On the Capacity of Markov Sources over Noisy Channels

Aleksandar Kavčić
Division of Engineering and Applied Sciences
Harvard University, Cambridge, MA 02138

Abstract— **We present an expectation-maximization method for optimizing Markov process transition probabilities to increase the mutual information rate achievable when the Markov process is transmitted over a noisy finite-state machine channel. The method provides a tight lower bound on the achievable information rate of a Markov process over a noisy channel and it is conjectured that it actually maximizes this information rate. The latter statement is supported by empirical evidence (not shown in this paper) obtained through brute-force optimization methods on low-order Markov processes. The proposed expectation-maximization procedure can be used to find tight lower bounds on the capacities of finite-state machine channels (say, partial response channels) or the noisy capacities of constrained (say, run-length limited) sequences, with the bounds becoming arbitrarily tight as the memory-length of the input Markov process approaches infinity. The method links the Arimoto-Blahut algorithm to Shannon's noise-free entropy maximization by introducing the *noisy adjacency matrix*.**

I. Introduction

In his landmark paper [1], Shannon computed the maximal entropy rate of a discrete-time Markov process. This result is widely used to determine the noise-free capacity of constrained sequences (such as the run-length limited sequences) [2], [3]. In the presence of noise, the computation of the capacity of a Markov source when transmitted over a noisy channel has remained an open problem [2], [3]. A related problem is the computation of capacity bounds for partial response channels. In [4], Hirt proposed a Monte-Carlo method to evaluate lower and upper bounds on the i.i.d. rate. Shamai et al. have computed closed-form upper bounds on the capacity, upper and lower bounds on the i.i.d. rate, and a simple i.i.d. rate lower-bound conjecture, which seems to be very tight [5], [6]. Recently, Arnold and Loeliger [7], and independently Pfister et al. [8], devised a Monte-Carlo method to compute the exact value of the information rate of any Markov process transmitted over the channel, which when optimized can deliver a tight lower bound on the capacity. Here we introduce an iterative Markov process optimization method.

Structure: Section II describes the source/channel model and the problem to be addressed. In Section III we reformulate the well-known Arimoto-Blahut algorithm [9], [10] as a *stochastic* expectation-maximization procedure. Motivated by this development, in Section IV we construct a similar stochastic expectation-maximization for Markov sources, linking the method to a *noisy adjacency matrix*, whose computation is shown in Section V. Section VI gives a numeric example by computing a lower bound on the capacity of run-length limited sequences over the binary symmetric channel. Section VII concludes the paper.

Notation: The superscript $^{\mathrm{T}}$ denotes matrix and vector transposition. Random variables are denoted by uppercase letters, while their realizations are denoted by lowercase letters. If a random variable is a member of a random sequence, an index "t" is used to denote time, e.g., X_t. A vector of random variables $[X_i, X_{i+1}, \cdots, X_j]^{\mathrm{T}}$ is shortly denoted by X_i^j, while its realization is shortly denoted by x_i^j. The letter H is used to denote the entropy, the letter I denotes mutual information, while the letter $\mathcal{I}$ denotes the mutual information rate.

II. Source/Channel Model

We assume that the source (channel input) is a stationary discrete-time Markov random process X_t whose realizations x_t take values from a finite-size source alphabet $\mathcal{X}$. It is assumed that the channel input process has memory $L \geq 0$, i.e., we have for any integer $m \geq 0$

$$\Pr\left(X_{t+1}|X_{t-L-m}^t\right) = \Pr\left(X_{t+1}|X_{t-L}^t\right). \tag{1}$$

We consider an indecomposable finite-state machine channel [11]. The channel state at time t is denoted by the random variable S_t whose realization is $s_t \in \mathcal{S} = \{1, 2, \ldots, M\}$. We choose the state alphabet size M to be the minimum integer $M > 0$ such that S_t forms a Markov process of memory 1, i.e., for any integer $m \geq 0$

$$\Pr\left(S_{t+1}|S_{t-m}^t\right) = \Pr\left(S_{t+1}|S_t\right). \tag{2}$$

For example, if the channel input X_t is a binary Markov process of memory 3 and the channel is PR4 (i.e., $1-D^2$) of memory 2, then $M = 2^{\max(3,2)} = 8$ guarantees that the state sequence is a Markov process of memory 1.

