# 11 Rate Distortion

In the previous chapters the aim is to have perfect reconstruction for source coding and arbitrary small error probability in coding. This is the basis of the source coding theorem and the channel coding theorem. However, in practical system design this is not always the case.

For example in image compression or voice coding, it is affordable to have a certain amount of losses in the reconstruction, as long as the perceived quality is not affected. This is the idea behind lossy coding instead of lossless coding, like Huffman coding or the LZ algorithms. The gain with allowing some distortion to the original image is that the compression ratio can be made much better. In image coding or video coding, algorithms like JPEG or MPEG are typical examples.

In his 1948 paper [60] Shannon started the study on how to incorporate an allowed distortion in the theory, but it was not until his paper in 1959 [61] for it to matured. In this paper the rate-distortion function is defined and shown to bound the compression capability in the same manner as the entropy does in the lossless case.

## 11.1 Rate-distortion function

To start the study of rate-distortion it must first be determined what is meant by distortion of a source. In Figure 11.1 a model for a source coding is depicted. The source symbol is a vector of length $n$, $\boldsymbol{X} = X_1, \dots X_n$. This is encoded to a length $\ell$ vector $\boldsymbol{Y} = Y_1 \dots Y_\ell$ which is then decoded back (reconstructed) to a length $n$ vector $\hat{\boldsymbol{X}} = \hat{X}_1, \dots \hat{X}_n$. The codeword length $\ell$ is regarded as a random variable and its expected value denoted $L = E[\ell]$. This is the same model as used in the lossless case in Chapter 4. The difference is that the mapping from

$\boldsymbol{X}$ to $\hat{\boldsymbol{X}}$ includes an allowance of a miss-match, i.e. in general they will not be equal. The rate of the code is defined as

$$R = \frac{L}{n} \tag{11.1}$$

This is the transmission rate and should not be confused with the compression ratio used earlier, which is its inverse. For simplicity, assume that $\boldsymbol{Y}$ is a binary vector.

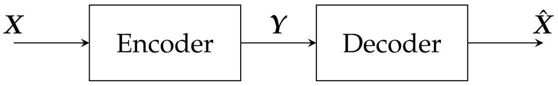

Figure 11.1: A communication model for introducing distortion.

To measure the introduced miss-match between the source symbol and the reconstructed symbol a *distortion measure* is required. It is here assumed that the distortion measure is additive, and can be written as

$$d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \sum_{i=1}^{n} d(x_i, \hat{x}_i) \tag{11.2}$$

where $d(x, \hat{x})$ is the single letter distortion. Without loss of generality it can be assumed that the minimum distortion is zero, $\min_{\hat{x}} d(x, \hat{x}) = 0$, for all $x$. There are several such measures but the two most well known are the Hamming distortion and the squared distance. The first one is typically used for discrete sources, especially for the binary case, while the second is mostly used for continuous sources.

**DEFINITION 11.1** The *Hamming distortion* between two discrete letters $x$ and $\hat{x}$ is

$$d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & x \neq \hat{x} \end{cases} \tag{11.3}$$

□

For the binary case the Hamming distortion can be written as

$$d(x, \hat{x}) = x \oplus \hat{x} \tag{11.4}$$

where $\oplus$ denotes addition modulo 2.

**DEFINITION 11.2** The *squared distance distortion* between two variables $x$ and $\hat{x}$ is

$$d(x, \hat{x}) = (x - \hat{x})^2 \tag{11.5}$$

□

In principle, all vector norms can be used as distortion measures, but for example the maximum as $d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \max_i |x_i - \hat{x}_i|$ does not work with the assumption of additive distortion measures. In the following, the derivations will be performed for discrete sources, but in most cases it is straight forward to generalise for continuous sources.

