Lower Limits of Discrete Universal Denoising

Krishnamurthy Viswanathan, Erik Ordentlich
HP Laboratories Palo Alto
HPL-2006-71
April 19, 2006*

denoising, universal algorithms, individual sequences, discrete memoryless channels

In the spirit of results on universal compression, we compare the performance of universal denoisers on discrete memoryless channels to that of the best performance obtained by a *k*-th order omniscient denoiser, namely one that is tuned to the transmitted noiseless sequence. We show that the additional loss incurred in the worst case by any universal denoiser on a length-*n* sequence grows at least like $\Omega\left(c^k n^{-\frac{1}{2}}\right)$, where *c* is a constant depending on the channel parameters and the loss function. This shows that for fixed *k* the additional loss incurred by the Discrete Universal Denoiser (DUDE) derived by Weissman *et al* is no larger than a constant multiplicative factor.

Furthermore we compare universal denoisers to denoisers that are aware of the distribution of the transmitted noiseless sequence. We show that, even for this weaker target loss, for any universal denoiser there exists some *i.i.d.* noiseless distribution whose optimum expected loss is lower than that incurred by the universal denoiser by $\Omega\left(n^{-\frac{1}{2}}\right)$.

* Internal Accession Date Only
Approved for External Publication

Lower Limits of Discrete Universal Denoising

Krishnamurthy Viswanathan and Erik Ordentlich
Hewlett Packard Labs, Palo Alto, CA 94304.

April 14, 2006

Abstract

In the spirit of results on universal compression, we compare the performance of universal denoisers on discrete memoryless channels to that of the best performance obtained by a k-th order omniscient denoiser, namely one that is tuned to the transmitted noiseless sequence. We show that the additional loss incurred in the worst case by any universal denoiser on a length-n sequence grows at least like $\Omega\left(\frac{c^k}{\sqrt{n}}\right)$, where c is a constant depending on the channel parameters and the loss function. This shows that for fixed k the additional loss incurred by the DUDE [1] is no larger than a constant multiplicative factor.

Furthermore we compare universal denoisers to denoisers that are aware of the distribution of the transmitted noiseless sequence. We show that, even for this weaker target loss, for any universal denoiser there exists some *i.i.d.* noiseless distribution whose optimum expected loss is lower than that incurred by the universal denoiser by $\Omega\left(\frac{1}{\sqrt{n}}\right)$.

1 Introduction

The problem of denoising is one of reproducing a signal based on observations obtained by passing it through a noisy channel, the quality of the reproduction being measured by a fidelity criterion. A version of this problem involving discrete memoryless channels was studied recently in [1]. In this setting, the clean and noisy signal are sequences of symbols belonging to the channel input and output alphabets respectively. In [1], a universal denoising algorithm, DUDE, was derived and its performance compared to the best sliding window denoiser for the noiseless-noisy pair of sequences in a semi-stochastic setting. It was shown that the additional loss incurred by the DUDE in this setting goes to zero as fast as $\mathcal{O}(kM^{2k}/\sqrt{n})$ where M is the size of the alphabet in question and k the order of the sliding window denoiser.

In this paper we derive lower bounds on the additional loss incurred by a denoiser in the worst-case when compared to the best kth-order sliding window denoiser for a given noiseless-noisy sequence pair. We show that for any denoiser and most channels and loss functions, this additional loss grows at least like $\Omega(c^k/\sqrt{n})$, where $c > 1$ is a function of the channel parameters and the loss function. This shows that for fixed k the additional loss incurred by the Discrete Universal Denoiser DUDE [1] is no larger than a constant multiplicative factor of the best possible.

We also prove a stronger result by deriving similar lower bounds for the excess loss incurred by a denoiser when measured against a benchmark that is a generalization of the one used in the compound decision problem [2], which can be viewed as a denoising problem over a binary input channel. In doing so we show that a certain rate of decay of excess loss, namely $\mathcal{O}(1/n)$, that can be achieved on continuous output channels cannot be achieved on discrete channels.

Extensions of these lower bounds to classes of sliding window denoisers based on a given bi-directional context set and to two-dimensionally indexed data are also derived. We also consider a stochastic variant of the same problem where for the class of *i.i.d.* noiseless sequences, we lower bound the additional average loss incurred by any denoiser when compared to the least loss incurred by any denoiser that is aware of the distribution.

We present the required notation in Section 2. Section 3 contains the main result of the paper as well as some of the preliminary results that lead to it. Section 4 states the correponding result for the benchmark considered in the compound decision problem. The above results and their implications are discussed in Section 5. Extensions to arbitrary sliding window denoisers based on arbitrary context sets and to two dimensionally indexed data are handled in Section 6. Finally, we consider a stochastic variant of the problem in Section 7.

2 Notation

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 and for an invertible matrix A^{-T} will denote the transpose of its inverse A^{-1}.

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 $s_i \in \mathcal{A}^{k_i}$, let $s_1 s_2$ denote the string formed by the concatenation of s_1 and s_2.

We 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 $\mathbf{\Pi} = \{\mathbf{\Pi}(i,j)\}_{i,j \in \mathcal{A}}$, 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 $i \in \mathcal{A}$, π_i denotes the ith column of $\mathbf{\Pi}$. Let $M = |\mathcal{A}|$ denote the size of the alphabet and $\mathcal{M}$ the simplex of M-dimensional probability vectors.

The denoiser outputs a reconstruction sequence $\{\hat{X}_t\}_{t=1}^n \in \mathcal{A}^n$. The loss function associated with the denoising problem is denoted by the *loss matrix* $\mathbf{\Lambda} = \{\Lambda(i,j)\}_{i,j \in \mathcal{A}}$, whose components are

also indexed by elements of $\mathcal{A}$, where $\Lambda(i,j)$ denotes the loss incurred by a denoiser that outputs j when the channel input was i. For $i \in \mathcal{A}$, let λ_i denote 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

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

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 kth-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 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). Note that we have slightly modified the definition of normalized cumulative loss to accomodate noiseless and noisy sequences of differing lengths. 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). This quantity will be one of the benchmarks against which we will compare the loss incurred by other denoisers.

The compound decision problem [3], as pointed out in [1], can be viewed as a denoising problem over a binary input channel. In work related to the compound decision problem "denoisers" are

measured against the best 0-th order denoiser that is aware of the noiseless sequence x^n but not tuned to the output sequence. This benchmark has been generalized [4] to

$$\begin{aligned}\bar{D}_k(x^n) &\stackrel{\text{def}}{=} \min_{\hat{X}^n \in \mathcal{S}_k} E\Big[L_{\hat{X}^n}\Big(x_{k+1}^{n-k}, Z^n\Big)\Big] \\ &= \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} \frac{1}{n-2k} \sum_{i=k+1}^{n-k} E\Big[\Lambda\Big(x_i, f\Big(Z_{i-k}^{i+k}\Big)\Big)\Big], \end{aligned} \tag{2}$$

the minimum expected loss incurred by any k-th order sliding window denoiser when the noiseless sequence is x^n. Clearly for all $x^n \in \mathcal{A}^n$,

$$\hat{D}_k(x^n) = E\left[\min_{\hat{X}^n \in \mathcal{S}_k} L_{\hat{X}^n}\Big(x_{k+1}^{n-k}, Z^n\Big)\right] \le \min_{\hat{X}^n \in \mathcal{S}_k} E\Big[L_{\hat{X}^n}\Big(x_{k+1}^{n-k}, Z^n\Big)\Big] = \bar{D}_k(x^n). \tag{3}$$

For any n-block denoiser $\hat{X}^n$ we can define two different *regret* functions,

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

and

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

to be the additional loss incurred in the worst-case, over the benchmarks defined in (1) and (2) respectively. From (3) for all n-block denoisers $\hat{X}^n$

$$\hat{R}_k\Big(\hat{X}^n\Big) \ge \bar{R}_k\Big(\hat{X}^n\Big).$$

The Discrete Universal Denoiser (DUDE) was proposed in [1] and it was shown that the regret of a sequence $\{\hat{X}_{\text{univ}}^{n,k}\}$[1] of such denoisers converges to zero with n. More precisely

$$\hat{R}_k\Big(\hat{X}_{\text{univ}}^{n,k}\Big) = \mathcal{O}\left(\sqrt{\frac{kM^{2k}}{n}}\right).$$

In this paper we investigate if this is the best possible rate of convergence. To do so we derive lower bounds on $\hat{R}_k\Big(\hat{X}^n\Big)$ and $\bar{R}_k\Big(\hat{X}^n\Big)$ for any n-block denoiser $\hat{X}^n$.

The above definitions measure the performance of a denoiser against the best sliding window denoisers in a semi-stochastic setting where x^n is a fixed individual noiseless sequence and Z^n is a random noisy sequence obtained when x^n is transmitted over the channel. One could also judge the universality of denoisers by considering the setting where the noiseless sequence X^n is a random process generated according to an unknown distribution $\mathbf{P}$ from some known class $\mathcal{P}$. Then the performance of denoisers could be measured against that obtained by a denoiser that knows $\mathbf{P}$. Formally, let $\mathbf{P}$ denote the distribution of the noiseless sequence X^n, and let

$$\mathbf{D}(\mathbf{P}) \stackrel{\text{def}}{=} \min_{\hat{X}^n \in \mathcal{D}_n} E_{\mathbf{P}}\big[L_{\hat{X}^n}(X^n, Z^n)\big],$$

[1] $\hat{X}_{\text{univ}}^{n,k}$ refers to the DUDE with parameter k

denote the minimum expected loss incurred by any n-block denoiser where the expectation is over all X^n distributed according to $\mathbf{P}$ and all Z^n that are outputs of the channel when X^n is the input. In this stochastic setting, the regret of an n-block denoiser $\hat{X}^n$ for a class of distributions $\mathcal{P}$ is defined to be

$$\mathbf{R}_{\mathcal{P}}\left(\hat{X}^n\right) \stackrel{\text{def}}{=} \max_{\mathbf{P}\in\mathcal{P}} E_{\mathbf{P}}\left[L_{\hat{X}^n}(X^n, Z^n)\right] - \mathbf{D}(\mathbf{P}).$$

It was shown in [1] that for the collection of all stationary processes the regret of the DUDE asymptotically tended to zero. In this paper we consider the subclass $\mathcal{I}_n$ of *i.i.d.* distributions over $\mathcal{A}^n$ and derive lower bounds on $\mathbf{R}_{\mathcal{I}_n}\left(\hat{X}^n\right)$ for any $\hat{X}^n \in \mathcal{D}_n$.

3 Main Result

The main result of the paper is that for most discrete memoryless channels and all $\hat{X}^n \in \mathcal{D}_n$,

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

where $c > 1$ is a constant that depends on the channel transition probability matrix $\mathbf{\Pi}$ and the loss function Λ. As we show later this applies to all non-trivial $(\mathbf{\Pi}, \Lambda)$ pairs. We also derive similar results for $\bar{R}_k\left(\hat{X}^n\right)$ and $\mathbf{R}_{\mathcal{I}_n}\left(\hat{X}^n\right)$. We consider the special case of 0th order denoisers in subsection 3.2, move on to the general case of kth order denoisers in subsection 3.3. The case of $\bar{R}_k\left(\hat{X}^n\right)$ and $\mathbf{R}_{\mathcal{I}_n}\left(\hat{X}^n\right)$ are handled in subsequent sections. To derive these results we first require a few preliminary Lemmas on denoisers that minimize expected loss when the noiseless sequence x^n is drawn according to a known *i.i.d.* distribution. These are presented in subsection 3.1

3.1 Bayes Response for *i.i.d.* Distributions

Given a loss matrix $\mathbf{\Lambda}$, the *Bayes response* (cf., *e.g.,* [5]) $\hat{x}(\mathbf{P})$ of any $\mathbf{P} \in \mathcal{M}$ is

$$\hat{x}(\mathbf{P}) = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{P},$$

and the corresponding *Bayes envelope* is

$$U(\mathbf{P}) = \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{P}.$$

Let

$$\hat{X}^n_{\text{opt}} \stackrel{\text{def}}{=} \arg\min_{\hat{X}^n\in\mathcal{D}_n} E\left[L_{\hat{X}^n}(X^n, Z^n)\right]$$

denote the Bayes-optimal denoiser, the n-block denoiser that minimizes the expected loss and let D_{opt} denote the minimum loss. Let $\mathbf{P}_{X_i|z^n}$ denote the column vector whose α-th component is $Pr(X_i = \alpha | Z^n = z^n)$. Then it is easy to see that

$$\hat{X}^n_{\text{opt}}(z^n)[i] = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{P}_{X_i|z^n} = \hat{x}\left(\mathbf{P}_{X_t|z^n}\right), \tag{4}$$

the Bayes response to $\mathbf{P}_{X_t|z^n}$ and the minimum expected loss is

$$D_{\text{opt}} = \frac{1}{n}\sum_{i=1}^{n} E\big[U\big(\mathbf{P}_{X_i|Z^n}\big)\big], \tag{5}$$

the expected value of the corresponding Bayes envelope.

Example 1. Let $\mathcal{A} = \{0, 1\}$. Let

$$\mathbf{\Pi}_{\text{BSC}} = \begin{bmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{bmatrix}$$

be the transition probability matrix of a binary symmetric channel with crossover probability δ, and let

$$\Lambda_{\text{Ham}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

represent the Hamming loss function. This example will be reexamined repeatedly in the paper.

The optimal denoiser for this example is the *Maximum Aposteriori Denoiser* (MAP) given by

$$\hat{X}^n_{\text{opt}}(z^n)[i] = \arg\max_{\hat{x}\in\{0,1\}} \mathbf{P}_{X_i|z^n}[\hat{x}],$$

and the corresponding optimal loss is

$$D_{\text{opt}} = \frac{1}{n}\sum_{i=1}^{n} E\left[1 - \max_{\hat{x}\in\{0,1\}} \mathbf{P}_{X_i|Z^n}[\hat{x}]\right]. \qquad \square$$

In the following Lemma we restate the well known fact that if X^n is drawn *i.i.d.* then $\hat{X}^n_{\text{opt}}$ is a 0-th order sliding window denoiser, *i.e.,*

$$\hat{X}^n_{\text{opt}}(z^n)[i] = \hat{X}^n_{\text{opt}}(y^n)[j]$$

if $z_i = y_j$. In other words the denoiser defines a function $f : \mathcal{A} \to \mathcal{A}$. Recall that the columns of $\mathbf{\Pi}$ are denoted by π_α, $\alpha \in \mathcal{A}$.

Lemma 1. If X^n is drawn *i.i.d.* according to $\mathbf{P}$ then

$$\hat{X}^n_{\text{opt}}(z^n)[i] = \arg\min_{\hat{x}\in\mathcal{A}} \frac{\lambda_{\hat{x}}^T(\mathbf{P}\odot\pi_{z_i})}{\mathbf{P}^T\pi_{z_i}},$$

and

$$D_{\text{opt}} = \sum_{z\in\mathcal{A}} \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T(\mathbf{P}\odot\pi_z).$$

Example 2. Continuing from Example 1 which deals with the BSC and Hamming loss, let the noiseless sequence X^n be an *i.i.d.* Bernoulli process with parameter p, namely, $\mathbf{P} = [1-p \quad p]^T$. Then from Lemma 1 the optimal denoiser is given by

$$\hat{X}^n_{\text{opt}}(z^n)[i] = \arg\min_{\hat{x}\in\{0,1\}} \frac{\lambda_{\hat{x}}^T(\mathbf{P}\odot\pi_{z_i})}{\mathbf{P}^T\pi_{z_i}} = \arg\max_{\hat{x}\in\{0,1\}} (\mathbf{P}\odot\pi_{z_i})[\hat{x}].$$

This can be further reduced to the following: if $z_i = 0$

$$\hat{X}^n_{\text{opt}}(z^n)[i] = \begin{cases} 0 & p < 1-\delta \\ 1 & p > 1-\delta \\ \text{either} & p = 1-\delta \end{cases}$$

and if $z_i = 1$

$$\hat{X}^n_{\text{opt}}(z^n)[i] = \begin{cases} 0 & p < \delta \\ 1 & p > \delta \\ \text{either} & p = \delta. \end{cases}$$

The optimal loss is given by

$$D_{\text{opt}} = \sum_{z \in \{0,1\}} \min_{\hat{x} \in \mathcal{A}} (\mathbf{P} \odot \pi_z)[\hat{x}].$$

If $\delta \leq 1/2$ then this reduces to

$$D_{\text{opt}} = \begin{cases} p & 0 \leq p \leq \delta \\ \delta & \delta < p \leq 1-\delta \\ 1-p & 1-\delta < p \leq 1. \end{cases}$$

If $\delta > 1/2$ the loss can be obtained by replacing δ with $1-\delta$ in the above expression. □

In the subsequent subsections we employ Lemma 1 to derive lower bounds on $\hat{R}_k\left(\hat{X}^n\right)$ and $\bar{R}_k\left(\hat{X}^n\right)$ for any $\hat{X}^n \in \mathcal{D}_n$.

3.2 Zeroth order

The pair $(\mathbf{\Pi}, \mathbf{\Lambda})$, comprising a $M \times M$ channel transition probability matrix $\mathbf{\Pi}$ and a $M \times M$ loss matrix $\mathbf{\Lambda}$ is *neutralizable* if there exist $t, i, j \in \mathcal{A}$ such that for some distribution $\mathbf{P} \in \mathcal{M}$, $\mathbf{P} \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$, and

$$\mathbf{P}^T(\lambda_i \odot \pi_t) = \mathbf{P}^T(\lambda_j \odot \pi_t) = \min_{k \in \mathcal{A}} \mathbf{P}^T(\lambda_k \odot \pi_t). \tag{6}$$

The distribution $\mathbf{P}$ is said to be *loss-neutral* with respect to $(\pi_t, \lambda_i, \lambda_j)$.

