The Compound Channel Capacity of a Class of Finite-State Channels

Amos Lapidoth, *Member, IEEE*, and İ. Emre Telatar, *Member, IEEE*

***Abstract*— A transmitter and receiver need to be designed to guarantee reliable communication on any channel belonging to a given family of finite-state channels defined over common finite input, output, and state alphabets. Both the transmitter and receiver are assumed to be ignorant of the channel over which transmission is carried out and also ignorant of its initial state. For this scenario we derive an expression for the highest achievable rate. As a special case we derive the compound channel capacity of a class of Gilbert–Elliott channels.**

***Index Terms*—Compound channel, error exponents, finite-state channels, Gilbert–Elliott channel, universal decoding.**

I. Introduction

THIS paper deals with a situation where a communication system consisting of a transmitter and a receiver needs to be designed for an unknown channel. The only information available to the designer is the family of channels over which the system will be used, and based on this information the designer must design a transmitter (codebook) and a receiver (decoder) that will guarantee reliable communication over any channel in the family. No feedback mechanism is available to the transmitter, and the codebook must therefore be fixed before transmission begins. At the receiver end, the decoding rule must not depend on the channel over which communication takes place since the receiver too is assumed ignorant of the channel. This situation is commonly referred to as coding for the compound channel [1] or coding for a class of channels [2], [3].

The highest rate at which robust reliable communication is achievable in this setup is called the compound channel capacity of the family, or the capacity of the family. It has been studied in the case where the family of channels consists of memoryless channels in [2] and in the case where the family consists of linear dispersive channels with an unknown distorting filter in [4].

In this paper we study the compound channel capacity for families of finite-state channels. Such channels exhibit memory and are often used to model wireless communications in the presence of fading [5]–[7].

Before turning to a precise statement of the problem and results, we explain some of the difficulties encountered in the compound channel by considering the case where the family of channels consists of memoryless channels only. For this case the compound channel capacity is given by [1]–[3]

$$C = \max_Q \inf_{P \in \mathcal{F}} I(Q; P) \tag{1}$$

where $\mathcal{F}$ is the family of discrete memoryless channels under consideration (defined over common finite input and output alphabets), the maximization is over the set of input distributions, and $I(Q; P)$ is the mutual information between the input and output of the channel P when the input distribution to the channel is Q.

First note that the compound channel capacity is, in general, not equal to the infimum of the capacities of the different channels in the family, as the capacity-achieving input distributions may be different for different channels in the family. Once this is noticed, one soon realizes that the best one can hope for is to achieve (1). Notice, however, that in order to demonstrate that C of (1) is achievable one must demonstrate that one can achieve $I(Q; P)$ with codes such that neither the codebook nor the decoder depend on the channel being used. One cannot employ maximum-likelihood or joint-typicality decoding with respect to the law of the channel in use as this law is unknown. Moreover, the classical random coding argument is based on computing the average probability of error for a random ensemble of codes and then inferring that there exists at least one code that performs as well as this average. The choice of the codebook from the ensemble typically depends on the channel, and one of the difficulties in proving (1) is in showing that there exists a codebook that is simultaneously good for all the channels in the family.

Showing that one cannot guarantee reliable communication at any rate higher than C is usually simpler, but requires work as it does not follow directly from the single-channel converse theorem: C may be smaller than the capacity of every channel in the family. To prove a converse one assumes that a good codebook of rate R is given, and from it one then derives a distribution $Q(x)$ that satisfies

$$I(Q; P) > R - \epsilon, \qquad \forall P \in \mathcal{F} \tag{2}$$

where $\epsilon > 0$ is small and is related to the blocklength and the probability of error attained by the code. The distribution $Q(x)$ is typically related to the empirical distribution of the

Manuscript received January 2, 1997; revised June 17, 1997. The work of A. Lapidoth was supported in part by the NSF Faculty Early Career Development (CAREER) Program. The material in the paper was presented in part at the International Conference on Combinatorics, Information Theory and Statistics, University of Southern Maine, Portland, ME, July 18–20, 1997. This research was conducted in part while A. Lapidoth was visiting Lucent Technologies, Bell Labs.

A. Lapidoth is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

İ. E. Telatar is with Lucent Technologies, Bell Labs, Murray Hill, NJ 0797 USA.

Publisher Item Identifier S 0018-9448(98)02347-5.

codebook, and (2) is usually shown using Fano's inequality and some convexity arguments.

In this paper we shall derive a result analogous to (1) for finite-state channels. The result will be derived by demonstrating that for appropriate rates and maximum-likelihood decoding the ensemble average probability of error for all the channels in the family can be uniformly bounded by an exponentially decreasing function of the blocklength, and by invoking a result from [8] on the existence of a universal decoder for the class of finite-state channels that performs asymptotically as well as a maximum-likelihood decoder tuned to the channel in use.

The rest of the paper is organized as follows. We conclude this introductory section with a more precise statement of the problem and with a presentation of our main result—the capacity of the compound channel consisting of finite-state channels. The coding theorem and the converse theorem needed to prove this result are presented in Section II. In Section III we use this result to derive the capacity of a class of Gilbert–Elliott channels. The paper concludes with a brief discussion in Section IV.

Precise Statement of the Problem and Main Result

For a given set $\mathcal{A}$, let $\mathcal{P}(\mathcal{A})$ denote the set of probability distributions on $\mathcal{A}$. We will only deal with probability distributions on finite sets, and so, we will identify an element Q of $\mathcal{P}(\mathcal{A})$ with a nonnegative function $Q: \mathcal{A} \mapsto \mathbb{R}$ such that

$$\sum_{a\in\mathcal{A}} Q(a) = 1.$$

For a positive integer n, denote by $\mathcal{P}_n(\mathcal{A})$ the set of probability distributions Q on $\mathcal{A}$ with the property that $nQ(a)$ is an integer for all $a \in \mathcal{A}$. We will call such a Q an n-type on $\mathcal{A}$.

A discrete finite-state channel with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and state space $\mathcal{S}$ is characterized by a conditional probability assignment

$$P(y, s'|x, s), \qquad y \in \mathcal{Y},\ x \in \mathcal{X},\ s, s' \in \mathcal{S}.$$

Operationally, if at time $n-1$ the state of the channel is s_{n-1} and the input to the channel at time n is x_n, then the output of the channel y_n at time n and the state s_n of the channel at time n are determined according to the distribution

$$P(y_n, s_n|x_n, s_{n-1}).$$

For such a channel, the probability $P_n(\boldsymbol{y}, s_n|\boldsymbol{x}, s_0)$ that the channel output is $\boldsymbol{y} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$ and the final channel state is s_n conditional on the initial state $s_0 \in \mathcal{S}$ and the channel input $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$ is given by

$$P_n(\boldsymbol{y}, s_n|\boldsymbol{x}, s_0) = \sum_{s_1, \cdots, s_{n-1}} \prod_{i=1}^{n} P(y_i, s_i|x_i, s_{i-1}). \tag{3}$$

We can sum this probability over s_n to obtain the probability that the channel output is $\boldsymbol{y} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$ conditional on the initial state $s_0 \in \mathcal{S}$ and the channel input $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$

$$P_n(\boldsymbol{y}|\boldsymbol{x}, s_0) = \sum_{s_1, \cdots, s_n} \prod_{i=1}^{n} P(y_i, s_i|x_i, s_{i-1}). \tag{4}$$

Given an initial state $s_0 \in \mathcal{S}$ and a distribution Q_n on $\mathcal{X}^n$, the joint distribution of the channel input and output is well-defined, and the mutual information between the input and the output is given by

$$\begin{aligned} I(\boldsymbol{X}; \boldsymbol{Y}|s_0) &= I(Q_n; P_n(\cdot|\cdot, s_0)) \\ &= \sum_{\boldsymbol{x}, \boldsymbol{y}} Q_n(\boldsymbol{x}) P_n(\boldsymbol{y}|\boldsymbol{x}, s_0) \\ &\quad \cdot \ln \frac{P_n(\boldsymbol{y}|\boldsymbol{x}, s_0)}{\sum_{\boldsymbol{x}'} Q_n(\boldsymbol{x}') P_n(\boldsymbol{y}|\boldsymbol{x}', s_0)}. \end{aligned}$$

We are abusing the notation to let I take as arguments both the random variables and the distributions.

