

II. NEW FEEDBACK SCHEME

Consider a memoryless channel $F_{Y|X}(y|x)$ with noiseless feedback. The transmitted symbol at time i , X_i , is a function of the message $M \in [1 : e^{nR}]$ and past received symbols Y^{i-1} . Initially, each message m is represented by the subinterval $[(m-1)e^{-nR}, me^{-nR}]$ of $[0, 1]$, where the length of the interval represents the posterior probability of its corresponding message. As more symbols are received, the encoder and decoder update the length of the interval for each message to an estimate of its posterior probability. At the end of transmission, the decoder outputs the message corresponding to the longest interval.

We describe our scheme with the aid of Figure 1. We fix F_X , the cdf of the input symbols (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 will be described later). The encoder, which has Y^{i-1} from 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 $t_{i-1}(m, y^{i-1}, u^{i-1})$ and denote the length of this subinterval by $s_{i-1}(m, y^{i-1}, u^{i-1})$. All subintervals are cyclically shifted by an amount $U_i \sim U[0, 1]$, which is generated independently for each i and is known to both the encoder and the decoder. A point w_i is then selected in the subinterval corresponding to the transmitted message m according to $w_i = (v_i \cdot s_{i-1}(m, y^{i-1}, u^{i-1}) + t_{i-1}(m, y^{i-1}, u^{i-1})) + u_i \text{ mod } 1$, 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 symbol corresponding to the subinterval in \mathcal{I} that contains w_i . At the end of communication, the decoder outputs the message m corresponding to the subinterval with the greatest length $s_n(m, y^{n-1}, u^{n-1})$.

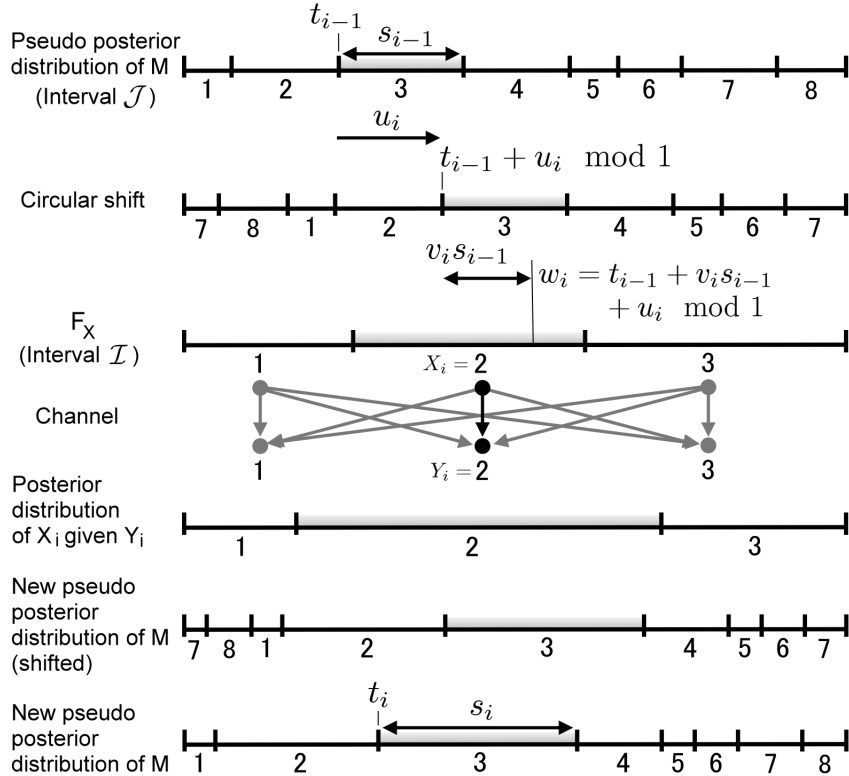


Figure 1. Illustration of the new feedback scheme for a DMC with input and output alphabet $\{1, 2, 3\}$. The message $m = 3$ is transmitted. In time i , symbol $X_i = 2$ is transmitted, and symbol $Y_i = 2$ is received.

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

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

where

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

where s_{i-1} and t_{i-1} are shorthands of $s_{i-1}(m, y^{i-1}, u^{i-1})$ and $t_{i-1}(m, y^{i-1}, u^{i-1})$ respectively. Note that in the above integral we used the notation $[t, t+s] + u \bmod 1$ to mean the set $\{x + u \bmod 1 : x \in [t, t+s]\}$.

The function $F_{W|Y}(w|y_i)$ (assuming y is discrete) is the cdf of W conditioned on Y . Let

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

Then,

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

Assume message m is transmitted. The encoder selects $V_i \in [0, 1]$ using a *maximal information gain* greedy rule, which maximizes the gain in information for each channel use, as

$$V_i = \arg \max_{v \in [0, 1]} \mathbb{E} \left[\ln s_i(m, (y^{i-1}, Y_i), u^i) \mid W_i = w_i(m, y^{i-1}, u^i, v) \right], \tag{2}$$

where Y_i is distributed according to $F_{Y|W}(y|w_i)$.

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

To explain the rule for selecting X_i , note that at time i , both the encoder and the decoder know Y^{i-1} . The encoder would generate $X_i(M, Y^{i-1})$ 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 would 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.