Consider a denoiser $\hat{X}^n$ with the property $\hat{X}^n(z^n)[i] = k$ if $z_i = t$. If X^n is drawn *i.i.d.* according to $\mathbf{P}$, $\mathbf{P}^T(\lambda_k \odot \pi_t)$ is the average loss incurred by this denoiser in reconstructing the symbols whose noisy version $Z_i = t$. If (6) is satisfied, then it implies that there are two Bayes optimal denoisers. One returns i on observing t and the other returns j. The condition $\mathbf{P} \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$ ensures that the denoisers differ in a non-trivial fashion. Therefore a loss-neutral distribution is an iid distribution on the clean sequence that results in at least two distinct Bayes-optimal denoisers.

The class of neutralizable $(\mathbf{\Pi}, \mathbf{\Lambda})$ pairs is rich enough to accomodate commonly encountered non-trivial channels and loss functions, *e.g.,* the Binary Symmetric Channel and the Hamming loss function.

Example 3. The BSC-Hamming loss pair is neutralizable. Indexing columns by elements of $\{0,1\}$, $\pi_1 = [\delta \quad 1-\delta]^T$, $\lambda_0 = [0 \quad 1]^T$ and $\lambda_1 = [1 \quad 0]^T$. Choosing $\mathbf{P} = [1-\delta \quad \delta]^T$ we obtain

$$\mathbf{P}^T(\lambda_0 \odot \pi_1) = \mathbf{P}^T(\lambda_1 \odot \pi_1) = (1-\delta)\delta.$$

Hence $(\mathbf{\Pi}_{\text{BSC}}, \Lambda_{\text{Ham}})$ is neutralizable and $\mathbf{P} = [1-\delta \quad \delta]^T$ is a loss-neutral distribution. Note that if $\delta \neq 1/2$, the uniform distribution is not loss-neutral. □

In fact, if $(\mathbf{\Pi}, \mathbf{\Lambda})$ is not neutralizable then the denoising problem is trivial, *i.e.,* there exists a symbol-by-symbol denoiser $\hat{X}$ whose loss incurred is the least possible for all noiseless sequences. The arguments that justify this claim follow.

Suppose for all t, there exists $\ell(t)$ such that, for all $\mathbf{P} \in \mathcal{M}$, and all $k \in \mathcal{A}$

$$\mathbf{P}^T\big(\lambda_{\ell(t)} \odot \pi_t\big) \leq \mathbf{P}^T(\lambda_k \odot \pi_t). \tag{7}$$

Then for all $\alpha \in \mathcal{A}$ and any k

$$\mathbf{\Lambda}(\alpha, \ell(t))\mathbf{\Pi}(\alpha, t) \leq \mathbf{\Lambda}(\alpha, k)\mathbf{\Pi}(\alpha, t).$$

Then the denoiser $\hat{X}^*(z^n)[i] = \ell(z_i)$ is optimal in a strong sense, namely, for all $x^n, Z^n \in \mathcal{A}^n$ that have a non-zero probability, and any $\hat{X}^n \in \mathcal{D}_n$

$$L_{\hat{X}^*}(x^n, Z^n) \leq L_{\hat{X}^n}(x^n, Z^n).$$

This implies that no class of denoisers is rich enough to ensure a positive regret.

We will show that if $(\mathbf{\Pi}, \mathbf{\Lambda})$ is not neutralizable, then (7) is satisfied. Observe that if $(\lambda_i - \lambda_j) \odot \pi_t = 0$, then $\mathbf{P}^T(\lambda_i \odot \pi_t) = \mathbf{P}^T(\lambda_j \odot \pi_t)$ for all $\mathbf{P} \in \mathcal{M}$. Therefore, for the subsequent arguments, it suffices to consider distinct columns $\lambda_i \odot \pi_t$.

For $i \in \mathcal{M}$, let

$$\mathcal{M}_i \stackrel{\text{def}}{=} \big\{\mathbf{P} \in \mathcal{M} : \forall k \in \mathcal{A},\ \mathbf{P}^T(\lambda_i \odot \pi_t) \leq \mathbf{P}^T(\lambda_k \odot \pi_t)\big\}.$$

Observe that $\mathcal{M}_i$ is an intersection of halfspaces and is therefore a convex polytope defined by a subset of the hyperplanes $\mathcal{H}_{i,j}$, $i \neq j$, where

$$\mathcal{H}_{i,j} = \big\{x \in \mathbb{R}^M : x^T((\lambda_i - \lambda_j) \odot \pi_t) = 0\big\}.$$

Note that, since $(\mathbf{\Pi}, \mathbf{\Lambda})$ is not neutralizable, none of the hyperplanes that define the boundary of any polytope $\mathcal{M}_i$ can intersect the interior of $\mathcal{M}$. Otherwise, (6) will be satisfied for some i, j, and $\mathbf{P}$ in the interior of $\mathcal{M}$. Since $\mathbf{P}$ is in the interior, $\mathbf{P}(\alpha) > 0$ for all α, and since $\lambda_i \odot \pi_t$ and $\lambda_j \odot \pi_t$ are distinct, $\mathbf{P} \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$, which contradicts the non-neutralizability of $(\mathbf{\Pi}, \mathbf{\Lambda})$. Since

$$\bigcup_{i \in \mathcal{A}} \mathcal{M}_i = \mathcal{M}$$

and none of the hyperplanes that define the boundaries of $\mathcal{M}_i$ intersect the interior of $\mathcal{M}$, there exists i_0 such that $\mathcal{M} = \mathcal{M}_{i_0}$ for some i_0. By setting $\ell(t) = i_0$, (7) is satisfied.

Thus, we have shown that the class of non-neutralizable $(\mathbf{\Pi}, \mathbf{\Lambda})$ pairs poses trivial denoising problems. We are now in a position to state our theorem

Theorem 2. For any neutralizable pair $(\mathbf{\Pi}, \mathbf{\Lambda})$, and any sequence $\{\hat{X}^n \in \mathcal{D}_n\}$ of denoisers, as n tends to infinity

$$\hat{R}_0\Big(\hat{X}^n\Big) \geq \frac{c}{\sqrt{n}}(1 + o(1)),$$

where c is a positive function of $\mathbf{\Pi}, \mathbf{\Lambda}$ and a loss-neutral distribution $\mathbf{P}^*$. □

In order to provide an intuition we first consider the example of the BSC and Hamming loss and provide an outline of the proof for that specific case. The more general proof will be presented after the example.

Example 4. As shown in Example 3, $(\mathbf{\Pi}_{\text{BSC}}, \Lambda_{\text{Ham}})$ is neutralizable and $\mathbf{P}^* = [1-\delta \quad \delta]^T$ is a loss-neutral distribution. Observe that for all $\hat{X}^n \in \mathcal{D}_n$ and any distribution $\mathbf{P}$ on X^n

$$\hat{R}_0\Big(\hat{X}^n\Big) \geq E_{\mathbf{P}}\Big[E\big[L_{\hat{X}^n}(X_1^n, Z^n)\big] - \hat{D}_0(X^n)\Big].$$

This is true in particular for $\mathbf{P}$ *i.i.d.* with marginal distribution $\mathbf{P}^*$. Setting $p = \delta$ in Example 2

$$E_{\mathbf{P}^*}\big[E\big[L_{\hat{X}^n}(X_1^n, Z^n)\big]\big] \geq D_{\text{opt}} = \delta, \tag{8}$$

and the lower bound is achieved by the 0-th order sliding window denoiser

$$\hat{X}^n_{\text{opt}}(z^n)[t] = \begin{cases} 0 & z_t = 0 \\ \text{either} & z_t = 1 \end{cases}$$

where δ is assumed to be less than $1/2$.

To upper bound $\hat{D}_0(x^n) = E[D_0(x^n, Z^n)]$ we construct a 0-th order sliding window denoiser. The normalized *Hamming weight* $d_H(x^n)$ of x^n is the fraction of 1s in x^n, and the normalized *Hamming distance* $d_H(x^n, z^n)$ between x^n and z^n is the fraction of bit positions in which they differ. Then for each (x^n, z^n) define $\hat{X}^n$, a 0-th order sliding window denoiser, to be

$$\hat{X}^n(z^n)[t] = \begin{cases} 0 & z_t = 0 \\ 0 & z_t = 1, d_H(x^n, z^n) \geq d_H(x^n) \\ 1 & z_t = 1, d_H(x^n, z^n) < d_H(x^n). \end{cases}$$

By definition

$$D_0(x^n, z^n) \leq L_{\hat{X}^n}(x^n, z^n) = \min\{d_H(x^n, z^n), d_H(x^n)\}.$$

Therefore

$$E_{\mathbf{P}^*}\Big[\hat{D}_0(X^n)\Big] = E_{\mathbf{P}^*}[D_0(X^n, Z^n)] \leq E_{\mathbf{P}^*}[\min\{d_H(X^n, Z^n), d_H(X^n)\}].$$

It is easy to verify that as n tends to infinity both $\sqrt{n}(d_H(X^n, Z^n) - \delta)$ and $\sqrt{n}(d_H(X^n) - \delta)$ tend to independent and identically distributed Gaussian random variables with mean 0 and variance $\delta(1-\delta)$. Hence

$$E_{\mathbf{P}^*}[\min\{d_H(X^n, Z^n), d_H(X^n)\}] = \delta - \sqrt{\frac{\delta(1-\delta)}{n}}c(1 + o(1))$$

where c is the expected value of the maximum of two independent zero mean Gaussian random variables with unit variance, and therefore positive. Substituting in (8)

$$\hat{R}_0\Big(\hat{X}^n\Big) \geq E_{\mathbf{P}^*}\Big[E\big[L_{\hat{X}^n}(X_1^n, Z^n)\big] - \hat{D}_0(X^n)\Big] \geq \sqrt{\frac{\delta(1-\delta)}{n}}c(1+o(1)). \qquad \square$$

Now we present a more rigorous version of the argument in the example, that is general enough to address any neutralizable pair $(\mathbf{\Pi}, \Lambda)$. To characterize the benchmark in $\hat{R}_0\Big(\hat{X}^n\Big)$, namely $\hat{D}_0(x^n)$, we require some notation for the frequency of occurrence of symbols in x^n and z^n. We employ the following notation that was employed in [1]. For $x^n, z^n \in \mathcal{A}^n$, $c \in \mathcal{A}$ let $\mathbf{q}(z^n, z^n, c)$ denote the M-dimensional column vector whose j-th component, $j \in \mathcal{A}$, is

$$\mathbf{q}(z^n, x^n, c)[j] = \frac{1}{n}\left|\{i : 1 \leq i \leq n, z_i = c, x_i = j\}\right|$$

the frequency of the occurrence of c in z^n along with j in the corresponding location in x^n. If X^n is drawn *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$ and Z^n represents the noisy output from the channel then

$$E[\mathbf{q}(Z^n, X^n, c)] = \mathbf{P} \odot \pi_c. \tag{9}$$

We express the best 0-th order mimimum loss $D_0(x^n, z^n)$ for the pair (x^n, z^n) in terms of the vectors $\mathbf{q}(z^n, x^n, c)$, $c \in \mathcal{A}$. Observe that for all $x^n, z^n \in \mathcal{A}^n$

$$\begin{aligned} D_0(x^n, z^n) = \min_{f:\mathcal{A}\to\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\Lambda(x_i, f(z_i)) &= \sum_{c\in\mathcal{A}}\min_{\hat{x}\in\mathcal{A}}\sum_{j\in\mathcal{A}}\Lambda(j,\hat{x})\mathbf{q}(z^n, x^n, c)[j] \\ &= \sum_{c\in\mathcal{A}}\min_{\hat{x}\in\mathcal{A}}\lambda_{\hat{x}}^T\mathbf{q}(z^n, x^n, c). \end{aligned} \tag{10}$$

To prove Theorem 2 we require the following lemma on the asymptotics of $\mathbf{q}(Z^n, X^n, c)$.

Lemma 3. If X^n is generated *i.i.d.* according to some $\mathbf{P}$ in $\mathcal{M}$, then for any column vector $\alpha \in \mathbb{R}^M$, and any $c \in \mathcal{A}$,

$$\lim_{n\to\infty} E_{\mathbf{P}}\big[\sqrt{n}\big(\big|\alpha^T\mathbf{q}(Z^n, X^n, c) - \alpha^T(\mathbf{P}\odot\pi_c)\big|\big)\big] = \sqrt{\frac{2V}{\pi}}$$

where $V = (\alpha\odot\alpha)^T(\mathbf{P}\odot\pi_c) - \big(\alpha^T(\mathbf{P}\odot\pi_c)\big)^2$.

Proof We first show that when X^n is generated *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$, for any column vector $\alpha \in \mathbb{R}^M$, and any $c \in \mathcal{A}$, as n tends to infinity, $\alpha^T\mathbf{q}(Z^n, X^n, c)$ suitably normalized converges in distribution to a Gaussian random variable, namely,

$$\sqrt{n}\big(\alpha^T\mathbf{q}(Z^n, X^n, c) - \alpha^T(\mathbf{P}\odot\pi_c)\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V) \tag{11}$$

where $V = (\alpha\odot\alpha)^T(\mathbf{P}\odot\pi_c) - \big(\alpha^T(\mathbf{P}\odot\pi_c)\big)^2$.

For a given $\alpha \in \mathbb{R}^M$ and $c \in \mathcal{A}$ we define the sequence Y^n of random variables as

$$Y_i \stackrel{\text{def}}{=} \sum_{j\in\mathcal{A}}\alpha(j)1(X_i = j, Z_i = c)$$

where $1(\cdot)$ is the indicator function. Then

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \alpha^T \mathbf{q}(Z^n, X^n, c)$$

and if the sequence X^n is drawn *i.i.d.* according to $\mathbf{P}$, then the sequence Y^n is also *i.i.d.* and has finite moments. Hence the central limit theorem applies to it. Therefore

$$\sqrt{n}\big(\alpha^T \mathbf{q}(Z^n, X^n, c) - E_{\mathbf{P}}[Y_i]\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{var}[Y_i])$$

where $\mathrm{var}[Y_i]$ denotes the variance of Y_i. Equation (11) follows as

$$E_{\mathbf{P}}[Y_i] = \sum_{j\in\mathcal{A}} \alpha(j) Pr(X_i = j, Z_i = c) = \alpha^T(\mathbf{P} \odot \pi_c)$$

and

$$E_{\mathbf{P}}\big[Y_i^2\big] = \sum_{j\in\mathcal{A}} \alpha(j)^2 Pr(X_i = j, Z_i = c) = (\alpha \odot \alpha)^T(\mathbf{P} \odot \pi_c).$$

A straightforward extension yields

$$\sqrt{n}\big(\big|\alpha^T \mathbf{q}(Z^n, X^n, c) - \alpha^T(\mathbf{P} \odot \pi_c)\big|\big) \xrightarrow{\mathcal{L}} |G|$$

where $G \sim \mathcal{N}(0, V)$.

To prove the lemma from this point onwards we use the fact (cf. *e.g.,* Theorem 25.12 [6]) that if a sequence of random variables A^n are *uniformly integrable*, *i.e.,*

$$\lim_{\beta\to\infty} \sup_n \int_{[|A_n|\geq\beta]} |A_n| dP = 0$$

and if $A_n \xrightarrow{\mathcal{L}} A$ then $\lim_{n\to\infty} E[A_n] = E[A]$, where in our case

$$A_n = \sqrt{n}\big(\big|\alpha^T \mathbf{q}(Z^n, X^n, c) - \alpha^T(\mathbf{P} \odot \pi_c)\big|\big).$$

Observe that

$$\int_{[|A_n|\geq\beta]} |A_n| dP \leq \int \frac{|A_n|}{\beta} |A_n| dP = \frac{E\big[|A_n|^2\big]}{\beta}.$$

Since

$$E\big[|A_n|^2\big] = n\frac{\mathrm{var}[Y_i]}{n} = V,$$

we obtain

$$\lim_{\beta\to\infty} \sup_n \int_{[|A_n|\geq\beta]} |A_n| dP \leq \lim_{\beta\to\infty} \sup_n \frac{V}{\beta} = 0.$$

Hence A^n is uniformly integrable. Therefore

$$\lim_{n\to\infty} E_{\mathbf{P}}\big[\sqrt{n}\big(\big|\alpha^T \mathbf{q}(Z^n, X^n, c) - \alpha^T(\mathbf{P} \odot \pi_c)\big|\big)\big] = E_{\mathbf{P}}[|G|] = \sqrt{\frac{2V}{\pi}}. \qquad \square$$

We are now in a position to prove Theorem 2. As in Example 4, we lower bound $\hat{R}_0\Big(\hat{X}^n\Big)$, the extra loss incurred in the worst-case, by the expected extra loss when the noiseless sequence X^n is drawn according an *i.i.d.* distribution. This proof technique is similar to the one employed for the problem of binary prediction in [7]. However choosing a uniform distribution, like in [7], for X^n does not yield the required results - the distribution chosen has to be loss-neutral.

