Generalizing Capacity: New Definitions and Capacity Theorems for Composite Channels

Michelle Effros, *Fellow, IEEE*, Andrea Goldsmith, *Fellow, IEEE*, and Yifan (Ethan) Liang

Abstract—We consider three capacity definitions for composite channels with channel side information at the receiver. A composite channel consists of a collection of different channels with a distribution characterizing the probability that each channel is in operation. The *Shannon capacity* of a channel is the highest rate asymptotically achievable with arbitrarily small error probability. Under this definition, the transmission strategy used to achieve the capacity must achieve arbitrarily small error probability for all channels in the collection comprising the composite channel. The resulting capacity is dominated by the worst channel in its collection, no matter how unlikely that channel is. We, therefore, broaden the definition of capacity to allow for some outage. The capacity versus outage is the highest rate asymptotically achievable with a given probability of decoder-recognized outage. The *expected capacity* is the highest average rate asymptotically achievable with a single encoder and multiple decoders, where channel side information determines the channel in use. The expected capacity is a generalization of capacity versus outage since codes designed for capacity versus outage decode at one of two rates (rate zero when the channel is in outage and the target rate otherwise) while codes designed for expected capacity can decode at many rates. Expected capacity equals Shannon capacity for channels governed by a stationary ergodic random process but is typically greater for general channels. The capacity versus outage and expected capacity definitions relax the constraint that all transmitted information must be decoded at the receiver. We derive channel coding theorems for these capacity definitions through information density and provide numerical examples to highlight their connections and differences. We also discuss the implications of these alternative capacity definitions for end-to-end distortion, source-channel coding, and separation.

Index Terms—Capacity versus outage, composite channel, expected capacity, information density, separation, Shannon capacity.

I. INTRODUCTION

CHANNEL capacity has a natural *operational definition* associated with reliable communication: the highest rate at which information can be sent over the channel with arbi-

M. Effros is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: effros@caltech.edu).
A. Goldsmith is with the Department of Electrical Engineering, Stanford University, Stanford CA 94305 USA (e-mail: andrea@wsl.stanford.edu).
Y. (E.) Liang is with Goldman, Sachs & Co., New York, NY 10004-1950 USA (e-mail: ethan.liang.el@gmail.com).

trarily small probability of error. Channel coding theorems and their converses prove that the capacity associated with a given channel model equals a formula that depends on the channel's probabilistic description. In his landmark paper [1], Shannon proved that the formula

$$C = \max_X I(X; Y) \tag{1}$$

equals the capacity of a *memoryless* channel, where X is the channel input, Y is the channel output, and $I(X; Y)$ is the average mutual information between X and Y. The formula (1), extended to the limiting expression

$$C = \lim_{n \to \infty} \sup_{X^n} \frac{1}{n} I(X^n; Y^n) \tag{2}$$

has been shown, when it exists, to equal the capacity of certain channels with memory, where now the input is the sequence X^n of length n with corresponding output sequence Y^n. In particular, Dobrushin proved that channel capacity is given by (2) for information stable channels. Channels for which the formula (2) does not converge to a limit are called information unstable channels, and for these channels this formula does not in general equal capacity. Examples of information unstable channels include the stationary regular decomposable channels [2], the stationary nonanticipatory channels [3], and the averaged memoryless channels [4].

In [5], Verdú and Han showed that the formula

$$C = \sup_X \underline{I}(X; Y) \tag{3}$$

equals capacity for general channels, where $\underline{I}(X; Y)$ is the liminf in probability of the normalized information density between input sequence X and output sequence Y. In particular, they showed that this formula equals capacity of very general channels without any assumptions such as memorylessness, information stability, stationarity, or causality on the channel's probabilistic description.

The focus of this paper is on one class of these general information unstable channels, the *composite channel*. A composite channel, formally defined in Section III, is a collection of channels $\{W_s : s \in \mathcal{S}\}$ parameterized by s, with each channel in the collection $W_s = \{W_s^n = \mathbf{P}_{\mathbf{Y^n}|\mathbf{X^n}}\}_{n=1}^{\infty}$ consisting of a sequence of n-dimensional conditional output distributions mapping from a common input alphabet to a common output alphabet. The channel input-output distribution is determined by the random variable S, independent of the channel input, which is chosen according to some channel state distribution $p_S(s)$ at the beginning of transmission and then held fixed. Thus, composite channels have an input-output distribution taken from a collection of possible channels $s \in \mathcal{S}$. A

compound channel [6] or class of channels [7], [8] is a similar model in that it also consists of a collection of possible channels $\{W_s : s \in \mathcal{S}\}$. However, these channel collections do not assume any distribution over the channels in the collection. Without this distribution, a transmission strategy can be unreliable with high probability unless it is designed to work reliably for all possible channels in the collection, since a "worst-case" channel may be realized with arbitrary probability. Another channel model associated with a collection of channels is the *average channel* defined by Ahlswede [4]. This channel, also called the *mixed channel* by Han [9], has a single input-output distribution associated with mixing over a composite channel distribution, i.e., the mixed channel has conditional distribution $W(\cdot|\cdot) = \sum_{s \in \mathcal{S}} \mathbf{p}_S(fs)W_s(\cdot|\cdot)$. While the mixed channel is defined over a collection of channels, its input-output distribution W, as a weighted sum, has the same or similar characteristics as each input-output distribution W_s in the collection. Hence, if the capacity of each channel in the collection is known, the capacity of the mixed channel is usually straightforward to determine (e.g., for discrete channel collections [4] or general channel collections [9]). Composite channels also differ from the arbitrarily varying channel [6], [10], where the channel state changes on each transmission in a manner that depends on the channel input, since the state S of a composite channel is independent of its input. The Shannon capacity for all of these channels has been explicitly characterized. In particular, the capacity of the composite channel is a special case of the general channel capacity derived by Verdú and Han. However, the distribution over the collection of channels is not used in this capacity calculation, since the definition of Shannon capacity requires reliable communication for all channels in the collection. Hence, the Verdú–Han (Shannon) capacity of a composite channel will be the same as the Shannon capacity of a compound channel over the same collection of possible channels $\{W_s : s \in \mathcal{S}\}$, regardless of the distribution $p_S(s)$ over the composite channel states. Our outage and expected capacity definitions take advantage of this distribution to allow unreliable transmission over the composite channel with some known probability. These capacity definitions are not possible for any of the other channel models described above, since either their channel state distribution is unknown, there is only a single channel state based on mixing, or the channel state depends on the channel input.

The composite channel model approximates many communication systems of practical interest, for instance, applications with fixed delay constraints such that a codeword may not experience all possible channel states. In particular, communications systems designed for slowly fading wireless channels typically have the channel coherence time much longer than the codeword duration. Hence, a codeword will experience one or just a few fading realizations out of all possible fading values. The composite channel model can also be generalized to channels for which the optimal input distribution induces a joint input-output distribution on which the ergodic decomposition theorem [11, Theorem 1.8.2] holds, e.g., stationary distributions defined on complete, separable metric spaces (Polish spaces). In this case the channel index s becomes the channel's *ergodic mode*.

In general, the capacity of a channel is formally defined as the maximal rate at which information can be reliably trans-

mitted over it (see, e.g., Definition 1.1 of [6]). Reliability is here used to mean that the maximum decoding error, over all codewords, can be made arbitrarily close to zero. Reliability is generally achieved by allowing the code blocklength (and, therefore, the delay) to grow without bound. This definition holds for all channels and has enabled great insight and design inspiration in communication systems. However, since the definition requires that all transmitted information be correctly decoded, for composite channels the capacity is dominated by the performance of the "worst" component channel, no matter how small its probability. This highlights the pessimistic nature of this capacity definition, which requires arbitrarily small error probability. An alternate capacity definition that allows errors is ϵ-capacity, denoted C_ϵ, which is given in Definition 1.1 of [6] as the maximal rate at which information can be transmitted over a channel such that the maximum decoding error, over all codewords, does not exceed ϵ. The relationship between a channel's ϵ-capacity and its capacity, denoted C, is thus $\lim_{\epsilon \to 0} C_\epsilon = C$. Several alternate capacity definitions have been considered that incorporate a notion of channel outage. In particular, outage capacity has been applied to channels that vary over multiple channel states with perfect state information available at the transmitter(s) and receiver(s). The traditional definition of capacity for this model, also called the ergodic or Shannon capacity, equals the maximum mutual information for each channel state averaged probabilistically over all channel states. This average capacity is achieved by adapting the transmission rate to approach capacity in each channel state. Outage capacity is typically applied to systems where a variable transmission rate is not possible or desirable, for example in fixed-rate delay-constrained applications such as voice. In this setting, the transmitter declares an outage for certain states and transmits at the same fixed rate in nonoutage states. The outage capacity defines the maximal (fixed) rate achievable in all nonoutage states with asymptotically small error probability [12], [13]. For an outage probability of zero, this definition reduces to zero-outage or delay-limited capacity [14]. For the same channel model with state information available only at the receiver, capacity versus outage is a common metric, whereby the transmitter sends at a fixed rate, and the receiver declares an outage when the channel cannot support this rate with asymptotically small error probability [15]. These outage notions in capacity definitions are most commonly applied to wireless channel models [16], [17], as they reflect common design practice, where some loss of data is tolerated in exchange for higher overall data rates. Our definition of capacity versus outage given below extends these ideas to composite channels with receiver knowledge of the channel state.

Throughout this paper we assume the channel state information (CSI) is revealed to the receiver (CSIR), but no channel state information is available at the transmitter (CSIT). We assume CSIR since otherwise capacity analysis must be based on *mismatch capacity*, where the decoding rule is mismatched to the actual channel realization, as we discuss in more detail in Section III. Typically CSI is measured at the receiver to obtain CSIR and, when a feedback link is available, sent back to the transmitter to obtain CSIT. A downlink satellite communication system is an example where this feedback link is not avail-

able since the terrestrial receivers do not typically have sufficient transmit power to feed back the channel knowledge to the satellite transmitter. In other cases, the feedback channel may be available but incur too much delay to be useful, or too much overhead to utilize. Alternatively the transmitter may opt for simplified strategies which do not implement any adaptive transmission based on channel state, and, therefore, CSIT becomes irrelevant.

The first alternative definition we consider is capacity versus outage. Under this definition, without CSIT the transmitter is forced to use a single code, but the decoder may decide whether the information can be reliably decoded based on CSIR. Capacity versus outage assumes that the transmission works well most of the time, but with some maximal probability q, the decoder sees a bad channel and declares an outage; in this case, the transmitted information is lost. The encoding scheme is designed to maximize the capacity for nonoutage states. Capacity versus outage was previously examined in [15] for single-antenna cellular systems, and later became a common performance criterion for multiple-antenna wireless fading channels [17]–[19]. In this work we formalize the operational definition of capacity versus outage and show via a coding theorem and converse that the formula for this capacity is based on the distribution of the normalized information density.

Another method for dealing with channels of variable quality is to allow the receiver to decode *partial* transmitted information. This idea can be illustrated using the broadcast strategy suggested by Cover [20]. The transmitter views the composite channel as a broadcast channel with a collection of virtual receivers indexed by channel realization S. The encoder uses a broadcast code and encodes information as if it were broadcasting to the virtual receivers. The receiver chooses the appropriate decoder for the broadcast code based on the channel W_S in action. The goal is to identify the point in the broadcast rate region that maximizes the *expected rate*, where the expectation is taken with respect to the state distribution $p(S)$ on $\mathcal{S}$. The expected capacity for Gaussian slowly fading channels was first obtained in [21], and these results were later extended to MIMO fading channels in [22]. The formal definition of expected capacity was introduced in our earlier work [23], where upper and lower bounds were also derived for the expected capacity of any composite channel. Details of the proofs together with a numerical example of a composite binary symmetric channel (BSC) appeared in [24]. Application of the broadcast strategy to minimize the end-to-end expected distortion is also considered in [25]–[27].

The alternative capacity definitions are of particular interest for applications where it is desirable to maximize average received rate even if it means that part of the transmitted information is lost and the encoder does not know the exact delivered rate. In this case the receiver either tolerates the information loss or has a mechanism to recover the lost information. Examples include scenarios with some acceptable outage probability, communication systems using multiresolution or multiple description source codes such that partial received information leads to a coarse but still useful source reconstruction at a larger distortion level, feedback channels where the receiver tells the transmitter which symbols to resend, or appli-

cations where lost source symbols are well approximated by surrounding samples. The received rate averaged over multiple transmissions is a meaningful metric when there are two time horizons involved: a short time horizon at the end of which decoding has to be performed because of stringent delay or decoder complexity constraints, and a long time horizon at the end of which the overall throughput is evaluated. For example, consider a wireless service subscriber billed based on usage. Whenever the user requests a voice or data transmission over the network, he usually expects the information to be delivered with small delay, i.e., the short time horizon. However, the service charge is typically calculated on a monthly basis depending on the total or average throughput within the entire period, i.e., the long time horizon.

The remainder of this paper is structured as follows. In Section II, we provide background on channel capacity definitions leading up to the generalized capacity definition of Verdú and Han. Composite channels are defined in Section III. Shannon capacity of composite channels is considered in Section IV, where we show that this capacity only depends on the support set of the component channel distribution. In Section V, we give a formal definition of the capacity versus outage and compare it with the closely-related concept of ϵ-capacity. In Section VI, we introduce the expected capacity and establish a bijection between the expected-rate code and the broadcast channel code. In Section VII, we compare capacity definitions and their implications through two examples: the Gilbert–Elliott channel and the BSC with random crossover probabilities. The implication of these alternative capacity definitions for end-to-end distortion, source-channel coding and separation is briefly discussed in Section VIII. Conclusions are given in Section IX.

II. BACKGROUND

Shannon in [1] defined the channel capacity as the supremum of all achievable rates R for which there exists a sequence of $(2^{nR}, n)$ codes such that the probability of error tends to zero as the blocklength n approaches infinity, and showed the capacity formula (1)

$$C = \max_X I(X; Y)$$

holds for *memoryless* channels. Although the capacity formula (1) is a single-letter expression, the direct channel coding theorem requires coding over long blocklengths to achieve arbitrarily small error probability. The receiver decodes by joint typicality with the typical set defined as [28, pp. 195]

$$A_\epsilon^{(n)} = \Big\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n :$$

$$\left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon$$

$$\left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon$$

$$\left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \Big\} \quad (4)$$

which relies on the law of large numbers to obtain the asymptotic equipartition property (AEP).

