Capacity and Coding for the Gilbert–Elliott Channels

MORDECHAI MUSHKIN AND ISRAEL BAR-DAVID, FELLOW, IEEE

Abstract—**The Gilbert–Elliott channel is a varying binary symmetric channel with crossover probabilities determined by a binary-state Markov process. In general, such a channel has a memory that depends on the transition probabilities between the states. First, a method of calculating the capacity of this channel is introduced and applied to several examples; then the question of coding is addressed. In the conventional usage of varying channels, a code suitable for memoryless channels is used in conjunction with an interleaver, with the decoder considering the deinterleaved symbol stream as the output of a derived memoryless channel. The transmission rate in such uses is limited by the capacity of this memoryless channel, which is, however, often considerably less than the capacity of the original channel. In this work a decision-feedback decoding algorithm, that completely recovers this capacity loss, is introduced. It is shown that the performance of a system incorporating such an algorithm is determined by an equivalent genie-aided channel, the capacity of which equals that of the original channel. Also, the calculated random coding exponent of the genie-aided channel indicates a considerable increase in the cutoff rate over the conventionally derived memoryless channel.**

I. Introduction and Review of Main Results

The GILBERT–ELLIOTT channel [1] is a varying binary symmetric channel, the crossover probabilities of which are determined by the current state of a discrete-time stationary binary Markov process (see Fig. 1). The states are appropriately designated G for good and B for bad. Due to the underlying Markov nature of the channel, it has memory that depends on the transition probabilities between the states. Section II of this paper is devoted to the calculation of the capacity C_μ of the channel where μ is a measure of memory. It is shown that, when the one-dimensional statistics of the channel are fixed, C_μ increases monotonically with μ and converges asymptotically to a value C^{SI} which is the capacity of the same channel when side information about its instantaneous state is available to the receiver. Section III is devoted to the efficient use of codes, originally designed for memoryless channels, over this channel with memory.

Manuscript received June 10, 1988; revised April 19, 1989. An early version of this work was presented at the 1986 Symposium on Information Theory, Ann Arbor, MI, under the title, "Benefiting from Hidden Memory in Interleaved Codes." This work was supported in part by a grant from the Israeli Ministry of Communication, in part by Elisra Electronic Systems Ltd., in part by Mr. and Mrs. I. E. Berezin, and in part by the Fund for Promotion of Research at the Technion.
M. Mushkin is with Elisra, Electronic Systems, Ltd., 48 Mivtza Kadesh Street, Bene-Beraq 51203, Israel.
I. Bar-David is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.
IEEE Log Number 8931526.

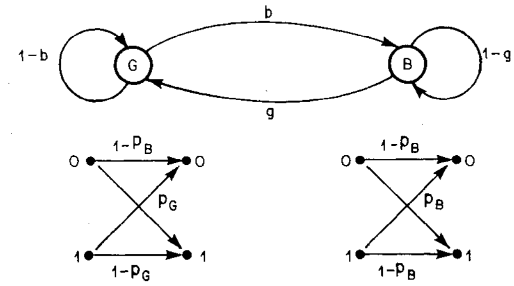

Fig. 1. Gilbert–Elliott channel model. p_G and p_B are channel error probabilities in "good" and "bad" states, respectively, and g and b are transition probabilities between states.

It is known that reliable communication over a finite-state channel is theoretically possible at any rate below capacity[1] [2, pp. 176–182]. In practical uses, however, two major difficulties arise. First, much less is known about good codes for such channels than for memoryless ones; second, the length (and therefore the decoding complexity) of such codes would per force depend on the length of the channel memory. This is apparent from the fact that the error exponent for channels with memory [2, p. 178] depends on the block length N whereas for memoryless channels it is independent of N.

The conventional solution to these two problems is to fragment and disperse the channel memory by interleaving the encoded stream of symbols prior to transmission and to deinterleave the corresponding stream of received symbols [3, pp. 364–366]. If the interleaving span is long then the *interleaved channel* (the cascade of interleaver, channel, and deinterleaver) may be considered memoryless, and therefore efficient coding techniques for memoryless channels may be used. We denote the capacity of the interleaved channel, under the assumption of no memory, by C^{NM}. The disadvantage of such a solution is that the capacity C^{NM} of the memoryless interleaved channel is lower than the capacity C_μ of the original channel. This fact is demonstrated in Section II and illustrated in Fig. 2, where a typical curve of C_μ is drawn as a function of μ between its limits C^{NM} and C^{SI}. The other curves in this figure will be explained further. In reality, however, the interleaving process, being invertible, does not remove the channel memory but only transfers it into a *latent* frag-

[1]Strictly speaking, at any rate below $\underline{C}$ [2, pp. 180–181]. The Gilbert–Elliott channel is indecomposable and therefore $\underline{C} = \overline{C} = C$ [2, p. 109].

0018-9448/89/1100-1277$01.00 ©1989 IEEE

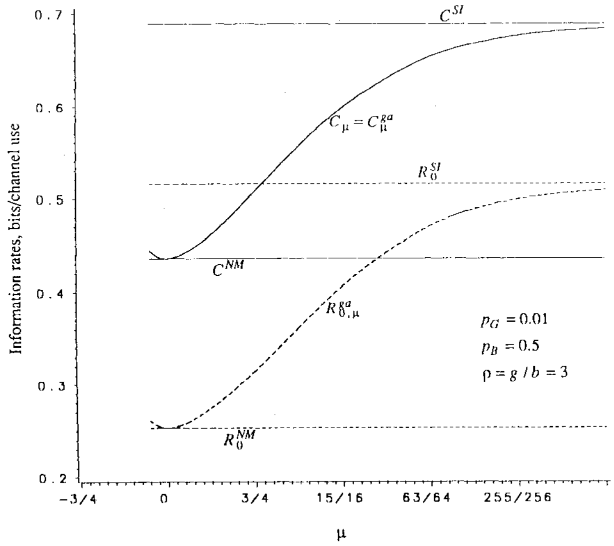

Fig. 2. Information rates for Gilbert–Elliott channels as function of channel memory μ, for fixed p_G, p_B, and ρ. C^{SI} and R_0^{SI} are capacity and the cutoff rate of channel, respectively, when side-information is available. C^{NM} and R_0^{NM} are the respective values of interleaved channel, when considered memoryless.

mented form which does not interfere with the operation of standard error correcting coders and decoders. The data processing theorem [2, p. 80] implies that an invertible operation does not reduce the channel capacity, and therefore the interleaved channel has the same capacity C_μ as the original one. Thus it is not the interleaving operation that causes the capacity reduction in conventional systems but rather the decoding algorithm that ignores the latent memory in the interleaved channel.

In Section III a decision-feedback decoder that *does* utilize the latent memory is introduced (see Fig. 3). As shown later, its use with interleaved codes enables transmission at rates comparable to those over memoryless channels with capacity C_μ. The decision-feedback decoder is composed of a novel metric calculator and a soft-decision decoder, such as is used with memoryless channels. The metric calculator operates on the received channel symbols y and and on feedback decoded data $\hat{x}$ to estimate the probability q that the next channel symbol is in error, conditioned on the previous channel errors. The estimate $\hat{q}$ of this probability is used to produce a log-likelihood metric m which is supplied to the soft-decision decoder. The additional complexity of the proposed decision-feedback decoder, over a conventional decoder, is due solely to the metric calculator, which is shown in Section III to be a simple recursive operation.

The performance of a system that includes the decision-feedback decoder is evaluated by first establishing its equivalence to a system that includes a genie-aided channel which is composed of the interleaved channel, as defined before, and of a genie-aided metric calculator. The input of the genie-aided channel is the encoded symbol stream, while its output consists of the deinterleaved symbol stream and the perfectly estimated probabilities q. The genie-aided channel is shown to be binary-input, output-symmetric, discrete and essentially memoryless, provided the interleaving is sufficiently deep. Its capacity is shown to be equal to that of the original Gilbert–Elliott channel. Both the capacity and the random coding exponent are given in terms of the probability distribution of the random variables q. The advantage of the decision-feedback decoding algorithm, over a conventional one, in terms of capacity and cutoff rate, is shown via numerical examples in Fig. 2. In this figure C_μ^{ga} and $R_{o,\mu}^{\mathrm{ga}}$ denote, respectively, the capacity and the cutoff rate of the genie-aided channel derived from the Gilbert–Elliott channel of memory μ, while C^{NM} and R_o^{NM} denote the respective quantities of the interleaved channel when considered memoryless.

Previous work on this subject includes the following: Capacity calculation for the Gilbert channel model [4], which specifies that, in one of the states, the channel is

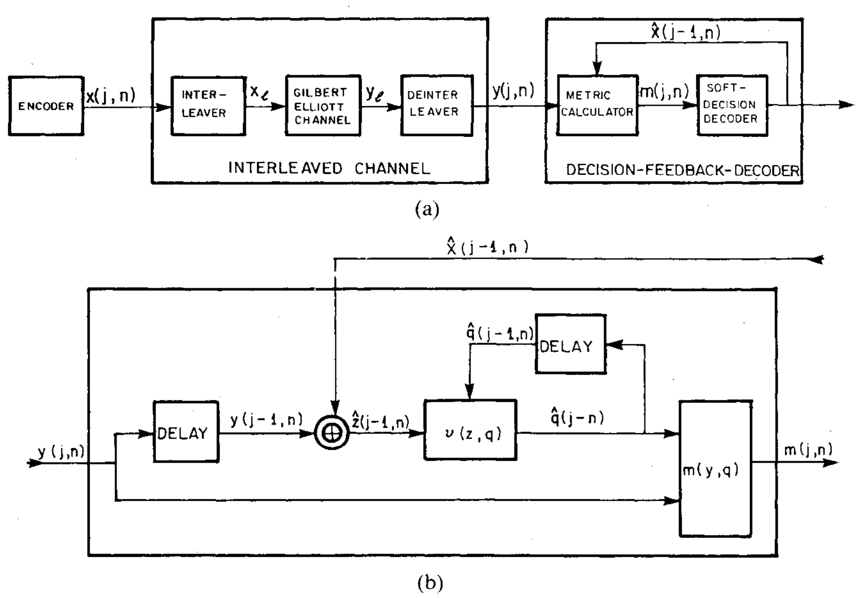

Fig. 3. Communication system with decision-feedback decoder. (a) General description. (b) Metric calculator.

error-free and historically precedes the Gilbert–Elliott model; capacity calculation for the Gilbert–Elliott channel under the assumption that imperfect side-information about the channel states is available [5]; presentation of a version of a decision-feedback decoder for an interleaved Gilbert–Elliott channel, using an *ad hoc* binary estimate of the channel state [6]; and presentation and analysis of a decoding algorithm for an interleaved Gilbert channel [7].

II. The Capacity of the Gilbert–Elliott Channel

Overview of the Section

The Gilbert–Elliott channel is formally introduced in Definition 1. Proposition 4 gives the expression for the capacity of this channel in terms of the expectation of functions of the random variables q_l, interpreted as the probability of a channel error at the lth use of the channel, conditioned on the channel errors at its previous uses, and q_l^* which in addition is conditioned also on the initial state of the channel. Definition 2 and Propositions 1–3 establish properties of q_l and q_l^*. Propositions 5 and 6 establish that within the class of channels with the same one-dimensional statistics, the channel capacity increases monotonically with an appropriately defined channel memory and converges asymptotically to a quantity equal to the capacity of the channel when side information about its instantaneous state is provided.

Throughout this paper the subscripts and superscripts to vectors are to be interpreted as follows: $\mathbf{w}_n^m \triangleq (w_m, w_{m+1}, \cdots, w_n)$ for $m \le n$ and $\mathbf{w}_n \triangleq \mathbf{w}_n^1$. Logarithms are to base 2. The symbol $\oplus$ denotes addition modulo 2.