Proof of Theorem 2 Let $t, i, j \in \mathcal{A}$ and let the distribution $\mathbf{P}^* \in \mathcal{M}$ be loss-neutral with respect to $(\pi_t, \lambda_i, \lambda_j)$, so that $\mathbf{P}^* \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$ and

$$(\mathbf{P}^*)^T(\lambda_i \odot \pi_t) = (\mathbf{P}^*)^T(\lambda_j \odot \pi_t) = \min_{k \in \mathcal{A}} (\mathbf{P}^*)^T(\lambda_k \odot \pi_t). \tag{12}$$

By definition

$$\hat{R}_0\Big(\hat{X}^n\Big) = \max_{x^n \in \mathcal{A}^n} E\big[L_{\hat{X}^n}(x_1^n, Z^n)\big] - \hat{D}_0(x^n),$$

hence for any *i.i.d.* distribution $\mathbf{P} \in \mathcal{M}$ on X^n and for all $\hat{X}^n \in \mathcal{D}_n$

$$\hat{R}_0\Big(\hat{X}^n\Big) \geq E_{\mathbf{P}}\Big[E\big[L_{\hat{X}^n}(X_1^n, Z^n)\big] - \hat{D}_0(X^n)\Big]. \tag{13}$$

In particular this is true for $\mathbf{P}^*$. Since X^n is generated *i.i.d.* accroding to $\mathbf{P}^*$ it follows from Lemma 1 that for all $\hat{X}^n \in \mathcal{D}_n$

$$E_{\mathbf{P}^*}\big[E\big[L_{\hat{X}^n}(X_1^n, Z^n)\big]\big] \geq D_{\text{opt}} = \sum_{z \in \mathcal{A}} \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T(\mathbf{P}^* \odot \pi_z). \tag{14}$$

Now we upper bound $E_{\mathbf{P}^*}\Big[\hat{D}_0(X^n)\Big]$. From (10)

$$\begin{aligned}
E_{\mathbf{P}^*}\Big[\hat{D}_0(X^n)\Big] &= E_{\mathbf{P}^*}[D_0(X^n, Z^n)] \\
&= E_{\mathbf{P}^*}\left[\sum_{c \in \mathcal{A}} \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}(Z^n, X^n, c)\right] \\
&= E_{\mathbf{P}^*}\left[\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}(Z^n, X^n, t)\right] + \sum_{c \in \mathcal{A}, c \neq t} E_{\mathbf{P}^*}\left[\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}(Z^n, X^n, c)\right] \\
&\leq E_{\mathbf{P}^*}\left[\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}(Z^n, X^n, t)\right] + \sum_{c \in \mathcal{A}, c \neq t} \min_{\hat{x} \in \mathcal{A}} E_{\mathbf{P}^*}\left[\lambda_{\hat{x}}^T \mathbf{q}(Z^n, X^n, c)\right] \\
&= E_{\mathbf{P}^*}\left[\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}(Z^n, X^n, t)\right] + \sum_{c \in \mathcal{A}, c \neq t} \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T(\mathbf{P}^* \odot \pi_c).
\end{aligned} \tag{15}$$

In the sequence of inequalities and equalities above, the inequality holds since for any set of real random variables $\{X_a : a \in \mathcal{A}\}$

$$E\left[\min_{a \in A} X_a\right] \leq \min_{a \in A} E[X_a],$$

and the last equality follows from (9).

Substituting (15) and (14) in (13) we obtain for all $\hat{X}^n \in \mathcal{D}_n$

$$\hat{R}_0\Big(\hat{X}^n\Big) \geq \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T(\mathbf{P}^* \odot \pi_t) - E_{\mathbf{P}^*}\left[\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}(Z^n, X^n, t)\right].$$

Noting that for any M-dimensional vector $\mathbf{P}$ and any $a, t \in \mathcal{A}$, $\mathbf{P}^T(\lambda_a \odot \pi_t) = \lambda_a^T(\mathbf{P} \odot \pi_t)$ and considering (12)

$$\hat{R}_0\Big(\hat{X}^n\Big) \geq \lambda_i^T(\mathbf{P}^* \odot \pi_t) - E_{\mathbf{P}^*}\big[\min\big\{\lambda_i^T\mathbf{q}(Z^n, X^n, t), \lambda_j^T\mathbf{q}(Z^n, X^n, t)\big\}\big].$$

Since for any $x, y \in \mathbb{R}$, $2\min\{x, y\} = (x + y - |x - y|)$, and since $\lambda_i^T(\mathbf{P}^* \odot \pi_t) = \lambda_j^T(\mathbf{P}^* \odot \pi_t)$,

$$E_{\mathbf{P}^*}\big[\min\big\{\lambda_i^T\mathbf{q}(Z^n, X^n, t), \lambda_j^T\mathbf{q}(Z^n, X^n, t)\big\}\big] = \lambda_i^T(\mathbf{P}^* \odot \pi_t) - \frac{1}{2}\Big(E_{\mathbf{P}^*}\Big[\Big|(\lambda_i - \lambda_j)^T\mathbf{q}(Z^n, X^n, t)\Big|\Big]\Big).$$

Therefore

$$\hat{R}_0\Big(\hat{X}^n\Big) \geq \frac{1}{2}\Big(E_{\mathbf{P}^*}\Big[\Big|(\lambda_i - \lambda_j)^T\mathbf{q}(Z^n, X^n, t)\Big|\Big]\Big).$$

Applying Lemma 3 with $\mathbf{P} = \mathbf{P}^*$, $c = t$, and $\alpha = \lambda_i - \lambda_j$ and observing from (12) that

$$\alpha^T(\mathbf{P} \odot \pi_c) = (\lambda_i - \lambda_j)^T(\mathbf{P}^* \odot \pi_t) = 0,$$

we obtain

$$\lim_{n\to\infty} \sqrt{n} E_{\mathbf{P}^*}\Big[\Big|(\lambda_i - \lambda_j)^T\mathbf{q}(Z^n, X^n, t)\Big|\Big] = \sqrt{\frac{2V}{\pi}}$$

where

$$V = ((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T(\mathbf{P}^* \odot \pi_t).$$

Note that since $\mathbf{P}^*$ is a loss-neutral distribution $(\lambda_i - \lambda_j) \odot \mathbf{P}^* \odot \pi_t \neq 0$ and therefore $V > 0$. This proves the theorem. □

3.3 Higher order

Theorem 2 can be extended to $\hat{R}_k\Big(\hat{X}^n\Big)$. There we compare the loss incurred by a denoiser to $\hat{D}_k(x^n)$, a smaller quantity, and consequently the lower bound that we obtain on $\hat{R}_k\Big(\hat{X}^n\Big)$ is larger. In fact, the lower bound increases exponentially with k. The proof is very much along the lines of that for Theorem 2 and therefore we refer to it frequently here. As in Theorem 2 we require a few preliminaries.

Following [1], we extend the definition of $\mathbf{q}(\cdot)$ to count the frequency of sequences of length $2k + 1$. For $x^n, z^n \in \mathcal{A}^n$, $c_{-k}^k \in \mathcal{A}^{2k+1}$ let $\mathbf{q}\big(z^n, x^n, c_{-k}^k\big)$ denote the M-dimensional column vector whose j-th component, $j \in \mathcal{A}$, is

$$\mathbf{q}\Big(z^n, x^n, c_{-k}^k\Big)[j] = \frac{1}{n - 2k}\Big|\Big\{i : k + 1 \leq i \leq n - k, z_{i-k}^{i+k} = c_{-k}^k, x_i = j\Big\}\Big|$$

the frequency of occurrence of the sequence c_{-k}^k in z^n along with j in x^n at the location corresponding to c_0 in z^n. Also note that if X^n is drawn *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$ and Z^n represents the noisy output from the channel then Z^n is also an *i.i.d.* sequence and

$$E\Big[\mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\Big] = (\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k, i\neq 0}^{k} \mathbf{P}^T\pi_{c_i}. \tag{16}$$

We express the best k-th order mimimum loss $D_k(x^n, z^n)$ for the pair (x^n, z^n) in terms of the vectors $\mathbf{q}\big(z^n, x^n, c_{-k}^k\big)$, $c_{-k}^k \in \mathcal{A}^{2k+1}$. Observe that for all $x^n, z^n \in \mathcal{A}^n$

$$D_k(x^n, z^n) = \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}} \min_{\hat{x}\in\mathcal{A}} \sum_{j\in\mathcal{A}} \Lambda(j,\hat{x}) \mathbf{q}\Big(z^n, x^n, c_{-k}^k\Big)[j] = \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}} \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}\Big(z^n, x^n, c_{-k}^k\Big). \quad (17)$$

To prove Theorem 2, we required a lemma on the asymptotics of $\mathbf{q}(Z^n, X^n, c_0)$, which can be written as a sum of *i.i.d.* random variables. Note, however, that for $k \geq 1$, $\mathbf{q}\big(Z^n, X^n, c_{-k}^k\big)$ can no longer be written as a sum of *i.i.d.* random variables and, therefore, the standard Central Limit Theorem, which was used in Lemma 3, does not apply. To address this problem, we require a Central Limit Theorem for dependent random variables such as the one proved by Hoeffding *et al* [8]. We state the theorem below. A sequence X^n of random variables is m-dependent if for all $s > r + m$ $(X_1, X_2, \ldots, X_r)$ and $(X_s, X_{s+1}, \ldots, X_n)$ are independent.

Theorem 4. [8] For a stationary and m-dependent sequence X^n of random variables such that $E[X_1] = 0$, and $E\big[|X_1|^3\big] < \infty$, as n tends to infinity

$$n^{-1/2} \sum_{i=1}^{n} X_i \xrightarrow{\mathcal{L}} \mathcal{N}(0, V)$$

where

$$V = E\big[X_1^2\big] + 2 \sum_{i=2}^{m+1} E[X_1 X_i]. \qquad \square$$

Applying this theorem to $\mathbf{q}\big(Z^n, X^n, c_{-k}^k\big)$ results in the following lemma.

Lemma 5. If X^n is generated *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$, then for any column vector $\alpha \in \mathbb{R}^M$, and any $c_{-k}^k \in \mathcal{A}^{2k+1}$, such that $\alpha^T(\mathbf{P} \odot \pi_{c_0}) = 0$

$$\lim_{n\to\infty} E_{\mathbf{P}}\Big[\sqrt{n}\,\Big|\alpha^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\Big|\Big] = \sqrt{\frac{2V}{\pi}}$$

where $V = (\alpha \odot \alpha)^T (\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k, i\neq 0}^{k} \mathbf{P}^T \pi_{c_i}$.

Proof We first show that when X^n is generated *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$, for any column vector $\alpha \in \mathbb{R}^M$, and any $c_{-k}^k \in \mathcal{A}^{2k+1}$ satisfying $\alpha^T(\mathbf{P} \odot \pi_{c_0}) = 0$, as n tends to infinity, $\alpha^T \mathbf{q}\big(Z^n, X^n, c_{-k}^k\big)$ suitably normalized converges in distribution to a Gaussian random variable, namely,

$$\sqrt{n}\Big(\alpha^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V) \quad (18)$$

where $V = (\alpha \odot \alpha)^T (\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k, i\neq 0}^{k} \mathbf{P}^T \pi_{c_i}$.

For a given $\alpha \in \mathbb{R}^M$ and $c_{-k}^k \in \mathcal{A}^{2k+1}$ we define the sequence Y_{k+1}^{n-k} of random variables as

$$Y_i \stackrel{\text{def}}{=} \sum_{\ell\in\mathcal{A}} \alpha(\ell) 1\Big(X_i = \ell, Z_{i-k}^{i+k} = c_{-k}^k\Big), \qquad k+1 \leq i \leq n-k.$$

Then

$$\frac{1}{n-2k} \sum_{i=k+1}^{n-k} Y_i = \alpha^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big).$$

If the sequence X^n is drawn *i.i.d.* according to $\mathbf{P}$, Y_{k+1}^{n-k} is stationary and since each Y_i is a function of $2k+1$ consecutive X_i's, it is easy to verify that Y_{k+1}^{n-k} is a $2k$-dependent sequence. Furthermore

$$E_{\mathbf{P}}[Y_i] = \sum_{\ell \in \mathcal{A}} \alpha(\ell) Pr\Big(X_i = \ell, Z_{i-k}^{i+k} = c\Big) = \alpha^T (\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k, i\neq 0}^{k} \mathbf{P}^T \pi_{c_i} = 0$$

where the last equality follows from the choice of α and c_0. Furthermore the higher moments of Y_i exist, hence Theorem 4 applies and therefore

$$\sqrt{n}\Big(\alpha^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, E\big[Y_{k+1}^2\big] + 2\sum_{i=1}^{2k} E[Y_{k+1}Y_{k+1+i}]\right). \tag{19}$$

Observe that

$$E\big[Y_{k+1}^2\big] = \sum_{\ell \in \mathcal{A}} \alpha(\ell)^2 Pr\Big(X_{k+1} = \ell, Z_1^{2k+1} = c_{-k}^k\Big) = V = (\alpha \odot \alpha)^T (\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k, i\neq 0}^{k} \mathbf{P}^T \pi_{c_i}$$

and we will show that for all $i \geq 1$

$$E[Y_{k+1}Y_{k+1+i}] = 0.$$

Substituting these in (19) establishes (18). Define the collection of indicator functions $S_r : \mathcal{A}^{2k+1} \to \{0,1\}$, $1 \leq r \leq 2k$, as

$$S_r\Big(c_{-k}^k\Big) = 1(c_i = c_{i+r} \text{ for all } -k \leq i \leq k-r)$$

that indicate whether the sequence c_{-k}^k partially matches a shifted version of itself. For example, $S_1\big(c_{-k}^k\big) = 1$ iff $c_{-k}^k = 00\ldots0$ or $11\ldots1$, and $S_2(10101) = 1$, $S_2(10110) = 0$ etc. In general S_r indicates if the sequence is periodic with period r. Also define the transformation $A_r : \mathcal{A}^{2k+1} \to \mathcal{A}^{2k+1+r}$ as

$$A_r\Big(c_{-k}^k\Big) = c_{-k}^k c_{k-r+1}^k,$$

namely, the concatenation of c_{-k}^k and the last r symbols of c_{-k}^k. Then for any $1 \leq i \leq 2k$ and $c_{-k}^k \in \mathcal{A}^{2k+1}$, letting $a_1^{2k+1+i} = A_i\big(c_{-k}^k\big)$ we obtain

$$Y_{k+1}Y_{k+1+i} = S_i\Big(c_{-k}^k\Big) \sum_{\ell \in \mathcal{A}} \sum_{m \in \mathcal{A}} \alpha(\ell)\alpha(m) 1\Big(X_{k+1} = \ell, X_{k+1+i} = m, Z_1^{2k+i+1} = a_1^{2k+1+i}\Big).$$

Since X^n is drawn *i.i.d.* according to $\mathbf{P}$, for $i \geq 1$

$$\begin{aligned}
E[Y_{k+1}Y_{k+1+i}] &= S_i\Big(c_{-k}^k\Big)\left(\sum_{\ell \in \mathcal{A}} \alpha(\ell) Pr(X_{k+1} = \ell, Z_{k+1} = c_0)\right)^2 \prod_{j=1, j\neq k+1, k+1+i}^{2k+i+1} Pr(Z_j = a_j) \\
&= S_i\Big(c_{-k}^k\Big)\big(\alpha^T(\mathbf{P} \odot \pi_{c_0})\big)^2 \prod_{j=1, j\neq k+1, k+1+i}^{2k+i+1} \mathbf{P}^T \pi_{a_j} \\
&= 0
\end{aligned}$$

where the last equality follows from the fact that $\alpha^T(\mathbf{P} \odot \pi_{c_0}) = 0$, by assumption.

A straightforward extension of (18) yields

$$\sqrt{n}\Big(\Big|\alpha^T\mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\Big|\Big) \xrightarrow{\mathcal{L}} |G|$$

where $G \sim \mathcal{N}(0, V)$. Arguments similar to the ones used in Lemma 3 can be used to show that $\sqrt{n}(|\alpha^T\mathbf{q}(Z^n, X^n, c_{-k}^k)|)$ is uniformly integrable and therefore

$$\lim_{n\to\infty} E_{\mathbf{P}}\Big[\sqrt{n}\Big(\Big|\alpha^T\mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\Big|\Big)\Big] = E_{\mathbf{P}}[|G|] = \sqrt{\frac{2V}{\pi}}. \qquad \square$$

Using Lemma 5 and arguments similar to those in the proof of Theorem 2 we prove the following.

Theorem 6. For any neutralizable pair $(\mathbf{\Pi}, \Lambda)$, and any sequence $\{\hat{X}^n \in \mathcal{D}_n\}$ of denoisers, as n tends to infinity

$$\hat{R}_k\Big(\hat{X}^n\Big) \geq \frac{c}{\sqrt{n}}\left(\sum_{a\in\mathcal{A}}\sqrt{(\mathbf{P}^*)^T\pi_a}\right)^{2k}(1+o(1))$$

where $\mathbf{P}^*$ is any loss-neutral distribution and c is a positive function of $(\mathbf{\Pi}, \Lambda)$ and $\mathbf{P}^*$.

