Denoising and Filtering Under the Probability of Excess Loss Criterion

Stephanie Pereira and Tsachy Weissman, *Member, IEEE*

***Abstract*—Subclasses of finite alphabet denoising and filtering (causal denoising) schemes are compared. Performance is measured by the normalized cumulative loss (a.k.a. distortion), as measured by a single-letter loss function. We aim to minimize the probability that the normalized cumulative loss exceeds a given threshold. We call this quantity the probability of excess loss. Specifically, we consider a scheme to be optimal if it attains the maximal exponential decay rate of the probability of excess loss. This provides another way of comparing schemes that complements and contrasts previous work which considered the expected value of the normalized cumulative loss.**

In particular, the question of whether the optimal denoiser is symbol-by-symbol for an independent and identically distributed (i.i.d.) source and a discrete memoryless channel (DMC) is investigated. For Hamming loss, the optimal denoiser is proven to be symbol-by-symbol. Perhaps somewhat counterintuitively, for a general single letter loss function, the optimal scheme need not be symbol-by-symbol.

The optimal denoiser requires unbounded delay and unbounded look-ahead while symbol-by-symbol schemes mandate zero delay and look-ahead. It is natural to wonder about the effect of limited delay and limited look-ahead. Consequently, finite sliding-window denoisers and finite block denoisers are defined. They are shown to perform no better than symbol-by-symbol denoisers.

Finally, the effect of causality is investigated. While it is difficult to characterize the performance of filters with unbounded memory explicitly, it is shown that finite memory filters perform no better than symbol-by-symbol filters.

***Index Terms*—Causality, delay, denoising, filtering, large deviations, look-ahead, memory, probability of excess loss, single letter loss, sliding-block, Stein's paradox, symbol-by-symbol, time-invariant schemes.**

I. Introduction

The denoising and filtering problems have a long history focussed on the continuous alphabet case. Recently, there has been work on the discrete alphabet case (cf., [1], [2]). To our knowledge, only the problem of minimizing *expected* loss has been considered. We study the probability that the loss *exceeds* a particular threshold, first considered by Marton in [3] in the context of lossy source codes. This *excess loss* criterion enables us to design denoisers and filters that have loss less than some target level with high probability. Further, even if a denoiser/filter has low expected loss, the spread of this loss may be high. The excess loss criterion provides a handle on the spread of the loss. Our work was partially inspired by results in lossy source coding (cf. [3], [4], [5]).

In particular, we analyze the asymptotic excess loss probability by establishing a large deviations principle (LDP) for denoisers and determining the corresponding rate function. Large deviations characterizations have been used as a performance metric both in the information theory and statistics literature (see [3], [6], [4], [7], and [8], respectively).

The LDP for denoising is a special case of the lossy source coding LDP discussed in [4] and [7]. However, while [4] and [7] are concerned with characterizing the performance of the optimal scheme, the basic question we ask in this work is how different subclasses of schemes compare to the optimal scheme. In other words, how much, if anything, is lost by restricting the class of allowable schemes? There is a clear practical motivation to this question. The subclasses we consider are those that limit the amount of noisy observations that the denoiser "sees." In practice, a denoiser may not have an unbounded horizon so it is important to ascertain whether/when such practical schemes are close to the optimal bound. Further, we demonstrate that there are cases where symbol-by-symbol denoising is stricly suboptimal. This result is qualitatively similar to Stein's paradox [9], [10] where it is shown that an admissible estimate of an individual sequence corrupted by independent and identically distributed (i.i.d.) Gaussian noise (alternately estimating the parametric mean of a multivariate) under mean-square error loss requires that the estimate for each sequence component be based on the entire observation sequence. Note, however, that in our problem we are estimating an i.i.d. source (as opposed to an individual sequence or parametric estimation) and optimizing the exponent of the probability of the excess loss (as opposed to the minimum mean-square error).

We further note that, while the derivations of [4] and [7] are information-theoretic, our results have more of a large deviations flavor. That is, while the characterizations in [4] and [7] are given in terms of minimum Kullback–Leibler divergences, in this work we emphasize the Fenchel–Legendre transform representation of the exponents. This representation makes the comparison of the rate functions for different subclasses more transparent and helps us to establish cases of strict suboptimality of symbol-by-symbol and other classes of schemes.

II. Setup

The setup (see Fig. 1) is as follows: a source generates n i.i.d. symbols, $\{X_i\}_{i=1}^n$, that take values in a discrete alphabet

Manuscript received March 26, 2006; revised November 14, 2006. This work was supported in part by a Texas Instruments Stanford Graduate Fellowship, an Intel startup grant, and by the National Science Foundation under Grant CCR-0311633. The material in this paper was presented at the 43rd Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, October 2005.

The authors are with Stanford University, Stanford, CA 94305-9510 USA (e-mail: sfp@stanford.edu; tsachy@stanford.edu).

Communicated by X. Wang, Associate Editor for Detection and Estimation.

Color versions of Figures 8 and 10 in this paper is available online at http://ieeexplore.org.

Digital Object Identifier 10.1109/TIT.2007.892772

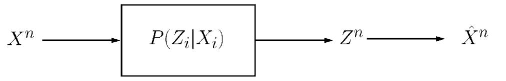

Fig. 1. Denoising/filtering setup.

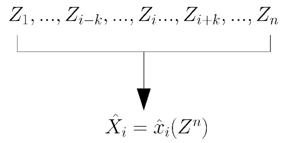
$Z_1, ..., Z_{i-k}, ..., Z_i ..., Z_{i+k}, ..., Z_n$
$\hat{X}_i = \hat{x}_i(Z^n)$

Fig. 2. Denoiser.

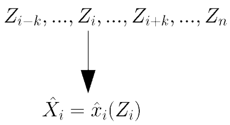
$Z_{i-k}, ..., Z_i, ..., Z_{i+k}, ..., Z_n$
$\hat{X}_i = \hat{x}_i(Z_i)$

Fig. 3. Symbol-by-symbol denoiser/filter.

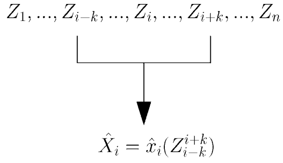
$Z_1, ..., Z_{i-k}, ..., Z_i, ..., Z_{i+k}, ..., Z_n$
$\hat{X}_i = \hat{x}_i(Z_{i-k}^{i+k})$

Fig. 4. k-finite sliding-window denoiser.

$\mathcal{X}$ offinite cardinality. These source symbols X_i pass through a discrete memoryless channel (DMC) $\Pi(z \mid x)$ to produce Z_i that take values in a discrete alphabet $\mathcal{Z}$ of finite cardinality. Denote the distribution of the Z_i by Q.

A. Denoising

The goal of a denoiser is to estimate $X^n = (X_1, X_2, \ldots, X_n)$ from $Z^n = (Z_1, Z_2, \ldots, Z_n)$. The vector Z^n is *denoised* to produce the symbol $\hat{X}_i = \hat{x}_i(Z^n)$, for each i, where the *denoising functions* $\hat{x}_i(\cdot)$ are general, deterministic functions of the random vector Z^n with range $\hat{\mathcal{X}} = \{1, 2, \ldots, |\hat{\mathcal{X}}|\}$, $|\hat{\mathcal{X}}| < \infty$. We note that while $\hat{x}_i(\cdot)$ is a deterministic function, $\hat{X}_i$ is a random variable. The *denoiser* is the collection of *denoising functions* $(\hat{x}_1(\cdot), \hat{x}_2(\cdot), \ldots, \hat{x}_n(\cdot))$ and is denoted by $\hat{x}^n(\cdot)$. We illustrate a general denoiser in Fig. 2. If the denoising functions satisfy $\hat{x}_i(Z^n) = \hat{x}(Z^n, Z_i)$, $1 \leq i \leq n$ for some deterministic function $\hat{x} : \mathcal{Z}^n \times \mathcal{Z} \to \hat{\mathcal{X}}$, we call the denoiser *time invariant*.

We refer to a denoiser with denoising functions that depend only on Z_i as a *symbol-by-symbol* denoiser, so that $\hat{X}_i = \hat{x}_i(Z^n) = \hat{x}_i(Z_i)$ for a symbol-by-symbol denoiser. A symbol-by-symbol denoiser is shown in Fig. 3. Note that the $\hat{x}_i(\cdot)$ may vary with time (hence the subscript i). Applying the above definition and the definition of a symbol-by-symbol denoiser, we can see that a symbol-by-symbol denoiser is time invariant if $\hat{x}_i(z_i) = \hat{x}(z_i)$ for some function $x : \mathcal{Z} \to \hat{\mathcal{X}}$.

We define the k-finite sliding-window denoiser to allow $\hat{X}_i$ to depend on Z_{i-k}^{i+k} (i.e., $\hat{X}_i = \hat{x}_i(Z_{i-k}^{i+k})$) (see Fig. 4). As above, a time-invariant k-finite sliding-window denoiser satisfies $\hat{x}_i(Z_{i-k}^{i+k}) = \hat{x}(Z_{i-k}^{i+k}, Z_i)$ for $\hat{x} : \mathcal{Z}^{2k+1} \times \mathcal{Z} \to \hat{\mathcal{X}}$. We can view k-blocks of symbols as *supersymbols* to be denoised. We thus define the *k-finite block denoiser* to divide the output sequence Z^n sequentially into $\lfloor \frac{n}{k} \rfloor$ blocks of k symbols and up to one remainder block of less than k symbols: $Z_{M(k-1)+1}^{Mk}$, $M = 1, 2, \ldots, \lfloor \frac{n}{k} \rfloor$, and $Z_{\lfloor \frac{n}{k} \rfloor k+1}^{n}$. A block of reconstruction symbols is produced after observing each output block, so that

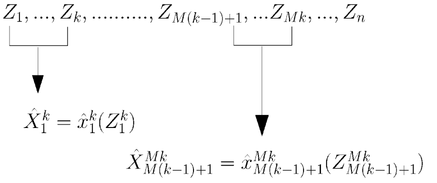
$Z_1, ..., Z_k, ..., Z_{M(k-1)+1}, ...Z_{Mk}, ..., Z_n$
$\hat{X}_1^k = \hat{x}_1^k(Z_1^k)$
$\hat{X}_{M(k-1)+1}^{Mk} = \hat{x}_{M(k-1)+1}^{Mk}(Z_{M(k-1)+1}^{Mk})$

Fig. 5. k-finite block denoiser.

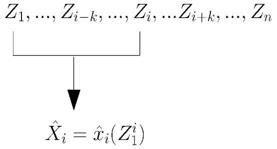
$Z_1, ..., Z_{i-k}, ..., Z_i, ...Z_{i+k}, ..., Z_n$
$\hat{X}_i = \hat{x}_i(Z_1^i)$

Fig. 6. Filter.

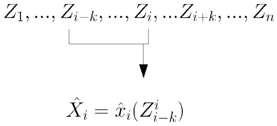
$Z_1, ..., Z_{i-k}, ..., Z_i, ...Z_{i+k}, ..., Z_n$
$\hat{X}_i = \hat{x}_i(Z_{i-k}^i)$

Fig. 7. k-finite memory filter.

$$\hat{X}_{M(k-1)+1}^{Mk} = \hat{x}_{M(k-1)+1}^{Mk}(Z_{M(k-1)+1}^{Mk}), \quad M = 1, 2, \ldots, \lfloor \frac{n}{k} \rfloor$$

and

$$\hat{X}_{\lfloor \frac{n}{k} \rfloor k+1}^{n} = \hat{x}_{\lfloor \frac{n}{k} \rfloor k+1}^{n}(Z_{\lfloor \frac{n}{k} \rfloor k+1}^{n})$$

(see Fig. 5). It is straightforward to deduce the form of a time-invariant k-finite block filter from the above definition.

B. Filtering

The basic idea in filtering is to reconstruct X^n *causally*. That is, the filtering function $\hat{x}_i(\cdot)$ at time i may depend only on Z^i so that $\hat{X}_i = \hat{x}_i(Z^i)$ as in Fig. 6. So, the most general filtering functions can make use of all of Z^i when deciding on output $\hat{x}_i$. The memory of a general filter is unbounded in that the number of observation symbols Z_i used to make the decision on $\hat{x}_i$ grows arbitrarily large with i. We call this most general class of filters the class of *infinite* filters. It turns out to be difficult to analyze such filters so we now define some classes of filters with*finite* memory (i.e., their output at time i depends on a fixed number of past output symbols) which are interesting in their own right. In particular, we consider the *symbol-by-symbol filter*, which is the same as the symbol-by-symbol denoiser (i.e., $\hat{X}_i = \hat{x}_i(Z_i)$) of Fig. 3, and the *$k$-finite memory filter*, which allows $\hat{X}_i$ to depend on Z_{i-k}^i (i.e., $\hat{X}_i = \hat{x}_i(Z_{i-k}^i)$) as shown in Fig. 7. Similar to the above, a time-invariant k-finite memory filter satisfies $\hat{x}_i(Z_{i-k}^i) = \hat{x}(Z_{i-k}^i, Z_i)$ for $\hat{x} : \mathcal{Z}^{k+1} \times \mathcal{Z} \to \hat{\mathcal{X}}$.

C. Criterion for Optimality

We assume a given single-letter loss function $\Lambda(x, \hat{x}) : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ such that there exists a maximum loss $\Lambda_{\max} = \max_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} \Lambda(x, \hat{x})$. Note that what we call loss is also referred to as distortion, particularly in the context of source coding. An example of such a single letter loss function is Hamming loss, where $\Lambda(x,\hat{x})=0$ if $x=\hat{x}$ and $\Lambda(x,\hat{x})=1$ if $x\neq\hat{x}$. We denote the maximum value of the single-letter loss function by $\Lambda_{\max}$.

We consider the normalized cumulative loss

$$L_n=\frac{1}{n}\sum_{i=1}^{n}\Lambda(X_i,\hat{X}_i). \tag{1}$$

For the general denoising setup

$$L_n=\frac{1}{n}\sum_{i=1}^{n}\Lambda(X_i,\hat{x}_i(Z^n)). \tag{2}$$

The cumulative loss for the other denoisers and for the filters is defined analogously, with the appropriate restrictions on the functions $\hat{x}_i$. Note that L_n depends on the particular denoiser/filter as well as on X^n and Z^n. We omit these from the notation for readability.

