Computation of Information Rates from Finite-State Source/Channel Models

Dieter Arnold
arnold@isi.ee.ethz.ch

Hans-Andrea Loeliger
loeliger@isi.ee.ethz.ch

Pascal O. Vontobel
vontobel@isi.ee.ethz.ch

Signal & Information Proc. Lab. (ISI)
ETH Zentrum
CH-8092 Zürich, Switzerland

Abstract

It has recently become feasible to compute information rates of finite-state source/channel models with not too many states. We review such methods and demonstrate their extension to compute upper and lower bounds on the information rate of very general (non-finite-state) channels by means of finite-state approximations.

1 Introduction

We consider the problem of computing the information rate

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

between the input process $X = (X_1, X_2, \ldots)$ and the output process $Y = (Y_1, Y_2, \ldots)$ of a time-invariant 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. (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, but the sequence $I_1, I_2, I_3, \ldots$ converges rather slowly even for very simple examples.

For *finite-state* channels (to be defined in Section 2), a practical method for the computation of (1) was recently presented independently by Arnold and Loeliger [2], by Sharma and Singh [17], and by Pfister et al. [16]. The new method consists essentially of sampling both a long input sequence $x^n \triangleq (x_1,\ldots,x_n)$ and the corresponding output sequence $y^n \triangleq (y_1,\ldots,y_n)$, followed by the computation of $\log p(y^n)$ (and, if necessary, of $\log p(y^n|x^n)$) by means of a forward sum-product recursion on the joint source/channel trellis. We will review this method in Section 3.

In Section 4, we show that essentially the same method can be used to compute upper and lower bounds on the information rate of very general channels with memory. (The upper bound was presented in [3].) The basic idea is to approximate the given "difficult" channel by a finite-state model; we then use simulated (or measured) input/output pairs from the actual channel as inputs to a computation on the trellis of the finite-state model. 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 $h(Y|X)$. A numerical example is given in Section 5.

To conclude this introduction, we wish to mention same earlier and some related recent work on similar topics. Hirt [10] proposed a Monte-Carlo method to evaluate lower and upper bounds on the i.u.d. rate of binary-input intersymbol interference channels (see Example 1 below). Shamai et al. [18] [19] 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. Mushkin and Bar-David [15] analyzed the Gilbert-Elliot channel and Goldsmith and Varaiya [9] extended that work to general channels with a freely evolving state (see Example 2 below); they gave expressions for the channel capacity and the information rate as well as recursive methods for their evaluation.

Subsequent to [2] [17], Kavčić presented a highly nontrivial generalization of the Blahut-Arimoto algorithm to maximize the information rate over finite-state Markov sources [11]. Vontobel and Arnold [21] proposed an algorithm to compute an upper bound on the capacity of finite-state channels; that algorithm appears to be practical only for small examples, however. Many of these topics are discussed in [4]; none of these topics will be further considered in the present paper.

2 Finite-State Source/Channel Models

We will assume that X, Y, and $S = (S_0, S_1, S_2, \ldots)$ are stochastic processes such that

$$p(x_1, \ldots, x_n, y_1, \ldots, y_n, s_0, \ldots, s_n) = p(s_0) \prod_{k=1}^{n} p(x_k, y_k, s_k | s_{k-1}) \tag{3}$$

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 a *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.

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

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

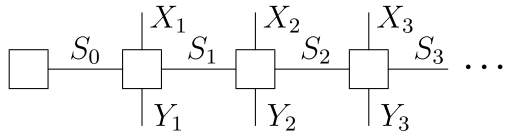

Figure 1: The (Forney-style) factor graph of (3).

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.,

$$p(x_k|x_{k-1}, x_{k-2}, \ldots) = p(x_k|x_{k-1}, \ldots, x_{k-L}),$$

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

Example 2 (Channels 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

$$p(y_1, \ldots, y_n, s'_0, \ldots, s'_n|x_1, \ldots, x_n) = p(s'_0) \prod_{k=1}^{n} p(y_k|x_k, s'_{k-1})\, p(s'_k|s'_{k-1})$$

for all $n > 0$. If X is Markov of order L, then (3) holds for $S_k \triangleq (S'_k, X_k, \ldots, X_{k-L+1})$. This class of channels, which includes the Gilbert-Elliot channel, was investigated in [9].

Under the stated assumptions, 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. [6] [13].

