\Big)\Big] - \hat{D}_k(x^n)\Big\}.$$

Notice that this quantity differs from the regret defined in (2) in that the maximization is restricted to $\mathcal{N}_n$. In analogy with (5), we shall say that the TU-DUDE of order k_n is twice–univeral with

[4] To reduce notational clutter, we suppress the dependence of $\mathcal{N}_n$ on (α, γ, k_n).

(α, γ, k_n)-non-pathological penalty $\epsilon(k, n)$ if

$$\hat{R}_{NP,k}\Big(\hat{X}^{n,k_n}_{\text{TU-DUDE}}\Big) \leq C\sqrt{\frac{M^{2k}(k+1)}{n}} + \epsilon(k,n) \tag{81}$$

for all sufficiently large n, and all k simultaneously. In this context, we shall refer to $\epsilon(k, n)$ as the (α, γ, k_n)-non-pathological twice–universality penalty of the TU-DUDE of order k_n.

Theorem 11. *For all channels such that the transition matrix $\mathbf{\Pi}$ has only non-zero entries, all sequences k_n satisfying $k_n = O(\log n)$, any $\gamma > 0$, any sufficiently large $\alpha > 0$, and all sufficiently large n, $\hat{X}^{n,k_n}_{TU\text{-}DUDE}$ is twice–universal with (α, γ, k_n)-non-pathological penalty*

$$\epsilon_{TU\text{-}NP,k_n}(k,n) = C\sqrt{\frac{\alpha(k_n+1)^3 \log n}{n}}$$

for a constant C and $k \leq k_n$.

We also show that for many channels and loss functions, the fraction of sequences that are (α, γ, k_n)-non-pathological tends to one as long as $k_n \leq \tau \log n$, where τ depends on the channel. Let $\mathbf{P}^*$ denote the M-dimensional column vector whose entries are all $\frac{1}{M}$. We prove this result for all channels with transition probability matrices $\mathbf{\Pi}$ such that $\mathbf{\Pi}^T\mathbf{P}^*$ does not fall within $\mathcal{M}^*$. Many channels and loss functions, *e.g.*, many symmetric channels, including the BSC with crossover probability $0 < \delta < \frac{1}{2}$, and the Hamming loss function, satisfy this requirement.

For all $\mathbf{\Pi}$, let

$$\Delta(\mathbf{\Pi}, \mathbf{\Lambda}) \stackrel{\text{def}}{=} \inf_{\mathbf{m}^* \in \mathcal{M}^*} ||\mathbf{m}^* - \mathbf{\Pi}^T\mathbf{P}^*||_1, \tag{82}$$

where the dependence on $\mathbf{\Lambda}$ is through $\mathcal{M}^*$. Recalling that $\mathcal{N}_n$ denotes the set of all (α, γ, k_n)-non-pathological clean sequences, we prove the following theorem.

Theorem 12. *If $\mathbf{\Pi}, \mathbf{\Lambda}$ are such that $\mathbf{\Pi}^T\mathbf{P}^* \notin \mathcal{M}^*$, then for $k_n \leq \tau \log n$, where*

$$\tau < \frac{1}{-2\log\left(\min_{a \in \mathcal{A}}(\mathbf{\Pi}^T\mathbf{P}^*)[a]\right)},$$

all $\gamma < \Delta(\mathbf{\Pi}, \mathbf{\Lambda})$ and all $\alpha > 0$, we have $|\mathcal{N}_n| = M^n(1 - o(1))$.

Observe that for symmetric channels, $\mathbf{\Pi}$ is such that $\mathbf{\Pi}^T\mathbf{P}^*[a] = M^{-1}$ for all $a \in \mathcal{A}$. Therefore, it follows from Theorems 11 and 12 that, for all $k_n \leq \tau \log n$, where $\tau < (2\log M)^{-1}$, all $\gamma < \Delta(\mathbf{\Pi}, \mathbf{\Lambda})$ and all sufficiently large α, the (α, γ, k_n)-non-pathological twice–universality penalty of the TU-DUDE with parameter k_n is $C\sqrt{(\alpha(k_n+1)^3 \log n)/n}$ for a fraction of clean sequences $(|\mathcal{N}_n|/M^n)$ tending to one. Recall from Section 3 that the sequence κ_n is approximately $\log n/(2\log M)$. Thus, for a fraction of clean sequences tending to one, the above more favorable twice–universality penalty holds for the TU-DUDE with parameter arbitrarily close to this most "ambitious" sequence. Notice that the above (α, γ, k_n)-non-pathological twice–universality penalty is negligible relative to the DUDE redundancy for k over a wider range, roughly $O(\log\log n) < k < k_n$, as compared to $\kappa_n/2 < k \leq k_n$ in the unconstrained case. To get this result, however, we cannot quite set k_n to κ_n, as in the unconstrained case, but only to $(1-\epsilon)\kappa_n$, for an arbitrarily small, but fixed $\epsilon > 0$.

The proofs of the above theorems constitute the rest of this section. For $0 \le j \le k$, we shall consider a set of "bad sequences" $z^n \in \mathcal{A}^n$ for which there exists a sufficiently populated context c_{-k}^{-1}, c_1^k for which $\tilde{\mathbf{m}}_j(z^n, c_{-k}^{-1}, c_1^k)$ falls within $\mathcal{M}^*$ if at most one coordinate z_i, with $i = j \bmod (k+1)$, is changed. Recall that the complement of the set of indices $\{i : i = j \bmod (k+1)\}$ was denoted as $\mathcal{S}_j$ in the proof of Theorem 1, and this notation shall appear again below. Formally, for $\alpha > 0$, the set of "bad sequences" will be taken to be

$$\begin{aligned}\mathcal{B}_{k,\alpha}(j) \stackrel{\text{def}}{=} \{z^n :& \exists c_{-k}^{-1}, c_1^k \in \mathcal{A}^k, \text{ s.t. } ||\tilde{\mathbf{m}}_j(z^n, c_{-k}^{-1}, c_1^k)||_1 \ge \alpha \log n \text{ and} \\ & \exists\, \mathbf{m}^* \in \mathcal{M}^*, ||\tilde{\mathbf{m}}_j(z^n, c_{-k}^{-1}, c_1^k) - \mathbf{m}^*||_1 \le 2\}. \end{aligned} \tag{83}$$

The following lemma implies that for all (α, γ, k_n)-non-pathological sequences, the probability that the noisy sequence Z^n is in $\mathcal{B}_{k,\alpha}(j)$ vanishes for a sufficiently large choice of α.

