Copyright Notice

Simulation-Based Computation of Information Rates for Channels With Memory

Dieter M. Arnold, *Member, IEEE*, Hans-Andrea Loeliger, *Fellow, IEEE*, Pascal O. Vontobel, *Member, IEEE*, Aleksandar Kavčić, *Senior Member, IEEE*, and Wei Zeng, *Student Member, IEEE*

***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.**

***Index Terms*—Bounds, channel capacity, finite-state models, hidden-Markov models, information rate, sum–product algorithm, trellises.**

I. Introduction

We consider the problem of computing the information rate

$$I(X;Y) \triangleq \lim_{n\to\infty} \frac{1}{n} I(X_1,\ldots,X_n;Y_1,\ldots,Y_n) \qquad (1)$$

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.

Manuscript received January 12, 2004; revised April 29, 2006. The work of P. O. Vontobel was supported by the National Science Foundation under Grants CCR 99-84515 and CCR 01-05719. The material in this paper was presented in part at IEEE International Conference on Communications, Helsinki, Finland, June 2001; IEEE International Symposium on Information Theory, Lausanne, Switzerland, June/July 2002; the 40th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, October 2002; and IEEE International Symposium on Information Theory, Yokohama, Japan, June/July 2003.

D. M. Arnold was with the Department of Information Technology and Electrical Engineering, ETH Zurich, CH-8092 Zurich, Switzerland, and with IBM Research Laboratories, Rüschlikon, Zürich, Switzerland. He is now with Siemens Switzerland AG, 8047 Zürich, Switzerland (e-mail: Dieter.M.Arnold@siemens.com).

H.-A. Loeliger is with the Department of Information Technology and Electrical Engineering, ETH Zurich, CH-8092 Zürich, Switzerland (e-mail: loeliger@isi.ee.ethz.ch).

P. O. Vontobel was with the Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland, and later with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. He is now with Hewlett-Packard Laboratories, Palo Alto, CA 94304 (e-mail: pascal.vontobel@ ieee.org).

A. Kavčić and W. Zeng are with the Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: kavcic@deas.harvard.edu; wzeng@deas.harvard.edu).

Communicated by G. Battail, Associate Editor At Large.

Digital Object Identifier 10.1109/TIT.2006.878110

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

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

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 (ISI) channels, ii) generalizations of the Gilbert–Elliott channel, and iii) channels with constrained input (cf. the examples in Section II). The binary-input linear ISI channel was investigated by Hirt [21], who proposed a Monte Carlo method to evaluate certain quantities closely related to the i.u.d. information rate (cf. Section IV). Shamai *et al.* [36], [37] also investigated the ISI 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 [29]. Goldsmith and Varaiya extended that work to general channels with a freely evolving state [18] (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 [46]; they derived a set of lower bounds for Markovian input and demonstrated some numerical results. 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 [35]. A related topic is the continuous-time additive white Gaussian noise (AWGN) channel with peak-amplitude-constrained input, which was addressed by Heegard *et al.* [19], [20].

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 II). This algorithm was independently discovered also by Sharma and Singh [38] and by Pfister *et al.* [32]. We will review this algorithm in Section III.

Since the original submission of this paper, this algorithm has been used and extended in various ways. For example, Zhang *et al.* investigate information rates both of magnetic recording

channels [48] and of fading multiple-input multiple-output (MIMO) channels with ISI [49]; magnetic recording is also considered by Ryan *et al.* [34] as well as by Pighi *et al.* [33]. Two-dimensional ISI channels are considered by Siegel *et al.* [11], [40] and by Shental *et al.* [39]. Related analytical results were presented by Sharma and Singh [38] as well as by Holliday *et al.* [22], [23]; the latter explore, in particular, the relation to Lyapunov exponents of the product of random matrices. Further related work by the authors of the present paper (not covered here) includes [42], [13], [47]; see also [43] and [5].

In this paper, after describing the basic algorithm, we extend the method to very general (non-finite-state) channels with memory. In Section V-C and Appendix III, we demonstrate the use of reduced-state recursions to compute upper and lower bounds on the information rate. In Section VI, 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 the process X over finite-state hidden-Markov sources (cf. [24]) will be reported in a separate paper [43]. Computational upper bounds on the channel capacity were proposed in [42] and [45].

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

II. 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

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

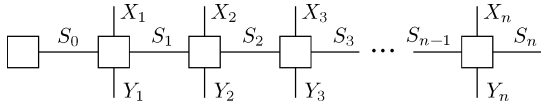

Fig. 1. The factor graph of (3).

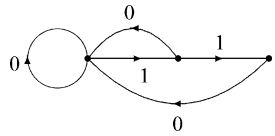

Fig. 2. Finite-state machine describing a run-length constraint.

for all s_0, s_1, x_1, in order to guarantee the existence of certain limits, cf. [27]. 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 [16], [28]; add a circle on each branch to obtain a factor graph as in [26].)

Example 1 (Channel With Binary-Input Finite Impulse Response (FIR) Filter and 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.,

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

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

As shown in Appendix II, 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 $S_k \triangleq (S'_k, X_{k-L+1}, \ldots, X_{k-1}, X_k)$. This class of channels includes the Gilbert–Elliott channel [29].