In the model for the lossy source coding scheme the distortion is introduced in the encoder/decoder mapping. A mathematical counterpart of the decoder is the probability for the reconstructed symbol $\hat{X}$ conditioned on the source symbol $X$, $p(\hat{x}|x)$. Then the distortion is modeled as a probabilistic mapping between the inputs and outputs. From the assumption of additive distortion, the average distortion over a vector of length $n$ is

$$E\big[d(\boldsymbol{X}, \hat{\boldsymbol{X}})\big] = E\Big[\sum_{i=1}^{n} d(X_i, \hat{X}_i)\Big] = \sum_{i=1}^{n} E\big[d(X_i, \hat{X}_i)\big] = nE\big[d(X, \hat{X})\big] \tag{11.6}$$

By specifying a maximum distortion per symbol as $\delta$ the averaged distortion should be bounded by $E\big[d(\boldsymbol{X}, \hat{\boldsymbol{X}})\big] \leq n\delta$. The expected distortion is averaged over the joint probability of the input sequence and the output sequence, $p(\boldsymbol{x}, \hat{\boldsymbol{x}}) = p(\boldsymbol{x})p(\hat{\boldsymbol{x}}|\boldsymbol{x})$. Among those the input distortion is fixed by the source, meaning that the requirement of a maximum symbol distortion gives a set of conditional distributions as

$$\big\{p(\hat{\boldsymbol{x}}|\boldsymbol{x}) : E[d(\boldsymbol{X}, \hat{\boldsymbol{X}})] \leq n\delta\big\} \tag{11.7}$$

According to (11.6) this can for an additive distortion measure be written as

$$\big\{p(\hat{x}|x) : E[d(X, \hat{X})] \leq \delta\big\} \tag{11.8}$$

From the assumption that $\boldsymbol{Y}$ is a binary vector with average length $L$, the number of codewords is $2^L = 2^{nR}$. Each code vector is decoded to an estimated reconstruction vector $\hat{\boldsymbol{X}}$, and there are equally many possible reconstructed vectors. Thus, the mutual information between the input and the output can be bounded as

$$I(\boldsymbol{X}; \hat{\boldsymbol{X}}) = H(\hat{\boldsymbol{X}}) - H(\hat{\boldsymbol{X}}|\boldsymbol{X}) \leq H(\hat{\boldsymbol{X}}) \leq \log 2^{nR} = nR \tag{11.9}$$

Equivalently, the rate can be bounded by the mutual information as

$$R \geq \frac{1}{n} I(\boldsymbol{X}; \hat{\boldsymbol{X}}) \tag{11.10}$$

That is, to get a measure of the lowest possible rate, the mutual information should be minimised with respect to a certain maximum distortion. Since the maximum distortion level corresponds to a set of conditional distributions the following definition is reasonable.

**DEFINITION 11.3** The *rate-distortion function* for a source with output vector $\boldsymbol{X}$ and a distortion measure $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ is

$$R(\delta) = \min_{p(\hat{\boldsymbol{x}}|\boldsymbol{x}): E[d(\boldsymbol{X},\hat{\boldsymbol{X}})] \le n\delta} \frac{1}{n} I(\boldsymbol{X}; \hat{\boldsymbol{X}}) \qquad (11.11)$$

□

For an i.i.d. source, i.e. memoryless and equally distributed symbols, $I(\boldsymbol{X}; \hat{\boldsymbol{X}}) = \frac{1}{n} I(X; \hat{X})$, together with (11.8) gives the following theorem.

**THEOREM 11.1** The rate-distortion function for an i.i.d. source with output variable $X$, and the distortion measure $d(x, \hat{x})$ is

$$R(\delta) = \min_{p(\hat{x}|x): E[d(X,\hat{X})] \le \delta} I(X; \hat{X}) \qquad (11.12)$$

□

Before showing that $R(\delta)$ is the minimum average number of bits needed to represent a source symbol when the acceptable distortion is $\delta$, a closer look on the actual derivation of the rate-distortion function and some of its properties is in place. If $\delta_1 \le \delta_2$, the set of distributions $\{p(\hat{x}|x) : E[d(X, \hat{X})] \le \delta_1\}$ is a subset of $\{p(\hat{x}|x) : E[d(X, \hat{X})] \le \delta_2\}$, and

$$R(\delta_1) \ge R(\delta_2) \qquad (11.13)$$

Hence, the rate-distortion function is a decreasing function in $\delta$. To see how the rate-distortion function can behave the next example derives it for a binary source.