We conclude this section by noting that the factorization (3) may be expressed by the graph of Fig. 1. (This graph is a Forney-style factor graph, see [8] [14]; add a circle on each branch to obtain a factor graph as in [12].) From this graph, the computations described in the next section will be obvious.

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

From the above remarks, 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.

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

$$\hat{I}(X;Y) = \frac{1}{n}\log p(x^n, y^n) - \frac{1}{n}\log p(x^n) - \frac{1}{n}\log p(y^n), \tag{4}$$

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

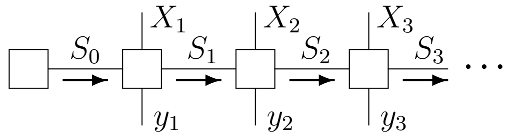

Figure 2: Computation of $p(y^n)$ by message passing through Fig. 1.

Obviously, this algorithm is practical only if the computations in Step 2 are feasible. For finite-state source/channel models as defined in Section 2, these computations can be carried out by forward sum-product message passing through the graph of Fig. 1, as illustrated in Fig. 2. Since Fig. 1 represents a trellis, this computation is just the forward sum-product recursion of the BCJR algorithm [5].

Consider, for example, the computation of

$$p(y^n) = \sum_{x^n, s^n} p(x^n, y^n, s^n) \tag{5}$$

with $s^n \triangleq (s_0, s_1, \ldots, s_n)$. By straightforward application of the sum-product algorithm (cf. [12] [8]), we recursively compute the messages (i.e., state metrics)

$$\begin{aligned} \mu_{\mathrm{f}}(s_k) &= \sum_{x_k, s_{k-1}} \mu_{\mathrm{f}}(s_{k-1})\, p(x_k, y_k, s_k | s_{k-1}) & (6) \\ &= \sum_{x^k, s^{k-1}} p(x^k, y^k, s^k) & (7) \end{aligned}$$

for $k = 1, 2, 3, \ldots$, as illustrated in Fig. 2. The desired quantity (5) is then obtained as

$$p(y^n) = \sum_{s_n} \mu_{\mathrm{f}}(s_n), \tag{8}$$

the sum of all final state metrics.

For large n, the state metrics $\mu_{\mathrm{f}}(.)$ computed according to (6) quickly tend to zero. In practice, the recursion rule (6) is therefore changed to

$$\mu'_{\mathrm{f}}(s_k) = \lambda_k \sum_{x_k, s_{k-1}} \mu'_{\mathrm{f}}(s_{k-1})\, p(x_k, y_k, s_k | s_{k-1}) \tag{9}$$

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

$$\frac{1}{n} \sum_{k=1}^{n} \log \lambda_k = -\frac{1}{n} \log p(y^n). \tag{10}$$

The quantity $-\frac{1}{n} \log p(y^n)$ thus appears as the sum 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. If there is no feedback from the channel to the source, the computation of $\log p(x^n)$ uses only the source model rather than the joint source/channel model.

4 Bounds on $I(X;Y)$ for General Channels

The methods of the previous section can be extended to compute upper and lower bounds on the information rate of very general (non-finite-state) channels. For the sake of clarity, we begin by stating the bounds for the discrete memoryless case.

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

$$q_p(y) \triangleq \sum_x p(x)\, q(y|x). \tag{11}$$

Theorem (Upper-Bound):

$$\begin{aligned} I(X;Y) &\leq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q_p(y)} && (12)\\ &= \mathrm{E}_{p(x,y)}\big[\log p(Y|X) - \log q_p(Y)\big]\,. && (13) \end{aligned}$$

This bound appears to have been observed first by Topsøe [20]. (It was brought to our attention by recent work of A. Lapidoth.) The proof is straightforward. Let $\overline{I}_q(X;Y)$ be the right-hand side of (12). Then

$$\begin{aligned} \overline{I}_q(X;Y) - I(X;Y) &= \sum_{x,y} p(x,y) \left[\log \frac{p(y|x)}{q_p(y)} - \log \frac{p(y|x)}{p(y)}\right] && (14)\\ &= \sum_{x,y} p(x,y) \log \frac{p(y)}{q_p(y)} && (15)\\ &= \sum_{y} p(y) \log \frac{p(y)}{q_p(y)} && (16)\\ &= D\big(p(y)||q_p(y)\big) && (17)\\ &\geq 0. && (18) \end{aligned}$$