Suppose now that we are given a class Θ of discrete finite-state channels with common finite-state space $\mathcal{S}$, and common finite input and finite output alphabets $\mathcal{X}$ and $\mathcal{Y}$. Each channel $\theta \in \Theta$ is characterized by

$$P(y, s'|x, s, \theta), \qquad y \in \mathcal{Y},\ x \in \mathcal{X},\ s, s' \in \mathcal{S} \tag{5}$$

and in analogy to (3) and (4) we will denote by $P_n(\boldsymbol{y}, s_n|\boldsymbol{x}, s_0, \theta)$ the probability that the output of channel θ is $\boldsymbol{y} \in \mathcal{Y}^n$ and the final state is $s_n \in \mathcal{S}$ conditional on the input $\boldsymbol{x} \in \mathcal{X}^n$ and initial state $s_0 \in \mathcal{S}$, and by $P_n(\boldsymbol{y}|\boldsymbol{x}, s_0, \theta)$ the probability that the output of channel θ is $\boldsymbol{y} \in \mathcal{Y}^n$ under the same conditioning. Given a channel $\theta \in \Theta$, an initial state s_0, and a distribution Q_n on $\mathcal{X}^n$, the mutual information between the input and output of the channel θ is given by

$$\begin{aligned} I(\boldsymbol{X}; \boldsymbol{Y}|s_0, \theta) &= I(Q_n; P_n(\cdot|\cdot, s_0, \theta)) \\ &= \sum_{\boldsymbol{x}, \boldsymbol{y}} Q_n(\boldsymbol{x}) P_n(\boldsymbol{y}|\boldsymbol{x}, s_0, \theta) \\ &\quad \cdot \ln \frac{P_n(\boldsymbol{y}|\boldsymbol{x}, s_0, \theta)}{\sum_{\boldsymbol{x}'} Q_n(\boldsymbol{x}') P_n(\boldsymbol{y}|\boldsymbol{x}', s_0, \theta)}. \end{aligned}$$

Definition 1: A rate R is said to be *achievable* for the family of channels Θ, if for any $\epsilon > 0$, there exists an $n(\epsilon)$ such that for all $n > n(\epsilon)$ there exists an encoder

$$f: \{1, \cdots, \lceil e^{nR} \rceil\} \to \mathcal{X}^n$$

and a decoder

$$\phi: \mathcal{Y}^n \to \{1, \cdots, \lceil e^{nR} \rceil\}$$

such that the average probability of error is less than ϵ irrespective of the initial state $s_0 \in \mathcal{S}$ and the channel $\theta \in \Theta$ over which the transmission is carried out. That is,

$$\frac{1}{\lceil e^{nR} \rceil} \sum_{i=1}^{\lceil e^{nR} \rceil} \sum_{\substack{\boldsymbol{y} \in \mathcal{Y}^n: \\ \phi(\boldsymbol{y}) \neq i}} P_n(\boldsymbol{y}|f(i), s_0, \theta) < \epsilon,$$

$$\text{for all } s_0 \in \mathcal{S} \text{ and } \theta \in \Theta.$$

Since the state is not observed by the encoder or decoder, we will assume that the channels in the class have a common state space $\mathcal{S} = \{1, \cdots, |\mathcal{S}|\}$ as long as each channel in the class has the same number of states. When we are presented with a class of finite-state channels with common input and output alphabets where each channel has at most but possibly less than S states, we can equip the class with a state space with S states that is common to all the channels in the family by augmenting the state space of those channels that have less than S states. However, one has to be careful not to artificially introduce bad states in this process. An approach that will avoid this is the following. If a channel θ has $S_\theta < S$ states, pick one of its states, say state S_θ, and define a new channel with S states by

$$\begin{aligned} &P'(y, s|x, s', \theta) \\ &= \begin{cases} P(y, s|x, \min\{s', S_\theta\}, \theta), & 1 \le s < S_\theta \\ \dfrac{1}{S - S_\theta} P(y, S_\theta|x, \min\{s', S_\theta\}, \theta), & s \ge S_\theta. \end{cases} \end{aligned}$$

The new channel P' and the old channel P satisfy

$$P'_n(\boldsymbol{y}|\boldsymbol{x}, s_0, \theta) = P_n(\boldsymbol{y}|\boldsymbol{x}, \min\{s_0, S_\theta\}, \theta)$$

and thus a code that is good for one is good for the other.

Let $C(\Theta)$ denote the compound channel capacity of the class Θ, i.e., the supremum of all achievable rates for the class. We will prove the following theorem.

Theorem 1: The compound channel capacity of the class Θ of finite-state channels defined over common finite input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$ is given by

$$C(\Theta) = \lim_{n\to\infty} \max_{Q_n \in \mathcal{P}(\mathcal{X}^n)} \inf_{s_0 \in \mathcal{S}, \theta\in\Theta} \frac{1}{n} I(Q_n; P_n(\cdot|\cdot, s_0, \theta)). \tag{6}$$

Analytic calculation of this limit is possible only in special cases, e.g., for a class of Gilbert–Elliott channels which will be discussed below, and, in general, the limit cannot even be computed numerically. Nonetheless, in the course of the paper we will establish a sequence of lower bounds monotonically increasing to $C(\Theta)$ (Proposition 1).

A relatively simple finite-state channel that is often used to model communication in the presence of fading is the Gilbert–Elliott channel, which has been studied in [6], [11], and in references therein. In the Gilbert–Elliott channel the channel input and output alphabets are binary $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, and the channel state is also binary, but for convenience we set $\mathcal{S} = \{B, G\}$ corresponding to a "bad" state and a "good" state. In this model the channel output sequence $\boldsymbol{y}$ is related to the channel input sequence $\boldsymbol{x}$ by

$$\boldsymbol{y} = \boldsymbol{x} \oplus \boldsymbol{z}$$

where $\oplus$ denotes binary $(\bmod 2)$ addition, and $\boldsymbol{z}$ is the realization of a binary hidden Markov noise process with internal state set $\mathcal{S}$, see Section III.

The capacity of the Gilbert–Elliott channel is achieved by an independent and identically distributed (i.i.d.) Bernoulli $1/2$ input distribution, irrespective of the channel parameters [6]. Using Theorem 1 we shall show in Section III that under relatively mild conditions outlined in Theorem 2 the compound channel capacity of a class of Gilbert–Elliott channels is equal to the infimum of the capacities of the members of the family. The highest rate at which reliable communication can be guaranteed is thus not reduced due to the ignorance of the transmitter and receiver of the channel in use.

II. Proof of the Converse and a Coding Theorem

Before we can even start to prove the theorem, we need to show that the right-hand side of (6) exists. This is a consequence of the following proposition, which is proved in Appendix I.

Proposition 1: The maximum over $Q_n \in \mathcal{P}(\mathcal{X}^n)$ in defining the sequence

$$C_n(\Theta) = \max_{Q_n \in \mathcal{P}(\mathcal{X}^n)} \inf_{s_0 \in \mathcal{S}, \theta\in\Theta} \frac{1}{n} I(Q_n; P_n(\cdot\,|\,\cdot, s_0, \theta)), \quad n = 1, 2, \cdots \tag{7}$$

is well-defined, and the sequence converges. Moreover,

$$\lim_{n\to\infty} C_n(\Theta) = \sup_n \hat{C}_n(\Theta) \tag{8}$$

where

$$\hat{C}_n(\Theta) = C_n(\Theta) - (\ln|\mathcal{S}|)/n. \tag{9}$$

A. Converse

Given a code $\mathcal{C}$ of blocklength n and rate R, and error probability not exceeding ϵ for any channel $\theta \in \Theta$ and initial state $s_0 \in \mathcal{S}$, define $Q_n \in \mathcal{P}(\mathcal{X}^n)$ as

$$Q_n(\boldsymbol{x}) = \begin{cases} 1/|\mathcal{C}|, & \boldsymbol{x} \in \mathcal{C} \\ 0, & \text{otherwise.} \end{cases}$$

Then by Fano's inequality we get, for all $\theta \in \Theta$ and $s_0 \in \mathcal{S}$

$$I(Q_n; P_n(\cdot|\cdot, s_0, \theta)) \ge n(1-\epsilon)R - \ln 2 \tag{10}$$

and thus by (7)–(9)

$$C(\Theta) \ge \hat{C}_n(\Theta) \ge (1-\epsilon)R - \frac{\ln 2}{n} - \frac{\ln|\mathcal{S}|}{n}$$

where the first inequality follows from Proposition 1, and the second follows from the definition of $\hat{C}_n$ and (10). We then conclude that no rate above $C(\Theta)$ is achievable.

B. Coding Theorem

We will prove the coding theorem in a sequence of steps. The first step is to quote a result from [5] to show that if R is less than $C(\Theta)$ and if we employ maximum-likelihood decoding tuned to the channel $\theta \in \Theta$ over which the transmission is carried out, then we can find an m such that if we view the channel inputs and outputs in blocks of m (that is, to consider an equivalent channel with input alphabet $\mathcal{X}^m$ and output alphabet $\mathcal{Y}^m$), and construct i.i.d. random codes for this channel (that is, codes where each symbol of each codeword is chosen independently according to some distribution), we achieve exponentially decaying probability of error in increasing blocklength. The decay of the probability of error depends, of course, on the channel $\theta \in \Theta$ in use. The second step is to show that the resulting error exponent is bounded from below *uniformly* over the class Θ. The third step is to convert this i.i.d. random coding result to a different kind of random coding, where the codewords are chosen independently but uniformly over some set of input sequences. The last step is to invoke a theorem of [8] on the existence of a universal decoders for the class of finite-state channels to show that this is sufficient to construct good codes and decoders (that do not require knowledge of the channel in use) for the compound channel.

