CHAPTER 8

Channel coding: capacity, random coding exponents

8.1. Channel coding problem

In the channel coding problem, we are interested in transmitting an element of set of messages $\{1, \ldots, M\}$, over a possibly unfaithful medium called a channel. Concrete examples of channels may be provided by Hertzian channels that convey mobile phones conversations, optic fibers carrying cableTV, cooper lines carrying telephone conversations or IP traffic through ADSL techniques. As far as we are concerned, channels will be modelled as transition probabilities.

A channel is defined by an input alphabet $\mathcal{X}$ and an output alphabet $\mathcal{Y}$, and a conditional distribution $\mathbb{Q}_{Y|X}$, where $\mathbb{Q}_{Y|X}\{y \mid x\}$ denotes the probability of receiving $y \in \mathcal{Y}$ when $x \in \mathcal{X}$ has been transmitted. In this lecture, $\mathbb{Q}_{Y|X}$ will denote a channel. We will assume that input and output alphabets are finite.

Channels are not used in a single shot way. A sequence of input symbols x_1^n is transmitted and a sequence of output symbols y_1^n is received (one output symbol is received for each input symbol). For the sake of simplicity, we will exclusively deal with memoryless channels. The probability of receiving y_1^n when x_1^n has been transmitted is given by

$$\prod_{i=1}^{n} \mathbb{Q}_{Y|X}\{y_i \mid x_i\} \,.$$

Examples: The Binary Symmetric Channel (BSC) and Binary Erasure Channel (BEC) provide two simple and useful examples of channels. The BSC with parameter $p \in [0, 1]$ has input and output alphabets equal to $\{0, 1\}$. The transition probability is defined by:

$$\mathbb{Q}_{Y|X}\{Y = x \mid x\} = 1 - p \,.$$

When facing a BSC, the receiver does not know with certainty which symbol was sent.

The binary erasure channel has output alphabet $\{0, 1, e\}$ and transition probability defined by

$$\mathbb{Q}_{Y|X}\{Y = x \mid x\} = 1 - p \quad \text{and} \quad \mathbb{Q}_{Y|X}\{Y = e \mid x\} = p \quad \text{for } x \in \{0, 1\}.$$

When facing a BEC, if the receiver receives either 0 or 1, he knows for certain that this was the transmitted symbol.

The channel coding problem consists in enabling reliable transmission of messages over unreliable channels. In order to enable reliable information transmission, we will use forward-error-correcting codes (FEC). FEC add redundancy to the message so as to help the receiver

DEFINITION 8.1.1. [CHANNEL CODE] A channel code with block-length n is a pair of mappings (f, ϕ) where the coder f maps the set of messages $\{1, \ldots, M\}$ toward $\mathcal{X}^n$ and the decoder ϕ maps $\mathcal{Y}^n$ toward $\{1, \ldots, M\}$. The rate R of the code is defined as

$$\frac{1}{n} \log_{|\mathcal{X}|} M = \mathsf{R}.$$

The set of sequences $f(\{1, \ldots, M\})$ is often called the codebook or even the code.

The rate R of a FEC code is smaller than 1, it may be considered as the fraction of information carrying symbols among transmitted symbols.

8.2. Channel codes and Errors

In order to assess the transmission capabilities of channels, we will deal with two criteria that capture the error-correcting capabilities of FEC codes. The first one will prove useful when proving positive results.

DEFINITION 8.2.1. [AVERAGE ERROR OF A CHANNEL CODE] Let (f, ϕ) denote a forward error correcting code with block-length n, the average of (f, ϕ) over channel $\mathbb{Q}_{Y|X}$ is equal to

$$\sum_{\omega \in \{1,\ldots,M\}} \frac{1}{M} \otimes_{i=1}^{n} \mathbb{Q}_{Y|X} \left\{ \omega \neq \phi(Y_1^m) \mid x_1^n = f(\omega) \right\} .$$

But communication engineers cannot be happy with a FEC code that has only low average error. The communication engineer has no control on the messages that are actually transmitted through the channel. From a realistic viewpoint, what matters is the following criterion.

DEFINITION 8.2.2. [MAXIMAL ERROR OF A CHANNEL CODE] Let (f, ϕ) denote a forward error correcting code with block-length n, the maximal error of (f, ϕ) over channel $\mathbb{Q}_{Y|X}$ is equal to

$$\max_{\omega \in \{1,\ldots,M\}} \otimes_{i=1}^{n} \mathbb{Q}_{Y|X} \left\{ \omega \neq \phi(Y_1^m) \mid x_1^n = f(\omega) \right\} .$$

The main goal of channel coding is to design codes with high rates and small maximal error for given channels.

DEFINITION 8.2.3. [ACHIEVABLE RATE OVER A CHANNEL] Rate R is achievable over channel $\mathbb{Q}_{Y|X}$if and only if there exists a sequence (f_n, ϕ_n) of FEC codes, each with block-length n, limiting rate R and vanishing maximal error. The supremum over all achievable rates defines the limiting reliable transmission ratio of the channel.

The limiting reliable transmission ratio over a channel is defined in an operational way. As usual in Information Theory, we will first attempt to characterize this operationally defined quantity as the solution of an optimization problem. The quantity that will arise from this endeavor is the capacity of the channel. The rest of this Lecture is dedicated to the proof of the Noisy Channel Coding Theorem.