**EXAMPLE 11.1** Consider a binary i.i.d. source with output symbol $X \in \{0, 1\}$ and $p(X = 0) = p$, where $p \le 1/2$. The aim of this example is to derive the rate-distortion function for binary source and Hamming distortion of maximum $\delta \le 1/2$. To derive the rate-distortion function it is possible to apply standard optimisation technology, but already in this simple case it becomes relatively complex. Instead first note that $E[d(x, \hat{x})] = P(X \ne \hat{X}) = P(X \oplus \hat{X} = 1) \le \delta$. Then a lower bound on the mutual information can be derived as

$$\begin{aligned} I(X; \hat{X}) &= H(X) - H(X|\hat{X}) \\ &= h(p) - H(X \oplus \hat{X}|\hat{X}) \\ &\ge h(p) - H(X \oplus \hat{X}) \ge h(p) - h(\delta) \end{aligned} \qquad (11.14)$$

For this lower bound to equal the rate-distortion function it is needed that $H(X|\hat{X}) = h(\delta)$, which gives the distribution

| | | $X$ | |
|---|---|---|---|
| $p(x\mid\hat{x})$ | | 0 | 1 |
| $\hat{X}$ | 0 | $1-\delta$ | $\delta$ |
| | 1 | $\delta$ | $1-\delta$ |

To get the distribution on $\hat{X}$ assign $P(\hat{X} = 0) = q$,

$$\begin{aligned} p &= P(X = 0) \\ &= P(X = 0|\hat{X} = 0)P(\hat{X} = 0) + P(X = 0|\hat{X} = 1)P(\hat{X} = 1) \\ &= (1-\delta)q + \delta(1-q) = (1-2\delta)q + \delta \end{aligned} \tag{11.15}$$

or, equivalently,

$$q = \frac{p-\delta}{1-2\delta} \quad \text{and} \quad 1-q = \frac{1-p-\delta}{1-2\delta} \tag{11.16}$$

For the case when $0 \leq \delta \leq p \leq 1/2$ the probability of $\hat{X}$ in (11.16) is bounded by $0 \leq q \leq p$.Thus, $q$ and $1-q$ forms a distribution, and according to (11.14) the rate-distortion function is $R(\delta) = h(p) - h(\delta)$, $0 \leq \delta \leq p$.

For the case when $p < \delta \leq 1/2$ let $P(\hat{X} = 1|X) = 1$, $q$ and $1-q$ does not form a distribution since $p - q < 0$. Instead, always set the reconstructed symbol to $\hat{X} = 1$ to get $E[d(X, \hat{X})] = p \leq \delta$ and the distortion requirement is fulfilled. Since $\hat{X} = 1$ independent of $X$ the mutual information is $I(X; \hat{X}) = 0$ which gives $R(\delta) = 0$. Summarising, for a binary i.i.d. source with $P(X = 0) = p$ the rate-distortion function is

$$R(\delta) = \begin{cases} h(p) - h(\delta), & 0 \leq \delta \leq p \leq 1/2 \\ 0, & p < \delta \leq 1/2 \end{cases} \tag{11.17}$$

In Figure 11.2 this function is shown as a plot.

It is interesting to notice in Figure 11.2 that for no distortion, i.e. $\delta = 0$, the rate-distortion function equals the entropy for the source. Since the rate-distortion function was defined as a lower bound on the transmission rate, and that the symbols are binary, this is the amount of information in one source symbol. Thus, it falls back to the lossless case and the source coding theorem as seen before.