Given $Q_n \in \mathcal{P}(\mathcal{X}^n)$, let $P_e^r(n, R, Q_n; P_n(\cdot\,|\,\cdot, s_0, \theta))$ denote the average (over codebooks and messages) probability of error that is incurred when a blocklength-n rate-R code whose codewords are drawn independently according to the distribution Q_n is used over the channel $P_n(\cdot|\cdot, s_0, \theta)$ and is decoded using a maximum-likelihood decoder tuned to the channel. We know [5, pp. 176–182] that for any $\rho \in [0, 1]$

$$P_e^r(n, R, Q_n; P_n(\cdot\,|\,\cdot, s_0, \theta)) \le |\mathcal{S}| \exp\{-n[F_n(\rho, Q_n; P_n(\cdot\,|\,\cdot, \cdot, \theta)) - \rho R]\}$$

where

$$\begin{aligned} &F_n(\rho, Q_n; P_n(\cdot\,|\,\cdot, \cdot, \theta)) \\ &\quad = -\rho \frac{\ln|\mathcal{S}|}{n} + \min_{s_0 \in \mathcal{S}} \frac{1}{n} E_0(\rho, Q_n; P_n(\cdot\,|\,\cdot, s_0, \theta)) \\ &E_0(\rho, Q_n; P_n(\cdot\,|\,\cdot, s_0, \theta)) \\ &\quad = \ln \sum_{\boldsymbol{y}} \left[\sum_{\boldsymbol{x}} Q_n(\boldsymbol{x}) P_n(\boldsymbol{y}|\boldsymbol{x}, s_0, \theta)^{1/(1+\rho)} \right]^{1+\rho}. \end{aligned}$$

Lemma 1: Given $Q_k \in \mathcal{P}(\mathcal{X}^k)$ and $Q_m \in \mathcal{P}(\mathcal{X}^m)$. Let $n = k + m$ and define $Q_n \in \mathcal{P}(\mathcal{X}^n)$ by

$$Q_n(\boldsymbol{x}) = Q_k(\boldsymbol{x}_1) Q_m(\boldsymbol{x}_2)$$

where

$$\begin{aligned} \boldsymbol{x} &= (x_1, \cdots, x_n) \\ \boldsymbol{x}_1 &= (x_1, \cdots, x_k) \\ \boldsymbol{x}_2 &= (x_{k+1}, \cdots, x_n). \end{aligned}$$

Then

$$F_n(\rho, Q_n; P_n(\cdot|\cdot, \cdot, \theta)) \ge \frac{k}{n} F_k(\rho, Q_k; P_k(\cdot|\cdot, \cdot, \theta)) + \frac{m}{n} F_m(\rho, Q_m; P_m(\cdot|\cdot, \cdot, \theta)).$$

Proof: Identical to [5, pp. 179–180, proof of Lemma 5.9.1]. □

Lemma 2: For any channel P with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, and for any distribution Q on $\mathcal{X}$

$$E_0(\rho, Q; P) \ge \rho I(Q; P) - \tfrac{1}{2}\rho^2 [\ln(eY)]^2$$

where $Y = |\mathcal{Y}|$ is the size of the output alphabet.

Proof: See Appendix II. □

Given $R < C(\Theta)$, set $\epsilon = (C(\Theta) - R)/2$. Choose m such that $\hat{C}_m(\Theta) \ge R + \epsilon$ and let $Q_m \in \mathcal{P}(\mathcal{X}^m)$ be the distribution that achieves the supremum $\hat{C}_m(\Theta)$

$$\inf_{s_0 \in \mathcal{S}, \theta \in \Theta} \frac{1}{m} I(Q_m; P_m(\cdot|\cdot, s_0, \theta)) - \frac{\ln|\mathcal{S}|}{m} \ge R + \epsilon.$$

For $N \ge 1$, let $Q_m^N \in \mathcal{P}(\mathcal{X}^{Nm})$ denote the distribution

$$Q_m^N(\boldsymbol{x}) = \prod_{i=1}^N Q_m(\boldsymbol{x}_i)$$

where $\boldsymbol{x} = (x_1, \cdots, x_{Nm})$ and $\boldsymbol{x}_i = (x_{(i-1)m+1}, \cdots, x_{im})$, $i = 1, \cdots, N$. From Lemma 1 we have

$$\begin{aligned} &F_{Nm}(\rho, Q_m^N; P_{Nm}(\cdot\,|\,\cdot, \cdot, \theta)) - \rho R \\ &\quad \ge F_m(\rho, Q_m; P_m(\cdot\,|\,\cdot, \cdot, \theta)) - \rho R \\ &\quad = \min_{s_0} \frac{1}{m} E_0(\rho, Q_m; P_m(\cdot\,|\,\cdot, s_0, \theta)) - \rho\left(R + \frac{\ln|\mathcal{S}|}{m}\right) \end{aligned}$$

applying Lemma 2 to the channel $P_m(\cdot\,|\,\cdot, s_0, \theta)$

$$\begin{aligned} &\ge \rho\left[\min_{s_0} \frac{1}{m} I(Q_m; P_m(\cdot\,|\,\cdot, s_0, \theta)) - \left(R + \frac{\ln|\mathcal{S}|}{m}\right)\right] - \frac{1}{2m}\rho^2[\ln(eY^m)]^2 \\ &\ge \rho\epsilon - \frac{1}{2m}\rho^2[\ln(eY^m)]^2. \end{aligned}$$

Choosing $\rho = \min\{1, m\epsilon/[\ln(eY^m)]^2\}$ maximizes the right-hand side in the range $0 \le \rho \le 1$ with the value

$$\beta(\epsilon, m, Y) = \begin{cases} m\epsilon^2/(2[\ln(eY^m)]^2), & \epsilon < \dfrac{1}{m}[\ln(eY^m)]^2 \\ \epsilon - \dfrac{1}{2m}[\ln(eY^m)]^2, & \text{else.} \end{cases}$$

Notice that for $\epsilon > 0$, $\beta(\epsilon, m, Y) > 0$. We conclude that for all $N \ge 1$

$$P_e^r(Nm, R, Q_m^N; P_{Nm}(\cdot\,|\,\cdot, s_0, \theta)) \le |\mathcal{S}| \exp\{-Nm\beta(\epsilon, m, Y)\}. \quad (11)$$

The important thing to note is that the right-hand side is independent of both θ and s_0.

We have thus seen that if one constructs a code of blocklength Nm by constructing a random code by choosing codewords independently according to Q_m^N, then the probability of error decays exponentially in N as long as the rate of the code is below $C(\Theta)$. We now show that this implies that random coding by choosing codewords independently and uniformly over a type class has similar performance:

Lemma 3: Given $Q \in \mathcal{P}(\mathcal{X})$, let $Q^n \in \mathcal{P}(\mathcal{X}^n)$ denote the distribution that is the n-fold product of Q, i.e.,

$$Q_n(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i).$$

For a given type $\hat{Q} \in \mathcal{P}_n(\mathcal{X})$, let $\hat{Q}^{(n)} \in \mathcal{P}(\mathcal{X}^n)$ denote the distribution that is uniform over the n length sequences of type $\hat{Q}$.

For every distribution $Q \in \mathcal{P}(\mathcal{X})$ there exists a type $\hat{Q} \in \mathcal{P}_n(\mathcal{X})$ whose choice depends on Q and n but not on P such that

$$P_e^r(n, R - \delta(n), \hat{Q}^{(n)}; P) \le \exp(2n\delta(n)) P_e^r(n, R, Q^n; P)$$

for all P. Here $\delta(n) = |\mathcal{X}| \ln(n+1)/n$ tends to 0 as n tends to infinity.

Proof: Given a code $\mathcal{C}$ of rate R chosen according to the product distribution, consider the following procedure to construct an equitype subcode $\mathcal{C}'$ of rate $R-\delta(n)$: Find the type Q' with the highest occurrence in $\mathcal{C}$ (resolving ties according to some arbitrary but fixed rule). The number of codewords of this type will be lower-bounded by $e^{n(R-\delta(n))}$ since the number of types is upper-bounded by $e^{n\delta(n)}$. Construct the code $\mathcal{C}'$ by picking the first $e^{n(R-\delta(n))}$ codewords in $\mathcal{C}$ of type Q'. Since $\mathcal{C}'$ is a subcode of $\mathcal{C}$, its average probability of error when used over the channel P is upper-bounded by that of $\mathcal{C}$ times $|\mathcal{C}|/|\mathcal{C}'| = e^{n\delta(n)}$.