DEFINITION 8.2.4. [CAPACITY OF A MEMORYLESS CHANNEL] Let a memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ be defined by the transition probability $\mathbb{Q}_{Y|X}$, then the capacity of the channel is defined as

$$C \stackrel{\Delta}{=} \sup_{\mathbb{Q}_X \in \mathfrak{M}_1(\mathcal{X})} I\left(\mathbb{Q}_X; \mathbb{Q}_{Y|X}\right) = \sup_{\mathbb{Q}_X \in \mathfrak{M}_1(\mathcal{X})} \mathbb{E}_{\mathbb{Q}_X}\left[D\left(\mathbb{Q}_{Y|X} \| \mathbb{E}_{\mathbb{Q}_X}\left[\mathbb{Q}_{Y|X}\right]\right)\right] .$$

Recall that, as a function on the convex and compact set$\mathfrak{M}_1(\mathcal{X})$, the functional $I\left(\cdot; \mathbb{Q}_{Y|X}\right)$ is concave and continuous. Hence the supremum in the definition of C is achieved by some distribution (sometimes called the capacity-achieving distribution, see Lemma 8.2.6 for a characterization).

In order to enable full comparison with the rate/distortion function of a source, we may define the capacity under input constraints. Let ρ denote a function that maps $\mathcal{X}$ on the reals.

DEFINITION 8.2.5. [CAPACITY UNDER INPUT CONSTRAINTS] Let a memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ be defined by the transition probability $\mathbb{Q}_{Y|X}$, then the capacity of the channel under ρ-constraint B is defined as

$$\begin{aligned} C(B) &\stackrel{\Delta}{=} \sup_{\mathbb{Q}_X \in \mathfrak{M}_1(\mathcal{X}), \mathbb{E}_{\mathbb{Q}_X}[\rho(X)] \leq B} I\left(\mathbb{Q}_X; \mathbb{Q}_{Y|X}\right) \\ &= \sup_{\mathbb{Q}_X \in \mathfrak{M}_1(\mathcal{X}), \mathbb{E}_{\mathbb{Q}_X}[\rho(X)] \leq B} \mathbb{E}_{\mathbb{Q}_X}\left[D\left(\mathbb{Q}_{Y|X} \| \mathbb{E}_{\mathbb{Q}_X}\left[\mathbb{Q}_{Y|X}\right]\right)\right] . \end{aligned}$$

The following Lemma characterizes the capacity achieving input distribution.

LEMMA 8.2.6. [CAPACITY ACHIEVING INPUT DISTRIBUTIONS] *The distribution $\mathbb{Q}_X$ achieves capacity C over the channel $\mathbb{Q}_{Y|X}$ if and only if:*

$$\forall x \in \mathcal{X}: \quad D\left(\mathbb{Q}_{Y|X=x} \| \mathbb{E}_{\mathbb{Q}_X}\left[\mathbb{Q}_{Y|x}\right]\right) \begin{cases} = C & \text{if } \mathbb{Q}_X\{x\} > 0 \\ \leq C & \text{if } \mathbb{Q}_X\{x\} = 0 . \end{cases}$$

This Lemma asserts that under the capacity achieving input distributions, all conditional output distributions are at the same "Information distance" of the marginal output distributions. This is why the capacity is sometimes called the *information radius* of the channel.

The proof of the Lemma relies on some of the arguments that led to the algorithm for computing the rate-distortion function of a memoryless source. As a matter of fact, the channel capacity can also be characterized as a saddlepoint.

PROOF. As

$$I(\mathbb{Q}_X; \mathbb{Q}_{Y|X}) = \inf_{Q_Y \in \mathfrak{M}_1(\mathcal{Y})} \mathbb{E}_{\mathbb{Q}_X} \left[D\left(\mathbb{Q}_{Y|X} \mid Q_Y \right) \right]$$

we get (for example from Sion Minmax Theorem) the following saddlepoint characterization

$$C = \sup_{Q_X \in \mathfrak{M}_1(\mathcal{X})} \inf_{Q_Y \in \mathfrak{M}_1(\mathcal{Y})} \mathbb{E}_{\mathbb{Q}_X} \left[D\left(\mathbb{Q}_{Y|X} \mid Q_Y \right) \right] = \inf_{Q_Y \in \mathfrak{M}_1(\mathcal{Y})} \max_{Q_X \in \mathfrak{M}_1(\mathcal{X})} \mathbb{E}_{\mathbb{Q}_X} \left[D\left(\mathbb{Q}_{Y|X} \mid Q_Y \right) \right] .$$

Moreover, as the sets $\mathfrak{M}_1(\mathcal{X})$ and $\mathfrak{M}_1(\mathcal{Y})$ are compact (for all usual topologies) the infimum is attained for some probability $Q_Y^* \in \mathfrak{M}_1(\mathcal{Y})$. Then

$$\mathbb{E}_{\mathbb{Q}_X} \left[D\left(\mathbb{Q}_{Y|X} \mid Q_Y^* \right) \right]$$

is maximized by choosing $\mathbb{Q}_X$ in such a way that $\mathbb{Q}_X$ puts all its weight on those $x \in \mathcal{X}$ that maximize $D\left(\mathbb{Q}_{Y|X}(\cdot \mid X = x) \mid Q_Y^* \right)$. The latter (random) quantity is constant (and equal to C) on the support set of the optimal input distribution $\mathbb{Q}_X^*$. □

