Efficient algorithms for discrete universal denoising for channels with memory

Ciprian Doru Giurcăneanu
Institute of Signal Processing
Tampere University of Technology
P.O. Box 553, FIN-33101 Tampere, Finland
Email: ciprian.giurcaneanu@tut.fi

Bin Yu
Department of Statistics
University of California at Berkeley
Berkeley, CA 94720-3860 USA
Email: binyu@stat.berkeley.edu

***Abstract*— The paper is focused on the problem of discrete universal denoising: one estimates the input sequence to a discrete channel based on the observation of the entire output signal, and without assuming any particular knowledge on the statistical properties of the input sequence. A $2k+1$ sliding window denoiser (DUDE) has recently been introduced, and its asymptotic optimality was proven in the case of memoryless channels and additive channels with memory. However, DUDE is computationally infeasible for large values of its context parameter k. The purpose of this paper is to further investigate DUDE in the case of channels with memory. First, for the important family of binary additive channels, we propose H-DUDE, a computationally feasible implementation of DUDE. It modifies the DUDE algorithm to exploit the property of the block transition probability matrix to be diagonalized by the Hadamard transform. H-DUDE accommodates large values of k, and we demonstrate this for the particular case of the finite-memory contagion channel. Second, we apply DUDE for a non-additive channel model that was previously used in design of stack filters to show its favorable performance.**

I. Introduction

Denoising is a time honored problem, and many approaches have been already proposed, but most of them assume that the models are either continuous or are known in the discrete case for both the signal and the noise. The problem of discrete universal denoising is to estimate the input sequence to a discrete channel based on the observation of the entire output signal, and without assuming any particular knowledge on the statistical properties of the input sequence. The state of the art algorithm for discrete universal denoising is dubbed DUDE and it was originally introduced in [1]. DUDE assumes that the channel is memoryless and known a priori. In [2] DUDE is extended for additive memory channels, and experimental results are reported for the particular case of burst noise channel [3][4]. We will use in the sequel the name DUDE for the denoising algorithm described in [2]. More theoretical results and extensions for this algorithm can be found in [5].

The kernel of DUDE is a $2k+1$ sliding window denoiser. Since the probabilities for the $2k+1$-tuple of the original signal are not known a priori, DUDE estimates them from the noisy signal. It was pointed out in [2] that increasing the value of k can improve the denoising performances, but the results are reported only for small values of k due to computational burdens. When the symbols at the input and the output of the channel are from a binary alphabet, the direct application of DUDE leads to the computation of the inverse for a $2^{2k+1} \times 2^{2k+1}$ matrix, which prevents the use of the denoiser for k larger than five. The solution proposed in [2] to lower the computational complexity relies on a supplementary hypothesis that the noiseless source is Markov with order no greater than $2k'$, where the value of k' is empirically chosen. Since it was already proven that DUDE is asymptotically optimal under mild conditions for the noise sequence [2], it is highly desirable to investigate experimentally the performances of DUDE for large values of the parameter k and for various channel models.

In this paper we assume that the symbols in the noiseless and noisy sequences are from the binary alphabet $\mathcal{A} = \{0, 1\}$. Our choice is supported by the fundamental role of binary case in discrete denoising. We show that for additive channels the computation of the inverse of a huge matrix can be circumvented by exploiting the property of the block transition probability matrix to be diagonalized by the Hadamard transform. To illustrate the procedure, we apply it for the finite-memory contagion channel [6], which was used during the last years in several joint source-channel coding studies, see for example [7] and the references therein. We investigate also the use of DUDE in the case when the original signal and the channel errors are assumed to be statistically independent, but the output of the channel is not simply the sum modulo two of the two sequences.

The rest of the paper is organized as follows. In Section II we revisit the DUDE algorithm for memory channels relying on results from [1] and [2]. For the family of binary additive channels, Section III-A outlines a modified DUDE algorithm that accommodates large values of the context parameter k. The newly proposed algorithm is applied in Section III-B to the particular case of the finite-memory contagion channel. To complete the analysis, we investigate in Section IV the DUDE performances for a non-additive binary channel.