For channels with memory, the formula (1) generalizes to the limit (2) when it exists

$$C = \lim_{n \to \infty} \sup_{X^n} \frac{1}{n} I(X^n; Y^n).$$

This formula equals capacity for information stable channels. This class of channels, which includes memoryless channels as a special case, can be roughly described as having the property that the input maximizing the mutual information $I(X^n; Y^n)$ and its corresponding output behave ergodically. In a sense, an ergodic sequence is the most general dependent sequence for which the strong law of large numbers holds [28, p. 474]. The coding theorem of information stable channels follows similarly as that of memoryless channels.

However, the joint typicality decoding technique cannot be generalized to information unstable channels. For general channels, the set $A_\epsilon^{(n)}$ defined in (4) does not have the AEP. As evidence, the probability of $A_\epsilon^{(n)}$ does not approach 1 for large n. Thus, we may not construct channel codes which have small error probability and meanwhile have a rate arbitrarily close to (2). Therefore, the right-hand side of (2), although still a valid upper bound for channel capacity, is not necessarily tight. In [5], Verdú and Han presented a tight upper bound for general channels and showed its achievability through Feinstein's lemma [29]. As discussed in Section IV, achievability can also be shown based on an appropriately-modified notion of typical sets.

This information stable condition can be illustrated using the concept of *information density*.

Definition 1 (Information Density): Given a joint distribution $P_{X^n Y^n}$ on $\mathcal{X}^n \times \mathcal{Y}^n$ with marginal distributions P_{X^n} and P_{Y^n}, the information density is defined as [30]

$$i_{X^n Y^n}(x^n; y^n) = \log \frac{P_{X^n Y^n}(x^n, y^n)}{P_{X^n}(x^n) P_{Y^n}(y^n)}$$
$$= \log \frac{P_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)}. \qquad (5)$$

The distribution of the random variable $(1/n) i_{X^n Y^n}(x^n; y^n)$ is referred to as the *information spectrum* of $P_{X^n Y^n}$. It is observed that the normalized mutual information

$$\frac{1}{n} I(X^n; Y^n) = \sum_{(x^n, y^n)} p(x^n, y^n) \cdot \frac{1}{n} \log \frac{p(y^n|x^n)}{p(y^n)}$$

is the expectation of the normalized information density

$$\frac{1}{n} i(x^n; y^n) = \frac{1}{n} \log \frac{p(y^n|x^n)}{p(y^n)}$$

with respect to the underlying joint input-output distribution $p(x^n, y^n)$, i.e.,

$$\frac{1}{n} I(X^n; Y^n) = \mathbb{E}_{X^n Y^n} \left\{ \frac{1}{n} i_{X^n Y^n}(X^n; Y^n) \right\}.$$

Denote by X_*^n the input distribution that maximizes the mutual information $I(X^n; Y^n)$ and by Y_*^n the corresponding output distribution. The information stable condition [31,
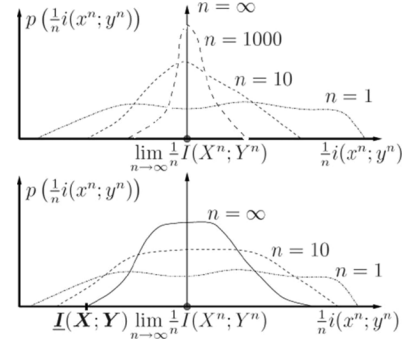

Fig. 1. Empirical distribution of normalized information density.

Definition 3] requires that the normalized information density $(1/n) i(X_*^n; Y_*^n)$, as a random variable, converges in distribution to a constant equal to the normalized mutual information $(1/n) I(X_*^n; Y_*^n)$ as the blocklength n approaches infinity.

In [5], Verdú and Han proved that the largest rate achievable for general channels with asymptotically small error probability is given by the formula (3)

$$C = \sup_{\boldsymbol{X}} \underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y})$$

where $\underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y})$ is the liminf in probability of the normalized information density. In contrast to information stable channels where the distribution of $(1/n) i(X^n; Y^n)$ converges to a single point, for information unstable channels, even with infinite blocklength the *support set*[1] of the distribution of $(1/n) i(X^n; Y^n)$ may still have multiple points or even contain an interval. The Shannon capacity equals the infimum of this support set.

The information spectrum of an information stable channel is demonstrated in the upper plot of Fig. 1. As the blocklength n increases, the convergence of the normalized information density to the channel capacity follows from the weak law of large numbers. In the lower plot of Fig. 1, we show the empirical distribution of $(1/n) i(X^n; Y^n)$ for an information unstable channel. The distribution of the normalized information density does not converge to a single point, so (2) does not equal the capacity, rather it is given by $\underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y})$ in (3).

III. COMPOSITE CHANNELS

We now formally define a composite channel.

Definition 2 (Composite Channel): A channel $\boldsymbol{W}$ is a composite channel if it is statistically modeled as a parameterized sequence of n-dimensional conditional distributions $\boldsymbol{W} = \{W_s^n = P_{Y^n|X^n}; s \in \mathcal{S}\}_{n=1}^\infty$. The channel input-output distribution W_S^n for a given n is determined by the random variable S, independent of the channel input, which is chosen according to some channel state distribution $p_S(s)$ at the beginning of transmission and is hence constant for all n. Thus, for any integer $n > 0$, W_S^n is the conditional distribution from the input space $\mathcal{X}^n$ to the output space $\mathcal{Y}^n$. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the input and output processes, respectively,

---

[1]The smallest closed set of which the complement set has probability measure zero.

for the given sequence of channels. Each process is specified by a sequence of finite-dimensional distributions, e.g., $\boldsymbol{X} = \{X^n = (X_1^{(n)}, \ldots, X_n^{(n)})\}_{n=1}^{\infty}$.

The composite channel model is a useful model (and well-models practical channels) regardless of assumptions about CSIR and CSIR. Indeed, the general capacity formula of Verdú and Han makes no assumptions about CSI at either the transmitter or receiver. Note, however, that without CSIR, capacity analysis must be based on *mismatch capacity*, i.e., capacity when the decoding rule is mismatched to the actual channel (see, e.g., [32] and the references therein). The Verdú–Han generalized capacity is a special case of mismatch capacity, where the decoding rule is based on the worst-case channel. To avoid the pessimistic nature of this capacity formula, our generalized capacity definitions are based on decoding rules matched to the channel, i.e., when CSIR is available. These generalized capacity definitions could be extended to mismatch capacity, in which case exact CSIR would not be needed and, in the limit of no information at all about the channel, capacity would revert to that of the Verdú–Han formula. To consider composite channels with CSIR but no CSIT, we represent the channel side information as an additional output of the channel. Specifically, we let $Z^n = (S, Y^n)$, where S is the channel side information and Y^n is the output of the channel described by parameter S. Throughout, we assume that S is a random variable independent of $\boldsymbol{X}$ and unknown to the encoder. Thus, for each n

$$P_{W^n}(z^n|x^n) = P_{Z^n|X^n}(s, y^n|x^n)$$
$$= P_S(s) P_{Y^n|X^n,S}(y^n|x^n, s)$$

and the information density (5) can be rewritten as

$$i_{X^n W^n}(x^n; z^n) = \log \frac{P_{W^n}(z^n|x^n)}{P_{Z^n}(z^n)}$$
$$= \log \frac{P_{Y^n|X^n,S}(y^n|x^n, s)}{P_{Y^n|S}(y^n|s)}$$
$$= i_{X^n W^n}(x^n; y^n|s). \tag{6}$$

In the following, we see that the generalized capacity definitions of composite channels depend crucially on information density instead of mutual information. We also denote by $F_{\boldsymbol{X}}(\alpha)$ the limit of the cumulative distribution function (cdf) of the normalized information density, i.e.,

$$F_{\boldsymbol{X}}(\alpha) = \lim_{n \to \infty} P_{X^n W^n} \left\{ \frac{1}{n} i_{X^n W^n}(X^n; Y^n|S) \leq \alpha \right\} \tag{7}$$

where the subscript emphasizes the input process $\boldsymbol{X}$.

IV. SHANNON CAPACITY

We now proceed to define the Shannon capacity of composite channels with CSIR.

Definition 3 (Shannon Capacity): Consider a sequence of $(2^{nR}, n)$ codes for channel $\boldsymbol{W}$, where for any $R > 0$, a $(2^{nR}, n)$ code is a collection of 2^{nR} blocklength-n channel codewords and the associated decoding regions. The Shannon capacity is defined as the supremum of all rates R for which there exists a sequence of $(2^{nR}, n)$ codes with vanishing error probability [1].

By this definition, the Shannon capacity $C(\boldsymbol{W})$ measures the rate that can be reliably transmitted from the encoder and also be reliably received at the decoder. We simplify this notation to C if the channel argument is clear from context.

The achievability and converse theorems for the Shannon capacity formula of a general channel, given by

$$C = \sup_{\boldsymbol{X}} \underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Z}) = \sup_{\boldsymbol{X}} \underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y}|S)$$
$$= \sup_{\boldsymbol{X}} \sup \{\alpha : F_{\boldsymbol{X}}(\alpha) = 0\} \tag{8}$$

are proved, respectively, by Theorems 2 and 5 of [5], using Feinstein's lemma [29] and the Verdú–Han lemma [5, Theorem 4]. The special case of a composite channel with CSIR follows immediately from this result. Achievability can also be proved with a modified notion of typical sets [9], [23], as follows (notation is simplified here by removing the explicit conditioning on the side information S):

Encoding: For any input distribution P_{X^n}, $\epsilon > 0$, and $R < \underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y}) - \epsilon$, generate the codebook by choosing $X^n(1), \ldots, X^n(2^{nR})$ i.i.d. according to the distribution $P_{X^n}(x^n)$.

Decoding: For any $\epsilon > 0$, the *typical set* $A_\epsilon^{(n)}$ is defined as

$$A_\epsilon^{(n)} = \left\{ (x^n, y^n) : \frac{1}{n} i_{X^n W^n}(x^n; y^n) \geq \underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y}) - \epsilon \right\}. \tag{9}$$

Channel output Y^n is decoded to $X^n(i)$ where i is the unique index for which $(X^n(i), Y^n) \in A_\epsilon^{(n)}$. An error is declared if more than one or no such index exists.

Error Analysis: We define the following events for all indices $1 \leq i, j \leq 2^{nR}$

$$E_{ji} = \left\{ (X^n(j), Y^n) \in A_\epsilon^{(n)} \, \middle| \, X^n(i) \text{ sent} \right\}. \tag{10}$$

Conditioned on codeword $X^n(i)$ being sent, the probability of the corresponding error event E_i

$$E_i = \bigcup_{j \neq i} E_{ji} \bigcup E_{ii}^c,$$

can be bounded by

$$\Pr(E_i) \leq \Pr(E_{ii}^c) + \sum_{j \neq i} \Pr(E_{ji}).$$

Since we generate i.i.d. codewords, $\Pr(E_{ii})$ and $\Pr(E_{ji}), j \neq i$, do not depend on the specific indices i, j. Assuming equiprobable inputs, the expected probability of error with respect to the randomly generated codebook is

$$P_e^{(n)}$$
$$= \Pr\{\text{error}|X^n(1) \text{ sent}\}$$
$$\leq \Pr(E_{11}^c) + \sum_{j=2}^{2^{nR}} \Pr(E_{j1})$$
$$\leq P_{X^n W^n} \left[\frac{1}{n} i_{X^n W^n}(X^n(1); Y^n) < \underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y}) - \epsilon \right]$$
$$\quad + 2^{nR} \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} P_{X^n}(x^n) P_{Y^n}(y^n)$$
$$\leq \epsilon_n + 2^{n[R - \underline{\boldsymbol{I}}(\boldsymbol{X}; \boldsymbol{Y}) + \epsilon]} \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} P_{X^n W^n}(x^n, y^n), \tag{11}$$

where by definition of $\underline{I}(\boldsymbol{X};\boldsymbol{Y})$ we have ϵ_n approaching 0 for n large enough. The last inequality uses (6), (9), and the fact that $(x^n, y^n) \in A_\epsilon^{(n)}$ implies

$$\frac{1}{n} i_{X^n W^n}(x^n; y^n) = \frac{1}{n} \log \frac{P_{X^n W^n}(x^n, y^n)}{P_{X^n}(x^n) P_{Y^n}(y^n)} \geq \underline{I}(\boldsymbol{X};\boldsymbol{Y}) - \epsilon$$

and consequently

$$P_{X^n}(x^n) P_{Y^n}(y^n) \leq 2^{-n[\underline{I}(\boldsymbol{X};\boldsymbol{Y}) - \epsilon]} P_{X^n W^n}(x^n, y^n).$$

From (11)

$$P_e^{(n)} \leq \epsilon_n + 2^{n[R - \underline{I}(\boldsymbol{X};\boldsymbol{Y}) + \epsilon]} \to 0$$

for all $R < \underline{I}(\boldsymbol{X};\boldsymbol{Y}) - \epsilon$ and arbitrary $\epsilon > 0$, which completes our proof.

Although a composite channel is characterized by the collection of component channels $\{W_s : s \in \mathcal{S}\}$ and the associated probability distribution $p(s)$ on $\mathcal{S}$, the Shannon capacity of a composite channel is solely determined by the support set of the channel state distribution $p(s)$. In the case of a discrete channel state set $\mathcal{S}$, we only need to know which channel states have positive probability. The exact positive value that the probability mass function $p(s)$ assigns to channel states is irrelevant in view of the Shannon capacity. In the case of a continuous channel state set $\mathcal{S}$, we only need to know the subset of channel states where the probability density function is strictly positive. This is formalized in Lemma 1. Before introducing the lemma we need the following definition [33, Appendix 8], which characterizes as *equivalent* probability measures for which the support set of one measure is the same as the support set of the other.

Definition 4 (Equivalent Probability Measures): A probability measure p_2 is absolutely continuous with respect to p_1, written as $p_2 \ll p_1$, if $p_1(A) = 0$ implies that $p_2(A) = 0$ for any event A. Here $p_i(A)$, $i = 1, 2$, is the probability of event A under probability measure p_i. p_1 and p_2 are equivalent probability measures if $p_1 \ll p_2$ and $p_2 \ll p_1$.