Proof Let $t, i, j \in \mathcal{A}$ and let the distribution $\mathbf{P}^* \in \mathcal{M}$ be loss-neutral with respect to $(\pi_t, \lambda_i, \lambda_j)$, so that $\mathbf{P}^* \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$ and

$$(\mathbf{P}^*)^T(\lambda_i \odot \pi_t) = (\mathbf{P}^*)^T(\lambda_j \odot \pi_t) = \min_{k\in\mathcal{A}}(\mathbf{P}^*)^T(\lambda_k \odot \pi_t). \tag{20}$$

As argued in the proof of Theorem 2

$$\hat{R}_k\Big(\hat{X}^n\Big) \geq E_{\mathbf{P}^*}\Big[E\big[L_{\hat{X}^n}(X_1^n, Z^n)\big] - \hat{D}_k(X^n)\Big], \tag{21}$$

and since X^n is generated *i.i.d.* according to $\mathbf{P}^*$ it follows from Lemma 1 that for all $\hat{X}^n \in \mathcal{D}_n$

$$E_{\mathbf{P}^*}\big[E\big[L_{\hat{X}^n}(X_1^n, Z^n)\big]\big] \geq D_{\text{opt}} = \sum_{z\in\mathcal{A}}\min_{\hat{x}\in\mathcal{A}}\lambda_{\hat{x}}^T(\mathbf{P}^* \odot \pi_z). \tag{22}$$

We upper bound the second term in (21), namely, $E_{\mathbf{P}^*}\Big[\hat{D}_k(X^n)\Big]$. Applying (17), the fact that the expectation of the minimum of a collection of random variables is lesser than the minimum of the expectations, and (16)

$$\begin{aligned}
E_{\mathbf{P}^*}\Big[\hat{D}_k(X^n)\Big] =& E_{\mathbf{P}^*}[D_k(X^n, Z^n)] \\
=& E_{\mathbf{P}^*}\left[\sum_{c_{-k}^k\in\mathcal{A}^{2k+1}}\min_{\hat{x}\in\mathcal{A}}\lambda_{\hat{x}}^T\mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right] \\
=& \sum_{c_{-k}^k\in\mathcal{A}^{2k+1}, c_0=t} E_{\mathbf{P}^*}\left[\min_{\hat{x}\in\mathcal{A}}\lambda_{\hat{x}}^T\mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right] \\
&+ \sum_{c_{-k}^k\in\mathcal{A}^{2k+1}, c_0\neq t} E_{\mathbf{P}^*}\left[\min_{\hat{x}\in\mathcal{A}}\lambda_{\hat{x}}^T\mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} E_{\mathbf{P}^*}\left[\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right] \\
&\quad + \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 \neq t} \min_{\hat{x}\in\mathcal{A}} E_{\mathbf{P}^*}\left[\lambda_{\hat{x}}^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right] \\
&= \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} E_{\mathbf{P}^*}\left[\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right] \\
&\quad + \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 \neq t} \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T (\mathbf{P}^* \odot \pi_{c_0}) \prod_{i=-k, i\neq 0}^{k} (\mathbf{P}^*)^T \pi_{c_i} \\
&= \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} E_{\mathbf{P}^*}\left[\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right] + \sum_{c_0 \neq t} \min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T (\mathbf{P}^* \odot \pi_{c_0}). \qquad (23)
\end{aligned}
$$

The last equality holds as

$$
\sum_{c_{-k}^{-1} \in \mathcal{A}^k, c_1^k \in \mathcal{A}^k} \prod_{i=-k, i\neq 0}^{k} (\mathbf{P}^*)^T \pi_{c_i} = 1.
$$

Substituting (23) and (22) in (21), combining with (20), and observing that the minimum over all $\hat{x} \in \mathcal{A}$ is less than the minimum over the set $\{i, j\} \subseteq \mathcal{A}$, we obtain for all $\hat{X}^n \in \mathcal{D}_n$

$$
\hat{R}_k\Big(\hat{X}^n\Big) \geq \lambda_i^T (\mathbf{P}^* \odot \pi_t) - \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} E_{\mathbf{P}^*}\left[\min\left\{\lambda_i^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big), \lambda_j^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right\}\right]. \quad (24)
$$

As in the proof of Theorem 2 we write $2\min\{x, y\} = (x + y - |x - y|)$, and note that the expectations of $\lambda_i^T \mathbf{q}(Z^n, X^n, c_{-k}^k)$ and $\lambda_j^T \mathbf{q}(Z^n, X^n, c_{-k}^k)$ are equal to obtain

$$
\begin{aligned}
&\sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} E_{\mathbf{P}^*}\left[\min\left\{\lambda_i^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big), \lambda_j^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right\}\right] \\
&\qquad = \lambda_i^T (\mathbf{P}^* \odot \pi_t) - \frac{1}{2} \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} E_{\mathbf{P}^*}\left[\left|(\lambda_i - \lambda_j)^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right|\right].
\end{aligned}
$$

Substituting this in (24)

$$
\hat{R}_k\Big(\hat{X}^n\Big) \geq \frac{1}{2} \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} E_{\mathbf{P}^*}\left[\left|(\lambda_i - \lambda_j)^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right|\right].
$$

From (20), $(\lambda_i - \lambda_j)^T (\mathbf{P}^* \odot \pi_t) = 0$ and therefore applying Lemma 5 for each $c_{-k}^k \in \mathcal{A}^{2k+1}$ with $c_0 = t$, and choosing $\mathbf{P} = \mathbf{P}^*$ and $\alpha = \lambda_i - \lambda_j$ we obtain

$$
\lim_{n\to\infty} \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} \sqrt{n} E_{\mathbf{P}^*}\left[\left|(\lambda_i - \lambda_j)^T \mathbf{q}\Big(Z^n, X^n, c_{-k}^k\Big)\right|\right] = \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} \sqrt{\frac{2V_{c_{-k}^k}}{\pi}}
$$

where $V_{c_{-k}^k} = ((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T (\mathbf{P}^* \odot \pi_t) \prod_{i=-k, i\neq 0}^{k} (\mathbf{P}^*)^T \pi_{c_i}$. Observe that

$$
\sum_{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t} \sqrt{\frac{2V_{c_{-k}^k}}{\pi}} = \sqrt{\frac{2}{\pi}} \sqrt{((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T (\mathbf{P}^* \odot \pi_t)} \sum_{a_1^{2k} \in \mathcal{A}^{2k}} \left(\prod_{i=1}^{2k} (\mathbf{P}^*)^T \pi_{a_i}\right)^{\frac{1}{2}}.
$$

Note that since $\mathbf{P}^*$ is a loss-neutral distribution $(\lambda_i - \lambda_j) \odot \mathbf{P}^* \odot \pi_t \neq 0$, and therefore

$$((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T (\mathbf{P} \odot \pi_t) > 0.$$

Also observe that

$$\sum_{a_1^{2k} \in \mathcal{A}^{2k}} \left(\prod_{i=1}^{2k} (\mathbf{P}^*)^T \pi_{a_i} \right)^{\frac{1}{2}} = \left(\sum_{a \in \mathcal{A}} \sqrt{(\mathbf{P}^*)^T \pi_a} \right)^{2k}. \qquad \square$$

A vector of dimension greater than 1 is *degenerate* if at most one of its components is non-zero. Note that

$$\sum_{a \in \mathcal{A}} \sqrt{(\mathbf{P}^*)^T \pi_a} \geq \sum_{a \in \mathcal{A}} (\mathbf{P}^*)^T \pi_a = 1$$

with equality iff $(\mathbf{P}^*)^T \mathbf{\Pi}$ is degenerate. Thus, if $(\mathbf{P}^*)^T \mathbf{\Pi}$ is non-degenerate, the lower bound in the statement of the theorem grows exponentially in k. For many $(\mathbf{\Pi}, \Lambda)$, *e.g.,* BSC and Hamming loss, the lower bound grows exponentially in k.

4 Compound Decision

The compound decision problem was first proposed by Robbins [2]. In this problem a sequence of n hypothesis tests each involving M possible hypothesis are to be solved simultaneously. As pointed out in [1], this is precisely the problem of denoising a length-n sequence over an alphabet of size M that has been transmitted over a memoryless channel $\mathbf{\Pi}$. The M distributions in the hypothesis testing problem correspond to the M rows of $\mathbf{\Pi}$. Robbins measures the performance of any scheme against a "symbol-by-symbol" decision rule that is aware of the true hypotheses. In the denoiser setting this corresponds to the best 0-th order denoiser for a given individual noiseless sequence. The loss of such a denoiser for a given sequence x^n is precisely $\bar{D}_0(x^n)$. Therefore the corresponding regret of any other denoiser $\hat{X}^n$ is $\bar{R}_0\left(\hat{X}^n\right)$.

Hannan and Van Ryzin [3] derived a scheme for the compond decision problem whose regret decreases like $\mathcal{O}(1/\sqrt{n})$. Furthermore, for certain types of hypothesis tests which, in the denoising setting, correspond to channels with continuous output, they showed that the regret decreases even faster - $\mathcal{O}(1/n)$. The need for a more stringent benchmark for these schemes was recognized by Johns [4] who considered sliding window denoisers and the corresponding benchmark $\bar{D}_k(x^n)$. In the denoising setting the regret $\hat{R}_k\left(\hat{X}^{n,k}_{\text{univ}}\right)$ of the DUDE [1] was upper bounded by $c^k/\sqrt{n}$ where $c > 1$ is a constant. It follows from (3) that this upper bound applies to $\bar{R}_k\left(\hat{X}^{n,k}_{\text{univ}}\right)$ as well.

We derive lowerbounds on $\bar{R}_k\left(\hat{X}^n\right)$, for all denoisers and all discrete channels, that scale like $\frac{c^k}{\sqrt{n}}$. This shows that the upper bounds derived in [3] for 0-th order regret and, for fixed k, those implied by [1] for the k-th order regret are tight up to a constant factor. This also shows that the rate of convergence result in [3] for continuous output channels does not extend to discrete memoryless channels. The proof for our lowerbound is along the lines of that for Theorems 2 and 6. However, unlike these proofs, to lower bound $\bar{R}_k\left(\hat{X}^n\right)$, we restrict the comparison of $\hat{X}^n$ to the class of one-sided k-th order sliding window denoisers. The following preliminaries are required.

A *k-th order one-sided sliding window denoiser* $\hat{X}^n$ is a denoiser with the property that for all $z^n \in \mathcal{A}^n$, if $z_{i-k}^i = z_{j-k}^j$ then

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

Thus the denoiser defines a mapping, $f : \mathcal{A}^{k+1} \to \mathcal{A}$ so that for all $z^n \in \mathcal{A}^n$

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

Let $\mathcal{S}_k^1$ denote the class of k-th order one-sided sliding window denoisers. For an individual noiseless sequence $x^n \in \mathcal{A}^n$, let

$$\bar{D}_k^1(x^n) \stackrel{\text{def}}{=} \min_{\hat{X}^n \in \mathcal{S}_k^1} E\Big[L_{\hat{X}^n}\Big(x_{k+1}^{n-k}, Z^n\Big)\Big]$$

denote the minimum expected loss incurred by any k-th order sliding window denoiser when the noiseless sequence is x^n. Clearly $\mathcal{S}_k^1 \subset \mathcal{S}_k$ and therefore for all x^n,

$$\bar{D}_k(x^n) \le \bar{D}_k^1(x^n). \tag{25}$$

For $x^n \in \mathcal{A}^n$, and $c_{-k}^0 \in \mathcal{A}^{k+1}$, let $\bar{\mathbf{q}}^1\big(x^n, c_{-k}^0\big)$ denote the M-dimensional column vector whose j-th component, $j \in \mathcal{A}$, is

$$\begin{aligned}\bar{\mathbf{q}}^1\big(x^n, c_{-k}^0\big)[j] &= \frac{1}{n-2k} E_{\mathbf{P}(Z^n|x^n)}\big[\big|\big\{i : k+1 \le i \le n-k, Z_{i-k}^i = c_{-k}^0, x_i = j\big\}\big|\big] \\ &= \frac{1}{n-2k} \sum_{i=k+1}^{n-k} 1(x_i = j) \prod_{j=-k}^{0} \mathbf{\Pi}(X_{i+j}, c_j)\end{aligned}$$

the expected frequency of occurrence of the sequence c_{-k}^0 in Z^n along with j in x^n at the location corresponding to c_0 in Z^n, when x^n is transmitted over a discrete memoryless channel $\mathbf{\Pi}$. Also note that if X^n is drawn *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$, then Z^n is also an *i.i.d.* sequence and

$$E\big[\bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)\big] = (\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k,}^{-1} \mathbf{P}^T \pi_{c_i}. \tag{26}$$

We express the minimum expected loss $\bar{D}_k^1(x^n)$ incurred by any k-th order one-sided sliding window denoiser when the noiseless sequence is x^n, in terms of the vectors $\bar{\mathbf{q}}^1\big(x^n, c_{-k}^0\big)$, $c_{-k}^0 \in \mathcal{A}^{k+1}$. Observe that for all $x^n \in \mathcal{A}^n$

$$\bar{D}_k^1(x^n) = \sum_{c_{-k}^0 \in \mathcal{A}^{k+1}} \min_{\hat{x} \in \mathcal{A}} \sum_{j \in \mathcal{A}} \Lambda(j, \hat{x}) \bar{\mathbf{q}}^1\big(x^n, c_{-k}^0\big)[j] = \sum_{c_{-k}^0 \in \mathcal{A}^{k+1}} \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \bar{\mathbf{q}}^1\big(x^n, c_{-k}^0\big). \tag{27}$$

To derive a lowerbound on $\bar{R}_k\Big(\hat{X}^n\Big)$ we require the following lemma on the asymptotics of $\bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)$. As in the analogous result for $\mathbf{q}\big(Z^n, X^n, c_{-k}^k\big)$ we invoke Theorem 4.

Lemma 7. If X^n is generated *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$, then for any column vector $\alpha \in \mathbb{R}^M$, and any $c_{-k}^0 \in \mathcal{A}^{k+1}$, such that $\alpha^T(\mathbf{P} \odot \pi_{c_0}) = 0$

$$\lim_{n \to \infty} E_{\mathbf{P}}\big[\sqrt{n}\,\big|\alpha^T \bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)\big|\big] = \sqrt{\frac{2V}{\pi}}$$

where $V = (\alpha \odot \alpha)^T (\mathbf{P} \odot \pi_{c_0} \odot \pi_{c_0}) \prod_{i=-k}^{-1} \mathbf{P}^T (\pi_{c_i} \odot \pi_{c_i})$.

Proof We first show that when X^n is generated *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$, for any column vector $\alpha \in \mathbb{R}^M$, and any $c_{-k}^0 \in \mathcal{A}^{k+1}$ satisfying $\alpha^T(\mathbf{P} \odot \pi_{c_0}) = 0$, as n converges to infinity, $\alpha^T \bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)$ suitably normalized tends in distribution to a Gaussian random variable, namely,

$$\sqrt{n}\big(\alpha^T \bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V) \tag{28}$$

where $V = (\alpha \odot \alpha)^T (\mathbf{P} \odot \pi_{c_0} \odot \pi_{c_0}) \prod_{i=-k}^{-1} \mathbf{P}^T (\pi_{c_i} \odot \pi_{c_i})$.

For a given $\alpha \in \mathbb{R}^M$ and $c_{-k}^0 \in \mathcal{A}^{k+1}$ we define the sequence Y_{k+1}^{n-k} of random variables as

$$\begin{aligned} Y_i &\stackrel{\text{def}}{=} \sum_{\ell \in \mathcal{A}} \alpha(\ell) 1(X_i = \ell) Pr\big(Z_{i-k}^i = c_{-k}^0 | X_{i-k}^i\big) \\ &= \sum_{\ell \in \mathcal{A}} \alpha(\ell) 1(X_i = \ell) \prod_{j=-k}^{0} \mathbf{\Pi}(X_{i+j}, c_j), \qquad k+1 \le i \le n-k. \end{aligned}$$

Then

$$\frac{1}{n-2k} \sum_{i=k+1}^{n-k} Y_i = \alpha^T \bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big).$$

If the sequence X^n is drawn *i.i.d.* according to $\mathbf{P}$, Y_{k+1}^{n-k} is stationary and since each Y_i is a function of $k+1$ consecutive X_i's, it is easy to verify that Y_{k+1}^{n-k} is a k-dependent sequence. Furthermore

$$E_{\mathbf{P}}[Y_i] = \sum_{\ell \in \mathcal{A}} \alpha(\ell) Pr\big(X_i = \ell, Z_{i-k}^i = c_{-k}^0\big) = \alpha^T (\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k}^{-1} \mathbf{P}^T \pi_{c_i} = 0,$$

where the last equality follows from the choice of α and c_0. Hence Theorem 4 applies and therefore

$$\sqrt{n}\big(\alpha^T \bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)\big) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, E\big[Y_{k+1}^2\big] + 2\sum_{i=1}^{k} E[Y_{k+1} Y_{k+1+i}]\right). \tag{29}$$

Observe that

$$\begin{aligned} E\big[Y_{k+1}^2\big] &= \sum_{\ell \in \mathcal{A}} \alpha(\ell)^2 E[1(X_{k+1} = \ell)] \mathbf{\Pi}(\ell, c_0)^2 \prod_{i=-k}^{-1} E[\mathbf{\Pi}(X_{i+j}, c_j)]^2 \\ &= (\alpha \odot \alpha)^T (\mathbf{P} \odot \pi_{c_0} \odot \pi_{c_0}) \prod_{i=-k}^{-1} \mathbf{P}^T (\pi_{c_i} \odot \pi_{c_i}) \end{aligned}$$

and we will show that for all $i \ge 1$

$$E[Y_{k+1} Y_{k+1+i}] = 0.$$

Substituting these in (29) establishes (28). Observe that

$$\begin{aligned} Y_{k+1} Y_{k+1+i} = &\sum_{\ell \in \mathcal{A}} \alpha(\ell) 1(X_{k+1} = \ell) \mathbf{\Pi}(\ell, c_0) \mathbf{\Pi}(\ell, c_{-i}) \sum_{m \in \mathcal{A}} \alpha(m) 1(X_{k+1+i} = m) \mathbf{\Pi}(m, c_0) \\ &\prod_{j=-k}^{-k+i-1} \mathbf{\Pi}(X_{k+1+j}, c_j) \prod_{j=-k+i}^{-1} \mathbf{\Pi}(X_{k+1+j}, c_j) \mathbf{\Pi}(X_{k+1+j}, c_{j-i}) \prod_{j=1}^{i-1} \mathbf{\Pi}(X_{k+1+j}, c_{j-i}). \end{aligned}$$

Since X^n is drawn *i.i.d.* according to $\mathbf{P}$, for $i \geq 1$

$$\begin{aligned} E[Y_{k+1}Y_{k+1+i}] =& \big(\alpha^T\big(\mathbf{P} \odot \pi_{c_0} \odot \pi_{c_{-i}}\big)\big)\big(\alpha^T(\mathbf{P} \odot \pi_{c_0})\big) \prod_{j=-k}^{-k+i-1} E[\mathbf{\Pi}(X_{k+1+j}, c_j)] \\ & \prod_{j=-k+i}^{-1} E[\mathbf{\Pi}(X_{k+1+j}, c_j)\mathbf{\Pi}(X_{k+1+j}, c_{j-i})] \prod_{j=1}^{i-1} E[\mathbf{\Pi}(X_{k+1+j}, c_{j-i})]. \\ =& \, 0 \end{aligned}$$

where the last equality follows from the fact that $\alpha^T(\mathbf{P} \odot \pi_{c_0}) = 0$.

