Reflections on the DUDE

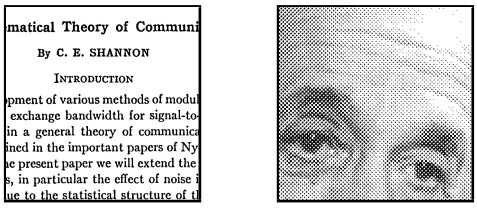
E. Ordentlich, G. Seroussi, S. Verdú, M. J. Weinberger, and T. Weissman

Dude:

1. *An inexperienced cowboy.*
2. *(slang) A man.*
3. *(slang) A term of address for a man.*
4. *(archaic) A dandy, a man who is very concerned about his dress and appearance.*
5. *(slang) A cool person of either sex.*
6. *(algorithm) Discrete Universal Denoiser*

1 Denoising

Suppose the binary images shown in Figure 1 were transmitted over a binary symmetric channel (BSC). What could be done to clean up or denoise the resulting noisy images? A quick scan of the image processing literature suggests the use of such algorithms as median filtering and morphological filtering. In the binary case a median filter replaces each pixel in the noisy image by the color of the majority of pixels in a neighborhood about that pixel, while a morphological filter carries out dilation and erosion operations to denoise a pixel, again based on the values of nearby pixels. Indeed, many of the simplest and most practical denoising algorithms in the image processing literature are of this form. They denoise any given pixel by applying a function to the pixel values in a nearby neighborhood. The denoising scenario above, however, raises an important issue with this class of algorithms, namely that a neighborhood function that works well for denoising one image may be disastrous for another. For example, a median filter would be an excellent choice for denoising the document image of Figure 1, but would obliterate the half-toning in the Einstein image, and thereby actually amplify the distortion. Thus, in general, we are faced with the following problem: Given a noisy binary image and a neighborhood size of k, which of the 2^{2^k} binary valued denoising functions should be used? If a genie were to reveal the underlying clean image we could determine the "best" function that would lead to the smallest distortion between the denoised and clean images. Can we hope to even come close to this level of performance with


matical Theory of Communi

By C. E. SHANNON

Introduction

pment of various methods of modul
exchange bandwidth for signal-to-
in a general theory of communica
ined in the important papers of Ny
he present paper we will extend the
s, in particular the effect of noise i
ue to the statistical structure of th

Figure 1: Binary images

access only to the noisy image? It turns out that with the additional knowledge of the BSC crossover probability, the answer is yes in a very strong sense, and, further, that it can be done using an eminently practical algorithm, which we call DUDE, as defined and analyzed in our paper [1].

Beyond binary images, denoising is a problem that permeates just about all branches of science and engineering involving some form of inference on a phenomenon of interest, based on noisy or incomplete observations. As we say in [1], denoising "has received significant attention for over half a century...Wiener...Kalman...Donoho and Johnstone...the amount of work and literature in between is far too extensive even to be given a representative sample of references", and we shall certainly not attempt to do it any more justice here.

In contrast to the majority of these works, our paper [1] approaches denoising from a universality perspective, as motivated by the above image denoising scenario. Indeed, one can view the framework of [1] as a broadening of the rich information theoretic framework of universality formalized in such works as [2, 3] on the Lempel-Ziv compression algorithm, [4] on universal modeling and arithmetic coding for tree sources, [5] on universal prediction, or more general decision problems as surveyed in [6]. The notion of universality in information theory, particularly the so called individual sequence formulation exemplified above, has much in common with work from the 1950's on the compound decision problem (starting with the seminal work [7]), its sequential version (e.g., [8, 9], cf. [10] for a comprehensive account of this literature), and on the repeated play of games [11].

