CHAPTER 6

Lossy source coding: rate-distortion theory

6.1. The lossy source coding problem. Distortions

Once a continuous-time signal has been sampled, the resulting discrete-time signal is not yet a digital entity, it is a sequence of values from some abstract alphabet (i.e. those values usually belong to some complete metric space, real numbers for example constitute an abstract alphabet). In order to store the signal on a digital computer, to send it through a digital transmission line, or to perform some transformations on the signal using a computer, it is necessary to build a discrete-valued approximation of the signal values. Such a process is called *quantization*. It consists in replacing $X_1, \ldots X_n$ by $\hat{X}_1 \ldots \hat{X}_n$ where $\hat{X}_i$ belongs to some finite set $\mathcal{Y}$. The original sequence is usually not recoverable from the quantized sequence, therefore that kind of coding is called *lossy*. Although we accept quantization to be lossy, we would like it to be faithful, i.e. we would like the quantized sequence $\hat{X}_1 \ldots \hat{X}_n$ to look like $X_1, \ldots X_n$. Looking like is certainly a very sloppy notion. It may be defined by empirical psycho-physical tests (nowadays no audio/video lossy coding technique can reach the status of standard without passing such tests). But those questions are outside the scope of a mathematically oriented course. In this chapter, we will continuously assume that there exists a single-letter distortion function ρ that maps $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}^+$. Our fidelity criterion will be:

$$\rho(x_1^n, \hat{x}_1^n) \triangleq \sum_{i=1}^{n} \rho(x_i, \hat{x}_i) \ .$$

A rate-distortion code with blocklength n is defined by a triplet (f, ψ, ϕ) where f maps $\mathcal{X}^n$ towards $\mathcal{Y}^n$, and (ψ, ϕ) defines a binary prefix code on $\mathcal{Y}^n$. The average rate of the code is defined by

$$\frac{1}{n}\mathbb{E}[\ell(\psi(f(X_1^n)))],$$

while the average distortion is defined by

$$\frac{1}{n}\mathbb{E}[\rho(X_1^n, f(X_1^n))].$$

Lossy source coding may be considered from different perspectives: one may fix the rate and optimize the distortion, or one may fix the allowed distortion and optimize the rate.

For example, rate R is achievable under distortion D if there exists a sequence of codes $\langle f_n, \psi_n, \phi_n \rangle$ with blocklength n (n tending towards ∞) with limiting rate less than R and limiting average distortion less than D.

Remark. In this chapter, the base of all logarithms is the size of the reconstruction alphabet $\mathcal{Y}$. Entropies, relative entropies, mutual informations, as well as code rates and rate-distortion functions are defined with respect to $\log_{|\mathcal{Y}|}$.

6.2. The rate-distortion function of a source

The concept of entropy rate of a stationary source proved to be the cornerstone of the theory of lossless compression. The *rate-distortion function* of a source plays a similar role in the theory of lossy compression.

Let $\mathbb{P}$ denote the joint probability of two random variables X and Y and $\mathbb{P}_X, \mathbb{P}_Y$ denote the corresponding marginals. Recall that the *mutual information* between X and Y is:

$$I\left(X;Y\right) = H(X) + H(Y) - H(X,Y) = H(X) - H(X \mid Y).$$

Definition 6.2.1. [Distortion compliant couplings] Let (X_n) denote an $\mathcal{X}$-valued process. Let $\mathbb{P}$ denote the law of (X_n). Let $\mathcal{U}_n(\mathsf{D})$ denote the family of laws $\mathbb{Q}$ of $\mathcal{X} \times \mathcal{Y}$-valued processes $(X_i, Y_i)_{i\in\mathbb{N}}$, which first marginal $(X_i)_{i\in\mathbb{N}}$ is distributed according to $\mathbb{P}$, and such that

$$\frac{1}{n}\mathbb{E}_{\mathbb{Q}}[\rho(X_1^n, Y_1^n] \leq \mathsf{D}.$$

Let us furthermore denote by $\mathsf{R}_n(\mathsf{D})$ the n^{th} order rate-distortion-function:

$$\mathsf{R}_n(\mathsf{D}) \stackrel{\Delta}{=} \inf_{\mathbb{Q}\in\mathcal{U}_n(\mathsf{D})} \frac{1}{n} I\left(X_1^n; Y_1^n\right).$$

Remark. 1) Joint processes whose marginals satisfy some conditions are sometimes called joinings or couplings.

2) If $\mathcal{X}$ and $\mathcal{Y}$ are finite, the infimum in the definition of $\mathsf{R}_n(\mathsf{D})$ is achieved.

Definition 6.2.2. [Rate-distortion function] The *rate-distortion function* of (X_n) at distortion D is denoted by $\mathsf{R}(\mathsf{D})$ and defined by:

$$\mathsf{R}(\mathsf{D}) \stackrel{\Delta}{=} \limsup_n \mathsf{R}_n(\mathsf{D}).$$

The limsup in the definition is actually a limit. This can be checked using a subadditivity argument and resorting to the Fekete Lemma.

6.3. Properties of the rate-distortion function

When considering couplings, it is useful to consider mutual information between X and Y as a functional of two arguments: this first one is the first marginal law $\mathbb{Q}_X$, the second argument is the conditional distribution of X given Y: $\mathbb{Q}_{Y|X}$. Overloading notations, we will denote the mutual information under the coupling defined by $\mathbb{Q}_X$and $\mathbb{Q}_{Y|X}$by $I(\mathbb{Q}_X;\mathbb{Q}_{Y|X})$.