A straightforward extension of (28) yields

$$\sqrt{n}\big(\big|\alpha^T \bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)\big|\big) \xrightarrow{\mathcal{L}} |G|$$

where $G \sim \mathcal{N}(0, V)$. Following arguments similar to the ones in Lemmas 3 and 5 we obtain

$$\lim_{n\to\infty} E_{\mathbf{P}}\big[\sqrt{n}\big(\big|\alpha^T \bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)\big|\big)\big] = E_{\mathbf{P}}[|G|] = \sqrt{\frac{2V}{\pi}}. \qquad \square$$

Using Lemma 7 and arguments similar to those in the proof of Theorem 6 we prove the following theorem.

Theorem 8. For any neutralizable pair $(\mathbf{\Pi}, \Lambda)$, and any sequence $\{\hat{X}^n \in \mathcal{D}_n\}$ of denoisers, as n tends to infinity

$$\bar{R}_k\Big(\hat{X}^n\Big) \geq \frac{c}{\sqrt{n}}\left(\sum_{a\in\mathcal{A}} \sqrt{(\mathbf{P}^*)^T(\pi_a \odot \pi_a)}\right)^k (1 + o(1))$$

where $\mathbf{P}^*$ is any loss-neutral distribution and c is a positive function of $(\mathbf{\Pi}, \Lambda)$ and $\mathbf{P}^*$.

Proof Let $t, i, j \in \mathcal{A}$ and let the distribution $\mathbf{P}^* \in \mathcal{M}$ be loss-neutral with respect to $(\pi_t, \lambda_i, \lambda_j)$, so that $\mathbf{P}^* \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$ and

$$(\mathbf{P}^*)^T(\lambda_i \odot \pi_t) = (\mathbf{P}^*)^T(\lambda_j \odot \pi_t) = \min_{k\in\mathcal{A}} (\mathbf{P}^*)^T(\lambda_k \odot \pi_t). \tag{30}$$

As argued in the proofs of Theorems 2 and 6

$$\begin{aligned} \bar{R}_k\Big(\hat{X}^n\Big) &\geq E_{\mathbf{P}^*}\big[E\big[L_{\hat{X}^n}(X_1^n, Z^n)\big] - \bar{D}_k(X^n)\big], \\ &\geq E_{\mathbf{P}^*}\big[E\big[L_{\hat{X}^n}(X_1^n, Z^n)\big] - \bar{D}_k^1(X^n)\big] \end{aligned} \tag{31}$$

where we have used (25).

Following the proof of Theorem 6, with $\hat{R}_k$ replaced by $\bar{R}_k$ and $\mathbf{q}(\cdot)$ replaced by $\bar{\mathbf{q}}^1(\cdot)$ we obtain

$$\bar{R}_k\Big(\hat{X}^n\Big) \geq \frac{1}{2} \sum_{c_{-k}^0 \in \mathcal{A}^{k+1}, c_0 = t} E_{\mathbf{P}^*}\Big[\Big|(\lambda_i - \lambda_j)^T \bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)\Big|\Big].$$

From (30), $(\lambda_i - \lambda_j)^T(\mathbf{P}^* \odot \pi_t) = 0$ and therefore applying Lemma 7 for each $c_{-k}^0 \in \mathcal{A}^{k+1}$ with $c_0 = t$, and choosing $\mathbf{P} = \mathbf{P}^*$ and $\alpha = \lambda_i - \lambda_j$ we obtain

$$\lim_{n\to\infty} \sum_{c_{-k}^0 \in \mathcal{A}^{k+1}, c_0 = t} \sqrt{n} E_{\mathbf{P}^*}\Big[\Big|(\lambda_i - \lambda_j)^T \bar{\mathbf{q}}^1\big(X^n, c_{-k}^0\big)\Big|\Big] = \sum_{c_{-k}^0 \in \mathcal{A}^{k+1}, c_0 = t} \sqrt{\frac{2V_{c_{-k}^0}}{\pi}}$$

where $V_{c_{-k}^0} = ((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T (\mathbf{P}^* \odot \pi_{c_0} \odot \pi_{c0}) \prod_{i=-k}^{-1} (\mathbf{P}^*)^T (\pi_{c_i} \odot \pi_{c_i})$.

Note that since $\mathbf{P}^*$ is a loss-neutral distribution $\mathbf{P}^* \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$ and therefore

$$((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T (\mathbf{P}^* \odot \pi_t \odot \pi_t) > 0.$$

Now, observe that

$$\sum_{\substack{c_{-k}^0 \in \mathcal{A}^{k+1}, \\ c_0 = t}} \sqrt{\frac{2V_{c_{-k}^0}}{\pi}} = \sqrt{\frac{2}{\pi}} \sqrt{((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T (\mathbf{P}^* \odot \pi_t \odot \pi_t)} \sum_{a_1^k \in \mathcal{A}^k} \left(\prod_{i=1}^{k} (\mathbf{P}^*)^T (\pi_{a_i} \odot \pi_{a_i}) \right)^{\frac{1}{2}}.$$

It can be easily verified that

$$\sum_{a_1^k \in \mathcal{A}^k} \left(\prod_{i=1}^{k} (\mathbf{P}^*)^T (\pi_{a_i} \odot \pi_{a_i}) \right)^{\frac{1}{2}} = \left(\sum_{a \in \mathcal{A}} \sqrt{(\mathbf{P}^*)^T (\pi_a \odot \pi_a)} \right)^k. \qquad \square$$

By Jensen's inequality

$$\sum_{a \in \mathcal{A}} \sqrt{(\mathbf{P}^*)^T (\pi_a \odot \pi_a)} \geq \sum_{a \in \mathcal{A}} (\mathbf{P}^*)^T \pi_a = 1$$

with equality only if for all $a \in \mathcal{A}$, all the non-zero entries of $\mathbf{P}^* \odot \pi_a$ are equal. Note that if they are unequal for some a, the lower bound in the statement of the theorem grows exponentially in k, and this is indeed the case for many $(\mathbf{\Pi}, \Lambda)$, *e.g.,* BSC and Hamming loss. However, using $\bar{D}_k^1$ instead of $\bar{D}_k$ in the proof cost us a factor of two in the exponent.

Instead of characterizing $\bar{D}_k^1(\hat{X}^n)$ we could have directly tried to characterize $\bar{D}_k(\hat{X}^n)$. In that case $\bar{\mathbf{q}}^1(\cdot)$ would have to be replaced by $\bar{\mathbf{q}}(\cdot)$ where for $x^n \in \mathcal{A}^n$, and $c_{-k}^k \in \mathcal{A}^{2k+1}$

$$\bar{\mathbf{q}}\Big(x^n, c_{-k}^k\Big)[j] = \frac{1}{n-2k} E_{\mathbf{P}(Z^n|x^n)} \Big[\Big| \Big\{ i : k+1 \leq i \leq n-k, Z_{i-k}^{i+k} = c_{-k}^k, x_i = j \Big\} \Big| \Big].$$

This can be written as a sum of Y_i's where

$$Y_i \stackrel{\text{def}}{=} \sum_{\ell \in \mathcal{A}} \alpha(\ell) 1(X_i = \ell) \Pi_{j=-k}^{k} \mathbf{\Pi}(X_{i+j}, c_j), \qquad k+1 \leq i \leq n-k.$$

This sequence is $2k$-dependent and therefore Theorem 4 applies. However, unlike in Lemmas 5 and 7, the Y_i's are not uncorrelated and therefore the expression for the asymptotic variance of $\sqrt{n}(\alpha^T \bar{\mathbf{q}}\big(X^n, c_{-k}^k\big))$ turns out to be cumbersome to handle in general. For $(\mathbf{\Pi}_{\text{BSC}}, \Lambda_{\text{Ham}})$, however, it can be shown that

$$\bar{R}_k\Big(\hat{X}^n\Big) \geq \frac{c}{\sqrt{n}} \left(\sum_{a \in \mathcal{A}} \sqrt{(\mathbf{P}^*)^T (\pi_a \odot \pi_a)} \right)^{2k} (1 + o(1))$$

where c is different from Theorem 8. We conjecture that this is true for all neutralizable $(\mathbf{\Pi}, \mathbf{\Lambda})$.

5 Discussion

We derived lower bounds for $\hat{R}_k(\hat{X}^n)$ and $\bar{R}_k(\hat{X}^n)$. These results imply that for all $\hat{X}^n \in \mathcal{D}_n$ there exists at least one individual noiseless sequence x^n for which the excess loss when compared to the best k-th order sliding window denoiser can be lower bounded by $\Omega(c^k/\sqrt{n})$. But, from the proofs, it is clear that this result can be strengthened slightly to apply to not just the worst-case sequence but also when averaging over noiseless sequences X^n when they are drawn according to a loss-neutral distribution.

Theorem 6 applied to the case where k was fixed and n tended to infinity. This result can be strengthened by deriving lower bounds for $\hat{R}_k(\hat{X}^n)$ when both k and n tend to infinity. For this purpose we require a result analogous to Lemma 5 that applies when k grows with n. To that end we prove the following lemma in the Appendix.

Lemma 9. If X^n is generated *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$, then for any column vector $\alpha \in \mathbb{R}^M$, and any sequence of integers $\{k_n\}$ and contexts $\{c_{-k_n}^{k_n}\} \in \mathcal{A}^{2k_n+1}$, such that $n > 2k_n$ and $\alpha^T(\mathbf{P} \odot \pi_{c_0}) = 0$, as n tends to infinity, if $k_n = o(\ln n)$

$$E_{\mathbf{P}}\left[\sqrt{n}\left|\alpha^T \mathbf{q}\Big(Z^n, X^n, c_{-k_n}^{k_n}\Big)\right|\right] = \sqrt{\frac{2V_n}{\pi}}(1+o(1))$$

where $V_n = (\alpha \odot \alpha)^T(\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k_n, i\neq 0}^{k_n} \mathbf{P}^T \pi_{c_i}$. □

Following the steps in the proof of Theorem 6 with Lemma 5 replaced by the above lemma we obtain the following.

Theorem 10. For any neutralizable pair $(\mathbf{\Pi}, \Lambda)$, any sequence $\{\hat{X}^n \in \mathcal{D}_n\}$ of denoisers, and any sequence $\{k_n\}$, as n tends to infinity, if $k_n = o(\ln n)$

$$\hat{R}_{k_n}\Big(\hat{X}^n\Big) \geq \frac{c}{\sqrt{n}}\left(\sum_{a\in\mathcal{A}} \sqrt{(\mathbf{P}^*)^T \pi_a}\right)^{2k_n} (1+o(1))$$

where $\mathbf{P}^*$ is any loss-neutral distribution and c is a positive function of $(\mathbf{\Pi}, \Lambda)$ and $\mathbf{P}^*$. □

It was shown in [1] that

$$\hat{R}_{k_n}\Big(\hat{X}^{n,k}_{\text{univ}}\Big) = \mathcal{O}\left(\sqrt{\frac{k_n M^{2k_n}}{n}}\right).$$

We pointed out in the introduction that this, in conjunction with Theorem 6, implies that asymptotically, for fixed k the regret of the DUDE is within a constant factor of the best possible. In light of the above theorem, a more general statement can be made for some $(\mathbf{\Pi}, \Lambda)$ pairs. If a loss-neutral distribution $\mathbf{P}^*$ induces a uniform distribution at the output of $\mathbf{\Pi}$, *i.e.,* the distribution $(\mathbf{P}^*)^T\mathbf{\Pi}$ is uniform, then the above theorem reduces to

$$\hat{R}_{k_n}\Big(\hat{X}^n\Big) \geq \frac{c}{\sqrt{n}} M^{k_n}(1+o(1)).$$

In this case asymptotically, when $k_n = o(\ln n)$, the regret of the DUDE is within a factor of $\sqrt{k_n}$ of the best possible. (We believe that the upper bound on the DUDE can be further strengthened to drop the $\sqrt{k_n}$ term, by employing results for k-dependent processes.) It is possible to construct $(\mathbf{\Pi}, \Lambda)$ pairs for which $(\mathbf{P}^*)^T\mathbf{\Pi}$ is uniform. For example, consider a Z-channel with

$$\mathbf{\Pi} = \begin{bmatrix} 1 & 0 \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

and Hamming loss, *i.e.,* $\Lambda = \Lambda_{\text{Ham}}$. Then it can be easily verified that $\mathbf{P}^* = [\frac{1}{4} \quad \frac{3}{4}]^T$ is a loss-neutral distribution and $(\mathbf{P}^*)^T\mathbf{\Pi} = [\frac{1}{2} \quad \frac{1}{2}]$.

Further strengthening Theorems 6 and 8 to obtain a Rissanen-style [9, 10] lower bound that applies to "most" individual sequences is not possible. This situation is similar to the one encountered in prediction problems [7]. To observe this, consider a binary symmetric channel with crossover probability δ and the Hamming loss function. If the Hamming weight of x^n, the noiseless sequence, is between δ and $1-\delta$, then with high probability the best 0-th order sliding window denoiser for the pair (x^n, Z^n) where Z^n is the noisy output of the channel when x^n is the input, is the "say-what-you-see" denoiser. For this denoiser $\hat{X}^n$ and a non-zero fraction of types of noiseless sequences, $E\big[L_{\hat{X}^n}(x^n, Z^n)\big] - \hat{D}_k(x^n)$ decays exponentially in n and $E\big[L_{\hat{X}^n}(x^n, Z^n)\big] - \bar{D}_k(x^n)$ is zero. Hence the lower bound we derived cannot apply to most individual sequences.

6 Extensions

In this section we extend the results to more general classes of sliding window denoisers as well as to multi-dimensionally indexed data. The results obtained here are of a similar nature.

6.1 Sliding window denoisers with arbitrary context

In Sections 3 and 4 we compared a given denoiser to the best loss obtainable with the class of k-th order sliding window denoisers. This comparison class can be generalized to sliding window denoisers where each symbol z_i is denoised based on its context where the context, unlike in the sliding window denoiser case, does not have to be all the k symbols to the left and right of z_i. In this section we derive lower bounds on the excess loss incurred by any denoiser when compared to the best sliding window denoiser that is based on a given context set. We begin with some preliminary definitions.

Bi-directional contexts have been defined and considered before in [11]. We present the definitions here for completeness. For a given alphabet $\mathcal{A}$, let $\mathcal{A}^*$ denote the set of all finite length strings over $\mathcal{A}$. For any string $x \in \mathcal{A}^*$ let $|x|$ denote the length of x and for any positive integer ℓ, let x^ℓ denote the ℓ-symbol prefix of x. Consider a finite collection $\mathcal{C} \subseteq \mathcal{A}^* \times \mathcal{A}^*$ of ordered pairs. Let k be the maximum length of any string in any of the ordered pairs in $\mathcal{C}$. The prefix set $\mathcal{P}(s)$ of any $s = (s_1, s_2) \in \mathcal{C}$ is given by

$$\mathcal{P}(s) = \Big\{(x, y) \in \mathcal{A}^k \times \mathcal{A}^k : x^{|s_1|} = s_1, y^{|s_2|} = s_2\Big\},$$

the set of all ordered pairs of length-k strings whose prefixes are s_1 and s_2. The set $\mathcal{C}$ is exhaustive if

$$\bigcup_{s\in\mathcal{C}} \mathcal{P}(s) = \mathcal{A}^k \times \mathcal{A}^k$$

and is disjoint if for all $s \neq s' \in \mathcal{C}$

$$\mathcal{P}(s) \cap \mathcal{P}(s') = \Phi.$$

A collection $\mathcal{C} \subseteq \mathcal{A}^* \times \mathcal{A}^*$ is a context set iff it is exhaustive and disjoint.

Based on the given bidirectional context set one can define a class of context-based sliding window denoisers. Formally let $\mathcal{C} \subseteq \mathcal{A}^* \times \mathcal{A}^*$ be a bidirectional context set and k the maximum length of any string in any of the ordered pairs in $\mathcal{C}$. Given a sequence z^n, the *left context* z_i^ℓ of the symbol z_i is the sequence $(z_{i-1}, z_{i-2}, \ldots)$ and the *right context* z_i^r of z_i is the sequence $(z_{i+1}, z_{i+2}, \ldots)$. Then z_i is *associated* with a context pair $(s_1, s_2) \in \mathcal{C}$ if

$$(z_i^\ell)^{|s_1|} = s_1 \text{ and } (z_i^r)^{|s_2|} = s_2.$$

The definition of $\mathcal{C}$ guarantees that for $k+1 \leq i \leq n-k$ every z_i is associated with exactly one context pair. A *$\mathcal{C}$-based sliding window denoiser* $\hat{X}^n$ is a denoiser with the property that for all $z^n \in \mathcal{A}^n$, if both z_i and z_j are associated with $(s_1, s_2) \in \mathcal{C}$ then

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

Thus the denoiser defines a mapping,

$$f : \mathcal{C} \to \mathcal{A}$$

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

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

where $s \in \mathcal{C}$ is the unique context associated with z_i. Let $\mathcal{S}_\mathcal{C}$ denote the class of $\mathcal{C}$-based sliding window denoisers. As in the case of k-th order sliding window denoisers we can define the best loss obtainable for a given pair of noiseless and noisy sequences with a $\mathcal{C}$-based sliding window denoiser.