A. Basic Properties of the Conditional Probabilities of Channel Errors

Definition 1: Let $x_l \in \{0,1\}$, $y_l \in \{0,1\}$ and $z_l \triangleq x_l \oplus y_l$ denote, respectively, the input, the output, and the error of the channel at the lth use, $l = 1, 2, \cdots$. The *error process* $\{z_l\}_{l=1}^{\infty}$ is independent of the channel input, that is,

$$\Pr[\mathbf{y}_l | \mathbf{x}_l] = \Pr[\mathbf{z}_l]. \tag{2.1}$$

The error process has memory in the sense that it depends on an underlying *state process* $\{s_l\}_{l=0}^{\infty}$, $s_l \in \{G, B\}$, where G stands for *good* and B for *bad*. When conditioned on the state process, the error process is memoryless, that is

$$\Pr[\mathbf{z}_l | \mathbf{s}_l] = \prod_{i=1}^{l} \Pr[z_i | s_i]. \tag{2.2}$$

The conditional probabilities $\Pr[z_l|s_l]$ do not depend on the time index l. The state process is a stationary first-order Markov process; that is,

$$\Pr[s_l | \mathbf{s}_{l-1}] = \Pr[s_l | s_{l-1}] \tag{2.3}$$

and the right side of (2.3) does not depend on the index l. The parameters of the channel are the crossover probabilities in the G and B states

$$p_G \triangleq \Pr[z_l = 1 | s_l = G], \qquad p_B \triangleq \Pr[z_l = 1 | s_l = B] \tag{2.4}$$

and the transition probabilities

$$g \triangleq \Pr[s_l = G | s_{l-1} = B], \qquad b \triangleq \Pr[s_l = B | s_{l-1} = G]. \tag{2.5}$$

The initial distribution of the state process is assumed to be

$$\Pr[s_0 = G] = g/(g+b), \qquad \Pr[s_0 = B] = b/(g+b), \tag{2.6}$$

which ensures its stationarity. We define the *good-to-bad ratio* ρ of the channel by

$$\rho \triangleq \Pr[s_l = G]/\Pr[s_l = B] = g/b. \tag{2.7}$$

In the singular cases $\rho = 0$ and $\rho = \infty$ the channel has actually one state only. These cases, being of no interest in the context of this paper, are not considered. It can be shown, by induction on l, following a procedure similar to that in Appendix II, that for $\xi \in \{G, B\}$,

$$\Pr[s_l = \xi | s_0 = \xi] - \Pr[s_l = \xi | s_0 \ne \xi] = (1 - g - b)^l. \tag{2.8}$$

This leads to the definition of the *channel memory* μ

$$\mu \triangleq 1 - g - b. \tag{2.9a}$$

When $\mu = 0$ the channel is memoryless, that is, the current state is independent of all previous states (see (2.3) and (2.5)). If $\mu > 0$, the channel has a *persistent* memory; that is, the probability of remaining in a given state is higher than the steady-state probability of being in that state. If $\mu < 0$, the channel has *oscillatory* memory; that is, the probability of remaining in a state is lower than the steady-state probability of being in that state. The extreme cases are $\mu = \pm 1$, in either of which the state process is completely determined by the initial state. If $\mu = 1$, the channel remains forever in the initial state; this case is not of interest in our context. If $\mu = -1$ the states alternate regularly. We therefore limit μ to the interval $[-1, 1)$. Note that when memory is persistent, any μ-values from zero to one can be associated with any good-to-bad ratio ρ. This is, however, not the case for oscillatory memory cases because it follows from (2.7) and (2.9a) that the restriction

$$\mu \ge \max\{-\rho, -\rho^{-1}\} \tag{2.9b}$$

holds; it is only for $\rho = 1$ that μ can reach the value -1.

Definition 2: For sample paths such that $\Pr[\mathbf{z}_{l-1}, s_0] \ne 0$, let $q_l^*(\mathbf{z}_{l-1}, s_0)$ denote the probability for a channel error at the lth use, conditioned on the initial state, and the previous channel errors, that is,

$$q_l^*(\mathbf{z}_{l-1}, s_0) \triangleq \Pr[x_l = 1 | \mathbf{z}_{l-1}, s_0], \tag{2.10a}$$

and let $q_l(\mathbf{z}_{l-1})$ denote the same probability conditioned only on the previous channel errors, that is,

$$q_l(\mathbf{z}_{l-1}) \triangleq E[q_{l-1}^*(\mathbf{z}_{l-1}, s_0) | \mathbf{z}_{l-1}] = \Pr[z_l = 1 | \mathbf{z}_{l-1}] \tag{2.10b}$$

where $E[\cdot]$ denotes expected value. The stationarity of the channel allows shifting the time reference and thus the following notation is valid

$$q_m^*\left(\mathbf{z}_{k+m-1}^{k+1}, s_k\right) = \Pr\left[z_{k+m} = 1 | \mathbf{z}_{k+m-1}^{k+1}, s_k\right] \tag{2.11a}$$

and

$$q_m(z_{k+m-1}^{k+1}) = \Pr[z_{k+m}=1|z_{k+m-1}^{k+1}]. \quad (2.11b)$$

The special cases $m=0,1$ mean that previous errors are not available, and therefore $q_0^*(z_{k-1}^{k+1}, s_k)$ and $q_1(z_{k-1}^{k+1})$ are denoted, for simplicity, by

$$q_0^*(s_k) \triangleq \Pr[z_k=1|s_k] \quad (2.12a)$$

and by

$$q_1 \triangleq \Pr[z_{k+1}=1], \quad (2.12b)$$

which, due to stationarity, do not depend on k. The following proposition, proved in Appendix I, gives recursions for these functions.

Proposition 1: The following recursions hold:

$$q_{l+1}^*(z_l, s_0) = \nu(z_l, q_l^*(z_{l-1}, s_0)) \quad (2.13)$$

and

$$q_{l+1}(z_l) = \nu(z_l, q_l(z_{l-1})) \quad (2.14)$$

where the function $\nu(\cdot,\cdot)$ is defined by

$$\nu(0,\mathrm{q}) \triangleq \begin{cases} p_G + b(p_B - p_G) + \mu(\mathrm{q} - p_G)[(1-p_B)/(1-\mathrm{q})], & p_B \neq 1 \\ (1-b)p_G + b, & p_B = 1, \mathrm{q} \neq 1 \end{cases} \quad (2.15)$$

and by

$$\nu(1,\mathrm{q}) \triangleq \begin{cases} p_G + b(p_B - p_G) + \mu(\mathrm{q} - p_G)(p_B/\mathrm{q}), & p_G \neq 0 \\ (1-g)p_B, & p_G = 0, \mathrm{q} \neq 0 \end{cases} \quad (2.16)$$

for $p_G \le \mathrm{q} \le p_B$. The initial values for these recursions

$$q_0^*(s_0) = \begin{cases} p_G, & s_0 = G \\ p_B, & s_0 = B \end{cases} \quad (2.17)$$

and

$$q_1(\rho p_G + p_B)(\rho+1)^{-1} \quad (2.18)$$

follow directly from Definitions 1 and 2.

Considering $\{q_l^*\}_{l=0}^{\infty}$ and $\{q_l\}_{l=1}^{\infty}$ as random sequences, the following holds.

Corollary: a) The sequence $\{q_l^*\}_{l=0}^{\infty}$ is a Markov process with initial distribution

$$\Pr[q_0^* = \alpha] = \begin{cases} \rho/(\rho+1), & \alpha = p_G \\ 1/(\rho+1), & \alpha = p_B \end{cases} \quad (2.19)$$

and transition probabilities

$$\Pr[q_{l+1}^* = \alpha | q_l^* = \beta] = \begin{cases} 1-\beta, & \alpha = \nu(0,\beta) \\ \beta, & \alpha = \nu(1,\beta) \end{cases} \quad (2.20)$$

where $\nu(\cdot,\cdot)$ is as in (2.15) and (2.16). b) The sequence $\{q_l\}_{l=1}^{\infty}$ is also a Markov process with the same transition probabilities (2.20) and with the initial distribution

$$\Pr[q_l = (\rho p_G + p_B)(\rho+1)^{-1}] = 1. \quad (2.21)$$

Proof: We prove here part a), the proof of part b) being similar. Equation (2.13) implies that for a given specific realization of q_l^*, the random variable q_{l+1}^* has exactly two possible outcomes, $\nu(0, q_l^*)$ and $\nu(1, q_l^*)$. Take any realization of (z_{l-1}, s_0) such that $q_l^*(z_{l-1}, s_0)$ is the above specific realization of q_l^*. Then

$$\Pr[q_{l+1}^* = \nu(1, q_l^*)|z_{l-1}, s_0] = \Pr[z_l = 1|z_{l-1}, s_0] = q_l^*. \quad (2.22)$$

Thus, when conditioned on q_l^*, q_{l+1}^* is independent of (z_{l-1}, s_0) and therefore of $\boldsymbol{q}_{l-1}^*$, which proves the Markov property of the process. The transition probabilities (2.20) follow directly from (2.22), and the initial distribution (2.19) follows from (2.4), (2.6), and (2.7).

Proposition 2: Let $f(\cdot)$ be continuous over $[p_G, p_B]$. Then the following limits exist and are equal:

$$\lim_{l\to\infty} E[f(q_l^*)] = \lim_{l\to\infty} E[f(q_l)]. \quad (2.23)$$

This proposition is proved in Appendix II.

In particular, let $\Phi_n^*(\cdot)$ and $\Phi_n(\cdot)$ be the characteristic functions of q_l^* and q_l, respectively. Then there exists $\Phi(\cdot)$ such that of all ω,

$$\lim_{l\to\infty} \Phi_l^*(\omega) = \lim_{l\to\infty} \Phi_l(\omega) = \Phi(\omega). \quad (2.24)$$

Corollary: Let $F_l^*(\cdot)$ and $F_l(\cdot)$ be the probability distribution functions of q_l^* and q_l, respectively. Then there exists a probability distribution function $F(\cdot)$ to which both sequences converge weakly, i.e.,

$$\lim_{l\to\infty} F_l^*(\mathrm{q}) = \lim_{l\to\infty} F_l(\mathrm{q}) = F(\mathrm{q}), \quad (2.25)$$

at each q where the limiting distribution $F(\mathrm{q})$ is continuous. The proof follows from (2.24) and [8, p. 203]. Parenthetically, it can be shown that a Markov process with transition probabilities (2.20) and any initial probability that vanishes outside $[p_G, p_B]$ will have the same limiting distribution $F(\cdot)$.

The probability distributions of q_l^* and q_l can be calculated recursively using (2.19)–(2.21). To avoid the exponential increase in the number of the computations, the q axis can be quantized in view of the continuity of $\nu(z,\mathrm{q})$ and of $\Pr[q_{l+1}|q_l = \mathrm{q}]$ in q. Fig. 4(a) is an example of the results of such a calculation with 50 quantization levels. For $l \ge 10$ the difference between successive probability functions is below the resolution of the graph, indicating that within these quantization levels q_l has practically reached its limiting distribution.