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 we are transmitting $M = m$, define $S_i = s_i(m, Y^i, U^i)$ to be the pseudo posterior probability of the transmitted message at time i and $T_i = t_i(m, Y^i, U^i)$. To see $\{S_i\}$ forms a Markov chain, note that by definition (1), S_i is a function of S_{i-1} , $T_{i-1} + U_i \bmod 1$, and Y_i . Also, Y_i depends only on W_i , which is a function of S_{i-1} and $T_{i-1} + U_i \bmod 1$, and $T_{i-1} + U_i \bmod 1 \sim U[0, 1]$ is independent of $\{S_1, \dots, S_{i-1}\}$. Hence, S_i is conditionally independent of $\{S_1, \dots, S_{i-2}\}$ given S_{i-1} , and to analyze the performance of our scheme, it suffices to study the behavior of the real-valued Markov chain $\{S_i\}$.

The reason why we use the complicated rule in (2) 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 (1) but V^n is an i.i.d. sequence with $V_i \sim U[0, 1]$ instead of using (2). The receiver uses this probability 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 (1) is obtained inductively using

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

$$= \frac{\tilde{\mathbb{P}}\{M = m | Y^{i-1} = y^{i-1}, U^i = u^i\} \cdot \tilde{\mathbb{P}}\{Y_i = y_i | M = m, Y^{i-1} = y^{i-1}, U^i = u^i\}}{\sum_{m'=1}^{|\mathcal{M}|} \tilde{\mathbb{P}}\{M = m' | Y^{i-1} = y^{i-1}, U^i = u^i\} \cdot \tilde{\mathbb{P}}\{Y_i = y_i | M = m', Y^{i-1} = y^{i-1}, U^i = u^i\}},$$

where

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

Hence,

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

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 1, hence we know the correct message is recovered when its pseudo posterior probability is greater than $1/2$.

III. ANALYSIS OF THE SCHEME

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

$$P_e^{(n)} = \mathbb{P} \left\{ M \neq \arg \max_m s_n(m, Y^n, U^n) \right\} \leq \mathbb{P} \{ S_n \leq 1/2 \}.$$

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

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

We use large deviation analysis to bound the error exponent of our scheme. As such, 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

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

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

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

The function $\ln \psi_s(\rho)$ is convex. As we will see in (3),

$$\psi'_s(0) = \left. \frac{d}{d\rho} \psi_s(\rho) \right|_{\rho=0} \leq -I(U_i; Y_i | M = m_0, S_{i-1} = s).$$

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 $\tau : (0, 1) \rightarrow [0, \infty)$. We have $\Psi \leq 1$ since we can take $\tau(s) = 0$.

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

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

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 in this section.

Theorem 2. 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

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

Proof: 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 $S_n > 1/2$, or equivalently, how $-\ln S_i$ decreases from $-\ln S_0 = nR$ to $-\ln S_n < \ln 2$. To show that our scheme achieves the capacity, that is, transmission is successful whenever $R < I(X; Y)$, on average, $-\ln S_i$ should decrease by about $I(X; Y)$ in each time step. Since the maximal information gain rule (2) 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,

$$\begin{aligned} & \mathbb{E}[-\ln S_i | S_{i-1} = s] \\ & \leq \tilde{\mathbb{E}}[-\ln S_i | S_{i-1} = s] \\ & \leq \tilde{\mathbb{E}} \left[-\ln \int_{[0,s]+U_i \bmod 1} \frac{f_{Y|W}(Y_i|w)}{f_Y(Y_i)} dw \right] \\ & = \int_0^1 \int_{[0,s]+u \bmod 1} \int \left(-\ln \int_{[0,s]+u \bmod 1} \frac{f_{Y|W}(y|w)}{f_Y(y)} dw \right) \cdot f_{Y|W}(y|w_0) dy \cdot s^{-1} dw_0 \cdot du \\ & = -\ln s + \int_0^1 \int \left(\int_{[0,s]+u \bmod 1} f_{Y|W}(y|w) \cdot s^{-1} dw \right) \left(-\ln \frac{\int_{[0,s]+u \bmod 1} f_{Y|W}(y|w) \cdot s^{-1} dw}{f_Y(y)} \right) dy \cdot du \\ & = -\ln s - I(U_i; Y_i | M = m, S_{i-1} = s) \\ & = -\ln s - I(X; Y) + I(W_i; Y_i | U_i, M = m, S_{i-1} = s). \end{aligned} \tag{3}$$

As can be seen, when S_i is small, the decrease in $\mathbb{E}[-\ln S_i]$ is close to $I(X; Y)$. However, in the latter part of transmission where S_i is not small, the decrease in $\mathbb{E}[-\ln S_i]$ becomes considerably smaller than $I(X; Y)$, and the above bound becomes loose. In this case, we need to consider the characteristics of the maximal information gain rule in order to establish a tighter bound.

Hence, the transitions of S_i can be quite different depending on the stage of transmission. As such, 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.

