Constraint Gain

John L. Fan, *Member, IEEE*, T. Lei Poo, *Student Member, IEEE*, and Brian H. Marcus, *Fellow, IEEE*

***Abstract*—In digital storage systems where the input to the noisy channel is required to satisfy a modulation constraint, the constrained code and error-control code (ECC) are typically designed and decoded independently. The achievable rate for this situation is evaluated as the rate of average intersection of the constraint and the ECC. The gap from the capacity of the noisy constrained channel is called the *constraint gain*, which represents the potential improvement in combining the design and decoding of the constrained code and the ECC. The constraint gain is computed for various constraints over the binary-input additive white Gaussian noise (AWGN) channel (BIAWGNC) as well as over intersymbol interference (ISI) channels. Finally, it is shown that an infinite cascade of reverse concatenation with independent decoding of constraint and ECC has a capacity equal to the rate of average intersection.**

***Index Terms*—Capacity, constrained codes, error-control codes (ECCs), noisy constrained channels, reverse concatenation.**

I. Introduction

In constrained coding, one encodes arbitrary data sequences into a restricted set of sequences that can be transmitted over a channel. Constraints such as runlength-limited (RLL) constraints are commonly used in magnetic and optical storage ([12], [17] and [11]). While the calculation of the noise-free capacity of constrained sequences is well known, the computation of the capacity of a constraint in the presence of noise (the "noisy constrained capacity") had been an unresolved problem in the half-century since Shannon's landmark paper [22], with only a handful of papers on this topic ([9], [16], [25], and [21]). Recently, significant progress has been made on this problem: papers by Arnold and Loeliger [1], Kavčić [13], Pfister, Soriaga, and Siegel [20], and Vontobel and Arnold [24] present practical methods for computing accurate bounds on the noisy capacity of Markov sources.

In this paper, we consider the performance of (d,k)-RLL modulation constraints on a binary-input additive white Gaussian noise (AWGN) channel (BIAWGNC) and over several intersymbol interference (ISI) channels. In particular, we evaluate the noisy capacity of such a system assuming a maxentropic input. This is compared with a lower bound on the capacity that corresponds to the rate of average intersection (defined in Section II-D) of the constraint and an error-control code (ECC). It is argued that this lower bound corresponds to the situation where the constraint is ignored in the design and decoding of the ECC. We introduce the concept of *constraint gain* (in analogy to "coding gain"), which corresponds to the additional robustness to error that is afforded by jointly designing the ECC and constrained code (also known as modulation code), and using the modulation constraint in the decoding process. Finally, we consider a method for combining the modulation constraint and the ECC using an infinite cascade of reverse concatenation. The achievable rate when the constraint and ECC are independently decoded is found to be equal to the rate of average intersection (when positive).

Manuscript received April 10, 2002; revised March 22, 2004. The material in this paper was presented in part at the IEEE International Symposia on Information Theory, Lausanne, Switzerland, June/July 2002 and Yokohama, Japan, June 2003.

J. L. Fan is with Flarion Technologies, Bedminster, NJ 07921 USA (e-mail: john@johnfan.com).

T. L. Poo is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305-9510 USA (e-mail: leipoo@stanford.edu).

B. H. Marcus was with IBM Almaden Research Center, San Jose, CA, USA. He is now with the Department of Mathematics, University of British Columbia, Vancouver, BC V6T IZ2, Canada (e-mail: marcus@math.ubc.ca).

Communicated by R. Urbanke, Associate Editor for Coding Techniques.

Digital Object Identifier 10.1109/TIT.2004.833349

II. Capacity

A. Basic Definitions

Let X represent the transmitted sequence, and Y represent the received sequence. (In storage, X would be the sequence of bits written to the disk, while Y would be the sequence of signals read from the disk.) These are infinite sequences of the form $X = (X_1, X_2, \ldots)$, and the first n elements are represented by $X_1^n = (X_1, X_2, \ldots, X_n)$.

Let θ represent the channel. In Sections II and III, we focus on two memoryless channels, the binary-symmetric channel (BSC) and the BIAWGNC. For the BSC, the input is $X_i \in \{0,1\}$ and the output is incorrect (i.e., $Y_i = 1 - X_i$) with crossover probability p and correct (i.e., $Y_i = X_i$) with probability $1-p$. For the BIAWGNC, we assume the input $X_i \in \{0,1\}$ is mapped using pulse amplitude modulation (PAM) to $\{1,-1\}$. The output of the channel is $Y_i = (-1)^{X_i} + N_i$, where N_i represents Gaussian noise with variance σ^2. In this case, we use the notation θ to also denote the signal-to-noise ratio (SNR) $\theta = \frac{1}{\sigma^2}$.

By an *unconstrained channel*, we mean a binary-input channel with no input constraints. For an unconstrained channel θ, the capacity is defined as the supremum of the mutual information rate over all stationary stochastic processes X

$$\text{cap}(\theta) = \sup_X I(X;Y).$$

The entropy rate of a stochastic process $X = (X_1, X_2, \ldots)$ is defined as

$$H(X) = \lim_{n\to\infty} \frac{1}{n} H\left(X_1^n\right).$$

The conditional entropy rate $H(X|Y)$ is defined similarly, and the mutual information rate is defined by

$$I(X;Y) = H(X) - H(X|Y).$$

Alternatively, the capacity can be given by

$$\mathrm{cap}(\theta) = \lim_{n\to\infty} \frac{1}{n} \sup_{p(X_1^n)} I\left(X_1^n; Y_1^n\right) \tag{1}$$

where the supremum is taken over all possible distributions $p\left(X_1^n\right)$ on the input sequence X_1^n.

From the channel coding theorem, $\mathrm{cap}(\theta)$ corresponds to the maximum number of bits that can be transmitted per symbol, and still have error-free performance for the channel. Specifically, for any $\epsilon > 0$, there exists an n and a subset $\boldsymbol{S}_{\mathrm{ECC}}^n$ consisting of codewords of length n such that the code rate $R_{\mathrm{ECC}} = \frac{1}{n} \log |\boldsymbol{S}_{\mathrm{ECC}}^n|$ satisfies

$$\mathrm{cap}(\theta) - R_{\mathrm{ECC}} \leq \epsilon$$

and the probability of block error (the maximum probability over all codewords) when using this code over the channel is less than ϵ. In this case, the code $\boldsymbol{S}_{\mathrm{ECC}}^n$ is said to be (n, ϵ) *-good*.

B. *Noisy Constrained Capacity*

The *constraint set* $\boldsymbol{S}_C$ is the set of sequences that are allowed as inputs to the channel. We restrict our attention to *constrained systems* ([17]), which are defined by directed graphs whose nodes correspond to the states and whose edges are labeled by symbols (e.g., 0 and 1). RLL codes form an important class of constrained systems. The (d, k)-RLL constraints restrict the number of 0's between adjacent 1's to be a minimum of d and a maximum of k. Following notation used in magnetic recording, we consider the (d, k)-RLL constraints to be applied to a nonreturn to zero NRZ inverse (NRZI) sequence where the 1's represent transitions. This yields an NRZ sequence as input to the channel. The minimum and maximum lengths of a run in the NRZ sequence are $d+1$ and $k+1$, respectively. For example, for the $(0,1)$-RLL constraint, the transition sequence cannot contain more than one nontransition between adjacent transitions, e.g., $S^{\mathrm{NRZI}} = 011101011010\ldots$; the corresponding input to the channel cannot contain more than two identical bits in a row, e.g., $S^{\mathrm{NRZ}} = 010110010011\ldots$. Note that $S_i^{\mathrm{NRZ}} = S_{i-1}^{\mathrm{NRZ}} \oplus S_i^{\mathrm{NRZI}}$.

The *Shannon capacity* (or noiseless capacity) of the constrained system is given by

$$\mathrm{cap}(C) = \lim_{n\to\infty} \frac{1}{n} \log |\boldsymbol{S}_C^n|$$

where $\boldsymbol{S}_C^n$ is the set of sequences of length n that meet the constraint. This capacity can be computed as the logarithm[1] of the largest real eigenvalue of the adjacency matrix of any deterministic[2] presentation of the constraint.

Define the noisy constrained capacity (also known as (a.k.a.) the noisy input-restricted channel capacity) as

$$\mathrm{cap}(C, \theta) = \sup_{X \in \boldsymbol{S}_C} I(X; Y) \tag{2}$$

where "$X \in \boldsymbol{S}_C$" means that X is a stationary process supported on the constraint $\boldsymbol{S}_C$, and Y is the corresponding output process. A process is said to be supported on a constraint if any finite sequence of strictly positive probability belongs to the constraint. Clearly

$$\mathrm{cap}(C, \theta) \leq \min\{\mathrm{cap}(C), \mathrm{cap}(\theta)\}.$$

Observe that the Shannon capacity is simply a limiting version of the noisy constrained capacities. For a BIAWGNC with channel parameter $\theta = \frac{1}{\sigma^2}$, we have

$$\mathrm{cap}(C) = \lim_{\theta\to\infty} \mathrm{cap}(C, \theta).$$

As in (1), it is possible to rewrite the noisy constrained capacity in (2) as

$$\mathrm{cap}(C, \theta) = \lim_{n\to\infty} \frac{1}{n} \sup_{\substack{p(X_1^n)\\ X_1^n \in \boldsymbol{S}_C^n}} I\left(X_1^n; Y_1^n\right).$$

The supremum is taken over all probability distributions $p\left(X_1^n\right)$ with support on $\boldsymbol{S}_C^n$.

We remark that the operational meaning of $\mathrm{cap}(C, \theta)$ is similar to that of $\mathrm{cap}(\theta)$: it corresponds to the maximum number of bits that can be transmitted over the noisy constrained channel, and still have error-free performance. This follows from the Shannon coding theorems for finite-state channels, see [8, Sec. 5.9].

