On the Optimality of Symbol by Symbol Filtering and Denoising

Erik Ordentlich, Tsachy Weissman
HP Laboratories Palo Alto
HPL-2003-254
December 5th , 2003*

E-mail: eord@hpl.hp.com, tsachy@stanford.edu

filtering, denoising, smoothing, state estimation, hidden Markov models, entropy rate, large deviations

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 non-trivial 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 $\{Z_{t'}\}_{t'=t}$) 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 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 $d > 0$ such that the "say-what-you-see" scheme is optimal provided the channel crossover probability is less than d. 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).

* Internal Accession Date Only
Approved for External Publication

On the Optimality of Symbol by Symbol Filtering and Denoising

Erik Ordentlich* Tsachy Weissman†

December 7, 2003

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 non-trivial 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 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).

Key words and phrases: Asymptotic entropy, Denoising, Discrete Memoryless Channels, Entropy rate, Estimation, Filtering, Hidden Markov processes, Large deviations performance, Noisy channels, Singlet decoding, Symbol by symbol schemes.

1 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, there is a given loss function measuring the goodness of the reconstruction 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* [Dev74]. When $\{X_t\}$ is not memoryless, on the other hand, the optimal estimate of each X_t will, in

*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, CA 94305 USA (e-mail: tsachy@stanford.edu). Part of this work was done while T. Weissman was visiting Hewlett-Packard Laboratories, Palo Alto, CA, USA. T. Weissman was partially supported by NSF grants DMS-0072331 and CCR-0312839.

general, depend on all the observations available to it in a non-trivial 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 non-linear filtering theory is devoted to the study of optimal estimation schemes for these problems (cf., e.g., [AZ97, ABK00, Kal80, Kun71, EM02] 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 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.

Qualitatively speaking, 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 present noisy observation[1], for all possible realizations of the noisy observations[2]. 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 the said 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 said distribution 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 [Dev74] (cf. also [Dra65, Sag70]) for the binary Markov source corrupted by a BSC. Though the characterization of situations where optimum performance is attained using symbol-by-symbol schemes has since been studied for other problems in information theory (e.g. [GRV03, NG82]), the optimality of singlet schemes for filtering and denoising has, to our knowledge, not been considered beyond the setting of [Dev74]. Our interest in the problem was triggered by the recently discovered Discrete Universal Denoiser (DUDE) [WOS$^+$03a, WOS$^+$03b]. 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 [DW03]

[1] Note that 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 only be the case if the underlying source was memoryless), but only that the corresponding optimal estimates do.

[2] More precisely, for source realizations in a set of probability one.

where, for example, in denoising a binary Markov chain with a strong enough bias towards 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 2 we introduce some notation and conventions that will be assumed throughout. Section 3 is dedicated to the case of a Markov chain corrupted by a memoryless channel. To fix notation and for completeness we start in subsection A by deriving classical results concerning the evolution of conditional distributions of the clean symbol given past and\or future observations. We then apply these results in subsection 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 4 (subsections A and B) we show that when the noise-free process is binary enough information about the support of the said 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 subsection C we further specialize the results to the BSC, characterizing all situations where singlet decoding is optimal (and thereby re-deriving the results of [Dev74] in a more explicit form). In subsection 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 5 we digress from the denoising problem and illustrate how the results of Section 4 can be used for obtaining bounds that appear to be new[3] on the entropy rate of a hidden Markov process. 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-level 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 [ADG94, Per] and the connection between Lyapunov exponents and entropy rate [HGG03, JSS03]. 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.

Section 6 is dedicated to the derivation of 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 is derived in a general setting encompassing arbitrarily distributed processes (or fields) corrupted by arbitrarily distributed noise. The remainder of that section details the application of the general condition to a few concrete scenarios. In subsection 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. Subsection 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 the said threshold).

In Section 7 we obtain the exponent associated with the large deviations performance of a singlet decoder, thus

[3] The closed form for the entropy rate of a hidden Markov process is still an open problem (cf. [EM02, HGG03] and references therein).

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 8 we summarize the paper and discuss a few directions for future research.

2 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 $\mathcal{A}$. $Z(T)$ will denote the noisy observation process, jointly distributed with $X(T)$ and having components taking values in $\mathcal{B}$. Formally, we define a *denoiser* to be a collection of measurable functions $\{\hat{X}_t\}_{t\in T}$, where $\hat{X}_t : \mathcal{B}^T \to \mathcal{A}$ 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) $\Lambda : \mathcal{A}^2 \to [0, \infty)$, represented by the matrix $\mathbf{\Lambda} = \{\Lambda(i, j)\}_{i,j\in\mathcal{A}}$, 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 the notation $\mathbf{X}$, $\{X_t\}_{t\in\mathbb{Z}}$ or $X_{-\infty}^{\infty}$ interchangeably with $X(T)$. We shall also let $X^t = \{X_{t'}\}_{t'\leq t}$. 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 $\mathbf{Z}$. 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 $\mathbf{P} \in \mathcal{M}(\mathcal{A})$, let $U(\mathbf{P})$ denote the Bayes envelope (cf., e.g., [Han57, Sam63, MF98]) associated with the loss function Λ defined by

$$U(\mathbf{P}) = \min_{\hat{x}\in\mathcal{A}} \sum_{a\in\mathcal{A}} \Lambda(a, \hat{x})\mathbf{P}(a) = \min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \mathbf{P}, \tag{1}$$

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. P will also be used for conditional probability with (when involving continuous alphabets or infinite index sets) the standard slight abuse of notation that goes with it: $P(X_i = a|Z_{-\infty}^i)$, 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

$$E\Lambda(X_t, \hat{X}_t(Z(T))) = EU(P(X_t = \cdot|Z(T))), \tag{2}$$

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

$$E\Lambda(X_t, \hat{X}_t(Z_{-\infty}^t)) = EU(P(X_t = \cdot|Z_{-\infty}^t)). \tag{3}$$

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 $f(\varepsilon) \stackrel{\sim}{<} g(\varepsilon)$ will stand for $\limsup_{\varepsilon\downarrow 0} \frac{f(\varepsilon)}{g(\varepsilon)} \leq 1$. $f(\varepsilon) = O(g(\varepsilon))$ will stand for $\limsup_{\varepsilon\downarrow 0} \frac{f(\varepsilon)}{g(\varepsilon)} < \infty$ and $f(\varepsilon) = \Omega(g(\varepsilon))$ will stand for $\liminf_{\varepsilon\downarrow 0} \frac{f(\varepsilon)}{g(\varepsilon)} > 0$. $f(\varepsilon) \asymp g(\varepsilon)$ 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.

3 Finite-Alphabet Markov Chain Corrupted by a Memoryless Channel

In this section we assume $\{X_t\}_{t\in\mathbb{Z}}$ to be a stationary ergodic first-order Markov process with the finite alphabet $\mathcal{A}$. Let $K : \mathcal{A}^2 \to [0, 1]$ be its transition kernel

$$K(a, b) = P(X_{i+1} = b | X_i = a), \tag{4}$$

K_r be the transition kernel of the time reversed process

$$K_r(a, b) = P(X_i = b | X_{i+1} = a), \tag{5}$$

and let μ denote its marginal distribution

$$\mu(a) = P(X_i = a), \tag{6}$$

which is the unique probability measure satisfying

$$\mu(b) = \sum_{a\in\mathcal{A}} \mu(a)K(a, b) \quad \forall b \in \mathcal{A}. \tag{7}$$

We assume, without loss of generality, that

$$\mu(a) = \Pr(X_t = a) > 0 \quad \forall a \in \mathcal{A}. \tag{8}$$

Throughout this section we assume that $\{Z_t\}$ is the noisy observation process of $\{X_t\}$ when corrupted by the memoryless channel C. For simplicity, we shall confine attention to one of two cases[4]:

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.

For concreteness in the derivations below, 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 PMF and $P(Z_i = z|Z_{-\infty}^i)$ as a conditional density rather than a conditional probability).

[4]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.

A Evolution of the Conditional Distributions

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

$$\beta_i(a) = P(X_i = a|Z_{-\infty}^i) \tag{9}$$

and

$$\gamma_i(a) = P(X_i = a|Z_i^\infty). \tag{10}$$

For $a \in \mathcal{A}$,

$$\begin{aligned}
\beta_i(a) &= P(X_i = a|Z_{-\infty}^i) \\
&= \frac{P(X_i = a, Z_i|Z_{-\infty}^{i-1})}{P(Z_i|Z_{-\infty}^{i-1})} \\
&= \frac{P(X_i = a|Z_{-\infty}^{i-1})P(Z_i|X_i = a)}{P(Z_i|Z_{-\infty}^{i-1})} \\
&= \frac{C(a, Z_i)\sum_{b\in\mathcal{A}} P(X_i = a|X_{i-1} = b)P(X_{i-1} = b|Z_{-\infty}^{i-1})}{P(Z_i|Z_{-\infty}^{i-1})} \\
&= \frac{C(a, Z_i)\sum_{b\in\mathcal{A}} K(b, a)\beta_{i-1}(b)}{P(Z_i|Z_{-\infty}^{i-1})} \\
&= \frac{C(a, Z_i)[K^T\beta_{i-1}](a)}{\sum_{a'\in\mathcal{A}} C(a', Z_i)[K^T\beta_{i-1}](a')},
\end{aligned} \tag{11}$$

where in the last line K^T denotes the transposed matrix representing the Markov kernel of (4). In vector form (11) becomes

$$\beta_i = \frac{1}{\mathbf{1}^T[\mathbf{c}_{Z_i} \odot [K^T\beta_{i-1}]]}\mathbf{c}_{Z_i} \odot [K^T\beta_{i-1}] = T(Z_i, \beta_{i-1}), \tag{12}$$

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

$$T(b, \beta) = \frac{1}{\mathbf{1}^T[\mathbf{c}_b \odot [K^T\beta]]}\mathbf{c}_b \odot [K^T\beta], \tag{13}$$

(12) assumes the form

$$\beta_i = T(Z_i, \beta_{i-1}). \tag{14}$$

An equivalent way of expressing (14) (which will be of convenience in the sequel) is in terms of the log-likelihoods: for $a, b \in \mathcal{A}$,

$$\begin{aligned}
\log\frac{\beta_i(a)}{\beta_i(b)} &= \log\frac{C(a, Z_i)}{C(b, Z_i)} + \log\frac{[K^T\beta_{i-1}](a)}{[K^T\beta_{i-1}](b)} \\
&= \log\frac{C(a, Z_i)}{C(b, Z_i)} + \log\frac{\sum_{c\in\mathcal{A}} K(c, a)\beta_{i-1}(c)}{\sum_{c\in\mathcal{A}} K(c, b)\beta_{i-1}(c)}.
\end{aligned} \tag{15}$$

By an analogous computation, we get

$$\gamma_{i-1} = T_r(Z_i, \gamma_i), \tag{16}$$

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

$$T_r(b, \gamma) = \frac{1}{\mathbf{1}^T[\mathbf{c}_b \odot [K_r^T\gamma]]}\mathbf{c}_b \odot [K_r^T\gamma]. \tag{17}$$

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 $E \subseteq \mathcal{M}(\mathcal{A})$ and $\beta, \gamma \in \mathcal{M}(\mathcal{A})$,

$$F(E,\beta) = \sum_{b:T(b,\beta)\in E} \left[\sum_{a\in\mathcal{A}} \mathbf{c}_b \odot [K^T\beta]\right], \quad F_r(E,\gamma) = \sum_{b:T_r(b,\gamma)\in E} \left[\sum_{a\in\mathcal{A}} \mathbf{c}_b \odot [K_r^T\gamma]\right], \tag{18}$$

the following result is implicit in [Bla57].