The actual moment generating function $\psi_s(\rho)$ describes the distribution of S_i when $S_{i-1} = s$. For the error exponent in Theorem 2 to be nonzero, we need $\Psi < 1$, therefore we need to find a nondecreasing function $\tau : (0, 1) \rightarrow [0, \infty)$ such that $\psi_s(\tau(s))$ is bounded above and away from 1. From the plot in Figure 2, we can see that $\psi_s(\rho)$ is well-behaved in the starting and ending phases, but not in the transition phase. We first analyze the starting and ending phase, and then argue that the transition phase does not affect the rate of the code.

During the starting phase, the length of the message interval $[T_{i-1} + U_i, T_{i-1} + S_{i-1} + U_i] \bmod 1$ is close to 0, and is very likely to overlap with the probability interval $[F_X(x-1), F_X(x)]$ for only one input symbol x . In this case, the maximal information gain rule would select 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.

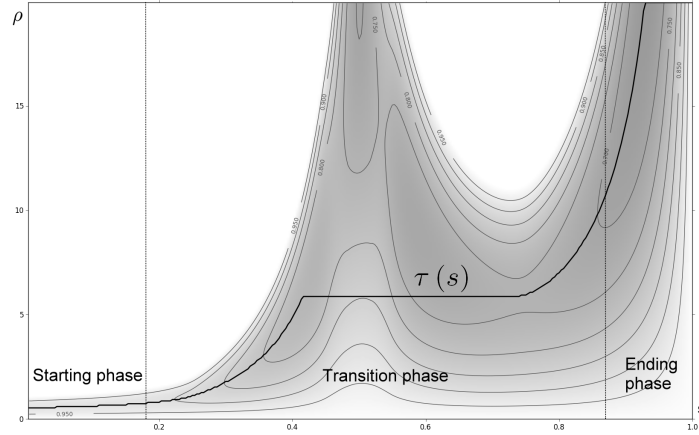


Figure 2. Contour plot of $\psi_s(\rho)$ for an example channel. Darker color indicates smaller $\psi_s(\rho)$. The minimizing function $\tau(s)$ is also plotted.

Lemma 3 (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 $\omega \geq 1$ such that*

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

for $s \leq S_{\text{start}}$ and $\rho \geq 0$.

The proof of the lemma is included in Section IV. In comparison, during the ending phase, the length of the message interval $[T_{i-1} + U_i, T_{i-1} + S_{i-1} + U_i] \bmod 1$ is close to 1. Hence, the maximal information gain rule is free to select practically any input symbol. However, the complement of the message interval is likely to overlap with only one symbol probability interval $[F_X(\bar{x}-1), F_X(\bar{x})]$. In this case, the maximal information gain rule would select x that is the opposite of \bar{x} , in the sense that the posterior probability of $X_i = \bar{x}$ is minimized when we send $X_i = x$, which would maximize the posterior probability of the message. We can bound the actual moment generating function during this phase as follows.

Lemma 4 (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 $s \geq S_{\text{end}}$,*

$$\psi_s(\gamma(1-s)^{-1}) \leq \Psi_{\text{end}}.$$

The proof of the lemma is included in Section IV. We now bound $\psi_s(\rho)$ in the transition phase. We have $\psi'_s(0) < 0$ for any s , and since we have bounded $\psi_s(\rho)$ in both ends, we can argue by continuity that we can find $\tau(s)$ such that $\psi_s(\tau(s))$ in the transition phase is bounded away from 1. This is formally stated in the following.

Lemma 5. *For a DMC $p(y|x)$ without redundant input symbols, we have $\Psi < 1$.*

The proof of the lemma is included in Section IV. To complete the proof of the theorem, let ρ^* be the maximizer of $-\rho R - \ln \max\{\phi(\rho), \Psi\}$, and define

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

and

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

where ϵ and ξ are suitable constants. The theorem follows by applying the Markov inequality to $g(S_n)$. Please refer to Section IV for details of this part of the proof. ■

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

Corollary 6. *The error exponent $E(R)$ satisfies*

$$\liminf_{R \rightarrow I(X;Y)} \frac{E(R)}{(I(X;Y) - R)^2} = \frac{1}{2\text{Var}[\ln(p(Y|X)/p(Y))]}.$$

Therefore $E(R)$ approaches $(I(X;Y) - R)^2 / (2\text{Var}[\ln(p(Y|X)/p(Y))])$ when R tends to $I(X;Y)$.

The quantity $\text{Var}[\ln(p(Y|X)/p(Y))]$ is known as the channel dispersion [14], [15]. 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.

