Reliable Communication Under Channel Uncertainty

Amos Lapidoth, *Member, IEEE*, and Prakash Narayan, *Senior Member, IEEE*

(Invited Paper)

***Abstract*— In many communication situations, the transmitter and the receiver must be designed without a complete knowledge of the probability law governing the channel over which transmission takes place. Various models for such channels and their corresponding capacities are surveyed. Special emphasis is placed on the encoders and decoders which enable reliable communication over these channels.**

***Index Terms*— Arbitrarily varying channel, compound channel, deterministic code, finite-state channel, Gaussian arbitrarily varying channel, jamming, MMI decoder, multiple-access channel, randomized code, robustness, typicality decoder, universal decoder, wireless.**

I. Introduction

SHANNON'S classic paper [111] treats the problem of communicating reliably over a channel when both the transmitter and the receiver are assumed to have full knowledge of the channel law so that selection of the codebook and the decoder structure can be optimized accordingly. We shall often refer to such channels, in loose terms, as known channels. However, there are a variety of situations in which either the codebook or the decoder must be selected without a complete knowledge of the law governing the channel over which transmission occurs. In subsequent work, Shannon and others have proposed several different channel models for such situations (e.g., the compound channel, the arbitrarily varying channel, etc.). Such channels will hereafter be referred to broadly as unknown channels.

Ultimate limits of communication over these channels in terms of capacities, reliability functions, and error exponents, as also the means of attaining them, have been extensively studied over the past 50 years. In this paper, we shall review some of these results, including recent unpublished work, in a unified framework, and also present directions for future research. Our emphasis is primarily on single-user channels. The important class of multiple-access channels is not treated in detail; instead, we provide a brief survey with pointers for further study.

There are, of course, a variety of situations, dual in nature to those examined in this paper, in which an information source must be compressed—losslessly or with some acceptable distortion—without a complete knowledge of the characteristics of the source. The body of literature on this subject is vast, and we refer the reader to [23], [25], [61], [71], and [128] in this issue.

In selecting a model for a communication situation, several factors must be considered. These include the physical and statistical nature of the channel disturbances, the information available to the transmitter, the information available to the receiver, the presence of any feedback link from the receiver to the transmitter, and the availability at the transmitter and receiver of a shared source of randomness (independent of the channel disturbances). The resulting capacity, reliability function, and error exponent will also rely crucially on the performance criteria adopted (e.g., average or worst case measures).

Consider, for example, a situation controlled by an adversarial jammer. Based on the physics of the channel, the received signal can often be modeled as the sum of the transmitted signal, ambient or receiver noise, and the jammer's signal. The transmitter and jammer are typically constrained in their average or peak power. The jammer's strategy can be described in terms of the probability law governing its signal. If the jammer's strategy is known to the system designer, then the resulting channel falls in the category studied by Shannon [111] and its extensions to channels with memory. The problem becomes more realistic if the jammer can select from a family of strategies, and the selected strategy, and hence the channel law, is not fully known to the system designer. Different statistical assumptions on the family of allowable jammer strategies will result in different channel models and, hence, in different capacities. Clearly, it is easier to guarantee reliable communication when the jammer's signal is independent and identically distributed (i.i.d.), albeit with unknown law, than when it is independently distributed but with arbitrarily varying and unknown distributions. The former situation leads to a "compound channel" model, and the latter to an "arbitrarily varying channel" model.

Next, various degrees of information about the jammer's strategy may be available to the transmitter or receiver, leading to yet more variations of such models. For example, if the jammer employs an i.i.d. strategy, the receiver may learn it from the signal received when the transmitter is silent, and yet be unable to convey its inference to the transmitter if the channel is one-way. The availability of a feedback link, on the other hand, may allow for suitable adaptation of the codebook, leading to an enhanced capacity value. Of course, in the extreme situation where the receiver has access to the pathwise realization of the jammer's signal and can

Manuscript received December 10, 1997; revised May 4, 1998.

A. Lapidoth is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4307 USA.

P. Narayan is with the Electrical Engineering Department and the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA.

Publisher Item Identifier S 0018-9448(98)05288-2.

subtract it from the received signal, the transmitter can ignore the jammer's presence. Another modeling issue concerns the availability of a source of common randomness which enables coordinated randomization at the encoder and decoder. For instance, such a resource allows the use of spread-spectrum techniques in combating jammer interference [117]. In fact, access to such a source of common randomness can sometimes enable reliable communication at rates that are strictly larger than those achievable without it [6], [48].

The capacity, reliability function, and error exponent for a given model will also depend on the precise notion of reliable communication adopted by the system designer with regard to the decoding error probability. For a given system the error probability will, in general, depend on the transmitted message and the jammer's strategy. The system designer may require that the error probability be small for all jammer strategies and for all messages; a less stringent requirement is that the error probability be small only as an (arithmetic) average over the message set. While these two different performance criteria yield the same capacity for a known channel, in the presence of a jammer the capacities may be different [20]. Rather than requiring the error probability to be small for every jammer strategy, we may average it over the set of all strategies with respect to a given prior. This Bayesian approach gives another notion of reliable communication, with yet another definition of capacity.

The notions of reliable communication mentioned above do not preclude the possibility that the system performance be governed by the worst (or average) jamming strategy even when a more benign strategy is employed. In some situations, such as when the jamming strategies are i.i.d., it is possible to design a decoder with error probability decaying asymptotically at a rate no worse than if the jammer strategy were known in advance. The performance of this "universal" decoder is thus governed not by the worst strategy but by the strategy that the jammer chooses to use.

Situations involving channel uncertainty are by no means limited to military applications, and arise naturally in several commercial applications as well. In mobile wireless communications, the varying locations of the mobile transmitter and receiver with respect to scatterers leads to an uncertainty in channel law. This application is discussed in the concluding section. Other situations arise in underwater acoustics, computer memories with defects, etc.

The remainder of the paper is organized as follows. Focusing on unknown channels with finite input and output alphabets, models for such channels without and with memory, as well as different performance criteria, are described in Section II. Key results on channel capacity for these models and performance criteria are presented in Section III. In Section IV, we survey some of the encoders and decoders which have been proposed for achieving reliable communication over such channels. While our primary focus is on channels with finite input and output alphabets, we shall consider in Section V the class of unknown channels whose output equals the sum of the transmitted signal, an unknown interference .and white Gaussian noise. Section VI consists of a brief review of unknown multiple-access channels. In the concluding Section VII, we examine the potential role in mobile wireless communications of the work surveyed in this paper.

II. Channel Models and Preliminaries

We now present a variety of mathematical models for communication under channel uncertainty. We shall assume throughout a discrete-time framework. For waveform channels with uncertainty, care must be exercised in formulating a suitable discrete-time model as it can sometimes lead to conservative designs. Throughout this paper, all logarithms and exponentiations are with respect to the base 2.

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets denoting the channel input and output alphabets, respectively. The probability law of a (known) channel is specified by a sequence of conditional probability mass functions (pmf's)

$$\{W_n(\boldsymbol{y}|\boldsymbol{x})\colon\ \boldsymbol{x}\in\mathcal{X}^n, \boldsymbol{y}\in\mathcal{Y}^n\}_{n=1}^{\infty} \tag{1}$$

where $W_n(\cdot|\cdot)$ denotes the conditional pmf governing channel use through n units of time, i.e., "n uses of the channel." If the known channel is a discrete memoryless channel (DMC), then its law is characterized in terms of a stochastic matrix $W\colon\ \mathcal{X}\mapsto\mathcal{Y}$ according to

$$W_n(\boldsymbol{y}|\boldsymbol{x})=\prod_{t=1}^{n}W(y_t|x_t) \tag{2}$$

where $\boldsymbol{x}=(x_1,\cdots,x_n)\in\mathcal{X}^n$ and $\boldsymbol{y}=(y_1,\cdots,y_n)\in\mathcal{Y}^n$. For notational convenience, we shall hereafter suppress the subscript n and use $W(\boldsymbol{y}|\boldsymbol{x})$ instead of $W_n(\boldsymbol{y}|\boldsymbol{x})$.

Example 1: The binary-symmetric channel (BSC) is a DMC with $\mathcal{X}=\mathcal{Y}=\{0,1\}$, and a stochastic matrix

$$W(y|x)=\begin{cases}p, & \text{if } y\neq x\\ 1-p, & \text{if } y=x\end{cases}$$

for a "crossover probability" $p\in[0,1]$. The BSC can also be described by writing

$$Y_t=x_t+Z_t$$

where $\{Z_t\}_{t=1}^{\infty}$ is a Bernoulli(p) process, and addition is $\bmod 2$.

A *family* of channels indexed by $\theta\in\Theta$ can be denoted by

$$\{W(\boldsymbol{y}|\boldsymbol{x};\theta),\boldsymbol{x}\in\mathcal{X}^n,\boldsymbol{y}\in\mathcal{Y}^n,\theta\in\Theta\}_{n=1}^{\infty} \tag{3}$$

for some parameter space Θ. For example, this family would correspond to a family of DMC's if

$$W(\boldsymbol{y}|\boldsymbol{x};\theta)=\prod_{t=1}^{n}W(y_t|x_t;\theta) \tag{4}$$

where $\{W(y|x;\theta),x\in\mathcal{X},y\in\mathcal{Y},\theta\in\Theta\}$ is a suitable subset of the set of all stochastic matrices $\mathcal{X}\mapsto\mathcal{Y}$. Such a family of channels, referred to as a compound DMC, is often used to model communication over a DMC whose law belongs to the family and remains unchanged during the course of a transmission, but is otherwise unknown.

Example 2: Consider a compound BSC with $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $\Theta \subset [0, 1]$ with

$$W(y|x; \theta) = \begin{cases} \theta, & \text{if } y \neq x \\ 1-\theta, & \text{if } y = x. \end{cases}$$

The case $\Theta = \{0.1, 0.9\}$, for instance, represents a compound BSC of unknown polarity.

A more severe situation arises when the channel parameters vary arbitrarily from symbol to symbol during the course of a transmission. This situation can sometimes be modeled by choosing $\Theta = \mathcal{S}^\infty$ where $\mathcal{S}$ is a finite set, often referred to as the state space, and by setting

$$W(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{s}) = \prod_{t=1}^{n} W(y_t|x_t; s_t) \tag{5}$$

where $\boldsymbol{s} = (s_1, \cdots, s_n)$, and $W\colon\ \mathcal{X} \times \mathcal{S} \mapsto \mathcal{Y}$ is a given stochastic matrix. This model is called a discrete memoryless arbitrarily varying channel and will hereafter be referred to simply as an AVC.

Example 3: Consider an AVC (5) with $\mathcal{X} = \mathcal{S} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, 2\}$, and

$$W(y|x; s) = \begin{cases} 1, & \text{if } y = x + s \\ 0, & \text{otherwise.} \end{cases}$$

This AVC can also be described by writing

$$y_t = x_t + s_t.$$

All additions above are arithmetic. Since the stochastic matrix $W\colon\ \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ has entries which are all $\{0, 1\}$-valued, such an AVC is sometimes called a deterministic AVC. This example is due to Blackwell *et al*. [31].

In some hybrid situations, certain channel parameters may be unknown but fixed during the course of a transmission, while other parameters may vary arbitrarily from symbol to symbol. Such a situation can often be modeled by setting $\Theta = \mathcal{S}^\infty \times \Xi$, where $\mathcal{S}$ is as above, Ξ connotes a subset of the stochastic matrices $\mathcal{X} \times \mathcal{S} \mapsto \mathcal{Y}$, and for $\xi \in \Xi$

$$W(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{s}, \xi) = \prod_{t=1}^{n} W(y_t|x_t; s_t, \xi). \tag{6}$$

We shall refer to this model as a hybrid DMC.

In some situations in which the channel law is *fully known,* memoryless channel models are inadequate and more elaborate models are needed. In wireless applications, a finite-state channel (FSC) model [64], [123] is often used. The memory in the transmission channel is captured by the introduction of a set of states Σ, and the probability law of the channel is given by

$$W(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\sigma_0 \in \Sigma} \pi(\sigma_0) \sum_{\sigma_1, \cdots, \sigma_n \in \Sigma^n} \prod_{t=1}^{n} W(y_t, \sigma_t|x_t, \sigma_{t-1}) \tag{7}$$

where π is a pmf on Σ, and $W\colon\ \mathcal{X} \times \Sigma \mapsto \mathcal{Y} \times \Sigma$ is a stochastic matrix. Operationally, if at time $t-1$ the state of the channel is σ_{t-1} and the input to the channel at time t is x_t, then the output of the channel y_t at time t and the state σ_t of the channel at time t are determined according to the conditional probability $W(y_t, \sigma_t|x_t, \sigma_{t-1})$.

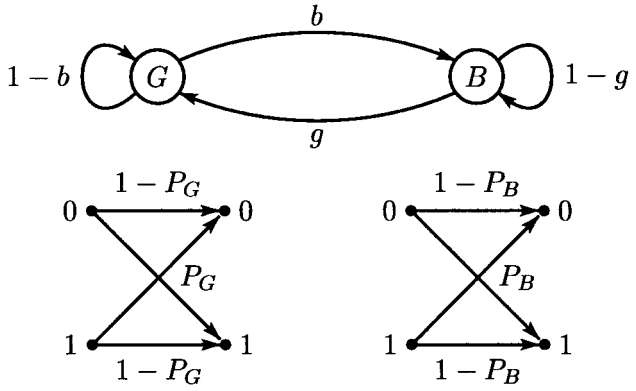

Fig. 1. Gilbert–Elliott channel model. P_G and P_B are the channel crossover probabilities in the "good" and "bad" states, and g and b are transition probabilities between states.

In wireless applications, the states often correspond to different fading levels which the channel may experience (cf. Section VII). It should be noted that the model (7) corresponds to a *known* channel, and the set of states Σ should not be confused with the state space $\mathcal{S}$ introduced in (5) in the definition of an AVC.

Example 4: The Gilbert–Elliott channel [57], [68], [69], [101] is a finite-state channel with two states $\Sigma = \{G, B\}$, the state G corresponding to the "good" state and state B corresponding to the "bad" state (see Fig. 1). The channel has input and output alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, and law

$$W(y, \sigma|x, \sigma') = q(y|x, \sigma')r(\sigma|\sigma')$$

where

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

and

$$\begin{aligned} q(1|0, B) &= 1 - q(0|0, B) = P_B \\ q(0|1, B) &= 1 - q(1|1, B) = P_B \\ q(1|0, G) &= 1 - q(0|0, G) = P_G \\ q(0|1, G) &= 1 - q(1|1, G) = P_G \end{aligned}$$

and where π is often taken as the stationary pmf of the state process, i.e.,

$$\pi(B) = 1 - \pi(G) = \frac{b}{b+g}.$$

The channel can also be described as

$$Y_t = x_t + Z_t$$

where addition is $\bmod 2$, and where $\{Z_t\}$ is a stationary binary hidden Markov process with two internal states.

We can, of course, consider a situation which involves an unknown channel with memory. If the matrix $W\colon \mathcal{X} \times \Sigma \mapsto \mathcal{Y} \times \Sigma$ is unknown but remains fixed during a transmission, the

channel can be modeled as a compound FSC [91] by setting Θ to be a set of pairs (π, W) of pmf's of the initial state and stochastic matrices $\mathcal{X} \times \Sigma \mapsto \mathcal{Y} \times \Sigma$ with

$$W(\boldsymbol{y}|\boldsymbol{x}; \theta) = \sum_{\sigma_0 \in \Sigma} \pi(\sigma_0; \theta) \sum_{\sigma_1, \cdots, \sigma_n \in \Sigma^n} \prod_{t=1}^{n} W(y_t, \sigma_t | x_t, \sigma_{t-1}; \theta) \quad (8)$$

where, with an abuse of notation, $(\pi(\cdot; \theta), W(\cdot, \cdot|\cdot, \cdot; \theta))$ denotes a generic element θ of Θ.

Example 5: A compound Gilbert–Elliott channel [91] is a family of Gilbert–Elliott channels indexed by some set Θ where each channel in the family has a different set of parameters $b(\theta)$, $g(\theta)$, $P_B(\theta)$, $P_G(\theta)$.

More severe yet is a situation where the channel parameters may vary in an arbitrary manner from symbol to symbol during a transmission. This situation can be modeled in terms of an arbitrarily varying FSC, which is described by introducing a state space $\mathcal{S}$ as above, setting $\Theta = \Gamma \times \mathcal{S}^\infty$ where Γ is a set of pmfs on Σ, and letting

$$W(\boldsymbol{y}|\boldsymbol{x}; \gamma, \boldsymbol{s}) = \sum_{\sigma_0 \in \Sigma} \gamma(\sigma_0) \sum_{\sigma_1, \cdots, \sigma_n \in \Sigma^n} \prod_{t=1}^{n} W(y_t, \sigma_t | x_t, \sigma_{t-1}; s_t) \quad (9)$$

where $\gamma \in \Gamma$, and

$$\{W(y, \sigma|x, \sigma'; s), x \in \mathcal{X}, y \in \mathcal{Y}, \sigma \in \Sigma, \sigma' \in \Sigma, s \in \mathcal{S}\}$$

is a family of stochastic matrices $\mathcal{X} \times \Sigma \times \mathcal{S} \mapsto \mathcal{Y} \times \Sigma$. To our knowledge, this channel model has not appeared heretofore in the literature, and is a subject of current investigation by the authors of the present paper.

The models described above for communication under channel uncertainty do not form an exhaustive list. They do, however, constitute a rich and varied class of channel descriptions.

We next provide precise descriptions of an encoder (transmitter) and a decoder (receiver). Let the set of messages be $\mathcal{M} = \{1, \cdots, M\}$. A length-$n$ block code is a pair of mappings (f, ϕ), where

$$f: \mathcal{M} \to \mathcal{X}^n \quad (10)$$

is the encoder, and

$$\phi: \mathcal{Y}^n \to \mathcal{M} \cup \{0\} \quad (11)$$

is the decoder. The rate of such a code is

$$\frac{1}{n} \log M. \quad (12)$$

Note that the encoder, as defined by (10), produces an output which is solely a function of the message. If the encoder is provided additional side information, this definition must be modified accordingly. A similar statement of appropriate nature applies to the decoder as well. Also, while 0 is allowed as a decoder output for the sake of convenience, it will signify a decoding failure and will always be taken to constitute an error.

The probability of error for the message $m \in \mathcal{M}$, when the code (f, ϕ) is used on a channel $\theta \in \Theta$ is given by

$$e(m, f, \phi, \theta) = \sum_{\boldsymbol{y}: \ \phi(\boldsymbol{y}) \neq m} W(\boldsymbol{y}|f(m); \theta). \quad (13)$$

The corresponding maximum probability of error is

$$e_{\max}(f, \phi, \theta) = \max_{m \in \mathcal{M}} e(m, f, \phi, \theta) \quad (14)$$

and the average probability of error is

$$\overline{e}(f, \phi, \theta) = \frac{1}{M} \sum_{m \in \mathcal{M}} e(m, f, \phi, \theta). \quad (15)$$

Obviously, the maximum probability of error will lead to a more stringent performance criterion than the average probability of error. In the case of known channels, both criteria result in the same capacity values. For certain unknown channels, however, these two criteria can yield different capacity results, as will be seen below [20].

For certain unknown channels, an improvement in performance can be obtained by using a *randomized code*. A randomized code constitutes a communication technique, the implementation of which requires the availability of a common source of randomness at the transmitter and receiver; the encoder and decoder outputs can now additionally depend on the outcome of a random experiment. Thus the set of allowed encoding–decoding strategies is enriched by permitting recourse to *mixed strategies*, in the parlance of game theory. The definition of a code in (10) and (11) must be suitably modified, and the potential enhancement in performance (e.g., in terms of the maximum or average probability of error in (14) and (15)) is assessed as an average with respect to the common source of randomness.

The notion of a randomized code should not be confused with the standard method of proof of coding theorems based on a *random-coding argument*. Whereas a randomized code constitutes a communication technique, a random-coding argument is a proof technique which is often used to establish the existence of a (single) deterministic code as in (10) and (11) which yields good performance on a known channel, without actually constructing the code. This is done by introducing a pmf on an ensemble of codes, computing the corresponding average performance over such an ensemble, and then invoking the argument to show that if this average performance is good, then there must exist at least one code in the ensemble with good performance. The random-coding argument is sometimes tricky to invoke when proving achievability results for families of channels. If for each channel in the family the average performance over the ensemble of codes is good, the argument cannot be used to guarantee the existence of a single code which is *simultaneously* good for all the channels in the family; for each channel, there may be a different code with performance as good as the ensemble average.

Precisely, a randomized code (F, Φ) is a random variable (rv) with values in the family of all length-n block codes (f, ϕ) defined by (10) and (11) with the same message set

$\mathcal{M} = \{1, \cdots, M\}$. While the pmf of the rv (F, Φ) may depend on a knowledge of the family of channels indexed by $\theta \in \Theta$, it is not allowed to depend on the actual value of $\theta \in \Theta$ governing a particular transmission or on the transmitted message $m \in \mathcal{M}$.

The maximum and average probabilities of error will be denoted, with an abuse of notation, by $e_{\max}(F, \Phi, \theta)$ and $\overline{e}(F, \Phi, \theta)$, respectively. These error probabilities are defined in a manner analogous to that of a deterministic code in (14) and (15), replacing $e(m, f, \phi, \theta)$ with $e(m, F, \Phi, \theta)$ given by

$$e(m, F, \Phi, \theta) = \mathbb{E}\left[\sum_{\boldsymbol{y}:\, \Phi(\boldsymbol{y}) \neq m} W(\boldsymbol{y}|F(m); \theta)\right] \tag{16}$$

where $\mathbb{E}$ denotes expectation with respect to the pmf of the rv (F, Φ). When randomized codes are allowed, the maximum and average error probability criteria lead to the same capacity value for any channel (known or unknown). This is easily seen since given a randomized code, a random permutation of the message set can be used to obtain a new randomized code of the same rate, whose maximum error probability equals the average error probability of the former (cf. e.g., [44, p. 223, Problem 5]).

While a randomized code is preferable for certain unknown channels owing to its ability to outperform deterministic codes by yielding larger capacity values, it may not be always possible to provide both the transmitter and the receiver with the needed access to a common source of randomness. In such situations, we can consider using a code in which the encoder alone can observe the outcome of a random experiment, whereas the decoder is deterministic. Such a code, referred to as a code with stochastic encoder, is defined as a pair (F, ϕ) where the encoder can be interpreted as a stochastic matrix F: $\mathcal{M} \to \mathcal{X}^n$, and the (deterministic) decoder is given by (11). In proving the achievability parts of coding theorems, the codewords $\{F(m), m \in \mathcal{M}\}$ are usually chosen independently, which completes the probabilistic description of the code (F, ϕ). The various error probabilities for such a code are defined in a manner analogous to that in (13)–(15). In comparison with deterministic codes, a code with stochastic encoder clearly cannot lead to larger capacity values for known channels (since even randomized codes cannot do so). However, for certain unknown channels, while deterministic codes may lead to a smaller capacity value for the maximum probability of error criterion than for the average probability of error criterion, codes with stochastic encoders may afford an improvement by yielding identical capacity values under both criteria.