Lemma 13. *For all (α, γ, k_n)-non-pathological x^n with γ, $0 < \gamma < 1$, all $k \le k_n$, all j, $0 \le j \le k+1$,*

$$P(Z^n \in \mathcal{B}_{k,\alpha}(j)) \le \beta_1 \left(\frac{n + 2k + 2}{(k+1)\alpha \log n} \right) e^{-\beta_2 \alpha \log n + \beta_3 k}$$

where $\beta_1, \beta_2, \beta_3$ are positive functions of γ and M.

Proof. We abbreviate $\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)$ by $\tilde{\mathbf{m}}_j$. Assuming $Z^n \in B_{k,\alpha}(j)$, let c_{-k}^{-1}, c_1^k be such that $||\tilde{\mathbf{m}}_j||_1 \ge \alpha \log n$, and let $\mathbf{m}^* \in \mathcal{M}^*$ be such that $||\tilde{\mathbf{m}}_j - \mathbf{m}^*||_1 \le 2$. By the triangle inequality we then have

$$||\mathbf{m}^* - E[\tilde{\mathbf{m}}_j]||_1 - ||E[\tilde{\mathbf{m}}_j] - \tilde{\mathbf{m}}_j||_1 \le 2. \tag{84}$$

If $E[||\tilde{\mathbf{m}}_j||_1] \ge \frac{\alpha \log n}{2(1+\gamma)}$, then since x^n is (α, γ, k_n)-non-pathological, by (80) and (84)

$$||E[\tilde{\mathbf{m}}_j] - \tilde{\mathbf{m}}_j||_1 > \gamma ||E[\tilde{\mathbf{m}}_j]||_1 + \frac{\alpha \log n}{2}. \tag{85}$$

If $E[||\tilde{\mathbf{m}}_j||_1] < \frac{\alpha \log n}{2(1+\gamma)}$, then since $||\tilde{\mathbf{m}}_j||_1 \ge \alpha \log n$

$$||E[\tilde{\mathbf{m}}_j] - \tilde{\mathbf{m}}_j||_1 > \alpha \log n \left(1 - \frac{1}{2(1+\gamma)}\right) = \alpha \log n \left(\frac{\gamma}{2(1+\gamma)}\right) + \frac{\alpha \log n}{2} \ge \gamma ||E[\tilde{\mathbf{m}}_j]||_1 + \frac{\alpha \log n}{2},$$

i.e., (85) still holds. Therefore, if $Z^n \in \mathcal{B}_{k,\alpha}(j)$, there exists c_{-k}^{-1}, c_1^k such that (85) holds, which implies that there exists $c \in \mathcal{A}$ such that

$$|\tilde{\mathbf{m}}_j[c] - E[\tilde{\mathbf{m}}_j[c]]| > \gamma E[\tilde{\mathbf{m}}_j[c]] + \frac{\alpha \log n}{2M}. \tag{86}$$

For $j = 0, \ldots, k$, and $p = 0, 1$, define $\tilde{\mathbf{m}}_j^p(z^n, c_{-k}^{-1}, c_1^k)$ by

$$\begin{aligned}&\tilde{\mathbf{m}}_j^p\left(z^n, c_{-k}^{-1}, c_1^k\right)[c_0] \\ &\quad = \left| \left\{ i : k+1 \le i \le n-k, i = (k+1)(2q+p) + j \text{ for some } q \in \mathbb{Z}, z_{i-k}^{i+k} = c_{-k}^k \right\} \right|,\end{aligned}$$

namely, the number of times $z_{i-k}^{i+k} = c_{-k}^k$ with $i-j$ being an even (for $p=0$) or an odd (for $p=1$) multiple of $k+1$, so that

$$\tilde{\mathbf{m}}_j\Big(z^n, c_{-k}^{-1}, c_1^k\Big) = \tilde{\mathbf{m}}_j^0\Big(z^n, c_{-k}^{-1}, c_1^k\Big) + \tilde{\mathbf{m}}_j^1\Big(z^n, c_{-k}^{-1}, c_1^k\Big).$$

Abbreviating $\tilde{\mathbf{m}}_j^p(Z^n, c_{-k}^{-1}, c_1^k)$ by $\tilde{\mathbf{m}}_j^p$, it follows that the existence of $c \in \mathcal{A}$ satisfying (86) implies that, for some $p \in \{0,1\}$,

$$\Big|\tilde{\mathbf{m}}_j^p[c] - E\Big[\tilde{\mathbf{m}}_j^p[c]\Big]\Big| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c]\Big] + \frac{\alpha \log n}{4M}. \tag{87}$$