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 $n > 2k$, $D_\mathcal{C}(x^n, z^n)$, the *$\mathcal{C}$-based minimum loss* of (x^n, z^n) is defined to be

$$D_\mathcal{C}(x^n, z^n) = \min_{\hat{X}^n \in \mathcal{S}_\mathcal{C}} L_{\hat{X}^n}\left(x_{k+1}^{n-k}, z^n\right)$$

and for a given channel $\mathbf{\Pi}$ and a noiseless sequence x^n define the expected $\mathcal{C}$-based minimum loss to be

$$\hat{D}_\mathcal{C}(x^n) \stackrel{\text{def}}{=} E[D_\mathcal{C}(x^n, Z^n)].$$

As in the case of k-th order sliding window denoisers, the regret of any denoiser $\hat{X}^n \in \mathcal{D}_n$ is defined to be

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

Since the maximum length of any string in any ordered pair in $\mathcal{C}$ is k, $\mathcal{S}_{\mathcal{C}} \subseteq \mathcal{S}_k$ and therefore for all $x^n \in \mathcal{A}^n$

$$\hat{D}_{\mathcal{C}}(x^n) \geq \hat{D}_k(x^n)$$

and hence for all $\hat{X}^n \in \mathcal{D}_n$

$$\hat{R}_{\mathcal{C}}\Big(\hat{X}^n\Big) \leq \hat{R}_k\Big(\hat{X}^n\Big).$$

In the following theorem we state a lower bound on $\hat{R}_{\mathcal{C}}\Big(\hat{X}^n\Big)$. We omit the proof since it is identical to that of Theorem 6.

Theorem 11. For any neutralizable pair $(\mathbf{\Pi}, \Lambda)$, and any sequence $\{\hat{X}^n \in \mathcal{D}_n\}$ of denoisers, as n tends to infinity

$$\hat{R}_k\Big(\hat{X}^n\Big) \geq \frac{c}{\sqrt{n}} \sum_{(\alpha,\beta)\in\mathcal{C}} \left(\left(\prod_{i=1}^{|\alpha|} (\mathbf{P}^*)^T \pi_{\alpha_i} \right) \left(\prod_{i=1}^{|\beta|} (\mathbf{P}^*)^T \pi_{\beta_i} \right) \right)^{\frac{1}{2}} (1 + o(1))$$

where $\mathbf{P}^*$ is any loss-neutral distribution and c is a positive function of $(\mathbf{\Pi}, \Lambda)$ and $\mathbf{P}^*$. □

This is a generalization of Theorem 6 in that the latter can be recovered from this result by setting $\mathcal{C} = \mathcal{A}^k \times \mathcal{A}^k$.

6.2 Two-dimensionally indexed data

So far, we considered denoising one-dimensionally indexed data. In this section, we extend some of the results to multi-dimensionally indexed data. Im many common applications, e.g., image denoising, the data is multi-dimensional. For such data, we define a class of sliding window denoisers, and derive lower bounds on the regret with respect to this class. For ease of exposition we present the results for two-dimensionally indexed data and a specific class of sliding window denoisers. But the results can be easily generalized to other classes as well.

Let $\mathcal{A}^{m\times n}$ denote the set of all two-dimensional arrays with m rows and n columns whose elements take values in $\mathcal{A}$. For any array $a^{m\times n}$ in $\mathcal{A}^{m\times n}$ we denote the entry in the ith row and jth column by $a_{i,j}$ or sometimes by $a[i,j]$. Further, for $i_1 \leq i_2$ and $j_1 \leq j_2$, let $a_{(i_1,j_1)}^{(i_2,j_2)}$ denote the rectangular array comprising of the elements $\{a_{i,j} : i_1 \leq i \leq i_2, j_1 \leq j \leq j_2\}$. Let $x^{m\times n} \in \mathcal{A}^{m\times n}$ denote the two-dimensional noiseless array that is input to a memoryless channel $\mathbf{\Pi}$ and $z^{m\times n} \in \mathcal{A}^{m\times n}$ the noisy output. Further let $x_{i,j}$ and $z_{i,j}$ respectively denote the symbol in the ith row and the j th column of the noiseless and noisy arrays.

An (m,n)-*block denoiser* is a mapping $\hat{X}^{m\times n} : \mathcal{A}^{m\times n} \to \mathcal{A}^{m\times n}$. For a given loss function $\mathbf{\Lambda}$, a noiseless input array $x^{m\times n}$ and the observed output sequence $z^{m\times n}$, the normalized cumulative loss $L_{\hat{X}^{m\times n}}(x^{m\times n}, z^{m\times n})$ of the denoiser $\hat{X}^{m\times n}$ is

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

Let $\mathcal{D}_{m\times n}$ denote the class of all (m,n)-block denoisers. A *k-th order sliding window denoiser* $\hat{X}^{m\times n}$ is a denoiser with the property that for all $z^{m\times n}\in\mathcal{A}^{m\times n}$, if

$$z_{(i_1-k,j_1-k)}^{(i_1+k,j_1+k)} = z_{(i_2-k,j_2-k)}^{(i_2+k,j_2+k)}$$

then

$$\hat{X}^{m\times n}(z^{m\times n})[i_1,j_1] = \hat{X}^{m\times n}(z^{m\times n})[i_2,j_2].$$

Thus the denoiser defines a mapping,

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

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

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

Let $\mathcal{S}_{k\times k}$ denote the class of 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 an individual noiseless array $x^{m\times n}\in\mathcal{A}^{m\times n}$ and a noisy array $z^{m\times n}\in\mathcal{A}^{m\times n}$, $k\ge 0$ and $m,n>2k$, $D_k(x^{m\times n},z^{m\times n})$, the *k-th order minimum loss* of $(x^{m\times n},z^{m\times n})$ is defined to be

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

the least loss incurred by any k-th order denoiser on the pair $(x^{m\times n},z^{m\times n})$. Note that we have slightly modified the definition of normalized cumulative loss to accomodate noiseless and noisy arrays of differing dimensions. For a given channel $\mathbf{\Pi}$ and a noiseless array $x^{m\times n}$ define

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

the expected k-th order minimum loss incurred when each random noisy array $Z^{m\times n}$ produced when $x^{m\times n}$ is input to the channel is denoised by the best k-th order denoiser for the pair $(x^{m\times n},Z^{m\times n})$. As for the one-dimensional case, for any (m,n)-block denoiser $\hat{X}^{m\times n}$ we define the regret function

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

Then we have the following lower bound on the regret.

Theorem 12. For any neutralizable pair $(\mathbf{\Pi},\Lambda)$, and any sequence $\{\hat{X}^{m\times n}\in\mathcal{D}_{m\times n}\}$ of denoisers, as n tends to infinity

$$\hat{R}_k\Big(\hat{X}^{m\times n}\Big)\ge\frac{c}{\sqrt{mn}}\left(\sum_{a\in\mathcal{A}}\sqrt{(\mathbf{P}^*)^T\pi_a}\right)^{(2k+1)^2-1}(1+o(1))$$

where $\mathbf{P}^*$ is any loss-neutral distribution and c is a positive function of $(\mathbf{\Pi},\Lambda)$ and $\mathbf{P}^*$. □

The proof of the theorem is identical to that of Theorem 6 except for the fact that we use a Central Limit Theorem for two-dimensionally indexed k-dependent random variables (*e.g.,* [12]) that lets us derive a lemma analogous to Lemma 5.

7 Stochastic Setting

So far, we dealt with the performance of universal denoisers in semi-stochastic settings, namely, compared their performance to denoisers tuned to the noiseless sequence that is input to the channel. In this section we try to derive similar lower bounds for universal denoisers in the stochastic setting. Recall that $\mathbf{D}(\mathbf{P})$ denotes the minimum expected loss incurred by any n-block denoiser where the expectation is over all X^n distributed according to $\mathbf{P}$ and all channel relaizations and that the regret of an n-block denoiser $\hat{X}^n$ for a class of distributions $\mathcal{P}$ is defined to be

$$\mathbf{R}_{\mathcal{P}}\left(\hat{X}^n\right) = \max_{\mathbf{P}\in\mathcal{P}} E\left[L_{\hat{X}^n}(X^n, Z^n)\right] - \mathbf{D}(\mathbf{P}).$$

It was shown in [1] that for the collection of all stationary processes, the regret of the DUDE asymptotically tends to zero. We consider the subclass $\mathcal{I}_n$ of *i.i.d.* distributions over $\mathcal{A}^n$ and derive lower bounds on $\mathbf{R}_{\mathcal{I}_n}\left(\hat{X}^n\right)$ for any $\hat{X}^n \in \mathcal{D}_n$, in Theorem 15. To do so we require a few preliminary results.

Lemma 13. Let $\mathbf{P}^* \in \mathcal{M}$ and let the sequences of distributions $\{\mathbf{P}^n \in \mathcal{M}\}$ and $\{\mathbf{Q}^n \in \mathcal{M}\}$ be such that for all $a \in \mathcal{A}$, both $|\mathbf{P}^n[a] - \mathbf{P}^*[a]| \leq \frac{c}{\sqrt{n}}$, and $|\mathbf{Q}^n[a] - \mathbf{P}^*[a]| \leq \frac{c}{\sqrt{n}}$ where $c > 0$. Also if the subset of $\mathcal{A}$ where $\mathbf{P}^*[a] > 0$ is identical to that of both $\mathbf{P}^n[a]$ and $\mathbf{Q}^n[a]$, then for all $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{A}^n$ such that $\mathcal{S}_1 \bigcup \mathcal{S}_2 = \mathcal{A}^n$, as n tends to infinity

$$\sum_{a^n\in\mathcal{S}_1}\prod_{i=1}^{n}\mathbf{P}^n[a_i] + \sum_{a^n\in\mathcal{S}_2}\prod_{i=1}^{n}\mathbf{Q}^n[a_i] = \Omega(1).$$

Proof See Appendix II □

Next we require a result on the nature of loss-neutral distributions.

Lemma 14. For any neutralizable $(\mathbf{\Pi}, \mathbf{\Lambda})$, there exists some distribution $\mathbf{P}^*$, and some $t \in \mathcal{A}$, and $i \neq j \in \mathcal{A}$, such that $\mathbf{P}^*$ is loss-neutral with respect to (t, i, j) and the following is true for some $\mathbf{v} \in \mathbb{R}^M$. For all $a \in \mathcal{A}$, $\mathbf{P}^*[a] = 0$ implies that $\mathbf{v}[a] = 0$ and for all sufficiently small $\epsilon > 0$, $\mathbf{P}^* + \epsilon\mathbf{v}, \mathbf{P}^* - \epsilon\mathbf{v} \in \mathcal{M}$. For all sufficiently small $\epsilon > 0$ and all $k \in \mathcal{A}$

$$(\mathbf{P}^* + \epsilon\mathbf{v})^T(\lambda_k \odot \pi_t) \geq (\mathbf{P}^* + \epsilon\mathbf{v})^T(\lambda_i \odot \pi_t) \tag{32}$$

with equality iff $\mathbf{P}^* \odot (\lambda_i - \lambda_k) \odot \pi_t = 0$, and for all $k \in \mathcal{A}$

$$(\mathbf{P}^* - \epsilon\mathbf{v})^T(\lambda_k \odot \pi_t) \geq (\mathbf{P}^* - \epsilon\mathbf{v})^T(\lambda_j \odot \pi_t) \tag{33}$$

with equality iff $\mathbf{P}^* \odot (\lambda_j - \lambda_k) \odot \pi_t = 0$. □

The optimal denoiser when the noiseless sequence X^n is drawn according to an *i.i.d.* distribution is a zeroth order sliding window denoiser. Here we derive the optimal denoiser when X^n is drawn according to a mixture of *i.i.d.* distributions. Let $\mathbf{P}_1, \mathbf{P}_2 \in \mathcal{I}_n$ and $\mathbf{P}$ be the mixture distribution

obtained by selecting one of $\mathbf{P}_1, \mathbf{P}_2$ uniformly. Formally let Γ be drawn uniformly from $\{1,2\}$ and X^n be drawn *i.i.d.* according to $\mathbf{P}_\Gamma$. Then

$$Pr(X^n = x^n) = \frac{1}{2}\prod_{i=1}^{n}\mathbf{P}_1[x_i] + \frac{1}{2}\prod_{i=1}^{n}\mathbf{P}_2[x_i].$$

We are interested in

$$\hat{X}^n_{\text{opt}} \stackrel{\text{def}}{=} \arg\min_{\hat{X}^n \in \mathcal{D}_n} E\big[L_{\hat{X}^n}(X^n, Z^n)\big],$$

the n-block denoiser that minimizes the expected loss, and D_{opt}, the minimum loss. Conditioned on Γ, $X_1, X_2, \ldots, X_n$ are drawn *i.i.d.* and since the channel is memoryless, for all $x^n, z^n \in \mathcal{A}^n$, $\gamma \in \{1,2\}$, and all $1 \le t \le n$

$$Pr(X_t = x_t | Z^n = z^n, \Gamma = \gamma) = Pr(X_t = x_t | Z_t = z_t, \Gamma = \gamma). \tag{34}$$

Recall that $\mathbf{P}_{X_t|z^n}$ denotes the column vector whose α-th component is $Pr(X_t = \alpha | Z^n = z^n)$. Then from (34)

$$\mathbf{P}_{X_t|z^n} = \frac{\frac{1}{2}\sum_\gamma \mathbf{P}_\gamma \odot \pi_{z_t} \prod_{i=1, i\neq t}^{n} \mathbf{P}_\gamma^T \pi_{z_i}}{\frac{1}{2}\sum_\gamma \prod_{i=1}^{n} \mathbf{P}_\gamma^T \pi_{z_i}}.$$

From (4)

$$\hat{X}^n_{\text{opt}}(z^n)[t] = \hat{x}\big(\mathbf{P}_{X_t|z^n}\big) = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{P}_{X_t|z^n} = \arg\min_{\hat{x}\in\mathcal{A}} \frac{\sum_\gamma \lambda_{\hat{x}}^T (\mathbf{P}_\gamma \odot \pi_{z_t}) \prod_{i=1, i\neq t}^{n} \mathbf{P}_\gamma^T \pi_{z_i}}{\sum_\gamma \prod_{i=1}^{n} \mathbf{P}_\gamma^T \pi_{z_i}}.$$

Observe that

$$Pr(Z^n = z^n) = \frac{1}{2}\sum_\gamma \prod_{i=1}^{n} \mathbf{P}_\gamma^T \pi_{z_i},$$

and therefore the optimal expected loss incurred in denoising X_t is

$$\begin{aligned}
E\big[U\big(\mathbf{P}_{X_t|Z^n}\big)\big] &= E\left[\min_{\hat{x}\in\mathcal{A}} \frac{\sum_\gamma \lambda_{\hat{x}}^T (\mathbf{P}_\gamma \odot \pi_{Z_t}) \prod_{i=1, i\neq t}^{n} \mathbf{P}_\gamma^T \pi_{Z_i}}{\sum_\gamma \prod_{i=1}^{n} \mathbf{P}_\gamma^T \pi_{Z_i}}\right] \\
&= \frac{1}{2}\sum_{z^n \in \mathcal{A}^n} \min_{\hat{x}\in\mathcal{A}} \sum_\gamma \lambda_{\hat{x}}^T (\mathbf{P}_\gamma \odot \pi_{z_t}) \prod_{i=1, i\neq t}^{n} \mathbf{P}_\gamma^T \pi_{z_i} \\
&= \frac{1}{2}\sum_{z^{n-1} \in \mathcal{A}^{n-1}} \sum_{z\in\mathcal{A}} \min_{\hat{x}\in\mathcal{A}} \sum_{\gamma=1,2} \lambda_{\hat{x}}^T (\mathbf{P}_\gamma \odot \pi_z) \prod_{i=1}^{n-1} \mathbf{P}_\gamma^T \pi_{z_i}
\end{aligned}$$

which does not depend on the index t. Therefore by (5)

$$D_{\text{opt}} = \frac{1}{n}\sum_{t=1}^{n} E\big[U\big(\mathbf{P}_{X_t|Z^n}\big)\big] = E\big[U\big(\mathbf{P}_{X_t|Z^n}\big)\big] = \frac{1}{2}\sum_{z^{n-1} \in \mathcal{A}^{n-1}} \sum_{z\in\mathcal{A}} \min_{\hat{x}\in\mathcal{A}} \sum_\gamma \lambda_{\hat{x}}^T (\mathbf{P}_\gamma \odot \pi_z) \prod_{i=1}^{n-1} \mathbf{P}_\gamma^T \pi_{z_i}. \tag{35}$$

In the following theorem, we derive a lower bound on $\mathbf{R}_{\mathcal{I}_n}(\hat{X}^n)$ for any $\hat{X}^n \in \mathcal{D}_n$ using these preliminary results.

Theorem 15. For any neutralizable pair $(\mathbf{\Pi}, \Lambda)$, and any sequence $\{\hat{X}^n \in \mathcal{D}_n\}$ of denoisers, as n tends to infinity

$$\mathbf{R}_{\mathcal{I}_n}\Big(\hat{X}^n\Big) = \Omega\Big(n^{-\frac{1}{2}}\Big).$$

Proof Let $t, i, j \in \mathcal{A}$, $\mathbf{P}^* \in \mathcal{M}$, $\mathbf{v} \in \mathbb{R}^M$ and $\epsilon > 0$ be as in the statement of Lemma 14.