Lemma 1: Consider two composite channels $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ with component channels from the same collection $\{W_s : s \in \mathcal{S}\}$. Denote by $p_1(s)$ and $p_2(s)$, respectively, the corresponding channel state distribution of each composite channel. Then $p_2 \ll p_1$ implies $C(\boldsymbol{W}_1) \leq C(\boldsymbol{W}_2)$. Furthermore, if p_1 and p_2 are equivalent probability measures, then $C(\boldsymbol{W}_1) = C(\boldsymbol{W}_2)$.

Intuitively speaking, $p_2 \ll p_1$ if the support set for $\boldsymbol{W}_2$ is a subset of the support set for $\boldsymbol{W}_1$, so any input distribution that allows reliable transmission on $\boldsymbol{W}_1$ also allows reliable transmission on $\boldsymbol{W}_2$. In other words, if $\boldsymbol{W}_1$ has a larger support set than $\boldsymbol{W}_2$ then its capacity (8) cannot be larger. If p_1 and p_2 are equivalent probability measures then they share the same support set, and this guarantees that the corresponding composite channels have the same Shannon capacity. Details of the proof are given in Appendix A.

The equivalent probability measure is a sufficient but not necessary condition for two composite channels to have the same Shannon capacity. For example, consider two slow-fading Gaussian composite channels. It is possible that two probability measures have no support below the same channel gain, but one assigns nonzero probability to states with large capacity while the other does not. In this case, the probability measures are not equivalent; nevertheless the Shannon capacity of both composite channels are the same.

V. CAPACITY VERSUS OUTAGE

The Shannon capacity definition imposes the constraint that all transmitted information be decoded at the receiver with vanishing error probability, while in some real systems it is acceptable to lose a small portion of the transmitted information as long as there is a mechanism to cope with the packet loss. For example, in systems with a receiver complexity constraint, decoding over finite blocklength is necessary but in the case of packet loss, ARQ (automatic repeat request) protocols are implemented where the receiver requests retransmission of the lost information [34], [35]. If the system has a stringent delay constraint, lost information can be approximated from the context, for example the block-coded JPEG image transmission over noisy channels where missing blocks can be reconstructed in the frequency domain by interpolating the discrete cosine transformation (DCT) coefficients of available neighboring blocks [36]. These examples demonstrate a new notion of capacity versus outage: the transmitter sends information at a fixed rate, which is correctly received most of the time; with some maximal probability q, the decoder sees a bad channel and declares an outage, and the transmitted information is lost. This is formalized in the following definition:

Definition 5 (Capacity Versus Outage): Consider a composite channel $\boldsymbol{W}$ with CSIR. A $(2^{nR}, n)$ channel code for $\boldsymbol{W}$ consists of the following:
- an encoding function $X^n : \mathcal{U} = \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X}^n$, where $\mathcal{U}$ is the message index set and $\mathcal{X}$ is the input alphabet;
- an outage identification function $I : \mathcal{S} \to \{0, 1\}$, where $\mathcal{S}$ is the set of channel states;
- a decoding function $g_n : \mathcal{Y}^n \times \mathcal{S} \to \hat{\mathcal{U}} = \{1, 2, \ldots, 2^{nR}\}$, which only operates when $I = 1$.

Define the outage probability

$$P_o^{(n)} = \Pr\{I = 0\}$$

and the error probability in nonoutage states

$$P_e^{(n)} = \Pr\{U \neq \hat{U} | I = 1\}.$$

A rate R is outage-q achievable if there exists a sequence of $(2^{nR}, n)$ channel codes such that $\lim_{n \to \infty} P_o^{(n)} \leq q$ and $\lim_{n \to \infty} P_e^{(n)} = 0$. The capacity versus outage C_q of the channel $\boldsymbol{W}$ with CSIR is defined to be the supremum over all outage-q achievable rates.

In the above definition, $P_o^{(n)}$ is the probability that the decoder, using its side information about the channel, determines it cannot reliably decode the received channel output and declares an outage. In contrast, $P_e^{(n)}$ is the probability that the receiver decodes improperly given that an outage is not declared. The following theorem provides a formula for capacity versus outage as defined by Definition 5.
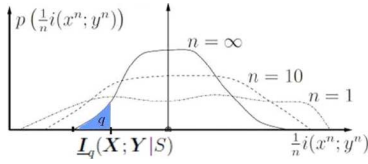
Fig. 2. Capacity versus outage.

Theorem 1 (Capacity Versus Outage): The capacity versus outage of a composite channel with outage probability q is given by

$$C_q = \sup_{\boldsymbol{X}} \boldsymbol{L}_q(\boldsymbol{X}; \boldsymbol{Y} | S)$$
$$= \sup_{\boldsymbol{X}} \sup \left\{ \alpha : F_{\boldsymbol{X}}(\alpha) \leq q \right\}. \quad (12)$$

The achievability proof follows the same typical-set argument given in Section IV. The converse result likewise follows [5]. Details are given in Appendix B. Notice that $C_0 = C$, so the capacity versus outage is a generalization of the Shannon capacity. We give an example of the capacity $C_q = \sup_{\boldsymbol{X}} \boldsymbol{L}_q(\boldsymbol{X}; \boldsymbol{Y}|S)$ versus outage q with respect to its information density in Fig. 2.

The concept of capacity versus outage was initially proposed in [15] for cellular mobile radios. See also [37, Ch. 4] and references therein for more details. A closely-related concept of ϵ-capacity applicable to composite channels was defined in [5] and, in fact, the formula for this ϵ-capacity is identical to the capacity versus outage formula (12). However, despite having the same formula for both, ϵ-capacity and capacity versus outage have fundamentally different definitions and have different assumptions about CSI. Specifically, in the definition of ϵ-capacity the nonzero error probability ϵ accounts for decoding errors *undetected* at the receiver, i.e., $0 < P_e^{(n)} < \epsilon$ does not approach zero with asymptotically large blocklength. In contrast, in the definition of capacity versus outage the receiver declares an outage when the channel state does not allow the receiver to decode with vanishing error probability. Asymptotically, the probability of outage P_o^n is bounded by some fixed constant q and all outages must be *recognized* at the decoder, while in nonoutage states P_e^n approaches zero with asymptotically large blocklength. As a consequence, no decoding is performed for outage states. If the power consumption to perform receiver decoding becomes an issue, as in the case of sensor networks with nonrechargeable nodes or power-conserving mobile devices, then we should distinguish between decoding with error and no decoding at all in view of energy conservation.

The difference in these capacity definitions also has important consequences when we consider end-to-end communication performance using source and channel coding. When the outage states are recognized by the receiver, it can request a retransmission or simply reconstruct the source symbol by its mean – giving an expected distortion equal to the source variance. In contrast, if the receiver cannot recognize the decoding error as in the case of an ϵ-capacity channel code, the reconstruction based on the incorrectly decoded symbol may lead to not only large distortion but also loss of synchronization in the source code's decoder.

We can further define the *outage capacity* $C_q^o = (1 - q)C_q$ as the long-term average rate of an extension of the composite channel to a *quasi-static* composite channel, where the composite channel code is used repeatedly over sufficiently long blocklengths to achieve a rate arbitrarily close to C_q (in nonoutage) with negligible error and, on each use, the channel state is drawn independently according to $p_{\mathcal{S}}(s)$. In this quasi-static composite channel the transmitter uses a single codebook and sends information at rate C_q on each channel use; the receiver can correctly decode the information a proportion $(1 - q)$ of the time and turns itself off a proportion q of the time. The outage capacity C_q^o is a meaningful metric if we are only interested in the fraction of correctly received packets and approximate the unreliable packets by surrounding samples. In this case, optimizing over the outage probability q to maximize C_q^o guarantees performance that is at least as good as the Shannon capacity and may be far better. As another example, if all information must be correctly decoded eventually, the packets that suffer an outage have to be retransmitted. This demands some repetition mechanism that is usually implemented in the link-layer error control of data communication. The number of channel uses K to transmit a packet of size $(N = C_q)$ bits has a geometric distribution

$$\Pr\{K = k\} = q^{k-1}(1 - q)$$

and the expected value is $\frac{1}{(1-q)} = \frac{N}{C_q^o}$, which also illustrates C_q^o as a measure of the long-term average throughput.

Next, we briefly analyze the capacity versus outage from a computational perspective. We need the following definition before we proceed.

Definition 6 (Probability-q Compatible Subchannel): Consider a composite channel $\boldsymbol{W}$ with state distribution $p(s), s \in \mathcal{S}$. Consider another channel $\boldsymbol{W}_q$ where the channel state set $\mathcal{S}_q$ is a subset of $\mathcal{S}$ ($\mathcal{S}_q \subseteq \mathcal{S}$). $\boldsymbol{W}_q$ is a probability-q compatible subchannel of $\boldsymbol{W}$ if $\Pr\{\mathcal{S}_q\} \geq 1 - q$.

Note that $\boldsymbol{W}_q$ is not exactly a composite channel since we only specify the state set $\mathcal{S}_q$ but not the corresponding state distribution over $\mathcal{S}_q$. However, we will only be interested in the Shannon capacity of $\boldsymbol{W}_q$, and as pointed out by Lemma 1, the exact distribution over $\mathcal{S}_q$ is irrelevant to determine this capacity. The capacity versus outage as defined in (12) requires a two-stage optimization. In the first step, we fix the input distribution $\boldsymbol{X}$ and find the probability-q compatible subchannel that yields the highest achievable rate. In the second step, we optimize over the distribution of $\boldsymbol{X}$. This view is more convenient if the optimal input distribution can be easily determined. We then evaluate the achievable rate of each component channel with this optimal input and declare outage for those with the lowest rates. As an example, consider a slow-fading MIMO channel with m transmit antennas. Assume the channel matrix $\boldsymbol{H}$ has i.i.d. Rayleigh fading coefficients. The outage probability associated with transmit rate R is known to be [38]

$$P_o(R) = \inf_{\boldsymbol{Q} \succeq 0, \operatorname{Tr}(\boldsymbol{Q}) \leq m} \Pr \left[\log \det \left(\boldsymbol{I} + \frac{\mathsf{SNR}}{m} \boldsymbol{H} \boldsymbol{Q} \boldsymbol{H}^\dagger \right) \leq R \right]$$

and the capacity versus outage is $C_q = \sup\{R : P_o(R) \leq q\}$. Although the optimal input covariance matrix $\boldsymbol{Q}$ is unknown in

general, it is shown in [19] that there is no loss of generality in assuming $\boldsymbol{Q} = \boldsymbol{I}$ in the high SNR regime and the corresponding capacity versus outage simplifies to

$$C_q = \sup \left\{ R : \Pr \left[\log \det \left(\boldsymbol{I} + \frac{\mathsf{SNR}'}{m} \boldsymbol{H}\boldsymbol{H}^\dagger \right) \leq R \right] \leq q \right\}$$

for SNR's in the high SNR regime. By reversing the order of the two optimization steps, we have another interpretation of capacity versus outage

$$C_q = \sup_{\boldsymbol{W}_q} C(\boldsymbol{W}_q). \tag{13}$$

Here, we first determine the Shannon capacity of each probability-q compatible subchannel, then optimize by choosing the one with the highest Shannon capacity. This view highlights the connection between C_q of a composite channel and the Shannon capacity of its probability-q compatible subchannels, and is more convenient if there is an intrinsic "ordering" of the component channels. For example consider a degraded collection of channels where for any channel states s_1 and s_2 there exists a transition probability $p(y_2^n|y_1^n)$ such that

$$p(y_2^n|x^n, s_2) = \sum_{y_1^n} p(y_1^n|x^n, s_1) p(y_2^n|y_1^n).$$

The degraded relationship can be extended to the *less noisy* and *more capable* conditions [39]. The more capable condition requires

$$I(X^n; Y_1^n|s_1) \geq I(X^n; Y_2^n|s_2) \tag{14}$$

for any input distribution $\boldsymbol{X}$. It is the weakest of all three but suffices to establish an ordering. The optimal probability-q compatible subchannel $\boldsymbol{W}_q^*$ has the smallest set of channel states $\mathcal{S}_q^*$ such that any component channel within $\mathcal{S}_q^*$ is more capable than a component channel not in $\mathcal{S}_q^*$. The Shannon capacity of $\boldsymbol{W}_q^*$ equals the capacity versus outage-q of the original channel $\boldsymbol{W}$.

VI. EXPECTED CAPACITY

The definition of capacity versus outage in Section V is essentially an all-or-nothing game: the receiver may declare outage for undesirable channel states but is otherwise required to decode all transmitted information. There are examples where *partial* received information is useful. Consider sending a multiresolution source code over a composite channel. Decoding all transmitted information leads to reconstruction with the lowest distortion. However, in the case of inferior channel quality, it still helps to decode partial information and get a coarse reconstruction. Although the transmitter sends information at a fixed rate, the notion of expected capacity allows the receiver to decide in expectation how much information can be correctly decoded based on channel realizations.

Next, we introduce some notation which is useful for the formal definition of the expected capacity. Conventionally, we represent information as a message index, cf. the Shannon capacity definition [28, p. 193] and the capacity versus outage definition in Section V. To deal with partial information, here we represent information as a block of bits $(b_i)_{i \in \mathcal{I}}$, where $\mathcal{I}$ is the set of bit indices. Denote by

$$\mathcal{M}(\mathcal{I}) = \{(b_i)_{i \in \mathcal{I}} : b_i \, \text{binary}\}$$

the set of all possible blocks of information bits with bit indices from the set $\mathcal{I}$. Each element in $\mathcal{M}(\mathcal{I})$ is a bit-vector of length $|\mathcal{I}|$, so the size of the set $\mathcal{M}(\mathcal{I})$ is $2^{|\mathcal{I}|}$. If another index set $\widetilde{\mathcal{I}}$ is a proper subset of $\mathcal{I}$ ($\widetilde{\mathcal{I}} \subset \mathcal{I}$), then $\mathcal{M}(\widetilde{\mathcal{I}})$ represents some partial information with respect to the full information $\mathcal{M}(\mathcal{I})$. This representation generalizes the conventional representation using message indices.