We briefly review a practical method to compute the mutual information rate for constrained channels. The Arnold–Loeliger approach [1] first breaks up the mutual information as follows:

$$I(X; Y) = H(Y) - H(Y|X). \tag{3}$$

For a channel with additive noise $(Y = X + N)$, the term $H(Y|X)$ is simply the entropy of the noise, e.g., for additive Gaussian noise

$$H(Y|X) = H(N) = \frac{1}{2} \log_2(2\pi e \sigma^2).$$

The key innovation in [1] is to use computations similar to the forward–backward algorithm (also known as the Bahl–Cocke–Jelinek–Raviv (BCJR) algorithm) to calculate the first term $H(Y)$ in (3). The entropy rate of the process Y is

$$H(Y) = -\lim_{n\to\infty} E\left[\log p\left(Y_1^n\right)\right]$$

where

$$p\left(Y_1^n\right) = \sum_s p\left(Y_1^n, S_n = s\right).$$

The forward messages

$$\alpha_n(s) = p\left(Y_1^n, S_n = s\right)$$

satisfy the following recursive sum:

$$\alpha_n(s) = \sum_{s'} p\left(Y_n \mid s, s'\right) p(s \mid s') \alpha_{n-1}(s')$$

where $p\left(Y_n \mid s, s'\right) = p\left(Y_n \mid X = x\right)$ is the probability of receiving Y_n given that x was transmitted (where x is the label for

[1]In this paper, all logarithms are taken base 2, so that the capacity is measured in bits.

[2]A labeling is deterministic if at each state, all outgoing edges are labeled distinctly.

transition from state s' to state s), and $p(s \mid s')$ is the transition probability from state s' to state s. With these computations, it is possible to evaluate $H(Y)$ and the mutual information $I(X, Y)$ for any stationary process X. Note that to compute $\mathrm{cap}(C, \theta)$, it may be necessary to search over all stationary processes X supported on $\boldsymbol{S}_C$ to find the one that maximizes the mutual information rate.

C. Noisy Maxentropic Constrained Capacity

For a given constraint $\boldsymbol{S}_C$, there is a particular source supported on $\boldsymbol{S}_C$ that deserves special attention: the stationary process $X_{\max}$ that maximizes the entropy rate among all stationary processes supported on the constraint, known as the *maxentropic distribution*. There are explicit formulas for this process ([17]), and its entropy rate coincides with the Shannon capacity of the constraint, i.e,

$$H(X_{\max}) = \mathrm{cap}(C).$$

The noisy maxentropic constrained capacity, or simply "*maxentropic capacity*," is defined as

$$\mathrm{cap}_{\mathrm{maxent}}(C, \theta) = I(X_{\max}; Y_{\max}) \tag{4}$$

where $Y_{\max}$ denotes the output distribution corresponding to $X_{\max}$. Note that the Arnold–Loeliger [1] method outlined in Section II-B can be used to evaluate $\mathrm{cap}_{\mathrm{maxent}}(C, \theta)$. This capacity in (4) defined by the maxentropic distribution is clearly upper-bounded by the noisy constrained capacity in (2)

$$\mathrm{cap}_{\mathrm{maxent}}(C, \theta) \leq \mathrm{cap}(C, \theta).$$

The noisy maxentropic constrained capacity in (4) is easier to compute than the noisy constrained capacity in (2) since it is not necessary to take the supremum over all input distributions. For some situations, this maxentropic capacity $\mathrm{cap}_{\mathrm{maxent}}(C, \theta)$ provides an approximation to the constrained capacity $\mathrm{cap}(C, \theta)$. Also, it should be noted that the maxentropic distribution is often used in practical implementations of constrained coding.

D. Intersection Rate

Let $\boldsymbol{S}_{\mathrm{ECC}}^n \subset F_2^n$ be a block binary error-control code of length n, where $F_2 = \{0, 1\}$. The *intersection rate* is defined as

$$\frac{1}{n} \log |\boldsymbol{S}_C^n \cap \boldsymbol{S}_{\mathrm{ECC}}^n|$$

which corresponds to the transmission rate that is possible by using the block code $\boldsymbol{S}_{\mathrm{ECC}}^n$ but restricting the transmission to codewords that also belong to the constraint set $\boldsymbol{S}_C^n$.

Patapoutian and Kumar [19] evaluated the expected size of intersection of a (d, k)-RLL constraint with a fixed *linear* error-correcting code (and its cosets). They gave the rate of intersection averaged over all cosets of the linear error correcting code as

$$\frac{1}{n} \log |\boldsymbol{S}_C^n| + \frac{1}{n} \log |\boldsymbol{S}_{\mathrm{ECC}}^n| - 1.$$

We define the rate of average intersection as

$$\mathrm{rate}_{\mathrm{avgECC}}(C, \theta) = \lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log E\left[|\boldsymbol{S}_C^n \cap \boldsymbol{S}_{\mathrm{ECC}}^n|\right]$$

where the expectation is taken over *all* (n, ϵ)-good codes $\boldsymbol{S}_{\mathrm{ECC}}^n$. This extends the definition of the rate of intersection from Patapoutian and Kumar [19] in that the codes considered are not necessarily linear.

Motivated by the preceding results and by Loeliger's averaging bounds for linear codes (see [14]), we prove the following.

Proposition 1:

$$\mathrm{rate}_{\mathrm{avgECC}}(C, \theta) = \mathrm{cap}(\theta) + \mathrm{cap}(C) - 1. \tag{5}$$

Proof: Denote $\boldsymbol{S}_{\mathrm{ECC}}^n$ by G for convenience. Let F_2^n be the vector space over the field $F_2 = \{0, 1\}$. Note that for any $v \in F_2^n$, the translated code $G + v$ is also (n, ϵ)-good for the unconstrained channel. This is because $|G| = |G+v|$, thus, they have the same rate. Also, $G + v$ has approximately the same error correcting ability as G since it is possible to communicate which fixed translate v was used at the expense of some initial bits in transmission, and the decoder could just subtract v from the received codewords and then use the decoding from G.

Since F_2^n is a vector space over F_2, it is an Abelian group under addition. Let F_2^n act on the set of all (n, ϵ)-good codes by translation, i.e., G maps to $G+v$ for any $v \in F_2^n$. This is a group action since $(G + v) + w = G + (v + w)$ for any $v, w \in F_2^n$ and $G + 0 = G$ where 0 is the zero vector in F_2^n. The orbit of G is $\{G + v : v \in F_2^n\}$, which we denote by $[G]$. This gives an equivalence relation ($G \sim G'$ if and only if G and G' are in the same orbit) and hence partitions the set of good codes.

Let $N(G+v) = |v' \in F_2^n : G+v+v' = G+v|$, i.e., the size of the stabilizer of $G+v$. We will show that $N(G+v) = N(G)$ for any $v \in F_2^n$, and in particular it follows that $N(G+v)$ does not depend on v.

If v' satisfies $G+v' = G$, then $G+v+v' = G+v'+v = G+v$; conversely, if $G+v+v' = G+v$, then $G+v'+v = G+v$, so $G+v' = G$. This shows that

$$\{v' \in F_2^n : G+v+v' = G+v\} = \{v' \in F_2^n : G+v' = G\}.$$

So $N(G + v) = N(G)$ for all $v \in F_2^n$.

Following the proof of Lemma 4 in [14], let $\mu : F_2^n \to \{0, 1\}$ be the indicator function of $\boldsymbol{S}_C^n$. Averaging over all $v \in F_2^n$ gives

$$\begin{aligned} \frac{1}{2^n} \sum_{v \in F_2^n} |(v + G) \cap \boldsymbol{S}_C^n| &= \frac{1}{2^n} \sum_{v \in F_2^n} \sum_{g \in G} \mu(v + G) \\ &= \frac{1}{2^n} \sum_{g \in G} \sum_{v \in F_2^n} \mu(v + G) \\ &= \frac{1}{2^n} \sum_{g \in G} |\boldsymbol{S}_C^n| \\ &= \frac{1}{2^n} |G| \, |\boldsymbol{S}_C^n|. \end{aligned}$$

Since $N(G+v) = N(G)$ for all v, the foregoing average can be interpreted as an average over the orbit $[G]$. Since the orbits partition the set of all (n, ϵ)-good codes, we can further average over the space of such codes of a fixed size T satisfying

$$2^{n(\mathrm{cap}(\theta)+\epsilon)} \geq T \geq 2^{n(\mathrm{cap}(\theta)-\epsilon)}.$$

Note that not all orbits of codes G of size T have the same number of codes $|[G]|$, but the average over each orbit is the same. Thus,

$$E\left[|G \cap \boldsymbol{S}_C^n|\right] = \frac{T}{2^n} \cdot |\boldsymbol{S}_C^n|$$

where the expected value is taken over *all* (n, ϵ)-good codes G of size T. Next, the expectation over all G can be found by averaging over all possible T. Finally, for sufficiently large n, the size of the constraint set approaches $2^{n\text{cap}(C)}$, e.g.,

$$2^{n(\text{cap}(C)+\epsilon)} \geq |S_C^n| \geq 2^{n(\text{cap}(C)-\epsilon)}.$$

Putting this all together, the rate of average intersection is $\text{cap}(C) + \text{cap}(\theta) - 1$. □

An interpretation for this result is that $\text{cap}(C) + \text{cap}(\theta) - 1$ is the rate of the average scheme over all schemes with independent design (and joint decoding) of the ECC and constraint code. We will give another interpretation for $\text{cap}(C) + \text{cap}(\theta) - 1$ in Section V.

By Proposition 1, the rate in (5) cannot exceed the noisy constrained capacity

$$\text{cap}(\theta) + \text{cap}(C) - 1 \leq \text{cap}(C, \theta). \tag{6}$$

The following result shows that (5) is also a lower bound for the maxentropic capacity.

