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

Cheuk Ting Li and Abbas El Gamal
Department of Electrical Engineering
Stanford University
Email: ctl@stanford.edu, abbas@ee.stanford.edu

Abstract—Existing feedback communication schemes are either specialized to particular channels (Schalkwijk–Kailath, Horstein), apply to general channels but have high coding complexity (block feedback schemes), or are difficult to analyze (posterior matching). This paper introduces a feedback coding scheme that achieves the capacity for all discrete memoryless channels with a bound on the error exponent that approaches the sphere packing bound as the rate approaches the capacity and coding complexity of only $O(n \log n)$. These benefits are attained by combining features from previous schemes with new randomization technique and decoding rule.

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]. However, feedback has other benefits, including greatly simplifying coding and increasing reliability. Early examples of feedback 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]. Schalkwijk and Kailath showed that the error probability for their scheme decays doubly exponentially in the block length. It is also known that the error exponent for symmetric discrete memoryless channels cannot exceed the sphere packing bound even when feedback is present [6]. Nevertheless, the schemes in [3], [4] can attain better error exponents than the best known achievable error exponent without feedback. D’yachkov [7] proposed a general feedback scheme for discrete memoryless channels. This scheme, however, appears to have a high computational complexity.

In addition to the traditional fixed-length setting where the number of channel uses is predetermined before transmission commences, there has been work on variable-length schemes in which transmission ends when the error probability is lower than a desired target. The optimal error exponent for this setting was given explicitly by Burnashev [8]. More recently, Shayevitz and Feder [9], [10], [11] 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 the error probability for variable-length schemes, including posterior matching, is given in a recent paper by Naghashvar, Javidi and Wigger [12].

In this paper, we propose a 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, and (iii) has coding complexity of only $O(n \log n)$. Our scheme is motivated by posterior matching. Unlike posterior matching, however, we assume a discrete message space as in the Burnashev scheme, and apply a random cyclic shift to the message points in each transmission 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 scheme. In Section III, we show that our scheme achieves the capacity of any DMC and establish a lower bound on its error exponent. In Section IV, we discuss scheme’s coding complexity. Details of the proof and the analysis of its coding complexity are given in [13].

Remark: Throughout this paper, we use nats instead of bits and ln instead of log to avoid adding normalization constants.

II. NEW FEEDBACK SCHEME

Consider a memoryless channel $F_{Y|X}(y|x)$ with noiseless feedback. The transmitted symbol at time $t$, $X_t$, is a function of the message $M \in [1 : e^{nR}]$ and past received symbols $Y^{t−1}$. 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 $t$ is determined as follows. The decoder, knowing $Y^{t−1}$, partitions another unit interval $\mathcal{J}$ according to the pseudo posterior probability distribution of $M$ given $Y^{t−1}$. This work is partially supported by Air Force grant FA9550-10-1-0124.
(the details of computing this distribution will be described later). The encoder, which has $Y^{t-1}$ 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 $w_i = (v_i \cdot s_{i-1}(m, y^{i-1}, u^{i-1}) + t_{i-1}(m, y^{i-1}, u^{i-1})) + u_i \mod 1$, 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 $\mathcal{J}$ that contains $w_i$. At the end of communication, the decoder outputs the message $m$ corresponding to the subinterval with the greatest length $s_n(m, y^{n-1}, u^{n-1})$.

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 $[t, t+s] + u \mod 1$ to mean the set $\{x + u \mod 1 : x \in [t, t+s]\}$.

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

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

Assume that message $m$ is transmitted. The encoder selects $V_i \in [0,1]$ using the maximal information gain greedy rule

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

where $Y_i\{W = w_i\} \sim F_{Y|W}(y|w_i)$. Note that this rule maximizes the gain in information for each channel use.