8.3. Channel coding theorem: Weak converse

As usual, the channel coding theorem is made of two parts: a direct part that asserts that provided the code rate does not exceed the channel capacity and block-length is sufficient, arbitrarily low decoding error can be achieved; and a converse part asserting that, if the code rate exceeds the channel capacity, whatever the block-length, arbitrarily low decoding probability cannot be achieved. As a matter of fact, there are two kinds of converses, a weak converse which we have just mentioned, and a strong converse which is much more subtle. The strong converse asserts that if code rate exceeds channel capacity, then as block-length increases, decoding error probability converges to 1.

We will first establish the weak converse. The latter is interesting per se even though it is weak, moreover its proof relies on a technical Lemma known as Fano's inequality. This Lemma has proved to be a very simple and valuable tool when dealing with multiple hypothesis problems and more generally when trying to prove negative results in statistical inference. For example, it can be used to

prove lower bounds on redundancy in universal source coding. Here follows a simple version of this Lemma.

LEMMA 8.3.1. [FANO INEQUALITY] *Let X and Y denote two random variables with finite support set $\mathcal{X}$, then we have*

$$H(X \mid Y) \leq \mathbb{P}\{X \neq Y\} \log(|\mathcal{X}| - 1)| + h\left(\mathbb{P}\{X \neq Y\}\right)$$

where $h(x) = -x \log x - (1-x)\log(1-x)$.

PROOF. Let Z denote the boolean random variable defined by $Z = \mathbb{1}_{X \neq Y}$. As Z is a function from X and Y:

$$\begin{aligned} H(X \mid Y) &= H(X, Z \mid Y) \\ &= H(Z \mid Y) + H(X \mid Y, Z) \\ &\leq H(Z) + \mathbb{P}\{Z = 0\} \times H(X \mid Y, Z = 0) + \mathbb{P}\{Z = 1\} \times H(X \mid Y, Z = 1) \\ &\leq H(Z) + \mathbb{P}\{X \neq Y\} \times \log(|\mathcal{X}| - 1), \end{aligned}$$

where the second equation comes from the chain rule for entropy, the first inequality from the fact that conditioning may only decrease entropy, the second inequality comes from the fact that conditionally on $Z = 1$, the conditional entropy of X with respect to Y is less than $\log(|\mathcal{X}| - 1)$ □

LEMMA 8.3.2. [Data-processing Lemma] *If X, Y, and Z are three random variables with joint distribution P such that $P_{Z|X,Y} = P_{Z|Y}$ then*

$$I(X; Z) \leq I(X; Y).$$

PROOF. Let X, Y, and Z be three random variables satisfying the conditions of the Lemma. We will first prove that

$$P_{X|Y,Z} = P_{X|Y}.$$

For any tuple x, y, z, such that $P\{x, y, z\} = P\{X = x, Y = y, Z = z\} > 0$, we have

$$\begin{aligned} P_{X|Y,Z}\{x \mid y, z\} &= \frac{P\{x, y, z\}}{P\{y, z\}} \\ &= \frac{P_{X,Y}\{x, y\}\, P_{Z|X,Y}\{z \mid x, y\}}{P_Y\{y\}\, P_{Z|Y}\{z \mid y\}} \\ &= \frac{P_{X,Y}\{x, y\}}{P_Y\{y\}} \\ &= P_{X|Y}\{x \mid y\}. \end{aligned}$$

The proof of the data-processing Lemma is now straightforward.

$$\begin{aligned} I(X;Y) &= \\ &= \mathbb{E}_{P_Y}[D(P_{X|Y} \mid P_X)] \\ &= \mathbb{E}_{P_Y}[D(P_{X|Y,Z} \mid P_X)] \\ &= \mathbb{E}_{P_{Y,Z}}[D(P_{X|Y,Z} \mid P_X)] \\ &= H(X) - H(X \mid Y,Z) \\ &\geq H(X) - H(X \mid Z) \\ &= I(X;Z)\,. \end{aligned}$$

□

THEOREM 8.3.3. [WEAK CONVERSE TO THE CHANNEL CODING THEOREM] *Let $\mathbb{Q}_{Y|X}$ denote a memoryless channel with capacity C,then for any $\epsilon > 0$, for any family (f_n, ϕ_n) of channel codes with rate $\mathsf{R} > C$, and block-length n, this family of codes cannot achieve arbitrarily low average decoding error probability.*

PROOF. Let us consider the two random variables X_1^n which is uniformly distributed over the codebook $\mathcal{C}_n$ and $\hat{X}_1^n \stackrel{\Delta}{=} f(\phi(Y_1^n))$, that is the codeword corresponding to the output of the decoder. The decoding error coincides with the probability that $X_1^n \neq \hat{X}_1^n$. From the Fano Inequality, we have (recall that e denotes the decoding error probability):

$$H(X_1^n \mid \hat{X}_1^n) \leq H(\mathsf{e}) + \mathsf{e} \log |\mathcal{C}_n| = H(\mathsf{e}) + n\,\mathsf{e}\;\mathsf{R}\,.$$