Therefore,

$$\begin{aligned}
&P(Z^n \in \mathcal{B}_{k,\alpha}(j)) \\
&\quad \le P\left(\exists c_{-k}^k \in \mathcal{A}^{2k+1}, p \in \{0,1\}, \text{ s.t. } \Big|\tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big]\Big| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] + \frac{\alpha \log n}{4M}\right) \\
&\quad \le \sum_{p\in\{0,1\}} \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}} P\left(\Big|\tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big]\Big| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] + \frac{\alpha \log n}{4M}\right) \qquad (88) \\
&\quad = \sum_{p\in\{0,1\}} \sum_{\substack{c_{-k}^k \in \mathcal{A}^{2k+1}: \\ E\left[\tilde{\mathbf{m}}_j^p[c_0]\right] > \frac{\alpha \log n}{8M}}} P\left(\Big|\tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big]\Big| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] + \frac{\alpha \log n}{4M}\right) \\
&\qquad + \sum_{p\in\{0,1\}} \sum_{\substack{c_{-k}^k \in \mathcal{A}^{2k+1}: \\ E\left[\tilde{\mathbf{m}}_j^p[c_0]\right] \le \frac{\alpha \log n}{8M}}} P\left(\Big|\tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big]\Big| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] + \frac{\alpha \log n}{4M}\right) \qquad (89)
\end{aligned}$$

where (88) follows from the union bound. Observe that for all j, c_{-k}^k, and $p \in \{0,1\}$, $\tilde{\mathbf{m}}_j^p[c_0]$ is a sum of independent $0-1$ random variables.[5] We can then apply Theorem 2.3(b) in [12], from which it follows that for a collection $Y_1, Y_2, \ldots, Y_n$ of independent random variables, $0 \le Y_i \le 1$, $S_n = \sum_i Y_i$, and any $\epsilon > 0$,

$$P(|S_n - E[S_n]| \ge \epsilon E[S_n]) \le 2e^{-\frac{\epsilon^2 E[S_n]}{2(1+\frac{\epsilon}{3})}}. \tag{90}$$

Thus, we obtain that for all j, c_{-k}^k, and $p \in \{0,1\}$,

$$\begin{aligned}
P\left(\Big|\tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big]\Big| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] + \frac{\alpha \log n}{4M}\right) &\le P\left(\Big|\tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big]\Big| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big]\right) \\
&\le 2e^{-\frac{\gamma^2 E\left[\tilde{\mathbf{m}}_j^p[c_0]\right]}{8(1+\frac{\gamma}{6})}} \qquad (91)
\end{aligned}$$

and, also,

$$\begin{aligned}
P\left(\Big|\tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big]\Big| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] + \frac{\alpha \log n}{4M}\right) &\le P\left(\Big|\tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big]\Big| > \frac{\alpha \log n}{4M}\right) \\
&\le 2e^{-\frac{\alpha^2 (\log n)^2}{32M^2\left(E\left[\tilde{\mathbf{m}}_j^p[c_0]\right] + \frac{\alpha \log n}{12M}\right)}}. \qquad (92)
\end{aligned}$$

[5] Observe that, due to overlap, this is not, in general, the case for $\tilde{m}_j[c_0]$.

Notice that for all z^n

$$\sum_{c_{-k}^k \in \mathcal{A}^{2k+1}} \tilde{\mathbf{m}}_j^p[c_0] \le \left\lceil \frac{n}{2(k+1)} \right\rceil. \tag{93}$$

Hence, the number of c_{-k}^k for which $E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] > \frac{\alpha \log n}{8M}$ is at most $\frac{8M(n+2k+2)}{(2k+2)\alpha \log n}$. Combining this fact with (91), we obtain that

$$\begin{aligned}
&\sum_{p\in\{0,1\}} \sum_{c_{-k}^k : E\left[\tilde{\mathbf{m}}_j^p[c_0]\right] > \frac{\alpha \log n}{8M}} P\left(\left| \tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] \right| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] + \frac{\alpha \log n}{4M} \right) \\
&\quad \le \sum_{p\in\{0,1\}} \sum_{c_{-k}^k : E\left[\tilde{\mathbf{m}}_j^p[c_0]\right] > \frac{\alpha \log n}{8M}} 2e^{-\frac{\gamma^2 E\left[\tilde{\mathbf{m}}_j^p[c_0]\right]}{8(1+\frac{\gamma}{6})}} \\
&\quad \le \sum_{p\in\{0,1\}} \sum_{c_{-k}^k : E\left[\tilde{\mathbf{m}}_j^p[c_0]\right] > \frac{\alpha \log n}{8M}} 2e^{-\frac{\gamma^2 \alpha \log n}{64M(1+\frac{\gamma}{6})}} \\
&\quad \le 4\left(\frac{8M(n+2k+2)}{(2k+2)\alpha \log n}\right) e^{-\frac{\gamma^2 \alpha \log n}{64M(1+\frac{\gamma}{6})}}.
\end{aligned} \tag{94}$$

On the other hand, by (92), we obtain that

$$\begin{aligned}
&\sum_{p\in\{0,1\}} \sum_{c_{-k}^k : E\left[\tilde{\mathbf{m}}_j^p[c_0]\right] \le \frac{\alpha \log n}{8M}} P\left(\left| \tilde{\mathbf{m}}_j^p[c_0] - E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] \right| > \frac{\gamma}{2} E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] + \frac{\alpha \log n}{4M} \right) \\
&\quad \le \sum_{p\in\{0,1\}} \sum_{c_{-k}^k : E\left[\tilde{\mathbf{m}}_j^p[c_0]\right] \le \frac{\alpha \log n}{8M}} 2e^{-\frac{\alpha^2 (\log n)^2}{32M^2\left(E\left[\tilde{\mathbf{m}}_j^p[c_0]\right] + \frac{\alpha \log n}{12M}\right)}} \\
&\quad \le 4M^{2k+1} e^{-\frac{\alpha^2 (\log n)^2}{32M^2\left(\frac{\alpha \log n}{8M} + \frac{\alpha \log n}{12M}\right)}} \\
&\quad = 4M^{2k+1} e^{-\frac{3\alpha \log n}{20M}}
\end{aligned} \tag{95}$$

