

Universal Discrete Denoising

Tsachy Weissman Erik Ordentlich Gadiel Seroussi Sergio Verdú Marcelo Weinberger¹

Abstract — We propose a discrete denoising algorithm, that, based on the observation of the output of a known Discrete Memoryless Channel (DMC), estimates the input sequence to minimize a given fidelity criterion. The algorithm is universal in the sense that it requires no knowledge of the input sequence or its statistical properties. Yet, asymptotically it performs as well as the optimum denoiser that knows the input sequence distribution. The proposed denoising algorithm is practical, and can be implemented in $O(n \log n)$ time and $O(n^{2/3} \log n)$ storage complexity. Extensions to the case of delay-constrained denoising, and to the case of channel uncertainty, are briefly discussed.

I. INTRODUCTION

Consider the problem of recovering a finite-alphabet signal corrupted by a DMC. The problem arises in a variety of situations ranging from typing and spelling correction to Hidden Markov model state estimation; from DNA sequence analysis and processing to enhancement of facsimile and other binary images; from blind equalization problems to joint source-channel decoding when a discrete source is sent unencoded through a DMC.

In this extended abstract we shall primarily focus on the case of a known channel which is assumed invertible in the sense that the distribution of the channel output uniquely determines the distribution of the input. Since this can only be the case when the output alphabet is at least as large as the input alphabet, and since when the input alphabet is strictly smaller than the output alphabet it can be artificially padded with dummy symbols, there is no loss of generality in assuming, as we do henceforth, the following:

1. The components of the clean, as well as of the noise-corrupted signal, take their values in the same M -ary alphabet $\mathcal{A} = \{1, \dots, M\}$.
2. The channel matrix $\Pi = \{\Pi(i, j)\}_{i, j \in \mathcal{A}}$, $\Pi(i, j)$ denoting the probability of an output symbol j when the input is i , is invertible.

¹Part of this work was performed while S. Verdú was an HP/MSRI visiting research professor; he is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: verdu@princeton.edu). The other authors are with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA (e-mail: tsachyw@hpl.hp.com; eord@hpl.hp.com; seroussi@hpl.hp.com; marcelo@hpl.hp.com).

3. There is a given single-letter loss function (fidelity criterion) $\Lambda : \mathcal{A}^2 \rightarrow [0, \infty)$, $\Lambda(i, j)$ denoting the loss incurred by estimating the symbol i with the symbol j , with respect to which the goodness of the reconstruction is evaluated.

For concreteness we assume one-dimensionally indexed data, though all that will be said and done extends to the multi-dimensional case in a straightforward way.

The basic criteria with which to assess a universal discrete denoising algorithm are:

1. *Theoretical*. How favorably does the denoising performance compare with the fundamentally optimal performance attainable?
2. *Algorithmic*. Can the denoiser be practically implemented?

There is no universal discrete denoiser attaining the fundamentally optimum performance available in the literature. As far as the second criterion goes, the compression-based approach proposed in [2] (cf. also [1] and references therein) was shown to yield an implementable denoiser, empirically observed to perform well on various data sets.

The main contribution of this work is a denoising algorithm performing propitiously with respect to the two criteria. We shall propose and analyze an algorithm that is:

1. Asymptotically optimal in
 - (a) *The semi-stochastic setting*. In this setting, we make no assumption on a probabilistic or any other type of mechanism that may be generating the underlying clean signal and assume it to be an "individual sequence". The randomness in this setting is due solely to the channel noise. We show that our denoising algorithm is guaranteed to attain the performance of the best finite-order sliding-window denoiser in an almost sure sense, regardless of the underlying individual sequence.
 - (b) *The stochastic setting*. We show that our denoising algorithm asymptotically attains the performance of the optimal distribution-dependent scheme, for any stationary ergodic source that may be generating the underlying signal. This follows easily from the result in the semi-stochastic setting.

2. Practical. Implementation of the denoiser requires

- (a) Time complexity which grows as $O(n \log n)$ if n is the data size.
- (b) Storage complexity which is sub-linear in the size of the data.