For any two *i.i.d.* distributions $\mathbf{P}_1$, $\mathbf{P}_2$

$$\begin{aligned}\mathbf{R}_{\mathcal{I}_n}\Big(\hat{X}^n\Big) &\geq \max_{\gamma\in\{1,2\}} E_{\mathbf{P}_\gamma}\big[L_{\hat{X}^n}(X_1^n, Z^n)\big] - \mathbf{D}(\mathbf{P}_\gamma) \\ &\geq \frac{1}{2}\sum_{\gamma=1,2}\big(E_{\mathbf{P}_\gamma}\big[L_{\hat{X}^n}(X_1^n, Z^n)\big] - \mathbf{D}(\mathbf{P}_\gamma)\big). \end{aligned} \tag{36}$$

Observe that if a random variable Γ is selected according to a uniform distribution over $\{1, 2\}$ and X^n is generated *i.i.d.* according to $\mathbf{P}_\Gamma$, then the expected loss incurred by $\hat{X}^n$ is

$$E_\Gamma\big[E_{\mathbf{P}_\Gamma}\big[L_{\hat{X}^n}(X^n, Z^n)\big]\big] = \frac{1}{2}\sum_{\gamma=1,2} E_{\mathbf{P}_\gamma}\big[L_{\hat{X}^n}(X_1^n, Z^n)\big].$$

From (35) this can be lower bounded as follows

$$\frac{1}{2}\sum_{\gamma=1,2} E_{\mathbf{P}_\gamma}\big[L_{\hat{X}^n}(X_1^n, Z^n)\big] \geq D_{\text{opt}} = \frac{1}{2}\sum_{z^{n-1}\in\mathcal{A}^{n-1}}\sum_{z\in\mathcal{A}}\min_{\hat{x}\in\mathcal{A}}\sum_{\gamma=1,2}\lambda_{\hat{x}}^T(\mathbf{P}_\gamma \odot \pi_z)\prod_{i=1}^{n-1}\mathbf{P}_\gamma^T\pi_{z_i}. \tag{37}$$

From Lemma 1

$$\mathbf{D}(\mathbf{P}_\gamma) = \sum_{z\in\mathcal{A}}\min_{\hat{x}\in\mathcal{A}}\lambda_{\hat{x}}^T(\mathbf{P}_\gamma \odot \pi_z) = \sum_{z\in\mathcal{A}}\min_{\hat{x}\in\mathcal{A}}\lambda_{\hat{x}}^T(\mathbf{P}_\gamma \odot \pi_z)\sum_{z^{n-1}\in\mathcal{A}^{n-1}}\prod_{i=1}^{n-1}\mathbf{P}_\gamma^T\pi_{z_i} \tag{38}$$

where the last equality holds as the extra term is merely the probability of all length-$n-1$ sequences over $\mathcal{A}$ and therefore 1. Substituting (38) and (37) in (36) we get for all *i.i.d.* distributions $\mathbf{P}_1$ and $\mathbf{P}_2$

$$\mathbf{R}_{\mathcal{I}_n}\Big(\hat{X}^n\Big) \geq \frac{1}{2}\sum_{z^{n-1}\in\mathcal{A}^{n-1}} g(z^{n-1}) \tag{39}$$

where

$$\begin{aligned} g(z^{n-1}) &= \sum_{z\in\mathcal{A}}\left(\min_{\hat{x}\in\mathcal{A}}\sum_{\gamma=1,2}\lambda_{\hat{x}}^T(\mathbf{P}_\gamma \odot \pi_z)\prod_{i=1}^{n-1}\mathbf{P}_\gamma^T\pi_{z_i} - \sum_{\gamma=1,2}\min_{\hat{x}\in\mathcal{A}}\lambda_{\hat{x}}^T(\mathbf{P}_\gamma \odot \pi_z)\prod_{i=1}^{n-1}\mathbf{P}_\gamma^T\pi_{z_i}\right) \\ &\geq \min_{\hat{x}\in\mathcal{A}}\sum_{\gamma=1,2}\lambda_{\hat{x}}^T(\mathbf{P}_\gamma \odot \pi_t)\prod_{i=1}^{n-1}\mathbf{P}_\gamma^T\pi_{z_i} - \sum_{\gamma=1,2}\min_{\hat{x}\in\mathcal{A}}\lambda_{\hat{x}}^T(\mathbf{P}_\gamma \odot \pi_t)\prod_{i=1}^{n-1}\mathbf{P}_\gamma^T\pi_{z_i} \end{aligned} \tag{40}$$

where we justify the inequality as follows. The minimum of a sum of functions is greater then the sum of the minimum values, hence each of the terms in the summation over z is non-negative and therefore dropping all the terms except the one correponding to $z = t$ gives a valid lower bound.

Let $\mathcal{A}^*$ be a subset of $\mathcal{A}$ that contains i and j and has the property that if $a_1 \neq a_2 \in \mathcal{A}^*$, then $\mathbf{P}^* \odot (\lambda_{a_1} - \lambda_{a_2}) \odot \pi_t \neq 0$. Furthermore let $\mathcal{A}^*$ be maximal, *i.e.,* for all $b \in \mathcal{A}/\mathcal{A}^*$, there exists some

$a \in \mathcal{A}^*$ such that $\mathbf{P}^* \odot (\lambda_a - \lambda_b) \odot \pi_t = 0$. Note that in that case the minimizations in (40) can be restricted to $\mathcal{A}^*$ instead of $\mathcal{A}$. If we choose $\mathbf{P}_1 = \mathbf{P}^* + \epsilon \mathbf{v}$ and $\mathbf{P}_2 = \mathbf{P}^* - \epsilon \mathbf{v}$, with ϵ satisfying (32) and (33), we obtain that

$$\min_{\hat{x} \in \mathcal{A}^*} \lambda_{\hat{x}}^T (\mathbf{P}_1 \odot \pi_t) = \lambda_i^T (\mathbf{P}_1 \odot \pi_t), \quad \text{and} \quad \min_{\hat{x} \in \mathcal{A}^*} \lambda_{\hat{x}}^T (\mathbf{P}_2 \odot \pi_t) = \lambda_j^T (\mathbf{P}_2 \odot \pi_t). \tag{41}$$

For $\alpha \in \mathcal{A}^*$, let

$$\mathcal{S}_\alpha \stackrel{\text{def}}{=} \left\{ z^{n-1} : \arg\min_{\hat{x} \in \mathcal{A}} \sum_{\gamma=1,2} \lambda_{\hat{x}}^T (\mathbf{P}_\gamma \odot \pi_t) \prod_{r=1}^{n-1} \mathbf{P}_\gamma^T \pi_{z_r} = \alpha \right\}$$

where the ties in the minimization are broken according to some fixed rule. The $\mathcal{S}_\alpha$'s partition the space $\mathcal{A}^{n-1}$. Using this fact and (40), (41), we can reduce (39) to

$$\begin{aligned} \mathbf{R}_{\mathcal{I}_n}\left(\hat{X}^n\right) \geq & \frac{1}{2} \sum_{\substack{\alpha \in \mathcal{A}^*, \\ \alpha \neq i}} (\lambda_\alpha - \lambda_i)^T (\mathbf{P}_1 \odot \pi_t) \sum_{z^{n-1} \in \mathcal{S}_\alpha} \prod_{r=1}^{n-1} \mathbf{P}_1^T \pi_{z_r} \\ & + \frac{1}{2} \sum_{\substack{\alpha \in \mathcal{A}^*, \\ \alpha \neq j}} (\lambda_\alpha - \lambda_j)^T (\mathbf{P}_2 \odot \pi_t) \sum_{z^{n-1} \in \mathcal{S}_\alpha} \prod_{r=1}^{n-1} \mathbf{P}_2^T \pi_{z_r}. \end{aligned}$$

The choice of $\mathbf{P}_1$ and $\mathbf{P}_2$ ensures that each term in both the summations over $\alpha \in \mathcal{A}^*$ are positive and therefore letting $\bar{\mathcal{S}}_a$ denote the set $\mathcal{A}^{n-1}/\mathcal{S}_\alpha$, we obtain

$$\begin{aligned} \mathbf{R}_{\mathcal{I}_n}\left(\hat{X}^n\right) \geq & \frac{1}{2} \min \left\{ \min_{a \in \mathcal{A}^*, a \neq i} \left\{ (\lambda_a - \lambda_i)^T (\mathbf{P}_1 \odot \pi_t) \right\}, \min_{a \in \mathcal{A}^*, a \neq j} \left\{ (\lambda_a - \lambda_j)^T (\mathbf{P}_2 \odot \pi_t) \right\} \right\} \\ & \left(\sum_{z^{n-1} \in \bar{\mathcal{S}}_i} \prod_{r=1}^{n-1} \mathbf{P}_1^T \pi_{z_r} + \sum_{z^{n-1} \in \bar{\mathcal{S}}_j} \prod_{r=1}^{n-1} \mathbf{P}_2^T \pi_{z_r} \right). \end{aligned} \tag{42}$$

Let $\epsilon = n^{-\frac{1}{2}}$ for some sufficiently large n. For all $\alpha \in \mathcal{A}^*, \alpha \neq i$, from (32)

$$(\lambda_\alpha - \lambda_i)^T (\mathbf{P}_1 \odot \pi_t) = (\lambda_\alpha - \lambda_i)^T (\mathbf{P}^* \odot \pi_t) + n^{-\frac{1}{2}} (\lambda_\alpha - \lambda_i)^T (\mathbf{v} \odot \pi_t) > 0.$$

Observe that $(\lambda_\alpha - \lambda_i)^T (\mathbf{P}^* \odot \pi_t)$ is non-negative for all α. If it is positive, then, as n tends to infinity, the quantity on the right hand side of the above equation is $\Omega(1)$ and if it is 0, then

$$(\lambda_\alpha - \lambda_i)^T (\mathbf{P}_1 \odot \pi_t) = \Omega\left(n^{-\frac{1}{2}}\right).$$

The same is true for $(\lambda_\alpha - \lambda_j)^T (\mathbf{P}_2 \odot \pi_t)$. Therefore

$$\min \left\{ \min_{a \neq i} \left\{ (\lambda_a - \lambda_i)^T (\mathbf{P}_1 \odot \pi_t) \right\}, \min_{a \neq j} \left\{ (\lambda_a - \lambda_j)^T (\mathbf{P}_2 \odot \pi_t) \right\} \right\} = \Omega\left(n^{-\frac{1}{2}}\right). \tag{43}$$

Let $\mathbf{Q}_1^T = \mathbf{P}_1^T \mathbf{\Pi}$ and $\mathbf{Q}^T = \mathbf{P}_2^T \mathbf{\Pi}$ be the distributions induced by $\mathbf{P}_1$ and $\mathbf{P}_2$ respectively at the output of the channel. Since $\epsilon = n^{-\frac{1}{2}}$, for all $a \in \mathcal{A}$,

$$|\mathbf{Q}_1[a] - \mathbf{Q}_2[a]| \leq \frac{c}{\sqrt{n}}$$

for some constant c. Also note that $\bar{\mathcal{S}}_i \bigcup \bar{\mathcal{S}}_j = \mathcal{A}^{n-1}$. Then, letting n tend to infinity, and applying Lemma 13 to the distributions $\mathbf{P}^*\mathbf{\Pi}$, $\mathbf{Q}_1$ and $\mathbf{Q}_2$ we obtain

$$\left(\sum_{z^{n-1} \in \bar{\mathcal{S}}_i} \prod_{r=1}^{n-1} \mathbf{P}_2^T \pi_{z_r} + \sum_{z^{n-1} \in \bar{\mathcal{S}}_j} \prod_{r=1}^{n-1} \mathbf{P}_1^T \pi_{z_r} \right) = \Omega(1).$$

Substituting this and (43) in (42) gives us the result. □

Acknowledgements

The authors would like to thank Gadiel Seroussi, Sergio Verdú, Marcelo Weinberger, and Tsachy Weissman for fruitful discussions and valuable comments.

References

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

[2] J. F. Hannan and H. Robbins. Asymptotic solutions of the compound decision problem for two completely specified distributions. *Annals of Mathematical Statistics*, 26:37–51, 1955.

[3] J. F. Hannan and J.R. Van Ryzin. Rate of convergence in the compound decision problem for two completely specified distributions. *Annals of Mathematical Statistics*, 36:1743–1752, 1965.

[4] M. V. Johns Jr. Two-action compound decision problems. In *Proc. 5th Berkeley Symposium on Mathematical and Statistical Probability*, pages 463–478, 1967.

[5] J. Hannan. Approximation to baye's risk in repeated play. In *Contributions to the theory of games*, volume 3, pages 97–139. Princeton University Press, Princeton, 1957.

[6] P. Billingsley. *Probability and Measure*. John Wiley and sons., 1986.

[7] T. M. Cover. Behaviour of sequential predictors of binary sequences. In *Trans. of the 4th Prague Conference on Information Theory, Statistical Decision functions, Random Processes*, 1965.

[8] W. Hoeffding and H. Robbins. The central limit theorem for dependent random variables. *Duke Math. Journal*, 15(3):773–780, 1948.

[9] J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, 1984.

[10] M. J. Weinberger, N. Merhav, and M. Feder. Optimal sequential probability assignment for individual sequences. *IEEE Transactions on Information Theory*, 40(2):384–396, 1994.

[11] E. Ordentlich, M. Weinberger, and T. Weissman. Multi-directional context sets with applications to universal denoising and compression. In *Proceedings of IEEE Symposium on Information Theory*, pages 1270–1274, 2005.

[12] L. Heinrich. Stable limit theorems for sums of multiply indexed m-dependent ranom variables. *Math. Nachr.*, 127:193–210, 1986.

[13] V. V. Shergin. The central limit theorem for finitely dependent random variables. In *Proc. 5th Vilnius conference on Probability Theory and Mathematical Statistics*, pages 424–431, 1990.

[14] W. J. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:713–721, 1963.

[15] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience Publication, 1991.

Appendix I: Stronger version of Lemma 5

To strengthen Lemma 5 in a way that it applies not just to fixed k, but also to a sequence $\{k_n\}$ that increases with n, we require a version of the Berry-Esseen inequality for m-dependent random variables.

Let $X_1, X_2, \ldots, X_n$ be a sequence of m-dependent random variables with zero means and finite variances. If the sequence is stationary then

$$V \stackrel{\text{def}}{=} E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = nE\left[X_1^2\right] + 2\sum_{j=1}^{m}(n-j)E[X_1 X_{1+j}].$$

Suppose

$$S_n \stackrel{\text{def}}{=} \frac{1}{\sqrt{V}}\sum_{i=1}^{n} X_i$$

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

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

Then it follows from the more general result by Shergin [13] that, under the finiteness of the sth absolute moment of X_1, $2 < s < 3$,

$$\sup_x |Pr(S_n < x) - \Phi(x)| \leq c(m+1)^{s-1} nE[|X_1|^s]V^{-\frac{s}{2}} \tag{44}$$

where c is a constant.

Using this inequality the following lemma may be obtained.

Lemma 16. If X^n is generated *i.i.d.* according to some $\mathbf{P} \in \mathcal{M}$, then for any column vector $\alpha \in \mathbb{R}^M$, and any sequence of integers $\{k_n\}$ and contexts $\{c_{-k_n}^{k_n}\} \in \mathcal{A}^{2k_n+1}$, such that $n > 2k_n$ and $\alpha^T(\mathbf{P} \odot \pi_{c_0}) = 0$, as n tends to infinity, if $k_n = o(\ln n)$

$$E_{\mathbf{P}}\Big[\sqrt{n}\Big|\alpha^T\mathbf{q}\Big(Z^n, X^n, c_{-k_n}^{k_n}\Big)\Big|\Big] = \sqrt{\frac{2V_n}{\pi}}(1+o(1))$$

where $V_n = (\alpha \odot \alpha)^T(\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k_n, i\neq 0}^{k_n} \mathbf{P}^T\pi_{c_i}$.

Proof As in the proof of Lemma 5, for a given $\alpha \in \mathbb{R}^M$ and $c_{-k_n}^{k_n} \in \mathcal{A}^{2k_n+1}$ we define the sequence $Y_{k_n+1}^{n-k_n}$ of random variables as

$$Y_i \stackrel{\text{def}}{=} \sum_{\ell\in\mathcal{A}} \alpha(\ell) 1\Big(X_i = \ell, Z_{i-k_n}^{i+k_n} = c_{-k_n}^{k_n}\Big), \qquad k_n + 1 \le i \le n - k_n.$$

Then

$$\frac{1}{n-2k_n}\sum_{i=k_n+1}^{n-k_n} Y_i = \alpha^T\mathbf{q}\Big(Z^n, X^n, c_{-k_n}^{k_n}\Big). \tag{45}$$

If the sequence X^n is drawn *i.i.d.* according to $\mathbf{P}$, $Y_{k_n+1}^{n-k_n}$ is stationary and since each Y_i is a function of $2k_n+1$ consecutive X_i's, it is easy to verify that $Y_{k_n+1}^{n-k_n}$ is a $2k_n$-dependent sequence. It may be recalled from the proof of Lemma 5 that

$$E\big[Y_{k_n+1}^2\big] = \sum_{\ell\in\mathcal{A}} \alpha(\ell)^2 Pr\Big(X_{k_n+1} = \ell, Z_1^{2k_n+1} = c_{-k_n}^{k_n}\Big) = V_n = (\alpha\odot\alpha)^T(\mathbf{P}\odot\pi_{c_0}) \prod_{i=-k_n, i\neq 0}^{k_n} \mathbf{P}^T\pi_{c_i}$$

and that for all $i \ge 1$

$$E[Y_{k_n+1}Y_{k_n+1+i}] = 0.$$

So the Lemma is trivially true if $V_n = 0$ and therefore we consider $V_n > 0$. Also note that

$$E[|Y_{k_n+1}|^s] = \sum_{\ell\in\mathcal{A}} |\alpha(\ell)|^s Pr\Big(X_{k_n+1} = \ell, Z_1^{2k_n+1} = c_{-k_n}^{k_n}\Big) = \alpha_s^T(\mathbf{P}\odot\pi_{c_0}) \prod_{i=-k_n, i\neq 0}^{k_n} \mathbf{P}^T\pi_{c_i}$$

where $\alpha_s(\ell) = |\alpha(\ell)|^s$ for all $\ell \in \mathcal{A}$. Therefore from (44)

$$\sup_x \left| Pr\left(\sqrt{\frac{n-2k_n}{V_n}}\alpha^T\mathbf{q}\Big(Z^n, X^n, c_{-k_n}^{k_n}\Big) < x\right) - \Phi(x)\right| \le c(2k_n+1)^{s-1}V_n^{-\frac{s}{2}}(n-2k_n)^{-(\frac{s}{2}-1)} \tag{46}$$

where c is a constant that depends on α, $\mathbf{P}$ and $\mathbf{\Pi}$. For all positive random variables X whose expectation is finite

$$E[X] = \int_0^\infty Pr(X \ge x)dx.$$

Applying this to $S_n \stackrel{\text{def}}{=} \sqrt{(n-2k_n)V_n^{-1}}|\alpha^T\mathbf{q}\Big(Z^n, X^n, c_{-k_n}^{k_n}\Big)|$ and comparing the expectation with the expected absolute value of a unit normal random variable, we obtain for any $\tau_n > 0$

$$\begin{aligned}\left|E_{\mathbf{P}}[S_n] - 2\int_0^\infty (1-\Phi(x))dx\right| &\le \int_0^{\tau_n} |Pr(S_n > x) - 2(1-\Phi(x))|\,dx \\ &\quad + \int_{\tau_n}^\infty Pr(S_n > x)dx + \int_{\tau_n}^\infty 2(1-\Phi(x))dx \end{aligned} \tag{47}$$