II. The DUDE algorithm revisited

To outline the main steps of the algorithm we need to introduce some notations and definitions. For the underlying signal we employ the notation $X^n = X_1 \dots X_n$ when referring to the corresponding stochastic process, and $x^n =$

$x_1 \dots x_n$ when referring to a specific realization. Similarly the stochastic process that models the output of the channel is $Z^n = Z_1 \dots Z_n$, and a specific realization is denoted $z^n = z_1 \dots z_n$. We use the notation $\hat{X}^k_{z^n}$ for the DUDE denoiser, with the convention that $\hat{X}^k_{z^n}(t)$ is the estimated binary symbol for an arbitrary index t between 1 and n. We emphasis here that for each t, $\hat{X}^k_{z^n}(t)$ is evaluated only after the whole sequence z^n was received.

The input sequence is estimated by DUDE such that to minimize a loss criterion. To be more precise let us assume that $\mathbf{\Lambda}$ is a square matrix with nonnegative entries, where $\Lambda(i,j)$ is the loss incurred by estimating the symbol i with the symbol j. To simplify the notations, we assume that indices in vectors and matrices are beginning from zero. DUDE is designed to minimize the normalized total loss criterion $\frac{1}{n}\sum_{t=1}^{n}\Lambda\left(x_t, \hat{X}^k_{z^n}(t)\right)$. The $2k+1$ sliding window denoiser that solves the optimization problem is $\hat{X}^k_{z^n}(t) = \mathcal{D}^k_{z_n}(z^{t+k}_{t-k})$, and the definition of the mapping $\mathcal{D}^k_{z^n} : \mathcal{A}^{2k+1} \to \mathcal{A}$ is given by $\mathcal{D}^k_{z^n}(z^k_{-k}) = \arg\min_x \boldsymbol{\lambda}_x^\top \mathbf{P}_{X|z^k_{-k}}$, where $\boldsymbol{\lambda}_x$ is the column of $\mathbf{\Lambda}$ that corresponds to symbol x, and the symbol $^\top$ denotes transposition. $\mathbf{P}_{X|z^k_{-k}}$ is a column vector with two entries: $\mathbf{P}_{X|z^k_{-k}}(x) = P(X = x|Z^k_{-k} = z^k_{-k})\forall x \in \mathcal{A}$.

When the loss function is measured by the Hamming distance, the columns of $\mathbf{\Lambda}$ matrix are $\boldsymbol{\lambda}_0 = [0\ 1]^\top$ and $\boldsymbol{\lambda}_1 = [1\ 0]^\top$. With the convention that $\oplus$ denotes the addition modulo two, the following identities are readily obtained:

$$\begin{aligned}\mathcal{D}^k_{z_n}(z^k_{-k}) &= \arg\min_x P(X = (x\oplus 1)|Z^k_{-k} = z^k_{-k}) \\ &= \arg\max_x P(X = x|Z^k_{-k} = z^k_{-k}) \\ &= \arg\max_x \sum_{x^k_{-k}:x_0=x} \mathbf{P}_{X^k_{-k}|z^k_{-k}}(x^k_{-k}) \qquad (1)\end{aligned}$$

The column vector $\mathbf{P}_{X^k_{-k}|z^k_{-k}}$ has 2^{2k+1} entries such that for each index $i \in \{0,\dots,2^{2k+1}-1\}$, $\mathbf{P}_{X^k_{-k}|z^k_{-k}}(i) = P(X^k_{-k} = \underline{i}|Z^k_{-k} = z^k_{-k})$, where $\underline{i}$ is the binary representation of i written with $2k+1$ bits. Whenever necessary to refer to the binary representation of a nonnegative integer, we will underline its symbol.