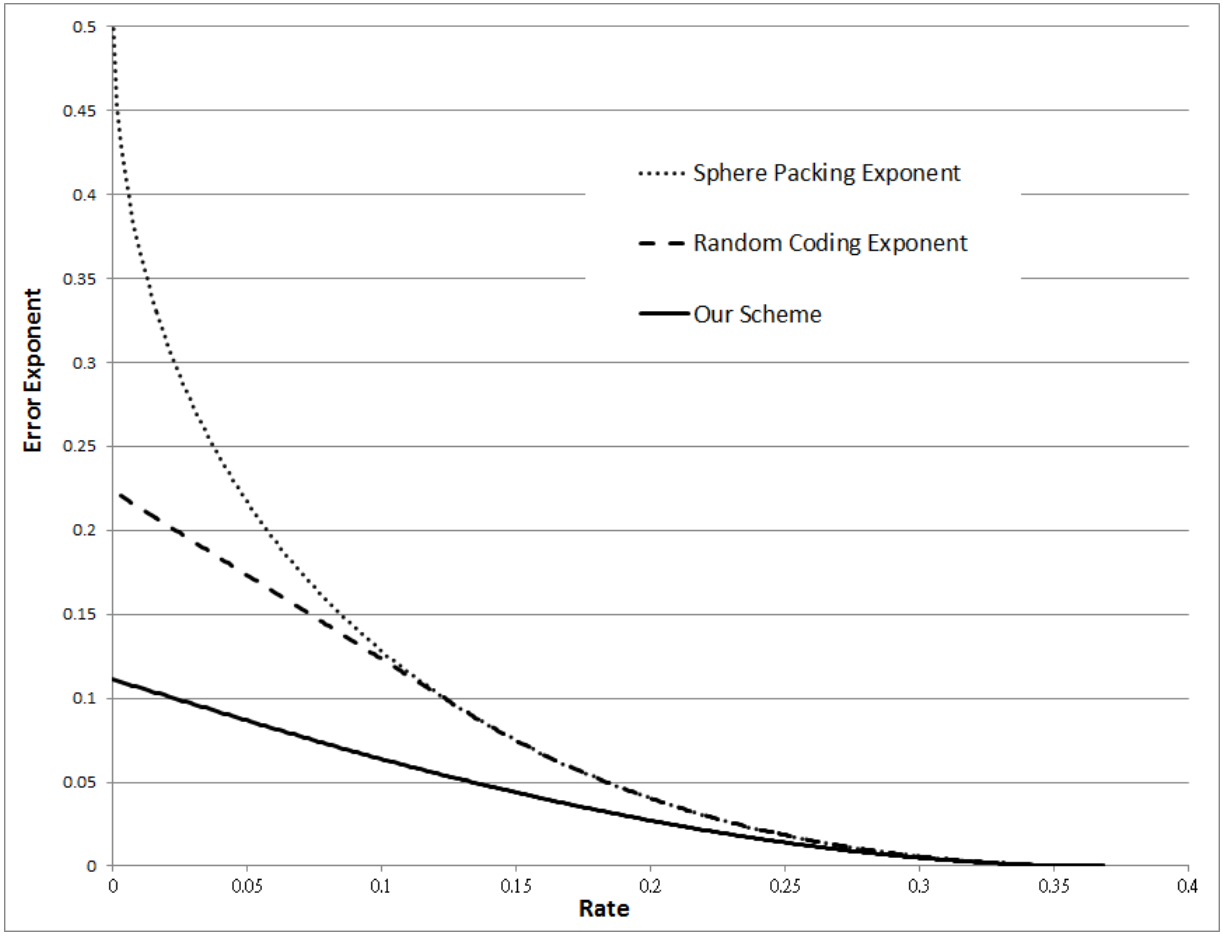


Figure 3. Comparison of error exponent for BSC(0.1)

A. Example: Binary Symmetric Channel

For a binary symmetric channel with crossover probability p , the capacity is achieved with the input $X \sim \text{Bern}(1/2)$. The maximal information gain rule always selects the input symbol whose probability interval has the larger overlap with the message interval. The actual moment generating function is

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

where $q = 1 - p$. The value of Ψ can be found using simple numerical methods.

For BSC(0.1), direct computation yields $\Psi \approx 0.8948$. Comparing this with $\inf_{\rho} \phi(\rho) \approx 0.8938$, we can see that the penalty caused by the transition and ending phases is very small.

Figure 3 compares the bound on the error exponent for our scheme to

- the sphere packing exponent $\max_Q \max_{\rho > 0} E_0(\rho, Q) - \rho R$, where $E_0(\rho, Q) = -\ln \sum_y \left(\sum_x Q(x) p(y|x)^{1/(1+\rho)} \right)^{1+\rho}$,
- the random coding exponent $\max_Q \max_{\rho \in (0,1]} E_0(\rho, Q) - \rho R$, which is a lower bound without feedback.

Note that our error exponent approaches the sphere packing exponent when R is close to C .

IV. PROOFS OF LEMMAS AND THEOREM

A. Proof of Lemma 3 (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 $F_X^{-1}(U'_1)$ (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 $S_1 = S_0 + \ln p_{X|Y}(F_X^{-1}(U'_1)|Y_1) / p_X(F_X^{-1}(U'_1))$.

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

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

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 $b_i(r) \rightarrow -\infty$ when $r \rightarrow 0$. 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 $S_0/S_1 \leq \omega_{\text{upper}} \stackrel{def}{=} (\max_{i,j} r_{i,j}^{-1})(\max_{x,y:p(x|y)>0} p(x)/p(x|y))$, and

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

The result follows by letting $\omega = \max\{\omega_{\text{upper}}, \omega_{\text{lower}}^{-1}\}$.