Proposition 2: For a BIAWGN channel

$$\text{rate}_{\text{avgECC}}(C,\theta) = \text{cap}(\theta) + \text{cap}(C) - 1 \leq \text{cap}_{\text{maxent}}(C,\theta). \tag{7}$$

Proof: See Appendix I. □

In addition, we consider the rate of maximum intersection, which is defined by the maximum intersection over all (n, ϵ)-good codes

$$\text{rate}_{\text{maxECC}}(C, \theta) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} \log_2 \max |\boldsymbol{S}_C^n \cap \boldsymbol{S}_{\text{ECC}}^n|.$$

Clearly, the maximum intersection rate satisfies

$$\text{rate}_{\text{maxECC}}(C, \theta) \leq \text{cap}(C, \theta).$$

The following proposition shows that this is an equality, i.e., there exist good (not necessarily linear) ECCs for the unconstrained channel θ where the intersection with the constraint has rate arbitrarily close to the capacity $\text{cap}(C, \theta)$.

Proposition 3:

$$\text{rate}_{\text{maxECC}}(C, \theta) = \text{cap}(C, \theta).$$

Proof: See Appendix II. □

For Sections III–V, we define $\text{rate}_{\text{lower}}(C, \theta)$ to be the maximum of $\text{rate}_{\text{avgECC}}(C, \theta)$ and 0

$$\text{rate}_{\text{lower}}(C, \theta) := \max\{\text{cap}(\theta) + \text{cap}(C) - 1, 0\}. \tag{8}$$

III. Constraint Gain

The noisy constrained capacity $\text{cap}(C, \theta)$ is achieved when the ECC and the constrained code are jointly designed and this knowledge is exploited in decoding. This is the theoretical maximum for the rate of any code for error-free transmission on the noisy constrained channel. Meanwhile, the lower bound $\text{rate}_{\text{lower}}(C, \theta)$ gives the average rate when the ECC is chosen without knowledge of the constraint. The difference between $\text{rate}_{\text{lower}}(C, \theta)$ and $\text{cap}(C, \theta)$ then gives an estimate of the increase in capacity that is theoretically available from a system that makes use of the constraint in designing the ECC and in decoding. We call this potential improvement the *constraint gain*.

Recall that a BIAWGNC with noise variance σ^2 is represented by the channel parameter $\theta = \frac{1}{\sigma^2}$. Constraint gain can be measured as a difference in thresholds for a given rate R, where the threshold is the minimum value of the channel parameter θ such that rate R can be achieved

$$\theta_{\text{lower}}(R) = \min\left\{\theta \mid \text{rate}_{\text{lower}}(C, \theta) \geq R\right\}$$
$$\theta(R) = \min\left\{\theta \mid \text{cap}(C, \theta) \geq R\right\}.$$

Measured in decibels, the constraint gain is given by

$$\Delta(R) = 10\log_{10}(\theta(R)) - 10\log_{10}(\theta_{\text{lower}}(R)).$$

The lower bound $\text{rate}_{\text{lower}}(C, \theta)$ is computed using (8), while the noisy constrained capacity can be computed using the Arnold–Loeliger technique [1].

An estimate of (and lower bound for) constraint gain can be obtained by substituting the more easily computed maxentropic constrained capacity $\text{cap}_{\text{maxent}}(C, \theta)$ for $\text{cap}(C, \theta)$

$$\theta_{\text{maxent}}(R) = \min\left\{\theta \mid \text{cap}_{\text{maxent}}(C, \theta) \geq R\right\}$$
$$\Delta'(R) = 10\log_{10}(\theta_{\text{maxent}}(R)) - 10\log_{10}(\theta_{\text{lower}}(R)).$$

Proposition 2 guarantees that the rate of average intersection (on which the lower bound is based) is at most the maxentropic capacity, so $\Delta'(R) \geq 0$.

In Fig. 1, various capacity curves are plotted for the $(0,1)$-RLL constraint as a function of θ. One curve shows an approximation to $\text{cap}(C, \theta)$ using only Markov chains of order 2 as input distributions. (The curve using Markov chains of order 1 is very similar, so it was omitted.) Over the range of SNRs on this AWGN channel and for this constraint, the noisy maxentropic constrained capacity $\text{cap}_{\text{maxent}}(C, \theta)$ is only very slightly less than the second-order approximation for $\text{cap}(C, \theta)$. Meanwhile, the lower bound $\text{rate}_{\text{lower}}(C, \theta)$ is given by the curve $\text{cap}(\theta)$ shifted down by $1 - \text{cap}(C)$. For some rate R, the constraint gain is estimated by the horizontal distance between the curves for $\text{cap}_{\text{maxent}}(C, \theta)$ and $\text{rate}_{\text{lower}}(C, \theta)$.

To compare the estimates for different constraints in a fair way, we make the somewhat arbitrary assumption that the rate R is a fixed proportion of $\text{cap}(C)$. Table I shows the estimates for the constraint gain $\Delta'(R)$ for rates of $R = 0.9\,\text{cap}(C)$ and $R = 0.75\,\text{cap}(C)$ for various constraints over a BIAWGNC.

We see that the gain $\Delta'(R)$ generally increases as the noiseless capacity $\text{cap}(C)$ decreases, but the exact value depends on the structure of the particular constraint. For example,

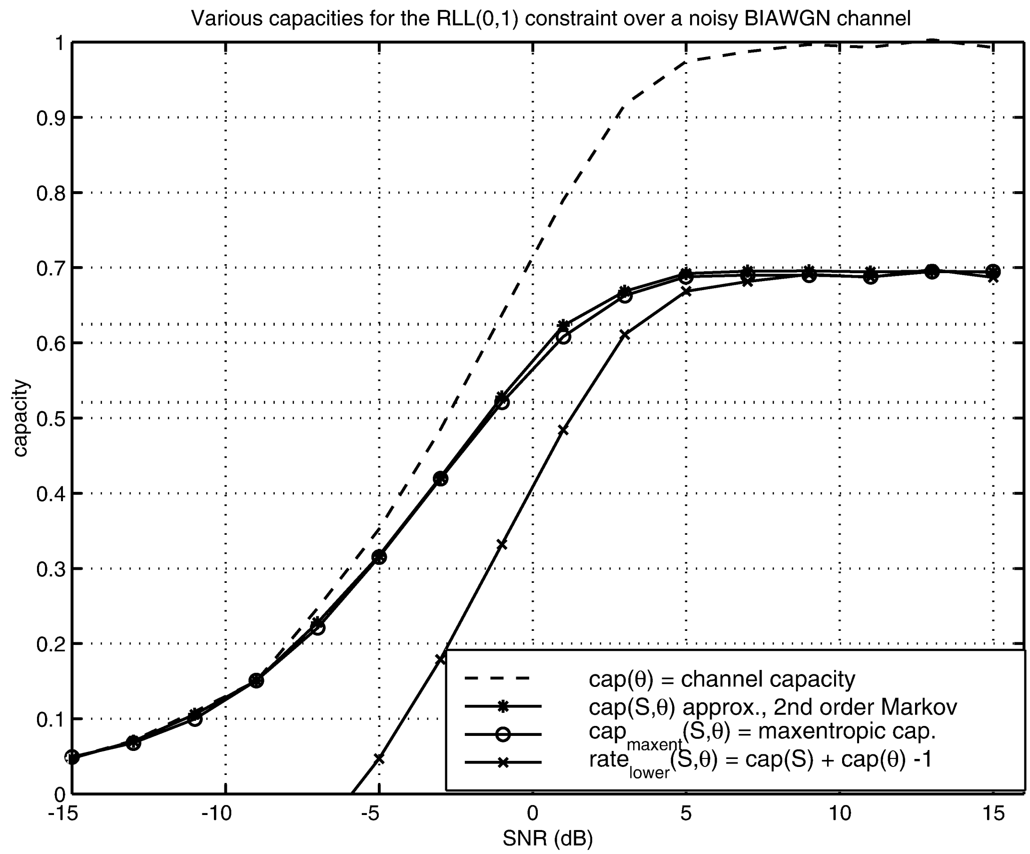

Fig. 1. Estimating the constraint gain for the $(0,1)$-RLL constraint.

TABLE I
LOWER BOUNDS ON CONSTRAINT GAIN FOR BINARY-INPUT AWGNC

constraint	cap(C)	Δ' (0.9 cap (C))	Δ' (0.75 cap (C))
$(0,3)$	0.9468	0.2 dB	0.3 dB
$(1,7)$	0.6793	1.2 dB	1.8 dB
$(0,1)$	0.6942	1.2 dB	1.5 dB
$(1,3)$	0.5515	3 dB	3.8 dB
$(2,7)$	0.5174	2.7 dB	3.4 dB

$(1,3)$-RLL has a higher noiseless capacity than $(2,7)$-RLL, yet its estimated constraint gain is larger.

These results are especially of interest in the case of optical storage, where a constraint with $d = 2$ is used in CDs and DVDs. In current systems, the error-control code is typically designed independently of the RLL constraint, and soft decoding is not used for the RLL constraint. These capacity results indicate that significant improvements in performance may be available from decoding the constraint.

IV. Noisy Constrained ISI Channels

These results extend to the case of a binary-input channel with ISI. In general, the methods for computing the noisy capacity apply to any situation where both the source and the channel have memory, such as a RLL constraint used in conjunction with a binary-input ISI channel with AWGN.

The ISI channel with AWGN can be represented by

$$Y_i = \sum_{j=0}^{\nu} h_j (-1)^{X_{i-j}} + N_i \tag{9}$$

where ν is the channel memory. We introduce a variable Z_i to represent the deterministic output of the channel before the random noise N_i has been added, and separate the channel into an ISI part and an AWGN part

$$Z_i = \sum_{j=0}^{\nu} h_j (-1)^{X_{i-j}}$$
$$Y_i = Z_i + N_i.$$

Let S' denote the set of 2^{ν} channel states that correspond to the last v bits, which is called the channel memory. A graph can be used to represent the transitions between channel states, with the deterministic output Z_i as edge labels.

