An Efficient Feedback Coding Scheme With Low Error Probability for Discrete Memoryless Channels

Cheuk Ting Li, *Student Member, IEEE*, and Abbas El Gamal, *Fellow, IEEE*

***Abstract*— Existing fixed-length feedback communication schemes are either specialized to particular channels (Schalkwijk–Kailath, Horstein), or apply to general channels but either have high coding complexity (block feedback schemes) or are difficult to analyze (posterior matching). This paper introduces a new fixed-length feedback coding scheme which achieves the capacity for all discrete memoryless channels, has an error exponent that approaches the sphere packing bound as the rate approaches the capacity, and has $O(n \log n)$ coding complexity. These benefits are achieved by judiciously combining features from previous schemes with new randomization technique and encoding/decoding rule. These new features make the analysis of the error probability for the new scheme easier than for posterior matching.**

***Index Terms*— Feedback, discrete memoryless channel, error exponent.**

I. Introduction

SHANNON showed that feedback does not increase the capacity of memoryless point-to-point channels [1]. Feedback, however, has many benefits, including simplifying coding and improving reliability. Early examples of feedback coding schemes that demonstrate these benefits include the Horstein [2], Zigangirov [3], and Burnashev [4] schemes for the binary symmetric channel; and the Schalkwijk–Kailath scheme for the Gaussian channel [5], [6]. Schalkwijk and Kailath showed that the error probability for their scheme decays doubly exponentially in the block length. It is known, however, that the error exponent for symmetric discrete memoryless channels with feedback cannot exceed the sphere packing bound [7]. Nevertheless, the schemes in [3] and [4] can attain better error exponents than the best known achievable error exponent without feedback. D'yachkov [8] proposed a general scheme for any discrete memoryless channel. The coding complexity for his scheme, however, appears to be very high.

In addition to the traditional fixed-length setting in which the number of channel uses is predetermined before transmission commences, there has been work on variable-length schemes in which transmission continues until the error probability is lower than a prescribed target. The optimal error exponent for this setting was given explicitly by Burnashev [9]. Recently, Shayevitz and Feder [10]–[12] introduced the posterior matching scheme, which unifies and extends the Schalkwijk–Kailath and the Horstein schemes to general memoryless channels. While they were able to show that the scheme achieves the capacity for most of these channels in the variable-length setting, their analysis of the error probability provides a lower bound that is applicable only for low rates. A more general analysis of error probability for variable-length schemes, including posterior matching, is given in a recent paper by Naghshvar *et al.* [13]. Note that our focus here is only on fixed-length coding schemes for which the optimal error exponent is not known in general.

In this paper, which is a more detailed version of our recent conference paper [14], we propose a new fixed-length feedback coding scheme for memoryless channels, which (i) achieves the capacity for all discrete memoryless channels (DMCs), (ii) achieves an error exponent that approaches the sphere packing bound for high rates (up to $O((I(X;Y) - R)^3)$), and (iii) has coding complexity of only $O(n \log n)$ for discrete memoryless channels. Our scheme is motivated by the posterior matching scheme. However, unlike posterior matching, we assume a discrete message space, e.g., as in the Burnashev scheme, apply a random cyclic shift to the message points in each transmission, and use a *maximal information gain coding* rule instead of the actual posterior probability to simplify the analysis of the probability of error. This simplicity of analysis, however, does not come at the expense of increased coding complexity relative to posterior matching.

The rest of the paper is organized as follows. In the next section, we describe our feedback coding scheme and explain in detail how it differs from posterior matching. In Section III, we show that our scheme achieves the capacity of any DMC, establish a lower bound on its error exponent, and compare this bound to the sphere packing bound and bounds for other schemes. In Section IV, we discuss the scheme's coding complexity. Details of the coding algorithm and its complexity analysis are given in [15].

Remark 1: Throughout this paper, we use nats instead of bits and ln instead of log to avoid adding normalization constants. We denote the cumulative distribution function (cdf), the probability mass function (pmf), and the probability density function (pdf) for a random variable X by F_X, p_X, and f_X, respectively. We denote the set of integers $\{a, a+1, \ldots, b\}$ as $[a:b]$. The uniform distribution over $[0, 1]$

Manuscript received November 1, 2013; revised September 11, 2014; accepted April 7, 2015. Date of publication April 30, 2015; date of current version May 15, 2015. This work was supported by the Air Force under Grant FA9550-10-1-0124. C. T. Li was supported by the Hong Kong Alumni Stanford Graduate Fellowship. This paper was presented at the 2014 IEEE International Symposium on Information Theory.

The authors are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: ctli@stanford.edu; abbas@ee.stanford.edu).

Communicated by Y. Liang, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2015.2428234

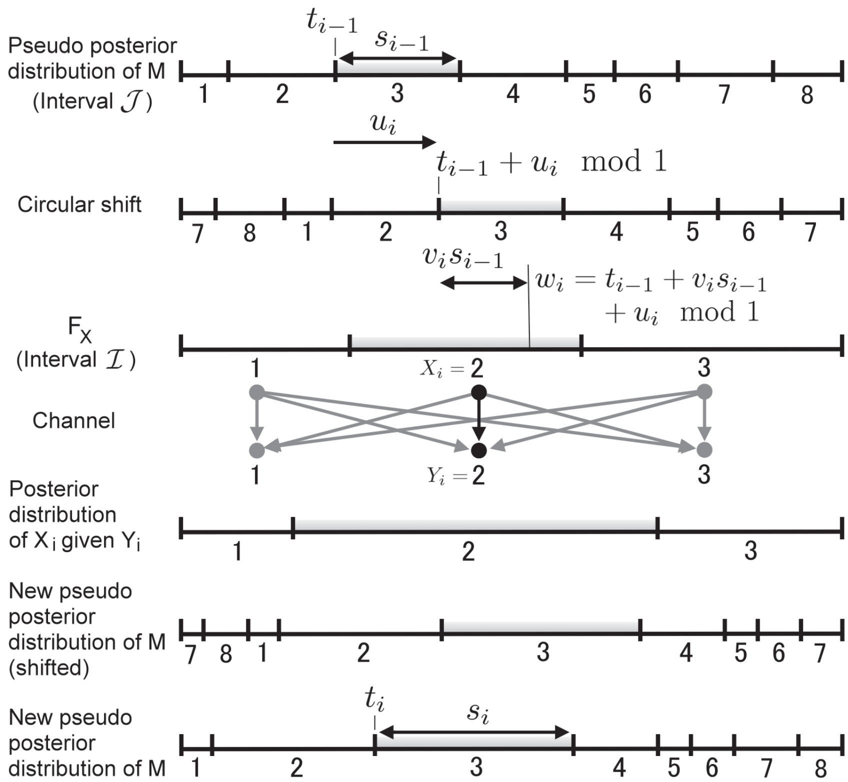

Fig. 1. Illustration of the new feedback scheme for a DMC with input and output alphabet $\{1, 2, 3\}$. The message $M = 3$ is transmitted. At time i, symbol $X_i = 2$ is transmitted and symbol $Y_i = 2$ is received.

is denoted by U[0, 1]. The fractional part of x is written as $x \bmod 1$.

II. New Feedback Coding Scheme

Our scheme is motivated and is most similar to the posterior matching scheme [12]. Hence we begin with a brief description of posterior matching and its limitations, which have led to the development of our scheme.

Posterior matching is a recursive coding scheme that achieves the capacity of memoryless channels. Consider a memoryless channel $F_{Y|X}(y|x)$ with causal noiseless feedback, i.e., the transmitted symbol X_i at time i is a function of the message and past received symbols Y^{i-1}. Fix a distribution F_X on the input symbols. The message is represented by a real number $\Theta \sim \mathrm{U}[0, 1]$. The transmitted symbol at time i is $X_i = F_X^{-1}(W_i)$, $W_i = F_{\Theta|Y^{i-1}}(\Theta|Y^{i-1})$, where $F_{\Theta|Y^{i-1}}$, the posterior cdf of Θ given the received symbols Y^{i-1}, is described recursively by

$$F_{\Theta|Y^0}(\theta) = \theta,$$
$$F_{\Theta|Y^i}(\theta|y^i) = F_{W|Y}(F_{\Theta|Y^{i-1}}(\theta|y^{i-1}) \mid y_i).$$

Here $F_{W|Y}$ is the cdf of W conditioned on Y assuming $W \sim \mathrm{U}[0, 1]$ and $X = F_X^{-1}(W)$. If Y is discrete, let

$$p_{Y|W}(y|w) = p_{Y|X}(y|F_X^{-1}(w)).$$

Then,

$$F_{W|Y}(w|y) = \frac{\int_0^w p_{Y|W}\left(y|w'\right) dw'}{\int_0^1 p_{Y|W}\left(y|w'\right) dw'}. \tag{1}$$

The expression for continuous Y can be given similarly.

Note that the posterior cdf $F_{\Theta|Y^i}$, which can be regarded as the state of the transmission, forms a Markov chain. To analyze the error probability, we can study the transition of this Markov chain. However, the posterior cdf is a complicated object. The analysis can be greatly simplified if a simpler object (e.g., the posterior probability of the transmitted message) can be used instead. As far as we know, this is not feasible due to the asymmetry of the scheme in Θ, in the sense that the behavior of the transition of $F_{\Theta|Y^i}$ depends on the transmitted value of Θ. Indeed, Shayevitz and Feder [12] needed to use iterated function system to study the transition of the entire posterior distribution, giving a rather complicated analysis of the error probability of posterior matching that is applicable only for rates below a certain threshold. Furthermore, this asymmetry results in messages having different error probabilities, which makes the maximal probability of error for the scheme worse than its average.

