Denoiser-loss estimators and twice-universal denoising

Erik Ordentlich, Krishnamurthy Viswanathan, Marcelo J. Weinberger

HP Laboratories
HPL-2009-343

Keyword(s):
Universal denoising, concentration inequalities, universal data compression

Abstract:
We study the concentration of denoiser loss estimators, with application to the selection of denoiser parameters for a given observed sequence (in particular, the window size k of the DUDE algorithm [1]) via minimization of the estimated loss. We show that for a loss estimator proposed earlier [2], it is not possible to derive strong concentration results for certain pathological input sequences. By modifying the estimator slightly we obtain a loss estimator for which the DUDE's estimated loss strongly concentrates around the true loss provided $kM^{2k} = o(n)$, where M is the size of the alphabet and n the sequence length. We also show that for certain channels, it is possible to estimate the best k using a combination of the two loss estimators. Moreover, for non-pathological sequences and $k = o(n^{1/4})$, we derive concentration results for the original loss estimator and all channels.

In a second set of results, we extend the notion of twice universality from universal data compression theory to the sliding window denoising setting. Given a sequence length n and a denoiser, we define the k-dependent twice-universality penalty of the denoiser as the worst case excess denoising loss relative to sliding window denoisers with window length k above and *beyond* the worst case excess loss of DUDE with parameter k. Given an increasing sequence of window parameters k_n in the data sequence length n, we use loss estimators and results from the analysis mentioned above to construct a sequence of (twice) universal denoisers that achieves a much smaller twice universality penalty for $k < k_n$ than the sequence of DUDEs with parameter k_n.

Internal Posting Date: October 21, 2009 [Fulltext]
Approved for External Publication
Internal Posting Date: October 21, 2009 [Fulltext]

Presented at IEEE International Symposium on Information Theory, June-July 2009.

Denoiser-loss estimators and twice-universal denoising

Erik Ordentlich, Krishnamurthy Viswanathan, Marcelo J. Weinberger

***Abstract*—We study the concentration of denoiser loss estimators, with application to the selection of denoiser parameters for a given observed sequence (in particular, the window size k of the DUDE algorithm [1]) via minimization of the estimated loss. We show that for a loss estimator proposed earlier [2], it is not possible to derive strong concentration results for certain pathological input sequences. By modifying the estimator slightly we obtain a loss estimator for which the DUDE's estimated loss strongly concentrates around the true loss provided $kM^{2k} = o(n)$, where M is the size of the alphabet and n the sequence length. We also show that for certain channels, it is possible to estimate the best k using a combination of the two loss estimators. Moreover, for non-pathological sequences and $k = o(n^{\frac{1}{4}})$, we derive concentration results for the original loss estimator and all channels.**

In a second set of results, we extend the notion of twice-universality from universal data compression theory to the sliding window denoising setting. Given a sequence length n and a denoiser, we define the k-dependent twice-universality penalty of the denoiser as the worst case excess denoising loss relative to sliding window denoisers with window length k *above and beyond* the worst case excess loss of DUDE with parameter k. Given an increasing sequence of window parameters k_n in the data sequence length n, we use loss estimators and results from the analysis mentioned above to construct a sequence of (twice) universal denoisers that achieves a much smaller twice-universality penalty for $k < k_n$ than the sequence of DUDEs with parameter k_n.

I. Introduction

The problem of denoising is one of reproducing a signal based on noisy observations, with the quality of the reproduction being measured by a fidelity criterion. In one version of this problem, a clean sequence x^n is passed through a known discrete memoryless channel to obtain a noisy sequence z^n and the goal of the denoiser is to produce a reconstruction $\hat{x}^n$. The quality of the reconstruction is measured by a single-letter loss function $\Lambda(\cdot)$. This problem was studied in [1] where a universal denoising algorithm, DUDE, was derived. The DUDE 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. In [1], it was shown that the performance of the DUDE with parameter k was close to that of the best k-th order sliding window denoiser for the pair x^n, z^n.

For a given channel, loss function, and noisy sequence z^n, the fundamental problem of identifying the best choice of k for the DUDE is still open. Unlike in corresponding problems in data compression where the outcome of the choice of k can be observed, and therefore the best k selected, in the denoising problem, the clean sequence x^n is not observable and therefore the loss is not computable. In [1], two approaches to this problem were proposed. One is to select the value of k that minimizes, in the worst case over x^n, the expected excess loss incurred by the DUDE with parameter k, over the loss that would have been incurred for the DUDE with the optimal parameter for that (x^n, z^n) pair. This value is hard to compute and therefore a second, heuristic approach was proposed: to select the value of k that results in the most compressible reconstruction.

In [2], it was proposed to select the value of k that minimizes an estimate of the loss. If one could prove that, for all clean sequences, the loss estimate concentrates around the true loss with high probability, then this technique would indeed guarantee that the loss incurred for the selected k is close to that incurred for the best k, with high probability. To this end, this paper studies the concentration of loss estimators.

A specific loss estimator was proposed in [2] and shown to be unbiased for all denoisers, not just the DUDE. Further, it was shown in [3] that when this estimator is applied to any *fixed* k-th order sliding window denoiser, the estimated loss strongly concentrates around the true loss provided $k = o(n/\log n)$. This result, however, does not extend to the DUDE, for which the denoising function is not fixed but depends on z^n. While we seek concentration results that do apply to the DUDE, we notice that it suffices to target $k = \mathcal{O}(\log n)$. This is because, for every channel, there exists a constant γ, which can be arbitrarily large depending on the channel transition probabilities, such that if $k > \gamma \log n$, then with high probability all contexts appear at most once in z^n, and the DUDE reduces to a fixed, zero-th order denoiser whose loss, by [3], can be estimated well. Further, it is easy to see that it suffices to obtain concentration bounds of the following type: For all $\tau > 0$, the probability that the estimated loss differs from the true loss by more than τ decreases exponentially (or at least super-polynomially) fast in n.