Substituting (94) and (95) in (89), we obtain that

$$P(Z^n \in \mathcal{B}_{k,\alpha}(j)) \le \beta_1 \left(\frac{n+2k+2}{(k+1)\alpha \log n}\right) e^{-\beta_2 \alpha \log n + \beta_3 k}$$

where β_1, β_2, and β_3 are positive functions of γ and M. □

The proof of Theorem 11 shall rely on the following definitions. For $0 \le j \le k$, define $g_j : \mathcal{A}^n \times \mathcal{A}^k \times \mathcal{A}^k \times \mathcal{A} \to \mathcal{A}$ to be

$$g_j(z^n, c_{-k}^{-1}, c_1^k, z') \stackrel{\text{def}}{=} \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \Big((\mathbf{\Pi}^{-T} \tilde{\mathbf{m}}_j(z^n, c_{-k}^{-1}, c_1^k)) \odot \pi_{z'} \Big)$$

where ties are broken based on some fixed ordering of the elements of $\mathcal{A}$, so that for $k+1 \le i \le n-k$

$$\hat{X}_{\text{D-DUDE}}^{n,k}(z^n)[i] = g_j(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}, z_i).$$

In analogy to (38), for $0 \le j \le k$, and $c_{-k}^{-1}, c_1^k \in \mathcal{A}^k$, a sequence z^n is said to be (j, c_{-k}^{-1}, c_1^k)-*continuous* if, for all $i \in \mathcal{S}_j^c$ and $z' \in \mathcal{A}$,

$$g_j\Big(z_1^{i-1} \cdot z' \cdot z_{i+1}^n, c_{-k}^{-1}, c_1^k, z'\Big) = g_j\Big(z_1^{i-1} \cdot z_i \cdot z_{i+1}^n, c_{-k}^{-1}, c_1^k, z'\Big). \tag{96}$$

It is said to be (j, c_{-k}^{-1}, c_1^k)-*discontinuous* otherwise. If $z^n \notin \mathcal{B}_{k,\alpha}(j)$ and $||\tilde{\mathbf{m}}_j(z^n, c_{-k}^{-1}, c_1^k)||_1 \geq \alpha \log n$, then it follows from (83) that z^n is (j, c_{-k}^{-1}, c_1^k)-*continuous*.

Proof of Theorem 11. The proof shall follow the reasoning of the proof of Theorem 1. Since $\hat{X}^{n,k}_{\text{D-DUDE}}$ satisfies (62), then by the obvious extension of Lemma 3 to (α, γ, k_n)-non-pathological twice–universality, it suffices to show that, under the assumptions of Theorem 11, for all $k \leq k_n$ and all (α, γ, k_n)-non-pathological x^n,

$$E\Big[\Big|L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(Z^n)\Big|\Big] \leq \frac{C}{2}\sqrt{\frac{\alpha(k_n+1)\log n}{n}} \tag{97}$$

for a constant C. For all k, $i \leq k+1$, $c_{-k}^{-1}, c_1^k \in \mathcal{A}^k$, let $\Delta_{k,j}(\cdot)$, $\hat{\Lambda}_i(\cdot)$, $\tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(\cdot)$ be as they were defined, respectively, in (66), (67), and (72) in Section 6. Then recall from the proof of Theorem 1 ((69) and (73)), that

$$E\Big[\Big|L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(Z^n)\Big|\Big] \leq \frac{1}{n}\sum_{j=0}^{k} E(|\Delta_{k,j}(x^n, Z^n)|) \tag{98}$$

and

$$\Delta_{k,j}(x^n, z^n) = \sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} \tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(x^n, z^n)$$

and that, for all j, the random variables $\{\tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(x^n, Z^n) : (c_{-k}^{-1}, c_1^k) \in \mathcal{A}^{2k}\}$ are conditionally zero mean and independent given $Z^{\mathcal{S}_j}$. Therefore, for all $z^{\mathcal{S}_j}$

$$E\big[\,\Delta_{k,j}^2(x^n, z^n)\big|\, Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\big] = \sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} E\left[\,\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n, z^n)\,\middle|\, Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\right] \tag{99}$$

(see left-most equation in (75)). We again abbreviate $\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)$ by $\tilde{\mathbf{m}}_j$. We depart from the proof of Theorem 1 by separately considering "good" and "bad" sequences z^n, observing that for all j, $0 \leq j \leq k$,

$$\begin{aligned}
&E\left[\,\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n, Z^n)\,\middle|\, Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\right] \\
&\quad= E\left[\,\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n, Z^n)1(Z^n \notin \mathcal{B}_{k,\alpha}(j))\,\middle|\, Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\right] \\
&\qquad\qquad + \sum_{z^n\in\mathcal{B}_{k,\alpha}(j)} P(Z^n = z^n | Z^{\mathcal{S}_j} = z^{\mathcal{S}_j})\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n, z^n) \\
&\quad\leq E\left[\,\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n, Z^n)1(Z^n \notin \mathcal{B}_{k,\alpha}(j))\,\middle|\, Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\right] + \nu_1 n^2 \sum_{z^n\in\mathcal{B}_{k,\alpha}(j)} P\big(Z^n = z^n | Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}\big)
\end{aligned} \tag{100}$$

where (100) follows from the fact that, for all $x^n, z^n, c_{-k}^{-1}, c_1^k$,

$$|\tilde{\Delta}_{k,j,c_{-k}^{-1},c_1^k}(x^n, z^n)| \leq \sqrt{\nu_1}||\tilde{\mathbf{m}}_j(z^n, c_{-k}^{-1}, c_1^k)||_1 \leq \sqrt{\nu_1} n \tag{101}$$

