January 12, 2004
Submitted to IEEE Trans. on Information Theory

Simulation-Based Computation of Information Rates for Channels with Memory

Dieter Arnold, Hans-Andrea Loeliger,
Pascal O. Vontobel, Aleksandar Kavčić, and Wei Zeng

Abstract—The information rate of finite-state source/channel models can be accurately estimated by sampling both a long channel input sequence and the corresponding channel output sequence, followed by a forward sum-product recursion on the joint source/channel trellis. This method is extended to compute upper and lower bounds on the information rate of very general channels with memory by means of finite-state approximations. Further upper and lower bounds can be computed by reduced-state methods.

Keywords—Bounds, channel capacity, finite-state models, hidden-Markov models, information rate, sum-product algorithm, trellises.

1 Introduction

We consider the problem of computing the information rate

$$I(X;Y) \stackrel{\triangle}{=} \lim_{n\to\infty} \frac{1}{n} I(X_1,\ldots,X_n;Y_1,\ldots,Y_n) \tag{1}$$

between the input process $X = (X_1, X_2, \ldots)$ and the output process $Y = (Y_1, Y_2, \ldots)$ of a time-invariant discrete-time channel with memory. We will assume that X is Markov or hidden Markov, and we will primarily be interested in the case where the channel input alphabet $\mathcal{X}$ (i.e., the set of possible values of X_k) is finite.

In many cases of practical interest, the computation of (1) is a problem. Analytical simplifications of (1) are usually not available even if the input symbols X_k are i.u.d.

D. Arnold was with the Dept. of Information Technology and Electrical Engineering, ETH Zürich, Switzerland, and with the IBM Research Lab., Rüschlikon, Zürich, Switzerland; he is now with Seagate Technology LLC, Pittsburgh, PA 15222-4215, USA. Email: Dieter.M.Arnold@seagate.com.

H.-A. Loeliger is with the Dept. of Information Technology and Electrical Engineering, ETH, CH-8092 Zürich, Switzerland. Email: loeliger@isi.ee.ethz.ch.

P. O. Vontobel is with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. Email: vontobel@ifp.uiuc.edu. Supported by NSF Grants CCR 99-84515 and CCR 01-05719.

A. Kavčić and W. Zeng are with the Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. Email: {kavcic, wzeng}@deas.harvard.edu.

Parts of this work were presented at several conferences, see [3], [4], [6], [7].

(independent and uniformly distributed). The complexity of the direct numerical computation of

$$I_n \triangleq \frac{1}{n} I(X_1, \ldots, X_n; Y_1, \ldots, Y_n) \tag{2}$$

is exponential in n, and the sequence $I_1, I_2, I_3, \ldots$ converges rather slowly even for very simple examples.

Prior work on this subject includes investigations of (i) linear intersymbol interference channels, (ii) generalizations of the Gilbert-Elliott channel, and (iii) channels with constrained input, cf. the examples in Section 2. The binary-input linear intersymbol interference channel was investigated by Hirt [19], who proposed a Monte-Carlo method to evaluate certain quantities closely related to the i.u.d. information rate (cf. Section 4). Shamai et al. [31], [32] also investigated the intersymbol interference channel and derived various closed-form bounds on the capacity and on the i.u.d. information rate as well as a lower-bound conjecture.

The Gilbert-Elliott channel was analyzed by Mushkin and Bar-David [26]. Goldsmith and Varaiya extended that work to general channels with a freely evolving state [16] (cf. Example 2); they gave expressions for the channel capacity and the information rate as well as recursive methods for their evaluation.

Zehavi and Wolf studied the binary symmetric channel with run-length limited input [40]; they derived a set of lower bounds for Markovian input and demonstrated some numerical results from brute-force computations. Both the binary symmetric channel and the Gaussian channel with run-length limited binary input were studied by Shamai and Kofman, who obtained upper and lower bounds on the i.u.d. information rate [30]. A related topic is the continuous-time AWGN channel with peak constraint input, which was addressed by Heegard et al. [17], [18].

Despite all this work, information rates of such channels could not be computed accurately enough for most engineering purposes except for the Gilbert-Elliott channel and its generalizations.

The first and main result of our own work (first reported in [3]) is a practical algorithm to compute information rates for general *finite-state* source/channel models (to be defined in Section 2). This algorithm was independently discovered also by Sharma and Singh [33], [34] and by Pfister et al. [29]. We will review this algorithm in Section 3. (Sharma and Singh [34] also gave various expressions for the information rate as well as proofs of convergence; expressions for the information rate were also given by Holliday et al. [20].)

We will then extend this method to very general (non-finite state) channels with memory. In Section 5.3 and Appendix C, we demonstrate the use of reduced-state recursions to compute upper and lower bounds on the information rate. In Section 6, we use finite-state approximations of the channel; by simulations of the actual source/channel and computations using the finite-state model, both an upper bound and a lower bound on the information rate of the actual channel are obtained. The bounds will be tight if the finite-state model is a good approximation of the actual channel. The lower bound holds under very weak assumptions; the upper bound requires a lower bound on the conditional entropy rate $h(Y|X)$.

In this paper, we will always assume that the channel input process X is given; in the numerical examples, we will often assume it to be i.u.d. Our parallel work on optimizing

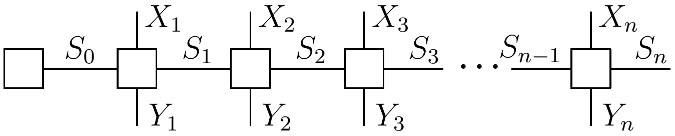

Figure 1: The factor graph of (3).

the process X over finite-state hidden Markov sources (cf. [21]) will be reported in a separate paper [37]. Computational upper bounds on the channel capacity were proposed in [36] and [39].

We will use the notation $x_k^n \triangleq (x_k, x_{k+1}, \ldots, x_n)$ and $x^n \triangleq (x_1, x_2, \ldots, x_n)$.

2 Finite-State Source/Channel Models

In this section, we will assume that the channel input process $X = (X_1, X_2, \ldots)$, the channel output process $Y = (Y_1, Y_2, \ldots)$, and some auxiliary state process $S = (S_0, S_1, S_2, \ldots)$ satisfy

$$p(x^n, y^n, s_0^n) = p(s_0) \prod_{k=1}^{n} p(x_k, y_k, s_k | s_{k-1}) \tag{3}$$

for all $n > 0$ and with $p(x_k, y_k, s_k | s_{k-1})$ not depending on k. We will assume that the state S_k takes values in some *finite* set and we will assume that the process S is ergodic; under the stated conditions, a sufficient condition for ergodicity is $p(s_k|s_0) > 0$ for all s_0, s_k for all sufficiently large k.

For the sake of clarity, we will further assume that the channel input alphabet $\mathcal{X}$ is a finite set and that the channel output Y_k takes values in $\mathbb{R}$; none of these assumptions is essential, however. With these assumptions, the left-hand side of (3) should be understood as a probability mass function in x_k and s_k and as a probability density in y_k. We will also assume that

$$\mathrm{E}\Big[\big|\log p(Y_1|s_0, s_1, x_1)\big|\Big] < \infty \tag{4}$$

for all s_0, s_1, x_1, in order to guarantee the existence of certain limits, cf. [24]. This condition formally excludes a finite channel output alphabet, but all results of this paper are easily reformulated to hold for that case.

The factorization (3) is expressed by the factor graph of Fig. 1. (This graph is a Forney-style factor graph, see [14], [25]; add a circle on each branch to obtain a factor graph as in [23].)

