Communication Over Individual Channels

Yuval Lomnitz and Meir Feder, *Fellow, IEEE*

***Abstract*—A communication problem in considered, where no mathematical model is specified for the channel. The achievable rates are determined as a function of the channel input and output sequences known a-posteriori, without assuming any a-priori relation between them. For discrete channels the empirical mutual information between the input and output sequences is shown to be achievable, while for continuous channels the achievable rate is based on the empirical correlation between the sequences. A rate-adaptive scheme employing feedback which achieves these rates asymptotically with a guaranteed reliability, without prior knowledge of the channel behavior, is presented.**

***Index Terms*—Channel uncertainty, communication, feedback communication, unknown channels, random coding, rateless coding.**

I. Introduction

COMMUNICATION over unknown channels is traditionally dealt with by using the framework of compound channels and arbitrarily varying channels [1]. In this framework, a statistical model of the channel is given, up to some unknown parameters or up to an unknown sequence of channel states, yet robust communication is required for all possible settings of these unknown parameters. This approach enables the design of robust communication systems; however there are two disadvantages in this setting. First, the resulting rates are often pessimistic, as they are tuned to the worst possible channel behavior. For many cases of interest, the compound or arbitrarily varying channel capacity may be zero, for example when there are some channels with zero capacity in the family. Second, a statistical model of the channel, defining the distribution of the output as a function of the input and the unknown parameters, is required, and assumed to be fully known.

A different communication model which overcomes the first issue was presented by Shayevitz and Feder [2], where the problem of communicating over a channel with an individual, predetermined noise sequence, which is unknown to the sender and receiver, was considered. Specifically, consider the simple example [3] of a binary channel $y_i = x_i \oplus e_i$ where the error sequence $\{e_i\}$ can be any arbitrary unknown sequence. Although the traditional capacity of this channel is zero, using perfect feedback and common randomness, communication was shown to be possible at a rate approaching $1 - h_b(\hat{\epsilon})$, i.e., the capacity of the binary symmetric channel (BSC) with cross-over probability $\hat{\epsilon}$, where $h_b(\cdot)$ is the binary entropy function and $\hat{\epsilon}$ is the relative number of '1'-s in $\{e_i\}$. The main idea is that although the noise sequence is unknown, if the rate can be adapted, one can opportunistically increase the rate when the channel is empirically "less noisy". Subsequently the authors extended the results to modulo-additive channels and adversary noise sequences [2]. The concept was further generalized by Eswaran *et al.* [4] to include general discrete channels with an individual state sequence, where they showed that the mutual information of an effective "state averaged channel" can be attained. While this model of communication avoids worst case assumptions by using feedback, it still requires the channel model to be mathematically specified and known.

In this paper we take this model one step further. We consider a channel where no specific probabilistic or mathematical relation between the input and the output is assumed. We term this channel an *individual channel*. In order to define positive communication rates without assumptions on the channel, we characterize the achievable rate using the specific input and output sequences. When there is a feedback link where the channel output or other information from the decoder can be sent back to the encoder, the rate of transmission is adapted to the empirical channel so that a small error probability is always guaranteed. Without feedback the rate of transmission cannot be matched to the quality of the channel so outage may occur.

As an example suppose the transmitter sends the input symbols $x_i \in \mathbb{R}, i = 1, \ldots, n$. For any input x_i the channel outputs a value $y_i \in \mathbb{R}$ in a way which is unknown to the encoder and decoder and may be adversarial. We may imagine that a demon is determining the channel output. Using feedback, y_i can be sent back to the encoder. Can we make any guarantee on the communication rate? Certainly, one cannot make any a-priori guarantee, since y_i may be unrelated to x_i. But as we will show here, one can guarantee a high rate if the input and output sequences, observed a-posteriori, have a good correlation. There is a system with feedback that adapts the communication rate to a value approaching $\frac{1}{2}\log(1 + S\hat{N}R)$, where $S\hat{N}R = \frac{\hat{\rho}^2}{1-\hat{\rho}^2}$ is an empirical measure of the effective *SNR* in the link between $\mathbf{x}$ and $\mathbf{y}$, and $\hat{\rho}$ is the empirical correlation factor. This system guarantees a small probability of error, without assuming any a-priori relation between the channel input and output.

We consider two classes of individual channels: discrete input and output channels and continuous real valued input and output channels, and two communication models: fixed rate (not requiring feedback) and adaptive rate (requiring feedback). In all cases we assume unlimited common randomness exists. The case of feedback is of higher interest, since by adapting the rate, outage is avoided. The case of fixed rate is used as an intermediate step, but the results are interesting since they can be

Manuscript received October 04, 2009; accepted June 23, 2011. Date of current version November 11, 2011. This work was supported in part by the Israeli Science Foundation under Grant 634/09, the Feder Family Award, and a fellowship from The Yitzhak and Chaya Weinstein Research Institute for Signal Processing at Tel Aviv University.

The authors are with the Department of Electrical Engineering—Systems, Tel-Aviv University, Ramat-Aviv 69978, Israel (e-mail: yuvall@eng.tau.ac.il; meir@eng.tau.ac.il).

Communicated by H. Yamamoto, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2011.2169130

used for analysis of semi-probabilistic models. The main result is that with a small amount of feedback, a communication at a rate close to the empirical mutual information (or its Gaussian equivalent for continuous channels) can be achieved, without any prior knowledge, or assumptions, about the channel structure.

The paper is organized as follows: in Section II we give a high level overview of the results. In Section III we define the model and notation. Section IV deals with communication without feedback where the results pertaining to discrete and continuous case are formalized and proven, and the choice of the rate function and the Gaussian prior for the continuous case is justified. Section V deals with the case where feedback is present. In this section we state the main result and the adaptive rate scheme that achieves it, and delay the proof to Section VI. Here, the error probability and the achieved rate are analyzed and bounded. Section VII gives several examples, Section VIII is dedicated to comments and Section IX highlights areas for further study.

II. Overview of the Main Results

We start with a high level overview of the definitions and results. The definitions below are conceptual rather than accurate, and detailed definitions follow in the next sections.

A rate function is a function $R_{\text{emp}} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ of the specific input and output sequences. Roughly speaking, the rate function represents an empirical "channel capacity". In the non-adaptive case, information is transmitted at a constant rate R. We expect the message to be received with an arbitrarily small error probability whenever R_{emp} meets or exceeds the rate of transmission, i.e., whenever $R_{\text{emp}}(\mathbf{x},\mathbf{y}) \geq R$. If this can be guaranteed for every $\mathbf{x},\mathbf{y}$ where $R_{\text{emp}}(\mathbf{x},\mathbf{y}) \geq R$, we say that R_{emp} is *achievable*. If $R_{\text{emp}}(\mathbf{x},\mathbf{y}) < R$ then we consider the channel to be in outage and no guarantee is made on the error probability.

In the rate adaptive case, we would like the system to adapt its transmission rate R using feedback, such that data at a rate $R \geq R_{\text{emp}}(\mathbf{x},\mathbf{y})$ would be transmitted and decoded with an arbitrarily small probability of error, for every $\mathbf{x},\mathbf{y}$. We allow excluding from this guarantee a small set of input sequences $\mathbf{x}$. If this can be done, we say that R_{emp} is *adaptively achievable*. Note that the fact that $R \geq R_{\text{emp}}(\mathbf{x},\mathbf{y})$ does not lead to outage in this case since the system controls the transmission rate, and keeps the guaranteed reliability for all sequences. Roughly speaking, this means that in any instance of the system operation, where a specific $\mathbf{x}$ was the input and a specific $\mathbf{y}$ was the output, the communication rate had been at least $R_{\text{emp}}(\mathbf{x},\mathbf{y})$. Note that we do not assume any relation between $\mathbf{x}$ and $\mathbf{y}$ and the only statistical assumptions are related to the common randomness. We consider the rate and error probability *conditioned* on a specific input and output, where the error probability is averaged over common randomness.

To make the concept clear, a trivial example of a rate function for a binary input—binary output channel is $R_{\text{emp}} = \text{Ind}(\mathbf{x} = \mathbf{y})$, i.e., $R_{\text{emp}} = 1$ iff the output is identical to the input. To attain this rate function non-adaptively, one would simply transmit the message un-coded, at a rate $R = 1$. If the channel output happened to equal the input, the communication had succeeded. If it happened to be different, $R_{\text{emp}} = 0 < R$ and thus no guarantee was made.

In a certain sense, the choice of rate functions is arbitrary: for any pair of encoder and decoder, we can tailor a function $R_{\text{emp}}(\mathbf{x},\mathbf{y})$ as a function equaling the transmitted rate whenever the error probability given the two sequences (averaged over messages and the common randomness) is sufficiently small, and 0 otherwise. However it is clear that there are certain rates which cannot be exceeded uniformly. Our interest will focus on simple functions of the input and output, and specifically in this paper we focus on functions of the instantaneous (zero order) empirical statistics. Extension to higher order models seems technical.

For the discrete channel we show that a rate

$$R_{\text{emp}} = \hat{I}(\mathbf{x};\mathbf{y}) \tag{1}$$

is asymptotically achievable for large block length $n \to \infty$, where $\hat{I}(\cdot;\cdot)$ denotes the empirical mutual information [5] (see definition in Section III-B, and Theorems 1, 3). This pertains to both fixed rate and adaptive rate systems according to the definitions above. For the fixed rate case this roughly means that it is possible to design an encoder and a decoder such that when $\hat{I}(\mathbf{x};\mathbf{y}) \geq R$, the message will be decoded correctly with high probability. For the adaptive case, this means that with feedback, it is possible to design a system in which the rate is adapted to achieve $R \geq \hat{I}(\mathbf{x};\mathbf{y})$ and the error probability is small at all times. All the inequalities above are up to asymptotically vanishing constants.

For the continuous (real valued) channel we show that a rate

$$R_{\text{emp}} = \frac{1}{2}\log\left(\frac{1}{1-\hat{\rho}(\mathbf{x},\mathbf{y})^2}\right) \tag{2}$$

is asymptotically achievable (in the fixed-rate and adaptive rate senses), where $\hat{\rho}$ is the empirical correlation factor between the input and output sequences (see Theorems 2, 4). We define the empirical correlation factor in a slightly non standard way as $\hat{\rho} = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|}$ (that is, without subtracting the mean). This is done only to simplify definitions and derivations, and similar claims can be made using the correlation factor defined in the standard way. Note that $\frac{1}{2}\log\left(\frac{1}{1-\rho^2}\right)$ is the mutual information between two jointly Gaussian random variables with a correlation factor ρ. Although the result regarding the continuous case is less tight, we show that this is the best achievable rate function that can be defined by second order moments, and is tight for the additive white Gaussian noise (AWGN) channel with signal power P and noise power N: for this channel $\hat{\rho}^2 \to \rho^2 = \frac{P}{P+N}$ therefore $R_{\text{emp}} \to \frac{1}{2}\log\left(1+\frac{P}{N}\right)$. This rate function (2) can be also formulated as $\frac{1}{2}\log(1+\hat{SNR})$, in analogy to the familiar expression for the AWGN capacity (see Section VII-D).

In all achievability results, we specify the rate function together with the set of input distributions for which the result holds. The reason is that the rate functions are a function of the channel input, which is determined by the scheme itself. This is an opening for possible falsity—the encoder may choose sequences for which the rate is attained more easily. For example, by setting $\mathbf{x} = 0$ one can attain the results above in a

void way, since the rate function will always be 0. We circumvent this difficulty by constraining the input distribution. Different from classical results in information theory, we do not use the input distribution only as a means to show the existence of good codes; taking advantage of the common randomness we *require* the encoder to emit input symbols that are random and distributed according to a defined prior (in the current paper we assume i.i.d. distribution). This definition breaks the circular dependence that might have been created, by specifying the input behavior together with the rate function. Specifically, in the discrete case, (1) is achievable with any i.i.d. input distribution, and in the continuous case (2) is achievable for an i.i.d. Gaussian input. Note that these results hold under the theoretical assumption that one may have access to a random variable of any desired distribution, which is in some cases un-feasible to generate in an exact manner—see discussion in Section VIII-F.

As will be seen, we achieve these rates by random coding and universal decoders. For the case of feedback we use iterated instances of rateless coding, i.e., we encode a fixed number of bits and the decision time depends on the channel. Although the theorems are stated in asymptotical terms, explicit expressions for rates guaranteed by the scheme for finite block lengths are shown in the proofs. With a small modification, the scheme is able to operate asymptotically with "zero rate" feedback (meaning any positive capacity of the feedback channel suffices). A similar but slightly more complicated scheme was used by Eswaran *et al.* [4] (see a comparison in the Appendix). The differences between the current framework and related models such as the arbitrarily varying channel (AVC) and channels with an individual noise sequence are examined in Section VIII-A.

Since the current paper was submitted, we have investigated possible extensions to the current results, some of which were published in conference papers [6], [7]. A more general framework which characterizes the set of achievable rate functions, with improved bounds on achievability, and includes the above examples as particular cases, is to appear in a follow up paper which is currently in preparation. We also considered [8] an alternative approach to universal communication which does not require determining the transmit distribution a-priori.

III. Definitions and Notation

A. Notation

In general we use uppercase letters to denote random variables, respective lowercase letters to denote their sample values and boldface letters to denote vectors, which are by default of length n. However we deviate from this practice when the change of case leads to confusion, and vectors are always denoted by lowercase letters even when they are random variables. $\mathbb{R}$ denotes the set of real numbers, and $\mathcal{N}(\mu,\sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2.

We denote the L_2 norm by $\|\mathbf{x}\| \triangleq \sqrt{\mathbf{x}^T\mathbf{x}}$. We denote by $P\circ Q$ the product of conditional probability functions e.g., $(P\circ Q)(x,y) = P(x)\cdot Q(y\,|\,x)$. $\mathbb{U}(A)$ denotes a uniform distribution over the set A.

A hat ($\hat{\square}$) denotes an estimated value. We denote the empirical distribution as $\hat{P}$ (e.g., $\hat{P}_{(\mathbf{x},\mathbf{y})}(x,y) \triangleq \frac{1}{n}\sum_{i=1}^{n}\delta_{(x_i-x),(y_i-y)}$). The source vectors $\mathbf{x},\mathbf{y}$ and/or the variables x,y are sometimes omitted when they are clear from the context. We denote by $\hat{H}(\,\cdot\,)$, $\hat{I}(\cdot;\cdot)$, $\hat{\rho}(\cdot;\cdot)$ the empirical entropy, the empirical mutual information and the empirical correlation factor, which are the respective values calculated for the empirical distribution. All expressions such as $\hat{H}(\mathbf{x})$, $\hat{H}(\mathbf{x}\,|\,\mathbf{y})$, $\hat{I}(\mathbf{x};\mathbf{y})$, $\hat{I}(\mathbf{x};\mathbf{y}\,|\,\mathbf{z})$, $\hat{I}(\mathbf{x};\mathbf{y}\,|\,\mathbf{z}=z_0)$ are interpreted as their respective probabilistic counterparts $H(X)$, $H(X\,|\,Y)$, $I(X;Y)$, $I(X;Y\,|\,Z)$, $I(X;Y\,|\,Z=z_0)$ where (X,Y,Z) are random variables distributed according to the empirical distribution of the vectors $\hat{P}_{(\mathbf{x},\mathbf{y},\mathbf{z})}$, or equivalently are defined as a random selection of an element of the vectors i.e., $(X,Y,Z)=(x_i,y_i,z_i), i\sim\mathbb{U}\{1,\ldots,n\}$. It is clear from this equivalence that relations on entropy and mutual information (e.g., positivity, chain rules) directly translate to relations on their empirical counterparts.

We apply superscript and subscript indices to vectors to define subsequences in the standard way, i.e., $\mathbf{x}_i^j \triangleq (x_i,x_{i+1},\ldots,x_j)$, $\mathbf{x}^i \triangleq \mathbf{x}_1^i$

We denote by $I(P,W)$ the mutual information $I(X;Y)$ when $(X,Y)\sim P(x)\cdot W(y\,|\,x)$. The Bernoulli distribution is denoted $\mathrm{Ber}(p)$, and $h_b(p)\triangleq H(\mathrm{Ber}(p)) = -p\log p-(1-p)\log(1-p)$ denotes the binary entropy function. The indicator function $\mathrm{Ind}(E)$ where E is a set or a probabilistic event is defined as 1 over the set (or when the event occurs) and 0 otherwise.

The functions $\log(\,\cdot\,)$ and $\exp(\,\cdot\,)$ refer to base 2 unless specified otherwise, and information theoretic quantities $H(\,\cdot\,), I(\cdot;\cdot), D(\cdot\|\cdot)$ are measured in bits.

We use Bachmann & Landau notations for orders of magnitude. Specifically, $f_n=\Theta(g_n)$, means $\exists n_0,\alpha,\beta>0:\forall n>n_0:\alpha g_n\le f_n\le\beta g_n$, $f_n\in o(g_n)$ or $f_n=o(g_n)$ means $\frac{f_n}{g_n}\underset{n\to\infty}{\longrightarrow}0$ and $f_n\in\omega(g_n)$ means $\frac{f_n}{g_n}\underset{n\to\infty}{\longrightarrow}\infty$.

Throughout this paper we use the term "continuous" to refer to the continuous *real valued* channel $\mathbb{R}\to\mathbb{R}$, although this definition does not cover all continuous input—continuous output channels. By the term "discrete" in this paper we always refer to finite alphabets (as opposed to countable ones).

B. Definitions

Definition 1 (Channel): A channel is defined by a pair of input and output alphabets $\mathcal{X},\mathcal{Y}$, and is denoted $\mathcal{X}\to\mathcal{Y}$.

Definition 2 (Fixed Rate Encoder, Decoder, Error Probability): A randomized block encoder and decoder pair for the channel $\mathcal{X}\to\mathcal{Y}$ with block length n and rate R without feedback is defined by a random variable S distributed over the set $\mathcal{S}$, a mapping $\phi:\{1,2,\ldots\exp(nR)\}\times\mathcal{S}\to\mathcal{X}^n$ and a mapping $\bar{\phi}:\mathcal{Y}^n\times\mathcal{S}\to\{1,2,\ldots\exp(nR)\}$. The error probability for message $w\in\{1,2,\ldots\exp(nR)\}$ is defined as

$$P_e^{(w)}(\mathbf{x},\mathbf{y}) = \Pr\left(\bar{\phi}(\mathbf{y},S)\neq w|\phi(w,S)=\mathbf{x}\right), \qquad (3)$$

where for $\mathbf{x}$ such that the conditioning in (3) cannot hold, we define $P_e^{(w)}(\mathbf{x},\mathbf{y})=0$.

This system is illustrated in Fig. 1. We treat $\mathbf{x}$ as a random variable and $\mathbf{y}$ as a deterministic sequence. This does not preclude applying the results to a channel whose output $\mathbf{y}$ is a

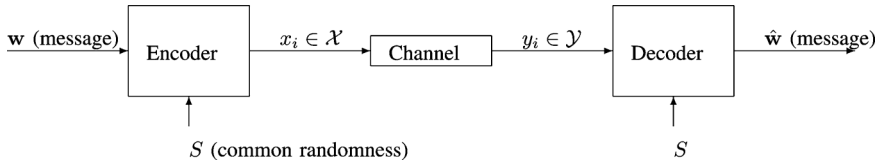

Fig. 1. Non rate adaptive encoder-decoder pair without feedback.

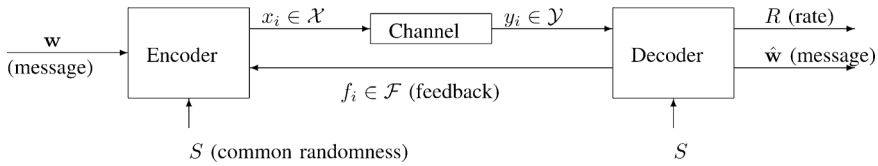

Fig. 2. Rate adaptive encoder-decoder pair with feedback.

random variable and depends on $\mathbf{x}$, since all results are conditioned on both $\mathbf{x}$ and $\mathbf{y}$. Note that the encoder rate must pertain to a discrete number of messages $\exp(nR) \in \mathbb{Z}_+$, but the empirical rates we refer to in the sequel may be any positive real numbers.

Definition 3 (Achievability): A rate function $R_{\text{emp}} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ is *achievable* with a prior $Q(\mathbf{x})$ defined over $\mathcal{X}^n$ and error probability ϵ if for any $R > 0$, there exist a pair of randomized encoder and decoder, with a rate of at least R such that for any $\mathbf{x}, \mathbf{y}$ where $R_{\text{emp}}(\mathbf{x}, \mathbf{y}) \geq R$ and any message w, $P_e^{(w)}(\mathbf{x}, \mathbf{y}) \leq \epsilon$.

Definition 4 (Adaptive Rate Encoder, Decoder, Error Probability): A randomized block encoder and decoder pair for the channel $\mathcal{X} \to \mathcal{Y}$ with block length n, adaptive rate and feedback is defined as follows:

- The message $\mathbf{w}$ is expressed by the infinite bit sequence $\mathbf{w}_1^\infty \in \{0,1\}^\infty$.
- The common randomness is defined as a random variable S distributed over the set $\mathcal{S}$.
- The feedback alphabet is denoted $\mathcal{F}$.
- The encoder is defined by a series of mappings $x_k = \phi_k(\mathbf{w}, s, \mathbf{f}^{k-1})$ where $\phi_k : \{0,1\}^\infty \times \mathcal{S} \times \mathcal{F}^{k-1} \to \mathcal{X}$.
- The decoder is defined by the feedback function $\varphi_k : \mathcal{Y}^{k-1} \times \mathcal{S} \to \mathcal{F}$, the decoding function $\bar{\phi} : \mathcal{Y}^n \times \mathcal{S} \to \{0,1\}^\infty$ and the rate function $r : \mathcal{Y}^n \times \mathcal{S} \to \mathbb{R}^+$ (where the rate is measured in bits), applied as follows:

$$f_k = \varphi_k(\mathbf{y}^k, S) \tag{4}$$
$$\hat{\mathbf{w}} = \bar{\phi}(\mathbf{y}, S) \tag{5}$$
$$R = r(\mathbf{y}, S) \tag{6}$$

The error probability for message $\mathbf{w}$ is defined as

$$P_e^{(\mathbf{w})}(\mathbf{x}, \mathbf{y}) = \Pr\left(\hat{\mathbf{w}}_1^{\lceil nR \rceil} \neq \mathbf{w}_1^{\lceil nR \rceil} \middle| \mathbf{x}, \mathbf{y}\right). \tag{7}$$

In other words, a recovery of the first $\lceil nR \rceil$ bits by the decoder is considered a successful reception. For $\mathbf{x}$ such that the conditioning in (7) cannot hold, we define $P_e^{(\mathbf{w})}(\mathbf{x}, \mathbf{y}) = 0$. The conditioning on $\mathbf{y}$ is mainly for clarification, since it is treated as a fixed vector. This system is illustrated in Fig. 2.