Lemma 6.3.1. [Convexity of mutual information] *If $\mathbb{Q}$ denotes the joint law of X and Y, while $\mathbb{Q}_X$ (resp. $\mathbb{Q}_Y$) denotes the marginal with respect to X (resp. Y) and $\mathbb{Q}_{Y|X}$ denotes the (a.s. defined) conditional distribution of Y with respect to X. If $\mathbb{Q}_{Y|X}$ is fixed, $I(X;Y) = I(\mathbb{Q}_X;\mathbb{Q}_{Y|X})$ is concave with respect to $\mathbb{Q}_X$ while if $\mathbb{Q}_X$ remains fixed $I(\mathbb{Q}_X;\mathbb{Q}_{Y|X})$ is convex with respect to $\mathbb{Q}_{Y|X}$.*

Proof. Let us assume that $\mathbb{Q}_X$is fixed and consider two conditional distribution $\mathbb{Q}^1_{Y|X}$ and $\mathbb{Q}^2_{Y|X}$, let $\mathbb{Q}^1$ and $\mathbb{Q}^2$ denote the corresponding joint distributions. If $\lambda \in [0,1]$, then $\lambda\mathbb{Q}^1 + (1-\lambda)\mathbb{Q}^2$ is the joint distribution associated with the conditional distribution $\lambda\mathbb{Q}^1_{Y|X} + (1-\lambda)\mathbb{Q}^2_{Y|X}$.To check that

$$I\left(\mathbb{Q}_X;\lambda\mathbb{Q}^1_{Y|X} + (1-\lambda)\mathbb{Q}^2_{Y|X}\right) \leq \lambda I\left(\mathbb{Q}_X;\mathbb{Q}^1_{Y|X}\right) + (1-\lambda) I\left(\mathbb{Q}_X;(1-\lambda)\mathbb{Q}^2_{Y|X}\right),$$

mutual information may be rewritten as:

$$\begin{aligned} I\left(\mathbb{Q}_X;\mathbb{Q}_{Y|X}\right) &= D(\mathbb{Q} \,\|\, \mathbb{Q}_X \otimes \mathbb{Q}_Y) \\ &= \mathbb{E}_{\mathbb{Q}_X}\left[D(\mathbb{Q}_{Y|X} \,\|\, \mathbb{E}_{\mathbb{Q}_X}\left[\mathbb{Q}_{Y|X}\right])\right]. \end{aligned}$$

Now the convexity of relative entropy with respect to its two arguments (which is a straightforward consequence of the variational representation of entropy, or in the simplest settings of the log-sum inequality) entails that conditionally on X:

$$\begin{aligned} & D(\lambda\mathbb{Q}^1_{Y|X} + (1-\lambda)\mathbb{Q}^2_{Y|X} \,\|\, \mathbb{E}_{\mathbb{Q}_X}\left[\lambda\mathbb{Q}^1_{Y|X} + (1-\lambda)\mathbb{Q}^2_{Y|X}\right]) \\ & \qquad \leq \lambda D(\mathbb{Q}^1_{Y|X} \,\|\, \mathbb{E}_{\mathbb{Q}_X}\left[\mathbb{Q}^1_{Y|X}\right]) + (1-\lambda) D(\mathbb{Q}^2_{Y|X} \,\|\, \mathbb{E}_{\mathbb{Q}_X}\left[\mathbb{Q}^2_{Y|X}\right]) \quad . \end{aligned}$$

Now taking expectations with respect to $\mathbb{Q}_X$ completes the proof of convexity.

The concavity property of mutual information with respect to $\mathbb{Q}_X$ is a straightforward consequence of the concavity of entropy and of the decomposition

$$I\left(\mathbb{Q}_X;\mathbb{Q}_{Y|X}\right) = H\left(\mathbb{E}_{\mathbb{Q}_X}\left[\mathbb{Q}_{Y|X}\right]\right) - \mathbb{E}_{\mathbb{Q}_X}\left[H\left(\mathbb{Q}_{Y|X}\right)\right] \quad .$$

□

Lemma 6.3.2. [Sub-additivity] *For any stationary source, for all distortion levels D, for all positive integers m and n:*

$$0 \leq (n+m)\mathsf{R}_{m+n}(\mathsf{D}) \leq m\mathsf{R}_m(\mathsf{D}) + n\mathsf{R}_n(\mathsf{D}) \quad .$$

PROOF. Let δ denote a small positive real. Let $\mathbb{Q} \in \mathcal{U}_n(\mathsf{D})$ and $\mathbb{Q}' \in \mathcal{U}_m(\mathsf{D}')$ satisfy:

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}_X}\left[D(\mathbb{Q}_{Y_1^n|X_1^n} \parallel \mathbb{Q}_{Y_1^n})\right] &\leq \mathsf{R}_n(\mathsf{D}) + \delta, \\ \mathbb{E}_{\mathbb{Q}'_X}\left[D(\mathbb{Q}'_{Y_1^n|X_1^n} \parallel \mathbb{Q}'_{Y_1^n})\right] &\leq \mathsf{R}_m(\mathsf{D}) + \delta. \end{aligned}$$