On the other hand

$$\begin{aligned} H(X_1^n \mid \hat{X}_1^n) &= H(X_1^n) - I(X_1^n; \hat{X}_1^n) \\ &= n\mathsf{R} - I(X_1^n; \hat{X}_1^n) \end{aligned}$$

while

$$\begin{aligned} I(X_1^n; \hat{X}_1^n) &\leq I(X_1^n; Y_1^n) \\ &= H(Y_1^n) - H(Y_1^n \mid X_1^n) \\ &= \sum_i H(Y_i \mid Y_1^{i-1}) - H(Y_i \mid X_1^n Y_1^{i-1}) \\ &\leq \sum_i H(Y_i) - H(Y_i \mid X_i^n Y_1^{i-1}) \\ &= \sum_i H(Y_i) - H(Y_i \mid X_i) \\ &= \sum_i I(X_i; Y_i) \\ &\leq nC\,, \end{aligned}$$

where the first inequality comes from the data-processing Lemma, the first equation matches the definition of mutual entropy, the second equation comes from the chain rule, the second inequality comes from the fact that conditioning does not increase entropy, the third equality from the fact that conditionally on X_i, Y_i is independent from X_1^i, X_{i+1}^n and Y_1^{i-1}.

Combining the three preceding results we get

$$n(\mathsf{R} - C) \leq H(\mathsf{e}) + n\mathsf{e}\mathsf{R},$$

that is

$$1 - \frac{C}{\mathsf{R}} \leq \mathsf{e} + \frac{H(\mathsf{e})}{n}\,.$$

If $\mathsf{R} > C$, then the decoding error probability e has to remain bounded away from 0. □

Exercise 8.4. [The Rissanen lower bound in universal coding] Recall the universal coding problem from Lecture II. Let Θ denote a set of probability distributions on $\mathcal{X}^n$. Assume Θ may be provided with the structure of a probability space. The maximin prior distribution $\hat{\mu}$, is the distribution over Θ which maximizes

$$\inf_{Q^n} \mathbb{E}_\mu \left[\mathbb{E}_\theta \left[\log \frac{P_\theta \{X_1^n\}}{Q^n \{X_1^n\}} \right] \right]\,.$$

Check that $\hat{\mu}$ is the distribution on parameter space that maximizes the mutual information between parameter and sample.

8.5. Channel coding Theorem: direct part

Just as for the Lossy Source Channel Coding Theorem, the proof of the direct part is non-constructive, it proceeds by a random selection argument. A distribution on the input alphabet $\mathbb{Q}_X$, a block-length n and a rate R are given and a codebook $\mathcal{C}_n$ is built by picking independently $M = |\mathcal{X}|^{n\mathsf{R}}$ words from $\mathcal{X}^n$ according to $\mathbb{Q}_X^{\otimes n}$, note that the construction does not require that $\mathbb{Q}_X$ coincides with the capacity achieving distribution. It is just optimized by using the capacity-achieving distribution.

Once the codebook is $\mathcal{C}_n$ is built, the encoding procedure is straightforward: message $m \in \{1, \ldots, M\}$ is encoded by the mth codeword in the codebook. The decoding procedure we will consider is called Maximum Likelihood decoding. For every $y_1^n \in \mathcal{Y}^n$, if the codebook is $\mathcal{C}_n$ then the decoded codeword is

$$\hat{m} \stackrel{\Delta}{=} \arg\min_{m \in \mathcal{C}_n} \otimes_{i=1}^n \mathbb{Q}_{Y|X}\{y_i \mid m_i\} .$$

Henceforth, we will use the following notion.

DEFINITION 8.5.1. [RANDOM CODING EXPONENT] The random coding exponent for coding distribution $\mathbb{Q}_X$ at rate R is defined as

$$\mathrm{RC}(\mathbb{Q}_X, \mathsf{R}) = \max_{s \in [0,1]} \left[-\log \left(\sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \mathbb{Q}_X\{x\} \left\{ \mathbb{Q}_{Y|x}\{y\} \right\}^{\frac{1}{1+s}} \right]^{1+s} \right) - s\mathsf{R} \right] .$$

LEMMA 8.5.2. *For any coding distribution $\mathbb{Q}_X$ the random coding exponent at rate R is positive if and only if*

$$I(\mathbb{Q}_X\{x\}; \mathbb{Q}_{Y|x}\{y\}) > \mathsf{R} .$$

PROOF. First note that for $s = 0$,

$$-\log \left(\sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \mathbb{Q}_X\{x\} \left\{ \mathbb{Q}_{Y|x}\{y\} \right\}^{\frac{1}{1+s}} \right]^{1+s} \right) - s\mathsf{R} = 0 ,$$

hence the random coding exponent is always non-negative. In order to prove that the random coding exponent is positive, it is enough to check when the derivative of

$$-\log \left(\sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \mathbb{Q}_X\{x\} \left\{ \mathbb{Q}_{Y|x}\{y\} \right\}^{\frac{1}{1+s}} \right]^{1+s} \right)$$

with respect to$s$ is larger than R at $s = 0$.