The normalized cumulative loss is a random variable. The performance of a denoiser is usually characterized by the expected value of the normalized cumulative loss. We take a different approach and examine the probability that the normalized cumulative loss exceeds some threshold D. For D less than or equal to the minimum achievable expected loss, this probability goes to 1 when using the optimal scheme, by the law of large numbers. Thus, we consider values of D that exceed the minimum achievable expected loss. In the sequel, a denoiser (i.e., the collection of denoising functions $\hat{x}^n(\cdot)=(\hat{x}_1(\cdot),\ldots,\hat{x}_n(\cdot))$ as described in Section II-A) will be said to be *optimal* if it achieves the best exponential rate of decay of $P(L_n>D)$. Similarly, the *optimal* symbol-by-symbol, k-finite block, k-finite sliding window and k-finite memory denoisers/filters maximize the exponential rate of decay of $P(L_n>D)$ among all symbol-by-symbol, k-finite sliding window, k-finite block, and k-finite memory schemes, respectively (i.e., among the schemes where the denoising/filtering functions $\hat{x}_i(\cdot)$ are chosen so that the denoiser/filter $\hat{x}^n(\cdot)$ is symbol-by-symbol, k-finite sliding window, k-finite block or k-finite memory, respectively, as in Sections II-A and II-B).

III. Main Results

In Section IV, we prove the following.

Theorem 1:

$$R(\mathcal{C},D)\triangleq\lim_{n\to\infty}-\frac{1}{n}\log\min_{\hat{x}^n(\cdot)\in\mathcal{C}}P(L_n>D) \tag{3}$$

exists for $\mathcal{C}$ equal to

- $\mathcal{C}_1$, the class of all denoisers;
- $\mathcal{C}_2$, the class of symbol-by-symbol denoisers;
- $\mathcal{C}_3(k)$, the set of k-finite block denoisers;
- $\mathcal{C}_4(k)$, the class of k-finite sliding window denoisers;
- $\mathcal{C}_5(k)$, the class of k-finite memory filters;

noting that $\mathcal{C}_i(k)$, $i=3,4$, or 5 defines a class of denoisers/filters for each value of k. We call $R(\mathcal{C},D)$ the *optimal rate function* for class $\mathcal{C}$. Furthermore, $\lim_{n\to\infty}-\frac{1}{n}\log P(L_n>D)$ exists for any symbol-by-symbol denoiser (i.e., $\hat{x}^n\in\mathcal{C}_2$) and we call it the *rate function for symbol-by-symbol denoisers*.

Theorem 2: There exist sources, channels, and distortion criteria for which $R(\mathcal{C}_1,D)>R(\mathcal{C}_2,D)$. That is, in general, symbol-by-symbol denoisers are suboptimal. Furthermore, the optimal denoiser and the optimal symbol-by-symbol denoiser are time invariant.

Theorem 3: Under Hamming loss, $R(\mathcal{C}_1,D)=R(\mathcal{C}_2,D)$, i.e., symbol-by-symbol denoising is optimal for all sources and channels.

Theorem 4: For any k

$$R(\mathcal{C}_2,D)=R(\mathcal{C}_3(k),D)=R(\mathcal{C}_4(k),D)=R(\mathcal{C}_5(k),D).$$

That is, in general, finite block denoisers, finite sliding-window denoisers, and finite memory filters do no better than symbol-by-symbol denoisers/filters. Since $\mathcal{C}_2\subset\mathcal{C}_i(k)$ for $i>2$, Theorem 2 implies that the optimal rate functions for $\mathcal{C}_i(k)$, $i>2$ are achieved by time-invariant symbol-by-symbol denoisers.

Remark 1: Establishing the LDP for the best denoiser in $\mathcal{C}_1$ in Theorem 1 is nontrivial because (as we elaborate upon in Section IV-A1), L_n is a sum of dependent random variables.

Remark 2: We give concrete examples where the inequality of Theorem 2 is strict.

Remark 3: Theorem 2 seems somewhat counterintuitive since the source is i.i.d., the channel is memoryless, and the distortion is single-letter.

Remark 4: To obtain Theorem 4, we first compute $R(\mathcal{C}_3(k),D)$ and show that $R(\mathcal{C}_2,D)=R(\mathcal{C}_3(k),D)$. We use these two results to obtain $R(\mathcal{C}_2,D)=R(\mathcal{C}_4(k),D)$ by an approximation argument. Finally, we show that $R(\mathcal{C}_2,D)=R(\mathcal{C}_5(k),D)$ by observing that $\mathcal{C}_4(k)\supset\mathcal{C}_5(k)$.

IV. Optimal Rate Functions and Optimality of Denoisers/Filters

We prove the first two parts of Theorem 1 in Section IV-A. We establish the last part of Theorem 1 and the time-invariance of the optimal symbol-by-symbol denoiser of Theorem 2 in Section IV-B. In Section IV-C, we show that the optimal denoisers in $\mathcal{C}_1$ and $\mathcal{C}_2$ are time invariant and find a set of examples where symbol-by-symbol denoising is strictly suboptimal thereby showing Theorem 2. Section IV-D characterizes the performance of k-finite block denoisers and shows it to be equivalent to the that of symbol-by-symbol denoisers, thus establishing the third part of Theorem 1 and part of Theorem 4. Section IV-E does the same for k-finite sliding-window denoisers. Finally, Section IV-F discusses filtering, explaining why it is difficult to analyze the performance of filters with unbounded memory and then characterizing finite memory filters in order to establish the remainders of Theorems 1 and 4.

A. $R(\mathcal{C}_1, D)$ and $R(\mathcal{C}_2, D)$

1) An Overview of Optimal Denoisers: We present upper and lower bounds on $P(L_n > D)$ for an optimal denoiser, preceded by some notation. First, however, we explain why obtaining the LDP for L_n is nontrivial. We note that L_n is a sum of random variables, $\Lambda(X_i, \hat{x}_i(Z^n))$, that are not in general independent of each other. This is because the estimate $\hat{x}_i$ is based on Z^n and Z^n is correlated to each of the X_i, $i = 1, 2, \ldots, n$ by the channel so that $\hat{x}_i(Z^n)$ is correlated to each of the X_i. It is thus nontrivial to show whether and when the sum L_n concentrates for general denoising functions.

If we restrict ourselves to symbol-by-symbol denoisers (as defined in Section II-A), we have that $\Lambda(X_i, \hat{x}_i(Z^n)) = \Lambda(X_i, \hat{x}_i(Z_i))$ and are independent but not identically distributed. It is reasonable to expect a sum of such random variables to concentrate, but the proof uses a key lemma from [5] which is a recent result. We show how to apply their *arbitrarily varying source* lemma directly to L_n in Section IV-B. Also, we will elaborate upon this lemma shortly.

For the case of a general denoiser, where the ith estimate $X_i = x_i(Z^n)$, we will show in Section IV-A3 that the sum concentrates for the *optimal* denoiser. We obtain the concentration result by conditioning on Z^n and expressing $P(L_n > D|Z^n)$ as a sum of conditionally independent but not identically distributed random variables. We can then use the arbitrarily varying source lemma of [5] to show that $P(L_n > D|Z^n)$ concentrates. However, to get $P(L_n > D)$ we must sum $P(L_n > D|Z^n)$ over an exponential set so that it is not clear that $P(L_n > D)$ concentrates. We will argue that the best denoiser depends only on the empirical type of Z^n and so the summation can be taken over the (polynomial) number of types of Z^n rather than the exponential number of Z^n. This will yield a concentration of $P(L_n > D)$ but only for the *optimal* denoiser.

Having established the difficulty of the problem and having summarized our approach, we now summarize and state formally an important lemma that we will use repeatedly in this paper.

2) Arbitrarily Varying Sources: Basically, the arbitrarily varying source lemma establishes an LDP for sums of n independent but not identically distributed random variables. It requires that the random variables take on a finite number of discrete values, have bounded support, and have probability distributions that lie in some finite set of distributions. In addition to establishing the LDP, the error of the LDP approximation is given and holds for sufficiently large but finite n. We now state the lemma formally.

Consider a set of r probability mass functions on the real line which we denote by $\{P_1, P_2, \ldots P_r\}$, $r < \infty$. Denote the support of distribution P_a by $S(P_a)$, $a = 1, 2, \ldots, r$. Suppose that for each a, there are a finite number of elements in $S(P_a)$ and that every element of $S(P_a)$, is upper-bounded by $\Delta_a \in \mathbb{R}$ and that $0 < P_a(\{\Delta_a\}) < 1$ (where the inequalities are strict). Now, let W_i be independent random variables with distribution in $\{P_a\}_{a=1}^r$. Then, following the terminology of [5], we call W_i an arbitrarily varying source (AVS). Define

$$S_n = \frac{1}{n}\sum_{i=1}^{n} W_i.$$

Denote the fraction of W_i in W^n with distribution P_a by Q_a^n. For example, if $n = 9$ and only five of the nine W_i have distribution P_1, then $Q_1^9 = \frac{5}{9}$. We now see the reason for using both a and n in the notation for Q_a^n. The a indexes the distribution P_a and the n is relevant because Q_a^n can only take values in $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. (In fact, we will want to optimize a function of the $\{Q_a^n\}$, $a = 1, 2, \ldots, r$ for a particular n and we will argue that as $n \to \infty$, we can equivalently solve the optimization over the continuous parameter space $[0, 1]^r$ instead of over the discrete valued parameter space $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}^r$ of the $\{Q_a^n\}$.) Let $\Psi_a(\cdot)$ be the moment generating function of a random variable with distribution P_a, i.e.,

$$\Psi_a(\lambda) = E_{P_a}(e^{\lambda W}) = \sum_{w \in S(P_a)} e^{\lambda w} P_a(w)$$

where we can write the expectation as a finite sum since $S(P_a)$ is a finite set.

Let $\Psi_{(n)}(\lambda) = \sum_{a=1}^{r} Q_a^n \log(\Psi_a(\lambda))$ and let

$$\begin{aligned}\Psi_{(n)}^*(x) &= \sup_{\lambda \geq 0} \left[\lambda x - \Psi_{(n)}(\lambda)\right] \\ &= \sup_{\lambda \geq 0} \left[\lambda x - \sum_{a=1}^{r} Q_a^n \log \Psi_a(\lambda)\right].\end{aligned}$$

Then, we have the following.

Lemma 1 (Large Deviations for AVS): For $D > 0$ and for n finite but sufficiently large,

$$P(S_n > D) \leq e^{-n\Psi_{(n)}^*(D)},$$
and
$$P(S_n > D) \geq e^{-n\left[\Psi_{(n)}^*(D) + o(n)\right]},$$

where $o(n)$ can be characterized explicitly as a function of n.

Proof: Although the proof is given in [5], for the convenience of the reader, we include in Appendix I of this paper a proof of the arbitrarily varying source lemma which includes more details than the proof in [5]. The precise expression for $o(n)$ can also be found in Appendix I. □

3) Optimal Denoisers: With Lemma 1, we can compute $P(L_n > D)$ for a general denoiser.

Consider an arbitrary denoiser $\hat{x}^n(\cdot) : \mathcal{Z}^n \to \mathcal{X}^n$. We examine the relation

$$P(L_n > D|Z^n = z^n) = P\left(\frac{1}{n}\sum_{i=1}^{n} \Lambda(X_i, \hat{x}_i(Z^n)) > D|Z^n = z^n\right).$$

Now, since $\hat{x}^n(\cdot)$ is deterministic, there is no randomness in $\hat{x}(Z^n)$ given $Z^n = z^n$. Given $Z^n = z^n$, consider the set $I_{\hat{x},z} = \{i : (\hat{x}_i, z_i) = (\hat{x}, z)\}$. We suppress the dependence of $I_{\hat{x},z}$ on n for simplicity. $I_{\hat{x},z}$ is the set of all noisy observations z_i (which are deterministic given Z^n) that are denoised to the estimate $\hat{x}_i = \hat{x}$. We index these pairs $(\hat{x}_i, z_i)$ by their time index i. One can see that for all $i \in I_{\hat{x},z}$, $\Lambda(X_i, \hat{x}_i)$ are conditionally independent given $Z^n = z^n$ and have the same distribution as $\Lambda(X, \hat{x})$ given $Z = z$, where X has the source distribution and Z has the distribution induced by the channel and by the distribution of X. So, $\Lambda(X_i, \hat{x}_i)$ given $Z^n = z^n$ is an AVS, where the finite set of possible distributions of this random variable is indexed by the possible values of $(\hat{x}_i, z_i)$ and has magnitude $|\hat{\mathcal{X}}||\mathcal{Z}|$. We can thus use Lemma 1 with r and the collection of distributions $\{P_a\}_{a=1}^r$ from the statement of the lemma corresponding to $|\hat{\mathcal{X}}||\mathcal{Z}|$ and the distributions associated with $\Lambda(X, \hat{x})$ given $Z = z$. Note that we have shown that, conditioned on $Z^n = z^n$, L_n is a sum of independent random variables.

Let Q_z^n be the fraction of occurrences of $z_i = z$ in z^n, and let $Q_{\hat{x}|z}^n$ be the fraction of occurrences of $(\hat{x}_i, z_i) = (\hat{x}, z)$ among the nQ_z^n pairs $(\hat{x}_i, z_i)$ with $z_i = z$. Given $Z^n = z^n$, the denoiser $\hat{x}^n(\cdot)$ induces a particular $Q_{\hat{x}|z}^n$. Also, the magnitude of the set $I_{\hat{x},z}$ defined above is $nQ_z^n Q_{\hat{x}|z}^n$. To simplify notation, we will not show the dependence of $Q_{\hat{x}|z}^n$ on $\hat{x}^n(\cdot)$ and Z^n. We have

$$P(L_n > D|Z^n = z^n) = P\left(\sum_{z\in\mathcal{Z}}\sum_{\hat{x}\in\hat{\mathcal{X}}}\sum_{i=1}^{nQ_z^n Q_{\hat{x}|z}^n} Y_i^{\hat{x},z} > nD\right)$$

where $\{Y_i^{\hat{x},z}\}$ are independent with $Y_i^{\hat{x},z}$ distributed as $\Lambda(X, \hat{x})$ given $Z = z$.