In addition to the individual sequence notion of universality pertaining to the scenario above, we can motivate the DUDE in a distributional setting as well. Suppose that the noiseless source is stochastic and suppose first that its distribution is given. In order to come up with its reconstruction sequence, the denoiser minimizing the expected cumulative loss (as measured by the given loss function) needs to compute the posterior distributions of each of the sequence components, given the noisy sequence it observes. Two salient aspects of the computation of these posterior distributions are their dependence on the distribution of the noiseless source, and the high complexity (exponential in the data size) associated with their computation. Thus, in general, beyond simple situations where the noiseless source distribution is memoryless or Markov (in which case the posterior distributions can be practically computed via the "forward-backward" dynamic programming [12]), it is impractical to implement the optimal denoiser for a given fully specified source. A seemingly logical conclusion then, which seems to have been implicit in the literature prior to our work, is that it is a fortiori not practical to attain optimum performance for large data size when the noiseless source is *not* specified. Fortunately, this conclusion turned out to be false, as long as we are willing to settle for *asymptotic* optimality: the DUDE (described below) does not depend on the source distribution (it is universal) and is practical (having linear complexity), while attaining the performance of the optimal denoiser for large data size.

2 DUDE

A Problem setting

Our setting for the discrete denoising problem is shown in Figure 2 for one-dimensional data. A discrete source emits a *clean* sequence $\mathbf{X} = X_1, X_2, \ldots, X_n$ of data symbols over a finite alphabet A. The clean sequence is transmitted over a discrete memoryless noisy channel, characterized by a

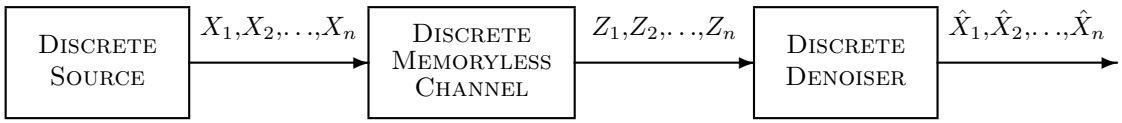


Figure 2: Discrete denoising setting

matrix of cross-over probabilities $P(\text{output}{=}z \,|\, \text{input}{=}x)$ for each pair (x, z) of symbols from A (here we assume, for simplicity, that the channel input and output alphabets are the same; a more general setting is considered in [1]). The channel output is a *noisy* sequence $\mathbf{Z} = Z_1, Z_2, \ldots, Z_n$, which is fed to a *discrete denoiser*. The denoiser, in turn, produces an estimate, $\hat{\mathbf{X}} = \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$, of the original clean sequence. Denoising performance is measured with a given (but otherwise arbitrary) single-letter *loss function* which judges how close the reproduction $\hat{\mathbf{X}}$ is to the clean sequence $\mathbf{X}$. For example, if the loss function is defined as the normalized Hamming distance between $\mathbf{X}$ and $\hat{\mathbf{X}}$, the overall loss incurred by the denoiser is the fraction of symbols of $\mathbf{X}$ that were not reconstructed perfectly. The goal of the denoiser is to produce an estimate $\hat{\mathbf{X}}$ that minimizes the given loss function.

To carry out its task, the denoiser has access *only* to the noisy sequence $\mathbf{Z}$. It has no information whatsoever on the clean sequence $\mathbf{X}$, its probability distribution, or even whether it has one.

On the face of it, the problem appears particularly difficult, or even ill-posed, as the denoiser must minimize a loss function it cannot measure. This hurdle distinguishes denoising from problems such as prediction or compression, where the data processing algorithm (predictor or compressor) can keep track of how well it is performing. But appearances are deceiving, and it turns out that a fairly simple algorithm can perform the denoising task very well, in fact, optimally well.

B The algorithm

The DUDE makes two passes over the data. In the first pass, it slides a window of length $2k{+}1$ over the sequence $\mathbf{Z}$, where k is a nonnegative integer we assume given for the time being. At time t, and ignoring border effects, the window contains the string $Z_{t-k} \ldots Z_{t-1} Z_t Z_{t+1} \ldots Z_{t+k}$. The symbol Z_t is said to appear in (*two-sided*) *context* $\mathbf{C}_t = [Z_{t-k} \ldots Z_{t-1} \,\square\, Z_{t+1} \ldots Z_{t+k}]$. The DUDE keeps counts of symbol occurrences in each context, incrementing, at time t, the count corresponding to the symbol Z_t in context $\mathbf{C}_t$. If the sequence is sufficiently long, symbol patterns will repeat, and we will have $\mathbf{C}_{t'} = \mathbf{C}_t$ for other (hopefully, many) values t'. At the end of the pass, we will have obtained a *conditional empirical distribution* $\hat{P}(Z|\mathbf{C}_t)$ for each context pattern encountered. Consider, for example, the noisy text in Figure 3. The figure shows all the occurrences of the $k{=}1$ context pattern $\mathbf{C}{=}[\sqcup\, \square\, \mathrm{i}]$ ($\sqcup$ represents a space). The corresponding conditional empirical distribution is