Our feedback coding scheme eliminates the aforementioned asymmetry of posterior matching resulting in all messages having the same error probability. As a result, we are able to greatly simplify the analysis of the error probability and obtain a bound on the error exponent for *all* rates.

Again consider a memoryless channel $F_{Y|X}(y|x)$ with causal noiseless feedback. We describe our scheme with the aid of Figure 1. We assume that the message M is uniformly distributed over $[1 : e^{nR}]$ and represent message $m \in [1 : e^{nR}]$ by the subinterval $[(m-1)e^{-nR}, me^{-nR}]$ in $[0, 1]$ (if the messages are not equally likely the subinterval length would be equal to the probability of the message). Fix the cdf $F_X(x)$ of the input symbol X (which may be the capacity achieving distribution for the channel), and partition the unit interval $\mathcal{I}$ according to this distribution. The symbol to be transmitted at time i is determined as follows. The decoder, knowing Y^{i-1}, partitions another unit interval $\mathcal{J}$ according to the pseudo posterior probability distribution of M given Y^{i-1} (the details of computing this distribution are described later). The encoder, which has Y^{i-1} via the feedback, also knows the partition of $\mathcal{J}$. We denote the location of the left edge of the subinterval corresponding to message m by $t_{i-1}(m, y^{i-1}, u^{i-1})$ (or $t_{i-1}(m)$ in short) and its length by $s_{i-1}(m, y^{i-1}, u^{i-1})$ (or $s_{i-1}(m)$ in short). All subintervals are *cyclically shifted* by an amount $U_i \sim \mathrm{U}[0,1]$, which is generated independently for each i and is known to both the encoder and the decoder. In practice, U_i can be generated using a random seed communicated to both the encoder and the decoder via the forward or feedback channel.

A point w_i is then selected in the subinterval corresponding to the transmitted message m according to $w_i = \big(v_i \cdot s_{i-1}(m) + t_{i-1}(m)\big) + u_i \bmod 1$, where $v_i \in [0,1]$ is selected using a greedy rule to be described later. The symbol to be transmitted at time i is the one corresponding to the subinterval in $\mathcal{I}$ which contains w_i. At the end of communication, the decoder outputs the message m corresponding to the subinterval with the greatest length $s_n(m)$.

We are now ready to formally describe our scheme. At time $i \in [1:n]$, the encoder transmits

$$X_i = F_X^{-1}(W_i), \quad W_i = w_i(M, Y^{i-1}, U^i, V_i),$$

where

$$\begin{aligned} w_i(m, y^{i-1}, u^i, v_i) &= v_i(m, y^{i-1}, u^i) \cdot s_{i-1}(m, y^{i-1}, u^{i-1}) \\ &\quad + t_{i-1}(m, y^{i-1}, u^{i-1}) + u_i \bmod 1, \\ s_0(m) &= e^{-nR}, \\ t_0(m) &= (m-1)\, e^{-nR}, \\ s_i(m, y^i, u^i) &= \int_{[t_{i-1}(m), t_{i-1}(m)+s_{i-1}(m)]+u_i \bmod 1} dF_{W|Y}(w|y_i), \\ t_i(m, y^i, u^i) &= \sum_{m' < m} s_i(m', y^i, u^i), \end{aligned} \tag{2}$$

where $F_{W|Y}$ is given in (1). Note that in the above integral we used the notation $[t, t+s] + u \bmod 1$ to mean the set $\{x + u \bmod 1 : x \in [t, t+s]\}$.

Assuming message m is transmitted, the encoder selects $v_i(m, y^{i-1}, u^i) \in [0,1]$ using the *maximal information gain* rule

$$v_i(m, y^{i-1}, u^i) = \arg\max_{v \in [0,1]} \mathbb{E}\Big[\ln s_i\Big(m, (y^{i-1}, Y_i), u^i\Big) \,\Big|\, W_i = w_i(m, y^{i-1}, u^i, v)\Big], \tag{3}$$

where Y_i is distributed according to $F_{Y|W}(y|w_i)$. Note that this is a greedy rule that maximizes the "information gain" for each channel use.

We now provide explanations for the main ingredients of our scheme.

1) To explain the rule for selecting X_i in (2), note that at time i, both the encoder and the decoder know Y^{i-1}. The encoder generates $X_i(M, Y^{i-1})$ that follows F_X as closely as possible. For a DMC,

$$\mathbb{P}\{X_i = x \mid Y^{i-1} = y^{i-1}\} = \sum_{m : x_i(m, y^{i-1}) = x} p_{M|Y^{i-1}}(m \mid y^{i-1}).$$

Therefore, the distribution of X_i is determined by how we divide the posterior probabilities of the message among the input symbols. If M is continuous, we use the same trick as in posterior matching, that is, $X_i = F_X^{-1} \circ F_{M|Y^{i-1}}(M \mid y^{i-1})$, and X_i would follow F_X. Since in our setting M is discrete, the posterior cdf $F_{M|Y^{i-1}}$ contains jumps, and each message m is mapped to an interval instead of a single point. We use V_i to select a point on the interval and map it by F_X^{-1} to obtain the input symbol.

2) To explain the need for the circular shift of the intervals via U_i, note that if we map a point on the interval directly to the input symbol, the chosen symbol would depend on both the position and the length of the interval corresponding to the correct message. While the length of the interval provides information about the posterior probability of the message, the position of the interval does not contain any useful information. By applying the random circular shift U_i, the analysis of the error probability involves only the interval lengths. Suppose m is sent, define $S_i = s_i(m, Y^i, U^i)$ to be the *pseudo* posterior probability of the transmitted message (the length of the interval) at time i and $T_i = t_i(m, Y^i, U^i)$ (the position of the interval). Note that $\{S_i\}$ forms a Markov chain, and its transition can be specified by

$$S_i = \int_{[0, S_{i-1}] + \tilde{U}_i \bmod 1} dF_{W|Y}(w|Y_i),$$

where $Y_i \sim p_{Y|W}(\cdot|W_i)$ is independent of $\tilde{U}^i$, S^{i-1} and Y^{i-1}, and

$$\begin{aligned} W_i &= V_i \cdot S_{i-1} + \tilde{U}_i \bmod 1, \\ V_i &= \arg\max_{v \in [0,1]} \mathbb{E}\left[\ln S_i \mid W_i = v \cdot S_{i-1} + \tilde{U}_i \bmod 1\right], \end{aligned}$$

Note that $\tilde{U}_i = T_{i-1} + U_i \sim \mathrm{U}[0,1]$ is independent of $\tilde{U}^{i-1}$, S^{i-1}, and Y^{i-1}.

As a result of the random circular shift, the analysis of error reduces to studying the real-valued Markov chain $\{S_i\}$. This is simpler than the analysis of posterior matching, which involves keeping track of the entire posterior distribution.

3) The reason we use the maximal information gain rule in (3) to select V_i is that it yields a better bound on the error exponent than the simpler rule of selecting V_i uniformly at random. With this complicated rule, however, it is very difficult to calculate the posterior probabilities. Hence, in our scheme, the interval length $s_i(m, y^i, u^i)$ is an *estimate* of the posterior probability assuming V_i is selected uniformly at random. In the following we explain the method of estimating the posterior probability in detail.

Define another probability distribution $\tilde{\mathbb{P}}$ on $(M, X^n, Y^n, W^n, U^n, V^n)$ in which X^n is also generated according to (2) but V^n is an i.i.d. sequence with $V_i \sim \mathrm{U}[0,1]$ instead of using (3). The receiver uses this distribution to estimate the posterior probability of each message, i.e.,

$$s_i(m, y^i, u^i) = \tilde{\mathbb{P}}\{M = m \mid Y^i = y^i,\, U^i = u^i\}.$$

The expression in (2) is obtained inductively using

$$\begin{aligned}&\tilde{\mathbb{P}}\{M=m \mid Y^i=y^i,\ U^i=u^i\}\\&\quad\propto \tilde{\mathbb{P}}\{M=m \mid Y^{i-1}=y^{i-1},\ U^i=u^i\}\\&\qquad\cdot\tilde{\mathbb{P}}\{Y_i=y_i \mid M=m,\ Y^{i-1}=y^{i-1},\ U^i=u^i\},\end{aligned}$$

where

$$\begin{aligned}&\tilde{\mathbb{P}}\{M=m \mid Y^{i-1}=y^{i-1},\ U^i=u^i\}\\&\quad\cdot\tilde{\mathbb{P}}\{Y_i=y_i \mid M=m,\ Y^{i-1}=y^{i-1},\ U^i=u^i\}\\&=s_{i-1}(m)\cdot\tilde{\mathbb{P}}\{Y_i=y_i \mid M=m, Y^{i-1}=y^{i-1}, U^i=u^i\}\\&=s_{i-1}(m)\cdot\int_0^1\tilde{\mathbb{P}}\{Y_i=y_i \mid M=m,\ Y^{i-1}=y^{i-1},\\&\qquad\qquad U^i=u^i,\ V_i=v\}\,dv\\&=s_{i-1}(m)\cdot\int_0^1\tilde{\mathbb{P}}\{Y_i=y_i \mid M=m,\ Y^{i-1}=y^{i-1},\\&\qquad U^i=u^i, W_i=v\cdot s_{i-1}(m)+t_{i-1}(m)+u_i \bmod 1\}dv\\&=\int_{[t_{i-1}(m),t_{i-1}(m)+s_{i-1}(m)]+u_i \bmod 1}\tilde{\mathbb{P}}\{Y_i=y_i \mid M=m, Y^{i-1}=y^{i-1}, U^i=u^i, W_i=w\}dw\\&=\int_{[t_{i-1}(m),t_{i-1}(m)+s_{i-1}(m)]+u_i \bmod 1}p_{Y|W}(y_i|w)\,dw.\end{aligned}$$