Example 1 (Binary-input FIR filter with AWGN). Let

$$Y_k = \sum_{i=0}^{m} g_i X_{k-i} + Z_k \tag{5}$$

with fixed real coefficients g_i, with X_k taking values in $\{+1, -1\}$, and where $Z = (Z_1, Z_2, \ldots)$ is white Gaussian noise. If X is Markov of order L, i.e.,

$$p(x_k|x^{k-1}) = p(x_k|x_{k-L}^{k-1}), \tag{6}$$

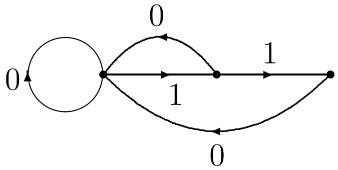

Figure 2: Finite-state machine describing a run-length constraint.

then (3) holds for $S_k \stackrel{\triangle}{=} (X_{k-M+1}, \ldots, X_{k-1}, X_k)$ with $M = \max\{m, L\}$.

As shown in Appendix B, the extension of this example to colored noise can be reduced to the case of white noise.

Example 2 (Channel with freely evolving state). Let $S' = (S'_0, S'_1, \ldots)$ be a first-order Markov process that is independent of X and with S'_k taking values in some finite set. Consider a channel with

$$p(y^n, s'_0, \ldots, s'_n \mid x^n) = p(s'_0) \prod_{k=1}^{n} p(y_k|x_k, s'_{k-1})\, p(s'_k|s'_{k-1}) \tag{7}$$

for all $n > 0$. If X is Markov of order L, then (3) holds for $S_k \stackrel{\triangle}{=} (S'_k, X_{k-L+1}, \ldots, X_{k-1}, X_k)$. This class of channels includes the Gilbert-Elliott channel [26].

Example 3 (Channel with constrained input). Consider a memoryless channel with input alphabet $\{0, 1\}$, and assume that no channel input sequence may contain more than two consecutive ones. Note that the admissible channel input sequences correspond to the walks through the directed graph shown in Fig. 2.

A finite-state process X that complies with these constraints may be obtained by assigning probabilities $p(s_k|s_{k-1})$ to the edges of Fig. 2 such that $\sum_{s_k} p(s_k|s_{k-1}) = 1$. (The problem of finding "good" branching probabilities $p(s_k|s_{k-1})$ is treated in [37]). We then have

$$p(x^n, y^n, s_0^n) = p(s_0) \prod_{k=1}^{n} p(s_k|s_{k-1}) p(x_k|s_k, s_{k-1}) p(y_k|x_k), \tag{8}$$

which is of the form (3).

Under the assumptions stated at the beginning of this section, the limit (1) exists. Moreover, the sequence $-\frac{1}{n} \log p(X^n)$ converges with probability 1 to the entropy rate $H(X)$, the sequence $-\frac{1}{n} \log p(Y^n)$ converges with probability 1 to the differential entropy rate $h(Y)$, and $-\frac{1}{n} \log p(X^n, Y^n)$ converges with probability 1 to $H(X) + h(Y|X)$, cf. [9], [24], and [12, Ch. IV-D]. The corresponding results for the case of a *finite* channel output alphabet are contained already in [28].

3 Computing $I(X;Y)$ for Finite-State Channels

From the above remarks, an obvious algorithm for the numerical computation of $I(X;Y) = h(Y) - h(Y|X)$ is as follows:

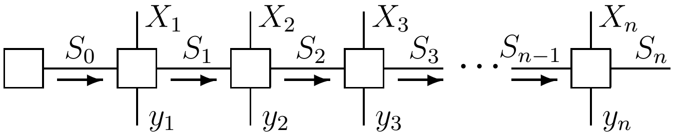

Figure 3: Computation of $p(y^n)$ by message passing through the factor graph of (3).

1. Sample two "very long" sequences x^n and y^n.
2. Compute $\log p(x^n)$, $\log p(y^n)$, and $\log p(x^n, y^n)$. If $h(Y|X)$ is known analytically, then it suffices to compute $\log p(y^n)$.
3. Conclude with the estimate

$$\hat{I}(X;Y) \triangleq -\frac{1}{n}\log p(y^n) - \frac{1}{n}\log p(x^n) + \frac{1}{n}\log p(x^n, y^n) \tag{9}$$

   or, if $h(Y|X)$ is known analytically, $\hat{I}(X;Y) \triangleq -\frac{1}{n}\log p(y^n) - h(Y|X)$.

The computations in Step 2 can be carried out by forward sum-product message passing through the factor graph of (3), as illustrated in Fig. 3. Since the graph represents a trellis, this computation is just the forward sum-product recursion of the BCJR algorithm [8].

Consider, for example, the computation of

$$p(y^n) = \sum_{x^n}\sum_{s_0^n} p(x^n, y^n, s_0^n). \tag{10}$$

Define the state metric $\mu_k(s_k) \triangleq p(s_k, y^k)$. By straightforward application of the sum-product algorithm [23], we recursively compute the messages (state metrics)

$$\begin{aligned} \mu_k(s_k) &= \sum_{x_k}\sum_{s_{k-1}} \mu_{k-1}(s_{k-1})\, p(x_k, y_k, s_k|s_{k-1}) && (11)\\ &= \sum_{x^k}\sum_{s_0^{k-1}} p(x^k, y^k, s^k) && (12)\end{aligned}$$

for $k = 1, 2, 3, \ldots$, as illustrated in Fig. 3. The desired quantity (10) is then obtained as

$$p(y^n) = \sum_{s_n} \mu_n(s_n), \tag{13}$$

the sum of all final state metrics.

For large k, the state metrics $\mu_k(.)$ computed according to (11) quickly tend to zero. In practice, the recursion (11) is therefore changed to

$$\mu_k(s_k) = \lambda_k \sum_{x_k}\sum_{s_{k-1}} \mu_{k-1}(s_{k-1})\, p(x_k, y_k, s_k|s_{k-1}) \tag{14}$$

where λ_1, λ_2, ... are positive scale factors. If these scale factors are chosen such that $\sum_{s_n} \mu_n(s_n) = 1$, then

$$\frac{1}{n} \sum_{k=1}^{n} \log \lambda_k = -\frac{1}{n} \log p(y^n). \tag{15}$$

The quantity $-\frac{1}{n} \log p(y^n)$ thus appears as the average of the logarithms of the scale factors, which converges (almost surely) to $h(Y)$.

If necessary, the quantities $\log p(x^n)$ and $\log p(x^n, y^n)$ can be computed by the same method: for $p(x^n)$, the recursion corresponding to (14) is

$$\mu_k(s_k) = \lambda_k \sum_{s_{k-1}} \mu_{k-1}(s_{k-1})\, p(x_k, s_k | s_{k-1}) \tag{16}$$

and for $p(x^n, y^n)$, the corresponding recursion is

$$\mu_k(s_k) = \lambda_k \sum_{s_{k-1}} \mu_{k-1}(s_{k-1})\, p(x_k, y_k, s_k | s_{k-1}). \tag{17}$$

If there is no feedback from the channel to the source, the computation (16) needs only the source model rather than the joint source/channel model. In this case, if (6) holds, $H(X)$ can be computed in closed form as the entropy of a Markov source [11].

4 Numerical Examples

We will here focus on channels as in Example 1. Further numerical examples (including channels as in Example 3 as well as the nonlinear channel of [2]) are given in [5] and [37].

The filter coefficients $g_0, g_1, \ldots, g_m$ in Example 1 are often compactly represented by the formal sum $G(D) \triangleq \sum_{k=0}^{m} g_k D^k$. The signal-to-noise ratio (SNR) will be defined as

$$\text{SNR} \triangleq \frac{\mathrm{E}[X_k^2]}{\mathrm{E}[Z_k^2] \sum_{k=0}^{m} g_k^2}. \tag{18}$$

(It is clear that this SNR definition is inadequate for some applications, but this qualification seems to apply also to all alternative definitions including that of [38].) For channels as in Example 1, $h(Y|X) = h(Z)$ is known analytically, which means that the algorithm of Section 3 is only needed to compute $h(Y)$.