We now apply Lemma 1 to the random variables $Y_i^{\hat{x},z}$. We will omit the explicit specification of the alphabets of $\hat{x}$ and z, for simplicity. Also, we use $\{Q_z^n\}$ to denote the empirical distribution of z, i.e., $\{Q_1^n, Q_2^n, \ldots, Q_{|\mathcal{Z}|}^n\}$. Similarly, $\{Q_{\hat{x}|z}^n\}$ denotes the collection of conditional empirical distributions of $\hat{x}$ given $z \in \{1, 2, \ldots |\mathcal{Z}|\}$. Then, using Lemma 1, we have that

$$P(L_n > D|Z^n = z^n) \le e^{-nI(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})} \tag{4}$$

and

$$P(L_n > D|Z^n = z^n) \ge e^{-n(I(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})+o(n))} \tag{5}$$

where $o(n)$ is independent of Z^n (and again, is given in Appendix I) and where, for probability distributions (using similar notation to that of the empirical distributions) $\{Q_z\}$ on Z and $\{Q_{\hat{x}|z}\}$ on $\hat{X}$

$$I(D, \{Q_z\}, \{Q_{\hat{x}|z}\}) \tag{6}$$

$$= \sup_{\lambda\ge 0}\left[\lambda D - \sum_{z,\hat{x}} Q_z Q_{\hat{x}|z} \log\left(\sum_x e^{\lambda\Lambda(x,\hat{x})} p(x|z)\right)\right] \tag{7}$$

where $p(x|z)$ is the given conditional distribution of the channel input given the channel output.

We now restrict our attention to the schemes that maximize the exponential rate of decay of

$$P(L_n > D) = \sum_{z^n} P(L_n > D|Z^n = z^n)P(Z^n = z^n) \tag{8}$$

i.e., those that achieve (3). Notice that the probabilities in (4) and (5) depend on z^n and the denoiser only through the joint empirical type of $(\hat{x}^n, z^n)$. We claim that the best (in the sense of maximizing the exponential rate of decay of $P(L_n > D)$), joint empirical type, $(\hat{x}^n, z^n)$, is constant for z^n of the same type. The reason is that the set of possible joint types of $(\hat{x}^n, z^n)$ is identical for z^n of the same type. This is easily seen by considering $z^{n,1}$ and $z^{n,2}$ to be of the same type and noting that, because they are of the same type, there is a bijection $\pi(\cdot)$ from $\mathcal{Z}^n \to \mathcal{Z}^n$ between $z^{n,2}$ and $z^{n,1}$. So, if a particular denoiser $\hat{x}^n(\cdot)$ produces joint type $\{Q_{\hat{x},z}^n\}$ when used on $z^{n,1}$, the denoiser resulting from the composition of $\hat{x}^n(\cdot)$ and $\pi(\cdot)$ on $z^{n,2}$ produces the same joint type $\{Q_{\hat{x},z}^n\}$. The opposite is clearly true since $\pi(\cdot)$ is a bijection.

Since the set of joint types is the same, the best exponent is the same (since it depends only on the joint type). So, rather than summing over all z^n in (8), a set with magnitude exponential in n, we can group the z^n according to their type and sum over the different types, $\{Q_z^n\}$ (a set with magnitude polynomial in n), and we may restrict our attention to denoising functions that depend only on $\{Q_z^n\}$. These facts will help us to express the desired probability as asymptotically equal to an exponential in n, that is, establish the LDP.

We will omit explicit dependence of the notation of the denoiser on $\{Q_z^n\}$ for brevity. Further, we can use the classical typical sequence bounds on the probability that Z^n has type $\{Q_z^n\}$ (c.f. [11]–[13]). Thus, for a particular choice of the denoising functions chosen among the set of optimal denoising functions, i.e., where $Q_{\hat{x}|z}^n$ induced by the denoiser depends on z^n only through its type Q_z^n

$$\begin{aligned} P(L_n > D) &= \sum_{z^n} P(L_n > D|Z^n = z^n)P(Z^n = z^n) \\ &\le \sum_{Z^n=z^n} e^{-nI(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})} P(Z^n = z^n) \\ &= \sum_{\{Q_z^n\}} e^{-nI(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})} P(Z^n \text{ has type } \{Q_z^n\}) \\ &\le \sum_{\{Q_z^n\}} e^{-n(I(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})+D(\{Q_z^n\}||Q))} \end{aligned} \tag{9}$$

where Q denotes the distribution of the channel output, $D(\cdot||\cdot)$ is standard Kullback–Leibler divergence, and

$$\begin{aligned} P(L_n > D) &\ge \sum_{\{Q_z^n\}} e^{-n(I(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})+o(n))} P(Z^n \text{ has type } \{Q_z^n\}) \\ &\ge \sum_{\{Q_z^n\}} e^{-n(I(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})+o(n)+D(\{Q_z^n\}||Q)+\frac{|\mathcal{X}|\log(n+1)}{n})}. \end{aligned} \tag{10}$$

Note that the summations in (9) and (10) are over the set of possible empirical distributions, which is of polynomial size.

The optimal denoiser chooses the best denoising functions given the type of z^n. Denoting the loss of the optimal denoiser by L_n^{opt}, we have

$$P\left(L_n^{\text{opt}} > D|Z^n = z^n\right) \le e^{-n\max_{\{Q_{\hat{x}|z}^n\}} I(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})} \tag{11}$$

and

$$P\left(L_n^{\text{opt}} > D|Z^n = z^n\right) \ge e^{-n(\max_{\{Q_{\hat{x}|z}^n\}} I(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})+o(n))}. \tag{12}$$

Therefore

$$P\left(L_n^{\text{opt}} > D\right) \doteq e^{-n\min_{\{Q_z^n\}}\max_{\{Q_{\hat{x}|z}^n\}}(I(D,\{Q_z^n\},\{Q_{\hat{x}|z}^n\})+D(\{Q_z^n\}||Q))} \tag{13}$$

where we use the notation $a_n \doteq b_n$ to denote that $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$.

4) Optimal Symbol-by-Symbol Denoisers: Now we derive the best performance among the class of symbol-by-symbol denoisers. As stated in Section IV-A1, one option is to use the AVS lemma immediately. We will do this in Section IV-B. Here, however, we will show how to derive the rate function in a manner similar to that of Section IV-A3. The optimal symbol-by-symbol denoiser must choose the denoising functions before observing the realized type so that $\{Q^n_{\hat{x}|z}\}$ is a deterministic mapping (i.e., it is the same for all types $\{Q^n_z\}$). It thus picks a set of denoising functions that maximize the exponent (or minimize $P(L_n > D|Z^n$ has type $\{Q^n_z\}$)) over all types $\{Q^n_z\}$. Denoting the loss of the symbol-by-symbol scheme by L^{ss}_n, we have from (9) and (10) the inequalities (14) and (15) shown at the bottom of the page, so that

$$P\left(L^{\text{ss}}_n > D\right) \doteq e^{-n \max_{\{Q^n_{\hat{x}|z}\}} \min_{\{Q^n_z\}} (I(D,\{Q^n_z\},\{Q^n_{\hat{x}|z}\})+D(\{Q^n_z\}\|Q))}. \tag{16}$$

5) Optimal Rate Functions: We compute (3) by showing that we can move the limit inside the optimizations and then optimizing over a continuum of distributions instead of over the discrete sets $\{Q^n_z\}$ and $\{Q^n_{\hat{x}|z}\}$ which take values in

$$\left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}^{|\mathcal{Z}|} \quad \text{and} \quad \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}^{|\mathcal{X}|}$$

respectively. We thus define a new domain of optimization variables, $\{Q_z\}$ and $\{Q_{\hat{x}|z}\}$, that are continuous valued in $[0,1]^{|\mathcal{Z}|}$ and $[0,1]^{|\mathcal{X}|}$, respectively. We will show that optimizing the rate function in terms of $\{Q_z\}$ and $\{Q_{\hat{x}|z}\}$ instead of $\{Q^n_z\}$ and $\{Q^n_{\hat{x}|z}\}$ is equivalent in the limit of large n. Specifically, we have the following statement.

Definition 1: For $z \in \mathcal{Z}$ and $\hat{x} \in \hat{\mathcal{X}}$, let $Q_z \in [0,1]$ and $Q_{\hat{x}|z} \in [0,1]$, such that $\sum_z Q_z = 1$, and $\sum_{\hat{x}|z} Q_{\hat{x}|z} = 1$ for each z. We can think of Q_z as the frequency of $Z = z$ in Z^n as $n \to \infty$ and, likewise, for $Q_{\hat{x}|z}$. We denote the collection of such frequencies by

$$\{Q_z\} = \{Q_1, Q_2, \ldots, Q_{|\mathcal{Z}|}\}$$

and

$$\{Q_{\hat{x}|z}\} = \{Q_{1\big|1}, Q_{2\big|1}, \ldots, Q_{\hat{\mathcal{X}}\big|1}, \ldots, Q_{|\hat{\mathcal{X}}|\big||\mathcal{Z}|}\}.$$

We now claim that we can move the limit inside the optimization. We let

$$\bar{\Lambda} = \sum_z \sum_{\hat{x}} Q_z Q_{\hat{x}|z} E Y^{\hat{x},z}$$

where $Y^{\hat{x},z}$ is distributed as $\Lambda(X,\hat{x})$ given $Z = z$, and

$$\Lambda = \sum_z \sum_{\hat{x}} Q_z Q_{\hat{x}|z} \max_{x\in\mathcal{X}} \Lambda(x,\hat{x}).$$

Then, we get the following.

Lemma 2: For $\bar{\Lambda} \le D < \Lambda$,

$$\begin{aligned}
&\lim_{n\to\infty} \max_{\{Q^n_{\hat{x}|z}\}} \min_{\{Q^n_z\}} \left(I\left(D, \{Q^n_z\}, \left\{Q^n_{\hat{x}|z}\right\}\right) + D\left(\{Q^n_z\}\|Q\right)\right) \\
&\quad = \max_{\{Q_{\hat{x}|z}\}} \min_{\{Q_z\}} (I(D,\{Q_z\},\{Q_{\hat{x}|z}\}) + D(\{Q_z\}\|Q))
\end{aligned} \tag{17}$$

and

$$\begin{aligned}
&\lim_{n\to\infty} \min_{\{Q^n_z\}} \max_{\left\{Q^n_{\hat{x}|z}\right\}} \left(I\left(D, \{Q^n_z\}, \left\{Q^n_{\hat{x}|z}\right\}\right) + D\left(\{Q^n_z\}\|Q\right)\right) \\
&\quad = \min_{\{Q_z\}} \max_{\{Q_{\hat{x}|z}\}} (I(D,\{Q_z\},\{Q_{\hat{x}|z}\}) + D(\{Q_z\}\|Q)).
\end{aligned} \tag{18}$$

We first need the following claims.

Claim 1: $I(D,\{Q_z\},\{Q_{\hat{x}|z}\})$ is convex in $\{Q_z\}$ and also in $\{Q_{\hat{x}|z}\}$.

Proof: For two different values of $\{Q^n_{\hat{x}|z}\}$ which we will denote $\{Q^n_{\hat{x}|z}\}^{(1)}$ and $\{Q^n_{\hat{x}|z}\}^{(2)}$, and for some $\alpha \in [0,1]$, $\bar{\alpha} = 1 - \alpha$, consider the linear combination

$$\alpha I(D,\{Q_z\},\{Q_{\hat{x}|z}\}^{(1)}) + \bar{\alpha} I(D,\{Q_z\},\{Q_{\hat{x}|z}\}^{(2)}).$$

We have that

$$\begin{aligned}
&\alpha I\left(D,\{Q_z\},\{Q_{\hat{x}|z}\}^{(1)}\right) + \bar{\alpha} I\left(D,\{Q_z\},\{Q_{\hat{x}|z}\}^{(2)}\right) \\
&= \alpha \left[\sup_{\lambda\ge 0} \lambda D - \sum_{z,\hat{x}} Q_z Q^{(1)}_{\hat{x}|z} \log\left(\sum_x e^{\lambda\Lambda(x,\hat{x}(z))} p(x|z)\right)\right] \\
&\quad + \bar{\alpha}\left[\sup_{\lambda\ge 0} \lambda D - \sum_{z,\hat{x}} Q_z Q^{(2)}_{\hat{x}|z} \log\left(\sum_x e^{\lambda\Lambda(x,\hat{x}(z))} p(x|z)\right)\right] \\
&= \sup_{\lambda\ge 0}\left[\alpha\lambda D - \sum_{z,\hat{x}} Q_z \alpha Q^{(1)}_{\hat{x}|z} \log\left(\sum_x e^{\lambda\Lambda(x,\hat{x}(z))} p(x|z)\right)\right] \\
&\quad + \sup_{\lambda\ge 0} [\bar{\alpha}\lambda D - \sum_{z,\hat{x}} Q_z \bar{\alpha} Q^{(2)}_{\hat{x}|z} \log\left(\sum_x e^{\lambda\Lambda(x,\hat{x}(z))} p(x|z)\right)\Bigg] \\
&\ge \alpha\lambda' D - \sum_{z,\hat{x}} Q_z \alpha Q^{(1)}_{\hat{x}|z} \log\left(\sum_x e^{\lambda'\Lambda(x,\hat{x}(z))} p(x|z)\right) \\
&\quad + \bar{\alpha}\lambda'' D - \sum_{z,\hat{x}} Q_z \bar{\alpha} Q^{(1)}_{\hat{x}|z} \log\left(\sum_x e^{\lambda''\Lambda(x,\hat{x}(z))} p(x|z)\right), \\
&\qquad \forall \lambda', \lambda''.
\end{aligned} \tag{19}$$

$$P\left(L^{\text{ss}}_n > D\right) \le \min_{\{Q^n_{\hat{x}|z}\}} \sum_{\{Q^n_z\}} e^{-n(I(D,\{Q^n_z\},\{Q^n_{\hat{x}|z}\})+D(\{Q^n_z\}\|Q))} \tag{14}$$

and

$$P\left(L^{\text{ss}}_n > D\right) \ge \min_{\{Q^n_{\hat{x}|z}\}} \sum_{\{Q^n_z\}} \left(e^{-n(I(D,\{Q^n_z\},\{Q^n_{\hat{x}|z}\})+o(n)+D(\{Q^n_z\}\|Q)+\frac{|\mathcal{X}|\log(n+1)}{n})}\right) \tag{15}$$

Thus, since λ' and λ'' are arbitrary, they can be chosen to be the λ that achieves

$$\sup_{\lambda\geq 0}\left[\lambda D-\sum_{z,\hat{x}}Q_z\left(\alpha Q^{(1)}_{\hat{x}|z}+\bar{\alpha}Q^{(2)}_{\hat{x}|z}\right)\times\log\left(\sum_{x}e^{\lambda\Lambda(x,\hat{x}(z))}p(x|z)\right)\right]\tag{20}$$

which exists since the objective is continuous in λ. Thus, (19) $\geq$ (20). Notice that (20) is simply $I(D,\{Q_z\},\alpha\{Q_{\hat{x}|z}\}+\bar{\alpha}\{Q_{\hat{x}|z}\})$. So we have that $I(D,\{Q_z\},\{Q_{\hat{x}|z}\})$ is convex in $\{Q_{\hat{x}|z}\}$. A similar argument holds for $\{Q_z\}$. Thus, we have Claim 1. □

Claim 2: $I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)$ is convex in $\{Q_z\}$ and also in $\{Q_{\hat{x}|z}\}$.