$$\hat{P}(\mathrm{w}|\mathbf{C}){=}3/7, \;\; \hat{P}(\mathrm{g}|\mathbf{C}){=}2/7, \;\; \hat{P}(\mathrm{y}|\mathbf{C}){=}1/7, \;\; \hat{P}(\mathrm{m}|\mathbf{C}){=}1/7, \;\; \hat{P}(\text{other}|\mathbf{C}){=}0.$$

As in other applications of context modeling, the goal of collecting conditional statistics is to capture high-order dependencies, which reveal structure in the data. The statistics collected in the first pass of the DUDE estimate the conditional probability of *noisy* symbols Z_t given their *noisy* contexts $\mathbf{C}_t$. Our goal is to estimate *clean* symbols X_t, given the observed sequence $\mathbf{Z}$. An important step towards this goal would be to obtain, for each t, an estimate of the conditional

"Whar g i ants?" said Sancho Panza. "Those thou seest theee," snswered y i s master, w i th the long arms, and spne have tgem ndarly two leagues long." "Look, ylur worship," sair Sancho; "what we see there zre not g i anrs but w i ndmills, and what seem to be their arms are the sails that turned by the w i nd make rhe m i llstpne go."

Figure 3: Contexts in noisy text, with $k = 1$ (a z b : sample z in context [a$\square$b]).

probability of the clean symbol X_t given the noisy window $\mathbf{C}_t^+ = [\mathbf{C}_t, Z_t]$ (the sliding window including the center sample). What we hope for is that the conditional statistics gathered from $\mathbf{Z}$ still allow us to glean some of the structure present in $\mathbf{X}$ (if any), which in turn could help us make good denoising decisions. But how reliable can the estimate of the conditional distribution of clean symbols be? The conditional structure of $\mathbf{X}$ is fogged by the noise in two ways: on one hand, we are taking counts of corrupted samples ("counting the wrong symbol in the right context"), and on the other hand, symbols that were in the same context in $\mathbf{X}$ might be scattered in different contexts in $\mathbf{Z}$, since the context patterns are also noisy ("counting the right symbol in the wrong context"). As it turns out, by requiring a mild non-degeneracy condition on the channel, the estimates of conditional distributions of clean symbols that can be derived from the statistics collected in the first pass are "reliable enough," in a well defined mathematical sense. The crux of the proof of optimality and universality of the DUDE in [1] lies in establishing this fact.

In the second pass, the DUDE scans the noisy data, again using a sliding window of size $2k+1$, and generates, at each instant of time t, the estimate $\hat{X}_t$ corresponding to the sample at the center of the window. In deciding on the estimate, the algorithm receives two types of "advice": the estimated conditional distribution $\hat{P}(X_t|\mathbf{C}_t)$ derived from the first pass informs about the likelihood of values of the clean symbol given the structure observed globally in the whole data sequence; the noisy symbol Z_t observed in the current location, and the known channel parameters, on the other hand, also provide information about the likelihood of clean symbol values, independently of other observations. Clearly, when the noise level is low, more weight should be given to Z_t, whereas at higher noise levels the global information might be more reliable. The two types of advice can be combined in the conditional distribution $\hat{P}(X_t|\mathbf{C}_t^+)$, which is a good estimate of the posterior distribution of X_t given the observed data. The DUDE uses a decision rule that takes into account the estimated probability distributions, the channel parameters, and the loss function, to determine $\hat{X}_t$. The decision rule is, in essence, a MAP estimator based on the estimated posterior of X_t, weighted by the loss function.