In all numerical examples reported in this paper, the sequence length $n = 10^6$ proved to be sufficient.

Our first example is a channel as in Example 1 with transfer function $G(D) = 1 - D$. In the magnetic recording literature, this channel is known as the *dicode* channel. Fig. 4 shows the following information rates for this channel:

1. The information rate for i.u.d. input.

2. The maximum information rate for X Markov of order $L = 1$.

3. The maximum information rate for X Markov of order $L = 2$.

The maximization of the information rate over the Markov sources can be done by the methods of [37] or (in this simple example) by brute force. For comparison, Fig. 4 also shows:

4. The capacity of the memoryless AWGN channel.

5. The capacity of the dicode channel for *Gaussian* (rather than binary) input.

The latter is obtained by the well-known waterfilling principle [11]. As the definition (18) allows the channel to provide a power gain for non-white input, the waterfilling capacity exceeds the capacity of the memoryless AWGN channel at low SNR.

The convergence behavior of the algorithm is illustrated by Fig. 5. The i.u.d.-input information rate for the dicode channel at 3.01 dB was computed 10+100+1000 times, each time by a simulation run of 10^6 symbols and with a new random seed. For each blocklength n, Fig. 5 shows the minimum and the maximum computed estimate of the information rate among the first 10, the next 100, and the remaining 1000 simulation runs.

Fig. 6 shows information rates for a channel as in Example 1 with $G(D) = 0.19 + 0.35D + 0.46D^2 + 0.5D^3 + 0.46D^4 + 0.35D^5 + 0.19D^6$. (This particular example was used by Hirt [19].) The following information rates are shown:

1. The information rate for i.u.d. input.

2. The maximum information rate for a Markov source of order $L = 6$.

3. The capacity of the memoryless AWGN channel.

4. The capacity of the channel for Gaussian (rather than binary) input.

Fig. 7 illustrates the performance of Hirt's method [19] as well as a conjectured lower bound on the channel capacity due to Shamai and Laroia [32]. The latter can be computed by evaluating a single one-dimensional integral. Fig. 7 shows several rates for the channel of Fig. 6, each for -5 dB, for 3 dB, and for 8 dB:

1. $I_{\mathrm{HL}}(n)$ (see below) as a function of n.

2. $I_{\mathrm{HU}}(n)$ (see below) as a function of n.

3. The Shamai-Laroia conjectured lower bound (SLLB).

4. The true information rate for i.u.d. input (computed by our algorithm).

As the figure shows, the Shamai-Laroia conjectured lower bound is extremely tight for low SNR.

Hirt defined

$$I_{\mathrm{HL}}(n) \triangleq \frac{1}{n} I(X^n; Y^n | X_0, X_{-1}, \ldots, X_{1-m}) \tag{19}$$

and

$$I_{\mathrm{HU}}(n) \triangleq \frac{1}{n} I(X^n; Y^{n+m} | X_0, \ldots, X_{1-m}, X_{n+1}, \ldots, X_{n+m}), \tag{20}$$

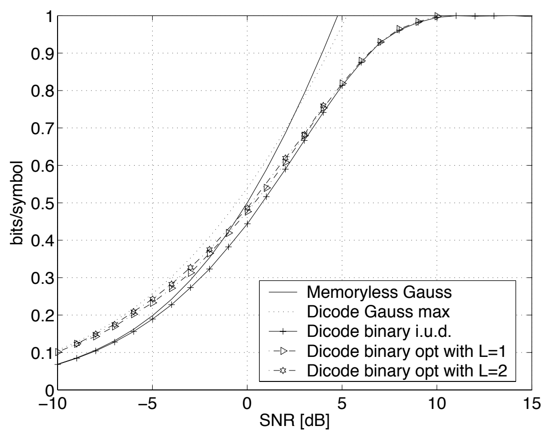

Figure 4: Information rates of the $1-D$ (dicode) channel.

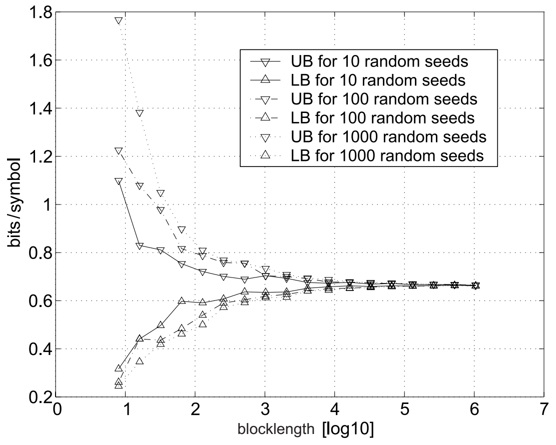

Figure 5: Convergence of the algorithm.

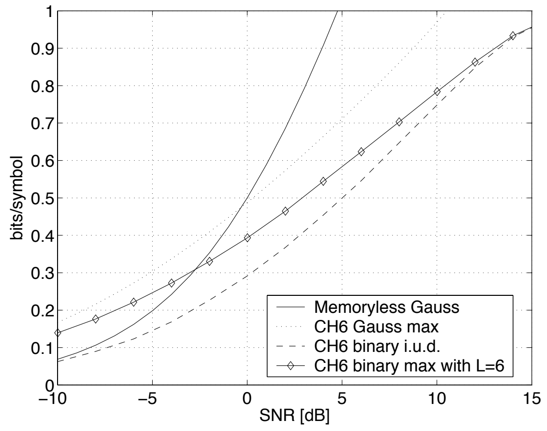

Figure 6: Information rates of an FIR channel with memory 6.

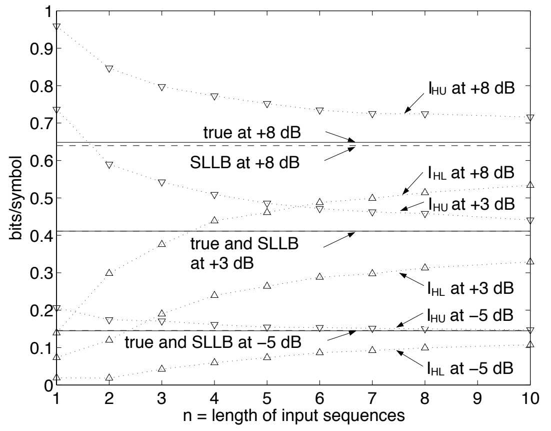

Figure 7: Comparison with Hirt's method and the Shamai-Laroia conjectured lower bound (SLLB) for the channel of Fig. 6.

where the input process X is assumed to be i.u.d. Hirt computed these quantities by numerical integration based on Monte-Carlo simulation. By standard arguments,

$$I_{\mathrm{HL}}(n) \leq I_{\mathrm{HU}}(n) \tag{21}$$

and

$$\lim_{n\to\infty} I_{\mathrm{HL}}(n) = \lim_{n\to\infty} I_{\mathrm{HU}}(n) = I(X;Y). \tag{22}$$

5 Extensions

5.1 Continuous Input Alphabet

As mentioned in Section 2, the assumption that the input alphabet $\mathcal{X}$ is finite is by no means essential. Assume, for example that $\mathcal{X} = \mathbb{R}$ and $p(x^n)$ is a probability density consistent with (3). If $p(x^n)$ is sufficiently nice (which we do not wish to discuss further), then the sequence $-\frac{1}{n}\log p(X^n)$ converges with probability 1 to the differential entropy rate $h(X)$ and the sequence $-\frac{1}{n}\log p(X^n, Y^n)$ converges with probability 1 to $h(X,Y)$. The only modification to the algorithm of Section 3 is that the recursion (14) becomes

$$\mu_k(s_k) = \lambda_k \sum_{s_{k-1}} \mu_{k-1}(s_{k-1}) \int_{-\infty}^{\infty} p(x_k, y_k, s_k|s_{k-1})\, dx_k, \tag{23}$$

which may be evaluated analytically or numerically.