For the estimator proposed in [2], we show in Section III that such concentration results may not be possible for all clean sequences. This negative result is based on the identification of pathological sequences for which the difference between estimated and true loss is at least a fixed constant with probability at least $O(n^{-\frac{1}{2}})$. We then show in Section IV that, with a slight modification, we obtain a loss estimator for which the DUDE's estimated loss concentrates around the true loss provided $kM^{2k} = o(n)$, where $M > 1$ is the size of the

All authors are with Hewlett-Packard Laboratories, Palo Alto, CA 94304, USA. (emails: {erik.ordentlich, krishnamurth.viswanathan, marcelo.weinberger}@hp.com)

Parts of this work were presented at the IEEE International Symposium on Information Theory, Seoul, Korea, June-July, 2009.

alphabet. This result is just a slight generalization of a result implicit in [3]. In Section V, we improve over this result by showing that for certain non-pathological clean sequences it is possible to derive concentration results for the original loss estimate and $k = o(n^{\frac{1}{4}})$. We also show that for many common channels and loss functions the fraction of non-pathological sequences approaches 1 for $k < \kappa \log n$, where κ depends on the channel and loss function. Finally, we point out that for certain channels, for all input sequences, it is possible to estimate the optimal value of k using a combination of the original and modified estimator.

By the above discussion, our ultimate goal is an estimator that concentrates for $k = \mathcal{O}(\log n)$, all x^n and all channels. While we fall short of that goal, for some channels, our concentration results apply to large enough k to permit us to estimate the optimal k for the DUDE. For other channels, we identify a key feature in the structure of the loss estimator proposed in [2] that prevents concentration results from applying to all clean sequences, and point to tools such as Kutin's inequality that provide partial concentration results and potential pathways to the ultimate goal.

The above summarizes our results through Section V. In Section VI, we introduce a definition of twice-universal denoising which seeks to capture the desirable property that the excess loss (over the best sliding window denoiser) of a universal denoiser be as small as possible simultaneously for all k in a growing range of sliding window parameters $k \leq k_n$, rather than targeting only a single sequence of parameter values k_n (one for each n). Some of the analysis of the preceding sections is leveraged to construct and analyze a new universal denoiser based on a deinterleaved version of the DUDE and the loss estimator of [2] that exhibits improved twice-universality behavior as compared to a DUDE with (growing) parameter k_n.

II. Notation and preliminaries

The notation used here is similar to the one in [1]. We first define the notation employed to refer to vectors, matrices and sequences. For any 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 and $||A||_\infty$ will denote the largest absolute value of any entry in the matrix or vector.

For any set $\mathcal{A}$, $\mathcal{A}^\infty$ denotes 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 $s_i \in \mathcal{A}^{k_i}$, 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 ith 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 whose input and output alphabet are both $\mathcal{A}$. Let the matrix $\mathbf{\Pi} = \{\mathbf{\Pi}(i,j)\}_{i,j \in \mathcal{A}}$, 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 jth column of $\mathbf{\Pi}$. We are interested in channels whose transition matrix $\mathbf{\Pi}$ is invertible. Let $M = |\mathcal{A}|$ denote the size of the alphabet.

Upon observing a noisy sequence $z^n \in \mathcal{A}^n$, the 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 ith 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 ith 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

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

The estimate of the loss incurred by a denoiser $\hat{X}$ proposed in [2] is given by

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

where we use $\hat{x}_i(z)$ to abbreviate $\hat{X}(z_1^{i-1} \cdot z \cdot z_{i+1}^n)[i]$. It was shown in [2] that this estimate is unbiased. For a denoiser $\hat{X}$ such as the DUDE, we are interested in upper-bounding the probability

$$\mathbf{Pr}\Big(\Big|L_{\hat{X}}(x^n, Z^n) - \hat{L}_{\hat{X}}(Z^n)\Big| \geq \tau\Big)$$

for all $\tau > 0$, uniformly in x^n, where the use of an upper case, *e.g.* Z, denotes that it is a random variable.

Concentration results have been derived for the class of sliding window denoisers in [3]. A *k-th order sliding window denoiser* $\hat{X}^n$ is a denoiser with the property that for all $a^n, b^n \in \mathcal{A}^n$, if $a_{i-k}^{i+k} = b_{j-k}^{j+k}$ then $\hat{X}^n(a^n)[i] = \hat{X}^n(b^n)[j]$. Thus, the denoiser is defined by a mapping, $f : \mathcal{A}^{2k+1} \to \mathcal{A}$ so that for all $z^n \in \mathcal{A}^n$,

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

A *k-th order sliding-window-based 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 $\hat{X}^n(z^n)[i] = \hat{X}^n(z^n)[j]$. This is less restrictive than a sliding window denoiser where, for a given context, the reconstruction cannot depend on the sequence z^n. For a given k-th order sliding-window-based denoiser $\hat{X}^n$, with each z^n, we can associate a mapping $f_{\hat{X}^n, z^n} : \mathcal{A}^k \times \mathcal{A} \times \mathcal{A}^k \to \mathcal{A}$ so that for all $z^n \in \mathcal{A}^n$

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

This mapping is made unique by insisting that for all contexts c^{2k+1} that do not appear in z^n, $f_{\hat{X}^n, z^n}(c_{-k}^{-1}, c_0, c_1^k) = c_0$.

The k-th order DUDE is a k-th order sliding-window-based denoiser but not a k-th order sliding window denoiser.

Let $\mathcal{S}_k$ denote the class of kth-order sliding window denoisers. It is shown in [3] that for all $\hat{X}^n \in \mathcal{S}_k$, $\mathbf{\Pi}$, $\mathbf{\Lambda}$, and x^n,

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

This 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 the loss estimator in (1) is well-behaved for sliding window denoisers, we will show that this is not always the case for the DUDE.

III. Pathological sequences for the DUDE

We show that in the case of the DUDE, for some clean sequences x^n, and some constant τ, the probability $\mathbf{Pr}\Big(\Big|L_{\hat{X}}(x^n, Z^n) - \hat{L}_{\hat{X}}(Z^n)\Big| \geq \tau\Big)$ is lower-bounded by a function that is $\Omega(n^{-\frac{1}{2}})$. Let $\hat{X}^{n,k}_{\text{DUDE}}$ denote the DUDE with parameter k. Let $\mathcal{A} = \{0,1\}$. Let $\mathbf{\Pi}$ correspond to the binary symmetric channel 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 zero order.