Hereafter, a deterministic code will be termed as such in those sections in which the AVC is treated; elsewhere, it will be referred to simply as a code. On the other hand, a code with stochastic encoder or a randomized code will be explicitly termed as such.

We now define the notion of the capacity of an unknown channel which, as the foregoing discussion might suggest, is more elaborate than the capacity of a known channel. For $0 < \epsilon < 1$, a number $R \geq 0$ is an ϵ-achievable rate on (an unknown) channel for maximum (resp., average) probability of error, if for every $\delta > 0$ and every n sufficiently large, there exists a length-n block code (f, ϕ) with rate

$$\frac{1}{n} \log M > R - \delta \tag{17}$$

and maximum (resp., average) probability of error satisfying

$$\sup_{\theta \in \Theta} e_{\max}(f, \phi, \theta) \leq \epsilon \tag{18}$$

$$(\text{resp., } \sup_{\theta \in \Theta} \overline{e}(f, \phi, \theta) \leq \epsilon). \tag{19}$$

A number $R \geq 0$ is an achievable rate for the maximum (resp., average) probability of error if it is ϵ-achievable for every $0 < \epsilon < 1$.

The ϵ-capacity of a channel for maximum (resp., average) probability of error is the largest ϵ-achievable rate as given by (17) and (18) (resp., (19)). It will be denoted C_ϵ^m (resp., C_ϵ^a) for those channels for which the two error probability criteria lead, in general, to different values of ϵ-capacity, in which cases, of course, $C_\epsilon^m < C_\epsilon^a$; otherwise, it will be denoted simply by C_ϵ.

The capacity of a channel for maximum or average probability of error is the largest achievable rate for that error criterion. It will be denoted by C^m or C^a for those channels for which the two error probability criteria lead, in general, to different capacity values, when, obviously, $C^m < C^a$; else, it will be denoted by C. Observe that the capacities C^m and C^a can be equivalently defined as the infima of the corresponding ϵ-capacities for $\epsilon > 0$, i.e.,

$$C^m = \lim_{\epsilon \to 0} C_\epsilon^m \quad \text{and} \quad C^a = \lim_{\epsilon \to 0} C_\epsilon^a.$$

Remark: If an ϵ-capacity of a channel (C_ϵ^m or C_ϵ^a) does not depend on ϵ, $0 < \epsilon < 1$, its value is called a strong capacity; such a result is often referred to as a strong converse. See [122] for conditions under which a strong converse holds for known channels.

When codes with stochastic encoders are allowed, analogous notions of ϵ-capacity (C_ϵ^m or C_ϵ^a) and capacity (C^m or C^a) of a channel are defined by modifying the previous definitions of these terms in an obvious way. In particular, the probabilities of error are understood in terms of expectations with respect to the probability law of the stochastic encoder. For randomized codes, too, analogous notions of ϵ-capacity and capacity are defined; note, however, that in this case the maximum and average probabilities of error will lead to the same results, as observed earlier.

While the fundamental notion of channel capacity provides the system designer with an indication of the ultimate coding rates at which reliable communication can be achieved over a channel, it is additionally very useful to assess coding performance in terms of the reductions attained in the various error probabilities by increasing the complexity and delay of a code as measured by its blocklength. This is done by determining the exponents with which the error probabilities can be made to vanish by increasing the blocklength n of the code, leading to the notions of reliability function, randomized code reliability function, and random-coding error exponent of a channel. Our survey does not address these important notions

for which we direct the reader to [43], [44], [46], [64], [65], [95], [115], [116], and references therein.

In the situations considered above, quite often the selection of codes is restricted in that the transmitted codewords must satisfy appropriate input constraints. Let g be a nonnegative-valued function on $\mathcal{X}$, and let

$$g(\boldsymbol{x}) = \frac{1}{n}\sum_{t=1}^{n} g(x_t), \qquad \boldsymbol{x} \in \mathcal{X}^n \tag{20}$$

where, for convenience, we assume that $\min_{x\in\mathcal{X}} g(x) = 0$. Given $\Gamma \geq 0$, a length-n block code (f, ϕ) given by (10) and (11), is said to satisfy input constraint Γ if the codewords $\{f(m), m \in \mathcal{M}\}$ satisfy

$$g(f(m)) \leq \Gamma, \qquad m \in \mathcal{M}. \tag{21}$$

Similarly, a randomized code (F, Φ) or a code with stochastic encoder (F, ϕ) satisfies input constraint Γ if

$$g(F(m)) \leq \Gamma \text{ almost surely (a.s.)}, \qquad m \in \mathcal{M}. \tag{22}$$

Of course, if $\Gamma \geq \max_{x\in\mathcal{X}} g(x)$, then the input constraint is inoperative.

Restrictions are often imposed also on the variations in the unknown channel parameters during the course of a transmission. For instance, in the AVC model (5), constraints can be imposed on the sequence of channel states $\boldsymbol{s} \in \mathcal{S}^n$ as follows. Let l be a nonnegative-valued function on $\mathcal{S}$, and let

$$l(\boldsymbol{s}) = \frac{1}{n}\sum_{t=1}^{n} l(s_t), \qquad \boldsymbol{s} \in \mathcal{S}^n \tag{23}$$

where we assume that $\min_{s\in\mathcal{S}} l(s) = 0$. Given $\Lambda \geq 0$, we shall say that $\boldsymbol{s} \in \mathcal{S}^n$ satisfies state constraint Λ if

$$l(\boldsymbol{s}) \leq \Lambda. \tag{24}$$

If $\Lambda \geq \max_{s\in\mathcal{S}} l(s)$, the state constraint is rendered inoperative.

If coding performance is to be assessed under input constraint Γ, then only such codes will be allowed as satisfy (21) or (22), as applicable. A similar consideration holds if the unknown channel parameters are subject to constraints. For instance, for the AVC model of (5) under state constraint Λ, the probabilities of error in (18) and (19) are computed with the maximization with respect to $\theta \in \Theta$ being now taken over all state sequences $\boldsymbol{s} \in \mathcal{S}^n$ which satisfy (24). Accordingly, the notion of capacity is defined.

The various notions of capacity for unknown channels described above are based on criteria involving error probabilities defined in terms of (18) and (19). The fact that these error probabilities are evaluated as being the largest with respect to the (unknown) parameter $\theta \in \Theta$ means that the resulting values of capacity can be attained when the channel uncertainty is at its severest during the course of a transmission, and, hence, in less severe instances as well. In the latter case, of course, these values may lead to a conservative assessment of coding performance.

In some situations, the system designer may have additional information concerning the vagaries of the unknown channel. For example, in a communication situation controlled by a jammer employing i.i.d. strategies, the system designer may have prior knowledge, based on past experience, of the jammer's relative predilections for the laws (indexed by θ) governing the i.i.d. strategies. In such cases, a Bayesian approach can be adopted where the previous model of the unknown channel comprising the family of channels (3) is augmented by considering θ to be a Θ-valued rv with a known (prior) probability distribution function (pdf) μ on Θ. Thus the transmitter and receiver, while unaware of the actual channel law (indexed by θ) governing a transmission, know the pdf μ of the rv θ. The corresponding maximum and average probabilities of error are now defined by suitably modifying (18) and (19); the maximization with respect to θ in (18) and (19) is replaced by expectation with respect to the law μ of the rv θ. When dealing with randomized codes or codes with stochastic encoders, we shall assume that all the rv's in the specification of such codes are independent of the rv θ. The associated notions of capacity are defined analogously as above, with appropriate modifications. For a given channel model, their values will obviously be no smaller than their counterparts for the more stringent criteria corresponding to (18) and (19), thereby providing a more optimistic assessment of coding performance. It should be noted, however, that this approach does not assure arbitrarily small probabilities of error for every channel in the family of channels (3); rather, probabilities of error are guaranteed to be small only when they are evaluated as averages over all the channels in the family (3) with respect to the (prior) law μ of θ. For this reason, in situations where there is a prior on Θ, the notion of "capacity versus outage" is sometimes preferred to the notion of capacity (see [102]).

Other kinds of situations can arise when the transmitter or receiver are provided with side information consisting of partial or complete knowledge of the exact parameter θ dictating a transmission, i.e., the channel law governing a transmission. We consider only a few such situations below; the reader is referred to [44, pp. 220–222 and 227–230] for a wider description. Consider first the case where the receiver alone knows the exact value of θ during a transmission. This situation can sometimes be reduced to that of an unknown channel without side information at the receiver, which has been described above, and hence does not lead to a new mathematical problem. This is seen by considering a new unknown channel with input alphabet $\mathcal{X}$, and with output alphabet $\tilde{\mathcal{Y}}$ which is an expanded version of the original output alphabet $\mathcal{Y}$, viz.

$$\tilde{\mathcal{Y}} = \mathcal{Y} \times \Theta \tag{25}$$

and specified by the family of channels

$$\{\tilde{W}(\boldsymbol{y}, \theta' | \boldsymbol{x}, \theta), \boldsymbol{x} \in \mathcal{X}^n, \boldsymbol{y} \in \mathcal{Y}^n, \theta \in \Theta, \theta' \in \Theta\}_{n=1}^{\infty} \tag{26}$$

where

$$\tilde{W}(\boldsymbol{y}, \theta' | \boldsymbol{x}, \theta) = \begin{cases} W(\boldsymbol{y}|\boldsymbol{x}; \theta), & \text{if } \theta' = \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

Of course, some structure may be lost in this construction (e.g., the finite cardinality of the output alphabet or the memory of the channel). A length-n block code for this channel is defined as a pair of mappings (f, ϕ), where the encoder f is defined in the usual manner by (10), while the decoder ϕ is a mapping

$$\phi: \mathcal{Y}^n \times \Theta \to \mathcal{M} \cup \{0\}. \tag{28}$$

We turn next to a case where the transmitter has partial or complete knowledge of θ prevalent during a transmission. For instance, consider communication over an AVC (5) with $\Theta = \mathcal{S}^\infty$ when the transmitter alone is provided, at each time instant $t = 1, \cdots, n$, a knowledge of all the past and present states $s_1, \cdots, s_t$ of the channel during a transmission. Then, a length-n block code is a pair of mappings (f, ϕ), where the decoder ϕ is defined as usual by (11), whereas the encoder f comprises a sequence of mappings $\{f_t\}_{t=1}^n$ with

$$f_t: \mathcal{M} \times \mathcal{S}^t \to \mathcal{X}. \tag{29}$$

This sequence of mappings determines the tth symbol of a codeword as a function of the transmitted message and the known past and present states of the channel. Significant benefits can be gained if the transmitter is provided state information in a noncausal manner (e.g., if the entire sequence of channel states $s_1, \cdots, s_n$ is known to the transmitter when transmission begins). The encoder f is then defined accordingly as a sequence of mappings $\{f_t\}_{t=1}^n$ with

$$f_t: \mathcal{M} \times \mathcal{S}^n \to \mathcal{X}. \tag{30}$$

Various combinations of the two cases just mentioned are, of course, possible with the transmitter and receiver possessing various degrees of knowledge about the exact value of θ during a transmission. In all these cases, the maximum and average probabilities of error are defined analogously as in (14) and (15), and the notion of capacity defined accordingly.

Yet another communication situation involves unknown channels with noiseless feedback from the receiver to the transmitter. At each time instant $t = 1, \cdots, n$, the transmitter knows the previous channel output symbols $y_1, \cdots, y_{t-1}$ through a noiseless feedback link. Now, in the formal definition of a length-n block code (f, ϕ), the decoder is given by (11) while the encoder f consists of a sequence of mappings $\{f_t\}_{t=1}^n$, where

$$f_t: \mathcal{M} \times \mathcal{Y}^{t-1} \mapsto \mathcal{X}. \tag{31}$$

Once again, the notion of capacity is defined accordingly.

We shall also consider the communication situation which obtains when list codes are used. Loosely speaking, in a list code, the decoder produces a list of messages, and the absence from the list of the message transmitted constitutes an error. When the size of the list is 1, the list coding problem reduces to the usual coding problem using codes as in (10) and (11). Formally, a length-n (block) list code of list size ν is a pair of mappings (f, ϕ), where the encoder f is defined by (10), while the (list) decoder ϕ is a mapping

$$\phi: \mathcal{Y}^n \to \mathcal{M}_\nu \cup \{0\} \tag{32}$$

where $\mathcal{M}_\nu$ is the set of all subsets of $\mathcal{M}$ with cardinality not exceeding ν. The rate of this list code with size ν is

$$\frac{1}{n} \log \frac{M}{\nu}. \tag{33}$$

The probability of error for the message $m \in \mathcal{M}$ when a list code (f, ϕ) with list size ν is used on a channel $\theta \in \Theta$ is defined analogously as in (13), with the modification that the sum in (13) is over those $\boldsymbol{y} \in \mathcal{Y}^n$ for which $m \notin \phi(\boldsymbol{y})$. The corresponding maximum and average probabilities of error are then defined accordingly, as is the notion of capacity.

III. Capacities

We now present some key results on channel capacity for the various channel models and performance criteria described in the previous section. Our presentation of results is not exhaustive, and seldom will the presented results be discussed in detail; instead, we shall often refer the reader to the bibliography for relevant treatments.

The literature on communication under channel uncertainty is vast, and our bibliography is by no means complete. Rather than directly citing all the literature relevant to a topic, we shall when possible, refer the reader to a textbook or a recent paper which contains a survey. The citations are thus intended to serve as pointers for further study of a topic, and not as indicators of where a result was first derived or where the most significant contribution to a subject was made.

In what follows, all channels will be assumed to have finite input and output alphabets, unless stated otherwise.

A. Discrete Memoryless Channels

We begin with the model originally treated by Shannon [111] of a known memoryless channel with finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. The channel law is given by (2) where $W(y|x)$ is known and fixed. For this model, the capacity is given by [111]

$$C = \max_{Q \in \mathcal{P}(\mathcal{X})} I(Q, W) \tag{34}$$

where $\mathcal{P}(\mathcal{X})$ denotes the set of all (input) pmf's on $\mathcal{X}$,

$$I(Q, W) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Q(x) W(y|x) \log \frac{W(y|x)}{(QW)(y)} \tag{35}$$

is the mutual information between the channel input and output, and

$$(QW)(y) = \sum_{x' \in \mathcal{X}} Q(x') W(y|x') \tag{36}$$

is the output pmf on $\mathcal{Y}$ which is induced when the channel input pmf is Q. This is the channel capacity regardless of whether the maximum or average probability of error criterion is used, and regardless of whether or not the transmitter and receiver have access to a common source of randomness. Moreover, a strong converse holds [124] so that

$$C = C_\epsilon, \qquad 0 < \epsilon < 1.$$

Upper and lower bounds on error exponents for the discrete memoryless channel can be found in [32], [44], [64], and in references therein.

Example 1 (Continued): The capacity C of a BSC with crossover probability p is given by [39], [44], [64]

$$C = 1 - h_b(p)$$

where

$$h_b(x) = -x \log x - (1-x)\log(1-x), \qquad x \in [0, 1]$$

is the binary entropy function.

In [114], Shannon considered a different model in which the channel law at time t depends on a state rv S_t, with values in a finite set $\mathcal{S}$, evolving in a memoryless (i.i.d.) fashion in accordance with a (known) pmf P_S on $\mathcal{S}$. When in state $s \in \mathcal{S}$, the channel obeys a transition law given by the stochastic matrix $\{W(y|x; s), x \in \mathcal{X}, y \in \mathcal{Y}\}$. The channel states are assumed to be known to the transmitter in a causal way, but unknown to the receiver. The symbol transmitted at time t may thus depend, not only on the message m, but also on present and past states $s_1^t = s_1, \cdots, s_t$ of the channel. A length-n block code (f, ϕ) consists of an encoder f which can be described as a sequence of mappings $\{f_t\}_{t=1}^n$ as in (29), while the decoder is defined as in (11). When such an encoding scheme is used, the probability $W(\boldsymbol{y}|f(m))$ that the channel output is $\boldsymbol{y}$ given that message m was transmitted, is

$$W(\boldsymbol{y}|f(m)) = \sum_{\boldsymbol{s}\in\mathcal{S}^n} \left(\prod_{i=1}^n P_S(s_i) \prod_{t=1}^n W\big(y_t|f_t(m, s_1^t); s_t\big) \right). \tag{37}$$

Shannon computed the capacity of this channel by observing that there is no loss in capacity if the output of the encoder is allowed to depend only on the message m and the *current* state s_t, and not on previous states s_1^{t-1}. As a consequence of this observation, we can compute channel capacity by considering a new memoryless channel $\tilde{W}$: $\mathcal{S}^{\mathcal{X}} \to \mathcal{Y}$ whose inputs are mappings from $\mathcal{S}$ to $\mathcal{X}$ and whose output is distributed for any input ψ: $\mathcal{S} \to \mathcal{X}$ according to

$$\tilde{W}(y|\psi) = \sum_{s\in\mathcal{S}} P_S(s) W(y|\psi(s); s). \tag{38}$$

Note that if neither transmitter nor receiver has access to state information, the channel becomes a simple memoryless one, and the results of [111] are directly applicable. Also note that in defining channel capacity, the probabilities of errors are averaged over the possible state sequences; performance is not guaranteed for every individual sequence of states. This problem is thus significantly different from the problem of computing the capacity of an AVC (5).

Regardless of whether or not the transmitter has state information, accounting for state information at the receiver poses no additional difficulty. The output alphabet can be augmented to account for this state information, e.g., by setting the new output alphabet to be

$$\tilde{\mathcal{Y}} = \mathcal{Y} \times \mathcal{S}. \tag{39}$$

For the corresponding new channel, with appropriate law, we can then use the results for the case where the receiver has no additional information. This technique also applies to situations where the receiver may have noisy observations of the channel states.

A variation of this problem was considered in [37], [67], [78], and in references therein, where state information is available to the transmitter in a *noncausal* way in that the entire realization of the i.i.d. state sequence is known when transmission begins. Such noncausal state information at the transmitter can be most beneficial (albeit rarely available) and can substantially increase capacity.

The inefficacy of feedback in increasing capacity was demonstrated by Shannon in [112]. For some of the results on list decoding, see [44], [55], [56], [62], [115], [120], and references therein.

1) The Compound Discrete Memoryless Channel: We now turn to the compound discrete memoryless channel, which models communication over a memoryless channel whose law is unknown but remains fixed throughout a transmission (see (4)). Both transmitter and receiver are assumed ignorant of the channel law governing the transmission; they only know the family Θ to which the law belongs. It should be emphasized that in this model no prior distribution on Θ is assumed, and in demonstrating the achievability of a rate R, we must therefore exhibit a code (f, ϕ) as in (10) and (11) which yields a small probability of error for *every* channel in the family.

Clearly, the highest achievable rate cannot exceed the capacity of any channel in the family, but this bound is not tight, as different channels in the family may have different capacity-achieving input pmf's. It is, however, true that the capacity of the compound channel is positive if and only if (iff) the infimum of the capacities of the channels in the family Θ is positive (see [126]).

The capacity of a compound DMC is given by the following theorem [30], [44], [52], [125], [126]:

Theorem 1: The capacity of the compound DMC (4), for both the average probability of error and the maximum probability of error, is given by

$$C = \max_{Q\in\mathcal{P}(\mathcal{X})} \inf_{\theta\in\Theta} I(Q, W(\cdot|\cdot; \theta)). \tag{40}$$

For the maximum probability of error, a strong converse holds so that

$$C = C_\epsilon^m, \qquad 0 < \epsilon < 1. \tag{41}$$

Note that the capacity value is not increased if the decoder knows the channel θ, but not the encoder. On the other hand, if the encoder knows the channel, then even if the decoder does not, the capacity is in general increased and is equal to the infimum of the capacities of the channels in the family [52], [125], [126].

Example 2 (Continued): The capacity of the compound DMC corresponding to a class of binary-symmetric channels is given by

$$C = \inf_{\theta\in\Theta} (1 - h_b(\theta)).$$

It is interesting to note that in this example the capacity of the family is the infimum of the capacities of the individual channels in the family. This always holds for memoryless families when the capacity-achieving input pmf is the same for all the channels in the family. In contrast, for families of channels with memory (Example 5), the capacity-achieving input pmf may be the same for all the channels in the family, and yet the capacity of the family can be strictly smaller than the infimum of the capacities of the individual channels.

Neither the direct part nor the converse of Theorem 1 follows immediately from the classical theorem on the capacity of a known DMC. The converse does not follow from (34) since the capacity in (40) may be strictly smaller than the capacity of any channel in the family. Nevertheless, an application of Fano's inequality and some convexity arguments [30] establishes the converse. A strong converse for the maximum probability of error criterion can be found in [44] and [126]. For the average probability of error, a strong converse need not hold [1], [44].

Proving the direct part requires showing that for any input pmf Q, any rate R, and any $\delta > 0$, there exists a sequence of encoders parametrized by the blocklength n that can be reliably decoded on any channel W that satisfies $I(Q, W) > R + \delta$. Moreover, the decoding rule must not depend on the channel. The receiver in [30] is a maximum-likelihood decoder with respect to a uniform mixture on a finite (but polynomial in the blocklength) set of DMC's which is in a sense dense in the class of all DMC's. The existence of a code is demonstrated using a random-coding argument. It is interesting to note [51], [119], that if the set of stochastic matrices $\{W(\cdot|\cdot;\, \theta), \theta \in \Theta\}$ is compact and convex, then the decoder can be chosen as the maximum-likelihood decoder for the DMC with stochastic matrix $W(\cdot|\cdot;\, \theta^*)$, where (Q^*, θ^*) is a saddle point for (40). The receiver can thus be a maximum-likelihood receiver with respect to the worst channel in the family.

Yet another decoder for the compound DMC is the maximum empirical mutual information (MMI) decoder [44]. This decoder will be discussed later in Section IV-B, when we discuss universal codes and the compound channel. The use of universal decoders for the compound channel is studied in [60] and [91], where a universal decoder for the class of finite-state channels is used to derive the capacity of a compound FSC. Another result on the compound channel capacity of a class of channels with memory can be found in [107] where the capacity of a class of Gaussian intersymbol interference channels is derived.

It should be noted that if the family of channels Θ is finite, then the problem is somewhat simplified and a Bayesian decoder [64, pp. 176–177] as well as a merged decoder, obtained by merging the maximum-likelihood decoders of each of the channels in the family [60], can be used to demonstrate achievability.

Cover [38] has shown interesting connections between communication over a compound channel and over a broadcast channel. An application of these ideas to communication over slowly varying flat-fading channels under the "capacity versus outage" criterion can be found in [109].

2) The Arbitrarily Varying Channel: The arbitrarily varying channel (AVC) was introduced by Blackwell, Breiman, and Thomasian [31] to model a memoryless channel whose law may vary with time in an arbitrary and unknown manner during the transmission of a codeword [cf. (5)]. The transmitter and receiver strive to construct codes for ensuring reliable communication, no matter which sequence of laws govern the channel during a transmission.