Definition 7 (Expected Capacity): Consider a composite channel $\boldsymbol{W}$ with channel state distribution $p(s)$. A $(2^{nR_t}, \{2^{nR_s}\}, n)$ code consists of the following:

- an encoding function

$$f_n : \mathcal{M}(\mathcal{I}_{n,t}) = \{(b_i)_{i \in \mathcal{I}_{n,t}}\} \rightarrow \mathcal{X}^n$$

  where $\mathcal{I}_{n,t} = \{1, 2, \ldots, nR_t\}$ is the index set of the transmitted information bits and $\mathcal{X}$ is the input alphabet;
- a collection of decoders, one for each channel state s,

$$g_{n,s} : \mathcal{Y}^n \times \mathcal{S} \rightarrow \mathcal{M}(\mathcal{I}_{n,s}) = \{(\hat{b}_i)_{i \in \mathcal{I}_{n,s}}\}$$

  where $\mathcal{I}_{n,s} \subseteq \mathcal{I}_{n,t}$ is the set of indices of the decodable information bits in channel state s. $|\mathcal{I}_{n,s}| = nR_s$.

Define the decoding error probability associated with channel state s as

$$P_e^{(n,s)} = \Pr \left\{ \cup_{i \in \mathcal{I}_{n,s}} (\hat{b}_i \neq b_i) \right\}$$

and the average error probability

$$P_e^{(n)} = \mathbb{E}_S P_e^{(n,S)} = \int P_e^{(n,s)} p(s) ds.$$

A rate $R = \mathbb{E}_S R_S$ is achievable in expectation if there exists a sequence of $(2^{nR_t}, \{2^{nR_s}\}, n)$ codes with average error probability $\lim_{n \to \infty} P_e^{(n)} = 0$. Each such rate R is associated with a collection of rates $\{R_s\}_{s \in S}$. The expected capacity $C^e(\boldsymbol{W})$ is the supremum of all rates R achievable in expectation.

We want to emphasize a few subtle points in the above definition. In channel state s, the receiver only decodes those information bits (b_i) with indices $i \in \mathcal{I}_{n,s}$. Decoding error occurs if any of the decoded information bits $(\hat{b}_i)$ is different from the transmitted information bit (b_i). No attempt is made to decode information bits with indices out of the index set $\mathcal{I}_{n,s}$; hence, these information bits are irrelevant to the error analysis for channel state s. The transmitted rate R_t, for which not all bits are decoded with vanishing error probability, has subscript t to differentiate it from the rate R_s, received with vanishing error probability when the channel is in state s.

The cardinality nR_s of the index set $\mathcal{I}_{n,s}$ depends only on the blocklength n and the channel state s. Among the transmitted nR_t information bits, the transmitter and the receiver can agree on the set of decodable information bits for each channel state before transmission starts, i.e., not only the cardinality of $\mathcal{I}_{n,s}$, but the set $\mathcal{I}_{n,s}$ itself is uniquely determined by the channel state s. Nevertheless, for the same channel state s, the receiver may

choose to decode different sets of information bits depending on the actual channel output $\mathcal{Y}^n$, although all these sets are of the same cardinality nR_s. In this case the set of decodable information bits for each channel state is unknown to the transmitter beforehand.

We first look at the case where the transmitter and the receiver agree on the set of decodable information bits for each channel state. In a composite channel, the transmitter can view the channel as a broadcast channel with a collection of virtual receivers indexed by channel realization S. The encoder uses a broadcast code to transmit to the virtual receivers. The receiver uses the side information S to choose the appropriate decoder. Before we proceed to establish a connection between the expected capacity of a composite channel and the capacity region of a broadcast channel, we state the following definition of the broadcast capacity region, which is a direct extension from the two-user case [28, p. 421] to the multiuser case.

Consider a broadcast channel with m receivers. The receivers are indexed by the set $\mathcal{S}$ with cardinality m, which is reminiscent of the index set of channel states in a composite channel. The set $\mathcal{P}(\mathcal{S})$ (or simply $\mathcal{P}$) is the set of all subsets of $\mathcal{S}$. The cardinality of the power set is $|\mathcal{P}(\mathcal{S})| = 2^m$.

Definition 8 (Broadcast Channel Capacity Region): A $(\{2^{nR_p}\}, n)$ code for a broadcast channel consists of the following:

- an encoder

$$f_n : \prod_{p \in \mathcal{P}, \, p \neq \phi} \mathcal{M}_p \to \mathcal{X}^n$$

  where ϕ is the empty set, $p \in \mathcal{P}(\mathcal{S})$ is a nonempty subset of users, and $\mathcal{M}_p = \{1, 2, \ldots, 2^{nR_p}\}$ is the message set intended for users within the subset p only. The short-hand notation $\prod_p \mathcal{M}_p$ denotes the Cartesian product of the corresponding message sets;

- a collection of m decoders, one for each user s

$$g_{n,s} : \mathcal{Y}_s^n \to \prod_{p \in \mathcal{P}, \, s \in p} \hat{\mathcal{M}}_p$$

  where $\mathcal{Y}_s^n$ is the channel output for user s.

Define the error event E_s for each user as

$$E_s = \left\{ g_{n,s}(Y_s^n) = \left(\hat{M}_p\right)_{p \in \mathcal{P}: s \in p} \neq \left(M_p\right)_{p \in \mathcal{P}: s \in p} \right\} \quad (15)$$

and the overall probability of error as

$$P_e^{(n)} = \Pr\left\{ \bigcup_s E_s \right\}.$$

A rate vector $\{R_p\}_{p \in \mathcal{P}}$ is broadcast achievable if there exists a sequence of $(\{2^{nR_p}\}, n)$ codes with $\lim_{n \to \infty} P_e^{(n)} = 0$. The broadcast channel capacity region $\mathcal{C}_{\text{BC}}$ is the convex closure of all broadcast achievable rate vectors.

In the above definition, we explicitly distinguish between private and common information. The message set $\mathcal{M}_p$ contains information decodable by all users $s \in p$ but no others. For instance, in a three-user BC we have private information $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, information for any pair of users $\mathcal{M}_{12}$, $\mathcal{M}_{23}$, $\mathcal{M}_{13}$,

and the common information $\mathcal{M}_{123}$. The total number of message sets is $2^3 - 1$ since the empty set ϕ is excluded.

We now establish a connection between the expected capacity of a composite channel and the capacity region of a broadcast channel through the following theorem. Specifically, the theorem establishes that for composite and broadcast channels related by certain properties, the capacity-achieving code for the broadcast channel achieves the expected capacity of the composite channel, and vice versa. For ease of notation, we state the theorem for a finite number of users (channel states). The result can be generalized to an infinite number of users (continuous channel state alphabets) using the standard technique of [40, Ch. 7], i.e., to first discretize the continuous channel state distribution and then take the limiting case, assuming there are no discontinuities in the limit.

Theorem 2: Consider a composite channel characterized by the joint distribution

$$P_{W^n}(s, y^n | x^n) = P_S(s) P_{Y^n | X^n, S}(y^n | x^n, s)$$

and the corresponding BC with the channel for each receiver satisfying

$$P_{Y_s^n | X^n}(y_s^n | x^n) = P_{Y^n | X^n, S}(y_s^n | x^n, s).$$

Denote by C^e the expected capacity of the composite channel and by $\mathcal{C}_{\text{BC}}$ the capacity region of the corresponding BC, as in Definitions 7 and 8, respectively. If the set of decodable information bits in the composite channel is uniquely determined by the channel state S, then the expected capacity satisfies

$$C^e = \sup_{\{R_p\}_{p \in \mathcal{P}} \in \mathcal{C}_{\text{BC}}} \sum_{p \in \mathcal{P}} R_p \sum_{s \in p} P_S(s) \quad (16)$$

where the supremum on the right side is taken over all vectors in the broadcast channel capacity region $\mathcal{C}_{\text{BC}}$. The proof, whose details are given in Appendix C, establishes a two-way mapping between codes for the corresponding broadcast and composite channels; any $(\{2^{nR_p}\}, n)$ code for the broadcast channel can be mapped to a $(2^{nR_t}, \{2^{nR_s}\}, n)$ expected-rate code for the composite channel and vice versa, where the mapping satisfies $R_s = \sum_{p: s \in p} R_p$ for channel state s. With this mapping, we have

$$C^e = \sup_{\{R_s\}_{s \in S}: \mathbb{E}_S R_S \in C^e} \sum_{s \in S} P_S(s) R_s$$

$$= \sup_{\{R_p\}_{p \in \mathcal{P}} \in \mathcal{C}_{\text{BC}}} \sum_{s \in S} P_S(s) \sum_{p: s \in p} R_p. \quad (17)$$

Although we have introduced a new notion of capacity, the connection established in Theorem 2 shows that the tools developed for broadcast codes can be applied to derive corresponding expected capacity results, with the addition of an optimization to choose the point on the BC rate region boundary that maximizes the expected rate. For example, in [22] some suboptimal approaches, including super-majorization and 1-D approximation, were introduced to analyze the expected capacity of a single-user slowly fading MIMO channel. After the full characterization of the MIMO BC capacity region through the work [13], [41]–[44], the expected capacity of a slowly
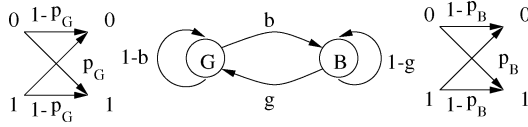
Fig. 3. Gilbert–Elliott channel.

fading MIMO channel can be obtained by choosing the optimal operating point on the boundary of the dirty-paper coding (DPC) region.

The connection in Theorem 2 also shows that any expected-rate code designed for a composite channel can be put into the framework of BC code design. Strategies like layered source coding with progressive transmission, proposed in [45], immediately generalize to the broadcast coding problem. Assuming there are only two channel states s_1 and s_2, this strategy divides the entire transmission block into two segments. The information transmitted in the first segment is intended for both states, and that in the second segment is intended for the better channel state s_2 only. This strategy can be easily mapped to a BC code with individual information $\mathcal{M}_2$ and common information $\mathcal{M}_{12}$, and orthogonal channel access. Furthermore, the complexity of deriving a single point on the BC region boundary is similar to that of deriving the expected capacity under a specific channel state distribution. The entire BC region boundary can be traced out by varying the channel state distributions.

We want to emphasize that in Theorem 2 the condition that the transmitter knows the set of decodable information bits in advance is not superfluous. If the receiver chooses to decode different sets of information bits depending on the actual channel output $\mathcal{Y}^n$, and consequently the transmitter does not know the set of decodable information bits for each state s, then the mapping between expected-rate codes and BC codes may not exist. In the following we give an example where the expected capacity exceeds the supremum of expected rates achievable by BC codes. Consider a binary erasure channel (BEC) where the erasure probability takes two equiprobable values $0 \leq \alpha_1 < \alpha_2 \leq 1$. In Appendix D we show that the maximum expected rate achievable by BC codes is

$$R = \max \left\{ 1 - \alpha_2, \frac{1 - \alpha_1}{2} \right\}. \tag{18}$$

However, we can transmit uncoded information bits directly over this composite BEC. In the limit of large blocklength n, the receiver can successfully decode $n(1 - \alpha_i)$ bits for channel states α_i, $i = 1, 2$, by simply inspecting the channel output, although these successfully decoded information bits cannot be determined at the transmitter *a priori*. Overall the expected capacity

$$C^e = 1 - \frac{\alpha_1 + \alpha_2}{2}$$

exceeds the maximum expected rate achievable by BC codes. Notice, however, these two channel codes are extremely different from an end-to-end coding perspective. The broadcast strategy may be combined with a multiresolution source code. In contrast, the source coding strategy required for the uncoded case is a multiple description source code with single-bit de-

scriptions. Due to this difference, it is not obvious which scenario yields the lower end-to-end distortion. The comparison depends on the channel state distribution and the rate-distortion function of the source.

Regardless of the transmitter's knowledge about decodable information bits, we show that C^e satisfies the lower bound $C^e \geq \sup_q C_q^o$ and the upper bound

$$C^e \leq \sup_{\boldsymbol{X}} \limsup_{n \to \infty} \mathbb{E}_S \mathbb{E}_{X^n Y^n | S} \left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n | S) \,\middle|\, S \right]. \tag{19}$$

The lower bound is achieved using the channel code for capacity versus outage-q, which achieves a rate C_q a proportion $(1 - q)$ of the time and zero otherwise. For the upper bound, we assume channel side information is provided to the transmitter (CSIT) so it can adapt the transmission rate to the channel state. In this case, the achievable expected rate can only be improved and the tightness of the bound depends on the channel characteristics. For example, in fading channels the lack of transmitter side information does not significantly decrease capacity at typical SNRs [37, Chapter 4.2]. However, we will illustrate in the next section that this side information can significantly increase the capacity of the Gilbert–Elliott channel. The proof of the upper bound (19) is given in Appendix E.

VII. EXAMPLES

In this section, we consider some examples to illustrate various capacity definitions.

A. Gilbert–Elliott Channel

The Gilbert–Elliott channel [46] is a two-state channel governed by a Markov chain, where each state is a BSC as shown in Fig. 3. The crossover probabilities for the "good" and "bad" BSCs satisfy $0 \leq p_G < p_B \leq 1/2$. The transition probabilities between the states are g and b, respectively. The initial state distribution is given by π_G and π_B for states G and B. We let $x_n \in \{0, 1\}$, $y_n \in \{0, 1\}$, and $z_n = x_n \oplus y_n$ denote the channel input, output, and error on the nth transmission. We then study capacity definitions when the channel characteristics of stationarity and ergodicity change with the parameters.

Example 1: Ergodic Case, Stationary, or Nonstationary: When $\pi_G = g/(g + b)$ and $\pi_B = b/(g + b)$, the Gilbert–Elliott channel is stationary and ergodic. In this case the information density $\frac{1}{n} i_{X^n W^n}(X^n; Y^n)$ under uniform inputs converges to a δ-function at the average mutual information, so capacity equals average mutual information as usual. Therefore, the Shannon capacity C is equal to the expected capacity $\pi_G C_G + \pi_B C_B$, where $C_G = 1 - h(p_G)$, $C_B = 1 - h(p_B)$ and $h(p) = -p \log p - (1 - p) \log(1 - p)$ is the binary entropy function.