Let us now compute this derivative:

$$-\sum_{y\in\mathcal{Y}} \frac{\left[\sum_x \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{1+s}}\right]^{1+s}}{\sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{1+s}}\right]^{1+s}}$$
$$\left(\log\left[\sum_x \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{1+s}}\right] - \frac{1}{1+s}\frac{\sum_x \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{1+s}}\log \mathbb{Q}_{Y|x}\{y\}}{\sum_x \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{1+s}}}\right).$$

For $s=0$, this derivative turns out to be equal to

$$-\sum_{y\in\mathcal{Y}}\left[\sum_x \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}\right]$$
$$\left(\log\left[\sum_x \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}\right] - \sum_x \frac{\mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}}{\sum_x \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}}\log \mathbb{Q}_{Y|x}\{y\}\right)$$
$$=$$
$$\sum_x \mathbb{Q}_X\{x\}\sum_y \mathbb{Q}_{Y|x}\{y\}\log\frac{\mathbb{Q}_{Y|x}\{y\}}{\mathbb{Q}_Y\{y\}}$$
$$= I\left(\mathbb{Q}_X;\mathbb{Q}_{Y|x}\right).$$

Hence, if $I\left(\mathbb{Q}_X;\mathbb{Q}_{Y|x}\right) > \mathsf{R}$, the random coding exponent $\textsc{rc}(\mathbb{Q}_X,\mathsf{R})$ is positive. In order to show that the random coding exponent is positive if and only if $I\left(\mathbb{Q}_X;\mathbb{Q}_{Y|x}\right) > \mathsf{R}$, it is enough to check that

$$-\log\left(\sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{1+s}}\right]^{1+s}\right)$$

is concave with respect to $1+s$, or that the second derivative of this function is negative.

In order to prove the concavity with respect to s, it is enough to resort to Hölder inequality. Indeed let $s=\theta s_1+(1-\theta)s_2$ with $\theta\in[0,1]$ and $s_1,s_2\in[0,\infty]$.Then

$$\frac{1}{s} = \frac{\theta s_1}{s}\frac{1}{s_1} + \frac{(1-\theta)s_2}{s}\frac{1}{s_2}$$

and letting $\mu=\theta s_1/s$, we have

$$\frac{1}{s} = \mu\frac{1}{s_1} + (1-\mu)\frac{1}{s_2}.$$

Hence for each $y \in \mathcal{Y}$, by Hölder inequality

$$\begin{aligned}
&\left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s}}\right]^s \\
&= \left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{\mu}{s_1}+\frac{1-\mu}{s_2}}\right]^s \\
&\leq \left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_1}}\right]^{\mu s} \left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_2}}\right]^{(1-\mu)s} \\
&= \left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_1}}\right]^{\theta s_1} \left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_2}}\right]^{(1-\theta)s_2}.
\end{aligned}$$

Applying Hölder inequality again,

$$\begin{aligned}
&\sum_y \left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_1}}\right]^{\theta s_1} \left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_2}}\right]^{(1-\theta)s_2} \\
&\leq \left(\sum_y \left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_1}}\right]^{s_1}\right)^{\theta} \left(\sum_y \left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_2}}\right]^{s_2}\right)^{(1-\theta)}.
\end{aligned}$$

Combining those two inequalities, we get for $s = \theta s_1 + (1-\theta)s_2$ with $\theta \in [0,1]$ and $s_1, s_2 \in [0,\infty]$.

$$\begin{aligned}
-\log\left(\sum_{y\in\mathcal{Y}}\left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{1+s}}\right]^{1+s}\right) \\
\geq -\theta\log\left(\sum_y\left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_1}}\right]^{s_1}\right) \\
-(1-\theta)\log\left(\sum_y\left[\sum_{x\in\mathcal{X}} \mathbb{Q}_X\{x\}\left\{\mathbb{Q}_{Y|x}\{y\}\right\}^{\frac{1}{s_2}}\right]^{s_2}\right),
\end{aligned}$$

which is exactly what we were looking for. □

THEOREM 8.5.3. [DIRECT CHANNEL CODING THEOREM] *Let the random codebook be constructed by using distribution $\mathbb{Q}_X$ on the input alphabet, block-length n, and rate R.*
Under the maximum likelihood decoding rule the average error of a random code with rate R is upper-bounded by

$$\exp\left(-n\,\mathrm{RC}(\mathbb{Q}_X, R)\right).$$

PROOF. In this proof all logarithms are defined with respect to base $|\mathcal{X}|$,$\log b \stackrel{\Delta}{=} \log_{|\mathcal{X}|} b$.

