a matrix satisfying the hypothesis of Lemma 1. Hence, Lemma 1 implies that $\mathsf{u}_i \neq \mathsf{v}_i$. Similarly, by considering $\mathsf{M}'_j(q, \mathsf{y}(q))$, we have $\mathsf{u}_j \neq \mathsf{v}_j$. This proves $d_H(\mathsf{u}, \mathsf{v}) = 4$. □

This shows that $A_q(q, 4, 3) = \binom{q}{3}$ for all $q \geq 3$. Theorem 4 now follows.

V. Conclusion

In this correspondence, we complete the determination of $A_q(n, 4, 3)$ by employing large sets with holes to construct optimal $(n, 4, 3)_q$-codes for $n \equiv 4$ or $5 \bmod 6$, $n \geq q - 1$, and by using a new technique based on special sequences to construct optimal $(q, 4, 3)_q$-codes. The results of this correspondence combine with those in [1] to give:

Main Theorem: $A_q(n, 4, 3) = \min\{U_q(n), \binom{n}{3}\}$ for all n and q.

Acknowledgment

The authors are grateful to Lingbo Yu for her help in locating literature. The authors are also grateful to the anonymous reviewer for comments that helped improve the presentation of this correspondence.

References

[1] Y. M. Chee and S. Ling, "Constructions for q-ary constant-weight codes," *IEEE Trans. Inf. Theory*, vol. 53, pp. 135–146, 2007.

[2] C. Ding, S. Ling, H. Wang, and J. Yuan, "Bounds on nonbinary constant weight codes," 2006, unpublished.

[3] G. Ge and D. Wu, "Some new optimal quaternary constant weight codes," *Sci. China Ser. F*, vol. 48, no. 2, pp. 192–200, 2005.

[4] D. Wu and G. Ge, "Generalized Steiner systems $\mathrm{Gs}_4(2, 4, v, 4)$," *J. Combin. Math. Combin. Comput.*, vol. 45, pp. 183–193, 2003.

[5] G. Ge and D. Wu, "4-$*$GDDs(3^n) and generalized Steiner systems $\mathrm{Gs}(2, 4, v, 3)$," *J. Combin. Des.*, vol. 11, no. 6, pp. 381–393, 2003.

[6] G. Ge and D. Wu, "Generalized Steiner triple systems with group size ten," *J. Math. Res. Expo.*, vol. 23, no. 3, pp. 391–396, 2003.

[7] P. R. J. Östergård and M. Svanström, "Ternary constant weight codes," *Electron. J. Combin.*, vol. 9, no. 1, 2002.

[8] G. Ge, "Further results on the existence of generalized Steiner triple systems with group size $g \equiv 1, 5 \bmod 6$," *Australas. J. Combin.*, vol. 25, pp. 19–27, 2002.

[9] G. Ge, "Generalized Steiner triple systems with group size $g \equiv 0, 3 \bmod 6$," *Acta Math. Appl. Sin. Engl. Ser.*, vol. 18, no. 4, pp. 561–568, 2002.

[10] D. S. Krotov, "Inductive constructions of perfect ternary constant-weight codes with distance 3," *Problemy Peredachi Informatsii*, vol. 37, no. 1, pp. 3–11, 2001.

[11] F.-W. Fu, T. Kløve, Y. Luo, and V. K. Wei, "On the Svanström bound for ternary constant-weight codes," *IEEE Trans. Inf. Theory*, vol. 47, pp. 2061–2064, 2001.

[12] D. Wu, G. Ge, and L. Zhu, "Generalized Steiner triple systems with group size $g = 7, 8$," *Ars Combin.*, vol. 57, pp. 175–191, 2000.

[13] G. Ge, "Generalized Steiner triple systems with group size $g \equiv 1, 5 \bmod 6$," *Australas. J. Combin.*, vol. 21, pp. 37–47, 2000.

[14] G. Bogdanova, "New bounds for the maximum size of ternary constant weight codes," *Serdica Math. J.*, vol. 26, no. 1, pp. 5–12, 2000.

[15] J. Yin, Y. Lu, and J. Wang, "Maximum distance holey packings and related codes," *Sci. China Ser. A*, vol. 42, no. 12, pp. 1262–1269, 1999.

[16] M. Svanström, "A class of perfect ternary constant-weight codes," *Des. Codes Cryptogr.*, vol. 18, no. 1-3, pp. 223–229, 1999.

[17] M. Svanström, "Ternary Codes With Weight Constraints," Ph.D., Linköpings Universitet, , Sweden, 1999.

[18] K. Phelps and C. Yin, "Generalized Steiner systems with block size three and group size four," *Ars. Combin.*, vol. 53, pp. 133–146, 1999.

[19] J. v. Lint and L. Tolhuizen, "On perfect ternary constant weight codes," *Des. Codes Cryptogr.*, vol. 18, no. 1-3, pp. 231–234, 1999.

[20] K. Chen, G. Ge, and L. Zhu, "Generalized Steiner triple systems with group size five," *J. Combin. Des.*, vol. 7, no. 6, pp. 441–452, 1999.

[21] S. Blake-Wilson and K. T. Phelps, "Constant weight codes and group divisible designs," *Des. Codes Cryptogr.*, vol. 16, no. 1, pp. 11–27, 1999.

[22] F.-W. Fu, A. J. H. Vinck, and S.-Y. Shen, "On the constructions of constant-weight codes," *IEEE Trans. Inf. Theory*, vol. 44, pp. 328–333, 1998.

[23] M. Svanström, "A lower bound for ternary constant weight codes," *IEEE Trans. Inf. Theory*, vol. 43, pp. 1630–1632, 1997.

[24] K. Phelps and C. Yin, "Generalized Steiner systems with block size three and group size $g \equiv 3 \bmod 6$," *J. Combin. Des.*, vol. 5, no. 6, pp. 417–432, 1997.

[25] T. Etzion, "Optimal constant weight codes over Z_k and generalized designs," *Discrete Math.*, vol. 169, no. 1-3, pp. 55–82, 1997.

[26] L. Teirlinck, "Large sets with holes," *J. Combin. Des.*, vol. 1, no. 1, pp. 69–94, 1993.

[27] D. L. Kreher and D. R. Stinson, *Combinatorial Algorithms: Generation, Enumeration and Search*. Boca Raton, FL: CRC, 1999.

Markov Processes Asymptotically Achieve the Capacity of Finite-State Intersymbol Interference Channels

Jiangxin Chen and Paul H. Siegel, *Fellow, IEEE*

***Abstract*—Recent progress in capacity evaluation has made it possible to compute a sequence of lower bounds on the capacity of a finite-state intersymbol-interference (ISI) channel by finding a sequence of optimized Markov input processes with increasing order r, for which the state of the process is the previous r input symbols. In this correspondence, we prove that, as the order r goes to infinity, the sequence of optimized Markov sources asymptotically achieves the capacity of the channel. The conclusion is extended to two-dimensional finite-state ISI channels, the binary-symmetric channel (BSC) with constrained inputs, and general indecomposable finite-state channels with a mild constraint.**

***Index Terms*—Capacity, finite-state channels, intersymbol interference (ISI) channels, Markov processes, run-length limited constraints, two-dimensional channels.**

I. Introduction

Magnetic recording channels are generally modeled as finite-state, linear intersymbol-interference (ISI) channels with additive Gaussian noise and a binary input constraint. While the capacity of a general Gaussian linear ISI channel can be evaluated with the water-filling formula [1], a formula for the capacity when the input is constrained to a finite alphabet remains unknown.

Manuscript received December 16, 2004; revised March 3, 2007. This work was supported in part by the National Science Foundation under Grant CCF-0514859 and by the Center for Magnetic Recording Research. The material in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Chicago, IL, June/July2004.

J. Chen was with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0407 USA. He is now with the Prediction Company, UBS, Santa Fe, NM, 87505 USA (e-mail: simple_address@yahoo.com).

P. H. Siegel is with the Center for Magnetic Recording Research, University of California, San Diego, La Jolla, CA, 92093-0401 USA (e-mail: psiegel@ucsd.edu).

Communicated by Y. Steinberg, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2007.915709

Recently, several groups independently proposed a numerical technique to estimate information rates of such finite-state ISI channels [2]–[5]. The method requires the generation of a long channel output realization and the application of the forward recursion of the Bahl–Cocke–Jelinek–Raviv (BCJR) algorithm [6].

Lower bounds on the channel capacity have been computed by using this technique to estimate the information rates $I^{(r)}$ of optimized, order-r Markovian input processes (see, e.g., [4], [7], [8]), where the Markov state of the source consists of the past r source symbols. As the order r increases, the optimized information rates $I^{(r)}$ form a nondecreasing sequence of capacity lower bounds. This approach has been further utilized to design channel codes with the property that the probability distribution of the codewords approximates that of the optimized, order-r Markov source. Such codes have been shown to approach the mutual information rate $I^{(r)}$ (see, e.g., [9]–[11]).