If the input process X is required to satisfy a constraint defined by a graph representation on a state set S'', then we can form the fiber product graph representation [17] for the combined source/channel. The states of this combined graph are the product states $S = S' \times S''$, with the transitions labeled by Z_i. (It is often possible to choose the constraint and the channel target such that the number of overall states can be made smaller than $|S' \times S''|$.)

The capacity of the noisy constrained ISI channel is

$$\operatorname{cap}(C,H,\theta) = \lim_{n\to\infty} \frac{1}{n} \sup_{\substack{p(X_1^n)\\ X_1^n \in \boldsymbol{S}_C^n}} I\left(X_1^n; Y_\nu^n\right) \tag{10}$$

where the supremum is over all possible distributions $p\left(X_1^n\right)$ with support on the constraint. In terms of stochastic processes, (10) can be rewritten as

$$\operatorname{cap}(C,H,\theta) = \sup_{X \in S_C} I(X;Y) \tag{11}$$

TABLE II
CONSTRAINT GAIN ESTIMATES $\Delta'(R)$ FOR ISI CHANNELS AT $R = 0.75\ \mathrm{cap}(C)$

constraint	cap	AWGN	Dicode	EPR4	E²PR4
$(0,1)$	0.6942	1.5 dB	2.6 dB	0.7 dB	1.9 dB
$(0,3)$	0.9468	0.3	0.1	0.6	0.1
$(1,7)$	0.6793	1.8	1.8	3.0	4.0
$(1,\infty)$	0.6942	2.0	0.8	2.9	3.5
$(2,\infty)$	0.5515	2.9	0.3	3.0	4.6

TABLE III
CONSTRAINT GAIN ESTIMATES $\Delta'(R)$ FOR ISI CHANNELS AT $R = 0.9\ \mathrm{cap}(C)$

constraint	cap	AWGN	Dicode	EPR4	E²PR4
$(0,1)$	0.6942	1.5 dB	4.0 dB	2.7 dB	1.3 dB
$(0,3)$	0.9468	0.3	1.3	0.7	0.2
$(1,7)$	0.6793	1.8	1.3	4.2	4.7
$(1,\infty)$	0.6942	2.7	1.0	4.3	4.6
$(2,\infty)$	0.5515	3.1	0.8	1.9	1.9

where Y is the output process corresponding to the input process X. Note that $I(X;Y) = I(Z;Y)$, where Z is the deterministic output of the ISI channel with input X.

As in (4), we consider the maxentropic capacity

$$\mathrm{cap}_{\mathrm{maxent}}(C,H,\theta) = I(Z_{\max};Y_{\max}) \tag{12}$$

where the subscript "max" refers to the corresponding distribution based on the maxentropic distribution $X_{\max}$ for the constraint. Expression (12) is more easily computed than (11) since it is not necessary to search over all input processes. The capacity can be found by applying the Arnold–Loeliger method [1] to the variables Z_i, which are the labels of the transitions between the finite set of product states S_i.

It is also possible to formulate a lower bound in analogy with (8). Suppose that the ECC is designed instead for a symmetric memoryless channel (i.e., the BIAWGNC) so that it is chosen from a uniform distribution (also known as the *i.i.d. distribution*). This suboptimal choice might be made because the transmitter does not have accurate knowledge of the channel, or for simplicity in implementation. Then the ECC has rate approaching

$$\mathrm{cap}_{\mathrm{i.i.d.}}(H,\theta) = I(X_{\mathrm{i.i.d.}};Y_{\mathrm{i.i.d.}})$$

which we call the *i.i.d. capacity* of an ISI channel. Again, $Y_{\mathrm{i.i.d.}}$ is the induced output process from $X_{\mathrm{i.i.d.}}$.

As in Proposition 1, it is possible to find the average intersection between the ECC and the constraint $\boldsymbol{S}_C^n$, and then to define the corresponding lower bound as

$$\mathrm{rate}_{\mathrm{lower}}(C,H,\theta) := \max\{\mathrm{cap}_{\mathrm{i.i.d.}}(H,\theta) + \mathrm{cap}(C) - 1, 0\}.$$

Constraint gain in this context can be defined analogously as the difference between $\mathrm{cap}(C,H,\theta)$ and $\mathrm{rate}_{\mathrm{lower}}(C,H,\theta)$. A lower bound estimate for constraint gain is obtained by using $\mathrm{cap}_{\mathrm{maxent}}(C,H,\theta)$ in place of $\mathrm{cap}(C,H,\theta)$. In Tables II and III, estimates of constraint gain are shown for various input-constrained ISI channels using the difference in thresholds between $\mathrm{cap}_{\mathrm{maxent}}(C,H,\theta)$ and $\mathrm{rate}_{\mathrm{lower}}(C,H,\theta)$. The dicode channel is $h(D) = (1-D)$, the EPR4 channel is $(1-D)(1+D)^2$, the E^2PR4 channel is $(1-D)(1+D)^2$.

V. REVERSE CONCATENATION AND THE RATE OF AVERAGE INSERSECTION

The lower bound $\mathrm{rate}_{\mathrm{lower}}(C,\theta)$ in (8) is based on restricting the transmission to ECC codewords that meet the constraint and then applying an ECC decoder that does not make use of the constraint. For this scenario, the decoding is straightforward, but it is a challenge in practical implementations to create a joint ECC and constraint encoder that specifies only the codewords of an ECC that meet the constraint. Instead, it is typical to concatenate an ECC encoder and constraint encoder to create a combined encoder.

The standard method of concatenation involves encoding the user sequence with an ECC encoder (ENC ECC), followed by a near-capacity constraint encoder (ENC C) that encodes the ECC encoder output into a constrained word, as shown in Fig. 2. The problem with standard concatenation is that the output of the constrained encoder is not necessarily an ECC codeword, and the decoding of the constraint often results in error propagation.

As an alternative approach, a technique known as reverse concatenation (a.k.a. "modified concatenation") has been known in the context of magnetic storage for many years (e.g., [2], [15]), and has been analyzed in [10], [6], and [7]. The idea is to first encode the user bits using the constraint encoder, and then apply a systematic ECC encoder to produce parity bits, which are then in turn encoded by another constraint encoder. (In addition, a variation in which the parity bits are inserted into the constrained sequence as unconstrained positions is considered in [4] and [23].)

In a typical implementation of reverse concatenation, the user sequence u is first encoded by a modulation encoder (ENC $\mathrm{C}^{(1)}$) to provide a constrained sequence $w^{(1)}$. Based on $w^{(1)}$, an ECC encoder (ENC ECC) produces a sequence $v^{(1)}$ of parity bits, which do not necessarily meet the modulation constraint. The parity bits are further encoded by a second modulation encoder (ENC $\mathrm{C}^{(2)}$) before being transmitted on the channel. This single stage of reverse concatenation is shown in Fig. 3.

It is natural to extend this single stage into a cascade of reverse concatenation by taking the unconstrained parity bits $(v^{(1)})$ from one stage of reverse concatenation as the user input for the next stage of reverse concatenation. In this manner, an infinite cascade of reverse concatenation can be constructed, as shown in Fig. 4. In this section, we will show that for the case of an infinite cascade where there is no cooperation between the ECC and modulation decoders, the asymptotic capacity is $\mathrm{rate}_{\mathrm{lower}}(C,\theta)$, corresponding to the rate of average intersection.

Considering the infinite cascade in more detail, the input to the ith stage is $v^{(i-1)}$ (where $v^{(0)} = u$), and the outputs are a message sequence $w^{(i)}$ that satisfies the constraint and a parity sequence $v^{(i)}$ that does not necessarily satisfy the constraint. The user data u is used as the first input $v^{(0)}$. Let S_i denote the ratio of the message $w^{(i)}$ to the entire output $w^{(i)}v^{(i)}$ for the ith stage

$$S_i = \frac{|w^{(i)}|}{|w^{(i)}| + |v^{(i)}|}.$$

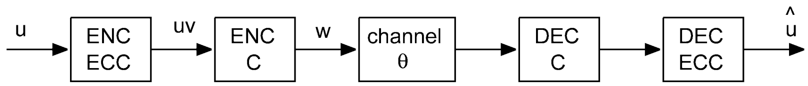

Fig. 2. Standard concatenation.

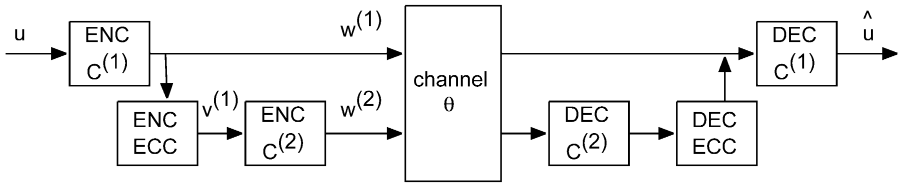

Fig. 3. A single stage of reverse concatenation.

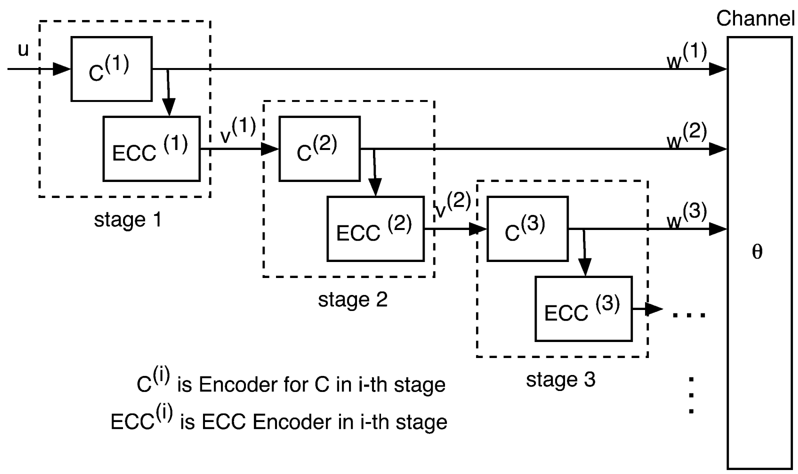

Fig. 4. Infinite cascade of reverse concatenation blocks.