Assume the first message is transmitted, that is the channel is fed with $m_1(1), ..., m_1(n)$, the maximum likelihood decoding rules errs if there exists a message $m_\ell(1), ..., m_\ell(n)$ with $\ell > 1$such that

$$\sum_{i=1}^{n} \log \frac{\mathbb{Q}_{Y|m_\ell(i)}\{y_i\}}{\mathbb{Q}_{Y|m_1(i)}\{y_i\}} > 0$$

that is, if message m_ℓ looks more likely than message m_1 given the channel output

$$y_1, ..., y_n \,.$$

Under the maximum likelihood decoding rule the average error of a random code is equal to the decoding error probability when the first message is transmitted:

$$\mathsf{e} = \mathbb{E}_{\mathcal{C}_n}\left[\sum_{y_1^n \in \mathcal{Y}^n} \otimes_{i=1}^n \mathbb{Q}_{Y|m_1(i)} \left\{ \max_{\ell>1} \sum_{i=1}^{n} \log \frac{\mathbb{Q}_{Y|m_\ell(i)}\{y_i\}}{\mathbb{Q}_{Y|m_1(i)}\{y_i\}} > 0 \right\}\right] .$$

We will refrain from applying a naive union bound. Now let s denote any number between 0 and 1, and λ denote any positive quantity. Then we have

$$\begin{aligned}
\mathsf{e} &= \sum_{y_1^n \in \mathcal{Y}^n} \sum_{m_1 \in \mathcal{X}^n} \mathbb{Q}_X^{\otimes n}\{m_1\} \otimes_{i=1}^n \mathbb{Q}_{Y|m_1(i)}\{y_i\} \left[\sum_{m_2,\ldots m_M} \otimes_{\ell=2}^M \mathbb{Q}_X^{\otimes n}\{m_\ell\} \mathbb{1}_{\left\{\max_{\ell>1} \frac{\otimes_{i=1}^n \mathbb{Q}_{Y|m_\ell(i)}\{y_i\}}{\otimes_{i=1}^n \mathbb{Q}_{Y|m_1(i)}\{y_i\}} > 1\right\}}\right] \\
&\leq \sum_{y_1^n \in \mathcal{Y}^n} \sum_{m_1 \in \mathcal{X}^n} \mathbb{Q}_X^{\otimes n}\{m_1\} \otimes_{i=1}^n \mathbb{Q}_{Y|m_1(i)}\{y_i\} \left[\sum_{m_2,\ldots m_M} \otimes_{\ell=2}^M \mathbb{Q}_X^{\otimes n}\{m_\ell\} \mathbb{1}_{\left\{\max_{\ell>2} \frac{\otimes_{i=1}^n \mathbb{Q}_{Y|m_\ell(i)}\{y_i\}}{\otimes_{i=1}^n \mathbb{Q}_{Y|m_1(i)}\{y_i\}} > 1\right\}}\right]^s \\
&\leq \sum_{y_1^n \in \mathcal{Y}^n} \sum_{m_1 \in \mathcal{X}^n} \mathbb{Q}_X^{\otimes n}\{m_1\} \otimes_{i=1}^n \mathbb{Q}_{Y|m_1(i)}\{y_i\} \left[\sum_{m_2,\ldots m_M} \otimes_{\ell=2}^M \mathbb{Q}_X^{\otimes n}\{m_\ell\} \max_{\ell>2} \left\{ \frac{\otimes_{i=1}^n \mathbb{Q}_{Y|m_\ell(i)}\{y_i\}}{\otimes_{i=1}^n \mathbb{Q}_{Y|m_1(i)}\{y_i\}} \right\}^\lambda\right]^s \\
&\leq M^s \sum_{y_1^n \in \mathcal{Y}^n} \sum_{m_1 \in \mathcal{X}^n} \mathbb{Q}_X^{\otimes n}\{m_1\} \left(\otimes_{i=1}^n \mathbb{Q}_{Y|m_1(i)}\{y_i\}\right)^{1-\lambda s} \left[\sum_{m_2} \mathbb{Q}_X^{\otimes n}\{m_2\} \left\{\otimes_{i=1}^n \mathbb{Q}_{Y|m_2(i)}\{y_i\}\right\}^\lambda\right]^s
\end{aligned}$$

where the equation follows by rearranging summations, the first inequality follows from the fact that $x \in [0,1]$ implies $x^s \geq x$ as $s \leq 1$, the second inequality comes from Markov inequality and the union bound, the last inequality comes from the exchangeability of codewords $m_2, \ldots m_M$.

Now taking $\lambda = \frac{1}{1+s}$ and using first the fact that m_1 and m_2 are identically distributed, and then the fact that we deal with product distributions:

$$\begin{aligned}
\mathsf{e} &\leq M^s \sum_{y_1^n \in \mathcal{Y}^n} \left[\sum_{m \in \mathcal{X}^n} \mathbb{Q}_X^{\otimes n}\{m\} \left\{ \otimes_{i=1}^n \mathbb{Q}_{Y|m(i)}\{y_i\} \right\}^{\frac{1}{1+s}} \right]^{1+s} \\
&\leq M^s \left(\sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \mathbb{Q}_X\{x\} \left\{ \mathbb{Q}_{Y|x}\{y\} \right\}^{\frac{1}{1+s}} \right]^{1+s} \right)^n .
\end{aligned}$$

In other terms:

$$\log \mathsf{e} \leq s \log M + n \log \left(\sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \mathbb{Q}_X\{x\} \left\{ \mathbb{Q}_{Y|x}\{y\} \right\}^{\frac{1}{1+s}} \right]^{1+s} \right) .$$

Define $\Lambda(s, \mathbb{Q}_X)$ by:

$$\Lambda(s, \mathbb{Q}_X) = -\log \left(\sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \mathbb{Q}_X\{x\} \left\{ \mathbb{Q}_{Y|x}\{y\} \right\}^{\frac{1}{1+s}} \right]^{1+s} \right) .$$

The last inequality translates into

$$\log \mathsf{e} \leq -n \sup_{s \in [0,1]} \left(-s\mathsf{R} + \Lambda(s, \mathbb{Q}_X) \right) = -n\mathrm{RC}(\mathbb{Q}_X, \mathsf{R}) .$$

□

8.6. Channel coding: Strong converse

The following result completes the weak converse that was derived from the Fano inequality.