5.2 Time-Varying and/or Non-Ergodic Source/Channel Model

If the factor $p(x_k, y_k, s_k|s_{k-1})$ in (3) depends on k, the quantity $\hat{I}(X;Y)$ defined by (9) may still be computed as described in Section 3, but there is no general guarantee that this estimate converges to $I(X;Y)$.

If the source/channel model is not ergodic, one may sample many sequences x^n and y^n and average over the corresponding estimates (9).

5.3 Bounds on Entropy Rates from Reduced-State Recursions

The basic recursion (11) can be modified to yield upper and lower bounds on $p(y^n)$ and thus on $h(Y)$ (and similarly for $H(X)$ and $h(Y|X)$). The modified recursions can be computed for channels where the number of states is large.

Let $\mathcal{S}'_k$ be a subset of the time-k states. If the sum in the recursion (11) is modified to

$$\mu_k(s_k) = \sum_{x_k} \sum_{s_{k-1}\in\mathcal{S}'_{k-1}} \mu_{k-1}(s_{k-1})\, p(x_k, y_k, s_k|s_{k-1}), \tag{24}$$

the sum of the final state metrics will be a lower bound on $p(y^n)$ and the corresponding estimate of $h(Y)$ will be increased. We thus have the following theorem.

Theorem (Reduced-State Upper Bound). Omitting states from the computation (11) yields an upper bound on $h(Y)$. □

The sets $\mathcal{S}'_k$ may be chosen arbitrarily. An obvious strategy is to keep only a fixed number of states with the largest metrics.

By a similar argument, one may also obtain also lower bounds on $h(Y)$. A particular case is worked out in Appendix C.

The upper bound can also be applied to certain non-finite-state channels as follows. Consider, e.g., the autoregressive channel of Fig. 8 and assume that, at time zero, the channel is in some fixed initial state. At time one, there will be two states; at time two, there will be four states, etc. We track all these states according to (11) until there are too many of them, and then we switch to the reduced-state recursion (24).

Some numerical examples for the upper bound of this section are given in Section 7.

6 Bounds on $I(X;Y)$ Using an Auxiliary Channel

Upper and lower bounds on the information rate of very general (non-finite) state channels can be computed by methods of the following general character.

1. Choose a finite-state (or otherwise tractable) auxiliary channel model that somehow approximates the actual (difficult) channel. (The accuracy of this approximation will affect the tightness, but not the validity of the bounds.)

2. Sample a "very long" channel input sequence and the corresponding channel output sequence of the actual channel.

3. Use these sequences for a computation (in the style of Sections 3–5) using the auxiliary channel model.

We begin by reviewing the underlying analytical bounds, which are well known. For the sake of clarity, we first state these bounds for a discrete memoryless channel. Let X and Y be two discrete random variables with joint probability mass function $p(x,y)$. We will call X the source and $p(y|x)$ the channel law. Let $q(y|x)$ be the law of an arbitrary auxiliary channel with the same input and output alphabets as the original channel. We will imagine that the auxiliary channel is connected to the same source X; its output distribution is then

$$q_p(y) \triangleq \sum_x p(x)\, q(y|x). \tag{25}$$

In the following, we will assume that $q(y|x)$ is chosen such that $q_p(y) > 0$ whenever $p(y) > 0$.

Theorem (Auxiliary-Channel Upper Bound).

$$\begin{aligned} I(X;Y) &\leq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q_p(y)} & (26) \\ &= \mathrm{E}_p\big[\log p(Y|X) - \log q_p(Y)\big]\,, & (27) \end{aligned}$$

where the sum in (26) should be read as running over the support of $p(x,y)$. Equality holds in (26) if and only if $p(y) = q_p(y)$ for all y. □

This bound appears to have been observed first by Topsøe [35]. The proof is straightforward. Let $\overline{I}_q(X;Y)$ be the right-hand side of (26). Then

$$\begin{aligned}
\overline{I}_q(X;Y) - I(X;Y) &= \sum_{x,y} p(x,y) \left[\log \frac{p(y|x)}{q_p(y)} - \log \frac{p(y|x)}{p(y)} \right] && (28) \\
&= \sum_{x,y} p(x,y) \log \frac{p(y)}{q_p(y)} && (29) \\
&= \sum_{y} p(y) \log \frac{p(y)}{q_p(y)} && (30) \\
&= D\big(p(y)||q_p(y)\big) && (31) \\
&\geq 0. && (32)
\end{aligned}$$

Theorem (Auxiliary-Channel Lower Bound).

$$\begin{aligned}
I(X;Y) &\geq \sum_{x,y} p(x,y) \log \frac{q(y|x)}{q_p(y)} && (33) \\
&= \mathrm{E}_p\big[\log q(Y|X) - \log q_p(Y)\big] && (34)
\end{aligned}$$

where the sum in (33) should be read as running over the support of $p(x,y)$. □

This bound is implicit in the classical papers by Blahut [10] and Arimoto [1]. Moreover, it may also be obtained as a special case of a bound due to Fischer [13] on mismatched decoding, which in turn is a special case of a general result by Ganti et al. [15, Equation (12) for $s = 1$]. It then follows from the results in [13] and [15] that the lower bound is achievable by a maximum-likelihood decoder for the auxiliary channel.

A simple proof of (33) goes as follows. Let $\underline{I}_q(X;Y)$ be the right-hand side of (33) and for y satisfying $p(y) > 0$ (which by the assumption after equation (25) implies $q_p(y) > 0$) let

$$r_p(x|y) \triangleq \frac{p(x)q(y|x)}{q_p(y)} \qquad (35)$$

be the "reverse channel" of the auxiliary channel. Then

$$\begin{aligned}
I(X;Y) - \underline{I}_q(X;Y) &= \sum_{x,y} p(x,y) \left[\log \frac{p(x,y)}{p(x)p(y)} - \log \frac{q(y|x)}{q_p(y)} \right] && (36) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)p(x)q(y|x)/q_p(y)} && (37) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)r_p(x|y)} && (38) \\
&= D\big(p(x,y)||p(y)r_p(x|y)\big) && (39) \\
&\geq 0. && (40)
\end{aligned}$$

As is easily verified, the difference between the above two bounds can be written as

$$\overline{I}_q(X;Y) - \underline{I}_q(X;Y) = D\big(p(x)p(y|x)||p(x)q(y|x)\big). \qquad (41)$$

The generalization of these bounds to the information rate of channels with memory is straightforward. For any finite $n > 0$, the bounds clearly apply to I_n as in (2). If the required limits for $n \to \infty$ exist, the upper bound becomes

$$\overline{I}_q(X;Y) \triangleq \lim_{n\to\infty} \mathrm{E}_p\Big[-\frac{1}{n}\log q_p(Y^n) + \frac{1}{n}\log p(Y^n|X^n)\Big] \tag{42}$$

and the lower bound becomes

$$\underline{I}_q(X;Y) \triangleq \lim_{n\to\infty} \mathrm{E}_p\Big[-\frac{1}{n}\log q_p(Y^n) + \frac{1}{n}\log q(Y^n|X^n)\Big]. \tag{43}$$

Now assume that $p(\cdot|\cdot)$ is some "difficult" (non-finite-state) ergodic channel. We can compute bounds on its information rate by the following algorithm:

1. Choose a finite-state source $p(\cdot)$ and an auxiliary finite-state channel $q(\cdot|\cdot)$ so that their concatenation is a finite-state source/channel model as defined in Section 3.

