On the Optimality of Symbol-by-Symbol Filtering and Denoising

Erik Ordentlich, *Member, IEEE,* and Tsachy Weissman, *Member, IEEE*

***Abstract*—We consider the problem of optimally recovering a finite-alphabet discrete-time stochastic process $\{X_t\}$ from its noise-corrupted observation process $\{Z_t\}$. In general, the optimal estimate of X_t will depend on all the components of $\{Z_t\}$ on which it can be based. We characterize nontrivial situations (i.e., beyond the case where (X_t, Z_t) are independent) for which optimum performance is attained using "symbol-by-symbol" operations (a.k.a. "singlet decoding"), meaning that the optimum estimate of X_t depends solely on Z_t. For the case where $\{X_t\}$ is a stationary binary Markov process corrupted by a memoryless channel, we characterize the necessary and sufficient condition for optimality of symbol-by-symbol operations, both for the filtering problem (where the estimate of X_t is allowed to depend only on $\{Z_{t'}\}_{t' \leq t}$) and the denoising problem (where the estimate of X_t is allowed dependence on the entire noisy process). It is then illustrated how our approach, which consists of characterizing the support of the conditional distribution of the noise-free symbol given the observations, can be used for characterizing the entropy rate of the binary Markov process corrupted by the binary-symmetric channel (BSC) in various asymptotic regimes. For general noise-free processes (not necessarily Markov), general noise processes (not necessarily memoryless), and general index sets (random fields) we obtain an easily verifiable sufficient condition for the optimality of symbol-by-symbol operations and illustrate its use in a few special cases. For example, for binary processes corrupted by a BSC, we establish, under mild conditions, the existence of a $\delta^* > 0$ such that the "say-what-you-see" scheme is optimal provided the channel crossover probability is less than δ^*. Finally, we show how for the case of a memoryless channel the large deviations (LD) performance of a symbol-by-symbol filter is easy to obtain, thus characterizing the LD behavior of the optimal schemes when these are singlet decoders (and constituting the only known cases where such explicit characterization is available).**

***Index Terms*—Asymptotic entropy, denoising, discrete memoryless channels (DMCs), entropy rate, estimation, filtering, hidden Markov processes (HMP), large deviations performance, noisy channels, singlet decoding, smoothing, symbol-by-symbol schemes.**

I. Introduction

Let $\{X_t\}_{t \in \mathbb{Z}}$ be a discrete-time stochastic process and $\{Z_t\}_{t \in \mathbb{Z}}$ be its noisy observation signal. The *denoising* problem is that of estimating $\{X_t\}$ from its noisy observations $\{Z_t\}$. Since perfect recovery is seldom possible, a loss function measuring the goodness of the reconstruction is given, and the goal is to estimate each X_t so as to minimize the expected loss. The *filtering* problem is the denoising problem restricted to causality, namely, when the estimate of X_t is allowed to depend on the noisy observation signal only through $\{Z_{t'}\}_{t' \leq t}$.

When $\{X_t\}$ is a memoryless signal corrupted by a memoryless channel, the optimal denoiser (and, *a fortiori*, the optimal filter) has the property that, for each t, the estimate of X_t depends on the noisy observation signal only through Z_t. A scheme with this property will be referred to as a *symbol-by-symbol* scheme or as a *singlet decoder* [12]. When $\{X_t\}$ is not memoryless, on the other hand, the optimal estimate of each X_t will, in general, depend on all the observations available to it in a nontrivial way. This is the case even when the noise-free signal is of limited memory (e.g., a first-order Markov process) and the noise is memoryless. Accordingly, much of nonlinear filtering theory is devoted to the study of optimal estimation schemes for these problems (cf., e.g., [3], [5], [24], [26], [14], and the many references therein), and basic questions such as the closed-form characterization of optimum performance (beyond the cases we characterize in this work where singlet decoding is optimum) remain open.

One pleasing feature of a singlet decoder is that its performance is amenable to analysis since its expected loss in estimating X_t depends only on the joint distribution of the pair (X_t, Z_t) (rather than in a complicated way on the distribution of the process pair $(\{X_t\}, \{Z_t\})$). Another of the obvious merits of a singlet decoder is the simplicity with which it can be implemented, which requires no memory and no delay. It is thus of practical value to be able to identify situations where no such memory and delay are required to perform optimally. Furthermore, it will be seen that in many cases of interest, where singlet decoding is optimal, it is the same scheme which is optimal across a wide range of sources and noise distributions. For example, for a binary source corrupted by a binary-symmetric channel (BSC) we shall establish, under mild conditions, the existence of a $\delta^* > 0$, such that the "say-what-you-see" scheme is optimal, provided the channel crossover probability is less than δ^*. This implies, in particular, the universality of this simple scheme with respect to the family of sources sharing this property, as well as with respect to all noise levels $\leq \delta^*$. Thus, the identification of situations where singlet decoding attains optimum performance is of interest from both the theoretical and the practical viewpoints, and is the motivation for our work.

A singlet decoder will be optimal if the value of the optimal estimate conditioned on all available observations coincides with the value of the optimal estimate conditioned on the

Manuscript received December 31, 2003; revised September 13, 2005. The work of T. Weissman was supported in part by the National Science Foundation under Grants DMS-0072331 and CCR-0312839. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004. Part of this work was performed while T. Weissman was visiting Hewlett-Packard Laboratories, Palo Alto, CA.

E. Ordentlich is with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA (e-mail: eord@hpl.hp.com).

T. Weissman is with the Electrical Engineering Department, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.edu).

Communicated by A. B. Nobel, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2005.860432

present noisy observation for almost all realizations of the noisy observations. This does not mean that the distribution of the clean symbol conditioned on all available observations coincides with its distribution conditioned on the present noisy observation (that would be the case if the underlying source was memoryless), but only that the corresponding optimal estimates do. This translates into a condition on the support of the distribution of the unobserved clean symbol given the observations (a measure-valued random variable measurable with respect to the observations). Indeed, for the Markov process corrupted by a memoryless channel this will lead to a necessary and sufficient condition for the optimality of singlet decoding in terms of the support of that distribution. In general, however, the support of this distribution (and, *a fortiori*, the distribution itself) is not explicitly characterizable, and, in turn, neither is the condition for optimality of singlet decoding. The support, however, can be bounded, leading to explicit sufficient conditions for this optimality. This will be our approach to obtaining sufficient conditions for the optimality of singlet decoding, which will be seen to lead to a complete characterization for the case of the corrupted binary Markov chain (where the upper and lower endpoints of the support can be obtained in closed form).

Characterization of cases where singlet decoding is optimal both for the filtering and the denoising problems was considered in [12] (cf. also [13], [33]) for the binary Markov source corrupted by a BSC. The closely related problem of singlet decoding optimality under the "word"-error probability criterion has been treated in [32] for a BSC-corrupted binary-symmetric Markov source and in [1] for the more general case of an asymmetric binary Markov chain corrupted by a binary Markov channel.

Our interest in the problem was triggered by the recently discovered discrete universal denoiser (DUDE) [39]. Experimentation has shown cases where the scheme applied to binary sources corrupted by a BSC of sufficiently small crossover probability remained idle (i.e., gave the noisy observation signal as its reconstruction). A similar phenomenon was observed with the extension of this denoiser to the finite-input continuous-output channel [10] where, for example, in denoising a binary Markov chain with a strong enough bias toward the 0 state, corrupted by additive white Laplacian noise, the reconstruction was the "all-zeros" sequence. As we shall see in this work, these phenomena are accounted for by the fact that the optimum distribution-dependent scheme in these cases is a singlet decoder (which the universal schemes identify and imitate).

An outline of the remainder of this work is as follows. In Section II, we introduce some notation and conventions that will be assumed throughout. Section III is dedicated to the case of a Markov chain corrupted by a memoryless channel. To fix notation and for completeness, we start in Section III-A by reviewing classical results concerning the evolution of conditional distributions of the clean symbol given past and \or future observations. We then apply these results in Section III-B to obtain necessary and sufficient conditions for the optimality of singlet decoding in both the filtering and the denoising problems. These conditions are not completely explicit in that they involve the support of a measure satisfying an integral equation whose closed-form solution is unknown.

In Section IV (subsections A and B), we show that, when the noise-free process is a binary Markov chain and corrupted by a memoryless channel, enough information about the support of that measure can be extracted for characterizing the optimality conditions for singlet decoding in closed form. Furthermore, the conditions both for the filtering and for the denoising problem are seen to depend on the statistics of the noise only through the support of the likelihood ratio between the channel output distributions associated with the two possible inputs. In Section IV-C, we further specialize the results to the BSC, characterizing all situations where singlet decoding is optimal (and thereby rederiving the results of [12] in a more explicit form). In Section IV-D, we point out a few immediate consequences of our analysis such as the fact that singlet decoding for the binary-input Laplace-output channel can only be optimal when the observations are useless, and that singlet decoding is never optimal for the binary-input Gaussian-output channel.

In Section V, we digress from the denoising problem and illustrate how the results of Section IV can be used for obtaining new bounds on the entropy rate of a hidden Markov process (HMP) (the closed form of which is still open, see [14], [22], and references therein). In particular, these bounds lead to a characterization of the behavior of the entropy rate of the BSC-corrupted binary Markov process in various asymptotic regimes (e.g., "rare-spikes," "rare-bursts," high "SNR," low "SNR," "almost memoryless"). The bounds also establish "graceful" dependence of the entropy rate on the parameters of the problem. Our results will imply continuity, differentiability, and in certain cases, higher order smoothness of the entropy rate in the process parameters. These results are new, even in view of existent results on analyticity of Lyapunov exponents in the entries of the random matrices [2], [31] and the connection between Lyapunov exponents and entropy rate [22], [23]. The reason is that, in the entropy rate, perturbations of the parameters affect both the matrices (corresponding to the associated Lyapunov exponent problem) *and* the distribution of the source generating them.

In Section VI, we derive a general and easily verifiable sufficient condition for the optimality of symbol-by-symbol schemes in both the filtering and the denoising problems. The condition (Theorem 8) is derived in a general setting, encompassing arbitrarily distributed processes (or fields) corrupted by arbitrarily distributed noise, and involves conditional probabilities associated with the channel, and the underlying noiseless process, that are usually easy to obtain (or bound) for a given process and noise specification. The remainder of that section details the application of the general condition to a few concrete scenarios. In Section VI-A, we look at the memoryless symmetric channel (with the same input and output alphabet) under Hamming loss. Our finding is that, under mild conditions on the noise-free source, there exists a positive threshold such that the "say-what-you-see" scheme is optimal whenever the level of the noise is below the threshold. Section VI-B shows that this continues to be the case for channels with memory, such as the Gilbert–Elliot channel (where this time it is the noise level associated with the "bad" state that need be below that threshold).

In Section VII, we obtain the exponent associated with the large deviations (LD) performance of a singlet decoder, thus

characterizing the LD behavior of the optimal schemes when these are singlet decoders (and constituting the only cases where the LD performance of the optimal filter is known). Finally, in Section VIII, we summarize the paper and discuss a few directions for future research.

II. Notation, Conventions, and Preliminaries

In general, we will assume a source $X(T)=\{X_t\}_{t\in T}$, where T is a countable index set. The components X_t will be assumed to take values in the finite alphabet $\mathcal{X}$. The noisy observation process, jointly distributed with $X(T)$, and having components taking values in $\mathcal{Z}$, will be denoted by $Z(T)$. Formally, we define a *denoiser* to be a collection of measurable functions $\{\hat{X}_t\}_{t\in T}$, where $\hat{X}_t:\mathcal{Z}^T\to\mathcal{X}$ and $\hat{X}_t=\hat{X}_t(Z(T))$ is the denoiser's estimate of X_t.

We assume a given loss function (fidelity criterion) $\Lambda:\mathcal{X}^2\to[0,\infty)$, represented by the matrix $\mathbf{\Lambda}=\{\Lambda(i,j)\}_{i,j\in\mathcal{X}}$, where $\Lambda(i,j)$ denotes the loss incurred by estimating the symbol i with the symbol j. Thus, the expected loss of a denoiser in estimating X_t is $E\Lambda(X_t,\hat{X}_t(Z(T)))$. A denoiser will be said to be *optimal* if, for each t, it attains the minimum of $E\Lambda(X_t,\hat{X}_t(Z(T)))$ among all denoisers.

In the case where $T=\mathbb{Z}$, we shall use $\boldsymbol{X},\{X_t\}_{t\in\mathbb{Z}}$ or $X_{-\infty}^{\infty}$ interchangeably with $X(T)$. We shall also let $X^t=\{X_{t'}\}_{t'\le t}$. In this setting, we define a *filter* analogously as a denoiser, only now $\hat{X}_t$ is a function only of $Z_{-\infty}^t$, rather than of the whole noisy signal $\boldsymbol{Z}$. The notion of an optimal filter is also extended from that of an optimal denoiser in an obvious way.

If $\{R_i\}_{i\in I}$ is any collection of random variables we let $\mathcal{F}(\{R_i\}_{i\in I})$ denote the associated sigma algebra. For any finite set $\mathcal{S}$, $\mathcal{M}(\mathcal{S})$ will denote the simplex of all $|\mathcal{S}|$-dimensional probability column vectors. For $v\in\mathcal{M}(\mathcal{S})$, $v(s)$ will denote the component of v corresponding to the symbol s according to some ordering of the elements of $\mathcal{S}$.

For $q\in\mathcal{M}(\mathcal{X})$, let $U(q)$ denote the Bayes envelope (cf., e.g., [21], [34], [27]) associated with the loss function Λ, defined by

$$U(q)=\min_{\hat{x}\in\mathcal{X}}\sum_{a\in\mathcal{X}}\Lambda(a,\hat{x})q(a)=\min_{\hat{x}\in\mathcal{X}}\boldsymbol{\lambda}_{\hat{x}}^T q$$

where $\boldsymbol{\lambda}_{\hat{x}}$ denotes the column of the loss matrix associated with the reconstruction $\hat{x}$.

We will generically use P to denote probability, and conditional probability: $P(X_i=a|Z_{-\infty}^i)$, for example, should be understood as the (random) probability of $X_i=a$ under a version of the conditional distribution of X_i given $\mathcal{F}(Z_{-\infty}^i)$. For a fixed individual $z_{-\infty}^i$, $P(X_i=a\,|\,z_{-\infty}^i)$ will denote that version of the conditional distribution evaluated for $Z_{-\infty}^i=z_{-\infty}^i$. Throughout the paper, statements involving random variables should be understood, when not explicitly indicated, in the almost sure sense.

Since the optimal estimate of X_t is the reconstruction symbol minimizing the expected loss given the observations, it follows that for an optimal denoiser

$$E\Lambda(X_t,\hat{X}_t(Z(T)))=EU(P(X_t=\cdot\,|\,Z(T))) \tag{1}$$

with $P(X_t=\cdot\,|\,Z(T))$ denoting the $\mathcal{M}(\mathcal{X})$-valued random variable whose ath component is $P(X_t=a\,|\,Z(T))$. Similarly, an optimal filter satisfies

$$E\Lambda\left(X_t,\hat{X}_t\left(Z_{-\infty}^t\right)\right)=EU\left(P\left(X_t=\cdot\,\middle|\,Z_{-\infty}^t\right)\right). \tag{2}$$

To unify and simplify statements of results, the following conventions will also be assumed: $0/0\equiv 1, 1/0\equiv\infty, 1/\infty\equiv 0$, $\log\infty\equiv\infty, \log 0\equiv-\infty, e^{\infty}\equiv\infty, e^{-\infty}\equiv 0, \infty+c\equiv\infty$. More generally, for a function $f:\mathbb{R}\to\mathbb{R}$, $f(\infty)$ will stand for $\lim_{x\to\infty}f(x)$ where the limit is assumed to exist (in the extended real line), and $f(-\infty)$ is defined similarly. For concreteness, logarithms are assumed throughout to be taken in the natural base.

For positive-valued functions f and g, $f(\varepsilon)\sim g(\varepsilon)$ will stand for $\lim_{\varepsilon\downarrow 0}\frac{f(\varepsilon)}{g(\varepsilon)}=1$, and $f(\varepsilon)\lesssim g(\varepsilon)$ will stand for $\limsup_{\varepsilon\downarrow 0}\frac{f(\varepsilon)}{g(\varepsilon)}\le 1$. Additionally, $f(\varepsilon)=O(g(\varepsilon))$ will stand for $\limsup_{\varepsilon\downarrow 0}\frac{f(\varepsilon)}{g(\varepsilon)}<\infty$, $f(\varepsilon)=\Omega(g(\varepsilon))$ will stand for $\liminf_{\varepsilon\downarrow 0}\frac{f(\varepsilon)}{g(\varepsilon)}>0$, and $f(\varepsilon)\asymp g(\varepsilon)$ will stand for the statement that both $f(\varepsilon)=O(g(\varepsilon))$ and $f(\varepsilon)=\Omega(g(\varepsilon))$ hold.

Finally, when dealing with the M-ary alphabet $\{0,1,\ldots,M-1\}$, $\oplus$ will denote modulo M addition.

III. Finite-Alphabet Markov Chain Corrupted by a Memoryless Channel

In this section, we assume $\{X_i\}_{i\in\mathbb{Z}}$ to be a stationary ergodic first-order Markov process with the finite alphabet $\mathcal{X}$, and $\{Z_i\}_{i\in\mathbb{Z}}$ to be its noisy observation process when corrupted by a memoryless channel. The process $\{Z_i\}_{i\in\mathbb{Z}}$ is known as a *hidden Markov process* (HMP). We therefore adopt notation which is standard in the literature on HMPs, such as that employed in [14]. Specifically, let $K:\mathcal{X}^2\to[0,1]$ be the transition kernel associated with $\{X_i\}_{i\in\mathbb{Z}}$

$$K(a,b)=P(X_{i+1}=b\,|\,X_i=a) \tag{3}$$

K_r be the transition kernel of the time reversed process

$$K_r(a,b)=P(X_i=b|X_{i+1}=a)$$

and let μ denote its marginal distribution

$$\mu(a)=P(X_i=a)$$

which is the unique probability measure satisfying

$$\mu(b)=\sum_{a\in\mathcal{X}}\mu(a)K(a,b),\qquad\forall b\in\mathcal{X}.$$

We assume, without loss of generality, that

$$\mu(a)=\Pr(X_i=a)>0,\qquad\forall a\in\mathcal{X}. \tag{4}$$

Throughout this section, we assume that $\{Z_i\}$ is the noisy observation process of $\{X_i\}$ when corrupted by the memoryless channel C. For simplicity, we shall confine attention to one of two cases.