THEOREM 8.6.1. [STRONG CONVERSE, WOLFOWITZ] *For any memoryless channel with alphabets $\mathcal{X}$ and $\mathcal{Y}$ and capacity C, for any family (f_n, ϕ_n) of block codes with block-length n and rate $\mathsf{R} > C$, the block-error probability tends to 1 as as n goes to infinity.*

PROOF. Let ϵ satisfy $\epsilon > \mathsf{R} - C$.

Let $\mathbb{Q}_X$ denote the input probability that achieves capacity. Then for any symbol $x \in \mathcal{X}$ we have (from Lemma 8.2.6)

$$\mathbb{E}_{\mathbb{Q}_{Y|x}} \left[\log \frac{\mathbb{Q}_{Y|x}\{Y\}}{\mathbb{E}_{\mathbb{Q}_X}[\mathbb{Q}_{Y|X}\{Y\}]} \right] \leq C .$$

We assume that all $M = |\mathcal{X}|^{n\mathsf{R}}$ messages have the same probability. $\mathcal{Y}^n$is divided into $M = |\mathcal{X}|^{n\mathsf{R}}$ decoding regions: $E_1, \ldots, E_M$ (note that those regions are pairwise disjoint). If $y_1^n \in E_r$ then $\phi(y_1^n) = r$. Let A_r denote the event:

$$\left\{ y_1^n : \sum_i \frac{1}{n} \log \frac{\mathbb{Q}_{Y|x}\{y_i|m_r\}}{\mathbb{E}_{\mathbb{Q}_X}[\mathbb{Q}_{Y|X}\{y_i\}]} > C + \epsilon \right\} .$$

Let us denote by $\mathbb{Q}_Y$ the probability distribution over $\mathcal{Y}$ defined by

$$\mathbb{Q}_Y\{y\} = \sum_{x \in \mathcal{X}} \mathbb{Q}_X\{x\}\, \mathbb{Q}_{Y|X}\{y \mid x\} = \mathbb{E}_{\mathbb{Q}_X}[\mathbb{Q}_{Y|X}\{y\}] .$$

Correct decoding occurs in two settings: in $A_r \cap E_r$ and in $A_r^c \cap E_r$when the r^{th}codeword $m_r$is transmitted.

Note first that $\cup_r A_r^c \cap E_r$has probability less than $|\mathcal{X}|^{(n(C+\epsilon-\mathsf{R}))}$.

$$\begin{aligned}
&\sum_r \frac{1}{M} \sum_{y_1,\ldots,y_n} \mathbb{Q}_{Y|X}^n \{y_1, ..., y_n \mid f_n(r)\}\ \mathbf{1}_{A_r^c \cap E_r} \\
&\leq \sum_r \frac{1}{M} \sum_{y_1,\ldots,y_n} \mathbb{Q}_Y^{\otimes n} \{y_1, ..., y_n\}\ |\mathcal{X}|^{n(C+\epsilon)}\, \mathbf{1}_{E_r} \\
&\leq |\mathcal{X}|^{n(C+\epsilon)} \frac{1}{M} \sum_{y_1,\ldots,y_n} \mathbb{Q}_Y^{\otimes n} \{y_1, ..., y_n\} \left[\sum_r \frac{1}{M}\, \mathbf{1}_{E_r}\{y_1 ... y_m\} \right] \\
&\leq |\mathcal{X}|^{n(C+\epsilon)} \frac{1}{M} \sum_{y_1,\ldots,y_n} \mathbb{Q}_Y^{\otimes n} \{y_1, ..., y_n\} \\
&= |\mathcal{X}|^{n(C+\epsilon)} \frac{1}{M} = |\mathcal{X}|^{n(C+\epsilon-\mathsf{R})} .
\end{aligned}$$

The probability of A_r given that $m_r$is transmitted is upper-bounded by the probability that a sum of n independent random variables with expectation less than C and variance bounded by some constant v is larger than $n(C+\epsilon)$. By Chebyshev inequality this is less than

$$v\epsilon^{-2}/n .$$

Hence the probability of correct decoding is upper-bounded by

$$\frac{v}{n\epsilon^2} + |\mathcal{X}|^{(n(C+\epsilon-\mathsf{R}))} .$$

This quantity goes to 0 as n goes to infinity. □

Note that we have just proved that if the rate of a family of codes exceeds the channel capacity, then for sufficiently large block-length, for at least one half of the messages the conditional probability a decoding error (conditioning on the fact that the channel is fed with codeword corresponding to those messages) exceeds one half.

8.7. Channel coding: Sphere packing exponents

Once equipped with the strong converse, it is possible to use a version of the change of measure argument to derive a lower bound on the probability of error when large blocklengths are considered.

DEFINITION 8.7.1. Given a word $x_1, ...x_n$on alphabet $\mathcal{X}$, the *type* of $x_1, ..., x_n$ is the probability P on $\mathcal{X}$ defined by

$$P\{a\} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_i=a}$$

for all $a \in \mathcal{X}$.