Formally, a discrete memoryless AVC with (finite) input alphabet $\mathcal{X}$ and (finite) output alphabet $\mathcal{Y}$ is determined by a family of channel laws $\{W(\cdot|\cdot;\, s), s \in \mathcal{S}\}$, each individual law in this family being identified by an index $s \in \mathcal{S}$ called the state. The state space $\mathcal{S}$, which is known to both transmitter and receiver, will be assumed to be also finite unless otherwise stated. The probability of receiving $\boldsymbol{y} \in \mathcal{Y}^n$, when $\boldsymbol{x} \in \mathcal{X}^n$ is transmitted and $\boldsymbol{s} \in \mathcal{S}^n$ is the channel state sequence, is given by (5). The standard AVC model introduced in [31], and subsequently studied by several authors (e.g., [2], [6], [10], [20], [45]), assumes that the transmitter and receiver are unaware of the actual state sequence $\boldsymbol{s} \in \mathcal{S}^n$ which governs a transmission. In the same vein, the "selector" of the state sequence $\boldsymbol{s}$, is ignorant of the actual message transmitted. However, the state "selector" is assumed to know the code when a deterministic code is used, and know the pmf generating the code when a randomized code is used (but not the actual codes chosen).[1]

There are a wide variety of challenging problems for the AVC. These depend on the nature of the performance criteria used (maximum or average probabilities of error), the permissible coding strategies (randomized codes, codes with stochastic encoders, or deterministic codes), and the degrees of knowledge of each other with which the codeword and state sequences are selected. For a summary of the work on AVC's through the late 1980's, and for basic results, we refer the reader to [6], [44], [47]–[49], and [126].

Before we turn to a presentation of key AVC results, it is useful to revisit the probability of error criteria in (18) and (19). Observe that in the definition of an ϵ-achievable rate (cf. Section II) on an AVC, the maximum (resp., average) probability of error criterion in (18) (resp., (19)) can be restated as

$$\max_{\boldsymbol{s} \in \mathcal{S}^n} e_{\max}(f, \phi, \boldsymbol{s}) \leq \epsilon \tag{42}$$

$$(\text{resp., } \max_{\boldsymbol{s} \in \mathcal{S}^n} \overline{e}(f, \phi, \boldsymbol{s}) \leq \epsilon) \tag{43}$$

with $e(m, f, \phi, \theta)$ in (13) now being replaced by

$$e(m, f, \phi, \boldsymbol{s}) = \sum_{\boldsymbol{y}:\ \phi(\boldsymbol{y}) \neq m} W(\boldsymbol{y}|f(m); \boldsymbol{s}). \tag{44}$$

In (42)–(44), recall that (f, ϕ) is a (deterministic) code of blocklength n. When a randomized code (F, Φ) is used, $e_{\max}(F, \Phi, \boldsymbol{s})$, $\overline{e}(F, \Phi, \boldsymbol{s})$, and $e(m, F, \Phi, \boldsymbol{s})$ will play the roles of $e_{\max}(f, \phi, \boldsymbol{s})$, $\overline{e}(f, \phi, \boldsymbol{s})$, and $e(m, f, \phi, \boldsymbol{s})$, respectively, in (42)–(44). Here, $e_{\max}(F, \Phi, \boldsymbol{s})$, $\overline{e}(F, \Phi, \boldsymbol{s})$, and $e(m, F, \Phi, \boldsymbol{s})$ are defined analogously as in (14)–(16), respectively.

[1] For the situation where a deterministic code is used and the state "selector" knows this code as well as the transmitted message, see [44, p. 233].

Given an AVC (5), let us denote by W_ζ, for any pmf ζ on $\mathcal{S}$, the "averaged" stochastic matrix $\mathcal{X} \to \mathcal{Y}$ defined by

$$W_\zeta(y|x) = \sum_{s\in\mathcal{S}} W(y|x;\, s)\zeta(s). \tag{45}$$

Further, let $\mathcal{P}(\mathcal{S})$ denote the set of all pmfs on $\mathcal{S}$.

The capacity of the AVC (5) for randomized codes is, of course, the same for the maximum and average probabilities of error, and is given by the following theorem [19], [31], [119].

Theorem 2: The randomized code capacity of the AVC (5) is given by

$$C = \max_{Q\in\mathcal{P}(\mathcal{X})} \min_{\zeta\in\mathcal{P}(\mathcal{S})} I(Q, W_\zeta) = \min_{\zeta\in\mathcal{P}(\mathcal{S})} \max_{Q\in\mathcal{P}(\mathcal{X})} I(Q, W_\zeta). \tag{46}$$

Further, a strong converse holds so that

$$C = C_\epsilon, \qquad 0 < \epsilon < 1. \tag{47}$$

The direct part of Theorem 2 can be proved [19] using a random-coding argument to show the existence of a suitable encoder. The receiver in [19] uses a (normalized) maximum-likelihood decoder for the DMC with stochastic matrix $W_{\zeta^*}\colon \mathcal{X} \to \mathcal{Y}$, where (Q^*, ζ^*) is a saddle point for (46). When input or state constraints are additionally imposed, the randomized code capacity $C(\Gamma, \Lambda)$ of the AVC (5), given below (cf. (48)), is achieved by a similar code with suitable modifications to accommodate the constraints [47].

The randomized code capacity of the AVC (5) under input constraint Γ and state constraint Λ (cf. (22), (24)), denoted $C(\Gamma, \Lambda)$, is determined in [47], and is given by

$$\begin{aligned} C(\Gamma, \Lambda) &= \max_{Q\in\mathcal{P}(\mathcal{X})\colon\ g(Q)\le\Gamma} \min_{\zeta\in\mathcal{P}(\mathcal{S})\colon\ l(\zeta)\le\Lambda} I(Q, W_\zeta) \\ &= \min_{\zeta\in\mathcal{P}(\mathcal{S})\colon\ l(\zeta)\le\Lambda} \max_{Q\in\mathcal{P}(\mathcal{X})\colon\ g(Q)\le\Gamma} I(Q, W_\zeta) \end{aligned} \tag{48}$$

where

$$g(Q) = \sum_{x\in\mathcal{X}} Q(x)g(x) \tag{49}$$

and

$$l(\zeta) = \sum_{s\in\mathcal{S}} \zeta(s)l(s). \tag{50}$$

Also, a strong converse exists. In the absence of input or state constraints, the corresponding value of the randomized code capacity of the AVC (5) is obtained from (48) by setting

$$\Gamma = g_{\max} \stackrel{\text{def}}{=} \max_{x\in\mathcal{X}} g(x)$$

or

$$\Lambda = l_{\max} \stackrel{\text{def}}{=} \max_{s\in\mathcal{S}} l(s).$$

It is further demonstrated in [47] that under weaker input and state constraints—in terms of expected values, rather than on individual codewords and state sequences as in (22) and (24)—a strong converse does not exist. (Similar results had been established earlier in [80] for a Gaussian AVC; see Section V below.)

Turning next to AVC performance using deterministic codes, recall that the capacity of a DMC (cf. (34)) or a compound channel (cf. (40)) is the same for randomized codes as well as for deterministic codes. An AVC, in sharp contrast, exhibits the characteristic that its deterministic code capacity is generally smaller than its randomized code capacity. In this context, it is useful to note that unlike in the case of a DMC (2), the existence of a randomized code (F, Φ) for an AVC (5) satisfying

$$\max_{\boldsymbol{s}\in\mathcal{S}^n} e_{\max}(F, \Phi, \boldsymbol{s}) \le \epsilon$$

or

$$\max_{\boldsymbol{s}\in\mathcal{S}^n} \overline{e}(F, \Phi, \boldsymbol{s}) \le \epsilon$$

does not imply the existence of a deterministic code (f, ϕ) (as a realization of the rv (F, Φ)) satisfying (42) and (43), respectively. Furthermore, in contrast to a DMC (2) or a compound channel (4), the deterministic code capacities C^m and C^a of the AVC (5) for the maximum and average probabilities of error, can be different;[2] specifically, C^m can be strictly smaller than C^a. An example [6] when $C^a > 0$ but $C^m = 0$ is the "deterministic" AVC with $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$, $\mathcal{S} = \{0, 1\}$, and $y = x + s$ modulo 3.

A computable characterization of C^m for an AVC (5) using deterministic codes, is a notoriously difficult problem which remains unsolved to date. Indeed, as observed by Ahlswede [2], it yields as a special case Shannon's famous graph-theoretic problem of determining the zero-error capacity of any DMC [96], [112], which remains a "holy grail" in information theory.

While C^m is unknown in general, a computable characterization is available in some special situations, which we next address. To this end, given an AVC (5), for any stochastic matrix $\tilde{\zeta}\colon \mathcal{X} \to \mathcal{S}$, we denote by $W_{\tilde{\zeta}}$ the "row-averaged" stochastic matrix $\mathcal{X} \to \mathcal{Y}$, defined by

$$W_{\tilde{\zeta}}(y|x) = \sum_{s\in\mathcal{S}} W(y|x;\, s)\,\tilde{\zeta}(s|x). \tag{51}$$

Further, let $\mathcal{P}(\mathcal{X} \to \mathcal{S})$ denote the set of stochastic matrices $\mathcal{X} \to \mathcal{S}$.

The capacity C^m of an AVC with a binary output alphabet was determined in [20] and is given by the following.

Theorem 3: The deterministic code capacity C^m of the AVC (5) for the maximum probability of error, under the condition $|\mathcal{Y}| = 2$, is given by

$$\begin{aligned} C^m &= \max_{Q\in\mathcal{P}(\mathcal{X})} \min_{\tilde{\zeta}\in\mathcal{P}(\mathcal{X}\to\mathcal{S})} I(Q, W_{\tilde{\zeta}}) \\ &= \min_{\tilde{\zeta}\in\mathcal{P}(\mathcal{X}\to\mathcal{S})} \max_{Q\in\mathcal{P}(\mathcal{X})} I(Q, W_{\tilde{\zeta}}). \end{aligned} \tag{52}$$

Further, a strong converse holds so that

$$C^m = C^m_\epsilon, \qquad 0 < \epsilon < 1. \tag{53}$$

[2] As a qualification, recall from Section III-A1) that for a compound channel (4), a strong converse holds for the maximum probability of error but not for the average probability of error.

The proof in [20] of Theorem 3 considers first the AVC (5) with binary input and output alphabets. A suitable code is identified for the DMC corresponding to the "worst row-averaged" stochastic matrix from among the family of stochastic matrices $W_{\tilde{\zeta}}\colon \mathcal{X} \to \mathcal{Y}$ (cf. 51) formed by varying $\tilde{\zeta} \in \mathcal{P}(\mathcal{X} \to \mathcal{S})$; this code is seen to perform no worse on any other DMC corresponding to a "row-averaged" stochastic matrix in said family. Finally, the case of a nonbinary input alphabet is reduced to that of a binary alphabet by using a notion of two "extremal" input symbols.

Ahlswede [10] showed that the formula for C^m in Theorem 3 is valid for a larger class of AVC's than in [20]. The direct part of the assertion in [10] entails a random selection of codewords combined with an expurgation, used in conjunction with a clever decoding rule.

The sharpest results on the problem of determining C^m for the AVC (5) are due to Csiszár and Körner [45], and are obtained by a combinatorial approach developed in [44] and [46]. The characterization of C^m in [45] requires additional terminology. Specifically, we shall say that the $\mathcal{X}$-valued rv's X, X', with the same pmf Q, are connected a.s. by the stochastic matrix $W\colon \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ appearing in (5), denoted $X \overset{W}{\sim} X'$, iff there exist pmf's ζ, ζ' on $\mathcal{S}$ such that

$$\Pr\left\{\sum_{s\in\mathcal{S}} W(y|X, s)\zeta(s) = \sum_{s\in\mathcal{S}} W(y|X', s)\zeta'(s) \quad \text{for every } y \in \mathcal{Y}\right\} = 1. \tag{54}$$

Also, define

$$D(Q) = \min_{X, X'\colon\ P_X = P_{X'} = Q,\, X \overset{W}{\sim} X'} I(X \wedge X') \tag{55}$$

where P_X denotes the pmf of the rv X. The following characterization of C^m in [45] is more general than previous characterizations in [10] and [20].

Theorem 4: For the AVC (5), for every pmf $Q \in \mathcal{P}(\mathcal{X})$

$$\min\left\{\min_{\tilde{\zeta}\in\mathcal{P}(\mathcal{X}\to\mathcal{S})} I(Q, W_{\tilde{\zeta}}), D(Q)\right\}$$

is an achievable rate for the maximum probability of error. In particular, for $(Q^*, \tilde{\zeta}^*)$, a saddle point for (52), if Q^* is such that $D(Q^*) \geq I(Q^*, W_{\tilde{\zeta}^*})$, then

$$C^m = I(Q^*, W_{\tilde{\zeta}^*}). \tag{56}$$

The direct part of Theorem 4 uses a code in which the codewords are identified by random selection from sequences of a fixed "type" (cf. e.g., [44, Sec. 1.2]), using suitable large deviation bounds. The decoder combines a "joint typicality" rule with a threshold decision rule based on empirical mutual information quantities (cf. Section IV-B6) below).

Upon easing the performance criterion to be now the average probability of error, the deterministic code capacity C^a of the AVC (5) is known. In a key paper, Ahlswede [6] observed that the AVC capacity C^a displays a dichotomy: it either equals the AVC randomized code capacity or else is zero. Ahlswede's alternatives [6] can be stated as

$$C^a = \max_{Q\in\mathcal{P}(\mathcal{X})} \min_{\zeta\in\mathcal{P}(\mathcal{S})} I(Q, W_\zeta), \qquad \text{or else } C^a = 0. \tag{57}$$

The proof of (57) in [6] used an "elimination" technique consisting of two key steps. The first step was the discovery of "random code reduction," namely, that the randomized code capacity of the AVC can be achieved by a randomized code restricted to random selections from "exponentially few" deterministic codes, e.g., from no more than n^2 deterministic codes, where n is the blocklength. Then, if $C^a > 0$, the second step entailing an "elimination of randomness," i.e., the conversion of this randomized code into a deterministic code, is performed by adding short prefixes to the original codewords so as to inform the decoder which of the n^2 deterministic codes is actually used; the overall rate of the deterministic code is, of course, only negligibly affected by the addition of the prefixes.

A necessary and sufficient computable characterization of AVC's for deciding between the alternatives in (57) was not provided in [6]. This lacuna was partially filled by Ericson [59] who gave a necessary condition for the deterministic code capacity C^a to be positive. By enlarging on an idea in [31], it was shown [59] that if the AVC state "selector" could emulate the channel input by means of a fictitious auxiliary channel (defined in terms of a suitable stochastic matrix $U\colon \mathcal{X} \to \mathcal{S}$), then the decoder fails to discern between the channel input and state, resulting in $C^a = 0$.

Formally, we say that an AVC (5) is *symmetrizable* if for some stochastic matrix $U\colon \mathcal{X} \to \mathcal{S}$

$$\sum_{s\in\mathcal{S}} W(y|x; s)U(s|x') = \sum_{s\in\mathcal{S}} W(y|x'; s)U(s|x), \qquad x \in \mathcal{X},\, x' \in \mathcal{X},\, y \in \mathcal{Y}. \tag{58}$$

Let $\mathcal{U}(\mathcal{X} \to \mathcal{S})$ denote the set of all "symmetrizing" stochastic matrices $U\colon \mathcal{X} \to \mathcal{S}$ which satisfy (58). An AVC (5) for which $\mathcal{U}(\mathcal{X} \to \mathcal{S}) = \emptyset$ is termed nonsymmetrizable. Thus it is shown in [59] that if an AVC (5) is such that its deterministic code capacity C^a is positive, then the AVC (5) must be nonsymmetrizable.

A computable characterization of AVC's with positive deterministic code capacity C^a was finally completed by Csiszár and Narayan [48], who showed that nonsymmetrizability is also a sufficient condition for $C^a > 0$. The proof technique in [48] does not rely on the existence of the dichotomy as asserted by (57); nor does it rely on the fact, used materially in [6] to establish (57), that

$$\max_{Q\in\mathcal{P}(\mathcal{X})} \min_{\zeta\in\mathcal{P}(\mathcal{S})} I(Q, W_\zeta)$$

is the randomized code capacity of the AVC (5). The direct part in [48] uses a code with the codewords chosen at random from sequences of a fixed type, and selectively identified by a generalized Chernoff bounding technique due to Dobrushin and Stambler [53]. The linchpin is a subtle decoding rule which decides on the basis of a joint typicality test together with a threshold test using empirical mutual information quantities, similarly as in [45]. A key step of the proof is to show

that the decoding rule is unambiguous as a consequence of the nonsymmetrizability condition. An adequate bound on the average probability of error is then obtained in a standard manner using the method of types (cf. e.g., [44]).

The results in [6], [48], and [59] collectively provide the following characterization of C^a in [48].

Theorem 5: The deterministic code capacity C^a of the AVC (5) for the average probability of error is positive iff the AVC (5) is nonsymmetrizable. If $C^a > 0$, it equals the randomized code capacity of the AVC (5) given by (46), i.e.,

$$C^a = \max_{Q \in \mathcal{P}(\mathcal{X})} \min_{\zeta \in \mathcal{P}(\mathcal{S})} I(Q, W_\zeta). \tag{59}$$

Furthermore, if the AVC (5) is nonsymmetrizable, a strong converse holds so that

$$C^a = C^a_\epsilon, \qquad 0 < \epsilon < 1. \tag{60}$$

It should be noted that sufficient conditions for the AVC (5) to have a positive deterministic code capacity C^a had been given earlier in [6] and [53]; these conditions, however, are not necessary in general. Also, a necessary and sufficient condition for $C^a > 0$, albeit in terms of noncomputable "product space characterization" (cf. [44, p. 259]) appeared in [6]. The nonsymmetrizability condition above can be regarded as "single-letterization" of the condition in [6]. For a comparison of conditions for $C^a > 0$, we refer the reader to [49, Appendix I].

Yet another means of determining the deterministic code capacity C^a of the AVC (5) is derived as a special case of recent work by Ahlswede and Cai [15] which completely resolves the deterministic code capacity problem for a multiple-access AVC for the average probability of error. For the AVC (5), the approach in [15], in effect, consists of elements drawn from both [6] and [48]. In short, by [15], if the AVC (5) is nonsymmetrizable, then a code with the decoding rule proposed in [48] can be used to achieve "small" positive rates. Thus $C^a > 0$, whereupon the "elimination technique" of [6] is applied to yield that C^a equals the randomized code capacity given by (46).

We consider next the deterministic code capacity of the AVC (5) for the average probability of error, under input and state constraints (cf. (21) and (24)). To begin with, assume the imposition of only a state constraint but no input constraint. Let $C^a(\Lambda)$ denote the capacity of the AVC (5) under state constraint Λ (cf. (24)). If the AVC is nonsymmetrizable then, by Theorem 5, its capacity C^a without state constraint is positive and, obviously, so too is its capacity $C^a(\Lambda)$ under state constraint Λ for every $0 \le \Lambda \le l_{\max}$. The elimination technique in [6] can be applied to show that $C^a(\Lambda)$ equals the corresponding randomized code capacity under state constraint Λ (and no input constraint) given by (48) as $C(g_{\max}, \Lambda)$. On the other hand, if the AVC (5) is symmetrizable, by Theorem 5, its capacity C^a without state constraint is zero. However, the capacity $C^a(\Lambda)$ under state constraint Λ may yet be positive. In order to determine $C^a(\Lambda)$, the elimination technique in [6] can no longer be applied; while the first step of "random code reduction" is valid, the second step of "elimination of randomness" cannot be performed unless the capacity without state constraint C^a is itself positive. The reason, loosely speaking, is that if C^a were zero, the state "selector" could prevent reliable communication by foiling reliable transmission of the prefix which identifies the codebook actually selected in the first step; to this end, the state "selector" could operate in an unconstrained manner during the (relatively) brief transmission of the prefix thereby denying it positive capacity, while still satisfying state constraint Λ over the entire duration of the transmission of the prefix and the codeword.

The capacity $C^a(\Lambda)$ of the AVC (5), in general, is determined in [48] by extending the approach used therein for characterizing C^a. A significant role is played by the functional $\Lambda_0(Q)$, $Q \in \mathcal{P}(\mathcal{X})$, defined by

$$\Lambda_0(Q) = \min_{U \in \mathcal{U}} \sum_{x \in \mathcal{X},\, s \in \mathcal{S}} Q(x) U(s|x) l(s) \tag{61}$$

with $\Lambda_0(Q) = \infty$ if $\mathcal{U} = \emptyset$, i.e., if the AVC (5) is nonsymmetrizable. The capacity $C^a(\Lambda)$ under state constraint Λ is shown in [48] to be zero if $\max_{Q \in \mathcal{P}(\mathcal{X})} \Lambda_0(Q)$ is smaller than Λ; on the other hand, $C^a(\Lambda)$ is positive and equals

$$C^a(\Lambda) = \max_{Q \in \mathcal{P}(\mathcal{X}):\ \Lambda_0(Q) \ge \Lambda} \min_{\zeta \in \mathcal{P}(\mathcal{S}):\ l(\zeta) \le \Lambda} I(Q, W_\zeta) \qquad \text{if } \max_{Q \in \mathcal{P}(\mathcal{X})} \Lambda_0(Q) > \Lambda. \tag{62}$$

In particular, it is possible that $C^a(\Lambda)$ lies strictly between zero and the randomized code capacity under state constraint Λ which, by (48), equals $C(g_{\max}, \Lambda)$; this represents a departure from the dichotomous behavior observed in the absence of any state constraint (cf. (57)). A comparison of (48) and (62) shows that if the maximum in (48) is not achieved by an input pmf $Q \in \mathcal{P}(\mathcal{X})$ which satisfies $\Lambda_0(Q) \ge \Lambda$, then $C^a(\Lambda)$ is strictly smaller than $C(g_{\max}, \Lambda)$, while still being positive if the hypothesis in (62) holds, i.e.,

$$\max_{Q \in \mathcal{P}(\mathcal{X})} \Lambda_0(Q) > \Lambda.$$

Next, if an input constraint Γ (cf. (21)) is also imposed, the capacity $C^a(\Gamma, \Lambda)$ is given in [48] by the following.