1. Discrete channel output alphabet, in which case $C(a,b)$ denotes the probability of a channel output symbol b when the channel input is a.
2. Continuous real-valued channel output alphabet, in which case $C(a,\cdot)$ will denote the density with respect to Lebesgue measure (assumed to exist) of the channel output distribution when the input is a.

The more general case of arbitrary channel output distributions can be handled by considering densities with respect to other dominating measures and the subsequent derivations remain valid up to obvious modifications. For concreteness in the following derivations, the notation should be understood in the sense of the first case whenever there is ambiguity. All the derivations, however, are readily verified to remain valid for the continuous-output channel with the obvious interpretations (e.g., of $C(a,\cdot)$ as a density rather than a probability mass function (PMF), and $P(Z_i = z \,|\, Z_{-\infty}^{i})$ as a conditional density rather than a conditional probability).

A. Evolution of the Conditional Distributions

Let $\{\beta_i\}, \{\gamma_i\}$, denote the processes with $\mathcal{M}(\mathcal{X})$-valued components defined, respectively, by

$$\beta_i(a) = P\left(X_i = a \,\middle|\, Z_{-\infty}^{i}\right) \tag{5}$$

and

$$\gamma_i(a) = P\left(X_i = a \,|\, Z_i^{\infty}\right). \tag{6}$$

For $a \in \mathcal{X}$, standard use of Bayes rule and the defining properties of HMPs leads to the "forward recursion" [14]

$$\beta_i(a) = \frac{C(a, Z_i)[K^T\beta_{i-1}](a)}{\sum_{a' \in \mathcal{X}} C(a', Z_i)[K^T\beta_{i-1}](a')} \tag{7}$$

where K^T denotes the transposed matrix representing the Markov kernel of (3). In vector form, (7) becomes

$$\beta_i = \frac{1}{\mathbf{1}^T[\boldsymbol{c}_{Z_i} \odot [K^T\beta_{i-1}]]} \boldsymbol{c}_{Z_i} \odot [K^T\beta_{i-1}], \tag{8}$$

where, for $b \in \mathcal{Z}$, $\boldsymbol{c}_b$ denotes the column vector whose ath component is $C(a,b)$ and $\odot$ denotes componentwise multiplication. Thus, defining the mapping $T : \mathcal{Z} \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ by

$$T(b,\beta) = \frac{1}{\mathbf{1}^T[\boldsymbol{c}_b \odot [K^T\beta]]} \boldsymbol{c}_b \odot [K^T\beta] \tag{9}$$

(8) assumes the form

$$\beta_i = T(Z_i, \beta_{i-1}). \tag{10}$$

An equivalent way of expressing (10) (which will be of convenience in the sequel) is in terms of the log likelihoods: for $a, b \in \mathcal{X}$

$$\begin{aligned}\log\frac{\beta_i(a)}{\beta_i(b)} &= \log\frac{C(a,Z_i)}{C(b,Z_i)} + \log\frac{[K^T\beta_{i-1}](a)}{[K^T\beta_{i-1}](b)} \\ &= \log\frac{C(a,Z_i)}{C(b,Z_i)} + \log\frac{\sum_{c\in\mathcal{X}} K(c,a)\beta_{i-1}(c)}{\sum_{c\in\mathcal{X}} K(c,b)\beta_{i-1}(c)}. \end{aligned} \tag{11}$$

By an analogous computation, we get

$$\gamma_{i-1} = T_r(Z_{i-1}, \gamma_i)$$

with the mapping $T_r : \mathcal{Z} \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ defined by

$$T_r(b,\gamma) = \frac{1}{\mathbf{1}^T\left[\boldsymbol{c}_b \odot \left[K_r^T\gamma\right]\right]} \boldsymbol{c}_b \odot \left[K_r^T\gamma\right].$$

By the definition of β_i, clearly $\beta_i \in \mathcal{F}(Z_{-\infty}^{i})$ and similarly $\gamma_i \in \mathcal{F}(Z_i^{\infty})$. Somewhat surprisingly, however, both $\{\beta_i\}$ and $\{\gamma_i\}$ turn out to be first-order Markov processes. Indeed, defining for $E \subseteq \mathcal{M}(\mathcal{X})$ and $\beta, \gamma \in \mathcal{M}(\mathcal{X})$

$$\begin{aligned} F(E,\beta) &= \sum_{b:T(b,\beta)\in E} \mathbf{1}^T[\boldsymbol{c}_b \odot [K^T\beta]] \\ F_r(E,\gamma) &= \sum_{b:T_r(b,\gamma)\in E} \mathbf{1}^T\left[\boldsymbol{c}_b \odot \left[K_r^T\gamma\right]\right] \end{aligned}$$

the following result is implicit in [6].

Claim 1 (Blackwell [6]): The processes $\{\beta_i\}$ and $\{\gamma_i\}$ defined, respectively, in (5) and (6) are both stationary first-order Markov processes. Furthermore, Q, the distribution of β_i, satisfies, for each Borel set $E \subseteq \mathcal{M}(\mathcal{X})$, the integral equation

$$Q(E) = \int_{\beta\in\mathcal{M}(\mathcal{X})} F(E,\beta)\, dQ(\beta) \tag{12}$$

and the distribution of γ_i, Q_r satisfies the integral equation

$$Q_r(E) = \int_{\gamma\in\mathcal{M}(\mathcal{X})} F_r(E,\gamma)\, dQ_r(\gamma).$$

We reproduce a proof in the spirit of [6] for completeness.

Proof of Claim 1: We prove the claim for $\{\beta_i\}$, the proof for $\{\gamma_i\}$ being analogous. Stationarity is clear. To prove the Markov property, note that

$$\begin{aligned} &P\left(\beta_i \in E \,\middle|\, Z_{-\infty}^{i-1}\right) \\ &\quad= P\left(T(Z_i,\beta_{i-1}) \in E \,\middle|\, Z_{-\infty}^{i-1}\right) \\ &\quad= \sum_{b:T(b,\beta_{i-1})\in E} P\left(Z_i = b \,\middle|\, Z_{-\infty}^{i-1}\right) \\ &\quad= \sum_{b:T(b,\beta_{i-1})\in E} \left[\sum_{a\in\mathcal{X}} P\left(Z_i = b, X_i = a \,\middle|\, Z_{-\infty}^{i-1}\right)\right] \\ &\quad= \sum_{b:T(b,\beta_{i-1})\in E} \mathbf{1}^T[\boldsymbol{c}_b \odot [K^T\beta_{i-1}]] \\ &\quad= F(E,\beta_{i-1}) \end{aligned}$$

where the last equality follows similarly as in the derivation of (10). Thus, we see that $P(\beta_i \in E \,|\, Z_{-\infty}^{i-1})$ depends on $Z_{-\infty}^{i-1}$ only through β_{i-1}, which, since $\mathcal{F}(\beta_{-\infty}^{i-1}) \subseteq \mathcal{F}(Z_{-\infty}^{i-1})$, implies that

$$P\left(\beta_i \in E \,\middle|\, \beta_{-\infty}^{i-1}\right) = F(E,\beta_{i-1}) \tag{13}$$

thus establishing the Markov property. Taking expectations in both sides of (13) gives (12). □

Note that the optimal filtering performance, $EU(\beta_i)$ (recall (2)), has a "closed-form" expression in terms of the distribution Q

$$EU(\beta_i) = \int_{\beta \in \mathcal{M}(\mathcal{X})} U(\beta)\, dQ(\beta). \tag{14}$$

Similarly, as was noted in [6], the entropy rate of $\boldsymbol{Z}$ can also be given a "closed-form" expression in terms of the distribution Q. To see this, note that

$$P\left(Z_{i+1} = z \,\middle|\, Z_{-\infty}^{i}\right) = \left[\beta_i^T \cdot K \cdot \mathcal{C}\right](z)$$

with K denoting the Markov transition matrix (with (a,b)th entry given by (3)), and $\mathcal{C}$ denoting the channel transition matrix (with (a,z)th entry given by $C(a,z)$). Thus, letting H denote the entropy functional $H(q) = -\sum_z q(z)\log q(z)$ and $\bar{H}(\boldsymbol{Z})$ denote the entropy rate of $\boldsymbol{Z}$

$$\begin{aligned}\bar{H}(\boldsymbol{Z}) &= EH\left(P\left(Z_{i+1} = \cdot | Z_{-\infty}^{i}\right)\right) = EH\left(\left[\beta_i^T \cdot K \cdot \mathcal{C}\right]\right) \\ &= \int_{\beta \in \mathcal{M}(\mathcal{X})} H([\beta^T \cdot K \cdot \mathcal{C}])\, dQ(\beta). \end{aligned} \tag{15}$$

Optimum denoising performance can also be characterized in terms of the measures Q and Q_r of Claim 1. For this, we define the $\mathcal{M}(\mathcal{X})$-valued process $\{\eta_i\}$ via

$$\eta_i(a) = P\left(X_i = a | Z_{-\infty}^{\infty}\right).$$

Standard use of Bayes rule and the statistical structure of the HMP gives [14]

$$\eta_i(a) = \frac{\frac{1}{\mu(a)}[K^T\beta_{i-1}](a)[K_r^T\gamma_{i+1}](a)C(a,Z_i)}{\sum_{a' \in \mathcal{X}} \frac{1}{\mu(a')}[K^T\beta_{i-1}](a')\left[K_r^T\gamma_{i+1}\right](a')C(a',Z_i)}$$

or, in vector notation

$$\begin{aligned}\eta_i &= \frac{[K^T\beta_{i-1}] \odot [K_r^T\gamma_{i+1}] \odot \boldsymbol{c}_{Z_i} \div \mu}{\mathbf{1}^T\left[[K^T\beta_{i-1}] \odot [K_r^T\gamma_{i+1}] \odot \boldsymbol{c}_{Z_i} \div \mu\right]} \\ &= G_{Z_i}(\beta_{i-1}, \gamma_{i+1}) \end{aligned} \tag{16}$$

where here $\div$ denotes componentwise division and, for $b \in \mathcal{Z}$, we define the mapping $G_b : \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ by

$$G_b(\beta,\gamma) = \frac{[K^T\beta] \odot \left[K_r^T\gamma\right] \odot \boldsymbol{c}_b \div \mu}{\mathbf{1}^T\left[[K^T\beta] \odot [K_r^T\gamma] \odot \boldsymbol{c}_b \div \mu\right]}.$$

Analogously to (11) we can write

$$\begin{aligned}\log\frac{\eta_i(a)}{\eta_i(b)} = \log\frac{C(a,Z_i)}{C(b,Z_i)} + \log\frac{[K^T\beta_{i-1}](a)}{[K^T\beta_{i-1}](b)} \\ + \log\frac{\left[K_r^T\gamma_{i+1}\right](a)}{\left[K_r^T\gamma_{i+1}\right](b)} - \log\frac{\mu(a)}{\mu(b)}. \end{aligned} \tag{17}$$

Note that, by (16), optimum denoising performance is given by $EU(\eta_i) = EU(G_{Z_i}(\beta_{i-1},\gamma_{i+1}))$, which can be expressed in terms of the measures Q and Q_r of Claim 1 analogously as in (14). More specifically, conditioned on X_i, Z_i, β_{i-1} and γ_{i+1} are independent. Thus, $EU(G_{Z_i}(\beta_{i-1},\gamma_{i+1}))$ is obtained by first conditioning on X_i. Then one needs to obtain the distribution of β_{i-1} and of γ_{i+1} conditioned on X_i, which can be done using calculations similar to those detailed.

The measure Q is hard to extract from the integral equation (12) and, unfortunately, is not explicitly known to date (cf. [6] for a discussion of some of its peculiar properties). Correspondingly, explicit expressions for optimum filtering performance (cf. [25]), denoising performance (cf. [35]), and for the entropy rate of the noisy process (cf. [6], [22], [14]) are unknown.

B. A Generic Condition for the Optimality of Symbol-by-Symbol Operations

We shall now see that the optimality of symbol-by-symbol operations for filtering and for denoising depends on the measures Q and Q_r (detailed in Claim 1) only through their supports. In what follows, we let C_Q and C_{Q_r} denote, respectively, the supports of Q and Q_r.

For $q \in \mathcal{M}(\mathcal{X})$ define $\hat{X}(q)$, the Bayes response to q, by

$$\hat{X}(q) = \left\{a \in \mathcal{X} : \boldsymbol{\lambda}_a^T q = \min_{\hat{x} \in \mathcal{X}} \boldsymbol{\lambda}_{\hat{x}}^T q\right\}. \tag{18}$$

Note that we have slightly deviated from common practice, letting $\hat{X}(q)$ be set-valued so that $|\hat{X}(q)| \geq 1$, with equality if and only if the minimizer of $\boldsymbol{\lambda}_{\hat{x}}^T q$ is unique. With this notation, the following is a direct consequence of the definition of β_i, and of the fact that an optimal scheme satisfies (respectively) (1) or (2).

Fact 1: A filtering scheme $\{\hat{X}_i(\,\cdot\,)\}$ is optimal if and only if for each i

$$P\left(\hat{X}_i\left(Z_{-\infty}^{i}\right) \in \hat{X}(\beta_i)\right) = 1 \tag{19}$$

where $\hat{X}(\beta_i)$ denotes the Bayes response to β_i, as defined in (18). A denoising scheme $\{\hat{X}_i(\,\cdot\,)\}$ is optimal if and only if, for each i

$$P\left(\hat{X}_i\left(Z_{-\infty}^{\infty}\right) \in \hat{X}(\eta_i)\right) = 1 \tag{20}$$

where $\hat{X}(\eta_i)$ denotes the Bayes response to η_i, as defined in (18).

For $f : \mathcal{Z} \to \mathcal{X}$, define $S_f \subseteq \mathcal{M}(\mathcal{X})$ by

$$S_f = \{s \in \mathcal{M}(\mathcal{X}) : f(b) \in \hat{X}(T(b,s))\ \forall b \in \mathcal{Z}\}. \tag{21}$$

In words, S_f is the set of distributions on the clean source alphabet sharing the property that $f(b)$ is the Bayes response to $T(b,s)$ for all $b \in \mathcal{Z}$. Somewhat less formally (neglecting the possibility that $|\hat{X}(T(b,s))| > 1$), S_f is the largest set with the property that the Bayes response to $T(\cdot,s)$ is $f(\,\cdot\,)$ regardless of the value of $s \in S_f$. It is thus clear, by (19) and (8), that singlet decoding with $f(\,\cdot\,)$ will result in optimal filtering for X_i if β_{i-1} is guaranteed to land in S_f. Conversely, if β_{i-1} can fall outside of S_f then, on that event, the Bayes response to $T(\cdot,\beta_{i-1})$ will not be $f(\,\cdot\,)$, so singlet decoding with f cannot be optimal. More formally, we have the following.

Theorem 1: Assume $C(a,b) > 0$ for all $a \in \mathcal{X}, b \in \mathcal{Z}$. The singlet decoding scheme $\hat{X}_i = f(Z_i)$ is an optimal filter if and only if $C_Q \subseteq S_f$.

Proof of Theorem 1: Suppose that $C_Q \subseteq S_f$. Then $P(\beta_{i-1} \in S_f) = 1$ and, by the definition of S_f

$$P(f(b) \in \hat{X}(T(b,\beta_{i-1})), \quad \forall b \in \mathcal{Z}) = 1.$$

Consequently

$$P(f(Z_i) \in \hat{X}(T(Z_i,\beta_{i-1}))) = P(f(Z_i) \in \hat{X}(\beta_i)) = 1$$

establishing optimality by (19).

For the other direction, suppose that $C_Q \not\subseteq S_f$. Then there exists a $J \subseteq \mathcal{M}(\mathcal{X})$ with $J \cap S_f = \emptyset$ such that $P(\beta_{i-1} \in J) > 0$. Since $J \cap S_f = \emptyset$, this implies that

$$P(f(b) \in \hat{X}(T(b,\beta_{i-1})), \quad \forall b \in \mathcal{Z}) < 1$$

implying the existence of $b \in \mathcal{Z}$ with

$$P(f(b) \in \hat{X}(T(b,\beta_{i-1}))) < 1$$

implying, in turn, when combined with (4), the existence of $a \in \mathcal{X}$ such that

$$P(f(b) \in \hat{X}(T(b,\beta_{i-1})) \,|\, X_i = a) < 1. \tag{22}$$

Now, Z_i and β_{i-1} are conditionally independent given X_i, and therefore,

$$\begin{aligned} &P(f(Z_i) \in \hat{X}(\beta_i) \,|\, X_i = a) \\ &\quad = P(f(Z_i) \in \hat{X}(T(Z_i,\beta_{i-1})) \,|\, X_i = a) \\ &\quad = \sum_{b' \in \mathcal{Z}} P(f(b') \in \hat{X}(T(b',\beta_{i-1})) \,|\, X_i = a) C(a,b'). \end{aligned} \tag{23}$$

Inequality (22), combined with (23) and the fact that $C(a,b) > 0$, implies $P(f(Z_i) \in \hat{X}(\beta_i) | X_i = a) < 1$, which, in turn, leads to $P(f(Z_i) \in \hat{X}(\beta_i)) < 1$. Thus, $\hat{X}_i(Z_{-\infty}^i) = f(Z_i)$ does not satisfy (19) and, consequently, is not an optimal filtering scheme. □

Remark: The assumption that all channel transitions have positive probabilities was made to avoid some technical nuisances in the proof. For the general case, the above proof can be slightly elaborated to show that Theorem 1 continues to hold upon slight modification of the definition of S_f in (21) to

$$\{s \in \mathcal{M}(\mathcal{X}) : f(b) \in \hat{X}(T(b,s)) \; \forall b \in \mathcal{S}(s)\}$$

where $\mathcal{S}(s) = \{b \in \mathcal{Z} : C(a,b) > 0 \text{ for some } a \in \mathcal{X} \text{ with } [s^T K](a) > 0\}$.