Let $\mathbb{Q}''$ be defined in the following way: the first marginal of $\mathbb{Q}''$ is $\mathbb{P}$; and conditionally on $X_1^{n+m} = x_1^{n+m}$,

$$\begin{aligned} &\mathbb{Q}''_{Y|X_1^{n+m}}\left\{Y_1^{n+m} = y_1^{n+m} \mid x_1^{n+m}\right\} \\ &\triangleq \mathbb{Q}_{Y|X}\left\{Y_1^n = y_1^n \mid x_1^n\right\} \times \mathbb{Q}'_{Y|X}\left\{Y_1^m = y_{n+1}^{n+m} \mid x_{n+1}^{n+m}\right\}. \end{aligned}$$

Using the stationarity of $\mathbb{P}$ and the additivity of the distortion measure ρ :

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}''}\left[\rho(X_1^{n+m}, Y_1^{n+m}\right] &= \mathbb{E}_{\mathbb{Q}}[\rho(X_1^n, Y_1^n)] + \mathbb{E}_{\mathbb{Q}''}[\rho(X_{n+1}^{n+m}, Y_{n+1}^{n+m})] \\ &= n\mathsf{D} + \mathbb{E}_{\mathbb{Q}'}[\rho(X_1^m, Y_1^m)] \\ &= (n+m)\mathsf{D}. \end{aligned}$$

Thus $\mathbb{Q}'' \in \mathcal{U}_{n+m}(\mathsf{D})$.

Now let us upper-bound the mutual information between X_1^{n+m} and Y_1^{n+m} under $\mathbb{Q}''$.

$$\begin{aligned} I\left(X_1^{n+m}; Y_1^{n+m}\right) &= I\left(X_1^n; Y_1^n\right) \\ &\quad + \mathbb{E}_{\mathbb{Q}''}\left[\mathbb{E}_{X_{n+1}^{n+m}}[D(\mathbb{Q}_{Y_{n+1}^{n+m}|X_{n+1}^{n+m}} \parallel \mathbb{Q}''_{Y_{n+1}^{n+m}|X_1^n}) \mid X_1^n]\right] . \end{aligned}$$

Now in the second summand, the conditional probability of Y_{n+1}^{n+m} given X_{n+1}^{n+m} is fixed and defined by $\mathbb{Q}'$, while the distribution of X_{n+1}^{n+m} is $\mathbb{P}\{X_{n+1}^{n+m} \mid x_1^n\}$, it is random with expectation equal to the distribution of X_1^m under $\mathbb{P}$ (by the stationarity assumption). We are now in a position to use the concavity property of the mutual information (Lemma 6.3.1) and Jensen inequality to conclude that the second summand is less than $m\mathsf{R}_m(\mathsf{D}) + \delta$. Thus under $\mathbb{Q}''$

$$I\left(X_1^{n+m}; Y_1^{n+m}\right) \leq n\mathsf{R}_n(\mathsf{D}) + m\mathsf{R}_m(\mathsf{D}) + 2\delta.$$

As δ may be chosen arbitrarily small, the Lemma is proved. □

LEMMA 6.3.3. [CONVEXITY OF RATE-DISTORTION FUNCTION]*For a given source, the rate/distortion function $R(D)$ is non-increasing, convex and continuous with respect to D.*

We will actually prove a stronger result:

Lemma 6.3.4. [Convexity of rate-distortion function at rank n]*For a given source, for all n, the n^{th} order rate-distortion function, $\mathsf{R}_n(\cdot)$ is convex and non-increasing.*

Lemma 6.3.3 follows from Lemmas 6.3.4 and the fact that the pointwise limit of convex functions is convex.

Proof. [Proof of Lemma 6.3.4.] The only point to check is the convexity. As the pointwise limit of convex functions on $\mathbb{R}$ is convex. It is enough to prove that for any n, $\mathsf{R}_n(\mathsf{D})$ is convex with respect to D.

Let ϵ be > 0. Let D_1 and D_2 denote two distortion levels. Assume that the pair of random variables (X_1^n, U_1^n) is such that

$$I\left(X_1^n; U_1^n\right) \leq \mathsf{R}_n(\mathsf{D}_1) + \epsilon,$$
$$\mathbb{E}[\rho(X_1^n, U_1^n)] \leq n\mathsf{D}_1.$$

And assume that the pair (X_1^n, V_1^n) satisfies:

$$I\left(X_1^n; V_1^n\right) \leq \mathsf{R}_n(\mathsf{D}_2) + \epsilon,$$
$$\mathbb{E}[\rho(X_1^n, V_1^n)] \leq n\mathsf{D}_2.$$

Now take $\lambda \in [0,1]$, assume there is probabilistic space where X_1^n, U_1^n, V_1^n live together with an independent Bernoulli random variable Z that equals 1 with probability λ, and let $W_1^n = U_1^n$ when $Z = 1$ and V_1^n otherwise. Then:

$$\begin{aligned} \mathbb{E}[\rho(X_1^n, W_1^n)] &= \lambda\mathbb{E}[\rho(X_1^n, U_1^n)] + (1-\lambda)\mathbb{E}[\rho(X_1^n, V_1^n)] \\ &\leq \lambda\mathsf{D}_1 + (1-\lambda)\mathsf{D}_2. \end{aligned}$$