B. Proof of Lemma 4 (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, and 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 $F_X^{-1}(U'_1)$ (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 $x_1 = F_X^{-1}(u'_1)$. Define the opposite symbol $\text{opp}(x_1)$ as the symbol \bar{x}_1 that minimizes

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

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

$$\begin{aligned} & \mathbb{E} [-\ln S_1 | U'_1 = u'_1, X = x] \\ &= \mathbb{E} \left[-\ln \left(1 - \int_{[-(1-s)/2, (1-s)/2] + u'_1 \bmod 1} dF_{W|Y}(w|Y) \right) \middle| X = x \right] \\ &= \mathbb{E} \left[-\ln \left(1 - \frac{1-s}{p_X(x_1)} p_{X|Y}(x_1|Y) \right) \middle| X = x \right] \\ &= \sum_{k=1}^K k^{-1} \left(\frac{1-s}{p_X(x_1)} \right)^k \mathbb{E} \left[(p_{X|Y}(x_1|Y))^k \middle| X = x \right] + O\left((1-s)^{K+1}\right) \end{aligned}$$

by the Taylor series expansion, we can find S_{end} such that the maximal information gain scheme chooses $\text{opp}(x_1) = \text{opp}(F_X^{-1}(u'_1))$ 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 $\mathbb{E} [p_{X|Y}(x_1|Y) | X = \bar{x}_1]$ 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 ,

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

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 $s \geq S_{\text{end}}$.

$$\begin{aligned} & \mathbb{E} \left[S_1^{-\gamma(1-s)^{-1}} \middle| A, S_0 = s \right] \\ &= \mathbb{E} \left[\left(1 - \frac{p_{X|Y}(F_X^{-1}(U'_1)|Y)}{p_X(F_X^{-1}(U'_1))} (1-s) \right)^{-\gamma(1-s)^{-1}} \middle| X = \text{opp}(F_X^{-1}(U'_1)), S_0 = s \right] \\ &\leq \max_{x \in \mathcal{X}} \mathbb{E} \left[\left(1 - \frac{p_{X|Y}(x|Y)}{p_X(x)} (1-s) \right)^{-\gamma(1-s)^{-1}} \middle| X = \text{opp}(x), S_0 = s \right] \\ &\leq \max_{x \in \mathcal{X}} \mathbb{E} \left[\exp \left(\gamma \left(\frac{p_{X|Y}(x|Y)}{p_X(x)} + \frac{\eta}{8} \right) \right) \middle| X = \text{opp}(x), S_0 = s \right] \\ &\leq \max_{x \in \mathcal{X}} \exp \left(\mathbb{E} \left[\gamma \left(\frac{p_{X|Y}(x|Y)}{p_X(x)} + \frac{\eta}{4} \right) \middle| X = \text{opp}(x), S_0 = s \right] \right) \\ &\leq \exp(\gamma((1-\eta) + \eta/4)) \\ &= e^{\gamma(1-\eta/4)} e^{-\eta\gamma/2} \end{aligned}$$

for S_{end} is close enough to 1 and $\gamma = \gamma(\eta, S_{\text{end}}) > 0$ 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

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

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

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

Hence,

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

for $1-s$ small enough.

C. Proof of Lemma 5

By Lemma 4, when $S_i \geq S_{\text{end}}$, the actual moment generating function can be bounded. It is left to bound the actual MGF for $S_i < S_{\text{end}}$. We first prove that $\psi'_s(0)$ for $s \leq S_{\text{end}}$ can be bounded above and away from 0. Recall that $\psi'_s(0) \leq -H(Y_i) + H(Y_i|U_i, M = m, S_{i-1} = s)$ assuming $V_i \sim \mathcal{U}[0, 1]$. Since $H(Y_i|U_i = u, M = m, S_{i-1} = s)$ is continuous in u , and entropy is strictly concave, to show $H(Y_i|U_i, M = m, S_{i-1} = s) < H(Y_i)$, 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 $V_i \sim \mathcal{U}[0, 1]$,

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

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 $H(Y_i|U_i, M = m, S_{i-1} = s) < H(Y_i)$ for $s < 1$. Since $H(Y_i|U_i, M = m, S_{i-1} = s)$ is continuous in $s \in [0, S_{\text{end}}]$ assuming $V_i \sim \mathcal{U}[0, 1]$, the expression is bounded above and away from $H(Y_i)$, and thus we have $\psi'_s(0) \leq \zeta$ for all $s \leq S_{\text{end}}$, 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 $[U_i, U_i + S_{i-1}] \bmod 1$, 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

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

where $\psi_{s, u, x}(\rho) = \mathbb{E}[S_i^{-\rho}/S_{i-1}^{-\rho} | S_{i-1} = s, U_i = u, X_i = x]$ is the moment generating function when the message interval is $[u, u + s] \bmod 1$ 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 $s \leq S_{\text{end}}$ and $\rho \leq 1$, the domain of the function is $[0, S_{\text{end}}] \times [0, 1] \times [0, 1]$ which is compact, and therefore the function is uniformly continuous in this domain. We can find $\bar{\rho}_x > 0$ such that $\psi'_{s, u, x}(\rho) - \psi'_{s, u, x}(0) \leq -\zeta/2$ for any $s \leq S_{\text{end}}$, $u \in [0, 1]$ and $\rho \leq \bar{\rho}_x$. Let $\bar{\rho} = \min_x \bar{\rho}_x$. For any $s \leq S_{\text{end}}$ and $\rho \leq \bar{\rho}$,