A similar line of argumentation leads to the denoising analogue of Theorem 1. For $f : \mathcal{Z} \to \mathcal{X}$ define $R_f \subseteq \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ by

$$\begin{aligned} R_f = \{(s_1,s_2) \in \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) : f(b) \\ \in \hat{X}(G_b(s_1,s_2)) \; \forall b \in \mathcal{Z}\}. \end{aligned}$$

Theorem 2: Assume $K(a,b) > 0$ and $C(a,z) > 0$ for all $a \in \mathcal{X}, b \in \mathcal{X}, z \in \mathcal{Z}$. The scalar scheme $\hat{X}_i = f(Z_i)$ is an optimal denoiser if and only if $C_Q \times C_{Q_r} \subseteq R_f$.

The proof, deferred to the Appendix, is similar to that of Theorem 1.

In general, even the supports C_Q and C_{Q_r} may be difficult to obtain explicitly. In such cases, however, outer and inner bounds on the supports may be manageable to obtain. For example, to get outer bounds, it is enough to bound the supports of the distributions of $\beta_i(a)$ for each a, which is a much simpler problem that can be handled using an approach similar to that underlying the results in Section IV. Then, the preceding theorems can be used to obtain, respectively, sufficient and necessary conditions for the optimality of symbol-by-symbol schemes. As we shall see in the next section, when the source alphabet is binary, Q and Q_r are effectively distributions over the unit interval, and enough information about their supports can be extracted to characterize the necessary and sufficient conditions for the optimality of symbol-by-symbol schemes. For this case, the conditions in Theorems 1 and 2 can be recast in terms of intersections of intervals with explicitly characterized endpoints.

IV. The Binary Markov Source

Throughout this section, we consider the special case of the setting of the previous section, when the underlying noiseless process is a first-order binary Markov process, and is corrupted by a memoryless channel. Specifically, assume now $\mathcal{X} = \{0,1\}$ and that $\{X_i\}$ is a stationary binary Markov source. Let $\pi_{01} = K(0,1)$ denote the probability of transition from 0 to 1 and $\pi_{10} = K(1,0)$. We assume, without loss of generality, that $0 < \pi_{01} \leq 1$ and $0 < \pi_{10} \leq 1$ (the remaining cases imply zero probability to one of the symbols and so are trivial). For concreteness, we shall assume Hamming loss, though it will be clear that the derivation (and analogous results) carry over to the general case.

For this case, (11) becomes

$$\begin{aligned} \log \frac{\beta_i(1)}{1-\beta_i(1)} &= \log \frac{C(1,Z_i)}{C(0,Z_i)} + \log \frac{\sum_{c \in \mathcal{X}} K(c,1)\beta_{i-1}(c)}{\sum_{c \in \mathcal{X}} K(c,0)\beta_{i-1}(c)} \\ &= \log \frac{C(1,Z_i)}{C(0,Z_i)} \\ &\quad + \log \frac{\pi_{01}(1-\beta_{i-1}(1)) + (1-\pi_{10})\beta_{i-1}(1)}{(1-\pi_{01})(1-\beta_{i-1}(1)) + \pi_{10}\beta_{i-1}(1)}. \end{aligned}$$

Equivalently, letting $l_i = \log \frac{\beta_i(1)}{1-\beta_i(1)}$, we obtain

$$l_i = \log \frac{C(1,Z_i)}{C(0,Z_i)} + h(l_{i-1}) \tag{24}$$

where

$$h(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x \pi_{10}}.$$

Denoting further $k_i = \log \frac{\gamma_i(1)}{1-\gamma_i(1)}, m_i = \log \frac{\eta_i(1)}{1-\eta_i(1)}$, and since the time-reversibility of the binary Markov process implies that $K_r = K$, (17) becomes

$$m_i = \log \frac{C(1,Z_i)}{C(0,Z_i)} + h(l_{i-1}) + h(k_{i+1}) - \log \frac{\pi_{01}}{\pi_{10}}. \tag{25}$$

Note that the above-defined l_i, k_i, and m_i are $\mathbb{R} \cup \{\infty, -\infty\}$-valued random variables.

A summary of our findings in this section is as follows. We will start in subsection A, by obtaining an explicit characterization of the upper and lower endpoints of the supports of l_i and

$$f(\pi_{01},\pi_{10},\alpha)=\log\left[\frac{-1+\alpha+\pi_{01}-\alpha\pi_{10}+\sqrt{4\alpha\pi_{01}\pi_{10}+(1-\alpha-\pi_{01}+\alpha\pi_{10})^2}}{2\pi_{10}}\right]. \tag{36}$$

m_i, respectively, in Theorems 3 and 4. We will then see, in subsection B (in Claims 2, 3, and 4), how necessary and sufficient conditions for the optimality of singlet decoding can be given solely in terms of these upper and lower endpoints. We will also give an example, in Corollary 1, of how the characterization in subsection A can be combined with the latter results to obtain an easily verifiable necessary and sufficient condition for the optimality of singlet decoding, dependent on the process parameters only through the Markov transition probabilities π_{10},π_{01}, and on the aforementioned upper and lower endpoints (which, in turn, are also explicitly given). In subsection C, we shall further illustrate the concrete use of these results for the case where the noisy channel is a BSC, and thereby recover, extend, and simplify the results of [12]. We will end this section in subsection D by noting that our results imply that singlet decoding for the Laplacian channel can only be optimal when the observations are useless, and that singlet decoding is never optimal for the binary input additive white Gaussian noise (BIAWGN) channel.

A. The Support of the Log Likelihoods

By differentiating, it is easily verified the following.

Fact 2: The function h is nondecreasing whenever $\pi_{10}+\pi_{01}\leq 1$, otherwise it is nonincreasing.

Define now

$$U_{\rm bs}=\operatorname{ess\,sup}\frac{C(1,Z_1)}{C(0,Z_1)} \tag{26}$$

and

$$L_{\rm bs}=\operatorname{ess\,inf}\frac{C(1,Z_1)}{C(0,Z_1)}. \tag{27}$$

For a general binary input channel, the ratios in (26) and (27) would be replaced by the Radon–Nykodim derivative of the output distribution given input symbol 1 with respect to (w.r.t.) the output distribution given input symbol 0.

Examples:

- BSC with $\delta\leq 1/2$. $U_{\rm bs}=\frac{1-\delta}{\delta}, L_{\rm bs}=\frac{\delta}{1-\delta}$.
- BIAWGN channel, where
$$C(0,z)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(z+\mu)^2}$$
$$C(1,z)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(z-\mu)^2},$$
$$\mu>0,\ \sigma^2>0,\ z\in\mathbb{R}. \tag{28}$$
In this case, it is easy to verify that $U_{\rm bs}=\infty, L_{\rm bs}=0$.
- Binary input additive Laplacian noise (BIALN) channel, where
$$C(0,z)=c(\alpha)e^{-\alpha|z+\mu|}$$
$$C(1,z)=c(\alpha)e^{-\alpha|z-\mu|},\qquad \mu>0,\ z\in\mathbb{R} \tag{29}$$
$c(\alpha)$ being the normalization factor. Here, it is easily seen that $U_{\rm bs}=e^{2\alpha\mu}, L_{\rm bs}=e^{-2\alpha\mu}$.

Define further

$$I_1=\operatorname{ess\,inf} l_i \tag{30}$$

and

$$I_2=\operatorname{ess\,sup} l_i. \tag{31}$$

The reason for our interest in I_1 and I_2 is that $[I_1,I_2]$ is the smallest interval containing the support of l_i. The sufficiency of symbol-by-symbol operations for the filtering problem, as will be seen later, depends on the support of l_i solely through this interval.

Theorem 3: The pair (I_1,I_2) (defined in (30) and (31)) is given by the unique solution (in the extended real line) to the following.

1. When $\pi_{10}+\pi_{01}\leq 1$: $I_1=\log L_{\rm bs}+h(I_1)$ and $I_2=\log U_{\rm bs}+h(I_2)$.
2. When $\pi_{10}+\pi_{01}>1$: $I_1=\log L_{\rm bs}+h(I_2)$ and $I_2=\log U_{\rm bs}+h(I_1)$.

Note, in particular, the dependence of (I_1,I_2) on the channel only through $L_{\rm bs}$ and $U_{\rm bs}$.

Proof of Theorem 3: We assume $\pi_{10}+\pi_{01}\leq 1$ (the proof for the case $\pi_{10}+\pi_{01}>1$ is analogous). Note that if $\pi_{01}=1$ (and therefore $\pi_{10}=0$) the claim follows trivially since in this case $h(x)\equiv\infty$ and $I_1=I_2=\infty$. Similarly, the claim is trivially true when $\pi_{10}=1$ and $\pi_{01}=0$. We can assume then that both $0<\pi_{10}<1$ and $0<\pi_{01}<1$. Monotonicity and continuity of h imply

$$\operatorname{ess\,inf} h(l_{i-1})=h(\operatorname{ess\,inf} l_{i-1})=h(\operatorname{ess\,inf} l_i)=h(I_1). \tag{32}$$

Thus,

$$\begin{aligned}
I_1&=\operatorname{ess\,inf} l_i\\
&=\operatorname{ess\,inf}\left[\log\frac{C(1,Z_i)}{C(0,Z_i)}+h(l_{i-1})\right] &(33)\\
&=\operatorname{ess\,inf}\left[\log\frac{C(1,Z_i)}{C(0,Z_i)}\right]+\operatorname{ess\,inf} h(l_{i-1}) &(34)\\
&=\log\left[\operatorname{ess\,inf}\frac{C(1,Z_i)}{C(0,Z_i)}\right]+h(I_1)\\
&=\log L_{\rm bs}+h(I_1) &(35)
\end{aligned}$$

where (33) follows from (24), (34) follows since all transitions of the Markov chain have positive probability, implying that the distribution of $\log\frac{C(1,Z_i)}{C(0,Z_i)}+h(l_{i-1})$ is absolutely continuous w.r.t., say, the distribution of R_1+R_2, where R_1 and R_2 are independent and $R_1\stackrel{d}{=}\log\frac{C(1,Z_i)}{C(0,Z_i)}$ and $R_2\stackrel{d}{=}h(l_{i-1})$, and equality (35) is due to (32). The relationship $I_2=\log U_{\rm bs}+h(I_2)$ is established similarly. □

Elementary algebra shows that for $\pi_{10}+\pi_{01}\leq 1$ and any $\alpha>0$, the unique real solution (for x) of the equation $x=\log\alpha+h(x)$ is given by $x=f(\pi_{01},\pi_{10},\alpha)$ where we get (36) at the top of the page. Thus, from Theorem 3 we get $I_1=f(\pi_{01},\pi_{10},L_{\rm bs})$ and $I_2=f(\pi_{01},\pi_{10},U_{\rm bs})$ when $\pi_{10}+\pi_{01}\leq 1$. An explicit form of the unique solution for the pair (I_1,I_2) when $\pi_{10}+\pi_{01}>1$ can also be obtained by solving the pair of equations in the second item of Theorem 3 (we omit the expressions which are somewhat more involved than that in (36)).

For the analogous quantities in the denoising problem

$$J_1 = \operatorname{ess\,inf} m_i$$

and

$$J_2 = \operatorname{ess\,sup} m_i$$

we have the following result.

Theorem 4:

1. When $\pi_{10} + \pi_{01} \leq 1$: $J_1 = \log L_{\rm bs} + 2h(I_1) - \log \frac{\pi_{01}}{\pi_{10}}$ and $J_2 = \log U_{\rm bs} + 2h(I_2) - \log \frac{\pi_{01}}{\pi_{10}}$.
2. When $\pi_{10} + \pi_{01} > 1$: $J_1 = \log L_{\rm bs} + 2h(I_2) - \log \frac{\pi_{01}}{\pi_{10}}$ and $J_2 = \log U_{\rm bs} + 2h(I_1) - \log \frac{\pi_{01}}{\pi_{10}}$.

Proof of Theorem 4: The proof is similar to that of Theorem 3, using (25) (instead of (24)), and the fact (by time reversibility) that l_{i-1} and k_{i+1} are equal in distribution, and, in particular, have equal supports. □

Thus, when $\pi_{10} + \pi_{01} \leq 1$, we get the explicit forms

$$J_1 = \log L_{\rm bs} + 2h\big(f(\pi_{01}, \pi_{10}, L_{\rm bs})\big) - \log \frac{\pi_{01}}{\pi_{10}}$$

and

$$J_2 = \log U_{\rm bs} + 2h\big(f(\pi_{01}, \pi_{10}, U_{\rm bs})\big) - \log \frac{\pi_{01}}{\pi_{10}}$$

where f was defined in (36). Explicit (though more cumbersome) expressions can also be obtained for the case $\pi_{10}+\pi_{01}>1$.

B. Conditions for Optimality of Singlet Decoding

When specialized to the present setting, Fact 1 asserts that in terms of the log-likelihood processes $\{l_i\}$ and $\{m_i\}$, $\hat{X}_i$ is an optimal filter if and only if it is of the form

$$\hat{X}_i\left(Z_{-\infty}^i\right) = f_{\rm opt}\left(Z_{-\infty}^i\right) = \begin{cases} 1, & \text{a.s. on } \{l_i > 0\} \\ 0. & \text{a.s. on } \{l_i < 0\} \\ \text{arbitrary} & \text{on } \{l_i = 0\}. \end{cases} \tag{37}$$

Similarly, a denoiser is optimal if and only if it is of the form

$$\hat{X}_i\left(Z_{-\infty}^\infty\right) = g_{\rm opt}\left(Z_{-\infty}^\infty\right) = \begin{cases} 1, & \text{a.s. on } \{m_i > 0\} \\ 0, & \text{a.s. on } \{m_i < 0\} \\ \text{arbitrary} & \text{on } \{m_i = 0\}. \end{cases} \tag{38}$$

The following is a direct consequence of (37) and (38) and the definitions of I_1, I_2, J_1, J_2.

Claim 2: The filter ignoring its observations and saying

- "all ones" is optimal if and only if $I_1 \geq 0$;
- "all zeros" is optimal if and only if $I_2 \leq 0$.

The denoiser ignoring its observations and saying

- "all ones" is optimal if and only if $J_1 \geq 0$;
- "all zeros" is optimal if and only if $J_2 \leq 0$.

Proof: To prove the first item, note that if $I_1 \geq 0$ then $l_i \geq 0$ a.s. Thus, by (37), $\hat{X}_i(Z_{-\infty}^i) \equiv 1$ is an optimal filter. Conversely, if $\hat{X}_i(Z_{-\infty}^i) \equiv 1$ is an optimal filter then, by (37), $l_i \geq 0$ a.s. which implies that $I_1 \geq 0$. The remaining items are proven similarly. □

Note that Theorems 3 and 4, together with Claim 2, provide complete and explicit characterization of the cases where the observations are "useless" for the filtering and denoising problems. For example, for the filtering problem, by recalling that $I_1 = f(\pi_{01}, \pi_{10}, L_{\rm bs})$ and $I_2 = f(\pi_{01}, \pi_{10}, U_{\rm bs})$ (with f given in (36)) we obtain the following.

Corollary 1: Assume $\pi_{10} + \pi_{01} \leq 1$. The filter ignoring its observations and saying

- "all-zeros" is optimal if and only if
$$\sqrt{4U_{\rm bs}\pi_{01}\pi_{10} + (1 - U_{\rm bs} - \pi_{01} + U_{\rm bs}\pi_{10})^2} \leq 1 - U_{\rm bs} - \pi_{01} + U_{\rm bs}\pi_{10} + 2\pi_{10};$$
- "all-ones" is optimal if and only if
$$\sqrt{4L_{\rm bs}\pi_{01}\pi_{10} + (1 - L_{\rm bs} - \pi_{01} + L_{\rm bs}\pi_{10})^2} \geq 1 - L_{\rm bs} - \pi_{01} + L_{\rm bs}\pi_{10} + 2\pi_{10}.$$

Explicit characterizations for the case $\pi_{1,0}+\pi_{0,1} > 1$ as well as for the denoising problem can be obtained similarly.

We now turn to a general characterization of the conditions under which the optimum scheme needs to base its estimate only on the present symbol.

Claim 3: Assume $\pi_{10}+\pi_{01} \leq 1$. Singlet decoding is optimal for the filtering problem if and only if

$$\log \frac{C(1, Z_1)}{C(0, Z_1)} \notin (\log U_{\rm bs} - I_2, \log L_{\rm bs} - I_1) \text{ a.s.} \tag{39}$$

or, in other words, the support of $\log \frac{C(1,Z_1)}{C(0,Z_1)}$ does not intersect the $(\log U_{\rm bs} - I_2, \log L_{\rm bs} - I_1)$ interval.

Proof of Claim 3: As in the proof of Theorem 3, the case $\pi_{10} = 1$ or $\pi_{01} = 1$ follows trivially, so we assume all transition probabilities are positive. From (37) it follows that the optimal filter is a singlet decoder if and only if the sign of l_i is, with probability one, determined solely by Z_i. From (24) (and the fact that all transitions of the underlying Markov chain have positive probability) it follows that this can be the case if and only if, with probability one

$$\log \frac{C(1, Z_i)}{C(0, Z_i)} \geq -\operatorname{ess\,inf} h(l_{i-1}) \quad \text{or}$$
$$\log \frac{C(1, Z_i)}{C(0, Z_i)} \leq -\operatorname{ess\,sup} h(l_{i-1}). \tag{40}$$

But, by Theorem 3,

$$(-\operatorname{ess\,sup} h(l_{i-1}), -\operatorname{ess\,inf} h(l_{i-1})) = (\log U_{\rm bs} - I_2, \log L_{\rm bs} - I_1)$$

so (40) is equivalent to (39). □

The analogous result for the denoising problem is the following.

