

DUDE: An Algorithm for Discrete Universal Denoising

Tsachy Weissman

Erik Ordentlich

Gadiel Seroussi

Sergio Verdú

Marcelo Weinberger¹

Finite-alphabet signals corrupted by discrete noisy channels arise naturally in a wide range of applications spanning fields such as statistics, engineering, computer science, image processing, astronomy, biology, cryptography, and information theory. In the Shannon paradigm, redundancy is added to the noiseless signal in order to protect it from the channel noise, and a decoder that knows the codebook and the channel statistics can recover the noiseless signal with arbitrary reliability, provided that the coding scheme respects the fundamental limits of information theory. In contrast, there are many applications where no channel coding is performed and the recovery of the corrupted signal can only be accomplished with a certain distortion. Examples include DNA sequence analysis and processing, text correction, Hidden Markov model state estimation, image denoising, and high-dimensional inverse problems arising in data-mining scenarios.

While the field of filtering or denoising of continuous-alphabet signals has a long history with notable achievements such as Wiener filters, Kalman filters, and Donoho-Johnstone denoisers, the field of discrete denoising has seen far less progress. The best known discrete denoiser is the dynamic programming algorithm either in forward (causal denoising) or backward-forward (noncausal denoising) mode. Unfortunately, the scope of applicability of this algorithm is severely limited to noiseless signals that are modelled by Markov chains with known statistics. Another existing approach to discrete denoising is to process the noisy signal with a lossy data compressor. The rationale is that the noise constitutes that part of the noisy signal which is hardest to compress. Thus, by lossily compressing the noisy signal and appropriately tuning the fidelity level of the compressor to match the noise level, it may be expected that the part of the noisy signal that will be lost will mainly consist of the noise, so that the reconstructed signal will, in effect, be a denoised version of the corrupted signal.

In many discrete denoising applications, a good model for the randomness of the noisy channel is known, whereas the statistical description of the noiseless signal is either unknown or too complex. It is therefore of considerable interest to pose the problem of discrete universal denoising where no knowledge exists about the statistics of the noiseless signal while the channel statistics are assumed known. Furthermore, in this work we restrict ourselves to the important special case of memoryless channels.

The DUDE algorithm for discrete universal denoising has very favorable complexity requirements. The number of arithmetic and other register-level operations grows linearly with the data size whereas, in addition to the need to store the data, the working storage requirements are sublinear in the

data size.

DUDE operates in two stages, making two passes through the noisy observation sequence. For a fixed *context* 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 (growing logarithmically with the size of the data set) yields a denoiser with the following key optimality properties in the asymptotic regime:

1. *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 the DUDE 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.
2. *The stochastic setting.* We show that the DUDE algorithm asymptotically attains the performance of the optimal distribution-dependent scheme, for any stationary ergodic source that may be generating the noiseless signal.

Those optimality properties are in sharp contrast to the performance attained by an ideal (nonimplementable) compression-based approach which selects the typical noise realization corresponding to the most compressible signal realization. This scheme is known to fall short of attaining the minimum loss achieved by the optimal distribution-dependent scheme.

In the special case of the erasure channel, the DUDE takes a particularly simple form: every erased symbol is replaced by the most frequent symbol seen within the same left and right k -contexts. In the special case of the binary symmetric channel with crossover probability δ , for every received bit b , the algorithm computes the fraction of the total number of bits received with the same left and right k -contexts that are equal to b . If this fraction is less than $2\delta(1 - \delta)$, the algorithm flips b to \bar{b} .

In the presentation, we will show the results of several experiments, including the comparison with backward-forward dynamic programming for denoising Markov chains, denoising of black-and-white images including the comparison with standard approaches such as median filtering and morphological filters, and the correction of corrupted English texts.

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, New Jersey. The other authors are with Hewlett-Packard Laboratories, Palo Alto, California. T. Weissman is also with the Department of Statistics, Stanford University, Palo Alto, California.