2. Concatenate the source to the *original* channel $p(\cdot|\cdot)$ and sample two "very long" sequences x^n and y^n.

3. Compute $\log q_p(y^n)$ and, if necessary, $\log p(x^n)$ and $\log q(y^n|x^n)p(x^n)$ by the method described in Section 3.

4. Conclude with the estimates

$$\hat{\overline{I}}_q(X;Y) \triangleq -\frac{1}{n}\log q_p(y^n) - h(Y|X) \tag{44}$$

and

$$\hat{\underline{I}}_q(X;Y) \triangleq -\frac{1}{n}\log q_p(y^n) - \frac{1}{n}\log p(x^n) + \frac{1}{n}\log q(y^n|x^n)p(x^n). \tag{45}$$

Note that the term $h(Y|X)$ in the upper bound (44) refers to the original channel and cannot be computed by means of the auxiliary channel. However, this term can often be determined analytically.

For this algorithm to work, (44) and (45) should converge with probability one to (42) and (43), respectively. Sufficient conditions for the existence of such limits are discussed in [28], [9], [24], [12, Ch. IV-D]. In particular, the following conditions are sufficient:

1. The original source/channel model $p(x, y)$ is of the form (3) with finite state space, with $p(x_k, y_k, s_k|s_{k-1})$ not depending on k, and with $p(s_k|s_0) > 0$ for all sufficiently large k.

2. The auxiliary channel model $q(y|x)$ (together with the original source $p(x)$) is of the same form.

3. In addition to (4), we also have $\mathrm{E}_p\big[\big|\log q_p(Y_k|s_{k-1}, s_k, x_k)\big|\big] < \infty$ for all s_{k-1}, s_k, x_k.

Quantities very similar to (42) and (43) seem to have been computed by essentially the same algorithm as far back as 1985, cf. [22].

7 Numerical Examples for the Bounds

We illustrate the methods of Sections 5.3 and 6 by some numerical examples. As in Section 4, we focus on channels as in Example 1 (and we will use the same definition of the SNR). The input process X will always be assumed to be i.u.d.

Our first example is a memory-10 FIR filter with $G(D) = \sum_{i=0}^{10} \frac{1}{1+(i-5)^2} D^i$. Fig. 10 shows the following curves.

1. Bottom: the exact information rate computed as described in Section 3.
2. Top: the reduced-state upper bound (RSUB) of Section 5.3, using the 100 (out of 1024) states with the largest state metric.
3. Middle: the reduced-state upper bound (still with 100 states) applied to an equivalent channel which is obtained by replacing $G(D)$ by the corresponding minimum-phase polynomial.

The notion of a minimum-phase filter is reviewed in Appendix A, and the justification for replacing $G(D)$ by the corresponding minimum-phase polynomial (i.e., the minimum-phase filter with the same amplitude response) is given in Appendix B. The motivation for this replacement is that minimum-phase filters concentrate the signal energy into the leading tap weights [27], which makes the reduced-state bound tighter.

It is obvious from Fig. 10 that the reduced-state upper bound works fine for high SNR and becomes useless for low SNR.

Our next example is the channel of Fig. 8 with an autoregressive filter

$$G(D) = 1/(1 - \alpha D) = (1 + \alpha D + \alpha^2 D^2 + \ldots)$$

for $\alpha = 0.8$. Both the reduced-state bound of Section 5.3 and the auxiliary-channel bound of Section 6 were applied. The auxiliary channel was obtained from the original channel by inserting a uniform quantizer in the feedback loop, which results in the finite-state channel of Fig. 9. Both the range of the quantizer and the noise variance of the auxiliary channel were numerically optimized to give as good bounds as possible. Fig. 11 shows the following curves.

1. Rightmost: the (indistinguishable) upper and lower bounds (AUB and ALB) using the auxiliary channel of Fig. 9 with 512 states.
2. Very slightly to the left: the reduced-state upper bound (RSUB) using only 4 states.
3. Leftmost: the memoryless binary-input (BPSK) channel.

In this example, the auxiliary-channel bounds yield the true information rate up to the accuracy of the plot. The reduced-state upper bound is extremely tight over the whole SNR range even for very few states.

For this same setup, Fig. 12 shows these three bounds as a function of the number of states (for SNR=7.45 dB). The superiority of the reduced-state bound is striking.

Our last example is an autoregressive filter with

$$G(D) = 1/(1.0000+0.3642{\cdot}D+0.0842{\cdot}D^2+0.2316{\cdot}D^3-0.2842{\cdot}D^4+0.2084{\cdot}D^5+0.2000{\cdot}D^6).$$

Fig. 13 shows (from left to right):

1. The capacity of the BPSK channel.
2. The reduced-state upper bound with only 2 states.
3. The reduced-state upper bound with 128 states.

8 Conclusions

We have presented a general method for the numerical computation of information rates of finite-state source/channel models. By extensions of this method, upper and lower bounds on the information rate can be computed for very general (non-finite state) channels. A lower bound can be computed from simulated (or measured) channel input/output data alone; for the corresponding upper bound, an additional assumption (such as a lower bound on $h(Y|X)$) is needed. Bounds from channel approximations and bounds from reduced-state trellis computations can be combined in several ways.

Acknowledgement

The authors are grateful to Amos Lapidoth from ETH Zürich for pointing out to us Topsøe's bound (26) as well as the references [13] and [15].

Appendix

A On Minimum-Phase Filters

This section summarizes some basic and well-known facts on discrete-time linear time-invariant (LTI) systems, cf. [27].

For a discrete-time signal $f : \mathbb{Z} \to \mathbb{R}$, we write $f_k \triangleq f(k)$. Such a signal is *left sided* if, for some $m \in \mathbb{Z}$, $f_k = 0$ for $k > m$; it is *right-sided* if, for some $m \in \mathbb{Z}$, $f_k = 0$ for $k < m$; and it is *causal* if $f_k = 0$ for $k < 0$.

An LTI system, or "filter", is specified by its impulse response g; the output signal y resulting from an arbitrary input signal x is given by $y_n = \sum_{k\in\mathbb{Z}} x_{n-k} g_k$. The filter is *stable* (bounded-input bounded-output) if and only if $\sum_{k\in\mathbb{Z}} |g_k| < \infty$. The filter is *causal* if and only if g is a causal signal.

The *transfer function* of such a filter is

$$G(z) \triangleq \sum_{k\in\mathbb{Z}} g_k z^{-k}, \tag{46}$$

which may be interpreted either as a formal series in the indeterminate z (i.e., $G(z) = G(D)$ for $D = z^{-1}$) or as a function $S_g \to \mathbb{C}$ with domain $S_g \subset \mathbb{C}$ (essentially the region of convergence of (46)) of the form $S_g = \{z \in \mathbb{C} : r_1 < |z| < r_2\}$, where r_1 is a nonnegative real number and $r_2 \in \mathbb{R} \cup \{\infty\}$. If the filter is right-sided, then $r_1 = 0$. If S_g contains the

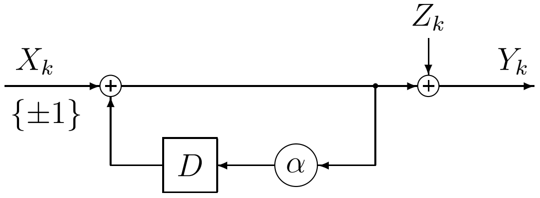

Figure 8: Autoregressive-filter channel.