Calculations based on Bayes' rule lead to

$$\mathbf{P}_{X^k_{-k}|z^k_{-k}} \propto \mathbf{P}_{z^k_{-k}|X^k_{-k}} \odot \mathbf{P}_{X^k_{-k}}, \qquad (2)$$

where $\propto$ indicates equality up to normalization and $\odot$ denotes componentwise multiplication of vectors. $\mathbf{P}_{X^k_{-k}}$ is the vector of probabilities for the $2k+1$-tuple of the noiseless signal, and its entries must be estimated from the received sequence z^n. The key role in this estimation is played by the block transition probability matrix $\mathbf{Q}$ with entries $Q(i,j) = P(Z^k_{-k} = \underline{j}|X^k_{-k} = \underline{i})$, where $\underline{i}$ and $\underline{j}$ are binary strings formed by the $2k+1$-bit representation of the nonnegative integers i and j, respectively. We remark that $\mathbf{P}_{z^k_{-k}|X^k_{-k}}$ is the column of $\mathbf{Q}$ with index z^k_{-k}, or equivalently we can write with Matlab-like notations that $\mathbf{P}_{z^k_{-k}|X^k_{-k}} = \mathbf{Q}(:, z^k_{-k})$. Under the hypothesis that $\mathbf{Q}$ is nonsingular, the distributions of $2k+1$-tuple for noiseless and noisy signals verify: $\mathbf{P}_{X^k_{-k}} = \mathbf{Q}^{-\top}\mathbf{P}_{Z^k_{-k}}$. The empirical distribution $\hat{\mathbf{P}}_{Z^k_{-k}}$ of the noisy $2k+1$-tuple can be easily obtained from z^n as

$$\hat{\mathbf{P}}_{Z^k_{-k}}(i) = \frac{1}{n-2k}\sum_{t=k+1}^{n-k}\mathbf{1}(z^{t+k}_{t-k}, \underline{i}) \qquad (3)$$

for each $i \in \{0,\dots,2^{2k+1}-1\}$. The function $\mathbf{1}(z^{t+k}_{t-k}, \underline{i})$ takes value one whenever the binary strings z^{t+k}_{t-k} and $\underline{i}$ are equal, and is zero otherwise. Consequently we can write $\hat{\mathbf{P}}_{X^k_{-k}} = \mathbf{Q}^{-\top}\hat{\mathbf{P}}_{Z^k_{-k}}$, and together with (2), we obtain:

$$\hat{\mathbf{P}}_{X^k_{-k}|z^k_{-k}} = \mathbf{Q}(:, z^k_{-k}) \odot \left[\mathbf{Q}^{-\top}\hat{\mathbf{P}}_{Z^k_{-k}}\right] \qquad (4)$$

The equations (1)-(4) suggest the steps of the denoising algorithm. During the first pass through data z^n, the number of occurrences of each binary context with length $2k+1$ is counted. The empirical distribution $\hat{\mathbf{P}}_{Z^{t+k}_{t-k}}$ is computed according to (3), and the output of the denoising mapping is evaluated by using (1) and (4) . In the second pass, the estimate $\hat{X}^k_{z^n}(t)$ is obtained for each t by observing the context z^{t+k}_{t-k} and applying the denoising mapping. We show in the next section how can be surmounted the difficulties related to the computation and storage of $\mathbf{Q}^{-\top}$.

III. Additive noise channel

A. Applying Hadamard transform in discrete denoising

We focus now on the case when the input-output dependence of the channel is described by $Z_t = X_t \oplus N_t$, where $t \in \{1,\dots,n\}$, and N^n is the noise process. We mention that this class of channel models is very important since encompasses famous models like the burst-error channel [3][4] and the finite-memory contagion channel [6]. The computational improvement that we propose for DUDE is based on the fundamental result below.

Proposition 3.1: [8] For any binary additive channel for which X^n and N^n are statistically independent, the block channel transition matrix $\mathbf{Q}$ has the following properties:

(a) $\mathbf{Q}$ is dyadic, i.e. two entries $Q(i,j)$ and $Q(i',j')$ are equal if $\underline{i}\oplus\underline{j} = \underline{i}'\oplus\underline{j}'$;

(b) The Hadamard transform of the column of $\mathbf{Q}$ indexed with zero is the vector having as entries the eigenvalues of $\mathbf{Q}$: $[\ell_0 \dots \ell_{2^{2k+1}-1}]^\top = \mathbf{H_{2^{2k+1}}}\mathbf{Q}(:,0)$, where $\mathbf{H}_{2^{2k+1}}$ is the Hadamard matrix of order 2^{2k+1};

(c) $\mathbf{Q} = \frac{1}{2^{2k+1}}\mathbf{H}_{2^{2k+1}} \times \mathbf{L} \times \mathbf{H}_{2^{2k+1}}$, where $\mathbf{L}$ is a diagonal matrix with entries $\ell_0,\dots,\ell_{2^{2k+1}-1}$.