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 [43].) 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], [27], and [14, Sec. IV-D]. The corresponding results for the case of a *finite* channel output alphabet are contained already in [31].

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

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

1) Sample two "very long" sequences x^n and y^n. (The meaning of "very long" is discussed in Section IV.)
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

$$\hat{I}(X;Y) \triangleq -\frac{1}{n}\log p(y^n) - h(Y|X). \tag{10}$$

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 Bahl–Cocke–Jelinek–Raviv (BCJR) algorithm [8].

Consider, for example, the computation of

$$p(y^n) = \sum_{x^n} \sum_{s_0^n} p(x^n, y^n, s_0^n). \tag{11}$$

Define the state metric $\mu_k(s_k) \triangleq p(s_k, y^k)$. By straightforward application of the sum–product algorithm [26], we recursively compute the messages (state metrics)

$$\mu_k(s_k) = \sum_{x_k} \sum_{s_{k-1}} \mu_{k-1}(s_{k-1})\, p(x_k, y_k, s_k|s_{k-1}) \tag{12}$$

$$= \sum_{x^k} \sum_{s_0^{k-1}} p(x^k, y^k, s_0^k) \tag{13}$$

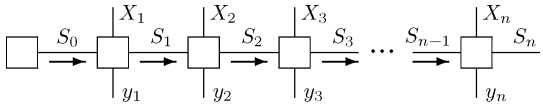

Fig. 3. Computation of $p(y^n)$ by message passing through the factor graph of (3).

for $k = 1, 2, 3, \ldots$, as illustrated in Fig. 3. The desired quantity (11) is then obtained as

$$p(y^n) = \sum_{s_n} \mu_n(s_n), \tag{14}$$

the sum of all final state metrics.

For large k, the state metrics $\mu_k(s_k)$ computed according to (12) quickly tend to zero. In practice, the recursion (12) is therefore changed to

$$\mu_k(s_k) = \lambda_k \sum_{x_k} \sum_{s_{k-1}} \mu_{k-1}(s_{k-1})\, p(x_k, y_k, s_k|s_{k-1}) \tag{15}$$

where $\lambda_1, \lambda_2, \ldots$ are positive scale factors. If these scale factors are chosen such that $\sum_{s_n} \mu_n(s_n) = 1$, then

$$\frac{1}{n}\sum_{k=1}^{n} \log \lambda_k = -\frac{1}{n}\log p(y^n). \tag{16}$$

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 (15) is

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

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

$$\mu_k(s_k) = \lambda_k \sum_{s_{k-1}} \mu_{k-1}(s_{k-1})\, p(x_k, y_k, s_k|s_{k-1}). \tag{18}$$

If there is no feedback from the channel to the source, the computation (17) 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 [12].

IV. Numerical Examples

We will focus here 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 [43].

The filter coefficients $g_0, g_1, \ldots, g_m$ in Example 1 are often compactly represented by the formal sum

$$G(D) \triangleq \sum_{k=0}^{m} g_k D^k.$$

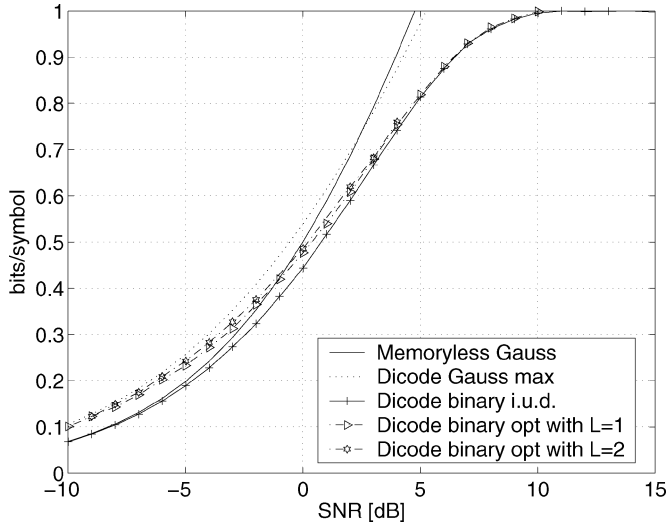

Fig. 4. Information rates of the $1 - D$ (dicode) channel.

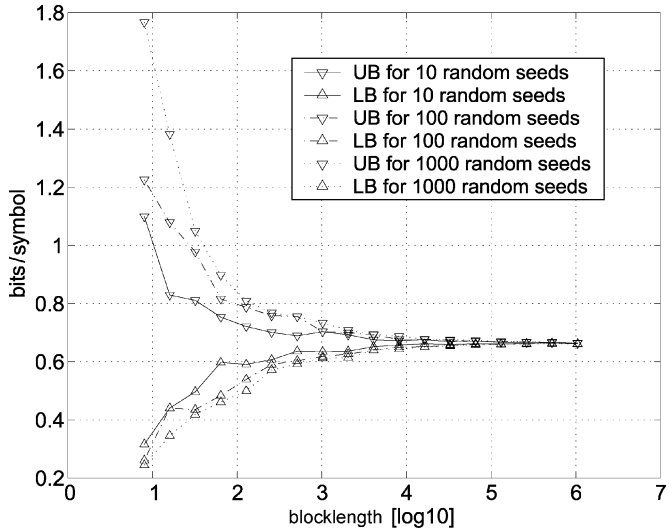

Fig. 5. Convergence of the algorithm.