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

and

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

Let

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

be a nondecreasing function, where γ is from Lemma 4. Then

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

D. Details of the last part of the proof of Theorem 2

We now use the lemmas to complete the proof of the theorem. The main idea is to design a function $g(s)$ and then to apply the Markov inequality to $g(S_n)$. Note that $\frac{d}{d\rho} \ln \phi(\rho)$ is continuous at $\rho = 0$ and $\frac{d}{d\rho} \ln \phi(\rho) \Big|_{\rho=0} = -C$, therefore $-\rho R - \ln \max\{\phi(\rho), \Psi\}$ is positive when ρ is small. If the proposed bound on the error exponent holds, then $E(R) > 0$ for any $R < C$, 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

$$\Psi e^\epsilon \geq \psi_s(\tau^*(s))$$

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

By Lemma 3, 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 3, 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

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

from the definition of ρ^* , which 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

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

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

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

The result follows by letting $\epsilon \rightarrow 0$.

E. Proof of Corollary 6

Recall that the error exponent given in Theorem 2 is

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

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

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

where

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

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

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

Hence

$$\liminf_{R \rightarrow I(X;Y)} \frac{E(R)}{(I(X;Y) - R)^2} \geq \frac{1}{2} \sigma^{-2}.$$

V. ALGORITHMS

In this section, we describe the algorithms for the scheme, and show that the encoding and decoding complexity of our scheme 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 a binary tree (which we call tree in message space) to keep track of boundary points of these intervals, and another self balancing binary (which we call tree in probability space) 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).

Each node a in the tree in message space T_m stores 5 attributes: its parent $a.parent$, its left child $a.lchild$, its right child $a.rchild$, the length of the corresponding interval in the message space $a.len$ (stored as a floating point number), and the corresponding node in the tree in probability space $a.pair$. The two children of a node corresponds to two sub-intervals

Each node b in the tree in probability space T_p stores 5 attributes: its parent $b.parent$, its left child $b.lchild$, its right child $b.rchild$, the ratio of the length of the interval of the left child to that of the interval of this node $b.mid$ (stored as a floating point number), and the corresponding node in the tree in message space $b.pair$. For example, if $b.lchild$ corresponds to an interval with pseudo posterior probability 0.1, and $b.rchild$ corresponds to an interval with pseudo posterior probability 0.3, then $b.mid = 1/4$. Note that $b.mid$ needs to be updated accordingly when we perform rotation on the tree. There is a one-to-one correspondence between nodes in T_m and nodes in T_p (i.e., $b.pair.pair = n$ for nodes b in T_m and T_p).

This data structure needs to support the following operations:

- 1) *Query the interval of the transmitted message.* Upon generation of an encoding symbol, the encoder queries the interval $[T_i, T_i + S_i]$ corresponding to the message. This can be done by locating the message in the first tree, and finding the corresponding node in the second tree.
- 2) *Updating interval lengths.* Upon receiving a symbol (and the the feedback from the decoder), the decoder (and the encoder) can calculate the posterior probability of each input symbol x by $p(x|y) = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}$, and update the pseudo posterior probabilities by multiplying a factor $p(x|y)/p(x)$ to the lengths of the intervals in the second tree corresponding to the input symbol x . This creates at most $|\mathcal{X}|$ new nodes in the tree.
- 3) *Query the message with the highest probability.* When transmission is finished, the decoder outputs the message with the highest pseudo posterior probability. Since the analysis shows that the pseudo posterior probability of the correct message is higher than 1/2 with high probability, it suffices to query the posterior median (or the point obtained similarly from the pseudo posterior probabilities) of the message point. This can be done by locating the median point in the second tree, and finding the corresponding node in the first tree.

We implemented the self balancing tree by a splay tree [16], which rotates a node to the root after it is accessed. 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 $O(n)$.

The pseudocode is given below. Given a point $x \in [0, 1]$ in the message space, the function `QueryMsgPoint` finds the corresponding position in the probability space (i.e. it gives the pseudo posterior probability that $M/|\mathcal{M}| \leq x$). To query the interval corresponding to message $m \in \{1, \dots, |\mathcal{M}|\}$, call `QueryMsgPoint` with $x = (m - 1) / |\mathcal{M}|$ and $x = m / |\mathcal{M}|$.

Function `QueryMsgPoint(T_m, T_p, x)`
Input: Trees T_m, T_p , position in message space $x \in [0, 1]$
Output: Corresponding position in probability space

```

a ← Tm.root
while IsNotLeaf(a)
  if x ≤ a.lchild.len
    a ← a.lchild
  else
    x ← x - a.lchild.len
    a ← a.rchild
b ← a.pair
while IsNotRoot(b)
  if b = b.parent.lchild
    x ← x · b.parent.mid
  else
    x ← x · (1 - b.parent.mid) + b.parent.mid
  b ← b.parent
return x

```

The procedure `ScaleInterval` updates the tree T_p so that all intervals in $[0, y]$ are scaled by r/y , and all intervals in $[y, 1]$ are scaled by $(1 - r) / (1 - y)$ (i.e., the interval $[0, y]$ becomes $[0, r]$, and the interval $[y, 1]$ becomes $[r, 1]$). We can update the interval length upon receiving a symbol by applying `ScaleInterval` multiple times. Note that `RotateToRoot(b)` is a procedure which performs tree rotation repeatedly until b becomes the root, and updates the `mid` attribute of the affected nodes.