Note that the conditional distribution of W_1^n given X_1^n is a convex combination of the conditional distribution of U_1^n given X_1^n and of the conditional distribution of V_1^n given X_1^n. Now one can check that $I\left(X;Y\right)$ is concave with respect the distribution of X when the conditional distribution of Y given X is fixed, while $I\left(X;Y\right)$ is convex with respect to the conditional distribution of Y given X when the distribution of X is fixed. This allows us to conclude:

$$\begin{aligned} I\left(X_1^n; W_1^n\right) &\leq \lambda I(X_1^n; U_1^n) + (1-\lambda) I(X_1^n; V_1^n) \\ &\leq \lambda\mathsf{R}(\mathsf{D}_1) + (1-\lambda)\mathsf{R}(\mathsf{D}_2) + \epsilon. \end{aligned}$$

Hence as ϵ may be chosen arbitrarily small:

$$\mathsf{R}(\lambda\mathsf{D}_1 + (1-\lambda)\mathsf{D}_2) \leq \lambda\mathsf{R}(\mathsf{D}_1) + (1-\lambda)\mathsf{R}(\mathsf{D}_2).$$

□

The rate distortion-function of memoryless sources turns out to have a single-letter characterization. This is checked using Lemma 6.3.4.

LEMMA 6.3.5. [SINGLE LETTER CHARACTERIZATION] *If $(X_n)_{n\in\mathbb{N}}$ denotes a memoryless source, then for all distortion levels:*

$$R(D) = R_1(D).$$

PROOF. It is enough to prove that for any $\epsilon > 0$, for any n, $\mathsf{R}(\mathsf{D}) \leq \mathsf{R}_n(\mathsf{D}) + \epsilon$. Let $\mathbb{Q}$ denote a joint distribution on $\mathcal{X}^n \times \mathcal{Y}^n$ that belongs to $\mathcal{U}_n(\mathsf{D})$ and satisfies:

$$I\left(X_1^n; Y_1^n\right) \leq n(\mathsf{R}_n(\mathsf{D}) + \epsilon).$$

Now:

$$\begin{aligned} I\left(X_1^n; Y_1^n\right) &\overset{(a)}{=} \sum_i [H(X_i) - H(X_i \mid X_1^{i-1}, Y_1^n)] \\ &\overset{(b)}{\geq} \sum_i [H(X_i) - H(X_i \mid Y_i)] \\ &\overset{(c)}{\geq} \sum_i I\left(X_i; Y_i\right), \end{aligned}$$

where (a) comes from the independence of the X_i's, and (b) from the fact that conditioning may only decrease entropy. Now let $\mathsf{D}_i \triangleq \mathbb{E}[\rho(X_i, Y_i)]$. From the definition of $\mathsf{R}_1(\cdot)$, we have for all $i \leq n$:

$$\mathsf{R}_1(\mathsf{D}_i) \leq I\left(X_i; Y_i\right),$$

while $\sum_{i\leq n} \mathsf{D}_i \leq n\mathsf{D}$. Combining with the previous inequality and the convexity of $\mathsf{R}_1(\cdot)$:

$$\mathsf{R}_1(\mathsf{D}) \leq \mathsf{R}_1(\sum \frac{1}{n}\mathsf{D}_i) \leq \sum_i \frac{1}{n}\mathsf{R}_1(\mathsf{D}_i) \leq \mathsf{R}_n(\mathsf{D}) + \epsilon.$$

□

The preceding lemma should not be misinterpreted. It might seem that for quantizing memoryless sources, considering long blocks of symbols does not help, i.e. that quantizing symbols independently is optimal. This is not true. Even when the source statistics are known, the rate-distortion function characterizes the achievable compression ratios under some fidelity criterion in a asymptotical way. That is, the rate-redundancy per symbol is no more $O(1/n)$ as it was for lossless coding. Proving such a result is beyond the scope of those notes. Nevertheless, in the next section, it should be clear from the proof of the direct part of the lossy source coding theorem (Theorem 6.6.1), that the rate-distortion function is used in an asymptotic manner.

6.4. The rate distortion function of Bernoulli source

Let us consider the memoryless Bernoulli source with success probability p :$\mathbb{P}\left\{X_n = 1\right\} = p$ for each n. The input and output alphabets are both $\{0,1\}$ .We will use the following distortion measure:

$$\rho\left(x,y\right) = \mathbb{1}_{x\neq y}.$$

The rate-distortion function can be computed analytically on this simple example. The tricks used to simplify this computation are of independent interest.

Note first that if $\mathsf{D} \geq \min(p, 1-p)$, $\mathsf{R}(\mathsf{D}) = 0$,it is enough to take Y equal to the most frequent value of X. Henceforth we assume $\mathsf{D} < \min(\rho, 1-\rho)$.

Let us first lower-bound the rate-distortion function. Assume that X and Y are jointly distributed according to a joining with distortion D :

$$\mathbb{E}\left[\rho(X,Y)\right] \leq \mathsf{D} .$$

Let us denote by $h_2(p)$ the binary Shannon entropy of a Bernoulli Random variable. The mutual information between X and Y is equal to

$$\begin{aligned} H(X) - H(X \mid Y) &= H(X) - H(X \oplus Y \mid Y) \\ &\geq H(X) - H(X \oplus Y) \\ &\geq h_2(p) - h_2(\mathsf{D}) . \end{aligned}$$

Now let us consider the joining defined by

$$\mathbb{Q}\left\{X = Y\right\} = 1 - \mathsf{D}$$

and

$$\mathbb{Q}\left\{Y = 0\right\} = (1 - p - \mathsf{D})/(1 - 2\mathsf{D}) .$$

It satisfies the distortion constraint. And as the entropy of $X \oplus Y$ given Y does not depend on the value of Y, the inequality in the above stated derivation is actually an equality.