With this choice of the states, it is apparent that the input sequence X_t and the state sequence S_t uniquely determine each other. Hence, from this point on, the term Markov process will be reserved for the state sequence S_t which is a Markov process of memory 1. The state transition probabilities of the Markov process are denoted by

$$P_{ij} = \Pr\left(S_{t+1} = j|S_t = i\right), \tag{3}$$

This work was supported by the National Science Foundation under Grant No. CCR-9904458 and by the National Storage Industry Consortium.

where $1 \leq i \leq M$ and $1 \leq j \leq M$. Clearly, we must have $\sum_{j=1}^{M} P_{ij} = 1$. A transition from state i to state j is considered to be *invalid* if the Markov state sequence cannot be taken from state i to state j. The transition probability for an invalid transition is thus zero. A *valid* transition is a transition that is not invalid. A trellis section, denoted by $\mathcal{T}$, is the set of all valid transitions, that is, a valid transition (i,j) satisfies $(i,j) \in \mathcal{T}$.

The channel output Y_t is a *hidden* Markov sequence induced by the state sequence S_t, i.e., for a discrete random variable Y_t, the probability mass function of Y_t satisfies

$$\Pr\left(Y_t | S_{-\infty}^{\infty}, Y_{-\infty}^{t-1}, Y_{t+1}^{\infty}\right) = \Pr\left(Y_t | S_{t-1}, S_t\right). \quad (4)$$

If Y_t is a continuous random variable, replace the probability mass functions in (4) by probability density functions.

For indecomposable channels, the choice of the initial state S_0 does not affect the mutual information rate [11], which may then be expressed as

$$\mathcal{I}\left(X_t; Y_t\right) = \mathcal{I}\left(S_t; Y_t\right) = \lim_{n\to\infty} \frac{1}{n} I\left(S_1^n; Y_1^n | S_0\right). \quad (5)$$

Assume the set $\mathcal{T}$ of valid state transitions is given and fixed, but we have the option of choosing the transition probabilities of the valid transitions in (3) to maximize the information rate in (5). That is, we can choose P_{ij} for all $(i,j) \in \mathcal{T}$ such that we achieve the capacity

$$C = \max_{\substack{P_{ij} \\ (i,j)\in\mathcal{T}}} \lim_{n\to\infty} \frac{1}{n} I\left(S_1^n; Y_1^n | S_0\right). \quad (6)$$

We are interested in evaluating the capacity C and the transition probabilities P_{ij} for all $(i,j) \in \mathcal{T}$ that achieve C. In the next section we illustrate the solution approach by considering a simpler problem (the computation of the discrete memoryless channel capacity) for which the solution is the Arimoto-Blahut algorithm [9], [10].

III. A Stochastic Arimoto-Blahut algorithm

We illustrate the main idea behind the new stochastic optimization algorithm by considering the well-known Arimoto-Blahut algorithm [9], [10] and turning it into a *stochastic* Arimoto-Blahut algorithm. Assume that we have a discrete memoryless channel, with an input alphabet $\mathcal{X} = \{1, 2, \ldots, M\}$ and an output alphabet $\mathcal{Y} = \{1, 2, \ldots, N\}$. The channel is defined by the cross-over probabilities $p_{ij} = \Pr\left(Y = j | X = i\right)$. The task is to determine the channel input probabilities $r_i = \Pr\left(X = i\right)$ to maximize the mutual information (i.e., to achieve the capacity) and is solved by the Arimoto-Blahut alternating maximization algorithm [9], [10]. We give here its equivalent expectation-maximization form.

Algorithm 1 The expectation-maximization version of the Arimoto-Blahut Algorithm

Initialization: Pick an arbitrary distribution r_i, such that $0 < r_i < 1$ and $\sum_{i=1}^{M} r_i = 1$.