Procedure `ScaleInterval(T_p, y, r)`
Input: Tree T_p , position at probability space $y \in [0, 1]$, ratio $r \in [0, 1]$

```

b ← Tp.root
while IsNotLeaf(b)
  if y ≤ b.mid
    y ← y/b.mid
    b ← b.lchild
  else
    y ← (y - b.mid)/(1 - b.mid)
    b ← b.rchild
b.mid ← y
b.lchild ← new node, b.rchild ← new node
RotateToRoot(b)
b.mid ← r

```

The function `PosteriorMedian` finds the posterior median of the message, which is the decoded message. Note that x is a high-precision floating point number, which has round-off error much less than $1/|\mathcal{M}|$.

Function PosteriorMedian(T_m, T_p)
Input: Trees T_m, T_p
Output: Decoded message $\hat{m} \in \{1, \dots, |\mathcal{M}|\}$

```

 $b \leftarrow T_p.root, y \leftarrow 1/2$ 
while IsNotLeaf( $b$ )
  if  $y \leq b.mid$ 
     $y \leftarrow y/b.mid$ 
     $b \leftarrow b.lchild$ 
  else
     $y \leftarrow (y - b.mid)/(1 - b.mid)$ 
     $b \leftarrow b.rchild$ 
 $a \leftarrow b.pair$ 
 $x \leftarrow HighPrecision(y \cdot a.len)$ 
while IsNotRoot( $a$ )
  if  $a = a.parent.rchild$ 
     $x \leftarrow x + a.parent.lchild.len$ 
   $a \leftarrow a.parent$ 
return  $\lfloor x \cdot |\mathcal{M}| + 1/2 \rfloor$ 

```

We would demonstrate the algorithm with an example. Consider BSC(0.3). Assume $U_1 = 0.1$ and $Y_1 = 0$. The decoder would know that the message point is more likely to lie on $[0, 0.4] \cup [0.9, 1]$, and less likely to lie on $[0.4, 0.9]$. The trees and the pseudo posterior cdf of message point are shown in the Figure 4. Assume $U_2 = 0.2$ and $Y_2 = 1$. The trees and the cdf in the next time step are shown in the Figure 5. Note that in the tree in message space, after each time step, some leaf nodes are created, while existing nodes are unchanged. In contrast, some nodes in the tree in probability space are modified due to tree rotation.

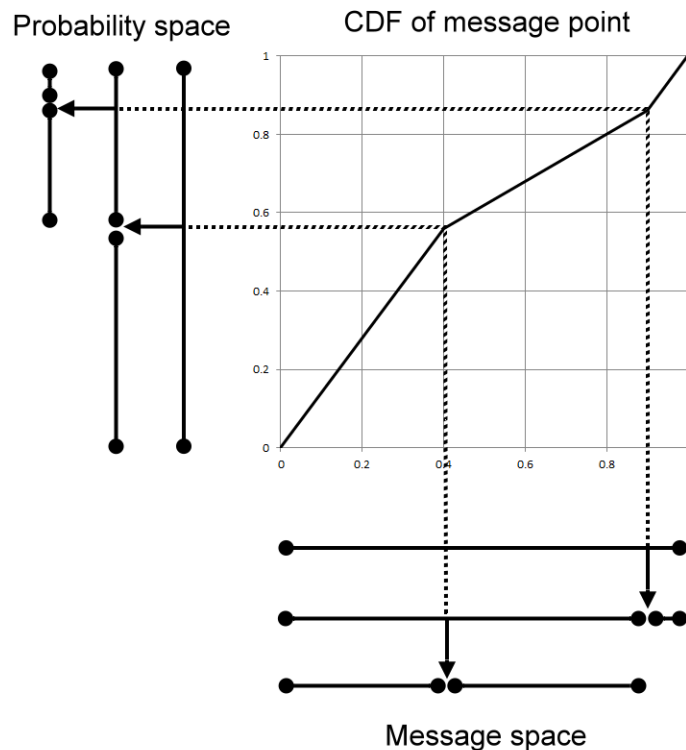


Figure 4. Top-left: Tree in probability space. Top-right: Pseudo posterior cdf of message point. Bottom: Tree in message space.