Claim 1 (Blackwell [Bla57]) *$\{\beta_i\}$, $\{\gamma_i\}$ are both stationary first-order Markov processes. Furthermore, the distribution of β_0, Q, satisfies, for each Borel set $E \subseteq \mathcal{M}(\mathcal{A})$, the integral equation*

$$Q(E) = \int_{\beta\in\mathcal{M}(\mathcal{A})} F(E,\beta)dQ(\beta) \tag{19}$$

and the distribution of γ_0, Q_r, satisfies the integral equation

$$Q_r(E) = \int_{\gamma\in\mathcal{M}(\mathcal{A})} F_r(E,\gamma)dQ_r(\gamma). \tag{20}$$

We reproduce a proof in the spirit of [Bla57] 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

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

where the last equality follows similarly as in the derivation of (14). 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

$$P(\beta_i \in E|\beta_{-\infty}^{i-1}) = F(E,\beta_{i-1}), \tag{21}$$

thus establishing the Markov property. Taking expectations in both sides of (21) gives (19). □

Note that the optimal filtering performance, $EU(\beta_i)$ (recall (3)), has a "closed form" expression in terms of the distribution Q:

$$EU(\beta_i) = \int_{\beta\in\mathcal{M}(\mathcal{A})} U(\beta)dQ(\beta). \tag{22}$$

Similarly, as was noted in [Bla57], the entropy rate (assuming a discrete channel output[5]) of $\mathbf{Z}$ can also be given a "closed form" expression in terms of the distribution Q. To see this note that

$$P(Z_{i+1} = z|Z_{-\infty}^i) = \left[\beta_i^T \cdot K \cdot \mathcal{C}\right](z), \tag{23}$$

[5]The case of a continuous-valued output (and differential entropy rate) would be handled analogously.

with K denoting the Markov transition matrix (with (a,b)-th entry given by (4)) 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 $H(\mathbf{P}) = -\sum_z \mathbf{P}(z)\log\mathbf{P}(z)$ and $\overline{H}(\mathbf{Z})$ denote the entropy rate of $\mathbf{Z}$,

$$\overline{H}(\mathbf{Z}) = EH(P(Z_{i+1}=\cdot|Z_{-\infty}^{i})) = EH\left(\left[\beta_i^T\cdot K\cdot\mathcal{C}\right]\right) = \int_{\beta\in\mathcal{M}(\mathcal{A})} H\left(\left[\beta^T\cdot K\cdot\mathcal{C}\right]\right) dQ(\beta). \tag{24}$$

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

$$\eta_i(a) = P(X_i=a|Z_{-\infty}^{\infty}) \tag{25}$$

and note that

$$\begin{aligned}
\eta_i(a) &= P(X_i=a|Z_{-\infty}^{\infty}) \\
&= \lim_{n\to\infty}\frac{P(X_i=a,Z_{-n}^{n})}{P(Z_{-n}^{n})} \\
&= \lim_{n\to\infty}\frac{P(Z_{-n}^{n}|X_i=a)\mu(a)}{P(Z_{-n}^{n})} \\
&= \lim_{n\to\infty}\frac{P(Z_{-n}^{i-1}|X_i=a)P(Z_{i+1}^{n}|X_i=a)P(Z_i|X_i=a)\mu(a)}{P(Z_{-n}^{n})} \\
&= \lim_{n\to\infty}\frac{\frac{P(X_i=a|Z_{-n}^{i-1})P(Z_{-n}^{i-1})}{\mu(a)}\cdot\frac{P(X_i=a|Z_{i+1}^{n})P(Z_{i+1}^{n})}{\mu(a)}P(Z_i|X_i=a)\mu(a)}{P(Z_{-n}^{n})} \\
&= \lim_{n\to\infty}\frac{P(X_i=a|Z_{-n}^{i-1})P(Z_{-n}^{i-1})P(X_i=a|Z_{i+1}^{n})P(Z_{i+1}^{n})P(Z_i|X_i=a)}{\mu(a)P(Z_{-n}^{n})} \\
&= \lim_{n\to\infty}\frac{\frac{1}{\mu(a)}P(X_i=a|Z_{-n}^{i-1})P(X_i=a|Z_{i+1}^{n})P(Z_i|X_i=a)}{\sum_{a'\in\mathcal{A}}\frac{1}{\mu(a')}P(X_i=a'|Z_{-n}^{i-1})P(X_i=a'|Z_{i+1}^{n})P(Z_i|X_i=a')} \\
&= \frac{\frac{1}{\mu(a)}P(X_i=a|Z_{-\infty}^{i-1})P(X_i=a|Z_{i+1}^{\infty})P(Z_i|X_i=a)}{\sum_{a'\in\mathcal{A}}\frac{1}{\mu(a')}P(X_i=a'|Z_{-\infty}^{i-1})P(X_i=a'|Z_{i+1}^{\infty})P(Z_i|X_i=a')} \\
&= \frac{\frac{1}{\mu(a)}[K^T\beta_{i-1}](a)[K_r^T\gamma_{i+1}](a)C(a,Z_i)}{\sum_{a'\in\mathcal{A}}\frac{1}{\mu(a')}[K^T\beta_{i-1}](a')[K_r^T\gamma_{i+1}](a')C(a',Z_i)}
\end{aligned}$$

or, in vector notation,

$$\eta_i = \frac{[K^T\beta_{i-1}]\odot[K_r^T\gamma_{i+1}]\odot\mathbf{c}_{Z_i}\div\mu}{\mathbf{1}^T\left[[K^T\beta_{i-1}]\odot[K_r^T\gamma_{i+1}]\odot\mathbf{c}_{Z_i}\div\mu\right]} = G_{Z_i}(\beta_{i-1},\gamma_{i+1}), \tag{26}$$

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

$$G_b(\beta,\gamma) = \frac{[K^T\beta]\odot[K_r^T\gamma]\odot\mathbf{c}_b\div\mu}{\mathbf{1}^T\left[[K^T\beta]\odot[K_r^T\gamma]\odot\mathbf{c}_b\div\mu\right]}. \tag{27}$$

Analogously to (15) we can write

$$\log\frac{\eta_i(a)}{\eta_i(b)} = \log\frac{C(a,Z_i)}{C(b,Z_i)} + \log\frac{[K^T\beta_{i-1}](a)}{[K^T\beta_{i-1}](b)} + \log\frac{[K_r^T\gamma_{i+1}](a)}{[K_r^T\gamma_{i+1}](b)} - \log\frac{\mu(a)}{\mu(b)}. \tag{28}$$

Note that, by (26), optimum denoising performance is given by $EU(\eta_i) = EU\left(G_{Z_i}(\beta_{i-1},\gamma_{i+1})\right)$ (which can be expressed in terms of the measures Q and Q_r of Claim 1 analogously as in (22)[6]).

[6]Conditioned on X_i, Z_i, β_{i-1} and γ_{i+1} are independent. Thus $EU\left(G_{Z_i}(\beta_{i-1},\gamma_{i+1})\right)$ 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 (19) and, unfortunately, is not explicitly known to date (cf. [Bla57] for a discussion of some of its peculiar properties). Correspondingly, explicit expressions for optimum filtering performance (cf. [KZ96]), denoising performance (cf. [SDAB01]), and for the entropy rate of the noisy process (cf. [Bla57, HGG03, EM02]) 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 denosing 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 support of Q and of Q_r.

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

$$\hat{X}(\mathbf{P}) = \{a \in \mathcal{A} : \boldsymbol{\lambda}_a^T \mathbf{P} = \min_{\hat{x} \in \mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \mathbf{P}\}. \tag{29}$$

Note that we have slightly deviated from common practice, letting $\hat{X}(\mathbf{P})$ be set-valued so that $|\hat{X}(\mathbf{P})| \geq 1$ with equality if and only if the minimizer of $\boldsymbol{\lambda}_{\hat{x}}^T \mathbf{P}$ 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) (2) or (3):

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

$$P(\hat{X}_i(Z_{-\infty}^i) \in \hat{X}(\beta_i)) = 1. \tag{30}$$

A denoising scheme $\{\hat{X}_i(\cdot)\}$ is optimal if and only if for each i

$$P(\hat{X}_i(Z_{-\infty}^\infty) \in \hat{X}(\eta_i)) = 1. \tag{31}$$

For $f : \mathcal{B} \to \mathcal{A}$ define $S_f \subseteq \mathcal{M}(\mathcal{A})$ by

$$S_f = \{s \in \mathcal{M}(\mathcal{A}) : f(b) \in \hat{X}(T(b, s)) \ \forall b \in \mathcal{B}\}. \tag{32}$$

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 $b \in \mathcal{B}$. Somewhat less formally[7], 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 (30) and (12), 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:

Theorem 1 *Assume $C(a, b) > 0$ for all $a \in \mathcal{A}, b \in \mathcal{B}$. 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,

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

Consequently $P\left(f(Z_i) \in \hat{X}(T(Z_i, \beta_{i-1}))\right) = P\left(f(Z_i) \in \hat{X}(\beta_i)\right) = 1$, establishing optimality by (30).

[7] Neglecting the possibility that $|\hat{X}(T(b, s))| > 1$.

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

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

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

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

implying, in turn, when combined with (8) the existence of $a \in \mathcal{A}$ such that

$$P\left(f(b) \in \hat{X}(T(b, \beta_{i-1}))|X_i = a\right) < 1. \tag{33}$$

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

$$P\left(f(Z_i) \in \hat{X}(\beta_i)|X_i = a\right) = P\left(f(Z_i) \in \hat{X}(T(Z_i, \beta_{i-1}))|X_i = a\right) = \sum_{b' \in \mathcal{B}} P\left(f(b') \in \hat{X}(T(b', \beta_{i-1}))|X_i = a\right) C(a, b'). \tag{34}$$

Inequality (33), combined with (34) and the fact that $C(a, b) > 0$, implies $P\left(f(Z_i) \in \hat{X}(\beta_i)|X_i = a\right) < 1$, which, in turn, leads to $P\left(f(Z_i) \in \hat{X}(\beta_i)\right) < 1$. Thus, $\hat{X}_i(Z^i_{-\infty}) = f(Z_i)$ does not satisfy (30) 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 (32) to

$$\{s \in \mathcal{M}(\mathcal{A}) : f(b) \in \hat{X}(T(b, s)) \ \forall b \in \mathcal{S}(s)\}, \tag{35}$$

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

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

$$R_f = \{(s_1, s_2) \in \mathcal{M}(\mathcal{A}) \times \mathcal{M}(\mathcal{A}) : f(b) \in \hat{X}(G_b(s_1, s_2)) \ \forall b \in \mathcal{B}\}. \tag{36}$$

Theorem 2 *Assume $C(a, b) > 0$ for all $a \in \mathcal{A}, b \in \mathcal{B}$. 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[8]. Then, the theorems above 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 Theorem 1 and Theorem 2 can be recast in terms of intersections of intervals with explicitly characterized endpoints.

[8] To get outer bounds, for example, 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 of Section 4.

4 The Binary Markov Source

Assume now $\mathcal{A} = \{0, 1\}$ 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[9], that $0 < \pi_{01} \leq 1$ and $0 < \pi_{10} \leq 1$. 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 (15) becomes

$$\begin{aligned} \log \frac{\beta_i(1)}{1-\beta_i(1)} &= \log \frac{C(1, Z_i)}{C(0, Z_i)} + \log \frac{\sum_{c\in\mathcal{A}} K(c, 1)\beta_{i-1}(c)}{\sum_{c\in\mathcal{A}} K(c, 0)\beta_{i-1}(c)} \\ &= \log \frac{C(1, Z_i)}{C(0, Z_i)} + \log \frac{\pi_{01}(1-\beta_{i-1}(1)) + (1-\pi_{10})\beta_{i-1}(1)}{(1-\pi_{01})(1-\beta_{i-1}(1)) + \pi_{10}\beta_{i-1}(1)}. \end{aligned} \tag{37}$$

Equivalently, letting[10] $l_i = \log \frac{\beta_i(1)}{1-\beta_i(1)}$, we obtain

$$l_i = \log \frac{C(1, Z_i)}{C(0, Z_i)} + h(l_{i-1}), \tag{38}$$

where

$$h(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x\pi_{10}}. \tag{39}$$

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 $K_r = K$, equation (28) becomes

$$m_i = \log \frac{C(1, Z_i)}{C(0, Z_i)} + h(l_{i-1}) + h(k_{i+1}) - \log \frac{\pi_{01}}{\pi_{10}}. \tag{40}$$

A The Support of the Log-Likelihoods

By differentiating it is easily verified that

Fact 2 *The function h is non-decreasing whenever $\pi_{10} + \pi_{01} \leq 1$, otherwise it is non-increasing.*

Define now[11]

$$U_{bs} = \operatorname{ess\,sup} \frac{C(1, Z_0)}{C(0, Z_0)} \tag{41}$$

and

$$L_{bs} = \operatorname{ess\,inf} \frac{C(1, Z_0)}{C(0, Z_0)}. \tag{42}$$

Examples:

- BSC with $\delta \leq 1/2$. $U_{bs} = \frac{1-\delta}{\delta}$, $L_{bs} = \frac{\delta}{1-\delta}$.
- Binary Input Additive White Gaussian Noise (BIAWGN) channel where

$$Z_i|X_i = 0 \sim \mathcal{N}(-1, 1), \quad Z_i|X_i = 1 \sim \mathcal{N}(1, 1). \tag{43}$$

  $U_{bs} = \infty$, $L_{bs} = 0$.

[9] The remaining cases imply zero probability to one of the symbols and so are trivial.

[10] l_i, as well as k_i and m_i defined below, are $\mathbb{R} \cup \{\infty, -\infty\}$-valued random variables.

[11] For a general binary input channel the ratios in equations (41) and (42) would be replaced by the Radon-Nykodim derivative of the output distribution given input symbol 1 w.r.t. the output distribution given input symbol 0.