As the type of a word of length n is completely defined by a mapping from $\mathcal{X}$ on $\{0, ..., n\}$, the words of length n on $\mathcal{X}$ define at most $(n+1)^{|\mathcal{X}|}$ different types. Hence if we consider a codebook of size $|\mathcal{X}|^{n\mathsf{R}}$, there exist at least one type P such that not less than $(n+1)^{-|\mathcal{X}|} |\mathcal{X}|^{n\mathsf{R}}$ codewords have type P.

Given a family of codes $(f_n, \phi_n)_n$ with rate R, it is thus possible to define a family of codes $(f'_n, \phi_n)_n$ such that for each n, there exists a type P_n which is the common type of all codewords of length n, and with rate not less than $\mathsf{R} - |\mathcal{X}| \log_{|\mathcal{X}|}(n+1)/n$.

Using the compactness of the set of probabilities over $\mathcal{X}$, it is possible to extract a sub-family of codes with (asymptotic) rate R, codewords of length n sharing a common type P_n and the sequence P_n converging to some probability P.

THEOREM 8.7.2. [SPHERE-PACKING BOUND] *Assume that $\mathbb{Q}_{Y|X}$ defines a memoryless channel with capacity C, let (f_n, ϕ_n)denote a sequence of codes with blocklength n and limiting rate $\mathsf{R} < C$, and limiting codeword type P then the sequence of maximum decoding error probabilities e_n satisfies:*

$$\lim_n \frac{1}{n} \log \mathsf{e}_n \geq - \inf_{\mathbb{Q}'_{Y|X}} \mathbb{E}_P \left[D\left(\mathbb{Q}'_{Y|X} \| \mathbb{Q}_{Y|X} \right) \right] ,$$

where $\mathbb{Q}'_{Y|X}$is chosen among channels with capacity less than R.

The proof proceeds by "change of channel" arguments that are closely related to the change of measure arguments used when proving the lower bound in large deviations principles (and that already proved useful when deriving the direct part of the lossy source coding theorem).

PROOF. Let $\mathbb{Q}'_{Y|X}$ be chosen among memoryless channels with capacity strictly less than R. For each n, let $m_n \in \{1, \dots \lfloor |\mathcal{X}|^{n\mathsf{R}} \rfloor\}$. Let A_n denote the event that m is not recovered after transmission of $f_n(m_n)$.

$$\begin{aligned} \otimes_{i=1}^n \mathbb{Q}_{Y|X} \left\{ \phi_n(Y_1^n) \neq m_n \mid f_n(m_n) \right\} &= \mathbb{E}_{\otimes_{i=1}^n \mathbb{Q}'_{Y|X}} \left[\frac{\otimes_{i=1}^n \mathbb{Q}_{Y|X}}{\otimes_{i=1}^n \mathbb{Q}'_{Y|X}} \mathbb{1}_{A_n} \right] \\ &= \mathbb{E}_{\otimes_{i=1}^n \mathbb{Q}'_{Y|X}} \left[\exp\left(-\log \frac{\otimes_{i=1}^n \mathbb{Q}'_{Y|X}}{\otimes_{i=1}^n \mathbb{Q}_{Y|X}} \right) \mathbb{1}_{A_n} \right]. \end{aligned}$$

Now observe that by the strong converse to the Noisy Channel Coding Theorem, for n sufficiently large, for at least half of the codewords, under $\otimes_{i=1}^n \mathbb{Q}'_{Y|X}$, A_n is realized with probability larger than $1/2$.

On the other hand under $\otimes_{i=1}^n \mathbb{Q}'_{Y|X}$, $\log \frac{\otimes_{i=1}^n \mathbb{Q}'_{Y|X}}{\otimes_{i=1}^n \mathbb{Q}_{Y|X}}$ is a sum of independent (but not necessarily identically distributed) bounded random variables. Its fluctuations around its expectation can be handled using Bienaymée-Chebychev inequality: for n sufficiently large, with probability larger than $1 - \epsilon$,

$$\otimes_{i=1}^n \mathbb{Q}'_{Y|X} \left\{ -\frac{1}{n} \log \frac{\otimes_{i=1}^n \mathbb{Q}'_{Y|X}}{\otimes_{i=1}^n \mathbb{Q}_{Y|X}} \leq -\mathbb{E}_{\mathbb{Q}_X} \left[D\left(\mathbb{Q}'_{Y|X} \mid \mathbb{Q}_{Y|X} \right) \right] - \epsilon \right\} \leq \frac{V}{n\epsilon^2} \leq \frac{1}{4}$$

where V is a constant that depends only on $\mathbb{Q}_X, \mathbb{Q}'_{Y|X}$ and $\mathbb{Q}_{Y|X}$.

$$\otimes_{i=1}^n \mathbb{Q}_{Y|X} \left\{ \phi_n(Y_1^n) \neq m_n \mid f_n(m_n) \right\} \geq (1/4) \exp\left(-n \left(\mathbb{E}_{\mathbb{Q}_X} \left[D\left(\mathbb{Q}'_{Y|X} \mid \mathbb{Q}_{Y|X} \right) \right] + \epsilon \right) \right)$$

□