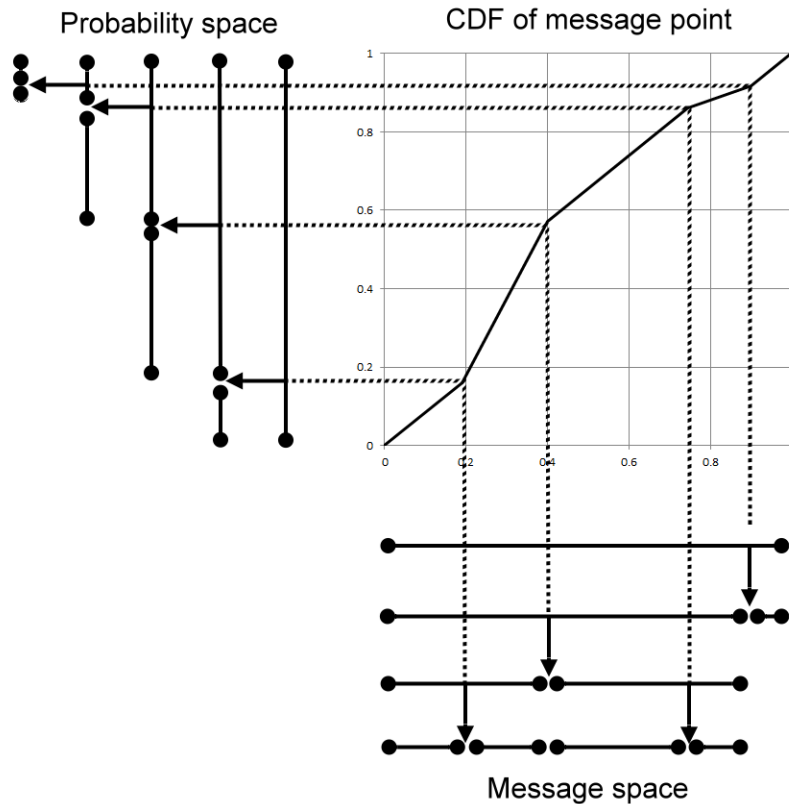


Figure 5. Top-left: Tree in probability space. Top-right: Pseudo posterior cdf of message point. Bottom: Tree in message space.

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 , 50 independent trials are run to obtain an average running time and an estimate of the error probability. Figure 6 shows that the average running time is close to linear.

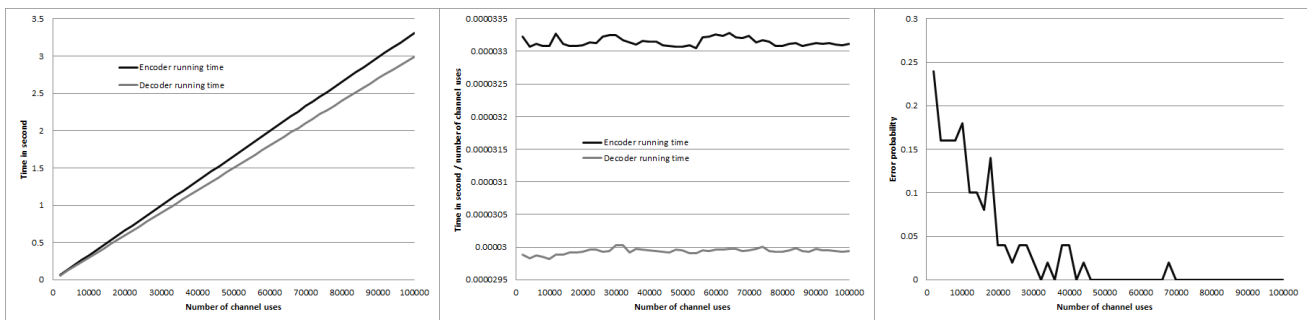


Figure 6. Left: Running time of the implemented encoding and decoding algorithm for BSC(0.1) versus the number of channel uses n . Middle: Running time divided by n . Right: Estimated error probability (the portion of trials where the decoded message does not match the transmitted one). The plot in the middle suggests that the running time complexity is close to linear.

VI. CONCLUSION

We proposed a new feedback coding scheme which achieves the capacity of all DMCs, has low complexity, and is 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.

VII. ACKNOWLEDGMENTS

This work is partially supported by Air Force grant FA9550-10-1-0124. Cheuk Ting Li is supported by a Stanford Graduate Fellowship. The authors are indebted to Young-Han Kim, Chandra Nair and Tsachy Weissman for valuable comments that have greatly improved the exposition of 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," *Probl. 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," *Probl. Peredachi Inf.*, vol. 24, no. 1, pp. 3–10, 1988.
- [5] J. P. M. 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. P. M. 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," *Probl. 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," *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 10–30, 1976.
- [10] O. Shayevitz and M. Feder, "Communication with feedback via posterior matching," in *Proc. of IEEE International Symposium of Information Theory*, June 2007.
- [11] —, "The posterior matching feedback scheme: Capacity achieving and error analysis," in *Proc. of IEEE International Symposium of Information Theory*, July 2008.
- [12] —, "Optimal feedback communication via posterior matching," *IEEE Trans. Info. Theory*, vol. 57, no. 3, pp. 1186–1222, Mar. 2011.
- [13] M. Naghshvar, T. Javidi, and M. A. Wigger, "Extrinsic jensen-shannon divergence: Applications to variable-length coding," 2013. [Online]. Available: <http://arxiv.org/abs/1307.0067>
- [14] V. Strassen, "Asymptotische abschätzungen in shannons informationstheorie," *Trans. Third Prague Conf. Information Theory*, pp. 689–723, 1962.
- [15] Y. Polyanskiy, H. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Info. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.
- [16] D. D. Sleator and R. E. Tarjan, "Self-adjusting binary search trees," *Journal of the ACM*, vol. 32, no. 3, pp. 652–686, 1985.