Similarly, let R_i denote the overall rate of the ith stage

$$R_i = \frac{|v^{(i-1)}|}{|w^{(i)}| + |v^{(i)}|}.$$

Note that

$$|w^{(i)}| = (S_i/R_i)|v^{(i-1)}| \text{ and } |v^{(i)}| = ((1-S_i)/R_i)|v^{(i-1)}|.$$

For a user input of $|u|$ bits, the resulting number of encoded bits from this construction is given by

$$\begin{aligned}\sum_{i=1}^{\infty} |w^{(i)}| &\approx |u| \left\{ \frac{S_1}{R_1} + \frac{S_2}{R_2}\left(\frac{1-S_1}{R_1}\right) + \right.\\ &\qquad \left. \frac{S_3}{R_3}\left(\frac{1-S_2}{R_2}\right)\left(\frac{1-S_1}{R_1}\right) + \cdots \right\}\\ &= |u| \sum_{i=1}^{\infty} \frac{S_i}{R_i} \prod_{j=1}^{i-1} \frac{1-S_j}{R_j}. \end{aligned} \tag{13}$$

This is only an approximate expression because the length of the sequence $|w^{(i)}|$ must be an integer.

The rate of the infinite cascade in (13) thus approaches

$$\mathcal{R}_\infty = \left(\sum_{i=1}^{\infty} \frac{S_i}{R_i} \prod_{j=1}^{i-1} \frac{1-S_j}{R_j} \right)^{-1} \tag{14}$$

if the summation converges.

If we assume that the rates S_i and R_i are the same for all i, i.e., $S_i = S$ and $R_i = R$, then (14) reduces to a reciprocal of a geometric series that converges when $(1-S)/R < 1$, yielding the rate

$$\mathcal{R}_\infty = \frac{R-1}{S} + 1. \tag{15}$$

A diagram illustrating independent decoding of the ECC and modulation constraint within the infinite cascade is shown in Fig. 5. This is the case where there is no passing of soft information between the ECC and modulation decoders and no iterative decoding. More specifically, in the decoding of the ith stage, the decoder receives signals $y^{(i)}$ corresponding to the constrained message $w^{(i)}$, and also receives a decoded version of $\hat{v}^{(i)}$ from the $(i+1)$th stage. What is meant by independent decoding of the ECC and modulation code is that the ECC decoder, without any knowledge of the modulation constraint, decodes based on the input $(y^{(i)}, \hat{v}^{(i)})$ to produce the decoded version of the constrained message bits $\hat{w}^{(i)}$. Assuming that the ECC has been appropriately designed for the channel, the decoder output $\hat{w}^{(i)}$ is correct with high probability. Finally, a constraint decoder is used to transform the constrained message $\hat{w}^{(i)}$ to a decoded version $\hat{v}^{(i-1)}$ of the input.

In this situation of independent decoding of the ECC and modulation code, we will later argue that the ECC essentially sees a mixture channel (see Fig. 6), where $w^{(i)}$ is transmitted over a noisy channel with capacity $\text{cap}(\theta)$ and $v^{(i)}$ is transmitted over a perfect channel with capacity 1. The capacity of a mixture

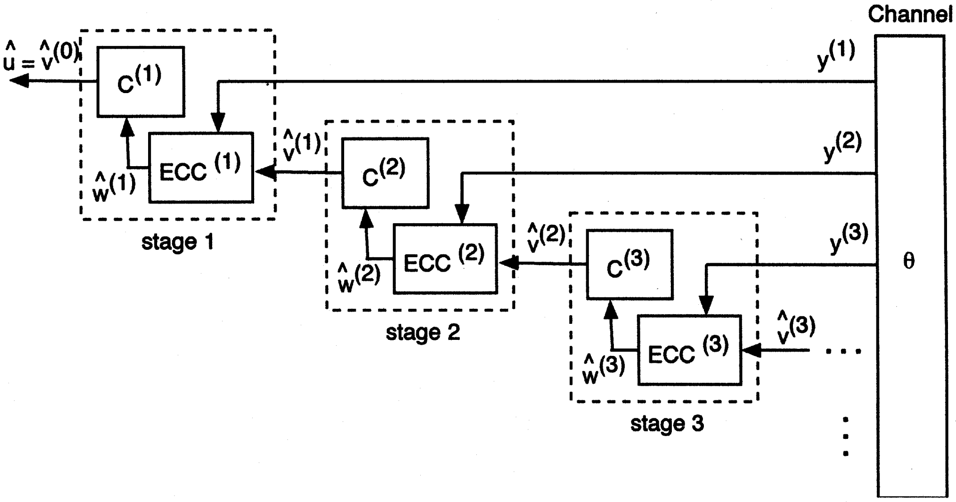

Fig. 5. Independent decoding of the infinite cascade.

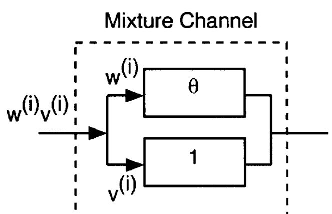

Fig. 6. Mixture channel.

channel is given by the weighted average of the capacities. The rate of the ECC S_i is limited by this mixture capacity, yielding the following upper bound on the rate S_i

$$S_i \le S_i \cdot \text{cap}(\theta) + (1 - S_i) \cdot 1$$

or equivalently

$$S_i \le 1/(2 - \text{cap}(\theta)). \tag{16}$$

On the other hand, the constraint encoder converts the unconstrained word $v^{(i-1)}$ (where $v^{(0)} = u$) to a constrained word $w^{(i)}$, so that its rate is bounded by the capacity of the constraint $\text{cap}(C)$. Thus, $|v^{(i-1)}|/|w^{(i)}| \le \text{cap}(C)$, so that R_i satisfies the bound

$$R_i \le S_i \cdot \text{cap}(C). \tag{17}$$

We claim that with sufficiently long codewords, the ratios S_i and R_i can both be made arbitrarily close to the upper bounds in (16) and (17). At the ith stage, the user input $u = v^{(i-1)}$ is encoded by the constraint encoder into $w = w^{(i)}$, which is transmitted across the channel θ. The goal of the systematic encoder is to produce a parity sequence $v = v^{(i)}$ for encoding by the next ($(i+1)$th) stage of the cascade. We give the following explicit construction for an ECC code that is suitable for the mixture channel.

Decompose w into a prefix t and suffix z (so that $w = tz$). Next consider a systematic encoder for an ECC whose input is t, and whose parity output p has the same length as z (so $|p| = |z|$). From this encoder, we will construct a systematic encoder for the mixture channel. The idea to transmit $w = tz$ across the channel θ, and then compute the XOR-ed sequence $v = z \oplus p$ for encoding by the next stage of the cascade.

At the decoder for the ith stage, $v = z \oplus p$ is assumed to be recovered perfectly from the inner stages. Let $\hat{w} = \hat{t}\hat{z}$ denote the received signal (corresponding to $w = tz$) from the channel θ. In the case of a BSC θ, let $\hat{p} = \hat{z} \oplus v = \hat{z} \oplus z \oplus p$. The noise that has been introduced in z is then passed on to p, i.e., $p_i \ne \hat{p}_i$ if and only if $z_i \ne \hat{z}_i$. Thus, the combined sequence $\hat{t}\hat{p}$ is equivalent to the result of passing tp through the BSC θ. If the ECC is good (in the sense of (n, ϵ)-good) for the channel θ, then with high probability it is possible to decode the sequence $\hat{t}\hat{p}$ to obtain t and p correctly. Taking the XOR of p and $v = z \oplus p$ yields the correct value of z. Recovering t and z yields w, which can then be decoded by a (deterministic) constraint decoder to obtain the original sequence u.

A similar result holds for the BIAWGNC. Let $\hat{w} = \hat{t}\hat{z}$ denote the received sequence from the BIAWGNC θ. It is a sequence of real-valued signals corresponding to a binary input $w = tz$ that has been modulated and corrupted with additive noise, i.e., $\hat{w}_i = (-1)^{w_i} + n_i$, where n_i is a Gaussian noise sample. On the other hand, the sequence v is received perfectly from the inner stages of the cascade. Instead of XOR-ing with the sequence v directly as in the BSC case, for the BIAWGNC it is necessary to reverse the sign of the real-valued sequence $\hat{z}$ according to v as follows: Let $\hat{p}_i = \hat{z}_i$ if $v_i = 0$, and $\hat{p}_i = -\hat{z}_i$ if $v_i = 1$. The resulting real-valued sequence $\hat{p}$ will then be equivalent to passing $p = v \oplus z$ through the channel θ. Thus, if the ECC is good for the channel θ, the decoder can with high probability recover the original t and p from the real-valued sequence $\hat{t}\hat{p}$. Again, from p and v, z can be found, and then from $w = tz$, the original sequence u can be obtained.

