# Sum-Product Algorithm and Feedback Capacity

**Shaohua Yang and Aleksandar Kavčić**
**Division of Engineering and Applied Sciences**
**Harvard University**
**33 Oxford Street**
**Cambridge, MA 02138**

**Abstract**

In this paper, we explore the link between the sum-product algorithm and the feedback capacity of a channel with memory. We show that the optimal (i.e., capacity-achieving) feedback is captured by the causal posterior state probabilities. For finite-state machine channels, the optimal feedback is captured by the forward recursion of the sum-product (Baum-Welch, BCJR) algorithm. This result drastically reduces the space over which the optimal feedback-dependent source distribution needs to be sought. Further, the feedback capacity computation may then be formulated as an average-reward-per-stage stochastic control problem, for which numerical solutions of Bellman's equation deliver the feedback-capacity-achieving source distribution. With the knowledge of the capacity-achieving source distribution, the value of the capacity is easily estimated using accurate Markov chain Monte Carlo methods. We demonstrate the applicability of the method by computing the feedback capacity of partial response channels and the feedback capacity of run-length-limited (RLL) sequences over binary symmetric channels (BSCs).

## 1 Introduction

The feedback capacity of a *memoryless* channel equals the feed-forward capacity of the same channel [1]. However, the computation (or characterization) of the feedback capacity of channels *with memory* has long remained an open problem [2]. In 1990, Massey [3] showed that the directed information rate is an upper bound on the feedback capacity. In 2000, Tatikonda [4] proved that any directed information rate is always achievable, and thus proved that the feedback capacity is the supremum of the directed information rate.

In this paper, we develop a numerical procedure to compute the feedback capacity for channels that can be represented as finite-state machines. We first give two theorems that drastically simplify the computation problem. Namely, 1) finite-memory feedback-dependent Markov sources achieve the feedback capacities, and 2) the optimal feedback is computed by the sum-product (Baum-Welch, BCJR) algorithm [5]. The feedback capacity is then evaluated by combining three tools: 1) the sum-product (Baum-Welch, BCJR) algorithm [5], 2) dynamic programming (for solving Belmann's equation [6]), and 3) Monte-Carlo evaluation of the entropy rate of a hidden Markov process [7, 8].

It is interesting to contrast the feedback capacity to the feed-forward capacity. For a *memoryless* channel, the feedback capacity equals the feed-forward capacity, and the capacity is achieved by a

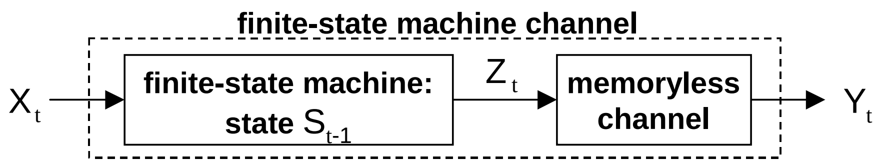


Figure 1: A finite state machine channel (concatenated finite state machine and a memoryless channel).

memoryless source [1]. It may be tempting to conjecture that the feed-forward capacity of a channel whose memory length is 1 symbol, is achieved by a 1-st order Markov process. This is clearly not correct as demonstrated in [7, 8, 9], where higher-order Markov processes are constructed to surpass the information rates of lower-order Markov processes. If, however, we utilize feedback, we show in this paper that a feedback-dependent Markov process *does* achieve the feedback capacity. We get a generalized statement (which does not hold for the feed-forward capacity): *the feedback capacity of a channel whose memory length is $L$, is achieved by a feedback-dependent $L$-th order Markov source.* This generalizes a known fact that a memoryless source (i.e, a 0-th order Markov source) achieves the capacity of a memoryless channel (i.e., a channel whose memory length is 0).

**Notation:** Uppercase letters represent random variables (or vectors), while lowercase letters represent their realizations. An index $t$ next to a random variable (e.g., $X_t$) denotes the random variable at time $t$. A vector of time-dependent variables is denoted as $X_t^\tau = [X_t, X_{t+1}, \ldots, X_{\tau-1}, X_\tau]$. The mutual information is denoted by the letter $I$, while the information *rate* is denoted by the letter $\mathcal{I}$. The entropy is denoted by the letter $H$, while the differential entropy is denoted by the letter $h$.