Note that if we are not interested in limiting the feedback rate, and perfect feedback can be assumed, the definition of feedback alphabet and feedback function is redundant (in this case $\mathcal{F} = \mathcal{Y}$ and $f_k = y_k$). The model in which the decoder determines the transmission rate is lenient in the sense that it gives the flexibility to exchange rate for error probability: the decoder may estimate the error probability and decrease it by reducing the decoding rate. In the scheme we discuss here the rate is determined during reception, but it's worth noting in this context the posterior matching scheme [9] for the known memoryless channel. In this scheme the message is represented as a real number $\theta \in [0,1)$ and the rate for a given error probability ϵ can be determined *after* reception by calculating $\Pr(\theta|\mathbf{y})$ and finding the smallest interval with probability at least $1 - \epsilon$.

Definition 5 (Adaptive Achievability): A rate function $R_{\text{emp}} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ is *adaptively achievable* with a prior $Q(\mathbf{x})$ defined over $\mathcal{X}^n$ and error probability ϵ, up to a subset $J \subset \mathcal{X}^n$, if there exist adaptive rate encoder and decoder with feedback such that $\mathbf{x} \sim Q$, and $\forall \mathbf{x} \notin J, \mathbf{y} \in \mathcal{Y}^n$:

$$\Pr\left\{\left(\hat{\mathbf{w}}_1^{\lceil nR \rceil} \neq \mathbf{w}_1^{\lceil nR \rceil}\right) \cup \left(R < R_{\text{emp}}(\mathbf{x}, \mathbf{y})\right) \middle| \mathbf{x}, \mathbf{y}\right\} \leq \epsilon. \tag{8}$$

In other words, with probability at least $1 - \epsilon$, a message with a rate of at least R_{emp} is decoded correctly.

Note that in the definition above we only require that for $\mathbf{x} \notin J, \mathbf{y} \in \mathcal{Y}^n$, $R \geq R_{\text{emp}}$ with a given probability, however in the two cases presented here, we show that $R \geq R_{\text{emp}}$ deterministically.

Definition 6 (Asymptotic Achievability): A sequence of rate functions defined for $n = 1, 2, \ldots$ is *asymptotically achievable* (adaptively/non adaptively) with a prior $Q(\mathbf{x})$ defined for vectors $\mathbf{x} \in \mathcal{X}^n$ of increasing size, if for all $\epsilon > 0$ there exists a sequence of functions $F_n(t)$, $n = 1, 2, \ldots$ with $F_n(t) \underset{n\to\infty}{\longrightarrow} t$, such that

1) For all n large enough, $F_n(R_{\text{emp}}(\mathbf{x}, \mathbf{y}))$ is achievable (adaptively/non adaptively, resp.) with the given ϵ and $Q(\mathbf{x})$.
2) In the adaptive case, the sequence of the sets J_n for which $F_n(R_{\text{emp}}(\mathbf{x}, \mathbf{y}))$ is adaptively achievable satisfies $\Pr(\mathbf{x} \in J) \underset{n\to\infty}{\longrightarrow} 0$.

IV. Fixed Rate Communication Without Feedback

In this section we show that the empirical mutual information (in the discrete case) and its Gaussian counterpart (in the continuous case) are asymptotically achievable. For the continuous case we justify the choice of the Gaussian distribution as the one yielding the maximum rate function that can be defined by second order moments.

A. The Discrete Channel Without Feedback

The following theorem formalizes the achievability of rate $\hat{I}(\mathbf{x};\mathbf{y})$ without feedback:

Theorem 1 (Non-Adaptive, Discrete Channel): Given discrete input and output alphabets $\mathcal{X},\mathcal{Y}$, for every $\epsilon>0$, $\delta>0$, prior $Q(x)$ over $\mathcal{X}$ and rate $R>0$ there exists n large enough and a random encoder-decoder pair of rate R over block size n, such that the distribution of the input sequence is $\mathbf{x}\sim Q^n$ and the probability of error for any message given an input sequence $\mathbf{x}\in\mathcal{X}^n$ and output sequence $\mathbf{y}\in\mathcal{Y}^n$ is not greater than ϵ if $\hat{I}(\mathbf{x};\mathbf{y})>R+\delta$.

Corollary 1: $R_{\text{emp}}=\hat{I}(\mathbf{x};\mathbf{y})$ is asymptotically achievable.

Theorem 1 follows almost immediately from the following lemma, which is proven in the Appendix using a simple calculation based on the method of types [10]:

Lemma 1: For any sequence $\mathbf{y}\in\mathcal{Y}^n$ the probability of a sequence $\mathbf{x}\in\mathcal{X}^n$ drawn independently according to Q^n to have $\hat{I}(\mathbf{x};\mathbf{y})\geq t$ is upper bounded by:

$$Q^n\left(\hat{I}(\mathbf{x};\mathbf{y})\geq t\right)\leq\exp\left(-n\left(t-\delta_n\right)\right), \tag{9}$$

where $\delta_n=|\mathcal{X}||\mathcal{Y}|\frac{\log(n+1)}{n}\to 0$.

Following notations used by Csiszár [10], $Q^n(A)$ denotes the probability of the event A or equivalently the set of sequences A under the i.i.d. distribution Q^n. Remarkably, the bound (9) does not depend on Q.

To prove Theorem 1, a codebook $\{\mathbf{x}_m\}_{m=1}^{\exp(nR)}$ is randomly generated by i.i.d. selection of its $L=\exp(nR)\cdot n$ letters. The common randomness $S\in\mathcal{X}^L$ is defined as the codebook itself and is distributed Q^L. The encoder sends the w-th codeword, and the decoder uses maximum mutual information decoding (MMI) i.e., chooses:

$$\hat{w}=\bar{\phi}(\mathbf{y},S)=\underset{m}{\operatorname{argmax}}\left[\hat{I}(\mathbf{x}_m;\mathbf{y})\right], \tag{10}$$

where ties are broken arbitrarily. If the message w was transmitted then $\mathbf{x}=\mathbf{x}_w$. Since the codewords are independent, conditioning on $\mathbf{x}$ does not change the distribution of the other codewords. By Lemma 1 and the union bound, the probability of error is bounded by:

$$\begin{aligned}P_e^{(w)}(\mathbf{x}_w,\mathbf{y})&\leq\Pr\left\{\bigcup_{m\neq w}\left(\hat{I}(\mathbf{x}_m;\mathbf{y})\geq\hat{I}(\mathbf{x}_w;\mathbf{y})\right)\middle|\,\mathbf{x}_w\right\}\\&\leq\exp(nR)\exp\left(-n\left(\hat{I}(\mathbf{x}_w;\mathbf{y})-\delta_n\right)\right)\\&=\exp\left(-n\left(\hat{I}(\mathbf{x}_w;\mathbf{y})-R-\delta_n\right)\right),\end{aligned} \tag{11}$$

where the probabilities above are with respect to the common randomness S (note that all codewords are random i.i.d., except $\mathbf{x}_w$ which is in the conditioning). For any δ there is n large enough such that $\frac{-\log(\epsilon)}{n}+\delta_n<\delta$. For this n, whenever $\hat{I}(\mathbf{x};\mathbf{y})>R+\delta$ we have

$$P_e^{(w)}(\mathbf{x},\mathbf{y})\leq\exp\left(-n\left(\delta-\delta_n\right)\right)<\epsilon, \tag{12}$$

which proves the theorem. □

Note that the MMI decoder used here is a popular universal decoder [5], [10], [11], and was shown to achieve the same error exponent as the maximum likelihood decoder for the discrete memoryless channel (DMC) with fixed composition codes. The error exponent obtained here (11) is better than the classical error exponent (slope of -1), and the reason is that the behavior of the channel is a-posteriori known, and therefore no errors occur as a result of non-typical channel behavior. Comparing for example with the derivation of the random coding error exponent for the probabilistic DMC [10] based on the method of types, in the later the error probability is summed across all potential "behaviors" (conditional types) of the channel accounting for their respective probabilities, resulting in one behavior, usually different from the typical behavior, dominating the bound. Here the behavior of the channel (the conditional distribution) is fixed, and therefore the error exponent is better. This relates to Hughes and Thomas' observation [12, Theorem 4] that the error exponent in an AVC with a constraint (on the empirical distribution of the state sequence) is in general better than the error exponent of a compound channel where the constraint applies to the (non empirical) distribution of the state sequence, which they explain by the fact that in AVC the constraint is strict. The error rate obtained here is not necessarily the best rate that can be achieved. It is known that random decision time and feedback may improve the error exponent for probabilistic and compound models [11], [13].

Note that the empirical mutual information is always well defined, even when some of the input and output symbols do not appear in the sequence, since at least one input symbol and one output symbol always appear. For the particular case of empirical mutual information measured over a single symbol, the empirical distributions become unit vectors (representing constants) and their mutual information is 0.

In this discussion we have not dealt with the issue of choosing the prior $Q(x)$. Since the channel behavior is unknown it makes sense to choose the maximum entropy, i.e., the uniform prior, which was shown to obtain a bounded loss from capacity [14].

B. The Continuous Channel Without Feedback

When turning to define empirical rates for the real valued alphabet case, the first obstacle we tackle is the definition of the empirical distribution. A potential approach is to use discrete approximations and turn the problem into the discrete problem discussed above (while limiting the growth rate of the discrete alphabet to still attain $\delta_n\to 0$), however we do not pursue this approach since it is somewhat arbitrary. We focus on empirical rates defined by the correlation factor. Although the later

approach is pessimistic and falls short of the mutual information for most channels, it is much simpler and elegant than discrete approximations. We believe this approach can be further extended to obtain results closer to the (probabilistic) mutual information.

1) Choosing the Input Distribution and Rate Function: First we justify our choice of the Gaussian input distribution and the aforementioned rate function. We take the point of view of a memoryless compound (probabilistic, unknown) channel. If a rate function cannot be attained for compound channel model, it cannot be attained also in the more stringent individual model. It is well known that for a memoryless additive noise channel with constraints on the transmit power and noise variance, the Gaussian noise is the worst noise when the prior is Gaussian, and the Gaussian prior is the best prior when the noise is Gaussian. Thus by choosing a Gaussian prior we choose the best prior for the worst noise, and can we guarantee the mutual information will equal, at least, the Gaussian channel capacity [15, Problem 9.21]. For the additive noise channel with an arbitrary i.i.d. noise distribution, Zamir and Erez [16] showed that the loss from capacity when using Gaussian distribution is limited to $\frac{1}{2}$ a bit. However the above is true only for additive noise channels. For the more general case where no additivity is assumed we show below (Lemma 3) that the rate function $R = -\frac{1}{2}\log(1-\rho^2)$ is the best rate function that can be defined by second order moments, and attained universally. Of course, this proof merely supplies the motivation to use a Gaussian distribution and does not rid us from the need to prove this rate is achievable for specific, individual sequences. We use the following technical lemma:

Lemma 2: Let X,Y be two continuous random variables with correlation factor $\rho \triangleq \frac{\mathbb{E}(XY)}{\sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}}$, where X is Gaussian $X \sim \mathcal{N}(0,P)$. Then $I(X;Y) \geq -\frac{1}{2}\log(1-\rho^2)$

Corollary 2: Equality holds iff X, Y are jointly Gaussian

Remark 3: The lemma does not hold for general X (not Gaussian)

The proof is given in the Appendix. Note that $-\frac{1}{2}\log(1-\rho^2)$ is the mutual information between two jointly Gaussian r.v-s [15, Example 8.5.1]. Also note the relation to Hassibi and Hochwald's result [17, Theorem 1] dealing with an additive channel with uncorrelated, but not necessarily independent noise. The following lemma justifies our selection of $R(\rho) = -\frac{1}{2}\log(1-\rho^2)$:

Lemma 3: Let $Q(x)$ be an input prior, $W(y\,|\,x)$ be an unknown channel, $\Lambda(Q,W)$ be the correlation matrix $\Lambda \triangleq \mathbb{E}\begin{pmatrix} X \\ Y \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}^T$ between X, Y induced by the joint probability $Q \circ W$ and $\rho(Q,W)$ be the correlation factor induced by Q, W ($\rho = \frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}}$). Then $R(\Lambda) = -\frac{1}{2}\log(1-\rho^2)$ is the largest function of Λ that satisfies the following condition: there exists a $Q(x)$ such that for every channel $W(y\,|\,x)$ inducing correlation Λ the mutual information is at least $R(\Lambda)$ (in other words all channels with such a correlation matrix can carry the rate $R(\Lambda)$). Alternatively this can be stated as:

$$R(\Lambda) \triangleq \max_{Q} \min_{W:\Lambda(Q,W)=\Lambda} I(Q,W) = -\frac{1}{2}\log(1-\rho^2). \tag{13}$$

Proof of Lemma 3: $R(\Lambda) = -\frac{1}{2}\log(1-\rho^2)$ satisfies the condition by selecting an input prior $Q = \mathcal{N}(0,P)$ and by Lemma 2 the mutual information is at least $R(\Lambda)$ for all channels. On the other hand, any function $R'(\Lambda)$ satisfying the conditions of the lemma satisfies $R'(\Lambda) \leq -\frac{1}{2}\log(1-\rho^2)$, since by writing the condition of the lemma for the additive white Gaussian noise (AWGN) channel W^* (a specific choice of W) and any Q, we have

$$\begin{aligned} R'(\Lambda) &\leq I(Q,W^*) \leq I(\mathcal{N}(0,E_Q(X^2)),W^*) \\ &= -\frac{1}{2}\log(1-\rho^2), \end{aligned} \tag{14}$$

where the inequalities follow from the conditions of the lemma and from the fact the Gaussian prior achieves the AWGN capacity. □

Note that since the mutual information between two Gaussian r.v-s is $-\frac{1}{2}\log(1-\rho^2)$, one can think of this value as a measure of mutual information under Gaussian assumptions. In the sequel we sometimes use the term "empirical mutual information" in a broad sense that includes also the metric $-\frac{1}{2}\log(1-\hat{\rho}^2)$.

2) A Communication Scheme for the Individual Channel: The following theorem is the analogue of Theorem 1 where the expression $-\frac{1}{2}\log(1-\rho^2)$ (interpreted as the Gaussian effective mutual information) plays the role of mutual information.

Theorem 2 (Non-Adaptive, Continuous Channel): Given the channel $\mathbb{R} \to \mathbb{R}$ for every $\epsilon > 0$, $\delta > 0$, power $P > 0$ and rate $R > 0$ there exists n large enough and a random encoder-decoder pair of rate R over block size n, such that the distribution of the input sequence is $\mathbf{x} \sim \mathcal{N}^n(0,P)$ and the probability of error for any message given an input sequence $\mathbf{x}$ and output sequence $\mathbf{y}$ with empirical correlation $\hat{\rho}$ is not greater than ϵ if $R_{\text{emp}} \triangleq \frac{1}{2}\log\left(\frac{1}{1-\hat{\rho}^2}\right) > R + \delta$

Corollary 4: R_{emp} is asymptotically achievable.

As before, the theorem will follow easily from the following lemma, proven in the Appendix.

Lemma 4: Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be two sequences, and let

$$\hat{\rho}(\mathbf{x},\mathbf{y}) \triangleq \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \tag{15}$$

be the empirical correlation factor. For any $\mathbf{y}$, the probability of $\mathbf{x}$ drawn according to $\mathcal{N}^n(0,P)$ to have $|\hat{\rho}| \geq t$ is bounded by:

$$\Pr\{|\hat{\rho}| \geq t\} \leq 2\exp\left(-(n-1)R_2(t)\right), \tag{16}$$

where

$$R_2(t) \triangleq \frac{1}{2}\log\left(\frac{1}{1-t^2}\right). \tag{17}$$

To prove Theorem 2, the codebook $\{\mathbf{x}_m\}_{m=1}^{\exp(nR)}$ is randomly generated by Gaussian i.i.d. selection of its $L = \exp(nR) \cdot n$ letters, and the common randomness $S \in \mathcal{X}^L$ is defined as the codebook itself and is distributed $\mathcal{N}^L(0, P)$. The encoder sends the w-th codeword, and the decoder uses maximum empirical correlation decoder i.e., chooses:

$$\hat{w} = \bar{\phi}(\mathbf{y}, S) = \underset{m}{\operatorname{argmax}} \, |\hat{\rho}(\mathbf{x}_m; \mathbf{y})| = \underset{m}{\operatorname{argmax}} \left[\frac{|\mathbf{x}_m^T \mathbf{y}|}{\|\mathbf{x}_m\|} \right], \tag{18}$$

where ties are broken arbitrarily. By Lemma 4 and the union bound, the probability of error is bounded by:

$$\begin{aligned} P_e^{(w)}(\mathbf{x}_w, \mathbf{y}) &\le \Pr\left\{ \bigcup_{m \neq w} \left(|\hat{\rho}(\mathbf{x}_m; \mathbf{y})| \ge |\hat{\rho}(\mathbf{x}_w; \mathbf{y})| \right) \middle| \, \mathbf{x}_w \right\} \\ &\le \exp(nR) \cdot 2 \exp\left(-(n-1) R_2(\hat{\rho}(\mathbf{x}_w; \mathbf{y})) \right) \\ &= 2 \exp(R) \cdot \exp\left(-(n-1) \left(R_2(\hat{\rho}) - R \right) \right), \end{aligned} \tag{19}$$

where the probabilities are with respect to the common randomness S. Choosing n large enough so that $\frac{1}{n-1}\left(R + \log\left(\frac{2}{\epsilon}\right)\right) \le \delta$ (where ϵ is from Theorem 2) we have that when $R_2(\hat{\rho}) > R + \delta$:

$$P_e^{(w)}(\mathbf{x}, \mathbf{y}) \le 2 \exp(R) \cdot \exp\left(-(n-1)\delta \right) \le \epsilon, \tag{20}$$

which proves the theorem. □

A note is due regarding the definition of $\hat{\rho}$ in singular cases where $\mathbf{x}$ or $\mathbf{y}$ are 0. The limit of $\hat{\rho}$ as $\mathbf{y} \to 0$ is undefined (the directional derivative may take any value in [0,1]), however for consistency we define $\hat{\rho} = 0$ when $\mathbf{y} = 0$. Since $\mathbf{x}$ is generated from a Gaussian distribution we do not worry about the event $\mathbf{x} = 0$ since the probability of this event is 0. We refer the reader to Section VIII-B, for a discussion of the connection between the decoding rules used in the discrete and continuous cases.

Combining Lemma 4 with the law of large numbers provides a simple proof for the achievability of the AWGN capacity ($\frac{1}{2}\log(1 + SNR)$), which uses more elementary tools than the popular proofs based on AEP or error exponents.

V. Rate Adaptive Communication With Feedback

In this section we present the rate-adaptive counterparts of Theorems 1, 2, and the scheme achieving them. The proof is delayed to the next section. The scheme we use in order to adaptively attain these rates is by iterating a rateless coding scheme. In other words, in each iteration we send a fixed number of bits K, by transmitting symbols from an n length codebook, until the receiver has enough information to decode. Then, the receiver sends an indication that the block is over and a new block starts. For background on rateless codes and comparisons with other schemes refer to Section VIII-C.

A. Statement of the Main Result

In this section and the next, we prove the following theorems, relating to the definitions given in Section III-B:

Theorem 3 (Rate Adaptive, Discrete Channels): Given discrete input and output alphabets $\mathcal{X}, \mathcal{Y}$, for every $\epsilon > 0$, $P_J > 0$, $\delta > 0$ and prior $Q(x)$ over $\mathcal{X}$ there is n large enough and random encoder and decoder with feedback and variable rate over block size n with a subset $J \subset \mathcal{X}^n$, such that:

- The distribution of the input sequence is $\mathbf{x} \sim Q^n$ independently of the feedback and the message
- The probability of error is smaller than ϵ for any $\mathbf{x}, \mathbf{y}$
- For any input sequence $\mathbf{x} \notin J$ and output sequence $\mathbf{y} \in \mathcal{Y}^n$ the rate is $R \ge \hat{I}(\mathbf{x}; \mathbf{y}) - \delta$
- The probability of J is bounded by $\Pr(\mathbf{x} \in J) \le P_J$

Corollary 5: $R_{\text{emp}} = \hat{I}(\mathbf{x}; \mathbf{y})$ is asymptotically adaptively achievable.

Theorem 4 (Rate Adaptive, Continuous Channels): Given the channel $\mathbb{R} \to \mathbb{R}$ for every $\epsilon > 0$, $P_J > 0$, $\delta > 0$, $\bar{R} > 0$, and power $P > 0$ there is n large enough and random encoder and decoder with feedback and variable rate over block size n with a subset $J \subset \mathbb{R}^n$, such that

- The distribution of the input sequence is $\mathbf{x} \sim \mathcal{N}(0, P)^n$ independently of the feedback and the message
- The probability of error is smaller than ϵ for any $\mathbf{x}, \mathbf{y}$
- For any input sequence $\mathbf{x} \notin J$ and output sequence $\mathbf{y} \in \mathbb{R}^n$ the rate is $R \ge \min\left[\frac{1}{2} \log\left(\frac{1}{1 - \hat{\rho}(\mathbf{x},\mathbf{y})^2} \right) - \delta, \bar{R} \right]$
- The probability of J is bounded by $\Pr(\mathbf{x} \in J) \le P_J$

Corollary 6: $R_{\text{emp}} = \frac{1}{2} \log\left(\frac{1}{1 - \hat{\rho}(\mathbf{x},\mathbf{y})^2} \right)$ is asymptotically adaptively achievable.