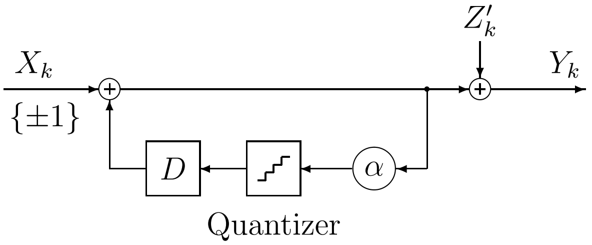

Figure 9: A quantized version of Fig. 8.

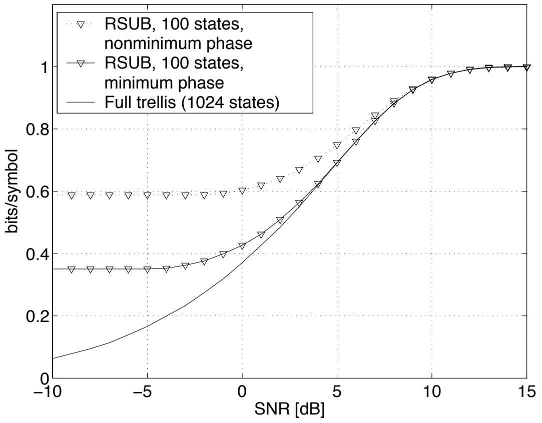

Figure 10: Memory-10 FIR filter.

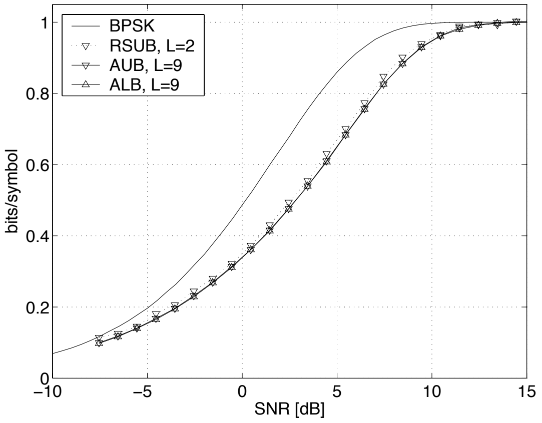

Figure 11: Bounds for Fig. 8 vs. SNR. $L \triangleq \log_2$(number of states).

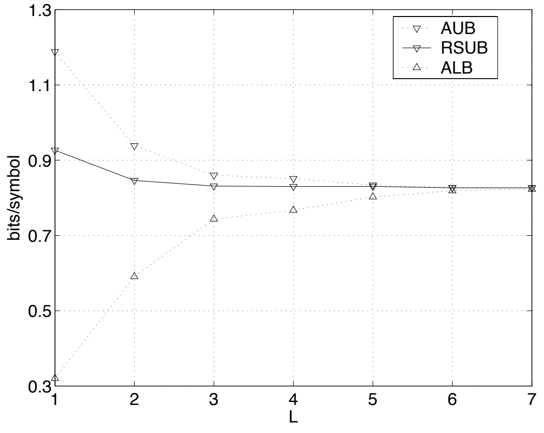

Figure 12: Bounds for Fig. 8 vs. $L \triangleq \log_2$(number of states).

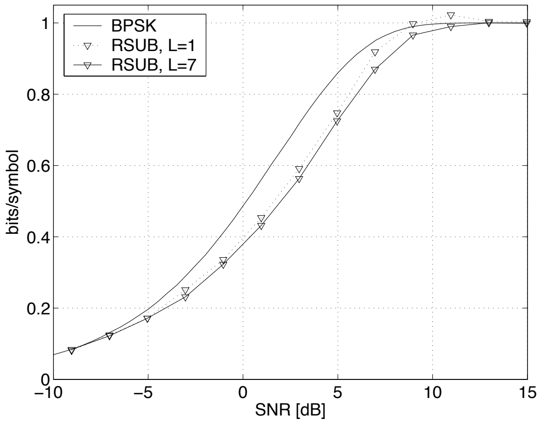

Figure 13: Order 6 autoregressive filter: upper bounds.

unit circle, then the filter is stable. An *inverse* to an LTI filter with transfer function G is an LTI filter with transfer function H such that $G(z)H(z) = 1$.

Now assume that $G(z)$ is a rational function. Then the following conditions are equivalent:

1. $G(z)$ is called *minimum-phase*.

2. All zeros and all poles of $G(z)$ are inside the unit circle, and the degree (in z) of the numerator equals the degree of the denominator.

3. The filter is causal and stable and has an inverse that is also causal and stable.

A filter $H(z)$ is an *all-pass* filter if $|H(e^{i\Omega})| = 1$ for all $\Omega \in \mathbb{R}$.

Theorem (Minimum-Phase/All-pass Decomposition). Let $F(z)$ be a rational function without zeros or poles on the unit circle. Then $F(z)$ can be written as

$$F(z) = G(z)H(z), \tag{47}$$

where $G(z)$ is minimum-phase and $H(z)$ is an all-pass. Moreover, $H(z)$ can be realized as a stable filter with a right-sided impulse response and $1/H(z)$ can be realized as a stable filter with a left-sided impulse response. □

Clearly, $|G(e^{i\Omega})| = |F(e^{i\Omega})|$ for all $\Omega \in \mathbb{R}$. For

$$F(z) = \frac{\prod_{k=1}^{m}(z - z_k)}{\prod_{\ell=1}^{n}(z - p_\ell)} \tag{48}$$

the corresponding minimum-phase filter $G(z)$ is

$$G(z) = \frac{z^{n-m}\prod_{k:|z_k|<1}(z - z_k)\prod_{k:|z_k|>1}(1 - z\overline{z_k})}{\prod_{\ell=1}^{n}(z - p_\ell)} \tag{49}$$

where $\overline{z_k}$ denotes the complex conjugate of z_k.

B On Linear Channels with Additive Noise

Consider the channel of Fig. 14.A: the input process X, which is assumed to be stationary, is filtered by a linear filter $F(z)$ and then the noise process W is added. The function $F(z)$ is assumed to be rational without poles or zeros on the unit circle. We will review the following facts.

1. If the noise is white Gaussian, replacing $F(z)$ by the corresponding minimum-phase filter $G(z)$ (as in (47) and (49)) does not change the information rate $I(X;Y)$.