Repeat until convergence

Step 1 - Expectation: For r_i fixed, compute

$$T_i = \mathrm{E}\left[\frac{\Pr\left(X = i|Y\right) \log \Pr\left(X = i|Y\right)}{r_i}\right].$$

Step 2 - Maximization: For T_i fixed, find r_i to maximize $\sum_i r_i \left[\log \frac{1}{r_i} + T_i\right]$, i.e., set $r_i = \frac{2^{T_i}}{\sum_k 2^{T_k}}$.

end

The expectation-maximization formulation is advantageous when T_i is hard (or impossible) to compute in closed form, because an estimate $\hat{T}_i$ can be easily found by Monte Carlo simulation. We create n channel input realizations according to the probability mass function r_i, and transmit them over the channel to collect n channel output realizations $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$. For n large,

$$\hat{T}_i = \frac{1}{n} \sum_{k=1}^{n} \frac{\Pr\left(X = i | Y = y^{(k)}\right) \log \Pr\left(X = i | Y = y^{(k)}\right)}{r_i}$$

converges with probability 1 to T_i. If we substitute $\hat{T}_i$ for T_i in the maximization step of Algorithm 1, we get a stochastic Arimoto-Blahut algorithm that converges (with probability 1) to the deterministic Arimoto-Blahut algorithm [9], [10] when $n \to \infty$, and hence it converges with probability 1 to the capacity-achieving input distribution.

IV. Markov process optimization

We now revert to the problem posed in Section II. Let $\mu_i = \Pr\left(S_t = i\right)$ and $P_{ij} = \Pr\left(S_t = j | S_{t-1} = i\right)$. By the stationarity requirement, μ_i and P_{ij} must satisfy $\mu_j = \sum_i \mu_i P_{ij}$. Our goal is to find the transition probabilities P_{ij} that achieve the maximization in (6).

Using the chain rule, the Markov property and stationarity, we rewrite the mutual information rate as

$$\begin{aligned} &\frac{1}{n} I\left(S_1^n; Y_1^n | S_0\right) \quad (7)\\ &= \frac{1}{n} \sum_{t=1}^{n} I\left(S_t; Y_1^n | S_0^{t-1}\right) \\ &= \frac{1}{n} \sum_{t=1}^{n} I\left(S_t; Y_1^n | S_{t-1}\right) \\ &= \underbrace{\frac{1}{n} \sum_{t=1}^{n} H\left(S_t | S_{t-1}\right)}_{\sum_{i,j:(i,j)\in\mathcal{T}} \mu_i P_{ij} \log \frac{1}{P_{ij}}} - \frac{1}{n} \sum_{t=1}^{n} H\left(S_t | S_{t-1}, Y_1^n\right), \end{aligned}$$

where $\sum_{i,j:(i,j)\in\mathcal{T}}$ denotes that the sum is taken over those values i, j such that the branch from state i to state j is a member of the trellis stage (i.e, the transition from input symbol i to input symbol j is valid). Further, express the entropy in the second term in (7) as

$$\begin{aligned} & -H\left(S_t|S_{t-1},Y_1^n\right) \qquad (8)\\ &= \mathrm{E}\left[\log \Pr\left(S_t|S_{t-1},Y_1^n\right)\right] \\ &= \mathrm{E}\left[\log \Pr\left(S_{t-1},S_t|Y_1^n\right)\right] - \mathrm{E}\left[\log \Pr\left(S_{t-1}|Y_1^n\right)\right] \\ &= \sum_{i,j:(i,j)\in\mathcal{T}} \mu_i P_{ij} \mathrm{E}_{Y_1^n|i,j}\left[\log \Pr\left(S_{t-1}=i,S_t=j|Y_1^n\right)\right] \\ &\quad -\sum_i \mu_i \mathrm{E}_{Y_1^n|i}\left[\log \Pr\left(S_{t-1}=i|Y_1^n\right)\right] \\ &= \sum_{i,j:(i,j)\in\mathcal{T}} \mu_i P_{ij} \mathrm{E}_{Y_1^n|i,j}\left[\log \Pr\left(S_{t-1}=i,S_t=j|Y_1^n\right)\right] \\ &\quad -\sum_{i,j:(i,j)\in\mathcal{T}} \mu_i P_{ij} \mathrm{E}_{Y_1^n|i}\left[\log \Pr\left(S_{t-1}=i|Y_1^n\right)\right]. \end{aligned}$$