We prove the two theorems together. First we define the scheme, and make some comments on the achievability results. In the next section we analyze the error performance and the rate and show the proposed scheme achieves the promise of the theorems.

B. A Proposed Rate Adaptive Scheme

The following communication scheme sends B blocks of K bits each, over n channel uses, where K is fixed, and B, which is the number of blocks, varies according to empirical channel behavior. The building block is a rateless transmission of K bits which is iterated until the n-th symbol is reached. Throughout this section and the following one we use n to denote the length of a complete transmission, and m to denote the length of a single block.

The transmit distribution Q is an arbitrary distribution for the discrete case and $Q = \mathcal{N}(0, P)$ for the continuous case. The empirical rate is:

$$R_{\text{emp}}(\mathbf{x}, \mathbf{y}) \triangleq \begin{cases} \hat{I}(\mathbf{x}; \mathbf{y}) & \text{discrete} \\ \frac{1}{2} \log\left(\frac{1}{1 - \hat{\rho}^2(\mathbf{x},\mathbf{y})} \right) & \text{continuous} \end{cases}. \tag{21}$$

The empirical rate is also used as a decoding metric. The codebook $C_{M \times n}$ consists of $M = \exp(K)$ codewords of length n, where all $M \times n$ symbols are drawn i.i.d. $\sim Q$ and known to the sender and receiver. k denotes the absolute time index $1 \le k \le n$. Block b starts from index k_b, where $k_1 = 1$. $m = k - k_b + 1$ denotes the time index inside the current block. For brevity of notation we denote the rate function measured

over m symbols ending at the current time k as $R_{\text{emp}}^{(m,k)}(\mathbf{x},\mathbf{y}) \triangleq R_{\text{emp}}(\mathbf{x}_{k-m+1}^{k},\mathbf{y}_{k-m+1}^{k})$.

In each rateless block $b = 1,2,\ldots$, a new index $i = i_b \in \{1,\ldots,M\}$ is sent to the receiver using the following procedure:

1) The encoder sends index i by sending the symbols of codeword i:
$$x_k = C_{i,k}. \tag{22}$$
Note that different blocks use different symbols from the codebook.
2) The encoder keeps sending symbols and incrementing k until the decoder announces the end of the block through the feedback link.
3) The decoder announces the end of the block after symbol m in the block if for any codeword x_i:
$$R_{\text{emp}}^{(m,k)}(\mathbf{x}_i,\mathbf{y}) \triangleq R_{\text{emp}}\left((\mathbf{x}_i)_{k_b}^{k},\mathbf{y}_{k_b}^{k}\right) \geq \mu_m^*, \tag{23}$$
where μ_m^* is a fixed threshold per symbol defined in (24) below.
4) When the end of block is announced one of the i fulfilling (23) is determined as the index of the decoded codeword $\hat{i}_b$ (breaking ties arbitrarily).
5) Otherwise the transmission continues, until the n-th symbol is reached. If symbol n is reached without fulfilling (23), then the last block is terminated without decoding.

After a block ends, b is incremented and if $k < n$ a new block starts at symbol $k_b = k+1$. After symbol n is reached the transmission stops and the number of blocks sent is $B = b-1$.

The threshold μ_m^* is defined as:
$$\begin{aligned}\mu_m^* &= \frac{K}{m-s} + \frac{1}{m-s}\log\left(\frac{n}{\epsilon}\right) + \delta_m \\ &= \begin{cases} \frac{K+\log\left(\frac{n}{\epsilon}\right)+|\mathcal{X}||\mathcal{Y}|\log(m+1)}{m} & \text{discrete} \\ \frac{K+\log\left(\frac{2n}{\epsilon}\right)}{m-1} & \text{continuous} \end{cases}, \end{aligned} \tag{24}$$
where
$$\delta_m = \begin{cases} |\mathcal{X}||\mathcal{Y}|\frac{\log(m+1)}{m} & \text{discrete} \\ \frac{\log 2}{m-1} & \text{continuous} \end{cases} \tag{25}$$
and
$$s = \begin{cases} 0 & \text{discrete} \\ 1 & \text{continuous} \end{cases}. \tag{26}$$

The threshold μ_m^* is tailored to achieve the designated error probability and is composed of 3 parts. The first part requires that the empirical rate R_{emp} would approximately equal the transmission rate of the block $\frac{K}{m}$, which guarantees there is approximately enough mutual information to send K bits. The second part is an offset responsible for guaranteeing an error probability bounded by ϵ over all the blocks in the transmission. The third part δ_m compensates the overhead terms in Lemmas 1,4.

The scheme achieves the claims of Theorems 3,4 with a proper choice of the parameters (discussed in Section VI-C). Note that the scheme uses feedback rate of 1 bit/use however it is easy to modify it to use any positive feedback rate (see Section VIII-G), therefore we can claim the theorems hold with "zero rate" feedback.

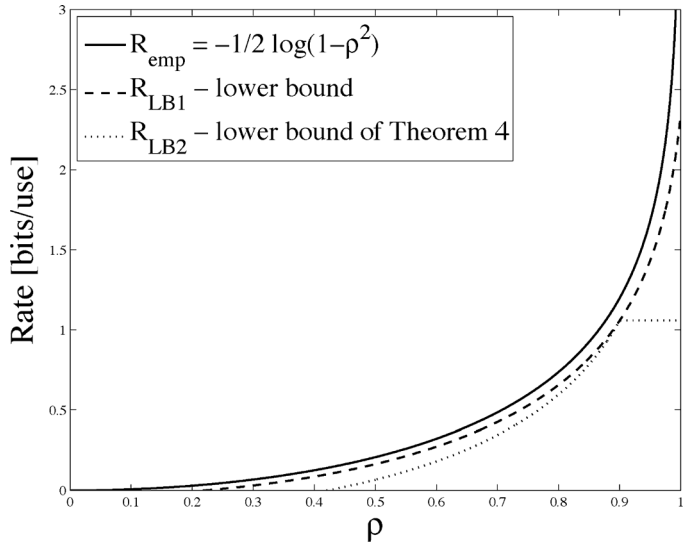

Fig. 3. Illustration of R_{emp} lower bound of Theorem 4 (R_{LB2}) and the lower bound R_{LB1} shown in the proof in Section VI-C2, as a function of $\hat{\rho}$. See the parameters in Table III in the Appendix.

C. Comments on the Results

In Theorem 4 we do not have uniform convergence of the rate function in $\mathbf{x},\mathbf{y}$, as opposed to other results in this paper. Unfortunately our scheme is limited by having a maximum rate for each n, and although the maximum rate tends to infinity as $n \to \infty$, we cannot guarantee uniform convergence for each n in the continuous case, where the target rate may be unbounded. The rates in the theorems are the minimal rates, and in certain conditions (e.g., a channel varying in time) higher rates may be achieved by the scheme.

Fig. 3 illustrates the lower bound for R_{emp} presented by Theorem 4 for a specific choice of parameters. The solid line presents $R_{\text{emp}} = -\frac{1}{2}\log(1-\hat{\rho}^2)$. Below it, R_{LB1} is a lower bound for the rate achieved by the proposed scheme (Section VI-C2, (65)). Observe that R_{LB1} follows the trend of R_{emp}, but is slightly lower, and reaches a bounded rate for $\hat{\rho} = 1$, where R_{emp} is unbounded. $R_{\text{LB2}} \leq R_{\text{LB1}}$ is a lower bound on R_{LB1} which obeys the structure defined in Theorem 4. The parameters generating these curves appear in Table III in the Appendix.

Regarding the set J as we shall see in the sequel there are some sequences for which poor rate is obtained, and since we committed to an input distribution we cannot avoid them (one example is the sequence of $\frac{1}{2}n$ zeros followed by $\frac{1}{2}n$ ones, in which as we shall see, at most one block will be sent). However there is an important distinction between the claim made in the theorems that "A failure may happen only when $\mathbf{x}$ belongs to a subset J with probability at most P_J", and a simpler, but a weaker claim that "For each $\mathbf{y}$ the probability of failure is at most P_J". This is demonstrated in Fig. 4, where each gray box indicates that $\mathbf{x}$ is a bad sequence for a specific $\mathbf{y}$. In Fig. 4(a) the probability of a bad sequence for each $\mathbf{y}$ is small, however for each $\mathbf{x}$ there is a $\mathbf{y}$ such that this $\mathbf{x}$ is bad for that $\mathbf{y}$, and

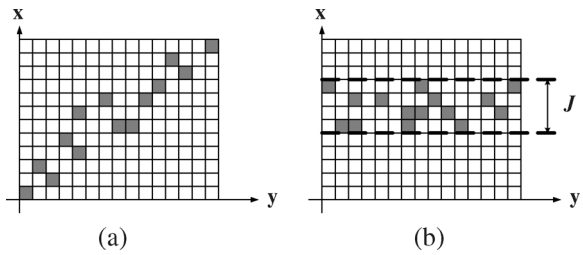

Fig. 4. Illustration of bad sequences and Lemma 5. In (a) the probability of bad sequences, denoted by grey boxes, is low for each $\mathbf{y}$, whereas in (b) in addition, the set J of $\mathbf{x}$ for which bad sequences may occur has low probability irrespective of $\mathbf{y}$.

therefore by choosing $\mathbf{y}$ as a function of $\mathbf{x}$ one may increase the failure probability to 1. In Fig. 4(b) the set of bad sequences in $\mathbf{x}$ is limited independently of $\mathbf{y}$, therefore this is avoided. A consequence of the definition that a failure may only happen if $x \in J$ is that the failure probability is bounded by P_J for any conditional probability $\Pr(\mathbf{y} \mid \mathbf{x})$ on the sequences. This issue is further discussed in Section VI-A.

Note that the probability P_J could be absorbed into ϵ by a simple trick, but this seems to make the theorem less insightful. After reception the receiver knows the input sequence with probability of at least $1-\epsilon$ and may calculate the empirical mutual information $\hat{I}(\mathbf{x};\mathbf{y})$. If the rate achieved by the scheme falls short of $\hat{I}(\mathbf{x};\mathbf{y})$ it may declare a rate of $R = \hat{I}(\mathbf{x};\mathbf{y})$ (which will most likely result in a decoding error). This way the receiver will never declare a rate which is lower than $\hat{I}(\mathbf{x};\mathbf{y})$ unless there is an error, and we could avoid the restriction $\mathbf{x} \notin J$ required for achieving R_{emp}, but on the other hand, the error probability becomes conditioned on the set J.

VI. Proof of the Main Result

In this section we analyze the adaptive rate scheme presented and show it achieves Theorems 3,4. Before analyzing the scheme we develop some general results pertaining to the convexity of the mutual information and correlation factors over sub-vectors. The proof of the error probability is common to the two cases, while the analysis of the achieved rate is performed separately for each case.

A. Preliminaries

1) Likely Convexity of the Mutual Information: A property which would be useful for the analysis is $\cup$-convexity of the empirical mutual information with respect to joint empirical distributions $\hat{P}_{(\mathbf{x},\mathbf{y})}(x,y)$ measured over different sub-vectors. The main application of such a property is to show that obtaining the empirical mutual information over each sub-block in the rateless scheme, yields a rate equal at least to the empirical mutual information measured over the entire transmission. Had the rate been averaged over multiple sequences $\mathbf{x}$ rather than obtained for a specific sequence, the regular convexity of the mutual information with respect to the channel distribution would have been sufficient (as in the case of an individual state sequence [4]). However for a specific sequence, this property does not hold. Instead, we show the desired convexity approximately holds, apart from a vanishing set of input sequences. The property is formalized in the following lemma:

Lemma 5 (Likely Convexity of Mutual Information): Let $\{A_i\}_{i=1}^{p}$ define a disjoint partitioning of the index set $\{1,\ldots,n\}$ into p subsets, i.e., $\bigcup_i A_i = \{1,\ldots,n\}$ and $A_i \cap A_j = \emptyset$ for $i \neq j$. $\mathbf{x}$, $\mathbf{y}$ are n-length sequences, and $\mathbf{x}_A$, $\mathbf{y}_A$ denote the sub-sequences of $\mathbf{x}$, $\mathbf{y}$ (resp.) over the index set A. Let the elements of $\mathbf{x}$ be chosen i.i.d. with distribution Q. Then for any $\Delta > 0$ there is a subset $J_\Delta \subset \mathcal{X}^n$ such that:

$$\forall \mathbf{x} \notin J_\Delta, \mathbf{y} \in \mathcal{Y}^n : \sum_{i=1}^{p} \frac{|A_i|}{n} \hat{I}(\mathbf{x}_{A_i};\mathbf{y}_{A_i}) \geq \hat{I}(\mathbf{x};\mathbf{y}) - \Delta, \tag{27}$$

and

$$Q^n\{J_\Delta\} \leq \exp\left(-n\left(\Delta - \tilde{\delta}_n\right)\right), \tag{28}$$

with $\tilde{\delta}_n = p \cdot |\mathcal{X}| \cdot \frac{\log(n+1)}{n} \to 0$.

The lemma does not claim that convexity holds with high probability, but rather that any positive deviation from convexity may happen only on a subset of $\mathbf{x}$ with vanishing probability. Surprisingly, the bound does not depend on $\mathbf{y}$, Q and the size of the subsets, and only weakly depends on the number of subsets. In the sequel we use the lemma to:

1) Show that the average rate (empirical mutual information) over multiple blocks equals at least the mutual information measured over the blocks together.
2) Bound the loss due to insufficient utilization of the last symbol in each rateless encoding block.
3) Bound the loss due to not completing the last rateless encoding block.

Note that the set J_Δ in Lemma 5 corresponds to the set J in Theorem 3. At this point, we can return to the discussion regarding the set J in Section V-C and make it more concrete. We can see that the scheme would fail for the sequence $\mathbf{x}$ of $\frac{1}{2}n$ zeros followed by $\frac{1}{2}n$ ones. This sequence guarantees that at most one block will be received (since at most one block includes both 0-s and 1-s at the input which is necessary for having $\hat{I} > 0$). On the other hand the zero order empirical input distribution of this sequence is good $(\text{Ber}(\frac{1}{2}))$, and by setting $\mathbf{y} = \mathbf{x}$ we can have $\hat{I}(\mathbf{x};\mathbf{y}) = 1$, i.e., high empirical mutual information together with a low communication rate. This sequence belongs to the set J_Δ where the "likely convexity" does not hold. As was stressed in Section V-A and illustrated in Fig. 4, it is important that Lemma 5 is stated in terms of a failure set in $\mathbf{x}$, rather than in terms of the probability of failure for each $\mathbf{y}$ separately.

Proof of Lemma 5: Define the vector $\mathbf{u}$ denoting the subset number of each element $\mathbf{u}_k = i \ \forall k \in A_i$. Then the empirical distribution of $\mathbf{u}$ is $\hat{P}_{\mathbf{u}}(i) = \frac{|A_i|}{n}$ and $\hat{I}(\mathbf{x}_{A_i};\mathbf{y}_{A_i}) = \hat{I}(\mathbf{x};\mathbf{y} \mid \mathbf{u} = i)$ (refer to the definitions in Section III-A). Therefore we can write the weighted sum of empirical mutual information over the partitions, as a conditional empirical mutual information:

$$\begin{aligned}\sum_{i=1}^{p}\left(\frac{|A_i|}{n}\hat{I}(\mathbf{x}_{A_i};\mathbf{y}_{A_i})\right) &= \sum_{i=1}^{p}\hat{P}_{\mathbf{u}}(i)\hat{I}(\mathbf{x};\mathbf{y} \mid \mathbf{u} = i) \\ &= \hat{I}(\mathbf{x};\mathbf{y} \mid \mathbf{u}).\end{aligned} \tag{29}$$

Using the chain rule for mutual information [15, Sec. 2.5]:

$$\begin{aligned}\hat{I}(\mathbf{x};\mathbf{y})-\hat{I}(\mathbf{x};\mathbf{y}|\mathbf{u}) &= \hat{I}(\mathbf{x};\mathbf{y})-\left(\hat{I}(\mathbf{x};\mathbf{y}\mathbf{u})-\hat{I}(\mathbf{x};\mathbf{u})\right)\\ &= \hat{I}(\mathbf{x};\mathbf{u})-\hat{I}(\mathbf{x};\mathbf{u}|\mathbf{y})\le \hat{I}(\mathbf{x};\mathbf{u}).\end{aligned} \tag{30}$$

Define the set $J_\Delta=\{\mathbf{x}:\hat{I}(\mathbf{x};\mathbf{u})>\Delta\}$, then

$$\forall \mathbf{x}\notin J_\Delta,\mathbf{y}:\hat{I}(\mathbf{x};\mathbf{y})-\hat{I}(\mathbf{x};\mathbf{y}|\mathbf{u})\le\hat{I}(\mathbf{x};\mathbf{u})\le\Delta, \tag{31}$$

and since $\mathbf{x}$ is chosen i.i.d. and $\mathbf{u}$ is a fixed vector, we have from Lemma 1:

$$\Pr\left(\mathbf{x}\in J_\Delta\right)\le\exp\left(-n\left(\Delta-\tilde{\delta}_n\right)\right), \tag{32}$$

with $\tilde{\delta}_n=|\mathcal{X}||\{1,\ldots,p\}|\frac{\log(n+1)}{n}$. □

Note that if the empirical distribution of $\mathbf{x}$ is the same over all partitions then $\hat{H}(\mathbf{x}|\mathbf{u})=\hat{H}(\mathbf{x})$ therefore $\hat{I}(\mathbf{x};\mathbf{u})=0$ and the empirical mutual information will be truly convex.

2) Likely Convexity of the Correlation Factor: For the continuous case we use the following property which somewhat parallels Lemma 5. The reasons for not following the same path as the discrete case will be explained in the sequel (Subsection VI-C). The proof appears in Appendix E. Note that again the bound does not depend on the size of the subsets.

Lemma 6 (Likely Convexity of $\hat{\rho}^2$): Define $\{A_i\}_{i=1}^p$ as in Lemma 5. Let $\mathbf{x}$, $\mathbf{y}$ be n-length sequences and define the correlation factors of the sub-sequences, and the overall correlation factor as

$$\hat{\rho}_i=\frac{\left|\mathbf{x}_{A_i}^T\mathbf{y}_{A_i}\right|}{\|\mathbf{x}_{A_i}\|\cdot\|\mathbf{y}_{A_i}\|}\quad\text{and}\quad\hat{\rho}=\frac{\left|\mathbf{x}^T\mathbf{y}\right|}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|}, \tag{33}$$

respectively. Let $\mathbf{x}$ be drawn i.i.d from a Gaussian distribution $\mathbf{x}\sim\mathcal{N}(0,P)$. Then for any $0<\Delta\le\frac{1}{7}$ there is a subset $J_\Delta\subset\mathbb{R}^n$ such that:

$$\forall\mathbf{x}\notin J_\Delta,\mathbf{y}\in\mathbb{R}^n:\sum_{i=1}^p\frac{|A_i|}{n}\hat{\rho}_i^2\ge\hat{\rho}^2-\Delta, \tag{34}$$

and

$$\Pr\{\mathbf{x}\in J_\Delta\}\le 2^p e^{-n\Delta^2/8}. \tag{35}$$

In other words, there is a subset with high probability on which the mean of the correlation factors does not fall considerably below the overall correlation factor.

3) Likely Convexity With Dependencies: The properties of likely convexity defined in the previous sections pertain to a case where the partition of the n block is fixed and $\mathbf{x}$ is drawn i.i.d. However in the transmission scheme we described, the partition varies in a way that depends on the value of $\mathbf{x}$ (through the decoding decisions and the empirical mutual information), which may, in general, change the probability of the convexity property with a given Δ to occur. Although it stands to reason that the variability of the block sizes in the decoding process reduces the probability to deviate from convexity since it tends to equalize the amount of mutual information in each rateless block, for the analysis we assume an arbitrary dependence, and assume that the size of the set J increases by factor of the number of possible partitions, as explained below.

Denote a partition by $\pi=\{A_i\}_{i=1}^p$ (as defined in Lemmas 5, 6) and the group of all possible partitions (for a given encoder-decoder) by Π. We assume that the partition $\pi\in\Pi$ is selected arbitrarily. For each partition π from Lemmas 5, 6 there is a subset $J(\pi)$ with probability bounded by p_J outside which approximate convexity (as defined in the lemmas) holds. Then approximate convexity is guaranteed to hold for $\mathbf{x}\notin J\triangleq\bigcup_{\pi\in\Pi}J(\pi)$, where the probability of the set J is bounded by the union bound:

$$\Pr(\mathbf{x}\in J)=\Pr\left(\bigcup_{\pi\in\Pi}(\mathbf{x}\in J(\pi))\right)\le|\Pi|\cdot p_J. \tag{36}$$

Now we bound the number of partitions. In the two cases we will deal with in Section VI-C the number of subsets can be bounded by a value $p_{\max}$, and all subsets but one contain continuous indices. Therefore the partition is completely defined by the start and end indices of $p_{\max}-1$ subsets (allowed to overlap if there are less than $p_{\max}$ subsets), thus $|\Pi|\le n^{2p_{\max}-2}<n^{2p_{\max}}$ and we have

$$\Pr(J)\le n^{2p_{\max}}\cdot p_J=\exp(2p_{\max}\log(n))\cdot p_J, \tag{37}$$

where p_J is defined in the previous lemmas. So for our purposes we may say that these lemmas hold even if the partition depends on $\mathbf{x}$ with an appropriate change in the probability of J.

B. Error Probability Analysis

In this subsection we show the probability (with respect to the common randomness) to decode incorrectly any of the B indices is smaller than ϵ.