Proposition 3: Let $f(\cdot)$ be convex C over $[p_G, p_B]$. Then the sequence $\{E[f(q_l)]\}_{l=1}^{\infty}$ is monotonically increasing, the sequence $\{E[f(q_l^*)]\}_{l=0}^{\infty}$ is monotonically decreasing,

$$E[f(q_l)] \le E[f(q_{l+1})] \le E[f(q_{l+1}^*)] \le E[f(q_l^*)], \quad (2.26)$$

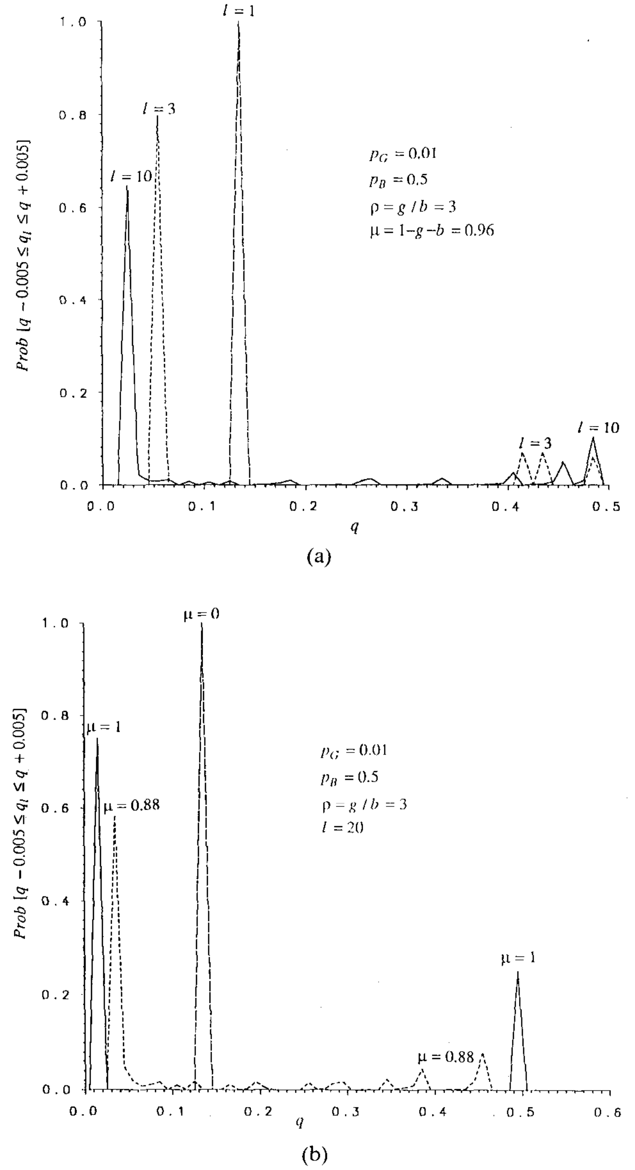

Fig. 4. Distribution of random variables q_l. (a) For various values of temporal channel index l. (b) For various values of channel memory μ.

and both converge to the same limit

$$\lim_{l \to \infty} E[f(q_l^*)] = \lim_{l \to \infty} E[f(q_l)]. \tag{2.27}$$

Inequalities (2.26) are proved in Appendix III. The convexity of $f(\cdot)$ implies its continuity over $[p_G, p_B]$, and therefore (2.27) follows from Proposition 2.

B. Evaluation of Capacity

Having established the basic properties of q_l and q_l^*, we turn to the problem of evaluating the capacity. The issues involved in the definition of the capacity of a finite-state time-varying channel are discussed in [2, pp. 97–111]. By (2.8) the Gilbert–Elliott channel is indecomposable for $-1 < \mu < 1$. Thus its capacity C can be formulated in terms of input and output sequences as follows:

$$C = \lim_{l \to \infty} \frac{1}{l} \max_{P(\boldsymbol{x}_l)} I(\boldsymbol{x}_l; \boldsymbol{y}_l) \tag{2.28}$$

where $I(\cdot\,;\cdot)$ denotes mutual information and where the maximum is taken over all possible probability functions $P(\boldsymbol{x}_l)$ of the input sequence $\boldsymbol{x}_l$. The capacity in the singular case $\mu = -1$ is still covered by (2.28), even though the channel is not indecomposable in the sense of [2, p. 105], because the equality $\underline{C} = \overline{C}$ [2] holds in this case as well, as can easily be recognized.

Proposition 4: The capacity of the Gilbert–Elliott channel, in bits per channel use, is given by

$$C = 1 - \lim_{l \to \infty} E[h(q_l)] = 1 - \lim_{l \to \infty} E[h(q_l^*)] \tag{2.29}$$

where $h(\cdot)$ is the binary entropy function

$$h(\mathrm{q}) \triangleq -\mathrm{q}\log\mathrm{q} - (1-\mathrm{q})\log(1-\mathrm{q}). \tag{2.30}$$

For the proof see Appendix IV. Successive approximations to the limit C can be calculated using the recursions for q_l and q_l^* given in (2.19) to (2.21). A bound on the truncation error, evident from (2.26) and the convexity of $h(\cdot)$, is readily calculable from the value of $E[h(q_l)] - E[h(q_l^*)]$.

We proceed to investigate the dependence of the capacity on the memory μ when p_G, p_B, and ρ are fixed and use the explicit notation C_μ. Notice that the three fixed parameters are one-dimensional statistics of the channel and, due to stationarity, invariant under interleaving. Propositions 3 and 4 imply that

$$C^{\mathrm{NM}} \triangleq 1 - h(q_1) \le C_\mu \le 1 - E[h(q_0^*)] \triangleq C^{\mathrm{SI}} \tag{2.31}$$

where the superscripts NM and SI to the capacity stand for "no memory" and "side information," respectively. It follows from definition (2.12) that the lower and upper bounds C^{NM} and C^{SI} are also one-dimensional statistics of the channel and therefore independent of μ and invariant under interleaving. Observe that C^{NM} is the capacity of the "memoryless" interleaved channel in the sense discussed in the third paragraph of the Introduction. It is therefore also the capacity C_0 of the memoryless channel with the same p_G, p_B, ρ and with $\mu = 0$. Also, it follows from the definition of $q_0^*(s_k)$ in (2.12a) that C^{SI} is the capacity of a channel with the same p_G, p_B, ρ and with arbitrary μ, when side information about the current state s_k is available; this justifies the superscript SI. An example illustrating C_μ and its bounds is presented in Fig. 2. The monotonic convergence of C_μ, as $\mu \to 1$, to its upper bound C^{SI} is not specific to the example in Fig. 2 but rather is a general property which follows from the next two propositions, proven in Appendices V and VI, and their corollaries.

Proposition 5: Let p_G, p_B, and ρ be fixed, denote by $q_l^{(\mu)}, l = 1, 2, \cdots$, the random variable q_l associated with the Gilbert–Elliott channel with memory μ, and let $f(\cdot)$ be convex-U over $[p_G, p_B]$. Then $|\mu_0| \le |\mu_1|$ and $\mu_0 \mu_1 \ge 0$ imply that

$$E\left[f\left(q_l^{(\mu_0)}\right)\right] \le E\left[f\left(q_l^{(\mu_1)}\right)\right] \tag{2.32}$$

for $l = 1, 2, \cdots$.

Proposition 6: Let p_G, p_B, ρ, and $q_l^{(\mu)}$ be as in Proposition 5 and let $f(\cdot)$ be continuous over $[p_G, p_B]$. Then

$$\lim_{\mu\to 1}\lim_{l\to\infty} E\left[f\left(q_l^{(\mu)}\right)\right] = E\left[f\left(q_0^*(s_k)\right)\right] \quad (2.33)$$

for $k = 1, 2, \cdots$. The existence of the inner limit in the left side of (2.33) was established in Proposition 2.

Corollary 1: Let $F_l^\mu(\cdot)$ and $F_0{}^*(\cdot)$ denote, respectively, the probability distribution functions of $q_l^{(\mu)}$ and q_0^*. Then

$$\lim_{\mu\to 1}\lim_{l\to\infty} F_l^\mu(q) = F_0{}^*(q) \quad (2.34)$$

where the convergences are in the weak sense. The existence of the inner limit was established in the corollary to Proposition 2. The convergence of the outer limit to $F_0{}^*(\cdot)$ follows from a similar argument in which Proposition 6 is used instead of Proposition 2. See Fig. 4(b) for an example.

Corollary 2: a) If $|\mu_0| \le |\mu_1|$ and $\mu_0\mu_1 \ge 0$, then

$$C_{\mu_0} \le C_{\mu_1}. \quad (2.35)$$

b)

$$\lim_{\mu\to 1} C_\mu = C^{\text{SI}}. \quad (2.36)$$

Proof: By Proposition 4

$$C_\mu = \lim_{l\to\infty} E\left[1 - h\left(q_l^{(\mu)}\right)\right]. \quad (2.37)$$

Since $1 - h(\cdot)$ is convex and continuous over $[p_G, p_B]$, Proposition 5 implies that (2.35) holds and Proposition 6 implies that

$$\lim_{\mu\to 1} C_\mu = E\left[1 - h(q_0^*)\right]. \quad (2.38)$$

Recall that, by its definition (2.31) C^{SI} is equal to the right side of (2.38), and thus (2.36) also holds.

The monotonic convergence of C_μ to C^{SI} when the memory is persistent is intuitively satisfactory because for larger μ the expected dwell time in each state is larger and the next state can be better predicted. The quality of this prediction is asymptotically equal to that of a perfect predictor, which is equivalent to side information. As discussed before, when memory is oscillatory, μ can reach the extreme value -1 only for $\rho = 1$. In that case, (2.33), (2.34), and (2.36) hold also for $\mu \to -1$.

III. The Decision-Feedback Decoder

A general description of a communication system employing the decision-feedback decoding algorithm, as introduced in Section I, is presented in Fig. 3(a). The system is composed of a conventional encoder, block interleaver, Gilbert–Elliott channel, deinterleaver, and a decision-feedback decoder. The latter includes the metric calculator and a conventional soft-decision decoder. The interleaver operates as follows. A stream of JN encoded symbols is stored in the rows of a J by N matrix and then transmitted over the channel column by column; the deinterleaver performs the inverse operation. The temporal index l used in Section II is broken up into the indices j and n such that $l = (n-1)\cdot J + j$ where $1 \le j \le J$ and $1 \le n \le N$. Thus j and n denote the row and column indices, respectively, of the symbol transmitted at the lth channel use. For convenience we define the doubly indexed variables by the correspondence

$$w_l \leftrightarrow w(j,n) \quad (3.1)$$

where w stands for x, y, z or s. We also denote $\mathbf{w}(j) \triangleq (w(j,1),\cdots,w(j,N))$. For a Gilbert–Elliott channel with memory $|\mu| < 1$ the interleaved channel can be considered memoryless in the following sense, where the influence of μ and J is apparent. Since the state variables $s(j,n)$ and $s(j,n+1)$ are J channel uses apart, it follows from (2.8) and (2.9a) that

$$\Pr[s(j,n+1)=\xi|s(j,n)=\xi] - \Pr[s(j,n+1)=\xi|s(j,n)\ne\xi] = \mu^J \quad (3.2\text{a})$$

for $\xi \in \{G, B\}$. When conditioned on $s(j,n+1)$, $z(j,n+1)$ is independent of $s(j,n)$, $z(j,n),\cdots,z(j,1)$, and when conditioned on $s(j,n)$, $s(j,n+1)$ is independent of $z(j,n),\cdots,z(j,1)$. Therefore,

$$\begin{aligned}
&|\Pr[z(j,n+1)|z(j,n),\cdots,z(j,1)] - \Pr[z(j,n+1)]|\\
&\le \sum_{s(j,n+1)} \Pr[z(j,n+1)|s(j,n+1)]\\
&\quad\cdot\Big|\sum_{s(j,n)} \Pr[s(j,n+1)|s(j,n)]\\
&\quad\cdot\Pr[s(j,n)|z(j,n),\cdots,z(j,1)] - \Pr[s(j,n+1)]\Big|\\
&\le \sum_{s(j,n+1)} \Pr[z(j,n+1)|s(j,n+1)]\\
&\quad\cdot\max_{s(j,n)}|\Pr[s(j,n+1)|s(j,n)] - \Pr[s(j,n+1)]|\\
&\le |\mu|^J
\end{aligned} \quad (3.2\text{b})$$

and for $|\mu| < 1$ the last quantity converges to zero with J. Thus, for large enough J, the errors $\mathbf{z}(j)$ in the jth row of the interleaver matrix are effectively independent of each other. On the other hand, because the symbols $x(1,n)$, $x(2,n),\cdots,x(J,n)$ in the columns of the interleaver matrix are transmitted at consecutive channel uses, the corresponding errors $z(1,n)$, $z(2,n),\cdots,z(J,n)$ are highly dependent. This dependence is referred to as *latent* memory, and the purpose of the metric calculator is to enable its utilization by a conventional soft-decision decoder. In the singular case $\mu = -1$ the memory can be utilized with or without interleaving in a rather straightforward manner, not discussed here.

