

Lower Limits of Discrete Universal Denoising

Krishnamurthy Viswanathan and Erik Ordentlich

Hewlett Packard Labs, Palo Alto, CA 94304.

Email: krishnamurthy.viswanathan@hp.com, eord@hpl.hp.com

Abstract—Following recent work by Weissman *et al* , in the spirit of results on universal compression, we compare the performance of universal denoisers on discrete memoryless channels to that of the best 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 denoiser on a length- n sequence grows like $\Omega(c^k/\sqrt{n})$, where $c > 1$ 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) is no larger than a constant multiplicative factor of the best possible.

I. 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 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.

We present the required notation in Section II. Section III contains the main result of the paper as well as some of the preliminary results that lead to it. Section IV states the corresponding result for the benchmark considered in the

compound decision problem. We conclude with a discussion in Section V.

II. NOTATION

The notation used here is similar to the one in [1]. We first define the notation employed to refer to vectors, matrices and sequences. For two vectors \mathbf{u} and \mathbf{v} of the same dimension, $\mathbf{u} \odot \mathbf{v}$ will denote the vector obtained from componentwise multiplication. For any vector or matrix A , A^T will denote transposition.

For any set \mathcal{A} , \mathcal{A}^∞ denotes the set of one-sided infinite sequences with \mathcal{A} -valued components, *i.e.*, $\mathbf{a} \in \mathcal{A}^\infty$ is of the form $\mathbf{a} = (a_1, a_2, \dots)$, $a_i \in \mathcal{A}$, $i \geq 1$. For $\mathbf{a} \in \mathcal{A}^\infty$, let $a^n = (a_1, a_2, \dots, a_n)$ and $a_i^j = (a_i, a_{i+1}, \dots, a_j)$. More generally we will permit the indices to be negative as well, for example, $u_{-k}^k = (u_{-k}, \dots, u_0, \dots, u_k)$. 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 i th column of $\mathbf{\Pi}$. Let $M = |\mathcal{A}|$ denote the size of the alphabet and \mathcal{M} the simplex of M -dimensional probability vectors.

Upon observing a noisy sequence $Z^n \in \mathcal{A}^n$, the denoiser outputs a reconstruction sequence $\{\hat{X}_t\}_{t=1}^n \in \mathcal{A}^n$. The *loss matrix* $\mathbf{\Lambda} = \{\Lambda(i, j)\}_{i, j \in \mathcal{A}}$, whose components are also indexed by elements of \mathcal{A} , represents the loss function associated with the denoising problem, namely, $\Lambda(i, j)$ denotes the loss incurred by a denoiser when its output $\hat{X} = j$ when the channel input $X = i$. Also, for $i \in \mathcal{A}$, λ_i denotes the i th column of $\mathbf{\Lambda}$.

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

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

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, $f : \mathcal{A}^{2k+1} \rightarrow \mathcal{A}$ so that for all $z^n \in \mathcal{A}^n$

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

Let \mathcal{S}_k denote the class of k th-order sliding window denoisers. In the sequel we define the best loss obtainable for a given pair of noiseless and noisy sequences with a k -th order sliding window denoiser.

For 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

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

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 accommodate noiseless and noisy sequences of differing lengths. For a given channel $\mathbf{\Pi}$ and a noiseless sequence x^n define

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

to be 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 x^n , but not z^n . This benchmark has been generalized [4] to

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

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

$$\begin{aligned} \hat{D}_k(x^n) &= E \left[\min_{\hat{X}^n \in \mathcal{S}_k} L_{\hat{X}^n}(x_{k+1}^{n-k}, Z^n) \right] \\ &\leq \min_{\hat{X}^n \in \mathcal{S}_k} E[L_{\hat{X}^n}(x_{k+1}^{n-k}, Z^n)] = \bar{D}_k(x^n). \end{aligned} \quad (3)$$

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

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

and

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

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

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

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

$$\hat{R}_k(\hat{X}_{\text{univ}}^{n,k}) = \mathcal{O} \left(\sqrt{\frac{kM^{2k}}{n}} \right). \quad (4)$$

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

III. MAIN RESULT

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

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

where $c > 1$ is a constant that depends on the channel transition probability matrix $\mathbf{\Pi}$ and the loss function $\mathbf{\Lambda}$. This applies to all non-trivial $(\mathbf{\Pi}, \mathbf{\Lambda})$ pairs. We also derive similar results for $\bar{R}_k(\hat{X}^n)$. To derive these results we first require a preliminary result on denoisers that minimize expected loss when the noiseless sequence x^n is drawn according to a known *i.i.d.* distribution. This is presented in subsection III-A

A. Optimal denoisers for *i.i.d.* sequences

Given a noiseless random sequence X^n drawn according to a distribution \mathbf{P} let

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

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

$$\hat{X}_{\text{opt}}^n(z^n)[t] = \arg \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{P}_{X_t|z^n}$$

and the minimum expected loss is

$$D_{\text{opt}} = \frac{1}{n} \sum_{t=1}^n E \left[\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{P}_{X_t|Z^n} \right].$$

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

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

if $z_i = y_j$. In other words the denoiser defines a function $f : \mathcal{A} \rightarrow \mathcal{A}$.