for some constant ν_1. Now

$$
\begin{aligned}
&E\left[\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n,Z^n)1(Z^n\notin\mathcal{B}_{k,\alpha}(j))\,\middle|\, Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\right]\\
&\quad=E\left[\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n,Z^n)1(Z^n\notin\mathcal{B}_{k,\alpha}(j),||\tilde{\mathbf{m}}_j||_1\geq\alpha\log n)\,\middle|\, Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\right]\\
&\qquad+E\left[\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n,Z^n)1(Z^n\notin\mathcal{B}_{k,\alpha}(j),||\tilde{\mathbf{m}}_j||_1<\alpha\log n)\,\middle|\, Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\right]\\
&\quad\leq E\left[\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n,Z^n)1\Big(Z^n\text{ is }(j,c_{-k}^{-1},c_1^k)\text{-}\textit{continuous}\Big)\,\middle|\, Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\right]\\
&\qquad+E\left[\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n,Z^n)1(Z^n\notin\mathcal{B}_{k,\alpha}(j),||\tilde{\mathbf{m}}_j||_1<\alpha\log n)\,\middle|\, Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\right] \qquad (102)\\
&\quad\leq E\left[\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n,Z^n)1\Big(Z^n\text{ is }(j,c_{-k}^{-1},c_1^k)\text{-}\textit{continuous}\Big)\,\middle|\, Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\right]\\
&\qquad+E\left[\nu_1\Big(||\tilde{\mathbf{m}}_j(z^n,c_{-k}^{-1},c_1^k)||_1\Big)^2 1(||\tilde{\mathbf{m}}_j||_1<\alpha\log n)\,\middle|\, Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\right] \qquad (103)
\end{aligned}
$$

where (102) follows from the fact that $z^n\notin\mathcal{B}_{k,\alpha}(j)$ and $||\tilde{\mathbf{m}}_j||_1\geq\alpha\log n$ imply that z^n is (j,c_{-k}^{-1},c_1^k)-*continuous* and (103) from (101). It was shown in the proof of Lemma 9 (see (60)) that for all x^n

$$
E\left[\Big(L_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(x^n,Z^n)-\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}(Z^n)\Big)^2 1\Big(Z^n\text{ is }\hat{L}_{\hat{X}^{n,0}_{\mathrm{DUDE}}}\text{-}\textit{continuous}\Big)\right]\leq\frac{c_1||\mathbf{\Lambda}||_\infty^2M^4}{n}.
$$

Following the same steps, we obtain that, for all $z^{\mathcal{S}_j}$,

$$
E\left(\tilde{\Delta}^2_{k,j,c_{-k}^{-1},c_1^k}(x^n,Z^n)1\Big(Z^n\text{ is }(j,c_{-k}^{-1},c_1^k)\text{-}\textit{continuous}\Big)\,\middle|\, Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\right)\leq c_1||\mathbf{\Lambda}||_\infty^2M^4\left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\right|, \qquad (104)
$$

where the set of indices $\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)$ was defined in (71) and is deterministic when conditioned on $Z^{\mathcal{S}_j}$.

Observe that

$$
||\tilde{\mathbf{m}}_j(z^n,c_{-k}^{-1},c_1^k)||_1=\left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\right|.
$$

Therefore,

$$
\begin{aligned}
&E\left[\nu_1\Big(||\tilde{\mathbf{m}}_j(Z^n,c_{-k}^{-1},c_1^k)||_1\Big)^2 1\Big(||\tilde{\mathbf{m}}_j(Z^n,c_{-k}^{-1},c_1^k)||_1<\alpha\log n\Big)\,\middle|\, Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\right]\\
&\quad=\nu_1\left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\right|^2 1\Big(\left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\right|<\alpha\log n\Big). \qquad (105)
\end{aligned}
$$

Substituting (104) and (105) in (103), combining with (100) and (99), and noting that

$$
\sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}}\left|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\right|=|\mathcal{S}_j^c|\leq\frac{n+k+1}{k+1}, \qquad (106)
$$

we obtain that

$$
\begin{aligned}
&E\big[\,\Delta_{k,j}^2(x^n,z^n)\big|\,Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\big]\\
&\quad\le \nu_2\frac{n+k+1}{k+1}+\nu_1\sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}}\Big|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\Big|^2\,1\Big(\Big|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\Big|<\alpha\log n\Big)\\
&\qquad+\nu_1 n^2\sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}}\ \sum_{z^n\in\mathcal{B}_{k,\alpha}(j)}P\big(Z^n=z^n|Z^{\mathcal{S}_j}=z^{\mathcal{S}_j}\big)
\end{aligned}
\tag{107}
$$

for some constant ν_2. Note that

$$
\begin{aligned}
&\sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}}\Big|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\Big|^2\,1\Big(\Big|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\Big|<\alpha\log n\Big)\\
&\qquad\le(\alpha\log n)\sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}}\Big|\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}\big(z^{\mathcal{S}_j}\big)\Big|\\
&\qquad\le(\alpha\log n)\frac{n+k+1}{k+1}
\end{aligned}
\tag{108}
$$

where the last inequality follows from (106). Substituting (108) in (107), and taking expectation over $Z^{\mathcal{S}_j}$, we obtain

$$
\begin{aligned}
E\big[\Delta_{k,j}^2(x^n,z^n)\big]&\le\nu_2\frac{n+k+1}{k+1}+\nu_1\frac{n+k+1}{k+1}(\alpha\log n)+\nu_1 n^2M^{2k}P(Z^n\in\mathcal{B}_{k,\alpha}(j))\\
&\le\nu_2\frac{n+k+1}{k+1}+\nu_1\frac{n+k+1}{k+1}(\alpha\log n)+\nu_1 n^2M^{2k}\frac{\beta_1(n+2k+2)}{(k+1)\alpha\log n}e^{-\beta_2\alpha\log n+\beta_3 k}
\end{aligned}
\tag{109}
$$

$$
\le C'\left(\frac{\alpha n\log n}{k+1}\right)
\tag{110}
$$

for $k\le k_n=O(\log n)$, some constant C', sufficiently large α (depending on the implicit constant in the $O(\log n)$ bound on k_n), and sufficiently large n, where (109) follows from Lemma 13, and (110) follows from $k\le k_n=O(\log n)$ for, as noted, α and n sufficiently large. Applying Jensen's inequality, we obtain that, for all j, $0\le j\le k$,

$$
E[|\Delta_{k,j}(x^n,z^n)|]\le\sqrt{C'\left(\frac{\alpha n\log n}{k+1}\right)}.
$$