The denoising algorithm we propose basically operates in two stages, making two passes through the noisy observation sequence. For a fixed length k , counts of the number of occurrences of all the strings of length $2k + 1$ appearing along the noisy observation sequence are accumulated in the first pass. The actual denoising is done in the second pass, where at each location along the noisy sequence an easily implementable metric computation is carried out (based on the known channel matrix, the loss function, and the counts from the previous pass) to determine whether the noisy symbol at that location should be changed and, if so, the symbol it should be changed to. A judicious choice of k (as a function of the size of the data set) yields a denoiser with the above properties.

The remainder of this abstract is organized as follows. Section II presents some notation and conventions. In Section III we introduce our denoising algorithm and detail its explicit form for a few special cases. Section IV gives an upper bound on the algorithm's time and storage complexity. Section V presents results assessing the performance of the proposed algorithm in the strong sense of the semi-stochastic setting, as well as the fully stochastic setting. Proofs of all stated results can be found in [5]. In Section VI we report the result of an initial experiment where our algorithm was employed on noisy data. Finally, in Section VII we briefly discuss extensions of the problem in two directions. The first is that of causal and delay-constrained denoising, where algorithms based on an idea similar to that underlying the denoising algorithm from previous sections are considered. The second is the case of channel uncertainty.

II. NOTATION AND CONVENTIONS

Throughout we let $\mathcal{A} = \{1, \dots, M\}$ and assume that the components of the clean signal, as well as those of the noisy observation sequence and the reconstruction sequence, take their values in \mathcal{A} . We let $\Lambda = \{\Lambda(i, j)\}_{i, j \in \mathcal{A}}$ denote the loss matrix. The i -th column of Π will be denoted by π_i and the j -th column of Λ by λ_j :

$$\Pi = [\pi_1 \mid \dots \mid \pi_M], \quad \Lambda = [\lambda_1 \mid \dots \mid \lambda_M].$$

For a vector or matrix Γ , Γ^T will denote transposition. For M -dimensional vectors \mathbf{u} and \mathbf{v} , $\mathbf{u} \odot \mathbf{v}$ will denote the vector obtained from componentwise multiplication.

We let \mathcal{A}^∞ 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 $\mathbf{a} = (a_1, a_2, \dots)$, $a_i \in \mathcal{A}$, $i \geq 1$. For $\mathbf{a} \in \mathcal{A}^\infty$ let $\mathbf{a}^n = (a_1, \dots, a_n)$ and $\mathbf{a}_i^j = (a_i, \dots, a_j)$. More generally, we will allow the indices of vector components to be negative as well, so, for example, $\mathbf{u}_{-k}^k = (u_{-k}, \dots, u_0, \dots, u_k)$. For positive integers k_1, k_2, k_3 and strings $s_i \in \mathcal{A}^{k_i}$, we let $s_1 s_2 s_3$ denote the string of length $k_1 + k_2 + k_3$ formed by concatenation.

For $\mathbf{a} \in \mathcal{A}^n$, $\mathbf{b}, \mathbf{c} \in \mathcal{A}^k$ let $\mathbf{m}[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ denote the M -dimensional column vector whose α -th component, $1 \leq \alpha \leq M$, is equal to

$$\mathbf{m}[\mathbf{a}, \mathbf{b}, \mathbf{c}](\alpha) = \sum_{i=k+1}^{n-k} \mathbf{1}_{\{a_{i-k}^{i-1} = \mathbf{b}, a_i = \alpha, a_{i+1}^{i+k} = \mathbf{c}\}}. \quad (1)$$

Since $\mathbf{m}[\mathbf{a}, \mathbf{b}, \mathbf{c}](\alpha)$ is the number of appearances of the string $\mathbf{b}\alpha\mathbf{c}$ along the sequence \mathbf{a} , the normalized (unit sum) version of the vector $\mathbf{m}[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ gives the empirical conditional distribution of a single letter given that the left context is the string \mathbf{b} and the right context is the string \mathbf{c} .