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

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

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

and

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

B. Lower Bound on $\hat{R}_k(\hat{X}^n)$

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

$$\mathbf{P}^T (\lambda_i \odot \pi_t) = \mathbf{P}^T (\lambda_j \odot \pi_t) = \min_{k \in \mathcal{A}} \mathbf{P}^T (\lambda_k \odot \pi_t). \quad (5)$$

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 (5) 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.

It is shown in [5] that 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. Most commonly encountered non-trivial channels and loss functions are neutralizable, *e.g.*, the Binary Symmetric Channel and the Hamming loss function.

Example 1: Let the alphabet be $\mathcal{A} = \{0, 1\}$. 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 $0 < \delta < 1$, and let

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

represent the Hamming loss function. Indexing columns by elements of $\{0, 1\}$, $\pi_1 = [\delta \ 1 - \delta]^T$, $\lambda_0 = [0 \ 1]^T$ and $\lambda_1 = [1 \ 0]^T$. Choosing $\mathbf{P} = [1 - \delta \ \delta]^T$ we obtain

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

Furthermore $\mathbf{P} \odot (\lambda_0 - \lambda_1) \odot \pi_1 \neq 0$ Hence $(\mathbf{\Pi}_{\text{BSC}}, \Lambda_{\text{Ham}})$ is neutralizable and $\mathbf{P} = [1 - \delta \ \delta]^T$ is a loss-neutral distribution. Note that if $\delta \neq 1/2$, the uniform distribution is not loss-neutral. \square

For $x^n, z^n \in \mathcal{A}^n$, $c_{-k}^k \in \mathcal{A}^{2k+1}$ let $\mathbf{q}_{z^n, x^n}(c_{-k}^k)$ denote the M -dimensional column vector whose j -th component, $j \in \mathcal{A}$, is

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

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 $\mathbf{P} \in \mathcal{M}$ then Z^n , the noisy output from the channel, is also an *i.i.d.* sequence and

$$E[\mathbf{q}_{Z^n, X^n}(c_{-k}^k)] = (\mathbf{P} \odot \pi_{c_0}) \prod_{i=-k, i \neq 0}^k \mathbf{P}^T \pi_{c_i}. \quad (6)$$

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

$$\begin{aligned} D_k(x^n, z^n) &= \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}} \min_{\hat{x} \in \mathcal{A}} \sum_{j \in \mathcal{A}} \Lambda(j, \hat{x}) \mathbf{q}_{z^n, x^n}(c_{-k}^k)[j] \\ &= \sum_{c_{-k}^k \in \mathcal{A}^{2k+1}} \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}_{z^n, x^n}(c_{-k}^k). \end{aligned} \quad (7)$$

To prove the required result we characterize the asymptotic behavior of $\mathbf{q}_{Z^n, X^n}(c_{-k}^k)$ when X^n is drawn according to an *i.i.d.* distribution. Note that while Z^n is *i.i.d.*, for $k \geq 1$, $\mathbf{q}_{Z^n, X^n}(c_{-k}^k)$ cannot be written as a sum of *i.i.d.* random variables and therefore the standard Central Limit Theorem which can be employed to characterize the asymptotic behavior when $k = 0$ 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* [6]. 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, \dots, X_r) and $(X_s, X_{s+1}, \dots, X_n)$ are independent.

Theorem 2: [6] For a stationary and m -dependent sequence X^n of random variables such that $E[X_1] = 0$, and $E[|X_1|^3] < \infty$, as n tends to infinity $n^{-1/2} \sum_{i=1}^n X_i \xrightarrow{\mathcal{L}} \mathcal{N}(0, V)$ where $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution and $V = E[X_1^2] + 2 \sum_{i=2}^{m+1} E[X_1 X_i]$. \square

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

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 $c_{-k}^k \in \mathcal{A}^{2k+1}$, such that $\alpha^T (\mathbf{P} \odot \pi_{c_0}) = 0$

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

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 will 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, $\alpha^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k)$ suitably normalized tends in distribution to a Gaussian random variable, namely,

$$\sqrt{n} (\alpha^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V) \quad (8)$$

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 vector $\alpha \in \mathbb{R}^M$, indexed by members of \mathcal{A} , and $c_{-k}^k \in \mathcal{A}^{2k+1}$ we define the sequence Y_{k+1}^{n-k} to be