A subtle, but important question is: Does the sequence $I^{(r)}$ converge to the true capacity, or does it remain bounded away from the capacity for some or all finite-state ISI channels? The answer will not only determine whether the capacity lower bound $I^{(r)}$ is asymptotically tight, but also verify that the code design methodologies based upon the optimized Markov source distribution produce good codes. While empirical evidence suggests that the sequence converges to the capacity for certain simple recording channel models [13], [14], a rigorous analysis has been lacking.

We remark that Blackwell *et al.* [12] proved that a periodic Markovian input of period n which generates consecutive, optimized block codewords of length n can achieve the capacity of a one-dimensional finite-state indecomposable channel as the order (and the period) of the Markov process goes to infinity. However, the state of such a Markov process is not observable, whereas we are considering the class of Markov processes for which the state is explicitly determined by the last r symbols.

We can reformulate the question above as follows: For a finite-state ISI channel as described above, can a sequence of finite-order Markov sources, whose states are determined by the last r symbols, be used to approximate arbitrarily closely a capacity-achieving, discrete input process, in the sense that the information rates achieved by the Markovian inputs asymptotically achieve the channel capacity?

In this correspondence, we give an affirmative answer to this question. We then extend the result to another input-constrained, finite-state channel model, namely, the one-dimensional binary-symmetric channel (BSC) with a (d,k)-run-length limited (RLL) input. Information rates for this channel have been addressed in [8], [15]. Next, we discuss the proof of an analogous result for two-dimensional Gaussian linear finite-state ISI channels.[1] Finally, we demonstrate that this is a general result applicable to all finite-state indecomposable channels.

The correspondence is organized as follows. In Section II, we give the proof of the convergence result for one-dimensional ISI channels. We then discuss the relatively straightforward extension to the RLL-constrained BSC channel. In Section III, we consider the issue of Markovian approximations for capacity-achieving processes for two-dimensional Gaussian linear finite-state ISI channels and state a corresponding convergence result. The extension to general finite-state indecomposable channels is given in Section IV. Concluding remarks are made in Section V.

[1]The one-dimensional BSC with a (d,k) RLL-constrained input is also a recording channel model for traditional magnetic storage devices. The two-dimensional finite-state ISI channel is used to model newly developed, page-oriented storage devices such as holographic memory [21]. In the two-dimensional channel model, the input, output, and channel impulse response are all two-dimensional arrays (see Section III for more detailed definition of two-dimensional finite-state ISI channels). In general, we denote a channel to be D-dimensional if the channel input, output, and the channel impulse response are all D-dimensional arrays.

II. Convergence for One-Dimensional Finite-State ISI Channels

A. Linear Gaussian ISI Channels

We first consider the following one-dimensional finite-state ISI channel model

$$y_k = \sum_{l=0}^{m-1} h_l x_{k-l} + n_k$$

where x_k is the discrete-time finite-alphabet channel input, y_k is the channel output, n_k is the additive white Gaussian noise with zero mean, variance σ^2, and h_k is the channel impulse response. The memory length of this ISI channel is $m-1$. Without loss of generality, we assume the channel is causal.

The capacity of such a channel model is generally defined as

$$C = \lim_{n\to\infty} \sup_{p(X_1^n)} \frac{1}{n} I(X_1^n; Y_1^n). \tag{1}$$

As pointed out in [16], this definition requires that the input maximizing the mutual information behave ergodically. Indeed, Feinstein [17] proved that for a one-dimensional finite-memory channel with discrete input and discrete or continuous output, the capacity C can be achieved by a stationary ergodic input process. Since the one-dimensional finite-state ISI channel is a special case of the one-dimensional finite-memory channel, the same conclusion holds. Therefore, we will restrict our attention to stationary and ergodic input processes. Note that even though the definition of capacity in (1) only refers to the portion of the input and output processes for positive indices $k > 0$, we assume that all processes considered in the correspondence are bi-infinite.

The mutual information rate achieved by a stationary input process $\mathcal{X}$, with corresponding output process $\mathcal{Y}$, is defined as

$$\begin{aligned} I(\mathcal{X},\mathcal{Y}) &= \lim_{n\to\infty} \frac{1}{n} I(X_1^n; Y_1^n) \\ &= \lim_{n\to\infty} \frac{1}{n} H(X_1^n) + \lim_{n\to\infty} \frac{1}{n} H(Y_1^n) \\ &\quad - \lim_{n\to\infty} \frac{1}{n} H(X_1^n, Y_1^n) \\ &= H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y}). \end{aligned}$$

The capacity is therefore the supremum of the mutual information rates over all the stationary ergodic input processes.

To prove the convergence result, we need to answer two questions.

1) Given a capacity-achieving input process $\mathcal{X}$ and a Markov order r, how do we find a good approximating Markov input process $\tilde{\mathcal{X}}$ whose state is its past r input symbols?
2) Given such a Markov input $\tilde{\mathcal{X}}$, how do we evaluate the difference between the mutual information rate achieved by the capacity-achieving input process $\mathcal{X}$ and that achieved by $\tilde{\mathcal{X}}$?

Therefore, we can summarize the proof procedure as follows. To answer the first question, we will use a systematic procedure to construct a Markov approximation, given a stationary input process $\mathcal{X}$ and a Markov order r. Although the resulting order-r Markov approximation may not achieve the maximum mutual information rate for the given Markov order r, we show that the mutual information rates of the approximating processes approach the channel capacity as the Markov order r goes to infinity.

To answer the second question, we first establish some equalities between the entropies related to the desired process $\mathcal{X}$ and those related to its approximation $\tilde{\mathcal{X}}$. More specifically, for an order-r ($r \geq m-1$) Markov approximation, we will establish

$$\begin{aligned} H\left(X_1^{r+1}\right) &= H\left(\tilde{X}_1^{r+1}\right) \\ H\left(X_1^{r-m+2} \middle| Y_1^{r-m+2}\right) &= H\left(\tilde{X}_1^{r-m+2} \middle| \tilde{Y}_1^{r-m+2}\right). \end{aligned} \tag{2}$$

Then we will show that by choosing a sufficiently large Markov order r during the approximation, the mutual information rate induced by $\tilde{\mathcal{X}}$ is lower-bounded by $I(\mathcal{X};\mathcal{Y}) - \delta$ for any prespecified $\delta > 0$.

We now consider how to approximate a discrete stationary ergodic input process by a Markov chain of order r whose Markov state is the past r input symbols. Denote a cylinder set of length $r+1$ in a discrete, stationary ergodic process $\mathcal{X}$ with finite alphabet by $X_1^{r+1} = [X_1, \ldots, X_{r+1}]$. Using the same approximation approach as in [19], we define an order-r Markov chain $\tilde{\mathcal{X}}$ as follows. Set the transition probabilities to

$$\begin{aligned}\mathsf{Pr}\left(\tilde{X}_{r+k+1} = x_{r+1} \left| \tilde{X}_{1+k}^{r+k} = x_1^r \right.\right) &= \mathsf{Pr}\left(\tilde{X}_{r+1} = x_{r+1} \left| \tilde{X}_1^r = x_1^r \right.\right) \\ &= \frac{\mathsf{Pr}\left(X_1^{r+1} = x_1^{r+1}\right)}{\mathsf{Pr}\left(X_1^r = x_1^r\right)}\end{aligned}$$

for all k, and the initial probability distribution to

$$\mathsf{Pr}\left(\tilde{X}_1^r = x_1^r\right) = \mathsf{Pr}\left(X_1^r = x_1^r\right)$$

for every set X_1^r whose probability $\mathsf{Pr}\left(X_1^r = x_1^r\right) > 0$. Thus, the set of states of this Markov process is $S_{\tilde{\mathcal{X}}} = \{x_1^r : \mathsf{Pr}(X_1^r = x_1^r) > 0\}$. It can be shown that this defines a stationary, ergodic Markov process whose finite distribution satisfies $\mathsf{Pr}(\tilde{X}_d^{d+j-1} = x_1^j) = \mathsf{Pr}\left(X_1^j = x_1^j\right)$ whenever $j \leqslant r+1$. This is due to the fact that, for every state in $S_{\tilde{\mathcal{X}}}$

$$\begin{aligned}&\mathsf{Pr}\left(\tilde{X}_1^{r+1} = x_1^{r+1}\right) \\ &\quad = \mathsf{Pr}\left(\tilde{X}_1^r = x_1^r\right) \mathsf{Pr}\left(\tilde{X}_{r+1} = x_{r+1} \left| \tilde{X}_1^r = x_1^r \right.\right) \\ &\quad = \mathsf{Pr}\left(X_1^r = x_1^r\right) \frac{\mathsf{Pr}\left(X_1^{r+1} = x_1^{r+1}\right)}{\mathsf{Pr}\left(X_1^r = x_1^r\right)} \\ &\quad = \mathsf{Pr}\left(X_1^{r+1} = x_1^{r+1}\right). \end{aligned} \tag{3}$$