# 2 Channel Model

We assume that the channel is represented by a finite-state machine observed through a memoryless noisy channel, see Figure 1. The state of the channel at time $t$ is denoted by $S_t$, and the state realization is denoted by $s_t$. The state alphabet $\mathcal{S}$ is finite, $|\mathcal{S}| = M < \infty$, i.e., $s_t \in \mathcal{S} = \{0, 1, \ldots, M-1\}$. The channel input process is denoted by $X_t$, while the input process realization is denoted by $x_t$. The process $X_t$ is drawn from a finite alphabet $\mathcal{X}$, i.e., $x_t \in \mathcal{X}$, and $|\mathcal{X}| < \infty$. The channel satisfies the following assumptions:

1. The state $s_t$ is a function of $s_{t-1}$ and $x_t$.

2. The finite-state machine output $z_t$ is a function of $s_{t-1}$ and $s_t$.

3. If the finite-state machine output $Z_t = z_t$ is known, then the channel output $Y_t$ is stochasti-

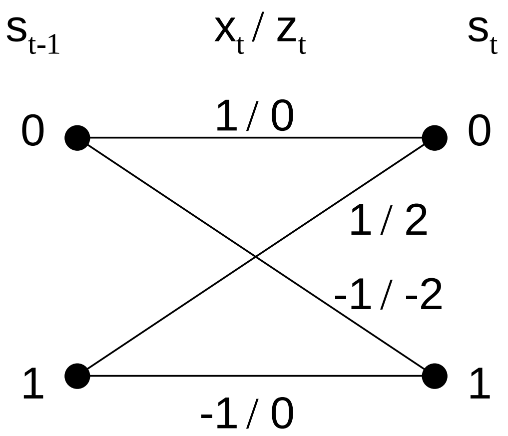


Figure 2: Trellis representation of the dicode partial response channel.

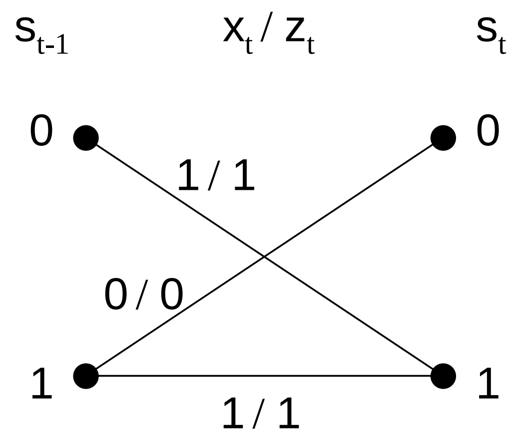


Figure 3: The trellis of the RLL(0,1) constraint.

cally independent of all other random variables, i.e.,

$$f_{Y_t \left| Z_t, S_{-\infty}^t, Y_{-\infty}^{t-1} \right.} \left( y_t \left| z_t, s_{-\infty}^t, y_{-\infty}^{t-1} \right. \right) = f_{Y_t|Z_t} \left( y_t \left| z_t \right. \right). \tag{2.1}$$

We generally assume that the finite-state machine is indecomposable, and that the initial state $S_0 = s_0$ is known. Then, from assumptions 1-3, we have that

**I)** There is a 1-to-1 correspondence between $(S_0, X_1^t)$ and $S_0^t$

$$\left(S_0, X_1^t\right) \stackrel{1:1}{\longleftrightarrow} S_0^t. \tag{2.2}$$

**II)** The conditional probability density function (pdf) of the channel output satisfies

$$f_{Y_t \left| S_{-\infty}^t, Y_{-\infty}^{t-1} \right.} \left( y_t \left| s_{-\infty}^t, y_{-\infty}^{t-1} \right. \right) = f_{Y_t|S_{t-1},S_t} \left( y_t \left| s_{t-1}, s_t \right. \right). \tag{2.3}$$

In a finite-state machine, some state pairs $(s_{t-1}, s_t) = (i, j)$ may not be valid, that is, the machine cannot be taken from state $i$ to state $j$. We denote the set of all *valid* state pairs by $\mathcal{T}$.