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

To explain the rule for selecting $X_i$, note that after transmission $i$, 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, i.e.,

$$P\{X_i = x \mid Y^{i-1} = y^{i-1}\} = \sum_{m : x \in F^{-1}_M(y^{i-1})} P_M(Y^{i-1}(m) = y^{i-1}).$$

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, $X_i = F^{-1}_X \circ F_{M|Y^{i-1}}(M \mid y^{i-1})$, 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 $S_{i-1}$, $T_{i-1} + U_i \mod 1$, and $Y_i$. Also, $Y_i$ depends only on $W_i$, which is a function of $S_{i-1}$ and $T_{i-1} + U_i \mod 1$, and $T_{i-1} + U_i \mod 1 \sim U[0,1]$ is independent of $\{S_1, \ldots, S_{i-1}\}$. Hence, $S_i$ is conditionally independent of $\{S_1, \ldots, S_{i-2}\}$ 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. To explain this more precisely, define a probability distribution $\mathbb{P}$ 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., $s_i(m, y^i, u^i) = \mathbb{P}(M = m | Y^i = y^i, U^i = u^i)$.

The expression in (1) is obtained inductively using

$$
\mathbb{P}(M = m | Y^i = y^i, U^i = u^i) = \mathbb{P}(Y_i = y_i | M = m, Y^{i-1} = y^{i-1}, U^i = u^i) \\
= \left( \sum_{m'} \mathbb{P}(M = m' | Y^i = y^{i-1}, U^i = u^i) \right)^{-1}
$$

where

$$
\mathbb{P}(M = m | Y^i = y^{i-1}, U^i = u^i) = s_{i-1}(m, y^{i-1}, u^{i-1})
$$

$$
\mathbb{P}(Y_i = y_i | M = m, Y^{i-1} = y^{i-1}, U^i = u^i, V_i = v) = \int_0^1 \mathbb{P}(Y_i = y_i | M = m, Y^{i-1} = y^{i-1}, U^i = u^i, V_i = v) dv.
$$

Hence,

$$
\mathbb{P}(M = m | Y^{i-1} = y^{i-1}, U^i = u^i) = \int_{y_i = y_i} \mathbb{P}(Y_i = y_i | M = m, Y^{i-1} = y^{i-1}, U^i = u^i) \prod_{l=1}^{i-1} m_l(y_l | u_{l}) \, dw
$$

and

$$
\mathbb{P}(M = m | Y^i = y^i, U^i = u^i) = \left( \sum_{m'} \mathbb{P}(M = m' | Y^i = y^i, U^i = u^i) \right)^{-1}
$$

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. Outline of the Analysis of the Scheme

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

In our scheme, the decoder declares $\hat{n} = \arg \max_m s_n(m', y^n, u^n)$. 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

$$
P(\epsilon(n)) = \mathbb{P}(M \neq \hat{n} = \arg \max_m s_n(m, y^n, u^n)) \leq \mathbb{P}(S_n > 1/2).
$$

We consider the error exponent

$$
E(\epsilon) = \lim_{n \to \infty} -n^{-1} \ln P(\epsilon(n)).
$$

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

$$
\phi(\rho) = \sum_{x} p(x) \sum_{y} p(y|x) \left( \frac{p(x|y)}{p(x)} \right)^{-\rho}
$$

$$
= \sum_{x} p(x) \sum_{y} p(y|x) \left( \frac{p(y|x)}{\sum_{z} p(z)p(y|z)} \right)^{-\rho}.
$$

The function $\ln \phi(\rho)$ is convex and it can be shown that

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

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

$$
\psi_\epsilon(s) = \mathbb{E} \left[ s_{\epsilon}^{i-\rho} | S_{i-1} = s \right].
$$

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

$$
\psi_\epsilon'(0) = \frac{d}{d\rho} \psi_\epsilon(\rho) \bigg|_{\rho = 0} \leq -I(U_i; Y_i | M = m_0, S_{i-1} = s).
$$