Here, $\mathrm{E}_{Y_1^n|i,j}$ denotes the conditional expectation taken over the variable Y_1^n when the pair (S_{t-1}, S_t) equals (i, j). Similarly, $\mathrm{E}_{Y_1^n|i}$ is the conditional expectation taken over Y_1^n when $S_{t-1} = i$. Define the expectation T_{ij} as

$$\begin{aligned} T_{ij} = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\Big\{ & \mathrm{E}_{Y_1^n|i,j}\left[\log \Pr\left(S_{t-1}=i,S_t=j|Y_1^n\right)\right] \\ & -\mathrm{E}_{Y_1^n|i}\left[\log \Pr\left(S_{t-1}=i|Y_1^n\right)\right]\Big\}. \qquad (9) \end{aligned}$$

Using the Bayes rule, the expectation T_{ij} may be alternatively expressed as

$$T_{ij} = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left[\log\frac{\mathrm{P}_t\left(i,j|Y_1^n\right)^{\frac{\mathrm{P}_t\left(i,j|Y_1^n\right)}{\mu_i P_{ij}}}}{\mathrm{P}_t\left(i|Y_1^n\right)^{\frac{\mathrm{P}_t\left(i|Y_1^n\right)}{\mu_i}}}\right], \qquad (10)$$

where $\mathrm{P}_t\left(i,j|Y_1^n\right)$ is short for $\Pr\left(S_{t-1}=i,S_t=j|Y_1^n\right)$ and $\mathrm{P}_t\left(i|Y_1^n\right)$ is short for $\Pr\left(S_{t-1}=i|Y_1^n\right)$. The expression in (10) is advantageous for numerical evaluations because it does not involve the conditional expectation as does (9). For now, we assume that we have a method for computing T_{ij} (it will be shown in a subsequent section that the value T_{ij} can be accurately estimated using the Arnold-Loeliger sum-product approach [7]). Combining (5), (7), (8) and (9), we may express the mutual information rate as

$$\begin{aligned} \mathcal{I}\left(S_t;Y_t\right) &= \lim_{n\to\infty}\frac{1}{n}I\left(S_1^n;Y_1^n|S_0\right) \\ &= \sum_{i,j:(i,j)\in\mathcal{T}}\mu_i P_{ij}\left[\log\frac{1}{P_{ij}}+T_{ij}\right]. \qquad (11) \end{aligned}$$

We now formulate the expectation-maximization procedure in the spirit of Section III, but for the case that the state sequence is a Markov process.

Algorithm 2 EXPECTATION-MAXIMIZATION FOR OPTIMIZING MARKOV PROCESS TRANSITION PROBABILITIES

Initialization: Pick an arbitrary distribution P_{ij} that satisfies the following two constraints:
1) if $(i,j)\in\mathcal{T}$ then $0 < P_{ij} < 1$, otherwise $P_{ij} = 0$ and
2) for any i, require that $\sum_j P_{ij} = 1$.

Repeat until convergence

Step 1 - Expectation: While keeping all P_{ij} fixed, for $(i,j)\in\mathcal{T}$ compute T_{ij} using (9) or (10).

Step 2 - Maximization: While keeping all T_{ij} fixed, find all P_{ij} (and the corresponding values $\mu_j = \sum_i \mu_i P_{ij}$) to achieve the maximization of (11), i.e., set

$$[P_{ij}] = \arg\max_{[P_{ij}]}\sum_{i,j:(i,j)\in\mathcal{T}}\mu_i P_{ij}\left[\log\frac{1}{P_{ij}}+T_{ij}\right]. \qquad (12)$$

end

There is an obvious similarity between Algorithms 1 and 2. While the convergence of Algorithm 1 to the capacity-achieving distribution can be proved by modifying the proofs in [9], [10], it is not clear whether the same strategy can be used to prove the convergence of Algorithm 2 to the capacity-achieving Markov process transition probabilities. However, numerical evidence (obtained via comparison to time-consuming brute-force optimization methods) seems to suggest that Algorithm 2 does converge to the capacity-achieving transition probabilities.