With R_{emp} defined in (21), we have from Lemma 4 that under the conditions of the lemma $\Pr(R_{\text{emp}}\ge t)=\Pr(|\hat{\rho}|\ge R_2^{-1}(t))\le 2\exp(-(n-1)t)$. Then combining Lemmas 1 and 4, we may say that for any $\mathbf{y}_1^m$ the probability of $\mathbf{x}_1^m$ generated i.i.d. from the relevant prior to have $R_{\text{emp}}\ge t$ is bounded by:

$$Q^m(R_{\text{emp}}(\mathbf{x}_1^m,\mathbf{y}_1^m)\ge t)\le\exp\left(-(m-s)(t-\delta_m)\right), \tag{38}$$

where δ_m is defined in (25) and s is defined in (26).

An error might occur if at any symbol $1\le k\le n$ an incorrect codeword meets the termination condition (24). The probability that codeword $j\ne i$ meets (24) at a specific symbol k which is the m-th symbol of a rateless block is bounded by:

$$\begin{aligned}\Pr\left(R_{\text{emp}}^{(m,k)}(\mathbf{x}_j,\mathbf{y})\ge\mu_m^*\right) &\le\exp\left(-(m-s)(\mu_m^*-\delta_m)\right)\\ &=\exp\left(-\left[K+\log\left(\frac{n}{\epsilon}\right)\right]\right)\\ &=\frac{\epsilon}{n\exp(K)}=\frac{\epsilon}{Mn}.\end{aligned} \tag{39}$$

The probability of any erroneous codeword to meet the threshold at any symbol is bounded by the union bound:

$$\Pr(\text{error}) \le \Pr\left\{\bigcup_{k=1}^{n}\bigcup_{j\neq i_b}\left(R_{\text{emp}}^{(m,k)}(\mathbf{x}_j,\mathbf{y}) \ge \mu_m^*\right)\right\} \le n(M-1)\frac{\epsilon}{Mn} < \epsilon, \tag{40}$$

where probabilities above are with respect to the common randomness S. The first inequality is since the correct codeword might be decoded even if an erroneous codeword met the threshold. Although the index m in the expression above depends on k and the specific sequences $\mathbf{x},\mathbf{y}$ in an unspecified way, the assertion is true since the probability of the event in the union has an upper bound independent of m.

C. Rate Analysis

We now turn to prove the achieved rate. The total amount of information sent (with or without error) is $B\cdot K$. Therefore the actual rate is

$$R_{\text{act}} = \frac{BK}{n}. \tag{41}$$

We now endeavor to show this rate is close to or higher than the empirical mutual information with probability of at least P_J over the sequences $\mathbf{x}$, regardless of $\mathbf{y}$ and of whether a decoding error occurred.

The following definition of index sets in $\{1,\dots,n\}$ is used for both the discrete and the continuous cases: $U_b = \{k\}_{k=k_b}^{k_{b+1}-2}$ denotes the channel uses of block b *except* the last one, L_0 collects the last channel uses of all the blocks $L_0 = \{k_b - 1 : b > 1\}$, and U_{B+1} denotes the indices of the un-decoded (last) block $U_{B+1} = \{k\}_{k=k_{B+1}}^{n}$ (including its last symbol), and is an empty set if the last block is decoded. The sets $\{U_b\}_{b=1}^{B+1}, L_0$ are disjoint and their union is $\{1,\dots,n\}$. We denote the length of each block not including the last symbol by $m_b \triangleq |U_b|$. From this point on we split the discussion and we start with the discrete case which is simpler.

Roughly speaking, since $\mu_m^* \approx \frac{K}{m}$, if no error occurs, the correct codeword crossed the threshold when $R_{\text{emp}}^{(m,k)}(\mathbf{x}_i,\mathbf{y}) \approx \frac{K}{m}$ therefore the rate achieved over a rateless block is $R_b = \frac{K}{m} \approx R_{\text{emp}}^{(m,k)}(\mathbf{x}_i,\mathbf{y})$, and due to the approximate convexity by achieving the above rate on each block separately we meet or exceed the rate $R_{\text{emp}}(\mathbf{x},\mathbf{y})$ over the entire transmission. However in a detailed analysis we have the following sources of rate loss:

1) The offsets inserted in μ_m^* to meet the desired error probability
2) The offset from convexity (Lemma 5) introduced by the slight differences in empirical distribution of $\mathbf{x}$ between the blocks
3) Unused symbols:
   a) The last symbol of each block, which is not fully utilized, as explained below
   b) The last (unfinished) block, which is not utilized

The proof is given in the next two subsections, separately for the discrete and continuous case. Following the proof, we give some comments regarding the proof technique.

1) Rate Analysis for the Discrete Case: We write the threshold μ_m^* (24) as $\mu_m^* = \frac{K+\Delta_m}{m} \le \frac{K+\Delta_\mu}{m}$ where Δ_m, Δ_μ are defined below:

$$\Delta_m = \log\left(\frac{n}{\epsilon}\right) + m\delta_m = \log\left(\frac{n}{\epsilon}\right) + |\mathcal{X}||\mathcal{Y}|\log(m+1) \le \log\left(\frac{n}{\epsilon}\right) + |\mathcal{X}||\mathcal{Y}|\log(n+1) \triangleq \Delta_\mu. \tag{42}$$

From Lemma 5 and (37) we have that the following equation:

$$\hat{I}(\mathbf{x};\mathbf{y}) - \Delta \le \sum_{b=1}^{B+1}\left(\frac{m_b}{n}\hat{I}(\mathbf{x}_{B_b};\mathbf{y}_{B_b})\right) + \frac{|L_0|}{n}\hat{I}(\mathbf{x}_{L_0};\mathbf{y}_{L_0}) \tag{43}$$

is satisfied when $\mathbf{x}$ is outside a set J_Δ with probability of at most $\exp\left(-n\left(\Delta - \tilde{\delta}_n\right)\right)$ where $\tilde{\delta}_n = (B+2)|\mathcal{X}|\cdot\frac{\log(n+1)}{n} + 2B_{\max}\frac{\log(n)}{n}$. We shall find the maximum number of blocks $B_{\max}$ later on. To make sure the probability of J is less than P_J we require $\exp\left(-n\left(\Delta-\tilde{\delta}_n\right)\right) \le P_J$ therefore

$$\Delta \ge \tilde{\delta}_n - \frac{1}{n}\log(P_J) = (B+2)|\mathcal{X}|\cdot\frac{\log(n+1)}{n} + 2B_{\max}\frac{\log(n)}{n} - \frac{1}{n}\log(P_J), \tag{44}$$

and we choose

$$\Delta = (3B_{\max}+2)|\mathcal{X}|\cdot\frac{\log(n+1)}{n} - \frac{1}{n}\log(P_J). \tag{45}$$

We now bound each element of (43). Consider block b with m_b+1 symbols. At the last symbol before decoding (symbol $m_b \triangleq |U_b|$) none of the codewords, including the correct one crosses the threshold μ_m^*, therefore:

$$\mu_{m_b}^* = \frac{K+\Delta_{m_b}}{m_b} > \hat{I}(\mathbf{x}_{U_b};\mathbf{y}_{U_b}). \tag{46}$$

Specifically for the unfinished block we have at symbol n:

$$\mu_{m_{B+1}}^* = \frac{K+\Delta_{m_{B+1}}}{m_{B+1}} > \hat{I}\left(\mathbf{x}_{U_{B+1}};\mathbf{y}_{U_{B+1}}\right). \tag{47}$$

The way to understand these bounds is as a guarantee on the shortness of the blocks given sufficient mutual information. On the other hand, at the end of each block *including* the last symbol (i.e., channel uses $\{k_b,\dots,k_b+m_b\}$), since one of the sequences was decoded we have:

$$\mu_{m_b+1}^* = \frac{K+\Delta_{m_b+1}}{m_b+1} \le \max_i \hat{I}\left((\mathbf{x}_i)_{k_b}^{k_b+m_b};\mathbf{y}_{k_b}^{k_b+m_b}\right) \le \log\min(|\mathcal{X}|,|\mathcal{Y}|) \triangleq h_0, \tag{48}$$

which we can use to bound the number of blocks, since $m_b+1 \geq \frac{K}{h_0}$ therefore

$$B \leq \sum_{b=1}^{B} \left(\frac{h_0}{K}(m_b+1) \right) \leq \frac{h_0 \cdot n}{K} \triangleq B_{\max}. \quad (49)$$

As for the unused last symbols we bound:

$$\hat{I}(\mathbf{x}_{L_0}; \mathbf{y}_{L_0}) \leq h_0. \quad (50)$$

Combining (49) and (45) we have:

$$\Delta \leq \left(\frac{3h_0}{K} + \frac{2}{n} \right) |\mathcal{X}| \cdot \log(n+1) - \frac{1}{n} \log(P_J). \quad (51)$$

Combining (46), (47), (50) with (43) and substituting $\Delta_m \leq \Delta_\mu$ yields:

$$\begin{aligned} \hat{I}(\mathbf{x};\mathbf{y}) &< \Delta + \sum_{b=1}^{B+1} \frac{m_b}{n} \left(\frac{K+\Delta_{m_b}}{m_b} \right) + \frac{B}{n} h_0 \\ &\leq \Delta + \sum_{b=1}^{B+1} \frac{1}{n} (K+\Delta_\mu) + \frac{B}{n} h_0 \\ &= \Delta + \frac{B+1}{n} (K+\Delta_\mu) + \frac{B}{n} h_0. \end{aligned} \quad (52)$$

From (52) B and consequently R_{act} can be lower bounded:

$$\begin{aligned} R_{\text{act}} = \frac{B}{n} \cdot K &> \frac{\hat{I}(\mathbf{x};\mathbf{y}) - \Delta - \frac{1}{n}(K+\Delta_\mu)}{K+\Delta_\mu+h_0} \cdot K \\ &= \frac{\hat{I}(\mathbf{x};\mathbf{y}) - \Delta - \frac{K}{n}\left(1+\frac{\Delta_\mu}{K}\right)}{1+\frac{\Delta_\mu+h_0}{K}}. \end{aligned} \quad (53)$$

Now if we increase K with n such that $K \in o(n)$ and $K \in \omega(\log n)$, for example by choosing $K = n^\alpha$, $0 < \alpha < 1$, then $\frac{K}{n} \to 0$ as $n \to \infty$, since $\Delta_\mu = \Theta(\log(n))$ (see (42)) we have $\frac{\Delta_\mu}{K} \to 0$ and from (51) we have $\Delta \to 0$ thus for any δ_0 we have n large enough so that:

$$\begin{aligned} R_{\text{act}} &> \frac{\hat{I}(\mathbf{x};\mathbf{y}) - \delta_0}{1+\delta_0} > \left(\hat{I}(\mathbf{x};\mathbf{y}) - \delta_0 \right)(1-\delta_0) \\ &> \hat{I}(\mathbf{x};\mathbf{y}) - \underbrace{(1+h_0)\delta_0}_{\delta} = R_{\text{emp}} - \delta, \end{aligned} \quad (54)$$

outside the set J, where the last inequality is due to the fact $\hat{I}$ is bounded by h_0. To prove Theorem 3, given δ, we just need to plug in the above equation $\delta_0 = (1+h_0)^{-1}\delta$. Hence we proved our claim that the rate exceeds a rate function which converges uniformly to the empirical mutual information and the proof of Theorem 3 is complete. □

2) Rate Analysis for the Continuous Case: We denote $\hat{\rho}_b \triangleq \hat{\rho}(\mathbf{x}_{U_b}, \mathbf{y}_{U_b})$ and $\hat{\rho} \triangleq \hat{\rho}(\mathbf{x},\mathbf{y})$ the correlation factor measured on a rateless block and on the entire transmission block, respectively. We define a threshold T on the block size (which we will choose later on) and denote by $B_S = \{b : m_b \leq T\}$ and $B_L = \{b : m_b > T\}$ the indices of the small and the large blocks respectively (the last unfinished block included). The total number of symbols in the large blocks is denoted $m_L \triangleq \sum_{b \in B_L} m_b$. The number of large blocks is bounded by $|B_L| < \frac{n}{T}$.

The decoding threshold is written as

$$\mu^*_m = \frac{K}{m-1} + \frac{1}{m-1} \log\left(\frac{n}{\epsilon}\right) + \frac{\log(2)}{m-1} = \frac{K+\Delta_\mu}{m-1}, \quad (55)$$

where we denoted $\Delta_\mu \triangleq \log\left(\frac{2n}{\epsilon}\right)$. We consider the partitioning of the index set $\{1,\ldots,n\}$ into at most $p = \frac{n}{T}$ sets: the first $\frac{n}{T}-1$ (or less) sets are the large blocks except their last symbol $\bigcup_{b \in B_L} U_b$ (each with at least $T+1$ symbols by definition), and the last set denoted L_1 includes the rest of the symbols (last symbols of these blocks and all symbols of small blocks), and has $|L_1| = n - m_L$. Since this partitioning has a bounded number of sets, by applying Lemma 6 and (37) with $p = \frac{n}{T}$ we have that the likely convexity condition (57) below is satisfied when $\mathbf{x}$ is outside a set J with probability at most:

$$\begin{aligned} \Pr(J) &\leq n^{2p} \cdot 2^p e^{-n\Delta^2/8} = \left(\sqrt{2}n\right)^{2\frac{n}{T}} e^{-n\Delta^2/8} \\ &= \exp\left[-n \left(\log(e)\Delta^2/8 - \frac{2}{T}\log\left(\sqrt{2}n\right) \right) \right] \end{aligned} \quad (56)$$

for any $0 < \Delta \leq \frac{1}{7}$. This bound tends to 0 if $T \in \omega(\log(n))$ (since $\log(e)\Delta^2/8 - \frac{2}{T}\log\left(\sqrt{2}n\right) \to \log(e)\Delta^2/8 > 0$) therefore for any such Δ there is n large enough such that this probability falls below the required P_J. The convexity condition is:

$$\begin{aligned} \hat{\rho}^2 - \Delta &\leq \sum_{b \in B_L} \frac{m_b}{n} \hat{\rho}_b^2 + \frac{|L_1|}{n} \hat{\rho}(\mathbf{x}_{L_1}; \mathbf{y}_{L_1})^2 \\ &\leq \sum_{b \in B_L} \frac{m_b}{n} \hat{\rho}_b^2 + \frac{n-m_L}{n}, \end{aligned} \quad (57)$$

where Δ can be made arbitrarily close to 0. We define a factor $\eta_1 < 1$ and apply the function $\left(-\frac{1}{2}\right)\log(1-\eta_1 t)$ to both sides of the above equation. Since the function is monotonically increasing and convex $\cup$ over $t \in [0,1)$ (stemming from concavity $\cap$ of $\log(t)$), we have:

$$\begin{aligned} r_0 &\triangleq \left(-\frac{1}{2}\right)\log(1-\eta_1 \cdot (\hat{\rho}^2 - \Delta)) \\ &\overset{(57)}{\leq} \left(-\frac{1}{2}\right)\log\left[1 - \eta_1 \left(\sum_{b \in B_L} \frac{m_b}{n} \hat{\rho}_b^2 + \frac{n-m_L}{n} \cdot 1 \right) \right] \\ &\leq \sum_{b \in B_L} \frac{m_b}{n} \left(-\frac{1}{2}\right) \log\left(1 - \eta_1 \hat{\rho}_b^2\right) \\ &\quad + \frac{n-m_L}{n}\left(-\frac{1}{2}\right)\log(1-\eta_1 \cdot 1). \end{aligned} \quad (58)$$

We start by bounding the terms related to the large blocks. At the last symbol before decoding in each block (or symbol n for the unfinished block) none of the codewords, including the correct one crosses the threshold μ^*_m, therefore we have for $b = 1,\ldots,B+1$:

$$\mu^*_{m_b} = \frac{K+\Delta_\mu}{m_b - 1} > R_{\text{emp}}(\mathbf{x}_{U_b}, \mathbf{y}_{U_b}) = -\frac{1}{2}\log(1-\hat{\rho}_b^2) \quad (59)$$

and since for a large block $m_b \geq T+1$:

$$\frac{m_b}{n}\left(-\frac{1}{2}\right)\log\left(1-\eta_1\hat{\rho}_b^2\right) \leq \frac{m_b}{n}\left(-\frac{1}{2}\right)\log\left(1-\hat{\rho}_b^2\right) \overset{(59)}{<} \frac{m_b}{n}\cdot\frac{K+\Delta_\mu}{m_b-1} = \left(1+\frac{1}{m_b-1}\right)\frac{K+\Delta_\mu}{n} \leq \left(1+\tfrac{1}{T}\right)\tfrac{K+\Delta_\mu}{n}. \quad (60)$$

For the small blocks we use $n \leq \sum_{b\in B_L}(m_b+1) + \sum_{b\in B_S}(m_b+1) \leq m_L + |B_L| + (T+1)|B_S|$ (where the inequality is since the unterminated block has length m_b) to bound $n - m_L \leq |B_L| + (T+1)|B_S|$.

Combining (58) with these bounds we have:

$$r_0 \leq |B_L|\left(1+\frac{1}{T}\right)\frac{K+\Delta_\mu}{n} + \frac{|B_L|+(T+1)|B_S|}{n}\left[-\frac{1}{2}\log\left(1-\eta_1\right)\right]. \quad (61)$$

The last equation is a lower bound on a linear combination of $|B_L|$ and $|B_S|$. Since the total information sent depends on $|B_L|+|B_S|$ we equalize the coefficients multiplying $|B_L|$ and $|B_S|$ by determining η_1 so that:

$$-\frac{1}{2}\log\left(1-\eta_1\right) = \left(1+\frac{1}{T}\right)\frac{K+\Delta_\mu}{T}. \quad (62)$$

This is always possible since the RHS is positive and the LHS maps $\eta_1 \in (0,1)$ to $(0,\infty)$. Then

$$\begin{aligned} r_0 &\leq \left(|B_L| + \frac{|B_L|+(T+1)|B_S|}{T}\right)\left(1+\frac{1}{T}\right)\frac{K+\Delta_\mu}{n} \\ &= \left(|B_L|+|B_S|\right)\left(1+\frac{1}{T}\right)^2\frac{K+\Delta_\mu}{n} \\ &= (B+1)\frac{(K+\Delta_\mu)\left(1+\frac{1}{T}\right)^2}{n}. \end{aligned} \quad (63)$$

Extracting a lower bound on B from (63) yields a bound on the empirical rate:

$$\begin{aligned} R_{\text{act}} &= \frac{K}{n}\cdot B \\ &\geq \frac{K}{n}\cdot\left(\frac{r_0\cdot n}{(K+\Delta_\mu)\left(1+\frac{1}{T}\right)^2} - 1\right) \\ &= \frac{r_0}{(1+K^{-1}\Delta_\mu)\left(1+\frac{1}{T}\right)^2} - \frac{K}{n} \\ &= \frac{\left(-\frac{1}{2}\right)\log(1-\eta_1(\hat{\rho}^2-\Delta))}{(1+K^{-1}\Delta_\mu)\left(1+\frac{1}{T}\right)^2} - \frac{K}{n} \triangleq R_{\text{LB1}}. \end{aligned} \quad (64)$$

Equation (64) may be optimized with respect to T to obtain a tighter bound, but this is not necessary to prove the theorem. Recall that $\Delta_\mu = \Theta(\log(n))$. By choosing $K \in \omega(\log(n)) \cap o(n)$ the factor $(1+K^{-1}\Delta_\mu)$ in (64) can be made arbitrarily close to 1 and $\frac{K}{n}$ can be made arbitrarily close to 0. As we saw above choosing $T \in \omega(\log(n)) \cap o(n)$ enables us to have $P_J \to 0$ with Δ arbitrarily close to 0, and such a choice will result in $\left(1+\frac{1}{T}\right)^2 \to 1$. Finally if $K \in \omega(T)$ then the RHS of (62) tends to ∞ and therefore we can choose η_1 arbitrarily close to 1. Summarizing the above, by selecting $T \in \omega(\log n), K \in \omega(T), K \in o(n)$ we can write the rate as

$$R_{\text{act}} \geq R_{\text{LB1}} \triangleq \left(-\frac{1}{2}\right)\log(1-\eta_1\cdot(\hat{\rho}^2-\Delta))\cdot\eta_2 - \delta_1, \quad (65)$$

with $\eta_1, \eta_2 \underset{n\to\infty}{\longrightarrow} 1^-$ and $\delta_1, \Delta \underset{n\to\infty}{\longrightarrow} 0^+$. R_{LB1} tends to the target rate $R_2(\hat{\rho}) \triangleq \frac{1}{2}\log\left(\frac{1}{1-\hat{\rho}^2}\right)$ for each point $\hat{\rho} \in [0,1)$ (but not uniformly), and it remains to show that for any $\bar{R}, \delta$ there is n large enough such that $R_{\text{LB1}} \geq R_{\text{LB2}} \triangleq \min(R_2(\hat{\rho}) - \delta, \bar{R})$.

The functions $R_2(\rho)$ and $R_{\text{LB1}}(\rho)$ are monotonically increasing (for fixed η_1, η_2 and δ_1) and it is easy to verify by differentiation that the difference $R_2(\rho) - R_{\text{LB1}}(\rho)$ is also monotonically increasing. Given $\bar{R}, \delta$, choose ρ_0 such that $R_2(\rho_0) = \bar{R} + \delta$. Since $R_{\text{LB1}}(\rho_0) \underset{n\to\infty}{\longrightarrow} R_2(\rho_0)$, for n large enough we have $R_2(\rho_0) - R_{\text{LB1}}(\rho_0) \leq \delta$, and therefore $R_{\text{LB1}}(\rho_0) \geq R_2(\rho_0) - \delta = \bar{R}$. For this n, for any $\rho \leq \rho_0$ from the monotonicity of the difference we have that $R_2(\rho) - R_{\text{LB1}}(\rho) \leq \delta$, and therefore $R_{\text{LB1}} \geq R_{\text{LB2}}$, and for any $\rho \geq \rho_0$ we have from the monotonicity of $R_{\text{LB1}}(\rho)$ that $R_{\text{LB1}}(\rho) \geq \bar{R} = R_{\text{LB2}}(\rho)$, therefore $R_{\text{LB1}} \geq R_{\text{LB2}}$, which completes the proof of Theorem 4. □

D. Comments Regarding the Proof Technique

In this subsection we highlight some points relating to the proof technique, in order to help understand the motivation for some of the steps.