The signal-to-noise ratio (SNR) will be defined as

$$\mathrm{SNR} \triangleq \frac{\mathrm{E}\left[X_k^2\right] \sum_{k=0}^{m} g_k^2}{\mathrm{E}\left[Z_k^2\right]}. \tag{19}$$

(It is clear that this SNR definition is inadequate for some applications, but this qualification seems to apply also to alternative definitions including that of [44].) For channels, as in Example 1, $h(Y|X) = h(Z)$ is known analytically, which means that the algorithm of Section III 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 to obtain reliable plots (cf. the discussion of Fig. 5 below).

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 [43] or (in this simple example) by brute force. For comparison, Fig. 4 also shows:

1) The capacity of the memoryless AWGN channel.
2) The capacity of the dicode channel for *Gaussian* (rather than binary) input.

The latter is obtained by the well-known waterfilling principle [12]. As the definition (19) allows the channel to provide a power gain for nonwhite 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 1110 times, each time by a simulation run of 10^6 symbols and with a new random seed (both for the pseudorandom channel input sequence and for the pseudorandom noise sample sequence). For every block length 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. As the figure shows, all these 1110 independent estimates converge very nicely to the same value, up to the accuracy of the plot. This kind of good-natured convergence was encountered in all our numerical experiments with many different channels.

The convergence slows down, of course, if a higher accuracy is required; in fact, for most channels, it is not feasible to obtain more than three decimal digits of the information rate.

Note that plots such as Fig. 5 give a partial answer to the practical need to choose a sequence length n: for some candidate length n, run the algorithm about 10 times (each time with a new random seed) and check whether all estimates of the information rate agree up to the desired accuracy.

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 [21].) 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 [21] as well as a conjectured lower bound on the channel capacity due to Shamai and Laroia [37]. The latter can be computed by evaluating a single one-dimensional integral. Fig. 7 shows several rates for the channel of Fig. 6, each evaluated at -5, 3, and 8 dB.

1) $I_{\mathrm{H1}}(n)$ (see below) as a function of n.
2) $I_{\mathrm{H2}}(n)$ (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 the algorithm of Section III).

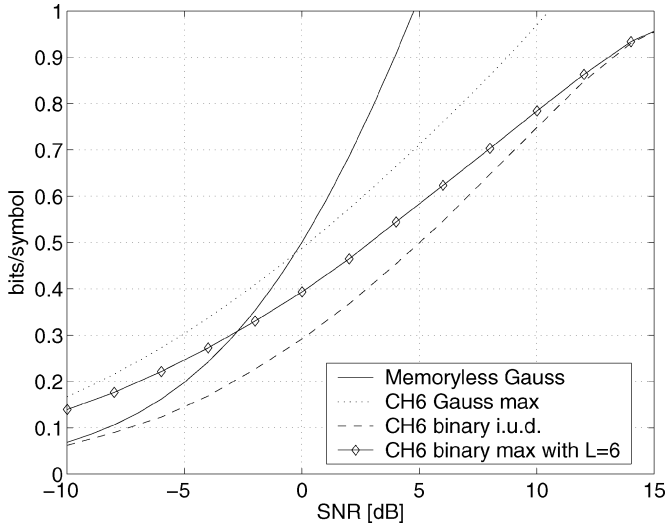

Fig. 6. Information rates of an FIR channel with memory 6.

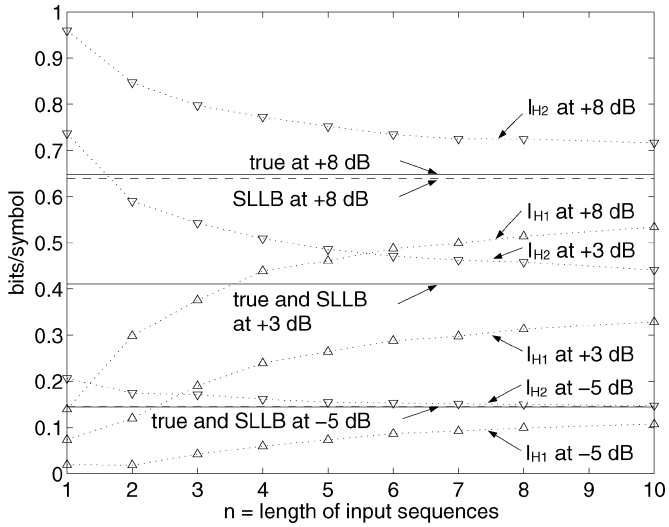

Fig. 7. Comparison with Hirt's method and the Shamai–Laroia conjectured lower bound (SLLB) for the channel of Fig. 6.

As the figure shows, the SLLB is extremely tight for low SNR.