From (46)

$$\int_0^{\tau_n} |Pr(S_n > x) - 2(1-\Phi(x))|\, dx \leq 2\tau_n c(2k_n+1)^{s-1} V_n^{-\frac{s}{2}} (n-2k_n)^{-(\frac{s}{2}-1)}. \tag{48}$$

Observe from (45) that

$$\begin{aligned} Pr(S_n > x) &= Pr\left(\left| \sum_{i=k_n+1}^{n-k_n} Y_i \right| > x\sqrt{(n-2k_n)V_n} \right) \\ &= Pr\left(\left| \sum_{l=0}^{2k_n} \sum_{\substack{i=k_n+1, \\ i \mod 2k_n+1=\ell}}^{n-k_n} Y_i \right| > x\sqrt{(n-2k_n)V_n} \right) \\ &\leq \sum_{l=0}^{2k_n} Pr\left(\left| \sum_{\substack{i=k_n+1, \\ i \mod 2k_n+1=\ell}}^{n-k_n} Y_i \right| > \frac{x\sqrt{(n-2k_n)V_n}}{2k_n+1} \right). \end{aligned}$$

Since for each $0 \leq \ell \leq 2k_n$, $\{Y_i : i \mod 2k_n+1 = \ell\}$ is a collection of iid random variables with zero mean, we can apply Hoeffding's inequality [14] to obtain

$$Pr(S_n > x) \leq 2(2k_n+1) e^{-2x^2(n-2k_n)V_n(2k_n+1)^{-2}\left(\left\lceil \frac{n-2k_n}{2k_n+1} \right\rceil\right)^{-1}} \leq 2(2k_n+1) e^{-2V_n x^2 \left(\frac{n-2k_n}{(n+1)(2k_n+1)}\right)}.$$

Therefore

$$\int_{\tau_n}^{\infty} Pr(S_n > x)dx \leq 2(2k_n+1) \int_{\tau_n}^{\infty} e^{-2V_n x^2 \left(\frac{n-2k_n}{(n+1)(2k_n+1)}\right)} dx \leq \frac{(2k_n+1) e^{-2V_n \tau_n^2 \left(\frac{n-2k_n}{(n+1)(2k_n+1)}\right)}}{2V_n \tau_n \left(\frac{n-2k_n}{(n+1)(2k_n+1)}\right)} \tag{49}$$

where we have used the fact that

$$\int_x^{\infty} e^{-\frac{\alpha t^2}{2}} dt \leq \frac{e^{-\frac{\alpha x^2}{2}}}{\alpha x}.$$

Also

$$\int_{\tau_n}^{\infty} 2(1-\Phi(x))dx = \int_{\tau_n}^{\infty} 2 \int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \leq \int_{\tau_n}^{\infty} \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{x^2}{2}}}{x} dx \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{\tau_n^2}{2}}}{\tau_n^2}. \tag{50}$$

Since $V_n > 0$ observe that

$$V_n = (\alpha \odot \alpha)^T (\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k_n, i\neq 0}^{k_n} \mathbf{P}^T \pi_{c_i} \geq \beta \gamma^{2k_n}$$

where $\beta = (\alpha \odot \alpha)^T (\mathbf{P} \odot \pi_{c_0})$ and

$$\gamma = \min_{a \in \mathcal{A}: \mathbf{P}^T \pi_a > 0} \mathbf{P}^T \pi_a.$$

Substituting the above inequality and (48), (49), and (50) into (47) and setting τ_n to be $k_n\gamma^{-k_n}\ln n$ and we obtain

$$\left|E_{\mathbf{P}}[S_n] - 2\int_0^\infty (1-\Phi(x))dx\right| \leq 2c\beta^{-\frac{s}{2}}(2k_n+1)^{s-1}k_n\gamma^{-k_n(s+1)}(n-2k_n)^{-(\frac{s}{2}-1)}\ln n + \frac{(2k_n+1)e^{-2\beta(k_n\ln n)^2\left(\frac{n-2k_n}{(n+1)(2k_n+1)}\right)}}{2\beta\gamma^{k_n}k_n\ln n\left(\frac{n-2k_n}{(n+1)(2k_n+1)}\right)} + \sqrt{\frac{2}{\pi}}\frac{e^{-\frac{(k_n\ln n\gamma^{-k_n})^2}{2}}}{(k_n\ln n\gamma^{-k_n})^2}.$$

It is easy to verify that when n tends to infinity and $k_n = o(\ln n)$, the terms on the right hand side tend to zero. The proof is complete on observing that

$$2\int_0^\infty (1-\Phi(x))dx = \sqrt{\frac{2}{\pi}}. \qquad \square$$

Appendix II: Proof of Lemma 13

Lemma. Let $\mathbf{P}^* \in \mathcal{M}$ and let the sequences of distributions $\{\mathbf{P}^n \in \mathcal{M}\}$ and $\{\mathbf{Q}^n \in \mathcal{M}\}$ be such that for all $a \in \mathcal{A}$, both $|\mathbf{P}^n[a] - \mathbf{P}^*[a]| \leq \frac{c}{\sqrt{n}}$, and $|\mathbf{Q}^n[a] - \mathbf{P}^*[a]| \leq \frac{c}{\sqrt{n}}$ where $c > 0$. Also if the subset of $\mathcal{A}$ where $\mathbf{P}^*[a] > 0$ is identical to that of both $\mathbf{P}^n[a]$ and $\mathbf{Q}^n[a]$, then for all $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{A}^n$ such that $\mathcal{S}_1 \bigcup \mathcal{S}_2 = \mathcal{A}^n$, as n tends to infinity

$$\sum_{a^n\in\mathcal{S}_1}\prod_{i=1}^n \mathbf{P}^n[a_i] + \sum_{a^n\in\mathcal{S}_2}\prod_{i=1}^n \mathbf{Q}^n[a_i] = \Omega(1).$$

Proof Since the sum on the left hand side is minimum when $\mathcal{S}_2 = \mathcal{A}/\mathcal{S}_1$, it suffices to consider that case. Observe that

$$\begin{aligned}
1 - \left(\sum_{a^n\in\mathcal{S}_1}\prod_{i=1}^n \mathbf{P}^n[a_i] + \sum_{a^n\in\mathcal{A}^n/\mathcal{S}_1}\prod_{i=1}^n \mathbf{Q}^n[a_i]\right) &= \sum_{a^n\in\mathcal{S}_1}\prod_{i=1}^n \mathbf{Q}^n[a_i] - \sum_{a^n\in\mathcal{S}_1}\prod_{i=1}^n \mathbf{P}^n[a_i] \\
&\leq \max_{\mathcal{S}_1\subseteq\mathcal{A}^n}\left(\sum_{a^n\in\mathcal{S}_1}\prod_{i=1}^n \mathbf{Q}^n[a_i] - \sum_{a^n\in\mathcal{S}_1}\prod_{i=1}^n \mathbf{P}^n[a_i]\right) \\
&= \frac{1}{2}\sum_{a^n\in\mathcal{A}^n}\left|\prod_{i=1}^n \mathbf{P}^n[a_i] - \prod_{i=1}^n \mathbf{Q}^n[a_i]\right| \\
&\leq \frac{1}{2}\sum_{a^n\in\mathcal{A}^n}\left|\prod_{i=1}^n \mathbf{P}^n[a_i] - \prod_{i=1}^n \mathbf{P}^*[a_i]\right| \\
&\quad + \frac{1}{2}\sum_{a^n\in\mathcal{A}^n}\left|\prod_{i=1}^n \mathbf{P}^*[a_i] - \prod_{i=1}^n \mathbf{Q}^n[a_i]\right| \\
&\leq \left(\frac{\ln 2}{2}\right)^{\frac{1}{2}}\left(\sqrt{nD(\mathbf{P}^*||\mathbf{P}^n)} + \sqrt{nD(\mathbf{P}^*||\mathbf{Q}^n)}\right)
\end{aligned} \tag{51}$$

where $D(\cdot)$ is the Kullback-Leibler distance and we have used Pinsker's inequality (see *e.g.,* Lemma 12.6.1 in [15]).

Let $\mathbf{P}^*[a] > 0$ for k values of a and let $\mathcal{K}$ denote the set of k-dimensional probability distributions. Then we define the function $f : \mathcal{K} \to \mathbb{R}$ to be

$$f(\mathbf{P}) = D(\mathbf{P}^*||\mathbf{P}) = \sum_{\alpha \in \mathcal{A}: \mathbf{P}^*[\alpha] > 0} \mathbf{P}^*[\alpha] \log \frac{\mathbf{P}^*[\alpha]}{\mathbf{P}[\alpha]}.$$

Applying Taylor's formula about $\mathbf{P}^*$, we get

$$f(\mathbf{P}) = f(\mathbf{P}^*) + f'(\mathbf{P}^*)^T(\mathbf{P} - \mathbf{P}^*) + \frac{1}{2}(\mathbf{P} - \mathbf{P}^*)^T f''(\tilde{\mathbf{P}})(\mathbf{P} - \mathbf{P}^*) \tag{52}$$

where $\tilde{\mathbf{P}} = \lambda \mathbf{P}^* + (1 - \lambda)\mathbf{P}$ for some $\lambda \in (0, 1)$, and where $f'(\mathbf{P})$ is the column vector whose ith term is

$$f'(\mathbf{P})[i] = \frac{\partial f(\mathbf{P})}{\partial \mathbf{P}[i]},$$

and $f''(\mathbf{P})$ is the square matrix whose (i, j)th entry is

$$f''(\mathbf{P})[i, j] = \frac{\partial^2 f(\mathbf{P})}{\partial \mathbf{P}[i] \partial \mathbf{P}[j]}$$

where the indices i and j run over the set $\{a \in \mathcal{A} : \mathbf{P}^*[a] > 0\}$. Let $\mathbf{P} = \mathbf{P}^n$ in (52). Note that $f(\mathbf{P}^*) = 0$ and that $f'(\mathbf{P}^*)(\mathbf{P}^n - \mathbf{P}^*) = 0$. It can be verified that

$$(\mathbf{P}^n - \mathbf{P}^*)^T f''(\tilde{\mathbf{P}})(\mathbf{P}^n - \mathbf{P}^*) = \sum_{a \in \mathcal{K}} \frac{(\mathbf{P}^n[a] - \mathbf{P}^*[a])^2}{(\tilde{\mathbf{P}}[a])^2}.$$

Since $\tilde{\mathbf{P}}$ is in the interior of $\mathcal{K}$, for all $a \in \mathcal{K}$, $\tilde{\mathbf{P}}[a]$ can be lower bounded by a constant independent of n. Combining this with the fact that $|\mathbf{P}^*[a] - \mathbf{P}^n[a]| \leq \frac{c}{\sqrt{n}}$ for all $a \in \mathcal{K}$,

$$(\mathbf{P}^n - \mathbf{P}^*)^T f''(\tilde{\mathbf{P}})(\mathbf{P}^n - \mathbf{P}^*) = \mathcal{O}\big(n^{-1}\big).$$

Therefore (52) can be reduced to $f(\mathbf{P}^n) = \mathcal{O}\big(n^{-1}\big)$. Similarly $f(\mathbf{Q}^n) = \mathcal{O}\big(n^{-1}\big)$. Combining these with (51), we obtain the Lemma. □

Appendix III: Proof of Lemma 14

Lemma. For any neutralizable $(\mathbf{\Pi}, \mathbf{\Lambda})$, there exists some distribution $\mathbf{P}^*$, and some $t \in \mathcal{A}$, and $i \neq j \in \mathcal{A}$, such that $\mathbf{P}^*$ is loss-neutral with respect to (t, i, j) and the following is true for some $\mathbf{v} \in \mathbb{R}^M$. For all $a \in \mathcal{A}$, $\mathbf{P}^*[a] = 0$ implies that $\mathbf{v}[a] = 0$ and for all sufficiently small $\epsilon > 0$, $\mathbf{P}^* + \epsilon\mathbf{v}, \mathbf{P}^* - \epsilon\mathbf{v} \in \mathcal{M}$. For all sufficiently small $\epsilon > 0$ and all $k \in \mathcal{A}$

$$(\mathbf{P}^* + \epsilon\mathbf{v})^T(\lambda_k \odot \pi_t) \geq (\mathbf{P}^* + \epsilon\mathbf{v})^T(\lambda_i \odot \pi_t)$$

with equality iff $\mathbf{P}^* \odot (\lambda_i - \lambda_k) \odot \pi_t = 0$, and for all $k \in \mathcal{A}$

$$(\mathbf{P}^* - \epsilon\mathbf{v})^T(\lambda_k \odot \pi_t) \geq (\mathbf{P}^* - \epsilon\mathbf{v})^T(\lambda_j \odot \pi_t)$$

with equality iff $\mathbf{P}^* \odot (\lambda_j - \lambda_k) \odot \pi_t = 0$.

Proof Let $i \neq j$ be such that $\mathbf{P}^*$ is loss-neutral with respect to (t, i, j). Let $\mathcal{I}_+ = \{a \in \mathcal{A} : \mathbf{P}^* > 0\}$ denote the set of all symbols which are not assigned 0 probability by $\mathbf{P}^*$ and let

$$\mathcal{K} = \left\{ \mathbf{v} \in \mathbb{R}^M : \forall a \notin \mathcal{I}_+, \mathbf{v}[a] = 0, \sum_a \mathbf{v}[a] = 1 \right\}.$$

For all $\alpha \neq \beta \in \mathcal{A}$ let

$$\mathcal{H}_{\alpha,\beta} = \big\{ \mathbf{v} \in \mathcal{K} : \mathbf{v}^T((\lambda_\alpha - \lambda_\beta) \odot \pi_t) = 0 \big\}.$$

For all i', j' such that $\mathbf{P}^*$ is loss-neutral with respect to (t, i', j'), the dimension of the affine space $\mathcal{H}_{i',j'}$ is at least one less than the dimension of the space $\mathcal{K}$. Hence, if

$$\mathcal{L}_{\mathbf{P}^*,t} \stackrel{\text{def}}{=} \big\{ (i,j) \in \mathcal{A}^2 : i \neq j, \quad \mathbf{P}^* \text{ is loss-neutral with respect to } (t,i,j) \big\},$$

there exists $\mathbf{v}$, such that $\mathbf{P}^* + \mathbf{v} \in \mathcal{M} \cap \mathcal{K}$ and

$$\mathbf{P}^* + \epsilon \mathbf{v} \notin \bigcup_{(i,j) \in \mathcal{L}_{\mathbf{P}^*,t}} \mathcal{H}_{i,j}$$

for all sufficiently small non-zero $\epsilon \in \mathbb{R}$. For any $(i,j) \notin \mathcal{L}_{\mathbf{P}^*,t}$ such that $\mathbf{P}^* \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$,

$$(\mathbf{P}^*)^T((\lambda_i - \lambda_j) \odot \pi_t) \neq 0,$$

and hence for all sufficiently small ϵ, $\mathbf{P}^* + \epsilon \mathbf{v} \notin \mathcal{H}_{i,j}$. Therefore for all i, j, such that $\mathbf{P}^* \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$, $(\mathbf{P}^* + \epsilon \mathbf{v})^T((\lambda_i - \lambda_j) \odot \pi_t)$ is either strictly positive or negative and therefore there exists an $i_0(\epsilon)$ that satisfies for all $k \in \mathcal{A}$

$$(\mathbf{P}^* + \epsilon \mathbf{v})^T(\lambda_k \odot \pi_t) \geq (\mathbf{P}^* + \epsilon \mathbf{v})^T\big(\lambda_{i_0(\epsilon)} \odot \pi_t\big) \tag{53}$$

with equality iff $\mathbf{P}^* \odot (\lambda_{i_0(\epsilon)} - \lambda_k) \odot \pi_t = 0$. Clearly, for all sufficiently small $\epsilon > 0$, $i_0(\epsilon)$ is a constant, say i_0. In particular

$$\min_{k \in \mathcal{A}} (\mathbf{P}^*)^T(\lambda_k \odot \pi_t) = (\mathbf{P}^*)^T(\lambda_{i_0} \odot \pi_t). \tag{54}$$

Since $\mathbf{P}^*$ is loss-neutral, it is not hard to see that there exists a j that achieves the above minimum and $\mathbf{P}^* \odot (\lambda_{i_0} - \lambda_j) \odot \pi_t \neq 0$. Note that for all $\epsilon > 0$

$$(\mathbf{P}^* - \epsilon \mathbf{v})^T((\lambda_j - \lambda_{i_0}) \odot \pi_t) = -\epsilon \mathbf{v}^T((\lambda_j - \lambda_{i_0}) \odot \pi_t) < 0.$$

Therefore, for all sufficiently small $\epsilon > 0$, the $i_0(-\epsilon)$ that satisfies (53) when $\mathbf{P}^* + \epsilon \mathbf{v}$ is replaced by $\mathbf{P}^* - \epsilon \mathbf{v}$, is a constant j_0 satisfying $\mathbf{P}^* \odot (\lambda_{i_0} - \lambda_{j_0}) \odot \pi_t \neq 0$. In particular, like i_0, j_0 also satisfies (54). Therefore $\mathbf{P}^*$ is loss-neutral with respect to (t, i_0, j_0). □