Example 1. Suppose the data is binary and is transmitted through a binary symmetric channel (BSC) with crossover probability p, and that the loss function is the Hamming distance. Define the threshold $T = \big(2p(1-p)\big)^{-1} - 1$, and assume the current noisy sample is $Z_t = b$, where b is a binary value whose complement will be denoted $\bar{b}$. The DUDE's decision rule is: if $P(\,\bar{b}|\mathbf{C}_t)/P(\,b|\mathbf{C}_t) > T$ then flip Z_t, else leave it alone. Notice that the threshold T tends to infinity as p tends to zero — the denoiser will be unlikely to flip Z_t when p is small, since it trusts the "advice" of Z_t in that situation. Conversely, T tends to one as p tends to $1/2$ — the denoiser gives more credence to the global information when the channel is very noisy.

Example 2. The text of Figure 3 is part of a complete noisy version of a famous literary piece. The full piece was DUDE-denoised with $k = 2$, and the denoised passage corresponding to Figure 3

"What giants?" said Sancho Panza. "Those thou seest there," answered his master, "with the long arms, and spne have them nearly two leagues long." "Look, your worship," said Sancho; "what we see there are not giants but windmills, and what seem to be their arms are the sails that turned by the wind make the millstone go."

Figure 4: Denoised text (the two errors remaining out of the original 14 are underlined).

is shown in Figure 4. Only two out of fourteen original errors are left.

C Properties and highlights

Universality. In [1], the DUDE is proven to be universal in two different settings. In the *stochastic* setting, it is assumed that the clean sequence $\mathbf{X}$ is emitted by a probabilistic stationary source, and is transmitted through a probabilistic channel. It is proved that, with a choice $k = k_n$ that grows with n, but such that $k_n < c \log_{|A|} n$ for $c < 1/2$, the DUDE will denoise the input sequence $\mathbf{Z}$ asymptotically (as $n \to \infty$) as well as the *best* denoiser designed with full knowledge of the probability law governing $\mathbf{X}$. In the *semi-stochastic* setting, it is assumed that $\mathbf{X}$ is an individual sequence, with no governing probability law, while the channel remains probabilistic as before. Universality, in this case, is established by comparing the DUDE's performance with that of the class of kth order sliding window denoisers. Each such denoiser scans the data with a sliding window of size $2k + 1$, as the DUDE does, and replaces the sample at the center of the window with a function $f_k : A^{2k+1} \to A$ of the $2k+1$ samples in the window. Each such function defines a denoiser. Notice that, in particular, one of the denoisers in the class is the one that we would obtain, in principle, if we had full knowledge of the clean sequence $\mathbf{X}$ (in addition to the noisy $\mathbf{Z}$), and we exhaustively tried all possible functions f_k and picked the one giving the least loss for the given pair $(\mathbf{X}, \mathbf{Z})$. It is proved in [1] that for a choice of $k = k_n$ as specified above, the DUDE performs, asymptotically, no worse than the best k_n-th order sliding window denoiser. Notice that most useful denoisers used in practice are of the sliding window kind, e.g., the median filter mentioned in our motivating binary image denoising example.

The universality of the DUDE in both the semi-stochastic (individual sequence) and stochastic settings is analogous to that established, in the case of data compression, by the original Lempel-Ziv algorithms [2, 3]. As in LZ and in other cases in information theory, the individual-sequence ("pointwise") universality result for the DUDE is the stronger one, and the stochastic result follows as a corollary.

Choice of the parameter k. From the asymptotic point of view, the choice of $k = k_n$ as described above guarantees convergence of the DUDE's performance to the optimum denoising performance. This statement still leaves a very broad range of choices for k, so broad in fact, that the statement is not very useful in practice when we are faced with the task of denoising a given data sequence of finite length. In the latter setting, it makes sense to ask the question "what is *the best* value of k for *this particular* sequence?" Notice that analogous questions have well defined answers, for example, in data compression, and the answers can be found efficiently. For example, in various settings, we know how to implement the MDL principle and find the best Markov model order, or more generally, the best context tree to compress a given sequence [4]. For denoising, the question remains formally open. The most obvious difficulty is that since we cannot measure denoising

performance directly, we have no direct way of telling whether one value of k is better or worse than another, a task that is easy in data compression. Nevertheless, various heuristics for choosing the best value of k for the DUDE have proven very effective in practice. These heuristics are based on using an observable parameter as a proxy for the denoising performance, and optimizing the value of k based on the proxy. A heuristic described in [1] suggests using the *compressibility* of the denoised sequence (using a universal compressor), and is based on the empirical observation that when the sequence is denoised with the optimal value of k, the denoised sequence exhibits a local minimum in compression ratio. This heuristic has proven effective, in practice, in finding the best values of k for a wide range of practical data sets. More principled approaches are discussed later in this note.