**Example 2.1 (The dicode partial response channel).** *Let* $\mathcal{X} = \{-1, 1\}$ *and let*

$$Z_t = X_t - X_{t-1} \tag{2.4}$$
$$Y_t = Z_t + W_t, \tag{2.5}$$

*where* $W_t$ *is white Gaussian noise with variance* $\sigma^2$*, shortly denoted by* $W_t \sim \mathcal{N}(0, \sigma^2)$*. Denote the state by* $S_t = \frac{1 - X_t}{2}$*. So,* $s_t \in \mathcal{S} = \{0, 1\}$*. Clearly,* $Z_t = 2S_{t-1} - 2S_t$*, and*

$$f_{Y_t|S_{t-1},S_t} \left( y_t \left| s_{t-1}, s_t \right. \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{[y_t - (2s_{t-1} - 2s_t)]^2}{2\sigma^2}}. \tag{2.6}$$

*The channel is represented by the trellis in Figure 2. As evident from the figure, the set of valid state pairs is* $\mathcal{T} = \{(0,0),(0,1),(1,0),(1,1)\}$.

**Example 2.2 (RLL(0,1) sequences over binary symmetric channels).** *Let* $X_t$ *be a run-length-limited (RLL) sequence with the constraint RLL(0,1), i.e., no consecutive zeros appear in the sequence* $X_0, X_1, \cdots, X_t, \cdots$. *The memory in the channel is exhibited by the constraint that if* $X_{t-1} = 0$*, then necessarily* $X_t = 1$*. The sequence* $X_t$ *is transmitted over a binary symmetric channel (BSC) with cross-over probability* $p$*. We may define the state as* $S_t = X_t$*, so* $s_t \in \mathcal{S} = \{0,1\}$*. The channel is represented by*

$$\begin{aligned} Z_t &= X_t && (2.7) \\ Y_t &= Z_t \oplus W_t, && (2.8) \end{aligned}$$

*where* $\oplus$ *denotes binary addition, and* $W_t$ *represents a sequence of binary independent and identically distributed (i.i.d.) random variables with* $\Pr(W_t = 1) = p$*. The channel law is*

$$\Pr\left(Y_t = y_t \left| S_{t-1} = s_{t-1}, S_t = s_t\right.\right) = \begin{cases} p & \text{if } y_t = s_t \\ 1-p & \text{otherwise} \end{cases}. \tag{2.9}$$

*The channel is represented by the trellis in Figure 3. As evident from the figure, the set of valid state pairs is* $\mathcal{T} = \{(0,1),(1,0),(1,1)\}$.

# 3 The Feedback Capacity

Tatikonda [4] proved that the feedback capacity of a channel is the supremum of the directed information rate. The directed information rate is defined as

$$\begin{aligned} \mathcal{I}(X \to Y) &= \lim_{n\to\infty} \frac{1}{n} \mathrm{I}\left(X_1^n \to Y_1^n\right) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathrm{I}\left(X_1^t; Y_t \left| Y_1^{t-1}\right.\right). && (3.10) \end{aligned}$$

For the channel model given in Section 2, we know that the sequence $s_0^t$ is in a 1-to-1 relationship with the sequence $(s_0, x_1^t)$. For this reason, the feedback capacity of the finite-state machine channel may be expressed as

$$\begin{aligned} C^{(\mathrm{fb})} &= \sup_{\mathcal{P}} \mathcal{I}(S \to Y) \\ &= \sup_{\mathcal{P}} \lim_{n\to\infty} \frac{1}{n} \mathrm{I}\left(S_1^n \to Y_1^n \right| S_0 = s_0\left.\right). && (3.11) \end{aligned}$$

In (3.11), the supremum is taken over the set $\mathcal{P}$, where the set $\mathcal{P}$ is the collection of all (causal conditional) probability measure functions

$$\mathcal{P} = \left\{\Pr\left(S_1 \left| S_0\right.\right), \Pr\left(S_2 \left| S_0^1, Y_1\right.\right), \cdots, \Pr\left(S_t \left| S_0^{t-1}, Y_1^{t-1}\right.\right), \cdots\right\} \tag{3.12}$$

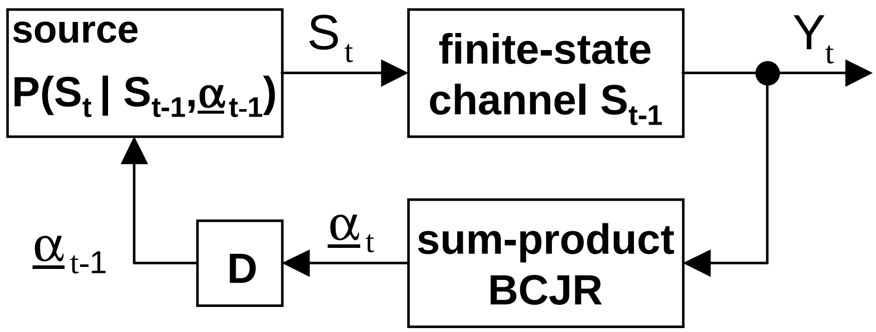


Figure 4: Feedback loop that achieves the feedback capacity.