This is a single-state composite channel; hence, the composite channel CSIR is constant and known, although the state of the Gilbert–Elliott channel varies. Since any transmission may experience either a good or a bad channel condition, and the receiver does not know the state of the Gilbert–Elliott channel, it has no basis for choosing to declare an outage on certain transmissions and not on others. Thus, capacity versus outage equals Shannon capacity in this case.
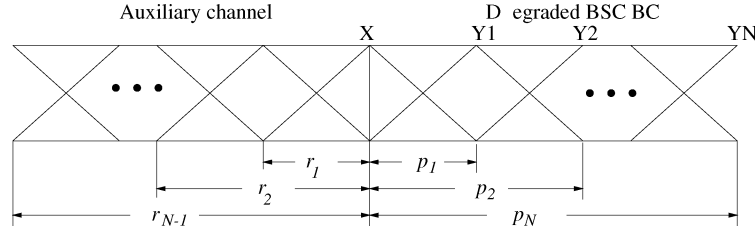
Fig. 4. BSC broadcast channel with auxiliary channels for random coding.

If $\pi_G \neq g/(g+b)$ but b and g are nonzero, then the Gilbert–Elliott channel is ergodic but not stationary. However, the distribution on the states G and B converges to a stationary distribution under uniform inputs. Thus, the channel is asymptotically mean stationary, and the definitions of capacity have the same values as in the stationary case.

Example 2: Stationary and Nonergodic Case: We now set $g = b = 0$. So the initial channel state is chosen according to probabilities $\{\pi_G, \pi_B\}$ and then remains fixed for all time. This is a two-state composite channel and we assume the CSIR is known at the receiver. The Shannon capacity equals that of the bad channel ($C = C_B$). The capacity versus outage-q $C_q = C_B$ if the outage probability $q < \pi_B$ and $C_q = C_G$ otherwise. The loss incurred from lack of side information at the encoder is that the expected capacity is strictly less than the average of individual capacities $\pi_B C_B + \pi_G C_G$ and is equal to [20]

$$\max_{0 \leq r \leq 1/2} 1 - h(r * p_B) + \pi_G[h(r * p_G) - h(p_G)] \quad (20)$$

where $\alpha * \beta = \alpha(1-\beta) + (1-\alpha)\beta$. The interpretation here is that the broadcast code achieves rate $1 - h(r * p_B)$ for the bad channel and an additional rate $h(r * p_G) - h(p_G)$ for the good channel, so the average rate is the expected capacity. Using the Lagrangian multiplier method, we can obtain r^* which maximizes (20). Namely, if we define

$$k = \frac{\pi_G}{\pi_B}, \quad A = \frac{1 - 2p_B}{1 - 2p_G}, \quad f(p_1, p_2) = \frac{\log(1/p_1 - 1)}{\log(1/p_2 - 1)}$$

then $r^* = 0$ if $k \leq Af(p_B, p_G)$; $r^* = 1/2$ if $k \geq A^2$ and r^* solves $f(r * p_G, r * p_B) = A/k$ otherwise. With this solution we can determine the looseness of the upper bound (19) for this particular channel model. Specifically, for $\pi_G = \pi_B = 0.5$, $p_B = .5$ and $p_G = .1$, we get an expected capacity of .0974. With transmitter side information this becomes .2655, more than 2.5 times larger than the expected capacity without the encoder knowledge of the channel state.

B. BSC With Random Crossover Probabilities

In the nonergodic case, the Gilbert–Elliott Channel is a two-state channel, where each state corresponds to a BSC with a different crossover probability. We now generalize that example to allow more than two states. We consider a BSC with random crossover probability $0 \leq p \leq 1/2$. At the beginning of time, p is chosen according to some distribution $f(p)$ and then held fixed. We also use $F(p) = \int_0^p f(s)ds$ to denote the cumulative distribution function. Like the nonergodic Gilbert–Elliott channel, this is a multistate composite channel provided $\{p :$

$f(p) > 0\}$ has cardinality at least two. The Shannon capacity is $C = 1 - h(p^*)$ where

$$p^* = \sup\{p : f(p) > 0\} = \inf\{p : F(p) = 1\}$$

and the capacity versus outage-q is $C_q = 1 - h(p_q)$ where $p_q = \inf\{p : F(p) \geq 1 - q\}$.

We consider a broadcast approach on this channel to achieve the expected capacity, assuming CSIR, i.e., knowledge of p, at the receiver. The receiver is equivalent to a continuum of ordered users, each indexed by the BSC crossover probability p and occurring with probability $f(p)dp$. If the set $\{p : f(p) > 0\}$ is infinite, then the transmitter sends an infinite number of layers of coded information and each user decodes an incremental rate $|dR(p)|$ corresponding to its own layer. Since the BSC broadcast channel is degraded, a user with crossover probability p can also decode layers indexed by larger crossover probabilities; therefore, we achieve a rate of

$$R(p) = -\int_p^{1/2} dR(p) \quad (21)$$

for a receiver associated with crossover probability p. The problem of determining the expected capacity then boils down to the characterization of the broadcast rate region and the choice of the point on that region that maximizes $\int_p R(p)f(p)dp$.

In the discrete case with N users, assuming $0 \leq p_1 \leq \cdots \leq p_N \leq 1/2$, the capacity region is shown to be [47]

$$\{\boldsymbol{R} = (R_i)_{1 \leq i \leq N} : R_i = R(p_i) = h(r_i * p_i) - h(r_{i-1} * p_i)\} \quad (22)$$

where $0 = r_0 \leq r_1 \leq \cdots \leq r_N = 1/2$. Since the original broadcast channel is stochastically degraded it has the same capacity region as a cascade of N BSC's. The capacity region boundary is traced out by augmenting $(N-1)$ auxiliary channels [47] and varying the crossover probabilities of each. For each i, r_i equals the overall crossover probability for auxiliary channels 1 up to i. See Fig. 4 for an illustration. The resulting expected capacity is

$$C^e = \max_{0 = r_0 \leq \cdots \leq r_N = 1/2} \sum_{i=1}^N f(p_i) \sum_{j=i}^N [h(r_i * p_i) - h(r_{i-1} * p_i)].$$

We extend the above result to the continuous case with an infinite number of auxiliary channels. In this case, we define a monotonically increasing function $r(p)$ equal to the overall crossover probability of auxiliary channels up to that indexed

by p. In the following, we use $r(p)$ and r_p interchangeably. For the layer indexed by p, the incremental rate is

$$-dR(p) = h(p * r_p) - h(p * r_{p-dp}).$$

Using the first order approximation $r_{p-dp} \approx r_p - r'_p dp$ and $h(x - \delta) \approx h(x) - h'(x)\delta$ for small δ, we obtain

$$\begin{aligned}
-dR(p) &= h(p * r_p) - h(p * r_{p-dp}) \\
&\approx h(p * r_p) - h(p * r_p - (1 - 2p)r'_p dp) \\
&\approx \log\left(\frac{1}{p * r_p} - 1\right)(1 - 2p)r'_p dp.
\end{aligned}$$

Note here $\delta = (1 - 2p)r'_p dp$ is a small variation, and we do not explicitly address the problematic limiting case $h'(x) \to \infty$ as x approaches zero.[2]

Overall the expected rate is

$$\begin{aligned}
C^e &= \int_0^{1/2} f(p)R(p)dp = -\int_0^{1/2} F(p)dR(p) \\
&\approx \int_0^{1/2} F(p)\log\left(\frac{1}{p * r_p} - 1\right)(1 - 2p)r'_p dp. \quad (23)
\end{aligned}$$

The optimal $r(p)$ maximizing the expected rate can be solved through calculus of functional variation. Define $S(p, r_p, r'_p)$ as

$$S(p, r_p, r'_p) = F(p)\log\left(\frac{1}{p * r_p} - 1\right)(1 - 2p)r'_p. \quad (24)$$

The optimal $r(p)$ should satisfy the Euler equation [48]

$$S_r - \frac{d}{dp}S_{r'} = 0 \quad (25)$$

where

$$\begin{aligned}
S_r &= \frac{\partial S}{\partial r} = -\frac{(1 - 2p)^2 F(p)r'_p}{p * r_p - (p * r_p)^2} \\
S_{r'} &= \frac{\partial S}{\partial r'} = (1 - 2p)F(p)\log\left[\frac{1 - p * r_p}{p * r_p}\right] \\
\frac{dS_{r'}}{dp} &= [(1 - 2p)f(p) - 2F(p)]\log\left[\frac{1 - p * r_p}{p * r_p}\right] \\
&\quad - \frac{(1 - 2p)F(p)}{p * r_p - (p * r_p)^2}[1 - 2r_p + (1 - 2p)r'_p].
\end{aligned}$$

After some algebra (25) simplifies to

$$\frac{(p * r_p)^{-1} - (1 - p * r_p)^{-1}}{\log(1 - p * r_p) - \log(p * r_p)} = \frac{(1 - 2p)f(p) - 2F(p)}{F(p)}. \quad (26)$$

In general, (26) has no closed-form solution but there exist obvious numerical approaches.

As an example, suppose that the crossover probability is uniformly distributed on $[0, 1/2]$. The Shannon capacity is limited

Fig. 5. Capacity under different definitions of BSC with random crossover probability.

by the worst channel state ($p = 1/2$), giving $C = 0$. The capacity versus outage-q is $C_q = \left[1 - h\left(\frac{1-q}{2}\right)\right]$. To approximate the expected capacity, we solve for $r(p)$ in (26) for each p. It is seen that $0 \le r_p \le 1/2$ only for $p_l \le p \le p_u$, where the two cutoff probabilities satisfy $r(p_l) = 0$ and $r(p_u) = 1/2$. For the uniform distribution case, $p_l = 0.136$ and $p_u = 1/6$, which demonstrates that it is unnecessary to use the channel all the time to achieve the expected capacity. In fact, no information is sent for $p \ge 1/6$.

In Fig. 5 we plot the expected capacity, the outage-q capacity, and the capacity versus outage-q. Although the capacity versus outage-q exceeds the expected capacity C^e for some values of q, the outage-q capacity C_q^o is always dominated by the expected capacity C^e, since an outage-q code is one of many possible codes for the expected capacity. Define cutoff outage probabilities $q_l = 1 - 2p_l$ and $q_u = 1 - 2p_u$. Note that $C_q^o \approx C^e$ for all $q \in [q_l, q_u]$. In this range, an outage code gives almost the same expected rate as a broadcast code.

In Fig. 6, we plot the rate used in each state by the expected capacity code and the capacity versus outage codes at outage probabilities q_l, q_u and $1/2$. We see that the code for outage capacity achieves a constant rate for nonoutage states and a rate 0, otherwise. For this example, the incremental rates $|dR(p)|$ are nonzero only for $p_l \le p \le p_u$. Therefore, the code for expected capacity achieves a rate 0 when $p > p_u$. As p decreases from p_u to p_l, the rate gradually increases from 0 to 0.38 bits per channel use, and stays at this constant level for $p < p_l$. Since all channels are equally probable, the area under each curve is the expected rate of that strategy. The area under the expected capacity curve is the largest. The expected capacity curve is, in some places, lower than the curve for outage-q_l capacity. Although the outage-q_l code achieves a rate higher than the broadcast code for expected capacity when $p < p_l$, the same code has decoding rate 0 for all other channel states $p > p_l$, giving a lower area under the total curve.

A potential advantage of the outage code is its simplicity. The transmission rate is fixed, so the code may be coupled with a

---

[2]The achievable rate $R(p)$ for any state is bounded by one; therefore, $\int_{\epsilon}^{1/2} f(p)R(p)dp$, as a function of ϵ, is right continuous at $\epsilon = 0$. We can avoid the problematic limiting case by focusing on strictly positive ϵ and obtain the expected capacity (23) by continuity.
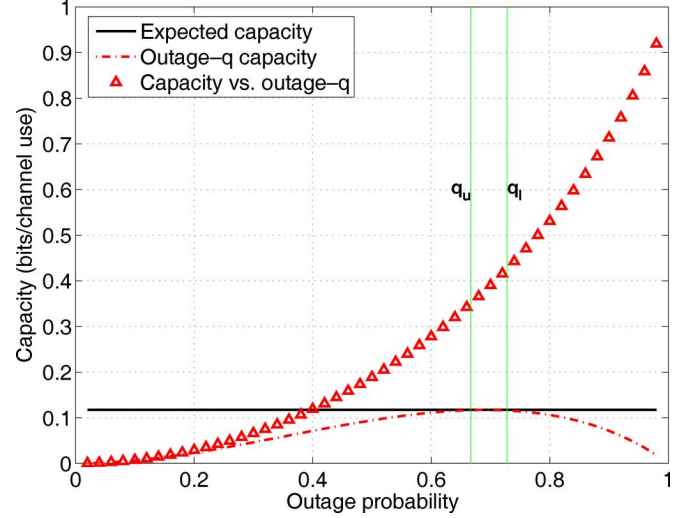
Fig. 6. Achievable rate for each channel state.

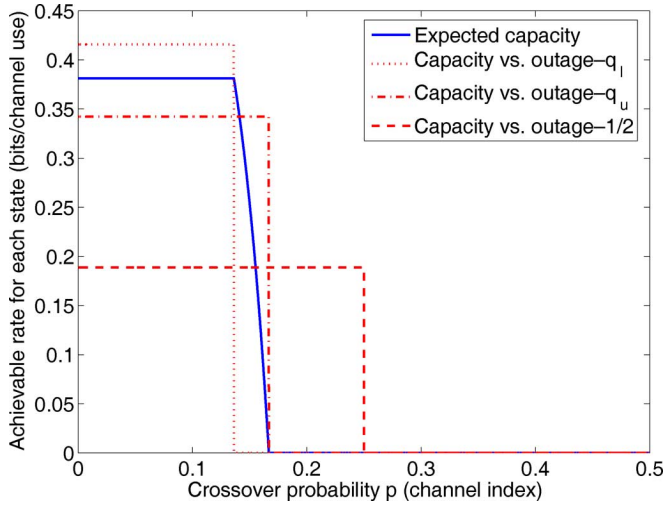Fig. 7. Effect of cutoff range.

conventional source code. The advantage of the expected capacity code is its higher expected rate. The code may be coupled with a multiresolution source code. It is not obvious which strategy yields better end-to-end coding performance in this example. In general, an expected rate code is required to achieve the optimal end-to-end distortion, but this code may use a rate vector on the boundary of the BC capacity region which is different from the rate vector used by the code that achieves the expected capacity [26]. The procedure to solve for the expected capacity is computationally intensive. In the above example, when looking for the optimal $r(p)$ which leads to the expected capacity, we first identify the cutoff probabilities (p_l, p_u) and then solve (26) for each p in this range. We want to emphasize that the correct cutoff range, although seemingly a very coarse characterization of the optimal solution, is crucial to the expected rate. Consider the following closed-form suboptimal solutions for $r(p)$ based on approximating the solution to (26).