Note that for each $k \geq 1$, $\mathbf{H}_{2^{2k+1}} = \underbrace{\mathbf{H}_2 \otimes \cdots \otimes \mathbf{H}_2}_{2k+1}$, where $\mathbf{H}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and $\otimes$ denotes the Kronecker product. We will use the notation $\mathcal{H}(\mathbf{v})$ for the Hadamard transform of an arbitrary vector $\mathbf{v}$. We outline in Figure 1 the new H-DUDE algorithm (H for Hadarmard). Since each dyadic matrix is symmetric, $\mathbf{Q}^\top$ is replaced with $\mathbf{Q}$ in the description of H-DUDE. Remark that the complexity of calculating $\mathbf{Q}^{-\top}\hat{\mathbf{P}}_{Z^k_{-k}}$

Input: The context size k and the noisy sequence z^n.
1. Apply (3) to calculate $\hat{\mathbf{P}}_{Z^k_{-k}}$;
2. Compute the entries of $\mathbf{Q}(:,0)$;
$[\ell_0 \dots \ell_{2^{2k+1}-1}]^\top = \mathcal{H}\left(\mathbf{Q}(:,0)\right)$;
$\hat{\mathbf{P}}_{X^k_{-k}} = \frac{1}{2^{2k+1}} \mathcal{H}\left([1/\ell_0 \dots 1/\ell_{2^{2k+1}-1}]^\top \odot \mathcal{H}\left(\hat{\mathbf{P}}_{Z^k_{-k}}\right)\right)$;
3. Evaluate the mapping $\mathcal{D}^k_{z^n}(z^k_{-k})$ for each context z^k_{-k}:
For $z^k_{-k} = 0 : 2^{2k+1}-1$,
    If $\hat{\mathbf{P}}_{Z^k_{-k}}(z^k_{-k}) > 0$,
        For $i = 0 : 2^{2k+1}-1$,
            $\mathbf{Q}(i, z^k_{-k}) = \mathbf{Q}(\underline{i} \oplus z^k_{-k}, 0)$;
        End
        $\hat{\mathbf{P}}_{X^k_{-k}|z^k_{-k}} = \mathbf{Q}(:, z^k_{-k}) \odot \hat{\mathbf{P}}_{X^k_{-k}}$;
        Apply (1) to compute $\mathcal{D}^k_{z^n}(z^k_{-k})$;
    End
End
4. Perform the denoising:
For t=k+1:n-k,
    Evaluate $\hat{X}^k_{z^n}(t) = \mathcal{D}^k_{z_n}(z^{t+k}_{t-k})$;
End

Fig. 1. H-DUDE: The DUDE algorithm modified to exploit the property of binary additive channels that the block transition matrix $\mathbf{Q}$ is diagonalized by the Hadamard transform. The channel is assumed to be known.

in H-DUDE is $\mathcal{O}(k2^{2k})$ instead of $\mathcal{O}(2^{6k})$ in DUDE, and this is the major computational improvement introduced by H-DUDE. Moreover, we observe from Figure 1 that H-DUDE does not need to store all 2^{4k+2} entries of the matrix $\mathbf{Q}$, but only the 2^{2k+1} entries of the vector $\mathbf{Q}(:,0)$.

B. *Finite-memory contagion channel*

We investigate the application of the H-DUDE algorithm for the finite-memory contagion channel [6]. The noise process is an M-th order Markov chain for which $P(N_t = 1|N^{t-1}_{t-M} = n^{t-1}_{t-M}) = \left(\varepsilon + \delta \sum_{i=1}^{M} n_{t-i}\right) / (1 + M\delta)$. The parameter ε is the channel bit error rate, $\varepsilon = P(N = 1)$, and takes values in the interval $(0, 1/2]$. The nonnegative parameter δ determines the correlation between errors at different time moments. More precisely, the correlation coefficient between N_t and N_{t-i} is given by $\frac{\delta}{1+\delta}$ for $i \in \{1, \dots, M\}$. M describes the range of dependence in the noise process. The larger the M, the longer the dependence range. δ is a measure on the strength of this dependence. When $\delta \to 0$, the channel reduces to the binary symmetric channel with parameter ε.