An n -block denoiser is a mapping $\hat{X}^n : \mathcal{A}^n \rightarrow \mathcal{A}^n$. We let $L_{\hat{X}^n}(x^n, z^n)$ denote the normalized cumulative loss, as measured by Λ , of the denoiser \hat{X}^n when the observed sequence is $z^n \in \mathcal{A}^n$ and the underlying one is $x^n \in \mathcal{A}^n$, i.e.,

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

with $\hat{X}_i^n(z^n)$ denoting the i -th coordinate of $\hat{X}^n(z^n)$.

III. THE UNIVERSAL DENOISER

A The Algorithm

For $\mathbf{a} \in \mathcal{A}^n$, $\mathbf{b}, \mathbf{c} \in \mathcal{A}^k$ and $\alpha \in \mathcal{A}$ let

$$g_{\mathbf{a}}^k(\mathbf{b}, \alpha, \mathbf{c}) = \arg \min_{\hat{\alpha} \in \mathcal{A}} \mathbf{m}^T[\mathbf{a}, \mathbf{b}, \mathbf{c}] \Pi^{-1} [\lambda_{\hat{\alpha}} \odot \pi_{\alpha}]. \quad (2)$$

For n and k let $\hat{X}^{n,k}$ denote the n -block denoiser given by

$$\hat{X}_i^{n,k}(z^n) = g_{z^n}^k(z_{i-k}^{i-1}, z_i, z_{i+1}^{i+k}) \quad k+1 \leq i \leq n-k. \quad (3)$$

The value of $\hat{X}_i^{n,k}(z^n)$ for $i \leq k$ and $i > n-k$ will be (asymptotically) inconsequential in subsequent development but for concreteness can be assumed to identically be given by the symbol¹ 1. Finally, for each n we define the n -block denoiser \hat{X}_{univ}^n by $\hat{X}_{\text{univ}}^n = \hat{X}^{n, k_n}$, where $k_n = \left\lceil \frac{\log n}{3 \log M} \right\rceil$.

B A few Special Cases

We detail the explicit form of the denoiser for a few cases of special interest. Hamming loss is assumed (with equal loss for any errors in the non binary case) in all the examples below.

- *Binary Symmetric Channel (BSC)*: For a BSC with crossover probability δ the denoiser assumes the form

$$g_{z^n}^k(\mathbf{u}_{-k}^{-1} 0 \mathbf{u}_1^k) = \begin{cases} 0 & \text{if } \frac{2\delta(1-\delta)}{(1-\delta)^2 + \delta^2} < \frac{\mathbf{m}[\mathbf{z}^n, \mathbf{u}_{-k}^{-1}, \mathbf{u}_1^k](0)}{\mathbf{m}[\mathbf{z}^n, \mathbf{u}_{-k}^{-1}, \mathbf{u}_1^k](1)} \\ 1 & \text{otherwise} \end{cases}$$

¹The choice of the symbol 1 is taken merely to emphasize its inconsequential nature. In practice a more judicious choice may be the optimal estimate of the channel input symbol based on its observed output, on a symbol-by-symbol basis.

and

$$g_{z^n}^k(u_{-k}^{-1}1u_1^k) = \begin{cases} 0 & \text{if } \frac{(1-\delta)^2 + \delta^2}{2\delta(1-\delta)} < \frac{m[z^n, u_{-k}^{-1}, u_1^k](0)}{m[z^n, u_{-k}^{-1}, u_1^k](1)} \\ 1 & \text{otherwise.} \end{cases}$$

- *The Z Channel:* The channel Matrix, and its inverse, for this case are given by

$$\Pi = \begin{pmatrix} 1-\delta & \delta \\ 0 & 1 \end{pmatrix}, \quad \Pi^{-1} = \begin{pmatrix} \frac{1}{1-\delta} & \frac{-\delta}{1-\delta} \\ 0 & 1 \end{pmatrix}.$$