A detailed description of the metric-calculator is presented in Fig. 3(b). The comparison of the channel output $y(j-1,n)$ with the decoder decision $\hat{x}(j-1,n)$ yields an estimate $\hat{z}(j-1,n)$ of the channel error $z(j-1,n)$. We rewrite (2.11b) in the deinterleaved indexing (3.1). Let $q(j,n)$ denote the probability of a channel error in the received symbol $y(j,n)$, conditioned on the previous channel errors $z(1,n),\cdots,z(j-1,n)$ in the *same* column of the deinterleaver. Thus

$$q(j,n) \triangleq q_j(z(1,n),\cdots,z(j-1,n)). \tag{3.3}$$

The metric calculator yields the estimate $\hat{q}(j,n)$ of $q(j,n)$ by substituting the estimates $\hat{z}(1,n),\cdots,\hat{z}(j-1,n)$, instead of the actual $z(1,n),\cdots,z(j-1,n)$, in (3.3). The metric calculator has a simple, recursive implementation, as shown by Proposition 1. We also define $\boldsymbol{q}(j) \triangleq (q(j,1),\cdots,q(j,N))$ and $\hat{\boldsymbol{q}}(j) \triangleq (\hat{q}(j,1),\cdots,\hat{q}(j,N))$.

Conventional soft-decision decoders operate on metrics rather than probabilities, and therefore $y(j,n)$ and $\hat{q}(j,n)$ are combined into a log-likelihood metric $m(j,n) = m(y(j,n),\ \hat{q}(j,n))$ which is supplied to the soft-decision decoder. The function $m(\cdot,\cdot)$ is defined, for $0 \le q \le 1$, by

$$m(y,q) \triangleq \begin{cases} \log \dfrac{q}{1-q}, & \text{for } y=0 \\ \log \dfrac{1-q}{q}, & \text{for } y=1 \end{cases} \tag{3.4}$$

where the quantities $\log 0$ and $\log(1/0)$ are understood as $-\infty$ and $+\infty$, respectively.

To evaluate the performance of the communication system that employs the decision-feedback decoding algorithm, let $P_e(j)$ be the probability that the jth row is erroneously decoded, that is, $\hat{\boldsymbol{x}}(j) \neq \boldsymbol{x}(j)$. We found it impractical to obtain a direct evaluation of $P_e(j)$ because of its dependence on the random vector $\hat{\boldsymbol{q}}(j)$ which, in turn, depends on the previous decoding results. The following proposition, however, provides an upper bound on $P_e(j)$ in terms of the corresponding error probability $P_e^{ga}(j)$ over the following *genie-aided channel.*

Definition 3: For a given j, the genie-aided channel is defined by the input $x(j,n)$ and the output pair $(y(j,n),\ q(j,n))$. It will be apparent that, for given j, the channel is the same for $n=1,2,\cdots,N$, and therefore there are J different genie-aided channels, each one of them used exactly N times. Denote by $P_e^{ga}(j)$ the probability of a decoding error at the jth genie-aided channel, using the same encoder and the same soft-decision decoder.

Proposition 7: The following inequality holds

$$P_e(j) \le \frac{J\hat{P}_e}{1-J\hat{P}_e} \tag{3.5}$$

where

$$\hat{P}_e \triangleq \max_{1\le j\le J} P_e^{ga}(j). \tag{3.6}$$

This proposition is proved in Appendix VII. We establish further below that the genie-aided channel is discrete, asymptotically memoryless for deep interleaving and output-symmetric. For such channels, codes are available for which $P_e^{ga}(j)$ decreases exponentially with code length. Therefore, for small $J\hat{P}_e$, the deviation of the denominator in (3.5) from unity is negligible and an exponential behavior, similar to that of $P_e^{ga}(j)$, applies to $P_e(j)$ as well, for any fixed J.

We proceed to prove the above-claimed attributes of the genie-aided channel. Discreteness follows from the fact that $x(j,n)$ and $y(j,n)$ are binary variables while $q(j,n) \triangleq q_j(z(1,n),\cdots,z(j-1,n))$ can obtain at most 2^{j-1} different values. The asymptotic lack of memory of the jth channel over its N uses is justifiable by the following argument. The crossover probabilities $\Pr[\boldsymbol{y}(j),\boldsymbol{q}(j)|\boldsymbol{x}(j)]$ are equal to $\Pr[\boldsymbol{q}(j),\boldsymbol{z}(j)]$ which, in turn, can be decomposed as follows:

$$\begin{aligned}\Pr[\boldsymbol{q}(j),\boldsymbol{z}(j)] = \prod_{n=1}^{N} \{&\Pr[q(j,n)|q(j,n-1),\cdots,\\ &\qquad q(j,1),\ z(j,n-1),\cdots,z(j,1)]\\ &\cdot \Pr[z(j,n)|q(j,n),\ q(j,n-1),\cdots,q(j,1),\\ &\qquad z(j,n-1),\cdots,z(j,1)]\}. \end{aligned} \tag{3.7}$$

The following proposition, proved in Appendix VIII, asserts that the nth factor of the right side of (3.7) converges to $\Pr[q(j,n)]\Pr[z(j,n)|q(j,n)]$, as $J\to\infty$, establishing asymptotic memorylessness.

Proposition 8: For $1 \le j \le J$ and $1 \le n \le N$, let

$$\begin{aligned} F_{j,J,n}(q|q(j,n-1),\cdots,q(j,1),\ z(j,n-1),\cdots,z(j,1)) \\ \triangleq \Pr[q(j,n) \le q|q(j,n-1),\cdots,q(j,1),\\ z(j,n-1),\cdots,z(j,1)] \end{aligned} \tag{3.8}$$

and let

$$F_j(q) \triangleq \Pr[q(j,n) \le q]. \tag{3.9}$$

Then for any q

$$\begin{aligned}\lim_{J\to\infty} \max|F_{j,J,n}(q|q(j,n-1),\cdots,q(j,1),\\ z(j,n-1),\cdots,z(j,1)) - F_j(q)| = 0\end{aligned} \tag{3.10}$$

and

$$\begin{aligned}\lim_{J\to\infty} \max|\Pr[z(j,n)|q(j,n),\ q(j,n-1),\cdots,q(j,1),\\ z(j,n-1),\cdots,z(j,1)]\\ -\Pr[z(j,n)|q(j,n)]| = 0\end{aligned} \tag{3.11}$$

where the maxima are taken over $1 \le j \le J$, over $1 \le n \le N$ and over the domains of $q(j,n-1),\cdots,q(j,1),\ z(j,n-1),\cdots,\ z(j,1)$.

Finally, a discrete memoryless channel is called output-symmetric if its set of outputs can be partitioned into subsets in such a way that every submatrix of the transition probabilities matrix is symmetric [2, p. 94], [9, p. 86]. The genie-aided channel is output-symmetric because

$$\begin{aligned}\Pr[y(j,n),\ q(j,n)|x(j,n)]\\ = \Pr[y(j,n)|x(j,n),\ q(j,n)]\Pr[q(j,n)]\end{aligned} \tag{3.12}$$

while

$$\Pr[y(j,n)|x(j,n),\ q(j,n)] = \begin{cases} 1-q(j,n), & y(j,n)=x(j,n) \\ q(j,n), & y(j,n)\neq x(j,n) \end{cases}. \tag{3.13}$$

Having established the atributes of the genie-aided channel, we now turn to the evaluation of its information rates. The capacity and the random coding exponent of a discrete memoryless channel are defined in [2, pp. 74, 138–139]. For a binary-input output-symmetric channel these values are obtained for the uniform input distribution [9, p. 141]. Substituting (3.12), (3.13), and $\Pr[x(j,n)=0]=\Pr[x(j,n)=1]=1/2$ in the above mentioned definitions in [2] gives, for the capacity and the random coding exponent of the genie-aided channel, appropriately superscripted, the following equations:

$$C^{ga(j)}=1-E[h(q(j,n))] \tag{3.14}$$

where $h(\cdot)$ is the entropy function (2.30) and

$$E_r^{ga(j)}(R)=\max_{0\le\lambda\le 1}\{E_0^{ga(j)}(\lambda)-\lambda R\} \tag{3.15}$$

where

$$E_0^{ga(j)}(\lambda)\triangleq\lambda-\log\{E[f_e(q(j,n);\lambda)]\} \tag{3.16}$$

and

$$f_e(q;\lambda)=\{(1-q)^{1/(1+\lambda)}+q^{1/(1+\lambda)}\}^{1+\lambda}. \tag{3.17}$$

We proceed to derive properties of these descriptors of the genie-aided channel. Notice that due to the stationarity of the Gilbert–Elliott channel the distribution of the random variables $q(j,n)$ is independent of n and equal to the distribution of the random variable q_l, considered in Section II, for $l=j$. Thus the results of Section II are applicable here.

Definition 4: Let $C^{*(l)}$ and $E_0^{*(l)}$ denote the following quantities

$$C^{*(l)}\triangleq 1-E[h(q_l^*)] \tag{3.18}$$

$$E_0^{*(l)}(\lambda)\triangleq\lambda-\log\{E[f_e(q_l^*;\lambda)]\} \tag{3.19}$$

where q_l^* is as defined in (2.10a).

Proposition 9: a) The sequence $\{C^{ga(j)}\}_{j=1}^{\infty}$ increases with j, the sequence $\{C^{*(j)}\}_{j=0}^{\infty}$ decreases with j,

$$C^{ga(j)}\le C^{ga(j+1)}\le C^{*(j+1)}\le C^{*(j)} \tag{3.20}$$

and both converge to the same limit

$$\lim_{j\to\infty}C^{ga(j)}=\lim_{j\to\infty}C^{*(j)}\triangleq C^{ga}. \tag{3.21}$$

b) The same applies to $\{E_0^{ga(j)}\}_{j=1}^{\infty}$ and $\{E_0^{*(j)}\}_{j=0}^{\infty}$.

The proof follows from (3.14)–(3.19), Proposition 3, and the convexity of $h(\cdot)$ and of $f_e(\cdot;\lambda)$. The curve for $C^{ga(j)}$ in Fig. 5 is an example of this monotonic convergence.

Corollary: The capacity of the genie-aided channel equals that of the original Gilbert–Elliott channel, i.e.,

$$C^{ga}=C. \tag{3.22}$$

The proof follows from (3.21), (3.14), and (2.29).

The interpretation of this result is that by interleaving and decision-feedback decoding the problem of reliable communication over the time-varying binary-input binary-output Gilbert–Elliott channel is transformed into the problem of reliable communication over a binary-input multiple-output essentially memoryless channel that possesses the same capacity. However, the following qualifications need to be added to this statement.

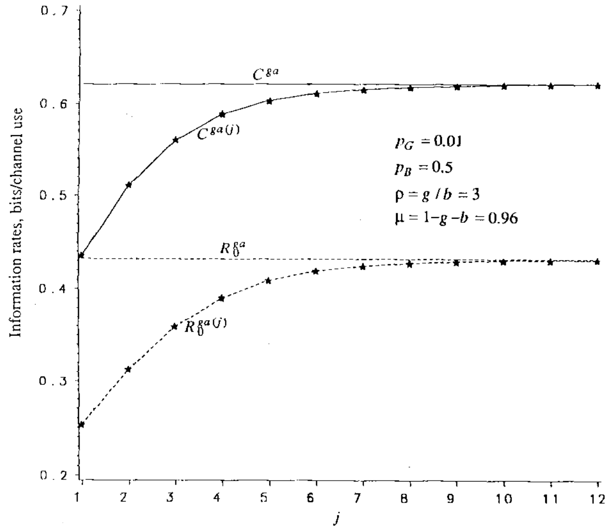

Fig. 5. Increase in capacity $C^{ga(j)}$ and cutoff rate $R_0^{ga(j)}$ of genie-aided channel, with deinterleaver row index j, to respective limits C^{ga} and R_0^{ga}.

1) Since the capacity and the random coding exponent are only upper bounds on the achievable performance, it is not obvious that a specific code will have the same performance over the multioutput channel as over a binary symmetric channel (BSC) of similar capacity. It is for the latter channel that code performance data are generally available.

2) The advantage of the decision-feedback decoder over a conventional decoder depends on the proper utilization of the calculated metrics by the incorporated soft-decision decoder. Standard soft-decision decoders have a small number of predetermined weights which might not match the calculated metrics of the multioutput channel, thus resulting in inferior performance.