When the channel input is a stationary process, the channel output is also stationary. Let the channel output processes induced by the input processes $\tilde{\mathcal{X}}$ and $\mathcal{X}$ be denoted by $\tilde{\mathcal{Y}}$ and $\mathcal{Y}$, respectively. It is straightforward to show that the joint probability density function (pdf) $f\left(\tilde{X}_1^{r+1} = x_1^{r+1}, \tilde{Y}_m^{r+1} = y_m^{r+1}\right)$, for $r \geq m-1$, is also equal to $f\left(X_1^{r+1} = x_1^{r+1}, Y_m^{r+1} = y_m^{r+1}\right)$ when (3) holds. Specifically

$$\begin{aligned}&f\left(\tilde{X}_1^{r+1} = x_1^{r+1}, \tilde{Y}_m^{r+1} = y_m^{r+1}\right) \\ &\quad = \mathsf{Pr}\left(\tilde{X}_1^{r+1} = x_1^{r+1}\right) f\left(\tilde{Y}_m^{r+1} = y_m^{r+1} \left| \tilde{X}_1^{r+1} = x_1^{r+1} \right.\right) \\ &\quad = \frac{\mathsf{Pr}\left(X_1^{r+1} = x_1^{r+1}\right)}{(2\pi\sigma^2)^{\frac{r-m+2}{2}}} \exp\left\{ -\frac{\sum_{l=m}^{r+1}\left(y_l - \sum_{k=0}^{m-1} h_k x_{l-k}\right)^2}{2\sigma^2} \right\} \\ &\quad = f\left(X_1^{r+1} = x_1^{r+1}, Y_m^{r+1} = y_m^{r+1}\right). \end{aligned} \tag{4}$$

From (4), we conclude that

$$\begin{aligned}H\left(X_1^{r+1}\right) &= H\left(\tilde{X}_1^{r+1}\right) \\ H\left(X_m^{r+1} \left| Y_m^{r+1} \right.\right) &= H\left(\tilde{X}_m^{r+1} \left| \tilde{Y}_m^{r+1} \right.\right).\end{aligned}$$

Due to the stationarity of both the input and the output processes, we can also have

$$H\left(X_1^{r-m+2} \left| Y_1^{r-m+2} \right.\right) = H\left(\tilde{X}_1^{r-m+2} \left| \tilde{Y}_1^{r-m+2} \right.\right).$$

if we look at the input/output segments $\left(X_{-m+2}^{r-m+2}, Y_1^{r-m+2}\right)$ and $\left(\tilde{X}_{-m+2}^{r-m+2}, \tilde{Y}_1^{r-m+2}\right)$ instead.

We now state and prove the main result of this section.

Theorem 1: Let C denote the capacity of a finite-state ISI channel with finite input alphabet and additive Gaussian noise. Then, for any $\epsilon > 0$, there exists a finite-order Markov process $\tilde{\mathcal{X}}$ whose state is the previous r symbols, with corresponding output process $\tilde{\mathcal{Y}}$, such that

$$I(\tilde{\mathcal{X}};\tilde{\mathcal{Y}}) > C - \epsilon.$$

Proof: From the definition of the capacity and Feinstein's result, we know that for any $\epsilon > 0$, there exists a stationary ergodic input process $\mathcal{X}$ such that the mutual information rate satisfies $I(\mathcal{X};\mathcal{Y}) > C - (\epsilon/2)$. Since both input $\mathcal{X}$ and output $\mathcal{Y}$ are stationary processes, we also have that for any $\delta > 0$, there exists $N > 0$ such that when $n > N$

$$\left| H(\mathcal{X}|\mathcal{Y}) - \frac{1}{n} H\left(X_1^n | Y_1^n\right) \right| < \delta. \tag{5}$$

If we approximate the stationary input process $\mathcal{X}$ with an order-r Markov process $\tilde{\mathcal{X}}$, with $r \geqslant m-1$, we get

$$\begin{aligned}&p\left(X_{-m+2}^{r-m+2} = x_{-m+2}^{r-m+2}, Y_1^{r-m+2} = y_1^{r-m+2}\right) \\ &\qquad = p\left(\tilde{X}_{-m+2}^{r-m+2} = x_{-m+2}^{r-m+2}, \tilde{Y}_1^{r-m+2} = y_1^{r-m+2}\right).\end{aligned}$$

The stationarity of the input/output processes and the above equation lead to the following relationships:

$$\begin{aligned}H\left(\tilde{\mathcal{X}}\right) &= H\left(\tilde{X}_{r+1} \left| \tilde{X}_1^r \right.\right) = H\left(X_{r+1} | X_1^r\right) \\ &\geqslant H\left(X_{r+1} | X_{-\infty}^r\right) = H(\mathcal{X}),\end{aligned} \tag{6}$$

and

$$H\left(\tilde{X}_1^n \left| \tilde{Y}_1^n \right.\right) = H\left(X_1^n | Y_1^n\right) \tag{7}$$

where $1 \leqslant n \leqslant r-m+2$.

Using the chain rule, the stationarity of the processes, and the fact that conditioning reduces entropy, we get the following:

$$\begin{aligned}H\left(\tilde{X}_1^{sn} \left| \tilde{Y}_1^{sn} \right.\right) &= \sum_{i=1}^{s} H\left(\tilde{X}_{(i-1)n+1}^{in} \left| \tilde{Y}_1^{sn}, \tilde{X}_1^{(i-1)n} \right.\right) \\ &\leqslant \sum_{i=1}^{s} H\left(\tilde{X}_{(i-1)n+1}^{in} \left| \tilde{Y}_{(i-1)n+1}^{in} \right.\right) \\ &= sH\left(\tilde{X}_1^n \left| \tilde{Y}_1^n \right.\right).\end{aligned}$$

As $s \to \infty$

$$\begin{aligned}H\left(\tilde{\mathcal{X}} \left| \tilde{\mathcal{Y}} \right.\right) &= \lim_{s\to\infty} \frac{1}{sn} H\left(\tilde{X}_1^{sn} \left| \tilde{Y}_1^{sn} \right.\right) \\ &\leqslant \frac{1}{n} H\left(\tilde{X}_1^n \left| \tilde{Y}_1^n \right.\right).\end{aligned} \tag{8}$$

Choosing $r \geqslant N+m-1$, $n = r-m+2$, and combining (5), (6), (7), (8), we can bound $I\left(\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}\right)$ from below by

$$\begin{aligned}I\left(\tilde{\mathcal{X}};\tilde{\mathcal{Y}}\right) &= H\left(\tilde{\mathcal{X}}\right) - H\left(\tilde{\mathcal{X}} \left| \tilde{\mathcal{Y}} \right.\right) \\ &\geqslant H(\mathcal{X}) - \frac{1}{n} H\left(\tilde{X}_1^n \left| \tilde{Y}_1^n \right.\right) \\ &= H(\mathcal{X}) - \frac{1}{n} H\left(X_1^n | Y_1^n\right) \\ &> H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}) - \delta \\ &= I(\mathcal{X};\mathcal{Y}) - \delta.\end{aligned}$$

Setting $\delta = \epsilon/2$, we see that

$$I(\tilde{\mathcal{X}};\tilde{\mathcal{Y}}) > I(\mathcal{X};\mathcal{Y}) - \epsilon/2 > C - \epsilon.$$

It follows that the mutual information rates of the approximating Markov processes constructed in the proof converge to the channel capacity as their order goes to infinity.

B. BSC With RLL-(d,k) Inputs

RLL (d,k)-constrained binary sequences have seen wide use in data storage devices. The parameters d and k represent the minimum number and maximum number, respectively, of 0's separating two consecutive 1's. The capacity of an RLL-(d,k), input-constrained BSC is defined as

$$C = \sup_{\mathcal{X}} I(\mathcal{X},\mathcal{Y})$$

where the supremum is over all stationary input processes which satisfy the RLL-(d,k) constraint. We have the following analogue of Theorem 1.

Theorem 2: Let C denote the capacity of an RLL-(d,k), input-constrained BSC. Then, for any $\epsilon > 0$, there exists a finite-order, RLL-(d,k)-constrained Markov process $\tilde{\mathcal{X}}$, with corresponding output process $\tilde{\mathcal{Y}}$, such that

$$I(\tilde{\mathcal{X}};\tilde{\mathcal{Y}}) > C - \epsilon.$$

The proof of Theorem 2 follows from reasoning similar to that used in the proof of Theorem 1. The only significant difference is that we need to ensure that the approximating Markov processes generate sequences that satisfy the RLL-(d,k) constraint. This will be the case if the finite-dimensional distribution of $\tilde{\mathcal{X}}$ satisfies $\mathsf{Pr}\left(\tilde{X}_l^{l+k} = x_l^{l+k}\right) = 0$ for sequences x_l^{l+k} that violate the constraint. If $k = \infty$, we substitute d for k in this condition.

If we choose the Markov order r large enough to satisfy $r > \max\{N + m - 1, k\}$, where N has the same meaning as in Section II-A, then we can define an approximating Markov process with a finite-dimensional distribution such that

$$\mathsf{Pr}(\tilde{X}_l^{l+r} = x_l^{l+r}) = \mathsf{Pr}(X_l^{l+r} = x_l^{l+r})$$

for all length-$(r+1)$ words. (If $k = \infty$, then the corresponding condition is $r > \max\{N + m - 1, d\}$.) This condition guarantees that all patterns forbidden by the RLL-(d,k) constraint will have probability zero, since the process $\mathcal{X}$ obeys that constraint. Moreover, if N is chosen to be sufficiently large, the approximating Markov process can achieve a mutual information rate $I(\tilde{\mathcal{X}};\tilde{\mathcal{Y}}) > C - \epsilon$, for any prespecified $\epsilon > 0$.