Note that we write $s_{i-1}(m)=s_{i-1}(m,y^{i-1},u^{i-1})$ and $t_{i-1}(m)=t_{i-1}(m,y^{i-1},u^{i-1})$ for simplicity. Hence,

$$\begin{aligned}&\tilde{\mathbb{P}}\{M=m \mid Y^i=y^i,\ U^i=u^i\}\\&\quad=\frac{\int_{[t_{i-1}(m),t_{i-1}(m)+s_{i-1}(m)]+u_i \bmod 1}p_{Y|W}(y_i|w)\,dw}{\sum_{\tilde{m}=1}^{|\mathcal{M}|}\int_{[t_{i-1}(\tilde{m}),t_{i-1}(\tilde{m})+s_{i-1}(\tilde{m})]+u_i \bmod 1}p_{Y|W}(y_i|w)dw}\\&\quad=\frac{\int_{[t_{i-1}(m),t_{i-1}(m)+s_{i-1}(m)]+u_i \bmod 1}p_{Y|W}(y_i|w)\,dw}{\int_0^1 p_{Y|W}(y_i|w)\,dw}\\&\quad=\int_{[t_{i-1}(m),t_{i-1}(m)+s_{i-1}(m)]+u_i \bmod 1}dF_{W|Y}(w|y_i).\end{aligned}$$

The quantity $s_i(m,y^i,u^i)$ can be viewed as a *pseudo posterior probability* of message m. Note that the pseudo posterior probabilities of all the messages still sum up to one, hence we know the correct message is recovered when its pseudo posterior probability is greater than $1/2$.

From the above description, the key differences between our scheme and posterior matching are as follows:

1) We apply a random circular shift U_i to reduce the analysis of error to studying the behavior of the Markov chain $\{S_i\}$.

2) The message is an integer $M\in[1:e^{nR}]$ rather than a real number $\Theta\in[0,1]$. This again simplifies the analysis.

3) Instead of using the posterior probability of the message as in posterior matching, we use the maximal information gain rule, which is crucial to the analysis of the scheme.

As a result of these differences, our scheme can achieve good error exponent over the entire rate range using a simpler error probability analysis.

III. Analysis of the Probability of Error

In this section, we analyze the rate and the error exponent of our scheme for DMCs. Note that in this case, $W_i=w\in[0,1]$ is mapped to $X_i=x=F_X^{-1}(w)$ if $F_X(x-1)<w\le F_X(x)$. As we discussed in the previous section, the pseudo posterior probability of the transmitted message $\{S_i\}$ forms a Markov chain. We obtain the bound on the error exponent by analyzing this Markov chain.

In our scheme, the decoder declares $\hat{m}=\arg\max_{m'}s_n(m',y^n,u^n)$. Since the pseudo posterior probabilities of all the messages sum up to one, if the pseudo posterior probability of the transmitted message $S_n=s_n(m,Y^n,U^n)>1/2$, we can be sure that the message is recovered correctly. Hence, the probability of error is upper bounded as

$$\begin{aligned}P_e^{(n)}&=\mathbb{P}\left\{M\ne\arg\max_m s_n(m,Y^n,U^n)\right\}\\&\le\mathbb{P}\{S_n\le 1/2\}.\end{aligned}$$

Remark 2: An alternative approach would be to use a threshold decoder [16], which decodes to the message with posterior probability greater than a threshold γ. However, this would introduce another error event when there is a message other than the correct one with pseudo posterior probability greater than γ. As a result, we cannot analyze the error probability by studying S_n only. Therefore we fix the threshold at $1/2$ to simplify the analysis.

To study how the error probability decays with n, we consider the error exponent

$$E(R)=\limsup_{n\to\infty}-n^{-1}\ln P_e^{(n)}(R).$$

We define the moment generating function of the ideal increment of information (or *ideal moment generating function* in short) for DMC as

$$\begin{aligned}\phi(\rho)&=\sum_x p(x)\sum_y p(y|x)\left(\frac{p(x|y)}{p(x)}\right)^{-\rho}\\&=\sum_x p(x)\sum_y p(y|x)\left(\frac{p(y|x)}{\sum_{x'}p(x')p(y|x')}\right)^{-\rho}.\end{aligned}$$

The function $\ln\phi(\rho)$ is convex, and it is not difficult to check that

$$\phi'(0)=\frac{d}{d\rho}\phi(\rho)\bigg|_{\rho=0}=\frac{d}{d\rho}\ln\phi(\rho)\bigg|_{\rho=0}=-I(X;Y).$$

Similarly, we define the moment generating function of the actual increment of information at s (or *actual moment generating function* in short) as

$$\psi_s(\rho)=\mathbb{E}\left[S_i^{-\rho}/S_{i-1}^{-\rho}\,\middle|\,S_{i-1}=s\right].$$

The function $\ln\psi_s(\rho)$ is convex. To obtain the bound on the error exponent, we also need the quantity

$$\Psi=\inf_{\tau(s)}\sup_{s\in(0,1)}\psi_s(\tau(s)),$$

where $\tau(s)$ is nondecreasing and the infimum is taken over all nondecreasing functions $\tau:(0,1)\to[0,\infty)$. We have $\Psi\le 1$, since we can take $\tau(s)=0$.

We introduce the following condition on a DMC, which is sufficient for our scheme to achieve the capacity.

Definition 1: A pair of input symbols $x_1\ne x_2$ in a DMC $p(y|x)$ is said to be redundant if $p(y|x_1)=p(y|x_2)$ for all y.

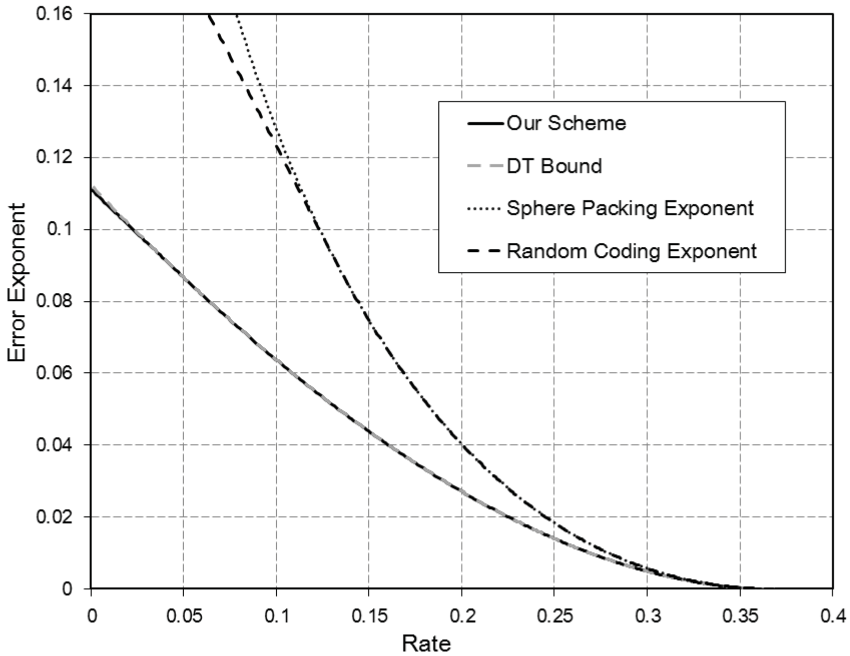

Fig. 2. Comparisons of the bound on the error error exponent for a BSC(0.1).

Note that if the channel has redundant input symbols, we can always use only one of these symbols and ignore the others. Therefore we can assume without loss of generality that the channel has no redundant input symbols.

We are now ready to state the main result of this paper.

Theorem 1: For any DMC $p(y|x)$ without redundant input symbols, we have $\Psi < 1$, and the maximal information gain scheme can achieve the capacity. Further, for any $R < I(X;Y)$, the error exponent is lower bound as

$$E(R) \geq \sup_{\rho>0} \{-\rho R - \ln \max(\phi(\rho), \Psi)\}.$$

The proof of this theorem is detailed in the following subsection.

The bound on the error exponent of our scheme becomes quite tight as the rate tends to the capacity.

Corollary 1: The error exponent $E(R)$ satisfies

$$E(R) = \frac{(I(X;Y) - R)^2}{2\mathrm{Var}[\ln(p(Y|X)/p(Y))]} - O\left((I(X;Y) - R)^3\right)$$

as R tends to $I(X;Y)$.

The quantity $\mathrm{Var}[\ln(p(Y|X)/p(Y))]$ is known as the channel dispersion [17], [18]. Note that this is the same limit as for the sphere packing bound. Hence the error exponent of our scheme tends to the sphere packing bound when the rate tends to the capacity. The proof of this corollary is given in Appendix A.