Finally, substituting into (98), we obtain that, for all $k\le k_n=O(\log n)$,

$$
E\Big(|L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n,Z^n)-\hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(Z^n)|\Big)\le\sqrt{C'\left(\frac{\alpha(k+1)\log n}{n}\right)}\le\sqrt{C'\left(\frac{\alpha(k_n+1)\log n}{n}\right)}.
$$

which proves (97) with $C=2\sqrt{C'}$, as desired. □

Proof of Theorem 12. Let $X^n\in\mathcal{A}^n$ be a random *i.i.d.* sequence with uniform distribution $\mathbf{P}^*$. We will show that under certain conditions on k_n, the probability that X^n is *not* (α,γ,k_n)-non-pathological vanishes with n, when $\mathbf{\Pi}^T\mathbf{P}^*\notin\mathcal{M}^*$. It will follow that $|\mathcal{N}_n|=M^n(1-o(1))$.

The outline of the proof is as follows. We first obtain a bound on the probability that the conditional expectation $E\big[\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)|X^n\big]$ deviates by more than a certain amount from its mean $E\big[\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)\big]$. We then show that if the complements of these events occur for all j and all (c_{-k}^{-1}, c_1^k) then X^n is (α, γ, k_n)-non-pathological, for γ and k_n satisfying the conditions of the theorem. The reverse inclusion of the complements of these events is then combined, via a union bound, with the first step, to prove the theorem.

For all j, $0 \le j \le k$, and all $c_{-k}^{-1}, c_1^k \in \mathcal{A}^k$, since Z^n is *i.i.d.* (as induced by X^n being *i.i.d.*),

$$E\Big[\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)\Big] = |\mathcal{S}_j^c| P\Big(Z_1^{2k} = c_{-k}^{-1} c_1^k\Big) \mathbf{\Pi}^T \mathbf{P}^*$$

where the expectation is over all random choices of X^n and channel realizations Z^n. From the fact that X^n is a sequence of *i.i.d.* random variables, it follows that

$$p_{c_{-k}^{-1}c_1^k} \stackrel{\text{def}}{=} P(Z_1^{2k} = c_{-k}^{-1} c_1^k) \ge e^{-2\lambda_1 k} \tag{111}$$

where

$$\lambda_1 = -\log\left(\min_{a \in \mathcal{A}} (\mathbf{\Pi}^T \mathbf{P}^*)[a]\right), \tag{112}$$

which is well defined, since the assumed invertibility of $\mathbf{\Pi}$ implies that there are no all-zero columns, so that, for $1 \le i \le n$, and all $a \in \mathcal{A}$,

$$P(Z_i = a) = (\mathbf{\Pi}^T \mathbf{P}^*)[a] > 0.$$

For $p \in \{0, 1\}$, let $\tilde{\mathbf{m}}_j^p(z^n, c_{-k}^1, c_1^k)$ be as in the proof of Lemma 13. In the sequel, we will abbreviate $\tilde{\mathbf{m}}_j^p(z^n, c_{-k}^1, c_1^k)$ by $\tilde{\mathbf{m}}_j^p$. Also for $p \in \{0, 1\}$, let

$$\mathcal{S}_j^{c,p} = \{i \in \mathcal{S}_j^c : i = (k+1)(2q+p) + j \text{ for some } q \in \mathbb{Z}\}$$

so that $\mathcal{S}_j^c = \mathcal{S}_j^{c,0} \bigcup \mathcal{S}_j^{c,1}$. By the triangle inequality, repeated applications of the union bound (over $p \in \{0, 1\}$ and $c_0 \in \mathcal{A}$), and the fact that $\sum_{c_0 \in \mathcal{A}} (\mathbf{\Pi}^T P^*)[c_0] = ||\mathbf{\Pi}^T P^*||_1 = 1$, we obtain that, for all j, $0 \le j \le k$, $c_{-k}^{-1}, c_1^k \in \mathcal{A}^k$, and $p \in \{0, 1\}$, and all sufficiently large n,

$$\begin{aligned}
&P\left(\left|\left|E\Big[\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)|X^n\Big] - |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k} \mathbf{\Pi}^T \mathbf{P}^*\right|\right|_1 \ge \frac{|\mathcal{S}_j^c|}{\log n} p_{c_{-k}^{-1}c_1^k}\right) \\
&\le \sum_{p \in \{0,1\}} P\left(\left|\left|E\Big[\tilde{\mathbf{m}}_j^p \Big| X^n\Big] - |\mathcal{S}_j^{c,p}| p_{c_{-k}^{-1}c_1^k} \mathbf{\Pi}^T \mathbf{P}^*\right|\right|_1 \ge \frac{|\mathcal{S}_j^{c,p}|}{\log n} p_{c_{-k}^{-1}c_1^k}\right) \\
&\le \sum_{p \in \{0,1\}} \sum_{c_0 \in \mathcal{A}} P\left(\left|E\Big[\tilde{\mathbf{m}}_j^p[c_0] \Big| X^n\Big] - |\mathcal{S}_j^{c,p}| p_{c_{-k}^{-1}c_1^k} (\mathbf{\Pi}^T \mathbf{P}^*)[c_0]\right| \ge \frac{|\mathcal{S}_j^{c,p}|}{\log n} p_{c_{-k}^{-1}c_1^k} (\mathbf{\Pi}^T \mathbf{P}^*)[c_0]\right). \quad (113)
\end{aligned}$$