We remark that Kavčić [8] described a sequence of lower bounds on the capacity of the BSC with stationary RLL-(d,k)-constrained inputs. These were obtained by applying an iterative algorithm that is conjectured to produce a constrained Markov input process of specified order that maximizes the mutual information. If the conjecture is true, then our result shows that this sequence of lower bounds converges to the constrained channel capacity.

III. Convergence for Two-Dimensional Finite-State ISI Channels

In this section, we extend the convergence result to two-dimensional finite-state ISI channels

$$y_{k_1,k_2} = \sum_{l_1=0}^{m_1-1}\sum_{l_2=0}^{m_2-1} h_{l_1,l_2} x_{k_1-l_1,k_2-l_2} + n_{k_1,k_2} \tag{9}$$

where x_{k_1,k_2} is the discrete-time finite-alphabet channel input, y_{k_1,k_2} is the channel output, n_{k_1,k_2} is the additive white Gaussian noise with zero mean, variance σ^2, and h_{k_1,k_2} is the channel impulse response. Again, we assume the channel is causal. This channel model is applicable to page-oriented storage technologies [21] or to certain image processing applications.

Equation (9) defines a two-dimensional finite-state ISI model with a very regular structure. That is, the local relationships of the channel input, the interference, the noise, and the channel output are identical throughout the two-dimensional space. The regular channel structure makes it possible to extend Shannon's coding theorem to this two-dimensional setting.

More specifically, let $X_{i,j}^{k,l}$ be a rectangular array of $\mathcal{X}$ whose upper left corner position is (i,j) and lower right corner position is (k,l) $(k \geqslant i, l \geqslant j)$. A code is defined as a collection of two-dimensional arrays of size $u \times v$ with probability distribution $p\left(X_{1,1}^{u,v}\right)$. As the dimensions of the arrays go to infinity, the achievable coding rate is

$$\lim_{u,v\to\infty} \frac{1}{uv} I\left(X_{1,1}^{u,v}; Y_{1,1}^{u,v}\right).$$

The capacity of this two-dimensional channel is defined as

$$C = \lim_{u,v\to\infty} \sup_{p\left(X_{1,1}^{u,v}\right)} \frac{1}{uv} I\left(X_{1,1}^{u,v}; Y_{1,1}^{u,v}\right)$$

which is a direct extension of the capacity definition for one-dimensional ISI channels in (1). However, in contrast to the one-dimensional case, it is not known, to the best of our knowledge, whether the capacity of a two-dimensional ISI channel is achieved by a stationary ergodic input process. Therefore, we will focus on the *stationary capacity*, meaning the supremum of the mutual information rates achievable with stationary input processes, defined by

$$\begin{aligned} C_s &= \sup_{\mathcal{X}} I(\mathcal{X},\mathcal{Y}) \\ &= \sup_{\mathcal{X}} [H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y})] \end{aligned}$$

where $\mathcal{X}$ is the stationary two-dimensional input process, and $H(\cdot)$ is the entropy rate of the corresponding two-dimensional stationary process. (For a stationary two-dimensional process $\mathcal{W}$, the subadditivity of $H\left(W_{1,1}^{u,v}\right)$ in each dimension ensures that the entropy rate exists [22].)

We still denote the desired two-dimensional stationary discrete input process by $\mathcal{X}$, the approximating Markov input by $\tilde{\mathcal{X}}$, and the corresponding outputs by $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$, respectively. Using the notation in [23], we define the "past" of element $X_{i,j}$ in the row direction as

$$\mathrm{Past}_R\{X_{i,j}\} = \{X_{u,v} : u < i\} \cup \{X_{u,v} : u = i, v < j\}. \tag{10}$$

This definition applies to arrays of finite size as well as arrays of infinite size. For arrays of infinite size, $\mathrm{Past}_R\{X_{i,j}\}$ consists of infinite elements. We denote by $\mathrm{Past}_{R,\underline{l}}\{X_{i,j}\}$ a finite region in the past of $X_{i,j}$ in the row direction

$$\begin{aligned} \mathrm{Past}_{R,\underline{l}}\{X_{i,j}\} = \{X_{u,v} : i - l_1 \leqslant u < i, j - l_2 \leqslant v \leqslant j + l_3\} \\ \cup \{X_{u,v} : u = i, j - l_2 \leqslant v < j\} \end{aligned}$$

where $\underline{l} = [l_1, l_2, l_3]$ and $l_i \geqslant 0$, $i = 1,2,3$. The r-close neighborhood of $X_{i,j}$ is defined as[2]

$$L_R^{(r)}(X_{i,j}) = \mathrm{Past}_{R,[r,r,0]}\{X_{i,j}\}.$$

[2]This is only one of many ways of defining a two-dimensional neighborhood. Many other neighborhood definitions can be used to prove the results in this section.

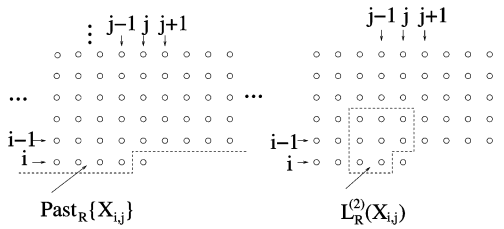

Fig. 1. The regions $\mathrm{Past}_R\{X_{i,j}\}$ and $L_R^{(2)}\{X_{i,j}\}$.

The definitions above are illustrated in Fig. 1. A two-dimensional Markov chain of order r in the row direction[3] is a two-dimensional process $\tilde{\mathcal{X}}$ that satisfies

$$\begin{aligned}\mathsf{Pr}\left(\tilde{X}_{i,j}=x_{i,j}\left|\mathrm{Past}_{R,\underline{\infty}}\left(\tilde{X}_{i,j}\right)=\mathrm{Past}_{R,\underline{\infty}}\left(x_{i,j}\right)\right.\right)\\ =\mathsf{Pr}\left(\tilde{X}_{i,j}=x_{i,j}\left|L_R^{(r)}(\tilde{X}_{i,j})=L_R^{(r)}(x_{i,j})\right.\right),\quad\forall\,(i,j)\end{aligned}$$

where $\underline{\infty}=[\infty,\infty,\infty]$.

Given a two-dimensional stationary process $\mathcal{X}$, we can approximate it with a Markov chain $\tilde{\mathcal{X}}$ of order r by specifying the transition probabilities as shown in the equation at the bottom of the page, and assigning the boundary (initial) probability distribution

$$\begin{aligned}&\mathsf{Pr}\left(\tilde{X}^{0,\infty}_{-r+1,-r+1}=x^{0,\infty}_{-r+1,-r+1}\right)\\&\quad=\mathsf{Pr}\left(X^{0,\infty}_{-r+1,-r+1}=x^{0,\infty}_{-r+1,-r+1}\right)\\&\mathsf{Pr}\left(\tilde{X}^{\infty,0}_{-r+1,-r+1}=x^{\infty,0}_{-r+1,-r+1}\right)\\&\quad=\mathsf{Pr}\left(X^{\infty,0}_{-r+1,-r+1}=x^{\infty,0}_{-r+1,-r+1}\right).\end{aligned}\tag{11}$$

It is easy to see that $\tilde{\mathcal{X}}$ is a two-dimensional stationary process with the property that

$$\begin{aligned}&\mathsf{Pr}\left(\tilde{X}_{i,j}=x_{i,j},L_R^{(r)}(\tilde{X}_{i,j})=L_R^{(r)}(x_{i,j})\right)\\&\quad=\mathsf{Pr}\left(X_{i,j}=x_{i,j},L_R^{(r)}(X_{i,j})=L_R^{(r)}(x_{i,j})\right),\quad\forall\,i>0,\ j>0.\end{aligned}\tag{12}$$

A stationary two-dimensional input also produces a stationary channel output, and (12) leads to an equality similar to (4) for the joint probability density of the finite-dimensional input/output array, which further leads to equalities similar to (2).

To extend the one-dimensional proof to two-dimensional ISI channels, we first invoke a theorem on the entropy rate of d-dimensional stationary processes on the Z^d lattice due to Katznelson and Weiss [24]. Anastassiou and Sakrison [25] obtained a similar result for stationary two-dimensional processes on Z^2. For simplicity, we state this result in the context of the Z^2 lattice.

Theorem 3: For a stationary two-dimensional random process $\mathcal{Y}$ on Z^2, the entropy rate $H(\mathcal{Y})$ satisfies the following equality:

$$H(\mathcal{Y})=H\left(Y_{i,j}\left|\mathrm{Past}_{R,\underline{\infty}}\{Y_{i,j}\}\right.\right),$$

meaning that for any given $\epsilon>0$, there exist $M_i>0$, $i=1,\ldots,3$, such that when $l_i>M_i$ for every $1\leqslant i\leqslant 3$

$$\left|H(\mathcal{Y})-H\left(Y_{i,j}\left|\mathrm{Past}_{R,\underline{l}}\{Y_{i,j}\}\right.\right)\right|<\epsilon.$$

With Theorem 3, we can extend the proof for Theorem 1 and obtain the following convergence result for two-dimensional finite-state ISI channels.