To illustrate the above results, consider the following.

Example 1 (Binary Symmetric Channel): Consider a binary symmetric channel with crossover probability p. It is well known that the capacity of this channel is achieved with $X \sim \mathrm{Bern}(1/2)$. The maximal information gain rule always selects the input symbol whose probability interval has the larger overlapping area with the message interval. The actual moment generating function is $\psi_s(\rho) = p\alpha + q\beta$, where $q = 1 - p$, and

$$\alpha = \begin{cases} (2p)^{-\rho}\left(1 - 2s + \dfrac{4sp}{(\rho-1)(q-p)}\right) + \dfrac{2s}{\rho-1} & s \leq \dfrac{1}{2}, \\ (2s-1)\left(2q - \dfrac{q-p}{s}\right)^{-\rho} \\ \qquad - \dfrac{2s(1-(2q-(q-p)/s)^{1-\rho})}{(\rho-1)(q-p)} & s > \dfrac{1}{2}, \end{cases}$$

$$\beta = \begin{cases} (2q)^{-\rho}\left(1 - 2s - \dfrac{4sq}{(\rho-1)(q-p)}\right) + \dfrac{2s}{\rho-1} & s \leq \dfrac{1}{2}, \\ (2s-1)\left(2p + \dfrac{q-p}{s}\right)^{-\rho} \\ \qquad + \dfrac{2s(1-(2p-(q-p)/s)^{1-\rho})}{(\rho-1)(q-p)} & s > \dfrac{1}{2}. \end{cases}$$

The value of Ψ can be found approximately using dynamic programming. For example, for BSC(0.1), $\Psi \approx 0.8948$.

Figure 2 compares the bound on the error exponent for our scheme to the following.

1) The sphere packing exponent $\max_Q \max_{\rho>0} E_0(\rho, Q) - \rho R$, where

$$E_0(\rho, Q) = -\ln \sum_y \left(\sum_x Q(x)\, p(y|x)^{1/(1+\rho)}\right)^{1+\rho}.$$

2) The random coding exponent $\max_Q \max_{\rho\in(0,1]} E_0(\rho, Q) - \rho R$, which is a lower bound without feedback.

3) The dependence-testing (DT) bound [16], which is the error exponent for random coding without feedback when a threshold decoder is used.

Note that our error exponent approaches the sphere packing exponent when R is close to the capacity. Also our exponent almost coincides with the DT bound, with noticeable difference only when the rate is close to zero.

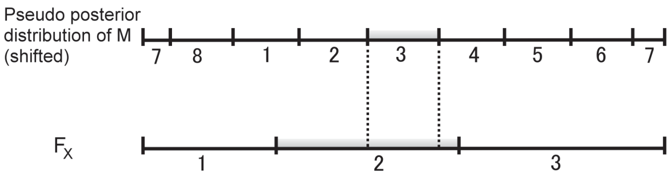

Fig. 3. Illustration of the shifted pseudo posterior distribution during the starting phase (assuming $M = 3$ is sent).

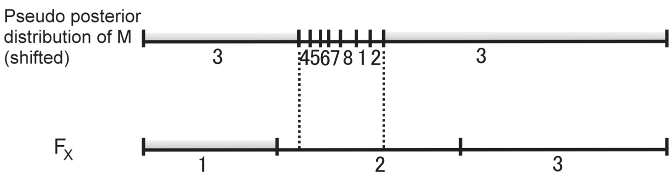

Fig. 4. Illustration of the shifted pseudo posterior distribution during the ending phase (assuming $M = 3$ is transmitted).

A. Proof of Theorem 1

We first outline the main ideas of the proof of the theorem. As we discussed, to analyze the error probability of our scheme, it suffices to study how S_i increases from $S_0 = e^{-nR}$ to $S_n > 1/2$. We divide the analysis of the scheme by the stage of transmissions into: the *starting phase*, where S_i is small, the *transition phase*, where S_i is not close to 0 or 1, and the *ending phase,* where S_i is close to 1. We outline the proof for each phase.

The starting phase refers to the transmission period in which $S_i \le S_{\text{start}}$, where S_{start} is a constant that depends on the channel. During this phase, the length of the message interval $[T_{i-1} + U_i, T_{i-1} + S_{i-1} + U_i] \mod 1$ is close to 0 and is very likely to overlap with the probability interval $[F_X(x-1), F_X(x)]$ for only a single input symbol x as illustrated in Figure 3. In this case, the maximal information gain rule selects x and the probability of X_i would be close to $p(x)$. The following lemma shows that in this regime the actual moment generating function is close to the ideal one.

Lemma 1 (Starting Phase MGF): For any DMC $p(y|x)$ with input pmf $p(x)$, let $S_{\text{start}} = \min_{x : p(x)>0} p(x)$, then there exists $\omega \ge 1$ such that

$$\left(1 - \frac{s}{S_{\text{start}}}\right)\phi(\rho) + \frac{s}{S_{\text{start}}}\omega^{-\rho} \le \psi_s(\rho) \le \left(1 - \frac{s}{S_{\text{start}}}\right)\phi(\rho) + \frac{s}{S_{\text{start}}}\omega^{\rho}$$

for $s \le S_{\text{start}}$ and $\rho \ge 0$.

The proof of the lemma is given in Appendix B.

The ending phase refers to the transmission period in which $S_i \ge S_{\text{end}}$, where S_{end} is a constant that depends on the channel. During the ending phase, the length of the message interval $[T_{i-1} + U_i, T_{i-1} + S_{i-1} + U_i] \mod 1$ is close to one. Hence, the maximal information gain rule is free to select any input symbol. However, the complement of the message interval is likely to overlap with only one symbol probability interval $[F_X(\bar{x}-1), F_X(\bar{x})]$ as illustrated in Figure 4. In this case, the maximal information gain rule selects the input symbol x, which is the “opposite” of $\bar{x}$ in the sense that the posterior probability of $X_i = \bar{x}$ is minimized when $X_i = x$ is transmitted. This would maximize the posterior probability of the message. We can bound the actual moment generating function during this phase as follows.

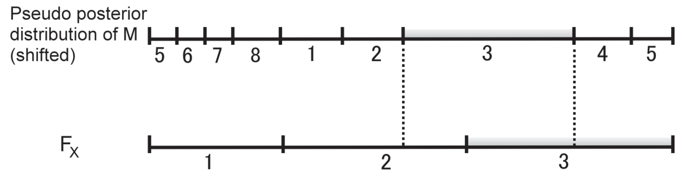

Fig. 5. Illustration of the shifted pseudo posterior distribution during the transition phase (assuming $M = 3$ is sent).

Lemma 2 (Ending Phase MGF): For any DMC $p(y|x)$, there exists $0 < S_{\text{end}} < 1$, $\gamma > 0$ and $\Psi_{\text{end}} < 1$ such that when $s \ge S_{\text{end}}$,

$$\psi_s\left(\gamma\,(1-s)^{-1}\right) \le \Psi_{\text{end}}.$$

The proof of the lemma is given in Appendix C.

The transition phase refers to the transmission period in which $S_{\text{start}} < S_i < S_{\text{end}}$ as illustrated in Figure 5. For the error exponent in Theorem 1 to be nonzero, we need $\Psi < 1$, therefore we need to find a nondecreasing function $\tau : (0,1) \to [0,\infty)$ such that $\psi_s(\tau(s))$ is bounded above and away from 1. From the plot in Figure 6, we can see that $\psi_s(\rho)$ is well-behaved in the starting and ending phases, but not in the transition phase. Nevertheless, it is possible to construct τ satisfying the requirement, as shown in the following lemma.

Lemma 3: For a DMC $p(y|x)$ without redundant input symbols, we have $\Psi < 1$.

The proof of the lemma is given in Appendix D.

To show that our scheme achieves the capacity, recall that S_i should increase from $S_0 = e^{-nR}$ to $S_n > 1/2$, or equivalently, $-\ln S_i$ should decrease from $-\ln S_0 = nR$ to $-\ln S_n < \ln 2$. When n is large, S_0 is close to zero; hence the time spent in the starting phase would dominate. Since the actual moment generating function is close to the ideal one during this phase, we expect the decrease in $-\ln S_i$ for each time step to be close to $-\phi'(0) = I(X;Y)$. Therefore as long as $R < I(X;Y)$, $-\ln S_i$ would decrease from nR to a value smaller than $\ln 2$ in n time steps. However, we still need to show that the transition and the ending phase would not affect the performance of the code. As we will see in the proof of the theorem, the fact that $\Psi < 1$ is sufficient for this purpose.

We now discuss the details of the proof of Theorem 1.

Let ρ^* be the maximizer of $-\rho R - \ln\max\{\phi(\rho), \Psi\}$, and define

$$\tau_2(s) = \begin{cases} \rho^* - \epsilon & \text{when } s < \xi \\ \tau(s) & \text{when } s \ge \xi, \end{cases}$$

and

$$g(s) = \exp\left(-\int_{\xi}^{s} \tau_2(r)\, r^{-1}\, dr\right),$$

where ϵ and ξ are suitable constants.