If we were to apply (3.11) directly, the computation of the feedback capacity would be an enormously complex problem since we would need to specify $\Pr\left(S_t \left| S_0^{t-1}, Y_1^{t-1}\right.\right)$ for every time $t$ and every realization $\left(S_0^t, Y_1^{t-1}\right) = \left(s_0^t, y_1^{t-1}\right)$. As $t$ gets larger, the number of realizations $\left(S_0^t, Y_1^{t-1}\right) = \left(s_0^t, y_1^{t-1}\right)$ becomes unbounded, and the problem would become intractable. The next two theorems help us to reduce the problem to a manageable complexity.

**Theorem 3.1.** *For a finite-state channel, the feedback capacity is achieved by a feedback-dependent Markov process determined by the following collection of state transition probabilities*

$$\mathcal{P}^{\text{Markov}} = \left\{\Pr\left(S_1 \left| S_0\right.\right), \Pr\left(S_2 \left| S_1, Y_1\right.\right), \cdots, \Pr\left(S_t \left| S_{t-1}, Y_1^{t-1}\right.\right), \cdots\right\} \tag{3.13}$$

**Theorem 3.2.** *For a finite-state channel, after observing the channel output sequence $Y_1^t = y_1^t$, the optimal feedback is the vector of causal posterior state probabilities*

$$\underline{\alpha}_t = \left[\alpha_t(0), \alpha_t(1), \cdots, \alpha_t(M-1)\right]^{\mathrm{T}}, \tag{3.14}$$

*where*

$$\alpha_t(i) = \Pr\left(S_t = i \left| Y_1^t = y_1^t\right.\right). \tag{3.15}$$

(Note that the values $\alpha_t(i)$ can be recursively computed using the sum-product (Baum-Welch, BCJR) algorithm [5].)

Combining Theorems 3.1 and 3.2, we conclude that the feedback capacity can be formulated as

$$C^{(\mathrm{fb})} = \sup_{\mathcal{P}_{\alpha}^{(\text{Markov})}} \mathcal{I}\left(S \to Y\right), \tag{3.16}$$

where the supremum is taken over the set

$$\mathcal{P}_{\alpha}^{(\text{Markov})} = \left\{\Pr\left(S_1 \left| S_0, \underline{\alpha}_0\right.\right), \Pr\left(S_2 \left| S_1, \underline{\alpha}_1\right.\right), \cdots, \Pr\left(S_t \left| S_{t-1}, \underline{\alpha}_{t-1}\right.\right), \cdots\right\}. \tag{3.17}$$

We note that all elements of the set $\mathcal{P}_{\alpha}^{(\text{Markov})}$ are time-invariant. We may therefore write

$$C^{(\mathrm{fb})} = \sup_{\Pr\left(S_t \left| S_{t-1}, \underline{\alpha}_{t-1}\right.\right)} \mathcal{I}\left(S \to Y\right). \tag{3.18}$$

The block diagram of a system that achieves the capacity is given in Figure 4.

# 4 Stochastic Control Formulation

We denote by $\{P_{ij}(\underline{a})\}$ the set of all probabilities $P_{ij}(\underline{a}) = \Pr\left(S_t = j \left| S_{t-1} = i, \underline{\alpha}_{t-1} = \underline{a}\right.\right)$ for all pairs $(i,j) \in \mathcal{T}$, and all possible realizations $\underline{\alpha}_{t-1} = \underline{a}$. If the set $\{P_{ij}(\underline{a})\}$ is known (and fixed), given the knowledge of the vector $\underline{\alpha}_{t-1}$ and the channel output realization $y_t$, the vector $\underline{\alpha}_t$ is a function of $\{P_{ij}(\underline{a})\}$, $\underline{\alpha}_{t-1}$ and $y_t$, i.e.,

$$\underline{\alpha}_t = F_{\mathrm{BCJR}}\left(\underline{\alpha}_{t-1}, \{P_{ij}(\underline{a})\}, y_t\right). \tag{4.19}$$