Regarding the existence of an error-control code that is good for the channel, the channel coding theorem says that with sufficiently large block lengths, such a code exists with rate arbitrarily close to the channel capacity $\text{cap}(\theta)$. Formally, for $\epsilon > 0$, there is some N_0 such that for all $n \ge N_0$, there exists an (n, ϵ)-good ECC for the BSC or BIAWGNC θ such that the rate $|t|/(|t| + |p|)$ is greater than $\text{cap}(\theta) - \delta$ (where $\delta = \delta(\epsilon)$ is to

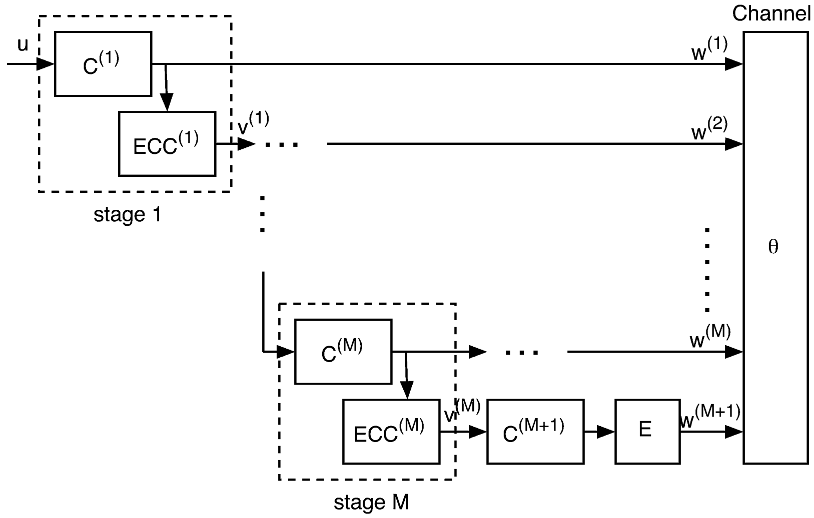

Fig. 7. M-stage truncation of the reverse concatenation cascade.

be chosen later). Moreover, it is possible to choose this code to be a linear systematic code.

Then for this construction, the ratio S_i is given by

$$\begin{aligned} S_i &= \frac{|w|}{|w|+|v|} \\ &= \frac{|t|+|p|}{|t|+2|p|} \\ &= \frac{1}{2-\frac{|t|}{|t|+|p|}} \\ &> \frac{1}{2-\mathrm{cap}(\theta)+\delta}. \end{aligned}$$

Now choose δ such that S_i is within $\epsilon/2$ of $\frac{1}{2-\mathrm{cap}(\theta)}$.

Next, the ratio $R_i = \frac{|u|}{|w|+|v|}$ is related to S_i by the rate $|u|/|w|$ of the constraint encoder. Since for sufficiently large block length, this ratio can be chosen to be arbitrarily close to the capacity of the constraint, and in particular, we choose it such that $|u|/|w| > \mathrm{cap}(C) - \epsilon/2$. Then

$$\begin{aligned} R_i &\geq \left(\mathrm{cap}(C) - \frac{\epsilon}{2}\right)\left(\frac{1}{2-\mathrm{cap}(\theta)} - \frac{\epsilon}{2}\right) \\ &\geq \frac{\mathrm{cap}(C)}{2-\mathrm{cap}(\theta)} - \frac{\epsilon}{2}\left(\mathrm{cap}(C) + \frac{1}{2-\mathrm{cap}(\theta)}\right) \\ &\geq \frac{\mathrm{cap}(C)}{2-\mathrm{cap}(\theta)} - \epsilon \end{aligned}$$

since both $\mathrm{cap}(C)$ and $\frac{1}{2-\mathrm{cap}(\theta)}$ are at most 1.

Thus, for any ϵ, it is possible to construct an encoder and decoder for this mixture channel such that the ratios S_i and R_i are within ϵ of their maximum rates in the inequalities (16) and (17), i.e., $S_i = S$ and $R_i = R$, for all i, where $S = 1/(2-\mathrm{cap}(\theta))$ and $R = S \cdot \mathrm{cap}(C)$ in (15). We will now use this to show in Proposition 4 that an infinite cascade of reverse concatenation can achieve an overall rate of

$$\mathcal{R}_\infty = \mathrm{cap}(C) + \mathrm{cap}(\theta) - 1$$

when $\mathrm{cap}(C) + \mathrm{cap}(\theta) - 1 > 0$.

To prove this result, it will be helpful to define an *M-stage truncation*, as shown in Fig. 7. An M-stage truncation can be thought of as an infinite cascade that has been truncated at the Mth stage. The output of the Mth joint encoder block $(v^{(M)})$ is fed into a combined ECC and modulation encoder E before transmission on the channel. This encoder E should have nonzero rate $R_E > 0$ and is designed for the channel to allow decoding with low probability of error.

Proposition 4: For a channel θ and constraint C, if $\mathrm{cap}(C) + \mathrm{cap}(\theta) - 1 > 0$, the maximum rate of an infinite cascade of reverse concatenation with independent decoding is given by $\mathrm{cap}(C) + \mathrm{cap}(\theta) - 1$.

In other words, for any $\epsilon > 0$, for sufficiently large number of stages M and sufficiently long user input $|u|$, there exists an M-stage truncation that has probability of error $<\epsilon$ and has rate within ϵ of $\mathrm{cap}(C) + \mathrm{cap}(\theta) - 1$.

Proof: Fix $\epsilon > 0$. The number of stages M, the allowed variation in rate δ, and the user input length $|u|$ will be chosen later based on ϵ. Typically, M and $|u|$ will be large integers, while δ will be a real number close to 0.

The resulting number of encoded bits for an M-stage truncation is given by

$$|u|\left\{\sum_{i=1}^{M-1}\frac{S_i}{R_i}\prod_{j=1}^{i-1}\frac{1-S_j}{R_j} + \left(\frac{S_M}{R_M}\prod_{j=1}^{M-1}\frac{1-S_j}{R_j}\right)\frac{1}{R_E}\right\}$$

where R_E is the rate of the combined modulation and ECC encoder E in the Mth stage. Thus, the rate $\mathcal{R}_M^E$ of the M-stage truncation as a function of S_i and R_i for $i = 1, \ldots, M$ is

$$\begin{aligned} &\mathcal{R}_M^E(S_1,\ldots,S_M,R_1,\ldots,R_M) \\ &= \left\{\sum_{i=1}^{M-1}\frac{S_i}{R_i}\prod_{j=1}^{i-1}\frac{1-S_j}{R_j} + \left(\frac{S_M}{R_M}\prod_{j=1}^{M-1}\frac{1-S_j}{R_j}\right)\frac{1}{R_E}\right\}^{-1}. \end{aligned} \tag{18}$$

Let $\mathcal{R}_M(S_1,\ldots,S_M,R_1,\ldots,R_M)$ denote the rate of the M-stage truncation without the final ECC/modulation encoder E in the Mth stage

$$\mathcal{R}_M(S_1,\ldots,S_M,R_1,\ldots,R_M) = \left\{\sum_{i=1}^{M}\frac{S_i}{R_i}\prod_{j=1}^{i-1}\frac{1-S_j}{R_j}\right\}^{-1}. \tag{19}$$

For simplicity, the expressions in (18) and (19) will be represented by $\mathcal{R}_M^E(S_i, R_i)$ and $\mathcal{R}_M(S_i, R_i)$, respectively. In addition, for the case where the rates take their maximum allowed values according to (16) and (17), i.e., $S = 1/(2 - \text{cap}(\theta))$ and $R = S \cdot \text{cap}(C)$, let the expressions in (18) and (19) be represented by $\mathcal{R}_M^E(S, R)$ and $\mathcal{R}_M(S, R)$, respectively.

Consider the following three observations.

i) We can choose an M_1 large enough such that for all $M \geq M_1$
$$|\mathcal{R}_M(S,R) - \mathcal{R}_\infty| \leq \frac{\epsilon}{3}$$
since $\mathcal{R}_M(S,R)$ is a partial sum of $\mathcal{R}_\infty$.

ii) With $\left|\frac{1-S}{R}\right| < 1$ (this is equivalent to $\text{cap}(C) + \text{cap}(\theta) - 1 > 0$), we can fix an encoder E with rate $R_E > 0$ and probability of error arbitrarily small ($< \epsilon/2$), and choose an M_2 such that for all $M \geq M_2$, R_E satisfies
$$\left(\frac{1-S}{R}\right)^{M-1} \cdot \frac{1}{R_E} \ll \sum_{i=1}^{M-1} \left(\frac{1-S}{R}\right)^i.$$
Thus, for $M \geq M_2$
$$\left|\mathcal{R}_M^E(S,R) - \mathcal{R}_M(S,R)\right| \leq \frac{\epsilon}{3}.$$

iii) Since $\mathcal{R}_M^E(S_i, R_i)$ is continuous in variables S_i and R_i about S and R, respectively, we can choose δ so small that if $|S - S_i| < \delta$ and $|R - R_i| < \delta$, then
$$\left|\mathcal{R}_M^E(S_i,R_i) - \mathcal{R}_M^E(S,R)\right| \leq \frac{\epsilon}{3}.$$

Thus, for $M = \max(M_1, M_2)$, we can realize the cascade with block codes of sufficiently long block lengths such that S_i and R_i are within δ of their ideal values. The deviation of the rate $\mathcal{R}_M^E(S_i, R_i)$ from $\mathcal{R}_\infty$ is
$$\begin{aligned}
&\left|\mathcal{R}_M^E(S_i,R_i) - \mathcal{R}_\infty\right| \\
&\quad \leq \left|\mathcal{R}_M^E(S_i,R_i) - \mathcal{R}_M^E(S,R)\right| + \left|\mathcal{R}_M^E(S,R) - \mathcal{R}_M(S,R)\right| \\
&\qquad + |\mathcal{R}_M(S,R) - \mathcal{R}_\infty| \\
&\quad \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\
&\quad \leq \epsilon.
\end{aligned}$$

Also, we may assume that each stage of the M-stage truncation is decodable with probability of error at most $\epsilon/(2M)$ provided the parity is known perfectly, and that the probability of error for E is $< \epsilon/2$. So, with probability $> 1 - \epsilon$, we can decode $w^{(M+1)}$ correctly, and hence $v^{(M)}$. Each stage of the truncated cascade with received $w^{(i)}, i = 1, \ldots, M$ can thus be decoded correctly when we backtrack up the cascade using the correctly decoded $v^{(M)}$ to obtain $w^{(M)}$, then $v^{(M-1)}$, etc., until we get u. □

Note that this proposition gives an additional justification for the use of
$$\text{rate}_{\text{lower}} = (C, \theta) \max\{\text{cap}(C) + \text{cap}(\theta) - 1, 0\}$$
as the rate that can be achieved when a code designed for the channel θ is used for a noisy constrained channel.