Theorem 4: Let C_s denote the stationary capacity of a two-dimensional finite-state linear ISI channel with finite input alphabet and additive Gaussian noise. Then, for any $\epsilon>0$, there exists a finite-order two-dimensional Markov process in the row direction $\tilde{\mathcal{X}}$, with corresponding output process $\tilde{\mathcal{Y}}$, such that

$$I(\tilde{\mathcal{X}};\tilde{\mathcal{Y}})>C_s-\epsilon.$$

Schetch of Proof: We will only highlight the differences from the proof of Theorem 1.

Similar to the one-dimensional case, for the stationary processes $\mathcal{X}$ and $\mathcal{Y}$, we have that for any $\delta>0$, there exists $N_1>0$ and $N_2>0$ such that when $u>N_1$, $v>N_2$

$$\left|H\left(\mathcal{X}\,|\mathcal{Y}\right)-\frac{1}{uv}H\left(X_{1,1}^{u,v}\left|Y_{1,1}^{u,v}\right.\right)\right|<\delta.\tag{13}$$

We approximate the desired two-dimensional input $\mathcal{X}$ by a Markov chain $\tilde{\mathcal{X}}$ of order r. The approximation ensures that

$$H\left(\tilde{X}_{1,1}^{u,v}\left|\tilde{Y}_{1,1}^{u,v}\right.\right)=H\left(X_{1,1}^{u,v}\left|Y_{1,1}^{u,v}\right.\right)\tag{14}$$

where $1\leqslant u\leqslant r-m_1+2$, and $1\leqslant v\leqslant r-m_2+2$.

The two-dimensional Markov property and Theorem 3 lead to the following inequality:

$$\begin{aligned}H\left(\tilde{\mathcal{X}}\right)&=H\left(\tilde{X}_{u,v}\left|L_R^{(r)}\left(\tilde{X}_{u,v}\right)\right.\right)\\&=H\left(X_{u,v}\left|L_R^{(r)}\left(X_{u,v}\right)\right.\right)\\&\geqslant H\left(X_{u,v}\left|\mathrm{Past}_{R,\underline{\infty}}\{X_{u,v}\}\right.\right)=H\left(\mathcal{X}\right)\end{aligned}\tag{15}$$

Now we consider the channel input array of size $n_1u\times n_2v$, $\tilde{X}_{1,1}^{n_1u,n_2v}$, and divide it into disjoint $u\times v$ subarrays of the form $\tilde{X}_{(i-1)u+1,(j-1)v+1}^{iu,jv}$, denoted by $\tilde{X}_{i,j}^{B}$. We apply the same procedure

[3]In the rest of this section, we will for simplicity refer to this as a Markov chain of order r.

$$\begin{aligned}&\mathsf{Pr}\left(\tilde{X}_{r+k+1,r+l+1}=x_{r+1,r+1}|L_R^{(r)}(\tilde{X}_{r+k+1,r+l+1})=L_R^{(r)}(x_{r+1,r+1})\right)\\&\quad=\mathsf{Pr}\left(\tilde{X}_{r+1,r+1}=x_{r+1,r+1}|L_R^{(r)}(\tilde{X}_{r+1,r+1})=L_R^{(r)}(x_{r+1,r+1})\right)\\&\quad=\frac{\mathsf{Pr}\left(X_{r+1,r+1}=x_{r+1,r+1},L_R^{(r)}(X_{r+1,r+1})=L_R^{(r)}(x_{r+1,r+1})\right)}{\mathsf{Pr}\left(L_R^{(r)}(X_{r+1,r+1})=L_R^{(r)}(x_{r+1,r+1})\right)},\qquad\forall\,k,l\end{aligned}$$

to the $n_1u \times n_2v$ channel output array $\tilde{Y}_{1,1}^{n_1u,n_2v}$. Thus, we have the decompositions

$$\tilde{X}_{1,1}^{n_1u,n_2v} = \left\{\tilde{X}_{1,1}^B, \tilde{X}_{1,2}^B, \ldots, \tilde{X}_{1,n_2}^B, \tilde{X}_{2,1}^B, \ldots, \tilde{X}_{n_1,n_2}^B\right\}$$

and

$$\tilde{Y}_{1,1}^{n_1u,n_2v} = \left\{\tilde{Y}_{1,1}^B, \tilde{Y}_{1,2}^B, \ldots, \tilde{Y}_{1,n_2}^B, \tilde{Y}_{2,1}^B, \ldots, \tilde{Y}_{n_1,n_2}^B\right\}.$$

If we treat the regions $\tilde{X}_{i,j}^B$ and $\tilde{Y}_{i,j}^B$ as elements of the corresponding two-dimensional arrays, we can order them according to the row direction within their respective arrays and define the "past" as in (10). Then we can apply the chain rule based on this two-dimensional ordering as follows:

$$\begin{aligned} H\left(\tilde{X}_{1,1}^{n_1u,n_2v}\middle|\tilde{Y}_{1,1}^{n_1u,n_2v}\right) &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} H\left(\tilde{X}_{i,j}^B\middle|\tilde{Y}_{1,1}^{n_1u,n_2v}, \mathrm{Past}_R\left\{\tilde{X}_{i,j}^B\right\}\right) \\ &\leqslant \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} H\left(\tilde{X}_{i,j}^B\middle|\tilde{Y}_{i,j}^B\right) = n_1n_2H\left(\tilde{X}_{1,1}^B\middle|\tilde{Y}_{1,1}^B\right). \end{aligned}$$

As $n_1, n_2 \to \infty$

$$\begin{aligned} H\left(\tilde{\mathcal{X}}\middle|\tilde{\mathcal{Y}}\right) &= \lim_{n_1,n_2\to\infty} \frac{1}{n_1n_2uv} H\left(\tilde{X}_{1,1}^{n_1u,n_2v}\middle|\tilde{Y}_{1,1}^{n_1u,n_2v}\right) \\ &\leqslant \frac{1}{uv} H\left(\tilde{X}_{1,1}^B\middle|\tilde{Y}_{1,1}^B\right) \\ &= \frac{1}{uv} H\left(\tilde{X}_{1,1}^{u,v}\middle|\tilde{Y}_{1,1}^{u,v}\right). \end{aligned}$$

The rest follows the same reasoning as the proof of Theorem 1.

IV. Extension to General Indecomposable Finite-State Channels

We now extend our previous results to the more general indecomposable finite-state channels with a mild constraint that the entropy of the channel output is finite.

In the one-dimensional scenario, let X_k be the channel input at time k, Y_k be the corresponding channel output, and Z_k be the channel state at time k, respectively. (While X_k and Z_k are discrete with finite alphabets, Y_k can be either discrete or continuous.) A one-dimensional finite-state channel can be defined as

$$\begin{aligned} Z_k &= g_1\left(Z_{k-1}, X_k\right) \\ Y_k &= w_1\left(Z_{k-1}, X_k\right) \end{aligned}$$

where $g_1(\cdot)$ is characterized by the transition probability

$$\mathsf{Pr}\left(Z_k = z_k \middle| Z_{k-1} = z_{k-1}, X_k = x_k\right) = q\left(z_k \middle| z_{k-1}, x_k\right)$$

and $w_1(\cdot)$ is described by the conditional probability distribution

$$\mathsf{Pr}\left(Y_k = y_k \middle| Z_{k-1} = z_{k-1}, X_k = x_k\right) = p\left(y_k \middle| z_{k-1}, x_k\right).$$

(If Y_k is continuous, the left-hand side of the equation is replaced by its corresponding pdf.) Unlike the finite-state ISI channel model which is also a finite-memory channel, the channel state Z_k in this general finite-state channel model may not be completely determined by a finite number of past channel inputs X_{k-l}^k. Thus, (7) may not hold. However, the indecomposability of the channel provides the following property [26]: For an arbitrary $\delta > 0$, there exists $M > 0$ such that when $m > M$

$$\left|q\left(z_m \middle| x_1^m, z_0\right) - q\left(z_m \middle| x_1^m, \tilde{z}_0\right)\right| < \delta \tag{16}$$

for all $z_0(\tilde{z}_0)$, z_m, and x_1^m. This property allows us to bound the difference of the two quantities in (7) to be arbitrarily small by choosing a Markov order large enough. Thus, the generalization of the convergence theorem, Theorem 1, holds for this class of channels. More details can be found in Appendix A.