In the previous example, the relation between $p(x)$, $p(x|\hat{x})$ and $p(\hat{x})$ is often described by using a *backward test channel* from $\hat{X}$ to $X$, as in Figure 11.3. It

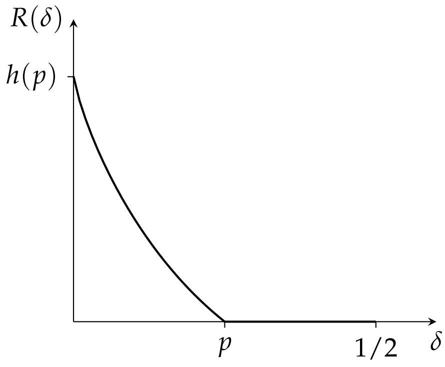

Figure 11.2: Rate-distortion function for a binary i.i.d. source.

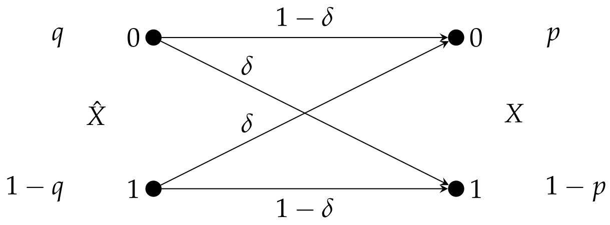

Figure 11.3: Test channel for describing the distributions in Example 11.1.

should be noted that this channel does not have anything to do with transmission, it should be seen as a mathematical model showing the relations. It has its purpose in giving an overview of the distributions involved in the problem.

It turns out that the rete-distortion function plotted in Figure 11.2 has a typical behaviour. It starts at some value for $\delta = 0$ decreases as a convex function down until $\delta = \delta_{\text{max}}$ where $R(\delta_{\text{max}}) = 0$. As was seen in (11.13), the rate-distortion function $R(\delta)$ is a decreasing function, but not necessarily strictly decreasing. At some value $\delta_{\text{max}}$ the allowed distortion is so large the reconstructed value $\hat{X}$ can take a pre-determined value. Then it is not needed to transmit any codeword and the rate becomes $R(\delta_{\text{max}}) = 0$. Since the rate-distortion function is decreasing and the mutual information is non-negative, $R(\delta) = 0$, $\delta \geq \delta_{\text{max}}$. To determine $\delta_{\text{max}}$ notice that since the output is pre-determined, the input $X$ and output $\hat{X}$ are independent, giving $p(\hat{x}|x) = p(\hat{x})$, and

$$\begin{aligned} E\big[d(X,\hat{X})\big] &= \sum_{x,\hat{x}} p(x,\hat{x})d(x,\hat{x}) \\ &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) \\ &= \sum_{\hat{x}} p(\hat{x}) \sum_{x} p(x)d(x,\hat{x}) \end{aligned} \tag{11.18}$$

To get the minimum, find an $\hat{x}$ that minimises $\sum_x p(x)d(x,\hat{x})$ and set $p(\hat{x}) = 1$ for this value, yielding

$$\delta_{\max} = \min_{\hat{x}} \sum_{x} p(x)d(x,\hat{x}) \tag{11.19}$$

To show that the rate-distortion function is convex let $p_1(\hat{x}|x)$ and $p_2(\hat{x}|x)$ denote the distributions achieving $R(\delta_1)$ and $R(\delta_2)$, i.e.

$$R(\delta_1) = I_{p_1}(X;\hat{X}) \quad \text{where} \quad E_{p_1}\big[d(X,\hat{X})\big] \le \delta_1 \tag{11.20}$$

$$R(\delta_2) = I_{p_2}(X;\hat{X}) \quad \text{where} \quad E_{p_2}\big[d(X,\hat{X})\big] \le \delta_2 \tag{11.21}$$