Practicality. Aside from its asymptotic theoretical properties of optimality and universality, the DUDE is a very practical algorithm. It can be implemented to run in linear time complexity, with simple data structures. The scheme has been tried on a variety of data types—experiments on synthetic sources, bi-tone images, and text are reported on in [1] for a "plain-vanilla," verbatim implementation of the algorithm. Adaptations to other, more difficult data types have been reported on in the literature and are discussed later in this note. It has been said that "the DUDE is very practical despite being optimal."

Soft output. The estimated posterior $\hat{P}(X_t|\mathbf{C}_t^+)$ is used in [1] as a means to produce a "hard" decision on the denoised sample $\hat{X}_t$. The distribution, however, is valuable on its own as "soft output" from the algorithm. Applications of this soft-output DUDE (termed sDUDE) are discussed later in this note.

3 History

In February 2002, Sergio Verdú started a research sojourn at the Mathematical Sciences Research Institute (MSRI) at Berkeley, California. As part of his appointment, funded by HP Research Labs, he also started a weekly collaboration with Gadiel Seroussi and Marcelo Weinberger of the Information Theory group at HP Labs. Shortly thereafter, in March 2002, Erik Ordentlich and Tsachy Weissman joined HP Labs, and at once became involved in the research efforts of the group.

Searching for research areas of common interest led to the consideration of problems of universal statistical inference based on the noisy observation of finite-alphabet redundant signals such as text or images. Gadiel and Sergio organized a workshop at MSRI on February 25th through March 1st of 2002, in which David Donoho presented "The Kolmogorov Sampler," a work that adhered to the principle of choosing the signal realization that can be explained by the noise realization with least "complexity". Donoho proposed a scheme that, while not practically implementable, exhibits asymptotic performance within a factor of the optimal nonuniversal Bayesian scheme. This factor represented a penalty for universality that is not incurred in other problems in information theory, such as lossless compression and prediction.

We quickly zeroed in on the issue of assessing the fundamental penalty for universality, i.e., the asymptotic performance penalty incurred by the optimal universal scheme over the optimal non-universal Bayesian scheme. Our progress was hampered by an overly ambitious goal of allowing statistical uncertainty in both the signal and the noisy channel. However, in many cases, the noise mechanism is far easier to accurately model than the signal, whose redundancy structure is often

quite intricate and unknown. So, at some point, we narrowed our scope to the case of a known channel. In addition to the problem of theoretical limits, we also became interested in the more practically-minded problem of coming up with implementable algorithms for discrete universal denoising. Our literature search did not discover much beyond B. K. Natarajan's work, developed also at HP Labs in the early 90's [13]. In this work, lossy data compressors such as JPEG are used as noise filters.

Despite being preoccupied primarily with the 2002 World Cup, we came up with the idea of estimating the individual letter probabilities of the noisy signal conditioned on the noisy contexts which could then be projected back to the input (thanks to the assumption that the channel matrix is not only known but invertible). We were then delighted to discover that not only we could implement that idea with a linear-complexity algorithm, but it paid no asymptotic penalty for universality: with large enough data size and a certain context length growth with the data size the scheme was able to perform as well as the optimal Bayesian algorithm. We filed a patent for the algorithm [14] and we also came up with its acronym (DUDE) after narrowly defeating alternatives such as DUD or STUPID (STatistical Universal Probabilistic Inversion Denoiser).

We first presented our results on November 2002 at the IEEE Information Theory Workshop that took place in Bangalore. Despite this previous presentation, we were fortunate that our submission to the 2003 ISIT was not rejected, and presented the paper in Yokohama.