Since only locations i where $z_i = 1$ may need correction, we are only interested in evaluation of $g_{z^n}^k(u_{-k}^{-1}1u_1^k)$ for contexts of the form $u_{-k}^{-1}1u_1^k$. Plugging into (2) easily gives

$$g_{z^n}^k(u_{-k}^{-1}1u_1^k) = \begin{cases} 0 & \text{if } \frac{1-\delta}{2\delta} < \frac{m[z^n, u_{-k}^{-1}, u_1^k](0)}{m[z^n, u_{-k}^{-1}, u_1^k](1)} \\ 1 & \text{otherwise.} \end{cases}$$

- *The Erasure Channel:* For the alphabet $\mathcal{A} = \{1, \dots, M\}$ and erasure probability δ we obtain

$$g_{z^n}^k(u_{-k}^{-1}e u_1^k) = \arg \max_{\hat{x} \in \{1, \dots, M\}} m[z^n, u_{-k}^{-1}, u_1^k](\hat{x}).$$

Note that this denoiser does not depend on the channel parameter δ !

IV. COMPLEXITY OF THE UNIVERSAL DENOISER

Our implementation of the universal denoiser $g_{z^n}^k(\cdot)$ makes two passes through the observations z^n . The empirical counts $m[z^n, u_{-k}^{-1}, u_1^k](u_0)$, for the various strings u_{-k}^k appearing along the sequence z^n , are accumulated and stored in the first pass while the actual application of $g_{z^n}^k(\cdot)$, as determined by the accumulated statistics, is performed in the second pass. Execution of the first pass is similar in spirit to practical implementations of context-based universal data compression algorithms (cf., e.g., [3]). The time complexity can be shown (cf. [5, Subsection 3.B]) to be generously upper bounded by $c_1 2kn + c_2 M^2 \min(M^{2k}, n)$ and the storage complexity (in bits) by $c_3 (\log_2 n + \log_2 [\min(M^{2k}, 2kn)]) M \min(M^{2k+1}, 2kn)$, c_1, c_2 and c_3 being implementation dependent constants that are independent of k, n , and M . Taking $k = k_n = \lceil \frac{\log n}{3 \log M} \rceil$ gives a time complexity of $O(n \log n)$ and storage complexity of $O(n^{2/3} \log n)$.

V. OPTIMALITY OF THE DENOISER

A The Semi-Stochastic Setting

For $\mathbf{x}, \mathbf{z} \in \mathcal{A}^\infty$, $k \geq 0$, and $n > 2k$, we define the k -th order sliding-window minimum loss of (x^n, z^n) by

$$D_k(x^n, z^n) = \min_{f: \mathcal{A}^{2k+1} \rightarrow \mathcal{A}} \left[\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(x_i, f(z_{i-k}^{i+k})) \right]$$

and that of (\mathbf{x}, \mathbf{z}) by

$$D_k(\mathbf{x}, \mathbf{z}) = \limsup_{n \rightarrow \infty} D_k(x^n, z^n).$$

Finally, we define the sliding-window minimum loss by

$$D(\mathbf{x}, \mathbf{z}) = \lim_{k \rightarrow \infty} D_k(\mathbf{x}, \mathbf{z}).$$

Note that $D_k(\mathbf{x}, \mathbf{z})$ is non-increasing with k so that $D(\mathbf{x}, \mathbf{z})$ is well-defined.

By *Semi-Stochastic Setting* we refer to the case where there is an individual sequence \mathbf{x} and \mathbf{Z} is its noise-corrupted version, i.e., \mathbf{Z} is the output of the memoryless channel, Π , whose input is \mathbf{x} . In the statement of the theorem that follows we assume the semi-stochastic setting. Our discrete denoiser is optimal in the following sense.

Theorem 1 *The sequence of denoisers $\{\hat{X}_{\text{univ}}^n\}$ satisfies*

$$\limsup_{n \rightarrow \infty} L_{\hat{X}_{\text{univ}}^n}(x^n, Z^n) \leq D(\mathbf{x}, \mathbf{Z}) \quad \text{a.s. } \forall \mathbf{x} \in \mathcal{A}^\infty. \quad (4)$$

Note that $D(\mathbf{x}, \mathbf{Z})$, on the right side of (4), is a random variable, dependent on the noise realization. It can easily be shown, however, to be a degenerate one in the sense of being a.s. equal to a deterministic constant dependent, of course, on the individual sequence \mathbf{x} and on the channel Π (cf. [5, Claim 1]).