Assume for now that we have a method to compute T_{ij}. (In Section V, we show that T_{ij} can be accurately estimated using the Arnold-Loeliger sum-product approach [7].) Then the solution to the maximization step (12) is given by the following generalization of Shannon's result for the maximal achievable (noise-free) entropy rate of a Markov process [1]. Form a *noisy adjacency matrix* $\mathbf{A}$ whose elements are defined as

$$A_{ij} = \begin{cases} 2^{T_{ij}} & \text{if } (i,j)\in\mathcal{T} \\ 0 & \text{otherwise} \end{cases}. \qquad (13)$$

Let W_{max} be the maximal real eigenvalue of $\mathbf{A}$, and let $\underline{b} = [b_1, b_2, \ldots, b_M]^{\mathrm{T}}$ be the corresponding eigenvector. Then the the maximization in (12) is achieved by

$$P_{ij} = \frac{b_j}{b_i}\cdot\frac{A_{ij}}{W_{max}}, \qquad (14)$$

and the maximal value of $\sum_{i,j:(i,j)\in\mathcal{T}}\mu_i P_{ij}\left[\log\frac{1}{P_{ij}}+T_{ij}\right]$ is

$$C\left(T_{ij}\right) = \log W_{max}. \qquad (15)$$

A few comments are in order here. If the channel is *noiseless*, then the matrix $\mathbf{A}$ is the standard *noise-free*

adjacency matrix [3]. The logarithm of the maximal real eigenvalue of this noise-free adjacency matrix is the maximal achievable noise-free entropy rate of a Markov process, as was shown by Shannon [1]. If we accept the conjecture that Algorithm 2 converges to the capacity achieving transition probabilities as true, and if T_{ij} is the globally optimal expectation (computed after many iterations of Algorithm 2), then the channel capacity (the maximal mutual information rate) is given by $C = \log W_{max}$. In [12], Khayrallah and Neuhoff relate the capacity of a *soft*-constrained sequences to the eigenvalue of (what we rename here) a *soft* adjacency matrix **B**. It is somewhat speculative (though likely) that the soft adjacency matrix and the noisy adjacency matrix could be unified into a *soft-noisy* adjacency matrix which could be related to the capacity of soft-constrained sequences over noisy channels.

V. Computing the noisy adjacency matrix

The noisy adjacency matrix can be computed using the sum-product algorithm since in (9) and (10) the probabilities $\Pr(S_{t-1} = i, S_t = j|Y_1^n)$ and $\Pr(S_{t-1} = i|Y_1^n)$ are exactly the outputs of the sum-product (BCJR, Baum-Welch) algorithm [13]. Assume that the transition probabilities P_{ij} are given. For n large, generate a realization s_0^n of the state sequence S_0^n according to these transition probabilities P_{ij}. Pass the realization s_0^n of the state sequence through the noisy channel to get a realization y_1^n of the output sequence Y_1^n. Now run the sum-product (BCJR) algorithm [13] and compute the outputs $\mathrm{P}_t(i,j|y_1^n) = \Pr(S_{t-1} = i, S_t = j|y_1^n)$ and $\mathrm{P}_t(i|y_1^n) = \Pr(S_{t-1} = i|y_1^n)$ for all $1 \le t \le n$, all $1 \le i \le M$ and all pairs $(i,j) \in \mathcal{T}$. Next for $(i,j) \in \mathcal{T}$ estimate (10) as the empirical expectation

$$\hat{T}_{ij} = \frac{1}{n}\sum_{t=1}^{n}\left[\log \frac{\mathrm{P}_t(i,j|y_1^n)^{\frac{\mathrm{P}_t(i,j|y_1^n)}{\mu_i P_{ij}}}}{\mathrm{P}_t(i|y_1^n)^{\frac{\mathrm{P}_t(i|y_1^n)}{\mu_i}}}\right]. \qquad (16)$$

By the ergodicity assumption, invoking the law of large numbers, we have (with probability 1) $\lim_{n\to\infty} \hat{T}_{ij} = T_{ij}$. Thus, we have a method for implementing Algorithm 2.

Algorithm 3 A stochastic method for optimizing Markov process transition probabilities

Initialization: Pick an arbitrary distribution P_{ij} that satisfies the following two constraints:
1) if $(i,j) \in \mathcal{T}$ then $0 < P_{ij} < 1$, otherwise $P_{ij} = 0$ and
2) for any i, require that $\sum_j P_{ij} = 1$.

Repeat until convergence

Step 1: For n large, generate s_0^n according to the transition probabilities P_{ij} and pass them through the noisy channel to get y_1^n.