Proof: The claim follows from the previous claim and the fact that $D(\{Q_z\}\|Q)$ is convex in $\{Q_z\}$ and independent of $\{Q_{\hat{x}|z}\}$. □

Claim 3: For $\bar{\Lambda}\leq D<\Lambda$, and M satisfying $0<M<\infty$, $I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)$ is uniformly continuous in $\{Q_{\hat{x}|z}\}$ and uniformly continuous in $\{\{Q_z\}:D(\{Q_z\}\|Q)\leq M\}$.

Proof: The set of allowable $\{Q_{\hat{x}|z}\}$ is closed and bounded and thus compact. Since $I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)$ is convex in $\{Q_{\hat{x}|z}\}$, it is uniformly continuous in $\{Q_{\hat{x}|z}\}$.

Since $D(\{Q_z\}\|Q)$ is continuous where finite, the set $\{\{Q_z\}:D(\{Q_z\}\|Q)\leq M\}$ is closed. Clearly, this set is also bounded since the range of $\{Q_z\}$ is bounded. Uniform continuity follows from the compactness of this set and the convexity of $I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)$. It is also clear that the types $\{Q_z\}$ such that $D(\{Q_z\}\|Q)=\infty$ cannot minimize the rate function, so we may assume the existence of some $M>0$ such that the optimization is equivalent to optimizing over $\{\{Q_z\}:D(\{Q_z\}\|Q)\leq M\}$. To simplify the notation, we will not state this restricted range of values of $\{Q_z\}$ explicitly in the following. □

We are now ready to prove Lemma 2.

Proof of Lemma 2: For all $\eta_1>0$, there exists an N such that $n>N$ implies

$$\max\left(\min_{\{Q^n_z\}}\|\{Q^n_z\}-\{Q_z\}\|_2,\ \min_{\left\{Q^n_{\hat{x}|z}\right\}}\|\left\{Q^n_{\hat{x}|z}\right\}-\{Q_{\hat{x}|z}\}\|_2\right)<\eta_1$$

for all $\{Q_z\},\{Q_{\hat{x}|z}\}$ since we can approximate a point in $[0,1]$ arbitrarily well by a point in $\{0,\frac{1}{n},\frac{2}{n},\ldots,1\}$ as $n\to\infty$.

Since $I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)$ is uniformly continuous in $\{Q_{\hat{x}|z}\}$, for all $\eta_2>0$, there is an $\eta_3>0$ such that

$$\min_{\left\{Q^n_{\hat{x}|z}\right\}}\|\left\{Q^n_{\hat{x}|z}\right\}-\{Q_{\hat{x}|z}\}\|_2<\eta_3$$

implies

$$|(I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)) - (I(D,\{Q_z\},\{Q^n_{\hat{x}|z}\})+D(\{Q_z\}\|Q))|<\eta_2.$$

The same is true using $\{Q_z\}$ and $\{Q^n_z\}$. So, there is an N such that for all $n>N$

$$|\left(I(D,\{Q_z\},\{Q_{\hat{x}|z}\}\right)+D(\{Q_z\}\|Q)) - \left(I(D,\{Q_z\},\{Q^n_{\hat{x}|z}\}\right)+D(\{Q_z\}\|Q))|<\eta_2$$

and

$$|\left(I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)\right) -(I(D,\{Q^n_z\},\{Q_{\hat{x}|z}\})+D(\{Q^n_z\}\|Q))\,|<\eta_2.$$

Thus, for $n>N$

$$\begin{aligned}\max_{\{Q_{\hat{x}|z}\}}\min_{\{Q_z\}}&I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)\\ &\leq\max_{\{Q_{\hat{x}|z}\}}\min_{\{Q^n_z\}}I(D,\{Q^n_z\},\{Q_{\hat{x}|z}\})+D(\{Q^n_z\}\|Q)\\ &\leq\max_{\left\{Q^n_{\hat{x}|z}\right\}}\min_{\{Q^n_z\}}I\left(D,\{Q^n_z\},\left\{Q^n_{\hat{x}|z}\right\}\right)+D(\{Q^n_z\}\|Q)+\eta_2\end{aligned}$$

and

$$\begin{aligned}\max_{\{Q_{\hat{x}|z}\}}\min_{\{Q_z\}}&I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)\\ &\geq\max_{\left\{Q^n_{\hat{x}|z}\right\}}\min_{\{Q_z\}}I(D,\{Q_z\},\left\{Q^n_{\hat{x}|z}\right\})+D(\{Q_z\}\|Q)\\ &\geq\max_{\left\{Q^n_{\hat{x}|z}\right\}}\min_{\{Q^n_z\}}I(D,\{Q^n_z\},\left\{Q^n_{\hat{x}|z}\right\})+D\left(\{Q^n_z\}\|Q\right)-\eta_2.\end{aligned}$$

Since η_2 is arbitrary, we have the first part of the lemma. The second part follows analogously. □

Thus, combining Lemma 2 with (13) and (16) shows that $R(\mathcal{C}_1,D)$ and $R(\mathcal{C}_2,D)$ are well defined and that

$$R(\mathcal{C}_1,D)=\min_{\{Q_z\}}\max_{\{Q_{\hat{x}|z}\}}(I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q))\tag{21}$$

and

$$R(\mathcal{C}_2,D)=\max_{\{Q_{\hat{x}|z}\}}\min_{\{Q_z\}}(I(D,\{Q_z\},\{Q_{\hat{x}|z}\})+D(\{Q_z\}\|Q)).\tag{22}$$

B. Alternate Derivation of $R(\mathcal{C}_2,D)$ and Rate Function for Symbol-by-Symbol Denoisers

We can find the rate function for a symbol-by-symbol denoiser by noting that $\Lambda(X_i,\hat{x}_i(Z_i))$ is an AVS with distribution depending only on the denoising function $\hat{x}_i(\,\cdot\,)$. There are a finite number, $|\hat{\mathcal{X}}|^{|\mathcal{Z}|}$, of different denoising functions $\hat{x}(\,\cdot\,)$ which we will now label $\hat{x}^{(j)}(\,\cdot\,)$, $j=1,\ldots,|\hat{\mathcal{X}}|^{|\mathcal{Z}|}$. Let $\theta^{(n)}_j$ be the fraction of times $\hat{x}^{(j)}(\,\cdot\,)$ appears (there is a dependence on n since there are n total observations being denoised). Then, we can apply Lemma 1 to conclude that, for any symbol-by-symbol denoiser, we have (23) which is shown at the bottom of the page, thus establishing the last part of Theorem 1.

We get an alternate derivation of the optimal rate function by optimizing over the $\{\theta_j^{(n)}\}$ to get (24), which is also shown at the bottom of the page. Equation (24) follows from the fact that the optimal denoising function, $\hat{x}_j(\,\cdot\,)$, depends only on the distribution of (X, Z) for all j and is therefore the same function for all j. In other words, the optimal symbol-by-symbol scheme is time-invariant, establishing part of Theorem 2. We now have another expression for $R(\mathcal{C}_2, D)$, namely

$$R(\mathcal{C}_2, D) = \max_{\hat{x}(\,\cdot\,)} \sup_{\lambda>0} [\lambda D - \log E_{X,Z} \exp(\lambda\Lambda(X, \hat{x}(Z)))]. \tag{25}$$

C. Optimal Denoising and Theorem 2

1) Theory: Our goal is to investigate whether $P(L_n^{\text{opt}} > D) \doteq P(L_n^{\text{ss}} > D)$, i.e., whether $R(\mathcal{C}_1, D) = R(\mathcal{C}_2, D)$. We can see the following from (21).

Lemma 3: For Hamming loss, symbol-by-symbol denoising is optimal.

Proof: For all types $\{Q_z\}, I(D, \{Q_z\}, \{Q_{\hat{x}|z}\})$ (see (6)) is maximized over $\{Q_{\hat{x}|z}\}$ by $Q_{\hat{x}|z} = \mathbb{1}(\{\hat{x} = \arg\max_x p(x|z)\})$, i.e., the deterministic conditional distribution that sets $\hat{x}$ as the most likely x given z. □

We thus have Theorem 3. Using (21) and (22), we now show the following.

Lemma 4: The best denoisers and the best symbol-by-symbol denoisers are time invariant.

Proof: The best denoiser picks a conditional distribution $Q_{\hat{x}|z}$ based on Q_z. It is easy to extend Claim 1 to continuous distributions. Thus, for a fixed $\{Q_z\}$, $I(D, \{Q_z\}, \{Q_{\hat{x}|z}\}) + D(\{Q_z\}\|Q)$ is convex in $Q_{\hat{x}|z}$. Thus, for each $z \in \mathcal{Z}$, the best choice of $Q_{\hat{x}|z}$ sets 1 for some $\hat{x}$ and 0 otherwise, i.e., the best denoiser is time invariant. The best symbol-by-symbol denoiser chooses $\{Q_{\hat{x}|z}\}$ to maximize $\min_{\{Q_z\}}(I(D, \{Q_z\}, \{Q_{\hat{x}|z}\}) + D(\{Q_z\}\|Q))$. It is easy to see that this expression is convex in $\{Q_{\hat{x}|z}\}$. So, for each $z \in \mathcal{Z}$, the best symbol-by-symbol denoiser has $Q_{\hat{x}|z}$ equal to 1 for some $\hat{x}$ and equal to 0 otherwise. Thus, it is time invariant. □

2) Concrete Examples of Suboptimality: We now show that there are cases for which symbol-by-symbol denoising is strictly suboptimal, i.e., the inequality between (21) and (22) is strict.

We will consider a binary-symmetric channel (BSC) with crossover probability δ. We use the notation $\mathrm{BSC}(\delta)$ to refer to such a channel. Consider a $\mathrm{Bernoulli}(p), p = 0.5$, source that passes through a $\mathrm{BSC}(\delta)$ with $\delta < 0.5$. We define an asymmetric loss function where a loss of 1 is incurred when we decode 0 as 1 and a loss of 2 is incurred when we decode 1 as 0.

By Lemma 4, it is clear that the best denoiser has $Q_{0|z} = 0$ or 1 for $z = 0, 1$. So, for a given $\{Q_z\}$, there is no need to time-share; the best denoiser makes the same decision at each time for the same output symbol z. Thus, there are only four possible optimal denoising schemes: say-what-you see (SWYS), say-the-opposite ($\overline{\text{SWYS}}$), decode all ones (ONES), and decode all zeros (ZEROS). We represent the denoising decision by $\hat{x}(\,\cdot\,)$ which takes a single symbol z as an argument. So, for the SWYS denoiser, $\hat{x}(z) = z$ and for the ONES denoiser, $\hat{x}(z) = 1$.

Since we are in the binary setting, we can simplify the notation for the frequency/distribution vectors. Instead of writing $\{Q_z\}$ or Q, we specify the frequency/distribution, respectively, by Q_0 or $Q(0)$. In our setting, the objective function, $I(D, \{Q_z^n\}, \{Q_{\hat{x}|z}^n\}) + D(\{Q_z^n\}\|Q)$, for a particular denoiser, $\hat{x}(\,\cdot\,)$, is

$$\begin{aligned}
&\sup_{\lambda\geq 0} \lambda D - Q_0 \log\left(e^{\lambda\Lambda(0,\hat{x}(0))}\frac{p(0,0)}{Q(0)} + e^{\lambda\Lambda(1,\hat{x}(0))}\frac{p(1,0)}{Q(0)}\right)\\
&\quad - Q_1 \log\left(e^{\lambda\Lambda(0,\hat{x}(1))}\frac{p(0,1)}{Q(1)} + e^{\lambda\Lambda(1,\hat{x}(1))}\frac{p(1,1)}{Q(1)}\right)\\
&\qquad + Q_0 \log\frac{Q_0}{Q(0)} + Q_1 \log\frac{Q_1}{Q(1)} \qquad (26)\\
&= \sup_{\lambda\geq 0} \lambda D - \log E_{X,Z} e^{\lambda\Lambda(X,\hat{x}(Z))}\\
&\qquad + D_b\left(Q_0 \,\middle\|\, \frac{\sum_x p(x,0) e^{\lambda\Lambda(X,\hat{x}(0))}}{E_{X,Z} e^{\lambda\Lambda(X,\hat{x}(Z))}}\right) \qquad (27)
\end{aligned}$$

where D_b denotes binary divergence. That is

$$D_b(p(0)\|q(0)) = p(0) \log\frac{p(0)}{q(0)} + p(1) \log\frac{p(1)}{q(1)}$$

where p and q are probability distributions on $\{0, 1\}$.

We can now make the following claim.