Hirt defined

$$I_{\mathrm{H1}}(n) \triangleq \frac{1}{n} I(X^n; Y^n | X_0, X_{-1}, \ldots, X_{1-m}) \tag{20}$$

and

$$I_{\mathrm{H2}}(n) \triangleq \frac{1}{n} I(X^n; Y^{n+m} | X_0, \ldots, X_{1-m}, X_{n+1}, \ldots, X_{n+m}), \tag{21}$$

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

$$I_{\mathrm{H1}}(n) \le I_{\mathrm{H2}}(n) \tag{22}$$

and

$$\lim_{n\to\infty} I_{\mathrm{H1}}(n) = \lim_{n\to\infty} I_{\mathrm{H2}}(n) = I(X;Y). \tag{23}$$

V. Extensions

A. Continuous Input Alphabet

As mentioned in Section II, the assumption that the input alphabet $\mathcal{X}$ is finite is by no means essential. Assume, for example, that $\mathcal{X} = \mathbb{R}$ and that $p(x^n)$ 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 III is that the recursion (15) becomes

$$\mu_k(s_k) = \lambda_k \sum_{s_{k-1}} \mu_{k-1}(s_{k-1}) \int_{-\infty}^{\infty} p(x_k, y_k, s_k | s_{k-1})\, dx_k \tag{24}$$

which may be evaluated analytically or numerically.

B. Time-Varying And/Or Nonergodic 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 III, 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 compute $\log p(x^n)$, $\log p(y^n)$, and $\log p(x^n, y^n)$ for each sample sequence. By averaging over these quantities, we obtain estimates of $\frac{1}{n}H(X^n)$, of $\frac{1}{n}H(Y^n)$, and of $\frac{1}{n}I(X^n;Y^n)$. The significance of these quantities depends on the application.

C. Bounds on Entropy Rates From Reduced-State Recursions

The basic recursion (12) 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 (12) is modified to

$$\mu_k(s_k) = \sum_{x_k} \sum_{s_{k-1} \in \mathcal{S}'_{k-1}} \mu_{k-1}(s_{k-1})\, p(x_k, y_k, s_k | s_{k-1}) \tag{25}$$

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 (12) 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 lower bounds on $h(Y)$. A particular case is worked out in Appendix III.

The upper bound can also be applied to certain nonfinite-state channels as follows. Consider, e.g., the autoregressive channel of Fig. 9 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 (12) until there are too many of them, and then we switch to the reduced-state recursion (25).

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

VI. 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 III–V) 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

$$q_p(y) \triangleq \sum_x p(x)\, q(y|x). \tag{26}$$

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):

$$I(X;Y) \le \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q_p(y)} \tag{27}$$

$$= \mathrm{E}_p\left[\log p(Y|X) - \log q_p(Y)\right] \tag{28}$$

where the sum in (27) should be read as running over the support of $p(x,y)$. Equality holds in (27) if and only if $p(y) = q_p(y)$ for all y. □

This bound appears to have been observed first by Topsøe [41]. The proof is straightforward. Let $\overline{I}_q(X;Y)$ be the right-hand side of (27). Then

$$\overline{I}_q(X;Y) - I(X;Y)$$

$$= \sum_{x,y} p(x,y) \left[\log \frac{p(y|x)}{q_p(y)} - \log \frac{p(y|x)}{p(y)}\right] \tag{29}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(y)}{q_p(y)} \tag{30}$$

$$= \sum_{y} p(y) \log \frac{p(y)}{q_p(y)} \tag{31}$$

$$= D\left(p(y) \| q_p(y)\right) \tag{32}$$

$$\ge 0. \tag{33}$$

Theorem (Auxiliary-Channel Lower Bound):

$$I(X;Y) \ge \sum_{x,y} p(x,y) \log \frac{q(y|x)}{q_p(y)} \tag{34}$$

$$= \mathrm{E}_p\left[\log q(Y|X) - \log q_p(Y)\right] \tag{35}$$

where the sum in (34) should be read as running over the support of $p(x,y)$. □

This bound is implicit in the classic papers by Blahut [10] and Arimoto [1]. Moreover, it may also be obtained as a special case of a bound due to Fischer [15] on mismatched decoding, which in turn is a special case of a general result by Ganti *et al.* [17, eq. (12) for $s = 1$]. It then follows from the results in [15] and [17] that the lower bound is achievable by a maximum-likelihood decoder for the auxiliary channel.

A simple proof of (34) goes as follows. Let $\underline{I}_q(X;Y)$ be the right-hand side of (34) and for y satisfying $p(y) > 0$ (which by the assumption after (26) implies $q_p(y) > 0$) let

$$r_p(x|y) \triangleq \frac{p(x)q(y|x)}{q_p(y)} \tag{36}$$

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

$$I(X;Y) - \underline{I}_q(X;Y)$$

$$= \sum_{x,y} p(x,y) \left[\log \frac{p(x,y)}{p(x)p(y)} - \log \frac{q(y|x)}{q_p(y)}\right] \tag{37}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)p(x)q(y|x)/q_p(y)} \tag{38}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)r_p(x|y)} \tag{39}$$

$$= D\left(p(x,y) \| p(y)r_p(x|y)\right) \tag{40}$$

$$\ge 0. \tag{41}$$

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

$$\overline{I}_q(X;Y) - \underline{I}_q(X;Y) = D\left(p(x)p(y|x) \| p(x)q(y|x)\right). \tag{42}$$