If expanded, equation (4.19) takes the form

$$\begin{bmatrix} \alpha_t(0) \\ \alpha_t(1) \\ \vdots \\ \alpha_t(M-1) \end{bmatrix} = \begin{bmatrix} F^{(0)}_{\mathrm{BCJR}}\left(\underline{\alpha}_{t-1}, \{P_{ij}(\underline{a})\}, y_t\right) \\ F^{(1)}_{\mathrm{BCJR}}\left(\underline{\alpha}_{t-1}, \{P_{ij}(\underline{a})\}, y_t\right) \\ \vdots \\ F^{(M-1)}_{\mathrm{BCJR}}\left(\underline{\alpha}_{t-1}, \{P_{ij}(\underline{a})\}, y_t\right) \end{bmatrix}, \tag{4.20}$$

where from [5] we have

$$\alpha_t(\ell) = F^{(\ell)}_{\mathrm{BCJR}}\left(\underline{\alpha}_{t-1}, \{P_{ij}(\underline{a})\}, y_t\right) = \frac{\sum\limits_{i:(i,\ell)\in\mathcal{T}} \alpha_{t-1}(i) P_{i\ell}\left(\underline{\alpha}_{t-1}\right) f_{Y_t|S_{t-1},S_t}\left(y_t \left| i, \ell\right.\right)}{\sum\limits_{i,j:(i,j)\in\mathcal{T}} \alpha_{t-1}(i) P_{ij}\left(\underline{\alpha}_{t-1}\right) f_{Y_t|S_{t-1},S_t}\left(y_t \left| i, j\right.\right)}. \tag{4.21}$$

Since $y_t$ is a realization of a random variable $Y_t$, we can regard $\underline{\alpha}_t$ as a realization of a random vector $\underline{A}_t$. Using (4.19), the random vector $\underline{A}_t$ is recursively described as

$$\underline{A}_t = F_{\mathrm{BCJR}}\left(\underline{A}_{t-1}, \{P_{ij}(\underline{a})\}, Y_t\right). \tag{4.22}$$

The process $\underline{A}_t$ is a Markov process since

$$f_{\underline{A}_t|\underline{A}_1^{t-1}}\left(\underline{\alpha}_t \left| \underline{\alpha}_1^{t-1}\right.\right) = f_{\underline{A}_t|\underline{A}_{t-1}}\left(\underline{\alpha}_t \left| \underline{\alpha}_{t-1}\right.\right). \tag{4.23}$$

The Markov process $\underline{A}_t$ is fully determined by the set $\{P_{ij}(\underline{a})\}$ and by the channel law (conditional pdf) $f_{Y_t|S_{t-1},S_t}\left(y_t \left| i, j\right.\right)$. Thereby, the channel law is fixed, but we have the freedom of choosing the values in the set $\{P_{ij}(\underline{a})\}$. Our task is to find the optimal set $\{P_{ij}(\underline{a})\}$ that maximizes the directed information rate, and evaluate this maximized directed information rate, i.e., we seek to find

$$\left\{P^*_{ij}(\underline{a})\right\} = \arg \max_{\{P_{ij}(\underline{a})\}} \mathcal{I}(S \to Y) \tag{4.24}$$

$$C^{(\mathrm{fb})} = \left.\mathcal{I}(S \to Y)\right|_{\{P_{ij}(\underline{a})\} = \left\{P^*_{ij}(\underline{a})\right\}}. \tag{4.25}$$

Problem (4.24) can be formulated as a stochastic control problem as follows. We rewrite the directed information rate $\mathcal{I}(S \to Y)$ as