Theorem 1: There exists a clean sequence x^n and constants K and τ_0 such that

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

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 (with a little abuse of notation) the DUDE of zero order is given by

$$\hat{X}^{n,0}_{\text{DUDE}}(z^n) = \hat{X}^{n,0}_{\text{DUDE}}(z_1)\cdots\hat{X}^{n,0}_{\text{DUDE}}(z_i)\cdots\hat{X}^{n,0}_{\text{DUDE}}(z_n)$$

where for $z \in \mathcal{A}$

$$\hat{X}^{n,0}_{\text{DUDE}}(z) = \begin{cases} 0, & T(z^n) \leq n2\delta(1-\delta) \\ z, & n2\delta(1-\delta) < T(z^n) \leq n(1-2\delta(1-\delta)) \\ 1, & T(z^n) > n(1-2\delta(1-\delta)). \end{cases}$$

Let $T(z^n) = \lfloor n2\delta(1-\delta)\rfloor$. Then we have

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

Observe that for the same z^n, if $z_i = 0$, then $\hat{X}^{n,0}_{\text{DUDE}}(z_1^{i-1}\cdot z\cdot z_{i+1}^n)[i] = z, \quad z = 0,1$. Therefore, for all $x \in \{0,1\}$,

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

Hence,

$$\sum_{x\in\{0,1\}}\mathbf{\Pi}^{-T}(x,z_i)\sum_{z\in\{0,1\}}\Lambda(x,\hat{X}^{n,0}_{\text{DUDE}}(z_1^{i-1}\cdot z\cdot z_{i+1}^n)[i])\mathbf{\Pi}(x,z) = \delta\sum_{x\in\{0,1\}}\mathbf{\Pi}^{-T}(x,0) = \delta. \quad (3)$$

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

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

Hence,

$$\sum_{x\in\{0,1\}}\mathbf{\Pi}^{-T}(x,z_i)\sum_{z\in\{0,1\}}\Lambda(x,\hat{X}^{n,0}_{\text{DUDE}}(z_1^{i-1}\cdot z\cdot z_{i+1}^n)[i]))\mathbf{\Pi}(x,z) = \sum_{x\in\{0,1\}}\mathbf{\Pi}^{-T}(x,1)\Lambda(x,0) = \mathbf{\Pi}^{-T}(1,1) = \frac{1-\delta}{1-2\delta}. \quad (4)$$

Combining (1), (3), and (4), we obtain that if $T(z^n) = \lfloor n2\delta(1-\delta)\rfloor$, and $\delta < \frac{1}{2}$, then

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

for all sufficiently large n. Since for all $\hat{X}$ and z^n, $L_{\hat{X}}(x^n, z^n) \leq 1$, we obtain that if $T(z^n) = \lfloor n2\delta(1-\delta)\rfloor$

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

The probability of this pathological type depends on the type $T(x^n)$ of the clean sequence. If $T(x^n) \approx n\delta$, then it can be shown that for some constant K

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

which proves the theorem. ∎

IV. Modified loss-estimator for the DUDE

We derive a concentration result for a class of denoisers that includes the DUDE. This result is a slight generalization of a result implicit in [3]. In [3], the DUDE's loss was estimated as the minimum of the loss estimates as computed using (1), where the minimization is over all sliding window denoisers with the same window size. This estimate is identical to the loss estimate for the DUDE as given by the modified estimator in this section. For a sliding-window-based denoiser $\hat{X}^n$, let

$$\tilde{L}_{\hat{X}^n}(z^n) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\sum_{x\in\mathcal{A}}\mathbf{\Pi}^{-T}(x,z_i) \times \sum_{z\in\mathcal{A}}\Lambda(x,\tilde{x}_i(z))\mathbf{\Pi}(x,z) \quad (5)$$

where $\tilde{x}_i(z) = f_{\hat{X}^n,z^n}\big(z_{i-k}^{i-1}\cdot z\cdot z_{i+1}^{i+k}\big)$. Note that this estimator is no longer unbiased. Let $\tilde{\mathcal{S}}_k$ denote the set of all kth order sliding-window-based denoisers.

Theorem 2: For all $\hat{X}^n \in \tilde{\mathcal{S}}_k$, $\tau > 0$, and all x^n

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

Proof: Let $\hat{X}^n$ be defined by the collection of functions $\{f_{\hat{X}^n,z^n}\}$, $f_{\hat{X}^n,z^n} : \mathcal{A}^{2k+1} \to \mathcal{A}$. Then, letting $L_{f_{\hat{X}^n,z^n}}$

denote the loss incurred by the k-th order sliding window denoiser defined by $f_{\hat{X}^n,z^n}$, for all $z^n \in \mathcal{A}^n$ we have $L_{\hat{X}^n}(x^n,z^n) = L_{f_{\hat{X}^n,z^n}}(x^n,z^n)$. Similarly, letting $\hat{L}_{f_{\hat{X}^n,z^n}}$ denote the loss estimate, given by (1), for the k-th order sliding window denoiser defined by $f_{\hat{X}^n,z^n}$, it follows from (5) that, for all $z^n \in \mathcal{A}^n$, $\tilde{L}_{\hat{X}^n}(z^n) = \hat{L}_{f_{\hat{X}^n,z^n}}(z^n)$. Therefore, for all clean sequences $x^n \in \mathcal{A}^n$, we have

$$\begin{aligned}&\mathbf{Pr}\Big(\Big|L_{\hat{X}^n}(x^n,Z^n) - \tilde{L}_{\hat{X}^n}(Z^n)\Big| \geq \tau\Big)\\ &\quad = \mathbf{Pr}\Big(\Big|L_{f_{\hat{X}^n,Z^n}}(x^n,Z^n) - \hat{L}_{f_{\hat{X}^n,Z^n}}(Z^n)\Big| \geq \tau\Big). \quad (6)\end{aligned}$$

With a slight abuse of notation, let $\mathcal{S}_k$ also denote the set of all functions $\{g : \mathcal{A}^{2k+1} \to \mathcal{A}\}$, and let $\mathcal{E}_g$ denote the event that $f_{\hat{X}^n,Z^n}$ is g. Then

$$\begin{aligned}&\mathbf{Pr}\Big(\Big|L_{f_{\hat{X}^n,Z^n}}(x^n,Z^n) - \hat{L}_{f_{\hat{X}^n,Z^n}}(Z^n)\Big| \geq \tau\Big)\\ &\quad = \sum_{g\in\mathcal{S}_k} \mathbf{Pr}\Big(\mathcal{E}_g \cap \Big|L_g(x^n,Z^n) - \hat{L}_g(Z^n)\Big| \geq \tau\Big)\\ &\quad \leq \sum_{g\in\mathcal{S}_k} \mathbf{Pr}\Big(\Big|L_g(x^n,Z^n) - \hat{L}_g(Z^n)\Big| \geq \tau\Big)\\ &\quad \leq M^{M^{2k+1}}(k+1)e^{\frac{-2(n-2k)\tau^2}{(k+1)||\mathbf{\Lambda}||_\infty^2(1+M||\mathbf{\Pi}^{-1}||_\infty)^2}}\end{aligned}$$

where the last inequality follows from (2) and the fact that $|\mathcal{S}_k| \leq M^{M^{2k+1}}$. Substituting this in (6) completes the proof. ∎

Theorem 2 implies that for all $\tau > 0$, and all sliding-window-based denoisers the probability that the estimation error exceeds τ vanishes with n as long as $kM^{2k+1} = o(n)$, which includes $k = \gamma_1 \log n$ where $\gamma_1 < (2\log M)^{-1}$. We note that Theorem 2 also holds for an estimator that is based on the posterior probability as computed by the DUDE.