Claim 4: Singlet decoding is optimal for the denoising problem if and only if

$$\frac{1}{2}\left[\log \frac{C(1, Z_1)}{C(0, Z_1)} - \log \frac{\pi_{01}}{\pi_{10}}\right] \notin (\log U_{\rm bs} - I_2, \log L_{\rm bs} - I_1) \text{ a.s.} \tag{41}$$

or, in other words, the support of $\frac{1}{2}\left[\log \frac{C(1,Z_1)}{C(0,Z_1)} - \log \frac{\pi_{01}}{\pi_{10}}\right]$ does not intersect the $(\log U_{\rm bs} - I_2, \log L_{\rm bs} - I_1)$ interval.

Proof of Claim 4: The proof follows from (25) and Theorem 4, analogously as Claim 3 followed from (24) and Theorem 3. □

Remarks:

1. Claim 3 (resp., 4) explicitly characterizes the conditions under which singlet decoding is optimal for the filtering (resp., denoising) problem. The proof idea, however, is readily seen to imply, more generally, even in cases where singlet decoding is not optimal, that if the observation Z_i happens to be such that $\log \frac{C(1,Z_i)}{C(0,Z_i)}$ (resp., $\frac{1}{2}\left[\log \frac{C(1,Z_i)}{C(0,Z_i)} - \log \frac{\pi_{01}}{\pi_{10}}\right]$) falls outside the $(\log U_{\rm bs} - I_2, \log L_{\rm bs} - I_1)$ interval, then the optimal estimate of X_i will be independent of the other observations (namely, it will be 0 if that quantity falls below the interval and 1 if it falls above it, irrespective of other observations).
2. Note that the optimality of singlet decoding depends on the noisy channel only through the support of $\log \frac{C(1,Z_1)}{C(0,Z_1)}$ (or, equivalently, of $\frac{C(1,Z_1)}{C(0,Z_1)}$).

C. Optimality of Singlet Decoding for the BSC

We now show that the results of the previous subsection, when specialized to the BSC, give the explicit characterization of optimality of singlet decoding derived initially in [12]. The results below refine and extend those of [12] in that they provide the explicit conditions for optimality of the "say-what-you-see" scheme in the nonsymmetric Markov chain case as well. Also, our derivation (of the results in the previous subsection) avoids the need to explicitly find the "worst observation sequence" (the approach on which the results of [12] are based). Finally, due to a parametrization different than that in [12], the region of optimality of singlet decoding for this setting admits a simple form.

We assume here a BSC $(\delta)(\delta \le 1/2)$, and restrict attention throughout to the case $\pi_{10} + \pi_{01} \le 1$. In this case, $U_{\rm bs} = (1-\delta)/\delta$, and $L_{\rm bs} = \delta/(1-\delta)$ so that, by Theorem 3 (and the remark following its proof), the smallest interval containing the support of l_i is

$$[f(\pi_{01}, \pi_{10}, \delta/(1-\delta)), f(\pi_{01}, \pi_{10}, (1-\delta)/\delta)]$$

where f is given in (36).

Corollary 2: For the BSC (δ), we have the following.

1. Filtering: The "say-what-you-see" filter is optimal if and only if either
$$\pi_{01} \le \pi_{10} \text{ and } 2\log\frac{1-\delta}{\delta} \ge -f(\pi_{01}, \pi_{10}, \delta/(1-\delta))$$
or
$$\pi_{01} > \pi_{10} \text{ and } 2\log\frac{1-\delta}{\delta} \ge f(\pi_{01}, \pi_{10}, (1-\delta)/\delta).$$
2. Denoising: The "say-what-you-see" denoiser is optimal if and only if
$$\frac{3}{2}\log\frac{1-\delta}{\delta} \ge \max\left\{-\frac{1}{2}\log\frac{\pi_{10}}{\pi_{01}} - f(\pi_{01}, \pi_{10}, \delta/(1-\delta)),\ \frac{1}{2}\log\frac{\pi_{10}}{\pi_{01}} + f(\pi_{01}, \pi_{10}, (1-\delta)/\delta)\right\}.$$

Remark: Note that Corollary 2, together with Corollary 1, completely characterize the cases of optimality of singlet filtering for the BSC. Optimality of singlet denoising is similarly characterized by the second part of Corollary 2, together with (for the "all-zero" and "all-one" schemes) the conditions $J_2 \le 0$ and $J_1 \ge 0$ (recall Claim 2), using the expressions for J_1 and J_2, as characterized by Theorem 4.

Proof of Corollary 2: Claim 3 and the remark closing the previous subsection imply that the condition for optimality of the "say-what-you-see" filter is that $\log[\frac{1-\delta}{\delta}]$ be above the interval on the right-hand side of (39) and $\log[\frac{\delta}{1-\delta}]$ be below that interval. More compactly, the condition is

$$\log\left[\frac{1-\delta}{\delta}\right] \ge \max\Big\{\log L_{\rm bs} - f(\pi_{01}, \pi_{10}, \delta/(1-\delta)),\ -\log U_{\rm bs} + f(\pi_{01}, \pi_{10}, (1-\delta)/\delta)\Big\}$$

or, since for this case $\frac{1-\delta}{\delta} = U_{\rm bs} = 1/L_{\rm bs}$

$$2\log\frac{1-\delta}{\delta} \ge \max\Big\{-f(\pi_{01}, \pi_{10}, \delta/(1-\delta)),\ f(\pi_{01}, \pi_{10}, (1-\delta)/\delta)\Big\}. \quad (42)$$

Now, it is straightforward to check that when $\pi_{01} \le \pi_{10}$, it is the left branch which attains the maximum in (42), otherwise it is the right branch. This establishes the first part. The second part follows from Claim 4, similarly as the first part followed from Claim 3. □

For the symmetric Markov source, where $\pi_{01} = \pi_{10} = \pi$, the "all-zeros" and "all-ones" schemes are clearly always suboptimal, except for the trivial case $\delta = 1/2$. For the optimality of the "say-what-you-see" scheme, the conditions in Corollary 2 simplify, following elementary algebra, to give the following.

Corollary 3: For the symmetric Markov source with $\pi \le 1/2$, corrupted by the BSC (δ), we have the following conditions.

1. The "say-what-you-see" scheme is an optimal filter if and only if either $\pi \ge 1/4$ (and all $0 \le \delta \le 1/2$), or $\pi < 1/4$ and $\delta \le \frac{1}{2}(1-\sqrt{1-4\pi})$. More compactly, if and only if
$$\delta \le \frac{1}{2}\left(1 - \sqrt{\max\{1-4\pi, 0\}}\right).$$
2. The "say-what-you-see" scheme is an optimal denoiser if and only if either $\pi \ge 1/3$ (and all $0 \le \delta \le 1/2$), or $\pi < 1/3$ and $\delta \le \frac{1}{2}(1-\sqrt{1-4(\frac{\pi}{1-\pi})^2})$. More compactly, if and only if
$$\delta \le \frac{1}{2}\left(1 - \sqrt{\max\{1-4(\frac{\pi}{1-\pi})^2, 0\}}\right).$$

Note that Corollary 3, both for the filtering and the denoising problems, completely characterizes the region in the square $0 \le \pi \le 1/2, 0 \le \delta \le 1/2$, where the minimum attainable error rate is δ. The minimum error rate at all points outside that region remains unknown. The asymptotic behavior of the minimum error rate as $\pi \to 0$ has been characterized, respectively, for the filtering and denoising problems in [25] and [35].

Corollary 3 carries over to cover the whole $0 \le \pi \le 1$, $0 \le \delta \le 1/2$ region as follows. The idea is to show a one-to-one correspondence between an optimal scheme for the Markov

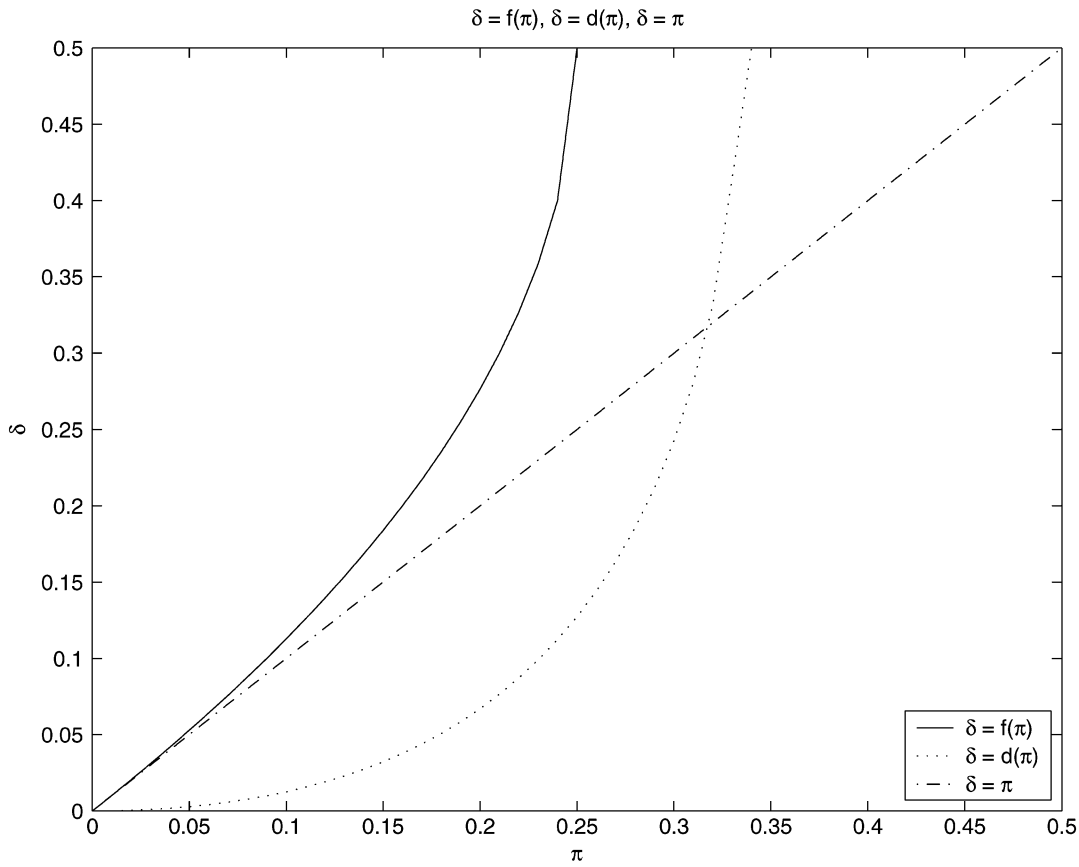

Fig. 1. Optimality region in $\delta - \pi$ plane for singlet decoding of a binary-symmetric Markov chain with transition probability π corrupted by a BSC (δ) : $\delta \leq f(\pi) = \frac{1}{2}(1 - \sqrt{\max\{1-4\pi, 0\}})$ (solid line) for filtering and $\delta \leq d(\pi) = \frac{1}{2}(1 - \sqrt{\max\{1-4(\frac{\pi}{1-\pi})^2, 0\}})$ (dotted line) for denoising. Dashed line is the $\delta = \pi$ curve which is totally contained in the singlet filtering region.

chain with transition probability π and that for the chain with transition probability $1 - \pi$. Let $\hat{X}_t(Z^t)$ be a filter and b^t be the alternating sequence, e.g., $b^t = (\cdots 010101)$. Consider the filter $\hat{Y}_t$ given by $\hat{Y}_t(Z^t) = b_t \oplus \hat{X}_t(Z^t \oplus b^t)$. We argue that the error rate of the filter $\hat{X}_t$ on the chain with transition probability π equals that of the filter $\hat{Y}_t$ on the chain with $1 - \pi$. To see this, note that $\hat{Y}_t(z^t)$ makes an error if and only if $\hat{X}_t(z^t \oplus b^t) \neq x_t \oplus b_t$. The claim follows since the distribution of $\{(X_t, Z_t)\}$ under π is equal to the distribution of $\{(X_t \oplus b_t, Z_t \oplus b_t)\}$ under $1 - \pi$. The same argument applies for denoisers. Thus, the overall region of optimality of the "say-what-you-see" scheme in the $0 \leq \pi \leq 1, 0 \leq \delta \leq 1/2$ rectangle is symmetric about $\pi = 1/2$. Note finally that the characterization for the $0 \leq \pi \leq 1, 0 \leq \delta \leq 1/2$ region trivially extends to that of the whole $0 \leq \pi \leq 1, 0 \leq \delta \leq 1$ square by looking at the complement of each bit when $\delta > 1/2$.

Fig. 1 plots the two curves associated with Corollary 3, as well as the line $\delta = \pi$. All points on or below the solid curve (and only such points) correspond to a value of the pair (π, δ) for which the "say-what-you-see" scheme is an optimal filter. All points below the dotted curve correspond to values of this pair where the "say-what-you-see" scheme is an optimal denoiser. The latter region is, of course, contained in the former. A few additional observations in the context of Fig. 1 are as follows.

1. The region $\delta \leq \pi$ is entirely contained in the region of optimality of the "say-what-you-see" filter. This can be understood by considering the genie-aided filter allowed to base its estimate of X_i on X_{i-1} (in addition to the noisy observations), which reduces to a singlet decoder when $\delta \leq \pi$. Note that this implies, in particular, that
$$\text{optimum filtering performance} \begin{cases} = \delta, & \text{for } \delta \leq \pi \\ > \pi, & \text{for } \delta > \pi. \end{cases} \tag{43}$$
2. On the other hand, the $\delta \leq \pi$ region is *not* entirely contained in the region of optimality of the "say-what-you-see" denoiser. This implies, in particular, that
$$\text{optimum denoiser performance} \begin{cases} < \delta, & \text{for a nonempty subset of the } \delta < \pi \text{ region} \\ > \pi, & \text{for a nonempty subset of the } \delta > \pi \text{ region.} \end{cases} \tag{44}$$
3. For filtering or denoising a Bernoulli (π) process corrupted by a BSC (δ), we have
$$\text{optimum filtering/denoising Bernoulli } (\pi) \begin{cases} = \delta, & \text{for } \delta \leq \pi \\ = \pi, & \text{for } \delta > \pi. \end{cases}$$
Comparing with (43) and (44) we reach the following conclusions.
   - Filtering of the symmetric Markov chain is *always* (i.e., for all values of (δ, π)) harder (not everywhere strictly) than filtering of the Bernoulli process with the same entropy rate.
   - For some regions of the parameter space, denoising of a Markov chain is harder than denoising a

Bernoulli process with the same entropy rate, while for other regions it is easier.

In particular, this implies that the entropy rate of the clean source is not completely indicative of its "filterability" and "denoisability" properties.

4. It is interesting to note that both for the filtering and the denoising problems, for π large enough ($\geq 1/4$ and $\geq 1/3$, respectively) the "say-what-you-see" scheme is optimal, no matter how noisy the observations.

D. Singlet Decoding Is Optimal for the Laplacian Channel Only When Observations Are Useless

For the Laplacian channel detailed in (29), we now argue that singlet decoding can be optimal only when the observations are useless, namely, when the optimal scheme is either the "all-zeros" or the "all-ones" (and hence, never in the case of a symmetric Markov chain). To see this, consider first the filtering problem. As is readily verified, for this channel the support of $\log \frac{C(1,Z_1)}{C(0,Z_1)}$ is the interval $[-2\alpha\mu, 2\alpha\mu]$. Thus, for the support not to intersect the interval $(\log U_{\rm bs} - I_2, \log L_{\rm bs} - I_1)$, it must lie either entirely below this interval (in which case the "all-zeros" filter would be optimal) or entirely above it (in which case the "all-ones" filter would be optimal). A similar argument applies to the denoising problem.

A similar conclusion extends to any continuous output channel when, say, the densities associated with the output distributions for the two possible inputs are everywhere positive and continuous. In this case, the support of $\log \frac{C(1,Z_1)}{C(0,Z_1)}$ will be a (not necessarily finite) interval. In particular, if it does not intersect the $(\log U_{\rm bs} - I_2, \log L_{\rm bs} - I_1)$ interval; it must be either entirely below or entirely above it.

Finally, we note that by a similar argument, for the BI-AWGN channel detailed in (28), singlet decoding can never be optimal since the support of $\log \frac{C(1,Z_1)}{C(0,Z_1)}$ (resp., $\frac{1}{2}\left[\log \frac{C(1,Z_1)}{C(0,Z_1)} - \log \frac{\pi_{01}}{\pi_{10}}\right]$) is the entire real line (so, in particular, intersects that interval).

V. Asymptotics of the Entropy Rate of the Noisy Observation Process

In this section, we digress from the filtering and denoising problems to illustrate how the bounds on the support of the log likelihoods developed in Section IV can be used to bound the entropy rate of the noisy observation process, the precise form of which is unknown (cf. [14], [22], [16], [29], and references therein). Proofs are deferred to the Appendix, part B.