One difficulty stems from the losses due to not fully utilizing the last symbol of each block, and the last block. Regarding the last symbol of each block, note that after receiving the previous symbol the empirical mutual information is below the threshold, and at the last symbol it meets or exceeds the threshold. However the proposed scheme does not gain additional rate from the difference between the mutual information and the threshold, and thus it loses with respect to its target (the mutual information over the block) when this difference is large. Here a "good" channel becomes disadvantageous. Since we operate under an individual channel regime, the increase of the mutual information at the last symbol is not bounded by the mutual information contributed by a single symbol. This is especially evident in the continuous case where the empirical mutual information is unbounded. A high value of y together with high value of x at the last symbol causes an unbounded increase in R_{emp}: if we choose $\mathbf{x}_m, \mathbf{y}_m \to \infty$ then $\hat{\rho}(\mathbf{x}_1^m, \mathbf{y}_1^m) \to 1$ regardless of the history $\mathbf{x}_1^{m-1}, \mathbf{y}_1^{m-1}$. Therefore over a single block we might have an arbitrarily low rate ($|\hat{\rho}|$ is small over the $m-1$ first symbols) and arbitrarily large R_{emp}. In the discrete case this phenomenon exists but is less accented (consider for example the sequences $\mathbf{x} = \mathbf{y} = 0^{n-1}1 = (0,\ldots,0,1)$). Similarly regarding the last block, the fact that the length of the block may be bounded does not immediately indicate the increase in the empirical mutual information can be bounded as well. We use the likely convexity (Lemma 5) to show the last two losses are bounded for most $\mathbf{x}$ sequences.

The continuous case is more difficult for several reasons. One is that the error probability exponent has a missing degree of freedom ($\approx \exp((n-1)t)$). This results in a rate loss (through s in the definition of μ_m^*), which is larger for small blocks, and

can be bounded only when assuming the number of blocks does not grow linearly with n. Since the effective mutual information $R_{\rm emp}(\mathbf{x},\mathbf{y})$ is unbounded we cannot simply bound the loss of mutual information over the unused symbols. Specifically for a single symbol, $\hat{\rho}=1$ and $R_{\rm emp}=\infty$. Therefore we use the convexity of the correlation factor and the fact it is bounded by 1. As a result, the loss introduced in order to attain convexity (over the rateless blocks) is in the correlation factor rather than the empirical mutual information. A loss in the correlation factor induces an unbounded loss in the rate function for $\rho\approx 1$, leading to a maximum rate. In order to cope with these difficulties we use the threshold T on the number of symbols in a block, and treat large and small blocks differently: the large blocks are analyzed through their correlation factor and for the small blocks the correlation factor is upper bounded by 1 and only the number of blocks is accounted for.

VII. Examples

In this section we give some examples to illustrate the model developed in this paper. In this section we use a slightly less formal notation.

A. Constant Outputs and Other Illustrative Cases

The statement that a rate which is determined by the input and output sequences can be attained without assuming any dependence between them may seem paradoxical at first. Some insight can be gained by looking at the specific case where the output sequence is fixed and does not depend on the input. In this case, obviously, no information can be transferred. Since the encoder uses random sequences, the result of fixing the output is that the probability to have an empirical mutual information larger than $\delta>0$ tends to 0, therefore most of the time the rate will be 0. Infrequently, however, the input sequence accidentally has empirical mutual information larger than $\delta>0$ with the output sequence. In this case the decoder will set a positive rate, but will very likely fail to decode. These cases occur with vanishing probability and constitute part of the error probability. So in this case we will transmit rate $R=0$ with probability of at least $1-\epsilon$ and $R>0$ with probability at most ϵ. Conversely, if the channel appears to be good according to the input and output sequences (suppose for example $y_k=x_k$), the decoder does not know if it is facing a good channel or just a coincidence, however it takes a small risk by assuming it is indeed a good channel and attempting to decode, since the chances of high mutual information appearing accidentally are small (and uniformly bounded for all output sequences).

Another point that appears paradoxical at first sight is that the decoder is able to determine a rate $R\geq R_{\rm emp}$ without knowing $\mathbf{x}$ for any $\mathbf{x}\notin J$. First observe that although it is an output of the decoder, the rate R is not controlled by the encoder and therefore cannot convey information. Since the decoder knows the codebook, and given the codebook the sequence $\mathbf{x}$ is limited to a number of possibilities (determined by the possible messages and block locations), it is easy to find an $R(\mathbf{y})\geq R_{\rm emp}(\mathbf{x},\mathbf{y})$ by maximizing $R_{\rm emp}$ over all possible sequences $\mathbf{x}$. Vaguely speaking, the decoding process is indeed a maximization of $R_{\rm emp}$ over multiple $\mathbf{x}$ sequences and by Lemmas 1, 4 such a decoding process guarantees a small probability of error.

B. Using Individual Channel Model to Analyze Adversarial Individual Sequence

As we noted in the overview, the results obtained for the individual channel model constitute a convenient starting point for analyzing channel models which have a full or partial probabilistic behavior. It is clear that results regarding achievable rates in fully probabilistic, compound, arbitrarily varying and individual noise sequence models can be obtained by applying the weak law of large numbers to the theorems discussed here (in general, common randomness would have to be assumed).

E.g. for a compound channel model $W_\theta(y|x)$ with an unknown parameter θ since $\hat{P}(\mathbf{x};\mathbf{y})\underset{n\to\infty}{\longrightarrow}P_\theta(x,y)=W_\theta(y|x)Q(x)$ in probability for every θ and since $I(\cdot;\cdot)$ is continuous $\hat{I}(\mathbf{x};\mathbf{y})\underset{n\to\infty}{\longrightarrow}I_\theta(X;Y)$. Hence from Theorem 1 rate $\min_\theta I_\theta(X;Y)$ can be obtained without feedback, and from Theorem 3 rate $I_\theta(X;Y)$ can be obtained with feedback. These results are not new for the non-adaptive case [18], [19], and for the rate adaptive case can be obtained as a special case of results on channel with an individual state sequence [2], [4] since the individual noise sequence model can be degenerated into a compound model. They are given only to show the ease of using the individual model once established.

To show the strength of the model we analyze a problem considered also by Shayevitz and Feder [2] of an individual sequence which is determined by an adversary and allowed to depend in a fixed or randomized way on the past channel inputs and outputs. For simplicity we start with the binary channel $Y_k=X_k\oplus Z_k$ where Z_k is allowed to depend on $\mathbf{X}_1^{k-1}$ and $\mathbf{Y}_1^{k-1}$ (possibly in a random fashion), and the target is to show the empirical capacity is still achievable in this scenario. Note that here, unlike in most of this paper, the noise Z_k is a random variable but not assumed to be i.i.d. We denote the relative number of errors by $\hat{\epsilon}\stackrel{\Delta}{=}\frac{1}{n}\sum_{k=1}^n Z_k$. We would like to show the communication scheme achieves a rate close to $1_{\rm bit}-h_b(\hat{\epsilon})$ with high probability, regardless of the adversary's policy. Note that both the achieved rate and the target $1_{\rm bit}-h_b(\hat{\epsilon})$ are random variables and the claim is that they are close with high probability (i.e., that the difference converges in probability to 0 when $n\to\infty$)

Applying the scheme achieving Theorem 3 with $Q=\mathrm{Ber}(\frac{1}{2})$ we can asymptotically approach (or exceed) for almost every $\mathbf{X}^n,\mathbf{Y}^n$ the rate:

$$\begin{aligned}\hat{I}(\mathbf{X}^n;\mathbf{Y}^n)&=\hat{H}(\mathbf{Y}^n)-\hat{H}(\mathbf{Y}^n|\mathbf{X}^n)=\hat{H}(\mathbf{Y}^n)-\hat{H}(\mathbf{Z}^n|\mathbf{X}^n)\\&\geq\hat{H}(\mathbf{Y}^n)-\hat{H}(\mathbf{Z}^n)=\hat{H}(\mathbf{Y}^n)-h_b(\hat{\epsilon}).\end{aligned}\tag{66}$$

Note that unlike in the probabilistic BSC where we have $I(X;Y)=H(Y)-H(Z)$, here the empirical distribution of $\mathbf{Z}^n$ is not necessarily independent of $\mathbf{X}^n$, therefore the entropies are only related by the inequality $\hat{H}(\mathbf{Z}^n|\mathbf{X}^n)\leq\hat{H}(\mathbf{Z}^n)$ (conditioning reduces entropy). In order to show a rate of $1_{\rm bit}-h_b(\hat{\epsilon})$ is achieved, we only need to show $\hat{H}(\mathbf{Y}^n)\underset{n\to\infty,prob.}{\longrightarrow}1_{\rm bit}$.

Since X_k is independent of $\mathbf{X}^{k-1}, \mathbf{Y}^{k-1}$ and therefore also of Z_k we have:

$$\begin{aligned}\Pr(Y_k=0|\mathbf{Y}^{k-1}) &= \sum_{X_k}\Pr(Y_k=0|\mathbf{Y}^{k-1},X_k)\Pr(X_k|\mathbf{Y}^{k-1}) \\ &= \sum_{X_k}\Pr(Z_k=X_k|\mathbf{Y}^{k-1})\underbrace{\Pr(X_k)}_{\frac{1}{2}} \\ &= \frac{1}{2}\sum_{X_k}\Pr(Z_k=X_k|\mathbf{Y}^{k-1}) = \frac{1}{2}. \end{aligned} \quad (67)$$

Therefore Y_1^n is distributed i.i.d. $\mathrm{Ber}(\frac{1}{2})$ and from the law of large numbers and the continuity of $H(\cdot)$ we have the desired result.

We can extend the example above to general discrete channels and perform a consolidation of the adversarial sequence model considered by Shayevitz and Feder [2] (for modulo additive channels) with the general discrete channel with fixed sequence considered by Eswaran *et al.* [4]. We address the channel $W_s(y|x)$ with state sequence s_k potentially determined by an adversary knowing all past inputs and outputs. We would like to show that the rate $I(Q,\sum_s W_s(y|x)\hat{P}_{\mathbf{s}}(s))$ (the mutual information of the state-averaged channel) can be asymptotically attained in the sense defined above.

This result is a superset of the previous results [4], [2]. It overlaps with the first [4] in the case $\mathbf{s}$ is a fixed sequence and with the other [2] for the case of modulo-additive channel (or when the target rate is based on the modulo additive model).

Since Theorem 3 shows the rate $\hat{I}(\mathbf{x};\mathbf{y}) \triangleq I(\hat{P}(\mathbf{x}),\hat{P}(\mathbf{y}|\mathbf{x}))$ can be approached or exceeded asymptotically, it remains to show that the empirical distribution $\hat{P}(\mathbf{x},\mathbf{y})$ is asymptotically close to the state-averaged distribution $P_{avg}(x,y) \triangleq \sum_s W_s(y|x)\hat{P}_{\mathbf{s}}(s)Q(x) = \frac{1}{n}\sum_k W_{S_k}(y|x)Q(x)$, and the result will follow from continuity of the mutual information. Note that the later value is a random variable (function) depending on the behavior of the adversary. Here we do not use the law of large numbers because of the interdependencies between the signals $\mathbf{x},\mathbf{y}$ and $\mathbf{s}$.

Our purpose is to prove that the difference $\Delta(t,r)$ defined below converges in probability to 0 for every t,r:

$$\begin{aligned}\Delta(t,r) &\triangleq \hat{P}_{(\mathbf{x},\mathbf{y})}(t,r) - P_{avg}(t,r) \\ &= \frac{1}{n}\sum_k \mathrm{Ind}(X_k=t,Y_k=r) - \frac{1}{n}\sum_k W_{S_k}(r|t)Q(t) \\ &\triangleq \frac{1}{n}\sum_k \varphi_k(t,r), \end{aligned} \quad (68)$$

where $\varphi_k(t,r) \triangleq \mathrm{Ind}(X_k=t,Y_k=r) - W_{S_k}(r|t)Q(t)$. For brevity of notation we omit the argument (t,r) from $\varphi_k(t,r)$ since from this point on it takes a fixed value. Then

$$\begin{aligned}&\mathbb{E}(\mathrm{Ind}(X_k=t,Y_k=r)\,|\,X^{k-1},Y^{k-1},S^k) \\ &= \Pr(X_k=t,Y_k=r\,|\,X^{k-1},Y^{k-1},S^k) \\ &= \Pr(X_k=t\,|\,X^{k-1},Y^{k-1},S^k) \\ &\quad\cdot\Pr(Y_k=r\,|\,X_k=t,X^{k-1},Y^{k-1},S^k) \\ &\overset{(a)}{=} \Pr(X_k=t)\cdot\Pr(Y_k=r\,|\,X_k=t,S_k) \overset{(b)}{=} Q(t)W_{S_k}(r\,|\,t), \end{aligned} \quad (69)$$

where (a) is due to the independent drawing of X_k (when not conditioned on the codebook), the fact S^k is independent of X_k, and the memoryless channel (defining the Markov chain $(X^{k-1},Y^{k-1},S^{k-1}) \leftrightarrow (X_k,S_k) \leftrightarrow Y_k$), and (b) is due to the i.i.d drawing of X_k from Q and the definition of W. From (69) we have that:

$$\mathbb{E}(\varphi_k|X^{k-1},Y^{k-1},S^k) = 0. \quad (70)$$

By taking an expected value from both sides of (70) and using the law of iterated expectations we also have that φ_k has zero mean $\mathbb{E}(\varphi_k)=0$. We now show that φ_k are uncorrelated. Consider two different indices $j<k$ (without loss of generality) then

$$\begin{aligned}\mathbb{E}(\varphi_k\cdot\varphi_j) &= \mathbb{E}\left[\mathbb{E}(\varphi_k\cdot\varphi_j|X^{k-1},Y^{k-1},S^k)\right] \\ &= \mathbb{E}\left[\varphi_j\cdot\mathbb{E}(\varphi_k|X^{k-1},Y^{k-1},S^k)\right] = 0, \end{aligned} \quad (71)$$

where we used the law of iterated expectations and the fact φ_j is completely determined by X_j,Y_j,S_j which are given. In addition since by definition $-1 \le \varphi_k \le 1$, $\mathbb{E}(\varphi_k^2) \le 1$. Therefore

$$\mathbb{E}(\Delta^2) = \frac{1}{n^2}\sum_{j,k=1}^n \mathbb{E}(\varphi_k\cdot\varphi_j) \le \frac{1}{n^2}\sum_{j,k=1}^n \delta_{jk} = \frac{1}{n}, \quad (72)$$

and by Chebyshev inequality for any $\delta>0$:

$$\Pr(|\Delta(t,r)|>\delta) \le \frac{\mathbb{E}(\Delta^2)}{\delta^2} \le \frac{1}{n\delta^2} \underset{n\to\infty}{\longrightarrow} 0, \quad (73)$$

which proves the claim. □

This result is new, to our knowledge, however the main point here is the relative simplicity in which it is attained when relying on the empirical channel model (note that most of the proof did not require any information-theoretic argument).

C. Employing the Continuous Channel Scheme Over a BSC

When operated over a channel different than the Gaussian additive noise channel, the rates achieved with the scheme we described in the continuous case are suboptimal compared to the channel capacity. The loss depends on the channel in question. As an example, suppose the communication system is used over a BSC with crossover probability p, i.e., the continuous input value X is translated to a binary value by $\mathrm{sign}(X)$, and the output is $Y = \mathrm{sign}(X)\cdot(-1)^{\mathrm{Ber}(p)}$. The capacity of this channel is $C = 1_{\mathrm{bit}} - h_b(p)$ and we are interested to calculate the rate which would be achieved by our scheme (which does not know the channel) for this channel behavior. For this channel with Gaussian $\mathcal{N}(0,P)$ input we have (through a simple calculation):

$$\mathbb{E}(XY) = (1-2p)\sqrt{\frac{2P}{\pi}}, \quad (74)$$

hence

$$\rho^2 = \frac{\mathbb{E}(XY)^2}{P\cdot\mathbb{E}(1^2)} = \frac{2}{\pi}(1-2p)^2, \quad (75)$$

and

$$R = \frac{1}{2}\log\left(\frac{1}{1-\frac{2}{\pi}(1-2p)^2}\right). \quad (76)$$

The comparison between C and R is presented in Fig. 5. It can be shown that $R \ge \frac{2}{\pi}C$, thus the maximum loss is 36%.

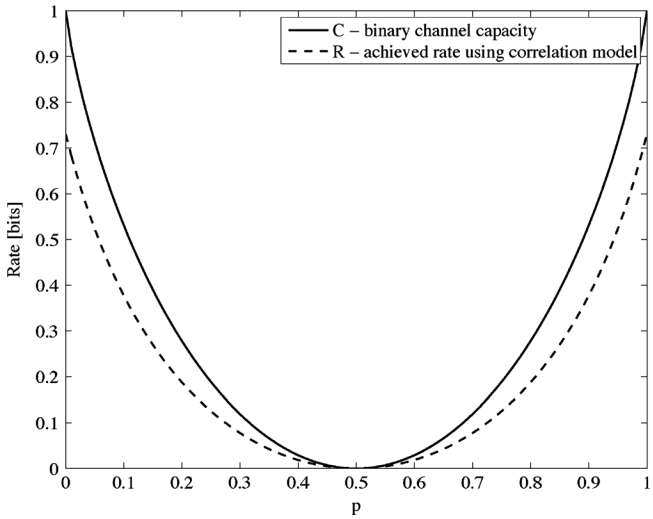

Fig. 5. Comparison of C,R for the BSC.

D. An Effective AWGN Channel

The rate function $R_{\text{emp}} = -\frac{1}{2}\log(1-\hat{\rho}^2)$ in the continuous case, can be also written in the familiar form similar to the AWGN capacity

$$R_{\text{emp}} = \frac{1}{2}\log(1+S\hat{N}R), \tag{77}$$

where

$$S\hat{N}R \triangleq \frac{\hat{\rho}^2}{1-\hat{\rho}^2}. \tag{78}$$

$S\hat{N}R$ represents the *SNR* measured using a reference channel: given the input and output sequences, the output can be described by the following virtual additive channel:

$$y_i = \alpha x_i + v_i, \tag{79}$$

so the effective noise sequence is $v_i = y_i - \alpha x_i$, where α is chosen such that $\mathbf{v} \perp \mathbf{x}$, i.e., $\frac{1}{n}\sum_i v_i x_i = 0$. An equivalent condition is that α minimizes $\|\mathbf{v}\|^2$. The resulting α is the LMMSE coefficient in estimation of $\mathbf{y}$ from $\mathbf{x}$ (assuming zero mean), i.e.,

$$\alpha = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|^2}. \tag{80}$$

Define the effective signal power and noise power as $P = \frac{1}{n}\sum_{i=1}^n x_i^2$, and $N = \frac{1}{n}\sum_{i=1}^n v_i^2$, respectively. $S\hat{N}R$ can now be written as the *SNR* of this effective channel:

$$S\hat{N}R = \frac{\alpha^2 P}{N} = \frac{\hat{\rho}^2}{1-\hat{\rho}^2}. \tag{81}$$

This yields an alternative representation of (2), i.e., the rate $R_{\text{emp}} = \frac{1}{2}\log\left(1+S\hat{N}R\right)$ is asymptotically adaptively achievable.

E. Non Linear Channels

In analyzing probabilistic channels, the correlation model determines the rate $\frac{1}{2}\log\left(\frac{1}{1-\rho^2}\right)$ is always achievable using a Gaussian code (no randomization is needed if the channel is probabilistic, as can be shown by the standard argument about the existence of a good code). This is actually a result of Lemma 2.

This expression is useful for analyzing channels in which the noise is not additive or non-linearities exist. As an example, transmitter noise is usually modeled as an additive noise. However large part of this noise is due to distortion (e.g., in the power amplifier), and therefore depends on the transmitted signal and is inversely correlated to it. Consider the non linear channel $Y = f(X) + V$ with $V \sim \mathcal{N}(0, N)$. In this case if we define the effective *SNR* as $SNR = \frac{\rho^2}{1-\rho^2}$ then rate $R = \frac{1}{2}\log(1+SNR)$ is achievable. The correlation factor is:

$$\rho^2 = \frac{\mathbb{E}(XY)^2}{\mathbb{E}(X^2)\mathbb{E}(Y^2)} = \frac{\mathbb{E}(Xf(X))^2}{\mathbb{E}(X^2)(\mathbb{E}(f(X)^2)+N)}. \tag{82}$$

Therefore the effective *SNR* can be written as:

$$SNR = \frac{\rho^2}{1-\rho^2} = \frac{P_{\text{eff}}}{N+N_{\text{eff}}}, \tag{83}$$

where we defined the effective gain γ, the effective power P_{eff} and the effective noise N_{eff} as:

$$\gamma \triangleq \frac{\mathbb{E}(Xf(X))}{\mathbb{E}(X^2)} \tag{84}$$

$$P_{\text{eff}} \triangleq \mathbb{E}\left[(\gamma X)^2\right] \tag{85}$$

$$N_{\text{eff}} \triangleq \mathbb{E}\left[(f(X)-\gamma X)^2\right] \tag{86}$$

This yields a simple characterization of the degradation caused by the non-linearity, which is independent of the noise power. This model enables to characterize the transmitter distortions by the two parameters $P_{\text{eff}}, N_{\text{eff}}$, a characterization which is more convenient and practical to calculate than the channel capacity, and on the other hand guarantees that transmitter noise evaluated this way never degrades the channel capacity in more than determined by (83).

Another interesting application of this bound is in treating receiver estimation errors, since it is sometimes simpler to calculate the loss in the correlation factor induced due to the imperfect knowledge of the channel parameters than the loss in capacity (see Hassibi's bounds for the loss due to channel estimation from training [17]).

VIII. Comments and Discussion

A. Relation to Similar Models

Below we examine the differences between the framework proposed here and two similar models: the arbitrarily varying channel (AVC) and the channel with an individual noise sequence.