The IT Transactions paper was submitted on February 2003 and suffered the excruciatingly long refereeing delays for which this journal is notorious. We were relieved to finally see the paper in print in page 1 of the 2005 volume. Unfortunately, the IT Transactions printing process of our image denoising examples rendered them so different from what we could see in the pdf version of the paper that they must have raised quite a bit of skepticism about the usefulness of the DUDE for image denoising.

4 Subsequent Work

The DUDE has motivated a growing body of related research, which has proceeded in roughly three directions: generalizing and relaxing some of the assumptions underlying the DUDE, optimizing context selection, and addressing applications. Next, we overview some of this work.

As noted above, two of the defining steps of the DUDE algorithm are computing the conditional distribution of a clean symbol given its noisy context through a simple matrix multiplication involving the corresponding conditional distribution of the noisy symbol, and estimating the latter using context dependent counts. These operations are tied very closely to the assumptions of finite signal alphabets and memoryless channels. For example, the very notion of noisy context dependent counts breaks down for continuous valued channels, since, in general, a context value occurs only once with probability one. These two assumptions have been relaxed in more recent work and several corresponding extensions of the DUDE have been proposed. A common tool in these approaches is the processing of signals in overlapping super-symbols, consisting of the noisy symbol and its context. Joint probability distributions of such super-symbols are estimated from the noisy signal and computed for the clean signal using super-symbol versions of the matrix multiplication step. In [15, 16], this tool is combined with techniques from non-parametric density estimation, quantization, and convex optimization to extend the DUDE algorithm and the corresponding op-

timality results to discrete and continuous valued signals corrupted by continuous valued channels. In [17, 18] the super-symbol approach is applied in the finite alphabet setting to handle certain channels with memory. One emphasis in these latter works is on exploiting symmetries in the super-symbol channel transition probabilities to more efficiently compute the super-symbol version of the matrix multiplication step.

Another key assumption of the DUDE, namely that the channel transition matrix is known, has also been relaxed [19, 20]. When both the channel and source distribution are unknown, strong universality results of the type in [1] are no longer possible since, in general, the source and channel distributions are not identifiable from the channel output distribution alone. Nevertheless, as shown in [19, 20], a notion of minimax universality can be defined which leads to efficient and robust generalizations of the DUDE that continue to work well even under channel (and source) uncertainty.

In a different direction, the notion of a two-sided context has been generalized from the fixed length symmetrical kind considered in [1] to two-sided analogues of the context tree models from the universal compression literature. The challenge in the denoising setting is to determine, in a data dependent fashion, a good set of two-sided contexts, without, as in the case of compression, actually measuring the performance of a choice of contexts. As noted above, the latter is impossible here since the performance or loss can only be assessed with certainty using the unavailable clean signal. In [21], a loss estimator is combined with a dynamic programming pruning technique for selecting a good choice of contexts, while in [22] modeling techniques from universal compression are used to build two-sided context models from single sided models. These approaches can also be used to select the fixed context length k of the DUDE in a data dependent fashion more efficiently than the compressibility heuristic mentioned in [1]. The selection of k is also addressed in [23] and [24]. The first of these references also proposes an alternative linear time implementation of the DUDE, based on merging suffix arrays.

On the applications front, [25] elaborates on the binary image denoising application described in [1], formally extending the DUDE to two-dimensionally indexed data. A further extension to achieve practical results with gray-scale images is not straightforward, since the theoretical guarantees of the baseline DUDE in terms of convergence to optimal performance are tied to obtaining a sufficient number of occurrences of sufficiently many contexts for the law of large numbers to "kick-in." For useful context sizes and the large alphabet size (256 for 8 bit gray scale images), this calls for signal sizes significantly larger than the nominal resolutions of mainstream digital imagery. In fact, we encountered some initial skepticism in the image processing community. However, once we pose the key problem as one of modeling a distribution of the noisy symbol given its noisy context (not necessarily by means of just counting symbol occurrences in a given context), as mentioned, this challenge is not too different from the "context dilution" problem encountered in image compression. In [26], denoising of gray-scale images corrupted by various noise processes is considered. The approach taken in [26] is inspired by work in image compression [27] and is based on exploiting prior knowledge about digital images to "merge" statistics from multiple contexts to denoise in any given context. The exploitation of prior knowledge can be seen as partially "backing away" from universality, or as a reduction in the model class for which universality is claimed. The denoising performance of such an image-informed DUDE is found to be competitive with state-of-the-art approaches based on wavelets, but unlike such approaches, which are tied fairly closely

to additive Gaussian noise, it can be more easily tailored to handle different noise processes. In particular, the image-informed DUDE has unsurpassed performance for salt-and-pepper noise.