Claim 4: $\overline{\text{SWYS}}$ and ZEROS are suboptimal.

$$P(L_n > D) \doteq e^{-n \sup_{\lambda>0}\left[\lambda D - \sum_{j=1}^{|\hat{\mathcal{X}}|^{|\mathcal{Z}|}} \theta_j^{(n)} \log E_{X,Z} \exp(\lambda\Lambda(X, \hat{x}^{(j)}(Z)))\right]} \tag{23}$$

$$\begin{aligned}
P(L_n^{\text{ss}} > D) &\doteq \min_{\{\theta_j^n\}} e^{-n \sup_{\lambda>0}\left[\lambda D - \sum_{j=1}^{|\hat{\mathcal{X}}|^{|\mathcal{Z}|}} \theta_j^{(n)} \log E_{X,Z} \exp(\lambda\Lambda(X, \hat{x}^{(j)}(Z)))\right]}\\
&\doteq e^{-n \max_{\{\theta_j^n\}} \sup_{\lambda>0}\left[\sum_{j=1}^{|\hat{\mathcal{X}}|^{|\mathcal{Z}|}} \theta_j^{(n)}\left(\lambda D - \log E_{X,Z} \exp(\lambda\Lambda(X, \hat{x}^{(j)}(Z)))\right)\right]}\\
&\doteq e^{-n \max_{\{\theta_j^n\}} \sum_{j=1}^{|\hat{\mathcal{X}}|^{|\mathcal{Z}|}} \theta_j^{(n)} \sup_{\lambda>0}\left[\lambda D - \log E_{X,Z} \exp(\lambda\Lambda(X, \hat{x}^{(j)}(Z)))\right]}\\
&\doteq e^{-n \max_{\hat{x}(\,\cdot\,)} \sup_{\lambda>0}[\lambda D - \log E_{X,Z} \exp(\lambda\Lambda(X, \hat{x}(Z)))]}.
\end{aligned} \tag{24}$$

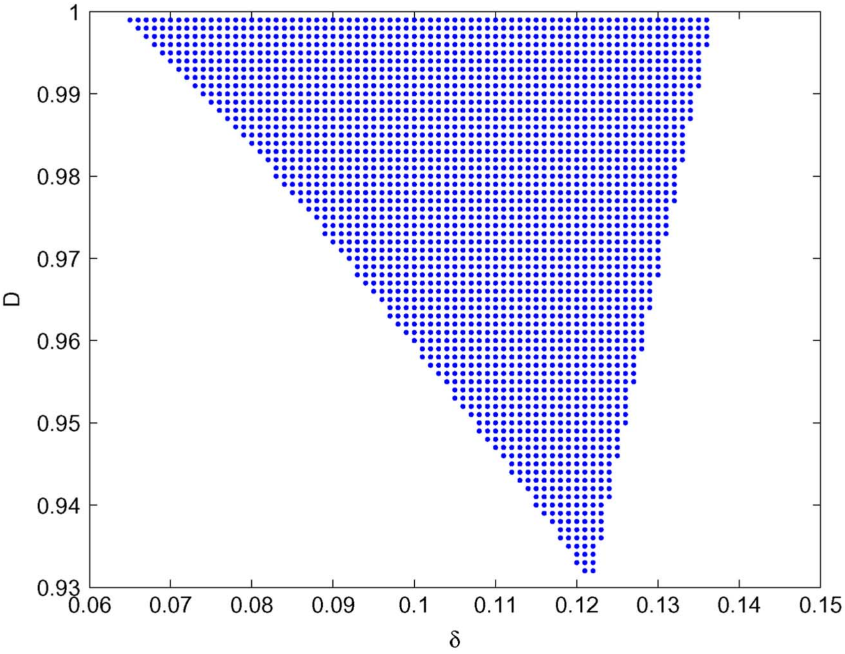

Fig. 8. Region of symbol-by-symbol suboptimality.

Proof: For SWYS, the quantity inside the supremum of (26) is

$$\lambda D - Q_0 \log(\bar{\delta} + e^{2\lambda}\delta) - Q_1 \log(e^{\lambda}\delta + \bar{\delta}). \tag{28}$$

For $\overline{\text{SWYS}}$, it is

$$\lambda D - Q_1 \log(\delta + e^{2\lambda}\bar{\delta}) - Q_0 \log(e^{\lambda}\bar{\delta} + \delta). \tag{29}$$

Since $\delta < 0.5$, (28) for $Q_0 = Q_0^1$ is greater than or equal to (29) for $Q_0 = 1 - Q_0^1$, for each $\lambda \geq 0$. So, for any fixed denoiser and $Q_0 = Q_0^1$, (26) for SWYS is better than for $\overline{\text{SWYS}}$ for that same denoiser and $Q_0 = 1 - Q_0^1$. Thus, the performance of SWYS is better than that of $\overline{\text{SWYS}}$.

Since $p = 0.5$, we are as likely to incorrectly decode a 1 as we are to incorrectly decode a 0. Since it is more costly to mistake a 1, the ONES denoiser is better than the ZEROS denoiser. □

We also have the following claim.

Claim 5:

$$\lambda D - \log E_{X,Z} e^{\lambda\Lambda(X,\hat{x}(Z))} + D_b\left(Q_0 \middle\| \frac{\sum_x p(x,0)e^{\lambda\Lambda(X,\hat{x}(0))}}{E_{X,Z}e^{\lambda\Lambda(X,\hat{x}(Z))}}\right) \tag{30}$$

is concave in λ.

Proof: Since log convexity is preserved under sums and $e^{\lambda a}, a \in \mathbb{R}$ is log convex in λ, the terms inside the logarithm of (26) are log convex. Hence, the log terms are convex and so (30) is concave in λ. □

Now, (17) can be expressed as follows:

$$\max_{\hat{x}(\cdot)} \min_{Q_0} \sup_{\lambda\geq 0} \lambda D - \log E_{X,Z} e^{\lambda\Lambda(X,\hat{x}(Z))} + D_b\left(Q_0 \middle\| \frac{\sum_x p(x,0)e^{\lambda\Lambda(X,\hat{x}(0))}}{E_{X,Z}e^{\lambda\Lambda(X,\hat{x}(Z))}}\right) \tag{31}$$

$$= \max_{\hat{x}(\cdot)} \sup_{\lambda\geq 0} \min_{Q_0} \lambda D - \log E_{X,Z} e^{\lambda\Lambda(X,\hat{x}(Z))} + D_b\left(Q_0 \middle\| \frac{\sum_x p(x,0)e^{\lambda\Lambda(X,\hat{x}(0))}}{E_{X,Z}e^{\lambda\Lambda(X,\hat{x}(Z))}}\right) \tag{32}$$

$$= \max_{\hat{x}(\cdot)} \sup_{\lambda\geq 0} \lambda D - \log E_{X,Z} e^{\lambda\Lambda(X,\hat{x}(Z))} \tag{33}$$

where we can switch the min and sup in (31) to get (32) since the objective is convex in the minimization variable Q_0 and concave in λ, the variable over which the supremum is taken. Equality (33) follows by setting

$$Q_0 = \frac{\sum_x p(x,0)e^{\lambda\Lambda(X,\hat{x}(0))}}{E_{X,Z}e^{\lambda\Lambda(X,\hat{x}(Z))}}$$

so that the binary divergence term is minimized. We could have obtained (33) directly from (23) but we re-derived it here because we use the form of (32) in the following.

Since only the SWYS and ONES denoisers can be optimal, the problem reduces to comparing the exponents of these two denoisers. That is, we use (33) and substitute either the SWYS or ONES function for $\hat{x}(\cdot)$. We use a Matlab program to search the space of channels in terms of δ and the range of thresholds, D, to determine for which channels and threshold values symbol-by-symbol denoising is strictly suboptimal. Although the region of such (δ, D) pairs is computed numerically, each point in the region can be verified analytically to show the suboptimality of symbol-by-symbol denoising. The details of our method and an explanation of why each point in the region can be verified analytically can be found in Appendix II. Fig. 8 shows a plot of the region of (δ, D) for which symbol-by-symbol denoising is suboptimal. This concludes the proof of Theorem 2.

D. $R(\mathcal{C}_3(k), D)$

We derive an expression for the exponent of the probability of excess loss for the k-finite block denoiser and show that the best

k-finite block denoiser does no better than the best symbol-by-symbol denoiser.

For simplicity, assume that $n = mk$, for integer m. Since we take $\lim_{n\to\infty} -\frac{1}{n}\log \mathrm{P}(L_n > D)$, it will be clear that this does not affect the validity of the derivation. We index the set of denoising functions $\hat{x}^k(\cdot) : \mathcal{Z}^k \to \hat{\mathcal{X}}^k$ by $i \in \mathcal{I}_k = \{1, 2, \ldots, |\hat{\mathcal{X}}|^{k|\mathcal{Z}|^k}\}$ to get $\hat{x}^{k,(i)}(\cdot)$. Note that for fixed k, this set is finite. Now, given $n = mk$, the most general deterministic scheme will use a certain fraction of each type of denoising function. Denote the fraction of time denoiser i is used by $\theta_i^{(m)} \in \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$. Since the $(X_{(j-1)k+1}^{jk}, Z_{(j-1)k+1}^{jk})$ are i.i.d., for a particular i, $\Lambda(X_{(j-1)k+1}^{jk}, \hat{x}^{k,(i)}(Z_{(j-1)k+1}^{jk}))$ are i.i.d. So, $\Lambda(X^k, \hat{x}^{k,(i)}(Z^k))$ is an AVS. Thus

$$
\begin{aligned}
P(L_n > D) &= P\left(\sum_{j=1}^{n} \Lambda(X_j, \hat{X}_j) > nD\right)\\
&= P\left(\sum_{j=1}^{m} \Lambda\left(X_{(j-1)k+1}^{jk}, \hat{X}_{(j-1)k+1}^{jk}\right) > nD\right)\\
&= P\left(\frac{1}{m}\sum_{j=1}^{m} \Lambda\left(X_{(j-1)k+1}^{jk}, \hat{X}_{(j-1)k+1}^{jk}\right) > kD\right)\\
&= P\left(\frac{1}{m}\sum_{i=1}^{|\mathcal{I}_k|}\sum_{j=1}^{m\theta_i^{(m)}} \Lambda\left(X^{k,(i,j)}, \hat{x}^{k,(i)}(Z^{k,(i,j)})\right) > kD\right)
\end{aligned}
$$

where for each (i,j), $(X^{k,(i,j)}, Z^{k,(i,j)})$ has the same distribution as (X^k, Z^k). We can therefore use Lemma 1 on the AVS $\Lambda(X^k, \hat{x}^{k,(i)}(Z^k))$ to compute

$$
\begin{aligned}
&-\frac{1}{n}\log P(L_n > D)\\
&\quad = -\frac{1}{mk}\log P\left(\frac{1}{m}\sum_{i=1}^{|\mathcal{I}_k|}\sum_{j=1}^{m\theta_i^{(m)}} \right.\\
&\qquad\qquad \left.\Lambda\left(X^{k,(i,j)}, \hat{x}^{k,(i)}\left(Z^{k,(i,j)}\right)\right) > kD\right)\\
&\quad = \frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda kD - \sum_{i=1}^{|\mathcal{I}_k|}\theta_i^{(m)}\log E_{X^k,Z^k} e^{\lambda\Lambda\left(X^k,\hat{x}^{k,(i)}(Z^k)\right)}\right]\\
&\qquad + o(m)
\end{aligned}
$$

where $o(m) \to 0$ as $m \to \infty$. Since we are concerned about the behavior for fixed k as $n \to \infty$, we have $m \to \infty$. So, we can neglect the $o(m)$ term and optimize over $\{\theta_i\}$, $\theta_i \in [0,1]$ instead of over the $\{\theta_i^{(m)}\}$. We can rewrite this as

$$
\begin{aligned}
&\frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda kD - \sum_{i=1}^{|\mathcal{I}_k|}\theta_i \log\right.\\
&\qquad \left.\times\left(\sum_{x^k,z^k}\left(p(x^k,z^k)e^{\lambda\Lambda\left(x^k,\hat{x}^{k,(i)}(z^k)\right)}\right)\right)\right]\\
&= \frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda kD - \sum_{i=1}^{|\mathcal{I}_k|}\theta_i\log\left(\sum_{x^k,z^k}\left(p(z^k)p(x^k|z^k)\right.\right.\right.\\
&\qquad \left.\left.\left.\times\, e^{\lambda\Lambda\left(x^k,\hat{x}^{k,(i)}(z^k)\right)}\right)\right)\right]\\
&= \frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda kD - \sum_{i=1}^{|\mathcal{I}_k|}\theta_i\log\left(\sum_{x^k,z^k}\left(p(z^k)\right.\right.\right.\\
&\qquad \left.\left.\left.\times\prod_{j=1}^{k}\left(p(x_j|z_j)\right)e^{\sum_{j=1}^{k}\lambda\Lambda\left(x_j,\hat{x}_j^{(i)}(z^k)\right)}\right)\right)\right]\\
&= \frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda kD - \sum_{i=1}^{|\mathcal{I}_k|}\theta_i\log\left(\sum_{x^k,z^k}\left(p(z^k)\right.\right.\right.\\
&\qquad \left.\left.\left.\times\prod_{j=1}^{k}\left(p(x_j|z_j)e^{\lambda\Lambda\left(x_j,\hat{x}_j^{(i)}(z^k)\right)}\right)\right)\right)\right]\\
&= \frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda kD - \sum_{i=1}^{|\mathcal{I}_k|}\theta_i\log\left(\sum_{z^k}\left(p(z^k)\right.\right.\right.\\
&\qquad \left.\left.\left.\times\prod_{j=1}^{k}\sum_{x_j}\left(p(x_j|z_j)e^{\lambda\Lambda\left(x_j,\hat{x}_j^{(i)}(z^k)\right)}\right)\right)\right)\right]\\
&= \frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda kD - \sum_{i=1}^{|\mathcal{I}_k|}\theta_i\log\left(\sum_{z^k}\left(p(z^k)\right.\right.\right.\\
&\qquad \left.\left.\left.\times\prod_{j=1}^{k}\sum_{x}\left(p(x|z_j)e^{\lambda\Lambda\left(x,\hat{x}_j^{(i)}(z^k)\right)}\right)\right)\right)\right].
\end{aligned}
$$