Remark 6.4.1. The reasoning relies on the fact that X can be recovered from the value of Y and from the distortion measure. If we were dealing with two dimensional random variables and with squared error distortion, this would not be feasible anymore.

6.5. The converse to the rate-distortion teorem

Theorem 6.5.1. [converse to the rate-distortion teorem] *Let $(X_n)_{n\in\mathbb{N}}$ be a stationary process over some complete metric space , $\mathcal{Y}$ a reconstruction alphabet, ρ a positive bounded distortion function. Let $R(D)$ denote the associated rate-distortion function, then for all distortion levels D,*
rate R is achievable under distortion D only if

$$R > R(D).$$

Proof. [Proof of converse rate-distortion teorem] Assume that rate R is achievable at distortion D on source (X_n) distributed according to $\mathbb{P}$. Then for each $\epsilon > 0$, there exists an n, and a rate distortion code $\langle f, \psi, \phi\rangle$ operating at rate less $\mathsf{R}+\epsilon$ and average distortion less than $\mathsf{D}+\epsilon$. Let us consider the random vector $(X_1^n, f_n(X_1^n))$, obviously from the definition of rate-distortion codes:

$$\frac{1}{n}\mathbb{E}_{\mathbb{P}}[\rho(X_1^n, f_n(X_1^n))] \leq \mathsf{D}+\epsilon.$$

On the other hand:

$$\begin{aligned} I(X_1^n; f_n(X_1^n)) &\overset{(a)}{\leq} H(f_n(X_1^n)) \\ &\overset{(b)}{\leq} \mathbb{E}_{\mathbb{P}}[\ell(\phi\circ\psi(X_1^n))] \\ &\overset{(c)}{\leq} n(\mathsf{R}+\epsilon) \ . \end{aligned}$$

Thus $\mathsf{R}_n(\mathsf{D}+\epsilon) \leq \mathsf{R}+\epsilon$. As ϵ may be chosen arbitrarily small, we can conclude that, $\mathsf{R}(\mathsf{D}) \leq \mathsf{R}$. □

6.6. The direct source coding theorem (memoryless sources)

The direct *source coding theorem* asserts the existence of good rate-distortion codes provided the bounds prescribed by the rate-distortion function are met. The argument leading to this theorem matured between 1948 and 1959.The toy version given here is concerned with memoryless sources on finite alphabets with bounded distortion function. Such restrictions allow to use easily the weak law of large numbers. Despite this very limited ambition, the proof of the toy direct source coding theorem outlines two fruitful ideas:

(1) Random selection procedures which prove the existence of objects satisfying a given property by checking that the property has non-null probability in some probability space. Since the inception of Information Theory in 1948, this device became a standard tool in Combinatorics under the influence of Erdös and Rényi.

(2) Change of measure arguments. Such arguments are at the core of Large Deviations lower bounds.

THEOREM 6.6.1. [DIRECT SOURCE CODING THEOREM FOR MEMORYLESS SOURCES] *Let $\mathbb{P}$ denote a memoryless source on the finite alphabet $\mathcal{X}$. Let $\mathcal{Y}$ denote a finite reconstruction alphabet, and ρa bounded distortion function on $\mathcal{X} \times \mathcal{Y}$. Let $\mathsf{R}(\mathsf{D})$ denote the rate-distortion function o f $\mathbb{P}$. for any D and R such that*

$$\mathsf{R}(\mathsf{D}) < \mathsf{R},$$

for any $\epsilon > 0$, there exists some $n(\epsilon)$ and a rate-distortion code with block-length $n(\epsilon)$, rate R and average distortion less than $\mathsf{D} + \epsilon$ for $\mathbb{P}$.

6.6.1. Proof of the direct source coding theorem. Let D^*denote an upper-bound on $\rho(\cdot,\cdot)$.

Let $\mathbb{Q}$ denote a joining such that $\mathbb{E}_{\mathbb{Q}}[\rho(X,Y)] \leq \mathsf{D}$ and

$$D(\mathbb{Q} \,||\, \mathbb{Q}_X \otimes \mathbb{Q}_Y) \leq \mathsf{R}\,(\mathsf{D}) \leq \mathsf{R} - \alpha.$$

During the proof of the source coding Theorem, we will rely on the following Lemma which combines laws of large numbers in different ways.

LEMMA 6.6.2. [TYPICALITY LEMMA] *For every $\epsilon > 0$, there exists some $n(\epsilon)$ such that for all $n > n(\epsilon)$:*

$$\begin{aligned}
\mathbb{Q}_{X_1^n}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\left(D\left(\mathbb{Q}_{Y|x_i}\|\mathbb{Q}_Y\right) - D\left(\mathbb{Q}\|\mathbb{Q}_X\otimes\mathbb{Q}_Y\right)\right)\right| > \epsilon\right\} &\leq \epsilon \\
\otimes_{i=1}^n \mathbb{Q}_{Y|x_i}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\left(\log\frac{\mathbb{Q}_{Y|x_i}\{Y_i\}}{\mathbb{Q}_Y\{Y_i\}} - D\left(\mathbb{Q}_{Y|x_i}\|\mathbb{Q}_Y\right)\right)\right| > \epsilon\right\} &\leq \epsilon \\
\mathbb{Q}_{X_1^n}\left\{\otimes_{i=1}^n \mathbb{Q}_{Y|x_i}\left\{\frac{1}{n}\rho\left(X_1^n, Y_1^n\right) > \mathsf{D} + \epsilon\right\} > \epsilon\right\} &\leq \epsilon\,.
\end{aligned}$$