Since $\mathcal{C}$ is a random code, the type Q' is also random with a distribution that depends on Q, n, and R but not on P. Also, conditional on Q', the codewords in $\mathcal{C}'$ are mutually independent and uniformly distributed over a set of sequences of length n and type Q'. Denoting the distribution of Q' by π, we choose $\hat{Q}$ to satisfy $\pi(\hat{Q}) \geq e^{-n\delta(n)}$. Again, this is possible since the number of types is upper-bounded by $e^{n\delta(n)}$. Then

$$\begin{aligned}\pi(\hat{Q})P_e^r(n, R-\delta(n), \hat{Q}^{(n)}; P) &\leq \sum_{Q'} \pi(Q')P_e^r(n, R-\delta(n), Q'^{(n)}; P)\\ &\leq e^{n\delta(n)}P_e^r(n, R, Q^n; P)\end{aligned}$$

and thus

$$P_e^r(n, R-\delta(n), \hat{Q}^{(n)}; P) \leq \exp(2n\delta(n))P_e^r(n, R, Q^n; P).$$

□

Combining this lemma with (11), we see that for each $N \geq 1$, there is a type $\hat{Q}_m \in \mathcal{P}_N(\mathcal{X}^m)$ such that

$$\begin{aligned}&P_e^r(Nm, R, \hat{Q}_m^{(N)}; P_{Nm}(\cdot\,|\,\cdot, s_0, \theta))\\ &\quad\leq |\mathcal{S}| \exp\left\{-Nm[\beta\left(\epsilon - |\mathcal{X}|^m \frac{\ln(N+1)}{Nm}, m, |\mathcal{Y}|\right)\right.\\ &\qquad\left. -2|\mathcal{X}|^m \frac{\ln(N+1)}{Nm}]\right\}.\end{aligned}$$

Choose N_0 such that for all $N > N_0$

$$|\mathcal{X}|^m \frac{\ln(N+1)}{Nm} < \epsilon/2$$

and

$$2|\mathcal{X}|^m \frac{\ln(N+1)}{Nm} < \frac{1}{2}\beta(\epsilon/2, m, Y).$$

Then, for all $N > N_0$

$$\begin{aligned}&P_e^r(Nm, R, \hat{Q}_m^{(N)}; P_{Nm}(\cdot\,|\,\cdot, s_0, \theta))\\ &\qquad\leq |\mathcal{S}| \exp\{-Nm\tfrac{1}{2}\beta(\epsilon/2, m, Y)\}. \qquad (12)\end{aligned}$$

We will now invoke the following theorem proved in [8], to show that (12) implies the existence of a code and a decoder (neither depending on θ or s_0) that perform well for every θ and s_0:

Theorem [8]: Given an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$, and a finite state space $\mathcal{S}$, let Ψ index the class of all finite-state channels with these input, output, and state alphabets. Consider now a random codebook of rate R and blocklength n whose codewords are drawn independently and uniformly over a set $B_n \subset \mathcal{X}^n$, and let $\overline{P}_{n,\mathrm{ML}}(\mathrm{error}\,|\,\psi, s_0)$ denote the average (over messages and codebooks) probability of error incurred over the channel $\psi \in \Psi$ with initial state s_0 when maximum-likelihood decoding is performed (with the knowledge of ψ and s_0 at the decoder). Similarly, for any decoder ϕ and codebook $\mathcal{C}$, let $P_{n,\phi,\mathcal{C}}(\mathrm{error}|\psi, s_0)$ denote the average (over message) probability of error incurred over the channel ψ with initial state s_0 when the codebook $\mathcal{C}$ and decoder ϕ are used.

There then exists a sequence of rate R codes $\mathcal{C}_n \subset B_n$, and a sequence of decoders ϕ_n (both sequences not depending on the specific channel ψ or initial state s_0) such that

$$\lim_{n\to\infty} \sup_{\psi\in\Psi,\, s_0\in\mathcal{S}} \frac{1}{n} \ln\left(\frac{P_{n,\phi_n,\mathcal{C}_n}(\mathrm{error}\,|\,\psi, s_0)}{\overline{P}_{n,\mathrm{ML}}(\mathrm{error}\,|\,\psi, s_0)}\right) = 0.$$

For a detailed description of the structure of the universal decoder and the family of codes of this theorem the reader is referred to [8]. Loosely speaking, the decoder is constructed by "merging" the maximum-likelihood decoders that correspond to each of a set of finite-state channels. To within a factor which is no bigger than the cardinality of the set, the merged decoder is shown to have a probability of error that is no worse than the probability of error incurred by any of the maximum-likelihood decoders. The set is chosen so as to have a polynomial cardinality (in the blocklength) and to be "dense" in the sense that every finite-state channel is within "distance" $\epsilon_n \to 0$ of some channel in the set, where the notion of distance between laws is made precise in [8].

We apply this theorem with $n = Nm$, B_n the set of sequences of type $\hat{Q}_m$, to conclude that for $R < C(\Theta)$ there exists a sequence of rate R codes $\mathcal{C}_{Nm}$ of blocklength Nm, $N = 1, 2, \cdots$, and a sequence of decoders ϕ_{Nm} such that the probability of error decays exponentially in N for any $\theta \in \Theta$ and $s_0 \in \mathcal{S}$. This approach can be easily extended to treat the case where the blocklength is of the form $mN + \nu$, where $0 \leq \nu < m$, by appending a constant string of length ν to each codeword in the codebook designed for the blocklength mN. One must, of course, guarantee that ν is negligible compared to mN.

III. A Class of Gilbert–Elliott Channels

In this section we study the compound channel capacity of a class of Gilbert–Elliott channels. Such channels have two internal states G, B and binary input and output alphabets, $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. The channel law of a Gilbert–Elliott channel θ is determined by four parameters, $b(\theta)$, $g(\theta)$, $P_G(\theta)$, $P_B(\theta)$ which define the channel law through (3) where $P(y, s|x, s', \theta)$ is defined as follows:

$$P(y, s|x, s', \theta) = q_\theta(y|x, s')r_\theta(s|s')$$

where

$$\begin{aligned}r_\theta(G|B) &= 1 - r_\theta(B|B) = g(\theta)\\ r_\theta(B|G) &= 1 - r_\theta(G|G) = b(\theta)\end{aligned}$$

and

$$\begin{aligned}q_\theta(1|0, B) &= 1 - q_\theta(0|0, B) = P_B(\theta)\\ q_\theta(0|1, B) &= 1 - q_\theta(1|1, B) = P_B(\theta)\end{aligned}$$

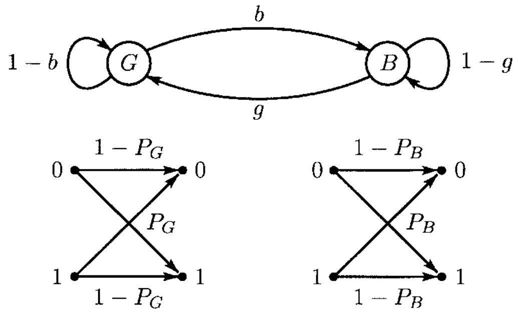

Fig. 1. The Gilbert–Elliott channel model. P_G and P_B are the channel error probabilities in the "good" and "bad" states, and g and b are transition probabilities between states.

$$\begin{aligned} q_\theta(1|0, G) &= 1 - q_\theta(0|0, G) = P_G(\theta) \\ q_\theta(0|1, G) &= 1 - q_\theta(1|1, G) = P_G(\theta) \end{aligned}$$

see Fig. 1. It is straightforward to verify that the noise sequence

$$\boldsymbol{z} = \boldsymbol{x} \oplus \boldsymbol{y}$$

(where $\oplus$ denotes $\bmod 2$ addition) is independent of the input $\boldsymbol{x}$. Thus

$$Y_i = X_i \oplus Z_i$$

where Z_i is independent of the inputs and is Bernoulli P_B distributed if $s_i = B$, and is Bernoulli P_G distributed if $s_i = G$.

The capacity $C(\theta)$ of a single Gilbert–Elliott channel is derived in [6] and is achieved by a memoryless Bernoulli $1/2$ input distribution, irrespective of the channel parameters. It can be expressed in the following form. Let $\{\tilde{S}_k\}$ denote a stationary Markov process taking value in $\{G, B\}$ with the law

$$\begin{aligned} \Pr(\tilde{S}_{k+1} = G|\tilde{S}_k = B) &= 1 - \Pr(\tilde{S}_{k+1} = B|\tilde{S}_k = B) \\ &= g(\theta) \\ \Pr(\tilde{S}_{k+1} = B|\tilde{S}_k = G) &= 1 - \Pr(\tilde{S}_{k+1} = G|\tilde{S}_k = G) \\ &= b(\theta). \end{aligned}$$

Let $\{\tilde{Z}_k\}$ be a $\{0, 1\}$-valued random process, where conditional on the process $\tilde{S}$, $\{\tilde{Z}_k\}$ forms an independent sequence of random variables with

$$\Pr(\tilde{Z}_k = 1|\tilde{S}_{-\infty}^{+\infty}) = \begin{cases} P_B, & \text{if } \tilde{S}_k = B \\ P_G, & \text{if } \tilde{S}_k = G \end{cases}$$

where if $\{X_k\}$ is a sequence of random variables then X_l^m denotes $X_l, \cdots, X_m$. We now have that the capacity $C(\theta)$ of the Gilbert–Elliott channel of parameters $(b(\theta), g(\theta), P_B(\theta), P_G(\theta))$ is given by

$$C(\theta) = \ln 2 - h_\theta(\tilde{Z})$$

provided that

$$0 < b(\theta), g(\theta) < 1$$

where $h_\theta(\tilde{Z})$ is the entropy rate of the process $\{\tilde{Z}_k\}$.

Using Theorem 1 we shall now establish the compound channel capacity of a class of Gilbert–Elliott channels.

Theorem 2: Consider a family of Gilbert–Elliott channels, and let

$$\delta = \inf_{\theta \in \Theta} \min\{g(\theta), b(\theta), 1 - g(\theta), 1 - b(\theta)\}. \tag{13}$$

If $\delta > 0$, then the compound channel capacity $C(\Theta)$ of the family is given by

$$C(\Theta) = \inf_{\theta \in \Theta} C(\theta)$$

where $C(\theta)$ is the capacity of the Gilbert–Elliott channel with parameters $(b(\theta), g(\theta), P_G(\theta), P_B(\theta))$.