V. Non-pathological sequences

In this section, we derive a concentration bound for all "non-pathological" clean sequences. The bound vanishes with n even when k is only required to be $o(n^{\frac{1}{4}})$. We employ a martingale inequality derived by Kutin [4], a simplified version of which is stated below. Let f be a function of random variables $Z^n \in \mathcal{A}^n$. We say that f is *strongly difference bounded* by (τ_1,τ_2,δ) if for any $z^n,\tilde{z}^n \in \mathcal{A}^n$ differing only in one co-ordinate, we have $|f(z^n)-f(\tilde{z}^n)| \leq \tau_1$, and moreover, $|f(z^n)-f(\tilde{z}^n)| \leq \tau_2$ provided z^n is not in a "bad" subset B of $\mathcal{A}^n$ such that $\mathbf{Pr}(Z^n \in B) = \delta$. For any f that is strongly difference bounded by (τ_1,τ_2,δ), Kutin's inequality states that for any $\tau > 0$,

$$\mathbf{Pr}(|f(Z^n) - E(f(Z^n))| \geq \tau) \leq 2e^{-\frac{\tau^2}{8n\tau_2^2}} + \frac{2n\tau_1\delta}{\tau_2}.$$

Observe that τ_2 must tend to zero with n for this bound to be non-trivial. We will show that for all non-pathological sequences x^n, $f(z^n)$ given by $L_{\hat{X}^{n,k}_{\text{DUDE}}}(x^n,z^n) - \tilde{L}_{\hat{X}^{n,k}_{\text{DUDE}}}(z^n)$ is strongly difference bounded and therefore a concentration result such as the one above holds.

First, we define the "bad" set B, for which we require the following notation. For $2k < n$, $a^n \in \mathcal{A}^n$, $c_{-k}^{-1}, c_1^k \in \mathcal{A}^k$, let $\mathbf{m}(a^n, c_{-k}^{-1}, c_1^k)$ denote the M-dimensional column vector whose c_0-th component, $c_0 \in \mathcal{A}$, is $\mathbf{m}(a^n, c_{-k}^{-1}, c_1^k)[c_0] = |\{i : k+1 \leq i \leq n-k, a_{i-k}^{i+k} = c_{-k}^k\}|$, the number of occurrences of the string c_{-k}^k in a^n. These counts are the basis for DUDE's denoising decision. Let $\mathcal{M}$ denote the set of all M-dimensional vectors with non-negative components that sum to at most n. 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 DUDE involving reconstruction symbols a,b and noisy symbol c. The "bad" set B is defined to be the set of all $z^n \in \mathcal{A}^n$ such that at least one of the counts $\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k)$ could fall within $\mathcal{M}^*$ if at most one co-ordinate of z^n is modified. Formally, given $\alpha > 0$,

$$\begin{aligned}B \stackrel{\text{def}}{=} \{&z^n : \exists c_{-k}^k, a, b, \text{ s.t. } \mathbf{m}(z^n, c_{-k}^{-1}, c_1^k)[c_0] \geq n^\alpha \text{ and}\\ &\exists\, \mathbf{m}^* \in \mathcal{M}^*(a,b,c_0), ||\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k) - \mathbf{m}^*||_1 \leq 2k+1\}.\end{aligned}$$

We term a clean sequence "non-pathological" if the expected values of the counts $\mathbf{m}$, if significant, are bounded away from $\mathcal{M}^*(a,b,c)$ for all a,b,c. Formally, x^n is said to be "non-pathological" with parameter γ, if for all c_{-k}^k, if $E\big[\mathbf{m}(Z^n, c_{-k}^{-1}, c_1^k)[c_0]\big] \geq \frac{n^\alpha}{2(1+\gamma)}$, then for all a,b, and $\mathbf{m}^* \in \mathcal{M}^*(a,b,c_0)$,

$$\begin{aligned}||E\big[\mathbf{m}(Z^n, c_{-k}^{-1}, c_1^k)\big] - \mathbf{m}^*||_1 >& \gamma||E\big[\mathbf{m}(Z^n, c_{-k}^{-1}, c_1^k)\big]||_1\\ &+ (2k+1) + \frac{Mn^\alpha}{2}. \quad (7)\end{aligned}$$

We will show that for non-pathological sequences, the probability that the noisy sequence Z^n is in B vanishes.

Lemma 3: For all "non-pathological" x^n with $0 < \gamma < 1$

$$\mathbf{Pr}(Z^n \in B) \leq \beta_1 n^{1-\alpha} e^{-\beta_2 \frac{n^\alpha}{2k+1}}$$

where β_1, β_2 are positive functions of γ.

Proof outline: We abbreviate $\mathbf{m}(Z^n, c_{-k}^{-1}, c_1^k)$ by $\mathbf{m}$. Let $Z \in B$ and let c_{-k}^k be such that $\mathbf{m}[c_0] \geq n^\alpha$, and let $a,b \in \mathcal{A}$ and $m^* \in \mathcal{M}(a,b,c_0)$ be such that $||\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k) - \mathbf{m}^*||_1 \leq 2k+1$. Applying the triangle inequality this implies that

$$||\mathbf{m}^* - E[\mathbf{m}]||_1 - ||E[\mathbf{m}] - \mathbf{m}||_1 \leq 2k+1.$$

If $E[\mathbf{m}[c_0]] \geq \frac{n^\alpha}{2(1+\gamma)}$ then since x^n is "non-pathological", by (7)

$$||E[\mathbf{m}] - \mathbf{m}||_1 > \gamma||E[\mathbf{m}]||_1 + \frac{Mn^\alpha}{2}.$$

Therefore there exists $c \in \mathcal{A}$ such that

$$|\mathbf{m}[c] - E[\mathbf{m}[c]]| > \gamma E[\mathbf{m}[c]] + \frac{n^\alpha}{2}. \quad (8)$$

If for all $c \in \mathcal{A}$, $E[\mathbf{m}[c_0]] < \frac{n^\alpha}{2(1+\gamma)}$, using the fact that $\mathbf{m}[c_0] \geq n^\alpha$ it is easy to see that (8) holds for $c = c_0$.