The finite-memory contagion channel has only three parameters (M, ε, δ) and theoretical results on its properties are already known. For example, closed form formulae are given in [8] for the entries of $\mathbf{Q}(:,0)$, and the computational complexity for each entry is linear in the context parameter k. We assess experimentally the performances of H-DUDE in four different scenarios. Since H-DUDE accommodates large values for k, the first goal of the experiments is to demonstrate the impact of k on the denoising results. We gain more insights by comparing the results of DUDE with those achieved by a $2k+1$ sliding window denoiser that has complete knowledge on the probability distribution for the $2k+1$-tuple of the underlying process X^n. Following the nomenclature from [2] we call this denoiser "genie-aided". The genie-aided denoiser outperforms all $2k+1$ sliding window denoisers that do no assume any knowledge on the source distribution. The difference between the algorithm listed in Figure 1 and the genie-aided denoiser consists in that the last one does not estimate from data the entries of the vector $\mathbf{P}_{X^k_{-k}}$. We extend the comparisons by reporting results obtained with the median filter.

In the first experiment, the samples $x_1, \dots, x_n$ are outcomes from a Bernoulli model with parameter $p \triangleq P(X = 0) = 0.95$, and $(M, \varepsilon) = (1, 0.01)$ for the finite-memory contagion channel. We test the algorithm for various values of the parameter δ that determines the correlation between two adjacent noise samples. It is proven in [9] that the singlet "say-what-you-see" is an optimal sequence maximum a posteriori (MAP) detection rule if $\delta \leq \frac{1-\varepsilon-p}{2p-1}$. In our settings, the condition becomes $\delta \leq 0.044$, and we choose to experiment with four different values of δ, namely $\delta \in \{0.5, 1, 2, 10\}$. For each value of δ, one single realization with length 10^6 is generated. The denoising results expressed in terms of bit error rate (BER) are shown in Figure 2. For easy comparison, we plot in the same figure the results achieved by H-DUDE, the genie-aided algorithm and the median filter, together with BER of the noisy sequence (raw data). For all denoisers, the window length parameter k is varied between one and seven, and the best BER is shown in Figure 2. Note that the median filter fails to denoise the observed sequence, and the results of H-DUDE are very close to those of the genie-aided algorithm. For better understanding this last fact, let us note that for all values of δ the best BER is obtained with the H-DUDE when $k = 4$, and the same is true for the genie-aided algorithm. The difference is that for the genie-aided algorithm the performances remain constant for all values of $k \geq 4$, while for H-DUDE the performances decrease when k takes values larger than four. Remark for both H-DUDE and genie-aided algorithm that the larger the value of δ, the better the performances. This observation agrees with the statement from [9] that the increase of noise correlation improves the results of the MAP detector when x^n is a Bernoulli string.

In the next experiments the underlying signal is a order-1 symmetric Markov chain with parameter $q \triangleq P(X_{t+1} = X_t) = 0.95$. For the results shown in Figure 3 the parameters of the finite memory contagion channel are $M = 1$ and $\varepsilon = 0.01$, the same as in the first experiment. In [9] it is proven that in the case of symmetric Markov sources, the singlet "say-what-you-see" is an optimal sequence MAP detector if $\delta \geq \frac{q}{1-q}\sqrt{\varepsilon(1-\varepsilon)} + \varepsilon - 1$. Since for $q = 0.95$ and $\varepsilon = 0.01$, the threshold is approximately 0.9, we evaluate the performances of the denoising algorithms for $\delta \in \{0.1, \dots, 0.8\}$. The same values of the correlation coefficient are used also for the results plotted in Figures 4 and 5. In all cases one single realization with length 10^6 is considered for each value of δ. We observe in Figure 3 that H-DUDE performs very similarly with the genie-aided algorithm. We note that like in the first experiment the best results are obtained for both denoisers with small values of the context parameter: $k = 2$ or $k = 3$. Unlike the first experiment, the median filter has good results achieved with the window parameter $k = 1$.