- Binary input additive Laplacian noise (BIALN) channel where

$$C(0,z) = c(\alpha)e^{-\alpha|z+\mu|}, \quad C(1,z) = c(\alpha)e^{-\alpha|z-\mu|}, \quad \mu > 0, \quad z \in \mathbb{R}, \tag{44}$$

  $c(\alpha)$ being the normalization factor. $U_{bs} = e^{2\alpha\mu}$, $L_{bs} = e^{-2\alpha\mu}$.

Define further

$$I_1 = \operatorname{ess\,inf} l_i \tag{45}$$

and

$$I_2 = \operatorname{ess\,sup} l_i. \tag{46}$$

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 below, depends on the support of l_i solely through this interval.

Theorem 3 *The pair (I_1, I_2) (defined in (45) and (46)) is given by the unique solution (in the extended real line) to*

1. *When $\pi_{1,0} + \pi_{0,1} \leq 1$: $I_1 = \log L_{bs} + h(I_1)$ and $I_2 = \log U_{bs} + h(I_2)$.*
2. *When $\pi_{1,0} + \pi_{0,1} > 1$: $I_1 = \log L_{bs} + h(I_2)$ and $I_2 = \log U_{bs} + h(I_1)$.*

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

Proof of Theorem 3: We assume $\pi_{1,0} + \pi_{0,1} \leq 1$ (the proof for the case $\pi_{1,0} + \pi_{0,1} > 1$ is analogous). Monotonicity and continuity of h imply

$$\operatorname{ess\,inf} h(l_{i-1}) = h(\operatorname{ess\,inf} l_{i-1}) = h(\operatorname{ess\,inf} l_i) = h(I_1). \tag{47}$$

Thus

$$\begin{aligned}
I_1 &= \operatorname{ess\,inf} l_i && (48)\\
&= \operatorname{ess\,inf}\left[\log\frac{C(1,Z_i)}{C(0,Z_i)} + h(l_{i-1})\right] && (49)\\
&= \operatorname{ess\,inf}\left[\log\frac{C(1,Z_i)}{C(0,Z_i)}\right] + \operatorname{ess\,inf} h(l_{i-1}) && (50)\\
&= \log\left[\operatorname{ess\,inf}\frac{C(1,Z_i)}{C(0,Z_i)}\right] + h(I_1) && (51)\\
&= \log L_{bs} + h(I_1), && (52)
\end{aligned}$$

where (49) follows from (38), (50) follows since all transitions of the Markov chain have positive probability, and (51) is due to (47). The relationship $I_2 = \log U_{bs} + h(I_2)$ is established similarly. □

Elementary algebra shows that for $\pi_{1,0} + \pi_{0,1} \leq 1$ 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

$$f(\pi_{01}, \pi_{10}, \alpha) = \log\left[\frac{-1 + \alpha + \pi_{01} - \alpha\pi_{10} + \sqrt{4\alpha\pi_{01}\pi_{10} + (1 - \alpha - \pi_{01} + \alpha\pi_{10})^2}}{2\pi_{10}}\right]. \tag{53}$$

Thus, from Theorem 3 we get $I_1 = f(\pi_{01}, \pi_{10}, L_{bs})$ and $I_2 = f(\pi_{01}, \pi_{10}, U_{bs})$ when $\pi_{1,0} + \pi_{0,1} \leq 1$. An explicit form of the unique solution for the pair (I_1, I_2) when $\pi_{1,0} + \pi_{0,1} > 1$ can also be obtained[12] by solving the pair of equations in the second item of Theorem 3.

[12] We omit the expressions which are somewhat more involved than that in (53).

For the analogous quantities in the denosing problem

$$J_1 = \operatorname{ess\,inf} m_i \tag{54}$$

and

$$J_2 = \operatorname{ess\,sup} m_i, \tag{55}$$

we have the following:

Theorem 4 *1. When $\pi_{1,0} + \pi_{0,1} \leq 1$: $J_1 = \log L_{bs} + 2h(I_1) - \log \frac{\pi_{01}}{\pi_{10}}$ and $J_2 = \log U_{bs} + 2h(I_2) - \log \frac{\pi_{01}}{\pi_{10}}$.*

2. When $\pi_{1,0} + \pi_{0,1} > 1$: $J_1 = \log L_{bs} + 2h(I_2) - \log \frac{\pi_{01}}{\pi_{10}}$ and $J_2 = \log U_{bs} + 2h(I_1) - \log \frac{\pi_{01}}{\pi_{10}}$.

Proof: The proof is similar to that of Theorem 3, using (40) (instead of (38)) 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 $\pi_{1,0} + \pi_{0,1} \leq 1$, we get the explicit forms $J_1 = \log L_{bs} + 2h(f(\pi_{01}, \pi_{10}, L_{bs})) - \log \frac{\pi_{01}}{\pi_{10}}$ and $J_2 = \log U_{bs} + 2h(f(\pi_{01}, \pi_{10}, U_{bs})) - \log \frac{\pi_{01}}{\pi_{10}}$ where f was defined in (53). Explicit (though more cumbersome) expressions can also be obtained for the case $\pi_{1,0} + \pi_{0,1} > 1$.

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

$$\hat{X}_i(Z^i_{-\infty}) = f_{opt}(Z^i_{-\infty}) = \begin{cases} 1 & \text{a.s. on } \{l_i > 0\} \\ 0 & \text{a.s. on } \{l_i < 0\} \\ \text{arbitrary} & \text{on } \{l_i = 0\}. \end{cases} \tag{56}$$

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

$$\hat{X}_i(Z^\infty_{-\infty}) = g_{opt}(Z^\infty_{-\infty}) = \begin{cases} 1 & \text{a.s. on } \{m_i > 0\} \\ 0 & \text{a.s. on } \{m_i < 0\} \\ \text{arbitrary} & \text{on } \{m_i = 0\}. \end{cases} \tag{57}$$

The following is a direct consequence of equations (56) and (57) 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 (56), $\hat{X}_i(Z^i_{-\infty}) \equiv 1$ is an optimal filter. Conversely, if $\hat{X}_i(Z^i_{-\infty}) \equiv 1$ is an optimal filter then, by (56), $l_i \geq 0$ 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 $I_1 = f(\pi_{01}, \pi_{10}, L_{bs})$ and $I_2 = f(\pi_{01}, \pi_{10}, U_{bs})$ (with f given in (53)) we obtain:

Corollary 1 *Assume $\pi_{1,0}+\pi_{0,1}\le 1$. The filter ignoring its observations and saying*

- *"all-zeros" is optimal if and only if*

$$\sqrt{4U_{bs}\pi_{01}\pi_{10}+(1-U_{bs}-\pi_{01}+U_{bs}\pi_{10})^2}\le 1-U_{bs}-\pi_{01}+U_{bs}\pi_{10}+2\pi_{10}.$$

- *"all-ones" is optimal if and only if*

$$\sqrt{4L_{bs}\pi_{01}\pi_{10}+(1-L_{bs}-\pi_{01}+L_{bs}\pi_{10})^2}\ge 1-L_{bs}-\pi_{01}+L_{bs}\pi_{10}+2\pi_{10}.$$

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 *Singlet decoding is optimal for the filtering problem if and only if*

$$\log\frac{C(1,Z_1)}{C(0,Z_1)}\notin(\log U_{bs}-I_2,\log L_{bs}-I_1)\quad a.s. \tag{58}$$

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

Remark: Elementary algebra shows that, for $\pi_{1,0}+\pi_{0,1}\le 1$,

$$h(I_1)=\log\frac{-L_{bs}(1-\pi_{10})^2-\pi_{01}(1+\pi_{10})+(-1+\pi_{10})\left(-1+\sqrt{4L_{bs}\pi_{01}\pi_{10}+(-1+\pi_{01}+L_{bs}(1-\pi_{10}))^2}\right)}{\pi_{10}\left(-1+\pi_{01}-L_{bs}(1-\pi_{10})-\sqrt{4L_{bs}\pi_{01}\pi_{10}+(-1+\pi_{01}+L_{bs}(1-\pi_{10}))^2}\right)}. \tag{59}$$

Thus the condition in this case is that $\frac{C(1,Z_1)}{C(0,Z_1)}$ lie outside the interval whose left endpoint is

$$\frac{\pi_{10}\left(-1+\pi_{01}-U_{bs}(1-\pi_{10})-\sqrt{4U_{bs}\pi_{01}\pi_{10}+(-1+\pi_{01}+U_{bs}(1-\pi_{10}))^2}\right)}{-U_{bs}(1-\pi_{10})^2-\pi_{01}(1+\pi_{10})+(-1+\pi_{10})\left(-1+\sqrt{4U_{bs}\pi_{01}\pi_{10}+(-1+\pi_{01}+U_{bs}(1-\pi_{10}))^2}\right)} \tag{60}$$

and right endpoint is

$$\frac{\pi_{10}\left(-1+\pi_{01}-L_{bs}(1-\pi_{10})-\sqrt{4L_{bs}\pi_{01}\pi_{10}+(-1+\pi_{01}+L_{bs}(1-\pi_{10}))^2}\right)}{-L_{bs}(1-\pi_{10})^2-\pi_{01}(1+\pi_{10})+(-1+\pi_{10})\left(-1+\sqrt{4L_{bs}\pi_{01}\pi_{10}+(-1+\pi_{01}+L_{bs}(1-\pi_{10}))^2}\right)}. \tag{61}$$

Proof of Claim 3: From (56) 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 (38) (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,

$$\log\frac{C(1,Z_i)}{C(0,Z_i)}\ge -\operatorname{ess\,inf}h(l_{i-1})\quad\text{or}\quad\log\frac{C(1,Z_i)}{C(0,Z_i)}\le -\operatorname{ess\,sup}h(l_{i-1}). \tag{62}$$

But, by Theorem 3, $(-\operatorname{ess\,sup}h(l_{i-1}),-\operatorname{ess\,inf}h(l_{i-1}))=(\log U_{bs}-I_2,\log L_{bs}-I_1)$ so (62) is equivalent to (58). □

The analogous result for the denoising problem is the following:

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

$$\frac{1}{2}\left[\log\frac{C(1,Z_1)}{C(0,Z_1)}-\log\frac{\pi_{01}}{\pi_{10}}\right]\notin(\log U_{bs}-I_2,\log L_{bs}-I_1)\quad a.s. \tag{63}$$

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* $(\log U_{bs}-I_2,\log L_{bs}-I_1)$ *interval.*

Proof: The proof follows from (40) and Theorem 4 analogously as Claim 3 followed from (38) and Theorem 3. □

Remarks:

1. For the case $\pi_{1,0} + \pi_{0,1} \leq 1$, the condition (63) is equivalent to $\sqrt{\frac{C(1,Z_1)\pi_{10}}{C(0,Z_1)\pi_{01}}}$ lying outside the interval whose endpoints are given, respectively, by (60) and (61).

2. Claim 3 (resp. 4) explicitly characterizes the conditions under which singlet decoding is optimal for the filtering (resp. denoising) problems. 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 $(\log U_{bs} - I_2, \log L_{bs} - I_1)$ interval, then the optimal estimate of X_i will be independent of the other observations (namely, it will be 0 if the said quantity falls below the interval and 1 if it falls above it irrespective of other observations).

3. 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 [Dev74]. The results below refine and extend those of [Dev74] in that they provide the explicit conditions for optimality of the "say-what-you-see" scheme in the non-symmetric 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 [Dev74] are based). Finally, due to a parametrization different than that in [Dev74], the region of optimality of singlet decoding for this setting admits a simple form.

We assume here a BSC(δ), ($\delta \leq 1/2$), and restrict attention throughout to the case $\pi_{1,0} + \pi_{0,1} \leq 1$. In this case, $U_{bs} = (1-\delta)/\delta$, and $L_{bs} = \delta/(1-\delta)$ 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 (53).