Theorem 6: The deterministic code capacity $C^a(\Gamma, \Lambda)$ of the AVC (5) under input constraint Γ and state constraint Λ, for the average probability of error, is given by (63) at the bottom of this page. Further, in the cases considered in (63), a strong converse holds so that

$$C^a(\Gamma, \Lambda) = C^a_\epsilon(\Gamma, \Lambda), \qquad 0 < \epsilon < 1. \tag{64}$$

$$C^a(\Gamma, \Lambda) = \begin{cases} \max\limits_{Q \in \mathcal{P}(\mathcal{X}):\ \Lambda_0(Q) \ge \Lambda,\, g(Q) \le \Gamma} \min\limits_{\zeta \in \mathcal{P}(\mathcal{S}):\ l(\zeta) \le \Lambda} I(Q, W_\zeta) > 0, & \text{if } \max\limits_{Q \in \mathcal{P}(\mathcal{X}):\ g(Q) \le \Gamma} \Lambda_0(Q) > \Lambda \\ 0, & \text{if } \max\limits_{Q \in \mathcal{P}(\mathcal{X}):\ g(Q) \le \Gamma} \Lambda_0(Q) < \Lambda. \end{cases} \tag{63}$$

The case

$$\max_{Q\in\mathcal{P}(X):\ g(Q)\le\Gamma} \Lambda_0(Q) = \Lambda$$

remains unresolved in general; for certain AVC's, $C^a(\Gamma, \Lambda)$ equals zero in this case too (cf. [48, remark following the proof of Theorem 3]). Again, it is possible that $C^a(\Gamma, \Lambda)$ lies strictly between zero and the randomized code capacity $C(\Gamma, \Lambda)$ under input constraint Γ and state constraint Λ given by (48).

The results of Theorem 6 lead to some interesting combinatorial interpretations (cf. [48, Example 1] and [49, Sec. III]).

Example 3 (Continued): We refer the reader to [48, Example 2] for a full treatment of this example. For a pmf $Q = (1-q, q)$ on the input alphabet $\{0, 1\}$, and a pmf $\zeta = (1-r, r)$ on the state space $\{0, 1\}$, we obtain from (35) and (45) that

$$\begin{aligned} I(Q, W_\zeta) &= \tilde{I}(q, r) \\ &= H(qr, (1-q)(1-r), q+r-2qr) - h_b(r) \end{aligned}$$

where H denotes entropy. The randomized code capacity of the AVC in Example 3 is then given by Theorem 2 as (cf. (46))

$$C = \max_q \min_r \tilde{I}(q, r) + 1/2 \tag{65}$$

where $q = r = 1/2$ is a saddle point for $\tilde{I}(q, r)$. Turning to the deterministic code capacity C^a for the average probability of error, note that the symmetrizability condition (58) is satisfied iff the stochastic matrix $U\colon \mathcal{X} \to \mathcal{S}$ is the identity matrix. By Theorem 5, we have $C^a = 0$; obviously, the deterministic code capacity for the maximum probability of error is then $C^m = 0$. Thus in the absence of input or state constraints, the randomized code capacity C is positive while the deterministic code capacities C^a and C^m are zero.

We now consider AVC performance under input and state constraints. Let the functions $g(x) = x$, $x \in \{0, 1\}$, and $l(s) = s$, $s \in \{0, 1\}$, be used in the input and state constraints (cf. (20)–(24)). Thus $g(\boldsymbol{x})$ in (20) and $l(\boldsymbol{s})$ in (23) are the normalized Hamming weights of the n-length binary sequences $\boldsymbol{x}$ and $\boldsymbol{s}$. Then the randomized code capacity $C(\Gamma, \Lambda)$ under the input constraint Γ and state constraint Λ, $0 \le \Gamma, \Lambda \le 1$, is given by (48) as

$$C(\Gamma, \Lambda) = \max_{q\le\Gamma} \min_{r\le\Lambda} \tilde{I}(q, r). \tag{66}$$

In particular

$$C(\Gamma, \Lambda) = C = 1/2, \qquad \text{if } \Gamma, \Lambda \ge 1/2. \tag{67}$$

Next, we turn to the deterministic code capacity $C^a(\Gamma, \Lambda)$ for the average probability of error. It is readily seen from (49), (50), and (61) that $\Lambda_0(Q) = q$, $g(Q) = q$, and $l(\zeta) = r$, respectively. It then follows from Theorem 6 that (cf. [47, Example 2])

$$C^a(\Gamma, \Lambda) = \begin{cases} 0, & \text{if } \Gamma \le \Lambda \\ \max\limits_{\Lambda\le q\le\Gamma} \min\limits_{r\le\Lambda} \tilde{I}(q, r), & \text{if } \Gamma > \Lambda. \end{cases} \tag{68}$$

We can conclude from (66)–(68) (cf. [48, Example 2]) that for $1/2 \le \Gamma \le \Lambda$, it holds that $C^a(\Gamma, \Lambda) = 0$ while $C(\Gamma, \Lambda) = 1/2$. Next, if $\Gamma > \Lambda > 1/2$, we have that $C^a(\Gamma, \Lambda)$ is positive but smaller that $C(\Gamma, \Lambda)$. On the other hand, if $\Lambda \le 1/2$, $\Gamma > \Lambda$, then $C^a(\Gamma, \Lambda) = C(\Gamma, \Lambda)$. Thus under state constraint Λ, several situations exist depending on the value of Λ, $0 \le \Lambda \le 1$. The deterministic code capacity for the average probability of error can be zero while the corresponding randomized code capacity is positive. Further, the former can be positive and yet smaller than the latter; or it could equal the latter.

Several of the results described above from [47]–[49] on the randomized as well as the deterministic code capacities of the AVC (5) with input constraint Γ and state constraint Λ have been extended by Csiszár to AVC's with general input and output alphabets and state space; see [41].

It remains to characterize AVC performance using codes with stochastic encoders. For the AVC (5) without input or state constraints, the following result is due to Ahlswede [6].

Theorem 7: For codes with stochastic encoders, the capacities of the AVC (5) for the maximum as well as the average probabilities of error equal its deterministic code capacity for the average probability of error.

Thus by Theorem 7, when the average probability of error criterion is used, codes with stochastic encoders offer no advantage over deterministic codes in terms of yielding a larger capacity value. However, for the maximum probability of error criterion, the former can afford an improvement over the latter, since the AVC capacity is now raised to its value under the (less stringent) average probability of error criterion. The previous assertion is proved in [6] using the "elimination technique." If state constraints (cf. (24)) are additionally imposed on the AVC (5), the previous assertion still remains true even though the "elimination technique" does not apply in the presence of state constraints (cf. [48, Sec. V]).

We next address AVC performance when the transmitter or receiver are provided with side information. Consider first the situation where this side information consists of partial or complete knowledge of the sequence of states $s_1, \cdots, s_n$ prevalent during a transmission. The reader is referred to [44, pp. 220–222 and 227–230] for a compendium of several relevant problems and results. We cite here a paper of Ahlswede [11] in which, using previous results of Gel'fand and Pinsker [67], the deterministic code capacity problem is fully solved in the case when the state sequence is known to the transmitter in a noncausal manner. Specifically, the deterministic code capacity of the AVC (5) for the maximum probability of error, when the transmitter alone is aware of the entire sequence of channel states $s_1, \cdots, s_n$ when transmission begins (cf. (30)), is characterized in terms of mutual information quantities obtained in [67]. Further, this capacity is shown to coincide with the corresponding deterministic code capacity for the average probability of error. The proof entails a combination of the aforementioned "elimination technique" with the "robustification technique" developed in [8] and [9]. The situation considered above in [11] is to be contrasted with that in [13], [67], and [78] where the channel states $s_1, \cdots, s_n$, which are known to the transmitter alone at the

commencement of transmission, constitute a realization of an i.i.d. sequence with (known) pmf μ on $\mathcal{S}$. The corresponding maximum probability of error is now defined by replacing the maximization with respect to $\boldsymbol{s} \in \mathcal{S}^n$ in (42) by expectation with respect to the pmf on S^n induced by μ.

If the state sequence $s_1, \cdots, s_n$ is known to the receiver alone, the resulting AVC performance can be readily characterized in terms of that of a new AVC with an expanded output alphabet but without any side information, and hence does not lead to a new mathematical problem as observed earlier in Section II (cf. (25)–(28)). Note that the decoder ϕ of a length-n block code (f, ϕ) is now of the form

$$\phi\colon \mathcal{Y}^n \times \mathcal{S}^n \to \mathcal{M} \cup \{0\} \tag{69}$$

while the encoder f is as usually defined by (10). The deterministic code capacities of the AVC (5), with the channel states $s_1, \cdots, s_n$ known to the receiver, for the maximum and average probabilities of error, can then be seen to be the same as the corresponding capacities—without any side information at the receiver—of a new AVC with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y} \times \mathcal{S}$, and stochastic matrix $\tilde{W}\colon \mathcal{X} \times \mathcal{S} \to \mathcal{Y} \times \mathcal{S}$ defined by

$$\tilde{W}(y, s|x, s') = \begin{cases} W(y|x, s), & s' = s \\ 0, & s' \neq s. \end{cases} \tag{70}$$

Using this technique, it was shown by Stambler [118] that this deterministic code capacity for the average probability of error equals

$$\max_{Q \in \mathcal{P}(\mathcal{X})} \min_{s \in \mathcal{S}} I(Q, W(\cdot|\cdot; s))$$

which is the capacity of the compound DMC (cf. (3) and (4)) corresponding to the family of DMC's with stochastic matrices $\{W(y|x; s), x \in \mathcal{X}, y \in \mathcal{Y}, s \in \mathcal{S}\}$ (cf. Theorem 1).

Other forms of side information provided to the transmitter or receiver can significantly improve AVC performance. For instance, if noiseless feedback is available from the receiver to the transmitter (cf. (31)), it can be used to establish "common randomness" between them (whereby they have access to a common source of randomness with probability close to 1), so that the deterministic code capacity C^a of the AVC (5) for the average probability of error equals its randomized code capacity given by Theorem 2. For more on this result due to Ahlswede and Csiszár, as also implications of "common randomness" for AVC capacity, see [18]. Ahlswede and Cai [17] have examined another situation in which the transmitter and receiver observe the components $U_1, \cdots, U_n$ and $V_1, \cdots, V_n$, respectively, of a memoryless correlated source $\{(U_t, V_t)\}_{t=1}^{\infty}$ (i.e., an i.i.d. process with generic rv's (U, V) which satisfy $I(U \wedge V) > 0$), and have shown that C^a equals the randomized code capacity given by Theorem 2.

The performance of an AVC (5) using deterministic list codes (cf. (32) and (33)) is examined in [5], [12], [14], [33]–[35], [82], and [83]. The value of this capacity for the maximum probability of error and vanishingly small list rate was determined by Ahlswede [5]. Lower bounds on the sizes of constant lists for a given average probability of error and an arbitrarily small maximum probability of error, respectively, were obtained by Ahlswede [5] and Ahlswede and Cai [14]. The fact that the deterministic list code capacity of an AVC (5) for the average probability of error displays a dichotomy similar to that described by (57) was observed by Blinovsky and Pinsker [34] who also determined a threshold for the list size above which said capacity equals the randomized code capacity given by Theorem 2. A complete characterization of the deterministic list code capacity for the average probability of error, based on an extended notion of symmetrizability (cf. (58)), was obtained by Blinovsky, Narayan, and Pinsker [33] and, independently, by Hughes [82], [83].

We conclude this section by noting the role of compound DMC's and AVC's in the study of communication situations partially controlled by an adversarial jammer. For dealing with such situations, several authors (cf. e.g., [36], [79], and [97]) have proposed a game-theoretic approach which involves a two-person zero-sum game between the "communicator" and the "jammer" with mutual information as the payoff function. An analysis of the merits and limitations of this approach from the viewpoint of AVC theory is provided in [49, Sec. VI]. See also [44, pp. 219–222 and 226–233].

B. Finite-State Channels

The capacity of a finite-state channel (7) has been studied under various conditions in [29], [64], [113], and [126]. Of particular importance is [64], where error exponents for a general finite-state channel are also computed. Before stating the capacity theorem for this channel, we introduce some notation [64], [91]. A (known) finite-state channel is specified by a pmf π on the initial state[3] in Σ and a conditional pmf $\{W(y, \sigma'|x, \sigma), x \in \mathcal{X}, y \in \mathcal{Y}, \sigma, \sigma' \in \Sigma\}$ as in (7). For such a channel, the probability $W_n(\boldsymbol{y}, \sigma_n|\boldsymbol{x}, \sigma_0)$ that the channel output is $\boldsymbol{y} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$ and the final channel state is σ_n, conditioned on the initial state $\sigma_0 \in \Sigma$ and the channel input $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$, is given by

$$W_n(\boldsymbol{y}, \sigma_n|\boldsymbol{x}, \sigma_0) = \sum_{\sigma_1, \cdots, \sigma_{n-1}} \prod_{i=1}^{n} W(y_i, \sigma_i|x_i, \sigma_{i-1}). \tag{71}$$

We can sum this probability over σ_n to obtain the probability that the channel output is $\boldsymbol{y} = (y_1, \cdots, y_n) \in \mathcal{Y}^n$ conditioned on the initial state $\sigma_0 \in \Sigma$ and the channel input $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$

$$W_n(\boldsymbol{y}|\boldsymbol{x}, \sigma_0) = \sum_{\sigma_1, \cdots, \sigma_n} \prod_{i=1}^{n} W(y_i, \sigma_i|x_i, \sigma_{i-1}). \tag{72}$$

Averaging (72) with respect to the pmf π of the initial state yields (7).

Given an initial state $\sigma_0 \in \Sigma$ and a pmf Q_n on $\mathcal{X}^n$, the joint pmf of the channel input and output is well-defined, and the mutual information between the input and the output is

[3] In [64], no prior pmf on the initial state is assumed and the finite-state channel is treated as a family of channels, corresponding to different initial states which may or may not be known to the transmitter or receiver.

given by

$$I(\boldsymbol{X} \wedge \boldsymbol{Y}|\sigma_0) = I(Q_n, W_n(\cdot|\cdot, \sigma_0)) = \sum_{\boldsymbol{x}, \boldsymbol{y}} Q_n(\boldsymbol{x}) W_n(\boldsymbol{y}|\boldsymbol{x}, \sigma_0) \cdot \ln \frac{W_n(\boldsymbol{y}|\boldsymbol{x}, \sigma_0)}{\sum_{\boldsymbol{x}'} Q_n(\boldsymbol{x}') W_n(\boldsymbol{y}|\boldsymbol{x}', \sigma_0)}.$$

Similarly, a family of finite-state channels, as in (8), can be specified in terms of a family of conditional pmf's $\{W(y, \sigma'|x, \sigma; \theta), x \in \mathcal{X}, y \in \mathcal{Y}, \sigma, \sigma' \in \Sigma, \theta \in \Theta\}$, and in analogy with (71) and (72), we denote by $W_n(\boldsymbol{y}, \sigma_n|\boldsymbol{x}, \sigma_0; \theta)$ the probability that the output of channel θ is $\boldsymbol{y} \in \mathcal{Y}^n$ and the final state is $\sigma_n \in \Sigma$ conditioned on the input $\boldsymbol{x} \in \mathcal{X}^n$ and initial state $\sigma_0 \in \Sigma$, and by $W_n(\boldsymbol{y}|\boldsymbol{x}, \sigma_0; \theta)$ 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 σ_0, and a pmf Q_n on $\mathcal{X}^n$, the mutual information between the input and output of the channel θ is given by

$$I(\boldsymbol{X} \wedge \boldsymbol{Y}|\sigma_0, \theta) = I(Q_n; W_n(\cdot|\cdot, \sigma_0; \theta)) = \sum_{\boldsymbol{x}, \boldsymbol{y}} Q_n(\boldsymbol{x}) W_n(\boldsymbol{y}|\boldsymbol{x}, \sigma_0; \theta) \cdot \ln \frac{W_n(\boldsymbol{y}|\boldsymbol{x}, \sigma_0; \theta)}{\sum_{\boldsymbol{x}'} Q_n(\boldsymbol{x}') W_n(\boldsymbol{y}|\boldsymbol{x}', \sigma_0; \theta)}.$$

The following is proved in [64].

Theorem 8: If a finite-state channel (7) is indecomposable [64] or if $\pi(\sigma_0) > 0$ for every $\sigma_0 \in \Sigma$, then its capacity C is given by

$$C = \lim_{N \to \infty} \max_{Q_N \in \mathcal{P}(\mathcal{X}^N)} \min_{\sigma_0 \in \Sigma} I(Q_N, W_N(\cdot|\cdot; \sigma_0)).$$

It should be noted that the capacity of the finite-state channel [64] can be estimated arbitrarily well, since there exist a sequence of lower bounds and a sequence of upper bounds which converge to it [64].

Example 4 (Continued): Assuming that neither b nor g takes the extreme values 0 or 1, the capacity of the Gilbert–Elliott channel [101] is given by

$$C = 1 - h(Z)$$

where $h(Z)$ is the entropy rate of the hidden Markov process $\{Z_t\}$.

Theorem 8 can also be used when the sequence of states $\sigma_0, \cdots, \sigma_n$ of the channel during a transmission is known to the receiver (but not to the transmitter). We can consider a new output alphabet $\mathcal{Y} \times \Sigma$, with corresponding transitions probabilities. The resulting channel is still a finite-state channel.

The capacity of the channel when the sequence of states is unknown to the receiver but known to the transmitter in a causal manner, was found in [86], thus extending the results of [114] to finite-state channels. Once again, knowledge at the receiver can be treated by augmenting the output alphabet. A special case of the transmitter and receiver both knowing the state sequence in a causal manner, obtains when the state is "computable at both terminals," which was studied by Shannon [113]. In this situation, given the initial state (assumed known to both transmitter and receiver), the transmitter can compute the subsequent states based on the channel input, and the receiver can compute the subsequent states based on the received signal.

1) The Compound Finite-State Channel: In computing the capacity of a class of finite-state channels (8), we shall assume that for every pair $(\pi, W) \in \Theta$ of pmf π of the initial state and conditional pmf $W(y, \sigma|x, \sigma')$, we have

$$(\pi, W) \in \Theta \quad \text{implies} \quad (\pi_u, W) \in \Theta \tag{73}$$

where π_u is the uniform distribution on Σ. We are, thus, assuming that reliable communication must be guaranteed for every initial state and any transition law, and that neither is known to the transmitter and receiver. Under this assumption we have the following [91].

Theorem 9: Under the assumption (73), the capacity C of a family Θ of finite-state channels (8) with common (finite) input, output, and state alphabets $\mathcal{X}$, $\mathcal{Y}$, Σ, is given by

$$C = \lim_{n \to \infty} \max_{Q_n \in \mathcal{P}(\mathcal{X}^n)} \inf_{\theta, \sigma_0} \frac{1}{n} I(Q_n; W_n(\cdot|\cdot, \sigma_0, \theta)). \tag{74}$$

Example 5 (Continued): If the transition probabilities of the underlying Markov chains of the different channels are uniformly bounded away from zero, i.e.,

$$\inf_{\theta \in \Theta} \min\{g(\theta), b(\theta), 1 - g(\theta), 1 - b(\theta)\} > 0 \tag{75}$$

then the capacity of the family is the infimum of the capacities of the individual channels in the family [91].

The following example demonstrates that if (75) is violated, the capacity of the family may be smaller than the infimum of the capacities of its members [91]. 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 $k \geq 1$. For any given k, we can achieve rates exceeding $1 - h_b(1/4)$ 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

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

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

The proof of Theorem 9 relies on the existence of a universal decoder for the class of finite-state channels [60], and on the fact that for rates below C the random-coding error probability

(for the natural choice of codebook distribution) is bounded above *uniformly* for all the channels in Θ by an exponentially decreasing function of the blocklength.

The similarity of the expressions in (40) and (74) should not lead to a mistaken belief that the capacity of any family of channels is given by a max inf expression. A counterexample is given in [31], and [52], and is repeated in [91].

IV. Encoders and Decoders

A variety of encoders and decoders have been proposed for achieving reliable communication over the different channel models described in Section II, and, in particular, for establishing the direct parts of the results on capacities described in Section III. The choices run the gamut from standard codes with randomly selected codewords together with a "joint typicality" decoder or a maximum-likelihood decoder for known channels, to codes consisting of fairly complex decoders for certain models of unknown channels. We shall survey below some of the proposed encoders and decoders, with special emphasis on the latter. While it is customary to study the combined performance of an encoder–decoder pair in a given communication situation, we shall—for the sake of expository convenience—describe encoders and decoders separately.

A. Encoders

The encoders chosen for establishing the capacity results stated in Section III, for various models of known and unknown channels described in Section II, often use randomly selected codewords in one form or another [111]. The notion of random selection of codewords affords several uses. The classical application, of course, involves randomly selected codewords as a mathematical artifice in proving, by means of the random-coding argument technique, the existence of deterministic codes for the direct parts of capacity results for known channels and certain types of unknown channels. Second, codewords chosen by random selection afford an obvious means of constructing randomized codes or codes with stochastic encoders for enhancing reliable communication over some unknown channels (cf. Section IV-A2)), thereby serving as models of practical engineering devices. Furthermore, the notion of random selection can lead to the selective identification of deterministic codewords with refined properties which are useful for determining the deterministic code capacities of certain unknown channels (cf. Section IV-A3)).

We first present a brief description of some standard methods of picking codewords by random selection.

1) Encoding by Random Selection of Codewords: One standard method of random selection of codewords entails picking them in an i.i.d. manner according to a fixed pmf Q_n on $\mathcal{X}^n$. Specifically, let $\boldsymbol{X}_m$, $m \in \mathcal{M}$, be i.i.d. $\mathcal{X}^n$-valued rv's, each with (common) pmf Q_n. The encoder F of a (length-n block) randomized code or a code with stochastic encoder is obtained by setting

$$F(m) = \boldsymbol{X}_m, \qquad m \in \mathcal{M}. \tag{76}$$

In some situations, a random selection of codewords involves choosing them with a uniform distribution from a fixed subset of $\mathcal{X}^n$. Precisely, for a given subset $B_n \subseteq \mathcal{X}^n$, the encoder F of a randomized code or code with stochastic encoder is obtained as

$$F(m) = \boldsymbol{X}'_m, \qquad m \in \mathcal{M} \tag{77}$$

where $\boldsymbol{X}'_m$, $m \in \mathcal{M}$, are i.i.d. B_n-valued rv's, each distributed uniformly on B_n. This corresponds to Q_n being the uniform pmf on B_n. For memoryless channels (known or unknown), the random codewords in (76) are usually chosen to have a simple structure, namely, to consist of i.i.d. components, i.e., for a fixed pmf Q on $\mathcal{X}$, we set

$$F(m) = \boldsymbol{X}_m = (X_{m1}, \cdots, X_{mn}), \qquad m \in \mathcal{M} \tag{78}$$

where $X_{m1}, \cdots, X_{mn}$ are i.i.d. $\mathcal{X}$-valued rv's with (common) pmf Q on $\mathcal{X}$. This corresponds to choosing Q_n to be the n-fold product pmf on $\mathcal{X}^n$ with marginal pmf Q on $\mathcal{X}$.