Assuming a discrete-valued noisy process, from (15) it is clear that lower and upper bounds on its entropy rate are given by

$$\min_{\beta\in C_Q} H([\beta^T \cdot K \cdot \mathcal{C}]) \leq \bar{H}(\boldsymbol{Z}) \leq \max_{\beta\in C_Q} H([\beta^T \cdot K \cdot \mathcal{C}]) \quad (45)$$

with C_Q denoting the support of β_i. We now illustrate the use of (45) to derive an explicit bound for the entropy rate of the binary Markov chain corrupted by a BSC (considered in Section IV-C). For this case

$$H([\beta^T \cdot K \cdot \mathcal{C}]) = h_b([\beta(1)(1-\pi_{10}) + (1-\beta(1))\pi_{01}] * \delta) \quad (46)$$

where h_b is the binary entropy function

$$h_b(x) = -[x \log x + (1-x)\log(1-x)]$$

and $*$ denotes binary convolution defined, for $p, \delta \in [0,1]$, by

$$p * \delta = p(1-\delta) + (1-p)\delta.$$

A. Entropy Rate in the "Rare-Spikes" Regime

Assuming $\delta \leq 1/2, \pi_{01}+\pi_{10} \leq 1$, and $\pi_{01} \leq \pi_{10}$ it is readily verified that $[\beta(1)(1-\pi_{10}) + (1-\beta(1))\pi_{01}] * \delta$ is increasing with $\beta(1)$ and is in $[0, 1/2]$ when $\beta(1) \in [0, 1/2]$. Thus, since $\beta_i(1) = e^{l_i}/(1+e^{l_i})$, it follows that the right side of (45) in this case becomes

$$h_b\left(\left[\frac{e^{I_2}}{1+e^{I_2}}(1-\pi_{10}) + \frac{1}{1+e^{I_2}}\pi_{01}\right] * \delta\right) \quad (47)$$

provided $I_2 \leq 0$ (since then $e^{I_2}/(1+e^{I_2}) \leq 1/2$), as $h_b(x)$ is increasing for $0 \leq x \leq 1/2$. Hence, the expression in (47), with I_2 given explicitly in (36) with $\alpha = (1-\delta)/\delta$, is an upper bound to the entropy rate of the noisy process for all $\delta \leq 1/2$ and all π_{01}, π_{10} satisfying $\pi_{01} + \pi_{10} \leq 1$ and $\pi_{01} \leq \pi_{10}$, provided $I_2 \leq 0$. Arguing analogously, for the parameters in this region, the expression in (47) with I_1 replaced by I_2 is a lower bound on the entropy rate, yielding

$$\begin{aligned} &h_b\left(\left[\frac{e^{I_1}}{1+e^{I_1}}(1-\pi_{10}) + \frac{1}{1+e^{I_1}}\pi_{01}\right] * \delta\right) \\ &\leq \bar{H}(\pi_{10}, \pi_{01}, \delta) \\ &\leq h_b\left(\left[\frac{e^{I_2}}{1+e^{I_2}}(1-\pi_{10}) + \frac{1}{1+e^{I_2}}\pi_{01}\right] * \delta\right) \end{aligned} \quad (48)$$

where we let $\bar{H}(\pi_{10}, \pi_{01}, \delta)$ denote the entropy rate of the noisy process associated with these parameters. It is evident from (48) that the bounds become tight as I_1 and I_2 grow closer to each other or very negative.

One regime where this happens (and the conditions $I_2 \leq 0, \pi_{01} + \pi_{10} \leq 1$, and $\pi_{01} \leq \pi_{10}$ are maintained) is when the Markov chain tends to concentrate on state 0 by jumping from 1 to 0 with high probability and from 0 to 1 with low probability (the "rare-spikes" regime). More concretely, for

$$\pi_{10} = 1 - \varepsilon \quad \text{and} \quad \pi_{01} = a(\varepsilon) \quad (49)$$

where $a(\cdot)$ is an arbitrary function satisfying $0 \leq a(\varepsilon) \leq \varepsilon$ (and all ε sufficiently small so that $I_2 \leq 0$), (48) becomes

$$\begin{aligned} &h_b\left(\left[\frac{e^{I_1}}{1+e^{I_1}}\varepsilon + \frac{1}{1+e^{I_1}}a(\varepsilon)\right] * \delta\right) \\ &\leq \bar{H}(1-\varepsilon, a(\varepsilon), \delta) \\ &\leq h_b\left(\left[\frac{e^{I_2}}{1+e^{I_2}}\varepsilon + \frac{1}{1+e^{I_2}}a(\varepsilon)\right] * \delta\right). \end{aligned} \quad (50)$$

Note, in particular, that for $a(\varepsilon) = \varepsilon$, (50) gives $\bar{H}(1-\varepsilon, \varepsilon, \delta) = h_b(\varepsilon * \delta)$, as it should since for this case, the clean source is Bernoulli (ε). Furthermore, as ε becomes small, the noise-free source with parameters given in (49) tends to the "all-zeros" source so it is natural to expect that

$$\lim_{\varepsilon \downarrow 0} \bar{H}(1-\varepsilon, a(\varepsilon), \delta) = h_b(\delta). \quad (51)$$

We now use the bounds in (50), combined with the characterization of I_1 and I_2 from Section IV-A, to show that not only does (51) hold, but the convergence rate is linear in $a(\varepsilon)$ with the constant identified as well.

Theorem 5: For $0 \le \delta \le 1/2$ and an arbitrary function $a(\cdot)$ satisfying $0 < a(\varepsilon) \le \varepsilon$

$$\begin{aligned}\lim_{\varepsilon \downarrow 0} \frac{\bar{H}(a(\varepsilon), 1-\varepsilon, \delta) - h_b(\delta)}{a(\varepsilon)} \\ = \lim_{\varepsilon \downarrow 0} \frac{\bar{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta)}{a(\varepsilon)} \\ = (1-2\delta)\log\frac{1-\delta}{\delta}. \end{aligned} \tag{52}$$

B. Entropy Rate in the "Rare-Bursts" Regime

The bounds in (48) are valid also in the "rare-bursts" regime where $0 < \pi_{10} < 1$ remains fixed and $\pi_{01} = \varepsilon$ is small (since for ε small $\pi_{10} + \pi_{01} \le 1, \pi_{01} \le \pi_{10}$, and $I_2 \le 0$ will be satisfied).

For this case, we get

$$e^{I_2} = \frac{-1+\alpha+\varepsilon-\alpha\pi_{10}+\sqrt{4\alpha\varepsilon\pi_{10}+(1-\alpha-\varepsilon+\alpha\pi_{10})^2}}{2\pi_{10}} \tag{53}$$

with $\alpha = \frac{1-\delta}{\delta}$. It follows via Taylor expansions from (53) that, as $\varepsilon \downarrow 0$

$$e^{I_2} \sim \begin{cases} \frac{(1-\delta)\varepsilon}{\delta-(1-\delta)(1-\pi_{10})}, & \text{for } 1 > \pi_{10} > \frac{1-2\delta}{1-\delta} \\ \sqrt{\frac{1-\delta}{\delta\pi_{10}}}\sqrt{\varepsilon}, & \text{for } \pi_{10} = \frac{1-2\delta}{1-\delta} \\ \frac{-1+\frac{1-\delta}{\delta}(1-\pi_{10})}{\pi_{10}}, & \text{for } 0 < \pi_{10} < \frac{1-2\delta}{1-\delta} \end{cases} \tag{54}$$

or, since $\operatorname{ess\,sup}\beta_i(1) = e^{I_2}/(1+e^{I_2})$, that

$$\operatorname{ess\,sup}\beta_i(1) \sim \begin{cases} \frac{(1-\delta)\varepsilon}{\delta-(1-\delta)(1-\pi_{10})}, & \text{for } 1 > \pi_{10} > \frac{1-2\delta}{1-\delta} \\ \sqrt{\frac{1-\delta}{\delta\pi_{10}}}\sqrt{\varepsilon}, & \text{for } \pi_{10} = \frac{1-2\delta}{1-\delta} \\ 1 - \frac{\delta\pi_{10}}{(1-2\delta)(1-\pi_{10})}, & \text{for } 0 < \pi_{10} < \frac{1-2\delta}{1-\delta}. \end{cases} \tag{55}$$

Remark: Note, in particular, that

$$\lim_{\varepsilon \downarrow 0} \operatorname{ess\,sup}\beta_i(1) = \begin{cases} 0, & \text{for } \pi_{10} > \frac{1-2\delta}{1-\delta} \\ 1 - \frac{\delta\pi_{10}}{(1-2\delta)(1-\pi_{10})}, & \text{for } \pi_{10} < \frac{1-2\delta}{1-\delta}. \end{cases} \tag{56}$$

A possible intuition behind this phase transition is as follows: In the "rare-bursts" regime, the noise-free signal consists of a long stretch of zeros followed by a stretch of a few ones (a "burst") followed by another long stretch of zeros, etc. Accordingly, $\beta_i(1)$ is, with high probability, close to zero. There is always, however, positive probability of observing, say, a very long stretch of ones in the noisy signal. When that happens, there are two extremal explanations for it. One is that this is the result of a large deviations event in the channel noise (namely, that all noise components $= 1$ while the underlying signal is at zero). The other extreme is that this is the result of a long burst of ones in the noise-free signal (while all noise components are zero). If the length of the burst is l then the first possibility has probability $\approx \delta^l(1-\varepsilon)^l \approx \delta^l$ while the second one $\approx (1-\delta)^l(1-\pi_{10})^l$ (up to factors that have subexponential dependence on l). Thus, when $\delta > (1-\delta)(1-\pi_{10})$, equivalently, when $\pi_{10} > \frac{1-2\delta}{1-\delta}$, even when observing a long stretch of ones in the noisy signal the underlying clean symbol is still overwhelmingly more likely to be a zero than a one. Thus, $\beta_i(1)$ will always be close to zero (and hence $\operatorname{ess\,sup}\beta_i(1)$ will be close to zero). On the other hand, when $\pi_{10} < \frac{1-2\delta}{1-\delta}$, a very long stretch of ones in the noisy signal is more likely to be due to a long burst (with noise components at zero) than to a fluctuation in the noise components, and therefore, when such bursts occur, the value of $\beta_i(1)$ will rise significantly above zero (so $\operatorname{ess\,sup}\beta_i(1)$ is significantly above zero).

Continuing the derivation, e^{I_1} is given by the right side of (53) with $\alpha = \frac{\delta}{1-\delta}$, so

$$\operatorname{ess\,inf}\beta_i(1) \sim e^{I_1} \sim \frac{\delta\varepsilon}{(1-\delta)-\delta(1-\pi_{10})} \tag{57}$$

(since $\pi_{10} > \frac{2\delta-1}{\delta}$ always holds for $\delta \le 1/2$). Combining (54), (57), and (48) leads to the following.

Theorem 6:

1. For $0 \le \delta \le 1/2$ and $0 < \pi_{10} < 1$

$$\liminf_{\varepsilon \downarrow 0} \frac{\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\varepsilon} \ge \frac{(1-\delta)(1-2\delta)}{(1-\delta)-\delta(1-\pi_{10})}\log\frac{1-\delta}{\delta}. \tag{58}$$

2. For $0 \le \delta \le 1/2$ and $\frac{1-2\delta}{1-\delta} < \pi_{10} < 1$

$$\limsup_{\varepsilon \downarrow 0} \frac{\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\varepsilon} \le \frac{\delta(1-2\delta)}{\delta-(1-\delta)(1-\pi_{10})}\log\frac{1-\delta}{\delta}.$$

3. For $0 \le \delta \le 1/2$ and $\pi_{10} = \frac{1-2\delta}{1-\delta}$

$$\limsup_{\varepsilon \downarrow 0} \frac{\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\sqrt{\varepsilon}} \le \sqrt{\frac{1-\delta}{\delta\cdot\pi_{10}}}(1-\pi_{10})(1-2\delta)\log\frac{1-\delta}{\delta}.$$

Note that Theorem 6 implies, in particular, the following.

1. For $0 \le \delta \le 1/2$ and $\frac{1-2\delta}{1-\delta} < \pi_{10} < 1$

$$\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) \asymp \varepsilon. \tag{59}$$

2. For $0 \le \delta \le 1/2$ and $\pi_{10} = \frac{1-2\delta}{1-\delta}$

$$\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) = O(\sqrt{\varepsilon})$$

and

$$\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) = \Omega(\varepsilon).$$

3. For $0 \le \delta \le 1/2$ and $0 < \pi_{10} < \frac{1-2\delta}{1-\delta}$

$$\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) = \Omega(\varepsilon).$$

It is the authors' conjecture that (59) holds for values of (δ, π_{10}) in the other two regions. Our proof technique, which sandwiches the entropy rate using (48), fails to give a nontrivial upper bound on $\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)$ in the region $0 \le \delta \le 1/2$ and

$0 < \pi_{10} < \frac{1-2\delta}{1-\delta}$. In this region, since the third branch in (55) does not approach 0 as $\varepsilon \downarrow 0$, the upper bound in (48) would not even imply the trivial fact that

$$\limsup_{\varepsilon \downarrow 0} \bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) \leq 0.$$

Note also that Theorem 6 implies

$$\begin{aligned} &\lim_{\pi_{10} \uparrow 1} \liminf_{\varepsilon \downarrow 0} \frac{\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\varepsilon} \\ &= \lim_{\pi_{10} \uparrow 1} \limsup_{\varepsilon \downarrow 0} \frac{\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\varepsilon} \\ &= (1-2\delta) \log \frac{1-\delta}{\delta} \end{aligned}$$

which is consistent with Theorem 5 (though does not imply it).

C. Entropy Rate When the Underlying Markov Chain Is Symmetric

When the clean source is a binary-symmetric Markov process $\pi_{10} = \pi_{01} = \pi$ with $\pi \leq 1/2$, (46) implies

$$\bar{H}(\pi, \pi, \delta) = E h_b(\beta_i(1) * \pi * \delta).$$

Equation (36) for this case implies

$$e^{I_1} = \frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi}$$

with $\alpha = \delta/(1-\delta)$. Thus, we get the first equation at the bottom of the page. Also, it follows from the first item in Corollary 3 that when $\delta \leq \frac{1}{2}(1-\sqrt{\max\{1-4\pi, 0\}})$, $h(I_1) + \log\frac{1-\delta}{\delta}$ is the lowest point of the support of l_i in the positive part of the real line (the fact that $h(I_1) + \log\frac{1-\delta}{\delta} \geq 0$ follows from (24) and the optimality of the "say-what-you-see" scheme, while the fact that this is the lowest point in the support of l_i in the positive part of the real line follows from the monotonicity of h and the definition of I_1). Now, the first part of Theorem 3 implies that $h(I_1) + \log\frac{1-\delta}{\delta} = I_1 + 2\log\frac{1-\delta}{\delta}$. Translating to the β_i domain, this implies that the lowest point of the support of $\beta_i(1)$ above $1/2$ is

$$\frac{e^{I_1+2\log\frac{1-\delta}{\delta}}}{1+e^{I_1+2\log\frac{1-\delta}{\delta}}} = \frac{\left(\frac{1-\delta}{\delta}\right)^2 e^{I_1}}{1+\left(\frac{1-\delta}{\delta}\right)^2 e^{I_1}}$$

implying, by symmetry, that the highest point of the support of $\beta_i(1)$ below $1/2$ is as shown in the second equation at the bottom of the page. Summarizing, we obtain the following result.

Theorem 7: For all $0 \leq \pi \leq 1/2$ and

$$0 \leq \delta \leq \frac{1}{2}(1-\sqrt{\max\{1-4\pi, 0\}})$$

we have (60) at the bottom of the page, where $\alpha = \delta/(1-\delta)$.

It is instructive to compare the bounds of Theorem 7 to those obtained by bounding the entropy rate from above by $H(Z_0 \,|\, Z_{-1})$ and from below by $H(Z_0 \,|\, X_{-1})$, which leads to

$$h_b(\pi * \delta) \leq \bar{H}(\pi, \pi, \delta) \leq h_b(\delta * \pi * \delta). \tag{61}$$

Evidently, the lower bound of Theorem 7 is always better than that in (61). The upper bound is better whenever

$$\frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} < \delta.$$

It will be seen later that there are asymptotic regimes (see, e.g., Corollary 5) where the bounds of Theorem 7 are tight while those in (61) are not. Similarly, there are regimes where the bounds of Theorem 7 would be tight, whereas bounds of the form $H(Z_0 \,|\, Z_{-k}^{-1}, X_{-(k+1)}) \leq \bar{H} \leq H(Z_0 \,|\, Z_{-k}^{-1})$, for fixed k, would not. For example, in the setting of Corollary 6 below, it can be shown (cf. discussion below) that any lower bound of the form $H(Z_0 \,|\, Z_{-k}^{-1}, X_{-(k+1)}) \leq \bar{H}$ would not give the right order.

The High SNR Regime:

Corollary 4: For $0 \leq \pi \leq 1/2$

$$\begin{aligned} \frac{1-2\pi}{1-\pi}\log\frac{1-\pi}{\pi} &\leq \liminf_{\delta \downarrow 0} \frac{\bar{H}(\pi,\pi,\delta) - h_b(\pi)}{\delta} \\ &\leq \limsup_{\delta \downarrow 0} \frac{\bar{H}(\pi,\pi,\delta) - h_b(\pi)}{\delta} \\ &\leq \frac{1-2\pi}{\pi}\log\frac{1-\pi}{\pi} \end{aligned} \tag{62}$$

in particular

$$\bar{H}(\pi,\pi,\delta) - h_b(\pi) \asymp \delta, \qquad \text{as } \delta \to 0. \tag{63}$$

$$\operatorname{ess\,inf} \beta_i = \frac{e^{I_1}}{1+e^{I_1}} = \frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}.$$

$$1 - \frac{e^{I_1+2\log\frac{1-\delta}{\delta}}}{1+e^{I_1+2\log\frac{1-\delta}{\delta}}} = \frac{\alpha^2}{\alpha^2+e^{I_1}} = \frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}.$$

$$\begin{aligned} h_b\left(\frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} * \pi * \delta\right) &\leq \bar{H}(\pi,\pi,\delta) \\ &\leq h_b\left(\frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} * \pi * \delta\right) \end{aligned} \tag{60}$$

It is easy to check that, for this regime, the bounds of (61) would give

$$(1-2\pi)\log\frac{1-\pi}{\pi} \le \liminf_{\delta\downarrow 0}\frac{\bar{H}(\pi,\pi,\delta)-h_b(\pi)}{\delta} \le \limsup_{\delta\downarrow 0}\frac{\bar{H}(\pi,\pi,\delta)-h_b(\pi)}{\delta} \le 2(1-2\pi)\log\frac{1-\pi}{\pi} \quad (64)$$

which is a slightly better upper bound and a slightly worse lower bound than in (62) (but implies (63) just the same). The bounds in (64) and (62) are consistent with the main result of [23] (see also [30]), which established

$$\frac{\bar{H}(\pi_{10},\pi_{01},\delta)-\bar{H}(\pi_{10},\pi_{01},0)}{\delta}\sim \upsilon(\pi_{10},\pi_{01})$$

explicitly identifying $\upsilon(\pi_{10},\pi_{01})$. It turns out that the upper bound in (64) coincides with $\upsilon(\pi_{10},\pi_{01})$.

The "Almost Memoryless" Regime:

In Corollary 5 and Corollary 6 to follow, the entropy rate is given in bits.