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*

$$\pi_{01} \leq \pi_{10} \quad \textit{and} \quad 2\log\frac{1-\delta}{\delta} \geq -f(\pi_{01}, \pi_{10}, \delta/(1-\delta))$$

or

$$\pi_{01} > \pi_{10} \quad \textit{and} \quad 2\log\frac{1-\delta}{\delta} \geq f(\pi_{01}, \pi_{10}, (1-\delta)/\delta).$$

2. *Denoising: The "say-what-you-see" denoiser is optimal if and only if*

$$\frac{3}{2}\log\frac{1-\delta}{\delta} \geq \max\left\{-\frac{1}{2}\log\frac{\pi_{10}}{\pi_{01}} - f(\pi_{01}, \pi_{10}, \delta/(1-\delta)), \frac{1}{2}\log\frac{\pi_{10}}{\pi_{01}} + f(\pi_{01}, \pi_{10}, (1-\delta)/\delta)\right\}.$$

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 as well as (for the "all-zero" and "all-one" schemes) by writing out the conditions $J_2 \leq 0$ and $J_1 \geq 0$ (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 $\log\left[\frac{1-\delta}{\delta}\right]$ be above the interval on the right side of (58) and $\log\left[\frac{\delta}{1-\delta}\right]$ be below that interval. More compactly, the condition is

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

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

$$2\log\frac{1-\delta}{\delta} \geq \max\left\{-f(\pi_{01},\pi_{10},\delta/(1-\delta)), f(\pi_{01},\pi_{10},(1-\delta)/\delta)\right\}. \tag{64}$$

Now, it is straightforward to check that when $\pi_{01} \leq \pi_{10}$, it is the left branch which attains the maximum in (64), 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 $\pi \leq 1/2$, corrupted by the BSC(δ), we have:*

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

Note that Corollary 3, both for the filtering and the denoising problems, completely characterizes the region in the square $0 \leq \pi \leq 1/2, 0 \leq \delta \leq 1/2$ where the minimum attainable error rate is δ. The minimum error rate at all points outside that region remains unknown[13].

This characterization carries over to cover the whole $0 \leq \pi \leq 1, 0 \leq \delta \leq 1/2$ region[14] as follows. The idea is to show a one-to-one correspondence between an optimal scheme for the Markov 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 $\pi = 1/2$.

Figure 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 Figure 1 are as follows:

[13] Other than that, the asymptotic behavior of the minimum error rate as $\pi \to 0$ has been characterized, respectively, for the filtering and denoising problems in [KZ96] and [SDAB01].

[14] 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 $\delta > 1/2$.

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

$$\text{optimum filtering performance} \begin{cases} = \delta & \text{for } \delta \leq \pi \\ > \pi & \text{for } \delta > \pi. \end{cases} \tag{65}$$

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

$$\text{optimum denoiser performance} \begin{cases} < \delta & \text{for a non-empty subset of the } \delta < \pi \text{ region} \\ > \pi & \text{for a non-empty subset of the } \delta > \pi \text{ region} \end{cases} \tag{66}$$

3. For filtering or denoising a Bernoulli(π) process corrupted by a BSC(δ) we have

$$\text{optimum filtering/denoising Bernoulli}(\pi) \begin{cases} = \delta & \text{for } \delta \leq \pi \\ = \pi & \text{for } \delta > \pi \end{cases} \tag{67}$$

   Comparing with (65) and (66) 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 (44) 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 $(\log U_{bs} - I_2, \log L_{bs} - I_1)$, 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 $(\log U_{bs} - I_2, \log L_{bs} - I_1)$ interval it must be either entirely below or entirely above it.

Finally, we note that by a similar argument, for the BIAWGN channel detailed in (43), 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 the said interval).

5 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 4 can be used to bound the entropy rate of the noisy observation process, the precise form of which is unknown (cf. [EM02, HGG03, GV96, MBD89] and references therein).

Assuming a discrete-valued noisy process, from equation (24) it is clear that a lower and an upper bound on its entropy rate is given by

$$\min_{\beta\in C_Q} H\left(\left[\beta^T\cdot K\cdot\mathcal{C}\right]\right) \le \overline{H}(\mathbf{Z}) \le \max_{\beta\in C_Q} H\left(\left[\beta^T\cdot K\cdot\mathcal{C}\right]\right), \tag{68}$$

with C_Q denoting the support of β_i. We now illustrate the use of (68) to derive an explicit bound for the entropy rate of the binary Markov chain corrupted by a BSC (considered in subsection 4.C). For this case,

$$H\left(\left[\beta^T\cdot K\cdot\mathcal{C}\right]\right) = h_b\left([\beta(1)(1-\pi_{10})+(1-\beta(1))\pi_{01}]*\delta\right), \tag{69}$$

where h_b is the binary entropy function

$$h_b(x) = -[x\log x + (1-x)\log(1-x)] \tag{70}$$

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

$$p*\delta = p(1-\delta)+(1-p)\delta. \tag{71}$$

A Entropy Rate in the "Rare-Spikes" Regime

Assuming $\delta\le 1/2$, $\pi_{01}+\pi_{10}\le 1$, and $\pi_{01}\le\pi_{10}$ it is readily verified that $[\beta(1)(1-\pi_{10})+(1-\beta(1))\pi_{01}]*\delta$ is increasing with $\beta(1)$ and is $\le 1/2$ 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 (68) in this case becomes

$$h_b\left(\left[\frac{e^{I_2}}{1+e^{I_2}}(1-\pi_{10})+\frac{1}{1+e^{I_2}}\pi_{01}\right]*\delta\right) \tag{72}$$

provided $I_2\le 0$ (since then $e^{I_2}/(1+e^{I_2})\le 1/2$), as $h_b(x)$ is increasing for $0\le x\le 1/2$. Hence, the expression in (72), with I_2 given explicitly in (53) with $\alpha=(1-\delta)/\delta$, is an upper bound to the entropy rate of the noisy process for all $\delta\le 1/2$ and all π_{01},π_{10} satisfying $\pi_{01}+\pi_{10}\le 1$ and $\pi_{01}\le\pi_{10}$, provided $I_2\le 0$. Arguing analogously, for the parameters in this region the expression in (72) with I_1 replaced by I_2 is a lower bound on the entropy rate so we get

$$h_b\left(\left[\frac{e^{I_1}}{1+e^{I_1}}(1-\pi_{10})+\frac{1}{1+e^{I_1}}\pi_{01}\right]*\delta\right) \le \overline{H}(\pi_{10},\pi_{01},\delta) \le h_b\left(\left[\frac{e^{I_2}}{1+e^{I_2}}(1-\pi_{10})+\frac{1}{1+e^{I_2}}\pi_{01}\right]*\delta\right), \tag{73}$$

where we let $\overline{H}(\pi_{10},\pi_{01},\delta)$ denote the entropy rate of the noisy process associated with these parameters. It is evident from (73) 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 $I_2\le 0$, $\pi_{01}+\pi_{10}\le 1$, and $\pi_{01}\le\pi_{10}$ 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

$$\pi_{10}=1-\varepsilon \quad\text{and}\quad \pi_{01}=a(\varepsilon), \tag{74}$$

where $a(\cdot)$ is an arbitrary function satisfying $0\le a(\varepsilon)\le\varepsilon$ (and all ε sufficiently small so that $I_2\le 0$), (73) becomes

$$h_b\left(\left[\frac{e^{I_1}}{1+e^{I_1}}\varepsilon+\frac{1}{1+e^{I_1}}a(\varepsilon)\right]*\delta\right) \le \overline{H}(1-\varepsilon,a(\varepsilon),\delta) \le h_b\left(\left[\frac{e^{I_2}}{1+e^{I_2}}\varepsilon+\frac{1}{1+e^{I_2}}a(\varepsilon)\right]*\delta\right). \tag{75}$$

Note, in particular, that for $a(\varepsilon) = \varepsilon$ (75) gives $\overline{H}(1-\varepsilon, \varepsilon, \delta) = h_b(\varepsilon * \delta)$ 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 (74) tends to the "all-zeros" source so it is natural to expect that

$$\lim_{\varepsilon \downarrow 0} \overline{H}(1-\varepsilon, a(\varepsilon), \delta) = h_b(\delta). \tag{76}$$

We now use the bounds in (75), combined with the characterization of I_1 and I_2 from subsection 4.A, to show that not only does (76) hold, but the convergence rate is linear in $a(\varepsilon)$ with the constant identified as well.

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

$$\lim_{\varepsilon \downarrow 0} \frac{\overline{H}(a(\varepsilon), 1-\varepsilon, \delta) - h_b(\delta)}{a(\varepsilon)} = \lim_{\varepsilon \downarrow 0} \frac{\overline{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta)}{a(\varepsilon)} = (1-2\delta) \log \frac{1-\delta}{\delta}. \tag{77}$$

Proof: The first equality in (77) follows trivially by symmetry, thus we turn to establish the second equality. Substituting into (53) we obtain

$$e^{I_2} = \frac{-1 + \alpha + a(\varepsilon) - \alpha(1-\varepsilon) + \sqrt{4\alpha a(\varepsilon)(1-\varepsilon) + (1 - \alpha - a(\varepsilon) + \alpha(1-\varepsilon))^2}}{2(1-\varepsilon)}, \tag{78}$$

where $\alpha = \frac{1-\delta}{\delta}$. It follows from (78) (using a first-order McLaurin approximation to $\sqrt{1+\varepsilon}$) that

$$\lim_{\varepsilon \downarrow 0} \frac{e^{I_2}}{a(\varepsilon)\frac{1-\delta}{\delta}} = 1. \tag{79}$$

It thus follows from the upper bound in (75) that for $\eta > 0$ and all sufficiently small $\varepsilon > 0$

$$\overline{H}(1-\varepsilon, a(\varepsilon), \delta) \leq h_b\left([(1+\eta)a(\varepsilon)] * \delta\right) \tag{80}$$

and, consequently,

$$\overline{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta) \quad \leq \quad h_b\left([(1+\eta)a(\varepsilon)] * \delta\right) - h_b(\delta). \tag{81}$$

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

$$\begin{aligned} h_b\left([(1+\eta)a(\varepsilon)] * \delta\right) - h_b(\delta) &= (1+\eta)a(\varepsilon)(1-2\delta)h_b'(\delta) + o(a(\varepsilon)) & (82)\\ &= (1+\eta)a(\varepsilon)(1-2\delta)\log\frac{1-\delta}{\delta} + o(a(\varepsilon)). & (83) \end{aligned}$$

Combining (81) with (83) gives

$$\limsup_{\varepsilon \downarrow 0} \frac{\overline{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta)}{a(\varepsilon)} \leq (1+\eta)(1-2\delta)\log\frac{1-\delta}{\delta}, \tag{84}$$

implying

$$\limsup_{\varepsilon \downarrow 0} \frac{\overline{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta)}{a(\varepsilon)} \leq (1-2\delta)\log\frac{1-\delta}{\delta} \tag{85}$$

by the arbitrariness of η. The inequality

$$\liminf_{\varepsilon \downarrow 0} \frac{\overline{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta)}{a(\varepsilon)} \geq (1-2\delta)\log\frac{1-\delta}{\delta} \tag{86}$$

is established similarly. □

B Entropy Rate in the "Rare-Bursts" Regime

The bounds in (73) 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 $\pi_{10} + \pi_{01} \leq 1$, $\pi_{01} \leq \pi_{10}$, and $I_2 \leq 0$ will be satisfied).

For this case we get

$$e^{I_2} = \frac{-1 + \alpha + \varepsilon - \alpha\pi_{10} + \sqrt{4\alpha\varepsilon\pi_{10} + (1 - \alpha - \varepsilon + \alpha\pi_{10})^2}}{2\pi_{10}}, \tag{87}$$

with $\alpha = \frac{1-\delta}{\delta}$. It follows via Taylor expansions from (87) that, as $\varepsilon \downarrow 0$,

$$e^{I_2} \sim \begin{cases} \frac{(1-\delta)\varepsilon}{\delta - (1-\delta)(1-\pi_{10})} & \text{for } 1 > \pi_{10} > \frac{1-2\delta}{1-\delta} \\ \sqrt{\frac{1-\delta}{\delta\pi_{10}}}\sqrt{\varepsilon} & \text{for } \pi_{10} = \frac{1-2\delta}{1-\delta} \\ \frac{-1 + \frac{1-\delta}{\delta}(1-\pi_{10})}{\pi_{10}} & \text{for } 0 < \pi_{10} < \frac{1-2\delta}{1-\delta}, \end{cases} \tag{88}$$

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

$$\operatorname{ess\,sup} \beta_i(1) \sim \begin{cases} \frac{(1-\delta)\varepsilon}{\delta - (1-\delta)(1-\pi_{10})} & \text{for } 1 > \pi_{10} > \frac{1-2\delta}{1-\delta} \\ \sqrt{\frac{1-\delta}{\delta\pi_{10}}}\sqrt{\varepsilon} & \text{for } \pi_{10} = \frac{1-2\delta}{1-\delta} \\ 1 - \frac{\delta\pi_{10}}{(1-2\delta)(1-\pi_{10})} & \text{for } 0 < \pi_{10} < \frac{1-2\delta}{1-\delta}. \end{cases} \tag{89}$$

Remark: Note, in particular, that

$$\lim_{\varepsilon \downarrow 0} \operatorname{ess\,sup} \beta_i(1) = \begin{cases} 0 & \text{for } \pi_{10} > \frac{1-2\delta}{1-\delta} \\ 1 - \frac{\delta\pi_{10}}{(1-2\delta)(1-\pi_{10})} & \text{for } \pi_{10} < \frac{1-2\delta}{1-\delta}. \end{cases} \tag{90}$$