$$\begin{aligned}
\mathcal{I}(S \to Y) &= \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathrm{I}\left(S_1^t; Y_t \left| S_0, Y_1^{t-1}\right.\right) && (4.26)\\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left[\mathrm{h}\left(Y_t \left| S_0, Y_1^{t-1}\right.\right) - \mathrm{h}\left(Y_t \left| S_0^t, Y_1^{t-1}\right.\right)\right] && (4.27)\\
&= \lim_{n\to\infty} \frac{1}{n} \left[\sum_{t=1}^{n} \mathrm{h}\left(Y_t \left| S_0, Y_1^{t-1}\right.\right)\right] - \mathrm{h}\left(Y_t \left| S_{t-1}, S_t\right.\right) && (4.28)\\
&= \lim_{n\to\infty} \frac{1}{n} \mathrm{E}\left[\sum_{t=1}^{n} - \log f_{Y_t|S_0,\underline{A}_{t-1}}\left(Y_t \left| S_0, \underline{A}_{t-1}\right.\right)\right] - \mathrm{h}\left(Y_t \left| S_{t-1}, S_t\right.\right), && (4.29)
\end{aligned}$$

where (4.26) and (4.27) are definitions of the directed information rate, (4.28) follows from the channel property (2.3), and (4.29) is shown by Theorem 3.2. From (4.29), problem (4.24) is equivalent to the following optimization problem

$$\left\{P_{ij}^*(\underline{a})\right\} = \arg \max_{\{P_{ij}(\underline{a})\}} \lim_{n\to\infty} \frac{1}{n} \mathrm{E}\left[\sum_{t=1}^{n} - \log f_{Y_t|\underline{A}_{t-1}}\left(Y_t \left| \underline{A}_{t-1}\right.\right)\right]. \tag{4.30}$$

This optimization problem can be regarded as an average-reward-per-stage stochastic control problem [6]. For this problem, at each stage (or time) $t$, the *state* is the vector $\underline{\alpha}_t$, which evolves according to (4.19), the *control* is the set of Markov transition probabilities $\{P_{ij}(\underline{a})\}$ which is assumed to be stationary, and the *reward* is taken to be $-\log f_{Y_t|\underline{A}_{t-1}}\left(y_t \left| \underline{\alpha}_{t-1}\right.\right)$.

Let $\lambda^*$ be the *maximum average reward*, and let $h^*(\underline{a})$ be the *optimal reward-to-go* (or return) function. Bellman's equation for this stochastic control problem (4.30) is the following [6]

$$\lambda^* + h^*(\underline{a}) = \max_{\{P_{ij}(\underline{a})\}} \mathrm{E}\left\{- \log f_{Y_t|\underline{A}_{t-1}}\left(Y_t \left| \underline{a}\right.\right) + h^*(F_{\mathrm{BCJR}}\left(\underline{a}, \{P_{ij}(\underline{a})\}, Y_t\right))\right\}. \tag{4.31}$$

Under the assumption that the state process $\underline{A}_t$ forms a steady state recurrent class, there exists at least one optimal stationary control $\left\{P_{ij}^*(\underline{a})\right\}$ which satisfies Bellman's equation (4.31). Further, there exist efficient dynamic programming algorithms, e.g., value iteration and policy iteration [6], which solve Bellman's equation (4.31) and thus find an optimal feedback-dependent source distribution $\left\{P_{ij}^*(\underline{a})\right\}$ which achieves the feedback capacity.

Once the optimal source distribution $\left\{P_{ij}^*(\underline{a})\right\}$ for problem (4.24) is found, we can use the Monte Carlo method in [7, 8] to compute the feedback capacity.

# 5 Numerical Results

We get numerical solutions to (4.31) by applying the dynamic programming algorithm [6]. Since the state $\underline{\alpha}_t$ and the control $\{P_{ij}(\underline{a})\}$ are real-valued vectors, we quantize them to get a finite state space and a finite control space. We use the value iteration algorithm to get the optimal (actually, optimal to within the quantization accuracy) Markov channel input transition probabilities