3) A decoding error causes the true value of $q(j,n)$ to change into the estimate $\hat{q}(j,n)$ and this mismatch increases the probability of further decoding errors above that of the first decoding error. However, the associated loss in performance is also negligible when the system is designed to operate at low probability of decoding errors, as shown by Proposition 7.

An important simple descriptor of the quality of a memoryless channel is the cutoff rate parameter $R_0\triangleq E_0(\lambda=1)$. In [9, p. 88] a bound for the decoding error probability of a maximum likelihood decoder is given for any binary-input output-symmetric channel and *any specific linear code* in terms of the *Bhattacharyya distance* which is shown in [9, p. 153] to be a monotonic function of the cutoff rate. For the genie-aided channel, the cutoff rate is derived from (3.15) as

$$R_0^{ga(j)}=1-\log\{1+E[f_r(q(j,n))]\} \tag{3.23}$$

where

$$f_r(q) \triangleq 2 \cdot \sqrt{q(1-q)} . \tag{3.24}$$

As shown in Proposition 9 the genie-aided channel improves monotonically with j, thus theoretically allowing for a sequence of codes of increasing rate. In practical applications, however, a simpler strategy can be used. A K-row header, with content known to the receiver, is transmitted at the beginning of each interleaver frame. This header carries, of course, no information, and its sole use is to initialize the metric calculator. Encoded information is transmitted over the remaining $J-K$ rows, using a code suitable for the $(K+1)$th row and therefore suitable for all other rows. Fig. 5 indicates that for a typical case $R_0^{\text{ga}(j)}$ converges very quickly to its limit R_0^{ga} and therefore, even for moderate J, a small K can be used with $R_0^{\text{ga}(K+1)} \approx R_0^{\text{ga}}$ such that close to optimum performance is practical, with small overhead $(J-K)/J$.

We conclude this section by investigating the dependence of the quality of the genie-aided channel on memory μ when the one-dimensional parameters p_G, p_B, and ρ are fixed. It was established in (3.22) that $C_\mu^{\text{ga}} = C_\mu$. Therefore, (2.31) implies that the genie-aided channel is uniformly superior, in terms of capacity, to the interleaved channel when the latter is considered memoryless and (2.35) implies that this advantage of the decision-feedback decoder over a conventional one increases monotonically with $|\mu|$. When $\mu \to 1$, (2.36) implies that the genie-aided channel is asymptotically as good, in terms of capacity, as the channel with side information. Turning now to the random coding exponent and the cutoff rates, we recall from (3.15), (3.16), (3.17), (3.23), and (3.24) that they are given in terms of the expected values of the functions $f_e(q;\lambda)$ and $f_r(q)$, which are continuous and concave in $q \in [p_G, p_B]$ (for $\lambda \geq 0$). Thus (2.26) implies that

$$E_r^{\text{NM}}(R) \leq E_{r,\mu}^{\text{ga}}(R) \leq E_r^{\text{SI}}(R) \tag{3.25}$$

for every $0 \leq R < C$ and that

$$R_0^{\text{NM}} \leq R_{0,\mu}^{\text{ga}} \leq R_0^{\text{SI}} \tag{3.26}$$

where $E_r^{\text{NM}}(R)$ and R_0^{NM} denote, respectively, the random coding exponent and the cutoff rate of the interleaved channel, when considered memoryless, while $E_r^{\text{SI}}(R)$ and R_0^{SI} denote the respective parameters for the interleaved channel with side information. It deserves mention that, while the capacity C^{SI} is invariant under interleaving, the quantities $E_r^{\text{SI}}(R)$ and R_0^{SI} as defined here apply only to interleaved Gilbert–Elliott channels. It is expected that the corresponding quantities for the original (uninterleaved) Gilbert–Elliott channel would turn out to be considerably smaller upon numerical evaluation. As with capacity, (2.32) and (2.33) imply that $E_{r,\mu}^{\text{ga}}(R)$ and $R_{0,\mu}^{\text{ga}}$ increase uniformly with $|\mu|$ and converge asymptotically, as $\mu \to 1$, to their upper bounds $E_r^{\text{SI}}(R)$ and R_0^{SI}. Fig. 2 presents a numerical example for the capacity and the cutoff rate of the genie-aided channel as a function of μ together with their lower and upper limits. In the example considered the advantage of the decision-feedback decoder over a conventional one, in terms of the cutoff rates, reaches a factor of about 2.

IV. Discussion

The key for the calculation of the capacity of the Gilbert–Elliott channel is the recursive property of the conditional probability q_l, Proposition 1. The low complexity of the proposed decision-feedback decoder that operates on the deinterleaved channel output is also due to this property. The point of view emphasized in Section III was that of transforming a time-varying binary-input binary-output channel into an essentially memoryless binary-input multiple-output channel that has the same capacity. As an alternative point of view, consider the cascade of the encoder and the interleaver as a "supercoder" and the cascade of the deinterleaver and the decision-feedback decoder as an appropriate "superdecoder." Intuitively, the length of an efficient code over time-varying channels should be large in comparison with the mean duration of the channel memory. The well-known advantage of interleaving is that codes of any desirable length can be obtained from relatively simpler short *component codes* by increasing the interleaving depth alone. From this point of view the contribution of Section III is to show that restricting the choice of codes to those obtainable by interleaving component codes does not limit the achievable information rate over the Gilbert–Elliott channel, when the decision-feedback decoder is used. Furthermore, the advantage of the decision-feedback decoding algorithm is obtained without essential increase in the decoder complexity over that of a standard soft-decision decoder for the component code. The component codes operate over a practically memoryless channel, and therefore their lengths are, in principle, independent of the temporal variations of the Gilbert–Elliott channel, provided only that the interleaver depth J is of appropriate length. Note that the effective code length for calculating the performance of a system employing the decision-feedback decoder is the component code length N and not the super-code length NJ. This is a satisfying result since the essential decoding complexity is a function of N only and NJ is merely a measure on the decoding delay.

A system employing a similar decision-feedback decoder has been implemented at Elisra, Electronic System, Ltd. and tested over a channel similar to the Gilbert–Elliott one. Its performance was found to be in agreement with that predicted on the basis of the cutoff rate of the genie-aided channel when applied to the specific convolutional code used.

Acknowledgment

The authors wish to thank Dr. H. Kaspi for discussions on convergence properties of Markov processes and Dr. M. Sidi for suggesting we also consider negative values of μ.

Appendix I
Proof of Proposition 1

We prove the recursion for q_l^*, when $p_G \neq 0$ and $p_B \neq 1$; the proof for q_l is similar. The verification of (2.15) and (2.16) when $p_B = 1$ and $p_G = 0$, respectively, is simple.

By Definitions 1 and 2,

$$\begin{aligned} q_{l+1}^*(z_l, s_0) &= \Pr[z_{l+1} = 1 | z_l, s_0] \\ &= p_G \Pr[s_{l+1} = G | z_l, s_0] + p_B \Pr[s_{l+1} = B | z_l, s_0] \\ &= p_G + (p_B - p_G) \Pr[s_{l+1} = B | z_l, s_0] \end{aligned} \quad \text{(A.1a)}$$

and

$$q_{l+1}(z_l) = p_G + (p_B - p_G) \Pr[s_{l+1} = B | z_l]. \quad \text{(A.1b)}$$

When conditioned on s_l, s_{l+1} is independent of (z_l, s_0) and therefore

$$\begin{aligned} \Pr[s_{l+1} = B | z_l, s_0] &= \Pr[s_{l+1} = B | s_l = G] \Pr[s_l = G | z_l, s_0] \\ &\quad + \Pr[s_{l+1} = B | s_l = B] \Pr[s_l = B | z_l, s_0] \\ &= b \Pr[s_l = G | z_l, s_0] + (1 - g) \Pr[s_l = B | z_l, s_0] \\ &= b + (1 - g - b) \Pr[s_l = B | z_l, s_0] = b + \mu \Pr[s_l = B | z_l, s_0]. \end{aligned} \quad \text{(A.2)}$$

Substituting (A.2) into (A.1) gives

$$q_{l+1}^*(z_l, s_0) = p_G + b(p_B - p_G) + \mu(p_B - p_G) \Pr[s_l = B | z_l, s_0]. \quad \text{(A.3)}$$

By Bayes' rule

$$\Pr[s_l | z_l, s_0] = \frac{\Pr[s_l, z_l | z_{l-1}, s_0]}{\Pr[z_l | z_{l-1}, s_0]} = \frac{\Pr[z_l | s_l] \Pr[s_l | z_{l-1}, s_0]}{\Pr[z_l | z_{l-1}, s_0]}. \quad \text{(A.4)}$$

Rearranging (A.1) into

$$\Pr[s_l = B | z_{l-1}, s_0] = \frac{q_l^* - p_G}{p_B - p_G} \quad \text{(A.5)}$$

and substituting (A.5) into (A.4) gives

$$\Pr[s_l = B | z_{l-1}, s_0] = \frac{\Pr[z_l | s_l = B](q_l^* - p_G)}{\Pr[z_l | z_{l-1}, s_0](p_B - p_G)}. \quad \text{(A.6)}$$

Substituting (A.6) into (A.3) gives

$$\begin{aligned} q_{l+1}^*(z_l, s_0) &= p_G + b(p_B - p_G) + \mu(q_l^* - p_G) \frac{\Pr[z_l | s_l = B]}{\Pr[z_l | z_{l-1}, s_0]} \\ &= \nu(z_l, q_l^*(z_{l-1}, s_0)) \end{aligned} \quad \text{(A.7)}$$

where $\nu(\cdot, \cdot)$ is as defined in (2.15) and (2.16).

Appendix II
Proof of Proposition 2

We prove here that

$$\lim_{l \to \infty} \{ E[f(q_{l+k})] - E[f(q_l)] \} = 0 \quad \text{(A.8)}$$

uniformly over $k = 1, 2, \cdots$ and that

$$\lim_{l \to \infty} \{ E[f(q_l^*)] - E[f(q_l)] \} = 0 \quad \text{(A.9)}$$

for $f(\cdot)$ continuous over $[p_G, p_B]$. The proof is given first for $|\mu| < 1$ and the singular case $\mu = -1$ is discussed later.

Lemma A.1: For a Gilbert–Elliott channel, with $0 < \rho < \infty$ and $|\mu| < 1$, there exist $0 \leq \epsilon < 1$ such that

$$|\Pr[s_L = \xi | z_L, s_0 = \xi] - \Pr[s_L = \xi | z_L, s_0 \neq \xi]| \leq \epsilon^L \quad \text{(A.10)}$$

where $\xi \in \{G, B\}$, $L = 1, 2, \cdots$ and $z_L \in \{0,1\}^L$, provided that sample path z_L is consistent with either initial state, that is $\Pr[z_L, s_0 = G] \neq 0$ and $\Pr[z_L, s_0 = B] \neq 0$.