Similarly, a two-dimensional finite-state channel model is defined as

$$\begin{aligned} Z_{k,l} &= g_2\left(Z_{k-1,l}, Z_{k,l-1}, X_{k,l}\right) \\ Y_{k,l} &= w_2\left(Z_{k-1,l}, Z_{k,l-1}, X_{k,l}\right). \end{aligned}$$

For an arbitrary $\delta > 0$, if there exist $M > 0$ and $N > 0$ such that when $m > M$ and $n > N$

$$\begin{aligned} &\left|q\left(z_{m+1,n}^{m+u,n}, z_{m,n+1}^{m,n+v}\middle|z_{0,1}^{0,n+v}, z_{1,0}^{m+u,0}, x_{1,1}^{m+u,n}, x_{1,n+1}^{m,n+v}\right)\right. \\ &\quad \left. -q\left(z_{m+1,n}^{m+u,n}, z_{m,n+1}^{m,n+v}\middle|\tilde{z}_{0,1}^{0,n+v}, \tilde{z}_{1,0}^{m+u,0}, x_{1,1}^{m+u,n}, x_{1,n+1}^{m,n+v}\right)\right| < \delta \end{aligned}$$

for $\forall\ u > 0$, $v > 0$, we call it a two-dimensional indecomposable channel. Following a similar procedure, we can show that Theorem 4 also holds for such indecomposable two-dimensional finite-state channels.

V. Concluding Remarks

We have proved that an optimized Markov process whose state consists of the previous finite input symbols can approach the capacity of a finite-state ISI channel as the order of the process goes to infinity. This conclusion holds for both one-dimensional and two-dimensional channels. We then extend the result to the BSC channel with RLL input constraints and to the entire class of indecomposable finite-state channel models whose output has a finite entropy. These results confirm that simulation-based techniques for computing information rates of finite-state channels with Markovian inputs can in principle be used to estimate channel capacity to any desired degree of accuracy. They also provide justification for the design of channel codes based upon the optimized input distributions that emerge from the information rate calculations.

It is also worth noting that, in contrast to the one-dimensional case, there are many variations on the definition of a two-dimensional Markov chain, and results analogous to Theorem 4 hold for other classes of two-dimensional Markov processes, as well. We also remark that, when numerically optimizing the two-dimensional input Markov process to achieve the maximum mutual information, we need to consider both the transition probabilities and the stationary distribution in the search. This differs from the one-dimensional setting, where we can exploit the "forgetting" property of aperiodic, irreducible Markov chains—the stationary state distribution is achieved regardless of the initial state distribution—and simply optimize over the transition probabilities (see [28]).

Appendix
Extension to Indecomposable Finite-State Channels

In this appendix, we demonstrate how to extend our main result to indecomposable finite-state channel models whose output Y_k has a finite entropy (i.e., $|H\left(Y_k\right)| < \infty$). We will mainly focus on the one-dimensional case.

Given a stationary ergodic input process $\mathcal{X}$ such that the mutual information rate satisfies $I\left(\mathcal{X}, \mathcal{Y}\right) > C - \epsilon/2$, we can approximate it with an order-$(m + n - 1)$ Markov process $\tilde{\mathcal{X}}$. Thus, we have

$$\mathsf{Pr}\left(X_1^{m+n} = x_1^{m+n}\right) = \mathsf{Pr}\left(\tilde{X}_1^{m+n} = x_1^{m+n}\right). \tag{17}$$

However these two input processes may not achieve identical probability distributions for the channel state Z_m since a general finite-state channel may not be a finite-memory channel. Thus

$$\mathsf{Pr}\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n},X_{m+1}^{m+n}=x_{m+1}^{m+n}\right)\neq\mathsf{Pr}\left(\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n},\tilde{X}_{m+1}^{m+n}=x_{m+1}^{m+n}\right)$$

and (7) does not hold in general. But we can show that by utilizing the indecomposability property (16), we can make the difference between $H\left(X_{m+1}^{m+n}\left|Y_{m+1}^{m+n}\right.\right)$ and $H\left(\tilde{X}_{m+1}^{m+n}\left|\tilde{Y}_{m+1}^{m+n}\right.\right)$ arbitrarily small.

We first bound the difference between

$$\mathsf{Pr}\left(X_1^{m+n}=x_1^{m+n},Z_m=z_m\right)$$

and

$$\mathsf{Pr}\left(\tilde{X}_1^{m+n}=x_1^{m+n},\tilde{Z}_m=z_m\right).$$

It is easy to see that

$$\begin{aligned}&\mathsf{Pr}\left(X_1^{m+n}=x_1^{m+n},Z_m=z_m\right)\\&\quad=\sum_{z_0}\mathsf{Pr}\left(X_1^{m+n}=x_1^{m+n},Z_m=z_m,Z_0=z_0\right)\\&\quad=\mathsf{Pr}\left(X_1^{m+n}=x_1^{m+n}\right)\cdot\\&\sum_{z_0}\mathsf{Pr}\left(Z_m=z_m,Z_0=z_0\left|X_1^{m+n}=x_1^{m+n}\right.\right)\\&\quad=\mathsf{Pr}\left(X_1^{m+n}=x_1^{m+n}\right)\cdot\\&\sum_{z_0}q\left(z_m\left|x_1^m,z_0\right.\right)\mathsf{Pr}\left(Z_0=z_0\left|X_1^{m+n}=x_1^{m+n}\right.\right)\end{aligned}\tag{18}$$

where the third equality uses the property that Z_m is independent of the future channel inputs X_{m+1}^{m+n} given Z_0 and X_1^m. Similarly

$$\begin{aligned}&\mathsf{Pr}\left(\tilde{X}_1^{m+n}=x_1^{m+n},\tilde{Z}_m=z_m\right)\\&\quad=\mathsf{Pr}\left(\tilde{X}_1^{m+n}=x_1^{m+n}\right)\\&\qquad\cdot\sum_{z_0}q\left(z_m\left|x_1^m,z_0\right.\right)\mathsf{Pr}\left(\tilde{Z}_0=z_0\left|\tilde{X}_1^{m+n}=x_1^{m+n}\right.\right).\end{aligned}\tag{19}$$

From inequality (16), we know that $q\left(z_m\left|x_1^m,z_0\right.\right)$ satisfies

$$A\left(z_m,x_1^m\right)<q\left(z_m\left|x_1^m,z_0\right.\right)<A(z_m,x_1^m)+\delta$$

for all z_0, if $m>M$, where the quantity $A\left(z_m,x_1^m\right)$ depends only upon z_m and x_1^m, and is independent of z_0. Therefore, we have

$$\begin{aligned}A\left(z_m,x_1^m\right)&<\sum_{z_0}q\left(z_m\left|x_1^m,z_0\right.\right)\mathsf{Pr}\left(Z_0=z_0\left|X_1^{m+n}=x_1^{m+n}\right.\right)\\&<A\left(z_m,x_1^m\right)+\delta\end{aligned}\tag{20}$$

and

$$\begin{aligned}A\left(z_m,x_1^m\right)&<\sum_{\tilde{z}_0}q\left(z_m\left|x_1^m,\tilde{z}_0\right.\right)\mathsf{Pr}\left(\tilde{Z}_0=\tilde{z}_0\left|\tilde{X}_1^{m+n}=x_1^{m+n}\right.\right)\\&<A(z_m,x_1^m)+\delta.\end{aligned}\tag{21}$$

Combining (17), (18), (19), (20), and (21), we can show that

$$\begin{aligned}&\left|\mathsf{Pr}\left(X_1^{m+n}=x_1^{m+n},Z_m=z_m\right)\right.\\&\quad\left.-\mathsf{Pr}\left(\tilde{X}_1^{m+n}=x_1^{m+n},\tilde{Z}_m=z_m\right)\right|\\&\qquad\leqslant\mathsf{Pr}\left(X_1^{m+n}=x_1^{m+n}\right)\delta,\end{aligned}$$

and

$$\begin{aligned}&\left|\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n},Z_m=z_m\right)\right.\\&\quad\left.-\mathsf{Pr}\left(\tilde{X}_{m+1}^{m+n}=x_{m+1}^{m+n},\tilde{Z}_m=z_m\right)\right|\\&\qquad\leqslant\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n}\right)\delta.\end{aligned}$$

Similarly, since the joint probability

$$\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n},Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)$$

can be rewritten as

$$\begin{aligned}&\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n},Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\\&\quad=\sum_{z_m}\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n},Y_{m+1}^{m+n}=y_{m+1}^{m+n},Z_m=z_m\right)\\&\quad=\sum_{z_m}p\left(y_{m+1}^{m+n}\left|x_{m+1}^{m+n},z_m\right.\right)\\&\qquad\times\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n},Z_m=z_m\right),\end{aligned}$$

and the joint probability

$$\begin{aligned}&\mathsf{Pr}\left(\tilde{X}_{m+1}^{m+n}=x_{m+1}^{m+n},\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\\&\quad=\sum_{z_m}\mathsf{Pr}\left(\tilde{X}_{m+1}^{m+n}=x_{m+1}^{m+n},\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n},\tilde{Z}_m=z_m\right)\\&\quad=\sum_{z_m}p\left(y_{m+1}^{m+n}\left|x_{m+1}^{m+n},z_m\right.\right)\\&\qquad\times\mathsf{Pr}\left(\tilde{X}_{m+1}^{m+n}=x_{m+1}^{m+n},\tilde{Z}_m=z_m\right),\end{aligned}$$

we can bound the difference as

$$\begin{aligned}&\left|\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n},Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\right.\\&\quad\left.-\mathsf{Pr}\left(\tilde{X}_{m+1}^{m+n}=x_{m+1}^{m+n},\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\right|\\&\qquad\leqslant\sum_{z_m}p\left(y_{m+1}^{m+n}\left|x_{m+1}^{m+n},z_m\right.\right)\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n}\right)\delta\\&\qquad=|\mathcal{Z}|\,\mathsf{Pr}_u\left(X_{m+1}^{m+n}=x_{m+1}^{m+n},Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\delta,\end{aligned}\tag{22}$$

where $\mathsf{Pr}_u\left(X_{m+1}^{m+n}=x_{m+1}^{m+n},Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)$ is the joint channel input/output distribution when the channel state Z_m is initialized with equal probability for every possible value z_m.[4] Thus, the L_1-norm of the difference can be bounded as

$$\begin{aligned}&\left\|\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n},Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\right.\\&\quad\left.-\mathsf{Pr}\left(\tilde{X}_{m+1}^{m+n}=x_{m+1}^{m+n},\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\right\|_1\\&\qquad\leqslant|\mathcal{Z}|\,\delta.\end{aligned}\tag{23}$$