PROOF. [PROOF OF LEMMA] The first inequality is just a consequence of the weak law of large numbers. The mapping:

$$x \mapsto D\left(\mathbb{Q}_{Y|x}\|\mathbb{Q}_Y\right)$$

defines a bounded random variable since $\mathcal{X}$ is finite, this random variable has expectation $D\left(\mathbb{Q}\|\mathbb{Q}_X \otimes \mathbb{Q}_Y\right).$

The second inequality is also a consequence of the weak law of large numbers. Even though $\log\frac{\mathbb{Q}_{Y|x_i}\{Y_i\}}{\mathbb{Q}_Y\{Y_i\}} - D\left(\mathbb{Q}_{Y|x_i}\|\mathbb{Q}_Y\right)$ are not identically distributed, they are independent and centered. The variance of the summands is upper-bounded by

$$v = \max_{x \in \mathcal{X}} \mathrm{Var}\left[\log \frac{\mathbb{Q}_{Y|x}\{Y\}}{\mathbb{Q}_Y\{Y\}} - D\left(\mathbb{Q}_{Y|x} \| \mathbb{Q}_Y\right)\right]$$

where $Y$is distributed according to $\mathbb{Q}_{Y|x}$.Bienaymé-Chebyshev inequality implies

$$\otimes_{i=1}^n \mathbb{Q}_{Y|x_i}\left\{\frac{1}{n}\left|\sum_{i=1}^n \left(\log \frac{\mathbb{Q}_{Y|x_i}\{Y_i\}}{\mathbb{Q}_Y\{Y_i\}} - D\left(\mathbb{Q}_{Y|x_i} \| \mathbb{Q}_Y\right)\right)\right| > \epsilon\right\} \leq \frac{v}{n\epsilon^2}$$

it is enough to take $n(\epsilon) \geq v\epsilon^{-3}$.

The third inequality is proved using arguments *ejusdem farinae.* First note that

$$\sum_{i=1}^n \mathbb{E}_{\mathbb{Q}_{Y|X_i}}\left[\rho(X_i, Y_i)\right]$$

is a sum of independent random variables with variances less than D^{*2} and mean less than D. Hence by Bienaymé-Chebyshev inequality:

$$\mathbb{Q}_{X_1^n}\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{Q}_{Y|X_i}}\left[\rho(X_i, Y_i)\right] \geq \mathsf{D} + \frac{\epsilon}{2}\right\} \leq \frac{4\mathsf{D}^{*2}}{n\epsilon^2}.$$

Now assume that $\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{Q}_{Y|x_i}}\left[\rho(x_i, Y_i)\right] < \mathsf{D} + \frac{\epsilon}{2}$, then $\rho(x_1^n, Y_1^n)$ is a sum of independent random variables with mean less than $n(\mathsf{D} + \epsilon/2)$ and variance less than $n\mathsf{D}^{*2}$, resorting again to the Bienaymé-Chebyshev inequality

$$\otimes_{i=1}^n \mathbb{Q}_{Y|x_i}\left\{\frac{1}{n}\rho\left(X_1^n, Y_1^n\right) > \mathsf{D} + \epsilon\right\} \leq \frac{4\mathsf{D}^{*2}}{n\epsilon^2}.$$

Combining those two bounds, we get

$$\mathbb{Q}_{X_1^n}\left\{\otimes_{i=1}^n \mathbb{Q}_{Y|x_i}\left\{\frac{1}{n}\rho\left(X_1^n, Y_1^n\right) > \mathsf{D} + \epsilon\right\} > \frac{4\mathsf{D}^{*2}}{n\epsilon^2}\right\} \leq \frac{4\mathsf{D}^{*2}}{n\epsilon^2}.$$

Taking $n \geq 4^2\epsilon^{-3}$, we get the third part of the Lemma. □

Lemma 6.6.3. [Change of measure Lemma] *If $x_1^n \in \mathcal{X}^n$ is ϵ-typical, then*

$$\mathbb{Q}_Y^n\left\{\rho\left(x_1^n, Y_1^n\right) \leq n(D + \epsilon)\right\} \geq (1 - 2\epsilon) \times |\mathcal{Y}|^{(-n(R(D)+\epsilon))}.$$

In the proof of this lemma, we will use the following device. Assume that $\mathbb{P}$ and $\mathbb{Q}$ are two probability distributions over the same space, and furthermore that $\mathbb{P}\{E\} > 0$ entails $\mathbb{Q}\{E\} > 0$ for any event E. Let $\frac{d\mathbb{P}}{d\mathbb{Q}}$ and $\frac{d\mathbb{Q}}{d\mathbb{P}}$ denote respectively the densities of $\mathbb{P}$ (resp. $\mathbb{Q}$) with respect to $\mathbb{Q}$ (resp. $\mathbb{P}$) (when dealing with finite probability spaces, these are simply ratios of probabilities of elementary events, in more general contexts the existence of those densities follow from the Radon-Nykodim Theorem). Let A denote any event,

$$\begin{aligned}
\mathbb{P}\{A\} &= \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_A\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}\mathbb{1}_A\right] \\
&\geq \mathbb{E}_{\mathbb{Q}}\left[\exp\left(-\log\frac{d\mathbb{Q}}{d\mathbb{P}}\right)\mathbb{1}_A\right] .
\end{aligned}$$