Proof: Notice that

$$\begin{aligned} &\Pr[s_l = \xi | z_L, s_0 = \xi] - \Pr[s_l = \xi | z_L, s_0 \neq \xi] \\ &\quad = \Pr[s_l \neq \xi | z_L, s_0 \neq \xi] - \Pr[s_l \neq \xi | z_L, s_0 = \xi] \end{aligned} \quad \text{(A.11)}$$

and therefore

$$\begin{aligned} &\Pr[s_l = \xi | z_L, s_0 = \xi] - \Pr[s_l = \xi | z_L, s_0 \neq \xi] \\ &= \Pr[s_l = \xi | z_L, s_{l-1} = \xi] \Pr[s_{l-1} = \xi | z_L, s_0 = \xi] \\ &\quad + \Pr[s_l = \xi | z_L, s_{l-1} \neq \xi] \Pr[s_{l-1} \neq \xi | z_L, s_0 = \xi] \\ &\quad - \Pr[s_l = \xi | z_L, s_{l-1} = \xi] \Pr[s_{l-1} = \xi | z_L, s_0 \neq \xi] \\ &\quad - \Pr[s_l = \xi | z_L, s_{l-1} \neq \xi] \Pr[s_{l-1} \neq \xi | z_L, s_0 \neq \xi] \\ &= (\Pr[s_l = \xi | z_L, s_{l-1} = \xi] - \Pr[s_l = \xi | z_L, s_{l-1} \neq \xi]) \\ &\quad \cdot (\Pr[s_{l-1} = \xi | z_L, s_0 = \xi] - \Pr[s_{l-1} = \xi | z_L, s_0 \neq \xi]). \end{aligned} \quad \text{(A.12)}$$

Thus by induction

$$\Pr[s_L = \xi | z_L, s_0 = \xi] - \Pr[s_L = \xi | z_L, s_0 \neq \xi] = \prod_{l=1}^{L} A_l(z_L) \quad \text{(A.13)}$$

where

$$A_l(z_L) \triangleq \Pr[s_l = \xi | z_L, s_{l-1} = \xi] - \Pr[s_l = \xi | z_L, s_{l-1} \neq \xi] \quad \text{(A.14)}$$

for $l = 1, 2, \cdots, L$. The fact that $A_l(z_L)$ does not depend on the value of ξ is evident from (A.11). We proceed to show that there exists $0 \leq \epsilon < 1$ such that $|A_l(z_L)| < \epsilon$ uniformly in L, l and z_L. When conditioned on s_{l-1} and s_{l+1}, s_l is independent of z_{l-1}^1 and z_L^{l+1} and therefore

$$\begin{aligned} A_l(z_L) &= \sum_{s_{l+1}} \Pr[s_l = \xi | s_{l-1} = \xi, s_{l+1}, z_l] \Pr[s_{l+1} | s_{l-1} = \xi, z_L] \\ &\quad - \sum_{s_{l+1}} \Pr[s_l = \xi | s_{l-1} \neq \xi, s_{l+1}, z_l] \Pr[s_{l+1} | s_{l-1} \neq \xi, z_L]. \end{aligned} \quad \text{(A.15)}$$

The expected value of a random variable is always bounded by its extreme values. Therefore, $A_l(z_L)$, which is given in (A.15) as the difference between two expected values, is bounded, for any ξ and any z_L, by

$$|A_l(z_L)| \leq \max_{\xi_1, \xi_2} |U(\xi, \xi_1, \xi_2, z_l)| \quad \text{(A.16)}$$

where

$$U(\xi, \xi_1, \xi_2, z_l) \triangleq V(\xi, \xi_1, z_l) - W(\xi, \xi_2, z_l) \quad \text{(A.17)}$$

$$V(\xi, \xi_1, z_l) \triangleq \Pr[s_l = \xi | s_{l-1} = \xi, s_{l+1} = \xi_1, z_l] \quad \text{(A.18)}$$

and

$$W(\xi, \xi_2, z_l) \triangleq \Pr[s_l = \xi | s_{l-1} \neq \xi, s_{l+1} = \xi_2, z_l]. \quad \text{(A.19)}$$

Notice that, due to the stationarity of the channel, U, V, and W do not depend on the index l. We proceed to show that $|U(\xi, \xi_1, \xi_2, z_l)| < 1$ for any ξ, ξ_1, ξ_2, and any z_l by showing that the alternative contradicts the assumptions $0 < \rho < \infty$ and $|\mu| < 1$, stated in the lemma. Let $U(\xi, \xi_1, \xi_2, z_l) = 1$. Then $V(\xi, \xi_1, z_l) = 1$

and $W(\xi,\xi_2,z_l)=0$. By definition (A.18),

$$V(\xi,\xi_1,z_l)=\frac{\Pr[z_l|s_l=\xi]\Pr[s_l=\xi|s_{l-1}=\xi,s_{l+1}=\xi_1]}{\sum_{s_l}\Pr[z_l|s_l]\Pr[s_l|s_{l-1}=\xi,s_{l+1}=\xi_1]} \tag{A.20}$$

and therefore $V(\xi,\xi_1,z_l)=1$ implies $\Pr[s_l\neq\xi|s_{l-1}=\xi,s_{l+1}=\xi_1]=0$. The assumption that $0<\rho<\infty$ implies that $g,b\neq 0$ and by (2.2)

$$\Pr[s_l|s_{l-1},s_{l+1}]=\frac{\Pr[s_l|s_{l-1}]\Pr[s_{l+1}|s_l]}{\Pr[s_{l+1}|s_{l-1}]}. \tag{A.21}$$

Thus $\Pr[s_l\neq\xi|s_{l-1}=\xi,s_{l+1}=\xi_1]=0$ implies that $\xi_1\neq\xi$ and that $\Pr[s_{l+1}\neq\xi|s_l\neq\xi]=0$. The assumption that $V(\xi,\xi_1,z_l)=1$ also implies that $\Pr[z_l|s_l=\xi]\neq 0$ and therefore $W(\xi,\xi_2,z_l)=0$ implies that $\Pr[s_l=\xi|s_{l-1}\neq\xi,s_{l+1}=\xi_2]=0$, resulting in $\xi_2=\xi$ and $\Pr[s_{l+1}=\xi|s_l=\xi]=0$. Thus $U(\xi,\xi_1,\xi_2,z_l)=1$ implies that $g=b=1$, contradicting the assumption that $|\mu|<1$. In a similar way $U(\xi,\xi_1,\xi_2,z_l)=-1$ also leads to the same contradiction. Therefore, $\epsilon\triangleq\max_{\xi,\xi_1,\xi_2,z_l}|U(\xi,\xi_1,\xi_2,z_l|<1$ and the lemma follows from (A.13) and (A.16). Q.E.D.

We proceed to prove (A.9). From the lemma

$$\lim_{l\to\infty}|\Pr[s_l|z_l,s_0]-\Pr[s_l|z_l]|=0 \tag{A.22}$$

uniformly over the domains of all variables involved, provided that $\Pr[z_{l-1},s_0]\neq 0$. Therefore, (A.1) implies that

$$\lim_{l\to\infty}\{q_l^*(z_{l-1},s_0)-q_l(z_{l-1})\}=0 \tag{A.23}$$

uniformly over the domains of $\{z_l\}_{l=1}^{\infty}$ and s_0, provided that $\Pr[z_{l-1},s_0]\neq 0$. By assumption $f(\cdot)$ is uniformly continuous over the compact domain $[p_G,p_B]$. Since $p_G\le q_l^*\le p_B$ and $p_G\le q_l\le p_B$, (A.23) implies that

$$\lim_{l\to\infty}\{f(q_l^*(z_{l-1},s_0))-f(q_l(z_{l-1}))\}=0 \tag{A.24}$$

uniformly as above and (A.9) follows.

For the proof of (A.8) notice that

$$\begin{aligned}q_{k+l}(z_{k+l-1})&=\Pr[z_{k+l}=1|z_{k+l-1}]\\&=E\left[\Pr\left[z_{k+l}=1|z_{k+l-1}^{k+1},s_k\right]|z_{k+l-1}\right]\\&=E\left[q_l^*\left(z_{k+l-1}^{k+1},s_k\right)|z_{k+l-1}\right]\end{aligned} \tag{A.25}$$

Combining (A.23) with (A.25) gives

$$\lim_{l\to\infty}\left\{q_{k+l}(z_{k+l-1})-q_l\left(z_{k+l-1}^{k+1}\right)\right\}=0 \tag{A.26}$$

uniformly over the domains of k and of $\{z_i\}_{i=1}^{\infty}$ and (A.8) follows.

In the singular case $\mu=-1$ the current state s_l is determined by the initial state s_0. Furthermore, $\mu=-1$ implies, via (2.9a) and (2.6), that $\Pr[s_0=G]=\Pr[s_0=B]$, and therefore the distribution of q_l^* does not depend on the index l; this fixes the first term in (A.9) to a constant. Considering the second term, it can be shown, following the same reasoning as in Appendix VI that when $\mu=-1$,

$$\lim_{l\to\infty}E[f(q_l)]=E[f(q_0^*)], \tag{A.27}$$

satisfying (A.9). The first term in (A.8) converges to the same limit as the second one by the same argument.

Appendix III
Proof of Proposition 3

From Definition 2 it follows that

$$\begin{aligned}E[q_{l+1}(z_l)|z_l^2]&=E\left[\Pr[z_{l+1}=1|z_l]|z_l^2\right]\\&=\Pr\left[z_{l+1}=1|z_l^2\right]\\&=q_l\left(z_l^2\right).\end{aligned} \tag{A.28}$$

The left inequality in (2.26) is then proved as follows:

$$\begin{aligned}E[f(q_l(z_{l-1}))]&=E\left[f\left(q_l\left(z_l^2\right)\right)\right]\\&=E\left[f\left(E\left[q_{l+1}(z_l)|z_l^2\right]\right)\right]\\&\le E\left[E\left[f(q_{l+1}(z_l))|z_l^2\right]\right]\\&=E\left[f(q_{l+1}(z_l))\right].\end{aligned} \tag{A.29}$$

The first equality follows from definition (2.11b) and the stationarity of the channel while the second equality follows from (A.28). The inequality is an application of Jensen's inequality to the assumed convexity of $f(\cdot)$ over $[p_G,p_B]$. The proofs of the middle and the right inequalities in (2.26) are similar to the proof in (A.29), using

$$q_{l+1}(z_l)=E[q_{l+1}^*(z_l,s_0)|z_l] \tag{A.30}$$

and

$$q_{l+1}^*(z_l,s_0)=E\left[q_l^*\left(z_l^2,s_1\right)|z_l,s_0\right], \tag{A.31}$$

respectively, instead of (A.28). Equality (A.30) follows directly from definition (2.10b). The proof of (A.31) is similar to the one in (A.28).

Appendix IV
Proof of Proposition 4

For the proof of the left equality in (2.29) notice that

$$I(x_l;y_l)=H(y_l)-H(y_l|x_l)=H(y_l)-H(z_l) \tag{A.32}$$

where $H(\cdot)$ denoted the entropy. $H(y_l)$ achieves its maximum when x_l is uniformly distributed and $H(z_l)$ does not depend on the distribution of x_l. Therefore,

$$C=\lim_{l\to\infty}\frac{1}{l}\max_{p(x_l)}I(x_l;y_l)=1-\lim_{l\to\infty}\frac{1}{l}H(z_l). \tag{A.33}$$

Furthermore,

$$H(z_l)=\sum_{i=1}^{l}H(z_i|z_{i-1})=\sum_{i=1}^{l}E[h(q_i)] \tag{A.34}$$

where $h(\cdot)$ is as defined in (2.30). Since $h(\cdot)$ is convex over $[p_G,p_B]$, Proposition 3 implies that $\{E[h(q_i)]\}_{i=1}^{\infty}$ is monotonically decreasing and therefore

$$C=1-\lim_{l\to\infty}\frac{1}{l}\sum_{i=1}^{l}E[h(q_i)]=1-\lim_{l\to\infty}E[h(q_l)]. \tag{A.35}$$

The right equality in (2.29) follows from the left equality and (2.27).