Following the same line of reasoning, we can bound the channel output distributions induced by the two input processes as follows:

$$\begin{aligned}&\left|\mathsf{Pr}\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)-\mathsf{Pr}\left(\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\right|\\&\quad\leqslant\sum_{x_{m+1}^{m+n}}\sum_{z_m}p\left(y_{m+1}^{m+n}\left|x_{m+1}^{m+n},z_m\right.\right)\mathsf{Pr}\left(X_{m+1}^{m+n}=x_{m+1}^{m+n}\right)\delta\\&\quad=|\mathcal{Z}|\,\mathsf{Pr}_u\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\delta\end{aligned}\tag{24}$$

where $\mathsf{Pr}_u\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)$ is the channel output distribution when the channel state Z_m is initialized with equal probability for every possible value z_m. The L_1-norm of the difference satisfies

$$\left\|\mathsf{Pr}\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)-\mathsf{Pr}\left(\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\right\|_1\leqslant|\mathcal{Z}|\,\delta.\tag{25}$$

[4]Note that this distribution remains the same whether the input is the original process $\mathcal{X}$ or the approximating process $\tilde{\mathcal{X}}$, because the initial state distribution is the same and the finite-dimensional input distribution is also identical.

To complete the extension of Theorem 1, we consider three possible scenarios for the channel output characteristics.

A. Channel Output With Finite Alphabet

Applying [18, Theorem 16.3.2] and (25), we can obtain the following bound:

$$\left|H\left(Y_{m+1}^{m+n}\right)-H\left(\tilde{Y}_{m+1}^{m+n}\right)\right|\leqslant-|\mathcal{Z}|\,\delta\log\frac{|\mathcal{Z}|\,\delta}{|\mathcal{Y}|^{n}}.$$

Similarly, we can show that

$$\left|H\left(X_{m+1}^{m+n},Y_{m+1}^{m+n}\right)-H\left(\tilde{X}_{m+1}^{m+n},\tilde{Y}_{m+1}^{m+n}\right)\right|\leqslant-|\mathcal{Z}|\,\delta\log\frac{|\mathcal{Z}|\,\delta}{(|\mathcal{X}|\,|\mathcal{Y}|)^{n}}. \quad (26)$$

It is straightforward to see that the terms on the right-hand side of the two inequalities above go to 0 as δ goes to 0. Recalling that (17) implies

$$H\left(X_{m+1}^{m+n}\right)=H\left(\tilde{X}_{m+1}^{m+n}\right) \quad (27)$$

we can conclude that $\left|H\left(X_{m+1}^{m+n}\left|Y_{m+1}^{m+n}\right.\right)-H\left(\tilde{X}_{m+1}^{m+n}\left|\tilde{Y}_{m+1}^{m+n}\right.\right)\right|$ can indeed be made arbitrarily small if we choose m large enough in the approximation. Following the reasoning of the proof of Theorem 1, we can then reach the same conclusion as in Theorem 1.

B. Channel Output With Infinite Alphabet

The infinite alphabet of the channel output makes it impossible to apply [18, Theorem 16.3.2] directly. But we will show that the difference between the entropies induced by the two input processes is mainly determined by a finite subset of the infinite alphabet when $H(Y_k)<W$ for some finite value W. This implies that

$$\begin{aligned}H\left(Y_{m+1}^{m+n}\right)&=-\sum_{y_{m+1}^{m+n}}\mathsf{Pr}\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\\&\quad\times\log\mathsf{Pr}\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\\&\leqslant\sum_{i=1}^{n}H\left(Y_{m+i}\right)\leqslant nW.\end{aligned}$$

To see how we can bound the entropy difference by decomposing the infinite alphabet, we consider an alphabet in which the magnitudes of the elements are not bounded, while the number of elements with magnitude bounded by any specified positive value is finite. The other scenarios involving infinite alphabets can be treated similarly.

For any $n\geqslant 1$ and a positive constant Y_L, we denote the following event:

$$|Y_k|\leqslant Y_L,\quad\text{for every } k\ (m+1\leqslant k\leqslant m+n)$$

as event $\mathcal{O}$. Similarly, we define the event

$$\left|\tilde{Y}_k\right|\leqslant Y_L,\quad\text{for every } k\ (m+1\leqslant k\leqslant m+n)$$

as event $\tilde{\mathcal{O}}$. For any

$$0<\mu<\min\left\{H\left(Y_{m+1}^{m+n}\right),H\left(\tilde{Y}_{m+1}^{m+n}\right)\right\}$$

and $n\geqslant 1$, one can find a positive Y_L and corresponding events $\mathcal{O}$ and $\tilde{\mathcal{O}}$ such that for Y_{m+1}^{m+n} and $\tilde{Y}_{m+1}^{m+n}$, the finite partial sum

$$\begin{aligned}&H_1\left(Y_{m+1}^{m+n}\right)\\&=-\sum_{y_{m+1}^{m+n}\in\mathcal{O}}\mathsf{Pr}\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\log\mathsf{Pr}\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\\&\geqslant H\left(Y_{m+1}^{m+n}\right)-\mu,\\&H_1\left(\tilde{Y}_{m+1}^{m+n}\right)\\&=-\sum_{y_{m+1}^{m+n}\in\tilde{\mathcal{O}}}\mathsf{Pr}\left(\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\log\mathsf{Pr}\left(\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\\&\geqslant H\left(\tilde{Y}_{m+1}^{m+n}\right)-\mu,\end{aligned}$$

and

$$\begin{aligned}&H_2\left(Y_{m+1}^{m+n}\right)\\&=-\sum_{y_{m+1}^{m+n}\in\mathcal{O}^c}\mathsf{Pr}\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\log\mathsf{Pr}\left(Y_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\\&\leqslant\mu,\\&H_2\left(\tilde{Y}_{m+1}^{m+n}\right)\\&=-\sum_{y_{m+1}^{m+n}\in\tilde{\mathcal{O}}^c}\mathsf{Pr}\left(\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\log\mathsf{Pr}\left(\tilde{Y}_{m+1}^{m+n}=y_{m+1}^{m+n}\right)\\&\leqslant\mu\end{aligned}$$

where $\mathcal{O}^c$ and $\tilde{\mathcal{O}}^c$ are the complements of $\mathcal{O}$ and $\tilde{\mathcal{O}}$, respectively. This decomposition is made possible by the fact that $H\left(Y_{m+1}^{m+n}\right)$ and $H\left(\tilde{Y}_{m+1}^{m+n}\right)$ are finite. It is easy to see that

$$\begin{aligned}H\left(Y_{m+1}^{m+n}\right)&=H_1\left(Y_{m+1}^{m+n}\right)+H_2\left(Y_{m+1}^{m+n}\right)\\H\left(\tilde{Y}_{m+1}^{m+n}\right)&=H_1\left(\tilde{Y}_{m+1}^{m+n}\right)+H_2\left(\tilde{Y}_{m+1}^{m+n}\right).\end{aligned}$$

Given that events $\mathcal{O}$ and $\tilde{\mathcal{O}}$ are true, the conditional channel output has finite alphabet of size $|\mathcal{Y}_1|$. Following the approach in [18, the proof of Theorem 16.3.2], we can bound $\left|H_1\left(Y_{m+1}^{m+n}\right)-H_1\left(\tilde{Y}_{m+1}^{m+n}\right)\right|$ as follows:

$$\left|H_1\left(Y_{m+1}^{m+n}\right)-H_1\left(\tilde{Y}_{m+1}^{m+n}\right)\right|\leqslant-|\mathcal{Z}|\,\delta\log\frac{|\mathcal{Z}|\,\delta}{|\mathcal{Y}_1|^{n}}.$$