Proof: First note that the converse part of this theorem is trivial since no rate is achievable for a family of channels if it is unachievable for some member of the family. We thus focus on the direct part which we prove using Theorem 1 by choosing Q_n to be the i.i.d. Bernoulli$(1/2)$ input distribution. Noting that for this input distribution the channel output is also i.i.d. Bernoulli$(1/2)$ we deduce that

$$\begin{aligned} \frac{1}{n} I(X_1^n; Y_1^n|s_0, \theta) &= \ln 2 - \frac{1}{n} H(Y_1^n|X_1^n, s_0, \theta) \\ &= \ln 2 - \frac{1}{n} H(Z_1^n|s_0, \theta). \end{aligned}$$

The theorem will thus be established once we show that

$$\lim_{n\to\infty} \max_{s_0} \sup_{\theta \in \Theta} \left| \frac{1}{n} H(Z_1^n|s_0, \theta) - h_\theta(\tilde{Z}) \right| = 0. \tag{14}$$

Given a channel $\theta \in \Theta$, an initial state $s_0 \in \mathcal{S}$, and the channel input $\boldsymbol{x} \in \mathcal{X}^n$, we have by (3) that the probability $P_n(s_n|\boldsymbol{x}, s_0, \theta)$ of the channel state being s_n at time n is given by

$$P_n(s_n|\boldsymbol{x}, s_0, \theta) = \sum_{\boldsymbol{y} \in \mathcal{Y}^n} \sum_{s_1, \cdots, s_{n-1}} \prod_{i=1}^n P(y_i, s_i|x_i, s_{i-1}, \theta).$$

We can now extend the notion of indecomposability [5, p. 105] to families of channels as follows.

Definition 2: A family of channels defined over common finite input, output, and state alphabets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{S}$ is uniformly indecomposable if for any $\epsilon > 0$ there exists an N_0 such that for $n > N_0$

$$\sup_{\theta \in \Theta} |P_n(s_n|\boldsymbol{x}, s_0', \theta) - P_n(s_n|\boldsymbol{x}, s_0'', \theta)| < \epsilon$$

for all $s_n \in \mathcal{S}$, all $\boldsymbol{x} \in \mathcal{X}^n$, and all initial states $s_0', s_0'' \in \mathcal{S}$.

Extending [5, Theorem 4.6.4] to uniformly indecomposable families of channels we obtain that for uniformly indecomposable families of channels and any input distribution Q_n

$$\lim_{n\to\infty} \sup_{\theta \in \Theta} \left| \frac{1}{n} I(\boldsymbol{X}; \boldsymbol{Y}|s_0', \theta) - \frac{1}{n} I(\boldsymbol{X}; \boldsymbol{Y}|s_0'', \theta) \right| = 0. \tag{15}$$

Under the assumptions of the theorem the family of Gilbert–Elliott channels is uniformly indecomposable, and it thus follows that

$$\lim_{n\to\infty} \sup_{\theta \in \Theta} \left| \frac{1}{n} I(X_1^n; Y_1^n|s_0, \theta) - \left(\ln 2 - \frac{1}{n} H(\tilde{Z}_1^n|\tilde{S}_0, \theta) \right) \right| = 0$$

and (14) (and hence the theorem) will be established if we show

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}H(\tilde{Z}_1^n|\tilde{S}_0,\theta)-h_\theta(\tilde{Z})\right|=0. \tag{16}$$

We now show that to establish (16) it suffices to show that

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}|H(\tilde{Z}_n|\tilde{Z}_1^{n-1},\tilde{S}_0,\theta)-h_\theta(\tilde{Z})|=0. \tag{17}$$

This reduction follows by noting that $H(\tilde{Z}_n|\tilde{Z}_1^{n-1},\tilde{S}_0,\theta)$ is monotonically nonincreasing, and by noting that by the chain rule

$$H(\tilde{Z}_1^n|\tilde{S}_0,\theta)=\sum_{k=1}^{n}H(\tilde{Z}_k|\tilde{Z}_1^{k-1},\tilde{S}_0,\theta).$$

Next, writing

$$h_\theta(\tilde{Z})=H(\tilde{Z}_n|\tilde{Z}_{-\infty}^{n-1},\theta) \tag{18}$$

and recalling that given $\tilde{S}_0$ the future of $\tilde{Z}$ is independent of its past we conclude that

$$H(\tilde{Z}_n|\tilde{Z}_1^{n-1},\tilde{S}_0,\theta)\le h_\theta(\tilde{Z}). \tag{19}$$

On the other hand, by (18)

$$H(\tilde{Z}_n|\tilde{Z}_1^{n-1},\theta)\ge h_\theta(\tilde{Z}) \tag{20}$$

and thus, by (19) and (20), it follows that to prove (17) (and hence the theorem) it suffices to show

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}(H(\tilde{Z}_n|\tilde{Z}_1^{n-1},\theta)-H(\tilde{Z}_n|\tilde{Z}_1^{n-1},\tilde{S}_0,\theta))=0. \tag{21}$$

In words, we need to prove that given $\tilde{Z}_1^{n-1}$, knowing the initial state $\tilde{S}_0$ becomes immaterial to estimating $\tilde{Z}_n$ for sufficiently large n, with the convergence being uniform over the family.

Given an initial distribution $\pi_0=(\pi_0(G),\pi_0(B))$ on the channel state, let

$$\begin{aligned}&\pi_n(z_1,\cdots,z_{n-1})\\&\qquad=(\pi_n(G|z_1,\cdots,z_{n-1}),\pi_n(B|z_1,\cdots,z_{n-1}))\end{aligned}$$

be the distribution of the channel state at time n, after observing the noise process.

Let π'_n denote π_n when π_0 is chosen as the stationary distribution of $\tilde{S}_0$, let π_n^G denote π_n when $\pi_0=(1,0)$, and let π_n^B denote π_n when $\pi_0=(0,1)$.

To prove (21) write

$$\begin{aligned}H(\tilde{Z}_n|\tilde{Z}_1^{n-1},\theta)=&E_{\tilde{Z}_1^{n-1}}h_b(P_G\pi'_n(G|\tilde{Z}_1^{n-1})\\&+P_B\pi'_n(B|\tilde{Z}_1^{n-1}))\end{aligned}$$

where $h_b(x)$ is the binary entropy function, i.e.,

$$h_b(x)=x\ln\frac{1}{x}+(1-x)\ln\frac{1}{1-x}$$

and the expectation is taken over $\tilde{Z}_1^{n-1}$ when $\tilde{S}_0$ has its stationary distribution. Similarly, write

$$\begin{aligned}&H(\tilde{Z}_n|\tilde{Z}_1^{n-1},\tilde{S}_0,\theta)\\&\quad=E_{\tilde{S}_0,\tilde{Z}_1^{n-1}}h_b(P_G\pi_n^{\tilde{S}_0}(G|\tilde{Z}_1^{n-1})+P_B\pi_n^{\tilde{S}_0}(B|\tilde{Z}_1^{n-1}))\end{aligned}$$

where the expectation is taken over $\tilde{S}_0$ and $\tilde{Z}_1^{n-1}$ when $\tilde{S}_0$ has its stationary distribution. It follows that to prove (21) it suffices to prove that

$$\lim_{n\to\infty}\sup_{\theta\in\Theta}\max_{z_1,\cdots,z_{n-1}}\sum_{s\in\mathcal{S}}|\pi_n^G(s|\tilde{Z}_1^{n-1})-\pi_n^B(s|\tilde{Z}_1^{n-1})|=0 \tag{22}$$

which is proved in Appendix III.

A different proof of (22), which is valid with the additional condition

$$\inf_{\theta\in\Theta}\min\{P_G(\theta),P_B(\theta),1-P_G(\theta),1-P_B(\theta)\}>0$$

is outlined below. It has the advantage of being generalizable to finite-state channels with more states. The proof is based on the recursion equations for computing π_{n+1} from π_n, see [7, eq. (11)], [9, Lemma 2.1], and references therein.

$$\pi_{n+1}(z_1,\cdots,z_n)=\frac{\pi_n(z_1,\cdots,z_{n-1})A(z_n)}{\pi_n(z_1,\cdots,z_{n-1})A(z_n)\mathbf{1}} \tag{23}$$

where

$$A(0)=\begin{pmatrix}1-P_G & 0\\ 0 & 1-P_B\end{pmatrix}\begin{pmatrix}1-b & b\\ g & 1-g\end{pmatrix}$$
$$A(1)=\begin{pmatrix}P_G & 0\\ 0 & P_B\end{pmatrix}\begin{pmatrix}1-b & b\\ g & 1-g\end{pmatrix}$$

and $\mathbf{1}$ is a column all-one vector. The convergence in (22) now follows from [15, Proposition A.4] on normalized products of matrices that are each of the form of a diagonal matrix multiplied by a stochastic matrix. This proposition improves on the estimates of [9, Lemma 6.1] on normalized products of matrices by exploiting the structure of the matrices. □

It should be noted that irrespective of the parameters $b(\theta)$ and $g(\theta)$, and irrespective of the initial state s_0 we have that if $0\le P_G\le P_B\le 1/2$ and the input distribution is i.i.d. Bernoulli$(1/2)$ then

$$\frac{1}{n}I(X_1^n;Y_1^n|s_0,\theta)\ge\ln 2-h_b(P_B).$$

This observation could be used to slightly strengthen Theorem 2. In particular, if we wish to demonstrate that a rate R is achievable for the family Θ then in computing δ in (13) we may exclude those channels in Θ for which $\ln 2-h_b(P_B)\ge R$.