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 $\approx (1-\delta)^l(1-\pi_{10})^l$. 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 (87) with $\alpha = \frac{\delta}{1-\delta}$, so

$$\operatorname{ess\,inf} \beta_i(1) \sim e^{I_1} \sim \frac{\delta\varepsilon}{(1-\delta) - \delta(1-\pi_{10})} \tag{91}$$

(since $\pi_{10} > \frac{2\delta-1}{\delta}$ always holds for $\delta \leq 1/2$). Combining (88), (91), and (73) leads to the following:

Theorem 6 *1. For $0 \leq \delta \leq 1/2$ and $0 < \pi_{10} < 1$*

$$\liminf_{\varepsilon \downarrow 0} \frac{\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\varepsilon} \geq \frac{(1-\delta)(1-2\delta)}{(1-\delta) - \delta(1-\pi_{10})} \log \frac{1-\delta}{\delta}. \tag{92}$$

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

$$\limsup_{\varepsilon \downarrow 0} \frac{\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\varepsilon} \leq \frac{\delta(1-2\delta)}{\delta - (1-\delta)(1-\pi_{10})} \log \frac{1-\delta}{\delta}. \tag{93}$$

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

$$\limsup_{\varepsilon \downarrow 0} \frac{\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\sqrt{\varepsilon}} \leq \sqrt{\frac{1-\delta}{\delta \cdot \pi_{10}}} (1-\pi_{10})(1-2\delta) \log \frac{1-\delta}{\delta}. \tag{94}$$

Proof of Theorem 6: Since the claim trivially holds for $\delta = 1/2$ assume $0 \leq \delta < 1/2$. From the left inequality in (73) and (91) it follows that for fixed $\eta > 0$ and all sufficiently small $\varepsilon > 0$

$$\begin{aligned} \overline{H}(\pi_{10}, \varepsilon, \delta) &\geq h_b\left(\left\{\left[\frac{\delta\varepsilon}{(1-\delta) - \delta(1-\pi_{10})}(1-\pi_{10}) + \varepsilon\right](1-\eta)\right\} * \delta\right) &(95)\\ &= h_b\left(\left\{\left[\frac{(1-\delta)(1-\eta)}{(1-\delta) - \delta(1-\pi_{10})}\right]\varepsilon\right\} * \delta\right). &(96)\end{aligned}$$

Applying a Taylor's expansion to h_b around δ, similarly as in (83), (96) gives

$$\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) \geq \frac{(1-\delta)(1-\eta)}{(1-\delta) - \delta(1-\pi_{10})} \varepsilon (1-2\delta) \log \frac{1-\delta}{\delta} + o(\varepsilon), \tag{97}$$

implying (92) by the arbitrariness of η.

The second item is proven similarly (using (88) instead of (91)).

For the third item note that from the right inequality in (73) and (88) it follows that when $\pi_{10} = \frac{1-2\delta}{1-\delta}$, for fixed $\eta > 0$ and all sufficiently small $\varepsilon > 0$,

$$\overline{H}(\pi_{10}, \varepsilon, \delta) \leq h_b\left(\left[\sqrt{\frac{1-\delta}{\delta\pi_{10}}}\sqrt{\varepsilon}(1-\pi_{10})(1+\eta)\right] * \delta\right). \tag{98}$$

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

Note that Theorem 6 implies, in particular:

1. For $0 \leq \delta \leq 1/2$ and $\frac{1-2\delta}{1-\delta} < \pi_{10} < 1$,
$$\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) \asymp \varepsilon. \tag{99}$$
2. For $0 \leq \delta \leq 1/2$ and $\pi_{10} = \frac{1-2\delta}{1-\delta}$, $\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) = O(\sqrt{\varepsilon})$ and $\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) = \Omega(\varepsilon)$.
3. For $0 \leq \delta \leq 1/2$ and $0 < \pi_{10} < \frac{1-2\delta}{1-\delta}$, $\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) = \Omega(\varepsilon)$.

It is the authors' conjecture that (99) holds for values of (δ, π_{10}) in the other two regions. Our proof technique, which sandwiches the entropy rate using (73), fails to give a non-trivial upper bound on $\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)$ in the region $0 \leq \delta \leq 1/2$ and $0 < \pi_{10} < \frac{1-2\delta}{1-\delta}$. In this region, since the third branch in (89) does not approach 0 as $\varepsilon \downarrow 0$, the upper bound in (73) would not even imply the trivial fact that $\limsup_{\varepsilon \downarrow 0} \overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta) \leq 0$.

Note also that Theorem 6 implies

$$\lim_{\pi_{10} \uparrow 1} \liminf_{\varepsilon \downarrow 0} \frac{\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\varepsilon} = \lim_{\pi_{10} \uparrow 1} \limsup_{\varepsilon \downarrow 0} \frac{\overline{H}(\pi_{10}, \varepsilon, \delta) - h_b(\delta)}{\varepsilon} = (1-2\delta) \log \frac{1-\delta}{\delta}, \tag{100}$$

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 $\pi \leq 1/2$ equation (69) implies

$$\overline{H}(\pi, \pi, \delta) = Eh_b\left(\beta_i(1) * \pi * \delta\right). \tag{101}$$

Equation (53) for this case implies

$$e^{I_1} = \frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi}, \tag{102}$$

with $\alpha = \delta/(1-\delta)$. Thus,

$$\operatorname{ess\,inf}\beta_i = \frac{e^{I_1}}{1+e^{I_1}} = \frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}. \tag{103}$$

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[15]. 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

$$\frac{e^{I_1+2\log\frac{1-\delta}{\delta}}}{1+e^{I_1+2\log\frac{1-\delta}{\delta}}} = \frac{\left(\frac{1-\delta}{\delta}\right)^2 e^{I_1}}{1+\left(\frac{1-\delta}{\delta}\right)^2 e^{I_1}} \tag{104}$$

implying, by symmetry, that the highest point of the support of $\beta_i(1)$ below $1/2$ is

$$1 - \frac{e^{I_1+2\log\frac{1-\delta}{\delta}}}{1+e^{I_1+2\log\frac{1-\delta}{\delta}}} = \frac{\alpha^2}{\alpha^2+e^{I_1}} = \frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}. \tag{105}$$

Summarizing, we obtain the following:

Theorem 7 *For all* $0 \leq \pi \leq 1/2$ *and* $0 \leq \delta \leq \frac{1}{2}(1-\sqrt{\max\{1-4\pi, 0\}})$

$$h_b\left(\frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} * \pi * \delta\right) \tag{106}$$

$$\leq \quad \overline{H}(\pi, \pi, \delta) \tag{107}$$

$$\leq \quad h_b\left(\frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} * \pi * \delta\right), \tag{108}$$

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

$$h_b(\pi * \delta) \leq \overline{H}(\pi, \pi, \delta) \leq h_b(\delta * \pi * \delta). \tag{109}$$

Evidently, the lower bound of Theorem 7 is always better than that in (109) . 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 below that there are asymptotic regimes (e.g., Corollary 5) where the bounds of Theorem 7 are tight while those in (109) are not. Similarly, there are regimes where the bounds of Theorem 7 would be tight, whereas bounds of the form $H(Z_0|Z_{-k}^{-1}, X_{-(k+1)}) \leq \overline{H} \leq H(Z_0|Z_{-k}^{-1})$, 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 $H(Z_0|Z_{-k}^{-1}, X_{-(k+1)}) \leq \overline{H}$ would not give the right order.

The High SNR Regime:

[15] The fact that $h(I_1) + \log\frac{1-\delta}{\delta} \geq 0$ follows from (38) and the optimality of the "say-what-you-see" scheme. 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.

Corollary 4 *For* $0 \le \pi \le 1/2$

$$(1-2\pi)[(1-\pi)\pi+1]\log\frac{1-\pi}{\pi} \le \liminf_{\delta\downarrow 0}\frac{\overline{H}(\pi,\pi,\delta)-h_b(\pi)}{\delta} \le \limsup_{\delta\downarrow 0}\frac{\overline{H}(\pi,\pi,\delta)-h_b(\pi)}{\delta} \le (1-2\pi)\frac{1+(1-\pi)\pi}{(1-\pi)\pi}\log\frac{1-\pi}{\pi}, \tag{110}$$

in particular,

$$\overline{H}(\pi,\pi,\delta)-h_b(\pi) \asymp \delta \quad as \quad \delta\to 0. \tag{111}$$

Proof: Using the Taylor expansion for $\sqrt{1+\varepsilon}$ gives

$$\frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} \sim (1-\pi)\pi\alpha \quad \text{as} \quad \alpha\downarrow 0 \tag{112}$$

and

$$\frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} \sim \frac{\alpha}{(1-\pi)\pi} \quad \text{as} \quad \alpha\downarrow 0. \tag{113}$$

Since $\alpha\sim\delta$ as $\delta\downarrow 0$ it follows from (112) that for fixed $\eta>0$ and all sufficiently small δ

$$\begin{aligned}
& h_b\left(\frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} * \pi * \delta\right) && (114)\\
\ge\ & h_b\left([(1-\eta)(1-\pi)\pi\delta] * \pi * \delta\right) && (115)\\
\ge\ & h_b\left([(1-2\eta)[(1-\pi)\pi+1]\delta] * \pi\right) && (116)\\
=\ & h_b\left(\pi+(1-2\pi)(1-2\eta)[(1-\pi)\pi+1]\delta\right) && (117)\\
=\ & h_b(\pi)+(1-2\pi)(1-2\eta)[(1-\pi)\pi+1]\delta h_b'(\pi)+o(\delta) && (118)\\
=\ & h_b(\pi)+(1-2\pi)(1-2\eta)[(1-\pi)\pi+1]\delta\log\frac{1-\pi}{\pi}+o(\delta), && (119)
\end{aligned}$$

implying the left inequality in (110) via (107) and the arbitrariness of η. The right inequality in (110) follows from (113) in an analogous way. □

It is easy to check that, for this regime, the bounds of (109) would give

$$(1-2\pi)\log\frac{1-\pi}{\pi} \le \liminf_{\delta\downarrow 0}\frac{\overline{H}(\pi,\pi,\delta)-h_b(\pi)}{\delta} \le \limsup_{\delta\downarrow 0}\frac{\overline{H}(\pi,\pi,\delta)-h_b(\pi)}{\delta} \le 2(1-2\pi)\log\frac{1-\pi}{\pi}, \tag{120}$$

which is a slightly better upper bound and a slightly worse lower bound than in (110) (but implies (111) just the same). The bounds in (120) and (110) are consistent with the main result of the recent work [JSS03], which established

$$\frac{\overline{H}(\pi_{10},\pi_{01},\delta)-\overline{H}(\pi_{10},\pi_{01},0)}{\delta} \sim \upsilon(\pi_{10},\pi_{01}), \tag{121}$$

explicitly identifying $\upsilon(\pi_{10},\pi_{01})$.

The "Almost Memoryless" Regime:

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

$$\lim_{\varepsilon\downarrow 0}\frac{1-\overline{H}\left(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,\delta\right)}{\varepsilon^2} = 4(1-2\delta)^4. \tag{122}$$

Note that $\overline{H}\left(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,0\right) = h_b(\frac{1}{2}-\varepsilon)$ so $\lim_{\varepsilon\downarrow 0}\frac{1-\overline{H}\left(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,0\right)}{\varepsilon^2} = 4$ (namely, (122) at $\delta=0$) would follow from a Taylor expansion of h_b around 1/2. Equality (122) also trivially holds at $\delta = 1/2$, as $\overline{H}\left(\frac{1}{2}-\varepsilon,\frac{1}{2}-\varepsilon,1/2\right) = 1$. The simple intuition behind (122) 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 $1 - \overline{H}\left(\frac{1}{2}-\varepsilon, \frac{1}{2}-\varepsilon, \delta\right) \sim 1 - h_b\left((\delta \pm o(1)) * (1/2-\varepsilon) * \delta\right) \sim \varepsilon^2 4(1-2\delta)^4$. More formally:

Proof of Corollary 5: At $\pi = 1/2$ we have

$$\frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} = \frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} = \delta, \tag{123}$$

where $\alpha = \delta/(1-\delta)$. The claim now follows by continuity of the expressions in (123) at $\pi = 1/2$, the relationship $\frac{1}{2} - (\delta+\xi) * (\frac{1}{2}-\varepsilon) * \delta = \varepsilon(1-2(2\delta-2\delta^2+\xi(1-2\delta))) = \varepsilon(1-2\delta)^2(1+O(\xi))$, Theorem 7, and the fact that $h_b(1/2-\varepsilon) = 1 - 4\varepsilon^2 + o(\varepsilon^2)$. □

For this regime, the bounds of (109) would imply

$$1 - \overline{H}\left(\frac{1}{2}-\varepsilon, \frac{1}{2}-\varepsilon, \delta\right) \asymp \varepsilon^2, \tag{124}$$

but not give the constant characterized in (122).