For the last two experiments, the channel parameter M is selected to be three, and due to this choice the context parameter k is varied between two and seven such that the context length, $2k + 1$, to be larger than the channel memory. The difference between the two experiments is that for Figure 4, $\varepsilon = 0.01$, and for Figure 5, $\varepsilon = 0.1$. We can observe in both Figures 4 and 5 a gap between the results obtained with H-DUDE and genie-aided algorithm. This is easy understandable if we note that the best result for the genie-aided algorithm is obtained for $k = 6$ or $k = 7$, and the H-DUDE performs best for $k = 4$.

Therefore when the channel parameter M is large, the DUDE has to work with large values of the context parameter k. We have shown that this is computationally practical, but the drawback is that for large k, even if the sample is 10^6, the number of occurrences of each $2k + 1$-tuple in z^n is not enough to ensure an accurate estimate $\hat{\mathbf{P}}_{Z_{-k}^k}$. The use of this faulty $\hat{\mathbf{P}}_{Z_{-k}^k}$ in (4) leads to moderate denoising results.

IV. Non-additive binary channel

We extend the analysis of performances for the case when the output of the channel is not calculated as $X_t \oplus N_t$, where X_t is the input symbol and N_t is the error. To illustrate this case, we resort to a channel model previously applied in stack filters design [10]. The noise process is modelled by a Markov order 1 chain whose state space is denoted $\mathcal{S} = \{0, 1, 2\}$, and for which the transition matrix is $\mathbf{A} = \begin{bmatrix} 0.1 & 0.3 & 0.6 \\ 0.1 & 0.1 & 0.8 \\ 0.2 & 0.2 & 0.6 \end{bmatrix}$, with entries $a_{ij} = P(S_{t+1} = j | S_t = i)$, $0 \le i, j \le 2$. The output of the channel coincides with its input whenever the noise state is $s = 2$. If the noise state s is 0 or 1, the value of the input symbol is ignored, and the channel output is s.

The key role in the application of the DUDE algorithm is played by the computation of the $\mathbf{Q}$ matrix and its inverse. To gain more insights on the differences between the additive channel discussed in the previous section, and the channel model described above, we consider the example of computing in both cases the probability of observing as output of the channel the sequence 011 when the input sequence is 010. For the additive channel the calculation reduces to the probability of the noise sequence $011 \oplus 010 = 001$. For the non-additive channel, the possible sequences of noise states that produce the desired output for the given input are: 011, 021, 211 and 221.

Generally, for a context size k, the entries of the block transition matrix $\mathbf{Q}$ have the expressions $Q(i, j) = P(Z_{-k}^k = \underline{j} | X_{-k}^k = \underline{i}) = \sum_{\underline{s} \in \mathcal{S}^{2k+1}} \pi_{s_0} \prod_{m=1}^{2k} a_{s_{m-1}, s_m} \prod_{m=0}^{2k} b_{s_m}(i_m, j_m)$, where π_{s_0} denotes the stationary distribution for an arbitrary initial state s_0. The mapping $b_{s_m}(i_m, j_m)$ takes either value zero or one: $b_{s_m}(i_m, j_m) = 1$ if and only if j_m is the channel output symbol when the input symbol is i_m and the noise state is s_m. Since the matrix $\mathbf{Q}$ is not dyadic, we apply in experiments the DUDE algorithm.

In line with the experiments reported in the previous section, we investigate the influence of the context size k, and plot in Figures 6 and 7 the BER for $k \in \{1, \ldots, 5\}$. The results obtained with DUDE and the genie-aided algorithm are very similar. Remark how the performances depend on the parameter $P(X_{t+1} = X_t)$ of the underlying Markov source.