To obtain the bound on the error exponent, we define

$$
\Psi = \inf_{\tau} \sup_{\epsilon(0, 1)} \psi_\epsilon(\tau(s)),
$$

where $\tau(s)$ is nondecreasing and the infimum is taken over all nondecreasing functions $\tau : (0, 1) \to [0, \infty)$. We have $\Psi \leq 1$ 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 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

\[
E(R) \geq \sup_{\rho > 0} \left\{ -\rho R - \ln \max \left( \phi(\rho), \Psi \right) \right\}.
\]

**Proof outline:** As we discussed, to analyze the error probability of our scheme, it suffices to study how \( S_t \) increases from \( S_0 = e^{-\frac{nR}{2}} \) to \( S_n > 1/2 \), or equivalently, how \( -\ln S_t \) 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 \( R < I(X;Y) \), we compare, \( -\ln S_t \), should decrease by about \( I(X;Y) \) each time step. Since the maximal information gain rule (2) minimizes the expectation of \( -\ln S_t \), it has a smaller \( E[ -\ln S_t | S_{t-1} = s ] \) than any other rule of selecting \( V_t \). In particular, if we generate \( V_t \) according to \( U[0,1] \), the expectation would be \( E[ -\ln S_t | S_{t-1} = s ] \), where \( E \) denotes the expectation under \( P \). Thus,

\[
\begin{align*}
E[ -\ln S_t | S_{t-1} = s ] & \\
& \leq E[ -\ln S_t | S_{t-1} = s ] \\
& \leq E \left[ \ln \int_{\{s\} + U_t \bmod 1} f_{Y|Y}(Y_t | u) \frac{f_{V_t}(v_t)}{f_{V_t}(v_t)} \, du \right] \\
& = -\ln s - I(X;Y) + I(W_t; Y_t | U_t, M = m, S_{t-1} = s).
\end{align*}
\]

As can be seen, when \( S_t \) is small, the decrease in \( E[ -\ln S_t ] \) is close to \( I(X;Y) \). However, in the latter part of transmission where \( S_t \) is not small, the decrease in \( E[ -\ln S_t ] \) 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 \( S_t \) 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 \( S_t \) is small, the transition phase where \( S_t \) is not close to 0 or 1, and the ending phase where \( S_t \) is close to 1.

The actual moment generating function \( \psi_s(\rho) \) describes the distribution of \( S_t \) when \( S_{t-1} = s \). For the error exponent in Theorem 1 to be nonzero, we need \( \Psi < 1 \), therefore we need to find a nondecreasing function \( \tau : (0,1) \rightarrow [0,\infty) \) such that \( \psi_s(\tau(s)) \) is bounded above and away from 1. 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 \( [T_{t-1} + U_t, T_{t-1} + S_{t-1} + U_t] \bmod 1 \) 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 \( X_t \) 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 \( \omega \geq 1 \) such that

\[
\begin{align*}
\psi_s(\rho) & \geq \left( 1 - \frac{s}{S_{\text{start}}} \right) \phi(\rho) + \frac{s}{S_{\text{start}}} \omega^{-\rho}, \\
\psi_s(\rho) & \leq \left( 1 - \frac{s}{S_{\text{start}}} \right) \phi(\rho) + \frac{s}{S_{\text{start}}} \omega^\rho
\end{align*}
\]

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

During the ending phase, the length of the message interval \( [T_{t-1} + U_t, T_{t-1} + S_{t-1} + U_t] \bmod 1 \) 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 \( [F_X(x-1), F_X(x)] \). 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 \( X_t = \bar{x} \) is minimized when we send \( X_t = x \), which would maximize the posterior probability of the message. We bound the actual moment generating function during this phase as follows.

**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 \( s \geq S_{\text{end}} \),

\[
\psi_s(\gamma (1-s)^{-1}) \leq \psi_{\text{end}}.
\]

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

To complete the proof of the theorem, let \( \rho^* \) be the maximizer of \( -\rho R - \ln \max \{ \phi(\rho), \Psi \} \), and define

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

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

where \( \epsilon \) and \( \xi \) are suitable constants. The theorem follows by applying the Markov inequality to \( g(S_n) \).

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

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

VI. ACKNOWLEDGMENTS

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