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

$$\overline{I}_q(X;Y) \triangleq \lim_{n\to\infty} \mathrm{E}_p\left[-\frac{1}{n}\log q_p(Y^n) + \frac{1}{n}\log p(Y^n|X^n)\right] \tag{43}$$

and the lower bound becomes

$$\underline{I}_q(X;Y) \triangleq \lim_{n\to\infty} \mathrm{E}_p\left[-\frac{1}{n}\log q_p(Y^n) + \frac{1}{n}\log q(Y^n|X^n)\right]. \tag{44}$$

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 III.
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 III.
4) Conclude with the estimates

$$\hat{\bar{I}}_q(X;Y) \triangleq -\frac{1}{n}\log q_p(y^n) - h(Y|X) \tag{45}$$

and

$$\hat{\underline{I}}_q(X;Y) \triangleq -\frac{1}{n}\log q_p(y^n) - \frac{1}{n}\log p(x^n) + \frac{1}{n}\log q(y^n|x^n)p(x^n). \tag{46}$$

Note that the term $h(Y|X)$ in the upper bound (45) 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, (45) and (46) should converge with probability one to (43) and (44), respectively. Sufficient conditions for the existence of such limits are discussed in [31], [9], [27], [14, Sec. 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

$$\mathrm{E}_p\left[\left|\log q_p(Y_k|s_{k-1}, s_k, x_k)\right|\right] < \infty$$

for all s_{k-1}, s_k, x_k.

Quantities very similar to (43) and (44) seem to have been computed by essentially the same algorithm as far back as 1985, cf. [25].

VII. Numerical Examples for the Bounds

We illustrate the methods of Sections V-C and VI by some numerical examples. As in Section IV, 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. 8 shows the following curves.

1) Bottom: the exact information rate computed as described in Section III.
2) Top: the reduced-state upper bound (RSUB) of Section V-C, 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.

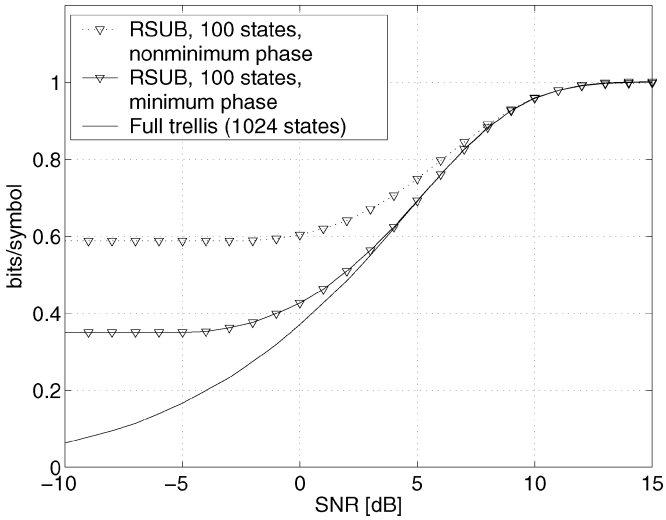

Fig. 8. Memory-10 FIR filter.

The notion of a minimum-phase filter is reviewed in Appendix I, 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 II. The motivation for this replacement is that minimum-phase filters concentrate the signal energy into the leading tap weights [30], which makes the reduced-state bound tighter.

It is obvious from Fig. 8 that the reduced-state upper bound works fine for high SNR and becomes useless for low SNR. This may be explained by noting that, for high SNR, only very few states carry substantial probability mass; for low SNR, however, the probability mass is spread over almost all states.

Our next example is the channel of Fig. 9 with an autoregressive filter

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

for $\alpha = 0.8$. We apply the auxiliary-channel bound of Section VI, where the auxiliary channel is obtained from the original channel by inserting a uniform quantizer in the feedback loop, which results in the finite-state channel of Fig. 10. Both the range of the quantizer and the noise variance of the auxiliary channel are 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. 10 with 512 states.
2) 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. For this same setup, Fig. 12 shows these two bounds as a function of the number of states (for SNR= 7.45 dB).

VIII. Conclusion

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

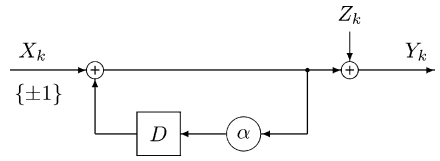

Fig. 9. Autoregressive-filter channel.

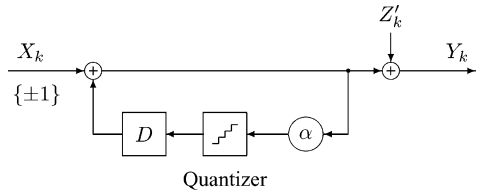

Fig. 10. A quantized version of Fig. 9.