To see that some condition on the transition probabilities of the state chain is necessary, consider a class of Gilbert–Elliott channels indexed by the positive integers. Specifically, let $P_G(k)=0$, $P_B(k)=1/2$, $b(k)=g(k)=2^{-k}$ for $k\ge 1$. For any given k, one can achieve rates exceeding $\ln 2-h_b(1/4)$ over the channel k by using a deep enough interleaver to make the channel look like a memoryless BSC with crossover probability $1/4$. Thus

$$\inf_{\theta\in\Theta}C(\theta)\ge\ln 2-h_b(1/4).$$

However, for any given blocklength n, the channel that corresponds to $\theta=n$ when started in the bad state will stay in the bad state for the duration of the transmission with probability exceeding $1 - n2^{-n} \geq \frac{1}{2}$. Since in the bad state the channel output is independent from the input, we conclude that reliable communication is not possible at any rate and

$$0 = C(\Theta) < \ln 2 - h_b(1/4) \leq \inf_{\theta \in \Theta} C(\theta).$$

IV. Discussion and Conclusions

In this paper we have derived the capacity of a class of finite-state channels. Comparing (6) with (1) we note that the expression for the capacity is very similar to the one that holds for a class of memoryless channels, as both take the form of a $\max - \inf$ of mutual informations, where the maximum is over all input distributions, and the infimum is over the channels in the family. The main difference is that in the expression for the capacity of a class of finite state channels there is an additional limit that allows for the dimension of the input distribution to go to infinity. This is not surprising, as such a limit is even needed to express the capacity of a single finite-state channel. This additional limit is significant, because if the convergence to the mutual information rate is not uniform over the family, unexpected compound channel capacities may result.

Theorem 1 should not lead one to conjecture that the compound channel capacity of any family is of a $\max - \inf$ form, as the following counterexample demonstrates [12], [13]. Let $\Theta = (0, 1)$ and let $\theta^{(1)}, \theta^{(2)}, \cdots$ be the digits in the binary expansion of $\theta \in \Theta$, i.e.,

$$\theta = \sum_{i>0} \theta^{(i)}/2^i, \qquad \theta^{(i)} \in \{0, 1\}.$$

(For definiteness, for a rational θ of the form $k/2^n$, take the terminating expansion among the two possible binary expansions.) Consider now the set of channels defined over common binary input and output alphabets where over the channel θ the conditional probability of the output sequence $\boldsymbol{y} = (y_1, \cdots, y_n)$ given the input sequence $\boldsymbol{x} = (x_1, \cdots, x_n)$ is given by

$$P_n(\boldsymbol{y}|\boldsymbol{x}, \theta) = \begin{cases} 1, & \text{if } y_i = x_i \oplus \theta^{(i)}, \quad \forall i \\ 0, & \text{otherwise} \end{cases}$$

where $\oplus$ denotes $\bmod 2$ addition. For this family of channels we have that the $\max - \inf$ expression yields a value of 1 irrespective of the blocklength, and yet the compound channel capacity of this family is 0, as can be easily verified using standard techniques from the theory of arbitrarily varying channels, see [13], and [14, Appendix]. This example can be explained by noting that for this family of channels there does not exist a universal decoder [8].

Appendix I

We prove Proposition 1 using the following sequence of lemmas.

Lemma 4: For each n, there exists $Q_n^* \in \mathcal{P}(\mathcal{X}^n)$ such that

$$\inf_{s_0 \in \mathcal{S}, \theta \in \Theta} \frac{1}{n} I(Q_n^*; P_n(\cdot|\cdot, s_0, \theta)) = \sup_{Q_n \in \mathcal{P}(\mathcal{X}^n)} \inf_{s_0 \in \mathcal{S}, \theta \in \Theta} \frac{1}{n} I(Q_n; P_n(\cdot, \cdot|s_0, \theta)).$$

Proof: For each n, $\theta \in \Theta$, and $s_0 \in \mathcal{S}$, the function

$$Q_n \in \mathcal{P}(\mathcal{X}^n) \mapsto I(Q_n; P_n(\cdot|\cdot, s_0, \theta)) \in \mathbb{R}$$

is continuous. Furthermore, for a fixed n, this class of functions is uniformly continuous in θ and s_0. (The modulus of continuity of $I(\cdot; P)$ can be bounded in terms of the size of the input and output alphabets alone, see, e.g., [2, Lemma 2].) Thus

$$Q_n \in \mathcal{P}(\mathcal{X}^n) \mapsto \inf_{s_0 \in \mathcal{S}, \theta \in \Theta} I(Q_n; P_n(\cdot|\cdot, s_0, \theta)) \in \mathbb{R}$$

is continuous. Since $\mathcal{P}(\mathcal{X}^n)$ is compact, the supremum over Q_n of this function is attained. □

Lemma 5: Given k, m, n, Q_k, and Q_m, Q_n as in Lemma 1. Then

$$\begin{aligned} &\inf_{s_0, \theta} I(Q_n; P_n(\cdot|\cdot, s_0, \theta)) \\ &\quad \geq \inf_{s_0, \theta} I(Q_k; P_k(\cdot|\cdot, s_0, \theta)) \\ &\qquad + \inf_{s_0, \theta} I(Q_m; P_m(\cdot|\cdot, s_0, \theta)) - \ln |\mathcal{S}|. \end{aligned}$$

Proof: For a given $\theta \in \Theta$ and $s_0 \in \mathcal{S}$, let $(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y}_1, \boldsymbol{Y}_2, S_k)$ be the distributed according to

$$Q_k(\boldsymbol{x}_1) Q_m(\boldsymbol{x}_2) P_k(\boldsymbol{y}_1, s_k|\boldsymbol{x}_1, s_0, \theta) P_m(\boldsymbol{y}_2|\boldsymbol{x}_2, s_k, \theta).$$

We want to show that

$$\begin{aligned} \inf_{s_0, \theta} I(\boldsymbol{X}_1\boldsymbol{X}_2; \boldsymbol{Y}_1\boldsymbol{Y}_2|s_0, \theta) \geq& \inf_{s_0, \theta} I(\boldsymbol{X}_1; \boldsymbol{Y}_1|s_0, \theta) \\ &+ \inf_{s_k, \theta} I(\boldsymbol{X}_2; \boldsymbol{Y}_2|s_k, \theta) - \ln |\mathcal{S}|. \end{aligned}$$

To that end, write

$$\begin{aligned} \inf_{s_0, \theta} I(\boldsymbol{X}_1\boldsymbol{X}_2; \boldsymbol{Y}_1\boldsymbol{Y}_2|s_0, \theta) =& \inf_{s_0, \theta} [I(\boldsymbol{X}_1; \boldsymbol{Y}_1\boldsymbol{Y}_2|s_0, \theta) \\ &+ I(\boldsymbol{X}_2; \boldsymbol{Y}_1\boldsymbol{Y}_2|\boldsymbol{X}_1, s_0, \theta)] \\ \geq& \inf_{s_0, \theta} I(\boldsymbol{X}_1; \boldsymbol{Y}_1\boldsymbol{Y}_2|s_0, \theta) \\ &+ \inf_{s_0, \theta} I(\boldsymbol{X}_2; \boldsymbol{Y}_1\boldsymbol{Y}_2|\boldsymbol{X}_1, s_0, \theta). \end{aligned}$$

The first term is lower-bounded by $\inf_{s_0, \theta} I(\boldsymbol{X}_1; \boldsymbol{Y}_1|s_0, \theta)$. For the second term, note that

$$\begin{aligned} I(\boldsymbol{X}_2; \boldsymbol{Y}_1\boldsymbol{Y}_2|\boldsymbol{X}_1, s_0, \theta) =& I(\boldsymbol{X}_2; \boldsymbol{Y}_1\boldsymbol{Y}_2\boldsymbol{X}_1|s_0, \theta) \\ &- I(\boldsymbol{X}_1; \boldsymbol{X}_2|s_0, \theta) \\ =& I(\boldsymbol{X}_2; \boldsymbol{Y}_1\boldsymbol{Y}_2\boldsymbol{X}_1|s_0, \theta) \\ \geq& I(\boldsymbol{X}_2; \boldsymbol{Y}_2|s_0, \theta) \end{aligned}$$

where we have used the independence of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. We now note that

$$I(\boldsymbol{X}_2; \boldsymbol{Y}_2|s_0, \theta) \geq I(\boldsymbol{X}_2; \boldsymbol{Y}_2|S_k, s_0, \theta) - \ln |\mathcal{S}|$$