Noting that, for all j, $0 \le j \le k$, and all $c_{-k}^{-1}, c_1^k \in \mathcal{A}^k$, and $p \in \{0, 1\}$

$$E\left(E\Big[\tilde{\mathbf{m}}_j^p[c_0] \Big| X^n\Big]\right) = E\Big[\tilde{\mathbf{m}}_j^p[c_0]\Big] = |\mathcal{S}_j^{c,p}| p_{c_{-k}^{-1}c_1^k} \mathbf{\Pi}^T \mathbf{P}^*[c_0],$$

and recalling that $\tilde{\mathbf{m}}_j^p[c_0]$, conditioned on X^n, is a sum of independent $0-1$ random variables, since X^n is *i.i.d.*, it follows analogously that $E\Big[\tilde{\mathbf{m}}_j^p[c_0] \Big| X^n\Big]$ is a sum of *i.i.d.* random variables

bounded between 0 and 1. Therefore, we can again invoke Theorem 2.3(b) in [12] (as in (90), with $\epsilon = 1/\log n$) to obtain

$$P\left(\left|E\left[\tilde{\mathbf{m}}_j^p[c_0]\right| X^n\right] - |\mathcal{S}_j^{c,p}| p_{c_{-k}^{-1}c_1^k}(\mathbf{\Pi}^T\mathbf{P}^*)[c_0]\right| \geq \frac{|\mathcal{S}_j^{c,p}|}{\log n} p_{c_{-k}^{-1}c_1^k}(\mathbf{\Pi}^T\mathbf{P}^*)[c_0]\right) \leq 2e^{-\frac{|\mathcal{S}_j^{c,p}| p_{c_{-k}^{-1}c_1^k}(\mathbf{\Pi}^T\mathbf{P}^*)[c_0]}{3(\log n)^2}}. \quad (114)$$

Incorporating (114) into (113) yields that, for all j, $0 \leq j \leq k$, $c_{-k}^{-1}, c_1^k \in \mathcal{A}^k$, and $p \in \{0,1\}$, and all sufficiently large n,

$$P\left(\left|\left|E\left[\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)|X^n\right] - |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k}\mathbf{\Pi}^T\mathbf{P}^*\right|\right|_1 \geq \frac{|\mathcal{S}_j^c|}{\log n} p_{c_{-k}^{-1}c_1^k}\right) \leq 4Me^{-\frac{\left(|\mathcal{S}_j^c|-1\right)e^{-(2k+1)\lambda_1}}{6(\log n)^2}} \quad (115)$$

where we used $|\mathcal{S}_j^{c,p}| \geq \frac{|\mathcal{S}_j^c|-1}{2}$, (111) and (112). This concludes the first step of the proof.

For the next step, we begin by observing that if, for some $x^n \in \mathcal{A}^n$, j, $0 \leq j \leq k$, and $c_{-k}^{-1}, c_1^k \in \mathcal{A}^k$,

$$||E[\tilde{\mathbf{m}}_j|\, X^n = x^n] - |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k}\mathbf{\Pi}^T\mathbf{P}^*||_1 < \frac{|\mathcal{S}_j^c|}{\log n} p_{c_{-k}^{-1}c_1^k} \quad (116)$$

then, since $||\mathbf{\Pi}^T P^*||_1 = 1$,

$$\left(1 - \frac{1}{\log n}\right)|\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k} < ||E[\tilde{\mathbf{m}}_j|\, X^n = x^n]||_1 < \left(1 + \frac{1}{\log n}\right)|\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k}. \quad (117)$$

Also, from (82), for all $\mathbf{m}^* \in \mathcal{M}^*$ and all c_{-k}^{-1}, c_1^k,

$$||\mathbf{m}^* - |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k}\mathbf{\Pi}^T\mathbf{P}^*||_1 \geq |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k}\Delta(\mathbf{\Pi}, \mathbf{\Lambda}). \quad (118)$$

Thus, if x^n, j, and c_{-k}^{-1}, c_1^k satisfy (116), then, by the triangle inequality, for all $\mathbf{m}^* \in \mathcal{M}^*$ and $\gamma < \frac{\Delta(\mathbf{\Pi},\mathbf{\Lambda})}{2}$, it follows that

$$\begin{aligned}
&||E[\tilde{\mathbf{m}}_j|\, X^n = x^n] - \mathbf{m}^*||_1 \\
&\quad \geq ||\mathbf{m}^* - |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k}\mathbf{\Pi}^T\mathbf{P}^*||_1 - ||E[\tilde{\mathbf{m}}_j|\, X^n = x^n] - |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k}\mathbf{\Pi}^T\mathbf{P}^*||_1 \\
&\quad > |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k}\left(\Delta(\mathbf{\Pi}, \mathbf{\Lambda}) - \frac{1}{\log n}\right) \quad (119) \\
&\quad = 2 \cdot \frac{|\mathcal{S}_j^c|}{2} p_{c_{-k}^{-1}c_1^k}\left(\Delta(\mathbf{\Pi}, \mathbf{\Lambda}) - \frac{1}{\log n}\right) \\
&\quad \geq \frac{||E[\tilde{\mathbf{m}}_j|\, X^n = x^n]||_1}{2(1 + \frac{1}{\log n})}\left(\Delta(\mathbf{\Pi}, \mathbf{\Lambda}) - \frac{1}{\log n}\right) + \frac{|\mathcal{S}_j^c|}{2} p_{c_{-k}^{-1}c_1^k}\left(\Delta(\mathbf{\Pi}, \mathbf{\Lambda}) - \frac{1}{\log n}\right) \quad (120) \\
&\quad \geq \gamma||E[\tilde{\mathbf{m}}_j|\, X^n = x^n]||_1 + \frac{|\mathcal{S}_j^c|}{2} p_{c_{-k}^{-1}c_1^k}\left(\Delta(\mathbf{\Pi}, \mathbf{\Lambda}) - \frac{1}{\log n}\right) \\
&\quad \geq \gamma||E[\tilde{\mathbf{m}}_j|\, X^n = x^n]||_1 + \frac{n-(k+1)}{2(k+1)} e^{-2\lambda_1 k}\left(\Delta(\mathbf{\Pi}, \mathbf{\Lambda}) - \frac{1}{\log n}\right) \quad (121)
\end{aligned}$$