On the other hand, if we allow a joint decoder for both the ECC and the constraint, it is possible to increase the capacity of the system. The reason is that with joint decoding, the decoder is allowed to use all the information about the ECC and constraint in decoding. If, in addition, arbitrary design is allowed in terms of joint encoding of modulation and ECC, then the capacity is $\text{cap}(C, \theta)$ by definition.

We also point out that the infinite cascade is a scheme that involves independent decoding of the ECC and modulation constraint. Yet Proposition 4 establishes that the rate of such a scheme is the same as that of the average scheme over all schemes with independent design (and joint decoding) of the ECC and constraint code.

VI. Conclusion

We summarize the capacity relations below as follows.
$$\begin{array}{ccc}
\text{rate}_{\text{maxECC}}(C,\theta) & & \text{rate}_{\text{avgECC}}(C,\theta) \\
\| & & \| \\
\text{cap}(C,\theta) \geq \text{cap}_{\text{maxent}} & (C,\theta) \geq & \text{cap}(C) + \text{cap}(\theta) - 1.
\end{array}$$

We introduced the notion of constraint gain as the potential improvement over a system in which the ECC is designed and decoded independently of the modulation constraint, as measured by the rate of average intersection. We defined constraint gain to be the difference between $\text{cap}(C, \theta)$ and $\text{rate}_{\text{lower}}(C, \theta)$. An accurate and easily computable lower bound estimate of constraint gain is given by comparing $\text{cap}_{\text{maxent}}(C, \theta)$ and $\text{rate}_{\text{lower}}(C, \theta)$. The constraint gain indicates the potential improvement in performance that can be obtained by making use of the constraint in the decoding. We also showed that an infinite cascade of reverse concatenation with independent decoding of constraint and ECC yields a capacity of $\text{rate}_{\text{lower}}(C, \theta)$, giving another interpretation.

Appendix I
Lower Bound

According to (6), $\text{rate}_{\text{lower}}(C, \theta) = \text{cap}(C) + \text{cap}(\theta) - 1$ is a lower bound for the noisy constrained capacity $\text{cap}(C, \theta)$. In this appendix, we prove Proposition 2, which claims that in the case of the BSC and BIAWGNC, it is actually a lower bound on the noisy maxentropic constrained capacity $\text{cap}_{\text{maxent}}(C, \theta)$.

Consider a binary-input channel that satisfies the following properties:

- memoryless, i.e.,
$$p(y_1 \ldots y_n \mid x_1 \ldots x_n) = p(y_1 \mid x_1) \cdots p(y_n \mid x_n)$$
(in the continuous case, this formula should be interpreted as a product of densities);
- symmetric with respect to inputs, i.e., $p(\cdot \mid x)$ is the same for all x;
- symmetric across zero, i.e., $p(y \mid x) = p(-y \mid x)$.

Let $X_{\text{i.i.d.}}$ be the uniform independent and identically distributed (i.i.d.) binary source, and $Y_{\text{i.i.d.}}$ be the corresponding output process. Let X be an arbitrary input process, and Y be the resulting output process. By symmetry, $H(Y|X)$ does not depend on X, and so in particular, it agrees with $H(Y_{\text{i.i.d.}}|X_{\text{i.i.d.}})$. Also, given our channel assumptions, the uniform input distribution $X_{\text{i.i.d.}}$ achieves the noisy channel capacity, i.e., $I(X_{\text{i.i.d.}}; Y_{\text{i.i.d.}}) = \text{cap}(\theta)$. Now, observe that the following four statements are equivalent:
$$I(X;Y) \geq H(X) + I(X_{\text{i.i.d.}}; Y_{\text{i.i.d.}}) - 1 \tag{20}$$

$$H(Y)-H(Y \mid X) \geq H(X)+H(Y_{\text{i.i.d.}})-H(Y_{\text{i.i.d.}}|X_{\text{i.i.d.}})-1$$
$$H(Y) \geq H(X)+H(Y_{\text{i.i.d.}})-H(X_{\text{i.i.d.}})$$
$$H(Y)-H(X) \geq H(Y_{\text{i.i.d.}})-H(X_{\text{i.i.d.}}). \tag{21}$$

In particular, if $X = X_{\max}$, the maxentropic process for a constraint C, then (20) becomes our desired lower bound

$$\text{cap}_{\text{maxent}}(C,\theta) = I(X_{\max}; Y_{\max}) \geq \text{cap}(C) + \text{cap}(\theta) - 1.$$

So, it suffices to prove (21) for $X = X_{\max}$. In fact, we show that (21) holds for any unifilar Markov source[3] X. Let S be the Markov chain on state sequences of the constraint graph that generates X. The output process Y is a function of a Markov chain (namely, a function of the independent joining of S and the memoryless noise process). It then follows from [3, Theorem 4.4.1] that

$$H(Y) \geq H(Y_1|S_0).$$

Since X is unifilar Markov, we also have $H(X) = H(S) = H(S_1|S_0)$, and so

$$H(Y) - H(X) \geq H(Y_1|S_0) - H(S_1|S_0).$$

Now, the right-hand side of this inequality is a weighted average of

$$H(Y_1|S_0 = s) - H(S_1|S_0 = s)$$

over all states s of the Markov chain, with weights equal to the stationary probability distribution. So, to prove (21) it suffices to show that for each state s

$$H(Y_1|S_0 = s) - H(S_1|S_0 = s) \geq H(Y_{\text{i.i.d.}}) - H(X_{\text{i.i.d.}}). \tag{22}$$

For the BSC with crossover probability ϵ, inequality (22) reduces to showing

$$H(Y_1|S_0 = s) \geq H(S_1|S_0 = s)$$

since $H(X_{\text{i.i.d.}}) = H(Y_{\text{i.i.d.}}) = 1$. Note that the random variable $(S_1|S_0 = s)$ is discretely distributed as $(q, 1-q)$ (where q may be dependent on the state s), since the Markov chain has at most two outgoing edges for each state. On the other hand, the random variable $(Y_1|S_0 = s)$ is distributed discretely as

$$((1-\epsilon)q + \epsilon(1-q), (1-\epsilon)(1-q) + \epsilon q)$$

which is closer to being uniform $\left(\frac{1}{2}, \frac{1}{2}\right)$ since it is a weighted average of q and $1-q$. Hence, $H(Y_1|S_0 = s) \geq H(S_1|S_0 = s)$, thus proving the desired lower bound

$$\text{cap}_{\text{maxent}}(C,\theta) \geq \text{cap}(C) + \text{cap}(\theta) - 1.$$

In the case of the BIAWGNC with $(S_1|S_0 = s)$ distributed discretely as $(q, 1-q)$, the output random variable $(Y_1|S_0 = s)$ is continuously distributed with probability density function

$$f(y,q) = q \cdot p(y|x=-1) + (1-q) \cdot p(y|x=1)$$

[3]A unifilar Markov source is a process on label sequences induced by a constraint graph such that: 1) the labeling is deterministic and 2) the graph is endowed with transition probabilities to form a first-order Markov chain.

where $p(y|x)$ is simply a Gaussian distribution. The differential entropy of $\{Y_1|S_0 = s\}$ is

$$h(Y_1|S_0 = s) = -\int f(y,q) \log f(y,q) dy$$

which is a function of q, so we denote $h(Y_1|S_0 = s)$ by $k(q)$. The inequality (22) is equivalent to the statement that the function

$$F(q) \equiv k(q) - H(q, 1-q) \tag{23}$$

is minimized at $q = 1/2$. (The second term, $H(.,.)$, is the binary entropy function.) It is straightforward to verify that $F'(1/2) = 0$ and that $F(q)$ is strictly convex. A plot of $F(q)$ versus q is shown in Fig. 8 with $\sigma = 0.8$.

APPENDIX II
MAXIMUM INTERSECTION RATE

We prove Proposition 3, which asserts that the maximum intersection rate $\text{rate}_{\text{maxECC}}(C,\theta)$ is equal to the noisy constrained capacity $\text{cap}(C,\theta)$.

First, the inequality

$$\text{rate}_{\text{maxECC}}(C,\theta) \leq \text{cap}(C,\theta)$$

is obvious. To show equality, we construct a sequence of channel-capacity-achieving block codes whose intersection with the constraint approaches the noisy constrained capacity. The idea is to take a code T which satisfies the constraint C whose rate nearly achieves $\text{cap}(C,\theta)$ and which can be decoded with probability of error $< \epsilon$, and combine it with an (n,ϵ)-good code U for the channel θ to obtain a code $K = T \cup U$. The code K will also be (n,ϵ)-good for the channel θ, and the intersection rate of K can be made arbitrarily close to the noisy constrained capacity.

To allow easy decoding of codewords from K, we form a new code K' by attaching a prefix of length m to each codeword. Codewords from the subset T will have a fixed prefix r that satisfies the constraint S_C, while codewords from U will have prefix $\bar{r}$ (the bitwise complement of r). This helps to distinguish between codewords derived from T and from U and thus bound the error probability during decoding. Denote the extended T code by T', and the extended U code by U'. A prefix r can be selected if the constraint graph is irreducible. The length m will be determined from channel statistics.