The Low SNR Regime:

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

$$4\left[\frac{(4\pi-1)(1-2\pi)}{\pi}\right]^2 \le \liminf_{\varepsilon\to 0} \frac{1-\overline{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right)}{\varepsilon^4} \le \limsup_{\varepsilon\to 0} \frac{1-\overline{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right)}{\varepsilon^4} \le 4\left[\frac{1-2\pi}{\pi}\right]^2. \tag{125}$$

In particular

$$1 - \overline{H}\left(\pi, \pi, \frac{1}{2}-\varepsilon\right) \asymp \varepsilon^4 \quad as \quad \varepsilon \to 0. \tag{126}$$

Note that the ratio between the upper and lower bound in (125) approaches 1 as $\pi \uparrow 1/2$. Also, it is not hard to see that for any k, a bound of the form $H(Z_0|Z_{-k}^{-1}, X_{-(k+1)}) \le \overline{H}$ would lead to $1 - \overline{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right) = O(\varepsilon^2)$, failing to capture the true ε^4 behavior.

Proof of Corollary 6: Since $\overline{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right) = \overline{H}\left(\pi,\pi,\frac{1}{2}+\varepsilon\right)$, we may assume the limits in (125) are taken along $\varepsilon \downarrow 0$. Letting $\delta = 1/2 - \varepsilon$, it is straightforward to show using a Taylor expansion that

$$\frac{1}{2} - \frac{-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}}{2\pi-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} \sim \frac{1}{2\pi}\varepsilon \tag{127}$$

and that

$$\frac{1}{2} - \frac{2\pi\alpha^2}{2\pi\alpha^2-1+\alpha+\pi-\alpha\pi+\sqrt{4\alpha\pi^2+(1-\alpha-\pi+\alpha\pi)^2}} \sim \left(2-\frac{1}{2\pi}\right)\varepsilon. \tag{128}$$

It thus follows from Theorem 7 that for every $1/4 \le \pi \le 1/2$ and $\varepsilon > 0$

$$1 - h_b\left(\left[\frac{1}{2} - \left(2-\frac{1}{2\pi}\right)\varepsilon\right] * \pi * \left(\frac{1}{2}-\varepsilon\right)\right) \tag{129}$$

$$\stackrel{\sim}{<} \quad 1 - \overline{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right) \tag{130}$$

$$\stackrel{\sim}{<} \quad 1 - h_b\left(\left[\frac{1}{2} - \frac{1}{2\pi}\varepsilon\right] * \pi * \left(\frac{1}{2}-\varepsilon\right)\right). \tag{131}$$

The claim now follows by

$$\left[\frac{1}{2} - \left(2-\frac{1}{2\pi}\right)\varepsilon\right] * \pi * \left(\frac{1}{2}-\varepsilon\right) = \frac{1}{2} - \frac{(4\pi-1)(1-2\pi)}{\pi}\varepsilon^2, \tag{132}$$

$$\left[\frac{1}{2}-\frac{1}{2\pi}\varepsilon\right]*\pi*\left(\frac{1}{2}-\varepsilon\right)=\frac{1}{2}-\frac{1-2\pi}{\pi}\varepsilon^2, \tag{133}$$

and the fact that $h_b(1/2-\varepsilon)=1-4\varepsilon^2+o(\varepsilon^2)$. □

In this regime, the bounds of (109) become

$$h_b(\pi*(1/2-\varepsilon))\le\overline{H}(\pi,\pi,\delta)\le h_b(\pi*(1/2-\varepsilon)*(1/2-\varepsilon)). \tag{134}$$

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

$$\liminf_{\varepsilon\to 0}\frac{1-\overline{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right)}{\varepsilon^4}\ge 16(1-2\pi)^2, \tag{135}$$

which gives a slightly better constant than the left side of (125). The lower bound in (134), however, would only imply $1-\overline{H}\left(\pi,\pi,\frac{1}{2}-\varepsilon\right)\overset{\sim}{<}\varepsilon^2$.

6 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 non-essential technicalities and simplify notation we assume both that the components of the clean and noisy process take values in the same finite alphabet $\mathcal{A}$, 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 $z(T)$

$$P(x_t|z(T))=\frac{P(x_t,z_t|z(T\setminus t))}{P(z_t|z(T\setminus t))}=\frac{P(x_t|z(T\setminus t))P(z_t|x_t,z(T\setminus t))}{P(z_t|z(T\setminus t))}, \tag{136}$$

so that

$$\log\frac{P(X_t=a|z(T))}{P(X_t=b|z(T))}=\log\frac{P(z_t|X_t=a,z(T\setminus t))}{P(z_t|X_t=b,z(T\setminus t))}+\log\frac{P(X_t=a|z(T\setminus t))}{P(X_t=b|z(T\setminus t))}. \tag{137}$$

Note that for a memoryless channel C, (137) particularizes to

$$\log\frac{P(X_t=a|z(T))}{P(X_t=b|z(T))}=\log\frac{C(z_t|a)}{C(z_t|b)}+\log\frac{P(X_t=a|z(T\setminus t))}{P(X_t=b|z(T\setminus t))}. \tag{138}$$

By observation of (137) 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 $a,b\in\mathcal{A}$ the sign of the right side of (137) 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 $\left\{\operatorname{ess\,sup}\log\frac{P(X_t=a|Z(T\setminus t))}{P(X_t=b|Z(T\setminus t))}\right\}_{a,b\in\mathcal{A}}$. While for the binary Markov chain these value were obtainable in closed form (Section 4), 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 below.

Returning to the general setting,

$$P(x_t = a|z(T \setminus t)) = \sum_{x(T\setminus t)} P(x_t = a, x(T \setminus t)|z(T \setminus t)) \tag{139}$$

$$= \sum_{x(T\setminus t)} P(x_t = a|x(T \setminus t), z(T \setminus t))P(x(T \setminus t)|z(T \setminus t)) \tag{140}$$

$$= \sum_{x(T\setminus t)} P(x_t = a|x(T \setminus t))P(x(T \setminus t)|z(T \setminus t)), \tag{141}$$

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

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

$$\max_{z(T\setminus t)} \frac{P(x_t = a|z(T \setminus t))}{P(x_t = b|z(T \setminus t))} \leq \max_{x(T\setminus t)} \frac{P(x_t = a|x(T \setminus t))}{P(x_t = b|x(T \setminus t))}. \tag{142}$$

Proof:

$$\frac{P(x_t = a|z(T \setminus t))}{P(x_t = b|z(T \setminus t))} = \frac{\sum_{x(T\setminus t)} P(x_t = a|x(T \setminus t))P(x(T \setminus t)|z(T \setminus t))}{\sum_{x(T\setminus t)} P(x_t = b|x(T \setminus t))P(x(T \setminus t)|z(T \setminus t))} \tag{143}$$

$$= \sum_{x(T\setminus t)} \frac{P(x_t = b|x(T \setminus t))P(x(T \setminus t)|z(T \setminus t))}{\left[\sum_{x'(T\setminus t)} P(x_t = b|x'(T \setminus t))P(x'(T \setminus t)|z(T \setminus t))\right]} \cdot \frac{P(x_t = a|x(T \setminus t))}{P(x_t = b|x(T \setminus t))} \tag{144}$$

$$\geq \left[\sum_{x(T\setminus t)} \frac{P(x_t = b|x(T \setminus t))P(x(T \setminus t)|z(T \setminus t))}{\left[\sum_{x'(T\setminus t)} P(x_t = b|x'(T \setminus t))P(x'(T \setminus t)|z(T \setminus t))\right]} \cdot \frac{P(x_t = b|x(T \setminus t))}{P(x_t = a|x(T \setminus t))}\right]^{-1}, \tag{145}$$

where the first equality follows from (141) and the inequality follows from Jensen's inequality (and convexity of $1/x$ for $x > 0$). Thus we get

$$\frac{P(x_t = b|z(T \setminus t))}{P(x_t = a|z(T \setminus t))} \leq \sum_{x(T\setminus t)} \frac{P(x_t = b|x(T \setminus t))P(x(T \setminus t)|z(T \setminus t))}{\left[\sum_{x'(T\setminus t)} P(x_t = b|x'(T \setminus t))P(x'(T \setminus t)|z(T \setminus t))\right]} \cdot \frac{P(x_t = b|x(T \setminus t))}{P(x_t = a|x(T \setminus t))} \tag{146}$$

$$\leq \max_{x(T\setminus t)} \frac{P(x_t = b|x(T \setminus t))}{P(x_t = a|x(T \setminus t))}, \tag{147}$$

implying (142) 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.

Claim 5 *Let T be an arbitrary index set and suppose for each $z_t \in \mathcal{A}$ there exists $a = a(z_t) \in \mathcal{A}$ such that for all $b \neq a$*

$$\operatorname{ess\,inf} \frac{P(z_t|X_t = a(z_t), Z(T \setminus t))}{P(z_t|X_t = b, Z(T \setminus t))} \geq \operatorname{ess\,sup} \frac{P(x_t = b|X(T \setminus t))}{P(x_t = a(z_t)|X(T \setminus t))}. \tag{148}$$

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

$$\hat{X}_t(Z(T)) = a(Z_t). \tag{149}$$

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

$$\min_{z(T\setminus t)} \frac{P(z_t|X_t = a, z(T \setminus t))}{P(z_t|X_t = b, z(T \setminus t))} \geq \max_{x(T\setminus t)} \frac{P(x_t = b|x(T \setminus t))}{P(x_t = a|x(T \setminus t))}, \tag{150}$$

then the estimate in (149) is an optimal estimate of X_t based on $Z(T)$. To see this note that if (150) holds then, for all $z(T)$, $a = a(z_t) \in \mathcal{A}$, and all $b \neq a$,

$$\begin{aligned} \frac{P(z_t|X_t = a, z(T \setminus t))}{P(z_t|X_t = b, z(T \setminus t))} &\geq \min_{z'(T\setminus t)} \frac{P(z_t|X_t = a, z'(T \setminus t))}{P(z_t|X_t = b, z'(T \setminus t))} &(151)\\ &\geq \max_{x(T\setminus t)} \frac{P(x_t = b|x(T \setminus t))}{P(x_t = a|x(T \setminus t))} &(152)\\ &\geq \frac{P(X_t = b|z(T \setminus t))}{P(X_t = a|z(T \setminus t))}, &(153) \end{aligned}$$

where (152) is due to (150) and (153) to (142). This implies, by (137), that $\log \frac{P(X_t=a|z(T))}{P(X_t=b|z(T))} \geq 0$ for all $b \neq a$ implying, in turn, that the optimal estimate of X_t based on $z(T)$ is $a = a(z_t)$. □

In what follows we illustrate the use of Claim 5 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 $|\mathcal{A}| - 1$ erroneous symbols. For this case we have for $b \neq a$ and $z_t = a$

$$\operatorname{ess\,inf} \frac{P(z_t|X_t = a, Z(T \setminus t))}{P(z_t|X_t = b, Z(T \setminus t))} = (|\mathcal{A}| - 1)\frac{1-\delta}{\delta}, \tag{154}$$

so Claim 5 implies

Corollary 7 *If*

$$(|\mathcal{A}| - 1)\frac{1-\delta}{\delta} \geq \max_{a,b\in\mathcal{A}} \operatorname{ess\,sup} \frac{P(x_t = b|X(T \setminus t))}{P(x_t = a|X(T \setminus t))} \tag{155}$$

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

The right side of (155) 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 [Geo88, Guy95]. A potential is *summable* if

$$\|\Phi\|_t \stackrel{\triangle}{=} \sum_{A\in\mathcal{S}, A\ni t} \|\Phi_A\|_\infty < \infty, \quad \forall t. \tag{156}$$

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

$$\operatorname{ess\,sup} \frac{P(x_t = b|X(T \setminus t))}{P(x_t = a|X(T \setminus t))} \leq e^{2\|\Phi\|_t}. \tag{157}$$

Combining Corollary 7 with (157) gives

Corollary 8 *The optimal estimate of X_t based on $Z(T)$ is Z_t if $\delta \leq \left(\frac{1}{|\mathcal{A}|-1}e^{2\|\Phi\|_t}+1\right)^{-1}$. In particular, singlet decoding with "say-what-you-see" is an optimal denoiser whenever $\delta \leq \left(\frac{1}{|\mathcal{A}|-1}e^{2\|\Phi\|_{max}}+1\right)^{-1}$, 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 for fixed $\delta < \frac{|\mathcal{A}|-1}{\mathcal{A}}$ and a potential satisfying $\|\Phi\|_{max} < \infty$, singlet decoding is optimal denoising for the field associated with $\beta\phi$ whenever $\beta \leq \frac{\log[(1/\delta-1)(|\mathcal{A}|-1)]}{2\|\Phi\|_{max}}$ (i.e., at sufficiently high temperatures [Geo88, Guy95]).