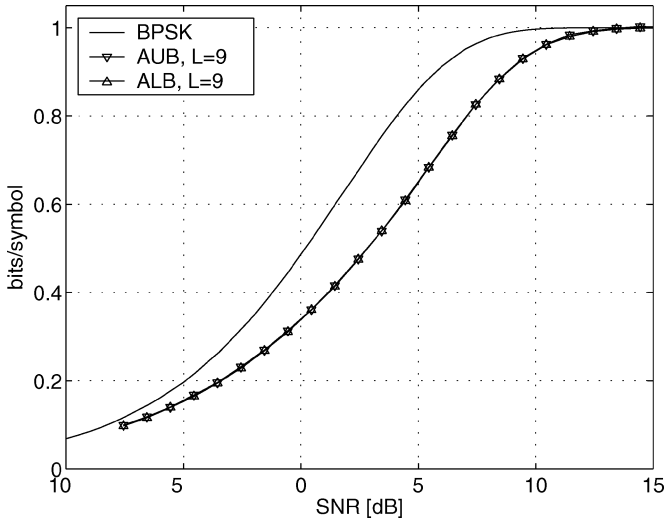

Fig. 11. Bounds for Fig. 9 versus SNR. $L \triangleq \log_2$ (number of states).

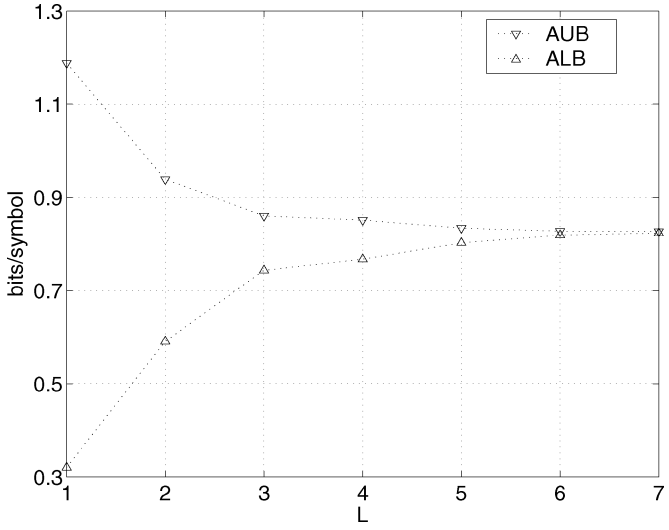

Fig. 12. Bounds for Fig. 9 versus $L \triangleq \log_2$ (number of states).

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.

APPENDIX I
ON MINIMUM-PHASE FILTERS

This appendix summarizes some basic and well-known facts on discrete-time linear time-invariant (LTI) systems, cf. [30].

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

$$G(z) \triangleq \sum_{k\in\mathbb{Z}} g_k z^{-k} \tag{47}$$

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 (47)) 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 g is right-sided, then $r_2 = \infty$. If S_g contains the 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 such that all poles of $F(z)$ are inside the unit circle and no zeros of $F(z)$ lie on the unit circle. Then $F(z)$ can be written as

$$F(z) = G(z)H(z) \tag{48}$$

where $G(z)$ is minimum-phase and $H(z)$ is an all-pass filter. 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. □

(A proof of this theorem may be obtained from the following observations.) Clearly, $|G(e^{i\Omega})| = |F(e^{i\Omega})|$ for all $\Omega \in \mathbb{R}$. For

$$F(z) = \frac{\prod_{k=1}^{m}(z - z_k)}{\prod_{\ell=1}^{n}(z - p_\ell)} \tag{49}$$

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

$$G(z) = \frac{z^{n-m}\prod_{k:|z_k|<1}(z - z_k)\prod_{k:|z_k|>1}(1 - z\overline{z_k})}{\prod_{\ell=1}^{n}(z - p_\ell)} \tag{50}$$

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

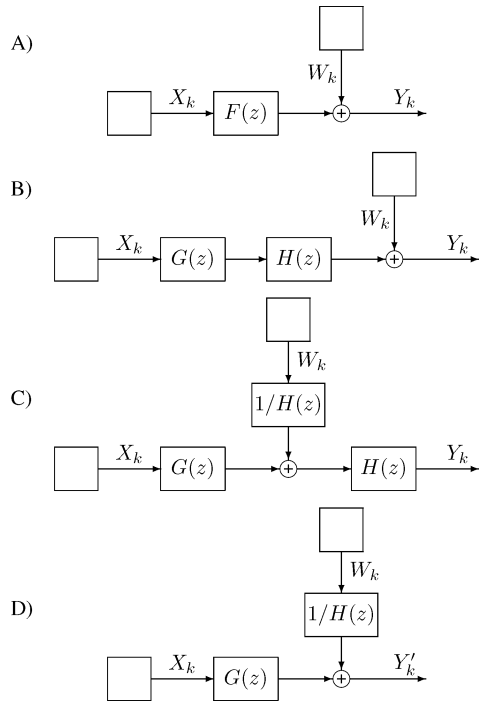

Fig. 13. On linear channels with additive noise.

Appendix II
On Linear Channels With Additive Noise

Consider the channel of Fig. 13 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 (48) and (50)) does not change the information rate $I(X;Y)$.
2) The case of colored Gaussian noise without a spectral null (as defined below) can be converted into the case of white Gaussian noise.