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

where $\alpha(\ell)$ denotes the ℓ -th component of α . Then

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

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,

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

where the last equality follows from the choice of α and c_0 . Furthermore, the higher moments of Y_i exist, hence Theorem 2 applies and therefore

$$\sqrt{n}(\alpha^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2) \quad (9)$$

where $\sigma^2 = E[Y_{k+1}^2] + 2 \sum_{i=1}^{2k} E[Y_{k+1} Y_{k+1+i}]$. Observe that

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

and it can be shown [5] that for all $i \geq 1$, $E[Y_{k+1} Y_{k+1+i}] = 0$. Substituting these in (9) establishes (8). A straightforward extension of (8) yields

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

where $G \sim \mathcal{N}(0, V)$. To complete the proof we use the fact (cf. e.g., Theorem 25.12, [7]) that if a sequence of random variables A^n is uniformly integrable and if $A_n \xrightarrow{\mathcal{L}} A$ then $\lim_{n \rightarrow \infty} E[A_n] = E[A]$, where in our case $A_n = \sqrt{n}|\alpha^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k)|$. It can be verified [5] that A^n is indeed uniformly integrable and therefore

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

Using Lemma 3 we derive a lower bound on $\hat{R}_k(\hat{X}^n)$ for all $\hat{X}^n \in \mathcal{D}_n$ in the following theorem. We lower bound the worst-case excess loss incurred by a denoiser over $\hat{D}_k(x^n)$ by the average excess loss incurred when X^n is drawn according to an *i.i.d.* distribution. This proof technique is similar to the one employed for the problem of binary prediction in [8]. However choosing a uniform distribution, like in [8], for X^n does not yield the required results - the distribution chosen has to be loss-neutral.

Theorem 4: For any neutralizable pair $(\mathbf{\Pi}, \Lambda)$, and any sequence $\{\hat{X}^n \in \mathcal{D}_n\}$ of denoisers, as n tends to infinity

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

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 $\mathbf{P}^* \in \mathcal{M}$ 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

$$(\mathbf{P}^*)^T (\lambda_i \odot \pi_t) = (\mathbf{P}^*)^T (\lambda_j \odot \pi_t) = \min_{k \in \mathcal{A}} (\mathbf{P}^*)^T (\lambda_k \odot \pi_t). \quad (10)$$

By definition

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

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

$$\hat{R}_k(\hat{X}^n) \geq E_{\mathbf{P}}[E[L_{\hat{X}^n}(X_1^n, Z^n)] - \hat{D}_k(X^n)]. \quad (11)$$

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

$$E_{\mathbf{P}^*}[E[L_{\hat{X}^n}(X_1^n, Z^n)]] \geq D_{\text{opt}} = \sum_{z \in \mathcal{A}} \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T (\mathbf{P}^* \odot \pi_z). \quad (12)$$

We upper bound $E_{\mathbf{P}^*}[\hat{D}_k(X^n)]$, the second term in (11). Applying (7), (6), the fact that the expectation of the minimum of a collection of random variables is lesser than the minimum of the expectations, and noting that

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

we obtain

$$\begin{aligned} E_{\mathbf{P}^*}[\hat{D}_k(X^n)] &= E_{\mathbf{P}^*} \left[\sum_{c_{-k}^k \in \mathcal{A}^{2k+1}} \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k) \right] \\ &\leq \sum_{c_{-k}^k \in \mathcal{S}_t} E_{\mathbf{P}^*} \left[\min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k) \right] \\ &\quad + \sum_{c_0 \neq t} \min_{\hat{x} \in \mathcal{A}} \lambda_{\hat{x}}^T (\mathbf{P}^* \odot \pi_{c_0}) \end{aligned} \quad (13)$$

where we use the abbreviation $\mathcal{S}_t = \{c_{-k}^k \in \mathcal{A}^{2k+1}, c_0 = t\}$. Substituting (13) and (12) in (11), combining with (10) 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 $\hat{X}^n \in \mathcal{D}_n$

$$\begin{aligned} \hat{R}_k(\hat{X}^n) &\geq \lambda_i^T (\mathbf{P}^* \odot \pi_t) \\ &\quad - \sum_{c_{-k}^k \in \mathcal{S}_t} E_{\mathbf{P}^*} [\min \{ \lambda_i^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k), \lambda_j^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k) \}]. \end{aligned} \quad (14)$$

Observe that (10) implies that

$$\begin{aligned} \lambda_i^T (\mathbf{P}^* \odot \pi_t) &= \sum_{c_{-k}^k \in \mathcal{S}_t} E_{\mathbf{P}^*} [\lambda_i^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k)] \\ &= \sum_{c_{-k}^k \in \mathcal{S}_t} E_{\mathbf{P}^*} [\lambda_j^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k)]. \end{aligned}$$