B The Stochastic Setting

Assume now that \mathbf{Z} is the output of the memoryless, invertible, channel Π whose input is the double-sided stationary ergodic \mathbf{X} . Letting $P_{X^n}, P_{\mathbf{X}}$ denote, respectively, the distributions of X^n, \mathbf{X} , and \mathcal{D}_n denote the class of all n -block denoisers, we define

$$\mathbb{D}(P_{X^n}, \Pi) = \min_{\hat{X}^n \in \mathcal{D}_n} EL_{\hat{X}^n}(X^n, Z^n),$$

the expectation on the right side assuming $X^n \sim P_{X^n}$ and

$$\mathbb{D}(P_{\mathbf{X}}, \Pi) = \lim_{n \rightarrow \infty} \mathbb{D}(P_{X^n}, \Pi). \quad (5)$$

The limit in (5) can be shown to exist and $\mathbb{D}(P_{\mathbf{X}}, \Pi) = \inf_{n \geq 1} \mathbb{D}(P_{X^n}, \Pi)$. Thus, $\mathbb{D}(P_{\mathbf{X}}, \Pi)$ is the fundamentally optimal asymptotic denoising performance attainable when the clean signal is emitted by the source $P_{\mathbf{X}}$ and corrupted by the channel Π . The following can be shown to be a consequence of Theorem 1.

Corollary 1 *The sequence of denoisers $\{\hat{X}_{\text{univ}}^n\}$ satisfies, for all stationary ergodic processes \mathbf{X} ,*

$$\limsup_{n \rightarrow \infty} L_{\hat{X}_{\text{univ}}^n}(X^n, Z^n) \leq \mathbb{D}(P_{\mathbf{X}}, \Pi) \quad \text{a.s.}$$

and

$$\lim_{n \rightarrow \infty} EL_{\hat{X}_{\text{univ}}^n}(X^n, Z^n) = \mathbb{D}(P_{\mathbf{X}}, \Pi). \quad (6)$$

An upper bound on the rate of convergence in (6) can be found in [5, Corollary 2].

VI. EMPIRICAL RESULTS

The described denoising scheme has been implemented in practice, and preliminary experiments have been run with simulated sources where the performance of the optimal distribution-dependent denoiser can be accurately estimated. For one typical example, the proposed scheme was tested on a binary Markov source with transition probability 0.01 corrupted by a BSC(0.1), for a moderate sequence length of $n = 10^6$. The optimal distribution-dependent denoiser for this case, implemented by the well-known forward-backward recursions, made 6081 errors. Our algorithm made 6535 errors, approximately 7.5% more. More extensive experiments on different types of simulated and real-life sources are under way.

VII. EXTENSIONS

A Causal and Delay-Constrained Denoisers

Define a *delay- d denoiser* to be a sequence of functions $\hat{X} = \{\hat{X}_t\}_{t \geq 1}$, where $\hat{X}_t : \mathcal{A}^{t+d} \rightarrow \mathcal{A}$. For each point in time, t , the delay- d denoiser outputs a reconstruction for X_t based on observing Z^{t+d} , namely, $\hat{X}_t(Z^{t+d})$. For general positive integers n, k, d , and $\mathbf{a} \in \mathcal{A}^n, \mathbf{b} \in \mathcal{A}^k, \mathbf{c} \in \mathcal{A}^d$, let $m[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ and $g_{\mathbf{a}}^{k,d}(\mathbf{b}, \mathbf{a}, \mathbf{c})$ denote the obvious extensions of the definitions in (1) and (2) to accommodate the possibility $k \neq d$. Consider now the delay- d denoiser \hat{X}^k given for $t > k$ by

$$\hat{X}_t^k(Z^{t+d}) = g_{z_{t+d}}^{k,d}(z_{t-k}^{t-1}, z_t, z_{t+1}^{t+d}) \quad (7)$$

and arbitrarily defined for $t \leq k$. It can be shown that

1. $\{\hat{X}^k\}$ is an asymptotically optimal sequence of delay- d denoisers in the sense that

$$\lim_{k \rightarrow \infty} \lim_{n \rightarrow \infty} E \left[\frac{1}{n} \sum_{t=1}^n \Lambda(X_t, \hat{X}_t^k(Z^{t+d})) \right]$$

is equal to the minimum asymptotic denoising loss attainable by a delay- d denoiser, for all stationary ergodic sources that may be generating the clean sequence.