We now use Lemma 1 to 3 to prove the theorem. The main idea is to design a function $g(s)$ and apply the Markov inequality to $g(S_n)$. Note that $\frac{d}{d\rho}\ln\phi(\rho)$ is continuous at

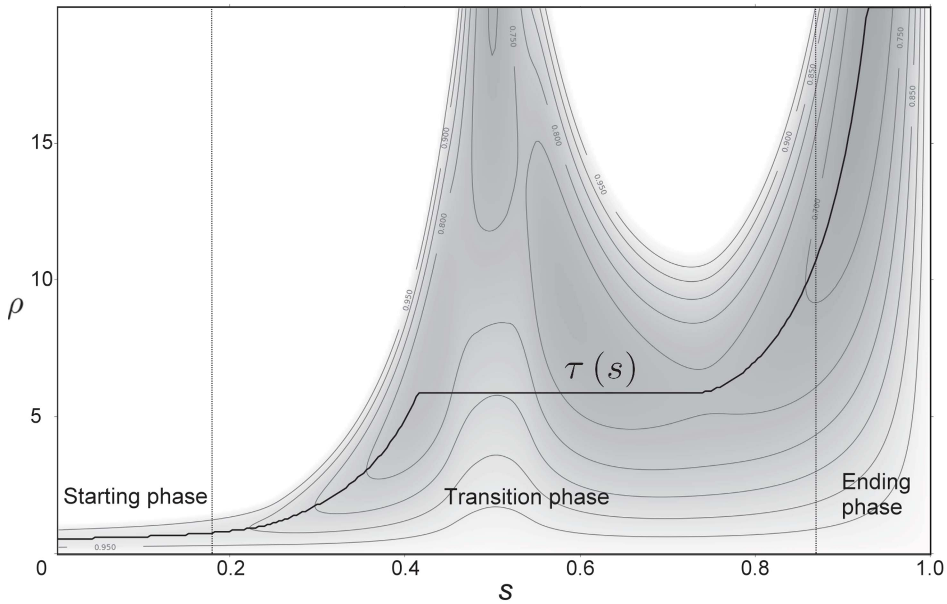

Fig. 6. Contour plot of $\psi_s(\rho)$ for an example channel. Darker color indicates smaller $\psi_s(\rho)$. The minimizing function $\tau(s)$ is also plotted.

$\rho = 0$ and $\frac{d}{d\rho}\ln\phi(\rho)\Big|_{\rho=0} = -I(X;Y)$, therefore, $-R\rho - \ln\max\{\phi(\rho), \psi\}$ is positive when ρ is small. If the proposed bound on the error exponent holds, then $E(R) > 0$ for any $R < I(X;Y)$, and thus capacity can be achieved.

Let ρ^* be the maximizer of $-\rho R - \ln\max(\phi(\rho), \Psi)$. Since $\phi(\rho)$ is continuous, we may assume $\phi(\rho^*) \geq \Psi$. Let $\epsilon > 0$ and $\tau^*(s) > 0$ be a nondecreasing function such that

$$\Psi e^{\epsilon} \geq \psi_s\left(\tau^*(s)\right)$$

for all $s \in (0,1)$.

By Lemma 1, there exists ξ_2 such that when $s \leq \xi_2$, we have $\psi_s(\rho) \leq \phi(\rho)e^{\epsilon}$ for $\rho \leq \rho^* - 4\epsilon/R$. Again by Lemma 1, there exists $\xi \leq \xi_2$ such that when $s \leq \xi$, we have $\phi(\rho) \leq \psi_s(\rho)e^{\epsilon}$ for $\rho \leq \tau^*(\xi_2)$. Define

$$\tau(s) = \begin{cases} \rho^* - 4\epsilon/R & \text{when } s < \xi \\ \tau^*(s) & \text{when } s \geq \xi. \end{cases}$$

Note that

$$\begin{aligned} \ln\phi\left(\tau^*(\xi)\right) &\leq \ln\psi_{\xi}\left(\tau^*(\xi)\right) + \epsilon \\ &\leq \ln\Psi + 2\epsilon \\ &\leq \ln\phi\left(\rho^*\right) + 2\epsilon \\ &< \ln\phi\left(\rho^* - 4\epsilon/R\right). \end{aligned}$$

This implies that $\tau^*(\xi) \geq \rho^* - 4\epsilon/R$ by the convexity of $\phi(\rho)$. Hence $\tau(s)$ is nondecreasing. Define

$$g(s) = \exp\left(-\int_{\xi}^{s} \tau(r)\,r^{-1}\,dr\right).$$

We then consider the quantity $\mathbb{E}[g(S_i)]$. Note that $g(s)$ is nonincreasing, hence

$$\begin{aligned} &\mathbb{E}\left[g(S_i)/g(S_{i-1})\middle|\, S_{i-1} = s\right] \\ &= \mathbb{E}\left[\exp\left(-\int_{s}^{S_i} \tau(r)\,r^{-1}\,dr\right)\middle|\, S_{i-1} = s\right] \\ &\leq \mathbb{E}\left[\exp\left(-\int_{s}^{S_i} \tau(s)\,r^{-1}\,dr\right)\middle|\, S_{i-1} = s\right] \\ &= \mathbb{E}\left[S_i^{-\tau(s)}/s^{-\tau(s)}\middle|\, S_{i-1} = s\right] \\ &\leq \max\left(\phi\left(\rho^* - 4\epsilon/R\right)e^{\epsilon},\ \Psi e^{\epsilon}\right) \\ &= \phi\left(\rho^* - 4\epsilon/R\right)e^{\epsilon}. \end{aligned}$$

Decoding succeeds if $S_n \geq 2/3 > 1/2$. Since $g(S_0) = e^{(\rho^* - 2\epsilon/R)nR}/\xi^{-(\rho^* - 2\epsilon/R)}$, we have

$$\begin{aligned} &\mathbb{P}\{S_n < 2/3\} \\ &\leq \mathbb{E}[g(S_n)]/g(2/3) \\ &\leq \frac{e^{(\rho^* - 4\epsilon/R)nR}}{\xi^{-(\rho^* - 4\epsilon/R)}} \cdot \frac{\left(\phi\left(\rho^* - 4\epsilon/R\right)e^{\epsilon}\right)^n}{g(2/3)} \\ &= \frac{1}{\xi^{-(\rho^* - 4\epsilon/R)}g(2/3)} \\ &\quad \cdot \exp\left(-n \cdot \left(-\rho^* R + \epsilon - \ln\left(\phi\left(\rho^* - 4\epsilon/R\right)\right)\right)\right). \end{aligned}$$

The proof of the theorem is completed by letting $\epsilon \to 0$.

IV. Coding Complexity

In this section, we briefly discuss the implementation of our coding algorithm and show that its computational complexity for DMCs is $O(n\log n)$ and its memory complexity is $O(n)$.

Although there are e^{nR} possible messages, most of them share the same pseudo posterior probability, so instead of storing the pseudo posterior probabilities of the messages separately, we store intervals of message points with the same pseudo posterior probability. We use one binary search tree to keep track of boundary points of these intervals, and another self balancing binary search tree to keep track of the cumulative pseudo posterior probabilities up to their boundary points. The encoder and the decoder both keep and update a copy of each tree (which holds the same content due to feedback).

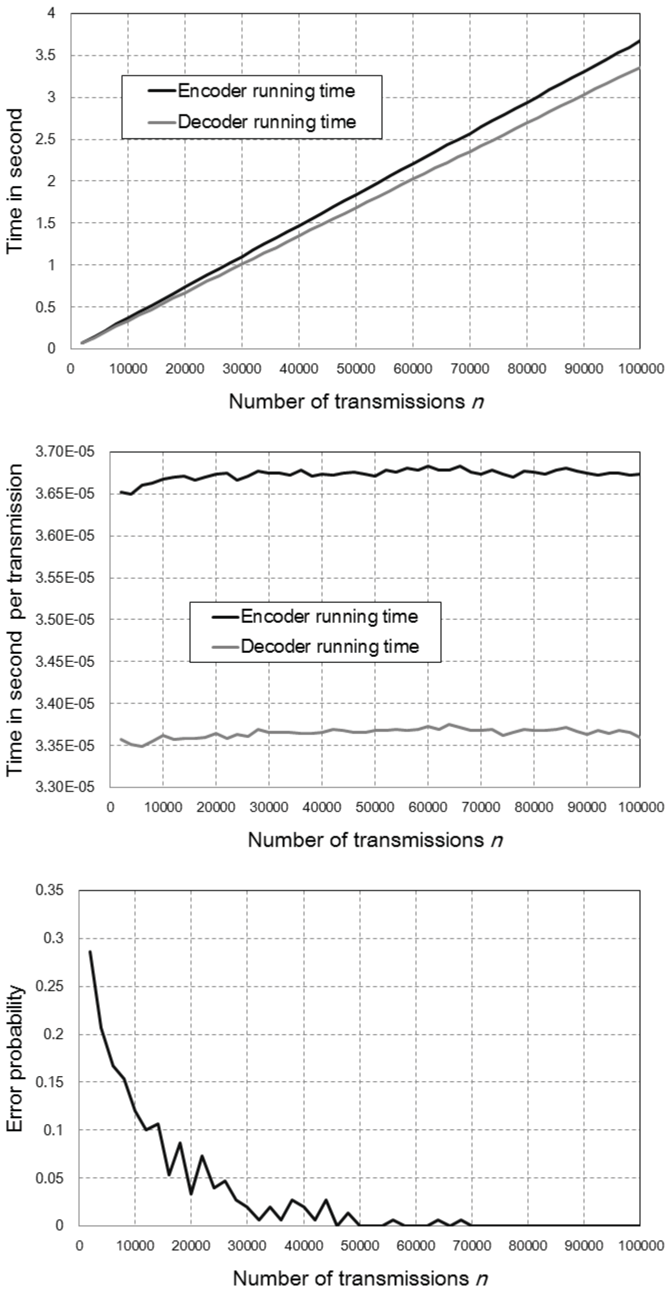

Fig. 7. Top: Running time of our coding algorithm for BSC(0.1) versus the number of channel uses n. Middle: Running time divided by n. Bottom: Empirical error probability (the portion of trials where the decoded message does not match the transmitted one).