In the AVC model [1], [20], the channel is defined by a probabilistic model which includes an unknown state sequence. Constraints on the sequence (such as power, number of errors) may be defined a-priori, and the target is to communicate equally well over all possible occurrences of the state sequence. In AVCs, the capacity depends on the existence of common randomness and on whether the average or maximum error probability (over the messages) is required to approach 0. Yet when sufficient common randomness is used, the capacities for maximum and average error probability are equal. Lapidoth

and Narayan's notes [1, p.2151] regarding common randomness and randomized encoders are also relevant to our case. As opposed to the current model, in the AVC any constraints on the state sequence have to be set in advance, and the rate considers the worst case conditions. In both AVCs and the current framework common randomness is important to achieve high communication rates. In the current framework common randomness plays a more vital role since we require a specific input distribution.

A result which may be considered as a different viewpoint on the AVC model was presented by Agarwal *et al.* [21]. Their motivations come from network coding theory. Their main result concerns communication over a black box which is only limited to a given level D of distortion according to a predefined metric, but has otherwise a block-wise undefined behavior. They show that it is possible to achieve a rate equal to the rate-distortion function of the input $R_X(D)$, if the black box guarantees average distortion D with high probability. This result is similar to our Theorem 1. The remarkable distinction from other results for AVC is that the rate is determined using a constraint on the channel inputs and outputs, rather than the channel state sequence. Langberg [22] proposed another characterization of an AVC, where common randomness is not assumed, however the "amount of information" the adversary may use about the transmitted sequence is limited.

Channels with individual noise (or state) sequence are treated by Shayevitz and Feder [2], [3] and Eswaran *et al.* [4]. The probabilistic setting is the same as in the AVC, and the difference is that instead of achieving a uniform (hence worst-case) rate, the target is to achieve a variable rate which depends on the particular sequence of noise, using a feedback link. In this setup, prior constraints on the state sequence can be relaxed. As opposed to AVC where the capacity is well defined, the target rate for each state sequence is determined in a somewhat arbitrary way (since many different constraints on the sequence can be defined). As an example, in the binary channel with an individual noise sequence [3], a rate of 0 would be obtained for the sequence $\mathbf{e} =' 01010101\ldots'$ since the empirical error probability is $\frac{1}{2}$, although obviously a scheme which favors this specific sequence and achieves a rate of 1 can be designed. On the other hand, with the AVC approach communication over this channel would not be possible without prior constraints on the noise sequence. Channels with individual noise sequence can be thought of as compound-AVCs (i.e., an AVC with unknown parameter, in this case, the constraint). As in the AVC model, existence of common randomness as well as the definition of error probability affect the achievable rates.

In all models discussed above, the achieved rates are related to some parameters of the problem which are outside the domain of the communication system itself. In contrast, in the individual channel model we use here, since no equation with state sequence connecting the input and output is given, the achievable rates cannot be defined without relating to the channel input. Therefore the definitions of achieved rates depend in a somewhat circular way on the channel input which is determined by the scheme itself. Currently we circumvent this difficulty by constraining the input distribution, as mentioned above.

In many aspects the model used in this paper is more stringent than the AVC and the individual noise sequence models, since it makes less assumptions on the channel, and the error probability is required to be met for (almost) every input and output sequence (rather than on average). In other aspects it is lenient since we may attribute "bad" channel behavior to the rate rather than suffer an error, therefore the error exponents are better than in probabilistic models, as was explained in Section IV-A.

B. The Decoding Rule

In the discrete case we used a maximum empirical mutual information receiver, and in the continuous case, a maximum empirical correlation receiver. In this section we draw connections between the two receivers and point out other potential receivers.

Since the mutual information between two Gaussian r.v-s is $-\frac{1}{2}\log(1-\rho^2)$, one can think of this value as a measure of mutual information under Gaussian assumptions. Thus, using this metric as an effective mutual information, since the mutual information is an increasing function of $|\rho|$ the MMI decoder becomes a maximum empirical correlation decoder. On the other hand, the receiver we used can be identified as the GLRT (generalized maximum likelihood ratio test) for the AWGN channel $Y = \alpha X + \mathcal{N}(0,\sigma^2)$ with α an unknown parameter, resulting from maximizing the likelihood of the codeword and the channel simultaneously:

$$\begin{aligned}\hat{w} &= \operatorname*{argmax}_{\mathbf{x}_m} \max_{\alpha} \log \Pr(\mathbf{y}|\mathbf{x};\alpha) \\ &= \operatorname*{argmin}_{m} \min_{\alpha} \|\mathbf{y}-\alpha\mathbf{x}_m\|^2 = \operatorname*{argmax}_{m} \frac{\left(\mathbf{x}_m^T\mathbf{y}\right)^2}{\|\mathbf{x}_m\|^2} \\ &= \operatorname*{argmax}_{m} \left[\hat{\rho}^2(\mathbf{x}_m,\mathbf{y})\right]. \end{aligned} \tag{87}$$

The choice of the GLRT is motivated by considering the individual channel as an effective additive channel with unknown gain (as presented in Section VII-D), combined with the fact Gaussian noise is the worse. For discrete memoryless channels it is easy to show that the GLRT (where the group of channels consists of all DMC-s) is synonymous with the MMI decoder [1]. Thus, we can identify the two decoders as GLRT decoders, or equivalently as variants of MMI decoders.

Regarding the receiver required to obtain the rates of Theorem 2, it is interesting to consider the simpler maximum projection receiver $\operatorname*{argmax}_{\mathbf{x}_m} |\mathbf{x}_m^T\mathbf{y}|$. This receiver seems to differ from the maximum correlation receiver only in the term $\|\mathbf{x}_m\|$ in (18), which is nearly constant for large n due to the law of large numbers. However surprisingly, the maximum rate achievable with the projection receiver is only $\frac{1}{2}\hat{\rho}^2$ as can be shown by a simple calculation equivalent to Lemma 4 (simpler, since $z = \mathbf{x}^T\mathbf{y}$ is Gaussian). The reason is that when $\mathbf{x}$ is chosen independently of $\mathbf{y}$, a large value of the projection (non typical event) is usually created by a sequence with power significantly exceeding the average (another non typical event). When one non-typical event occurs there is no reason to believe the sequence is typical in other senses thus the approximation $\|\mathbf{x}_m\| \approx \sqrt{nP}$ is invalid. The correlation receiver normalizes by the power of $\mathbf{x}$ and

compensates this effect. An alternative receiver which yields the rates of Theorem 2 and is similar to the AEP receiver looks for the codeword with the maximum absolute projection subject to power limited to $\frac{1}{n}\|\mathbf{x}_m\|^2 < P + \delta$. This can be shown by Sanov's theorem [10] or by using the Chernoff bound. The maximum correlation receiver was chosen because of its elegance and the simplicity of the proof of Lemma 4. Note that had we used a uniform distribution on an n dimensional sphere instead of the i.i.d. Gaussian distribution (i.e., using a constant power), then all these receivers would become equivalent (Lemma 4 in essence still holds in this case. See the comments in the proof of this lemma), however limiting the distribution to a sphere would not be appropriate to rateless coding since the stopping time is unknown in advance.

For the AWGN channel, combining Lemma 4 with the law of large numbers provides an alternative way to show the achievability of the AWGN capacity $\frac{1}{2}\log(1+SNR)$, without using the AEP receiver. The maximum correlation receiver has the technical advantage, compared to the AEP receiver, that it does not declare an error for codewords which have power deviating from the nominal power. This technical advantage is important in the context of rateless decoding since the power condition needs to be re-validated each symbol, thus increasing its contribution to the overall error probability.

Lapidoth [23] showed that the nearest neighbor receiver achieves a rate equal to the Gaussian capacity $\frac{1}{2}\log(1+P/N)$ over the additive channel $Y = X + V$ with an arbitrary noise distribution (with fixed noise power). This result parallels the result that the random code capacity of the AVC $Y = X + V$ with a power constraint on V equals the Gaussian capacity [24] (this stems directly from the characterization of the random code capacity of the AVC as $\max_{P_X(x)} \min_{P_S(s)} I(X;Y)$, [10, (V.4)].

C. Rateless Codes and Similar Schemes

The scheme proposed here is based on rateless codes. Here we give some background regarding the evolution of rateless codes, and the differences between the proposed techniques. Rateless codes have been used for two main purposes: reducing the error probability/exponent, and dealing with channel uncertainty. The earliest work is of Burnashev [13] who showed that for known channels, using feedback and a random decision time (i.e., decision time which depends on the channel output) yields an improved error exponent, which is attained by a 3 step protocol (best described by Tchamkerten and Telatar [11]) and shown to be optimal. Shulman [25] proposed to use random decision time as a means to deal with sending common information over broadcast channels (static broadcasting), and for unknown compound channels (which are treated as broadcast). In this scheme later described as "rateless coding" (or Incremental Redundancy Hybrid ARQ) a codebook of $\exp(K)$ infinite sequences is generated, and the sequence representing the message is sent to the receiver symbol by symbol, until the receiver decides to decode (and turn off, in case of a broadcast channel).

Tchamkerten and Telatar [11] connect the two results by showing that for some, but not all compound channels Burnashev error exponent can be attained universally using rateless coding and the 3 step protocol. Eswaran, Sarwate, Sahai and Gastpar [4] used iterated rateless coding to achieve the mutual information related to the empirical noise statistics on channels with individual noise sequences. The scheme we use here is most similar to the one used by Eswaran *et al.* [4] but less complicated. We do not use training symbols to learn the channel in order to decide on the decoding time but rely on the mutual information itself as the criterion (based on Lemmas 1,4) and the partitioning into blocks and the decision rules are simpler.

The later result [4] is an extension of Shayevitz and Feder's result [3] regarding the binary channel to general discrete channels with individual noise sequence. The original result [3] was obtained not by rateless codes but by a successive estimation scheme [26] which is a generalization of the Horstein [27] and Schalkwijk-Kailath [28] schemes. The same authors extend their results to discrete channels [2] using successive schemes (where the target rate is the capacity of the respective modulo-additive channel). The two concepts in achieving the empirical rates differ in various factors such as complexity and the amount of feedback and randomization required. The successive schemes require less common randomness but assume perfect feedback, while the schemes based on rateless coding require less (asymptotically 0 rate) feedback but potentially more randomness.

As noted, the technique we use here is similar to that of Eswaran *et al.* [4] in its high level structure, while the structure of the rateless decoder is similar to Shulman's [25, Chapter 3]. The application of this scheme to individual inputs and outputs and the extension to real-valued models requires proof and especially issues such as abnormal behavior of specific (e.g., last) symbols had to be treated carefully. The previous results [4] cannot be applied directly to individual channels since the channel model cannot be extracted based on the input and output sequences alone, and in the later both the model and the sequence are assumed to be fixed (over common randomness). Table II compares some attributes of the schemes. Another important factor is the overhead (i.e., the loss in the number of bits communicated with a given error exponent, compared to the target rate), which we were unable to compare. We conjecture that the current scheme may have a lower overhead due to its simplicity which results in a smaller number of parameters and constraints on their order of magnitude (compared to the previously suggested scheme [4] where relations between factors such as number of pilots and the minimum size of a chunk may require a large value of n).

D. Random Decision Time

In our discussion we have described two communication scenarios: fixed rate without feedback and variable rate with feedback, and in both we assumed a fixed block size n. Another scenario is that of random decision time or rateless coding [13], [25] in which the block size is not fixed but determined by the decoder. We did not include this scenario since the achievability result is less elegant in a way: the decoder indirectly affects the target rate (mutual information) through the block size. On the other hand this case may be of practical interest. Clearly the mutual information can be asymptotically attained for this communication scenario as well and its analysis is merely a simpler

TABLE I
SUMMARY OF DEFINITIONS AND REFERENCES FOR THE DISCRETE AND CONTINUOUS CASES

Item	Discrete case	Continuous case
Input distribution	Any Q	$Q=\mathcal{N}(0,P)$
Decoding metric	$R_{\mathrm{emp}}(\mathbf{x},\mathbf{y}) \triangleq \hat{I}(\mathbf{x};\mathbf{y})$	$R_{\mathrm{emp}}(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{2}\log\left(\frac{1}{1-\hat{\rho}^2(\mathbf{x},\mathbf{y})}\right)$
Decoder	maximize $R_{\mathrm{emp}}(\mathbf{x},\mathbf{y}) \Leftrightarrow$ maximize $\hat{I}(\mathbf{x};\mathbf{y})$	maximize $R_{\mathrm{emp}}(\mathbf{x},\mathbf{y}) \Leftrightarrow$ maximize $\lvert\hat{\rho}(\mathbf{x},\mathbf{y})\rvert$
Pairwise error probability $\Pr(R_{\mathrm{emp}} \geq t)$	$\leq \exp(-n(t-\delta_n))$ (Lemma 1)	$\leq 2\exp(-(n-1)t)$ (Lemma 4)
Likely convexity condition ($\forall \mathbf{x} \notin J_\Delta, \mathbf{y} \in \mathcal{Y}^n$ with $\lambda_i \triangleq \frac{1}{n}\lvert A_i\rvert$)	$\sum_{i=1}^{p}\lambda_i \hat{I}(\mathbf{x}_i;\mathbf{y}_i) \geq \hat{I}(\mathbf{x};\mathbf{y}) - \Delta$ (Lemma 5)	$\sum_{i=1}^{p}\lambda_i\hat{\rho}_i^2 \geq \hat{\rho}^2 - \Delta$ (Lemma 6)
Likely convexity probability ($\Pr(\mathbf{x} \notin J_\Delta)$, fixed partitioning)	$\geq 1-\exp\left(-n\left(\Delta-\tilde{\delta}_n\right)\right)$	$\geq 1-2^p e^{-n\Delta^2/8}$

version of the rate analysis performed in Section VI-C, since convexity is not required.

E. Limitations of the Zero Order Empirical Model

Although we did not assume anything about the channel, and specifically we did not assume the channel is memoryless, the fact we used the zero-order empirical distribution means the rates may fall short of the mutual information rate when the system is operated over channels with memory or, in the continuous case over non AWGN channels. Below we discuss several cases where the communication fails completely.

One example is when some delay is introduced between $\mathbf{x}$ and $\mathbf{y}$. In this case the proposed scheme may be suboptimal or fail completely. For example, for the channel $y_k = x_k + \frac{1}{2}x_{k-1} + v_k$ we would obtain positive rates and the intersymbol interference (ISI) $\frac{1}{2}x_{k-1}$ would be treated suboptimally as noise, but for the error free channel $y_k = x_{k-1}$ the achieved rate would be 0 with high probability. Similarly we can find a memoryless channel with infinite capacity but for which the correlation-based rate function we used for the continuous alphabet scheme fails: if $y_k = x_k^2$ (where x_k is Gaussian) then $\rho = 0$. Another example of practical importance is the fading channel (with memory) $y_n = h_n x_n + v_n$, where h_n is slowly fading with mean 0.

All these examples result from the simplicity of the models used, and can be solved by schemes employing higher order empirical distributions (e.g., by measuring the empirical mutual information over blocks of symbols, or by using Markov models, i.e., measuring conditional empirical probabilities), and by using higher order statistics in the continuous case. We have extended the results to MIMO [6], and presented a compression based model than can capture the time dependencies [7]. Furthermore, the current paper exhibits a considerable similarity between the continuous case and the discrete case which is not fully explored here, and we hope to present a unifying theory which will include the two as particular cases in a follow-up paper.

F. The Amount of Randomness Required

In this work we have assumed no restriction on the amount of common randomness available and have not attempted to minimize the amount of randomness required while maintaining the same rates. Furthermore, we have made a theoretical assumption that one may have access to random variables with any desired distribution, and specifically a Gaussian distribution. The total amount of randomness required is composed of

1) The amount of randomness in single random drawing of a letter $x_i \sim Q(x)$
2) The amount of random drawings $\sim Q(x)$ needed per n-block

The amount of randomness required to generate (simulate) a random variable is often measured by the number of random uniform i.i.d. bits necessary. Clearly, even the generation of a discrete random variable with an arbitrary distribution $Q(x)$ (in which $Q(x)$ are not multiplies of some 2^{-L}) cannot be accomplished using a finite number of random bits. However, as shown e.g., by Han and Hoshi [29, Theorem 3, Remark 8], it can be accomplished using a number of bits which is finite *on average* (the expected number of bits is bounded by $H(Q)+3$).

Another alternative is to use a fixed number L of bits, and approximate the distribution $Q(x)$ by its "rounded" version, i.e., a $\tilde{Q}(x)$ having $\tilde{Q}(x)\cdot 2^L \in \mathbb{Z}$, and $\lvert Q(x)-\tilde{Q}(x)\rvert \leq 2^{-L}$. In the context of individual channels there is no way to tell what is the result of changing the input distribution, as there is also no way to determine which is better, Q or $\tilde{Q}$, however we may use the mutual information over an unknown channel W as a figure of merit, i.e., compare $I(Q,W)$ with $I(\tilde{Q},W)$. It is easy to show that if $\lvert Q(x)-\tilde{Q}(x)\rvert \leq 2^{-L}$ then for any W, $I(Q,W)-I(\tilde{Q},W) \underset{L\to\infty}{\longrightarrow} 0$. Therefore a reasonable solution to limit the amount of randomness in the discrete case is to replace the input distribution by a distribution that can be simulated using a finite number of random bits.

In the continuous case, the number of random bits required to generate a Gaussian random variable is infinite (even the number of bits required to store a real number is infinite). Therefore any implementable system would have to use an approximation of the Gaussian distribution, and this will result in only approximately attaining the results shown here.

Regarding the number of random drawings, we have used $\exp(K)\cdot n$ random drawings, and since $K \in \omega(\log n)$, the number of random drawings is $\omega(n^2)$ (i.e., grows faster than n^2). This may be much larger than actually required. For the arbitrarily varying channel, Ahlswede [30] showed that selecting out of n^2 codebooks (i.e., $2\log n$ random bits) is sufficient. Similar results including lower bounds on the amount of randomness were obtained by Langberg [31] for a binary AVC. For the modulo-additive channel with an individual noise sequence, Shayevitz and Feder [2, Sec. V.5] had shown that less than $\Theta(n)$ random bits are required in some cases and $\Theta(n)$ are enough for others. These cases are not equivalent to the one discussed here, however the results seem to suggest the number of random

TABLE II
COMPARISON WITH THE RATE ADAPTIVE SCHEME IN [4]

Item	Eswaran *et al* [4]	Current Paper	Comments
Channel model	Individual sequence	Individual channel	
Mechanism for adaptivity	Repeated instanced of rateless coding	Repeated instanced of rateless coding	
Transmit format	Total time divided to rounds (=rateless blocks) which are divided to chunks	Total time divided to rateless blocks	Chunks in [4] used as feedback instances and expurgated code has constant type over chunks
Feedback	Ternary ("Bad Noise"/"Decoded"/"Keep Going"), once per chunk	Binary ("Decoded"/"Not Decoded") per symbol	Easy to generalize to once every q symbols (see VI-C)
Alphabet	Discrete	Discrete or real valued	
Training	Known symbols in random locations in each chunk	None	
Randomness	Full (num. codewords $\times$ n)	Full (num. codewords $\times$ n)	Might be reduced by selection from a smaller collection of codebooks (in both cases)
Codebook construction	Constant composition + expurgation + training insertion	Random i.i.d.	
Stopping condition	Threshold over mutual information of channel estimated from training	Threshold over empirical mutual information of best codeword	
Decoding	Maximum (empirical) mutual information	Maximum (empirical) mutual information	
Stopping location	End of chunk	Any symbol	

drawings can be reduced. Note however, that by making the basic requirement that input would behave randomly $\mathbf{x} \sim Q^n$ we inherently require n drawings from $Q(x)$.

G. Comments on the Scheme

1) Varying Channels: Although our target is the empirical mutual information over the n-block, an artifact of the partitioning to smaller blocks is that higher rates can be attained when the empirical conditional channel distribution varies over time, since by the convexity of mutual information with respect to the channel law the convex sum of mutual information over blocks may exceed the overall mutual information if the empirical channel behavior is not constant.

2) Zero Rate Feedback: It is easy to show that the feedback rate can be reduced to $\frac{1}{q}$ for some $q \in \mathbb{N}$, without significantly changing the results. The scheme of Section V-B is modified so that decoding and block termination (step (3)) is only performed once every q channel uses. This will result in having potentially q unused symbols instead of one. By a simple modification to the arguments that show the loss from not utilizing the last symbol vanishes asymptotically, it can be shown that this loss vanishes also for q unused symbols. As an example, for the discrete case, this modification would be expressed in replacing the factor h_0 in (54) by $q \cdot h_0$.

Hence the scheme can be modified to operate with "zero rate" feedback and would attain the same asymptotical rates. Similarly the scheme can operate with a noisy feedback channel by introducing in the feedback link a delay suitable to convey the decoder decisions with sufficiently low error rate over the noisy channel.

3) Maximal and Minimal Rate: The scheme has a minimal rate and a maximal rate for each block length. The minimal rate is $\frac{K}{n}$ resulting from sending a single block. If channel conditions are worse ($R_{\text{emp}} < \frac{K}{n}$), no information will be sent. A maximal rate exists since at best K bits could be sent every 2 symbols (since for the continuous case $\mu_1^* = \infty$ and for the discrete case $R_{\text{emp}}^{(1,k)}(\mathbf{x}_k^k, \mathbf{y}_k^k) = 0$ thus the decoding never terminates at the first symbol of the block), hence the maximum rate is $\frac{K}{2}$. As $n \to \infty$ we increase K so that the minimum rate (and the rate offsets) tend to 0 and the maximum rate tends to ∞. The maximum rate is the reason that the scheme cannot approach the target rate $R_{\text{emp}}(\mathbf{x}, \mathbf{y})$ uniformly in $\mathbf{x}, \mathbf{y}$ in the continuous case, since for some pairs of sequences the target rate may exceed the maximum rate by an unbounded factor. The maximum rate $\bar{R}$ that we achieve in the proof of Theorem 4 is much smaller than the absolute maximum $\frac{K}{2}$. Note that successive schemes (such as Schalkwijk's [28]) do not suffer from the problem of maximum rate. For the discrete case the target rate is bounded by $\log \min(|\mathcal{X}|, |\mathcal{Y}|)$ therefore for sufficiently large n the maximal rate $\frac{K}{2}$ exceeds $\log \min(|\mathcal{X}|, |\mathcal{Y}|)$ and we are able to show uniform convergence.