2. The case of colored Gaussian noise (as defined below) can be converted into the case of white Gaussian noise.

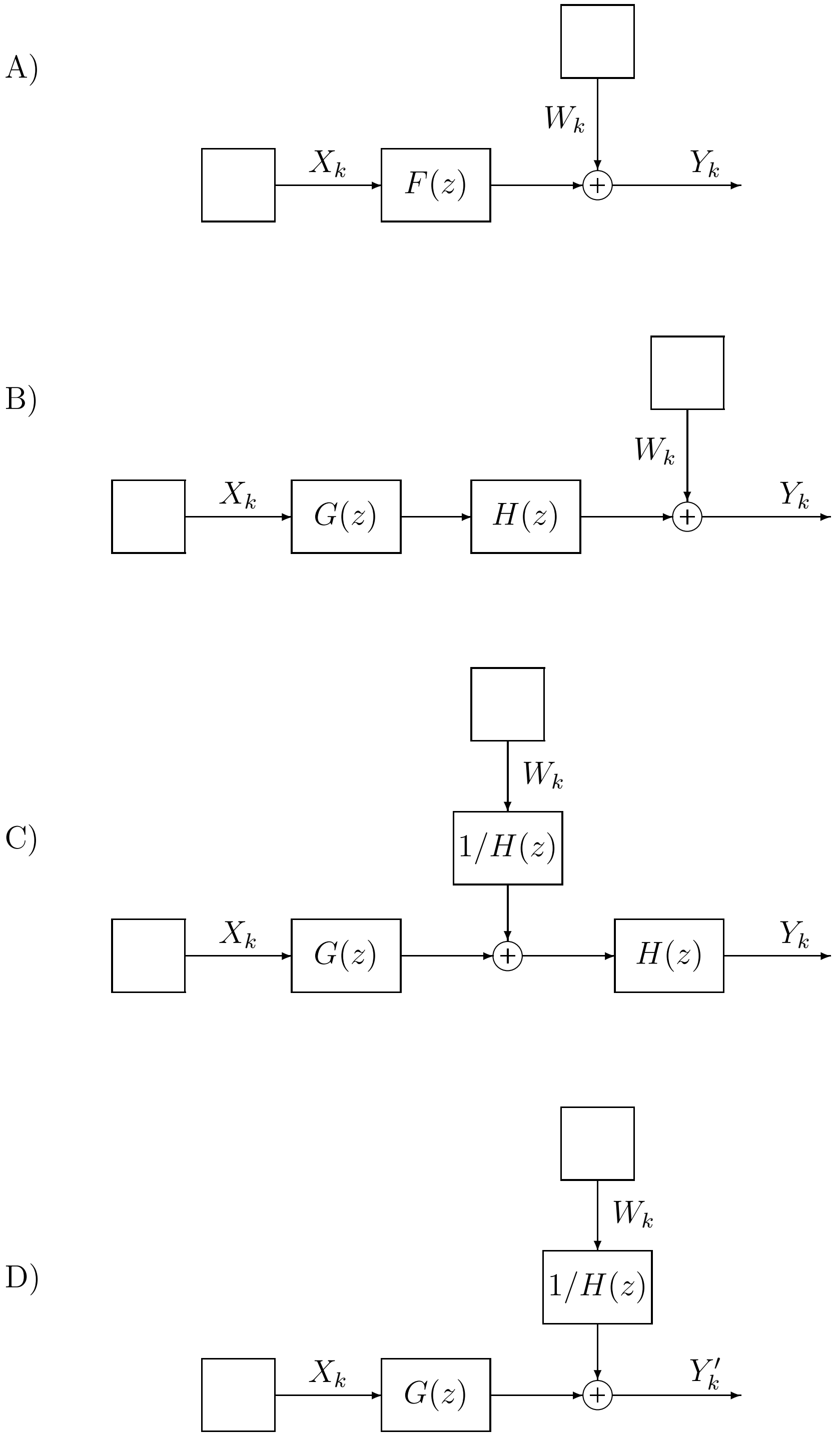

Figure 14: On linear channels with additive noise.

We begin with the first case. Clearly, when $F(z)$ is decomposed according to (47), the information rate $I(X;Y)$ remains unchanged (Fig. 14.B). It is then obvious that the channel of Fig. 14.C also has the same information rate $I(X;Y)$. Omitting the stable all-pass $H(z)$ at the output does not increase the information rate, and thus the information rate $I(X;Y')$ of the channel in Fig. 14.D equals $I(X;Y)$ of the original channel of Fig. 14.A. Finally, the (noncausal stable) all-pass filter $1/H(z)$ in Fig. 14.D transforms white Gaussian noise into white Gaussian noise and can be omitted without changing the information rate.

Now to the second case. Recall that colored Gaussian noise is filtered white Gaussian noise. This case may thus be represented by Fig. 14.D, where W is white Gaussian noise and where we now assume (without loss of generality) that the filter $H(z)$ is minimum-phase. The filter $G(z)$ is arbitrary; in particular, we could have $G(z) = 1$. Appending the minimum-phase filter $H(z)$ at the output (which results in Fig. 14.C) does not change the information rate. As before, Figures 14.C and 14.B are equivalent, and defining $F(z) = G(z)H(z)$, all channels in Fig. 14 have again the same information rate. If the noise-coloring filter $1/H(z)$ is autoregressive, $H(z)$ is an FIR filter.

C A Reduced-State Lower Bound on $I(X;Y)$

In Section 5.3, it was pointed out that omitting states in the basic recursion (11) yields an upper bound on the entropy rate $h(Y)$. Lower bounds on $h(Y)$ (and thus on $I(X;Y)$) may be obtained by merging states. In this section, we give a particular example of this type.

We consider a binary-input linear channel with

$$Y_k = \sum_{\ell=0}^{m} g_\ell X_{k-\ell} + Z_k, \tag{50}$$

with channel memory $m \in \mathbb{Z} \cup \{\infty\}$, with fixed known channel coefficients $g_0, g_1, \ldots, g_m \in \mathbb{R}$, and where $Z = (Z_1, Z_2, \ldots)$ is white Gaussian noise with variance σ^2. For the sake of clarity, the channel input process $X = (X_1, X_2, \ldots)$ is assumed to be a sequence of i.u.d. random variables taking values in $\{+1, -1\}$.

The channel state at time k is the m-tuple $(x_{k-1}, x_{k-2}, \ldots, x_{k-m})$ of the m past channel inputs. We will consider *merged states* of the form

$$(x_{k-1}, x_{k-2}, \ldots, x_{k-M}) \triangleq \bigcup_{x_{k-M-1},\ldots,x_{k-m}} \{(x_{k-1}, \ldots, x_{k-M}, x_{k-M-1}, \ldots, x_{k-m})\} \tag{51}$$

for some positive integer $M < m$ (which need not be the same for all merged states).

As in Section 5.3, we begin by assuming that the channel is in some known state at time zero. At time one, there will be two states; at time two, there will be four states, etc. We first compute the recursion (11) with all these states until there are too many of them. From that moment on, we merge states into the form (51), and we keep expanding and merging (merged) states according to some strategy that will not be detailed here. (One such strategy is described in [5].)

The crucial quantity in this computation is

$$p(y_k|x_k, x_{k-1}, \ldots, x_{k-m}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_k - w)^2/(2\sigma^2)} \tag{52}$$

with

$$w \triangleq w(x_k, \ldots, x_{k-m}) \triangleq g_0 x_k + \sum_{\ell=1}^{m} g_\ell x_{k-\ell}. \tag{53}$$

For each state $(x_{k-1}, \ldots, x_{k-m})$ in some merged state $(x_{k-1}, \ldots, x_{k-M})$, w lies in the interval $[w_\mathrm{L}, w_\mathrm{U}]$ with

$$\begin{aligned} w_\mathrm{U} &\triangleq w_\mathrm{U}(x_k, \ldots, x_{k-M}) && (54)\\ &\triangleq g_0 x_k + \sum_{\ell=1}^{M} g_\ell x_{k-\ell} + \sum_{\ell=M+1}^{m} |g_\ell| && (55) \end{aligned}$$

and

$$\begin{aligned} w_\mathrm{L} &\triangleq w_\mathrm{L}(x_k, \ldots, x_{k-M}) && (56)\\ &\triangleq g_0 x_k + \sum_{\ell=1}^{M} g_\ell x_{k-\ell} - \sum_{\ell=M+1}^{m} |g_\ell|. && (57) \end{aligned}$$

For each state $(x_{k-1}, \ldots, x_{k-m})$ in the merged state $(x_{k-1}, \ldots, x_{k-M})$, we thus have

$$p(y_k|x_k, x_{k-1}, \ldots, x_{k-m}) \leq \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_k - \hat{w})^2/(2\sigma^2)}, \tag{58}$$

where

$$\hat{w} \triangleq \hat{w}(x_k, \ldots, x_{k-M}, y_k) \triangleq \begin{cases} w_\mathrm{L} & \text{if } y_k < w_\mathrm{L} \\ w_\mathrm{U} & \text{if } y_k > w_\mathrm{U} \\ y_k & \text{else} \end{cases} \tag{59}$$

depends only on the merged state. Using the right-side of (58) in the recursion (11) yields a lower bound on $h(Y)$.