We implemented the self balancing tree by a splay tree [19]. For n transmissions, the number of nodes in the tree is at most $n|\mathcal{X}|$, and therefore the queries and the updates can be done in $O(n|\mathcal{X}|\log(n|\mathcal{X}|)) = O(n\log n)$, and the memory complexity is $O(n)$. Detailed description of this implementation can be found in [15].

To corroborate our analysis, we performed simulations of our algorithm assuming a BSC(0.1) and rate $R = 0.98C$ with n from 2000 to 100,000. For each n, 150 independent trials are run to obtain an average running time and an estimate of the error probability. Figure 7 shows that the average running time is close to linear.

V. Conclusion

We proposed a new low coding complexity feedback coding scheme which achieves the capacity of all DMCs. Our scheme is much easier to analyze than posterior matching, making it possible to establish a lower bound on the error exponent that is close to the sphere packing bound at high rate. It would be interesting to explore if our scheme can be modified so that the error exponent exactly coincides with the sphere packing bound when the rate is above a certain threshold. Another possible extension is to investigate whether our scheme achieves the channel dispersion given in [18]. Although variable-length coding with feedback can achieve zero dispersion [20], this may not be achievable using our scheme since it is fixed-length.

Appendix A
Proof of Corollary 1

Recall that the error exponent in Theorem 1 is

$$E(R) \geq \sup_{\rho>0}\{-\rho R - \ln\max(\phi(\rho), \Psi)\}.$$

Consider the Taylor expansion of $\ln\phi(\rho)$ at $\rho = 0$,

$$\begin{aligned}\ln\phi(\rho) &= \ln\sum_y\left(\sum_x p(x)p(y|x)^{1-\rho}\right)\left(\sum_x p(x)p(y|x)\right)^{\rho}\\ &= \rho\cdot -I(X;Y) + \frac{\rho^2}{2}\cdot\sigma^2 + O\left(\rho^3\right),\end{aligned}$$

where

$$\begin{aligned}\sigma^2 &= \mathrm{Var}[\ln(p(Y|X)/p(Y))]\\ &= \mathbb{E}\left[\ln(p(Y|X)/p(Y))^2\right] - \mathbb{E}\left[\ln(p(Y|X)/p(Y))\right]^2\\ &= \sum_x\sum_y p(x)p(y|x)\\ &\qquad\times\left(\ln\frac{p(y|x)}{\sum_{x'}p(x')p(y|x')}\right)^2 - I(X;Y)^2.\end{aligned}$$

Take $\rho = \sigma^{-2}(I(X;Y) - R)$. As $R \to I(X;Y)$, we have $\rho \to 0$, and therefore $\phi(\rho) \to 1$ will be larger than Ψ, and

$$\begin{aligned}E(R) &\geq -\rho R - \ln\phi(\rho)\\ &= \rho(I(X;Y) - R) - \frac{\rho^2}{2}\cdot\sigma^2 - O\left(\rho^3\right)\\ &= \frac{1}{2}\sigma^{-2}(I(X;Y) - R)^2 - O\left((I(X;Y) - R)^3\right).\end{aligned}$$

This completes the proof of the corollary.

Appendix B
Proof of Lemma 1 (Starting Phase MGF)

Assume $S_0 = s \leq S_{\text{start}}$, then

$$\begin{aligned}S_1 &= \int_{[0,s]+U_1 \bmod 1} dF_{W|Y}(w|Y_1)\\ &= \int_{[-s/2,s/2]+U_1' \bmod 1} dF_{W|Y}(w|Y_1),\end{aligned}$$

where $U_1' = U_1 + (s/2) \bmod 1$. Note that $F_X^{-1}(w) = x$ if $F_X(x-1) < w \leq F_X(x)$. Let $\alpha = s/\min_x p(x)$, and let A

be the event that

$$\left(1-\frac{\alpha}{2}\right) F_X(x-1)+\frac{\alpha}{2} F_X(x) < U_1' \leq \frac{\alpha}{2} F_X(x-1)+\left(1-\frac{\alpha}{2}\right) F_X(x)$$

for some x. Then $\mathbb{P}\{A\} = 1-\alpha$. Note that A is independent of $F_X^{-1}\left(U_1'\right)$ (the input symbol that U_1' is mapped to). Conditioned on A, the intervals $[-s/2, s/2]+U_1' \bmod 1$ does not cross the boundary points $F_X(x)$, and we have $S_1 = S_0 + \ln p_{X|Y}\left(F_X^{-1}\left(U_1'\right)|Y_1\right)/p_X\left(F_X^{-1}\left(U_1'\right)\right)$.

We now show that S_0/S_1 is almost surely bounded by a constant independent of $S_0 = s$. Note that

$$S_0/S_1 \geq \min_{x,y:p(x|y)>0} \frac{p(x)}{p(x|y)} \stackrel{def}{=} \omega_{\text{lower}}$$

almost surely. Next we establish an upper bound. If $p(x|y) > 0$ for any x, y, then

$$S_0/S_1 \leq \max_{x,y} \frac{p(x)}{p(x|y)}.$$

Note that when $S_0 \leq \min_x p(x)$, the interval $[0,s)+U_1 \bmod 1$ intersects at most one boundary point $F_X(x)$. Assume $F_X(i) \in ([0,s)+U_1 \bmod 1)$, and let $r = s^{-1}(F_X(i)-(U_1 \bmod 1))$ be the portion of the interval lying in the $X = i$ region, then the maximal information gain scheme would select X among $x \in \{i, i+1\}$ that gives a larger

$$\mathbb{E}\left[\ln\left(rp(i|Y)+(1-r)\,p(i+1|Y)\right)\middle| X=x\right] \stackrel{def}{=} b_x(r).$$

If we have $p_{Y|X}(y|i+1) > 0$ for any y with $p_{Y|X}(y|i) > 0$, then $S_0/S_1 \leq \max_{x,y:p(x|y)>0} p(x)/p(x|y)$ holds when $X = i$. Otherwise there exists a y such that $p_{Y|X}(y|i) > 0$, $p_{Y|X}(y|i+1) = 0$, then $b_i(r) \to -\infty$ when $r \to 0$. By continuity, assume $b_{i+1}(r) > b_i(r)$ for $r < r_{i,i+1}$, then when $X = i$ we have $r \geq r_{i,i+1}$, and

$$S_0/S_1 \leq r_{i,i+1}^{-1} \max_{x,y:p(x|y)>0} \frac{p(x)}{p(x|y)}$$

almost surely. Define $r_{i+1,i}$ similarly. Therefore $S_0/S_1 \leq \omega_{\text{upper}} \stackrel{def}{=} \left(\max_{i,j} r_{i,j}^{-1}\right)\left(\max_{x,y:p(x|y)>0} p(x)/p(x|y)\right)$, and

$$(1-\alpha)\,\phi(\rho)+\alpha\omega_{\text{lower}}^{\rho} \leq \mathbb{E}\left[S_1^{-\rho}/S_0^{-\rho}\middle| S_0=s\right] \leq (1-\alpha)\,\phi(\rho)+\alpha\omega_{\text{upper}}^{\rho}.$$

The proof of Lemma 1 is completed by letting $\omega = \max\left\{\omega_{\text{upper}}, \omega_{\text{lower}}^{-1}\right\}$.

APPENDIX C
PROOF OF LEMMA 2 (ENDING PHASE MGF)

Assume $S_0 = s \geq 1-\xi$. Note that the interval of the message $[0,s]+U_1 \bmod 1$ overlaps all the intervals corresponding to the input symbols. Therefore the encoder can choose among all symbols the one that minimizes the expected value of $-\ln S_1$.

$$\begin{aligned} S_1 &= \int_{[0,s]+U_1 \bmod 1} dF_{W|Y}(w|Y_1) \\ &= 1-\int_{[-(1-s)/2,(1-s)/2]+U_1' \bmod 1} dF_{W|Y}(w|Y_1), \end{aligned}$$

where $U_1' = U_1 + ((1+s)/2) \bmod 1$. Note that $F_X^{-1}(w) = x$ if $F_X(x-1) < w \leq F_X(x)$. Let $\alpha = (1-s)/\min_x p(x)$, and let A be the event that

$$\left(1-\frac{\alpha}{2}\right) F_X(x-1)+\frac{\alpha}{2} F_X(x) < U_1' \leq \frac{\alpha}{2} F_X(x-1)+\left(1-\frac{\alpha}{2}\right) F_X(x)$$

for some k. Then $\mathbb{P}\{A\} = 1-\alpha$. Note that A is independent of $F_X^{-1}\left(U_1'\right)$ (the input symbol that U_1' is mapped to).