- Optimal cutoff $[p_l, p_u]$ with suboptimal $r(p)$

$$r(p) = \begin{cases} \frac{(p-p_l)^\gamma}{2(p_u-p_l)^\gamma}, & p_l \le p \le p_u \\ 0, & \text{otherwise.} \end{cases} \quad (27)$$

- Cutoff range [0, 1/2]

$$r(p) = (1/2)(2p)^\gamma. \quad (28)$$

The first solution assumes the optimal cutoff range has been determined and approximates $r(p)$ as a function of this range. The second solution does not assume the optimal cutoff range is known. The choice of γ in both cases makes $r(p)$ convex $(\gamma > 1)$, linear $(\gamma = 1)$ or concave $(\gamma < 1)$ in both approaches. In Fig. 7, for γ ranging between zero and four, we plot the achievable expected rate using the optimal cutoff range $[p_l, p_u]$ and suboptimal $r(p)$ from (27), the expected rate using the cutoff range [0, 1/2] and suboptimal $r(p)$ from (28), and the expected capacity of this channel based on the optimal $r(p)$, which upper bounds these other plots. We observe that the optimal cutoff range with $r(p)$ heuristic (27) yields an expected rate very close to C^e, but the expected rate is clearly suboptimal if we use the heuristic (28) based on the generic cutoff range [0, 1/2]. While it may be possible to find better heuristics under both assumptions about the cutoff range to close the
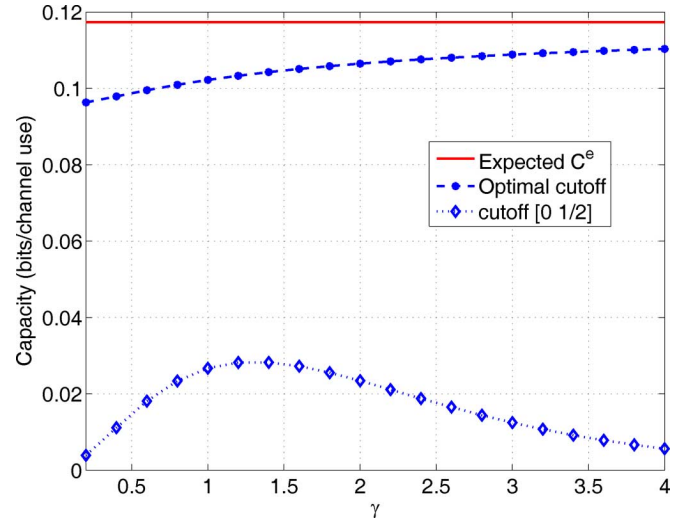
gap with the expected capacity, we see that the simple heuristic (27) based on optimizing the cutoff range captures most benefit of the expected-rate code as compared to the conventional code for Shannon capacity.

VIII. SOURCE-CHANNEL CODING AND SEPARATION

Channel capacity theorems deal with data transmission in a communication system. When extending the system to include the source of the data, we also need to consider the data compression problem. For the overall system, the end-to-end distortion is a well-accepted performance metric. When both the source and channel are stationary and ergodic, codes are usually designed to achieve the same end-to-end distortion level for any source sequence and channel realization. However, if the channel model is generalized to such scenarios as the composite channel above, it is natural to introduce generalized end-to-end distortion metrics such as the distortion versus outage and the *expected distortion* [49], similar to the development of alternative capacity definitions. These alternative distortion metrics are also considered in prior works [25]–[27], [50]–[53].

The renowned source-channel separation theorem [6, Theorem 2.4] asserts that a target distortion level D is achievable if and only if the channel capacity C exceeds the source rate distortion function $R(D)$, and a two-stage separate source-channel code suffices to meet the requirement.[3] This theorem enables separate design of source and channel codes and guarantees the optimal performance. However, there are a few underlying assumptions: a single-user channel; a stationary ergodic source and channel; a single distortion level maintained for all transmission. It is known that the separation theorem fails if the first two assumptions do not hold [31], [54]. In fact, the end-to-end distortion metrics also dictate whether the source-channel separation holds for a communication system. In [49], we showed the direct part of source-channel separation under the distortion versus outage metric and established the converse for certain systems. On the contrary, source-channel separation does not hold under the expected distortion metric.

---

[3]The separation theorem for lossless transmission [1] can be regarded as a special case of zero distortion.

Source-channel separation implies that the operation of source and channel coding does not depend on the statistics of the counterpart. Meanwhile, the source and channel do need to communicate with each other through an *interface*, which is a single number in the classical separation theorem. For generalized source/channel models and distortion metrics, the interface is not necessarily a single rate and may allow multiple parameters to be agreed on between the source and channel encoders and decoders. For example, under capacity versus outage, the source encoder might require knowledge of both the transmission rate and the outage probability to be robust against that percentage of lost data. If the source encoder also incorporated a rate versus outage metric, then both of these parameters should be made known to the channel encoder so that, for example, it could optimize its average power based on no transmission during source outages. Under expected capacity, the transmission strategy might utilize a multiresolution channel code that can be matched to a multiresolution source code to minimize expected end-to-end distortion. In this case, the source code must know the rates associated with each component channel in the composite channel. Since we expect a performance enhancement when source and channel exchange more information through more sophisticated interface, an interesting topic for future research would be to characterize the tradeoff between interface complexity and the achievable end-to-end performance [55].

IX. CONCLUSION

In view of the pessimistic nature of Shannon capacity for composite channels with CSIR, we propose alternative capacity definitions including capacity versus outage and expected capacity. These definitions lend insight to applications where side information at the receiver combined with appropriate source coding strategies can exploit these more flexible notions of capacity. We prove capacity theorems or bounds under each definition, and illustrate how expected achievable rates can be improved through examples of Gilbert–Elliot channels and a BSC with random crossover probabilities. While the use of capacity definitions inherently focuses our attention on achievable (expected) rates, we note that the existence of other meaningful measures of performance in the given coding environment. For example, since outage-q codes are compatible with conventional source codes while expected capacity codes require multiresolution or multiple description codes, depending on whether or not the corresponding broadcast channel is degraded, the fact that the expected rate of the expected capacity code exceeds that of the outage-q code does not guarantee lower end-to-end expected distortion. Furthermore, since a nonergodic channel experiences a single ergodic mode for all time, there is some justification for performance measures that take the probability of suffering a very low-rate state into account. These topics provide a wealth of interesting questions for future research with some initial work presented in [25], [26], and [49].

APPENDIX A
PROOF OF LEMMA 1

We prove $C(\mathbf{W}_1) \leq C(\mathbf{W}_2)$ if $p_2 \ll p_1$, and vice versa. Therefore, equivalent probability measures of p_1 and p_2 imply

identical Shannon capacity. The result is intuitive but we need to address a subtle technical issue: note that p_1 and p_2 are channel state distributions, while the Shannon capacity is defined through the information density distribution (7), which depends on both input and channel statistics.

Recall the Shannon capacity formula (8)

$$C(\mathbf{W}_1) = \sup_{\mathbf{X}} \sup\{\alpha : F_{\mathbf{X}}(\alpha) = 0\}.$$

Denote by $\mathbf{X}_*$ the input distribution that achieves the supremum in (8), and by $F_1(\alpha)$ the corresponding information density distribution. For arbitrary $\epsilon > 0$, we define

$$M_\epsilon(\alpha) = \left\{ s : \lim_{n\to\infty} P_{X_*^n Y^n|S}\left[\frac{1}{n} i_{X_*^n Y^n|S}(X^n; Y^n|s) \leq \alpha\right] \geq \epsilon \right\}.$$

Notice that

$$F_1(\alpha)$$
$$= \lim_{n\to\infty} P_{X_*^n W_1^n}\left\{\frac{1}{n} i_{X_*^n W_1^n}(X^n; Y^n|S) \leq \alpha\right\}$$
$$= \lim_{n\to\infty} \int P_{X_*^n Y^n|S}\left\{\frac{1}{n} i_{X_*^n Y^n|S}(X^n; Y^n|s) \leq \alpha\right\} \cdot p_1(s) ds$$
$$= \int \lim_{n\to\infty} P_{X_*^n Y^n|S}\left\{\frac{1}{n} i_{X_*^n Y^n|S}(X^n; Y^n|s) \leq \alpha\right\} \cdot p_1(s) ds$$
$$\geq \epsilon \int_{M_\epsilon(\alpha)} p_1(s) ds \qquad (29)$$

where we exchange the order of integral and limit according to dominated convergence theorem. From (29), we see that $F_1(\alpha) = 0$ implies

$$\int_{M_\epsilon(\alpha)} p_1(s) ds = 0.$$

Assuming $p_2 \ll p_1$, it follows that

$$\int_{M_\epsilon(\alpha)} p_2(s) ds = 0.$$

Now assuming $F_1(\alpha) = 0$, define $F_2(\alpha)$ as the information density distribution of channel $\mathbf{W}_2$ when evaluated at input $\mathbf{X}_*$, i.e., see the equation shown at the bottom of the next page. Since ϵ is arbitrary, we see that $F_1(\alpha) = 0$ implies $F_2(\alpha) = 0$, therefore

$$C(\mathbf{W}_1) = \sup\{\alpha : F_1(\alpha) = 0\}$$
$$\leq \sup\{\alpha : F_2(\alpha) = 0\}$$
$$\leq C(\mathbf{W}_2).$$

APPENDIX B
PROOF OF CAPACITY VERSUS OUTAGE THEOREM (12)

We first prove the achievability of the capacity versus outage theorem (12). Consider a fixed outage probability $q \geq 0$.

1) Encoding: For any input distribution P_{X^n}, $\epsilon > 0$, and $R < \underline{\mathbf{L}}_q(\mathbf{X}; \mathbf{Y}) - \epsilon$, generate the codebook by choosing $X^n(1), \ldots, X^n(2^{nR})$ i.i.d. according to the distribution $P_{X^n}(x^n)$.

2) Decoding: Define, for $\epsilon > 0$, the *typical set* $A_\epsilon^{(n)}$ as

$$A_\epsilon^{(n)} = \left\{ (x^n, y^n) : \frac{1}{n} i_{X^n W^n}(x^n; y^n) \geq \underline{\mathbf{L}}_q(\mathbf{X}; \mathbf{Y}) - \epsilon \right\}.$$

For any channel output Y^n, we decode as follows:
1) if $(X^n(i), Y^n) \notin A_\epsilon^{(n)}$ for all $i \in \{1, \ldots, 2^{nR}\}$, declare an outage;
2) otherwise, decode to the unique index $i \in \{1, \ldots, 2^{nR}\}$ such that $(X^n(i), Y^n) \in A_\epsilon^{(n)}$. An error is declared if more than one such index exists.

3) Outage and Error Analysis: We recall the definition of events E_{ji} in (10) as

$$E_{ji} = \left\{ (X^n(j), Y^n) \in A_\epsilon^{(n)} \,\middle|\, X^n(i) \text{ sent} \right\}.$$

Assuming equiprobable inputs, the expected probability of an outage using the above scheme is

$$
\begin{aligned}
P_o^{(n)} &= \Pr\left\{ \text{outage} | X^n(1) \text{ sent} \right\} \\
&= \Pr\left\{ \cap_{i=1}^{2^{nR}} E_{i1}^c \right\} \\
&\leq \Pr\left\{ E_{11}^c \right\} \\
&= P_{X^n W^n}\left\{ \frac{1}{n} i_{X^n W^n}(X^n(1); Y^n) < \underline{\boldsymbol{L}}_q(\boldsymbol{X}; \boldsymbol{Y}) - \epsilon \right\} \\
&\leq q + \epsilon_n
\end{aligned}
$$

where by definition of $\underline{\boldsymbol{L}}_q(\boldsymbol{X}; \boldsymbol{Y})$ we have ϵ_n approaching zero for n large enough. Likewise, when no outage is declared the expected probability of error is

$$
\begin{aligned}
P_e^{(n)} &= \Pr\left\{ \text{error} | X^n(1) \text{ sent and no outage declared} \right\} \\
&= \Pr\left\{ \bigcup_{i=2}^{2^{nR}} E_{i1} \right\} \\
&\leq 2^{nR} \Pr\left\{ E_{21} \right\} \\
&= 2^{nR} \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} P_{X^n}(x^n) P_{Y^n}(y^n) \\
&\leq 2^{n[R - \underline{\boldsymbol{L}}_q(\boldsymbol{X}; \boldsymbol{Y}) + \epsilon]} \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} P_{X^n W^n}(x^n, y^n) \quad (30)
\end{aligned}
$$

where the last inequality is obtained by noticing that $(x^n, y^n) \in A_\epsilon^{(n)}$ implies

$$\frac{1}{n} i_{X^n W^n}(x^n; y^n) = \frac{1}{n} \log \frac{P_{X^n W^n}(x^n, y^n)}{P_{X^n}(x^n) P_{Y^n}(y^n)} \geq \underline{\boldsymbol{L}}_q(\boldsymbol{X}; \boldsymbol{Y}) - \epsilon$$

or equivalently

$$P_{X^n}(x^n) P_{Y^n}(y^n) \leq 2^{-n[\underline{\boldsymbol{L}}_q(\boldsymbol{X}; \boldsymbol{Y}) - \epsilon]} P_{X^n W^n}(x^n, y^n).$$

From (30) we see that $P_e^{(n)} \to 0$ for all $R < \underline{\boldsymbol{L}}_q(\boldsymbol{X}; \boldsymbol{Y}) - \epsilon$ and arbitrary $\epsilon > 0$, which completes our proof.