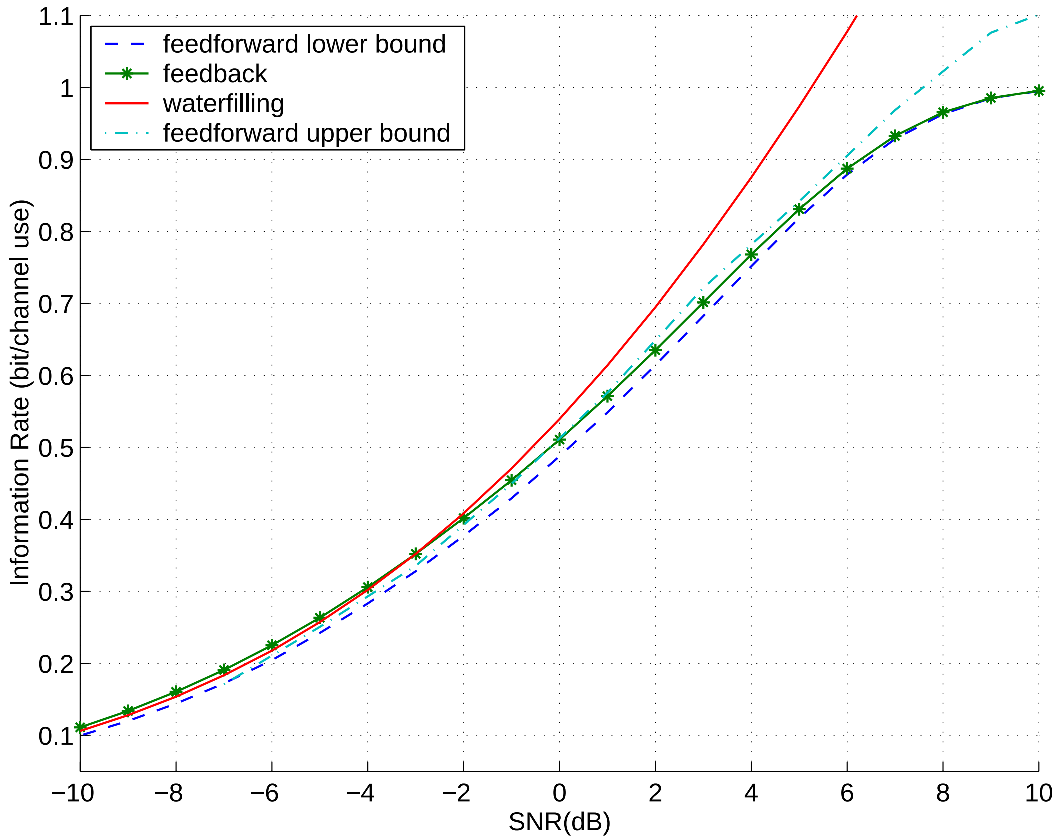


Figure 5: Capacity bounds for the dicode $(1-D)$ channel. At low SNRs, the feedback capacity surpasses the waterfilling channel capacity for continuous-valued channel inputs. At high SNRs, the feedback capacity is just above the feed-forward capacity estimate [9].

$\{P_{ij}(\underline{a})\}$. Then, we compute the optimized directed information rate $I(S \to Y)$ using the Monte Carlo method in [7, 8]. Strictly speaking, the directed information rate computed in this way is a lower bound on the feedback capacity, which converges to the feedback capacity only if we have an infinite number of quantization levels for $\underline{\alpha}_t$ and $\{P_{ij}(\underline{a})\}$.

Figure 5 shows the capacity bounds of the dicode partial response channel in Example 2.1. The feedback capacity is compared to the waterfilling capacity [1, 10], the tightest known feed-forward capacity lower bound computed by the EM (expectation-maximization) algorithm in [9], and the feed-forward capacity upper bound shown in [11]. At low SNRs, the feedback capacity surpasses the waterfilling capacity, which numerically verifies that feedback increases the capacity of channels with memory. At high SNRs, the feedback capacity is very close to the feed-forward capacity lower bound computed by the iterative algorithm given in [9].

In Figure 6, we show the capacity bounds for the RLL(0,1) sequences over BSCs in Example 2.2. The feedback capacity surpasses the tightest known feed-forward capacity lower bound computed by the iterative algorithm given in [9]. However, the difference is not significant (at most $10\%$ as seen in Figure 6. For comparison, shown in Figure 6 is the BSC capacity $C = 1 - H(p)$, where