IX. FURTHER RESEARCH

A. Determining the Behavior of the Transmitted Signal (Prior)

In this work we assumed a fixed prior (input probability distribution) and haven't dealt with the question of determining the prior, or more generally, how the encoder should adapt its behavior based on the feedback. Had the channel been a compound one, it stands to reason that a scheme using feedback may estimate the channel and adjust the input prior, and may asymptotically attain the channel capacity. However in the scope of individual channels (as well as individual sequence channels and AVC-s) it is not clear whether the approach of adjusting to the input distribution to the measured conditional distribution is of merit, if the empirical channel capacity can be attained for every sequence, and even the definition of achievability is unclear if the input distribution is allowed to vary.

Another related aspect is what we require from a communications system when considered under the individual channel framework. This question is relevant to all the requirements defined in the theorems (for example, is the existence of the failure

set J necessary ?), however the most outstanding requirement is related to the prior.

Currently we constrained the input sequence to be a random i.i.d. sequence chosen from a fixed prior, which seems to be an overly narrow definition. The rationale behind this choice is that without any constraint on the input, the theorems we presented can be attained in a void way by transmitting only bad (e.g., fixed) sequences that guarantee zero empirical rate. Furthermore, without this constraint, attainability results for probabilistic models, and in general any attainable rates which are not conditioned on the input sequence could not be derived from our individual sequence theorems. A weaker requirement from the encoder is to be able to emit any possible sequence. However this requirement is not sufficient, since from the existence of such encoders we could not infer the existence of encoders achieving any positive rate over a specific channel. Consider for example the encoder satisfying the requirement by transmitting bad sequences with probability $1-\delta$ and good sequences with probability $\delta \to 0$. Theorems 1,2,3 and 4 are existence theorems, i.e., they guarantee the existence of at least one system satisfying the conditions. Had we removed the requirement for fixed input prior these theorems would be attained by encoders that are unsatisfactory in other aspects. Once the theorem is satisfied by one encoder it cannot guarantee the existence of other (satisfactory) encoders, thus making it not useful. Therefore the requirement for fixed prior is necessary in the current framework. Although in the scope of the theorems presented here, this requirement only strengthens the theorems (since it reveals additional properties of the encoder attaining the other conditions of the theorem), we are still bothered by the question what should be the minimal requirements from a communication system, and these hopefully will not include a constraint on the input distribution.

This issue relates to a fundamental difficulty which aries in communication over individual channels: unlike universal source coding in which the sequence is given a-priori, here the sequences are given a-posteriori, and the actions of the encoder affect the outcome in an unspecified way. Currently we broke the tie by placing a constraint on the encoder, but we seek a more general definition of the problem.

B. Overhead and Error Exponent

An important aspect in universal communication is the overhead (or redundancy) associated with universality. Specifically, when considering extending the empirical distribution to include time dependencies, we expect that such extensions will increase the overhead. This overhead is related to the redundancy or regret associated with universal distributions [32]. Although we haven't performed a detailed analysis of the overheads and considered only the asymptotically achievable rates, it is obvious from comparing Lemmas 1 and 4 that the tighter rates we obtained for the discrete channel come at the cost of additional overhead ($\Theta(\log(n))$ compared to $\Theta(1)$ in the continuous case) which is associated with the richness of the channel family (describing a conditional probability as opposed to a single correlation factor). The issue of overheads requires additional analysis in order to determine the bounds on the overheads and the tradeoff between richness of the channel family and the rate, for a finite n. A further discussion on this subject is to appear in a follow-up paper.

As we noted in Section VI-D the bounds we currently have for the rate-adaptive scheme, especially in the continuous case are rather loose. The main reasons are given below. One reason is the fixed offset in the correlation factor in Lemma 6, which limits the maximum rate when $\hat{\rho} \to 1$, as visible in Fig. 3. Note that for $\hat{\rho}=1$, since this leads to $\hat{\rho}_i = 1$ for all rateless blocks, the scheme described in Subsection V-B would deliver a rate of $\frac{K}{2}$, which is much higher than the lower bound. Another reason is the use of the union bound in Subsection VI-A3 and in smaller extent in the proof of Lemma 6. Also, the intuitively appealing likely convexity property does not take into account the fact that the empirical mutual information over the blocks are almost equal, and may insert an additional loss in the bound. Alternative techniques that enable tighter bounds are to be presented in a follow up paper which is currently in preparation.

Since rate can be traded off for error probability, a related question is the error exponent, and the probability P_J. The scheme we described does not attempt to attain a good error exponent, specifically, since the block of n channel uses is broken into multiple smaller blocks. Note, however, that for rate adaptive schemes with feedback a good error exponent does not necessarily relate to the capability of sending a message with small probability of error, but rather to the capability to detect the errors (see Burnashev's error exponent and the three phase scheme achieving it [13], [11, IV.B]).

C. Upper Bounds

In this paper we focused on achievable rates and did not show a converse. An almost obvious statement is that any continuous rate function which depends only on the zero-order empirical statistics/correlation (respectively) cannot exceed uniformly the rate functions of Theorems 3, 4 respectively with vanishing error probability as $n \to \infty$, since otherwise it would be possible to achieve rates above the channel capacity of a memoryless channel. A further discussion on this subject is to appear in a follow-up paper.

X. Conclusion

We examined achievable transmission rates for channels with an unspecified model, and focused on rates determined by a channel's a-posteriori empirical behavior, and specifically on rate functions which are determined by the zero-order empirical distribution. This communication approach does not require a-priori specification of the channel model. The main result is that for discrete channels the empirical mutual information between the input and output sequences is attainable for any output sequence using feedback and common randomness, and for continuous real valued channels an effective "Gaussian capacity" $-\frac{1}{2}(1-\hat{\rho}^2)$ can be attained, while adapting the transmission rate to guarantee a prescribed error probability.

The framework proposed here is not completely satisfying, in two main senses. One issue is that there is no "channel capacity" in this framework (i.e., defining the maximum possible rate of communication is tricky), and another is that to complete the framework we need to fix an input distribution, which is not

TABLE III
PARAMETERS OF THE ADAPTIVE RATE SCHEME USED FOR FIG. 3

Item	Referrence	Parameter set 1 of Fig. 3	Parameter set 2
Transmission scheme	Section V-B	$n = 1e+008, K = 1e+006, P_J = 0.001, \epsilon = 0.001$	$n = 1e+020, K = 1e+017, P_J = 0.001, \epsilon = 0.001$
R_{LB1} parameters	Section VI-C2, Equation (65)	$T = 2.5e+005, \Delta_\mu = 37.5412, \Delta = 0.0345958, \eta_1 = 0.996007, \eta_2 = 0.999954, \delta_1 = 0.01$	$T = 7.5e+015, \Delta_\mu = 77.4043, \Delta = 3.14616e-007, \eta_1 = 1, \eta_2 = 1, \delta_1 = 0.001$
R_{LB2} parameters	Section VI-C2, Theorem 4	$\rho_0 = 0.9, \delta = 0.140878, \bar{R} = 1.05879$	$\rho_0 = 0.99998, \delta = 0.0068209, \bar{R} = 7.29818$

tuned to the channel. Although one may not be fully comfortable with this framework there is no question about the reality of the channel model itself, since the assumptions on the channel are minimized. In the traditional way of posing the communication problem, these issues are avoided, but at a price of assuming a statistical model for the channel, whose relation to the true channel may be questionable.

The current framework suggests a new viewpoint on the design of communication systems. The classical point of view first assumes a channel model and then devises a communication system optimized for it. Here we take the inverse direction: we devise a versatile, but not necessarily optimal, communication system without assumptions on the channel. Following these results, the individual channel approach becomes a very natural starting point for determining achievable rates for various probabilistic and arbitrary models (AVC-s, channels with an individual noise sequence, probabilistic models, compound channels) under the realm of randomized encoders, since the achievable rates for these models follow easily from the achievable rates for specific sequences, and the law of large numbers.

APPENDIX

A. Parameters of the Adaptive Rate Scheme Used for Fig. 3

Table III lists two sets of parameters for the continuous alphabet adaptive rate scheme. The first set was used for the curves in Fig. 3, and the second set shows the convergence of $\delta, \bar{R}$, for higher values of n, K. Note that the values of n, K are extremely high, and this is due to the looseness of the bounds used in the continuous case, specifically the constant offset Δ in the correlation factor due to Lemma 6 and the error exponent in this lemma.

B. Proof of Lemma 1

The proof is a rather standard calculation using the method of types. We use Csiszár's notations [10]. We divide the sequences according to their joint type $\mathcal{T}_{XY}$. The type $\mathcal{T}_{XY}$ is defined by the probability distribution $T_{XY} \in \mathcal{P}_n(\mathcal{XY})$. For notational purposes we define the dummy random variables $(\tilde{X}, \tilde{Y}) \sim T_{XY}$ and T_X, T_Y, $T_{Y|X}$ as the marginal and conditional distributions resulting from T_{XY}. The conditional type [10] is defined as $\mathcal{T}_{X|Y}(\mathbf{y}) \triangleq \{\mathbf{x} : (\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{XY}\}$. The empirical mutual information of sequences in the type $\mathcal{T}_{XY}$ is simply $I(\tilde{X}; \tilde{Y}) = I(T_Y, T_{Y|X})$. Define $T_t \triangleq \{T_{XY} \in \mathcal{P}_n(\mathcal{XY}) : I(T_Y, T_{Y|X}) \geq t\}$. Since all sequences in the conditional type have the same (marginal) type, we can write:

$$
\begin{aligned}
&Q^n\left(\hat{I}(\mathbf{x};\mathbf{y}) \geq t\right) \\
&= \sum_{\mathbf{x}:\hat{I}(\mathbf{x};\mathbf{y})\geq t} Q^n(\mathbf{x}) = \sum_{T_{XY}\in T_t} Q^n\left(\mathcal{T}_{X|Y}^{(T_{XY})}(\mathbf{y})\right) \\
&\overset{(a)}{=} \sum_{T_t} \left|\mathcal{T}_{X|Y}(\mathbf{y})\right| \exp\left\{-n\left[H(T_X) + D(T_X||Q)\right]\right\} \\
&\overset{(b)}{\leq} \sum_{T_t} \exp\left\{nH(\tilde{X}|\tilde{Y})\right\} \exp\left\{-n\left[H(\tilde{X}) + D(T_X||Q)\right]\right\} \\
&= \sum_{T_t} \exp\left\{-n\left[I(\tilde{X};\tilde{Y}) + D(T_X||Q)\right]\right\} \\
&\leq |\mathcal{P}_n(\mathcal{XY})| \cdot \exp\left\{-n\left(\min_{T_t}\left[I(T_Y, T_{X|Y}) + D(T_X||Q)\right]\right)\right\} \\
&\overset{(c)}{\leq} (n+1)^{|\mathcal{X}||\mathcal{Y}|} \cdot \exp(-nt) \\
&= \exp\left\{-n\left(t - |\mathcal{X}||\mathcal{Y}|\frac{\log(n+1)}{n}\right)\right\},
\end{aligned}
\tag{88}
$$

where (a) is due to [10, (II.1)], (b) results from (89) below which is an extension of (II.4) there to conditional types (and is a stronger version of Lemma II.3), based on the fact that in the conditional type $\mathcal{T}_{X|Y}(\mathbf{y})$ the values of $\mathbf{x}$ over the $n_a = nT_Y(a)$ indices for which $y_i = a$ have empirical distribution $T_{X|Y}$ and therefore the number of such sequences is limited to $\exp\left(n_a H(\tilde{X}|\tilde{Y} = a)\right)$, hence:

$$
\begin{aligned}
\left|\mathcal{T}_{X|Y}(\mathbf{y})\right| &\leq \prod_a \exp\left(nT_Y(a)H(\tilde{X}|\tilde{Y} = a)\right) \\
&= \exp\left(nH(\tilde{X}|\tilde{Y})\right).
\end{aligned}
\tag{89}
$$

(c) is based on bounding the number of types [15, Theorem 11.1.1], and the fact that in the minimization region $I(T_Y, T_{X|Y}) \geq t$ and $D(T_X||Q) \geq 0$ therefore the result of the minimum is at least t.

Note that for the proof of Theorem 1 we do not need the strict inequalities and equality in the error exponent would be sufficient, however these will be useful later for the rateless coding. An explanation for the fact that the bound does not depend on Q can be obtained by showing that $Q^n\left(\hat{I}(\mathbf{x};\mathbf{y}) \geq t\right)$ can be

bounded for each type of $\mathbf{x}$ separately. I.e. if $\mathbf{x}$ is drawn uniformly over the type $\mathcal{T}_X$ the probability of $\hat{I}(\mathbf{x};\mathbf{y}) \geq t$ is:

$$\frac{\sum_{T_{XY}\in T_t} \left|\mathcal{T}_{X|Y}(\mathbf{y})\right|}{|\mathcal{T}_X|} \doteq \frac{\sum_{T_{XY}\in T_t} \exp(nH(\tilde{X}|\tilde{Y}))}{\exp(nH(\tilde{X}))} = \sum_{T_{XY}\in T_t} \exp(-nI(\tilde{X};\tilde{Y})) \doteq \exp(-nt), \tag{90}$$

where $T_t \triangleq \left\{T_{XY} \in \mathcal{P}_n(\mathcal{X}\mathcal{Y}) : (T_{XY})_X = T_X, (T_{XY})_Y = T_Y, I(T_Y, T_{Y|X}) \geq t\right\}$ and since drawing $\mathbf{x} \sim Q^n$ is equivalent to first drawing the type of $\mathbf{x}$ and then drawing $\mathbf{x}$ uniformly over the type, the bound holds when $\mathbf{x} \sim Q^n$.

C. Proof of Lemma 2

For random variables X and Y where X is continuous (not necessarily Gaussian) we have the following bound on the conditional differential entropy ($\tilde{Y}$ denotes a dummy variable with the same distribution as Y and used for notational purposes):

$$\begin{aligned} h(X|Y) &= \mathop{\mathbb{E}}_{\tilde{Y}}\left[h\left(X\middle|Y=\tilde{Y}\right)\right] \\ &\overset{(a)}{\leq} \mathbb{E}\left[\frac{1}{2}\log\left(2\pi e\cdot \mathrm{Var}(X|Y)\right)\right] \\ &\overset{(b)}{\leq} \frac{1}{2}\log\left(2\pi e\mathbb{E}\left[\mathrm{Var}(X|Y)\right]\right) \\ &= \frac{1}{2}\log\left(2\pi e\mathbb{E}\left[\mathrm{Var}(X-\alpha\cdot Y|Y)\right]\right) \\ &\overset{(c)}{\leq} \frac{1}{2}\log\left(2\pi e\mathbb{E}(X-\alpha\cdot Y)^2\right) \underset{\alpha:=\frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}}{=} \\ &= \frac{1}{2}\log\left(2\pi e\left(\mathbb{E}(X^2)-\frac{\mathbb{E}(XY)^2}{\mathbb{E}(Y^2)}\right)\right) \\ &= \frac{1}{2}\log\left(2\pi e\mathbb{E}(X^2)(1-\rho^2)\right) \\ &= \frac{1}{2}\log\left(2\pi e\mathbb{E}(X^2)\right)+\frac{1}{2}\log\left(1-\rho^2\right), \end{aligned} \tag{91}$$

where the (a) is based on Gaussian bound for entropy and (b) on concavity of the $\log$ function (see also [15, (17.24)]) (c) is based on $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 \leq \mathbb{E}(X^2)$ and is similar to the assertion that $\mathbb{E}[\mathrm{Var}(X|Y)]$ which is the MMSE estimation error is not worse than the LMMSE estimation error (except our disregard for the mean).

Therefore for a Gaussian X:

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= \frac{1}{2}\log(2\pi e\mathbb{E}(X^2)) - h(X|Y) \overset{(91)}{\geq} -\frac{1}{2}\log(1-\rho^2). \end{aligned} \tag{92}$$

□

Proof of Corollary 2: Equality (a) holds only if $X|Y$ is Gaussian for every value of Y, (b) holds if X has fixed variance conditioned on every Y, and (c) if $\mathbb{E}(X-\alpha\cdot Y|Y) = 0 \Longrightarrow \mathbb{E}(X|Y) = \alpha\cdot Y$, therefore it results in $X|Y \sim \mathcal{N}(\alpha Y, \mathrm{const})$ which implies X, Y are jointly Gaussian (easy to check by calculating the pdf).

Note that if X, Y are jointly Gaussian then Y can be represented as a result of an additive white Gaussian noise channel (AWGN) with gain operating on X:

$$Y \sim \mathbb{E}(Y|X) + \mathcal{N}(0, \mathrm{Var}(Y|X)) = \tilde{\alpha}\cdot X + \mathcal{N}(0,\sigma^2) + \mathrm{const}. \tag{93}$$

To show the validity of Remark 3 consider $X = Y = \mathrm{Ber}(\frac{1}{2})$, in which case $I(X;Y) = 1$ and $\rho = 1$, therefore the assertion doesn't hold.

D. Proof of Lemma 4

Write the empirical correlation as

$$\hat{\rho} \triangleq \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} = \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^T\left(\frac{\mathbf{y}}{\|\mathbf{y}\|}\right). \tag{94}$$

From the expression above we can infer that $\hat{\rho}$ does not depend on the amplitude of $\mathbf{x}$ and $\mathbf{y}$ but only on their direction. Since $\mathbf{x}$ is isotropically distributed, the result does not depend on the direction of $\mathbf{y}$ (unless $\mathbf{y} = 0$ in which case it is trivially correct), therefore it is independent of $\mathbf{y}$ and we can conveniently choose $\mathbf{y} = (1,0,0,\ldots,0)$. To put the claim above more formally, for any unitary $n\times n$ matrix $\mathbf{U}$ we can write:

$$\begin{aligned} \hat{\rho} &= \frac{\mathbf{x}^T\mathbf{y}}{\sqrt{(\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})}} = \frac{\mathbf{x}^T\mathbf{U}^T\mathbf{U}\mathbf{y}}{\sqrt{(\mathbf{x}^T\mathbf{U}^T\mathbf{U}\mathbf{x})(\mathbf{y}^T\mathbf{U}^T\mathbf{U}\mathbf{y})}} \\ &= \left(\frac{\mathbf{U}\mathbf{x}}{\|\mathbf{U}\mathbf{x}\|}\right)^T\left(\frac{\mathbf{U}\mathbf{y}}{\|\mathbf{U}\mathbf{y}\|}\right). \end{aligned} \tag{95}$$

Since $\mathbf{x}$ is Gaussian, $\mathbf{U}\mathbf{x}$ has the same distribution of $\mathbf{x}$, thus the probability remains unchanged if we remove $\mathbf{U}$ from the left side and remain with $\hat{\rho}' = \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^T\left(\frac{\mathbf{U}\mathbf{y}}{\|\mathbf{U}\mathbf{y}\|}\right)$. For $\mathbf{y} \neq 0$, we may choose the unitary matrix $\mathbf{U}$ whose first row is $\frac{\mathbf{y}}{\|\mathbf{y}\|}$ and the other rows complete it to an orthonormal basis of the linear space $\mathbb{R}^n$. Then $\mathbf{U}\mathbf{y} = (\|\mathbf{y}\|, 0, 0, \ldots 0)$ and therefore $\left(\frac{\mathbf{U}\mathbf{y}}{\|\mathbf{U}\mathbf{y}\|}\right) = (1,0,0,\ldots 0)$. Thus the distribution of $\hat{\rho}' = (1,0,0,\ldots,0)\cdot\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right) = \frac{x_1}{\|\mathbf{x}\|}$ equals the distribution of $\hat{\rho}$. Assuming without loss of generality that $\mathbf{x} \sim \mathcal{N}^n(0,1)$ we have:

$$\begin{aligned} &\Pr(|\hat{\rho}| \geq t) = \Pr\left(\frac{x_1}{\|\mathbf{x}\|} \geq t\right) \\ &= \Pr\left(x_1^2 \geq t^2(\|\mathbf{x}_2^n\|^2 + x_1^2)\right) \\ &= \Pr\left(x_1^2 \geq \frac{t^2}{1-t^2}\|\mathbf{x}_2^n\|^2\right) \\ &= \mathbb{E}\left[\Pr\left(x_1^2 \geq \frac{t^2}{1-t^2}\|\mathbf{x}_2^n\|^2\right)\middle|\mathbf{x}_2^n\right] \\ &= \mathbb{E}\left[2Q\left(\sqrt{\frac{t^2}{1-t^2}\|\mathbf{x}_2^n\|^2}\right)\right] \leq \mathbb{E}\left[2e^{-\frac{1}{2}\frac{t^2}{1-t^2}\|\mathbf{x}_2^n\|^2}\right] \\ &= \int_{\mathbb{R}^{n-1}}\left(2e^{-\frac{1}{2}\frac{t^2}{1-t^2}\|\mathbf{x}_2^n\|^2}\right)\left(\frac{1}{(2\pi)^{(n-1)/2}}e^{-\frac{1}{2}\|\mathbf{x}_2^n\|^2}\right)d\mathbf{x}_2^n \\ &= 2\int_{\mathbb{R}^{n-1}}\frac{1}{(2\pi)^{(n-1)/2}}e^{-\frac{1}{2}\frac{1}{1-t^2}\|\mathbf{x}_2^n\|^2}\cdot d\mathbf{x}_2^n \\ &= 2(1-t^2)^{\frac{n-1}{2}}\int_{\mathbb{R}^{n-1}} f_{\mathcal{N}^{n-1}(0,1-t^2)}(\mathbf{x}_2^n)\cdot d\mathbf{x}_2^n \\ &= 2(1-t^2)^{\frac{n-1}{2}} = 2\exp\left(-(n-1)R_2(t)\right), \end{aligned} \tag{96}$$