Consider the probability $p(\hat{x}|x) = \alpha_1 p_1(\hat{x}|x) + \alpha_2 p_2(\hat{x}|x)$ where $\alpha_1 \ge 0$, $\alpha_2 \ge 0$ and $\alpha_1 + \alpha_2 = 1$. Then

$$\begin{aligned} E_p\big[d(X,\hat{X})\big] &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) \\ &= \sum_{x,\hat{x}} p(x)\big(\alpha_1 p_1(\hat{x}|x) + \alpha_2 p_2(\hat{x}|x)\big)d(x,\hat{x}) \\ &= \alpha_1 \sum_{x,\hat{x}} p(x)p_1(\hat{x}|x)d(x,\hat{x}) + \alpha_2 \sum_{x,\hat{x}} p(x)p_2(\hat{x}|x)d(x,\hat{x}) \\ &= \alpha_1 E_{p_1}\big[d(X,\hat{X})\big] + \alpha_1 E_{p_1}\big[d(X,\hat{X})\big] \\ &\le \alpha_1\delta_1 + \alpha_2\delta_2 \end{aligned} \tag{11.22}$$

With $\delta = \alpha_1\delta_1 + \alpha_2\delta_2$ this shows $p(\hat{x}|x)$ is one of the distribution in the minimisation to reach $R(\delta)$. From the convexity of the mutual information

$$\begin{aligned} R(\delta) \le I_p(X;\hat{X}) &\le \alpha_1 I_{p_1}(X;\hat{X}) + \alpha_2 I_{p_2}(X;\hat{X}) \\ &= \alpha_1 R(\delta_1) + \alpha_2 R(\delta_2) \end{aligned} \tag{11.23}$$

which shows the convexity of the rate-distortion function. To summarise the above reasoning the following theorem is stated.

**THEOREM 11.2** The rate-distortion function $R(\delta)$ is a convex and decreasing function. Furthermore, there exists a $\delta_{\max} = \min_{\hat{x}} \sum_x p(x)d(x,\hat{x})$ such that $R(\delta) = 0$, $\delta \ge \delta_{\max}$. □

So far the rate-distortion function has been considered for discrete random variables, but the same definition makes sense for continuous variables. The same theory as above will hold for this case. One important case is naturally the Gaussian distribution, that is treated in the next example.

---

**EXAMPLE 11.2** Consider an i.i.d. source where the output is a Gaussian variable $X \sim \mathrm{N}(0, \sigma_X)$. The reconstructed variable is $\hat{X}$, where it is assumed that $E\big[\hat{X}\big] = 0$ is inherited from $X$. It is also assumed that the squared distance distortion measure $d(x, \hat{x}) = (x - \hat{x})^2$ is used. Similar to the previous example with the binary source, instead of going directly to standard optimisation methods, derive a lower bound for the mutual information and find a distribution to fulfil it. Starting with the mutual information

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &= \frac{1}{2}\log(2\pi e\sigma_X^2) - H(X - \hat{X}|\hat{X}) \\ &\geq \frac{1}{2}\log(2\pi e\sigma_X^2) - H(X - \hat{X}) \end{aligned} \tag{11.24}$$

From $E\big[\hat{X}\big] = 0$ it follows that $E\big[X - \hat{X}\big] = 0$ and $V\big[X - \hat{X}\big] = E\big[(X - \hat{X})^2\big] = E\big[d(X, \hat{X})\big] \leq \delta$. Define a random variable $Z \sim \mathrm{N}(0, \sigma_Z)$ where $\sigma_Z^2 = V\big[X - \hat{X}\big]$. The rate-distortion function $R(\delta)$ is found by minimising $I(X;\hat{X})$ over the distributions $f(\hat{x}|x) : \sigma_Z^2 \leq \delta$. Since the Gaussian distribution maximises the differential entropy $H(X - \hat{X}) \leq \frac{1}{2}\log(2\pi e\sigma_Z^2) \leq \frac{1}{2}\log(2\pi e\delta)$. Hence, the bound on the mutual information becomes

$$I(X;\hat{X}) \geq \frac{1}{2}\log(2\pi e\sigma_X^2) - \frac{1}{2}\log(2\pi e\delta) = \frac{1}{2}\log\Big(\frac{\sigma_X^2}{\delta}\Big) \tag{11.25}$$