An alternative DUDE-based approach to gray scale image denoising is taken in [28], which treats digital images as real valued signals and applies the continuous alphabet DUDE [16] mentioned above. Though arrived at from completely different principles, the resulting algorithm bears a striking resemblance to the recently introduced non-local means (NL means) algorithm of [29] (see also [30]), which has been reported to achieve remarkable denoising performance.

The DUDE has also been applied [31] in an error correction setting to enhance the decoding of systematically encoded uncompressed sources. The idea is to run the DUDE on the noisy information symbols aggregated from multiple received codewords so that the corresponding denoised codewords are more likely to belong to correct decoding regions than the original noisy codewords. Depending on the "denoisability" of the source, which is closely related to its redundancy, such a DUDE enhanced communication system may be able to operate correctly under significantly noisier conditions than a DUDE-less approach. The universality of the DUDE allows the approach to work with a variety of source types, without requiring additional side-information from the encoder. It is thus "backward compatible" with deployed systems. Soft-input-soft-output extensions of the DUDE are also introduced in [31] to improve performance with soft channel decoders and to allow further iterations between the decoding and denoising stages.

Additional DUDE inspired works which we have only space to mention briefly include converse results [32], which show that the convergence rate of the performance of the DUDE to that of the optimal sliding window denoiser is the best possible up to a constant factor; the definition, analysis, and estimation of erasure entropy [33, 34], which is closely linked to asymptotic denoisability over an erasure channel in the limit of zero erasure probability; and universal versions of the filtering or causal denoising problem [35], aka causal DUDE.

References

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

[2] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, 23:337–343, May 1977.

[3] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, 24(5):530–536, September 1978.

[4] J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 29:656–664, September 1983.

[5] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. Inform. Theory*, 38:1258–1270, July 1992.

[6] N. Merhav and M. Feder. Universal prediction. *IEEE Trans. Inform. Theory*, 44(6):2124–2147, October 1998.

[7] H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statis. Prob.*, pages 131–148, 1951.

[8] J. Van Ryzin. The sequential compound decision problem with $m \times n$ finite loss matrix. *Ann. Math. Statist.*, 37:954–975, 1966.

[9] S. B. Vardeman. Admissible solutions of k-extended finite state set and the sequence compound decision problems. *J. Multiv. Anal.*, 10:426–441, 1980.

[10] C. H. Zhang. Compound decision theory and empirical Bayes methods. *Annals of Statistics*, 31(2):379–390, 2003.

[11] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, III:97–139, 1957. Princeton, NJ.

[12] F. Jelinek L. R. Bahl, J. Cocke and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Trans. Inform. Theory*, IT-20:284–287, March 1974.

[13] B. K. Natarajan, Filtering random noise from deterministic signals via data compression. *IEEE Trans. Signal Proc.*, 43(11):2595–2605, Nov. 1995.

[14] T. Weissman, E. Ordentlich, G. Seroussi, M.J. Weinberger, and S. Verdú,"Method for Correcting Noise Errors in a Digital Signal," U.S. Patent No. 7,047,472, filed Oct. 2003, issued May 2006.

[15] A. Dembo and T. Weissman. Universal denoising for the finite-input-general-output channel. *IEEE Trans. Inform. Theory*, 51(4):1507 – 1517, April 2005.

[16] K. Sivaramakrishnan and T. Weissman. Universal denoising of discrete-time continuous-amplitude signals. In *Proc. of the 2006 IEEE Intl. Symp. on Inform. Theory*, (ISIT'06), Seattle, WA, USA, July 2006.