Step 2: Run the forward-backward sum-product (Baum-Welch, BCJR) algorithm [13] and compute $\hat{T}_{ij}$ according to (16).

Step 3: Estimate the noisy adjacency matrix as

$$\hat{A}_{ij} = \begin{cases} 2^{\hat{T}_{ij}} & \text{if } (i,j) \in \mathcal{T} \\ 0 & \text{otherwise} \end{cases},$$

and find its maximal eigenvalue $\hat{W}_{max}$ and the corresponding eigenvector $\underline{\hat{b}} = \left[\hat{b}_1, \hat{b}_2, \cdots, \hat{b}_M\right]^{\mathrm{T}}$.

Step 4: Compute the entries of the new transition probability matrix for $(i,j) \in \mathcal{T}$ as $P_{ij} = \frac{\hat{b}_j}{\hat{b}_i} \cdot \frac{\hat{A}_{ij}}{\hat{W}_{max}}$

end

At the end of the execution of Algorithm 3, the optimal information rate can be evaluated using the Arnold-Loeliger method [7] or using expression (11). Numerical evaluations of Algorithm 3 on low-dimensional Markov processes (up to 4 trellis states) show that Algorithm 2 achieves the same Markov transition probabilities as a brute-force information rate maximization. The same numerical evaluations have shown that at the end of the execution of Algorithm 3, the optimized information rate can be just as accurately evaluated by $\hat{C} = \log \hat{W}_{max}$. This supports the conjecture that Algorithm 2 converges to the capacity-achieving distribution of the Markov transition probabilities and that Algorithm 3 achieves convergence to the same transition probabilities with probability 1.

VI. Lower bounds on the noisy capacity of run-length limited codes

We illustrate the applicability of Algorithm 3 by computing lower bounds on the capacity of run-length limited (RLL) sequences [3] when transmitted over a binary symmetric channel (BSC) [11] (but can easily be generalized for any finite-state machine channel)[1]. The channel's cross-over probability p is allowed to vary between 0 and 0.5. We present the result of the proposed procedure for a run-length limited sequence with parameters $(d,k) = (0,1)$, that is, the source output (channel input) is a sequence of 1s and 0s, where at least $d = 0$ and at most $k = 1$ consecutive 0s can appear in the sequence. We consider two cases: 1) the channel input is created by a 2-state Markov process - Figure 1a, and 2) the channel input is created by a 3-state Markov process - Figure 1b. In Figure 1, the notation P_{ij}/x denotes that the Markov process goes from state i to state j with probability P_{ij}, while producing x as the source output (channel input). Algorithm 3 with trellis length $n = 10^6$ was applied to optimize the transition probabilities P_{ij}, and the optimized information rates are depicted in Figure 2. The same figure

[1] The algorithm presented here can also be used to lower bound the capacity of partial response channels under binary input constraints, see [7] for this particular scenario.

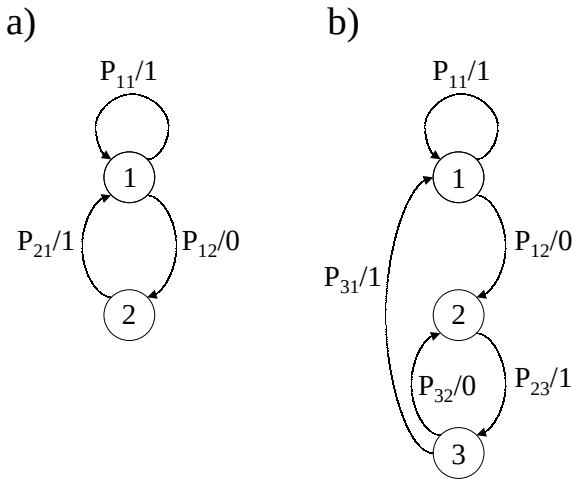

Fig. 1. Two Markov processes creating run-length limited binary sequences with parameters $(d,k) = (0,1)$.