Theorem (Lower Bound):

$$\begin{aligned} I(X;Y) &\geq \sum_{x,y} p(x,y) \log \frac{q(y|x)}{q_p(y)} && (19)\\ &= \mathrm{E}_{p(x,y)}\big[\log q(Y|X) - \log q_p(Y)\big]\,. && (20) \end{aligned}$$

This bound is implicit in the classical papers by Blahut [7] and Arimoto [1]. The proof goes as follows. Let $\underline{I}_q(X;Y)$ be the right-hand side of (19) and let

$$r_p(x|y) \triangleq \frac{p(x)q(y|x)}{q_p(y)} \tag{21}$$

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

$$I(X;Y) - \underline{I}_q(X;Y) = \sum_{x,y} p(x,y) \left[\log \frac{p(x,y)}{p(x)p(y)} - \log \frac{q(y|x)}{q_p(y)} \right] \tag{22}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)p(x)q(y|x)/q_p(y)} \tag{23}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)r_p(x|y)} \tag{24}$$

$$= D\big(p(x,y) || p(y)r_p(x|y)\big) \tag{25}$$

$$\geq 0. \tag{26}$$

It is obvious from these proofs that both the upper bound (12) and the lower bound (19) are tight if and only if $p(x)q(y|x) = p(x,y)$ for all x and y.

The generalization of these bounds to the information rate of channels with memory is straightforward: the upper bound becomes

$$\overline{I}_q(X;Y) \triangleq \lim_{n\to\infty} \mathrm{E}_{p(\cdot,\cdot)} \left[\frac{1}{n} \log p(Y^n|X^n) - \frac{1}{n} \log q_p(Y^n) \right] \tag{27}$$

and the lower bound becomes

$$\underline{I}_q(X;Y) \triangleq \lim_{n\to\infty} \mathrm{E}_{p(\cdot,\cdot)} \left[\frac{1}{n} \log q(Y^n|X^n) - \frac{1}{n} \log q_p(Y^n) \right]. \tag{28}$$

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 2.

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

$$\hat{\overline{I}}_q(X;Y) = -\frac{1}{n} \log q_p(y^n) - h(Y|X) \tag{29}$$

and

$$\hat{\underline{I}}_q(X;Y) = \frac{1}{n} \log q(y^n|x^n)p(x^n) - \frac{1}{n} \log p(x^n) - \frac{1}{n} \log q_p(y^n). \tag{30}$$

Note that the term $h(Y|X)$ in the upper bound (29) refers to the original channel and cannot be computed by means of the auxiliary channel.

5 An Example

Consider the channel consisting of a linear filter with impulse response

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

and additive white Gaussian noise with variance σ^2, as illustrated in Fig. 3. The channel input is restricted to $\{+1, -1\}$.

A natural finite-state approximation is obtained by truncating the impulse response. Another finite-state approximation is obtained by inserting a quantizer in the feedback loop as shown in Fig. 4. Note that the channel of Fig. 4 is nonlinear.

Some numerical results for this example are shown in Fig. 5. The figure shows the upper bound and the lower bound on the i.u.d. information rate, both for the truncated impulse response model and for the quantized-feedback model. The horizontal axis shows $\log_2 M$, where M is the number of states of the finite-state model. The particular numbers shown in Fig. 5 correspond to the values $\alpha = 0.8$ and $\sigma^2 = 1$. The quantizer in Fig. 4 was chosen to be a uniform quantizer optimized to give as good bounds as possible; the parameter σ' in Fig. 4 was also optimized. As Fig. 5 shows, the quantized-feedback model yields better bounds with less states than the truncated-impulse-response model.

6 Conclusions

Information rates of finite-state source/channel models (with not too many states) can now be computed accurately. By a new extension of such methods, we can compute upper and lower bounds on the information rate of very general non-finite-state channels (used with finite-state sources) by means of finite-state approximations of the channel. The bounds are tight if the approximation is good. The lower bound requires only that the channel is ergodic and can be simulated (or measured); the upper bound requires also a lower bound on $h(Y|X)$.

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 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.

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

[4] D. Arnold, *Computing Information Rates of Finite-State Models with Application to Magnetic Recording.* ETH-Diss no. 14760, ETH Zürich, Switzerland, 2002.

[5] 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.