To maximize this expression, we should set $\theta_i = 1$ for

$$
\hat{x}_j^{(i)}(z^k) = \mathrm{argmin}_{\hat{x}}\sum_{x} p(x|z_j)e^{\lambda\Lambda(x,\hat{x})}
$$

and 0 else, since this minimizes $\sum_x p(x|z_j)e^{\lambda\Lambda(x,\hat{x}_j^{(i)}(z^k))}$ for each j. So, the best denoiser uses the time-invariant, symbol-by-symbol function

$$
\hat{x}(z) = \mathrm{argmin}_{\hat{x}}\sum_{x} p(x|z)e^{\lambda\Lambda(x,\hat{x})}
$$

where λ achieves the above supremum. Since the function is continuous in λ, starts at 0 for $\lambda = 0$ and tends to $-\infty$ as $\lambda \to \infty$, the supremum is achieved and so our definition of the optimal function is valid. Letting $\hat{x}(\cdot)$ denote this choice of denoising function, we thus have

$$
\begin{aligned}
&\frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda kD - \log\left(\sum_{x^k,z^k} p(x^k,z^k)\prod_{j=1}^{k}\left(e^{\lambda\Lambda(x_j,\hat{x}(z_j))}\right)\right)\right]\\
&\quad = \frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda kD - \log\left(\sum_{x^{k-1},z^{k-1}} p(x^{k-1},z^{k-1})\right.\right.\\
&\qquad \left.\left.\times\prod_{j=1}^{k-1}\left(e^{\lambda\Lambda(x_j,\hat{x}(z_j))}\right)\sum_{x_k,z_k} p(x_k,z_k)e^{\lambda\Lambda(x_k,\hat{x}(z_k))}\right)\right]
\end{aligned}
$$

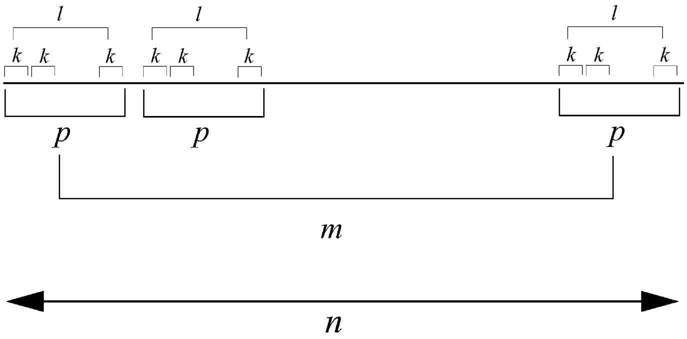

Fig. 9. Illustration for sliding-window denoiser proof.

$$
\begin{aligned}
&= \frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda k D - \log\left(\prod_{j=1}^{k}\sum_{x_j,z_j} p(x_j,z_j)e^{\lambda\Lambda(x_j,\hat{x}(z_j))}\right)\right]\\
&= \frac{1}{k}\sup_{\lambda\geq 0}\left[\lambda k D - \log\left(\sum_{x,z} p(x,z)e^{\lambda\Lambda(x,\hat{x}(z))}\right)^{k}\right]\\
&= \sup_{\lambda\geq 0}\left[\lambda D - \log\left(\sum_{x,z} p(x,z)e^{\lambda\Lambda(x,\hat{x}(z))}\right)\right]\\
&= R(\mathcal{C}_3(k),D).
\end{aligned}
$$

Clearly, the value of λ that maximizes this expression coincides with the value of λ that maximizes the symbol-by-symbol denoiser rate function (25). Therefore, finite block denoisers have the same performance as symbol-by-symbol denoisers, i.e., $R(\mathcal{C}_2,D)=R(\mathcal{C}_3(k),D)$, giving the third part of Theorem 1 and the first part of Theorem 4.

E. $R(\mathcal{C}_4(k),D)$

We show that, given a k-finite sliding-window denoiser, we can find a sequence of finite block denoisers of increasing order whose performance is a lower bound on $P(L_n > D)$ for the k-finite sliding-window denoiser. Consider a p-finite block denoiser, where $n=mp$, $p=lk$, and $l\gg 1$ and m,l are integers. It is straightforward to extend the argument to general m,l. We now show that

$$
\begin{aligned}
P(L_n > D) &= P\left(\sum_{i=1}^{n}\Lambda\left(X_i,\hat{x}_i\left(Z_{i-k}^{i+k}\right)\right) > nD\right)\\
&\geq P\left(\sum_{j=1}^{m}\Lambda\left(X_{(j-1)p+1}^{jp},\hat{x}_{(j-1)p+1}^{jp}\left(Z_{(j-1)p+1}^{jp}\right)\right)\right.\\
&\qquad \left. > n\left(D+\frac{2\Lambda_{\max}k}{p}\right)\right).
\end{aligned}
$$

This latter expression is the probability of excess loss of a p-finite block denoiser with threshold $D+\frac{2\Lambda_{\max}k}{p}$ and uses the notation for block denoisers defined in Section II-A. Fig. 9 illustrates the reason for the inequality. The inequality follows from the fact that the p-finite block denoiser has more information than the k-finite sliding-window denoiser for all indices except those of the form $jp-k+1, jp-k+2,\ldots,jp+k$. So the best block denoiser does at least as well as the sliding-window denoiser for indices that are not of this form. Furthermore, the loss for indices that are of this form cannot exceed $\Lambda_{\max}$. Since there are $2km$ such indices, increasing nD by $2km\Lambda_{\max}$ to get $n(D+\frac{2\Lambda_{\max}k}{p})$ gives the lower bound. We know the p-finite block denoiser can do no better than the optimal symbol-by-symbol denoiser with the same threshold, i.e., $D+\frac{2\Lambda_{\max}k}{p}$. Since the exponent is continuous in the threshold parameter and p was arbitrary, taking $p\to\infty$ gives us a tighter lower bound on the exponent associated with the probability of excess loss of a sliding-window denoiser. This lower bound is the probability of excess loss for an optimal symbol-by-symbol denoiser with parameter D.

It is obvious that the best finite sliding-window denoiser is no worse than a symbol-by-symbol denoiser of the same threshold. Thus, the performance of the best finite sliding-window denoiser is the same as the performance of the best symbol-by-symbol denoiser, i.e., $R(\mathcal{C}_4(k),D)=R(\mathcal{C}_2,D)$. So we have another part of Theorem 1 and of Theorem 4.

F. Filtering

1) Infinite Memory Filters: Explicit characterization of the performance of the infinite memory filter appears to be difficult. This characterization shares some intricacies with the characterization of the exponent of zero-delay, infinite memory source codes, which is mentioned but left open in [5]. It is not clear how to use the AVS lemma (Lemma 1) since the single-letter losses at different times are dependent on the infinite memory filter. This was also the case with the finite sliding-window denoiser, but because the memory and look-ahead were finite, we could get a handle on the rate function by using a series of finite block denoisers to upper-bound it. We are, however, able to characterize the performance of the finite memory filter, by sandwiching its performance between schemes whose performance we already know.

2) $R(\mathcal{C}_5(k),D)$: The analysis of finite memory filters is greatly simplified by the preceding results for denoisers. We observe that the set of k-finite sliding-window denoising functions includes the set of k-finite memory filtering functions which includes the set of symbol-by-symbol denoising/filtering functions. The equivalence of the best k-finite sliding-window filter and the best symbol-by-symbol denoiser/filter thus implies the performance of finite memory filters is the same as that of symbol-by-symbol filters, i.e., $R(\mathcal{C}_5(k),D)=R(\mathcal{C}_2,D)$. This gives the remaining parts of Theorems 1 and 4.

V. Conclusion and Future Work

We have studied the effect of limiting the domain of denoising and filtering functions under the probability of excess loss criterion. We established Theorems 1–4, which we now rephrase.

Symbol-by-symbol denoising of a DMC-corrupted memoryless source is found to be suboptimal using a general single-letter loss function, under the probability of excess loss criterion. In the case of Hamming loss, symbol-by-symbol denoising is optimal. In general, the best denoising and symbol-by-symbol denoising schemes are time invariant.

A region of suboptimality for a Bern($\frac{1}{2}$) source passing through a $\mathrm{BSC}(\delta)$ under an asymmetric loss function was found numerically. Each point of the region can be verified analytically, but an analytical characterization of the region of suboptimality is yet to be found and may be of interest.

We have shown that finite memory filters, finite sliding-window denoisers, and finite block denoisers all do no better than time-invariant symbol-by-symbol denoisers/filters.

We note that the case where the filter has unbounded memory is also of interest. An open question is how to characterize the performance of these infinite memory filters, or even to determine whether/when the performance is strictly better or worse than that of symbol-by-symbol filters and optimal denoisers, respectively.

Appendix I
Proof of the AVS Lemma

We use the notation given in the statement of Lemma 1. We also define the following quantities which are used in the proof. We let $\bar{w}_n = \sum_{a=1}^{r} Q_a^n E_{P_a} W$, $\Delta_{\max} = \max_a \Delta_a$, $\Delta_{(n)} = \sum_{a=1}^{r} Q_a^n \Delta_a$, and $\Psi_a^*(x) = \sup_{\lambda \geq 0} [\lambda x - \Psi_a(\lambda)]$.

We start by showing the following.

Claim 6: $\lambda x - \log \Psi_a(\lambda)$ is concave in λ.

Proof: Since $e^{\lambda w}$ is log convex in λ and log convexity is preserved under sums, $\Psi_a(\lambda)$ is a log-convex function of λ. This implies that $\log \Psi_a(\lambda)$ is convex in λ. Claim 6 follows. □

Clearly, $\lambda x - \log \Psi_a(\lambda)$ is 0 when $\lambda = 0$. If $x < \Delta_a$

$$\begin{aligned}
&\lim_{\lambda \to \infty} [\lambda x - \log \Psi_a(\lambda)] \\
&= \lim_{\lambda \to \infty} \log \frac{e^{\lambda x}}{\sum_{w \in S(P_a)} P_a(w) e^{\lambda w}} \\
&= \log \lim_{\lambda \to \infty} \frac{e^{\lambda(x - \Delta_a)}}{P_a(\Delta_a) + \sum_{w \in S(P_a), w \neq \Delta_a} P_a(w) e^{\lambda(w - \Delta_a)}} \\
&= -\infty,
\end{aligned}$$

since $P_a(\Delta_a) > 0$. Also,

$$\begin{aligned}
&\lim_{\lambda \to 0} \frac{d}{d\lambda} [\lambda x - \log \Psi_a(\lambda)] \\
&\quad = \lim_{\lambda \to 0} \left[x - \frac{\sum_{w \in S(P_a)} w P_a(w) e^{\lambda w}}{\sum_{w \in S(P_a)} P_a(w) e^{\lambda w}} \right] \\
&\quad = x - E_{P_a} W.
\end{aligned}$$

So, if $E_{P_a} W \leq x < \Delta_a$, the derivative is nonnegative for λ near zero and the function goes to $-\infty$ as $\lambda \to \infty$. Thus, since the function is concave by Claim 6, the sup is achieved for λ in $[0, \infty)$. Conversely, $x < E_{P_a} W$, then the supremum is achieved by $\lambda = 0$ and has value 0.

Thus, we have the following lemma:

Lemma 5: The $\sup_{\lambda \geq 0}$ of $\lambda x - \Psi_a(\lambda)$ is always achieved for $x < \Delta_a$ and is the solution of

$$\frac{d}{d\lambda} [\lambda x - \log \Psi_a(\lambda)] = 0$$

when $x > E_{P_a} W$. It is 0 otherwise.

We note that we will be using the variable δ in what follows. This δ is *not* the same as the parameter of the BSC mentioned in the body of this paper. We reuse the variable here to simplify the notation. There should be no ambiguity since this appendix is self-contained.

Lemma 5 implies the existence of

$$\lambda_a(\delta) = \arg\max_{\lambda} (\lambda(\Delta_a - \delta) - \Psi_a(\lambda))$$

the achiever of $\Psi_a^*(\Delta_a - \delta)$, for $0 < \delta < \Delta_a$.

Define $\lambda_0(\delta) = \max_a \lambda_a(\delta)$. Now, we have the following claim.

Claim 7: $\lambda x - \sum_{a=1}^{r} Q_a^n \log \Psi_a(\lambda)$ is concave in λ.

Proof: This follows easily since $\lambda x - \log \Psi_a(\lambda)$ is concave. □

Now, $\lambda x - \sum_{a=1}^{r} Q_a^n \log \Psi_a(\lambda)$ is 0 at $\lambda = 0$ and, if $x < \Delta_{(n)}$, we get the expression at the bottom of the page. Also

$$\begin{aligned}
&\lim_{\lambda \to 0} \frac{d}{d\lambda} \left[\lambda x - \sum_{a=1}^{r} Q_a^n \log \Psi_a(\lambda) \right] \\
&\quad = \lim_{\lambda \to 0} \left[x - \sum_{a=1}^{r} Q_a^n \frac{\sum_{w \in S(P_a)} w P_a(w) e^{\lambda w}}{\sum_{w \in S(P_a)} P_a(w) e^{\lambda w}} \right] \\
&\quad = x - \sum_{a=1}^{r} Q_a^n E_{P_a} W \\
&\quad = x - \bar{w}_n.
\end{aligned}$$

$$\begin{aligned}
\lim_{\lambda \to \infty} \left[\lambda x - \sum_{a=1}^{r} Q_a^n \log \Psi_a(\lambda) \right] &= \lim_{\lambda \to \infty} \log \frac{e^{\lambda x}}{\prod_{a=1}^{r} \Psi_a(\lambda)^{Q_a^n}} \\
&= \lim_{\lambda \to \infty} \log \left(\frac{e^{\lambda(x - \Delta_{(n)})}}{\prod_{a=1}^{r} (P_a(\Delta_a) + \sum_{w \in S(P_a), w \neq \Delta_a} P_a(w) e^{\lambda(w - \Delta_a)})^{Q_a^n}} \right) \\
&= -\infty.
\end{aligned}$$

So, if $\bar{w}_n \le x < \Delta_{(n)}$, the derivative is nonnegative near zero and the sup is achieved for λ in $[0, \infty)$. Conversely, if $x < \bar{w}_n$, the supremum is zero and is achieved by $\lambda = 0$.

Thus, we have the following lemma.

Lemma 6: The achiever $\lambda_{(n)}(\delta) = \arg\max_\lambda (\lambda x - \Psi_{(n)}(\lambda))$ always exists for $x < \Delta_{(n)}$ and is the solution of

$$\frac{d}{d\lambda}\left[\lambda x - \sum_{a=1}^{r} Q_a^n \log \Psi_a(\lambda)\right] = 0 \tag{34}$$

when $x > \bar{w}_n$. It is 0 otherwise.