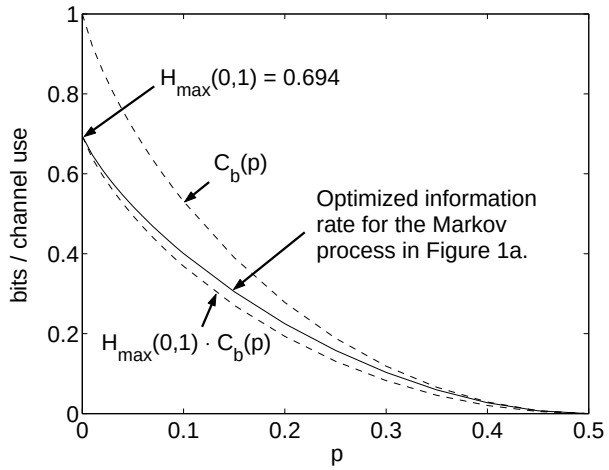

Fig. 2. Information rates as functions of the BSC cross-over probability p. The information rate for the memory 1 process in Figure 1b is not plotted because it is only marginally higher than the optimized information rate for Figure 1a.

shows the unconstrained capacity of the binary symmetric channel $C_b(p) = 1 + p\log p + (1-p)\log(1-p)$ and the maximal entropy rate of a noise-free $(d,k) = (0,1)$ run-length limited sequence $H_{max}(0,1) = \log_2\left(1+\sqrt{5}\right) - 1$. Also shown is the curve $H_{max}(0,1) \cdot C_b(p)$, which is numerically very close to the Zehavi-Wolf lower bound on the noisy capacity of the RLL(0,1) sequence over the binary symmetric channel [2]). The information rates calculated using Algorithm 3 lie above $H_{max}(0,1) \cdot C_b(p)$ and are getting tighter as the Markov memory of the run-length limited process increases. For the $(d,k) = (0,1)$ sequence, going from a Markov source of memory 1 (Figure 1a) to a source of memory 2 (Figure 1b) only marginally increases the information rate. The optimized input Markov process can actually be used to get a tight upper bound, as described in [14] (not shown here).

VII. Conclusion

We presented a stochastic method for optimizing information rates of Markov sequences over finite-state machine channels. No proof is known to show that this achieves the capacity of Markov sources over noisy channels, but results suggest that this is likely the case. At worst, the presented method provides a tight lower bound, which can be used to lower bound the capacities of partial response channels and the noisy capacities of constrained (say, run-length limited) sequences, as well as to guide the construction of tight upper bounds [14].

Acknowledgment

Special thanks to Dieter Arnold, Hans-Andrea Loeliger and Pascal Vontobel for helpful discussions and for an early disclosure of manuscript [7] that inspired this work.

References

[1] C. E. Shannon, "A mathematical theory of communications," *Bell Systems Technical Journal*, vol. 27, pp. 379–423 (part I) and 623–656 (part II), 1948.

[2] E. Zehavi and J. K. Wolf, "On runlength codes," *IEEE Transactions on Information Theory*, vol. 34, pp. 45–54, January 1988.

[3] K. A. S. Immink, P. H. Siegel, and J. K. Wolf, "Codes for digital recorders," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2260–2299, October 1998.

[4] W. Hirt, *Capacity and Information Rates of Discrete-Time Channels with Memory*. PhD thesis, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, 1988.

[5] 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 Transactions on Information Theory*, vol. 37, pp. 1527–1539, 1991.

[6] S. Shamai (Shitz) and R. Laroia, "The intersymbol interference channel: Lower bounds on capacity and channel precoding loss," *IEEE Transactions on Information Theory*, vol. 42, pp. 1388–1404, 1996.

[7] D. Arnold and H.-A. Loeliger, "On the information rate of binary-input channels with memory," in *Proceedings IEEE International Conference on Communications 2001*, (Helsinki, Finland), June 2001.

[8] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rates of finite state ISI channels," in *Proceedings IEEE Global communications Conference 2001*, (San Antonio, Texas), November 2001.

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

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

[11] R. G. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley and Sons, Inc., 1968.

[12] A. S. Khayrallah and D. L. Neuhoff, "Coding for channels with cost constraints," *IEEE Transactions on Information Theory*, vol. 42, pp. 854–867, May 1996.

[13] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. 20, pp. 284–287, Sept. 1974.

[14] P. Vontobel and D. M. Arnold, "An upper bound on the capacity of channels with memory and contraint input," presented at *IEEE Information Theory Workshop*, (Cairns, Australia), September 2001.