Writing $2 \min \{x, y\} = (x + y - |x - y|)$, and combining with the above equation we obtain

$$\begin{aligned} & \sum_{c_{-k}^k \in \mathcal{S}_t} 2E_{\mathbf{P}^*} [\min \{ \lambda_i^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k), \lambda_j^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k) \}] \\ &= 2\lambda_i^T (\mathbf{P}^* \odot \pi_t) - \sum_{c_{-k}^k \in \mathcal{S}_t} E_{\mathbf{P}^*} \left[\left| (\lambda_i - \lambda_j)^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k) \right| \right]. \end{aligned}$$

Substituting this in (14)

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

From (10) $(\lambda_i - \lambda_j)^T (\mathbf{P}^* \odot \pi_t) = 0$ and therefore applying Lemma 3 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

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sum_{c_{-k}^k \in \mathcal{S}_t} \sqrt{n} E_{\mathbf{P}^*} \left[\left| (\lambda_i - \lambda_j)^T \mathbf{q}_{Z^n, X^n}(c_{-k}^k) \right| \right] \\ &= V_t \sum_{c_{-k}^k \in \mathcal{S}_t} \prod_{\substack{i=-k, \\ i \neq 0}}^k ((\mathbf{P}^*)^T \pi_{c_i})^{\frac{1}{2}} = V_t \left(\sum_{a \in \mathcal{A}} ((\mathbf{P}^*)^T \pi_a)^{\frac{1}{2}} \right)^{2k} \end{aligned}$$

where $V_t = (2\pi^{-1}((\lambda_i - \lambda_j) \odot (\lambda_i - \lambda_j))^T (\mathbf{P}^* \odot \pi_t))^{\frac{1}{2}} > 0$ as $\mathbf{P}^* \odot (\lambda_i - \lambda_j) \odot \pi_t \neq 0$. \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 $(\mathbf{P}^*)^T \Pi$ is degenerate. Thus, if $(\mathbf{P}^*)^T \Pi$ is non-degenerate, the lower bound grows exponentially in k . This is the case for many (Π, Λ) , e.g., BSC and Hamming loss.

IV. THE COMPOUND DECISION BENCHMARK

In the compound decision problem, first proposed by Robbins [2], a sequence of n hypothesis tests each involving M possible hypotheses 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 passed through a memoryless channel Π . The M distributions in the hypothesis testing problem correspond to the M rows of Π . 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 $\bar{R}_0(\hat{X}^n)$.

Hannan and Van Ryzin [3] derived a scheme for the compound 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 $\hat{R}_k(\hat{X}_{\text{univ}}^{n,k})$ of the DUDE [1] was upper bounded as given in (4). From (3) this bound applies to $\bar{R}_k(\hat{X}_{\text{univ}}^{n,k})$ as well.

We derive lower bounds for $\bar{R}_k(\hat{X}^n)$ that scale like c^k/\sqrt{n} , for all denoisers and all neutralizable (Λ, Π) pairs where this constant c is smaller than the corresponding one in Theorem 4. 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 channels.

Theorem 5: For any neutralizable pair (Π, Λ) , and any sequence $\{\hat{X}^n \in \mathcal{D}_n\}$ of denoisers, as n tends to infinity

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

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

V. 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 applies not just to the worst-case sequence but also to the expected losses when the noiseless sequence X^n is drawn according to a loss-neutral distribution. Also, the results here are stated for fixed k and growing n . These can be extended to apply to $k = o(\log n)$ [5].

The results in this paper complement those derived in the semi-stochastic setting in [1]. Similar lower bounds can be derived for stochastic settings as well [5].

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*, vol. 51, no. 1, pp. 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*, vol. 26, pp. 37–51, 1955.
- [3] J. F. Hannan and J. V. Ryzin, ‘‘Rate of convergence in the compound decision problem for two completely specified distributions,’’ *Annals of Mathematical Statistics*, vol. 36, pp. 1743–1752, 1965.
- [4] M. V. J. Jr, ‘‘Two-action compound decision problems,’’ in *Proc. 5th Berkeley Symposium on Mathematical and Statistical Probability*, 1967, pp. 463–478.
- [5] K. Viswanathan and E. Ordentlich, ‘‘Lower limits of discrete universal denoising,’’ HP Laboratories, Tech. Rep. HPL-2006-71, Apr 2006.
- [6] W. Hoeffding and H. Robbins, ‘‘The central limit theorem for dependent random variables,’’ *Duke Math. Journal*, vol. 15, no. 3, pp. 773–780, 1948.
- [7] P. Billingsley, *Probability and Measure*. John Wiley and sons., 1986.
- [8] 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.