Corollary 5: For $0\le\delta\le 1/2$

$$\lim_{\varepsilon\downarrow 0}\frac{1-\bar{H}\left(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,\delta\right)}{\varepsilon^2}=\frac{2(1-2\delta)^4}{\ln 2}. \quad (65)$$

Note that $\bar{H}(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,0)=h_b(\frac{1}{2}-\varepsilon)$, so

$$\lim_{\varepsilon\downarrow 0}\frac{1-\bar{H}(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,0)}{\varepsilon^2}=\frac{2}{\ln 2}$$

(namely, (65) at $\delta=0$) would follow from a Taylor expansion of h_b around $1/2$. Equality (65) also trivially holds at $\delta=1/2$, as $\bar{H}(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,1/2)=1$. The simple intuition behind (65) is the following: when π is close to $1/2$, the support of $\beta_i(1)$ is highly concentrated around δ and $1-\delta$ (when $\pi=1/2$ it is exactly $\{\delta,1-\delta\}$). Thus,

$$1-\bar{H}\left(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,\delta\right) \sim 1-h_b((\delta\pm o(1))*(1/2-\varepsilon)*\delta) \sim \varepsilon^2\frac{2}{\ln 2}(1-2\delta)^4.$$

A formal proof can be found in the Appendix.

For this regime, the bounds of (61) would imply

$$1-\bar{H}\left(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,\delta\right)\asymp\varepsilon^2$$

but not give the constant characterized in (65).

The Low-SNR Regime:

Corollary 6: For $1/4\le\pi\le 1/2$

$$\frac{2}{\ln 2}\left[\frac{(4\pi-1)(1-2\pi)}{\pi}\right]^2 \le \liminf_{\varepsilon\to 0}\frac{1-\bar{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right)}{\varepsilon^4} \le \limsup_{\varepsilon\to 0}\frac{1-\bar{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right)}{\varepsilon^4}\le\frac{2}{\ln 2}\left[\frac{1-2\pi}{\pi}\right]^2. \quad (66)$$

In particular

$$1-\bar{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right)\asymp\varepsilon^4,\qquad \text{as } \varepsilon\to 0.$$

Note that the ratio between the upper and lower bounds in (66) approaches 1 as $\pi\uparrow 1/2$. Also, it can be shown that for any k, a bound of the form $H(Z_0\,|\,Z_{-k}^{-1},X_{-(k+1)})\le\bar{H}$ would lead to[1] $1-\bar{H}(\pi,\pi,\frac{1}{2}-\varepsilon)=O(\varepsilon^2)$, failing to capture the true ε^4 behavior.

In this regime, the bounds of (61) become

$$h_b(\pi*(1/2-\varepsilon))\le\bar{H}(\pi,\pi,\delta)\le h_b(\pi*(1/2-\varepsilon)*(1/2-\varepsilon)). \quad (67)$$

The upper bound implies, via a Taylor approximation (as in the above proof)

$$\liminf_{\varepsilon\to 0}\frac{1-\bar{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right)}{\varepsilon^4}\ge\frac{16(1-2\pi)^2}{\ln 2}$$

which gives a slightly better constant than the left-hand side of (66). The lower bound in (67), however, would only imply $1-\bar{H}(\pi,\pi,\frac{1}{2}-\varepsilon)\lesssim\varepsilon^2$.

VI. A Sufficient Condition for the Optimality of Singlet Decoding

In this section, we derive sufficient conditions for the optimality of singlet decoding for more general noise-free processes (not necessarily Markov), noise processes (not necessarily memoryless), and index sets. To minimize nonessential technicalities and simplify notation, we assume both that the components of the clean and noisy process take values in the same finite alphabet $\mathcal{X}$, and that the loss function is Hamming. It will be seen that under mild general conditions there exists a threshold such that if the noise level is below it the "say-what-you-see" scheme is optimal.

We start with the general setting of an arbitrarily distributed noise-free process $\{X_t\}$ corrupted by a noisy channel, i.e., there exists some process $\{N_t\}$ (the noise process, not necessarily of independent components) independent of $\{X_t\}$ and (deterministic) mappings $\{g_t\}$ such that the noisy observation process $\{Z_t\}$ is given by $Z_t=g_t(X_t,N_t)$, for all t. Observe first that, for all t,x_t, any finite index set T and $z(T)$

$$P(x_t|z(T))=\frac{P(x_t,z_t|z(T\setminus t))}{P(z_t|z(T\setminus t))}=\frac{P(x_t|z(T\setminus t))P(z_t|x_t,z(T\setminus t))}{P(z_t|z(T\setminus t))}$$

so that

$$\log\frac{P(X_t=a|z(T))}{P(X_t=b|z(T))}=\log\frac{P(z_t\,|\,X_t=a,z(T\setminus t))}{P(z_t\,|\,X_t=b,z(T\setminus t))}+\log\frac{P(X_t=a|z(T\setminus t))}{P(X_t=b|z(T\setminus t))}. \quad (68)$$

[1]This is true since $H(Z_0\,|\,Z_{-k}^{-1},X_{-(k+1)})\le H(Z_0|X_{-(k+1)})$ and $1-H(Z_0\,|\,X_{-(k+1)})=O(\varepsilon^2)$.

Note that for a memoryless channel C, (68) particularizes to

$$\log\frac{P(X_t=a|z(T))}{P(X_t=b|z(T))}=\log\frac{C(a,z_t)}{C(b,z_t)}+\log\frac{P(X_t=a|z(T\setminus t))}{P(X_t=b|z(T\setminus t))}.$$

By observation of (68), it is clear that an essentially necessary and sufficient condition for the optimal estimate of X_t to depend on $Z(T)$ only through Z_t is that, for all $a,b\in\mathcal{X}$, the sign of the right-hand side of (68) be determined by z_t, regardless of the value of $z(T\setminus t)$. This depends on the conditional distribution of X_t given $Z(T\setminus t)$ only through the values

$$\left\{\operatorname{ess\,sup}\log\frac{P(X_t=a|Z(T\setminus t))}{P(X_t=b|Z(T\setminus t))}\right\}_{a,b\in\mathcal{X}}.$$

While for the binary Markov chain these values were obtainable in closed form (Section IV), in general they are difficult to derive. They can, however, be bounded via the supports of the log likelihoods of the clean signal, leading to sufficient conditions for the optimality of singlet decoding. This is the approach taken in the following.

Returning to the general setting

$$\begin{aligned}&P(x_t=a\,|\,z(T\setminus t))\\&\quad=\sum_{x(T\backslash t)}P(x_t=a,x(T\backslash t)\,|\,z(T\backslash t))\\&\quad=\sum_{x(T\backslash t)}P(x_t=a\,|\,x(T\backslash t),z(T\backslash t))P(x(T\backslash t)\,|\,z(T\backslash t))\\&\quad=\sum_{x(T\backslash t)}P(x_t=a\,|\,x(T\backslash t))P(x(T\backslash t)\,|\,z(T\backslash t))\end{aligned}\tag{69}$$

where the last equality is due to the fact that $Z(T\backslash t)$ is a deterministic function of $X(T\backslash t)$ and $N(T\backslash t)$, so the independence of $\{X_t\}$ and $\{N_t\}$ implies the independence of X_t and $Z(T\backslash t)$ when conditioned on $X(T\backslash t)$. This leads to the following.

Lemma 1: For all $a,b\in\mathcal{X}$, and finite index set T

$$\max_{z(T\setminus t)}\frac{P(x_t=a\,|\,z(T\setminus t))}{P(x_t=b\,|\,z(T\setminus t))}\le\max_{x(T\setminus t)}\frac{P(x_t=a\,|\,x(T\setminus t))}{P(x_t=b\,|\,x(T\setminus t))}.\tag{70}$$

Proof: See the first equation at the bottom of the page, where the first equality follows from (69), and the inequality follows from Jensen's inequality (and convexity of $1/x$ for $x>0$). Thus, we get the second equation at the bottom of the page, implying (70) by the arbitrariness of $z(T\setminus t)$. □

Equipped with Lemma 1, we can obtain an easily verifiable sufficient condition for the optimality of singlet decoding in this general setting.

Theorem 8: Let T be an arbitrary index set, and suppose for each $z_t\in\mathcal{X}$ there exists $a=a(z_t)\in\mathcal{X}$ such that for all $b\neq a$

$$\begin{aligned}&\operatorname{ess\,inf}\frac{P(z_t\,|\,X_t=a(z_t),Z(T\setminus t))}{P(z_t\,|\,X_t=b,Z(T\setminus t))}\\&\qquad\ge\operatorname{ess\,sup}\frac{P(x_t=b\,|\,X(T\setminus t))}{P(x_t=a(z_t)\,|\,X(T\setminus t))}.\end{aligned}$$

Then an optimal estimate of X_t based on $Z(T)$ is

$$\hat{X}_t(Z(T))=a(Z_t).\tag{71}$$

Proof: By standard limiting and continuity arguments, it will suffice to assume T is a finite index set, and to show that if, for each $z_t\in\mathcal{X}$, there exists $a=a(z_t)\in\mathcal{X}$ such that for all $b\neq a$

$$\begin{aligned}&\min_{z(T\setminus t)}\frac{P(z_t\,|\,X_t=a,z(T\setminus t))}{P(z_t\,|\,X_t=b,z(T\setminus t))}\\&\qquad\ge\max_{x(T\setminus t)}\frac{P(x_t=b\,|\,x(T\setminus t))}{P(x_t=a\,|\,x(T\setminus t))};\end{aligned}\tag{72}$$

then the estimate in (71) is an optimal estimate of X_t based on $Z(T)$. To see this, note that if (72) holds then, for all $z(T)$, $a=a(z_t)\in\mathcal{X}$, and all $b\neq a$

$$\begin{aligned}\frac{P(z_t\,|\,X_t=a,z(T\setminus t))}{P(z_t\,|\,X_t=b,z(T\setminus t))}&\ge\min_{z'(T\setminus t)}\frac{P(z_t\,|\,X_t=a,z'(T\setminus t))}{P(z_t\,|\,X_t=b,z'(T\setminus t))}\\&\ge\max_{x(T\setminus t)}\frac{P(x_t=b\,|\,x(T\setminus t))}{P(x_t=a\,|\,x(T\setminus t))}\qquad(73)\\&\ge\frac{P(X_t=b\,|\,z(T\setminus t))}{P(X_t=a\,|\,z(T\setminus t))}\qquad(74)\end{aligned}$$

where (73) is due to (72) and (74) to (70). This implies, by (68), that $\log\frac{P(X_t=a\,|\,z(T))}{P(X_t=b|z(T))}\ge 0$ for all $b\neq a$ implying, in turn, that the optimal estimate of X_t based on $z(T)$ is $a=a(z_t)$. □

$$\begin{aligned}\frac{P(x_t=a\,|\,z(T\setminus t))}{P(x_t=b\,|\,z(T\setminus t))}&=\frac{\sum_{x(T\backslash t)}P(x_t=a\,|\,x(T\setminus t))P(x(T\setminus t)\,|\,z(T\setminus t))}{\sum_{x(T\backslash t)}P(x_t=b\,|\,x(T\setminus t))P(x(T\setminus t)\,|\,z(T\setminus t))}\\&=\sum_{x(T\backslash t)}\frac{P(x_t=b|x(T\setminus t))P(x(T\setminus t)|z(T\setminus t))}{\left[\sum_{x'(T\backslash t)}P(x_t=b\,|\,x'(T\setminus t))P(x'(T\setminus t)\,|\,z(T\setminus t))\right]}\frac{P(x_t=a\,|\,x(T\setminus t))}{P(x_t=b\,|\,x(T\setminus t))}\\&\ge\left[\sum_{x(T\backslash t)}\frac{P(x_t=b\,|\,x(T\setminus t))P(x(T\setminus t)\,|\,z(T\setminus t))}{\left[\sum_{x'(T\backslash t)}P(x_t=b\,|\,x'(T\setminus t))P(x'(T\setminus t)\,|\,z(T\setminus t))\right]}\frac{P(x_t=b\,|\,x(T\setminus t))}{P(x_t=a\,|\,x(T\setminus t))}\right]^{-1}\end{aligned}$$

$$\begin{aligned}\frac{P(x_t=b\,|\,z(T\setminus t))}{P(x_t=a\,|\,z(T\setminus t))}&\le\sum_{x(T\backslash t)}\frac{P(x_t=b\,|\,x(T\setminus t))P(x(T\setminus t)\,|\,z(T\setminus t))}{\left[\sum_{x'(T\backslash t)}P(x_t=b\,|\,x'(T\setminus t))P(x'(T\setminus t)\,|\,z(T\setminus t))\right]}\frac{P(x_t=b|x(T\setminus t))}{P(x_t=a\,|\,x(T\setminus t))}\\&\le\max_{x(T\backslash t)}\frac{P(x_t=b\,|\,x(T\setminus t))}{P(x_t=a\,|\,x(T\setminus t))}\end{aligned}$$

In what follows, we illustrate the use of Theorem 8 by deriving sufficient conditions for optimality of symbol-by-symbol filtering and denoising in a few specific settings.

A. The Memoryless Symmetric Channel

In this subsection, we assume the memoryless symmetric channel with error probability δ, uniformly distributed among the $|\mathcal{X}|-1$ erroneous symbols. For this case, we have for $b \neq a$ and $z_t = a$

$$\operatorname{ess\,inf} \frac{P(z_t \,|\, X_t = a, Z(T \setminus t))}{P(z_t \,|\, X_t = b, Z(T \setminus t))} = (|\mathcal{X}|-1)\frac{1-\delta}{\delta}$$

so Theorem 8 implies the following.

Corollary 7: If

$$(|\mathcal{X}|-1)\frac{1-\delta}{\delta} \geq \max_{a,b \in \mathcal{X}} \operatorname{ess\,sup} \frac{P(x_t = b \,|\, X(T \setminus t))}{P(x_t = a \,|\, X(T \setminus t))} \quad (75)$$

then $\hat{X}_t(Z(T)) = Z_t$ is an optimal estimate of X_t.

The right-hand side of (75) can readily be computed, or at least upper-bounded, for various processes and random fields, leading to a sufficient condition for the optimality of singlet decoding. A few examples follow.

Denoising a Gibbs Field: Let $T = \mathbb{Z}^d$ and $\mathcal{S}$ denote all finite subsets of T. Let $X(T)$ be the Gibbs field associated with the potential Φ [15], [19]. A potential is *summable* if

$$\|\Phi\|_t \triangleq \sum_{A \in \mathcal{S}, t \in A} \|\Phi_A\|_\infty < \infty, \qquad \forall t.$$

It is immediate from the definition of a Gibbs field that for all $a, b \in \mathcal{X}, t \in T$

$$\operatorname{ess\,sup} \frac{P(x_t = b \,|\, X(T \setminus t))}{P(x_t = a \,|\, X(T \setminus t))} \leq e^{2\|\Phi\|_t}. \quad (76)$$

Combining Corollary 7 with (76) gives the followng.

Corollary 8: The optimal estimate of X_t based on $Z(T)$ is Z_t if $\delta \leq \left(\frac{1}{|\mathcal{X}|-1} e^{2\|\Phi\|_t} + 1\right)^{-1}$. In particular, singlet decoding with "say-what-you-see" is an optimal denoiser whenever $\delta \leq \left(\frac{1}{|\mathcal{X}|-1} e^{2\|\Phi\|_{\max}} + 1\right)^{-1}$ where $\|\Phi\|_{\max} = \sup_{t \in T} \|\Phi\|_t$.

Note that $\|\Phi\|_{\max} < \infty$ for any spatially stationary (shift-invariant) Gibbs field with a summable potential. This includes, in particular, all Markov random fields (MRFs) with no restricted transitions (i.e., with the property that conditioned on any configuration of its neighborhood, all values at a given site have positive probability). The "say-what-you-see" denoiser is optimal for all such fields when δ is sufficiently small. Finally, we note that Corollary 8 implies that for fixed $\delta < \frac{|\mathcal{X}|-1}{|\mathcal{X}|}$ and a potential satisfying $\|\Phi\|_{\max} < \infty$, singlet decoding is optimal denoising for the field with potential $\beta\Phi$ whenever $\delta \leq \left(\frac{1}{|\mathcal{X}|-1} e^{2\beta\|\Phi\|_{\max}} + 1\right)^{-1}$ or, in other words, whenever

$$\beta \leq \frac{\log[(1/\delta - 1)(|\mathcal{X}|-1)]}{2\|\Phi\|_{\max}}$$

(i.e., at sufficiently high temperatures [15], [19]).

Filtering and Denoising a Stationary Source: If $X(T)$, $T = \mathbb{Z}$, is a stationary process then by defining

$$R(X(T)) \triangleq \max_{a,b \in \mathcal{X}} \operatorname{ess\,sup} \frac{P\left(X_0 = b \,|\, X_{-\infty}^{-1}\right)}{P\left(X_0 = a \,|\, X_{-\infty}^{-1}\right)}$$

and

$$S(X(T)) \triangleq \max_{a,b \in \mathcal{X}} \operatorname{ess\,sup} \frac{P\left(X_0 = b \,|\, X_{-\infty}^{-1}, X_1^{\infty}\right)}{P\left(X_0 = a \,|\, X_{-\infty}^{-1}, X_1^{\infty}\right)}$$

Corollary 7 implies the following.

Corollary 9: The "say-what-you-see" scheme is an optimal filter if $\delta \leq \left(\frac{1}{|\mathcal{X}|-1} R(X(T)) + 1\right)^{-1}$ and an optimal denoiser if $\delta \leq \left(\frac{1}{|\mathcal{X}|-1} S(X(T)) + 1\right)^{-1}$.