In order to describe the next standard method of random selection of codewords, we now define the notions of types and typical sequences (cf. e.g., [44, Sec. 1.2]). The type of a sequence $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$ is a pmf $P_{\boldsymbol{x}}$ on $\mathcal{X}$ where $P_{\boldsymbol{x}}(x)$ is the relative frequency of x in $\boldsymbol{x}$, i.e.,

$$P_{\boldsymbol{x}}(x) = \frac{1}{n} \sum_{t=1}^{n} I\{x_t = x\} \tag{79}$$

where $I\{\cdot\}$ denotes the indicator function:

$$I\{A\} = \begin{cases} 1, & \text{if statement } A \text{ is true} \\ 0, & \text{if statement } A \text{ is false.} \end{cases}$$

For a given type P of sequences in $\mathcal{X}^n$, let $\mathcal{T}_P^n$ denote the set of all sequences $\boldsymbol{x} \in \mathcal{X}^n$ with type P, i.e.,

$$\mathcal{T}_P^n = \{\boldsymbol{x} \in \mathcal{X}^n\colon\ P_{\boldsymbol{x}}(x) = P(x),\, x \in \mathcal{X}\}. \tag{80}$$

Next, for a given pmf Q on X, a sequence $\boldsymbol{x} \in \mathcal{X}^n$ is Q-typical with constant $\delta > 0$, or simply Q-typical (suppressing the explicit dependence on $\delta > 0$), if

$$\max_{x \in \mathcal{X}} |P_{\boldsymbol{x}}(x) - Q(x)| \le \delta, \qquad P_{\boldsymbol{x}}(x) = 0 \text{ if } Q(x) = 0. \tag{81}$$

Let $\mathcal{T}_{[Q]}^n$ denote the set of all sequences $\boldsymbol{x} \in \mathcal{X}^n$ which are Q-typical, i.e., the union of sets $\mathcal{T}_P^n$ for those types P of sequences in $\mathcal{X}^n$ which satisfy

$$\max_{x \in \mathcal{X}} |P(x) - Q(x)| \le \delta, \qquad P(x) = 0 \text{ if } Q(x) = 0. \tag{82}$$

Similarly, for later use, joint types are pmf's on product spaces. For example, the joint type of three given sequences $\boldsymbol{x} \in \mathcal{X}^n$, $\boldsymbol{s} \in \mathcal{S}^n$, $\boldsymbol{y} \in \mathcal{Y}^n$ is a pmf $P_{\boldsymbol{xsy}}$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$ where $P_{\boldsymbol{xsy}}(x, s, y)$ is the relative frequency of the triple (x, s, y) among the triples (x_t, s_t, y_t), $t = 1, \cdots, n$, i.e.,

$$P_{\boldsymbol{xsy}}(x, s, y) = \frac{1}{n} \sum_{t=1}^{n} I\{x_t = x,\, s_t = s,\, y_t = y\}. \tag{83}$$

A standard method of random selection of codewords now entails picking them from the set of sequences of a fixed type in accordance with a uniform pmf on that set. The resulting

random selection is a special case of (77) with the set B_n being $\mathcal{T}_P^n$. Precisely, for a fixed type P of sequences in $\mathcal{X}^n$, the encoder F of a randomized code or a code with stochastic encoder is obtained by setting

$$F(m) = \boldsymbol{X}''_m, \qquad m \in \mathcal{M} \tag{84}$$

where $\boldsymbol{X}''_m$, $m \in \mathcal{M}$, are i.i.d. $\mathcal{T}_P^n$-valued rv's, each distributed uniformly on $\mathcal{T}_P^n$. The codewords thus obtained are often referred to as "constant-composition" codewords. This method is sometimes preferable to that given by (78). For instance, in the case of a DMC (2), it is shown in [91] that for every randomized code comprising codewords selected according to (78) used in conjunction with a maximum-likelihood decoder (cf. Section IV-A2) below), there exists another randomized code with codewords as in (84) and maximum-likelihood decoder which yields a random-coding error exponent which is at least as good.

A modification of (84) is obtained when, for a fixed pmf Q on $\mathcal{X}$, the encoder F of a randomized code or a code with stochastic encoder is obtained by setting

$$F(m) = \boldsymbol{X}'''_m, \qquad m \in \mathcal{M} \tag{85}$$

where $\boldsymbol{X}'''_m$, $m \in \mathcal{M}$, are i.i.d. $\mathcal{T}_{[Q]}^n$-valued rv's, each distributed uniformly on $\mathcal{T}_{[Q]}^n$.

In the terminology of Section II, each set of randomly selected codewords $\{F(m), m \in \mathcal{M}\}$ chosen as in (76)–(85) constitutes a stochastic encoder.

Codes with randomly selected codewords as in (76)–(85), together with suitable decoders, can be used in random-coding argument techniques for establishing reliable communication over known channels. For instance, codewords for the DMC (2) can be selected according to (78) [111] or (85) [124], and for the finite-state channel (7) according to (76) [64]. In these cases, the existence of a code with deterministic encoder f, i.e., deterministic codewords $f(m)$, $m \in \mathcal{M}$, for establishing reliable communication, is obtained in terms of a realization of the random codewords $F(m)$, $m \in \mathcal{M}$, combined with a simple expurgation, to ensure a small maximum probability of error.

For certain types of unknown channels too, codewords chosen as in (76)–(85), without any additional refinement, suffice for achieving reliable communication. For instance, in the case of the AVC (5), random codewords chosen according to (5) were used [19], [119] to determine the randomized code capacity without input or state constraints in Theorem 2, and with such constraints (cf. (48)) [47].

2) Randomized Codes and Random Code Reduction: Randomly selected codewords $\{F(m), m \in \mathcal{M}\}$ as in (76)–(85), together with a decoder ϕ given by (11), obviously constitute a code with stochastic encoder (F, ϕ). They also enable the following elementary and standard construction of a (length-n block) randomized code (F, Φ). Associate with every realization $\{f(m), m \in \mathcal{M}\}$ of the randomly selected codewords $\{F(m), m \in \mathcal{M}\}$, a decoder $\phi_f \colon \mathcal{Y}^n \to \mathcal{M} \cup \{0\}$ which depends, in general, on said realization. This results in a randomized code (F, Φ), where the encoder F is as above, and the decoder Φ is defined by

$$\Phi(\boldsymbol{y}) = \phi_F(\boldsymbol{y}), \qquad \boldsymbol{y} \in \mathcal{Y}^n. \tag{86}$$

Such a randomized code (F, Φ), in addition to serving as an artifice in random-coding arguments for proving coding theorems as mentioned earlier, can lead to larger capacity values for the AVC (5) than those achieved by codes with stochastic encoders or deterministic codes (cf. Section III-A2) above). In fact, the randomized code capacity C of the AVC (5) given by Theorem 2 is achieved [19] using a randomized code (F, Φ) as above, where the encoder F is chosen as in (78) with pmf Q^* on $\mathcal{X}$ and the decoder Φ is given by (86) with ϕ_f being the (normalized) maximum-likelihood decoder (corresponding to the codewords $\{f(m), m \in \mathcal{M}\}$) for the DMC with stochastic matrix $W_{\zeta^*} \colon \mathcal{X} \to \mathcal{Y}$, where (Q^*, ζ^*) is a saddle point for (46). When input or state constraints are additionally imposed, the randomized code capacity $C(\Gamma, \Lambda)$ of the AVC (5) given by (48) is achieved by a similar code with suitable modifications to accommodate the constraints [47].

Consequently, randomized codes become significant as models of practical engineering devices; in fact, commonly used spread-spectrum techniques such as direct sequence and frequency hopping can be interpreted as practical implementations of randomized codes [58], employing synchronized random number generators at the transmitter and receiver. From a practical standpoint, however, a (length-n block) randomized code (F, Φ) of rate R bits per channel use, such as that just described above in the context of the randomized code capacity of the AVC (5), involves making a random selection from among a prohibitively large collection—of size $|\mathcal{X}|^{nM} = |\mathcal{X}|^{n\lceil \exp(nR)\rceil}$—of sets of codewords $\{f(m), m \in \mathcal{M}\}$, where $|\cdot|$ denotes cardinality. In addition, the outcome of this random selection must be observed by the receiver; else, it must be conveyed to the receiver requiring an infeasibly large overhead transmission of $\approx n\lceil \exp(nR)\rceil \log|\mathcal{X}|$ bits in order to ensure the reliable communication of $\approx nR$ information bits.

The practical feasibility of randomized codes, in particular for the AVC (5), is supported by Ahlswede's result on "random code reduction" [6], which establishes the existence of "good" randomized codes obtained by random selection from "exponentially few" (in blocklength n) deterministic codes. This result is stated below in a version which appears in [44, Sec. 2.6], and requires the following setup. For a fixed blocklength n, consider a family of channels indexed by $\theta \in \Theta$ as in (3), where Θ is now assumed to be a finite set. Let (F, Φ) be a given randomized code which results in a maximum probability of error $e_{\max}(F, \Phi, \theta)$ (cf. (14) and (16)) when used on the channel $\theta \in \Theta$.

Theorem 10: For any ϵ and K satisfying

$$\epsilon > 2\log\left(1 + \max_{\theta \in \Theta} e_{\max}(F, \Phi, \theta)\right), \quad K > \frac{2}{\epsilon}\left(\log M + \log|\Theta|\right) \tag{87}$$

there exists a randomized code (F', Φ') which is uniformly distributed on a family of K deterministic codes

$\{(f_k, \phi_k), k = 1, \cdots, K\}$ as in (10) and (11), and such that

$$e_{\max}(F', \Phi', \theta) < \epsilon, \qquad \theta \in \Theta. \tag{88}$$

The assertion in (88) concerning the performance of the randomized code (F', Φ') is equivalent to

$$\frac{1}{K} \sum_{k=1}^{K} e_m(f_k, \phi_k, \theta) < \epsilon, \qquad m \in \mathcal{M}, \theta \in \Theta. \tag{89}$$

Thus for every randomized code (F, Φ), there exists a "reduced" randomized code (F', Φ') which is uniformly distributed over K deterministic codes and has maximum probability of error on any channel not exceeding ϵ, provided the hypothesis (87) holds.

Theorem 10 above has two significant implications for AVC performance. First, for any randomized code (F, Φ) which achieves the randomized code capacity of the AVC (5) given by Theorem 2, there exists another randomized code (F', Φ') which does likewise; furthermore, (F', Φ') is obtained by random selection from no more than $K = n^2$ deterministic codes [6]. Hence, the outcome of the random selection of codewords at the transmitter can now be conveyed to the receiver using at most only $\approx 2 \log n$ bits, which represents a desirably drastic reduction in overhead transmission; the rate of this transmission, termed the "key rate" in [59], is arbitrarily small. Second, such a "reduced" randomized code (F', Φ') is amenable to conversion, by an "elimination of randomness" [6], into a deterministic code $((\hat{f}_n, f'), (\hat{\phi}_n, \phi'))$ (cf. e.g., [44, Sec. 2.6]) for the AVC (5), provided its deterministic code capacity C^a for the average probability of error is positive. Here, (f', ϕ') is as in (10) and (11), while $(\hat{f}_n, \hat{\phi}_n)$ represents a code for conveying to the receiver the outcome of the random selection at the transmitter, i.e.,

$$\hat{f}_n \colon \{1, \cdots, n^2\} \to \mathcal{X}^{k_n} \qquad \hat{\phi}_n \colon \mathcal{Y}^{k_n} \to \{1, \cdots, n^2\} \tag{90}$$

where k_n/n tends to 0 with increasing n. As a consequence, C^a equals the randomized code capacity C of the AVC (5) given by Theorem 2. This has been discussed earlier in Section III-A2).

3) Refined Codeword Sets by Random Selection: As stated earlier, the method of random selection can sometimes be used to prove the existence of codewords with special properties which are useful for determining the deterministic code capacities of certain unknown channels.

For instance, the deterministic code capacity of the AVC (5) for the maximum or average probability of error is sometimes established by a technique relying on the method of random selection as in (78), (84), and (85), used in such a manner as to assert the existence of codewords with select properties. A deterministic code comprising such codewords together with a suitably chosen decoder then leads to acceptable bounds for the probabilities of decoding errors. This artifice is generally not needed when using randomized codes or codes with stochastic encoders. Variants of this technique have been applied, for instance, in obtaining the deterministic code capacity of the AVC (5) for the maximum probability of error in [10] and in Theorem 4 [45], as well as for the average probability of error in Theorems 5 and 6 [48].

In determining the deterministic code capacity C^m for the maximum probability of error [10], random selection as in (78), together with an expurgation argument using Bernstein's version of Markov's inequality for i.i.d. rv's, is used to show in effect the existence of a codeword set with "spread-out" codewords, namely, every two codewords are separated by at least a certain Hamming distance. A codeword set with similar properties is also shown to result from alternative random selection as in (85). Such a codeword set, in conjunction with a decoder which decides on the basis of a threshold test using (normalized) likelihood ratios, leads to a bound for the maximum probability of error. A more general characterization of C^m in [45] relies on a code with codewords from the set $\mathcal{T}_P^n$ of sequences in $\mathcal{X}^n$ of type P (cf. (80)) which satisfy desirable "balance" properties with probability arbitrarily close to 1, together with a suitable decoding rule (cf. Section IV-B6)). The method of random selection in (84) combined with a large-deviation argument for i.i.d. rv's as in [10], is used in proving the existence of such codewords. Loosely speaking, the codewords are "balanced" in that for a transmitted codeword $\boldsymbol{x}$ and the (unknown) state sequence $\boldsymbol{s} \in \mathcal{S}^n$ which prevails during its transmission, the proportion of other codewords $\boldsymbol{x}'$ which have a specified joint type (cf. (83)) with $\boldsymbol{x}$ and $\boldsymbol{s}$ does not greatly exceed their overall "density" in $\mathcal{T}_P^n$. This limits, in effect, the number of spurious codewords which are jointly typical with $\boldsymbol{x}$, $\boldsymbol{s}$ and a received sequence $\boldsymbol{y} \in \mathcal{Y}^n$, leading to a satisfactory bound for the maximum probability of error.

The determination in [48] of the deterministic code capacity of the AVC (5) for the average probability of error, without or with input or state constraints (cf. Theorems 5 and 6) relies on codewords resulting from random selection as in (84) and a decoder described below in Section IV-B6). These codewords possess special properties in the spirit of [45], which are established using Chernoff bounding for dependent rv's as in [53].

B. Decoders

A variety of decoders have been proposed in order to achieve reliable communication in the different communication situations described in Section II. Some of these decoders will be surveyed below. We begin with decoders for known channels and describe the maximum-likelihood decoder and the various typicality decoders. We then consider the generalized likelihood-ratio test for unknown channels, the maximum-empirical mutual information (MMI) decoder, and more general universal decoders. The section ends with a discussion of decoders for the compound channel, mismatched decoders, and decoders for the arbitrarily varying channel.

1) Decoders for Known Channels: The most natural decoder for a known channel (1) is the maximum-likelihood decoder, which is optimal in the sense of minimizing the average probability of error (15). Given a set of codewords

$\{f(m), m \in \mathcal{M}\}$ in $\mathcal{X}^n$, the maximum-likelihood decoder ϕ_{ML} is defined by: $\phi_{\mathrm{ML}}(\boldsymbol{y}) = m$ only if

$$W(\boldsymbol{y}|f(m)) = \max_{m' \in \mathcal{M}} W(\boldsymbol{y}|f(m')). \tag{91}$$

If more than one $m \in \mathcal{M}$ satisfies (91), ties are resolved arbitrarily. While the maximum-likelihood rule is indeed a natural choice for decoding over a known channel, its analysis can be quite intricate [64], and was only conducted years after Shannon's original paper [111].

Several simpler decoders have been proposed for the DMC (2), under the name of "typicality" decoders. These decoders are usually classified as "weak typicality" decoders [39] (which are sometimes referred to as "entropy typicality" decoders [44]), and "joint typicality" decoders [24], [44], [126] (which are sometimes referred to as "strong" typicality decoders). We describe below the joint-type typicality decoder as well as a more stringent version which relies on a notion of typicality in terms of the Kullback–Leibler divergence (cf. e.g., [44]).

Given a set of codewords $\{f(m), m \in \mathcal{M}\}$ in $\mathcal{T}_P^n$, where P is a fixed type of sequences in $\mathcal{X}^n$, the joint typicality decoder ϕ_T for the DMC (2) is defined as follows: $\phi_T(\boldsymbol{y}) = m$ only if

$$\max_{x \in \mathcal{X}, y \in \mathcal{Y}} |P_{f(m)\boldsymbol{y}}(x, y) - (P \circ W)(x, y)| < \eta \tag{92}$$

where $W: \mathcal{X} \to \mathcal{Y}$ is the stochastic matrix in the definition of the DMC (2), $(P \circ W)(x, y) = P(x)W(y|x)$, and $\eta > 0$ is chosen sufficiently small. If more than one $m \in \mathcal{M}$ satisfies (92), or no $m \in \mathcal{M}$ satisfies (92), set $\phi(\boldsymbol{y}) = 0$. The capacity of a DMC (2) can be achieved by a joint typicality decoder ([111]; see also [44, Problem 7, p. 113]), but this decoder is suboptimal and does not generally achieve the channel reliability function $E(R)$.

Another version of a joint typicality decoder, which we term the divergence typicality decoder, has appeared in the literature (cf. e.g., [45] and [48]). It relies on a more stringent notion of typicality based on the Kullback–Leibler divergence (cf. e.g., [39] and [44]). Precisely, given a set of codewords $\{f(m), m \in \mathcal{M}\}$ in $\mathcal{T}_P^n$ as above, a divergence typicality decoder ϕ_{DT} for the DMC (2) is defined as follows: $\phi_{DT}(\boldsymbol{y}) = m$ only if

$$D(P_{f(m)\boldsymbol{y}} || P \circ W) < \eta \tag{93}$$

where $D(\cdot||\cdot)$ denotes Kullback–Leibler divergence and $\eta > 0$ is chosen sufficiently small. If more than one $m \in \mathcal{M}$, or no $m \in \mathcal{M}$, satisfies (93), we set $\phi_{DT}(\boldsymbol{y}) = 0$. The capacity of a DMC (2) can be achieved by the divergence typicality decoder.

2) The Generalized Likelihood Ratio Test: The maximum-likelihood decoding rules for channels governed by different laws are generally different mappings, and maximum-likelihood decoding with respect to the prevailing channel cannot therefore be applied if the channel law is unknown. The same is true of joint typicality decoding. A natural candidate for a decoder for a family of channels (3) is the generalized likelihood ratio test decoder.

The generalized likelihood ratio test (GLRT) decoder ϕ_{GLRT} can be defined as follows: given a set of codewords $\{f(m), m \in \mathcal{M}\}$, $\phi_{\mathrm{GLRT}}(\boldsymbol{y}) = m$ only if

$$\sup_{\theta \in \Theta} W(\boldsymbol{y}|f(m); \theta) = \max_{m' \in \mathcal{M}} \sup_{\theta \in \Theta} W(\boldsymbol{y}|f(m'); \theta)$$

where ties can be resolved arbitrarily among all $m \in \mathcal{M}$ which achieve the maximum.

If the family of channels corresponds to the family of *all* DMC's with finite input alphabet $\mathcal{X}$ and finite output alphabet $\mathcal{Y}$, then

$$\begin{aligned}
&\sup_{\theta \in \Theta} \frac{1}{n} \log W(\boldsymbol{y}|\boldsymbol{x}; \theta) \\
&\quad = \sup_{\theta \in \Theta} \sum_{x \in \mathcal{X}} P_{\boldsymbol{x}}(x) \sum_{y \in \mathcal{Y}} P_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \log W(y|x; \theta) \\
&\quad = \sum_{x \in \mathcal{X}} P_{\boldsymbol{x}}(x) \sum_{y \in \mathcal{Y}} P_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \log P_{\boldsymbol{y}|\boldsymbol{x}}(y|x) \\
&\quad = -H(\boldsymbol{y}|\boldsymbol{x}) \\
&\quad = I(\boldsymbol{x} \wedge \boldsymbol{y}) - H(\boldsymbol{y})
\end{aligned}$$

where the first equality follows by defining the condition empirical distribution $P_{\boldsymbol{y}|\boldsymbol{x}}(y|x)$ to satisfy

$$P_{\boldsymbol{x}\boldsymbol{y}}(x, y) = P_{\boldsymbol{x}}(x) P_{\boldsymbol{y}|\boldsymbol{x}}(y|x);$$

the second equality from the nonnegativity of relative entropy; the third equality by defining $H(\boldsymbol{y}|\boldsymbol{x})$ as the conditional entropy $H(Y|X)$, where X, Y are dummy rv's whose joint pmf on $\mathcal{X} \times \mathcal{Y}$ is the joint type $P_{\boldsymbol{x}\boldsymbol{y}}$; and the last equality by defining $I(\boldsymbol{x} \wedge \boldsymbol{y})$ as the mutual information $I(X \wedge Y)$, with X, Y as above.

Since the term $H(\boldsymbol{y})$ depends only on the output sequence $\boldsymbol{y}$, it is seen that for the family of all DMC's with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, the GLRT decoding rule is equivalent to the maximum empirical mutual information (MMI) decoder [44], which is defined by

$$\phi_{\mathrm{MMI}}(\boldsymbol{y}) = \arg \max_{m \in \mathcal{M}} I(f(m) \wedge \boldsymbol{y}). \tag{94}$$

Note that if the family under consideration is a subset of the class of all DMC's, then the GLRT will not necessarily coincide with the MMI decoder.

The MMI decoder is a universal decoder for the family of memoryless channels, in a sense that will be made precise in the next section.

3) Universal Decoding: Loosely speaking, a sequence of codes is universal for a family of channels if it achieves the same random-coding error exponent as the maximum-likelihood decoder without requiring knowledge of the specific channel in the family over which transmission takes place [44], [60], [92], [95], [98], [103], [129]. We now make this notion precise. Let $\{B_n\}_{n=1}^{\infty}$ denote a sequence of sets, with $B_n \subseteq \mathcal{X}^n$. Consider a randomized encoder $F_n: \mathcal{M} \to B_n$ whose codewords are drawn independently and uniformly from B_n as in (77). Let $\Phi_{F_n, \theta}$ denote a maximum-likelihood receiver for the encoder F_n and the channel $\theta \in \Theta$ as in (86) and (91). As in Section II we set $\overline{e}(F_n, \Phi_{F_n, \theta}, \theta)$ to be the average probability of error corresponding to the code $(F_n, \Phi_{F_n, \theta})$ for the channel θ. Note that the average is both with respect to the messages (as in (15)) and the pmf of the randomized code (as in (16)).

A sequence of codes $\{(f_n, u_n)\}_{n=1}^{\infty}$, of rate R, where $f_n: \mathcal{M} \to B_n$ and $u_n: \mathcal{Y}^n \to \mathcal{M} \cup \{0\}$ is said to be universal[4] for the input sets $\{B_n\}$ and the family (3) if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\overline{e}(f_n, u_n, \theta)}{\overline{e}(F_n, \Phi_{F_n,\theta}, \theta)} = 0. \tag{95}$$

Notice that neither encoder nor decoder is allowed to depend on the channel θ.

For families of DMC's the following result was proved by Csiszár and Körner [44].

Theorem 11: Assume that the input sets B_n correspond to type classes, i.e., $B_n = \mathcal{T}_P^n$ for some fixed type P of sequences in $\mathcal{X}^n$. Under this assumption, there exists a sequence of codes $\{(f_n, \phi_{\mathrm{MMI},n})\}_{n=1}^{\infty}$ with MMI decoder which is universal for any family of discrete memoryless channels.