PROOF. [Proof of Change of Measure Lemma] Let x_1^n be ϵ-typical. Then we have:

$$\begin{aligned}
&\mathbb{Q}_Y^n\left\{\rho\left(x_1^n, Y_1^n\right) \leq n(\mathsf{D}+\epsilon)\right\} \\
&= \mathbb{E}_{\mathbb{Q}_Y^n}\left[\mathbb{1}_{\rho\left(x_1^n,Y_1^n\right)\leq n(\mathsf{D}+\epsilon)}\right] \\
&= \mathbb{E}_{\otimes_{i=1}^n \mathbb{Q}_{Y|x_i}}\left[\exp\left(-n\sum_{i=1}^n \frac{1}{n}\log\frac{\mathbb{Q}_{Y|x_i}\{Y_i\}}{\mathbb{Q}_Y\{Y_i\}}\right)\mathbb{1}_{\rho\left(x_1^n,Y_1^n\right)\leq n(\mathsf{D}+\epsilon)}\right] \\
&\geq (1-2\epsilon)\times|\mathcal{Y}|^{|\left(-\sum_{i=1}^n D\left(\mathbb{Q}_{Y|x_i}\|\mathbb{Q}_Y\right)-n\epsilon\right)} \\
&\geq (1-2\epsilon)\times|\mathcal{Y}|^{(-nD(\mathbb{Q}\|\mathbb{Q}_X\otimes\mathbb{Q}_Y)-n\epsilon)} \\
&\geq (1-2\epsilon)\times|\mathcal{Y}|^{(-n(\mathsf{R}(\mathsf{D})+\epsilon))} ,
\end{aligned}$$

where we have used the second inequality in the Typicality Lemma which asserts that with $\otimes_{i=1}^n \mathbb{Q}_{Y|x_i}$-probability larger than $1-\epsilon$,

$$-n\sum_{i=1}^n \frac{1}{n}\log\frac{\mathbb{Q}_{Y|x_i}\{Y_i\}}{\mathbb{Q}_Y\{Y_i\}} \geq -\sum_{i=1}^n D\left(\mathbb{Q}_{Y|x_i}\|\mathbb{Q}_Y\right) - n\epsilon.$$

□

Equippped with those two lemmata, we may now prove the Theorem.

PROOF. [Proof of Direct Source Coding Theorem] Let $\mathcal{C}_n$denote an arbitrary collection of $2^{n\mathsf{R}}$words from $\mathcal{Y}^n$. The set $\mathcal{C}_n$ is often called a codebook. We associate with $\mathcal{C}_n$ the following quantization procedure:

$$f(X_1^n) = \arg\min_{Y_1^n\in\mathcal{C}_n} \rho\left(X_1^n, Y_1^n\right) ,$$

we are interested in upper-bounding the average distortion of this quantization procedure:

$$\mathbb{E}_{X_1^n}\left[\rho\left(X_1^n, f(X_1^n)\right)\right] \quad .$$

Rather than trying to construct explicitly a codebook $\mathcal{C}_n$, we use a randomized selection procedure: $\mathcal{C}_n$ is built by picking independently $2^{n\mathsf{R}}$ words from

$\mathcal{Y}^n$according to $\mathbb{Q}_Y^n$. In order to prove that there exists a codebook with average distortion less than D + 2ϵ,it is enough to prove that when averaging with respect to the codebook selection procedure, the average distortion is less than 2ϵ. As codebooks and codewords are picked independently, using the Tonelli-Fubini Theorem (if $\mathcal{X}$ and $\mathcal{Y}$ are not countable) or first principles (when $\mathcal{X}$and $\mathcal{Y}$ are countable) we have:

$$\mathbb{E}_{\mathcal{C}_n}\left[\mathbb{E}_{X_1^n}\left[\min_{Y_1^n\in\mathcal{C}_n}\rho\left(X_1^n,Y_1^n\right)\right]\right] \;=\; \mathbb{E}_{X_1^n}\left[\mathbb{E}_{\mathcal{C}_n}\left[\min_{Y_1^n\in\mathcal{C}_n}\rho\left(X_1^n,Y_1^n\right)\right]\right]$$

We may now focus on a fixed x_1^n, and consider the average distortion between x_1^n and its reconstruction when the codebook $\mathcal{C}_n$ is picked at random. We have:

$$\mathbb{E}_{\mathcal{C}_n}\left[\min_{Y_1^n\in\mathcal{C}_n}\rho\left(x_1^n,Y_1^n\right)\right] \;\le\; \mathsf{D}^*\times\mathbb{P}_{\mathcal{C}_n}\left\{\min_{Y_1^n\in\mathcal{C}_n}\rho\left(x_1^n,Y_1^n\right)>\mathsf{D}+\epsilon\right\}+\mathsf{D}+\epsilon\,.$$

$$\mathbb{E}_{\mathcal{C}_n}\left[\mathbb{E}_{X_1^n}\left[\min_{Y_1^n\in\mathcal{C}_n}\rho\left(X_1^n,Y_1^n\right)\right]\right]\le\mathsf{D}^*\times\mathbb{E}_{X_1^n}\left[\mathbb{P}_{\mathcal{C}_n}\left\{\min_{Y_1^n\in\mathcal{C}_n}\rho\left(X_1^n,Y_1^n\right)>\mathsf{D}+\epsilon\right\}\right]+\mathsf{D}+\epsilon$$