[17] C. D. Giurcaneanu and B. Yu. Efficient algorithms for discrete universal denoising for channels with memory. In *Proc. of the 2005 IEEE Intl. Symp. on Inform. Theory*, (ISIT'05), Adelaide, Australia, Sept. 2005.

[18] R. Zhang and T. Weissman. Discrete denoising for channels with memory. *Communications in Information and Systems (CIS)*, 5(2):257–288, 2005.

[19] G. M. Gemelos, S. Sigurjonsson, T. Weissman. Universal minimax discrete denoising under channel uncertainty. *IEEE Trans. Inform. Theory*, 52:3476–3497, 2006.

[20] G. M. Gemelos, S. Sigurjonsson and T. Weissman. Algorithms for discrete denoising under channel uncertainty. *IEEE Trans. Signal Processing*, 54(6):2263–2276, June 2006.

[21] E. Ordentlich, M.J. Weinberger, and T. Weissman. Multi-directional context sets with applications to universal denoising and compression. In *Proc. of the 2005 IEEE Intl. Symp. on Inform. Theory*, (ISIT'05), Adelaide, Australia, Sept. 2005.

[22] J. Yu and S. Verdú. Schemes for bidirectional modeling of discrete stationary sources. *IEEE Trans. Inform. Theory*, 52(11):4789–4807, 2006.

[23] S. Chen, S. N. Diggavi, S. Dusad and S. Muthukrishnan. Efficient string matching algorithms for combinatorial universal denoising. In *Proc. of IEEE Data Compression Conference (DCC)*, Snowbird, Utah, March 2005.

[24] G. Gimel'farb. Adaptive context for a discrete universal denoiser. In *Proc. Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshops, SSPR 2004 and SPR 2004,* Lisbon, Portugal, August 18-20, pp. 477–485.

[25] E. Ordentlich, G. Seroussi, S. Verdú, M.J. Weinberger, and T. Weissman. A universal discrete image denoiser and its application to binary images. In *Proc. IEEE International Conference on Image Processing*, Barcelona, Catalonia, Spain, September 2003.

[26] G. Motta, E. Ordentlich, I. Ramírez, G. Seroussi, and M. Weinberger. The DUDE framework for continuous tone image denoising. In *Proc. of IEEE International Conference on Image Processing*, Genoa, Italy, October 2005.

[27] M. J. Weinberger, G. Seroussi, and G. Sapiro. The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS. *IEEE Trans. on Image Processing*, Vol. 9, No. 8, August 2000.

[28] K. Sivaramakrishnan and T. Weissman. Universal denoising of continuous amplitude signals with applications to images. In *Proc. of IEEE International Conference on Image Processing*, Atlanta, GA, USA, October 2006, pp. 2609–2612.

[29] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In Proc. of 2005 IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recog.,(CVPR 2005) 2:60–65, June 2005.

[30] M. Mahmoudi and G. Sapiro. Fast image and video denoising via nonlocal means of similar neighborhoods. IEEE Signal Processing Letters, 12(12)839–842, Dec. 2005.

[31] E. Ordentlich, G. Seroussi, S. Verdú, and K. Viswanathan. Universal algorithms for channel decoding of uncompressed sources. Submitted to *IEEE Trans. Inform. Theory*, 2006.

[32] K. Viswanathan and E. Ordentlich. Lower limits of discrete universal denoising. In *Proc. of the 2006 IEEE Intl. Symp. on Inform. Theory*, (ISIT'06), Seattle, WA, USA, July 2006.

[33] S. Verdú and T. Weissman. Erasure entropy. In *Proc. of the 2006 IEEE Intl. Symp. on Inform. Theory*, (ISIT'06), Seattle, WA, USA, July 2006.

[34] J. Yu and S. Verdú. Universal erasure entropy estimation. In *Proc. of the 2006 IEEE Intl. Symp. on Inform. Theory*, (ISIT'06), Seattle, WA, USA, July 2006.

[35] T. Weissman, E. Ordentlich, M. Weinberger, A. Somekh-Baruch, N. Merhav. Universal filtering via prediction, *IEEE Trans. Inform. Theory*, 53(4):5–28, April 2007.