We thus have the existence of $\lambda_{(n)}(\delta)$, the achiever of $\Psi^*_{(n)}(\Delta_{(n)} - \delta)$, for $0 < \delta < \Delta_{(n)}$. We are now ready to prove the upper bound of Lemma 1.

Proof: We first assume that $\bar{w}_n \le D < \Delta_{(n)}$. Then, using $\mathbb{I}$ to denote the indicator function, for all $\lambda \ge 0$

$$\begin{aligned}
P(S_n > D) &= E(\mathbb{I}(S_n - D \ge 0))\\
&\le E\left(e^{n\lambda(S_n - D)}\right)\\
&= e^{-n\lambda D} \prod_{a=1}^{r} \prod_{i=1}^{nQ_a^n} E_{P_a}(e^{\lambda W})\\
&= e^{-n\lambda D} \prod_{a=1}^{r} E_{P_a}(e^{\lambda W})^{nQ_a^n}\\
&= e^{-n[\lambda D - \sum_{a=1}^{r} Q_a^n \log E_{P_a}(e^{\lambda W})]}\\
&= e^{-n[\lambda D - \sum_{a=1}^{r} Q_a^n \Psi_a(\lambda)]}.
\end{aligned}$$

Thus

$$\begin{aligned}
P(S_n > D) &\le e^{\left[-n \sup_{\lambda \ge 0}(\lambda D - \sum_{a=1}^{r} Q_a^n \Psi_a(\lambda))\right]}\\
&= e^{-n\Psi^*_{(n)}(D)}.
\end{aligned}$$

If $D \ge \Delta_{(n)}$

$$P(S_n > D) \le \sum_{a=1}^{r} P\left(\frac{1}{nQ_a^n}\sum_{i=1}^{nQ_a^n} W_i^a \ge \Delta_a\right) = 0$$

where we use the notation W_i^p, W_j^q, $i \ne j$ to denote independent random variables distributed according to P_p, P_q, respectively. Thus, $P(S_n > D) = 0$ for $D \ge \Delta_{(n)}$.

Since $\Psi_{(n)}(D) = \infty$ for $D \ge \Delta_{(n)}$, the upper bound holds.

Finally, for $0 < D < \bar{w}_n$, $\Psi_{(n)}(D) = 0$ and since $P(S_n > D) < 1$, the upper bound holds. This concludes the proof of the upper bound. □

We prove the lower bound of Lemma 1 by first showing a restricted version of the lemma.

Lemma 7: Fix $\delta > 0$ such that $\delta < \frac{\Delta_{(n)} - \bar{w}_n}{2}$. Choose D so that $\bar{w}_n \le D \le \Delta_{(n)} - 2\delta$. Choose an ϵ such that $0 < \epsilon < \delta$. For $n > \frac{r^2 \Delta_{\max}^2 \log 8r}{\delta^2}$

$$\begin{aligned}
&P(S_n > D)\\
&\ge \exp\left[-n\left(\Psi^*_{(n)}(D) + \frac{\lambda_0(\delta)\Delta_{\max} r\sqrt{2\log(8r)}}{\sqrt{n}} - \frac{\log\left(\frac{3}{4}\right)}{n}\right)\right].
\end{aligned}$$

Proof: Define the events

$$E = \left\{\sum_{t=1}^{n} W_t \ge nD\right\}$$

and

$$F = \left\{\sum_{t=1}^{n} W_t \le n(D + 2\epsilon)\right\}.$$

Define a new set of probability measures, $\{P^n_{a,\lambda}\}_{a=1}^r$, by $P^n_{a,\lambda}(w) = P_a(w)\frac{e^{\lambda w}}{\Psi_a(\lambda)}$.

Since $\bar{w}_n < D + \epsilon < D + \delta \le \Delta_{(n)} - \delta$, by Lemma 6, the achiever of $\Psi^*_{(n)}(D + \epsilon)$ exists. We denote it by λ^*_n. By Lemma 6, λ^*_n satisfies

$$\frac{d}{d\lambda}\left[\lambda(D + \epsilon) - \Psi_{(n)}(\lambda)\right] = 0 \tag{35}$$

so that

$$\begin{aligned}
D + \epsilon &= \frac{d}{d\lambda}\Psi_{(n)}(\lambda^*_n)\\
&= \sum_{a=1}^{r} Q_a^{(n)} \frac{\sum_{w \in S(P_a)} w P_a(w) e^{\lambda^*_n w}}{\sum_{w \in S(P_a)} P_a(w) e^{\lambda^*_n w}}\\
&= \sum_{a=1}^{r} Q_a^{(n)} \frac{\sum_{w \in S(P_a)} w P_a(w) e^{\lambda^*_n w}}{\Psi_a(\lambda^*_n)}\\
&= \sum_{a=1}^{r} Q_a^{(n)} E_{P_{a,\lambda^*_n}}(W).
\end{aligned}$$

Now

$$\begin{aligned}
&P\left(\frac{1}{n}\sum_{t=1}^{n} W_t > D\right) &(36)\\
&\quad\ge P(E \cap F)\\
&\quad= \sum_{w^n \in E \cap F} \prod_{t=1}^{n} P(w_t)\\
&\quad= \sum_{w^n \in E \cap F} \prod_{a=1}^{r} \prod_{t=1}^{nQ_a^n} P_a\left(w_{n\sum_{i=1}^{a} Q_i^n + t}\right). &(37)
\end{aligned}$$

Letting $h(a, t, n) = n\sum_{i=1}^{a} Q_i^n + t$, and substituting in the modified probability measure, (37) becomes

$$\begin{aligned}
&\sum_{w^n \in E \cap F} \prod_{a=1}^{r} \prod_{t=1}^{nQ_a^n} \Psi_a(\lambda^*_n) \exp\left[-\lambda^*_n w_{h(a,t,n)}\right] P_{a,\lambda^*_n}(w_{h(a,t,n)})\\
&\quad= \sum_{w^n \in E \cap F} \prod_{a=1}^{r} \exp\left(nQ_a^n \log(\Psi_a(\lambda^*_n)) - \lambda^*_n \sum_{t=1}^{n} Q_a^n w_{h(a,t,n)}\right)\\
&\qquad\times \prod_{a=1}^{r} \prod_{t=1}^{nQ_a^n} P_{a,\lambda^*_n}(w_{h(a,t,n)})\\
&\quad\ge \exp\left[n\left(\sum_{a=1}^{r} Q_a^n \log \Psi_a(\lambda^*_n) - \lambda^*_n(D + 2\epsilon)\right)\right] &(38)
\end{aligned}$$

$$\times \sum_{w^n \in E\cap F} \prod_{a=1}^{r} \prod_{t=1}^{nQ_a^n} P_{a,\lambda_n^*}\left(w_{h(a,t,n)}\right) \tag{39}$$

$$= \exp[-n(\lambda_n^*(D+2\epsilon) - \sum_{a=1}^{r} Q_a^n \log \Psi_a(\lambda_n^*))] \tag{40}$$

$$\times \sum_{w^n \in E\cap F} \prod_{a=1}^{r} \prod_{t=1}^{nQ_a^n} P_{a,\lambda_n^*}(w_{h(a,t,n)}) \tag{41}$$

where (38) follows from the fact that $w^n \in F$.

Now

$$\exp\left[-n\left(\lambda_n^*(D+2\epsilon) - \sum_{a=1}^{r} Q_a^n \log \Psi_a(\lambda_n^*)\right)\right] \tag{42}$$

$$\geq \exp\left[-n\left(2\lambda_n^*\epsilon + \Psi_{(n)}^*(D)\right)\right] \tag{43}$$

since $\lambda_n^* D - \Psi_{(n)}(\lambda_n^*) \leq \Psi_{(n)}^*(D)$. To proceed, we need the following.

Lemma 8: For all $x, y > 0$, if λ_1 achieves $\Psi_{(n)}^*(x)$ and λ_2 achieves $\Psi_{(n)}^*(x+y)$, then $\lambda_1 \leq \lambda_2$.

Proof:

$$\begin{aligned} \Psi_{(n)}^*(x+y) &= \lambda_2(x+y) - \Psi_n(\lambda_2) \qquad (44)\\ &\geq \lambda_1(x+y) - \Psi_n(\lambda_1)\\ &= \lambda_1 y + \Psi_{(n)}^*(x)\\ &\geq \lambda_1 y + \lambda_2 x - \Psi_n(\lambda_2) \qquad (45)\end{aligned}$$

and (44) $\geq$ (45) implies $\lambda_2 \geq \lambda_1$.

Similarly, if λ_1 achieves $\Psi_a^*(x)$ and λ_2 achieves $\Psi_a^*(x+y)$, then $\lambda_1 \leq \lambda_2$. □

It is straightforward to see that $\lambda_a(\delta)$ is continuous, and that, as $\delta \to 0$, $\lambda_a(\delta) \to \infty$, for all a. For a fixed $\delta > 0$, $\lambda_a(\delta) < \infty$ and is independent of n. Thus, $\lambda_0(\delta) = \max_a \lambda_a(\delta)$ achieves all $\Psi_a^*(\Delta_a - \phi_a)$ for some collection $\{\phi_a\}_{a=1}^r$ such that, for each a, $\phi_a \leq \delta$. We thus have that

$$\begin{aligned} &\sum_{a=1}^{r} Q_a^n \sup_{\lambda\geq 0}[\lambda(\Delta_a - \phi_a) - \Psi_a(\lambda)]\\ &\quad= \sum_{a=1}^{r} Q_a^n\left[\lambda_0(\delta)(\Delta_a - \phi_a) - \Psi_a(\lambda_0(\delta))\right]\\ &\quad= \sup_{\lambda\geq 0}\left[\sum_{a=1}^{r} Q_a^n\left(\lambda(\Delta_a - \phi_a) - \Psi_a(\lambda)\right)\right] \qquad (46)\end{aligned}$$

where the last equality follows by the choice of the $\{\phi_a\}$, which implies that $\lambda_0(\delta)$ achieves (46).

Now, by (46)

$$\begin{aligned} &= \sup_{\lambda\geq 0}\left\{\lambda\left[\sum_{a=1}^{r} Q_a^n(\Delta_a - \phi_a)\right] - \Psi_{(n)}(\lambda)\right\}\\ &= \sup_{\lambda\geq 0}\left[\lambda\left(\Delta_{(n)} - \sum_{a=1}^{r} Q_a^n\phi_a\right) - \Psi_{(n)}(\lambda)\right]\end{aligned}$$

and $\sum_{a=1}^{r} Q_a^n \phi_a \leq \delta$. Thus

$$\Delta_{(n)} - \sum_{a=1}^{r} Q_a^n\phi_a \geq \Delta_{(n)} - \delta.$$

And so, by Lemma 8, we have

$$\lambda_{(n)}(\delta) \leq \lambda_0(\delta) = \max_a \lambda_a(\delta) < \infty.$$

Also, $\epsilon < \delta$ and $D < \Delta_{(n)} - 2\delta$, so

$$D + \epsilon < \Delta_{(n)} - 2\delta + \epsilon < \Delta_{(n)} - \delta.$$

Thus

$$\lambda_n^* \leq \lambda_{(n)}(\delta) \leq \lambda_0(\delta) < \infty.$$

We use this in (42) to get

$$\begin{aligned} &\exp\left[-n\left(2\lambda_n^*\epsilon + \Psi_{(n)}^*(D)\right)\right]\\ &\qquad\qquad \geq \exp\left[-n\left(2\lambda_0(\delta)\epsilon + \Psi_{(n)}^*(D)\right)\right], \qquad (47)\end{aligned}$$

which is a lower bound on the first part of (40). We now bound the second part of (40)

$$\sum_{w^n \in E\cap F} \sum_{a=1}^{r} \prod_{t=1}^{nQ_a^n} P_{a,\lambda_n^*}\left(w_{h(a,t,n)}\right) \tag{48}$$

$$\geq 1 - \sum_{w^n \in E^C} \prod_{a=1}^{r} \prod_{t=1}^{nQ_a^n} P_{a,\lambda_n^*}\left(w_{h(a,t,n)}\right) \tag{49}$$

$$- \sum_{w^n \in F^C} \prod_{a=1}^{r} \prod_{t=1}^{nQ_a^n} P_{a,\lambda_n^*}\left(w_{h(a,t,n)}\right). \tag{50}$$

Also, note that $n(D+\epsilon) = \sum_{a=1}^{r} Q_a^n E_{P_{a,\lambda_n^*}}(W)$ from (36). Thus, we have

$$\begin{aligned} &\sum_{w^n \in F^C} \prod_{a=1}^{r} \prod_{t=1}^{nQ_a^n} P_{a,\lambda_n^*}\left(w_{h(a,t,n)}\right)\\ &\quad= P\left(\sum_{a=1}^{r}\sum_{t=1}^{nQ_a^n} w_{h(a,t,n)}^a - n(D+\epsilon) > n\epsilon\right)\\ &\quad= P\left(\sum_{a=1}^{r}\sum_{t=1}^{nQ_a^n}\left(w_{h(a,t,n)}^a - E_{P_{a,\lambda_n^*}}W\right) > n\epsilon\right)\\ &\quad\leq P\left(\bigcup_{a=1}^{r}\left[\sum_{t=1}^{nQ_a^n}\left(w_{h(a,t,n)}^a - E_{P_{a,\lambda_n^*}}W\right) > \frac{n\epsilon}{r}\right]\right)\\ &\quad\leq \sum_{a=1}^{r} P\left(\sum_{t=1}^{nQ_a^n}\left(w_{h(a,t,n)}^a - E_{P_{a,\lambda_n^*}}W\right) > \frac{(nQ_a^n)\epsilon}{Q_a^n r}\right) \qquad (51)\\ &\quad\leq \sum_{a=1}^{r}\exp\left[-\frac{2nQ_a^n\epsilon^2}{\left(Q_a^n\right)^2 r^2\Delta_a^2}\right] \qquad (52)\\ &\quad= \sum_{a=1}^{r}\exp\left[-\frac{2n\epsilon^2}{Q_a^n r^2\Delta_a^2}\right]\\ &\quad\leq r\exp\left[-\frac{2n\epsilon^2}{r^2\Delta_{\max}^2}\right], \qquad (53)\end{aligned}$$

where (51) follows from the union bound, (52) follows from the Hoeffding bound, and (53) follows from the facts that $Q_a^n \leq 1$ and $\Delta_a^2 \leq \Delta_{\max}^2$.