In order to complete our program, we need to prove that for most words x_1^n

$$\mathbb{P}_{\mathcal{C}_n}\left\{\min_{Y_1^n\in\mathcal{C}_n}\rho\left(x_1^n,Y_1^n\right)>\mathsf{D}+\epsilon\right\}$$

is small. Note that

$$\begin{aligned}\mathbb{P}_{\mathcal{C}_n}\left\{\min_{Y_1^n\in\mathcal{C}_n}\rho\left(x_1^n,Y_1^n\right)>\mathsf{D}+\epsilon\right\} &= \left(\mathbb{Q}_Y^n\left\{\rho\left(x_1^n,Y_1^n\right)>\mathsf{D}+\epsilon\right\}\right)^{|\mathcal{Y}|^{n\mathsf{R}}}\\ &= \left(1-\mathbb{Q}_Y^n\left\{\rho\left(x_1^n,Y_1^n\right)\le\mathsf{D}+\epsilon\right\}\right)^{|\mathcal{Y}|^{n\mathsf{R}}}\\ &\le \exp\;\left(-|\mathcal{Y}|^{n\mathsf{R}}\,\mathbb{Q}_Y^n\left\{\rho\left(x_1^n,Y_1^n\right)\le\mathsf{D}+\epsilon\right\}\right)\;.\end{aligned}$$

Hence,

$$\begin{aligned}(6.6.1)\quad &\mathbb{E}_{\mathcal{C}_n}\left[\mathbb{E}_{X_1^n}\left[\min_{Y_1^n\in\mathcal{C}_n}\rho\left(X_1^n,Y_1^n\right)\right]\right]\\ &\qquad\le\mathsf{D}^*\times\mathbb{E}_{X_1^n}\left[\exp\left(-|\mathcal{Y}|^{n\mathsf{R}}\,\mathbb{Q}_Y^n\left\{\rho\left(X_1^n,Y_1^n\right)\le\mathsf{D}+\epsilon\right\}\right)\right]+\mathsf{D}+\epsilon\end{aligned}$$

and if with high probability with respect to the distribution of X_1^n,

$$|\mathcal{Y}|^{n\mathsf{R}}\,\mathbb{Q}_Y^n\left\{\rho\left(x_1^n,Y_1^n\right)\le\mathsf{D}+\epsilon\right\}$$

is large, we will be in a good shape.

But,

$$|\mathcal{Y}|^{n\mathsf{R}}\,\mathbb{Q}_Y^n\left\{\rho\left(x_1^n,Y_1^n\right)\le\mathsf{D}+\epsilon\right\} \;=\; |\mathcal{Y}|^{\left(n\left(\mathsf{R}+\frac{1}{n}\log\mathbb{Q}_Y^n\left\{\rho\left(x_1^n,Y_1^n\right)\le\mathsf{D}+\epsilon\right\}\right)\right)}\,.$$

We will now use the Typicality Lemma and the Change of Measure Lemma. If x_1^n is ϵ-typical, then

$$\mathbb{Q}_Y^n\left\{\rho\left(x_1^n,Y_1^n\right)\le\mathsf{D}+\epsilon\right\}\ge(1-2\epsilon)\times|\mathcal{Y}|^{(-n(\mathsf{R}(\mathsf{D})+\epsilon))}\,,$$

And by the Typicality Lemma, this happens with probability larger than $1-2\epsilon$.

We can use those two points to get

$$\mathbb{E}_{X_1^n}\left[\exp\left(-|\mathcal{Y}|^{n\mathsf{R}}\,\mathbb{Q}_Y^n\left\{\rho\left(X_1^n,Y_1^n\right)\leq \mathsf{D}+\epsilon\right\}\right)\right] \\ \leq 2\epsilon+|\mathcal{Y}|^{(-(1-2\epsilon)n(\mathsf{R}-\mathsf{R}(\mathsf{D})+2\epsilon))} .$$

Plugging this last bound into Inequality (6.6.1), we get

$$\mathbb{E}_{X_1^n}\left[\rho\left(X_1^n,f(X_1^n)\right)\right] \quad \leq \mathsf{D}^* \times \left(2\epsilon+\exp\left(-(1-2\epsilon)n\left(\mathsf{R}-\mathsf{R}(\mathsf{D})+2\epsilon\right)\right)\right)+\mathsf{D}+\epsilon \quad .$$

Taking ϵ so that $\epsilon(1+3\mathsf{D}^*)<\eta$ and n sufficiently large so that such that $\exp\left(-(1-2\epsilon)n\left(\mathsf{R}-\mathsf{R}(\mathsf{D})+2\epsilon\right)\right)<\epsilon$,we get that the average distortion of the random codebook with block-length n and binary rate R is less than $\mathsf{D}+\eta$. □

The proof of the Direct Source Coding Theorem should be compared with the proof of the direct part of the Lossless Source Coding Theorem. In the lossless setting, we did not only prove the existence of entropy-achieving prefix codes, we also provided computationally efficient encoding and decoding methods. Moreover, once the source statistics are known, computing the entropy was easy.

We will see later thate there are quite efficient algorithms that enable to compute the rate-distortion function of a memoryless source. On the other hand we will not be able to exhibit efficient general quantization procedures.