Therefore, we can bound the entropy difference as

$$\begin{aligned}&\left|H\left(Y_{m+1}^{m+n}\right)-H\left(\tilde{Y}_{m+1}^{m+n}\right)\right|\\&\quad=\left|\left[H_1\left(Y_{m+1}^{m+n}\right)+H_2\left(Y_{m+1}^{m+n}\right)\right]\right.\\&\qquad\left.-\left[H_1\left(\tilde{Y}_{m+1}^{m+n}\right)+H_2\left(\tilde{Y}_{m+1}^{m+n}\right)\right]\right|\\&\quad\leqslant\left|H_1\left(Y_{m+1}^{m+n}\right)-H_1\left(\tilde{Y}_{m+1}^{m+n}\right)\right|\\&\qquad+\left|H_2\left(Y_{m+1}^{m+n}\right)-H_2\left(\tilde{Y}_{m+1}^{m+n}\right)\right|\\&\quad\leqslant-|\mathcal{Z}|\,\delta\log\frac{|\mathcal{Z}|\,\delta}{|\mathcal{Y}_1|^{n}}+2\mu.\end{aligned} \quad (28)$$

By choosing μ and δ (or m) appropriately, we can make the difference $\left|H\left(Y_{m+1}^{m+n}\right)-H\left(\tilde{Y}_{m+1}^{m+n}\right)\right|$ arbitrarily small. A similar conclusion can be derived for the difference

$$\left|H\left(X_{m+1}^{m+n},Y_{m+1}^{m+n}\right)-H\left(\tilde{X}_{m+1}^{m+n},\tilde{Y}_{m+1}^{m+n}\right)\right|.$$

Thus, we can conclude that

$$\left|H\left(X_{m+1}^{m+n}\left|Y_{m+1}^{m+n}\right.\right)-H\left(\tilde{X}_{m+1}^{m+n}\left|\tilde{Y}_{m+1}^{m+n}\right.\right)\right|$$

can be made arbitrarily small if we choose m large enough and μ sufficiently small. The rest of the proof then follows.

C. Continuous Channel Output

When the channel output Y_k is continuous, we first consider the conditional entropy $H\left(X_{m+1}^{m+n}\left|Y_{m+1}^{m+n}(\Delta)\right.\right)$ where $Y_k(\Delta)$ is the quantized output corresponding to quantization interval Δ. It is

well known that as $\Delta \to 0$, $I\left(X_{m+1}^{m+n}; Y_{m+1}^{m+n}(\Delta)\right)$ converges to $I\left(X_{m+1}^{m+n}; Y_{m+1}^{m+n}\right)$ (see [18, p. 231]). We can show similarly that $H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}(\Delta)\right.\right)$ converges to $H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}\right.\right)$. In other words, the difference between

$$H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}(\Delta)\right.\right) \quad \text{and} \quad H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}\right.\right)$$

can be made arbitrarily small if we choose Δ small enough. Furthermore, we have shown in the previous two scenarios that the difference

$$\left|H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}(\Delta)\right.\right) - H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}(\Delta)\right.\right)\right|$$

can be made arbitrarily small if we choose m and μ appropriately. Therefore, for any $\epsilon > 0$, we can choose an appropriate $\Delta > 0$ such that

$$\begin{aligned}\left|H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}(\Delta)\right.\right) - H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}\right.\right)\right| &< \epsilon/4\\ \left|H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}(\Delta)\right.\right) - H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}\right.\right)\right| &< \epsilon/4\end{aligned}$$

and we can find m and μ such that

$$\left|H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}(\Delta)\right.\right) - H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}(\Delta)\right.\right)\right| < \epsilon/2$$

for the chosen Δ. Thus, we can bound the difference in the conditional entropy as

$$\begin{aligned}&\left|H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}\right.\right) - H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}\right.\right)\right|\\ &\quad= \left|H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}\right.\right) - H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}(\Delta)\right.\right)\right.\\ &\qquad\left. + H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}(\Delta)\right.\right) - H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}(\Delta)\right.\right)\right.\\ &\qquad\left. + H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}(\Delta)\right.\right) - H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}\right.\right)\right|\\ &\quad\leqslant \left|H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}\right.\right) - H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}(\Delta)\right.\right)\right|\\ &\qquad + \left|H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}(\Delta)\right.\right) - H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}(\Delta)\right.\right)\right|\\ &\qquad + \left|H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}(\Delta)\right.\right) - H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}\right.\right)\right|\\ &\quad\leqslant \epsilon/4 + \epsilon/2 + \epsilon/4 = \epsilon.\end{aligned}$$

We can conclude that by properly choosing m, μ, and Δ, we can make the difference $\left|H\left(X_{m+1}^{m+n} \left| Y_{m+1}^{m+n}\right.\right) - H\left(\tilde{X}_{m+1}^{m+n} \left| \tilde{Y}_{m+1}^{m+n}\right.\right)\right|$ arbitrarily small. The rest of the proof then follows. □

For the two-dimensional indecomposable finite-state channel model, we can also show that if we choose the Markov order and other parameters properly, the difference

$$\left|H\left(X_{m+1,n+1}^{m+u,n+v} \left| Y_{m+1,n+1}^{m+u,n+v}\right.\right) - H\left(\tilde{X}_{m+1,n+1}^{m+u,n+v} \left| \tilde{Y}_{m+1,n+1}^{m+u,n+v}\right.\right)\right|$$

can become arbitrarily small. The proof is similar to its one-dimensional counterpart. We begin with the channel state $Z_{1,0}^{m+u,0} \bigcup Z_{0,1}^{0,n+v}$ and the input $X_{1,1}^{m+u,n} \bigcup X_{1,n+1}^{m,n+v}$, and make use of the indecomposability of the channel. The rest of the derivation follows the same line of reasoning as above.

ACKNOWLEDGMENT

The authors would like to thank Prof. Brian Marcus and Dr. Guangyue Han for pointing out the erroneous reliance on Breiman's result [20] in [27] and an earlier version of this correspondence.

REFERENCES

[1] W. Hirt and J. L. Massey, "Capacity of the discrete-time Gaussian channel with intersymbol interference," *IEEE Trans. Inf. Theory*, vol. 34, no. 3, pp. 380–388, May 1988.

[2] D. M. Arnold and H.-A. Loeliger, "On the information rate of binary-input channels with memory," in *Proc. IEEE Int. Conf. Communications*, Helsinki, Finland, Jun. 2001, vol. 9, pp. 2692–2695.

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

[4] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rate of finite state ISI channels," in *Proc. GLOBECOM 2001*, San Antonio, TX, Nov. 2001, vol. 5, pp. 2992–2996.

[5] D. M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavčić, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3498–3508, Aug. 2006.

[6] 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.

[7] S. Yang and A. Kavčić, "Capacity of partial response channels," in *Handbook of Coding and Signal Processing for Magnetic Recording Systems*. Boca Raton, FL: CRC , 2004, ch. 13.

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

[9] J. B. Soriaga and P. H. Siegel, "On near-capacity coding systems for partial response channels," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 267.

[10] A. Kavčić, X. Ma, and N. Varnica, "Matched information rate codes for partial response channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 973–989, Mar. 2005.

[11] N. Varnica, X. Ma, and A. Kavčić, "Capacity approaching codes for partial response channels," in *Handbook of Coding and Signal Processing for Magnetic Recording Systems*. Boca Raton, FL: CRC, 2004, ch. 23.

[12] D. Blackwell, L. Breiman, and A. J. Thomasian, "Proof of Shannon's theorem for finite-state indecomposable channels," *Ann. Math. Statist.*, vol. 29, no. 4, pp. 1209–1220, Dec. 1958.

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

[14] S. Yang, A. Kavčić, and S. Tatikonda, "The feedback capacity of finite-state machine channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 799–810, Mar. 2005.

[15] S. Shamai (Shitz) and Y. Kofman, "On the capacity of binary and Gaussian channels with run-length-limited inputs," *IEEE Trans. Inf. Theory*, vol. 38, no. 3, pp. 584–594, May 1990.

[16] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[17] A. Feinstein, "On the coding theorem and its converse for finite memory channels," *Inf. Contr.*, vol. 2, no. 1, pp. 25–44, 1959.

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

[19] D. S. Ornstein, *Ergodic Theory, Randomness and Dynamical Systems*. New Haven, CT: Yale Univ. Press, 1974.

[20] L. Breiman, "Finite-state channels," in *Proc. 2nd Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, Prague, Czechoslovakia, 1959, pp. 49–60.

[21] D. Psaltis and F. Mok, "Holographic memories," *Scient. Amer.*, pp. 70–76, Nov. 1995.

[22] H. O. Georgii, *Gibbs Measures and Phase Transitions*. Berlin, Germany: DeGruyter, vol. 9, Studies in Mathematics.

[23] D. N. Politis, "Markov chain in many dimensions," *Adv. Appl. Probab.*, vol. 26, pp. 756–774, 1994.

[24] Y. Katznelson and B. Weiss, "Commuting measure-preserving transformations," *Israel J. Math.*, vol. 12, pp. 161–173, 1972.

[25] D. Anastassiou and D. J. Sakrison, "Some results regarding the entropy rates of random fields," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 340–343, Mar. 1982.

[26] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[27] J. Chen and P. H. Siegel, "Markov processes asymptotically achieve the capacity of finite-state intersymbol interference channels," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 349.

[28] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications*. Providence, RI: Amer. Math. Soc., 1980, vol. 1, Contemporary Mathematics .