Appendix V
Proof of Proposition 5

Let $\{C^{(i)}\}_{i=1}^{\infty}$ be a set of statistically independent Gilbert–Elliott channels with *identical* parameters p_G, p_B, ρ, and $\mu=\mu_1$, and let $\{s_l^{(i)}\}_{l=0}^{\infty}$ and $\{z_l^{(i)}\}_{l=1}^{\infty}$ denote their state and error processes, respectively. We first show that from these processes one can construct a pair of processes $\{\tilde{s}_l\}_{l=0}^{\infty}$ and $\{\tilde{z}_l\}_{l=1}^{\infty}$, which can be interpreted as the state and the error processes, respectively, of a Gilbert–Elliott channel with parameters p_G, p_B, ρ and $\mu=\mu_0$. The proposition is then proved by showing that for $i=1,2,\cdots$ and $l=1,2,\cdots$

$$E\left[f\left(\Pr[\tilde{z}_l=1|\tilde{z}_{l-1}]\right)\right] \le E\left[f\left(\Pr\left[z_l^{(i)}=1|z_{l-1}^{(i)}\right]\right)\right]. \quad (A.36)$$

For the construction of $\{\tilde{s}_l\}_{l=0}^{\infty}$ and $\{\tilde{z}_l\}_{l=1}^{\infty}$, let $\{\omega_l\}_{l=0}^{\infty}$, $\omega_l \in \{1,2,\cdots\}$, denote the Markov process with the initial distribution $\Pr[\omega_0=1]=1$ and with the index-independent transition probabilities

$$\Pr[\omega_{l+1}|\omega_l] = \begin{cases} \alpha, & \text{for } \omega_{l+1}=\omega_l \\ 1-\alpha, & \text{for } \omega_{l+1}=\omega_l+1 \end{cases} \quad (A.37)$$

where

$$\alpha = \mu_0/\mu_1. \quad (A.38)$$

By assumption, $|\mu_0| \le |\mu_1|$ and $\mu_0\mu_1 \ge 0$ and therefore $0 \le \alpha \le 1$ is a valid probability. The process $\{\omega_l\}_{l=0}^{\infty}$ is statistically independent of $\{s_l^{(i)}\}_{l=0}^{\infty}$ and of $\{z_l^{(i)}\}_{l=1}^{\infty}$, $i=1,2\cdots$. We now define $\tilde{s}_l \triangleq s_l^{(\omega_l)}$ for $l=0,1,\cdots$ and $\tilde{z}_l \triangleq z_l^{(\omega_l)}$ for $l=1,2,\cdots$. Since the processes $\{s_l^{(i)}\}_{l=0}^{\infty}$, $i=1,2,\cdots$, are statistically independent, identically distributed and statistically independent of $\{\omega_l\}_{l=0}^{\infty}$, it follows from the construction that

$$\begin{aligned}\Pr[\tilde{s}_l=G|\tilde{s}_{l-1}=B,\tilde{s}_{l-2}] &= \sum_{j=1}^{\infty}\Pr\left[s_l^{(j)}=G|s_{l-1}^{(j)}=B\right]\Pr[\omega_l=j,\omega_{l-1}=j] \\ &\quad+\sum_{j=1}^{\infty}\Pr\left[s_l^{(j+1)}=G\right]\Pr[\omega_l=j+1,\omega_{l-1}=j] \\ &= g_1\alpha+\rho(\rho+1)^{-1}(1-\alpha) \\ &= (1-\mu_0)\rho(\rho+1)^{-1} \triangleq g_0 \end{aligned} \quad (A.39)$$

(which is the definition of g_0) where

$$g_1 \triangleq \Pr\left[s_l^{(i)}=G|s_{l-1}^{(i)}=B\right]=(1-\mu_1)\rho(\rho+1)^{-1}. \quad (A.40)$$

In a similar way it follows that

$$\Pr[\tilde{s}_l=B|\tilde{s}_{l-1}=G,\tilde{s}_{l-2}]=(1-\mu_0)(\rho+1)^{-1} \triangleq b_0 \quad (A.41)$$

which is the definition of b_0. Thus $\{s_l\}_{l=0}^{\infty}$ satisfies (2.3). Furthermore, $g_0/b_0=\rho$ and $1-g_0-b_0=\mu_0$. It is also obvious from the construction that $\{\tilde{z}_l\}_{l=1}^{\infty}$ and $\{\tilde{s}_l\}_{l=0}^{\infty}$ satisfy (2.2) with parameters p_G and p_B. Thus $\{\tilde{s}_l\}_{l=0}^{\infty}$ and $\{\tilde{z}_l\}_{l=1}^{\infty}$ can be interpreted as the state and the error processes, respectively, of a Gilbert–Elliott channel with the required parameters, as introduced at the beginning of the proof. To prove inequality (A.36) we use the following sequence which holds for any $l,i=1,2,\cdots$,

$$\begin{aligned} E[f(\Pr[\tilde{z}_l=1|\tilde{z}_{l-1}])] &= E\left[f\left(E\left[\Pr\left[z_l^{(\omega_l)}=1|\{z_{l-1}^{(j)}\}_{j=1}^{\infty},\omega_l\right]|\tilde{z}_{l-1}\right]\right)\right] \\ &\le E\left[f\left(\Pr\left[z_l^{(\omega_l)}=1|\{z_{l-1}^{(j)}\}_{j=1}^{\infty},\omega_l\right]\right)\right] \\ &= E\left[f\left(\Pr\left[z_l^{(\omega_l)}=1|z_{l-1}^{(\omega_l)},\omega_l\right]\right)\right] \\ &= E\left[f\left(\Pr\left[z_l^{(i)}=1|z_{l-1}^{(i)}\right]\right)\right]. \end{aligned} \quad (A.42)$$

The first equality follows from the fact that by definition $\tilde{z}_l=z_l^{(\omega_l)}$ and the fact that $\tilde{z}_{l-1}$ is determined by $\{z_{l-1}^{(j)}\}_{j=1}^{\infty}$ and ω_{l-1}. The inequality is an application of Jensen's inequality to the assumed convexity of $f(\cdot)$ over $[p_G,p_B]$. The second equality follows from the fact that, when conditioned on $z_{l-1}^{(\omega_l)}$, $z_l^{(\omega_l)}$ is independent of $\{z_{l-1}^{(j)}\}_{j=1,j\ne\omega_l}^{\infty}$. The last equality follows from the fact that all the processes $\{z_l^{(i)}\}_{l=1}^{\infty}$, $i=1,2,\cdots$ have identical statistics.

Appendix VI
Proof of Proposition 6

We assume here that $p_G<p_B$; if $p_G=p_B$ then the proposition holds trivially. For simplicity, the explicit dependence of the random variables on the value of μ will be omitted.Define

$$\alpha_l(z_{l-1}) \triangleq \Pr[s_l=B|z_{l-1}] \quad (A.43)$$

and

$$a_l^*(s_l) \triangleq \begin{cases}1, & s_l=B\\ 0, & s_l=G\end{cases} \quad (A.44)$$

for $l=1,2,\cdots$. Also define

$$\hat{a}_l^{(N)}(z_{l-1}) \triangleq \begin{cases}1, & \sum_{j=l-N}^{l-1} z_j \ge T_N\\ 0, & \text{otherwise}\end{cases} \quad (A.45)$$

for $N=1,2,\cdots$ and for $l=N+1,N+2,\cdots$, where

$$T_N \triangleq N(p_G+p_B)/2. \quad (A.46)$$

Lemma A.2: If $p_G<p_B$ then for $\alpha\in\{0,1\}$

$$\lim_{N\to\infty}\lim_{\mu\to1}\Pr\left[\hat{a}_l^{(N)}=\alpha|a_l^*=\alpha\right]=1 \quad (A.47)$$

and

$$\lim_{N\to\infty}\lim_{\mu\to1}\Pr\left[a_l^*=\alpha|\hat{a}_l^{(N)}=\alpha\right]=1. \quad (A.48)$$

Notice from (A.45) that $\hat{a}_l^{(N)}$ is defined only for $l>N$ and that the statistics of $\hat{a}_l^{(N)}$ and of a_l^* are independent of l, due to the stationarity of the channel.

Proof: It follows from (A.45) and from the assumption $p_G<p_B$ that

$$\lim_{N\to\infty}\Pr\left[\hat{a}_l^{(N)}=a_l^*(\xi)|s_l=s_{l-1}=\cdots s_{l-N}=\xi\right]=1 \quad (A.49)$$

for $\xi\in\{G,B\}$. Recall that $\mu\triangleq1-g-b$, and therefore $\mu\to1$ implies $g,b\to0$. It thus follows from the definition of $\{s_l\}_{l=0}^{\infty}$ (Definition 1) that for $N=1,2,\cdots$ and $\xi\in\{G,B\}$

$$\lim_{\mu\to1}\Pr[s_{l-1}=\cdots=s_{l-N}=\xi|s_l=\xi]=1 \quad (A.50)$$

and thus (A.47) follows from (A.49). For the proof of (A.48) note that since $0<\rho<\infty$, $\Pr[a_l^*=\alpha]>0$ for $\xi\in\{0,1\}$ and therefore (A.47) implies that

$$\lim_{N\to\infty}\lim_{\mu\to1}\frac{\Pr\left[\hat{a}_l^{(N)}=\alpha|a_l^*=\alpha\right]}{\Pr\left[a_l^*=\alpha|\hat{a}_l^{(N)}=\alpha\right]} = \lim_{N\to\infty}\lim_{\mu\to1}\frac{\Pr\left[\hat{a}_l^{(N)}=\alpha\right]}{\Pr[a_l^*=\alpha]}=1. \quad (A.51)$$

Since the limit of the numerator on the left side of (A.51) is unity, the limit of the corresponding denominator is also unity.

Lemma A.3:

$$\lim_{\mu\to1}\lim_{l\to\infty}E[|a_l-a_l^*|]=0. \quad (A.52)$$

Proof:

$$E[a_l|a_l^*=\alpha] = E\left[a_l|\hat{a}_l^{(N)}=\alpha, a_l^*=\alpha\right]\Pr\left[\hat{a}_l^{(N)}=\alpha|a_l^*=\alpha\right] + E\left[a_l|\hat{a}_l^{(N)}\neq\alpha, a_l^*=\alpha\right]\Pr\left[\hat{a}_l^{(N)}\neq\alpha|a_l^*=\alpha\right]$$
$$= E\left[a_l|\hat{a}_l^{(N)}=\alpha, a_l^*=\alpha\right] + \left(E\left[a_l|\hat{a}_l^{(N)}\neq\alpha, a_l^*=\alpha\right] - E\left[a_l|\hat{a}_l^{(N)}=\alpha, a_l^*=\alpha\right]\right)\cdot\Pr\left[\hat{a}_l^{(N)}\neq\alpha|a_l^*=\alpha\right] \tag{A.53}$$

for any $\alpha\in\{0,1\}$ any $N=1,2,\cdots$ and any $l>N$ and therefore (A.47) implies that

$$\lim_{N\to\infty}\lim_{\mu\to 1}\max_{l>N}\left|E[a_l|a_l^*=\alpha]-E\left[a_l|\hat{a}_l^{(N)}=\alpha, a_l^*=\alpha\right]\right|=0. \tag{A.54}$$

In a similar way,

$$E\left[a_l|\hat{a}_l^{(N)}=\alpha\right] = E\left[a_l|\hat{a}_l^{(N)}=\alpha, a_l^*=\alpha\right] + \left(E\left[a_l|a_l^*\neq\alpha, a_l^{(N)}=\alpha\right] - E\left[a_l|a_l^*=\alpha, a_l^{(N)}=\alpha\right]\right)\cdot\Pr\left[a_l^*\neq\alpha|a_l^{(N)}=\alpha\right] \tag{A.55}$$

and therefore (A.48) implies that

$$\lim_{N\to\infty}\lim_{\mu\to 1}\max_{l>N}\left|E\left[a_l|\hat{a}_l^{(N)}=\alpha\right]-E\left[a_l|\hat{a}_l^{(N)}=\alpha, a_l^*=\alpha\right]\right|=0. \tag{A.56}$$

Combining (A.54) and (A.56) gives

$$\lim_{N\to\infty}\lim_{\mu\to 1}\max_{l>N}\left|E[a_l|a_l^*=\alpha]-E\left[a_l|a_l^{(N)}=\alpha\right]\right|=0. \tag{A.57}$$

By definitions (A.43) and (A.44),

$$E\left[a_l|\hat{a}_l^{(N)}=\alpha\right] = E\left[\Pr[a_l^*=1|z_{l-1}]|\hat{a}_l^{(N)}=\alpha\right] = E\left[a_l^*|\hat{a}_l^{(N)}=\alpha\right]. \tag{A.58}$$

Equation (A.48) implies that

$$\lim_{N\to\infty}\lim_{\mu\to 1}\max_{l>N}\left|E\left[a_l^*|\hat{a}_l^{(N)}=\alpha\right]-\alpha\right|=0 \tag{A.59}$$

for $\alpha\in\{0,1\}$ and therefore (A.58) implies that

$$\lim_{N\to\infty}\lim_{\mu\to 1}\max_{l>N}\left|E\left[a_l|\hat{a}_l^{(N)}=\alpha\right]-\alpha\right|=0 \tag{A.60}$$

for $\alpha\in\{0,1\}$. Combining (A.57) and (A.60) gives

$$\lim_{N\to\infty}\lim_{\mu\to 1}\lim_{l\to\infty}\left\{E[a_l|a_l^*=\alpha]-\alpha\right\}=0 \tag{A.61}$$

for $\alpha\in\{0,1\}$ and (A.52) follows from the fact that the random variables in (A.61) do not depend on N. Q.E.D.