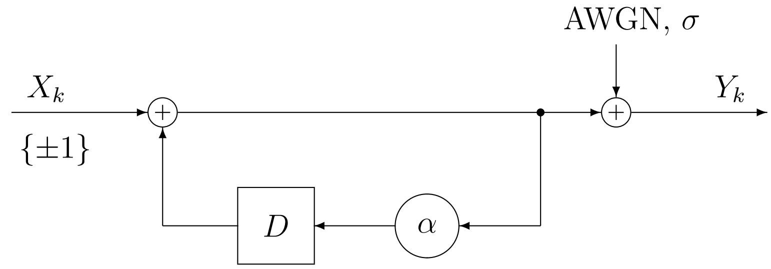

Figure 3: A simple non-finite-state binary-input linear channel.

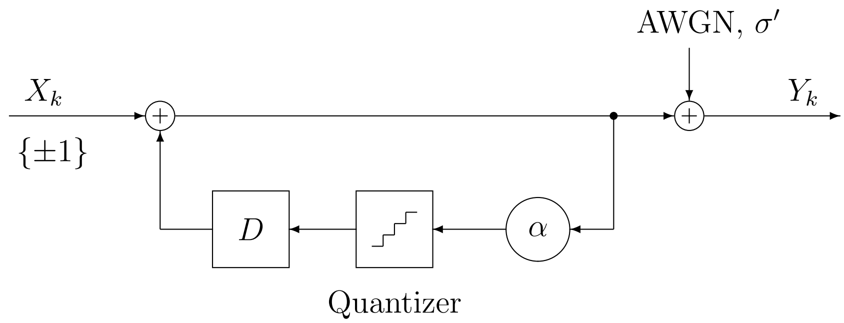

Figure 4: A quantized version of the channel of Figure 3.

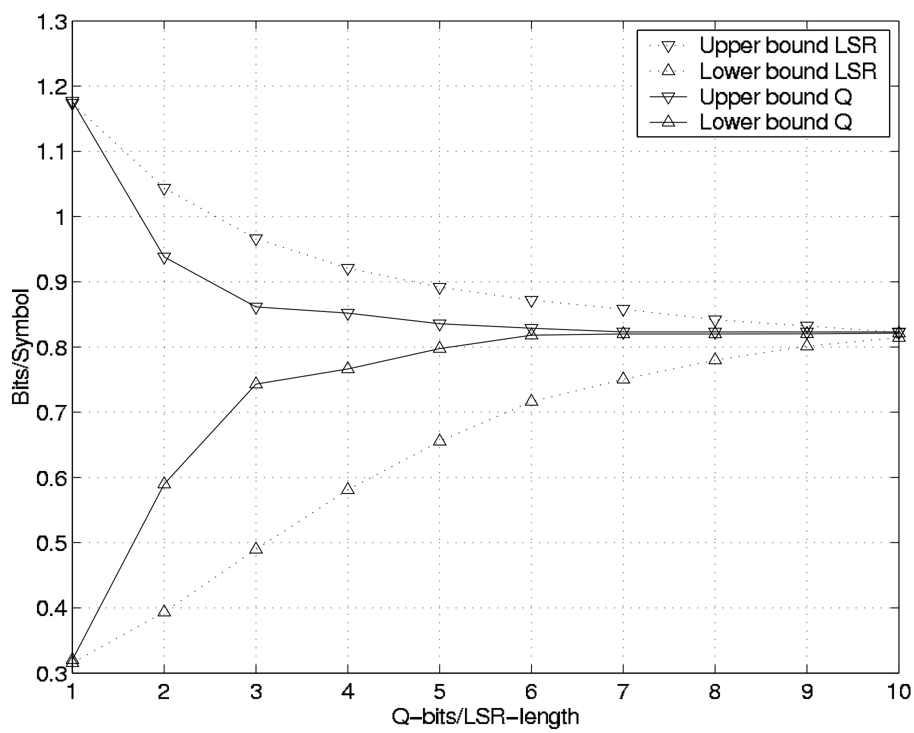

Figure 5: Bounds on the i.u.d. rate of the channel of Figure 3.

[6] A. Barron, "The strong ergodic theorem for densitities: generalized Shannon-McMillan-Breiman theorem," *Annals of Prob.,* vol. 13, no. 4, pp. 1292–1303, 1995.

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

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

[9] 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.

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

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

[12] 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.

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

[14] H.-A. Loeliger, "Least squares and Kalman filtering on Forney graphs," in *Codes, Graphs, and Systems,* (festschrift in honour of David Forney on the occasion of his 60^{th} birthday), R. E. Blahut and R. Koetter, eds., Kluwer, 2002, pp. 113–135.

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

[16] 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. 2001.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.