Let $\mathbf{m}_j(z^n, c_{-k}^{-1}, c_1^k)$, $0 \leq j \leq 2k$, abbreviated by $\mathbf{m}_j$, denote the M-dimensional column vector whose c_0-th component, $c_0 \in \mathcal{A}$, is $\mathbf{m}_j(z^n, c_{-k}^{-1}, c_1^k)[c_0] = |\{i : k+1 \leq i \leq n-k, i \mod (2k+1) = j, z_{i-k}^{i+k} = c_{-k}^k\}|$. From (8) it is easy to see that there exists $0 \leq j \leq 2k$ and $c_0 \in \mathcal{A}$ such that

$$|\mathbf{m}_j[c_0] - E[\mathbf{m}_j[c_0]]| > \max\{\gamma E[\mathbf{m}_j[c_0]], (4k+2)^{-1}n^\alpha\}.$$

Therefore,

$$\mathbf{Pr}(Z^n \in B) \leq \mathbf{Pr}\left(\exists c_{-k}^{-k}, j \text{ s.t. } |\mathbf{m}_j[c_0] - E[\mathbf{m}_j[c_0]]| \geq \max\{\gamma E[\mathbf{m}_j[c_0]], (4k+2)^{-1}n^\alpha\}\right).$$

Observing that each $\mathbf{m}_j[c_0]$ is the sum of independent $0-1$ random variables, and applying the union bound and Theorem 2.3(b) in [5], this can be reduced to

$$\mathbf{Pr}(Z^n \in B) \leq \beta_1 n^{1-\alpha} e^{-\beta_2 \frac{n^\alpha}{2k+1}}$$

where β_1 and β_2 are positive functions of γ. ∎

We will now show that both the loss of the DUDE, $L_{\hat{X}_{\text{DUDE}}^{n,k}}$, and the loss estimator, $\hat{L}_{\hat{X}_{\text{DUDE}}^{n,k}}$, are strongly difference bounded. Let z^n and $\tilde{z}^n$ be two sequences that differ only in one co-ordinate. The following lemma bounds the change in loss when a single co-ordinate of z^n is modified.

Lemma 4: For all x^n, and all z^n

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

and for all $z^n \notin B$

$$|L_{\hat{X}_{\text{DUDE}}^{n,k}}(x^n, z^n) - L_{\hat{X}_{\text{DUDE}}^{n,k}}(x^n, \tilde{z}^n)| \leq \frac{(4k+2)||\mathbf{\Lambda}||_\infty(n^\alpha M+1)}{n-2k}.$$

Proof: For all, x^n, z^n, and any denoiser $\hat{X}$, $0 \leq L_{\hat{X}}(x^n, z^n) \leq ||\mathbf{\Lambda}||_\infty$, hence the first result. Observe that changing one co-ordinate affects the reconstruction of a symbol in three ways: it might change the context of that symbol, the symbol itself or it might change the number of occurrences of the context of the symbol sufficiently to affect the DUDE's decision for that context. Since changing one co-ordinate changes the context of at most $2k+1$ symbols, the change in loss due to change in context or the symbol itself is at most $(2k+1)||\mathbf{\Lambda}||_\infty$.

We now bound the change in loss due to the third event. Modifying one symbol modifies the counts $\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k)[c_0]$ of at most $4k+2$ contexts c_{-k}^k. Let c_{-k}^k be one such affected context. **Case 1:** If for all $c \in \mathcal{A}$, $\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k)[c] < n^\alpha$, then the number of indices i such that $z_{i-k}^{i-1} = \tilde{z}_{i-k}^{i-1} = c_{-k}^{-1}$, $z_{i+1}^{i+k} = \tilde{z}_{i+1}^{i+k} = c_1^k$, and $\hat{X}_{\text{DUDE}}^{n,k}(z^n)[i] \neq \hat{X}_{\text{DUDE}}^{n,k}(\tilde{z}^n)[i]$ is at most Mn^α. **Case 2:** Suppose there exists $c \in \mathcal{A}$ such that $\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k)[c] \geq n^\alpha$. Note that for all c_{-k}^{-1}, c_1^k,

$$||\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k) - \mathbf{m}(\tilde{z}^n, c_{-k}^{-1}, c_1^k)||_1 \leq 2k+1.$$

Since $z^n \notin B$, for all a, b, c both $(\lambda_a - \lambda_b)^T\left((\mathbf{\Pi}^{-T}\mathbf{m}(z^n, c_{-k}^{-1}, c_1^k)) \odot \pi_c\right)$ and $(\lambda_a - \lambda_b)^T\left((\mathbf{\Pi}^{-T}\mathbf{m}(\tilde{z}^n, c_{-k}^{-1}, c_1^k)) \odot \pi_c\right)$ have the same sign. Since $\hat{X}_{\text{DUDE}}^{n,k}$ is given by

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

if i is an index such that $z_{i-k}^{i-1} = \tilde{z}_{i-k}^{i-1} = c_{-k}^{-1}$ and $z_{i+1}^{i+k} = \tilde{z}_{i+1}^{i+k} = c_1^k$, then $\hat{X}_{\text{DUDE}}^{n,k}(z^n)[i] = \hat{X}_{\text{DUDE}}^{n,k}(\tilde{z}^n)[i]$. In words, the reconstruction and therefore the loss is unaffected.

Combining both cases and the fact that the number of affected contexts is at most $4k+2$, we obtain the result. ∎

Following arguments similar to the proof of Lemma 4 we obtain the following lemma.

Lemma 5: For all x^n, and all z^n

$$|\hat{L}_{\hat{X}_{\text{DUDE}}^{n,k}}(z^n) - \hat{L}_{\hat{X}_{\text{DUDE}}^{n,k}}(\tilde{z}^n)| \leq 2M||\mathbf{\Pi}^{-1}||_\infty||\mathbf{\Lambda}||_\infty,$$

and for all $z^n \notin B$

$$|\hat{L}_{\hat{X}_{\text{DUDE}}^{n,k}}(x^n, z^n) - \hat{L}_{\hat{X}_{\text{DUDE}}^{n,k}}(x^n, \tilde{z}^n)| \leq \frac{(4k+2)M||\mathbf{\Pi}^{-1}||_\infty||\mathbf{\Lambda}||_\infty(n^\alpha M+1)}{n-2k}.$$

Combining Lemmas 3, 4 and 5, and Kutin's inequality we obtain the following result.