We begin with the first case. Clearly, when $F(z)$ is decomposed according to (48), the information rate $I(X;Y)$ remains unchanged (Fig. 13 B)). It is then obvious that the channel of Fig. 13 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. 13 D) equals $I(X;Y)$ of the original channel of Fig. 13 A). Finally, the (noncausal stable) all-pass filter $1/H(z)$ in Fig. 13 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. 13 D), where W is white Gaussian noise and where $1/H(z)$ is (the transfer function of) a suitable filter. The filter $G(z)$ is arbitrary; in particular, we could have $G(z) = 1$.

We now assume that $1/H(z)$ is rational with all poles inside the unit circle and without zeros on the unit circle. In this case, we can and we will assume without loss of generality that $1/H(z)$ (and thus also $H(z)$) is minimum-phase. Appending the minimum-phase filter $H(z)$ at the output (which results in Fig. 13 C)) does not change the information rate. As before, Fig. 13 C) and B) are equivalent, and defining $F(z) = G(z)H(z)$, all channels in Fig. 13 have again the same information rate. If the noise-coloring filter $1/H(z)$ is autoregressive, $H(z)$ is an FIR filter.

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

In Section V-C, it was pointed out that omitting states in the basic recursion (12) 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

$$Y_k = \sum_{\ell=0}^{m} g_\ell X_{k-\ell} + Z_k \tag{51}$$

with channel memory $m \in \mathbb{N}$, 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

$$\begin{aligned}&(x_{k-1}, x_{k-2}, \ldots, x_{k-M})\\ &\triangleq \bigcup_{x_{k-M-1},\ldots,x_{k-m}} \{(x_{k-1}, \ldots, x_{k-M}, x_{k-M-1}, \ldots, x_{k-m})\}\end{aligned} \tag{52}$$

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

As in Section V-C, 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 (12) with all these states until there are too many of them. From that moment on, we merge states into the form (52), 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

$$p(y_k | x_k, x_{k-1}, \ldots, x_{k-m}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_k - w)^2/(2\sigma^2)} \tag{53}$$

with

$$w \triangleq w(x_k, \ldots, x_{k-m}) \triangleq g_0 x_k + \sum_{\ell=1}^{m} g_\ell x_{k-\ell}. \tag{54}$$

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

$$w_{\mathrm{U}} \triangleq w_{\mathrm{U}}(x_k,\ldots,x_{k-M}) \tag{55}$$

$$\triangleq g_0 x_k + \sum_{\ell=1}^{M} g_\ell x_{k-\ell} + \sum_{\ell=M+1}^{m} |g_\ell| \tag{56}$$

and

$$w_{\mathrm{L}} \triangleq w_{\mathrm{L}}(x_k,\ldots,x_{k-M}) \tag{57}$$

$$\triangleq g_0 x_k + \sum_{\ell=1}^{M} g_\ell x_{k-\ell} - \sum_{\ell=M+1}^{m} |g_\ell|. \tag{58}$$

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

$$p(y_k|x_k,x_{k-1},\ldots,x_{k-m}) \le \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_k-\hat{w})^2/(2\sigma^2)}, \tag{59}$$

where

$$\hat{w} \triangleq \hat{w}(x_k,\ldots,x_{k-M},y_k) \triangleq \begin{cases} w_{\mathrm{L}}, & \text{if } y_k < w_{\mathrm{L}} \\ w_{\mathrm{U}}, & \text{if } y_k > w_{\mathrm{U}} \\ y_k, & \text{else} \end{cases} \tag{60}$$

depends only on the merged state. Using the right-hand side of (59) in the recursion (12) 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 VI. It should be noted, however, that the latter bound depends on the auxiliary channel; if no good auxiliary channel model is available, the bound of this section may be the method of choice.

ACKNOWLEDGMENT

We would like to thank Amos Lapidoth and Stefan M. Moser, whose work made us aware of Topsøe's inequality (cf. (27)). Moreover, we are grateful to Amos Lapidoth for pointing out to us [15] and [17].

REFERENCES

[1] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, 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," in *Proc. 2000 IEEE Int. Symp. Information Theory*, Sorrento, Italy, Jun. 2000, p. 433.

[3] D. Arnold and H.-A. Loeliger, "On the information rate of binary-input channels with memory," in *Proc. 2001 IEEE Int. Conf. Communications*, Helsinki, Finland, Jun. 2001, pp. 2692–2695.

[4] ——, "On finite-state information rates from channel simulations," in *Proc. 2002 IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, Jun./Jul. 2002, p. 164.

[5] D. M. Arnold, "Computing Information Rates of Finite-State Models with Application to Magnetic Recording," ETH-Diss 14760 (Ph.D dissertation), ETH Zurich, Zurich, Switzerland, 2003.

[6] D. Arnold, H.-A. Loeliger, and P. O. Vontobel, "Computation of information rates from finite-state source/channel models," in *Proc. 40th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 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," in *Proc. 2003 IEEE Int. Symp. Information Theory*, Yokohama, Japan, Jun./Jul. 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. Inf. Theory*, vol. IT-20, no. 2, pp. 284–287, Mar. 1974.

[9] A. R. Barron, "The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem," *Ann. Probab.*, vol. 13, no. 4, pp. 1292–1303, 1985.