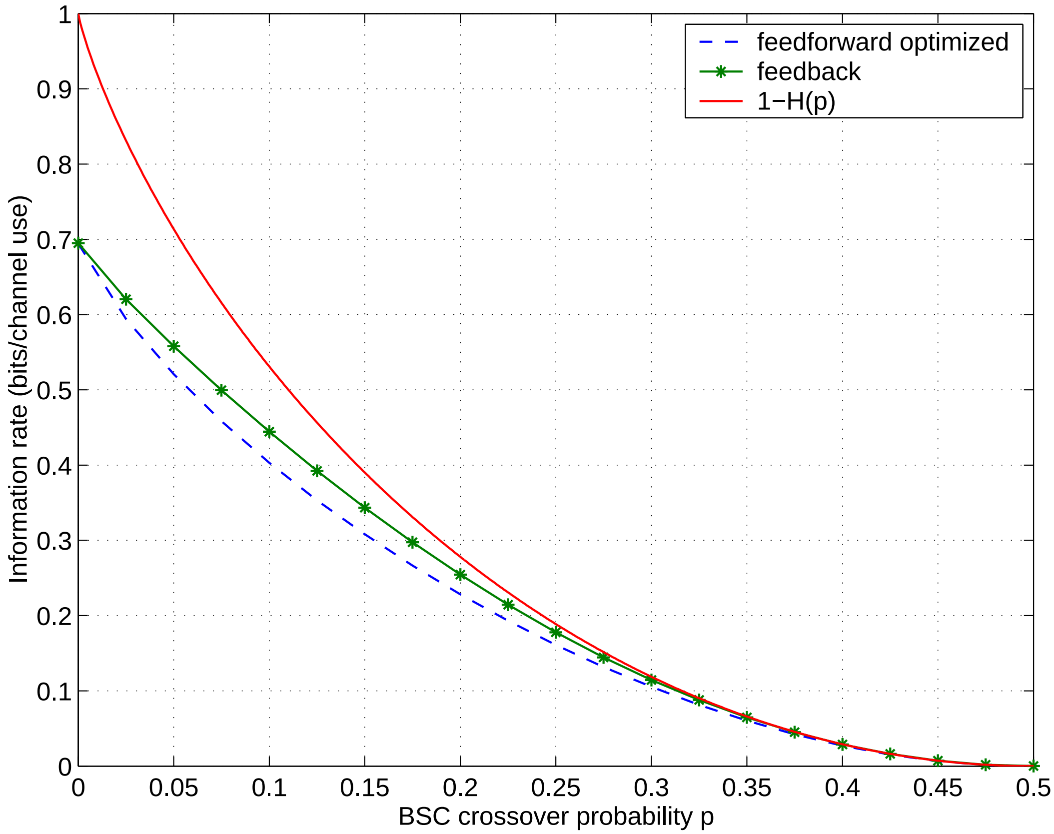


Figure 6: Capacity bounds for the $RLL(0,1)$ code over BSC channel. For this channel the feedback capacity is much higher than the tightest known feed-forward capacity lower bound [9].

$H(p)$ is the binary entropy function [1].

# 6 Conclusion

The feedback capacity of a finite-state channel is achieved by a feedback-dependent Markov channel input. The memory length of the feedback-capacity-achieving Markov channel input equals the channel memory length. The optimized transition probabilities are only dependent on the forward $\alpha$-coefficients computed using the Baum-Welch (BCJR) algorithm [5]. That is, the entire past of the observations is captured by the $\alpha$-coefficients. We formulated the feedback-dependent channel input distribution optimization problem as an average-reward-per-stage stochastic control problem and applied a dynamic programming algorithm to solve it numerically. Using the proposed methods, we computed the feedback capacities of partial response (PR) channels and BSC channels with run-length-limited (RLL) input constraints. Closed form computations of the optimal feedback-dependent Markov transition probabilities and the feedback capacity are still open problems.

**Acknowledgement**: Research supported in part by NSF Grant 9904458 and by the National Storage Industry Consortium.

# References

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

[2] S. Verdú, “Fifty years of Shannon theory,” *IEEE. Trans. of Inform. Theory*, vol. 44, pp. 2057–2077, October 1998.

[3] J. L. Massey, “Causality, feedback and directed information,” *Proc. 1990 Intl. Symp. on Info. Th. and its Applications*, pp. 27–30, Nov 1990.

[4] S. C. Tatikonda, *Control Under Communications Constraints*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, September 2000. available to dwnload at http://www.cs.berkeley.edu/˜tatikond/thesis.html.

[5] 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.

[6] D. P. Bertsekas, *Dynamic Programming and Optimal Control (Vol. I, 2nd Edition)*. Belmont, MA: Athena Scientific, 2001.

[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), pp. 2692–2695, 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] A. Kavčić, “On the capacity of Markov sources over noisy channels,” in *Proceedings IEEE Global Communications Conference 2001*, (San Antonio, Texas), November 2001. available at http://hrl.harvard.edu/˜kavcic/publications.html.

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

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