$$\geq \gamma||E[\tilde{\mathbf{m}}_j|\, X^n = x^n]||_1 + (2k+1) + \frac{\alpha \log n}{2}, \tag{122}$$

implying, if it holds for all j, c_{-k}^{-1}, c_1^k, and $k \leq k_n$, that x^n is (α, γ, k_n)-non-pathological for sufficiently large n, provided

$$\frac{n-(k+1)}{2(k+1)} e^{-2\lambda_1 k} \geq 2k+1+\frac{\alpha}{2}\log n, \tag{123}$$

where (119) follows from (116) and (118), (120) from (117), and (121) from (111) and the fact that

$$|\mathcal{S}_j^c| \geq \left\lfloor \frac{n}{k+1} \right\rfloor.$$

Note that the condition in (123) holds for all $\alpha > 0$ and sufficiently large n, provided that $k \leq k_n \leq \tau \log n$ where

$$\tau < \frac{1}{2\lambda_1}.$$

We complete the proof by combining the above two steps. In particular, since x^n satisfying (116) for all j, c_{-k}^{-1}, c_1^k, and $k \leq k_n$, implies that x^n is (α, γ, k_n)-non-pathological, it follows that, for $\gamma < \Delta(\mathbf{\Pi}, \mathbf{\Lambda})$ and $k_n \leq \tau \log n$,

$$\begin{aligned}
&P(X^n \notin \mathcal{N}_n) \\
&\leq P\bigg(\exists k \leq k_n, j, c_{-k}^{-1}, c_1^k, \text{ s. t. } \Big|\Big|E\Big[\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)|X^n\Big] - |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k} \mathbf{\Pi}^T \mathbf{P}^*\Big|\Big|_1 \geq \frac{|\mathcal{S}_j^c|}{\log n} p_{c_{-k}^{-1}c_1^k}\bigg) \\
&\leq \sum_{k=0}^{k_n}\sum_{j=0}^{k+1}\sum_{c_{-k}^{-1}, c_1^k \in \mathcal{A}^k} P\bigg(\Big|\Big|E\Big[\tilde{\mathbf{m}}_j(Z^n, c_{-k}^{-1}, c_1^k)|X^n\Big] - |\mathcal{S}_j^c| p_{c_{-k}^{-1}c_1^k} \mathbf{\Pi}^T \mathbf{P}^*\Big|\Big|_1 \geq \frac{|\mathcal{S}_j^c|}{\log n} p_{c_{-k}^{-1}c_1^k}\bigg) \\
&\leq k_n(k_n+1) M^{2k_n+1} 4 e^{-\frac{\left(\frac{n-2k_n-2}{k_n+1}\right)e^{-(2k_n+1)\lambda_1}}{6(\log n)^2}} \qquad (124) \\
&= o(1) \qquad (125)
\end{aligned}$$

when

$$\tau < \frac{1}{2\lambda_1}, \tag{126}$$

where (124) follows from (115), and (125) follows from the fact that, for $k_n \leq \tau \log n$, with τ satisfying (126), the factor $e^{(\cdot)}$ appearing in (124) is smaller than e^{-n^δ} for some sufficiently small $\delta > 0$ and sufficiently large n, and the other factors, again under the condition $k_n \leq \tau \log n$, are at most polynomially increasing. □

References

[1] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. on Info. Theory*, 51(1):5–28, 2005.

[2] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. 24, pp. 530–536, Sept. 1978.

[3] E. Ordentlich, K. Viswanathan, and M.J. Weinberger, "On concentration for denoiser-loss estimators," Proceedings of the 2009 IEEE International Symposium on Information Theory (ISIT'09), Seoul, Korea, June 2009.

[4] B. Ryabko, "Twice–universal coding," *Problems of Information Transmission*, vol. 20, pp. 173–177, July/September 1984.

[5] M. Feder and N. Merhav, "Hierarchical universal coding," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1354–1364, Sept. 1996.

[6] K. Viswanathan and E. Ordentlich, "Lower limits of discrete universal denoising," *IEEE Trans. Inform. Theory*, vol. 55, pp. 1374–1386, March 2009.

[7] E. Ordentlich, M. J. Weinberger, and T. Weissman, "Multi-directional context sets with applications to universal denoising and compression," in *Proc. of IEEE Symp. on Info. Theory*, pages 1270–1274, 2005.

[8] T. Moon and T. Weissman, "Discrete denoising with shifts," in *Proc. of 45th Annual Allerton Conf. on Communication, Control and Computation*, Monticello, Illinois, Sep 2007.

[9] T. Weissman, E. Ordentlich, M. J. Weinberger, A. Somekh-Baruch, and N. Merhav, "Universal filtering via prediction," *IEEE Trans. Inform. Theory*, vol. 53, pp. 1253–1264, April 2007.

[10] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, Wiley, 1971.

[11] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley, 1971.

[12] C. McDiarmid, "Concentration," *Probabilistic Methods for Algorithmic Discrete Mathematics*, Springer, 1998.