Note, in particular, that if $X(T)$ is a kth-order Markov source with no restricted sequences then

$$R(X(T)) = \max_{a,b,x_{-k}^{-1}} \frac{P\left(X_0 = a \,|\, X_{-k}^{-1} = x_{-k}^{-1}\right)}{P\left(X_0 = b \,|\, X_{-k}^{-1} = x_{-k}^{-1}\right)} > 0 \quad \text{and}$$

$$S(X(T)) = \max_{a,b,x_{-k}^{-1},x_1^k} \frac{P\left(X_0 = a \,|\, X_{-k}^{-1} = x_{-k}^{-1}, X_1^k = x_1^k\right)}{P\left(X_0 = b \,|\, X_{-k}^{-1} = x_{-k}^{-1}, X_1^k = x_1^k\right)} > 0$$

so the "say-what-you-see" scheme is optimal for all sufficiently small δ.

To get a feel for the tightness of these conditions, consider the symmetric binary Markov chain for which the optimality of the "say-what-you-see" scheme has been characterized in Corollary 3. Assuming $\pi \leq 1/2$, we have $R(X(T)) = (1-\pi)/\pi$ and $S(X(T)) = [(1-\pi)/\pi]^2$, so Corollary 9 would imply for this case that the "say-what-you-see" scheme is an optimal filter whenever $\delta \leq \pi$, and is an optimal denoiser whenever $\delta \leq \frac{\pi^2}{1-2\pi+2\pi^2}$. The solid and dashed curves in Fig. 1 display the curve characterizing the whole region of optimality of the singlet decoder for the filtering problem (from Corollary 3), together with the curve associated with the sufficient condition implied by Corollary 9, namely, the straight line $\delta = \pi$. Fig. 2 displays the analogous curves for the denoising problems. The region $\delta \leq \pi$ can be understood as the condition for optimality of singlet filtering when allowing a genie-aided filter to observe the clean symbol one step back. Similarly, the $\delta \leq \frac{\pi^2}{1-2\pi+2\pi^2}$ region is obtained by allowing the genie-aided denoiser to observe the clean symbols from both sides.

Denoising a Process or Field That Can Be Represented as Output of Discrete Memoryless Channel (DMC) (Hidden Markov Processes): Suppose that the noiseless process $X(T)$ was generated (or can be represented) as the output of a DMC whose input is some other process $U(T)$, which we assume for simplicity has components taking values in the same finite alphabet $\mathcal{X}$. Denote the DMC by W, i.e., $W(a \,|\, u) = \Pr(X_t = a \,|\, U_t = u)$. Thus, we have, assuming first T finite

$$P(X_t = a \,|\, X(T \setminus t) = x(T \setminus t)) = \sum_u P(U_t = u \,|\, X(T \setminus t) = x(T \setminus t)) W(a \,|\, u).$$

Consequently, for $a, b \in \mathcal{X}$, reasoning similarly as in the proof of Lemma 1, we obtain

$$\frac{P(X_t = a \,|\, X(T \setminus t) = x(T \setminus t))}{P(X_t = b \,|\, X(T \setminus t) = x(T \setminus t))} \leq \max_{u \in \mathcal{U}} \frac{W(a \,|\, u)}{W(b \,|\, u)}$$

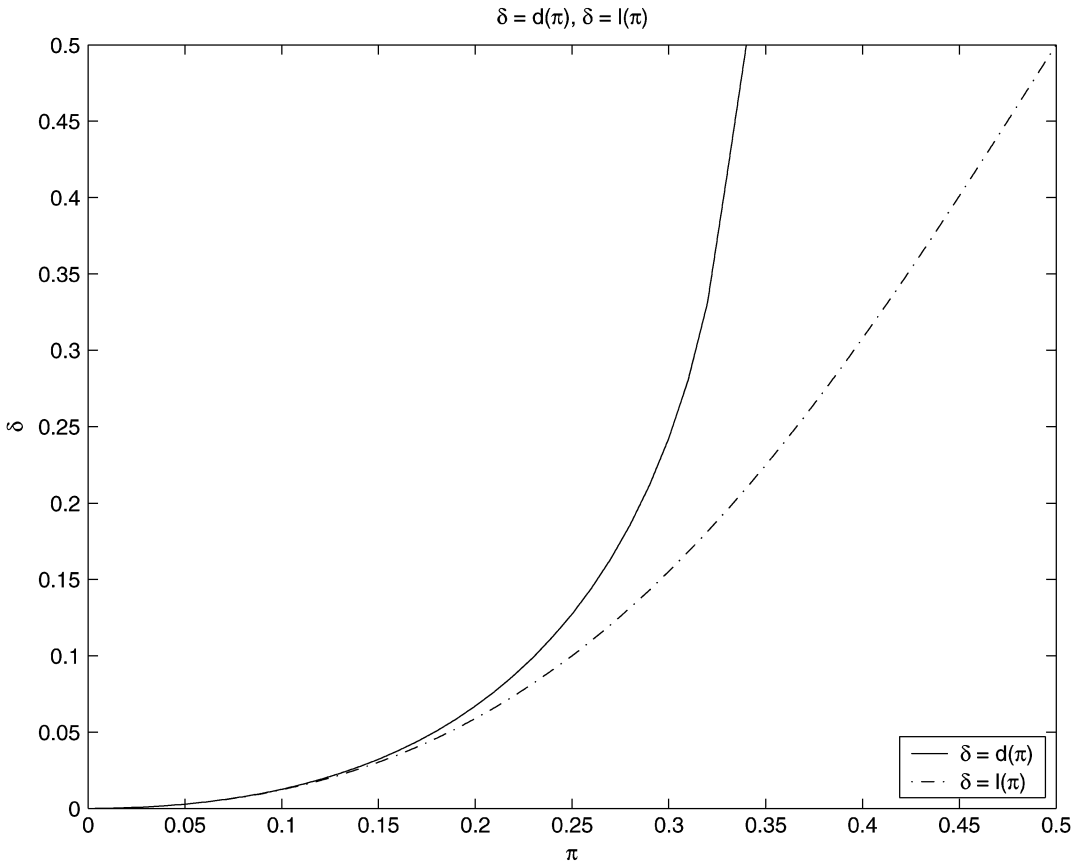

Fig. 2. Optimality region for singlet denoising: Solid line is the curve $d(\pi) = \frac{1}{2}(1 - \sqrt{\max\{1 - 4(\frac{\pi}{1-\pi})^2, 0\}})$ giving the precise region $\delta \leq d(\pi)$. Dashed line is the $l(\pi) = \frac{\pi^2}{1-2\pi+2\pi^2}$ curve associated with the sufficient condition in Corollary 9, $\delta \leq l(\pi)$.

for all $x(T \setminus t)$. By a standard limiting argument, we obtain for an arbitrary index set T

$$\operatorname{ess\,sup} \frac{P(X_t = a \,|\, X(T \setminus t))}{P(X_t = b \,|\, X(T \setminus t))} \leq \max_{u \in \mathcal{U}} \frac{W(a \,|\, u)}{W(b \,|\, u)}.$$

Combined with Corollary 7, this gives the following.

Corollary 10: If $X(T)$ is output of DMC W (for some input $U(T)$) then the "say-what-you-see" scheme is an optimal denoiser of $Z(T)$ provided

$$\delta \leq \left(\frac{1}{|\mathcal{X}| - 1} \max_{a \neq b, u \in \mathcal{U}} \frac{W(a \,|\, u)}{W(b \,|\, u)} + 1 \right)^{-1}.$$

Corollary 10 implies, in particular, for the case where the channel W is symmetric with parameter ε, that the "say-what-you-see" scheme is an optimal denoiser whenever $\delta \leq \varepsilon$. Note that $X(T)$ being the output of such a channel W is equivalent to its satisfying the Shannon lower bound (cf. [4], [7, Example 13.6]) with equality (under Hamming loss) for distortion levels $\leq \varepsilon$. It thus follows that any source or random field whose rate distortion function at distortion level D is given by the Shannon lower bound (cf., e.g., [20], [17], [18], [38] for examples of processes and fields with this property) is optimally denoised by the "say-what-you-see" scheme whenever $\delta \leq D$.

Filtering an Autoregressive Source: Let $T = \mathbb{Z}, \mathcal{X} = \{0, 1, \ldots, M-1\}$, and suppose the noiseless process can be represented by

$$X_t = g_t(X^{t-1}) \oplus W_t \tag{77}$$

where $\oplus$ denotes modulo-M addition and $\{W_t\}$ are independent and identically distributed (i.i.d.) (g_t and W_t take values in $\mathcal{X}$). For this process

$$P(X_t = a \,|\, X^{t-1} = x^{t-1}) = \Pr(W_t \oplus g_t(x^{t-1}) = a)$$

so, for $a \neq b$

$$\begin{aligned} \frac{P(X_t = a \,|\, X^{t-1} = x^{t-1})}{P(X_t = b \,|\, X^{t-1} = x^{t-1})} &\leq \max_{m \in \mathcal{X}} \frac{P(W_t = m)}{P(W_t = (b-a) \oplus m)} \\ &\leq \max_{a \neq 0, m \in \mathcal{X}} \frac{P(W_t = m)}{P(W_t = a \oplus m)} \\ &= \frac{\max_m P(W_t = m)}{\min_m P(W_t = m)}. \end{aligned} \tag{78}$$

Applied to this setting, and combined with (78), Corollary 7 gives the following.

Corollary 11: Let $\{X_t\}$ be given by (77), where $\{W_t\}$ is an i.i.d. sequence. The "say-what-you-see" scheme is an optimal filter provided

$$\delta \leq \left(\frac{1}{|\mathcal{X}| - 1} \frac{\max_m P(W_t = m)}{\min_m P(W_t = m)} + 1 \right)^{-1}.$$

Note, in particular, that for the commonly occurring case where the innovations are symmetric, namely,

$$P(W_t = a) = \begin{cases} 1-\varepsilon, & \text{for } a = 0 \\ \varepsilon/(|\mathcal{X}|-1), & \text{for } a \neq 0. \end{cases}$$

Corollary (11) implies optimality of the "say-what-you-see" filter provided $\delta \leq \varepsilon$.

B. Channels With Memory

The Gilbert–Elliot Channel: Assume $T = \mathbb{Z}$ and that $Z(T)$ is the noisy version of $X(T)$ when corrupted by the Gilbert–Elliot channel [29]. Let $S(T)$, with components in $\{B, G\}$, denote the first-order Markov channel state process, and let δ_g, δ_b denote the crossover probabilities associated, respectively, with the good and bad states, where $0 \leq \delta_g \leq \delta_b \leq 1/2$. In this case

$$\begin{aligned} &P(z_t \,|\, X_t = a, Z(T\setminus t)) \\ &\quad = \sum_{s_t} P(z_t \,|\, X_t = a, s_t, Z(T\setminus t)) P(s_t \,|\, X_t = a, Z(T\setminus t)) \\ &\quad = \sum_{s_t} P(z_t \,|\, X_t = a, s_t) P(s_t \,|\, X_t = a, Z(T\setminus t)) \end{aligned}$$

and thus we obtain for $a \neq b$ and $z_t = a$

$$\begin{aligned} \operatorname{ess\,inf} \frac{P(z_t \,|\, X_t = a, Z(T\setminus t))}{P(z_t \,|\, X_t = b, Z(T\setminus t))} &\geq \frac{\min_{s_t \in \{G,B\}} P(z_t \,|\, X_t = a, s_t)}{\max_{s_t \in \{G,B\}} P(z_t \,|\, X_t = b, s_t)} \\ &= \frac{1-\delta_b}{\delta_b}. \end{aligned} \tag{79}$$

A similar argument implies an inequality like (79), where in the left-hand side the conditioning is on the one-sided $Z_{-\infty}^{t-1}$ instead of on $Z(T \setminus t)$. Combining (79) (and its analogue for the one-sided conditioning) with Theorem 8 gives the following.

Corollary 12: The "say-what-you-see" scheme is an optimal denoiser (and a fortiori an optimal filter) for the Gilbert–Elliot channel if

$$\frac{1-\delta_b}{\delta_b} \geq \max_{b \neq a} \operatorname{ess\,sup} \frac{P(x_t = b \,|\, X(T\setminus t))}{P(x_t = a \,|\, X(T\setminus t))}. \tag{80}$$

It is also an optimal filter provided

$$\frac{1-\delta_b}{\delta_b} \geq \max_{b \neq a} \operatorname{ess\,sup} \frac{P(x_t = b \,|\, X_{-\infty}^{t-1})}{P(x_t = a \,|\, X_{-\infty}^{t-1})}. \tag{81}$$

Arbitrarily Distributed State Process: A first point to note is that the derivation of Corollary 12 did not depend in any way on the distribution of the state process. Also, the conclusion regarding the optimality condition for denoising did not rely on the fact that $T = \mathbb{Z}$ and would hold for any index set T. It is also readily checked that the binary alphabet can be replaced by any finite alphabet where δ_g, δ_b would denote the crossover parameters indexing the symmetric channels associated, respectively, with the good and bad states (and the left-hand side of (80) and (81) would be replaced by $(|\mathcal{X}|-1)\frac{1-\delta_b}{\delta_b}$). Finally, the state space need not be restricted to only two states; in general, each state will index a channel with a different parameter, in which case the definition of δ_b would be extended to $\delta_b = \max_{s \in \mathcal{S}} \delta_s$, $\mathcal{S}$ being the state space.

In this generality, of an arbitrarily distributed state process, a general state space, and a finite alphabet of any size, all the results of the previous subsection (namely, Corollaries 7 through 11) carry over with δ_b replacing δ.

Other channels with memory abound for which

$$\frac{P(z_t \,|\, X_t = a, Z(T \setminus t))}{P(z_t \,|\, X_t = b, Z(T \setminus t))}$$

can be lower-bounded leading, via Theorem 8, to sufficient conditions for the optimality of symbol-by-symbol schemes in denoising and filtering of various other processes and fields.

VII. Large Deviations Performance of the Optimal Filter

For concreteness, assume here $T = \mathbb{Z}$, that the components of $X(T)$ take values in the finite alphabet $\mathcal{X}$, and that $Z(T)$ is the output of a DMC C whose input is $X(T)$ with channel output alphabet $\mathcal{Z}$.

Using standard large deviations (LD) theory [11] or the method of types [8], [7], it is straightforward to show that, for every $f : \mathcal{Z} \to \mathcal{X}$ and $x^n \in \mathcal{X}^n$

$$\Pr\left(\frac{1}{n} \sum_{t=1}^{n} \Lambda(X_t, f(Z_t)) \geq d \,\middle|\, X^n = x^n \right) \approx \exp(-n J(P_{x^n}, d))$$

where

$$J(P, d) = \min_{Q : E_{P \otimes Q} \Lambda(X, f(Z)) \geq d} D(Q \| C | P),$$

with $D(Q\|C|P)$ denoting the conditional divergence (cf., e.g., [8, Sec. 2]) between conditional distributions (channels) Q and C (true channel) conditioned on a channel input distribution P, and $E_{P \otimes Q}$ denotes expectation assuming that $X \sim P$ and that Z is the output of the channel Q whose input is X.

More precisely, it can be shown (cf., e.g., [9], [28], [37], [36] for proofs of results in this spirit) that for any individual sequence $\boldsymbol{x} = \{x_t\}$

$$\left| -\frac{1}{n} \log \Pr\left(\frac{1}{n} \sum_{t=1}^{n} \Lambda(X_t, f(Z_t)) \geq d \,\middle|\, X^n = x^n \right) - J(P_{x^n}, d) \right| \longrightarrow 0. \tag{82}$$

This exponent can also be given in the form (cf., e.g., [9, Proposition 1])

$$J(P, d) = \sup_{\lambda \in \mathbb{R}} \left\{ \lambda d - \sum_{a \in \mathcal{X}} \left[\log \sum_{b \in \mathcal{Z}} e^{\lambda \Lambda(a, f(b))} C(a, b) \right] P(a) \right\}.$$

It follows from (82) that if the empirical measure associated with $X(T)$ satisfies a large deviations principle (LDP) [11] with the rate function I then

$$-\frac{1}{n}\log\Pr\left(\frac{1}{n}\sum_{t=1}^{n}\Lambda(X_t,f(Z_t))\geq d\right) \longrightarrow \min_{P\in\mathcal{M}(\mathcal{X})}\left[I(P)+\min_{Q:E_{P\otimes Q}\Lambda(X,f(Z))\geq d}D(Q\|C|P)\right]. \quad (83)$$

This gives a single-letter characterization of the "error exponent" associated with the optimal (in expectation sense) filter for all cases characterized in previous sections where the optimal scheme is a symbol-by-symbol filter and the underlying noise-free process satisfies an LDP with a known rate function (cf. [11] for wide range of processes for which this is the case). In particular, the error exponent is given by the right-hand side of (83), with f being the filtering function associated with the optimal scheme.

VIII. Conclusion and Open Directions

The goal of this work was to identify situations where optimal estimation of each signal component when observing a discrete signal corrupted by noise depends on available observations only via the noisy observation of that component. We obtained easily verifiable sufficient conditions for the optimality of such "symbol-by-symbol" schemes. For a binary Markov process corrupted by a general memoryless channel, an explicit necessary and sufficient condition was obtained. The condition for the optimality of singlet decoding was seen to depend on the channel only through the support of the Radon–Nikodym derivative between the distributions of the channel output associated with the two inputs (and, in fact, depend on this support only through its upper and lower ends). It was also observed that the large deviations behavior of a singlet filter can be easily characterized (provided the large deviations behavior of the noise-free process is known) when the noise is memoryless. Thus, the large deviations performance of the optimal scheme is characterized whenever it is a singlet decoder.

Characterization of the singlet filtering region for the corrupted binary Markov chain involved the computation of the lower and upper endpoints of the support of the distribution of the clean symbol conditioned on its noisy observation and noisy past. These bounds were seen to lead to new bounds on the entropy rate of the noisy observation process. The latter were shown to be tight and to characterize the precise behavior of the entropy rate in various asymptotic regimes. Further exploration of this approach to characterize the entropy rate in other asymptotic regimes, for larger alphabets, etc., is deferred to future work.

Two additional future research directions arise in the context of the LD performance analysis for a singlet scheme in Section VII. The first concerns the question of whether the expected-sense optimality of a singlet decoder (the criterion considered in this work) implies its optimality under the LD criterion as well. More generally, can conditions for the optimality of singlet decoding in the LD sense be obtained? The second interesting direction regards the characterization of the LD performance of a scheme which is not singlet. Even the characterization of the LD performance of a sliding window scheme of length 2 is currently open.