To see that this bound is actually tight, and equals $R(\delta)$, notice that $X = \hat{X} + Z$ and choose $\hat{X} \sim \mathrm{N}\big(0, \sqrt{\sigma_X^2 - \delta}\big)$ and $Z \sim \mathrm{N}(0, \sqrt{\delta})$. Then $X \sim \mathrm{N}(0, \sigma_X)$ and the average distortion $E\big[d(X, \hat{X})\big] = V\big[Z\big] = \delta$, meaning the minimisation criteria is fulfilled. Hence, for $0 \leq \delta \leq \sigma_X^2$ the rate-distortion function is $R(\delta) = \frac{1}{2}\log\big(\frac{\sigma_X^2}{\delta}\big)$. For $\delta \geq \sigma_X^2$ choose $\hat{X} = 0$ independently of $X$, implying $I(X;\hat{X}) = 0$. The minimisation criteria is fulfilled since $E\big[(X - \hat{X})^2\big] = E\big[X^2\big] = \sigma_X^2 \leq \delta$. Summarising, the rate distortion function for an i.i.d. Gaussian source is

$$R(\delta) = \begin{cases} \frac{1}{2}\log\Big(\frac{\sigma_X^2}{\delta}\Big), & 0 \leq \delta \leq \sigma_X^2 \\ 0, & \delta \geq \sigma_X^2 \end{cases} \tag{11.26}$$

The function is plotted in Figure 11.4.

---

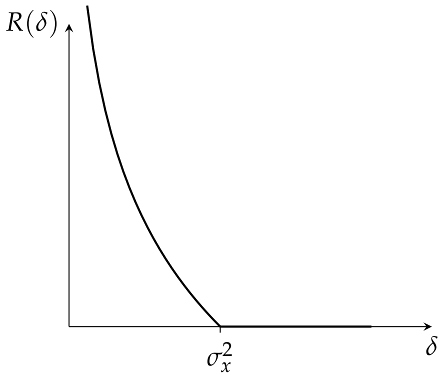

Figure 11.4: Rate-distortion function for a Gaussian source.

The importance of the rate-distortion function was partly seen in (11.10) where the rate is lower bounded by the mutual information. Together with the definition of the rate-distortion function it means for an i.i.d. source that $R(\delta) \leq I(X;\hat{X}) \leq R$. Hence, for any source code with average distortion $E\big[d(X,\hat{X})\big] \leq \delta$ the rate is bounded by $R \geq R(\delta)$. The next theorem, called the *rate-distortion theorem*, is the direct counterpart of the source coding theorem, stating also the existence of such code.

**THEOREM 11.3** Let $\boldsymbol{X} = X_1 X_2 \dots X_n$ be generated by an i.i.d. source, $\hat{\boldsymbol{X}} = \hat{X}_1 \hat{X}_2 \dots \hat{X}_n$ the reconstructed sequence after source coding, and $\delta$ the allowed distortion when the additive distortion measure $d(x,\hat{x})$ is used. Then there exists a source code with rate $R$ if and only if

$$R \geq R(\delta) = \min_{p(\hat{x}|x): E[d(X,\hat{X})] \leq \delta} I(\boldsymbol{X};\hat{\boldsymbol{X}}) \tag{11.27}$$

□

**Proof:** The first part of the theorem, that the rate of a given code satisfying the distortion requirement is bounded by the rate-distortion function, is already shown above. The existence part, that for a given rate satisfying the bound there exists a code, is a bit more tedious. The idea is to extend the concept of jointly typical sequences and construct an encoding/decoding pair satisfying the bound as the length of the source vector grows to infinity. As a start, a set of *distortion typical* sequences is defined.