Theorem 6: For all "non-pathological" x^n with $0 < \gamma < 1$, and $f(z^n) = L_{\hat{X}_{\text{DUDE}}^{n,k}}(x^n, z^n) - \hat{L}_{\hat{X}_{\text{DUDE}}^{n,k}}(z^n)$, we have

$$\mathbf{Pr}(|f(Z^n)| \geq \tau) \leq 2e^{-\frac{(n-2k)^{1-2\alpha}\tau^2}{8\left((4k+2)||\mathbf{\Lambda}||_\infty M\left(1+M||\mathbf{\Pi}^{-1}||_\infty\right)\right)^2}} + \frac{2\beta_1 n^{3-2\alpha}e^{-\beta_2\frac{n^\alpha}{2k+1}}}{(4k+2)M(1+M||\mathbf{\Pi}^{-1}||_\infty)}$$

where β_1, β_2 are positive functions of γ.
Observe that for all $\tau > 0$, the probability of estimation error $\geq \tau$, goes to zero as long as $k = o(n^\alpha)$ for $0 < \alpha < \frac{1}{4}$.

Further, we show that for many channels and loss functions, the fraction of sequences that are "non-pathological" tends to one as long as $k < \kappa \log n$ where κ depends on the channel. Let $\mathbf{P}^*$ denote the M-dimensional column vector whose entries are all $1/M$ and let $\mathcal{N}(\gamma)$ denote the set of non-pathological clean sequences with parameter $\gamma > 0$.

Theorem 7: If $\mathbf{\Pi}, \mathbf{\Lambda}$ are such that for all $a, b, c \in \mathcal{A}$, $\mathbf{\Pi}^T\mathbf{P}^* \notin \mathcal{M}^*(a, b, c)$, then there exists $\kappa > 0$ such that for $k < \kappa \log n$, $|\mathcal{N}(\gamma)| = M^n(1 - o(1))$.
Many channels and loss functions, *e.g.* BSC with crossover probability $0 < \delta < \frac{1}{2}$ and the Hamming loss function, satisfy the requirement of Theorem 7.

Theorem 2 implies that for the loss estimator $\tilde{L}_{\hat{X}_{\text{DUDE}}^{n,k}}$, if $k = \gamma_1 \log n$ where $\gamma_1 < (2\log M)^{-1}$, the estimated loss concentrates around the true loss. Extending k beyond this value is hard due to the difficulty in bounding the bias of $\tilde{L}_{\hat{X}_{\text{DUDE}}^{n,k}}$. However, one can redefine the notion of the "bad" set introduced in this section, to be the set of sequences where the number of occurrences of some context exceeds n^α, $\alpha < \frac{1}{2}$. In that case, if $k = \gamma_2 \log n$ where $\gamma_2 > \frac{1-\alpha}{-2\log||\mathbf{\Pi}||_\infty}$, then for the loss estimator $\hat{L}_{\hat{X}_{\text{DUDE}}^{n,k}}$ the estimated loss concentrates around the true loss. Thus, for channels where $||\mathbf{\Pi}||_\infty < M^{-\frac{1}{2}}$ (*e.g.*, BSC with crossover probability between $1 - 2^{-\frac{1}{2}}$ and $2^{-\frac{1}{2}}$), one can use a combination of $\tilde{L}_{\hat{X}_{\text{DUDE}}^{n,k}}$ and $\hat{L}_{\hat{X}_{\text{DUDE}}^{n,k}}$ to select the optimal k. For other channels, there remains a gap between the two critical values of k pointed out here and in this regime we can prove concentration only if the noiseless sequence is "non-pathological".

VI. Twice-universal denoising

It was shown in [1] that for all clean sequences x^n, the DUDE with parameter k satisfies, for some constant C independent of k and n,

$$E\left[L_{\hat{X}_{\text{DUDE}}^{n,k}}(x^n, Z^n) - D_k(x^n, Z^n)\right] \leq C\sqrt{\frac{M^{2k}k}{n}} \quad (9)$$

where $D_k(x^n, Z^n)$ denotes the loss of the best sliding denoiser for the clean and noisy sequences x^n and Z^n and the expectation is with respect to the noise distribution. Consider any increasing sequence k_n for which the bound (9) with $k = k_n$ converges to zero. The sequence of DUDEs with parameter k_n thus satisfies

$$E\left[L_{\hat{X}^{n,k_n}_{\text{DUDE}}}(x^n, Z^n) - D_k(x^n, Z^n)\right] \le C\sqrt{\frac{M^{2k_n}k_n}{n}} = o(1), \tag{10}$$

which holds simultaneously for all $k \le k_n$, since $D_k(x^n, Z^n)$ is decreasing in k for any fixed x^n and Z^n. Given such a sequence k_n, we shall say that a sequence of denoisers $\hat{X}^n$ is "twice-universal" with penalty $\epsilon(k, n)$ if its loss can be proved to satisfy

$$E\left[L_{\hat{X}^n}(x^n, Z^n) - D_k(x^n, Z^n)\right] \le C'\sqrt{\frac{M^{2k}k}{n}} + \epsilon(k, n) \tag{11}$$

for all sufficiently large n and all x^n and all $k \le k_n$ simultaneously and some constant C' independent of these parameters and x^n. Thus, the DUDE with parameter k_n is provably twice universal with penalty

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

where the latter approximation holds for $k << k_n$.

Given k_n as above, the main result of this section is a denoiser dubbed the TU-DUDE and denoted as $\hat{X}^n_{\text{TU}}$ that is provably twice universal with a penalty $\epsilon_{\text{TU}}(k, n) = \tilde{C}k_n(k_n/n)^{1/4}$, a big improvement over $\epsilon_{\text{D}}(n, k)$ when the bound (10) decreases substantially slower than $O(\log n(\log n/n)^{1/4})$.

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

$$\begin{aligned}\tilde{\mathbf{m}}_j(z^n, c_{-k}^{-1}, c_1^k)[c_0] = |\{&i : k+1 \le i \le n-k, \\ &i = j \,(\text{mod}\, k{+}1), z_{i-k}^{i+k} = c_{-k}^k\}|\end{aligned}$$

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

$$\begin{aligned}&\hat{X}^{n,k}_{\text{D-DUDE}}(z^n)[i] = \\ &\quad \arg\min_{\hat{x}\in\mathcal{X}} \lambda_{\hat{x}}^T\left((\mathbf{\Pi}^{-T}\tilde{\mathbf{m}}_{j(i)}(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})) \odot \pi_{z_i}\right),\end{aligned} \tag{12}$$

where $j(i) \in \{0, \ldots, k\}$ and satisfies $i = j(i) \,(\text{mod}\, k{+}1)$. Thus, the D-DUDE denoises the i-th symbol using only context occurrences for which the center-symbol index coincides with i modulo $k{+}1$. It can be shown that the D-DUDE satisfies all of the performance guarantees proved in [1] for the DUDE, including (9) above. In fact, the proofs in [1] actually involve a deinterleaving step.