In our numerical experiments so far, the lower bound of this section turned out to be consistently weaker than (a comparable version of) the lower bound of Section 6.

References

[1] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels", *IEEE Trans. Information Theory,* vol. 18, pp. 14–20, Jan. 1972.

[2] D. Arnold, A. Kavčić, R. Kötter, H.-A. Loeliger, and P. O. Vontobel, "The binary jitter channel: a new model for magnetic recording," *Proc. 2000 IEEE Int. Symp. Information Theory,* Sorrento, Italy, June 25–30, 2000, p. 433.

[3] D. Arnold and H.-A. Loeliger, “On the information rate of binary-input channels with memory,” *Proc. 2001 IEEE Int. Conf. on Communications,* Helsinki, Finland, June 11–14, 2001, pp. 2692–2695.

[4] D. Arnold and H.-A. Loeliger, “On finite-state information rates from channel simulations,” *Proc. 2002 IEEE Int. Symp. Information Theory,* Lausanne, Switzerland, June 30 – July 5, 2002, p. 164.

[5] D. Arnold, *Computing Information Rates of Finite-State Models with Application to Magnetic Recording.* ETH-Diss no. 14760. Hartung-Gorre Verlag, Konstanz, Germany, 2002.

[6] D. Arnold, H.-A. Loeliger, and P. O. Vontobel, “Computation of information rates from finite-state source/channel models,” *Proc. 40th Annual Allerton Conference on Communication, Control, and Computing,* (Allerton House, Monticello, Illinois), October 2 – October 4, 2002, pp. 457–466.

[7] D. Arnold, A. Kavčić, H.-A. Loeliger, P. O. Vontobel, and W. Zeng, “Simulation-based computation of information rates: upper and lower bounds,” *Proc. 2003 IEEE Int. Symp. Information Theory,* Yokohama, Japan, June 29 – July 4, 2003, p. 231.

[8] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, “Optimal decoding of linear codes for minimizing symbol error rate,” *IEEE Trans. Information Theory,* vol. 20, pp. 284–287, March 1974.

[9] A. Barron, “The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem,” *Annals of Prob.,* vol. 13, no. 4, pp. 1292–1303, 1985.

[10] R. E. Blahut, “Computation of channel capacity and rate-distortion functions,” *IEEE Trans. Information Theory,* vol. 18, pp. 460–473, July 1972.

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

[12] Y. Ephraim and N. Merhav, “Hidden Markov processes,” *IEEE Trans. Information Theory,* vol. 48, pp. 1518–1569, June 2002.

[13] T. R. M. Fischer, “Some remarks on the role of inaccuracy in Shannon’s theory of information transmission,” *Proc. 8th Prague Conf. Information Theory,* 1978, pp. 221–226.

[14] G. D. Forney, Jr., “Codes on graphs: normal realizations,” *IEEE Trans. Information Theory,* vol. 47, no. 2, pp. 520–548, 2001.

[15] A. Ganti, A. Lapidoth, and İ. E. Telatar, “Mismatched decoding revisited: general alphabets, channels with memory, and the wide-band limit,” *IEEE Trans. Information Theory,* vol. 46, pp. 2315–2328, Nov. 2000.

[16] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Information Theory,* vol. 42, pp. 868–886, May 1996.

[17] C. Heegard and A. Duel-Hallen, "On the capacity of the noisy run-length channel," *Proc. 1998 IEEE Information Theory Workshop,* Beijing, China, July 1988.

[18] C. Heegard, A. Duel-Hallen, and R. Krishnamoorthy, "On the capacity of the noisy runlength channel," *IEEE Trans. Information Theory,* vol. 37, pp. 712–720, May 1991.

[19] W. Hirt, *Capacity and Information Rates of Discrete-Time Channels with Memory.* ETH-Diss no. 8671, ETH Zurich, 1988.

[20] T. Holliday, P. Glynn, and A. Goldsmith, "On entropy and Lyapunov exponents for finite-state channels," submitted to *IEEE Trans. Information Theory.*

[21] A. Kavčić, "On the capacity of Markov sources over noisy channels," *Proc. 2001 IEEE Globecom,* San Antonio, TX, pp. 2997–3001, Nov. 25–29, 2001.

[22] B.-H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Technical J.,* vol. 64, pp. 391–408, Feb. 1985.

[23] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Information Theory,* vol. 47, pp. 498–519, Feb. 2001.

[24] B. G. Leroux, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Processes and their Applications,* vol. 40, pp. 127–143, 1992.

[25] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Proc. Mag.,* to appear in Jan. 2004.

[26] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliott channel," *IEEE Trans. Information Theory,* vol. 35, pp. 1277–1290, Nov. 1989.

[27] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing,* 2nd ed. Prentice Hall, 1999.

[28] T. Petrie, "Probabilistic functions of finite state Markov chains," *Ann. Math. Statist,* vol. 40, pp. 97–115, 1969.

[29] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rates of finite-state ISI channels," *Proc. 2001 IEEE Globecom,* San Antonio, TX, pp. 2992–2996, Nov. 25–29, 2001.

[30] Sh. Shamai and Y. Kofman, "On the capacity of binary and Gaussian channels with run-length-limited inputs," *IEEE Trans. Communications,* vol. 5, pp. 584–594, May 1990.

[31] Sh. Shamai, L. H. Ozarow, and A. D. Wyner, "Information rates for a discrete-time Gaussian channel with intersymbol interference and stationary inputs," *IEEE Trans. Information Theory,* vol. 37, pp. 1527–1539, Nov. 1991.

[32] Sh. Shamai and R. Laroia, "The intersymbol interference channel: lower bounds on capacity and channel precoding loss," *IEEE Trans. Information Theory,* vol. 42, pp. 1388–1404, Sept. 1996.

[33] V. Sharma and S. K. Singh, "Entropy and channel capacity in the regenerative setup with applications to Markov channels," *Proc. 2001 IEEE Int. Symp. Information Theory,* Washington, DC, USA, June 24–29, 2001, p. 283.

[34] V. Sharma and S. K. Singh, "Entropy and channel capacity in the regenerative setup with applications to Markov channels," *IEEE Trans. Information Theory,* to appear.

[35] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Studia Scientiarum Math. Hungarica,* vol. 2, pp. 291–292, 1967.

[36] P. O. Vontobel and D. Arnold, "An upper bound on the capacity of channels with memory and constraint input," *Proc. 2001 IEEE Information Theory Workshop,* Cairns, Australia, Sept. 2–7, 2001, pp. 147–149.

[37] P. O. Vontobel, A. Kavčić, D. Arnold, and H.-A. Loeliger, "Capacity of finite-state machine channels," to be submitted to *IEEE Trans. Information Theory.*

[38] W. Xiang and S. S. Pietrobon, "On the capacity and normalization of ISI channels," *IEEE Trans. Information Theory,* vol. 49, pp. 2263–2268, Sept. 2003.

[39] S. Yang, A. Kavčić, and S. Tatikonda, "Delayed feedback capacity of finite-state machine channels: upper bounds on the feedforward capacity," *Proc. 2003 IEEE Int. Symp. Information Theory,* Yokohama, Japan, June 29 – July 4, 2003, p. 290.

[40] E. Zehavi and J. K. Wolf, "On runlength codes," *IEEE Trans. Information Theory,* vol. 34, pp. 45–54, Jan. 1988.