We now show for sufficiently large n, the set $K' \equiv T' \cup U'$ is an $(m+n, 2\epsilon)$-good code for the unconstrained channel, i.e., it has rate close to $\text{cap}(\theta)$, and is decodable with small error probability. The rate of the code K' is given by

$$\begin{aligned}\frac{1}{n+m}\log|K'| &= \frac{1}{n+m}\log(|T'| + |U'|) \\ &\geq \frac{1}{n+m}\log(|U'|) \\ &= \frac{1}{n+m}\log|U| \\ &> \text{cap}(\theta) - (2\epsilon)\end{aligned}$$

for sufficiently large n, thus obtaining

$$\text{cap}(\theta) - \frac{1}{n+m}\log|K'| < 2\epsilon.$$

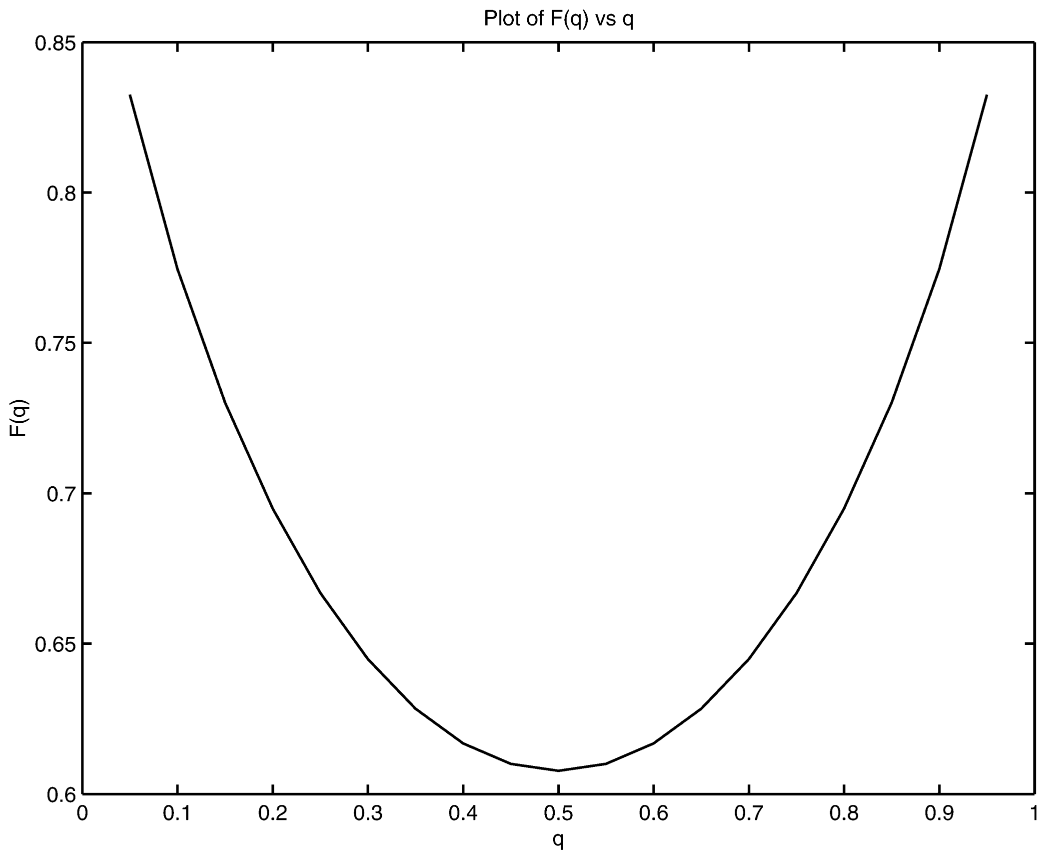

Fig. 8. Plot of $F(q)$ versus q.

Next, we will bound the probability of error when code K' is used over the BSC channel with crossover probability p (with $p < \frac{1}{2}$). It is straightforward to extend this to the case of the BIAWGNC when hard-decision decoding is employed for the prefix. When a codeword $\tilde{y}$ is received, the decoder for K' first decodes the first m bits to one of the two possible prefixes using nearest neighbor decoding. If the decoded prefix is r, the decoder for T is subsequently used in decoding the last n bits of $\tilde{y}$; otherwise, the decoder for U is used.

Suppose a codeword y from T' is sent and the received word is $\tilde{y}$. The possible error events fall into two cases: a) the decoded prefix is r, and $\tilde{y}$ is decoded to a codeword in T' but not y, and b) the decoded prefix is $\bar{w}$, and $\tilde{y}$ is decoded to a codeword in U'. In case a), if we denote the probability of the event by

$$P_{T',T'} = P(\tilde{y} \in T', \tilde{y} \neq y \mid y \in T')$$

then $P_{T',T'} \leq 2\epsilon$; in case b), the probability of error

$$P_{U',T'} = P(\tilde{y} \in U', \tilde{y} \neq y \mid y \in T')$$

is just the probability that more than half of the prefix bits have been flipped by the channel. By the Chernoff bound

$$\begin{aligned} P_{U',T'} &= P\left(w \text{ has} \geq \frac{m}{2} \text{ errors}\right) \\ &\leq \left(\frac{e^{\frac{1}{2p}-1}}{\frac{1}{2p}^{\frac{1}{2p}}} \right)^{mp} \\ &= \left(2pe^{(1-2p)}\right)^{\frac{m}{2}}. \end{aligned}$$

This can be made arbitrarily small with large enough m since $2pe^{1-2p} < 1$. (The latter can be shown using the well-known inequality $e^x < \frac{1}{1-x}$ for $x < 1$ by letting $x = 1 - 2p$.) Thus,

$$P_{U',T'} < (2pe^{(1-2p)})^{\frac{m}{2}} < 2\epsilon.$$

If y is sent from U', then by symmetry, if $P_{U',U'}$ is the probability that $\tilde{y}$ is decoded to a codeword in U' different from the one sent, then $P_{U',U'} \leq 2\epsilon$; also, if $P_{T',U'}$ denotes the probability that $\tilde{y}$ is decoded to a codeword in T' despite originating from U', then $P_{T',U'} < 2\epsilon$. This bounds the length of the required prefix to

$$\frac{2\log(2\epsilon)}{\log(2p) + 1 - 2p} < m$$

which is achievable for n large enough. Thus, the maximum block error probability $P_e(K')$ when using code K' is upper-bounded by

$$P_e(K') \leq \max\{P_{U',T'}, P_{T',T'}, P_{T',U'}, P_{U',U'}\} \leq 2\epsilon.$$

Hence, K' is an $(n+m, 2\epsilon)$-good code for the unconstrained channel.

Finally, $T' \subset K' \cap S_C^{m+n}$. This proves that K' is an $(m+n, 2\epsilon)$-good code for the unconstrained channel, whose intersection with the constraint yields a code with rate arbitrarily close to the noisy constrained capacity. Hence,

$$\text{rate}_{\text{maxECC}}(C, \theta) = \text{cap}(C, \theta).$$

ACKNOWLEDGMENT

The authors would like to acknowledge Panu Chaichanavong, Alek Kavčić, Tom Richardson, and Bruce Wilson for very helpful discussions. We also wish to thank John Gill, Jorge Campello, and Richard New for helpful comments at the 2002 IBM–Stanford Coding Seminar where parts of this work were first presented.

REFERENCES

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

[2] W. G. Bliss, "Circuitry for performing error correction calculations on baseband encoded data to eliminate error propagation," *IBM Tech. Discl. Bull.*, vol. 23, pp. 4633–4634, 1981.

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

[4] J. C. de Souza, B. H. Marcus, R. New, and B. A. Wilson, "Constrained systems with unconstrained positions," *IEEE Trans. Inform. Theory*, vol. 48, pp. 866–879, Apr. 2002.

[5] J. L. Fan, *Constrained Coding and Soft Iterative Decoding*. Norwell, MA: Kluwer Academic, 2001.

[6] J. L. Fan and A. R. Calderbank, "A modified concatenated coding scheme, with applications to magnetic storage," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1565–1574, July 1998.

[7] J. Fan and J. Cioffi, "Constrained coding techniques for soft iterative decoders," in *Proc. GLOBECOM*, Rio de Janeiro, Brazil, Dec. 5–9, 1999, pp. 723–727.

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

[9] C. D. Heegard, A. Duel-Hallen, and R. Krishnamoorthy, "On the capacity of the noisy runlength channel," *IEEE Trans. Inform. Theory*, vol. 37, pp. 712–720, May 1991.

[10] K. A. S. Immink, "A practical method for approaching the channel capacity of constrained channels," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1389–1399, Sept. 1997.

[11] ——, *Codes for Mass Data Storage Systems*. Amsterdam, The Netherlands: Shannon Foundation Publishers, 1999.

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

[13] A. Kavčić, "On the capacity of Markov sources over noisy channels," in *Proc. IEEE GLOBECOM*, San Antonio, TX, 2001.

[14] H. Loeliger, "Averaging bounds for lattices and linear codes," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1767–1773, Nov. 1997.

[15] M. Mansuripur, "Enumerative modulation coding with arbitrary constraints and post-modulation error correction coding and data storage systems," *Proc. SPIE*, pp. 72–86, 1991.

[16] B. H. Marcus, K. Petersen, and S. Williams, "Transmission rates and factors of Markov chains," *Contemp. Math.*, pp. 279–292, 1984.

[17] B. H. Marcus, R. M. Roth, and P. H. Siegel, *Handbook of Coding Theory*. Amsterdam, The Netherlands: Elsevier, 1998, ch. 20.

[18] B. H. Marcus, P. H. Siegel, and J. K. Wolf, "Finite-state modulation codes for data storage," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 5–37, 1992.

[19] A. Patapoutian and P. V. Kumar, "The (d, k) subcode of a linear block code," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1375–1382, July 1992.

[20] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rates of finite state isi channels," in *Proc. IEEE GLOBECOM*, San Antonio, TX, 2001.

[21] S. A. Raghavan, J. K. Wolf, and L. B. Milstein, "On the cutoff rate of a discrete memoryless channel with (d, k)-constrained input sequences," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 233–241, Jan. 1992.

[22] C. E. Shannon, "A mathematical theory of communications," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[23] A. J. van Wijngaarden and K. A. S. Immink, "Maximum runlength-limited codes with error control capabilities," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 602–611, Apr. 2001.

[24] P. Vontobel and D. Arnold, "An upper bound on the capacity of channels with memory and constrained input," in *Proc. IEEE Information Theory Workshop*, Cairns, Australia, Sept. 2001.

[25] E. Zehavi and J. K. Wolf, "On runlength codes," *IEEE Trans. Inform. Theory*, vol. 34, pp. 45–54, Jan. 1988.