Conditioned on A and $U_1' = u_1'$, the intervals $[-(1-s)/2, (1-s)/2]+U_1' \bmod 1$ does not cross the boundary points $F_X(x)$. Assume the interval maps to $x_1 = F_X^{-1}\left(u_1'\right)$. Define the opposite symbol $\text{opp}(x_1)$ as the symbol $\bar{x}_1$ that minimizes

$$\mathbb{E}\left[p_{X|Y}(x_1|Y)\middle| X=\bar{x}_1\right].$$

In case of a tie, choose the symbol that minimizes $\mathbb{E}\left[\left(p_{X|Y}(x_1|Y)\right)^2\middle| X=\bar{x}_1\right]$, and so on. Since

$$\begin{aligned} &\mathbb{E}\left[-\ln S_1 \mid U_1'=u_1', X=x\right] \\ &= \mathbb{E}\left[-\ln\left(1-\int_{[-\frac{1-s}{2},\frac{1-s}{2})+u_1' \bmod 1} dF_{W|Y}(w|Y)\right)\middle| X=x\right] \\ &= \mathbb{E}\left[-\ln\left(1-\frac{1-s}{p_X(x_1)} p_{X|Y}(x_1|Y)\right)\middle| X=x\right] \\ &= \sum_{k=1}^{K} k^{-1}\left(\frac{1-s}{p_X(x_1)}\right)^k \mathbb{E}\left[\left(p_{X|Y}(x_1|Y)\right)^k\middle| X=x\right] \\ &\quad + O\left((1-s)^{K+1}\right) \end{aligned}$$

by the Taylor series expansion, we can find S_{end} such that the maximal information gain scheme chooses $\text{opp}(x_1) = \text{opp}(F_X^{-1}\left(u_1'\right))$ whenever $s \geq S_{\text{end}}$ and u_1' satisfies the conditions of the event A.

Note that $p_X(x_1)$ is the weighted mean of $\mathbb{E}\left[p_{X|Y}(x_1|Y)\middle| X=\bar{x}_1\right]$ over $\bar{x}_1$, and those values are not all equal (or else the capacity of the channel is zero), we have, for any x_1,

$$\mathbb{E}\left[p_{X|Y}(x_1|Y)\middle| X=\text{opp}(x_1)\right] \leq (1-\eta)\,p_X(x_1)$$

for a constant $\eta > 0$ which does not depend on x_1.

Assume S_{end} is close enough to 1 such that $s^{-(1-s)^{-1}} \geq e^{1-\eta/4}$ for $s \geq S_{\text{end}}$.

$$\begin{aligned} &\mathbb{E}\left[S_1^{-\gamma(1-s)^{-1}}\middle| A, S_0=s\right] \\ &= \mathbb{E}\left[\left(1-\frac{p_{X|Y}\left(F_X^{-1}\left(U_1'\right)|Y\right)}{p_X\left(F_X^{-1}\left(U_1'\right)\right)}(1-s)\right)^{-\gamma(1-s)^{-1}}\right. \\ &\qquad \left.\middle| X=\text{opp}\left(F_X^{-1}\left(U_1'\right)\right), S_0=s\right] \\ &\leq \max_{x\in\mathcal{X}} \mathbb{E}\left[\left(1-\frac{p_{X|Y}(x|Y)}{p_X(x)}(1-s)\right)^{-\gamma(1-s)^{-1}}\right. \\ &\qquad \left.\middle| X=\text{opp}(x), S_0=s\right] \\ &\leq \max_{x\in\mathcal{X}} \mathbb{E}\left[\exp\left(\gamma\left(\frac{p_{X|Y}(x|Y)}{p_X(x)}+\frac{\eta}{8}\right)\right)\middle| X=\text{opp}(x), S_0=s\right] \end{aligned}$$

$$\begin{aligned}&\le \max_{x\in\mathcal{X}} \exp\left(\mathbb{E}\left[\gamma\left(\frac{p_{X|Y}(x|Y)}{p_X(x)}+\frac{\eta}{4}\right)\middle| X=\mathrm{opp}(x), S_0=s\right]\right)\\&\le \exp\left(\gamma\left((1-\eta)+\eta/4\right)\right)\\&= e^{\gamma(1-\eta/4)}e^{-\eta\gamma/2}\end{aligned}$$

for S_{end} is close enough to 1 and $\gamma = \gamma(\eta, S_{\mathrm{end}}) > 0$ small enough (depend only on the channel, η and S_{end}). The third line from the bottom can be shown by differentiating the expressions with respect to γ. We have

$$\begin{aligned}&\mathbb{E}\left[S_1^{-\gamma(1-s)^{-1}}/S_0^{-\gamma(1-s)^{-1}}\middle| A, S_0=s\right]\\&\quad\le e^{-\gamma(1-\eta/4)}\mathbb{E}\left[S_1^{-\gamma(1-s)^{-1}}\middle| A, S_0=s\right]\\&\quad\le e^{-\eta\gamma/2}.\end{aligned}$$

Define $\omega = \max_{x,y:p(x)>0} \frac{p(x|y)}{p(x)}$, then

$$\begin{aligned}S_1^{-\gamma(1-s)^{-1}}/S_0^{-\gamma(1-s)^{-1}} &\le (1-\omega(1-s))^{-\gamma(1-s)^{-1}} s^{\gamma(1-s)^{-1}}\\&\le e^{\gamma(1+\omega-\eta/4)}.\end{aligned}$$

Hence,

$$\begin{aligned}\psi_s\left(\gamma(1-s)^{-1}\right) &= (1-\alpha)e^{-\eta\gamma/2}+\alpha e^{\gamma(1+\omega-\eta/4)}\\&\le e^{-\eta\gamma/2}+\frac{1-s}{\min_{x_k} p(x_k)}e^{\gamma(1+\omega-\eta/4)}\\&\le e^{-\eta\gamma/4}\end{aligned}$$

for $1-s$ small enough. This completes the proof of Lemma 2.

Appendix D
Proof of Lemma 3

By Lemma 2, when $S_i \ge S_{\mathrm{end}}$, the actual moment generating function can be bounded. It is left to bound the actual MGF for $S_i < S_{\mathrm{end}}$. We first prove that $\psi_s'(0)$ for $s \le S_{\mathrm{end}}$ can be bounded above and away from 0.

Since the maximal information gain rule (3) minimizes the expectation of $-\ln S_i$, it has a smaller $\mathbb{E}[-\ln S_i | S_{i-1}=s]$ than any other rule of selecting V_i. In particular, if we generate V_i according to U[0, 1], the expectation would be $\tilde{\mathbb{E}}[-\ln S_i|S_{i-1}=s]$, where $\tilde{\mathbb{E}}$ denotes the expectation under the probability measure $\tilde{\mathbb{P}}$. Therefore,

$$\begin{aligned}&\mathbb{E}[-\ln S_i|S_{i-1}=s]\\&\le \tilde{\mathbb{E}}[-\ln S_i|S_{i-1}=s]\\&=\tilde{\mathbb{E}}\left[-\ln\int_{[T_{i-1},T_{i-1}+s]+U_i \bmod 1} f_{W|Y}(w|Y_i)\,dw\right]\\&=\tilde{\mathbb{E}}\left[-\ln\int_{[T_{i-1},T_{i-1}+s]+U_i \bmod 1} \frac{f_{Y|W}(Y_i|w)}{f_Y(Y_i)}\,dw\right]\\&=\tilde{\mathbb{E}}\left[-\ln\int_{[0,s]+U_i \bmod 1} \frac{f_{Y|W}(Y_i|w)}{f_Y(Y_i)}\,dw\,\middle|\,T_{i-1}=0\right]\\&=\int_0^1\int_{[0,s]+u \bmod 1}\int\left(-\ln\int_{[0,s]+u\bmod 1}\frac{f_{Y|W}(y|w)}{f_Y(y)}\,dw\right)\\&\qquad\cdot f_{Y|W}(y|w_0)\,dy\cdot s^{-1}\,dw_0\cdot du\\&=-\ln s+\int_0^1\int\left(\int_{[0,s]+u \bmod 1} f_{Y|W}(y|w)\cdot s^{-1}\,dw\right)\\&\qquad\left(-\ln\frac{\int_{[0,s]+u \bmod 1} f_{Y|W}(y|w)\cdot s^{-1}\,dw}{f_Y(y)}\right)dy\cdot du\\&=-\ln s - I(U_i;Y_i|M,S_{i-1}=s).\end{aligned}$$

Hence

$$\begin{aligned}\psi_s'(0) &= \frac{d}{d\rho}\psi_s(\rho)\Big|_{\rho=0}\\&=\mathbb{E}[-\ln S_i/s|S_{i-1}=s]\\&\le -I(U_i;Y_i|M,S_{i-1}=s)\\&=-H(Y_i)+H(Y_i|U_i,M=m,S_{i-1}=s).\end{aligned}$$

Since $H(Y_i|U_i=u, M, S_{i-1}=s)$ is continuous in u, and entropy is strictly concave, to show $H(Y_i|U_i, M, S_{i-1}=s) < H(Y_i)$, it suffices to show that Y_i does not have the same distribution conditioned on $U_i = u$ for different u. Assume the contrary, i.e., that there exists some $s < 1$ such that Y_i has the same distribution conditioned on $U_i = u$ and $S_{i-1} = s$ for all u. Note that if $V_i \sim$ U[0, 1],

$$\mathbb{P}\{Y_i=y|U_i=u\}=s^{-1}\int_{[0,s]+u \bmod 1} p_{Y|W}(y|w)\,dw.$$

Differentiating the expression with respect to u, we have

$$p_{Y|W}(y|w)=p_{Y|W}(y|w+s \bmod 1)$$

for all y and w. By $p_{Y|W}(y|w)=p_{Y|X}(y|F_X^{-1}(w))$ and the assumption that the channel has no redundant input symbols, we have

$$F_X^{-1}(w)=F_X^{-1}(w+s \bmod 1)$$

for all y and w. This implies $F_X^{-1}(w)$ is either constant or periodic, which leads to a contradiction since $F_X^{-1}(w)$ is nondecreasing and is not constant. Therefore we know that $H(Y_i|U_i, M, S_{i-1}=s) < H(Y_i)$ for $s < 1$. Since $H(Y_i|U_i, M, S_{i-1}=s)$ is continuous in $s \in [0, S_{\mathrm{end}}]$ assuming $V_i \sim$ U[0, 1], the expression is bounded above and away from $H(Y_i)$, and thus we have $\psi_s'(0) \le \zeta$ for all $s \le S_{\mathrm{end}}$, where $\zeta < 0$ is a constant.