V. Conclusion

Based on the Hadamard transform, a $2k+1$ sliding window denoiser, H-DUDE, was derived from the DUDE algorithm for the family of binary additive channel models. In contrast with the original DUDE, the new algorithm can be implemented with moderate computational resources even when the context parameter k is large. The major improvement is related to the computation of the inverse for the block channel transition matrix $\mathbf{Q}$: the complexity is reduced from $\mathcal{O}(2^{6k})$ to $\mathcal{O}(k2^k)$. Through experiments, the performances of H-DUDE were tested against the genie-aided $2k + 1$ sliding window denoiser and the median filter, for the case of finite-memory contagion channel. It was observed that the BER gap between the Hadamard transform-based algorithm and the genie-aided algorithm increases when the channel memory range M is larger, but the denoising results are superior to those achieved by the median filtering. The last remark can be easily explained since the median filter does not assume any knowledge on the statistics of the underlying signal, neither on the channel statistics. The experiments have been further extended for a non-additive channel model.

Acknowledgement

The work of C.D. Giurcăneanu was supported by Academy of Finland, project no. 44876. Bin Yu's research was partially supported by NSF grants CCR-0106656 and FD01-12731, ARO grant DAAD19-01-1-0643, and a Miller Research Professorship by Miller Institute at UC Berkeley in Spring 2004.

References

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

[2] R. Zhang and T. Weissman, "On discrete denoising for the burst-noise channel," in *Forty-second Annual Allerton Conf. on Communication, Control, and Computing*, Sep. 29 - Oct. 1 2004, http://www.stanford.edu/ tsachy/pdffiles/allertonpaper.pdf.

[3] E.O.Elliot, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, pp. 1977–1997, Sep. 1963.

[4] E. Gilbert, "Capacity of burst noise channels," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–1265, Sep. 1960.

[5] R. Zhang and T. Weissman, "Discrete denoising for channels with memory," Feb. 2005, preprint.

[6] F. Alajaji and T. Fuja, "A communication channel modeled on contagion," *IEEE Trans. Inform. Theory*, vol. 40, pp. 2035–2041, 1994.

[7] F.Behnamfar, F. Alajaji, and T. Linder, "Image transmission over the Polya channel via channel-optimized quantization," *IEEE Trans. Signal Processing*, vol. 53, no. 2, pp. 728 – 733, Feb. 2005.

[8] R. Iordache, I. Tabus, and J. Astola, "Robust index assignement using Hadamard transform for vector quantization transmission over finite-memory contagion channels," *Circuits, Systems, and Signal Processing*, vol. 21, no. 5, pp. 485–509, 2002.

[9] F. Alajaji, N. Phamdo, N. Farvardin, and T. Fuja, "Detection of binary Markov sources over channels with additive Markov noise," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 230–239, Jan. 1996.

[10] E. Coyle and J.-H. Lin, "Stack filters and the mean absolute error criterion," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 1244–1254, Aug. 1988.

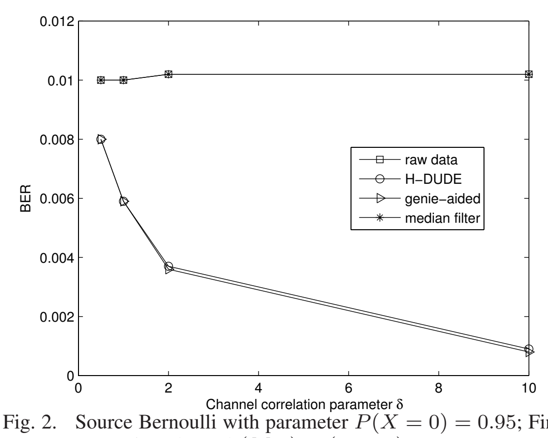

Fig. 2. Source Bernoulli with parameter $P(X=0)=0.95$; Finite-memory contagion channel $(M,\varepsilon)=(1,0.01)$.