The TU-DUDE, in turn, evaluates the estimated loss (1) of the D-DUDE for all parameter values $k \le k_n$ and denoises using the minimizing value. Formally, given k_n, the TU-DUDE is defined as

$$\hat{X}^n_{\text{TU}}(z^n)[i] = \hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}(z^n)[i] \tag{13}$$

where

$$\hat{k}_n^* = \arg\min_{k\le k_n} \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(z^n). \tag{14}$$

Theorem 8: For the binary case $M = 2$ and Hamming loss[1] the TU-DUDE, as defined in (13), is twice-universal with penalty $\epsilon_{\text{TU}}(k, n) = \tilde{C}k_n(k_n/n)^{1/4}$ for a constant $\tilde{C}$.

Proof: In the first part of the proof, a bound is obtained on the expected error in estimating $L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, z^n)$, the actual loss of $\hat{X}^{n,k}_{\text{D-DUDE}}$, using the estimate $\hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(z^n)$. This bound is then used, along with the above noted performance bounds on $\hat{X}^{n,k}_{\text{D-DUDE}}$, to establish the stated twice-universality properties of $\hat{X}^n_{\text{TU}}$.

For $j = 0, \ldots, k$, let

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

where

$$\begin{aligned}&\hat{\Lambda}_i(z^n) = \\ &\sum_{x\in\mathcal{A}} \mathbf{\Pi}^{-T}(x, z_i) \sum_{z\in\mathcal{A}} \Lambda(x, \hat{X}^{n,k}_{\text{D-DUDE}}(z_1^{i-1}, z, z_{i+1}^n))\mathbf{\Pi}(x, z).\end{aligned}$$

Note the identity

$$L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, z^n) - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(z^n) = \frac{1}{n}\sum_{j=0}^{k} \Delta_j(x^n, z^n). \tag{15}$$

From the above, it follows that

$$\begin{aligned}E(|L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(Z^n)|) \le \\ \frac{1}{n}\sum_{j=0}^{k} E(|\Delta_j(x^n, Z^n)|).\end{aligned} \tag{16}$$

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

$$\begin{aligned}E(|\Delta_j(x^n, Z^n)|) &= E(E(|\Delta_j(x^n, Z^n)|\big|Z^{\mathcal{S}_j})) \\ &\le \max_{z^{\mathcal{S}_j}} E(|\Delta_j(x^n, Z^n)|\big|Z^{\mathcal{S}_j} = z^{\mathcal{S}_j})\end{aligned} \tag{17}$$

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

$$\begin{aligned}&\tilde{\Delta}_{j,c_{-k}^{-1},c_1^k}(x^n, z^n) = \\ &\qquad \sum_{i\in\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})} \Lambda(x_i, \hat{X}^{n,k}_{\text{D-DUDE}}(z^n)[i]) - \hat{\Lambda}_i(z^n)\end{aligned} \tag{18}$$

so that

$$\Delta_j(x^n, z^n) = \sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} \tilde{\Delta}_{j,c_{-k}^{-1},c_1^k}(x^n, z^n). \tag{19}$$

It follows from the definitions of $\hat{X}^{n,k}_{\text{D-DUDE}}$ and the corresponding $\hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}$ that the random variables $\tilde{\Delta}_{j,c_{-k}^{-1},c_1^k}(x^n, Z^n)$ for fixed

[1] Some form, perhaps the same form, should hold for $M > 2$ and general distortion.

j and $(c_{-k}^{-1}, c_1^k) \in \mathcal{A}^{2k}$ are zero mean and conditionally independent given $Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}$. 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 $\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})$. An extension of the analysis in Section III shows that the conditional variance

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

satisfies

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

for some constant b_1. The conditional independence of the $\tilde{\Delta}_{j,c_{-k}^{-1},c_1^k}(x^n, Z^n)$ then implies (from (19)) that the conditional variance of $\Delta_j(x^n, Z^n)$ denoted by

$$\sigma_j^2 \stackrel{\triangle}{=} E(\Delta_j^2(x^n, Z^n)\big|Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}) \tag{20}$$

satisfies

$$\begin{aligned}\sigma_j^2 &= \sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} \sigma^2_{j,c_{-k}^{-1},c_1^k} \\ &\le b_1 \sum_{(c_{-k}^{-1},c_1^k)\in\mathcal{A}^{2k}} |\tilde{\mathcal{S}}_{j,c_{-k}^{-1},c_1^k}(z^{\mathcal{S}_j})|^{3/2} \end{aligned} \tag{21}$$

Noting that

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

it follows from the Schur convexity of the function $\sum_i x_i^{3/2}$ that, subject to this constraint, (21) can be no bigger than $((n/(k+1)) + 1)^{3/2}$, so that

$$\sigma_j^2 \le b_2 \left(\frac{n}{k}\right)^{3/2} \tag{22}$$

for some constant b_2 independent of k, n and j.

This, together with Jensen's inequality, implies

$$\begin{aligned} E(|\Delta_j(x^n, Z^n)|\big|Z^{\mathcal{S}_j} = z^{\mathcal{S}_j}) &\le (\sigma_j^2)^{1/2} && (23)\\ &\le (b_2)^{1/2} \left(\frac{n}{k}\right)^{3/4} && (24)\\ && (25)\end{aligned}$$

and this last step, together with (16) and (17), implies

$$\begin{aligned} E(|L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(Z^n)|) &\le (b_2)^{1/2} \frac{k+1}{n} \left(\frac{n}{k}\right)^{3/4} \\ &= b_3 \left(\frac{k}{n}\right)^{1/4} \end{aligned} \tag{26}$$

for some constant b_3.