We now proceed to prove (2.33). For $0\le a\le 1$ let

$$w(a)\triangleq p_G(1-a)+p_B a \tag{A.62}$$

and let

$$\tilde{f}(a)\triangleq f(w(a)). \tag{A.63}$$

Since $f(\cdot)$ is, by assumption, continuous over $[p_G,p_B]$, $\tilde{f}(\cdot)$ is continuous over [0,1]. Let $\delta>0$. By (A.52)

$$\lim_{\mu\to 1}\lim_{l\to\infty}\Pr[|a_l-a_l^*|>\delta]=0 \tag{A.64}$$

and since $\tilde{f}(\cdot)$ is uniformly continuous and bounded over [0,1],

$$\lim_{\mu\to 1}\lim_{l\to\infty}E\left[\tilde{f}(a_l)-\tilde{f}(a_l^*)\right]=0. \tag{A.65}$$

By Definitions 1, 2, (A.43), (A.44), and (A.62), $q_l=w(a_l)$ and $q_0^*(s_l)=w(a_l^*)$ and thus (A.63) and (A.65) imply (2.33).

Appendix VII
Proof of Proposition 7

Let PCD(j) and NPCD(j) denote the event that all previous rows up to and not including j have been correctly decoded, and the complementary event, respectively. For a hypothetical system, which includes the genie-aided channel, let $P_e^{\mathrm{PCD}}(j)$ and $P_e^{\mathrm{NPCD}}(j)$ denote the probability of the jth row being erroneously decoded, conditioned on PCD(j) and NPCD(j), respectively. Notice that under the PCD(j) condition $\hat{\boldsymbol{q}}(j)=\boldsymbol{q}(j)$ and therefore $P_e^{\mathrm{PCD}}(j)$ is equal to the equivalent probability in the actual system. By the union bound

$$P_e(j)\le\sum_{i=1}^{j}P_e^{\mathrm{PCD}}(i). \tag{A.66}$$

We proceed to upper-bound $P_e^{\mathrm{PCD}}(i)$. Since

$$P_e^{\mathrm{ga}}(i)=P_e^{\mathrm{PCD}}(i)\Pr[\mathrm{PCD}(i)]+P_e^{\mathrm{NPCD}}(i)\Pr[\mathrm{NPCD}(i)] \ge P_e^{\mathrm{PCD}}(i)\Pr[\mathrm{PCD}(i)]\ge P_e^{\mathrm{PCD}}(i)\left[1-\sum_{k=1}^{i-1}P_e^{\mathrm{ga}}(k)\right], \tag{A.67}$$

it follows that

$$P_e^{\mathrm{PCD}}(i)\le\frac{P_e^{\mathrm{ga}}(i)}{1-\sum_{k=1}^{i-1}P_e^{\mathrm{ga}}(k)}\le\frac{\hat{P}_e}{1-J\hat{P}_e} \tag{A.68}$$

where P_e is defined in (3.6). Substituting (A.68) into (A.66) implies (3.5).

Appendix VIII
Proof of Proposition 8

The proof of (3.10) is based on the following.

Lemma A.3: Let $f(\cdot)$ be continuous over $[p_G,p_B]$, and let $|\mu|<1$. Then

$$\lim_{L\to\infty}\max_{1\le l\le L}\left|E\left[f\left(q_l\left(z_{L-1}^{L-l+1}\right)\right)|s_0\right]-E\left[f\left(q_l\left(z_{L-1}^{L-l+1}\right)\right)\right]\right|=0 \tag{A.69}$$

for $s_0\in\{G,B\}$.

Proof: It follows from (2.8), (2.9a), and the assumption $|\mu|<1$ that

$$\lim_{l\to\infty}\left|\Pr[s_l|s_0]-\Pr[s_l]\right|=0 \tag{A.70}$$

for $s_l,s_0\in\{G,B\}$ and therefore

$$\lim_{L\to\infty}\max_{1\le l\le L/2}\left|\Pr[s_{L-l}|s_0]-\Pr[s_{L-l}]\right|=0 \tag{A.71}$$

for $s_{L-l},s_0\in\{G,B\}$. When conditioned on s_{L-l}, z_{L-l}^{L-l} is independent of s_0, and therefore

$$E\left[f\left(q_l\left(z_{L-l}^{L-l+1}\right)\right)|s_0\right]=E\left[A_{L,l}(\xi)|s_0\right] \tag{A.72a}$$

and

$$E\left[f\left(q_l\left(z_{L-1}^{L-l+1}\right)\right)\right] = E\left[A_{L,l}(\xi)\right] \quad \text{(A.72b)}$$

where

$$A_{L,l}(\xi) \triangleq E\left[f\left(q_l\left(z_{L-1}^{L-l+1}\right)\right)\mid s_{L-l} = \xi\right] \quad \text{(A.73)}$$

for $1 \le l \le L$ and $\xi \in \{G, B\}$. Since $q_l \in [p_G, p_B]$ and f is continuous over this compact domain, $f(q_l)$ is bounded and therefore $A_{L,l}(\xi)$ is also bounded uniformly in L and l. Thus (A.71) and (A.72) imply that

$$\lim_{L\to\infty} \max_{1\le l\le L/2} \left| E\left[f\left(q_l\left(z_{L-1}^{L-l+1}\right)\right)\mid s_0\right] - E\left[f\left(q_l\left(z_{L-1}^{L-l+1}\right)\right)\right]\right| = 0 \quad \text{(A.74)}$$

for $s_0 \in \{G, B\}$. We proceed to extend this result for l up to L. For $l = L/2$ (A.74) implies that

$$\lim_{L\to\infty} \left| E\left[f\left(q_{L/2}\left(z_{L-1}^{L-L/2+1}\right)\right)\mid s_0\right] - E\left[f\left(q_{L/2}\left(z_{L-1}^{L-L/2+1}\right)\right)\right]\right| = 0 \quad \text{(A.75)}$$

for $s_0 \in \{G, B\}$. It follows from the uniform convergence in (A.8) and (A.26), respectively, that

$$\lim_{L\to\infty} \max_{L/2\le l\le L} \left| E\left[f\left(q_l\left(z_{L-1}^{L-l+1}\right)\right)\right] - E\left[f\left(q_{L/2}\left(z_{L-1}^{L-L/2+1}\right)\right)\right]\right| = 0 \quad \text{(A.76)}$$

and

$$\lim_{L\to\infty} \max_{L/2\le l\le L} \left| E\left[f\left(q_l\left(z_{L-1}^{L-l+1}\right)\right)\mid s_0\right] - E\left[f\left(q_{L/2}\left(z_{L-1}^{L-L/2+1}\right)\right)\mid s_0\right]\right| = 0 \quad \text{(A.77)}$$

for $s_0 \in \{G, B\}$. Combining (A.75)–(A.77) gives the equivalent of (A.74) for $L/2 \le l \le L$ and then (A.69) follows. Q.E.D.

We proceed to prove (3.10). Replacing L by J and l by j, shifting the time index by $(n-2)J + j$ and rewriting (A.69) in the deinterleaved indexing (3.1) and (3.3) results in

$$\lim_{J\to\infty} \max_{1\le j\le J} \left| E[f(q(j,n))\mid s(j,n-1)] - E[f(q(j,n))]\right| = 0 \quad \text{(A.78)}$$

for $s(j,n-1) \in \{G, B\}$. Note that $q(j,n)$ is a function of $z(1,n),\cdots,z(j-1,n)$ only and the latter, when conditioned on $s(j,n-1)$, are independent of $q(j,1),\cdots,q(j,n-1)$ and $z(j,1),\cdots,z(j,n-1)$. Therefore

$$\begin{aligned}\left|E[f(q(j,n))\mid q(j,n-1),\cdots,q(j,1),\right.&\\ z(j,n-1),\cdots,z(j,1)]&\\ = E[E[f(q,(j,n))\mid s(j,n-1)]\mid q(j,n-1),\cdots,q(j,1),&\\ z(j,n-1),\cdots,z(j,1)]&\quad \text{(A.79)}\end{aligned}$$

and (A.78) implies

$$\lim_{J\to\infty} \max_{1\le j\le J} \left| E[f(q(j,n))\mid q(j,n-1),\cdots,q(j,1), z(j,n-1),\cdots,z(j,1)] - E[f(q(j,n))]\right| = 0, \quad \text{(A.80)}$$

uniformly over the domains of $q(j,1),\cdots,q(j,n-1)$, $z(j,1),\cdots,z(j,n-1)$. Using the same arguments as in the proof of the corollary of Proposition 2, (3.10) follows from (A.80).

To prove (3.11), notice that by Definition 2

$$\Pr\left[z_L = 1 \mid z_{L-1}^{L-l}, s_0\right] = E\left[q_L^*(z_{L-1}, s_0)\mid z_{L-1}^{L-l}, s_0\right] \quad \text{(A.81)}$$

and

$$\Pr\left[z_L = 1 \mid z_{L-1}^{L-l}\right] = E\left[q_L(z_{L-1})\mid z_{L-1}^{L-l}\right] \quad \text{(A.82)}$$

and therefore

$$\left|\Pr\left[z_L = 1\mid z_{L-1}^{L-l}, s_0\right] - \Pr\left[z_L = 1\mid z_{L-1}^{L-l}\right]\right| \le \max_{z_{l-1}} \left|q_L^*(z_{L-1}, s_0) - q_L(z_{L-1})\right|. \quad \text{(A.83)}$$

Thus (A.23) implies that

$$\lim_{L\to\infty} \max_{2\le l\le L} \left|\Pr\left[z_L\mid z_{L-1}^{L-l+1}, s_0\right] - \Pr\left[z_L\mid z_{L-1}^{L-l+1}\right]\right| = 0 \quad \text{(A.84)}$$

uniformly over the domain of all variables involved. It is simple to extend this result over $1 \le l \le L$ since for $l = 1$ the term z_{L-1}^{L-l+1} vanishes. Replacing L by J and l by j, shifting the time index by $k = (n-2)J + j$ and rewriting (A.84) in the deinterleaved indexing (3.1) results in

$$\lim_{J\to\infty} \max_{1\le j\le J} \left|\Pr[z(j,n)\mid z(1,n),\cdots,z(j-1,n), s(j,n-1)] - \Pr[z(j,n)\mid z(1,n),\cdots,z(j-1,n)]\right| = 0. \quad \text{(A.85)}$$

Since $q(j,n)$ is a function of $z(1,n),\cdots,z(j-1,n)$, (A.85) implies that

$$\lim_{J\to\infty} \max_{1\le j\le J} \left|\Pr[z(j,n)\mid q(j,n), s(j,n-1)] - \Pr[z(j,n)\mid q(j,n)]\right| = 0 \quad \text{(A.86)}$$

uniformly over the domains of all variables involved. When conditioned on $s(j,n-1)$, $q(j,n)$ is independent of $q(j,n-1),\cdots,q(j,1)$, $z(j,n-1),\cdots,z(j,1)$, and (3.11) follows.

References

[1] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, pp. 1977–1997, Sept. 1963.

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

[3] G. C. Clark and J. B. Cain, *Error Correcting Codes for Digital Communications*. New York: Plenum, 1981.

[4] E. N. Gilbert, "Capacity of burst-noise channels," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–1265, Sept. 1960.

[5] J. Hagernauer, "Zur Kanalkapazität bei Nachrichtenkanälen mit Fading und gebündelten Fehlern," *AEU Arch. Elektron. Uebertragung.*, vol. 34, pp. 229–237, 1980.

[6] ——, "Viterbi decoding of convolutional codes for fading and burst channels," in *Proc. 1980 Zurich Seminar on Digital Communication*, 1980, pp. G2.1–G2.7.

[7] K. S. Leung and L. R. Welch, "Erasure decoding in burst noise channels," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 160–167, Mar. 1981.

[8] M. Loève, *Probability Theory*, vol. I. New York: Springer, 1977.

[9] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*. New York: McGraw-Hill, 1979.