2. \hat{X}^k has a low complexity implementation which is similar to (and simpler than) that of the universal denoiser considered above. Here both acquisition of the statistics and the actual denoising can be performed sequentially in one pass.

Note that the statement in the first item is not as satisfying as having one denoiser asymptotically attaining optimal performance. A delay-constrained denoiser of the latter type, based on the mixture approach, can easily be constructed using [4, Theorem 5]. Such a denoiser, however, would be prohibitively complex. The question of the existence of a denoiser with the stronger property, implementable with low complexity, is under current investigation.

B Channel Uncertainty

For brevity, we limit the discussion here to binary sequences corrupted by a BSC. More concretely, suppose throughout this section that the clean binary sequence of interest is stationary ergodic and known to be corrupted by a BSC(δ), where $\delta \in [0, 1/2]$ is unknown. If P_X denotes the distribution of the noise-free process, we let $P_X * \delta$ denote the (stationary ergodic) channel output distribution. Let also $\mathbb{D}(P_X, \delta)$ denote $\mathbb{D}(P_X, \Pi)$ when Π is the channel matrix of the BSC(δ).

Definition 1 *The family of binary processes Ω will be said to be noble if for all $P_X, P'_X \in \Omega$ ($P_X \neq P'_X$)*

$$P_X * \delta \neq P'_X \quad \forall \delta \in [0, 1/2].$$

It is not hard to show that a family Ω being noble excludes also the possibility of the existence of (distinct) $P_X, P'_X \in \Omega$ and $\delta, \delta' \in [0, 1/2]$ such that $P_X * \delta \neq P'_X * \delta'$.

As we show in [6], *the nobility of Ω is a sufficient condition for the existence of a (sequence of) denoiser(s) which is universal in the sense of attaining $\mathbb{D}(P_X, \delta)$ for all $P_X \in \Omega$ and $\delta \in [0, 1/2]$* . A necessary condition is also provided in [6], which coincides with the nobility of Ω for non-degenerate cases.

One class of sources shown in [6] to be noble is that of *non-i.i.d* k -th order Markov processes.

This problem is also connected to rate-distortion theory: If the Shannon Lower Bound (SLB) is *not* tight (at any positive distortion) for the sources in Ω , then Ω is noble. For example, any source with restricted sequences, i.e., any source for which there exist one or more finite strings whose probability of appearance is zero, has this property. Thus, a class of sources with restricted sequences is noble. As we show in [6], if Ω is known to contain only sources for which the SLB is not tight, there exists a denoising algorithm which is a modified version of the one presented here, attaining $\mathbb{D}(P_X, \delta)$ for all $P_X \in \Omega$ and $\delta \in [0, 1/2]$.

REFERENCES

- [1] D. Donoho, "The Kolmogorov Sampler," January 2002 (preprint available at: <http://www-stat.stanford.edu/donoho/>).
- [2] B. Natarajan, "Filtering random noise from deterministic signals via data compression," *IEEE Trans. Signal Proc.*, 43(11):2595-2605, November 1995.
- [3] M. J. Weinberger, J. Rissanen, and R. Arps, "Application of universal context modeling to lossless compression of gray-scale images", *IEEE Trans. Image Processing*, 5(4):575-588, April 1996.
- [4] T. Weissman and N. Merhav, "Finite-delay lossy coding and filtering of individual sequences corrupted by noise", *IEEE Trans. Inform. Theory*, 48(3):721-733, March 2002.
- [5] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal Discrete Denoising: I. Known Channel", Preprint available, August 2002.
- [6] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal Discrete Denoising: II. Channel Uncertainty", In preparation, 2002.