The final step is to relate this bound on the expected error of the estimated loss to the performance of $\hat{X}^n_{\text{TU}}$. To this end, note that

$$\begin{aligned} &E(L_{\hat{X}^n_{\text{TU}}}(x^n, Z^n) - D_k(x^n, Z^n)) \\ &\quad = E(L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n) - D_k(x^n, Z^n)) \\ &\qquad + E(L_{\hat{X}^n_{\text{TU}}}(x^n, Z^n) - L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n)) \\ &\quad \le C\sqrt{\frac{M^{2k}k}{n}} + E(L_{\hat{X}^n_{\text{TU}}}(x^n, Z^n) - L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n)) \end{aligned} \tag{27}$$

where (27) follows from the fact noted above that the DUDE's performance bound (9) also holds for the D-DUDE.

In analogy to $\hat{k}_n^*$ defined in (14), let

$$k_n^* = \arg\min_{k\le k_n} L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, z^n), \tag{28}$$

be the actual loss minimizing context parameter for the D-DUDE. It is thus 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 the following

$$\begin{aligned} &E(L_{\hat{X}^n_{\text{TU}}}(x^n, Z^n) - L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n)) \\ &\quad = E(L_{\hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}}(x^n, Z^n) - L_{\hat{X}^{n,k_n^*}_{\text{D-DUDE}}}(x^n, Z^n)) + \\ &\qquad E(L_{\hat{X}^{n,k_n^*}_{\text{D-DUDE}}}(x^n, Z^n) - L_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n)) && (29)\\ &\quad \le E(L_{\hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}}(x^n, Z^n) - L_{\hat{X}^{n,k_n^*}_{\text{D-DUDE}}}(x^n, Z^n)) && (30)\\ &\quad = E(L_{\hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}}(x^n, Z^n)) + \\ &\qquad E(\hat{L}_{\hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}}(x^n, Z^n) - L_{\hat{X}^{n,k_n^*}_{\text{D-DUDE}}}(x^n, Z^n)) && (31)\\ &\quad \le E(L_{\hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}}(x^n, Z^n) - \hat{L}_{\hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}}(x^n, Z^n)) + \\ &\qquad E(\hat{L}_{\hat{X}^{n,\hat{k}_n^*}_{\text{D-DUDE}}}(x^n, Z^n) - L_{\hat{X}^{n,k_n^*}_{\text{D-DUDE}}}(x^n, Z^n)) && (32)\\ &\quad \le 2\sum_{k=0}^{k_n} E(|L_{\hat{X}^{n,k}_{\text{D-DUDE}}(x^n,Z^n)} - \hat{L}_{\hat{X}^{n,k}_{\text{D-DUDE}}}(x^n, Z^n)|) && (33)\\ &\quad \le 2b_3 k_n \left(\frac{k_n}{n}\right)^{1/4} && (34)\end{aligned}$$

where (29) follows from (13), (30) follows from (28), (32) follows from (14), (33) follows by taking absolute values and summing absolute estimated loss errors over all k, not just the two in the previous step, and (34) follows from (26). This proves the theorem with $\tilde{C} = 2b_3$. ∎

Improving the twice-universality penalty. The above proof suggests that it may be possible to improve the twice-universality penality through a more refined analysis. In particular, the worst case conditional variance bound identified in (22) corresponds to the case when the conditioning subsequence $z^{\mathcal{S}_j}$ forces all contexts to be the same. However, in this case, Jensen's inequality, as applied in (23) to bound the absolute deviation in terms of the variance, is weak. More specifically, an extension of the analysis of Section III reveals that the true behavior of the expected absolute deviation of the (unnormalized) estimated loss in this case is only $O(n^{1/2})$ and not $O(n^{3/4})$, as implied by (24).

The following intuitive considerations give some idea of the improvement in the twice-universality penalty that might be possible by tightening the above step in the proof. Consider

a conditioning subsequence $z^{\mathcal{S}_j}$ that results in m_n equally occuring contexts of length n/m_n. If m_n increases at a sufficient rate the m_n i.i.d. random variables $\tilde{\Delta}_{j,c_{-k}^{-1},c_1^k}(x^n, Z^n)$ corresponding to the occuring contexts will satisfy a central limit theorem. Let W_i, $i = 1, \ldots, m_n$ denote these random variables. We can use Lyapunov's condition [6] to determine such a central limit theorem inducing m_n. Letting σ_{W_i} denote the standard deviation of each W_i, Lyapunov's condition specialized to this application is that for some $\delta > 0$

$$\frac{m_n}{(\sqrt{m_n}\sigma_{W_i})^{2+\delta}} E(|W_i|^{2+\delta}) = o(1) \tag{35}$$

in which case $(1/\sqrt{m_n}) \sum_{i=1}^{m_n} W_i$ converges in distribution to $N(0, \sigma_{W_i}^2)$. As noted in the proof above,

$$\sigma_{W_i}^2 = O\left(\left[\frac{n}{m_n}\right]^{3/2}\right)$$

and through similar considerations, it can be shown that

$$E(|W_i|^{2+\delta}) = O\left(\left[\frac{n}{m_n}\right]^{2+\delta} \sqrt{\frac{m_n}{n}}\right).$$

Inserting these into (35) and simplifying reduces to the condition that m_n grows faster than $n^{1/5}$ in the sense that $n^{1/5}/m_n = o(1)$. The corresponding standard deviation of the sum of the W_i could thus grow almost as fast $\sqrt{m_n}\sigma_{W_i}$ $= O(n^{1/10}(n^{4/5})^{3/4}) = O(n^{7/10})$. After normalizing by n, this yields a standard deviation of $O(n^{-3/10})$. Since for a normally distributed random variable, the standard deviation and absolute deviation are of the same order, the above analysis would suggest, up to a polynomial factor in k_n, $O(n^{-3/10})$ (as opposed to $O(n^{-1/4})$ in the theorem statement) as a lower limit to the twice universality penalty that might be obtainable through a refinement of the Jensen's inequality/Schur convexity step in the proof of Theorem 8. We leave such refinements to future work and note that any additional improvements beyond $O(n^{-3/10})$ would likely require a significantly different approach.

Why D-DUDE and not DUDE? The technique underlying the proof of Theorem 8 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 $\tilde{\Delta}_{j,c_{-k}^{-1},c_1^k}(x^n, Z^n)$ 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 $\hat{X}_{\text{TU}}^n$ continues to yield the twice-universality properties of Theorem 8 is thus an open question.

ACKNOWLEDGMENT

We thank Tsachy Weissman for valuable discussions.

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] E. Ordentlich, M. 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.

[3] 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.

[4] S. Kutin, "Extensions to McDiarmid's inequality when differences are bounded with high probability," Technical Report TR 2002-04, University of Chicago, 2002.

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

[6] P. Billingsley, "Probability and Measure, 2nd Edition," John Wiley and Sons, 1986, p. 371.