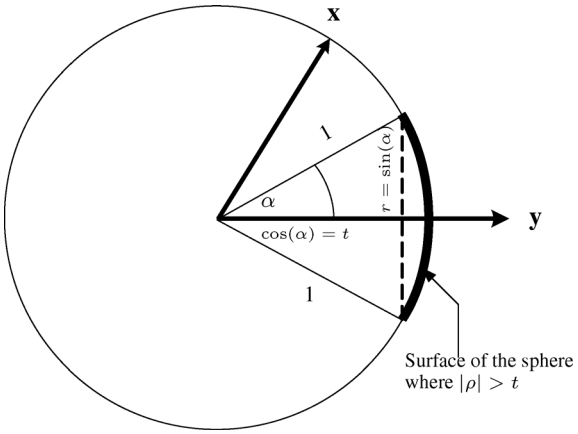

Fig. 6. A geometric interpretation of Lemma 4.

where we used the rough upper bound of the Gaussian error function $Q(x) \triangleq \Pr(\mathcal{N}(0,1) \geq x) \leq e^{-x^2/2}$, and $f_{\mathcal{N}^n(\mu,\sigma^2)}$ denotes the pdf of a Gaussian i.i.d. vector. □

Discussion: A result close to Lemma 4 can be obtained geometrically since $\Pr(|\hat{\rho}| > t)$ is related to the solid angle of the cone $\{\mathbf{x} : |\hat{\rho}| > t\}$. See Fig. 6. Since $\mathbf{x}$ is isotropically distributed, the probability to have $|\hat{\rho}| > t$ equals the relative surface determined by vectors having $|\hat{\rho}| > t$ on the unit n-ball (termed the solid angle). Since $\hat{\rho}$ is the cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$, the points where $|\hat{\rho}| > t$ generate a cone with inner angle 2α where $\cos(\alpha) = t$ and their intersection with the unit n-ball is a spherical cap (dome). We can obtain a similar bound as above using geometrical considerations. Write the volume of an n dimensional ball of radius r as $V_n r^n$ where V_n is a fixed factor $V_n = \frac{\pi^{n/2}}{\Gamma(1+n/2)}$ [33], and accordingly the surface of an n dimensional ball is (the derivative) $nV_n r^{n-1}$, then the relative surface of the spherical cap can be computed by integrating the surfaces of the $n-1$ dimensional balls with radius $\sin(\theta)$ that have a fixed angle θ with respect to $\mathbf{y}$, and can be bounded as follows:

$$\begin{aligned}
&\Pr(|\hat{\rho}| \geq t) = \frac{\text{Surface of cap}}{\text{Surface of sphere}} \\
&= \frac{1}{nV_n} \cdot \int_{\theta=0}^{\alpha} (n-1)V_{n-1} \sin^{n-2}(\theta) d\theta \\
&\leq \frac{V_{n-1}}{V_n} \cdot \sin^{n-3}(\alpha) \int_{\theta=0}^{\alpha} \sin(\theta) d\theta \\
&= \frac{V_{n-1}}{V_n} \cdot \sin^{n-3}(\alpha)(1-\cos(\alpha)) \\
&\overset{\alpha \leq \frac{\pi}{2}}{\leq} \frac{V_{n-1}}{V_n} \cdot \sin^{n-3}(\alpha)(1-\cos^2(\alpha)) \\
&= \Theta(\sqrt{n}) \cdot \sin^{n-1}(\alpha) = \Theta(\sqrt{n}) \cdot \sqrt{1-\cos^2(\alpha)}^{n-1} \\
&= \Theta(\sqrt{n}) \cdot (1-t^2)^{(n-1)/2}, \qquad (97)
\end{aligned}$$

where the asymptotic ratio $\frac{V_{n-1}}{\sqrt{n}V_n} \to 1$ is based on [34, (99)]. Compare the result with Lemma 4. An interesting observation is that the assumption of Gaussian distribution is not necessary and this bound is true for all isotopical distributions.

E. Proof of Lemma 6

We denote $\mathbf{x}_i, \mathbf{y}_i$ as the sub-vectors over A_i (i.e., $\mathbf{x}_i \triangleq \mathbf{x}_{A_i}, \mathbf{y}_i \triangleq \mathbf{y}_{A_i}$), their length by $n_i \triangleq |A_i|$ and their relative length by $\lambda_i = n_i/n$. We are interested to find a subset J of $\mathbf{x}$ with bounded probability such that outside the set $\sum_i \lambda_i \hat{\rho}_i^2 \geq \hat{\rho}^2 - \Delta$ for any $\mathbf{y}$. Consider the following inequality:

$$\begin{aligned}
&\|\mathbf{x}\|^2 \cdot \|\mathbf{y}\|^2 \cdot \hat{\rho}^2 = \left(\mathbf{x}^T\mathbf{y}\right)^2 = \left(\sum_i \mathbf{x}_i^T \mathbf{y}_i\right)^2 \\
&= \left(\sum_i \hat{\rho}_i \|\mathbf{x}_i\| \cdot \|\mathbf{y}_i\|\right)^2 \\
&\overset{(a)}{\leq} \left(\sum_i \hat{\rho}_i^2 \|\mathbf{x}_i\|^2\right) \cdot \left(\sum_i \|\mathbf{y}_i\|^2\right) \\
&= \left(\sum_i \lambda_i \hat{\rho}_i^2 + \sum_i \hat{\rho}_i^2 \left(\frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}\|^2} - \lambda_i\right)\right) \cdot \|\mathbf{x}\|^2 \cdot \|\mathbf{y}\|^2 \\
&\overset{(b)}{\leq} \left(\sum_i \lambda_i \hat{\rho}_i^2 + \sum_i \max\left(\frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}\|^2} - \lambda_i, 0\right)\right) \cdot \|\mathbf{x}\|^2 \cdot \|\mathbf{y}\|^2, \qquad (98)
\end{aligned}$$

where (a) is from Cauchy-Swartz inequality and (b) is since $\hat{\rho}_i^2 z_i \leq \max(\hat{\rho}_i^2 z_i, 0) = \hat{\rho}_i^2 \max(z_i, 0) \leq \max(z_i, 0)$ and is attained for $|\hat{\rho}_i| = \mathrm{Ind}(z_i > 0)$. Both inequalities are tight in the sense that for each $\mathbf{x}$ there is a sequence $\mathbf{y}$ (equivalent to choosing $\{\|\mathbf{y}_i\|^2, \hat{\rho}_i\}$) that meets them in equality, provided that $\forall i : n_i \geq 2, \mathbf{x}_i \neq 0$. Dividing by $\|\mathbf{x}\|^2 \cdot \|\mathbf{y}\|^2$ we have that

$$\hat{\rho}^2 - \sum_i \lambda_i \hat{\rho}_i^2 \leq \sum_i \max\left(\frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}\|^2} - \lambda_i, 0\right), \qquad (99)$$

where the RHS depends only on $\mathbf{x}$ and should be bounded by Δ. Thus the minimal set J_Δ is:

$$J_\Delta \triangleq \left\{\mathbf{x} : \sum_i \max\left(\frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}\|^2} - \lambda_i, 0\right) > \Delta\right\}. \qquad (100)$$

The set is minimal in the sense that none of its elements can be removed while meeting the conditions of the lemma. We would like to bound the probability of J_Δ. The result of $\sum_i \max(z_i, 0)$ is a partial sum of z_i, and since negative z_i are not summed, it is easy to see this is the maximal partial sum, i.e., we can write this sum alternatively as

$$\sum_i \max(z_i, 0) = \max_{I \in \mathcal{P}} \sum_{i \in I} z_i, \qquad (101)$$

where $\mathcal{P} \triangleq 2^{\{1,\ldots,p\}} \setminus \emptyset$ denotes all non empty sub-sets of $\{1, \ldots, p\}$, and its size is $2^p - 1$. Therefore from the union bound we have:

$$\begin{aligned}
\Pr\{J_\Delta\} &= \Pr\left\{\max_{I \in \mathcal{P}} \sum_{i \in I} \left(\frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}\|^2} - \lambda_i\right) > \Delta\right\} \\
&\leq \sum_{I \in \mathcal{P}} \Pr\left\{\sum_{i \in I} \left(\frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}\|^2} - \lambda_i\right) > \Delta\right\}. \qquad (102)
\end{aligned}$$

To bound the above probability we use the following bound on the probability $\Pr\left(\sum_i a_i \|\mathbf{x}_i\|^2 \le 0\right)$ for some coefficients a_i:

Lemma 7: Let $\mathbf{x} \sim \mathcal{N}(0,P)^n$, and $\mathbf{x}_i, \lambda_i$ defined as above. For coefficients $\{a_i\}_{i=1}^p$ with $\sum_i \lambda_i a_i = \bar{a} > 0$ and $|a_i| \le A$ where $|\bar{a}| \le \frac{1}{8}A$, we have

$$\Pr\left(\sum_i a_i \|\mathbf{x}_i\|^2 \le 0\right) \le e^{-nE_a}, \tag{103}$$

where

$$E_a = \frac{\bar{a}^2}{6A^2}. \tag{104}$$

The Lemma is proven at the end of this subsection. Now we apply the bound to the events in (102):

$$\begin{gathered}\sum_{i\in I}\left(\frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}\|^2} - \lambda_i\right) > \Delta \\ \Updownarrow \\ \sum_{i\in I}\|\mathbf{x}_i\|^2 - \sum_{i\in I}\lambda_i \cdot \sum_{i=1}^p \|\mathbf{x}_i\|^2 > \Delta \sum_{i=1}^p \|\mathbf{x}_i\|^2 \\ \Updownarrow \\ \sum_{i=1}^p \underbrace{\left(\Delta + \sum_{j\in I}\lambda_j - \mathrm{Ind}(i\in I)\right)}_{\triangleq a_i} \|\mathbf{x}_i\|^2 < 0\end{gathered}$$

We have:

$$\begin{aligned}\bar{a} &= \sum_{i=1}^p \lambda_i a_i = \Delta \cdot \underbrace{\sum_{i=1}^p \lambda_i}_{1} + \sum_{i\in I}\lambda_i \cdot \sum_{i=1}^p \lambda_i \\ &\quad - \sum_{i=1}^p \mathrm{Ind}(i\in I)\lambda_i \\ &= \Delta,\end{aligned} \tag{105}$$

and $|a_i| \le 1 + \Delta \triangleq A$, therefore for $\Delta \le 1/7$ we have $\bar{a} \le \frac{1}{8}A$ and by Lemma 7:

$$\Pr\left\{\sum_{i\in I}\left(\frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}\|^2} - \lambda_i\right) > \Delta\right\} \le e^{-nE_a} \le e^{-nE_0}, \tag{106}$$

where

$$E_a = \frac{\bar{a}^2}{6A^2} = \frac{\Delta^2}{6(1+\Delta)^2} \ge \frac{\Delta^2}{6(1+1/7)^2} \ge \frac{\Delta^2}{8} \triangleq E_0. \tag{107}$$

From (102) we have:

$$\begin{aligned}\Pr\{J_\Delta\} &\le \sum_{I\in\mathcal{P}} \Pr\left\{\sum_{i\in I}\left(\frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}\|^2} - \lambda_i\right) > \Delta\right\} \\ &\le |\mathcal{P}| \cdot e^{-nE_0} \le 2^p e^{-nE_0},\end{aligned} \tag{108}$$

which proves the lemma. Note that different bounds can be obtained by applying the bound on m smaller sets in $\{1,\ldots,p\}$ and requiring that the sum over each set will be bounded by Δ/m (as an example we could bound each $\max(z_i,0)$ separately by Δ/p), however this bound is most suitable for our purpose since when $p \ll n$ the element 2^p becomes negligible. □

Proof of Lemma 7: We assume without loss of generality that $\mathbf{x} \sim \mathcal{N}(0,1)^n$. For a Gaussian r.v. $X \sim \mathcal{N}(0,1)$ and $a < \frac{1}{2}$ we have:

$$\begin{aligned}\mathbb{E}(e^{aX^2}) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{(a-\frac{1}{2})x^2} dx \\ &= \frac{1}{\sqrt{1-2a}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-2a)^{-1}}} e^{-\frac{x^2}{2(1-2a)^{-1}}} dx \\ &= \frac{1}{\sqrt{1-2a}}.\end{aligned} \tag{109}$$

For coefficients $\{a_i\}_{i=1}^p$ with $\sum_i \lambda_i a_i = \bar{a} > 0$ and $|a_i| \le A$, $w > 0$ a positive constant of our choice, and $\mathbf{x} \sim \mathcal{N}(0,1)^n$ we have, using the Chernoff bounding technique:

$$\begin{aligned}&\ln \Pr\left(\sum_i a_i \|\mathbf{x}_i\|^2 \le 0\right) \le \ln \mathbb{E}\left[e^{-\frac{1}{2}w\cdot\sum_i a_i\|\mathbf{x}_i\|^2}\right] \\ &= \ln \mathbb{E}\left[e^{-\frac{1}{2}w\cdot\sum_i a_i \sum_{j\in A_i} x_j^2}\right] = \ln \prod_i \prod_{j\in A_i} \mathbb{E}\left[e^{-\frac{1}{2}w\cdot a_i\cdot x_j^2}\right] \\ &= \sum_i \sum_{j\in A_i} \ln\left((1+w\cdot a_i)^{-\frac{1}{2}}\right) = -\frac{1}{2}n\sum_i \lambda_i \ln(1+w\cdot a_i) \\ &\overset{(a)}{=} -\frac{1}{2}n\sum_i \lambda_i\left((w\cdot a_i) - \frac{1}{2}\frac{1}{(1+w\cdot t_i)^2}(w\cdot a_i)^2\right) \\ &\overset{(b)}{\le} -\frac{1}{2}n\sum_i \lambda_i\left((w\cdot a_i) - \frac{1}{2}\frac{1}{(1-w\cdot A)^2}(w\cdot A)^2\right) \\ &= -\frac{1}{2}n\left(\bar{a}w - \frac{A^2w^2}{2(1-w\cdot A)^2}\right),\end{aligned} \tag{110}$$

where (a) is based on the second order Tailor series of $\ln(1+wt)$ around $t=0$ with some $t_i \in [0,a_i] \cup [a_i,0]$ and (b) is since $|t_i| \le |a_i| \le A$. For simplicity we choose a sub-optimal $w^* = \frac{\bar{a}}{A^2}$ (which is obtained by assuming small a, w and optimizing the bound with respect to w ignoring the denominator) and obtain:

$$\begin{aligned}\bar{a}w^* - \frac{A^2w^{*2}}{2(1-w^*\cdot A)^2} &= \frac{\bar{a}^2}{A^2} - \frac{\bar{a}^2/A^2}{2(1-\bar{a}/A)^2} \\ &= \frac{\bar{a}^2}{A^2}\left(1 - \frac{A^2}{2(A-\bar{a})^2}\right).\end{aligned} \tag{111}$$

To simplify the bound, we make a further assumption that $|\bar{a}| \le \frac{1}{8}A$ therefore:

$$\begin{aligned}\frac{\bar{a}^2}{A^2}\left(1 - \frac{A^2}{2(A-\bar{a})^2}\right) &\ge \frac{\bar{a}^2}{A^2}\left(1 - \frac{A^2}{2\cdot(7/8)^2\cdot A^2}\right) \\ &= \frac{\bar{a}^2}{A^2}\cdot\frac{17}{49} \ge \frac{\bar{a}^2}{3A^2}.\end{aligned} \tag{112}$$

Combining (110), (111) and (112) we have the following bound: for $|\bar{a}| \leq \frac{1}{8}A$,

$$\Pr\left(\sum_i a_i \|\mathbf{x}_i\|^2 \leq 0\right) \leq e^{-nE_a}, \tag{113}$$

where $E_a = \frac{\bar{a}^2}{6A^2}$. Note that the bound is true for any $\mathbf{x} \sim \mathcal{N}(0,P)^n$. □

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers of the paper, as well as the reviewers of the ISIT 2009 conference paper on the subject for the helpful comments and references they contributed.

REFERENCES

[1] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.
[2] O. Shayevitz and M. Feder, "Achieving the empirical capacity using feedback: Memoryless additive models," *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1269–1295, Mar. 2009.
[3] O. Shayevitz and M. Feder, "Communicating using feedback over a binary channel with arbitrary noise sequence," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Sep. 2005.
[4] K. Eswaran, A. Sarwate, A. Sahai, and M. Gastpar, "Zero-rate feedback can achieve the empirical capacity," *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 25–39, Jan. 2010.
[5] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Control Inf. Theory*, vol. 4, pp. 97–102, 1975.
[6] Y. Lomnitz and M. Feder, "An achievable rate for the MIMO individual channel," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Aug. 2010.
[7] Y. Lomnitz and M. Feder, "Communicating over modulo-additive channels with compressible individual noise sequence," in *26th IEEE Conv. Electr. Electron. Eng. Israel (IEEEI)*, Nov. 2010.
[8] Y. Lomnitz and M. Feder, "Universal communication over modulo-additive individual noise sequence channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2011.
[9] O. Shayevitz and M. Feder, "The posterior matching feedback scheme: Capacity achieving and error analysis," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2008, pp. 900–904.
[10] I. Csiszár, "The method of types [information theory]," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.
[11] A. Tchamkerten and I. E. Telatar, "Variable length coding over an unknown channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2126–2145, May 2006.
[12] B. Hughes and T. Thomas, "On error exponents for arbitrarily varying channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 87–98, Jan. 1996.
[13] M. V. Burnashev, "Data transmission over a discrete channel with feedback: Random transmission time," *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 250–265, 1976.
[14] N. Shulman and M. Feder, "The uniform distribution as a universal prior," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1356–1362, Jun. 2004.
[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[16] R. Zamir and U. Erez, "A Gaussian input is not too bad," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1362–1367, Jun. 2004.
[17] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.
[18] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Stat.*, vol. 31, pp. 558–567, 1960.
[19] A. Lapidoth and I. Telatar, "The compound channel capacity of a class of finite-state channels," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 973–983, May 1998.
[20] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited : Positivity, constraints," *IEEE Trans. Inf. Theory*, vol. 34, no. 2, pp. 181–193, Mar. 1988.
[21] M. Agarwal, A. Sahai, and S. Mitter, "Coding into a source: A direct inverse rate-distortion theorem," presented at the 44th Allerton Conf. Commun., Control, Comput., Monticello, IL, Oct. 2006.
[22] M. Langberg, "Oblivious communication channels and their capacity," *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 424–429, Jan. 2008.
[23] A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1520–1529, Sep. 1996.
[24] B. Hughes and P. Narayan, "Gaussian arbitrarily varying channels," *IEEE Trans. Inf. Theory*, vol. 33, no. 2, pp. 267–284, Mar. 1987.
[25] N. Shulman, "Communication Over an Unknown Channel via Common Broadcasting," Ph.D. Dissertation, Tel Aviv University, , 2003.
[26] O. Shayevitz and M. Feder, "Communication with feedback via posterior matching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2007, pp. 391–395.
[27] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inf. Theory*, vol. IT-9, no. 3, pp. 136–143, Jul. 1963.
[28] J. P. M. Schalkwijk, "A coding scheme for additive noise channels with feedback part II: Band-limited signals," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 183–189, Apr. 1966.
[29] T. Han and M. Hoshi, "Interval algorithm for random number generation," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 599–611, Mar. 1997.
[30] R. Ahlswede, "Elimination of correlation in random codes for arbitrarily varying channels," *Probab. Theory Relat. Fields*, vol. 44, pp. 159–175, 1978.
[31] M. Langberg, "Private codes or succinct random codes that are (almost) perfect," in *Proc. 45th Annu. IEEE Symp. Found. Comput. Sci.*, Oct. 2004, pp. 325–334.
[32] A. Barron, J. Rissanen, and Y. Bin, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.
[33] E. W. Weisstein, Ball. From MathWorld—A Wolfram Web Resource [Online]. Available: http://mathworld.wolfram.com/Ball.html
[34] E. W. Weisstein, Gamma Function. From MathWorld—A Wolfram Web Resource [Online]. Available: http://mathworld.wolfram.com/GammaFunction.html

Yuval Lomnitz received the B.Sc. degree from the Technion, Israel in 1995 and the M.Sc. degree from Tel Aviv University, Israel in 2003, both in electrical engineering. Since 2007 he is a Ph.D. student in the Department of Electrical engineering—Systems in Tel Aviv University. His research focuses on communication over unknown channels, using feedback.

During 2001–2004 he was with Envara working on WiFi, and from 2004 he is with Intel working on wireless communication.

Mr. Lomnitz is a recipient of the Weinstein prize for a scientific publication in the field of signal processing (2010), the Weinstein prize for excellence in studies (2011), and the Feder Family prize for best student work in communication technology (2011).

Meir Feder (S'81–M'87–SM'93–F'99) received the B.Sc. and M.Sc. degrees from Tel-Aviv University, Israel, and the Sc.D. degree from the Massachusetts Institute of Technology (MIT) Cambridge, and the Woods Hole Oceanographic Institution, Woods Hole, MA, all in electrical engineering in 1980, 1984 and 1987, respectively.

After being a research associate and lecturer in MIT, he joined the Department of Electrical Engineering—Systems, Tel-Aviv University, where he is now a Professor. He had visiting appointments at the Woods Hole Oceanographic Institution, Scripps Institute, Bell Laboratories, and in 1995/1996 he has been a visiting professor at MIT. He is also extensively involved in the high-tech industry and co-founded several companies including Peach Networks, a developer of a unique server-based interactive TV solution which was acquired on March 2000 by Microsoft, and Amimon a leading provider of ASIC's for wireless high-definition A/V connectivity at the home.

Prof. Feder is a co-recipient of the 1993 IEEE Information Theory Best Paper Award. He also received the 1978 "creative thinking" award of the Israeli Defense Forces, the 1994 Tel-Aviv University prize for Excellent Young Scientists, the 1995 Research Prize of the Israeli Electronic Industry, and the research prize in applied electronics of the Ex-Serviceman Association, London, awarded by Ben-Gurion University. Between June 1993 and June 1996 he served as an Associate Editor for Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY.