Forty-Seventh Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 30 - October 2, 2009

Universal capacity of channels with given rate-distortion in absence of common randomness, and failure of universal source-channel separation

Mukul Agarwal Swastik Kopparty Sanjoy Mitter

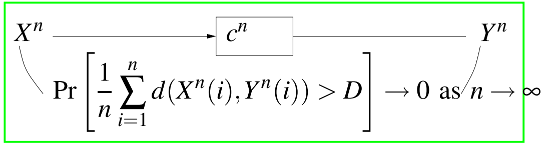
Abstract—**Recently, [3] studied the universal capacity of a set of channels where each channel in the set communicates a random source to within a distortion level D, when the transmitter and receiver have access to *common randomness*. In this paper, we study the universal capacity for this channel set in the case when there is no access to common randomness. We show that when the distortion level D is positive, the universal capacity is 0. This also leads to the conclusion that universal source-channel separation for rate-distortion as stated in [4], is false.**

I. Introduction

For a given distortion level D, let $\mathcal{C}$ be the set of channels which directly transmit a given source X such that

$$\lim_{n\to\infty} \Pr\left[\frac{1}{n}\sum_{i=1}^{n} d(X^n(i),Y^n(i)) > D\right] = 0 \qquad (1)$$

The above is a probabilistic criterion. See figure.


X^n
c^n
Y^n
$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} d(X^n(i),Y^n(i)) > D\right] \to 0$ as $n \to \infty$

c^n, above, refers to channel realization when sequence length (block length) is n. See Section III for precise definitions of c^n, X^n, Y^n, d

In [3], the problem of finding the univeral capacity of $\mathcal{C}$ was considered. They showed that, when the transmitter and receiver have access to common randomness, the universal capacity of $\mathcal{C}$ is precisely the rate-distortion function $R_X(D)$. The main idea of that result is that the common randomness can be used to generate a random code which is independent of the channel, and the transmitter and receiver can then communicate using this code.

In Shannon's random-coding argument, the existence of a random code for reliable communication implies the existence of deterministic code for reliable communication. This is true because there is only *one* channel, not a set of channels. In general, when asking the question of universal capacity of a set of channels, the existence of a random code does not imply the existence of a deterministic/individually stochastic code for the entire set of channels, unless there is common randomness at transmitter and receiver. We would use the phrases "random code", "common randomness,", and "stochastic-coupled encoder-decoder" (Section III) interchangeably.

In this paper we consider the case when there is no common randomness (although the transmitter and receiver can be independently stochastic). Our main result is that in the case of no common randomness, if the rate-distortion level D is > 0, then the universal capacity of $\mathcal{C}$ is *zero*.

The main motivation for this question comes from the the universal source-channel separation theorem for rate-distortion proved in [4]. We briefly recall the statement of this separation theorem:

> Let $\mathcal{A}$ be a set of channels. If there is common randomness at transmitter, in order to universally communicate i.i.d. X source to within a distortion level D in the sense of Equation 1 over $\mathcal{A}$, it is sufficient to consider architectures which consist of rate-distortion source-coding i.i.d. X source to within a distortion level in the sense of Equation 1 followed by universal reliable communication over $\mathcal{A}$.

The proof uses the fact that the universal capacity of $\mathcal{C}$ with common randomness is $R_X(D)$. Our result shows that without common randomness, if $D > 0$ the universal capacity of $\mathcal{C}$ is is zero, even though $R_X(D) > 0$. As a corollary, the universal source-channel separation theorem is false when there is no common randomness at transmitter and receiver.

Our proof uses some combinatorial ideas along with a powerful inequality due to Bonami and Beckner [5]. This inequality, which is a cornerstone of the modern study of boolean functions, was first applied to combinatorial situations by the highly influential paper of Kahn, Kalai and Linial [6]. We believe that these techniques could have applications to a wide range of information theoretic problems.

For simplicity of exposition, in this version of the paper we only consider the case of binary alphabets with Hamming distortion. The same ideas can also be used for the general case.

II. Past Work

Shannon had considered the same question of communication over channels which communicate i.i.d. X source within a

978-1-4244-5871-4/09/$26.00 ©2009 IEEE

distortion D [1]. Shannon considered just one channel, and instead of the "probabilistic criterion", considered an expectation criterion. The questions of universal channel capacity when there is no common randomness exist in literature. There are examples in literature where random codes perform much better than deterministic codes example [8].

III. Notation and Definitions

Sets, random variables, and distortion measure:

$\mathcal{X}$: finite set: channel input space

$\mathcal{Y}$: finite set: channel output space

$\mathcal{Y}$: finite set: channel output space.

X: random variable on $\mathcal{X}$.

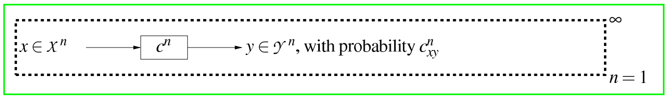
p_X: probability distribution of X. X^n: iid X sequence of random variables of length n

$d : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$ is a non-negative valued function. Think of $d(x,y)$ as the distortion incurred when $x \in \mathcal{X}$ is decoded as $y \in \mathcal{Y}$. For sequences of length n, $x^n \in \mathcal{X}^n$, $y^n \in \mathcal{Y}^n$, the average distortion measure $\frac{1}{n}\sum_{i=1}^{n} d(x^n(i), y^n(i))$. is used.

Notation 3.1 (superscript n): A superscript n will denote a variable when sequence length(or block length) is n.

Channel model: A channel is a sequence of transition probability matrices $\langle c^n_{xy}\rangle_1^\infty$. This channel will be denoted by $\langle c^n\rangle_1^\infty$. It's operation ' should be thought of, as follows:

when the length of the input sequence is n, the channel input space is $\mathcal{X}^n$, the channel output space is $\mathcal{Y}^n$, and the channel acts as c^n: c^n_{xy} denotes the probability that the channel output is $y \in \mathcal{Y}^n$ when the channel input is $x \in \mathcal{X}^n$. No causality or nestedness assumptions are assumed on $\langle c^n\rangle_1^\infty$.



This channel model is the same as the channel model in the paper of Verdu and Han [2].

Example 3.2 (Binary Symmetric Channel, BSC(D)): $\mathcal{X} = \mathcal{Y} = \{0,1\}$. $c^1_{0,0} = c^0_{1,1} = 1 - D$. $c^1_{0,1} = c^0_{1,0} = D$. The channel flips a bit with probability D. c^n_{ij} is the product of matrices c^1: the channel acts independently at each time.

Example 3.3 (Random walk channel, RWC(D)): The set $\{0,1\}^n$ can be thought of as the vertices of a hypercube. RWC is a random walk on the hypercube. Each point (sequence x) has n neighbors, $r_1, \ldots, r_n$: the sequences which are at a hamming distance 1 from the point. At the next time, the random walk jumps to one of these neighbors, each with probability $\frac{1}{n}$. This continues for nD jumps.

Definition 3.4 ($\mathcal{C}_{X,D}$): Consider a channel $\langle c^n\rangle_1^\infty$. If the input to the channel is i.i.d. X source X^n, the channel acts as c^n. The channel output is a random variable Y^n on $\mathcal{Y}^n$. A channel is said to belong to $\mathcal{C}_{X,D}$ if, under the joint distribution $p_{X^nY^n}$ on the input-output space,

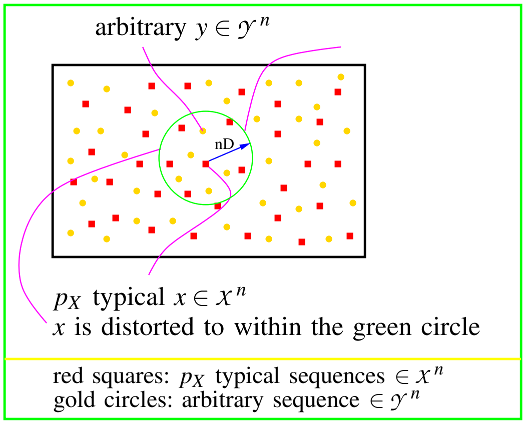
$$\Pr\left(\sum_{i=1}^{n}\frac{1}{n}d(X^n(i), Y^n(i)) > D\right) \to 0 \text{ as } n \to \infty \tag{2}$$

See figure in Section I.

Note 3.5: Y^n need not be i.i.d. Recall Notation 3.1.

The i.i.d. X sequence X_i is just a tool in the definition of the channel set $\mathcal{C}_{X,D}$. It does not mean that one is trying to communicate i.i.d. X source over the channel.

$\mathcal{C}_{X,D}$ is a set of channels. Intuitively, one can think of a channel in $\mathcal{C}_{X,D}$ as follows: a p_X-typical sequence of length n suffers a distortion $< nD$ after passing through the channel with high probability, for most p_X typical sequences.


arbitrary $y \in \mathcal{Y}^n$
nD
p_X typical $x \in \mathcal{X}^n$
x is distorted to within the green circle
red squares: p_X typical sequences $\in \mathcal{X}^n$
gold circles: arbitrary sequence $\in \mathcal{Y}^n$

Note 3.6: In what follows, whenever we talk about communicating a source to within a distortion level D, or compressing a source to within a distortion level D, it will be in sense (2).

Process of Communication and Universal channel capacity

Communication will be done using block codes. For block length n,

Channel input space $\mathcal{X}^n$, is the cartesian product of $\mathcal{X}$, n times.. $\mathcal{X}^n = \{x_1, x_2, \ldots, x_{|\mathcal{X}|^n}\}$.

Channel output space $\mathcal{Y}^n$, is the cartesian product of $\mathcal{Y}^n$, n times. $\mathcal{Y}^n = \{y_1, y_2, \ldots, y_{|\mathcal{X}|^n}\}$.

Suppose we want to communicate at rate R.

Message set $\mathcal{M}^n = \{m_1, m_2, \ldots, m_{2^{nR}}\}$. Message reproduction set is denoted by $\hat{\mathcal{M}}^n$. The elements of $\hat{\mathcal{M}}^n$ are the same as that of $\mathcal{M}^n$.

A *deterministic* encoder is a map $e : \mathcal{M}^n \to \mathcal{X}^n$. A deterministic decoder is a map $d : \mathcal{Y}^n \to \hat{\mathcal{M}}^n$. Deterministic encoder-decoder will be denoted as d-encoder-decoder.

A *Stochastic-decoupled encoder-decoder* is a pair of stochastic matrices:

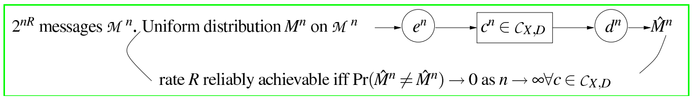
- *The encoder* is a stochastic matrix, p^n_{mx}, $m \in \mathcal{M}^n, x \in \mathcal{X}^n$. This should be interpreted as: message m is encoded to sequence x with probability p^n_{mx}.
- *The decoder* is a stochastic matrix $q^n_{y\hat{m}}$, $y \in \mathcal{Y}^n, \hat{m} \in \hat{\mathcal{M}}^n$ This should be interpreted as: channel output y is decoded to message m with probability $q^n_{y\hat{m}}$

The pair is called stochastic-decoupled encoder-decoder. This is because the encoder and decoder are individually stochastic and act independently of each other. These will be denoted by sd-encoder-decoder. *sd-encoder-decoder are the consideration in this paper.*

A *stochastic-coupled encoder-decoder* is the same as a random code. The encoder comes from a family of codes and the decoder has access to the realization of the encoder. One way in which encoder-decoder can generate a random code is through *common randomness*. Common randomness is defined as the encoder and decoder having access to the realizations of a continuous valued random variable. These realizations can then be used to generate random codes. *Stochastically coupled encoder-decoder were considered in [3].* They are *not* under consideration in this paper.

Encoders and decoders are sequences, $\langle e^n, d^n \rangle_1^\infty$, where n is the block length.

Universal Channel Capacity Consider a uniform distribution M^n on $\mathcal{M}^n$. Thus, $p_{M^n}(m) = \frac{1}{2^{nR}} \forall m \in \mathcal{M}^n$. The composition of the M^n, encoder, channel and decoder results in an output random variable $\hat{M}^n$ on $\hat{\mathcal{M}}$. This induces a joint probability distribution $p_{M^n\hat{M}^n}$ on the message-message reproduction space $\mathcal{M}^n \times \hat{\mathcal{M}}^n$. Rate R is universally achievable over $\mathcal{C}_{X,D}$ under the average block error probability criterion if there exist encoder-decoder pair such that under this joint probability distribution, $\Pr(\hat{M}^n \neq M^n) \to 0$ as $n \to \infty$ for each channel in $\mathcal{C}_{X,D}$. *Encoder-decoder should be independent of the channel. Supremum of achievable rates is called the universal channel capacity of* $\mathcal{C}_{X,D}$.


2^{nR} messages $\mathcal{M}^n$. Uniform distribution M^n on $\mathcal{M}^n$
e^n
$c^n \in \mathcal{C}_{X,D}$
d^n
$\hat{M}^n$
rate R reliably achievable iff $\Pr(\hat{M}^n \neq \hat{M}^n) \to 0$ as $n \to \infty \forall c \in \mathcal{C}_{X,D}$

Universal capacity can analogously be defined for any set $\mathcal{A}$.

The channel set can be interpreted as an adversary against reliable communication. First, the encoder and decoder are chosen and then, channel set acts on the output of the encoder. Thus, the channel set can choose the "worst" channel corresponding to this encoder-decoder.

The $\Pr(\hat{M}^n \neq M^n) \to 0$ as $n \to \infty$ is the average block error probability criterion. Other criteria exist. They will not be considered in this paper.

Definition 3.7 (C_d, C_{sd}, C_{sc}): : When encoder-decoder are required to be deterministic, universal capacity of $\mathcal{C}_{X,D}$ will be denoted by C_d. C_{sd} and C_{sc} are defined analogously.

IV. The main theorem

As mentioned in the introduction, here we will only deal with binary alphabets and Hamming distortion. We now fix some notation for binary channels and describe $\mathcal{C}_{X,D}$ in this case:

Example 4.1 (Binary input hamming distortion channels): • $\mathcal{X} = \mathcal{Y} = \{0,1\}$.

- d: Hamming distortion metric; i.e., $d(0,0) = d(1,1) = 0$ and $d(0,1) = d(1,0) = 1$
- X: a uniformly random bit: $p_X(0) = p_X(1) = \frac{1}{2}$
- $D \geq 0$: distortion level.
- $\mathcal{C}_{X,D}$: The definition from Equation ((2)), with the above choices for $\mathcal{X}, \mathcal{Y}, d, X$.

Roughly, for a channel in $\mathcal{C}_{X,D}$, the number of bit errors in an n length bit sequence is $< nD$ with high probability for most bit sequences.

Note 4.2: In what follows, $\mathcal{C}_{X,D}$ will refer to this example. C_d, C_{sd}, and C_{sc} will refer to the channel capacity for this particular $\mathcal{C}_{X,D}$.

We can now state our main theorem.

Theorem 4.3 (Main): Let $D \in [0,1]$ and let $\mathcal{C} = \mathcal{C}_{X,D}$, C_d, C_{sd}, C_{sc} be as above. Then:

1) a) If $D = 0$, rate 1 is achievable for d-encoder-decoder, whereas rates $0 < R < 1$ are not achievable.
   b) If $D = 0$, $C_{sd} = 1$.
2) If $D > 0$, $C_{sd} = 0$ (and thus, $C_d = 0$).

Note 4.4: Analogous theorem can we stated for general channel set $\mathcal{C}_{X,D}$. However, we do not do that in this paper.

V. Universal source-channel separation for rate-distortion

Theorem 4.3, (2) proves, by counterexample, that universal-source channel separation for rate-distortion, as stated and in [4], and recalled above in Section I, is false when there is no common randomness at transmitter and receiver. Note that in the universal source-channel separation theorem, universality is over the channel, that is, the channel is a set of channels, not just one channel.

The proof in [4] relies crucially on the fact that universal capacity of channel set $\mathcal{C}_{X,D}$ is $R_X(D)$ when there is common randomness at transmitter and receiver.

From the proof in [4], one sees the connection between universal source-channel separation to universal capacity of $\mathcal{C}_{X,D}$, and the crucial reliance of the proof on the fact that the universal capacity of $\mathcal{C}_{X,D}$ is $R_X(D)$ when there is common randomness at transmitter and receiver.

VI. Intuition

Recall that $\mathcal{C}_{X,D}$, C_d, C_{sd}, and C_{sc} will refer to the channel set in Example 4.1.

Intuition for Theorem 4.3, (1)

$C_{X,0}$ (Note, $C_{X,0}$, not $C_{X,D}$: this is $C_{X,D}$ with $D=0$) consists of channels where most bit sequences are, with high probability, perfectly received.

$C_{sd} \leq 1$ (and hence, $C_d \leq 1$)because there are only 2^n possible sequences in the input space $\mathcal{X}^n$.

Intuition for Theorem 4.3 (1a)

To transmit at rate $R=1$, there are 2^n messages. The encoder encodes each message to a bit sequence (both message set and channel input space have cardinality 2^n). The decoder decodes a received sequence to the corresponding message. This results in reliable communication. Thus, rate 1 is achievable with d encoder-decoder.

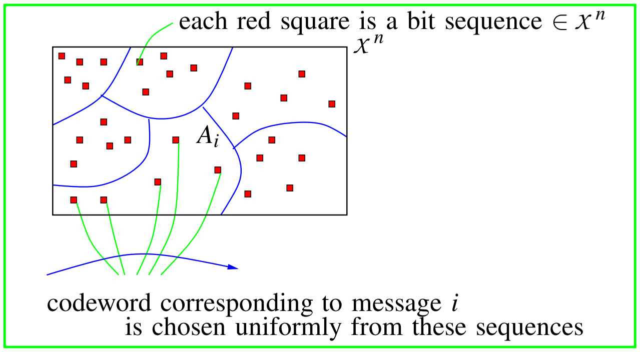
In what follows, a codeword is "killed" would mean that the output produced by the channel is the all zero sequence. Thus, there is no information transmission.

Suppose one tries to transmit at rate $R<1$ with d-encoder. Thus, each of the 2^{nR} messages is mapped to some bit sequence in $\mathcal{X}^n = \{0,1\}^n$. Consider the following channel: the channel "kills" each of these sequences which are codewords whereas rest of the sequences are transmitted perfectly. This channel $\in \mathcal{C}_{X,0}$. There is no information transmission. Thus, rates <1 are not achievable.

The question of capacity, thus, does not make sense for d-encoder-decoder. However, it does for sd encoder-decoder. Also, d and sd encoder-decoder are similar in the sense that encoder and decoder do not need to share any knowledge during communication.

Intuition for Theorem 4.3 (1b)

As said before, $C_{sd} \leq 1$. One can achieve rate $R \leq 1$ with sd-encoder-decoder in the following way: Divide $\mathcal{X}^n$ into 2^{nR} disjoint sets $A_i, 1 \leq i \leq 2^{nR}$ of cardinality $2^{n(1-R)}$ each. See figure.


each red square is a bit sequence $\in \mathcal{X}^n$
$\mathcal{X}^n$
A_i
codeword corresponding to message i
is chosen uniformly from these sequences

Each of the 2^{nR} messages is mapped to one of these 2^{nR} sets. Let m_i be mapped to the set A_i. When transmitting m_i, transmit $x \in \mathcal{A}_i$ with probability $2^{-n(1-R)}$. If y is received, decode it to the m_i such that $y \in \mathcal{A}_i$. It is easy to see that this results in reliable communication at rate R.

Definition 6.1 (HE encoder): The above encoder will be called HE encoder.

Think of HE as "high entropy." In some sense, a deterministic encoder has zero entropy. HE encoder has high entropy. Roughly, it is the opposite of a deterministic encoder and induces a high amount of randomness.

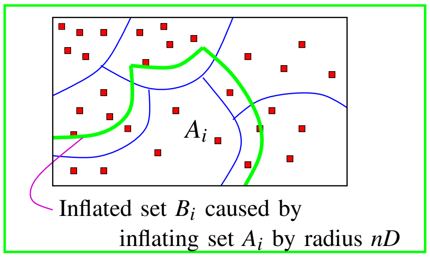
Intuition for Theorem 4.3 (2)

When $D>0$, a "bad" channel can "inflate" any set because output can be at any distance $<nD$ from the input.

For a deterministic encoder, as in Theorem 4.3 (1a), the channel can "kill" all the codewords and transmit rest of the sequences perfectly. Reliable communication will not be possible.

Now, consider the encoder HE (Definition 6.1). This encoder "uses the whole set of 2^n sequences as codewords." In some sense, it is the opposite of a deterministic encoder.

Let HE be used as the encoder. A channel in $\mathcal{C}_{X,D}$ can distort a sequence by nD. Thus, a "bad" channel will roughly, inflate each set $\mathcal{A}_i$ by radius nD. Call this inflated set $\mathcal{B}_i$. The set A_i has $2^{n(1-R)}$ elements. The set B_i will have $2^{n(1-R+\lambda)}$, $\lambda > 0$ elements (this is made rigorous in the next section). The sets $\mathcal{B}_i$ will "overlap significantly" since there are now 2^{nR} sets, each with cardinality $2^{n(1-R+\lambda)}$, and reliable communication is not possible. See figure


A_i
Inflated set B_i caused by
inflating set A_i by radius nD

The two extreme cases: d-encoder and HE-encoder provide all the intuition for proving Theorem 4.3 (2). In general, when the encoder lies somewhere "between" these two extreme encoders, a "bad" channel can be constructed such that it will "kill" some of the codewords which occur with high probability, and "inflate" others. One way of inflating a codeword by D is to "pass" the codeword through BSC(D). Another way is to "pass" it through RWC(D). There are others.

VII. Rigorous proofs

Rigorous proof for Theorem 4.3 (1): Rigorous proof is omitted because it is easy to make the intuition of the previous section precise.

Rigorous proof for Theorem 4.3, 2:

Suppose an sd encoder-decoder has been fixed for communication. We will construct a channel $\in \mathcal{C}_{X,D}$ over which reliable communication is not possible at any rate $R > 0$ for this encoder-decoder. This will prove that when $D > 0$, $C_{sd} = 0$.

The proof will consist of 2 parts in line with the intuition of the previous section:

1) use a channel which acts in a way that those codewords which occur with high probability are "killed".
2) Rest of the codewords are "inflated"to within a total distortion D by using BSC or RWC.

For block length n rate R, recall:

Message set $\mathcal{M}^n = \{m_1, m_2, \dots, m_{2^{nR}}\}$.

Channel input space $\mathcal{X}^n$. Arrange the 2^n sequences $\in \mathcal{X}^n$ in some order. Call them $x_1, x_2, \dots, x_{2^n}$.

Channel output space $\mathcal{Y}^n$. Arrange the 2^n sequences $\in \mathcal{Y}^n$ in some order. Call them $y_1, y_2, \dots, y_{2^n}$.

Let p^n_{ij}, $1 \le i \le 2^{nR}, 1 \le j \le 2^n$, denote the probability that m_i is encoded as x_j. p^n_{ij} are determined by the encoder alone and are independent of the decoder.

$$\begin{pmatrix} p^n_{11} & \cdots & p^n_{1j} & \cdots & p^n_{12^n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p^n_{i1} & \cdots & p^n_{ij} & \cdots & p^n_{i2^n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p^n_{2^{nR}1} & \cdots & p^n_{2^{nR}j} & \cdots & p^n_{2^{nR}2^n} \end{pmatrix}$$

p^n_{ij}: probability that m_i is encoded as x_j (bit sequences in codeword space are arranged in some order; it does not matter)

$$\begin{array}{ccc} | & | & | \\ \text{sum} & \text{each} & \text{column} \\ \downarrow & \downarrow & \downarrow \\ \gamma^n_1 & \gamma^n_i & \gamma^n_{2^n} \end{array}$$

Let

$$\gamma^n_j = \sum_{k=1}^{2^{nR}} p^n_{kj}, 1 \le j \le 2^n \text{ and } \alpha^n_j = \frac{\gamma^n_j}{2^{nR}} \tag{3}$$

α^n_j is the total probability that the j^{th} bit sequence x_j is used as codeword. This is because probability that message m_i is transmitted is $\frac{1}{2^{nR}}$ and the probability that sequence x_j is used as codeword given message m_i is transmitted is p^n_{ij}.

Consider a channel which "kills" a_n fraction of sequences in the input space. That is, the channel maps some $a_n 2^n$ of the 2^n possible input sequences to the all zero sequence. The values a_i will be fixed later.

As stated in the intuition, a "bad" channel will be constructed in a way that it will "kill" some of the sequences and inflate others. Intuitlvely, a "bad" channel will "kill" those bit sequences which transmit the maximum amount of information. Without loss of generality, α^n_i can be considered to be in descending order, that is, $\alpha^n_1 \ge \alpha^n_2 \ge \alpha^n_{2^n}$ (else, one can interchange and rename). With this re-ordering, probability that x_1 is used as codeword $\ge$ probability that x_2 is used as codeword, and so on. Consider a channel which "kills" $x_1, \dots, x_{a_n 2^n}$ and transmits rest of the sequences perfectly. This channel "kills" those a_n fraction of bit sequences which have the maximum probability of being codewords.

The only sequences which possibly transmit information are $x_{a_n 2^n + 1}, x_{a_n 2^n + 2}, \dots x_{2^n}$.

Define

$$\beta_n = \alpha^n_1 + \alpha^n_2 + \dots \alpha^n_{a_n 2^n} \tag{4}$$

β_n should be thought of as the "wasted probability": it is the total probability of those codewords which lead to no information transmission.

Since α_is are in descending order,

$$\alpha^n_k \le \frac{\beta_n}{a_n 2^n} \text{ for } k > a_n 2^n \tag{5}$$

Now, we fix a_n in such a way that it will be convenient for us to construct a channel over which reliable communication is not possible.

$$\text{Let } a_n = \frac{1}{n}. \text{ Note that } a_n \to 0 \text{ as } n \to \infty \tag{6}$$

Thus,

$$\alpha^n_k \le \frac{n\beta_n}{2^n} \text{ for } k > \frac{2^n}{n} \tag{7}$$

β_n, being a probability of an event, is < 1. It follows that

$$\boxed{\alpha^n_k \le n2^{-n} \text{ for } k > \frac{2^n}{n}} \tag{8}$$

It follows that those codewords which transmit useful information, each occurs with a probability $\le n2^{-n}$. $n2^{-n} \doteq 2^{-n}$. This encoder uses atleast $\frac{1}{n}$ fraction of $\mathcal{X}^n$ as codewords. $\frac{1}{n} \doteq 1$. This construction is in line with the intuition that a "good" encoder should use a significant fraction of sequences as codewords.

Consider the following channel:

Definition 7.1 (Modified BSC, mBSC):

1) Any sequence which is used as codeword with probability $> n2^{-n}$ is "killed"
These sequences will be called Type 1 sequences.
2) The channel acts as BSC$(D-\delta)$ on the rest of the sequences, $\delta > 0$.
These sequences will be called Type 2 sequences.

mBSC $\in \mathcal{C}_{X,D}$. mBSC depends on the encoder used for encoding. It does not depend on the decoder. mBSC is a modification of a BSC in that it "kills" some codewords and acts as BSC on others. So, it is called modified BSC, or mBSC. mBSC has been defined rigorously. mBSC will be the channel for which we will prove that reliable communication is not possible at any rate $R > 0$ (recall, $D > 0$).

Note 7.2: One can also define a modified random walk channel, mRWC, analogously to mBSC. mRWC "kills" every sequence which is used as a codeword with probability $> n2^{-n}$, and acts as RWC on all other codewords others.

Recall that the probability that a bit sequence x_j is used as codeword is α^n_j (Equation VII).

Let x_j be a Type 2 sequence. Then, $\alpha^n_j \le n2^{-n}$. Thus, $\gamma^n_j \le n2^{-n(1-R)}$. Thus,

$$p^n_{ij} \le n2^{-n(1-R)} \forall i \qquad (9)$$

Let rate R be reliably achievable over mBSC. By the standard information theoretic argument of going from average block error criterion to maximal block error criterion by throwing throwing away half the messages, it follows that

$$\exists \varepsilon_n \to 0 \text{ such that } \Pr(\text{error} \mid m_i \text{ is transmitted}) \le \varepsilon_n \qquad (10)$$
$$\text{for atleast } \frac{2^{nR}}{2} \text{ of the messages } m_i$$

Denote this subset of $\mathcal{M}^n$ by$\mathcal{M}^n_{\text{good}}$.

p^n_{ij}, $1 \le j \le 2^n$, i fixed, is a probability distribution on $\mathcal{X}^n$ By (9) and (10), it follows that for $m_i \in \mathcal{M}^n_{\text{good}}$, this probability distribution is such that atleast $(1-\varepsilon_n)$ of the probability is made up of individual probabilities, each of which is $\le n2^{-n(1-R)} \doteq 2^{-n(1-R)}$. Precisely, $\exists \mathcal{K} \subset \mathcal{X}^n$ such that $p^n_{ik} \le n2^{-n(1-R)}$ forall $k \in \mathcal{K}$, and $\sum_{k \in \mathcal{K}} p^n_{ik} \ge 1-\varepsilon_n$

Roughly, this says that a potentially "good" stochastic encoder encodes a message stochastically to atleast $\frac{1}{n}2^{n(1-R)} \doteq 2^{n(1-R)}$ sequences.

To recap, we have defined a channel mBSC. A necessary condition for reliable communication to be possible over mBSC at rate $R > 0$ is that atleast half of the 2^{nR} messages m_i, p^n_{ij}, $1 \le j \le 2^n$, is a probability distribution on $\mathcal{X}^n$ such that atleast $(1-\varepsilon_n)$ of the probability is made up of masses, each of which is $\le n2^{-n(1-R)} \doteq 2^{-n(1-R)}$.

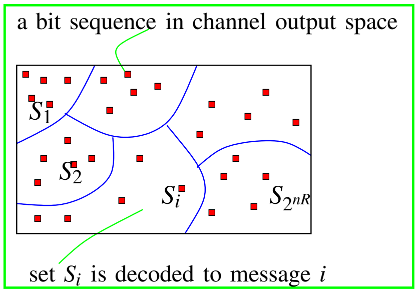
Note that the above property is saying that the encoder should act as being "close" to a HE encoder in the rigorous sense defined above for any hope of reliable communication However, the above condition is not sufficient for reliable communication over mBSC at rate R. "Inflations" described in the previous section kick in.

The rest of this section makes rigorous, the "inflations," and proves the fact that reliable communication is not possible over mBSC at any rate > 0.

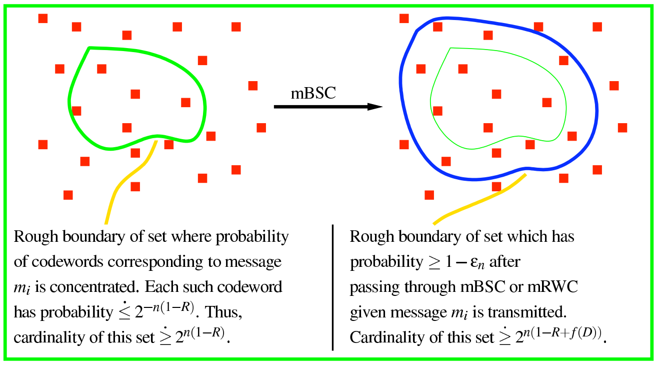
Since mBSC is one channel and not a set of channels, and the error probability criterion is average block, there exists a deterministic optimal decoder.

A deterministic decoder is a sub-division of channel output space $\mathcal{Y}^n = \{0,1\}^n$ into disjoint sets S_i such that if the channel output $\in S_i$, the estimate is m_i. See figure.

For $m_i \in \mathcal{M}_{n,\text{good}}$, $\Pr(y_i \in S_i | m_i) \ge 1-\varepsilon_n$.


a bit sequence in channel output space
S_1
S_2
S_i
$S_{2^{nR}}$
set S_i is decoded to message i

We want to make a rigorous statement saying that such a T_i should have cardinality exponentially larger than $2^{n(1-R)}$, that is, $2^{n(1-R+\lambda)}$ for some $\lambda > 0$.



Rough boundary of set where probability of codewords corresponding to message m_i is concentrated. Each such codeword has probability $\dot{\le} 2^{-n(1-R)}$. Thus, cardinality of this set $\dot{\ge} 2^{n(1-R)}$.

Rough boundary of set which has probability $\ge 1-\varepsilon_n$ after passing through mBSC or mRWC given message m_i is transmitted. Cardinality of this set $\dot{\ge} 2^{n(1-R+f(D))}$.

This is done using Theorem 8.3. in Section VIII. From this lemma, it will follow that sets T_i should have cardinality $> 2^{n(1-R)+\lambda}$ for suffiently large n for all i such that $m_i \in \mathcal{M}^n_{\text{good}}$, for some $\lambda > 0$. . This would imply that the number of sequences in the output space $\{0,1\}^n \ge \frac{2^{nR}}{2} \times 2^{n(1-R)+\lambda} > 2^n$, which is not true. The conclusion is that reliable communication is not possible for any rate $R > 0$ over mBSC.

This proves Theorem 4.3 (2).

The above argument could also have been carried out using the mRWC in place of the mBSC. The key technical fact needed in this case is Theorem 8.5, which shows that mRWC is also "inflating".

VIII. Inflating channels

A. The Bonami-Beckner inequality and other preliminaries

For this section, we will represent $\{0,1\}$ by $\mathbb{Z}_2$, the additive group on two elements. For $f : \mathbb{Z}_2^n \to \mathbb{R}$, we define the norm

$$\|f\|_p = \left(\frac{1}{2^n} \sum_{x \in \mathbb{Z}_2^n} |f(x)|^p \right)^{1/p}.$$

For $\varepsilon \in [0,1]$, the *Bonami-Beckner* operator, T_ε, acts on functions from $\mathbb{Z}_2^n$ to $\mathbb{R}$ as follows:

$$\mathrm{T}_\varepsilon(f)(x) = \sum_{y \in \mathbb{Z}_2^n} \left(\frac{1+\varepsilon}{2} \right)^{n-\mathrm{wt}(y)} \left(\frac{1-\varepsilon}{2} \right)^{\mathrm{wt}(y)} f(x+y).$$

$wt(y)$ is the weight of y: number of 1s in the bit sequence y. Note that T_ε is a convolution operator that "smooths" out the function f. This is made precise by the fundamental Bonami-Beckner inequality.

Theorem 8.1 (Bonami-Beckner): For any $f : \mathbb{Z}_2^n \to \mathbb{R}$, and any $\varepsilon \in [0,1]$ we have

$$\|\mathrm{T}_\varepsilon(f)\|_2 \leq \|f\|_{1+\varepsilon^2}.$$

It is instructive to compare the statement of the Bonami-Beckner inequality with the trivial observation that $\|\mathrm{T}_\varepsilon(f)\|_2 \leq \|f\|_2$ (note that $\|f\|_{1+\varepsilon^2} \leq \|f\|_2$ always, and $\|f\|_{1+\varepsilon^2}$ can be significantly smaller than $\|f\|_2$ in general).

a) The Fourier Transform on $\mathbb{Z}_2^n$.: We now introduce some basics of Fourier analysis on $\mathbb{Z}_2^n$. For $\xi \in \mathbb{Z}_2^n$, define the function $\chi_\xi : \mathbb{Z}_2^n \to \mathbb{R}$ by

$$\chi_\xi(x) := (-1)^{\sum_{i=1}^n \xi_i x_i}.$$

The functions χ_ξ are called the *characters* of $\mathbb{F}_2^n$.

We use this to define $\hat{f} : \mathbb{Z}_2^n \to \mathbb{R}$, the *Fourier transform* of f, by:

$$\hat{f}(\xi) = \frac{1}{2^n} \sum_{x \in \mathbb{Z}_2^n} f(x) \chi_\xi(x).$$

The Fourier inversion formula states that

$$f(x) = \sum_{\xi \in \mathbb{Z}_2^n} \hat{f}(\xi) \chi_\xi(x).$$

For future reference, we note that for any function f supported only on vectors of even weight, and for any $\xi \in \mathbb{Z}_2^n$, we have $\hat{f}(\xi) = \hat{f}(\bar{\xi})$, where $\bar{\xi}$ denotes the vector $\xi + (1,1,\ldots,1)$. Similarly for any function f supported only on vectors of odd weight, and for any $\xi \in \mathbb{Z}_2^n$, we have $\hat{f}(\xi) = -\hat{f}(\xi)$.

We have the basic Plancherel identity for any two functions $f, g : \mathbb{Z}_2^n \to \mathbb{R}$:

$$\frac{1}{2^n} \sum_{x \in \mathbb{Z}_2^n} f(x) g(x) = \sum_{\xi \in \mathbb{Z}_2^n} \hat{f}(\xi) \hat{g}(\xi)$$

As a special case, we get the Parseval equality:

$$\|f\|_2 = \sum_{\xi \in \mathbb{Z}_2^n} |\hat{f}(\xi)|^2.$$

The action of the Bonami-Beckner operator, T_ε also has a simple expression in the Fourier basis. It acts as a Fourier multiplier as follows:

$$\mathrm{T}_\varepsilon(f)(x) = \sum_{\xi \in \mathbb{Z}_2^n} \varepsilon^{\mathrm{wt}(\xi)} \hat{f}(\xi) \chi_\xi(x).$$

Finally, let $A \in \mathbb{R}^{2^n \times 2^n}$ be the transition probability matrix for the random walk on the hypercube, i.e., for $x, y \in \mathbb{Z}_2^n$, $A_{x,y} = 1/n$ if x and y differ in exactly one coordinate, and $A_{x,y} = 0$ otherwise. It can be checked that for any $f : \mathbb{Z}_2^n \to \mathbb{R}$, A also acts as a Fourier multiplier, as follows

$$Af(x) = \sum_{\xi \in \mathbb{Z}_2^n} \left(1 - 2\frac{\mathrm{wt}(\xi)}{n} \right) \hat{f}(\xi) \chi_\xi(x).$$

B. BSC is inflating

The following simple proposition relates the BSC to the Bonami-Beckner operator.

Proposition 8.2: Let μ be a probability distribution on input space $\mathbb{Z}_2^n$. Then, the distribution on the output space $\mathbb{Z}_2^n$ after passing through BSC is $T_{1-2D}(\mu)$.

Via the above proposition, the following theorem now shows that the BSC is inflating.

Theorem 8.3 (BSC is inflating): Let $D \in (0,1)$. Let μ be a probability measure on $\{0,1\}^n$ with $\mu(x) \leq 2^{-\alpha n}$ for all $x \in \{0,1\}^n$, and let $\nu = \mathrm{T}_{1-2D}(\mu)$. Then there exists a constant $\lambda_D \in (0,1)$, depending only on D, such that for any $S \subseteq \{0,1\}^n$ with $\sum_{x \in S} \nu(x) \geq \frac{1}{2}$, we have

$$|S| \geq \frac{1}{4} 2^{n(\alpha\lambda_D + (1-\lambda_D))}.$$

Proof: Let $\mathbf{1}_S : \{0,1\}^n \to \mathbb{R}$ be the indicator function of S. Let $\varepsilon = 1 - 2D$. We know that $\sum_{x \in \{0,1\}^n} \nu(x) \mathbf{1}_S(x) \geq \frac{1}{2}$.

$$\begin{aligned}
\frac{1}{2} &\leq^{*^1} \left(\sum_{x \in \{0,1\}^n} |\nu(x)|^2 \right)^{\frac{1}{2}} \left(\sum_{x \in \{0,1\}^n} |\mathbf{1}_S(x)|^2 \right)^{\frac{1}{2}} \\
&= 2^n \cdot \|\nu\|_2 \cdot \|\mathbf{1}_S\|_2 \\
&= 2^n \cdot \|\mathrm{T}_\varepsilon \mu\|_2 \cdot \|\mathbf{1}_S\|_2 \\
&\leq^{*^2} 2^n \cdot \|\mu\|_{1+\varepsilon^2} \cdot \left(\frac{|S|}{2^n} \right)^{\frac{1}{2}} \\
&\leq^{*^3} 2^n \cdot 2^{-\frac{(1+\alpha\varepsilon^2)n}{1+\varepsilon^2}} \cdot \left(\frac{|S|}{2^n} \right)^{1/2}
\end{aligned}$$

$(*^1)$ holds by Cauchy-Schwarz inequality, $(*^2)$ holds by Bonami-Beckner inequality, and $(*^3)$ holds since $\mu(x) \leq 2^{-\alpha n}$ for each x

Thus, $|S| \geq \frac{1}{4} \cdot 2^{n\left(\alpha \frac{2\varepsilon^2}{1+\varepsilon^2} + \frac{1-\varepsilon^2}{1+\varepsilon^2}\right)}$. The theorem follows. ■

C. RWC is inflating

The behaviour of the random-walk channel can be compactly described in terms of the matrix A via the following proposition.

Proposition 8.4: Let μ be a probability measure on output space $\mathbb{Z}_2^n$. Then, the distribution on the output space $\mathbb{Z}_2^n$ after passing through RWC is $A^{Dn}\mu$.

We now show that the RWC is inflating. The proof is a variation of an elegant argument due to Motwani, Naor and Panigrahy [7]. Following [7], by working in the Fourier domain, we relate the action of A to the action of the Bonami-Beckner operator, which then reduces us to the situation of Theorem 8.3.