[10] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.

[11] J. Chen and P. H. Siegel, "On the symmetric information rate of two-dimensional finite-state ISI channels," in *Proc. 2003 IEEE Information Theory Workshop*, Paris, France, Mar./Apr. 2003, pp. 320–323.

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

[13] J. Dauwels and H.-A. Loeliger, "Computation of information rates by particle methods," in *Proc. 2004 IEEE Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 178.

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

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

[16] G. D. Forney, Jr., "Codes on graphs: Normal realizations," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 520–548, Feb. 2001.

[17] A. Ganti, A. Lapidoth, and İ. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

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

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

[20] C. Heegard, A. Duel-Hallen, and R. Krishnamoorthy, "On the capacity of the noisy runlength channel," *IEEE Trans. Inf. Theory*, vol. 37, no. 3, pp. 712–720, May 1991.

[21] W. Hirt, "Capacity and Information Rates of Discrete-Time Channels with Memory," ETH-Diss 8671 (Ph.D. dissertation), ETH Zurich, Zurich, Switzerland, 1988.

[22] T. Holliday, A. Goldsmith, and P. Glynn, "Entropy and mutual information for Markov channels with general inputs," in *Proc. 40th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2002, pp. 824–833.

[23] T. Holliday, A. Goldsmith, and P. Glynn, "Capacity of finite-state channels based on Lyapunov exponents of random matrices," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3509–3532, Aug. 2006.

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

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

[26] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[27] B. G. Leroux, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Processes Their Applic.*, vol. 40, pp. 127–143, 1992.

[28] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Process. Mag.*, pp. 28–41, Jan. 2004.

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

[30] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[31] T. Petrie, "Probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 40, no. 1, pp. 97–115, 1969.

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

[33] R. Pighi, R. Raheli, and F. Cappelletti, "Information rates of multidimensional front-ends for digital storage channels with data-dependent transition noise," in *Proc. 2005 IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 1343–1347.

[34] W. E. Ryan, F. Wang, R. Wood, and Y. Li, "Optimal code rates for the Lorentzian channel: Shannon codes and LDPC codes," *IEEE Trans. Magn.*, vol. 40, no. 6, pp. 3559–3565, Nov. 2004.

[35] S. Shamai (Shitz) and Y. Kofman, "On the capacity of binary and Gaussian channels with run-length-limited inputs," *IEEE Trans. Commun.*, vol. 38, no. 5, pp. 584–594, May 1990.

[36] S. Shamai (Shitz), L. H. Ozarow, and A. D. Wyner, "Information rates for a discrete-time Gaussian channel with intersymbol interference and stationary inputs," *IEEE Trans. Inf. Theory*, vol. 37, no. 6, pp. 1527–1539, Nov. 1991.

[37] S. Shamai (Shitz) and R. Laroia, "The intersymbol interference channel: Lower bounds on capacity and channel precoding loss," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1388–1404, Sep. 1996.

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

[39] O. Shental, N. Shental, and S. Shamai (Shitz), "On the achievable information rates of finite-state input two-dimensional channels with memory," in *Proc. 2005 IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 2354–2358.

[40] J. B. Soriaga, P. H. Siegel, J. K. Wolf, and M. Marrow, "On achievable rates of multistage decoding on two-dimensional ISI channels," in *Proc. 2005 IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 1348–1352.

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

[42] P. O. Vontobel and D. M. Arnold, "An upper bound on the capacity of channels with memory and constrained input," in *Proc. 2001 IEEE Information Theory Workshop*, Cairns, Australia, Sep. 2001, pp. 147–149.

[43] P. O. Vontobel, A. Kavčić, D. M. Arnold, and H.-A. Loeliger, "A generalization of the Blahut-Arimoto algorithm to finite-state channels," *IEEE Trans. Inf. Theory*, to be published.

[44] W. Xiang and S. S. Pietrobon, "On the capacity and normalization of ISI channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 9, pp. 2263–2268, Sep. 2003.

[45] S. Yang, A. Kavčić, and S. Tatikonda, "Delayed feedback capacity of finite-state machine channels: Upper bounds on the feedforward capacity," in *Proc. 2003 IEEE Int. Symp. Information Theory*, Yokohama, Japan, Jun./Jul. 2003, p. 290.

[46] E. Zehavi and J. K. Wolf, "On runlength codes," *IEEE Trans. Inf. Theory*, vol. 34, no. 1, pp. 45–54, Jan. 1988.

[47] W. Zeng, J. Tokas, R. Motwani, and A. Kavčić, "Bounds on mutual information rates of noisy channels with timing errors," in *Proc. 2005 IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 709–713.

[48] Z. Zhang, T. M. Duman, and E. M. Kurtas, "Information rates of binary-input intersymbol interference channels with signal-dependent media noise," *IEEE Trans. Magn.*, vol. 39, no. 1, pp. 599–607, Jan. 2003.

[49] ——, "Achievable information rates and coding for MIMO systems over ISI channels and frequency-selective fading channels," *IEEE Trans. Commun.*, vol. 52, no. 10, pp. 1698–1710, Oct. 2004.