Next, we prove the converse of the capacity versus outage theorem (12). Consider any sequence of $(n, 2^{nR})$ codes with error probability $P_e^{(n)} \to 0$ and outage probability $\lim_{n \to \infty} P_o^{(n)} \leq q$. Let $\{X^n(1), \ldots, X^n(2^{nR})\}$ represent the nth code in the sequence, and assume a uniform input distribution

$$P_{X^n}(x^n) = \begin{cases} 2^{-nR}, & \forall\, x^n \in \{X^n(1), \ldots, X^n(2^{nR})\} \\ 0, & \text{otherwise.} \end{cases}$$

For each $i \in \{1, \ldots, 2^{nR}\}$, let D_i represent the decoding region associated with codeword $X^n(i)$ and B_i equal an analogy of the typical set, defined as

$$
\begin{aligned}
B_i &= \left\{ y^n \in \mathcal{Y}^n : \frac{1}{n} i_{X^n W^n}(X^n(i), y^n) \leq R - \gamma \right\} \\
&= \left\{ y^n \in \mathcal{Y}^n : \frac{1}{n} \log \frac{P_{X^n | Y^n}(X^n(i) | y^n)}{2^{-nR}} \leq R - \gamma \right\} \\
&= \left\{ y^n \in \mathcal{Y}^n : P_{X^n | Y^n}(X^n(i) | y^n) \leq 2^{-\gamma n} \right\}
\end{aligned}
$$

where $\gamma > 0$ is arbitrary. Then we have

$$
\begin{aligned}
&P_{X^n W^n}\left\{ \frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq R - \gamma \right\} \\
&= \sum_{i=1}^{2^{nR}} P_{X^n W^n}(X^n(i), B_i) \\
&= \sum_{i=1}^{2^{nR}} [P_{X^n W^n}(X^n(i), B_i \cap D_i) \\
&\qquad + P_{X^n W^n}(X^n(i), B_i \cap D_i^c)] \\
&\leq \sum_{i=1}^{2^{nR}} \sum_{y^n \in B_i \cap D_i} P_{X^n W^n}(X^n(i), y^n) + P_e^{(n)} + P_o^{(n)} \\
&\leq \sum_{i=1}^{2^{nR}} \sum_{y^n \in D_i} P_{Y^n}(y^n) 2^{-\gamma n} + P_e^{(n)} + P_o^{(n)} \\
&\leq 2^{-\gamma n} + P_e^{(n)} + P_o^{(n)}
\end{aligned}
$$

$$
\begin{aligned}
F_2(\alpha) &= \lim_{n \to \infty} P_{X_*^n W_2^n}\left\{ \frac{1}{n} i_{X_*^n W_2^n}(X^n; Y^n | S) \leq \alpha \right\} \\
&= \int_{\mathcal{S} - M_\epsilon(\alpha)} \lim_{n \to \infty} P_{X_*^n Y^n | S}\left\{ \frac{1}{n} i_{X_*^n Y^n | S}(X^n; Y^n | s) \leq \alpha \right\} \cdot p_2(s) ds \\
&\quad + \int_{M_\epsilon(\alpha)} \lim_{n \to \infty} P_{X_*^n Y^n | S}\left\{ \frac{1}{n} i_{X_*^n Y^n | S}(X^n; Y^n | s) \leq \alpha \right\} \cdot p_2(s) ds \\
&\leq \epsilon \int_{\mathcal{S} - M_\epsilon(\alpha)} p_2(s) ds + \int_{M_\epsilon(\alpha)} p_2(s) ds \\
&\leq \epsilon.
\end{aligned}
$$

since the decoding regions D_i cannot overlap. Thus

$$P_e^{(n)} \geq P_{X^n W^n} \left\{ \frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq R - \gamma \right\} - P_o^{(n)} - 2^{-\gamma n}$$

which goes to zero if and only if $R - \gamma \leq \underline{L}_q(\boldsymbol{X}; \boldsymbol{Y})$ by definition of $\underline{L}_q(\boldsymbol{X}; \boldsymbol{Y})$.

APPENDIX C
PROOF OF THEOREM 2

1) Mapping Broadcast Code to Expected-Rate Code: We first show that any broadcast code can be mapped to an expected-rate code, so

$$C^e \geq \sum_{p \in \mathcal{P}} R_p \sum_{s \in p} P_S(s) \tag{31}$$

for any $\{R_p\} \in \mathcal{C}_{\mathrm{BC}}$.

Given a $(\{2^{nR_p}\}, n)$ BC code as defined in Definition 8, we represent each message $M_p \in \mathcal{M}_p$ in a binary format consisting of nR_p bits and concatenate these bits to form an overall representation of nR_t bits, where

$$R_t = \sum_{p \in \mathcal{P}, p \neq \phi} R_p. \tag{32}$$

These nR_t information bits are indexed by the index set $\mathcal{I}_{n,t} = \{1, 2, \ldots, nR_t\}$. We denote by $\mathcal{I}_{n,p}$ the set of indices of the nR_p bits that correspond to the message set $\mathcal{M}_p$ in the BC code. Note that $\mathcal{I}_{n,p}$ may be empty for some $p \in \mathcal{P}$, for different p these index sets are mutually exclusive and

$$\mathcal{I}_{n,t} = \bigcup_{p \in \mathcal{P}, p \neq \phi} \mathcal{I}_{n,p}. \tag{33}$$

The $(\{2^{nR_p}\}, n)$ BC code can be mapped to the following expected-rate code with transmit rate R_t given by (32). For any $M_t \in \mathcal{M}(\mathcal{I}_{n,t})$, the bits (b_i) with $i \in \mathcal{I}_{n,p} \subseteq \mathcal{I}_{n,t}$ define a corresponding message M_p in the message set $\mathcal{M}_p$ of the BC code. The encoder for the expected rate code satisfies

$$f_n^e(M_t) = f_n^{\mathrm{BC}} \left(\prod_{p \in \mathcal{P}, p \neq \phi} M_p \right)$$

where the superscript e and BC distinguishes the encoder of the expected-rate code and the broadcast code. For a state s in the composite channel, the receiver decodes those information bits with indices in the set

$$\mathcal{I}_{n,s} = \bigcup_{p:s \in p} \mathcal{I}_{n,p} \tag{34}$$

and the decoding rate is $R_s = \sum_{p:s \in p} R_p$. For the composite channel, the decoder output

$$g_{n,s}^e(y^n) = (\hat{b}_i)_{i \in \mathcal{I}_{n,s}}$$

is obtained by concatenating the binary representations $(\hat{b}_i)_{i \in \mathcal{I}_{n,p}}$ of each $\hat{M}_p$, where $s \in p$ and

$$g_{n,s}^{\mathrm{BC}}(y^n) = \prod_{p:s \in p} \hat{M}_p$$

is the decoder output of receiver s in the broadcast channel. The decoding error probability for the expected-rate code in channel state s is

$$P_e^{(n,s)} = \Pr\{E_s\}$$

where the error event E_s for the broadcast code is defined in (15). Notice that

$$P_e^{(n,s)} = \Pr\{E_s\} \leq \Pr\{\cup_s E_s\} = P_e^{(n)}$$

so the expected error probability

$$\mathbb{E}_S P_e^{(n,S)} \leq P_e^{(n)} \to 0$$

as $n \to \infty$, according to the BC code definition. Therefore, the rate

$$R = \mathbb{E}_S R_S = \sum_s P_S(s) R_s = \sum_s P_S(s) \sum_{p:s \in p} R_p$$

is an achievable expected rate and (31) is proved.

2) Mapping Expected-Rate Code to Broadcast Code: Next we show that for any fixed $\epsilon > 0$,

$$C^e - \epsilon \leq \sup_{\{R_p\} \in \mathcal{C}_{\mathrm{BC}}} \sum_{p \in \mathcal{P}} R_p \sum_{s \in p} P_S(s). \tag{35}$$

According to the definition of the expected capacity, there exists a sequence of $\{(2^{nR_t}, \{2^{nR_s}\}, n)\}$ codes such that

$$\mathbb{E}_S R_S \to R \geq C^e - \epsilon \tag{36}$$

and $\mathbb{E}_S P_e^{(n,S)} \to 0$. The transmitted information bits are indexed by $\mathcal{I}_{n,t} = \{1, 2, \ldots, nR_t\}$. Since the transmitter and the receiver agree on the index set $\mathcal{I}_{n,s}$ of those information bits that can be reliably decoded in each channel state s, the transmitter can define, for each subset $p \in \mathcal{P}$ of channel states, the index set $\mathcal{I}_{n,p}$ of those information bits decodable exclusively for channel states within p, i.e.,

$$\mathcal{I}_{n,p} = \left(\bigcap_{s \in p} \mathcal{I}_{n,s} \right) \bigcap \left(\bigcap_{s \notin p} \bar{\mathcal{I}}_{n,s} \right)$$

where

$$\bar{\mathcal{I}}_{n,s} = \{i : i \in \mathcal{I}_{n,t}, i \notin \mathcal{I}_{n,s}\}$$

is the complement index set of $\mathcal{I}_{n,s}$. Denote by nR_p the cardinality of $\mathcal{I}_{n,p}$. We observe that $\mathcal{I}_{n,p}$ are mutually exclusive, the relationship (33) and (34) still hold and the decoding rate satisfies $R_s = \sum_{s \in p} R_p$.

The $\{(2^{nR_t}, \{2^{nR_s}\}, n)\}$ expected-rate code can be mapped to the following BC code. Define the message set of the BC code as

$$\mathcal{M}_p = \mathcal{M}(\mathcal{I}_{n,p})$$

in the sense that each message $M_p \in \mathcal{M}_p$ has the corresponding binary representation $(b_i)_{i \in \mathcal{I}_{n,p}}$. The encoder for the BC code satisfies

$$f_n^{\mathrm{BC}}\left(\prod_{p \in \mathcal{P}, p \neq \phi} M_p\right) = f_n^e(M_t)$$

where $M_t = (b_i)_{i \in \mathcal{I}_{n,t}}$ is obtained by concatenating the binary representations of each M_p. When the composite channel is in state s, the decoder output is

$$g_{n,s}^e(y^n) = \hat{M}_s = (\hat{b}_i)_{i \in \mathcal{I}_{n,s}}.$$

Since $\mathcal{I}_{n,p} \subseteq \mathcal{I}_{n,s}$ for any p satisfying $s \in p$, we define the decoder output for receiver s in the BC to be

$$g_{n,s}^{\mathrm{BC}}(y^n) = \prod_{p:s \in p} \hat{M}_p$$

where the binary representation $(b_i)_{i \in \mathcal{I}_{n,p}}$ of each $\hat{M}_p$ can be obtained by the corresponding bits in $\hat{M}_s$.

The error event E_s for receiver s of the BC is defined in (15) with the error probability

$$\Pr\{E_s\} = P_e^{(n,s)}$$

and the overall error probability

$$P_e^{(n)} = \Pr\{\cup_s E_s\} \leq \sum_s \Pr\{E_s\} = \sum_s P_e^{(n,s)}.$$

By definition of the expected-rate capacity

$$\mathbb{E}_S P_e^{(n,S)} = \sum_s P_S(s) P_e^{(n,s)} \geq \left(\min_{s \in \mathcal{S}} P_S(s)\right)\left(\sum_s P_e^{(n,s)}\right).$$

Assuming each channel state s occurs with strictly positive probability, i.e., $\min_{s \in \mathcal{S}} P(s) > 0$, then $\mathbb{E}_S P_e^{(n,S)} \to 0$ implies

$$P_e^{(n)} \leq \sum_s P_e^{(n,s)} \to 0.$$

Therefore, the code constructed above is a valid BC code, i.e., $\{R_p\} \in \mathcal{C}_{\mathrm{BC}}$, and we conclude

$$R = \mathbb{E}_S R_S = \sum_s P_S(s) R_s = \sum_s P_S(s) \sum_{p:s \in p} R_p$$

$$\leq \sup_{\{R_p\} \in \mathcal{C}_{\mathrm{BC}}} \sum_{p \in \mathcal{P}} R_p \sum_{s \in p} P_S(s). \qquad (37)$$

From (36) and (37) we see the inequality (35) is established. Since ϵ is arbitrary, Theorem 2 is a result of (31) and (35).
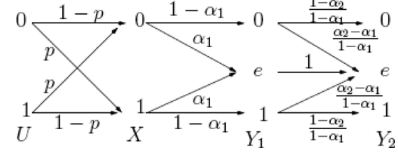

Fig. 8. Degraded binary erasure broadcast channel.

APPENDIX D
PROOF OF (18)

Consider a two-user BC where the channel to each user is a BEC with erasure probability α_i, $i = 1, 2$, i.e., the conditional marginal distribution satisfies

$$p(y_i|x) = \begin{cases} 1 - \alpha_i, & y_i = x \\ \alpha_i, & y_i = e. \end{cases}$$

Assuming $\alpha_1 < \alpha_2$, we observe that the BC is stochastically degraded since

$$p(y_2|x) = \sum_{y_1} p(y_1|x) p'(y_2|y_1)$$

where $p'(e|e) = 1$ and for $y_1 \neq e$

$$p'(y_2|y_1) = \begin{cases} \frac{1 - \alpha_2}{1 - \alpha_1}, & y_2 = y_1 \\ \frac{\alpha_2 - \alpha_1}{1 - \alpha_1}, & y_2 = e. \end{cases}$$

Therefore, the capacity region of the BEC-BC is the convex hull of the closure of all (R_1, R_{12}) satisfying

$$R_1 \leq I(X; Y_1|U)$$
$$R_{12} \leq I(U; Y_2) \qquad (38)$$

for some joint distribution $p(u)p(x|u)p(y_1, y_2|x)$. Since the cardinality of the random variable U is bounded by $|\mathcal{U}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}_1|, |\mathcal{Y}_2|\} = 2$ [28, p. 422] and the channel is symmetric with respect to the alphabet 0 and 1, we can take $p(u) \sim \mathrm{Bernoulli}(1/2)$ and $p(x|u)$ as the transition probability of a binary symmetric channel with crossover probability p. This stochastically degraded BEC-BC together with the auxiliary random variable U is illustrated in Fig. 8.

The capacity region (38) is evaluated to be

$$R_1 \leq (1 - \alpha_1)h(p)$$
$$R_{12} \leq (1 - \alpha_2)[1 - h(p)] \qquad (39)$$

where $h(p) = -p \log p - (1-p)\log(1-p)$ is the binary entropy function. Assuming the two ergodic components are equally probable in the composite channel, the achievable expected rate using a broadcast code is then

$$R = \sup_p \{R_{12} + R_1/2\}$$

$$= \max\left\{1 - \alpha_2, \frac{1 - \alpha_1}{2}\right\}.$$

APPENDIX E
PROOF OF UPPER BOUND FOR EXPECTED CAPACITY

Denote by $X_s^n(1), \ldots, X_s^n(2^{nR_s})$ and $D_s(1), \ldots, D_s(2^{nR_s})$ the set of codewords and decoding regions corresponding to channel s. We fix $\gamma > 0$ and define for each $s \in \mathcal{S}$ and $1 \leq i \leq 2^{nR_s}$

$$B_s(i) = \{Y^n \in \mathcal{Y}^n : \frac{1}{n} i_{X^n W^n}(X^n(i); Y^n | s) \leq R_s - \gamma\}$$
$$= \{Y^n \in \mathcal{Y}^n : P_{X^n | Y^n, S}(X^n(i) | Y^n, s) \leq 2^{-n\gamma}\} \quad (40)$$

where (40) follows from (6). Notice that for any s with $R_s > 0$

$$P_{X^n Y^n | S}\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n | s) \leq R_s - \gamma \middle| s\right]$$

$$\leq \sum_{i=1}^{2^{nR_s}} \left[2^{-nR_s} P_{Y^n | X^n, S}(D_s(i)^c | X^n(i), s)\right.$$
$$\left. + \sum_{y^n \in B_s(i) \cap D_s(i)} P_{X^n Y^n | S}(X^n(i), y^n | s)\right]$$

$$\leq P_e^{(n,s)} + \sum_{i=1}^{2^{nR_s}} \sum_{y^n \in B_s(i) \cap D_s(i)} 2^{-n\gamma} P_{Y^n | S}(y^n | s)$$

$$\leq P_e^{(n,s)} + 2^{-n\gamma}. \quad (41)$$

Furthermore, we have

$$\mathbb{E}_S \liminf_{n \to \infty} P_{X^n Y^n | S}\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n | S) \leq R_S - \gamma \middle| S\right]$$

$$\leq \lim_{n \to \infty} \mathbb{E}_S P_{X^n Y^n | S}\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n | S) \leq R_S - \gamma \middle| S\right]$$

$$\leq \lim_{n \to \infty} [\mathbb{E}_S P_e^{(n,S)} + 2^{-n\gamma}] = 0$$

where the chain of inequalities follows from Fatou's lemma, (41), and the code constraint $\mathbb{E}_S P_e^{(n,S)} \to 0$. Since the probability must be non-negative, we conclude

$$\liminf_{n \to \infty} P_{X^n Y^n | S}\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n | S) \leq R_S - \gamma \middle| S\right] = 0$$

almost surely (a.s.) in S. Thus, for any $\epsilon > 0$

$$P_{X^n Y^n | S}\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n | S) \leq R_S - \gamma \middle| S\right] < \epsilon$$

occurs infinitely often a.s. Assuming $|i_{X^n W^n}(X^n; Y^n | S)|$ is bounded by M, we then have

$$\mathbb{E}_{X^n Y^n | S}\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n | S) \middle| S\right] > (R_S - \gamma)(1 - \epsilon) - \epsilon M$$

also occurs infinitely often a.s. Since ϵ is arbitrary, we see that

$$\mathbb{E}_S \mathbb{E}_{X^n Y^n | S}\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n | S) \middle| S\right] \geq \mathbb{E}_S R_S - \gamma$$

occurs infinitely often for arbitrary γ, which gives us the upper bound (19) for expected capacity. Note that the expectation in the upper bound (19) is indeed $\frac{1}{n} I(X^n; Y^n | S)$, so the upper bound can also be proved using the standard technique of Fano's inequality.

REFERENCES

[1] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Jul.-Oct. 1948.

[2] K. Winkelbauer, "On the coding theorem for decomposable discrete information channels I," *Kybernetika*, vol. 7, no. 2, pp. 109–123, 1971.

[3] J. C. Kieffer, "A general formula for the capacity of stationary nonanticipatory channels," *Inf. Control*, vol. 26, no. 4, pp. 381–391, 1974.

[4] R. Ahlswede, "The weak capacity of averaged channels," *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, vol. 11, pp. 61–73, 1968.

[5] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 7, pp. 1147–1157, Jul. 1994.

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

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

[8] J. Wolfowitz, *Coding Theorems of Information Theory*. New York: Springer-Verlag, 1964.

[9] T. S. Han, *Information-Spectrum Method in Information Theory*. New York: Springer, 2003, Applications of mathematics.

[10] I. Csiszar and P. Narayan, "The capacity of the arbitrarily varying channel," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 18–26, Jan. 1991.

[11] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[12] L. Li and A. J. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels: Part II: Outage capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1103–1127, Mar. 2001.

[13] G. Caire and S. Shamai(Shitz), "On the achievable throughput of a multiple-antenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[14] S. V. Hanly and D. N. Tse, "Multiple-access fading channels: Part II: Delay-limited capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 11, pp. 2816–2831, Nov. 1998.

[15] L. Ozarow, S. Shamai, and A. Wyner, "Information theoretical considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 359–378, May 1994.

[16] J. P. E. Biglieri and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619–2692, Dec. 1998.

[17] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, Jun. 2003.

[18] G. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, pp. 311–335, Mar. 1998.

[19] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[20] T. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 2–14, Jan. 1972.

[21] S. Shamai(Shitz), "A broadcast strategy for the Gaussian slowly fading channel," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Ulm, Germany, Jun. 1997, p. 150.

[22] S. Shamai and A. Steiner, "A broadcast approach for a single-user slowly fading MIMO channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2617–2635, Oct. 2003.

[23] M. Effros and A. Goldsmith, "Capacity definitions and coding strategies for general channels with receiver side information," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Cambridge, MA, Aug. 1998, p. 39.

[24] M. Effros, A. Goldsmith, and Y. Liang, "Capacity definitions of general channels with receiver side information," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Nice, France, Jun. 2007, pp. 921–925.

[25] D. Gündüz and E. Erkip, "Joint source-channel codes for MIMO block-fading channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 116–134, Jan. 2008.

[26] C. T. K. Ng, D. Gündüz, A. Goldsmith, and E. Erkip, "Distortion minimization in Gaussian layered broadcast coding with successive refinement," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5074–5086, Nov. 2009.

[27] C. Tian, A. Steiner, S. Shamai, and S. Diggavi, "Successive refinement via broadcast: Optimizing expected distortion of a Gaussian source over a Gaussian fading channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 2903–2918, Jul. 2008.

[28] T. Cover and J. Thomas, *Elements of Information Theory*. Hoboken, NJ: Wiley, 2006.

[29] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inf. Theory*, vol. IT-4, pp. 2–22, 1954.

[30] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 752–772, May 1993.

[31] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 44–54, Jan. 1995.

[32] A. Ganti, A. Lapidoth, and I. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 11, pp. 2315–2328, Nov. 2000.

[33] R. Durrett, *Probability: Theory and Examples*, 3rd ed. Belmont, CA: Duxbury, 2005.

[34] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1971–1988, May 2001.

[35] T. Ghanim and M. C. Valenti, "The throughput of hybrid-ARQ in block fading under modulation constraints," in *Proc. Conf. Inform. Sciences and Systems (CISS)*, Princeton, NJ, Mar. 2006, pp. 253–258.

[36] M. Ancis and D. D. Giusto, "Reconstruction of missing blocks in JPEG picture transmission," in *Proc. IEEE Pacific Rim Conf. Comm., Computers and Signal Processing*, Victoria, BC, Canada, Aug. 1999, pp. 288–291.

[37] A. Goldsmith, *Wireless Communications*. New York: Cambridge Univ. Press, 2005.

[38] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Euro. Trans. Telecomm. (ETT)*, vol. 10, no. 6, pp. 585–596, Nov. 1999.

[39] A. El Gamal, "The capacity of a class of broadcast channels," *IEEE Trans. Inf. Theory*, vol. 25, no. 3, pp. 166–169, Mar. 1979.

[40] R. Gallager, *Information Theory and Reliable Communication*. Hoboken, NJ: Wiley, 1968.

[41] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

[42] W. Yu and J. Cioffi, "Sum capacity of Gaussian vector broadcast channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, Sep. 2004.

[43] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channels and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[44] H. Weingarten, Y. Steinberg, and S. Shamai(Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, 3936–3964, Sep. 2006.

[45] D. Gündüz and E. Erkip, "Source and channel coding for quasi-static fading channels," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2005, pp. 18–22.

[46] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert–Elliot channels," *IEEE Trans. Inf. Theory*, vol. 35, no. 11, pp. 1277–1290, Nov. 1989.

[47] P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Trans. Inf. Theory*, vol. 19, no. 3, pp. 197–207, Mar. 1973.

[48] D. Luenberger, *Optimization by Vector Space Methods*. Hoboken, NJ: Wiley, 1969.

[49] Y. Liang, A. Goldsmith, and M. Effros, "Distortion metrics of composite channels with receiver side information," in *Proc. IEEE Inform. Theory Workshop (ITW)*, Lake Tahoe, CA, Sep. 2007, pp. 559–564.

[50] S. Shamai, S. Verdú, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 564–579, Mar. 1998.

[51] Z. Reznic, M. Feder, and R. Zamir, "Distortion bounds for broadcasting with bandwidth expansion," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3778–3788, Aug. 2006.

[52] U. Mittal and N. Phamdo, "Hybrid digital-analog (HDA) joint source-channel codes for broadcasting and robust communications," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1082–1102, May 2002.

[53] K. Zachariadis, M. Honig, and A. Katsaggelos, "Source fidelity over a two-hop fading channel," in *Proc. IEEE MilCom*, Monterey, CA, Nov. 2004, pp. 134–139.

[54] T. Cover, A. El Gamal, and M. Salehi, "Muliple access channels with arbitrarily correlated sources," *IEEE Trans. Inf. Theory*, vol. 26, no. 11, pp. 648–657, Nov. 1980.

[55] Y. Liang, A. Goldsmith, and M. Effros, "Source-channel coding and separation for general communication systems," *IEEE Trans. Inf. Theory*, submitted for publication.

Michelle Effros (S'93–M'95–SM'03–F'09) received the B.S. degree with distinction in 1989, the M.S. degree in 1990, and the Ph.D. degree in 1994, all in electrical engineering from Stanford University, Stanford, CA.

During the summers of 1988 and 1989, she worked at Hughes Aircraft Company. She joined the faculty at the California Institute of Technology, Pasadena, in 1994 and is currently a Professor of Electrical Engineering. Her research interests include information theory, network coding, data compression, and communications.

Dr. Effros received Stanford's Frederick Emmons Terman Engineering Scholastic Award (for excellence in engineering) in 1989, the Hughes Masters Full-Study Fellowship in 1989, the National Science Foundation Graduate Fellowship in 1990, the AT&T Ph.D. Scholarship in 1993, the NSF CAREER Award in 1995, the Charles Lee Powell Foundation Award in 1997, the Richard Feynman–Hughes Fellowship in 1997, an Okawa Research Grant in 2000, and was cited by Technology Review as one of the world's top 100 young innovators in 2002. She and her coauthors received the 2009 Communications & Information Theory Society Joint Paper Award for the paper "A Random Linear Network Coding Approach to Multicast," which appeared in the IEEE TRANSACTIONS ON INFORMATION THEORY in October 2006. She is a member of Tau Beta Pi, Phi Beta Kappa, Sigma Xi, and a fellow of the IEEE. She served as the Editor of the IEEE Information Theory Society Newsletter from 1995 to 1998 and as a Member of the Board of Governors of the IEEE Information Theory Society from 1998–2003 and 2008–present and has been a member of the Advisory Committee for the Computer and Information Science and Engineering (CISE) Directorate at the National Science Foundation from 2009 to the present. She served on the IEEE Signal Processing Society Image and Multi-Dimensional Signal Processing (IMDSP) Technical Committee from 2001 to 2007 and on ISAT from 2006 to 2009. She was an Associate Editor for the joint special issue on Networking and Information Theory in the IEEE TRANSACTIONS ON INFORMATION THEORY and the IEEE/ACM TRANSACTIONS ON NETWORKING and as an Associate Editor for Source Coding for the IEEE TRANSACTIONS ON INFORMATION THEORY from 2004 to 2007. She is currently serving as an Associate Editor for the upcoming special issue of the IEEE TRANSACTIONS ON INFORMATION THEORY honoring Prof. R. Koetter. She has served on numerous technical program committees and review boards, including serving as general cochair for the 2009 Network Coding Workshop. She will serve as cochair of the Technical Program Committee for the International Symposium on Information Theory in 2012.

Andrea Goldsmith (S'90–M'93–SM'99–F'05) received the B. S., M.S., and Ph.D. degrees in electrical engineering from the University of California, Berkeley.

She is a Professor of electrical engineering at Stanford University, Stanford, CA, and was previously an Assistant Professor of electrical engineering at Caltech, Pasadena, CA. While on leave from Stanford, she founded and served as CTO of Quantenna Communications, Inc. She also previously held industry positions at Maxim Technologies, Memorylink Corporation, and AT&T Bell Laboratories. Her research includes work on wireless systems and networks, cognitive radios, sensor and green networks, and cross-layer network design. She is author of the book *Wireless Communications* and coauthor of the book *MIMO Wireless Communications*, both published by Cambridge University Press.

Dr. Goldsmith is a Fellow of the IEEE and of Stanford. She has received several awards for her work, including the National Academy of Engineering Gilbreth Lectureship, the IEEE Comsoc Wireless Communications Technical Committee Recognition Award, the Alfred P. Sloan Fellowship, the Stanford Terman Fellowship, the National Science Foundation CAREER Development Award, the Office of Naval Research Young Investigator Award, and the Silicon Valley Business Journal's Woman of Influence Award. In addition, she was a corecipient of the 2005 IEEE Communications Society and Information Theory Society joint paper award. She currently serves as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY and as editor for the *Journal on Foundations and Trends in Communications and Information Theory and in Networks*. She previously served as an editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and for the *IEEE Wireless Communications Magazine*, as well as guest editor for several IEEE journal and magazine special issues. She participates actively in committees and conference organization for the IEEE Information Theory and Communications Societies and has served on the Board of Governors for both societies. She is a Distinguished Lecturer for both societies and Past President of the IEEE Information Theory Society. She also founded the student committee of the IEEE Information Theory society, is an inaugural recipient of Stanford's Postdoc Mentoring Award, and serves as Stanford's Faculty Senate Chair for the 2009/2010 academic year.

Yifan (Ethan) Liang, received the B.E. degree (*summa cum laude*) from the University of Science and Technology of China in 2002 and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 2008.

In 2006, he served as an intern in the Crawford Hill Wireless Research Group at Lucent Bell Laboratories. He is now with Goldman, Sachs & Co., New York.