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

$$R(X(T)) \stackrel{\triangle}{=} \max_{a,b\in\mathcal{A}} \operatorname{ess\,sup} \frac{P(X_0 = b|X_{-\infty}^{-1})}{P(X_0 = a|X_{-\infty}^{-1})} \tag{158}$$

and

$$S(X(T)) \stackrel{\triangle}{=} \max_{a,b\in\mathcal{A}} \operatorname{ess\,sup} \frac{P(X_0 = b|X_{-\infty}^{-1}, X_1^{\infty})}{P(X_0 = a|X_{-\infty}^{-1}, X_1^{\infty})} \tag{159}$$

Corollary 7 implies

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

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

$$R(X(T)) = \max_{ab,x_{-k}^{-1}} \frac{P(X_0 = a|X_{-k}^{-1} = x_{-k}^{-1})}{P(X_0 = b|X_{-k}^{-1} = x_{-k}^{-1})} > 0 \quad \text{and} \quad S(X(T)) = \max_{ab,x_{-k}^{-1},x_1^k} \frac{P(X_0 = a|X_{-k}^{-1} = x_{-k}^{-1}, X_1^k = x_1^k)}{P(X_0 = b|X_{-k}^{-1} = x_{-k}^{-1}, X_1^k = x_1^k)} > 0 \tag{160}$$

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 figure 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 $\delta = \pi$. Figure 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 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 $\mathcal{A}$. Denote the DMC by W, i.e., $W(a|u) = \Pr(X_t = a|U_t = u)$. Thus we have, assuming first T finite,

$$P(X_t = a|X(T\setminus t) = x(T\setminus t)) = \sum_u P(U_t = u|X(T\setminus t) = x(T\setminus t))W(a|u). \tag{161}$$

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

$$\frac{P(X_t = a|X(T \setminus t) = x(T \setminus t))}{P(X_t = b|X(T \setminus t) = x(T \setminus t))} \leq \max_{u \in \mathcal{U}} \frac{W(a|u)}{W(b|u)} \tag{162}$$

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

$$\operatorname{ess\,sup} \frac{P(X_t = a|X(T \setminus t))}{P(X_t = b|X(T \setminus t))} \leq \max_{u \in \mathcal{U}} \frac{W(a|u)}{W(b|u)}. \tag{163}$$

Combined with Corollary 7 this gives

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*

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

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. [Ber71], [CT91, Ex. 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., [HB87, Gra70, Gra71, WM03] 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 Auto-Regressive Source: Let $T = \mathbb{Z}$, $\mathcal{A} = \{0, 1, \ldots, M-1\}$, and suppose the noiseless process can be represented by

$$X_t = g_t(X^{t-1}) \oplus W_t, \tag{164}$$

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

$$P(X_t = a|X^{t-1} = x^{t-1}) = \Pr(W_t \oplus g_t(x^{t-1}) = a) \tag{165}$$

so, for $a \neq b$,

$$\frac{P(X_t = a|X^{t-1} = x^{t-1})}{P(X_t = b|X^{t-1} = x^{t-1})} \leq \max_{m \in \mathcal{A}} \frac{P(W_t = m)}{P(W_t = (b-a) \oplus m)} \leq \max_{a \neq 0, m \in \mathcal{A}} \frac{P(W_t = m)}{P(W_t = a \oplus m)}. \tag{166}$$

Applied to this setting, and combined with (166), Corollary 7 gives

Corollary 11 *Let $\{X_t\}$ be given by (164), where $\{W_t\}$ is an i.i.d. sequence. The "say-what-you-see" scheme is an optimal filter provided*

$$\delta \leq \left(\frac{1}{|\mathcal{A}| - 1} \max_{a \neq 0, m \in \mathcal{A}} \frac{P(W_t = m)}{P(W_t = a \oplus m)} + 1 \right)^{-1}. \tag{167}$$

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

$$P(W_t = a) = \begin{cases} 1 - \varepsilon & \text{for } a = 0 \\ \varepsilon/(|\mathcal{A}| - 1) & \text{for } a \neq 0, \end{cases} \tag{168}$$

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 [MBD89]. 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

$$P(z_t|X_t = a, Z(T\backslash t)) = \sum_{s_t} P(z_t|X_t = a, s_t, Z(T\backslash t))P(s_t|X_t = a, Z(T\backslash t)) = \sum_{s_t} P(z_t|X_t = a, s_t)P(s_t|X_t = a, Z(T\backslash t)) \tag{169}$$

and thus, by an argument similar to that proving Lemma 1, we obtain for $a \neq b$ and $z_t = a$

$$\text{ess inf} \frac{P(z_t|X_t = a, Z(T \setminus t))}{P(z_t|X_t = b, Z(T \setminus t))} \geq \min_{s_t \in \{G,B\}} \frac{P(z_t|X_t = a, s_t)}{P(z_t|X_t = b, s_t)} = \frac{1 - \delta_b}{\delta_b}. \tag{170}$$

A similar argument implies an inequality like (170) where in the left side the conditioning is on the one-sided $Z_{-\infty}^{t-1}$ instead of on $Z(T \setminus t)$. Combining (170) (and its analogue for the one-sided conditioning) with Claim 5 gives

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

$$\frac{1 - \delta_b}{\delta_b} \geq \max_{b \neq a} \text{ess sup} \frac{P(x_t = b|X(T \setminus t))}{P(x_t = a|X(T \setminus t))}. \tag{171}$$

It is also an optimal filter provided

$$\frac{1 - \delta_b}{\delta_b} \geq \max_{b \neq a} \text{ess sup} \frac{P(x_t = b|X_{-\infty}^{t-1})}{P(x_t = a|X_{-\infty}^{t-1})}. \tag{172}$$

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 side of (171) and (172) would be replaced by $(|\mathcal{A}| - 1)\frac{1-\delta_b}{\delta_b}$). 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 $\frac{P(z_t|X_t=a,Z(T\backslash t))}{P(z_t|X_t=b,Z(T\backslash t))}$ can be lower bounded leading, via Claim 5, to sufficient conditions for the optimality of symbol-by-symbol schemes in denoising and filtering of various other processes and fields.

7 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 $\mathcal{A}$, and that $Z(T)$ is the output of a DMC C whose input is $X(T)$ with channel output alphabet $\mathcal{B}$.

Using standard large deviations theory [DZ98] or the method of types [CK81, CT91] it is straightforward to show that for every $f : \mathcal{B} \to \mathcal{A}$ and $x^n \in \mathcal{A}^n$

$$\Pr\left(\frac{1}{n} \sum_{t=1}^{n} \Lambda(X_t, f(Z_t)) \geq d \middle| X^n = x^n \right) \approx \exp\left(-nJ(P_{x^n}, d)\right), \tag{173}$$

where

$$J(P,d) = \min_{Q: E_{P\otimes Q}\Lambda(X,f(Z))\geq d} D(Q\|C|P), \tag{174}$$

with $D(Q\|C|P)$ denoting the conditional divergence (cf., e.g., Section 2 of [CK81]) 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., [DK99, MK03, WM02, Wei02] for proofs of results in this spirit) that for any individual sequence $\mathbf{x} = \{x_t\}$

$$\left| -\frac{1}{n}\log \Pr\left(\frac{1}{n}\sum_{t=1}^{n} \Lambda(X_t, f(Z_t)) \geq d \middle| X^n = x^n \right) - J(P_{x^n}, d) \right| \longrightarrow 0. \tag{175}$$

This exponent can also be given in the form (cf., e.g., [DK99, Prop. 1]):

$$J(P,d) = \sup_{\lambda\in\mathbb{R}} \left\{ \lambda d - \sum_{a\in\mathcal{A}} \left[\log \sum_{b\in\mathcal{B}} e^{\lambda\Lambda(a,f(b))} C(a,b) \right] P(a) \right\}. \tag{176}$$

It follows from (175) that if the empirical measure associated with $X(T)$ satisfies an LDP [DZ98] with the rate function I then

$$-\frac{1}{n}\log \Pr\left(\frac{1}{n}\sum_{t=1}^{n} \Lambda(X_t, f(Z_t)) \geq d \right) \longrightarrow \min_{P\in\mathcal{M}(\mathcal{A})} \left[I(P) + \min_{Q: E_{P\otimes Q}\Lambda(X,f(Z))\geq d} D(Q\|C|P) \right]. \tag{177}$$

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. [DZ98] for wide range of processes for which this is the case). In particular, the error exponent is given by the right side of (177) with f being the filtering function associated with the optimal scheme.

8 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 end points 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 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 7. The first regards 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 $l > 1$, to characterize 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 [Won65]), and whether there exist non-trivial situations where such singlet schemes are optimal.

Appendix

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 $P\left(f(b) \in \hat{X}(G_b(\beta_{i-1}, \gamma_{i+1}))\ \forall b \in \mathcal{B}\right) = P((\beta_{i-1}, \gamma_{i+1}) \in R_f) = 1$. Consequently, $1 = P\left(f(Z_i) \in \hat{X}(G_{Z_i}(\beta_{i-1}, \gamma_{i+1}))\right) = P\left(f(Z_i) \in \hat{X}(\eta_i)\right)$, establishing optimality by (31).

Conversely, suppose that $C_Q \times C_{Q_r} \not\subseteq R_f$. Then there exists $J \subseteq \mathcal{M}(\mathcal{A}) \times \mathcal{M}(\mathcal{A})$ such that $J \cap R_f = \emptyset$ and $P\left((\beta_{i-1}, \gamma_{i+1}) \in J\right) > 0$. This implies that $P\left((\beta_{i-1}, \gamma_{i+1}) \in R_f\right) = P\left(f(b) \in \hat{X}(G_b(\beta_{i-1}, \gamma_{i+1}))\ \forall b \in \mathcal{B}\right) < 1$, which implies the existence of $b \in \mathcal{B}$ with $P\left(f(b) \in \hat{X}(G_b(\beta_{i-1}, \gamma_{i+1}))\right) < 1$ implying, in turn, the existence of $a \in \mathcal{A}$ such that

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

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

$$P\left(f(Z_i) \in \hat{X}(\eta_i)|X_i = a\right) = P\left(f(Z_i) \in \hat{X}(G_b(\beta_{i-1}, \gamma_{i+1}))|X_i = a\right) = \sum_{b' \in \mathcal{B}} P\left(f(b') \in \hat{X}(G_{b'}(\beta_{i-1}, \gamma_{i+1}))|X_i = a\right) C(a, b'). \tag{A.2}$$

Inequality (A.1), combined with (A.2) and the fact that $C(a, b) > 0$ leads to $P\left(f(Z_i) \in \hat{X}(\eta_i)|X_i = a\right) < 1$ implying $P\left(f(Z_i) \in \hat{X}(\eta_i)\right) < 1$ and establishing the fact that (31) is not satisfied by $\hat{X}_i(Z_{-\infty}^{\infty}) = f(Z_i)$. □

References

[ABK00] A. Budhiraja A. Bhatt and R. Karandikar. Markov property and ergodicity of the nonlinear filter. *SIAM J. on Control and Optimization*, (39):928–949, 2000.

[ADG94] L. Arnold, L. Demetrius, and M. Gundlach. Evolutionary formalism for products of positive random matrices. *Annals of Applied Probability*, (4):859–901, 1994.

[AZ97] R. Atar and O. Zeitouni. Exponential stability for nonlinear filtering. *Annales de lInstitut H. Poincare Probabilites et Statistique*, (33):697725, 1997.

[Ber71] T. Berger. *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Englewood Cliffs, N.J., 1971.

[Bla57] D. Blackwell. The entropy of functions of finite-state markov chains. *Trans. First Prague Conf. Inf. Th., Statistical Decision Functions, Random Processes*, pages 13–20, 1957.

[CK81] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York, 1981.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[Dev74] J.L. Devore. A note on the observation of a markov source through a noisy channel. *IEEE Trans. Inform. Theory*, 20:762–764, November 1974.

[DK99] A. Dembo and I. Kontoyiannis. The asymptotics of waiting times between stationary processes, allowing distortion. *Ann. Appl. Probab.*, (9):413–429, 1999.

[Dra65] A. W. Drake. Observation of a Markov source through a noisy channel. In *IEEE Symp. on Signal Transmission and Processing*, pages 12–18, New York, 1965. Columbia Univ.