It should be noted that a singlet decoder is a sliding-window scheme of length 1. A natural extension of the characterization of optimal singlet decoding would be, for a given $l>1$, a characterization of conditions under which the optimal filter or denoiser is a sliding-window scheme of length l.

Finally, it may be interesting to see whether a meaningful analogue of the notion of a singlet scheme can be found for the continuous-time setting (say, for a Markov source corrupted by white noise, as in the setting of [40]), and whether there exist nontrivial situations where such singlet schemes are optimal.

Appendix

A. Proof of Theorem 2

The proof is similar to that of Theorem 1. Suppose that $C_Q\times C_{Q_r}\subseteq R_f$. The fact that

$$P(\beta_{i-1}\in C_Q)=P(\gamma_{i+1}\in C_{Q_r})=1$$

implies that

$$P(f(b)\in\hat{X}(G_b(\beta_{i-1},\gamma_{i+1}))\forall b\in\mathcal{Z}) = P((\beta_{i-1},\gamma_{i+1})\in R_f)=1.$$

Consequently,

$$1=P(f(Z_i)\in\hat{X}(G_{Z_i}(\beta_{i-1},\gamma_{i+1})))=P(f(Z_i)\in\hat{X}(\eta_i))$$

establishing optimality by (20).

Conversely, suppose that $C_Q\times C_{Q_r}\not\subseteq R_f$. Then, since[2]

$$\text{Support}(\beta_{i-1},\gamma_{i+1})=C_Q\times C_{Q_r}$$

there exists $J\subseteq\mathcal{M}(\mathcal{X})\times\mathcal{M}(\mathcal{X})$ such that $J\cap R_f=\emptyset$ and $P((\beta_{i-1},\gamma_{i+1})\in J)>0$. This implies that

$$P((\beta_{i-1},\gamma_{i+1})\in R_f) = P(f(b)\in\hat{X}(G_b(\beta_{i-1},\gamma_{i+1}))\forall b\in\mathcal{Z})<1$$

which implies the existence of $b\in\mathcal{Z}$ with

$$P(f(b)\in\hat{X}(G_b(\beta_{i-1},\gamma_{i+1})))<1$$

implying, in turn, the existence of $a\in\mathcal{X}$ such that

$$P(f(b)\in\hat{X}(G_b(\beta_{i-1},\gamma_{i+1}))\,|\,X_i=a)<1. \quad (A1)$$

Now, Z_i,β_{i-1} and γ_{i+1} are conditionally independent given X_i, and therefore,

$$\begin{aligned}P(f(Z_i)\in\hat{X}(\eta_i)\,|\,X_i=a)&=P(f(Z_i)\in\hat{X}(G_b(\beta_{i-1},\gamma_{i+1}))\,|\,X_i=a)\\&=\sum_{b'\in\mathcal{Z}}P(f(b')\in\hat{X}(G_{b'}(\beta_{i-1},\gamma_{i+1}))\,|\,X_i=a)C(a,b').\end{aligned} \quad (A2)$$

Inequality (A1), combined with (A2) and the fact that $C(a,b)>0$, leads to $P(f(Z_i)\in\hat{X}(\eta_i)|X_i=a)<1$, implying $P(f(Z_i)\in\hat{X}(\eta_i))<1$ and establishing the fact that (20) is not satisfied by $\hat{X}_i(Z_{-\infty}^{\infty})=f(Z_i)$. □

[2]The fact that $Support(\beta_{i-1},\gamma_{i+1})=C_Q\times C_{Q_r}$ is an immediate consequence of the assumed positivity of the kernel governing the noiseless process.

B. Proofs of Results From Section V

Proof of Theorem 5: The first equality in (52) follows trivially by symmetry, thus we turn to establish the second equality. Substituting into (36) we obtain (A3) at the bottom of the page, where $\alpha = \frac{1-\delta}{\delta}$. It follows from (A3) (using a first-order McLaurin approximation to $\sqrt{1+\varepsilon}$) that

$$\lim_{\varepsilon \downarrow 0} \frac{e^{I_2}}{a(\varepsilon)^{\frac{1-\delta}{\delta}}} = 1. \tag{A4}$$

It thus follows from the upper bound in (50) that for $\eta > 0$ and all sufficiently small $\varepsilon > 0$

$$\bar{H}(1-\varepsilon, a(\varepsilon), \delta) \leq h_b([(1+\eta)a(\varepsilon)] * \delta) \tag{A5}$$

and, consequently

$$\bar{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta) \leq h_b([1+\eta)a(\varepsilon)] * \delta) - h_b(\delta). \tag{A6}$$

Applying a Taylor's expansion around δ and noting that $[(1+\eta)a(\varepsilon)] * \delta - \delta = [(1+\eta)a(\varepsilon)](1-2\delta)$ gives

$$\begin{aligned} &h_b([(1+\eta)a(\varepsilon)] * \delta) - h_b(\delta) \\ &\quad = (1+\eta)a(\varepsilon)(1-2\delta)h_b'(\delta) + o(a(\varepsilon)) &\text{(A7)} \\ &\quad = (1+\eta)a(\varepsilon)(1-2\delta)\log\frac{1-\delta}{\delta} + o(a(\varepsilon)). &\text{(A8)} \end{aligned}$$

Combining (A6) with (A8) gives

$$\limsup_{\varepsilon \downarrow 0} \frac{\bar{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta)}{a(\varepsilon)} \leq (1+\eta)(1-2\delta)\log\frac{1-\delta}{\delta} \tag{A9}$$

implying

$$\limsup_{\varepsilon \downarrow 0} \frac{\bar{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta)}{a(\varepsilon)} \leq (1-2\delta)\log\frac{1-\delta}{\delta} \tag{A10}$$

by the arbitrariness of η. The inequality

$$\liminf_{\varepsilon \downarrow 0} \frac{\bar{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta)}{a(\varepsilon)} \geq (1-2\delta)\log\frac{1-\delta}{\delta} \tag{A11}$$

is established similarly. □

Proof of Theorem 6: Since the claim trivially holds for $\delta = 1/2$ assume $0 \leq \delta < 1/2$. From the left inequality in (48) and (57) it follows that for fixed $\eta > 0$ and all sufficiently small $\varepsilon > 0$

$$\begin{aligned} \bar{H}(\pi_{10}, \varepsilon, \delta) &\geq h_b\left(\left\{\left[\frac{\delta\varepsilon}{(1-\delta)-\delta(1-\pi_{10})}(1-\pi_{10}) + \varepsilon\right](1-\eta)\right\} * \delta\right) &\text{(A12)} \\ &= h_b\left(\left\{\left[\frac{(1-\delta)(1-\eta)}{(1-\delta)-\delta(1-\pi_{10})}\right]\varepsilon\right\} * \delta\right). &\text{(A13)} \end{aligned}$$

Applying a Taylor's expansion to h_b around δ, similarly as in (A8), (A13) gives

$$\begin{aligned} &\bar{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) \\ &\quad \geq \frac{(1-\delta)(1-\eta)}{(1-\delta)-\delta(1-\pi_{10})}\varepsilon(1-2\delta)\log\frac{1-\delta}{\delta} + o(\varepsilon) &\text{(A14)} \end{aligned}$$

implying (58) by the arbitrariness of η.

The second item is proven similarly (using (54) instead of (57)).

For the third item note that from the right inequality in (48) and (54) it follows that when $\pi_{10} = \frac{1-2\delta}{1-\delta}$, for fixed $\eta > 0$ and all sufficiently small $\varepsilon > 0$

$$\bar{H}(\pi_{10}, \varepsilon, \delta) \leq h_b\left(\left[\sqrt{\frac{1-\delta}{\delta\pi_{10}}}\sqrt{\varepsilon}(1-\pi_{10})(1+\eta)\right] * \delta\right). \tag{A15}$$

The claim now follows analogously as in proof of previous items via the Taylor approximation and the arbitrariness of η. □

Proof of Corollary 4: Using the Taylor expansion for $\sqrt{1+\varepsilon}$ gives (A16) and (A17) at the bottom of the page. Since $\alpha \sim \delta$ as $\delta \downarrow 0$ it follows from (A16) that for fixed $\eta > 0$ and all sufficiently small δ we get (A18)–(A23) at the top of the following page, implying the left inequality in (62) via (60) and the arbitrariness of η. The right inequality in (62) follows from (A17) in an analogous way. □

Proof of Corollary 5: At $\pi = 1/2$ we have (A24) at the top of the following page, where $\alpha = \delta/(1-\delta)$. The claim now follows by continuity of the expressions in (A24) at $\pi = 1/2$, the relationship

$$\begin{aligned} \frac{1}{2} - (\delta+\xi) * (\frac{1}{2} - \varepsilon) * \delta &= \varepsilon(1 - 2(2\delta - 2\delta^2 + \xi(1-2\delta))) \\ &= \varepsilon(1-2\delta)^2(1+O(\xi)) \end{aligned}$$

Theorem 7, and the fact that

$$h_b(1/2 - \varepsilon) = 1 - (2/\ln 2)\varepsilon^2 + o(\varepsilon^2).$$ □

$$e^{I_2} = \frac{-1+\alpha+a(\varepsilon)-\alpha(1-\varepsilon)+\sqrt{4\alpha a(\varepsilon)(1-\varepsilon)+(1-\alpha-a(\varepsilon)+\alpha(1-\varepsilon))^2}}{2(1-\varepsilon)} \tag{A3}$$

$$\frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} \sim \alpha\frac{\pi}{1-\pi}, \qquad \text{as } \alpha \downarrow 0 \tag{A16}$$

and

$$\frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} \sim \alpha\frac{1-\pi}{\pi}, \qquad \text{as } \alpha \downarrow 0. \tag{A17}$$

$$h_b\left(\frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} * \pi * \delta\right) \tag{A18}$$

$$\geq h_b\left(\left[(1-\eta)\frac{\pi}{1-\pi}\delta\right] * \pi * \delta\right) \tag{A19}$$

$$\geq h_b\left(\left[(1-2\eta)\left[\frac{\pi}{1-\pi}+1\right]\delta\right] * \pi\right) \tag{A20}$$

$$= h_b\left(\pi+(1-2\pi)(1-2\eta)\left[\frac{\pi}{1-\pi}+1\right]\delta\right) \tag{A21}$$

$$= h_b(\pi)+(1-2\pi)(1-2\eta)\left[\frac{\pi}{1-\pi}+1\right]\delta h_b'(\pi)+o(\delta) \tag{A22}$$

$$= h_b(\pi)+(1-2\pi)(1-2\eta)\left[\frac{\pi}{1-\pi}+1\right]\delta\log\frac{1-\pi}{\pi}+o(\delta) \tag{A23}$$

$$\frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} = \frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} = \delta \tag{A24}$$

$$\frac{1}{2}-\frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} \sim \frac{1}{2\pi}\varepsilon \tag{A25}$$

$$\frac{1}{2}-\frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} \sim \left(2-\frac{1}{2\pi}\right)\varepsilon. \tag{A26}$$

Proof of Corollary 6: Since

$$\bar{H}(\pi,\pi,\frac{1}{2}-\varepsilon) = \bar{H}(\pi,\pi,\frac{1}{2}+\varepsilon)$$

we may assume the limits in (66) are taken along $\varepsilon \downarrow 0$. Letting $\delta = 1/2-\varepsilon$, it is straightforward to show (A25) at the top of the page using a Taylor expansion, and (A26), also at the top of the page. It thus follows from Theorem 7 that for every $1/4 \leq \pi \leq 1/2$ and $\varepsilon > 0$

$$1-h_b\left(\left[\frac{1}{2}-\left(2-\frac{1}{2\pi}\right)\varepsilon\right] * \pi * \left(\frac{1}{2}-\varepsilon\right)\right) \tag{A27}$$

$$\lesssim 1-\bar{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right) \tag{A28}$$

$$\lesssim 1-h_b\left(\left[\frac{1}{2}-\frac{1}{2\pi}\varepsilon\right] * \pi * \left(\frac{1}{2}-\varepsilon\right)\right). \tag{A29}$$

The claim now follows by

$$\left[\frac{1}{2}-\left(2-\frac{1}{2\pi}\right)\varepsilon\right] * \pi * \left(\frac{1}{2}-\varepsilon\right) = \frac{1}{2}-\frac{(4\pi-1)(1-2\pi)}{\pi}\varepsilon^2 \tag{A30}$$

$$\left[\frac{1}{2}-\frac{1}{2\pi}\varepsilon\right] * \pi * \left(\frac{1}{2}-\varepsilon\right) = \frac{1}{2}-\frac{1-2\pi}{\pi}\varepsilon^2 \tag{A31}$$

and the fact that $h_b(1/2-\varepsilon) = 1-(2/\ln 2)\varepsilon^2+o(\varepsilon^2)$. □

REFERENCES

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

[2] L. Arnold, L. Demetrius, and M. Gundlach, "Evolutionary formalism for products of positive random matrices," *Ann. Appl. Probab.*, no. 4, pp. 859–901, 1994.

[3] R. Atar and O. Zeitouni, "Exponential stability for nonlinear filtering," *Annales de l'Institut H. Poincare Probabilites et Statistique*, no. 33, pp. 697–725, 1997.

[4] T. Berger, *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

[5] A. Bhatt, A. Budhiraja, and R. Karandikar, "Markov property and ergodicity of the nonlinear filter," *SIAM J. Control and Optimiz.*, no. 39, pp. 928–949, 2000.

[6] D. Blackwell, "The entropy of functions of finite-state Markov chains," in *Trans. 1st Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, Prague, Czechoslovakia, 1957, pp. 13–20.

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

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

[9] A. Dembo and I. Kontoyiannis, "The asymptotics of waiting times between stationary processes, allowing distortion," *Ann. Appl. Probab.*, no. 9, pp. 413–429, 1999.

[10] A. Dembo and T. Weissman, "Universal denoising for the finite-input-continuous-output channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1507–1517, Apr. 2005.

[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1998.

[12] J. L. Devore, "A note on the observation of a Markov source through a noisy channel," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 6, pp. 762–764, Nov. 1974.

[13] A. W. Drake, "Observation of a Markov source through a noisy channel," in *Proc. IEEE Symp. Signal Transmission and Processing*, Columbia Univ., New York, 1965, pp. 12–18.

[14] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1518–1569, Jun. 2002.

[15] H. O. Georgii, *Gibbs Measures and Phase Transitions*. Berlin, Germany: Walter de Gruyter, 1988.

[16] A. Goldsmith and P. Varaiya, "Capacity, mutual information, and coding for finite state Markov channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 868–886, May 1996.

[17] R. M. Gray, "Information rates of autoregressive processes," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 412–421, Jul. 1970.

[18] ——, "Rate distortion functions for finite-state finite-alphabet Markov sources," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 2, pp. 127–134, Mar. 1971.

[19] X. Guyon, *Random Fields on a Network*. New York: Springer-Verlag, 1995.

[20] B. E. Hajek and T. Berger, "A decomposition theorem for binary Markov random fields," *Ann. Probab.*, no. 15, pp. 1112–1125, 1987.

[21] J. Hannan, "Approximation to Bayes risk in repeated play," in *Contributions to the Theory of Games, Ann. Math. Study*. Princeton, NJ: Princeton Univ. Press, 1957, vol. III, pp. 97–139.

[22] T. Holliday, P. Glynn, and A. Goldsmith. (2003, Submitted for publication.) On Entropy and Lyapunov Exponents for Finite State Channels. [Online]. Available: http://wsl.stanford.edu/Publications/THolliday/Lyapunov.pdf

[23] P. Jacquet, G. Seroussi, and W. Szpankowski, "On the entropy of a hidden Markov process," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 2004, pp. 362–371.

[24] G. Kallianpur, *Stochastic Filtering Theory*. New York: Springer-Verlag, 1980.

[25] R. Khasminskii and O. Zeitouni, "Asymptotic filtering for finite state Markov chains," *Stochastic Processes and their Applications*, vol. 63, pp. 1–10, 1996.

[26] H. Kunita, "Asymptotic behavior of the nonlinear filtering errors of Markov processes," *J. Multivariate Anal.*, no. 1, pp. 365–393, 1971.

[27] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.

[28] N. Merhav and I. Kontoyiannis, "Source coding exponents for zero-delay coding with finite memory," *IEEE Trans. Inf. Theory*, no. 3, pp. 609–625, Mar. 2003.

[29] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert–Elliot channel," *IEEE Trans. Inf. Theory*, vol. 35, no. 6, pp. 1277–1290, Nov. 1989.

[30] E. Ordentlich and T. Weissman, "New bounds on the entropy rate of hidden Markov processes," in *Proc. IEEE Information Theory Workshop*, San Antonio, TX, Oct. 2004.

[31] Y. Peres, "Analytic dependence of Lyapunov exponents on transition probabilities," in *Proc. Oberwolfack Conf. (Lecture Notes in Mathematics)*. Berlin, Germany: Springer-Verlag, 1991, vol. 1486, pp. 64–80.

[32] N. Phamdo and N. Farvardin, "Optimal detection of discrete Markov sources over discrete memoryless channels—Applications to combined source-channel coding," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 186–193, Jan. 1994.

[33] D. Sagalowicz, "Hypothesis testing with finite memory," Ph.D. dissertation, Elec. Eng. Dept., Stanford Univ., Stanford, CA, 1970.

[34] E. Samuel, "An empirical Bayes approach to the testing of certain parametric hypotheses," *Ann. Math. Statist.*, vol. 34, no. 4, pp. 1370–1385, 1963.

[35] L. Shue, S. Dey, B. D. O. Anderson, and F. De Bruyne, "On state-estimation of a two-state hidden Markov model with quantization," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 202–208, Jan. 2001.

[36] 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.

[37] 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.

[38] ——, "On competitive prediction and its relationship to rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3185–3194, Dec. 2003.

[39] 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.

[40] W. M. Wonham, "Some applications of stochastic differential equations to optimal nonlinear filtering," *SIAM J. Control Optimiz.*, vol. 2, pp. 347–368, 1965.