(this follows from the inequality

$$|I(A; B|CD) - I(A; B|C)| \leq H(D)$$

see [5, p. 112]) and

$$\begin{aligned} I(\boldsymbol{X}_2; \boldsymbol{Y}_2|S_k, s_0, \theta) =& \sum_{s_k} q(s_k|s_0, \theta) I(\boldsymbol{X}_2; \boldsymbol{Y}_2|s_k, \theta) \\ \geq& \inf_{s_k} I(\boldsymbol{X}_2; \boldsymbol{Y}_2|s_k, \theta) \end{aligned}$$

where

$$q(s_k|s_0,\,\theta)=\sum_{\boldsymbol{x}_1,\,\boldsymbol{y}_1} P_n(\boldsymbol{y}_1,\,s_k|\boldsymbol{x}_1,\,s_0,\,\theta)Q_k(\boldsymbol{x}_1).$$

Combining the above

$$\inf_{s_0,\,\theta} I(\boldsymbol{X}_1\boldsymbol{X}_2;\boldsymbol{Y}_1\boldsymbol{Y}_2|s_0,\theta)\ge \inf_{s_0,\theta} I(\boldsymbol{X}_1;\,\boldsymbol{Y}_1|s_0,\,\theta) + \inf_{s_k,\,\theta} I(\boldsymbol{X}_2;\,\boldsymbol{Y}_2|s_k,\,\theta)-\ln|\mathcal{S}|.$$

□

Corollary 1: The sequence $C_n(\Theta)$ converges. Furthermore,

$$\lim_{n\to\infty} C_n(\Theta)=\sup_n \hat{C}_n(\Theta) \tag{24}$$

where

$$\hat{C}_n(\Theta)=C_n(\Theta)-(\ln|\mathcal{S}|)/n. \tag{25}$$

Proof: Given integers $k,\,m\ge 1$, let Q_k and Q_m achieve the maximizations in $C_k(\Theta)$ and $C_m(\Theta)$, respectively. Let $n=k+m$ and define Q_n as in the proof of Lemma 5. Then

$$\begin{aligned} nC_n(\Theta) &\ge \inf_{s_0,\theta} I(Q_n;\,P_n(\cdot|\cdot,\,s_0,\,\theta))\\ &\ge \inf_{s_0,\,\theta} I(Q_k;\,P_k(\cdot|\cdot,\,s_0,\,\theta))\\ &\quad + \inf_{s_0,\,\theta} I(Q_m;\,P_m(\cdot|\cdot,\,s_0,\,\theta))-\ln|\mathcal{S}|\\ &= kC_k(\Theta)+mC_m(\Theta)-\ln|\mathcal{S}|, \end{aligned}$$

or

$$(k+m)\hat{C}_{k+m}(\Theta)\ge k\hat{C}_k(\Theta)+m\hat{C}_m(\Theta).$$

Since $\hat{C}_n(\Theta)$ is also bounded (say by $\ln|\mathcal{X}|+\ln|\mathcal{S}|$), we conclude that (see [5, p. 112])

$$\lim_{n\to\infty} C_n(\Theta)=\lim_{n\to\infty}\hat{C}_n(\Theta)=\sup_n \hat{C}_n(\Theta).$$

□

Lemma 4 and the above corollary establish Proposition 1.

Appendix II

Proof of Lemma 2: This lemma is a slight modification of [5, Problem 5.23]. The reason we chose to include a proof is twofold: first, the solutions to the problems in [5] are not widely available, and second, the official solutions for this problem contain a small error.

By expanding $E_0(\rho,\,Q;\,P)$ in a power series around $\rho=0$ to second order, we obtain

$$\begin{aligned} E_0(\rho,\,Q;\,P) &=\rho E_0'(0,\,Q;\,P)+\tfrac{1}{2}\,\rho^2E_0''(\xi,\,Q;\,P),\\ &\qquad\qquad \text{for some } \xi\in[0,\,\rho]\\ &=\rho I(Q;\,P)+\tfrac{1}{2}\,\rho^2E_0''(\xi,\,Q;\,P) \end{aligned}$$

and so it suffices to show that

$$-E_0''(\rho,\,Q;\,P)\le[\ln(eY)]^2$$

for all $\rho\in[0,\,1]$. Differentiating

$$-E_0(\rho,\,Q;\,P)=\ln\sum_y \alpha(y)^{1+\rho}$$

$$\alpha(y)=\sum_x Q(x)P(y|x)^{1/(1+\rho)}$$

we obtain

$$\begin{aligned} -&E_0'(\rho,\,Q;\,P)\\ &=\frac{\sum_y \alpha(y)^{1+\rho}(\ln\alpha(y)+(1+\rho)\alpha'(y)/\alpha(y))}{\sum_y \alpha(y)^{1+\rho}}\\ &=\sum_y w(y)\left(\ln\alpha(y)-\sum_x q(x|y)\ln P(y|x)^{1/(1+\rho)}\right)\\ &=\sum_{x,\,y} w(y)q(x|y)\ln\frac{Q(x)}{q(x|y)} \end{aligned} \tag{26}$$

where

$$w(y)=\frac{\alpha(y)^{1+\rho}}{\sum_{y'}\alpha(y')^{1+\rho}}$$

and

$$q(x|y)=\frac{Q(x)P(y|x)^{1/(1+\rho)}}{\alpha(y)}.$$

We now differentiate $-E_0'$ to find $-E_0''$

$$\begin{aligned} -E_0''(\rho,\,Q;\,P) &=\sum_{x,\,y}\frac{\partial}{\partial\rho}[w(y)q(x|y)]\ln\frac{Q(x)}{q(x|y)}\\ &\quad -\sum_{x,\,y} w(y)\frac{\partial}{\partial\rho}q(x|y). \end{aligned} \tag{27}$$

Since $\sum_x q(x|y)=1$ for all y and ρ, the second term above is zero. We now compute

$$\begin{aligned} &\frac{\partial}{\partial\rho}[w(y)q(x|y)]\\ &\quad=\frac{\partial}{\partial\rho}\left[\frac{Q(x)P(y|x)^{1/(1+\rho)}\alpha(y)^{\rho}}{\sum_{y'}\alpha(y')^{1+\rho}}\right]\\ &\quad=w(y)q(x|y)\Bigg[-(1+\rho)^{-1}\ln P(y|x)^{1/(1+\rho)}+\ln\alpha(y)\\ &\qquad\qquad +\frac{\rho}{\alpha(y)}\frac{\partial}{\partial\rho}\alpha(y)+E_0'(\rho,\,Q;\,P)\Bigg]\\ &\quad=w(y)q(x|y)\Bigg[-(1+\rho)^{-1}\ln P(y|x)^{1/(1+\rho)}+\ln\alpha(y)\\ &\qquad\qquad -\frac{\rho}{1+\rho}\sum_{x'}q(x'|y)\ln P(y|x')^{1/(1+\rho)}\\ &\qquad\qquad +E_0'(\rho,\,Q;\,P)\Bigg]\\ &\quad=w(y)q(x|y)\Bigg[\frac{1}{1+\rho}\ln\frac{Q(x)}{q(x|y)}\\ &\qquad\qquad +\frac{\rho}{1+\rho}\sum_{x'}q(x'|y)\ln\frac{Q(x')}{q(x'|y)}\\ &\qquad\qquad +E_0'(\rho,\,Q;\,P)\Bigg]. \end{aligned}$$

Substituting this in (27) we obtain

$$\begin{aligned}-E_0''(\rho, Q; P) &= \frac{1}{1+\rho}\sum_{x,y} w(y)q(x|y)\left[\ln\frac{Q(x)}{q(x|y)}\right]^2\\ &\quad+\frac{\rho}{1+\rho}\sum_{y} w(y)\left[\sum_{x} q(x|y)\ln\frac{Q(x)}{q(x|y)}\right]^2\\ &\quad-[E_0'(\rho, Q; P)]^2. \qquad (28)\end{aligned}$$

Using the inequality $E[A]^2 \le E[A^2]$ on the second term in brackets and combining the first two terms

$$\begin{aligned}&-E_0''(\rho, Q; P)\\ &\qquad\le \sum_{x,y} w(y)q(x|y)\left[\ln\frac{Q(x)}{q(x|y)}\right]^2 - [E_0'(\rho, Q; P)]^2.\end{aligned}$$

Since the last term is nonpositive we can drop it to leave

$$-E_0''(\rho, Q; P) \le \sum_{x,y} w(y)q(x|y)\left[\ln\frac{Q(x)}{q(x|y)}\right]^2.$$

Observe now that for $z \ge 1$, $\ln(z)^2 \le (4/e^2)z \le z$. Thus

$$\begin{aligned}-E_0''(\rho, Q; P) &\le \sum_{\substack{x,y:\\ Q(x)\ge q(x|y)}} w(y)Q(x)\\ &\quad+\sum_{\substack{x,y:\\ Q(x)<q(x|y)}} w(y)q(x|y)\left[\ln\frac{Q(x)}{q(x|y)}\right]^2\\ &\le 1+\sum_{\substack{x,y:\\ Q(x)<q(x|y)}} w(y)q(x|y)\left[\ln\frac{Q(x)}{q(x|y)}\right]^2.\end{aligned}$$

Note that for $Q(x)/q(x|y) \le 1$

$$\left[\ln\frac{Q(x)}{q(x|y)}\right]^2 \le \left[\ln\frac{Q(x)P(y|x)^{1/(1+\rho)}}{q(x|y)}\right]^2 = (\ln\alpha(y))^2.$$

Furthermore,

$$\ln\alpha(y) = \frac{1}{1+\rho}(\ln w(y) - E_0(\rho, Q; P))$$

and

$$E_0(\rho, Q; P) \le \rho I(Q; P) \le \rho\ln Y$$

thus

$$(\ln\alpha(y))^2 \le \left(\frac{\ln w(y) - \rho\ln Y}{1+\rho}\right)^2$$

and

$$\begin{aligned}&-E_0''(\rho, Q; P)\\ &\le 1+\sum_{\substack{x,y:\\ Q(x)<q(x|y)}} w(y)q(x|y)\left(\frac{\ln w(y)-\rho\ln Y}{1+\rho}\right)^2\\ &\le 1+\sum_{y} w(y)\left(\frac{\ln w(y)-\rho\ln Y}{1+\rho}\right)^2.\end{aligned}$$

We will upper-bound the right-hand side further by maximizing over $w(y)$ subject to $w(y) \ge 0$, $\sum_y w(y) = 1$. To that end, expand the square and maximize each term

$$\begin{aligned}&-E_0''(\rho, Q; P)\\ &\quad\le 1+\frac{1}{(1+\rho)^2}\Bigg[\max_{w}\sum_{y} w(y)(\ln w(y))^2\\ &\qquad+2\rho\ln Y\max_{w}\sum_{y} w(y)\ln\frac{1}{w(y)}+(\rho\ln Y)^2\Bigg].\end{aligned}$$

The second maximization gives $\ln Y$. For the first, write

$$\begin{aligned}&\max_{w}\sum_{y} w(y)(\ln w(y))^2\\ &\quad=\max_{w}\sum_{y} w(y)\left(\ln\frac{w(y)}{e}+1\right)^2\\ &\quad=\max_{w}\sum_{y} w(y)\left[\left(\ln\frac{w(y)}{e}\right)^2+2\ln w(y)-1\right]\\ &\quad\le\max_{w}\sum_{y} w(y)\left(\ln\frac{w(y)}{e}\right)^2-1\\ &\quad=[\ln(eY)]^2-1\\ &\quad=[\ln Y]^2+2\ln Y\end{aligned}$$

where the next to last equality follows from the concavity of $z(\ln(z/e))^2$ in the range $0 \le z \le 1$, and thus the maximization taking place when $w(y) = 1/Y$. Thus

$$\begin{aligned}-E_0''(\rho, Q; P) &\le 1+[\ln Y]^2+\frac{2\ln Y}{(1+\rho)^2}\\ &\le 1+[\ln Y]^2+2\ln Y\\ &=[\ln(eY)]^2\end{aligned}$$

as was to be shown. □

Appendix III

In this Appendix we prove (22).

For a given Gilbert–Elliott channel θ define

$$U(s, s', s'', z) = V(s, s', z) - W(s, s'', z)$$

where

$$V(s, s', z) = \Pr(s_l = s | s_{l-1} = s, s_{s+1} = s', z_l = z)$$

and

$$W(s, s'', z) = \Pr(s_l = s | s_{l-1} \ne s, s_{s+1} = s'', z_l = z).$$

Mushkin and Bar-David [6, Appendix] show that the quantity in (22)

$$\max_{z_1,\cdots,z_{n-1}}\sum_{s\in\mathcal{S}}|\pi_n^G(s|\tilde{Z}_1^{n-1})-\pi_n^B(s|\tilde{Z}_1^{n-1})|$$

is upper-bounded by $(\max_{s,s',s'',z}|U(s, s', s'', z)|)^n$. We can thus prove (22) as long as for some $\epsilon > 0$

$$\max_{s,s',s'',z}|U(s, s', s'', z)| \le 1-\epsilon$$

for all channels θ in our class Θ. We will show that under the assumptions of Theorem 2, namely,

$$\delta = \inf_{\theta \in \Theta} \min\{b(\theta), g(\theta), 1 - b(\theta), 1 - g(\theta)\} > 0$$

(29) is satisfied and ϵ can be taken to be $\delta^2/2$.

To that end, suppose $|U| \geq 1 - \epsilon$ for some $\epsilon < 1$. This implies that either $V - W \geq 1 - \epsilon$, in which case $V \geq 1 - \epsilon$ and $W \leq \epsilon$, or else, $W - V \geq 1 - \epsilon$, in which case $W \geq 1 - \epsilon$ and $V \leq \epsilon$. Take the first alternative. $V \geq 1 - \epsilon$ implies $\Pr(z_l = z | s_l = s) > 0$, $\Pr(s_{l-1} = s, s_l = s, s_{l+1} = s') > 0$, and

$$\begin{aligned} &\Pr(z_l = z | s_l \neq s) \Pr(s_{l-1} = s, s_l \neq s, s_{l+1} = s') \\ &\quad \leq \frac{\epsilon}{1 - \epsilon} \Pr(z_l = z | s_l = s) \\ &\qquad \cdot \Pr(s_{l-1} = s, s_l = s, s_{l+1} = s'). \end{aligned} \tag{30}$$

$W \leq \epsilon$ implies $\Pr(z_l = z | s_l \neq s) > 0$, $\Pr(s_{l-1} \neq s, s_l \neq s, s_{l+1} = s'') > 0$ and

$$\begin{aligned} &\Pr(z_l = z | s_l = s) \Pr(s_{l-1} \neq s, s_l = s, s_{l+1} = s'') \\ &\quad \leq \frac{\epsilon}{1 - \epsilon} \Pr(z_l = z | s_l \neq s) \\ &\qquad \cdot \Pr(s_{l-1} \neq s, s_l \neq s, s_{l+1} = s''). \end{aligned} \tag{31}$$

From (30) and (31), we conclude that

$$\begin{aligned} &\Pr(s_{l-1} = s, s_l \neq s, s_{l+1} = s') \\ &\quad \cdot \Pr(s_{l-1} \neq s, s_l = s, s_{l+1} = s'') \\ &\leq \left(\frac{\epsilon}{1 - \epsilon}\right)^2 \Pr(s_{l-1} = s, s_l = s, s_{l+1} = s') \\ &\quad \cdot \Pr(s_{l-1} \neq s, s_l \neq s, s_{l+1} = s''). \end{aligned} \tag{32}$$

We can rewrite (32) as

$$\begin{aligned} \left(\frac{\epsilon}{1 - \epsilon}\right)^2 &\geq \frac{\Pr(s_{l-1} = s, s_l \neq s, s_{l+1} = s')}{\Pr(s_{l-1} = s, s_l = s, s_{l+1} = s')} \\ &\quad \cdot \frac{\Pr(s_{l-1} \neq s, s_l = s, s_{l+1} = s'')}{\Pr(s_{l-1} \neq s, s_l \neq s, s_{l+1} = s'')} \\ &\geq \delta^4 \end{aligned}$$

and thus $\epsilon/(1 - \epsilon) \geq \delta^2$, implying $\epsilon \geq \delta^2/2$.

Taking the second possibility yields the same lower bound. We thus conclude that $|U(s, s', s'', z)| \leq 1 - \delta^2/2$, which proves (22). □

Acknowledgment

Stimulating discussions with R. Gallager, S. Mitter, R. Urbanke, and J. Ziv are gratefully acknowledged. The authors wish to thank O. Zeitouni for bringing [15] to their attention, and the anonymous referees for their helpful comments.

References

[1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[2] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacity of a class of channels," *Ann. Math. Stat.*, vol. 30, pp. 1229–1241, Dec. 1959.

[3] J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. Berlin/Heidelberg: Springer-Verlag, 1978.

[4] W. L. Root and P. P. Varaiya, "Capacity of classes of Gaussian channels," *SIAM J. Appl. Math.*, vol. 16, pp. 1350–1393, Nov. 1968.

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

[6] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert–Elliott channel," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1277–1290, Nov. 1989.

[7] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inform. Theory*, vol. 42, pp. 868–886, May 1996.

[8] M. Feder and A. Lapidoth, "Universal decoding for noisy channels," Mass. Inst. Technol., Lab. Inform. Decision Syst., Tech. Rep. LIDS-P-2377, Dec. 1996; see also M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inform. Theory*, to be published.

[9] T. Kaijser, "A limit theorem for partially observed Markov chains," *Ann. Prob.*, vol. 3, no. 4, pp. 677–696, 1975.

[10] H. Furstenberg and H. Kesten, "Products of random matrices," *Ann. Math. Statist.*, vol. 31, pp. 457–469, 1960.

[11] G. Bratt, "Sequential decoding for the Gilbert–Elliott channel—Strategy and analysis," Ph.D. dissertation, Lund University, Lund, Sweden, June 1994.

[12] R. L. Dobrushin, "Optimum information transmission through a channel with unknown parameters," *Radiotekh. Elektron.*, vol. 4, no. 12, pp. 1951–1956, 1959.

[13] D. Blackwell, L. Breiman, and A. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Statist.*, vol. 31, pp. 558–567, 1960.

[14] I. Csiszár and P. Narayan, "Capacity of the Gaussian arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 37, pp. 18–26, Jan. 1991.

[15] F. LeGland and L. Mevel, "Geometric ergodicity in hidden Markov models," Institute de Recherche en Informatique et Systémes Aléatoires, IRISA pub. no. 1028, July 1996.