[DW03] A. Dembo and T. Weissman. Universal denoising for the finite-input-continuous-output channel. *Submitted*, October 2003. (available at: `http://www.stanford.edu/~tsachy/interest.htm`).

[DZ98] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer-Verlag, New York, 2nd edition, 1998.

[EM02] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, June 2002.

[Geo88] H. O. Georgii. *Gibbs Measures and Phase Transitions*. Walter de Gruyter, Berlin - New York, 1988.

[Gra70] R. M. Gray. Information rates of autoregressive processes. *IEEE Trans. Info. Theory*, IT-16:412–421, July 1970.

[Gra71] R. M. Gray. Rate distortion functions for finite-state finite-alphabet markov sources. *IEEE Trans. Inform. Theory*, IT-17(2):127–134, March 1971.

[GRV03] M. Gastpar, B. Rimoldi, and M. Vetterli. To code or not to code: Lossy source-channel communication revisited. *IEEE Trans. Inform. Theory*, 49(5):1147–1158, May 2003.

[Guy95] X. Guyon. *Random Fields on a Network*. Springer-Verlag, New York, 1995.

[GV96] A. Goldsmith and P. Varaiya. Capacity, mutual information, and coding for finite state Markov channels. *IEEE Trans. Information Theory*, 42:868–886, May 1996.

[Han57] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games, Ann. Math. Study*, III(39):97–139, 1957. Princeton University Press.

[HB87] B. E. Hajek and T. Berger. A decomposition theorem for binary Markov random fields. *The Annals of Probabiilty*, (15):1112–1125, 1987.

[HGG03] T. Holliday, P. Glynn, and A. Goldsmith. On entropy and Lyapunov exponents for finite state channels. *Submitted to IEEE Transactions on Information Theory*, 2003. (available at: `http://wsl.stanford.edu/Publications/THolliday/Lyapunov.pdf`).

[JSS03] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden Markov process. Preprint, 2003.

[Kal80] G. Kallianpur. *Stochastic filtering theory*. Springer-Verlag, New York, 1980.

[Kun71] H. Kunita. Asymptotic behavior of the nonlinear filtering errors of markov processes. *Journal of Multivariate Analysis*, (1):365–393, 1971.

[KZ96] R. Khasminskii and O. Zeitouni. Asymptotic filtering for finite state Markov chains. *Stochastic Processes and the Applications*, 63:1–10, 19996.

[MBD89] M. Mushkin and I. Bar-David. Capacity and coding for the GilbertElliot channel. *IEEE Trans. Inform. Theory*, November 1989.

[MF98] N. Merhav and M. Feder. Universal prediction. *IEEE Trans. Inform. Theory*, IT-44(6):2124–2147, October 1998.

[MK03] N. Merhav and I. Kontoyiannis. Source coding exponents for zero-delay coding with finite memory. *IEEE Transactions on Information Theory*, (49):609–625, March 2003.

[NG82] D. L. Neuhoff and R. K. Gilbert. Causal source codes. *IEEE Trans. Inform. Theory*, IT-28(5):701–713, September 1982.

[Per] Y. Peres.

[Sag70] D. Sagalowicz. *Hypothesis testing with finite memory*. PhD thesis, Stanford Univ, Elec. Eng. Dep., 1970.

[Sam63] E. Samuel. An empirical Bayes approach to the testing of certain parametric hypotheses. *Ann. Math. Statist.*, 34(4):1370–1385, 1963.

[SDAB01] 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 Processing*, 49(1):202–208, January 2001.

[Wei02] T. Weissman. Universally attainable error-exponents for rate-constrained denoising of noisy sources. *HP Laboratories Technical Report*, HPL-2002-214, August 2002. Also submitted to IEEE Trans. Inf. Th.

[WM02] T. Weissman and N. Merhav. Tradeoffs between the excess-code-length exponent and the excess-distortion exponent in lossy source coding. *IEEE Trans. Inform. Theory*, IT-48(2):396–415, February 2002.

[WM03] T. Weissman and N. Merhav. On competitive prediction and its relationship to rate-distortion theory. November 2003. to appear in IEEE Trans. Inform. Theory.

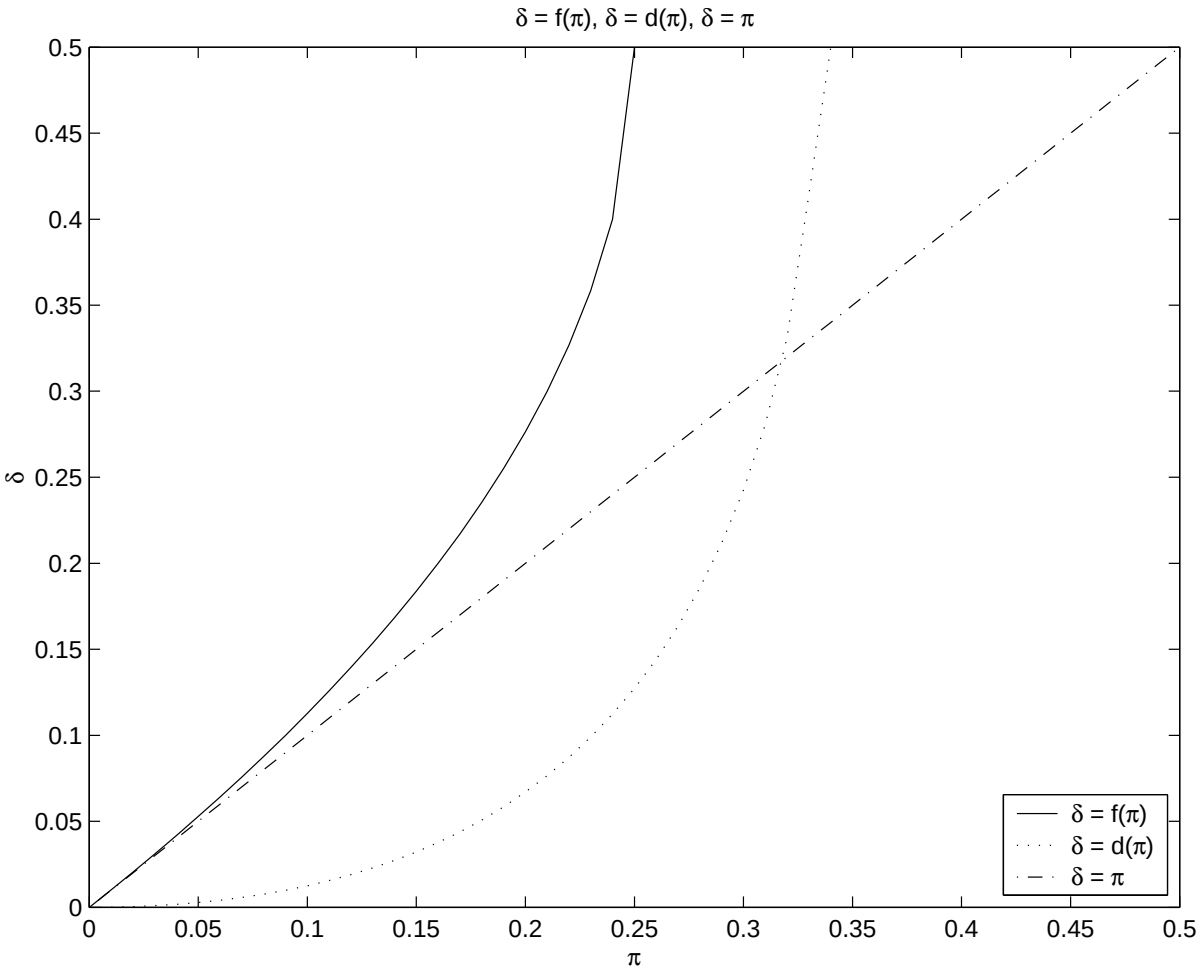

Figure 1: Optimality region in $\delta - \pi$ plane for singlet decoding of a binary symmetric Markov chain with transition probability π corrupted by a BSC(δ) : $\delta \leq f(\pi) = \frac{1}{2}(1 - \sqrt{\max\{1-4\pi, 0\}})$ for filtering and $\delta \leq d(\pi) = \frac{1}{2}\left(1 - \sqrt{\max\left\{1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

[Won65] W. M. Wonham. Some applications of stochastic differential equations to optimal nonlinear filtering. *SIAM J. Control Optim.*, 2:347–368, 1965.

[WOS+03a] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger. Universal discrete denoising: Known channel. 2003. Submitted to IEEE Trans. Inform. Theory (available at: `http://www.hpl.hp.com/techreports/2003/HPL-2003-29.pdf`).

[WOS+03b] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger. Universal discrete denoising: Known channel. In *IEEE Int. Symp. Information Theory*, Yokohama, Japan, June 29 - July 4 2003.

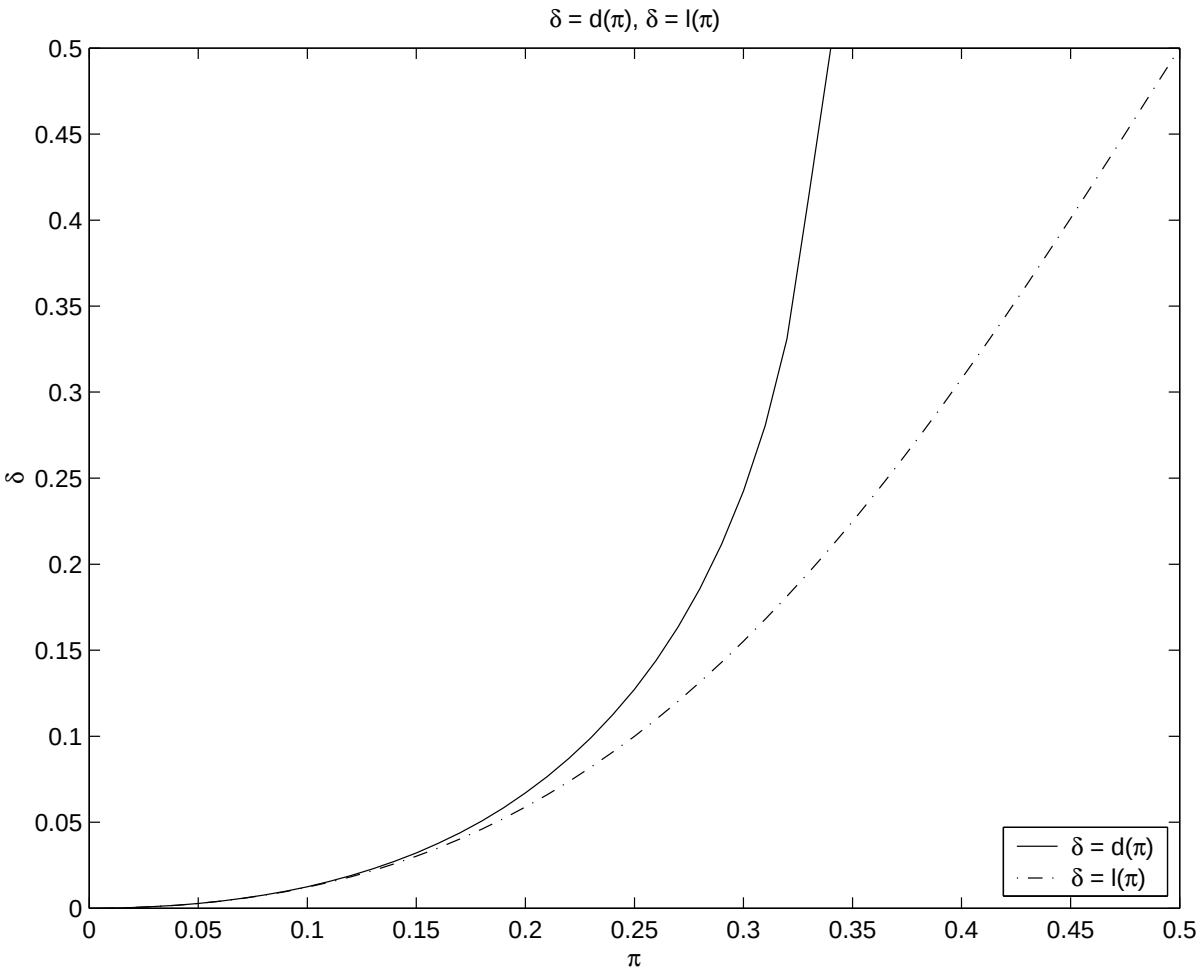

Figure 2: Optimality region for singlet denoising: Solid line is the curve $d(\pi) = \frac{1}{2}\left(1 - \sqrt{\max\left\{1 - 4\left(\frac{\pi}{1-\pi}\right)^2, 0\right\}}\right)$ giving the precise region. Dashed line is the $l(\pi) = \frac{\pi^2}{1-2\pi+2\pi^2}$ curve associated with the sufficient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