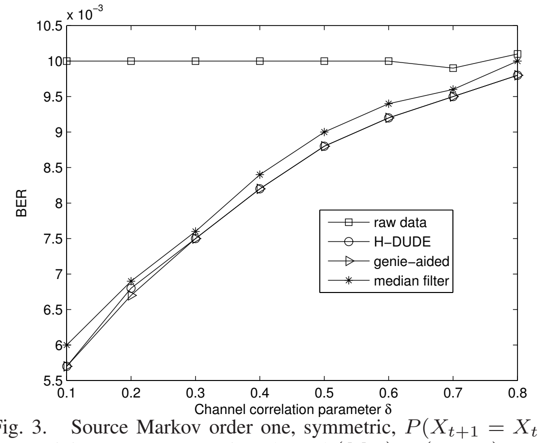

Fig. 3. Source Markov order one, symmetric, $P(X_{t+1}=X_t)=0.95$; Finite-memory contagion channel $(M,\varepsilon)=(1,0.01)$.

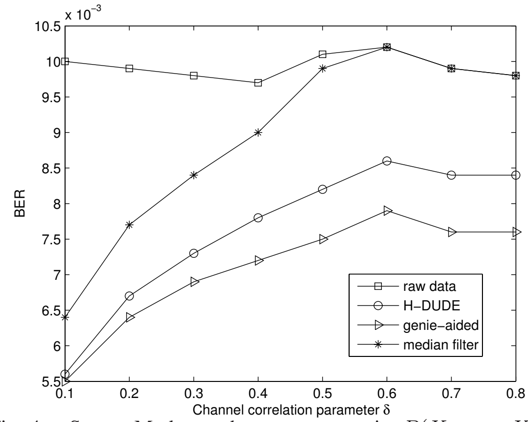

Fig. 4. Source Markov order one, symmetric, $P(X_{t+1}=X_t)=0.95$; Finite-memory contagion channel $(M,\varepsilon)=(3,0.01)$.

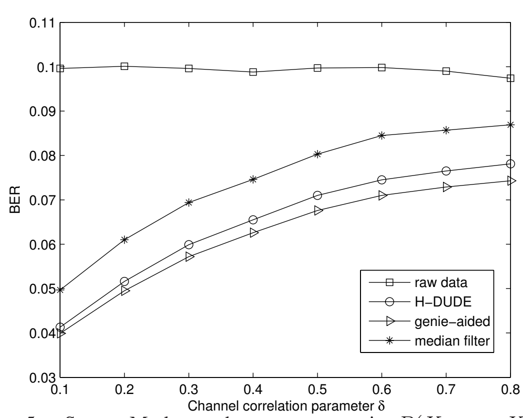

Fig. 5. Source Markov order one, symmetric, $P(X_{t+1}=X_t)=0.95$; Finite-memory contagion channel $(M,\varepsilon)=(3,0.1)$.

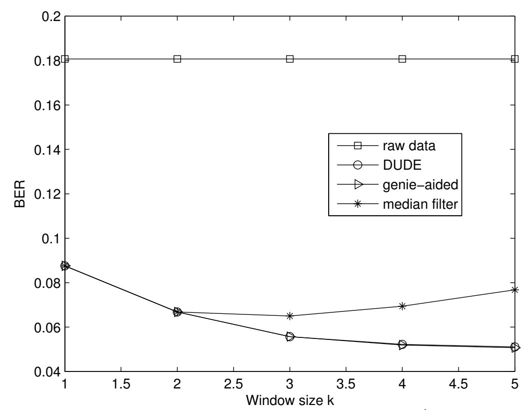

Fig. 6. Source Markov order one, symmetric, $P(X_{t+1}=X_t)=0.95$; Non-additive binary channel model.

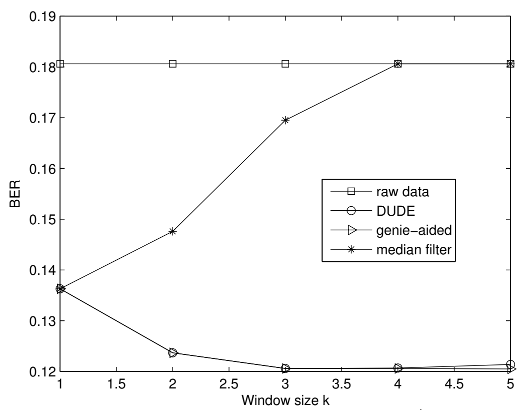

Fig. 7. Source Markov order one, symmetric, $P(X_{t+1}=X_t)=0.85$; Non-additive binary channel model.