Similarly, we have

$$\begin{aligned}
&\sum_{w^n\in E^C}\prod_{a=1}^{r}\prod_{t=1}^{nQ_a^n} P_{a,\lambda_n^*}\left(w_{h(a,t,n)}\right)\\
&= P\left(\sum_{a=1}^{r}\sum_{t=1}^{nQ_a^n} w^a_{h(a,t,n)} - n(D+\epsilon) < -n\epsilon\right)\\
&= P\left(\sum_{a=1}^{r}\sum_{t=1}^{nQ_a^n} \left(-w^a_{h(a,t,n)} - \left(-E_{P_{a,\lambda_n^*}}W\right)\right) > n\epsilon\right)\\
&\leq P\!\left(\bigcup_{a=1}^{r}\left[\sum_{t=1}^{nQ_a^n}\left(-w^a_{h(a,t,n)} - \left(-E_{P_{a,\lambda_n^*}}W\right)\right) > \frac{n\epsilon}{r}\right]\right)\\
&\leq \sum_{a=1}^{r} P\left(\sum_{t=1}^{nQ_a^n}\left(-w^a_{h(a,t,n)} - \left(-E_{P_{a,\lambda_n^*}}W\right)\right) > \frac{(nQ_a^n)\,\epsilon}{Q_a^n r}\right)\\
&\leq \sum_{a=1}^{r}\exp\left[-\frac{2nQ_a^n\epsilon^2}{(Q_a^n)^2 r^2\Delta_a^2}\right]\\
&= \sum_{a=1}^{r}\exp\left[-\frac{2n\epsilon^2}{Q_a^n r^2\Delta_a^2}\right]\\
&\leq r\exp\left[-\frac{2n\epsilon^2}{r^2\Delta_{\max}^2}\right].
\end{aligned}$$

Thus

$$\begin{aligned}
\sum_{w^n\in E\cap F}\prod_{a=1}^{r}\prod_{t=1}^{nQ_a^n} P_{a,\lambda_n^*}\left(w_{h(a,t,n)}\right)
&\geq 1 - 2r\exp\left[-\frac{2n\epsilon^2}{r^2\Delta_{\max}^2}\right]\\
&= \frac{3}{4} \qquad (54)
\end{aligned}$$

when $\epsilon = \frac{r\Delta_{\max}}{\sqrt{n}}\sqrt{\frac{\log(8r)}{2}}$. Since we assumed $\epsilon < \delta$, the proof is valid when

$$\delta > \frac{r\Delta_{\max}}{\sqrt{n}}\sqrt{\frac{\log(8r)}{2}}.$$

So we require

$$n > \frac{r^2\Delta_{\max}^2\log(8r)}{2\delta^2}. \qquad (55)$$

Thus, combining (47) and (54) yields

$$\begin{aligned}
P(E) &\geq \exp\left[-n\left(2\lambda_0(\delta)\epsilon + \Psi^*_{(n)}(D)\right)\right]\frac{3}{4}\\
&\geq \exp\left[-n\left(\Psi^*_{(n)}(D) + 2\lambda_0(\delta)\frac{r\Delta_{\max}}{\sqrt{n}}\sqrt{\frac{\log(8r)}{2}} - \frac{\log\left(\frac{3}{4}\right)}{n}\right)\right]\\
&= \exp\left[-n\left(\Psi^*_{(n)}(D) + \frac{\lambda_0(\delta)\Delta_{\max}r\sqrt{2\log(8r)}}{\sqrt{n}} - \frac{\log\left(\frac{3}{4}\right)}{n}\right)\right]
\end{aligned}$$

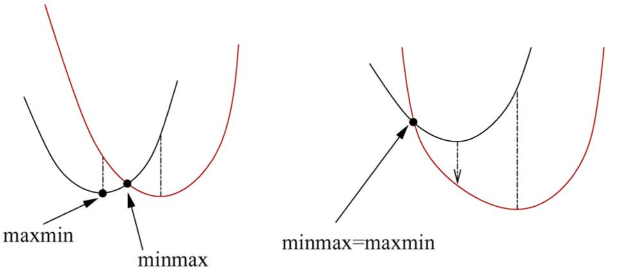

Fig. 10. Comparing two convex functions. The solid dots represent max-min or min-max points. The shaded dots represent the minimum of a function. The vertical lines connect the minimum of a particular function to the corresponding point on the other function. We see that when the max-min is strictly less than the min-max, the difference between each minimum point and the corresponding point on the other function is negative. When this does not hold, the max-min and min-max are equal. Of course, this is just an intuitive argument. Rigorous reasoning is given in the text.

concluding the proof of the restricted form of the lemma, with $N = \frac{r^2\Delta_{\max}^2\log(8r)}{2\delta^2}$ and $o(n) = \frac{\lambda_0(\delta)\Delta_{\max}r\sqrt{2\log(8r)}}{\sqrt{n}} - \frac{\log(\frac{3}{4})}{n}$. We thus have the lower bound of Lemma 1 for $\bar{w}_n \leq D < \Delta_{(n)}$ since δ was arbitrary. □

We now prove the lower bound of Lemma 1.

Proof: Observe that for $D \geq \Delta_{(n)}$, $P(S_n > D) = 0$. Since $\Psi^*_{(n)}(D) = \infty$ for such D, the lemma is true for all n and with $o(n) = 0$.

For $D < \bar{w}_n$, $P(S_n > D) \geq P(S_n > \bar{w}_n) = e^{-o(n)}$, for some $N, o(n)$ found in the manner described in the preceding proof. For such D, $\Psi_{(n)}(D) = 0$, so the lemma is true. This concludes the proof of the lower bound. □

Appendix II
Details of The Method to Compute The Region of Symbol-by-Symbol Suboptimality

We have shown that the only candidates for the optimal denoiser and optimal symbol-by-symbol denoiser are the SWYS and ONES denoisers. So, in order to determine whether symbol-by-symbol performance is optimal, we must determine when minimizing the rate function over all types Q_0 and then maximizing over the choice of denoiser, SWYS or ONES, is equivalent to maximizing the rate function over the choice of denoiser, SYWS or ONES, and then minimizing over the type Q_0. We have also demonstrated that the rate function for a particular denoiser is a convex function of Q_0.

So the question, illustrated in Fig. 10, becomes: when is the max (over the two schemes SWYS and ONES) of the min (over Q_0) of the two convex functions of Q_0 strictly less than the min (over Q_0) of the max (over SWYS and ONES)? We need only compute the value of the exponents at two points each in order to determine when the min-max equals the max-min since the functions are convex in the minimizing variable. We will soon show that the only way the min-max is not equal to the max-min is if the value of each function at its minimizing Q_0 is less than the value of the other function at that same Q_0 (see Fig. 10).

Although the region of such (δ, D) pairs is computed numerically, each point in the region can be verified analytically to show the suboptimality of symbol-by-symbol denoising.

Specifically, the program is used to compute the Q_0 that minimizes the exponent for each denoiser. This is easy to determine given the expression (32). We find the maximizing λ for (33) and then set Q_0 so the divergence term in (32) is zero. At this point, we need the following simple fact.

Claim 8: The minimizing Q_0 of the exponent for a particular denoiser is unique.

Proof: This follows from the fact that the divergence of two discrete distributions is zero if and only if the distributions are everywhere equal. □

The minimum value of the exponent for a given denoiser is compared to the value of the exponent using the same Q_0 in the other denoiser. If the minimum exponent for each denoiser is strictly less than the exponent using that Q_0 in the other denoiser, then symbol-by-symbol is strictly suboptimal by the uniqueness of the minimizers of the exponents of the denoisers, the continuity of the exponent, and the mean value theorem. These imply that the two exponent functions must cross for some value of Q_0 that lies strictly in between the minimizers of the two exponent functions. The value of the functions at this Q_0 is the min-max and is strictly greater than the max-min since the minimizers of the exponents of the denoisers are unique.

The value of the supremum over λ is computed by taking a derivative and setting it equal to zero. We solve the resulting equation numerically by using the roots function in Matlab. Since the functions are continuous in λ, the roots function will give an accurate value for λ. Each point in the region can be verified by computing the values of λ analytically and substituting into the corresponding rate function expressions.

1) Sample Calculation of Symbol by Symbol Suboptimality: We now provide an example of a particular channel and threshold where symbol-by-symbol denoising is suboptimal. In fact, we compute the max-min and min-max values and show the strict inequality.

Use the above problem setup and set $\delta = 0.1$ and $D = 0.98$. Fixing the denoiser to be SWYS, (33) becomes

$$\sup_{\lambda \geq 0} \lambda D - \log\left(\bar{\delta} + \frac{1}{2}\delta e^{\lambda} + \frac{1}{2}\delta e^{2\lambda}\right). \tag{56}$$

Differentiating with respect to λ and setting the result equal to zero, we get

$$\delta\left(\frac{D}{2} - 1\right)e^{2\lambda} + \delta\left(\frac{D-1}{2}\right)e^{\lambda} + D\bar{\delta} = 0. \tag{57}$$

Now, substituting the values for δ and D and scaling, we get

$$-51e^{2\lambda} - e^{\lambda} + 882 = 0. \tag{58}$$

Thus, the expression is maximized by the nonnegative root of this equation, so $e^{\lambda} \cong 4.1488$ and $\lambda \cong 1.4228$. This yields a value of 0.7173 for the objective. Now, as described above, (32) is minimized by setting

$$Q_0 = \frac{\sum_x p(x,0)e^{\lambda\Lambda(X,\hat{x}(0))}}{E_{X,Z}e^{\lambda\Lambda(X,\hat{x}(Z))}} \cong 0.6659.$$

We now compute the value of the objective for this value of Q_0 in the ONES denoiser. We use (26) to get

$$\sup_{\lambda \geq 0} \lambda D - Q_0 \log Q_0 - Q_0 \log\left(\bar{p}\bar{\delta}e^{\lambda} + p\delta\right) + Q_1 \log Q_1 - Q_1 \log\left(\bar{p}\delta e^{\lambda} + p\bar{\delta}\right).$$

We take a derivative, set it to zero, set $p = 0.5$, and simplify to get

$$\left(\delta\bar{\delta}\right)(D-1)e^{2\lambda} + \left(D\delta^2 + D\bar{\delta}^2 - Q_0\bar{\delta}^2 - Q_1\delta^2\right)e^{\lambda} + D\delta\bar{\delta} = 0.$$

We substitute the values for δ, D, and Q_0 to get

$$-18e^{2\lambda} + 2609e^{\lambda} + 882 = 0. \tag{59}$$

So that $e^{\lambda} \cong 145.28$, $\lambda = 4.9787$, and the objective is 0.7754. This is greater than the value of the objective for the SWYS denoiser.

Now, we find the optimal value of (33) for the ONES denoiser. We have

$$\sup_{\lambda \geq 0} \lambda D - \log\left(p + \bar{p}e^{\lambda}\right). \tag{60}$$

Differentiating with respect to λ and substituting $p = 0.5$ leads to

$$e^{\lambda} = \frac{D}{1-D}. \tag{61}$$

That is, $e^{\lambda} = 49$ and $\lambda \cong 3.8918$. Then, the objective is 0.5951. Furthermore, $Q_0 = 0.884$. We now compute the value of (32) for the SWYS denoiser and this value of Q_0. (32) becomes

$$\sup_{\lambda \geq 0} \lambda D + Q_0 \log Q_0 - Q_0 \log\left(\bar{p}\bar{\delta} + p\delta e^{2\lambda}\right) - Q_1 \log\left(p\bar{\delta} + p\delta e^{\lambda}\right) + Q_1 \log Q_1.$$

Differentiating with respect to λ and substituting $p = 0.5$ leads to

$$\left(D\delta^2 - Q_0\delta^2 - \delta^2\right)e^{3\lambda} + \left(D\delta\bar{\delta} - 2Q_0\delta\bar{\delta}\right)e^{2\lambda} + \left(D\delta\bar{\delta} - Q_1\delta\bar{\delta}\right)e^{\lambda} + D\bar{\delta}^2 = 0.$$

Substituting the δ, D, and Q_0 values and simplifying yields

$$-90.4e^{3\lambda} - 709.2e^{2\lambda} + 777.6e^{\lambda} + 7938 = 0. \tag{62}$$

Thus, the objective is maximized by $e^{\lambda} = 3.2306$, i.e., $\lambda = 1.1727$. The value of the objective is 0.8727, which is greater than that for the ONES denoiser.

By the results of this appendix, since the minimum value of the rate function of each denoiser is less than that of the other denoiser for the minimizing Q_0, the rate functions must cross and thus max-min < min-max. So, we have a concrete example of the suboptimality of symbol-by-symbol denoising schemes.

Acknowledgment

The authors would like to thank Prof. Robert Gray, Dr. James Mammen, and Prof. Young-Han Kim for helpful discussions, and Prof. Amos Lapidoth for pointing out the qualitative similarity between our results and Stein's paradox.

References

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

[2] E. Ordentlich and T. Weissman, "On the optimality of symbol by symbol filtering and denoising," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 19–40, Jan. 2006.

[3] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 197–199, Mar. 1974.

[4] T. Weissman and N. Merhav, "Tradeoffs between the excess-code-length exponent and the excess-distortion exponent in lossy source coding," *IEEE Trans. Inf. Theory*, vol. 48, no. 2, pp. 396–415, Feb. 2002.

[5] N. Merhav and I. Kontoyiannis, "Source coding exponents for zero-delay coding with finite memory," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 609–624, Mar. 2003.

[6] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code or not to code: Lossy source-channel communication revisited," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

[7] T. Weissman, "Universally attainable error-exponents for rate-distortion coding of noisy sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1229–1246, Jun. 2004.

[8] A. Puhalskii and V. Spokoiny, "On large deviation efficiency in statistical inference," *Bernoulli*, vol. 4, no. 2, pp. 203–272, 1998.

[9] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*. Berkeley, CA: Univ. California Press, 1956, vol. 1, pp. 197–206.

[10] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer-Verlag, 1998.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[12] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[13] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. New York: Springer-Verlag, 1998.