Without loss of generality, assume the message transmitted is $m = 1$, then the message interval at time i is $[U_i, U_i + S_{i-1}]$ mod 1, and the symbol selected by the maximal information gain scheme is a function $X_i = x^*(S_{i-1}, U_i)$ of S_{i-1} and U_i. Therefore

$$\begin{aligned}&\psi_s(\rho)\\&\quad=\mathbb{E}\left[S_i^{-\rho}/S_{i-1}^{-\rho}\middle| S_{i-1}=s\right]\\&\quad=\int_0^1\mathbb{E}\left[S_i^{-\rho}/S_{i-1}^{-\rho}\middle| S_{i-1}=s, U_i=u, X_i=x^*(s,u)\right]du\\&\quad=\int_0^1\psi_{s,u,x^*(s,u)}(\rho)\,du,\end{aligned}$$

where $\psi_{s,u,x}(\rho)=\mathbb{E}\left[S_i^{-\rho}/S_{i-1}^{-\rho}\middle| S_{i-1}=s, U_i=u, X_i=x\right]$ is the moment generating function when the message interval is $[u, u+s]$ mod 1 and the transmitted symbol is x.

It is easy to show that $\psi_{s,u,x}'(\rho)$, when treated as a function of (s, u, ρ), is continuous and strictly increasing in ρ. Restricted on $s \le S_{\mathrm{end}}$ and $\rho \le 1$, the domain of the function is $[0, S_{\mathrm{end}}] \times [0, 1] \times [0, 1]$ which is compact, and therefore the function is uniformly continuous in this domain. We can find $\bar{\rho}_x > 0$ such that $\psi_{s,u,x}'(\rho) - \psi_{s,u,x}'(0) \le -\zeta/2$ for any

$s \le S_{\text{end}}$, $u \in [0, 1]$ and $\rho \le \bar{\rho}_x$. Let $\bar{\rho} = \min_x \bar{\rho}_x$. For any $s \le S_{\text{end}}$ and $\rho \le \bar{\rho}$,

$$\begin{aligned}\psi_{s,u,x}(\rho) &= 1 + \int_0^{\rho} \psi'_{s,u,x}(r)\, dr \\ &\le 1 + \rho\left(\psi'_{s,u,x}(0) - \zeta/2\right),\end{aligned}$$

and

$$\begin{aligned}\psi_s(\rho) &= \int_0^1 \psi_{s,u,x^*(s,u)}(\rho)\, du \\ &\le \int_0^1 \left(1 + \rho\left(\psi'_{s,u,x^*(s,u)}(0) - \zeta/2\right)\right) du \\ &= 1 + \rho\left(\psi'_s(0) - \zeta/2\right) \\ &\le 1 + \rho\zeta/2.\end{aligned}$$

Let

$$\tau(s) = \begin{cases} \min\left(\bar{\rho},\ \gamma\,(1 - S_{\text{end}})^{-1}\right) & \text{when } s < S_{\text{end}} \\ \gamma\,(1 - s)^{-1} & \text{when } s \ge S_{\text{end}} \end{cases}$$

be a nondecreasing function, where γ is from Lemma 2. Then

$$\begin{aligned}\Psi &\le \sup_{s \in (0,1)} \psi_s(\tau(s)) \\ &\le \max\left(1 + \min\left(\bar{\rho},\ \gamma\,(1 - S_{\text{end}})^{-1}\right) \cdot \zeta/2,\ \Psi_{\text{end}}\right) \\ &< 1.\end{aligned}$$

This completest the proof of Lemma 3.

ACKNOWLEDGMENTS

The authors are indebted to Young-Han Kim, Chandra Nair, Tsachy Weissman, and the anonymous reviewers for invaluable comments that have greatly improved the exposition of the results in this paper.

REFERENCES

[1] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. 2, no. 3, pp. 8–19, Sep. 1956.

[2] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inf. Theory*, vol. 9, no. 3, pp. 136–143, Jul. 1963.

[3] K. S. Zigangirov, "Upper bounds for the error probability for channels with feedback," *Problemy Peredachi Inf.*, vol. 6, no. 2, pp. 87–92, 1970.

[4] M. V. Burnashev, "On the reliability function of a binary symmetrical channel with feedback," *Problemy Peredachi Inf.*, vol. 24, no. 1, pp. 3–10, 1988.

[5] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback—I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. 12, no. 2, pp. 172–182, Apr. 1966.

[6] J. Schalkwijk, "A coding scheme for additive noise channels with feedback—II: Band-limited signals," *IEEE Trans. Inf. Theory*, vol. 12, no. 2, pp. 183–189, Apr. 1966.

[7] R. L. Dobrushin, "An asymptotic bound for the probability error of information transmission through a channel without memory using the feedback," *Problemy Kibernetiki*, vol. 8, pp. 161–168, 1962.

[8] A. G. D'yachkov, "Upper bounds on the error probability for discrete memoryless channels with feedback," *Problemy Peredachi Inf.*, vol. 11, no. 4, pp. 13–28, 1975.

[9] M. V. Burnashev, "Data transmission over a discrete channel with feedback. Random transmission time," *Problemy Peredachi Inf.*, vol. 12, no. 4, pp. 10–30, 1976.

[10] O. Shayevitz and M. Feder, "Communication with feedback via posterior matching," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2007, pp. 391–395.

[11] O. Shayevitz and M. Feder, "The posterior matching feedback scheme: Capacity achieving and error analysis," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 900–904.

[12] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1222, Mar. 2011.

[13] M. Naghshvar, T. Javidi, and M. A. Wigger. (2013). "Extrinsic Jensen–Shannon divergence: Applications to variable-length coding." [Online]. Available: http://arxiv.org/abs/1307.0067

[14] C. T. Li and A. El Gamal, "An efficient feedback coding scheme with low error probability for discrete memoryless channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2014, pp. 416–420.

[15] C. T. Li and A. El Gamal. (2013). "An efficient feedback coding scheme with low error probability for discrete memoryless channels." [Online]. Available: http://arxiv.org/abs/1311.0100

[16] A. Martinez and A. Guillen i Fabregas, "Random-coding bounds for threshold decoders: Error exponent and saddlepoint approximation," in *Proc. ISIT*, Jul./Aug. 2011, pp. 2899–2903.

[17] V. Strassen, "Asymptotische abschätzungen in Shannons informationstheorie," in *Proc. Trans. 3rd Prague Conf. Inf. Theory*, 1962, pp. 689–723.

[18] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[19] D. D. Sleator and R. E. Tarjan, "Self-adjusting binary search trees," *J. ACM*, vol. 32, no. 3, pp. 652–686, Jul. 1985.

[20] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

Cheuk Ting Li (S'12) received the B.Sc. degree in mathematics and B.Eng. degree in information engineering from The Chinese University of Hong Kong in 2012, and the M.S. degree in electrical engineering from Stanford University in 2014. He is currently working toward the Ph.D. degree at Stanford University. His research interests include information theory and wireless communications.

Abbas El Gamal (S'71–M'73–SM'83–F'00) is the Hitachi America Professor in the School of Engineering and the Chair of the Department of Electrical Engineering at Stanford University. He received his B.Sc. Honors degree from Cairo University in 1972, and his M.S. in Statistics and Ph.D. in Electrical Engineering both from Stanford University in 1977 and 1978, respectively. From 1978 to 1980, he was an Assistant Professor of Electrical Engineering at USC. From 2004 to 2009, he was the Director of the Information Systems Laboratory at Stanford University. His research contributions have been in network information theory, FPGAs, and digital imaging devices and systems. He has authored or coauthored over 200 papers and holds 35 patents in these areas. He is coauthor of the book *Network Information Theory* (Cambridge Press 2011). He received several honors and awards for his research contributions, including the 2012 Claude E. Shannon Award and the 2004 INFOCOM Paper Award. He is a member of the US National Academy of Engineering and a Fellow of the IEEE. He has been serving on the Board of Governors of the Information Theory Society since 2009 and is currently its Junior Past President. He has played key roles in several semiconductor, EDA, and biotechnology startup companies.