Theorem 8.5: Let $D \in (0,1]$. Let μ be a probability measure on $\{0,1\}^n$ with $\mu(x) \leq 2^{-\alpha n}$ for all $x \in \{0,1\}^n$, and let $\nu = A^{Dn}\mu$. Then there exists a constant $\lambda_D \in (0,1)$, depending only on D, such that for any $S \subseteq \{0,1\}^n$ with $\sum_{x\in S} \nu(x) \geq \frac{1}{2}$, we have

$$|S| \geq \frac{1}{32} 2^{n(\alpha\lambda_D + (1-\lambda_D))}.$$

Proof: Let S_0 be the set of all even weight vectors in S and let S_1 be the set of odd weight vectors in S.

Let $\mathbf{1}_S : \{0,1\}^n \to \mathbb{R}$ be the indicator function of S. Similarly define $\mathbf{1}_{S_0}$ and $\mathbf{1}_{S_1}$. Note that since the support of S_0 is only on even weight vectors, $\hat{\mathbf{1}}_{S_0}(\xi) = \hat{\mathbf{1}}_{S_0}(\bar{\xi})$. Similarly, $\hat{\mathbf{1}}_{S_1}(\xi) = -\hat{\mathbf{1}}_{S_1}(\bar{\xi})$.

Let $\varepsilon = e^{-2D}$. We know that $\sum_{x\in\{0,1\}^n} \nu(x)\mathbf{1}_S(x) \geq \frac{1}{2}$. Therefore there is an $i \in \{0,1\}$ such that $\sum_{x\in\{0,1\}^n} \nu(x)\mathbf{1}_{S_i}(x) \geq \frac{1}{4}$.

$$\begin{aligned}
\frac{1}{4}\cdot\frac{1}{2^n} &\leq^{*1} \sum_{\xi\in\mathbb{Z}_2^n} \hat{\mu}(\xi)\hat{\mathbf{1}}_{S_i}(\xi)\left(1-2\frac{\mathrm{wt}(\xi)}{n}\right)^{Dn} \\
&\leq^{*2} \left(\sum_{\xi\in\mathbb{Z}_2^n} \hat{\mu}(\xi)^2\right)^{1/2} \left(\sum_{\xi\in\mathbb{Z}_2^n} \hat{\mathbf{1}}_{S_i}(\xi)^2 \left|1-2\frac{\mathrm{wt}(\xi)}{n}\right|^{2Dn}\right)^{1/2} \\
&\leq^{*^3} \|\mu\|_2 \left(\sum_{\xi\in\mathbb{Z}_2^n,\mathrm{wt}(\xi)\leq n/2} 2\hat{\mathbf{1}}_{S_i}(\xi)^2 \left|1-2\frac{\mathrm{wt}(\xi)}{n}\right|^{2Dn}\right)^{1/2} \\
&\leq^{*^4} \sqrt{2}\cdot\|\mu\|_2 \left(\sum_{\xi\in\mathbb{Z}_2^n,\mathrm{wt}(\xi)\leq n/2} \hat{\mathbf{1}}_{S_i}(\xi)^2 e^{-2\frac{\mathrm{wt}(\xi)}{n}\cdot 2Dn}\right)^{1/2} \\
&\leq \sqrt{2}\cdot\|\mu\|_2\cdot\|\mathrm{T}_\varepsilon(\mathbf{1}_{S_i})\|_2 \\
&\leq^{*^5} \sqrt{2}\cdot 2^{-n(1+\alpha)/2}\cdot\|\mathbf{1}_{S_i}\|_{1+\varepsilon^2} \\
&\leq \sqrt{2}\cdot 2^{-n(1+\alpha)/2}\cdot\left(\frac{|S_i|}{2^n}\right)^{\frac{1}{1+\varepsilon^2}}.
\end{aligned}$$

$(*^1)$ holds by Plancherel, and since $\nu = A^{Dn}\mu$, $(*^2)$ holds by Cauchy-Schwarz inequality, $(*^3)$ holds since $\left|1-2\frac{\mathrm{wt}(\xi)}{n}\right|^{2Dn} = \left|1-2\frac{\mathrm{wt}(\bar{\xi})}{n}\right|^{2Dn}$ and $\hat{\mathbf{1}}_{S_i}(\xi)^2 = \hat{\mathbf{1}}_{S_i}(\bar{\xi})^2$, $(*^4)$ holds since $e^{-x} \geq 1-x$, and $(*^5)$ holds by Bonami-Beckner inequality

Thus, $|S| \geq |S_i| \geq 2^{-5(1+\varepsilon^2)/2} \cdot 2^{n\left(\alpha\frac{1+\varepsilon^2}{2} + \frac{1-\varepsilon^2}{2}\right)}$. The theorem follows. ■

IX. Conclusion

We proved that the universal capacity of the set of channels where each channel in the set communicates i.i.d. X source to within a distortion level D, as defined rigorously in Section 4.3, when there is no common randomness at transmitter and receiver. is zero for $D > 0$. This then proves by counter-example that the universal source-channel separation theorem for rate-distortion as described in [4] is false when there is no common randomness at transmitter and receiver.

References

[1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record,* vol. 7, no. 4, pp. 142-163, 1959

[2] S. Verdu and T. S. Han, "A general formula for channel capacity," *IEEE Tran. Info. Th.*, vol. 40, no. 4, pp. 1147-1157, July 1994

[3] M. Agarwal, A. Sahai, S. Mitter, "Coding into a source: A direct inverse rate-distortion theorem" *in proceedings 44th Allerton conference*

[4] M. Agarwal, A. Sahai, S. Mitter, "A universal source-channel separation theorem and connections between source and channel coding,"*submitted, ITW 2010*, web.mit.edu/magar/www.

[5] W. Beckner. Inequalities in Fourier Analysis. *Annals of Mathematics* 102(1975).

[6] J. Kahn, G. Kalai, N. Linial. The Influence of Variables on Boolean Functions. *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*, 1988.

[7] R. Motwani, A. Naor, R. Panigrahy. Lower Bounds on Locality Sensitive Hashing. *SIAM J. Discrete Math*, 21(4), pp. 930-935, 2007.

[8] I Csiszar and P Narayan, The capacity of arbitrarily varying channels revisited: Positivity Constraints, *IEEE Trans. Info. Th.* vol 34, pp. 181-193, March 1988