As we have noted above, if the family of channels (3) is a *subset* of the set of all DMC's, then the GLRT for the family may be different from the MMI decoder. In fact, in this case the GLRT may not be universal for the family [90]. It is thus seen that the GLRT may not be universal for a family even when a universal decoder for the family exists [92]. The GLRT is therefore not "canonical."

Universal codes for families of finite-state channels (8) were proposed in [129] with subsequent refinements in [60] and [92]. The decoding rule proposed in [92] and [129] is based on the joint Lempel–Ziv parsing [130] of the received sequence $\boldsymbol{y}$ with each of the possible codewords $f(m)$, $m = 1, \cdots, M$.

A different approach to universal decoding can be found in [60], where a universal decoder based on the idea of "merging" maximum-likelihood decoders is proposed. This idea leads to existence results for fairly general families of channels including some with infinite alphabets (e.g., a family of Gaussian intersymbol interference channels). To state these results, we need the notion of a "strongly separable" family. Loosely speaking, a family is strongly separable if for any blocklength n there exists a subexponential number $K(n)$ of channels such that the law of any channel in the family can be approximated by one of these latter channels. The approximation is in the sense that except for rare sequences, the normalized log likelihood of an output sequence given any input sequence is similar under the two channels. Precisely:

A family of channels (3) with common finite input and output alphabets $\mathcal{X}, \mathcal{Y}$ is said to be *strongly separable* for the input sets $B_n \subseteq \mathcal{X}^n$ if there exists some (finite) $L > 0$ which serves as an upper bound for all the error exponents in the family, i.e.,

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} -\frac{1}{n} \log \overline{e}(F, \Phi_{\mathrm{ML}}, \theta) < L \tag{96}$$

such that for any $\epsilon > 0$ and blocklength n, there exists a subexponential number $K(n)$ (depending on L and ϵ) of channels $\{\theta_k^{(n)}\}_{k=1}^{K(n)} \subseteq \Theta$

$$\lim_{n\to\infty} \frac{1}{n} \log K(n) = 0 \tag{97}$$

which can approximate any $\theta \in \Theta$ in the following sense: for any $\theta \in \Theta$, there exists a channel $\theta_{k^*}^{(n)} \in \Theta$, $1 \le k^* \le K(n)$, satisfying

$$W(\boldsymbol{y}|\boldsymbol{x}; \theta) \le 2^{n\epsilon} W(\boldsymbol{y}|\boldsymbol{x}; \theta_{k^*}^{(n)}) \tag{98}$$

whenever $(\boldsymbol{x}, \boldsymbol{y}) \in B_n \times \mathcal{Y}^n$ is such that

$$W(\boldsymbol{y}|\boldsymbol{x}; \theta) > 2^{-n(L+\log|\mathcal{Y}|)}$$

and satisfying

$$W(\boldsymbol{y}|\boldsymbol{x}; \theta_{k^*}^{(n)}) \le 2^{n\epsilon} W(\boldsymbol{y}|\boldsymbol{x}; \theta) \tag{99}$$

whenever $(\boldsymbol{x}, \boldsymbol{y}) \in B_n \times \mathcal{Y}^n$ is such that

$$W(\boldsymbol{y}|\boldsymbol{x}; \theta_{k^*}^{(n)}) > 2^{-n(L+\log|\mathcal{Y}|)}.$$

For example, the family of all DMC's with finite input and output alphabets $\mathcal{X}, \mathcal{Y}$, is strongly separable for any sequence of input sets $\{B_n\}$. Likewise, the family of all finite-state channels with finite input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \Sigma$ is also strongly separable for any sequence of input sets $\{B_n\}$ [60]. For a definition of strong separability for channels with infinite alphabets see [60].

Theorem 12: If a family of channels (3) with common finite input and output alphabets $\mathcal{X}, \mathcal{Y}$ is strongly separable for the input sets $\{B_n\}$, then there exists a sequence of codes $\{(f_n, u_n)\}$ which is universal for the family.

Not surprisingly, in a nonparametric situation where nothing is known *a priori* about the channel statistics, universal decoding is not possible [99].

A slightly different notion of universality, referred to in [60] as "strong random-coding universality," requires that (95) hold for the "average encoder." More precisely, consider a decoding rule u which, given an encoder f, maps each possible received sequence $\boldsymbol{y}$ to some message $m \in \mathcal{M} \cup \{0\}$. We can then consider the random code (F_n, u_{F_n}) where, as before, F is a random encoder whose codewords are drawn independently and uniformly from the set B_n. The decoding rule u is strongly random coding universal for the input sets $\{B_n\}$ if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\overline{e}(F_n, u_{F_n}, \theta)}{\overline{e}(F_n, \Phi_{F_n,\theta}, \theta)} = 0. \tag{100}$$

It is shown in [60] that the hypothesis of Theorem 12 also implies strong random-coding universality.

We next demonstrate the role played by universal decoders in communicating over a compound channel, and also discuss some alternative decoders for this situation.

4) Decoders for the Compound Channel: Consider the problem of communicating reliably over a compound channel (3). Let $\{B_n\}$ be a sequence of input sets and let F_n be a randomized rate-R encoder which chooses the codewords independently and uniformly from the set B_n as in (77). Let $\Phi_{F_n,\theta}$ denote the maximum-likelihood decoder corresponding to the encoder F_n for the channel $\theta \in \Theta$. Suppose now that the code rate R is sufficiently low so that $\overline{e}(F, \Phi_{F_n,\theta}, \theta)$

[4]This form of universality is referred to as "strong deterministic coding universality" in [60]. See [60] for a discussion of other definitions for universality.

is uniformly bounded in θ by a function which decreases exponentially to zero with the blocklength n, i.e.,

$$\liminf_{n\to\infty} -\frac{1}{n}\log \sup_{\theta\in\Theta} \overline{e}(F_n, \Phi_{F_n,\theta}, \theta) > 0. \tag{101}$$

It then follows from (95) that if $\{(f_n, u_n)\}$ is a sequence of universal codes for the family Θ and input sets $\{B_n\}$, then R is an achievable rate and can be achieved with the decoders $\{u_n\}$.

It is, thus, seen that if a family of channels admits universal decoding, then the problem of demonstrating that a rate R is achievable only requires the study of random-coding error probabilities with maximum-likelihood decoding (101).

Indeed, the capacity of the compound DMC can be attained using an MMI decoder (Theorem 11) [44], and the capacity of a compound FSC can be attained using a universal decoder for that family [91].

The original decoder proposed for the compound DMC [30] is not universal; it is based on maximum-likelihood decoding with respect to a Bayesian mixture of a finite number of "representative" channels (polynomial in the blocklength) in the family [30], [64, pp. 176–178]. Nevertheless, if the "representatives" are chosen carefully, the resulting decoder is, indeed, universal.

A completely different approach to the design of a decoder for a family of DMC's can be adopted if the family (3) and (4) is compact and convex in the sense that for every θ', $\theta'' \in \Theta$ with corresponding stochastic matrices $W(\cdot|\cdot;\theta')$ and $W(\cdot|\cdot;\theta'')$, and for every $\alpha \in (0, 1)$, there exists $\theta^{(\alpha)} \in \Theta$ with corresponding stochastic matrix given by

$$W(y|x;\theta^{(\alpha)}) = \alpha W(y|x;\theta') + (1-\alpha)W(y|x;\theta''), \qquad x \in \mathcal{X},\ y \in \mathcal{Y}.$$

Under these assumptions of compactness and convexity, the capacity of the family is given by

$$\max_{Q\in\mathcal{P}(\mathcal{X})} \min_{\theta\in\Theta} I(Q, W(\cdot|\cdot;\theta)) = \min_{\theta\in\Theta} \max_{Q\in\mathcal{P}(\mathcal{X})} I(Q, W(\cdot|\cdot;\theta)). \tag{102}$$

Let (Q^*, θ^*) achieve the saddle point in (102). Then the capacity of this family of DMC's can be achieved by using a maximum-likelihood decoder for the DMC with stochastic matrix $W(\cdot|\cdot;\theta^*)$ [44], [51], [119].

The maximum-likelihood decoder with respect to $W(\cdot|\cdot;\theta^*)$ is generally much simpler to implement than a universal (e.g., MMI) decoder, particularly if the codes being used have a strong algebraic structure. A universal decoder, however, has some advantages. In particular, its performance on a channel $W(\cdot|\cdot;\tilde{\theta})$, for $\tilde{\theta} \neq \theta^*$, is generally better than the performance on the channel $W(\cdot|\cdot;\tilde{\theta})$ of the maximum-likelihood decoder for $W(\cdot|\cdot;\theta^*)$.

For example, on an average power-limited additive-noise channel with a prespecified noise variance, a Gaussian codebook and a Gaussian noise distribution form a saddle point for the mutual information functional. The maximum-likelihood decoder for the saddle-point channel is a minimum Eulidean distance decoder, which is suboptimal if the noise is not Gaussian. Indeed, if the noise is discrete rather than being Gaussian (which is worse), then a Gaussian codebook with universal decoding can achieve a positive random-coding error exponent at all positive rates; with minimum Euclidean distance decoding, however, the random-coding error exponent is positive only for rates below the saddle-point value of the mutual information [88]. In this sense, a Gaussian codebook and a minimum Euclidean distance decoder cause every noise distribution to appear as harmful as the worst (Gaussian) noise.

A situation in which transmission occurs over a channel $W(\cdot|\cdot;\theta)$, and yet decoding is performed as though the channel were $W(\cdot|\cdot;\theta')$, is sometimes referred to as "mismatched decoding." Generally, a decoder is mismatched with respect to the channel $W(\cdot|\cdot;\theta)$ if it chooses the codeword that minimizes a "metric" defined for sequences as the additive extension of a single-letter "metric" $d(\cdot,\cdot)$, where $d(\cdot,\cdot)$ is, in general, not equal to $-\log W(\cdot|\cdot;\theta)$ (see (103) below).

Mismatched decoding can arise when the receiver has a poor estimate of the channel law, or when complexity considerations restrict the metric of interest to take only a limited number of integer values. The "mismatch problem" entails determining the highest achievable rates with such a hindered decoder, and is discussed in the following subsection.

5) Mismatched Decoding: Consider a known DMC (2). Given a set of codewords $\{f(m), m \in \mathcal{M}\}$, define a decoder ϕ_d by: $\phi_d(\boldsymbol{y}) = m$ if

$$d(f(m), \boldsymbol{y}) < d(f(m'), \boldsymbol{y}), \qquad \text{for all } m' \neq m. \tag{103}$$

If no such $m \in \mathcal{M}$ exists (owing to a tie), set $\phi_d(\boldsymbol{y}) = 0$. Here

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{t=1}^{n} d(x_t, y_t)$$

and d: $\mathcal{X} \times \mathcal{Y} \to [0, \infty]$ is a given function which is often referred to as "decoding metric" (even though it may not be a metric in the topological sense). The decoder ϕ_d thus produces that message which is "nearest" to the received sequence $\boldsymbol{y}$ according to the additive "metric" $d(\boldsymbol{x}, \boldsymbol{y})$, resolving ties by declaring an error.

Setting

$$d(x, y) = \log \tilde{W}(y|x)$$

where $\tilde{W}(\cdot|\cdot)$ is a stochastic matrix $\mathcal{X} \to \mathcal{Y}$, corresponds to the study of a situation where the true channel law is $W(\cdot|\cdot)$ but the decoder being used is a maximum-likelihood decoder tuned to the channel $\tilde{W}(\cdot|\cdot)$. This situation may arise as discussed previously when $\tilde{W}(\cdot|\cdot)$ achieves the saddle point in (102) or when maximum-likelihood decoding with respect to $\tilde{W}(\cdot|\cdot)$ is simpler to implement than maximum-likelihood decoding with respect to the true channel $W(\cdot|\cdot)$. Complexity, for example, could be reduced by using integer metrics with a relatively small range [108].

The "mismatch problem" consists of finding the set of achievable rates for this situation, i.e., the supremum $C_d(W)$ of all rates that can be achieved over the DMC $W(\cdot|\cdot)$ with the decoder ϕ_d. This problem was studied extensively in [21], [22], [43], [51], [84], [87], and [100]. A lower bound on $C_d(W)$, which can be derived using a random-coding argument, is given by the following.

Theorem 13: Consider a DMC $W(\cdot|\cdot)$ with finite input and output alphabets $\mathcal{X}$, $\mathcal{Y}$. Then the rate

$$\max_{Q\in\mathcal{P}(\mathcal{X})} \min_{PY} I(X \wedge Y)$$

is achievable with the decoder ϕ_d defined in (103). Here $I(X \wedge Y)$ denotes the mutual information between X and Y with joint pmf P_{XY} on $\mathcal{X} \times \mathcal{Y}$, and the minimization is with respect to joint pmf's P_{XY} that satisfy

$$\sum_{y\in\mathcal{Y}} P_{XY}(x, y) = Q(x)$$

$$\sum_{x\in\mathcal{X}} P_{XY}(x, y) = \sum_{x\in\mathcal{X}} Q(x)W(y|x)$$

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x, y)d(x, y) \le \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} Q(x)W(y|x)\, d(x, y).$$

It should be noted that this bound is in general not tight [51]. This is not due to a loose analysis of the random-coding performance but rather because the best code for this situation may be much better than the "average" code [100].

Improved bounds on the mismatch capacity $C_d(W)$ can be found in [51] and [87]. It appears that the problem of precisely determining the capacity of this channel is very difficult; a solution to this problem would also yield a solution to the problem of determining the zero-error capacity of a graph as a special case [51]. Nevertheless, if the input alphabet is binary, Balakirsky has shown that the lower bound of Theorem 13 is tight [22]. Several interesting open problems related to mismatched decoding are posed in [51].

Extensions of the mismatch problem to the multiple-access channel are discussed in [87], and dual problems in rate distortion theory are discussed in [89].

6) Decoders for the Arbitrarily Varying Channel: Maximum-likelihood decoders can be used to achieve the randomized code capacity of an AVC (5), without or with input or state constraints (cf. Section IV-A2), passage following (86)). On the other hand, fairly complex decoders are generally needed to achieve its deterministic code capacity for the maximum or average probability of error. In fact, the first nonstandard decoder in Shannon theory appears, to our knowledge, in [10] in the study of AVC performance for deterministic codes and the maximum probability of error.

A significantly different decoder from that proposed in [10] is used in [45] to provide the characterization in Theorem 4 of the deterministic code capacity C^m of an AVC (5) for the maximum probability of error. The decoder in [45] operates in two steps. In the first step, a decision is made on the basis of a joint typicality condition which is a modified version of that used to define the divergence typicality decoder ϕ_{DT} in Section IV-B1). Any tie is broken in a second step by a threshold test which uses empirical mutual information quantities. Precisely, given a set of codewords $\{f(m), m \in \mathcal{M}\}$ in $\mathcal{T}_P^n$, for some fixed type P of sequences in $\mathcal{X}^n$ (cf. (80)), the decoder ϕ in [45] is defined as follows: $\phi(\boldsymbol{y}) = m$ iff

$$D(P_{f(m)\boldsymbol{s}\boldsymbol{y}}||P_{f(m)\boldsymbol{s}} \circ W) < \eta, \qquad \text{for some } \boldsymbol{s} \in \mathcal{S}^n \quad (104)$$

and for every $m' \neq m$ which satisfies (104) for some $\boldsymbol{s}' \in \mathcal{S}^n$, it holds that

$$I(f(m') \wedge \boldsymbol{y}|f(m), \boldsymbol{s}) < \eta \quad (105)$$

where W: $\mathcal{X}\times\mathcal{S} \to \mathcal{Y}$ is the stochastic matrix in the definition of the AVC (5), and $\eta > 0$ is chosen sufficiently small. Here, $I(f(m') \wedge \boldsymbol{y}|f(m), \boldsymbol{s})$ is the conditional mutual information $I(X' \wedge Y|X, S)$, where X, X', S, Y are dummy rv's whose joint pmf on $\mathcal{X} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$ is the joint type $P_{f(m)f(m')\boldsymbol{s}\boldsymbol{y}}$. In decoding for a DMC (2), a divergence typicality decoder of a simpler form than in (104) (viz. with the exclusion of the state sequence $\boldsymbol{s} \in \mathcal{S}^n$), defined by (93), suffices for achieving capacity. For an AVC (5), the additional tie-breaking step in (105) is interpreted as follows: the transmitted codeword $f(m)$, the state sequence $\boldsymbol{s} \in \mathcal{S}^n$ prevailing during its transmission, and the received sequence $\boldsymbol{y} \in \mathcal{Y}^n$, will satisfy (104) with high likelihood. If $f(m')$ is a spurious codeword which, for some $\boldsymbol{s}' \in \mathcal{S}^n$, also appears to be jointly typical with $\boldsymbol{y}$ in the sense of (104), then $f(m')$ can be expected to be only vanishingly dependent on $\boldsymbol{y}$ given $f(m)$ and $\boldsymbol{s}$, in the sense of (105). As stated in [40], the form of this decoder is, in fact, suggested by the procedure for bounding the maximum probability of error using the "method of types." An important element of the proof of Theorem 4 in [45] consists in showing that for a suitably chosen set of codewords $\{f(m), m \in \mathcal{M}\}$, the decoder in (104) and (105) for a sufficiently small $\eta > 0$ is unambiguous, i.e., it maps each received sequence into at most one message.

At this point, it is worth recalling that the joint typicality and divergence typicality decoders for known channels, described in Section IV-B1), are defined in terms of the joint types of $f(m)$ and $\boldsymbol{y}$, i.e., pairs of codewords and received sequences. Such decoders belong to the general class of α-decoders, studied in [43], which can be defined solely in terms of the joint types of pairs each consisting of a codeword and a received sequence. In contrast, for the deterministic code capacity problem for the AVC (5) under the maximum probability of error, the decoder in [45] defined by (104) and (105) involves the joint types of triples $(f(m), f(m'), \boldsymbol{y})$. This decoder, thus, belongs to a more general class of decoders, introduced in [42] under the name of β-decoders, which are based on pairwise comparisons of codewords relying on joint types of triples $(f(m), f(m'), \boldsymbol{y})$.

We turn next to decoders used for achieving the deterministic code capacity of the AVC (5) for the average probability of error, without or with input or state constraints. A comprehensive treatment is found in [49]. The decoder used in [48] to determine the AVC deterministic code capacity C^a for the average probability of error in Theorem 5 resembles that in (104) and (105), but has added complexity. It too does not belong to the class of α-decoders, but rather to the class of β-decoders. Precisely, given a set of codewords $\{f(m), m \in \mathcal{M}\}$ in $\mathcal{T}_P^n$ as above, the decoder ϕ in [48] is defined as follows: $\phi(\boldsymbol{y}) = m$ iff

$$D(P_{f(m)\boldsymbol{s}\boldsymbol{y}}||P \circ P_{\boldsymbol{s}} \circ W) < \eta, \qquad \text{for some } \boldsymbol{s} \in \mathcal{S}^n \quad (106)$$

and for every $m' \neq m$ which satisfies (106) for some $\boldsymbol{s}' \in \mathcal{S}^n$, it holds that

$$I(f(m') \wedge f(m), \boldsymbol{y}|\boldsymbol{s}) \leq \eta \tag{107}$$

where $\eta > 0$ is chosen sufficiently small. Here, $I(f(m') \wedge f(m), \boldsymbol{y}|\boldsymbol{s})$ is the conditional mutual information $I(X' \wedge X, Y|S)$, where X, X', S, Y are dummy rv's as arising above in (105). A main step of the proof of Theorem 5 in [48] is to show that this decoder is unambiguous if $\eta > 0$ is chosen sufficiently small. An obvious modification of the conditions in (106) and (107) by allowing only such state sequences $\boldsymbol{s} \in \mathcal{S}^n$ as satisfy state constraint Λ (cf. (24)), leads to a decoder used in [48] for determining the deterministic code capacity $C^a(\Gamma, \Lambda)$ of the AVC (5) under input constraint Γ and state constraint Λ (cf. Theorem 6).

It should be noted that the divergence typicality condition in (106) is alone inadequate for the purpose of establishing the AVC capacity result in Theorem 5. Indeed, a reliance on such a limited decoder prevented a complete solution from being reached in [53], where a characterization of C^a was provided under rather restrictive conditions; for details, see [49, Remark (i), p. 756].

A comparison of the decoder in (106) and (107) with that in (104) and (105) reveals two differences. First, the divergence quantity in (104) has, as its second argument, the joint type $P_{f(m)\boldsymbol{s}}$, whereas the analogous argument in (106) is the product of the associated marginal types $P, P_{\boldsymbol{s}}$. Second, in (105), $I(f(m') \wedge \boldsymbol{y}|f(m), \boldsymbol{s})$ is required to be small, whereas in (107) we additionally ask that $I(f(m') \wedge f(m)|\boldsymbol{s})$ also be small.

As a practical matter, the β-decoder in (106) and (107)—although indispensable for theoretical studies—is too complex to be implementable. On the other hand, finding a good decoder in the class of less complex α-decoders for every AVC appears unlikely. Nevertheless, under certain conditions, several common α-decoders suffice for achieving the deterministic code capacity of specific classes of AVC's for the average probability of error. For instance, C^a or $C^a(\Gamma, \Lambda)$ can be achieved under suitable conditions by the joint typicality decoder, the "independence" decoder, the MMI decoder (cf. Section IV-B2)) or the minimum-distance decoder. This issue is briefly addressed below; for a comprehensive treatment, see [49].

Given a set of codewords $\{f(m), m \in \mathcal{M}\}$ in $\mathcal{T}_P^n$ as above, the joint typicality decoder ϕ in [49] is defined as follows: $\phi(\boldsymbol{y}) = m$ iff

$$\max_{x \in \mathcal{X}, y \in \mathcal{Y}} |P_{f(m)\boldsymbol{y}}(x, y) - (P \circ W_\sigma)(x, y)| < \eta, \quad \text{for some } \sigma \in \mathcal{P}(\mathcal{S}) \tag{108}$$

where W_σ is defined by (45), and $\eta > 0$ is chosen suitably small. If more than one $m \in \mathcal{M}$ satisfies (108), or no $m \in \mathcal{M}$ satisfies (108), set $\phi(\boldsymbol{y}) = 0$. Observe that this decoder ϕ is akin to the joint typicality decoder in Section IV-B1), but relies on a less stringent notion of joint typicality than in (104). In a result closely related to that in [53], it is shown in [49] that for the AVC (5), if the input pmf Q^* (cf. paragraph following (47)) satisfies the rather restrictive "Condition DS" (named after Dobrushin and Stambler [53])—which is stronger than the nonsymmetrizability condition (cf. (58) and the subsequent passage)—then C^a can be achieved by the previous joint typicality decoder. An appropriate modification of (108) leads to a joint typicality decoder which achieves $C^a(\Gamma, \Lambda)$ under an analogous "Condition DS(Λ)" [49].

For the special case of additive AVC's, the joint typicality decoder in (108) is practically equivalent to the independence decoder [49]; the latter has the merit of being universal in that it does not rely on a knowledge of the stochastic matrix $W: \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ in (5). Loosely speaking, an AVC (5) with $\mathcal{X}$ and $\mathcal{Y}$ being subsets of a commutative group is called additive if $W(y|x, s)$ depends on x and y through the difference $y - x$ only. (For a formal definition of additive AVC's, see [49, Sec. II].) For a set of codewords $\{f(m), m \in \mathcal{M}\}$ in $\mathcal{T}_P^n$ as above, the independence decoder ϕ is defined as follows: $\phi(\boldsymbol{y}) = m$ iff

$$I(f(m) \wedge \boldsymbol{y} - f(m)) < \eta \tag{109}$$

where $I(f(m) \wedge \boldsymbol{y} - f(m))$ is the mutual information $I(X \wedge Y - X)$ involving dummy rv's X, Y with joint pmf on $\mathcal{X} \times \mathcal{Y}$ being the joint type $P_{f(m)\boldsymbol{y}}$, and $\eta > 0$ is chosen sufficiently small. If no $m \in \mathcal{M}$ or more than one $m \in \mathcal{M}$ satisfies (109), set $\phi(\boldsymbol{y}) = 0$. In effect, the independence decoder ϕ decodes a received sequence $\boldsymbol{y} \in \mathcal{Y}^n$ into a message $m \in \mathcal{M}$ whenever the codeword $f(m)$ is nearly "independent" of the "error" sequence $\boldsymbol{y} - f(m)$. This decoder is shown in [49] to achieve C^a and $C^a(\Gamma, \Lambda)$ under "Condition DS" and the analogous "Condition DS(Λ)," respectively.

The joint typicality decoder (108) reduces to an elementary form for certain subclasses of the class of deterministic AVC's, the latter class being characterized by stochastic matrices $W: \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ in (5) with $\{0, 1\}$-valued entries. This elementary decoder decodes a received sequence $\boldsymbol{y} \in \mathcal{Y}^n$ into a message $m \in \mathcal{M}$ iff the codeword $f(m)$ is "compatible" with $\boldsymbol{y}$. In this context, see [51, Theorem 4] for conditions under which the "erasures only" capacity of a deterministic AVC can be achieved by such a decoder.

The MMI decoder defined in Section IV-B2) can, under certain conditions, achieve C^a or $C^a(\Gamma, \Lambda)$. Specifically, let X, S, Y be dummy rv's with joint pmf $Q^* \circ \sigma^* \circ W$ on $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$, where (Q^*, σ^*) is a saddle point for (46). If the condition

$$I(X \wedge Y) > I(S \wedge Y) \tag{110}$$

is satisfied, then C^a can be achieved by the MMI decoder [49]. When input or state constraints are imposed, if (Q^*, σ^*) satisfies $C^a(\Gamma, \Lambda) = I(Q^*, W_{\sigma^*}) > 0$ in Theorem 6 as well as the condition (110) above, then $C^a(\Gamma, \Lambda)$ can be achieved by the MMI decoder [49]. Next, for any channel with binary input and output alphabets, the MMI decoder is related to the simple minimum (Hamming) distance decoder [49, Lemma 2]. Thus for AVC's with binary input and output alphabets, the minimum-distance decoder often suffices to achieve C^a or $C^a(\Gamma, \Lambda)$. See [49, Theorem 5] for conditions for the efficacy of this decoder.

V. The Gaussian Arbitrarily Varying Channel

While the discrete memoryless AVC (5) with finite input and output alphabets and finite-state space has been the beneficiary of extensive investigations, studies of AVC's with continuous alphabets and state space have been comparatively limited. In this section, we shall briefly review the special case of a Gaussian arbitrarily varying channel (Gaussian AVC). For additional results on the Gaussian AVC and generalizations, we refer the reader to [41]. (Other approaches to, and models for, the study of unknown channels with infinite alphabets can be found, for instance, in [63], [76], [106], and [107].)

A Gaussian AVC is formally defined as follows. Let the input and output alphabets, and the state space, be the real line. For any channel input sequence $\boldsymbol{x} = (x_1, \cdots, x_n)$ and state sequence $\boldsymbol{s} = (s_1, \cdots, s_n)$, the corresponding channel output sequence $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ is given by

$$\boldsymbol{Y} = \boldsymbol{x} + \boldsymbol{s} + \boldsymbol{Z} \tag{111}$$

where $\boldsymbol{Z} = (Z_1, \cdots, Z_n)$ is a sequence of i.i.d. Gaussian rv's with mean zero and variance $\sigma^2 > 0$, denoted $\mathcal{N}(0, \sigma^2)$. The state sequence $\boldsymbol{s}$ may be viewed as interference inserted by an intelligent and adversarial jammer attempting to disrupt the transmission of a codeword $\boldsymbol{x}$. As for the AVC (5), it will be understood that the transmitter and receiver are unaware of the actual state sequence $\boldsymbol{s}$. Likewise, in choosing $\boldsymbol{s}$, the jammer is assumed to be ignorant of the message actually transmitted. The jammer is , however, assumed to know the code when a deterministic code is used, and know the probability law generating the code when a randomized code is used (but not the actual codes chosen).

Power limitations of the transmitter and jammer will be described in terms of an input constraint Γ and state constraint Λ. Specifically, the codewords of a length-n deterministic code (f, ϕ) or a randomized code (F, Φ) will be required to satisfy, respectively,

$$\|f(m)\|^2 \leq n\Gamma, \qquad m \in \mathcal{M} \tag{112}$$

or

$$\|F(m)\|^2 \leq n\Gamma \text{ a.s.}, \qquad m \in \mathcal{M} \tag{113}$$

where $\Gamma > 0$ and $\|\cdot\|$ denotes Euclidean norm. Similarly, only those state sequences $\boldsymbol{s}$ will be permitted which satisfy

$$\|\boldsymbol{s}\|^2 \leq n\Lambda \tag{114}$$

where $\Lambda > 0$.

The corresponding maximum and average probabilities of error are defined as obvious analogs of (42)–(44) with appropriate modifications for randomized codes. The notions of ϵ-capacity and capacity are also defined in the obvious way.

The randomized code capacity of the Gaussian AVC (111), denoted $C_G(\Gamma, \Lambda)$, is given in [80] by the following theorem.

Theorem 14: The randomized code capacity $C_G(\Gamma, \Lambda)$ of the Gaussian AVC (111) under input constraint Γ and state constraint Λ, is given by

$$C_G(\Gamma, \Lambda) = \frac{1}{2} \log\left(1 + \frac{\Gamma}{\Lambda + \sigma^2}\right). \tag{115}$$

Further, a strong converse holds so that

$$C_G(\Gamma, \Lambda) = C_{G,\epsilon}(\Gamma, \Lambda), \qquad 0 < \epsilon < 1. \tag{116}$$

The formula in (115) appears without proof in Blachman [28, p. 58].

Observe that the value of $C_G(\Gamma, \Lambda)$ coincides with the capacity formula for the ordinary memoryless channel with additive Gaussian noise of power $\Lambda + \sigma^2$. Thus the arbitrary interference resulting from the state sequence $\boldsymbol{s}$ in (111) affects achievable rates no worse than i.i.d. Gaussian noise comprising $\mathcal{N}(0, \Lambda)$ rv's. The direct part of Theorem 14 is proved in [80] with the codewords $\{F(m), m \in \mathcal{M}\}$ being distributed independently and uniformly on an n-dimensional sphere of radius $\sqrt{n\Gamma}$. The receiver uses a minimum Euclidean distance decoder Φ_{MD}, namely $\Phi_{MD}(\boldsymbol{y}) = m$ iff

$$\|\boldsymbol{y} - F(m)\| < \|\boldsymbol{y} - F(m')\|, \qquad \text{for } m' \neq m \tag{117}$$

and we set $\Phi_{MD}(\boldsymbol{y}) = 0$ if no $m \in \mathcal{M}$ satisfies (117). The maximum probability of error is then bounded above using a geometric approach in the spirit of Shannon [116]. Theorem 14 can also be proved in an alternative manner analogous to that in [47] for determining the randomized code capacity $C(\Gamma, \Lambda)$ of the AVC (5) (cf. (48)–(50)). In particular, if (Q^*, W_{ζ^*}) is a saddle point for (48), then the counterpart of ζ^* in the present situation is a Gaussian distribution with mean zero and variance Λ; the counterpart of W_{ζ^*} is a Gaussian channel with variance $\Lambda + \sigma^2$.

If the input and state constraints in (112)–(114) on individual codewords and state sequences are weakened to restrictions on the expected values of the respective powers, the Gaussian AVC (111) ceases to have a strong converse; see [80]. The results of Theorem 14 can be extended to a "vector" Gaussian AVC [81] (see also [41]). Earlier work on the randomized code capacity of the Gaussian AVC (111) is due to Blachman [27], [28] who provided lower and upper bounds on capacity when the state sequence is allowed to depend on the actual codeword transmitted. Also, the randomized code capacity problem for the Gaussian AVC has presumably motivated the game-theoretic considerations of saddle points involving mutual information quantities in (cf. e.g., [36] and [97]).

If the state sequence $\boldsymbol{s}$ in (111) is replaced by a sequence $\boldsymbol{S} = (S_1, \cdots, S_n)$ of i.i.d. rv's with a probability distribution function which is unknown to the transmitter and receiver except that it satisfies the constraint

$$\mathbb{E}[S_i^2] \leq \Lambda, \qquad i = 1, \cdots, n \tag{118}$$

the resulting channel can be termed a Gaussian compound memoryless channel (cf. Section II, (3) and (4)). The parameter space Θ (cf. (3)) now corresponds to the set of distribution functions of real-valued rv's S with $\mathbb{E}[S^2] \leq \Lambda$. The capacity of this Gaussian compound channel follows from Dobrushin [52], and is given by the formula in (115). Thus ignorance of the true distribution of the i.i.d interference $\boldsymbol{S} = (S_1, \cdots, S_n)$, other than knowing that it satisfies (118), does not reduce achievable rates any more than i.i.d. Gaussian noise consisting of $\mathcal{N}(0, \Lambda)$ rv's.

We next turn to the performance of the Gaussian AVC (111) for deterministic codes and the average probability of error. Earlier work in this area is due to Ahlswede [3] who determined the capacity of an AVC comprising a Gaussian channel with noise variance arbitrarily varying but not exceeding a given bound. As for its discrete memoryless counterpart (5), the capacity $C_G^a(\Gamma, \Lambda)$ of the Gaussian AVC (111) shows a dichotomy: it either equals the randomized code capacity or else is zero, according to whether or not the transmitter power exceeds the power of the (arbitrary) interference $\boldsymbol{s}$. This result is proved in [50] as

Theorem 15: The deterministic code capacity of the Gaussian AVC (111) under input constraint Γ and state constraint Λ, for the average probability of error, is given by

$$C_G^a(\Gamma, \Lambda) = \begin{cases} \frac{1}{2}\log\left(1+\frac{\Gamma}{\Lambda+\sigma^2}\right), & \text{if } \Gamma > \Lambda \\ 0, & \text{if } \Gamma \le \Lambda. \end{cases} \tag{119}$$

Furthermore, if $\Gamma > \Lambda$, a strong converse holds so that

$$C_G^a(\Gamma, \Lambda) = C_{G,\epsilon}^a(\Gamma, \Lambda), \qquad 0 < \epsilon < 1. \tag{120}$$

Although $C_G^a(\Gamma, \Lambda)$ exhibits a dichotomy similar to the capacity $C^a(\Gamma, \Lambda)$ of the AVC (5) (cf. (57)), a proof of Theorem 15 using Ahlswede's "elimination" technique [7] is not apparent. Its proof in [50] is based on a straightforward albeit more computational approach akin to that in [48]. The direct part uses a code with codewords chosen at random from an n-dimensional spheres of radius $\sqrt{n\Gamma}$ and selectively identified as in [48]. Interestingly, simple minimum Euclidean distance decoding (cf. (117)) suffices to achieve capacity, in contrast with the complex decoding rule (cf. Section IV-B6)) used for the AVC (5) in [48].

In the absence of the Gaussian noise sequence $\boldsymbol{Z} = (Z_1, \cdots, Z_n)$ in (111), we obtain a noiseless additive AVC with output $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{s}$. The deterministic code capacity of this AVC under input constraint Γ and state constraint Λ, for the average probability of error, is, as expected, the limit of the capacity of the Gaussian AVC in Theorem 15 as $\sigma^2 \to 0$ [50]. While this is not a formal consequence of Theorem 15, it can be proved by the same method. Thus the capacity of this AVC equals $1/2 \log(1 + \Gamma/\Lambda)$ if $\Gamma > \Lambda$, and zero if $\Gamma \le \Lambda$, and can be achieved using the minimum Euclidean distance decoder (117). As noted in [50], this result provides a solution to a weakened version of an unsolved sphere-packing problem of purely combinatorial nature. This problem seeks the exponential rate of the maximum number of nonintersecting sphere of radius $\sqrt{n\Gamma}$ in n-dimensional Euclidean space with centers in a sphere of radius $\sqrt{n\Gamma}$. Consider instead a lesser problem in which the spheres are permitted to intersect, but for any given $\boldsymbol{s}$ of norm not exceeding $\sqrt{n\Gamma}$, only for a vanishingly small fraction of sphere centers $\boldsymbol{x}_i$ can $\boldsymbol{x}_i + \boldsymbol{s}$ be closer to another sphere center than to $\boldsymbol{x}_i$. The exponential rate of the maximum number of spheres satisfying this condition is given by the capacity of the noiseless additive AVC above.

Multiple-access counterparts of the single-user Gaussian AVC results surveyed in this section, remain largely unresolved issues.

We note that many of the issues that were described in previous sections for DMC's have natural counterparts for Gaussian channels and for more general channels with infinite alphabets. For example, universal decoding for Gaussian channels with a deterministic but unknown parametric interference was studied in [98], and more general universal decoding for channels with infinite alphabets was studied in [60]; the mismatch problem with minimum Euclidean distance decoding was studied in [100] and [88].

VI. Multiple-Access Channels

The study of reliable communication under channel uncertainty has not been restricted to the single-user channel; considerable attention has also been paid to the multiple-access channel (MAC). The MAC models a communication situation in which multiple users can simultaneously transmit to a single receiver, each user being ignorant of the messages of the other users [39], [44].

Many of the channel models for single-user communication under channel uncertainty have natural counterparts for the MAC. In this section, we shall briefly survey some of the studies of these models. We shall limit ourselves throughout to MAC's with two transmitters only; extensions to more users are usually straightforward.

A known discrete memoryless MAC is characterized by two finite input alphabets $\mathcal{X}_1$, $\mathcal{X}_2$, a finite output alphabet $\mathcal{Y}$, and a stochastic matrix W: $\mathcal{X}_1 \times \mathcal{X}_2 \mapsto \mathcal{Y}$. The rates R_1 and R_2 for the two users are defined analogously as in (12). The capacity region of the MAC for the average probability of error was derived independently by Ahlswede [4] and Liao [94]. A rate-pair (R_1, R_2) is achievable for the average probability of error iff

$$0 \le R_1 \le I(X_1 \wedge Y | X_2, V) \tag{121}$$

$$0 \le R_2 \le I(X_2 \wedge Y | X_1, V) \tag{122}$$

and

$$R_1 + R_2 \le I(X_1, X_2 \wedge Y | V) \tag{123}$$

for some joint pmf $P_{VX_1X_2Y}$ on $\mathcal{V} \times \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}$ of the form

$$P_{VX_1X_2Y}(v, x_1, x_2, y) = P_V(v) P_{X_1|V}(x_1|v) P_{X_2|V}(x_2|v) W(y|x_1, x_2)$$

where the "time-sharing" random variable V with values in the set $\mathcal{V}$ is arbitrary, but may be limited to assume two values, say $\{1, 2\}$ [44]. Extensions to account for average input constraints are discussed in [66], [121], and [127]. Low-complexity codes for the MAC are discussed in [70] and [105].

It is interesting to note that even for a *known* MAC, the average probability of error and the maximal probability of error criteria can lead to different capacity regions [54]; this is in contrast with the capacity of a known single-user channel.

The compound channel capacity region for a *finite* family of discrete memoryless MAC's has been computed by Han

in [77]. In the more general case where the family is not necessarily finite, it can be shown that a rate-pair (R_1, R_2) is achievable for the family

$$\{W(y|x_1, x_2; \theta), x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, y \in \mathcal{Y}, \theta \in \Theta\}$$

iff there exists a joint pmf $P_{X_1X_2V}$ of the form

$$P_{X_1X_2V}(x_1, x_2, v) = P_V(v)P_{X_1|V}(x_1|v)P_{X_2|V}(x_2|v)$$

so that (121)–(123) are satisfied for every $\theta \in \Theta$, where the mutual information quantities are computed with respect to the joint pmf

$$P_{VX_1X_2,Y;\theta}(v, x_1, x_2, y; \theta) = P_V(v)P_{X_1|V}(x_1|v)P_{X_2|V}(x_2|v)W(y|x_1, x_2; \theta).$$

The direct part of the proof of this claim follows from the code constructions in [95] and [103], in which neither the encoder nor the decoder depends on the channel law. The converse follows directly from [39, Sec. 14.3.4], where a converse is proved for the known MAC.

Mismatched decoding for the MAC has been studied in [87], and [88], and universal decoding in [60] and [95].

We turn next to the multiple-access AVC with stochastic matrix W: $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{S} \to \mathcal{Y}$ where $\mathcal{S}$ is a finite set. The deterministic code capacity region of this multiple-access AVC for the average probability of error, denoted $\mathcal{C}^a$, was determined by Jahn [85] assuming that it had a nonempty interior, i.e., $\text{int}\,(\mathcal{C}^a) \neq \emptyset$. A necessary and sufficient computable characterization of multiple-access AVC's for deciding when $\text{int}\,(\mathcal{C}^a) \neq \emptyset$ was not addressed in [85]. Further, assuming that $\text{int}\,(\mathcal{C}^a) \neq \emptyset$, Jahn [85] characterized the randomized code capacity region, denoted $\mathcal{C}$, for the average probability of error in terms of suitable mutual information quantities, and showed that $\mathcal{C}^a = \mathcal{C}$. The validity of this characterization of $\mathcal{C}$, even without the assumption in [85] that $\text{int}\,(\mathcal{C}^a) \neq \emptyset$, was demonstrated by Gubner and Hughes [75]. Observe that if $\text{int}\,(\mathcal{C}^a) = \emptyset$, at least one user and perhaps both users, cannot reliably transmit information over the channel using deterministic codes.

In order to characterize multiple-access AVC's with $\text{int}\,(\mathcal{C}^a) \neq \emptyset$, the notion of single-user symmetrizability (58) was extended by Gubner [72]. This extended notion of symmetrizability for the multiple-access AVC, in fact, involves three distinct conditions: symmetrizability with respect to each of the two individual users, and symmetrizability with respect to the two users jointly; these conditions are termed symmetrizability-$\mathcal{X}_1$, symmetrizability-$\mathcal{X}_2$, and symmetrizability-$\mathcal{X}_1\mathcal{X}_2$, respectively, [72]. Neither of the three conditions above need imply the others. It is readily seen in [72], by virtue of [59] and [48], that if a multiple-access AVC is such that $\text{int}\,(\mathcal{C}^a) \neq \emptyset$, then it must necessarily be nonsymmetrizable-$\mathcal{X}_1$, nonsymmetrizable-$\mathcal{X}_2$, and nonsymmetrizable-$\mathcal{X}_1\mathcal{X}_2$. The sufficiency of this set of nonsymmetrizability conditions for $\text{int}\,(\mathcal{C}^a) \neq \emptyset$ was conjectured in [72] and proved by Ahlswede and Cai [15], thereby completely resolving the problem of characterizing $\mathcal{C}^a$. (It was shown in [72] that $\text{int}\,(\mathcal{C}^a) \neq \emptyset$ under a set of conditions which are sufficient but not necessary.) Ahlswede and Cai [16] have further demonstrated that if the multiple-access AVC is only nonsymmetrizable-$\mathcal{X}_1\mathcal{X}_2$ (but can be symmetrizable-$\mathcal{X}_1$ or symmetrizable-$\mathcal{X}_2$), both users can still reliably transmit information over the channel using deterministic codes, if they have access to correlated side-information.

The randomized code capacity region of the multiple-access AVC under state constraint Λ (cf. (24)) for the maximum or average probability of error, denoted $\mathcal{C}(\Lambda)$, has been determined by Gubner and Hughes [75]. The presence of the state constraint renders $\mathcal{C}(\Lambda)$ nonconvex in general [75]; the corresponding capacity region $\mathcal{C}$ in the absence of any state constraint [85] is convex. Input constraints analogous to (22) are also considered in [75].

The deterministic code capacity region of the multiple-access AVC under state constraint Λ for the average probability of error remains unresolved. For preliminary results, see [73] and [74].

Indeed, multiple-access AVC counterparts of the single-user discrete memoryless AVC results of Section III-A2), which have not been mentioned above in this section, remain by and large unresolved issues.

VII. Discussion

We discuss below the potential role in mobile wireless communications of the work surveyed in this paper. Several situations in which information must be conveyed reliably under channel uncertainty are considered in light of the channel models described above. The difficulties encountered when attempting to draw practical guidelines concerning the design of transmitters and receivers for such situations are also examined. Suggested avenues for future research are indicated.

We limit our discussion to single-user channels, in which case the receiver for a given user treats all other users' signals (when present) as noise. (For some multiuser models see [26], [110], and references therein.) We do not, therefore, investigate the benefits of using the multiple-access transmitters and receivers suggested by the work mentioned in Section VI. We remark that the discrete channels surveyed above should be viewed as resulting from combinations of modulators, waveform transmission channels, and demodulators.

A few preliminary observations are in order. Considerations of delays in encoding and decoding as well as decoder complexity typically dictate the choice of blocklength n of codewords used in a given communication situation. Encoding delays result from the fact that a message must be buffered prior to transmission until an entire (block) codeword for it has been formed. Decoding delays are incurred since all the symbols in a codeword must be received before the operation of decoding can commence. Once a blocklength n has been fixed, the channel dictates a tradeoff between the transmitter power, the code rate, and the probability of decoding error. We note that if the choice of the blocklength n is determined by delay considerations rather than by those of complexity, the use of a complex decoder for enhancing channel coding performance becomes feasible. On the other hand, overriding concerns of complexity often inhibit the use of complex de-

coder structures. For instance, the universal MMI decoder (cf. Section III-A1)), which is known to achieve channel capacity and the random-coding error exponent in many situations, does not always afford a simple algorithmic implementation even when used in conjunction with an algebraically well-structured block code or a convolutional code on a DMC; however, see [92], [93], and [129]. Thus the task of finding universal decoders of manageable complexity constitutes a challenging research direction [93]. An alternative approach for designing receivers for use on unknown channels, which is widely used in practice, employs training sequences for estimating the parameters of the unknown channel followed by maximum-likelihood decoding (cf. Section IV-B1)) with respect to the estimated channel. In many situations, this approach leads to simple receiver designs. A drawback of this approach is that the code rate for information transmission is, in effect, reduced as the symbols of the training sequence appropriate a portion of blocklength n fixed by the considerations mentioned earlier. On the other hand, in situations where the unknown channel remains unchanged over multiple transmissions, viz. codewords, this approach is particularly attractive since channel parameters estimated with a training sequence during a transmission can be reused in subsequent transmissions.

An information signal transmitted over a mobile radio channel undergoes fading whose nature depends on the relation between the signal parameters (e.g., signal bandwidth) and the channel parameters (e.g., delay spread, Doppler spread). (For a comprehensive treatment, cf., e.g., [104, Ch. 4].) Four distinct types of fading can be experienced by an information signal, which are described next.

Doppler spread effects typically result in either "slow" fading or "fast" fading. Let T_n denote the transmission time (in seconds) of a codeword of blocklength n, and T_c the channel coherence time (in seconds). In slow fading, $T_n \ll T_c$, so that the channel remains effectively unchanged during the transmission of a codeword; hence, it can be modeled as a compound channel, without or with memory (cf. Section II). On the other hand, fast fading, when $T_n > T_c$, results in the channel undergoing changes during the transmission of a codeword, so that a compound channel model is no longer appropriate.

Independently of the previous effects, a multipath delay spread mechanism gives rise to either "flat" fading or "frequency-selective" fading. In flat fading, $T_n/n \gg \sigma_\tau$, where σ_τ is the root-mean-square (rms) delay spread (in seconds); in effect, the channel can be assumed to be memoryless from symbol to symbol of a codeword. In contrast, frequency-selective fading, when $T_n/n < \sigma_\tau$, results in intersymbol interference (ISI) which introduces memory into the channel, suggesting the use of finite-state models (cf. Section II).

The fading effects described above produce the four different combinations of slow flat fading, slow frequency-selective fading, fast flat fading and fast frequency-selective fading. It is argued below that the resulting channels can be described to various extents by the channel models of Section II; however, the work reviewed above may fail to provide satisfactory recommendations for transmitter–receiver designs which meet the delay and complexity requirements mentioned earlier.

For channels with slow flat fading, the compound DMC model (4) is an apt choice. The MMI decoder achieves the capacity of this channel (cf. Section IV-B4)); however, complexity considerations may preclude its use in practice. This situation is mitigated by the observation in [100] that a code with equi-energy codewords and minimum Euclidean distance decoder is often adequate. Alternatively, a training sequence can be used to estimate the prevailing state of the compound DMC, followed by maximum-likelihood decoding. A drawback of this approach, of course, is the effective loss of code rate alluded to earlier.

Channels characterized by slow frequency-selective fading can be described by a compound finite-state channel model (cf. Section III-A1)). The universal decoder in [60] achieves channel capacity and the random coding-error exponent. The high complexity of this decoder, however, renders it impractical if complexity is an overriding concern. In this situation, a training sequence approach as above offers a remedy, albeit at the cost of an effective reduction in code rate. A training sequence can be used to estimate the unknown ISI parameters of the compound FSC model followed by maximum-likelihood decoding; the special structure of the ISI channel renders both these operations fairly straightforward.

Channels with fast flat fading fluctuate between several different attenuation levels during the transmission of a codeword; during the period in which each such attenuation level prevails, the channels appear memoryless. A description of such a channel will depend on the severity of the fast fade. For instance, consider the case where different attenuation levels are experienced often enough during the transmission of a codeword. A compound finite-state model (cf. Section II) is a feasible candidate, where the set of states Σ corresponds to the set of attenuation levels, by dint of the fact that the "ergodicity time" T_e of the channel satisfies $T_e < T_n$. However, no encouraging options can be inferred from the work surveyed above for acceptable transmitter–receiver designs. A complex decoder [60] is generally needed to achieve channel capacity and the random-coding error exponent. Furthermore, the feasibility of the training sequence approach is also dubious owing to the inherent complexity of the estimation procedure and of the computation of the likelihood metric.[5] Next, if $T_e > T_n$, a compound FSC model is no longer appropriate, and even the task of finding an acceptable channel description from among the models surveyed appears difficult. Of course, an AVC model (5), with state space comprising the different attenuation levels, can be used provided the transitions between such levels occur in a memoryless manner; else, an arbitrarily varying FSC model (9) can be considered. When $T_e > T_n$, the choice of an arbitrarily varying channel model may, however, lead to overly conservative estimates of channel capacity. It must, however, be noted that in the former case, an AVC model does offer the feasibility of simpler transmitter and receiver designs through the use of randomized codes (with maximum-

[5] Even when the law of a finite-state channel is known, the maximum-likelihood decoder may be too complex to implement, since the computation of the likelihood of a received sequence given a codeword is exponential in the blocklength (7). A suboptimal decoder which does not necessarily achieve the random-coding error exponent, but does achieve capacity for some finite-state channels is discussed in [69] and [101].

likelihood decoder) for achieving channel capacity (cf. Section IV-A2)).

Finally, a channel with fast frequency-selective fading can be understood in a manner analogous to fast flat fading, with the difference that during the period of each prevalent attenuation level the channel possesses memory. Also, if $T_e < T_n$, such a channel can be similarly modeled by a compound FSC (cf. Section II), where the set of states—representing the various attenuation levels—now corresponds to a family of "smaller" FSC's with unknown parameters. Clearly, the practical feasibility of a decoder which achieves channel capacity or a receiver based on a training sequence approach appears remote. If $T_e > T_n$, similar comments apply as for the analogous situation in fast flat fading; each arbitrarily varying channel state, representing an attenuation level, will now correspond to a "smaller" FSC with unknown parameters.

Thus information-theoretic studies of unknown channels have produced classes of models which are rich enough to faithfully describe many situations arising in mobile wireless communications. There are, of course, some situations involving fast fading which yet lack satisfactory descriptions and for which new tractable channel models are needed. However, the shortcomings are acute in terms of providing acceptable guidelines for the design of transmitters and receivers which adhere to delay and complexity requirements. The feasibility of the training sequence approach is crucially reliant on the availability of good estimates of channel parameters and the ease of computation of the likelihood metric, which can pose serious difficulties especially for channels with memory. This provides an impetus for the study of efficient decoders which do not require a knowledge of the channel law and yet allow reliable communication at rates up to capacity with reasonable delay and complexity.

ACKNOWLEDGMENT

The authors are grateful to M. Pinsker for his careful reading of this paper and for his many helpful suggestions. They also thank S. Verdú and the reviewers for their useful comments.

REFERENCES

[1] R. Ahlswede, "Certain results in coding theory for compound channels," in *Proc. Coll. Inf. The. Debrecen 1967*, A. Rényi, Ed. Budapest, Hungary: J. Bolyai Math. Soc., 1968, vol. 1, pp. 35–60.

[2] ———, "A note on the existence of the weak capacity for channels with arbitrarily varying channel probability functions and its relation to Shannon's zero error capacity," *Ann. Math. Statist.*, vol. 41, pp. 1027–1033, 1970.

[3] ———, "The capacity of a channel with arbitrary varying Gaussian channel probability functions," in *Trans. 6th Prague Conf. Information Theory, Statistical Decision Functions and Random Processes*, Sept. 1971, pp. 13–21.

[4] ———, "Multi-way communication channels," in *Proc. 2nd. Int. Symp. Information Theory*. Budapest, Hungary: Hungarian Acad. Sci., 1971, pp. 23–52.

[5] ———, "Channel capacities for list codes," *J. Appl. Probab.*, vol. 10, pp. 824–836, 1973.

[6] ———, "Elimination of correlation in random codes for arbitrarily varying channels," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 44, pp. 159–175, 1978.

[7] ———, "Elimination of correlation in random codes for arbitrarily varying channels," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 44, pp. 159–175, 1978.

[8] ———, "Coloring hypergraphs: A new approach to multiuser source coding, Part I," *J. Combin., Inform. Syst. Sci.*, vol. 4, no. 1, pp. 76–115, 1979.

[9] ———, "Coloring hypergraphs: A new approach to multiuser source coding, Part II," *J. Combin., Inform. Syst. Sci.*, vol. 5, no. 3, pp. 220–268, 1980.

[10] ———, "A method of coding and an application to arbitrarily varying channels," *J. Comb., Inform. Syst. Sci..*, vol. 5, pp. 10–35, 1980.

[11] ———, "Arbitrarily varying channels with states sequence known to the sender," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 621–629, Sept. 1986.

[12] ———, "The maximal error capacity of arbitrarily varying channels for constant list sizes," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1416–1417, July 1993.

[13] R. Ahlswede, L. A. Bassalygo, and M. S. Pinsker, "Localized random and arbitrary errors in the light of arbitrarily varying channel theory," *IEEE Trans. Inform. Theory*, vol. 41, pp. 14–25, Jan. 1995.

[14] R. Ahlswede and N. Cai, "Two proofs of Pinsker's conjecture concerning arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1647–1649, Nov. 1991.

[15] ———, "Arbitrarily varying multiple-access channels Part I. Ericson's symmetrizability is adequate, Gubner's conjecture is true," in *Proc. IEEE Int. Symp. Information Theory* (Ulm, Germany, 1997), p. 22.

[16] ———, "Arbitrarily varying multiple-access channels, Part II: Correlated sender's side information, correlated messages, and ambiguous transmission," in *Proc. IEEE Int. Symp. Information Theory* (Ulm, Germany, 1997), p. 23.

[17] ———, "Correlated sources help transmission over an arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1254–1255, July 1997.

[18] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography: Part II: CR capacity," *IEEE Trans. Inform. Theory*, vol. 44, pp. 225–240, Jan 1998.

[19] R. Ahlswede and J. Wolfowitz, "Correlated decoding for channels with arbitrarily varying channel probability functions," *Inform. Contr.*, vol. 14, pp. 457–473, 1969.

[20] ———, "The capacity of a channel with arbitrarily varying channel probability functions and binary output alphabet," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 15, pp. 186–194, 1970.

[21] V. B. Balakirsky, "Coding theorem for discrete memoryless channels with given decision rules," in *Proc. 1st French–Soviet Workshop on Algebraic Coding* (Lecture Notes in Computer Science 573), G. Cohen, S. Litsyn, A. Lobstein, and G. Zémor, Eds. Berlin, Germany: Springer-Verlag, July 1991, pp. 142–150.

[22] ———, "A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1889–1902, Nov. 1995.

[23] A. Barron, J. Rissanen, and B. Yu, "Minimum description length principle in modeling and coding," this issue, pp. 2743–2760.

[24] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications* (CISM Course and Lecture Notes, no. 229), G. Longo, Ed. Berlin, Germany: Springer-Verlag, 1977, pp. 172–231.

[25] T. Berger and J. Gibson, "Lossy data compression," this issue, pp. 2693–2723.

[26] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information theoretic and communications aspects," this issue, pp. 2619–2692.

[27] N. M. Blachman, "The effect of statistically dependent interference upon channel capacity," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 553–557, Sept. 1962.

[28] ———, "On the capacity of a band-limited channel perturbed by statistically dependent interference," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 48–55, Jan. 1962.

[29] D. Blackwell, L. Breiman, and A. J. Thomasian, "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Statist.*, vol. 29, no. 4, pp. 1209–1220, 1958.

[30] ———, "The capacity of a class of channels," *Ann. Math. Statist.*, vol. 30, pp. 1229–1241, Dec. 1959.

[31] ———, "The capacities of certain channel classes under random coding," *Ann. Math. Statist.*, vol. 31, pp. 558–567, 1960.

[32] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.

[33] V. Blinovsky, P. Narayan, and M. Pinsker, "Capacity of an arbitrarily varying channel under list decoding," *Probl. Pered. Inform.*, vol. 31, pp. 99–113, 1995, English translation.

[34] V. Blinovsky and M. Pinsker, "Estimation of the size of the list when decoding over an arbitrarily varying channel," in *Proc. 1st French–Israeli Workshop on Algebraic Coding,* G. Cohen *et al.*, Eds.

(Paris, France, July 1993). Berlin, Germany: Springer, 1993, pp. 28–33.
[35] ——, "One method of the estimation of the size for list decoding in arbitrarily varying channel," in *Proc. of ISITA-94* (Sidney, Australia, 1994), pp. 607–609.
[36] J. M. Borden, D. J. Mason, and R. J. McEliece, "Some information theoretic saddlepoints," *SIAM Contr. Opt.*, vol. 23, no. 1, Jan. 1985.
[37] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 439–441, May 1983.
[38] T. M. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 2–14, Jan. 1972.
[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[40] I. Csiszár, "The method of types," this issue, pp. 2505–2523.
[41] ——, "Arbitrarily varying channels with general alphabets and states," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1725–1742, Nov. 1992.
[42] I. Csiszár and J. Körner, "Many coding theorems follow from an elementary combinatorial lemma," in *Proc. 3rd Czechoslovak–Soviet–Hungarian Sem. Information Theory* (Liblice, Czechoslovakia, 1980), pp. 25–44.
[43] ——, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 5–12, Jan. 1981.
[44] ——, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[45] ——, "On the capacity of the arbitrarily varying channel for maximum probability of error," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 57, pp. 87–101, 1981.
[46] I. Csiszár, J. Körner, and K. Marton, "A new look at the error exponent of discrete memoryless channels," in *IEEE Int. Symp. Information Theory* (Cornell Univ., Ithaca, NY, Oct. 1977), unpublished preprint.
[47] I. Csiszár and P. Narayan, "Arbitrarily varying channels with constrained inputs and states," *IEEE Trans. Inform. Theory*, vol. 34, pp. 27–34, Jan. 1988.
[48] ——, "The capacity of the arbitrarily varying channel revisited: Capacity, constraints," *IEEE Trans. Inform. Theory*, vol. 34, pp. 181–193, Jan. 1988.
[49] ——, "Capacity and decoding rules for classes of arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 752–769, July 1989.
[50] ——, "Capacity of the Gaussian arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 18–26, Jan. 1991.
[51] ——, "Channel capacity for a given decoding metric," *IEEE Trans. Inform. Theory*, vol. 41, pp. 35–43, Jan. 1995.
[52] R. L. Dobrushin, "Optimum information transmission through a channel with unknown parameters," *Radio Eng. Electron.*, vol. 4, no. 12, pp. 1–8, 1959.
[53] R. L. Dobrushin and S. Z. Stambler, "Coding theorems for classes of arbitrarily varying discrete memoryless channels," *Probl. Pered. Inform.*, vol. 11, no. 2, pp. 3–22, 1975, English translation.
[54] G. Dueck, "Maximal error capacity regions are smaller than average error capacity regions for multi-user channels," *Probl. Contr. Inform. Theory*, vol. 7, pp. 11–19, 1978.
[55] P. Elias, "List decoding for noisy channels," in *IRE WESCON Conv. Rec.*, 1957, vol. 2, pp. 94–104.
[56] ——, "Zero error capacity under list decoding," *IEEE Trans. Infom. Theory*, vol. 34, pp. 1070–1074, Sept. 1988.
[57] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, pp. 1977–1997, Sept. 1963.
[58] T. Ericson, "A min-max theorem for antijamming group codes," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 792–799, Nov. 1984.
[59] ——, "Exponential error bounds for random codes in the arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 42–48, Jan. 1985.
[60] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1726–1745, Sept. 1998.
[61] N. Merhav and M. Feder, "Universal prediction," this issue, pp. 2124–2147.
[62] G. D. Forney, "Exponential error bounds for erasure, list and decision feedback systems," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 206–220, Mar. 1968.
[63] L. J. Forys and P. P. Varaiya, "The ϵ-capacity of classes of unknown channels," *Inform. Contr.*, vol. 14, pp. 376–406, 1969.
[64] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[65] ——, "The random coding bound is tight for the average code," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 244–246, Mar. 1973.
[66] ——, "Energy limited channels: Coding, multiaccess, and spread spectrum," Tech. Rep. LIDS-P-1714, Lab. Inform. Decision Syst., Mass. Inst. Technol., Cambridge, MA, Nov. 1988.
[67] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Contr. Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.
[68] E. N. Gilbert, "Capacity of burst-noise channels," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–1265, Sept. 1960.
[69] 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.
[70] A. Grant, R. Rimoldi, R. Urbanke, and P. Whiting, "Rate-splitting multiple access for discrete memoryless channels," *IEEE Trans. Inform. Theory*, to be published.
[71] R. Gray and D. Neuhoff, "Quantization," this issue, pp. 2325–2383.
[72] J. A. Gubner, "On the deterministic-code capacity of the multiple-access arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 36, pp. 262–275, Mar. 1990.
[73] ——, "State constraints for the multiple-access arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 37, pp. 27–35, Jan. 1991.
[74] ——, "On the capacity region of the discrete additive multiple-access arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1344–1346, July 1992.
[75] J. A. Gubner and B. L. Hughes, "Nonconvexity of the capacity region of the multiple-access arbitrarily varying channel subject to constraints," *IEEE Trans. Inform. Theory*, vol. 41, pp. 3–13, Jan. 1995.
[76] D. Hajela and M. Honig, "Bounds on ϵ-rate for linear, time-invariant, multi-input/multi-output channels," *IEEE Trans. Inform. Theory*, vol. 36, Sept. 1990.
[77] T. S. Han, "Information-spectrum methods in information theory," Graduate School of Inform. Syst., Univ. Electro-Communications, Chofu, Tokyo 182 Japan, Tech. Rep., 1997.
[78] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 731–739, Sept. 1983.
[79] M. Hegde, W. E. Stark, and D. Teneketzis, "On the capacity of channels with unknown interference," *IEEE Trans. Inform. Theory*, vol. 35, pp. 770–783, July 1989.
[80] B. Hughes and P. Narayan, "Gaussian arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 267–284, Mar. 1987.
[81] ——, "The capacity of a vector Gaussian arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 34, pp. 995–1003, Sept. 1988.
[82] B. L. Hughes, "The smallest list size for the arbitrarily varying channel," in *Proc. 1993 IEEE Int. Symp. Information Theory* (San Antonio, TX, Jan. 1993).
[83] ——, "The smallest list for the arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 43, pp. 803–815, May 1997.
[84] J. Y. N. Hui, "Fundamental issues of multiple accessing," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 1983.
[85] J. H. Jahn, "Coding for arbitrarily varying multiuser channels," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 212–226, Mar. 1981.
[86] F. Jelinek, "Indecomposable channels with side information at the transmitter," *Inform. Contr.*, vol. 8, pp. 36–55, 1965.
[87] A. Lapidoth, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1439–1452, Sept. 1996.
[88] ——, "Nearest-neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1520–1529, Sept. 1996.
[89] ——, "On the role of mismatch in rate distortion theory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 38–47, Jan. 1997.
[90] A. Lapidoth and İ. E. Telatar, private communication, Dec. 1997.
[91] ——, "The compound channel capacity of a class of finite-state channels," *IEEE Trans. Inform. Theory*, vol. 44, pp. 973–983, May 1998.
[92] A. Lapidoth and J. Ziv, "On the universality of the LZ-based decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1746–1755, Sept. 1998.
[93] ——, "Universal sequential decoding," presented at the 1998 Information Theory Workshop, Kerry, Killarney Co., Ireland.
[94] H. Liao, "Multiple access channels," Ph.D. dissertation, Dept. Elec. Eng., Univ. Hawaii, 1972.
[95] Y.-S. Liu and B. L. Hughes, "A new universal random coding bound for the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 376–386, Mar. 1996.
[96] L. Lovász, "On the Shannon capacity of a graph," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 1–7, Jan. 1979.
[97] R. J. McEliece, "CISM courses and lectures," in *Communication in the Presence of Jamming–An Information Theory Approach*, no. 279. New York: Springer, 1983.
[98] N. Merhav, "Universal decoding for memoryless Gaussian channels with deterministic interference," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1261–1269, July 1993.
[99] ——, "How many information bits does a decoder need about the

channel statistics," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1707–1714, Sept. 1997.
[100] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1953–1967, Nov. 1994.
[101] 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.
[102] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, pp. 359–378, May 1994.
[103] J. Pokorny and H. M. Wallmeier, "Random coding bound and codes produced by permutations for the multiple access channel," *IEEE Trans. Inform. Theory*, 1985.
[104] T. S. Rappaport, *Wireless Communications, Principles and Practice*. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[105] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 364–375, Mar. 1996.
[106] W. L. Root, "Estimates of ϵ capacity for certain linear communication channels," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 361–369, May 1968.
[107] W. L. Root and P. P. Varaiya, "Capacity of classes of Gaussian channels," *SIAM J. Appl. Math.*, vol. 16, no. 6, pp. 1350–1393, Nov. 1968.
[108] J. Salz and E. Zehavi, "Decoding under integer metrics constraints," *IEEE Trans. Commun.*, vol. 43, nos. 2/3/4, pp. 307–317, Feb./Mar./Apr. 1995.
[109] S. Shamai, "A broadcast transmission strategy of the Gaussian slowly fading channel," in *Proc. Int. Symp. Information Theory ISIT'97* (Ulm, Germany, 1997), p. 150.
[110] S. Shamai (Shitz) and A. D. Wyner, "Information-theoretic considerations for systematic, cellular, multiple-access fading channels, Parts I and II," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1877–1894, Nov. 1997.
[111] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
[112] ———, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 8–19, 1956.
[113] ———, "Certain results in coding theory for noisy channels," *Inform. Contr.*, vol. 1, pp. 6–25, 1957.
[114] ———, "Channels with side information at the transmitter," *IBM J. Res. Develop.*, vol. 2, no. 4, pp. 289–293, 1958.
[115] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels," *Infom. Contr.*, vol. 10, pp. 65–103, pt. I, pp. 522–552, pt. II, 1967.
[116] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.*, vol. 38, pp. 611–656, May 1959.
[117] M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications Handbook*. New York: McGraw-Hill, 1994, revised edition.
[118] S. Z. Stambler, "Shannon theorems for a full class of channels with state known at the output," *Probl. Pered. Inform.*, vol. 14, no. 4, pp. 3–12, 1975, English translation.
[119] I. G. Stiglitz, "Coding for a class of unknown channels," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 189–195, Apr. 1966.
[120] İ. E. Telatar, "Zero-error list capacities of discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1977–1982, Nov. 1997.
[121] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1019–1030, Sept. 1990.
[122] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1147–1157, July 1994.
[123] H. S. Wang and N. Moayeri, "Finite-state Markov channel—A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, pp. 163–171, Feb. 1995.
[124] J. Wolfowitz, "The coding of messages subject to chance errors," *Illinois J. Math.*, vol. 1, pp. 591–606, Dec. 1957.
[125] ———, "Simultaneous channels," *Arch. Rat. Mech. Anal.*, vol. 4, pp. 371–386, 1960.
[126] ———, *Coding Theorems of Information Theory*, 3rd ed. Berlin, Germany: Springer-Verlag, 1978.
[127] A. D. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1713–1727, Nov. 1994.
[128] A. D. Wyner, J. Ziv, and A. J. Wyner, "On the role of pattern matching in information theory," this issue, pp. 2045–2056.
[129] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 453–460, July 1985.
[130] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.