**DEFINITION 11.4** The set of all *distortion typical* sequences $A_{\varepsilon,\delta}(X,\hat{X})$ is the set of all pairs of $n$-dimensional vectors of i.i.d. variables

$$\boldsymbol{x} = (x_1, x_2, \dots, x_n) \quad \text{and} \quad \hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n) \tag{11.28}$$

such that they are jointly typical $(\boldsymbol{x}, \hat{\boldsymbol{x}}) \in A_\varepsilon(X, \hat{X})$, see Definition 6.5 on page 127, and

$$\Big| \frac{1}{n} d(\boldsymbol{x}, \hat{\boldsymbol{x}}) - E\big[d(X, \hat{X})\big] \Big| \le \varepsilon \tag{11.29}$$

where $E\big[d(X, \hat{X})\big] \le \delta$. □

From the weak law of large numbers it follows directly that

$$\frac{1}{n} d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \xrightarrow{p} E\big[d(X, \hat{X})\big], \quad n \to \infty \tag{11.30}$$

Following Theorem 6.7 it can be seen that there exists a set of integers $n_i$, $i = 1, 2, 3, 4$, such that

$$P_1 = P\Big( \big| -\tfrac{1}{n} \log p(\boldsymbol{x}) - H(X) \big| > \varepsilon \Big) < \frac{\varepsilon}{4}, \qquad n > n_1 \tag{11.31}$$

$$P_2 = P\Big( \big| -\tfrac{1}{n} \log p(\boldsymbol{\hat{y}}) - H(\hat{X}) \big| > \varepsilon \Big) < \frac{\varepsilon}{4}, \qquad n > n_2 \tag{11.32}$$

$$P_3 = P\Big( \big| -\tfrac{1}{n} \log p(\boldsymbol{x}, \hat{\boldsymbol{x}}) - H(X, \hat{X}) \big| > \varepsilon \Big) < \frac{\varepsilon}{4}, \qquad n > n_3 \tag{11.33}$$

$$P_4 = P\Big( \big| \tfrac{1}{n} d(\boldsymbol{x}, \hat{\boldsymbol{x}}) - E\big[d(X, \hat{X})\big] \big| > \varepsilon \Big) < \frac{\varepsilon}{4}, \qquad n > n_4 \tag{11.34}$$

where $E\big[d(X, \hat{X})\big] \le \delta$. Then, for $n > \max\{n_1, n_2, n_3, n_4\}$, by the use of union bound,

$$P\Big( (\boldsymbol{x}, \hat{\boldsymbol{x}}) \notin A_{\varepsilon,\delta}(X, \hat{X}) \Big) < \varepsilon \tag{11.35}$$

and, hence, for arbitrary $\varepsilon > 0$ and sufficiently large $n$

$$P\Big( (\boldsymbol{x}, \hat{\boldsymbol{x}}) \in A_{\varepsilon,\delta}(X, \hat{X}) \Big) \ge 1 - \varepsilon \tag{11.36}$$

From the alternative definition of typical sequences, Definition 6.6, the conditional probability $P(\hat{\boldsymbol{x}}|\boldsymbol{x})$ is bounded as

$$\begin{aligned} p(\hat{\boldsymbol{x}}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}, \hat{\boldsymbol{x}})}{p(\boldsymbol{x})} = p(\hat{\boldsymbol{x}}) \frac{p(\boldsymbol{x}, \hat{\boldsymbol{x}})}{p(\boldsymbol{x}) p(\hat{\boldsymbol{x}})} \\ &\le p(\hat{\boldsymbol{x}}) \frac{2^{-n(H(X,\hat{X})-\varepsilon)}}{2^{-n(H(X)+\varepsilon)} 2^{-n(H(\hat{X})+\varepsilon)}} \\ &= p(\hat{\boldsymbol{x}}) 2^{n(H(X)+H(\hat{X})-H(X,\hat{X})+3\varepsilon)} = p(\hat{\boldsymbol{x}}) 2^{n(I(X;\hat{X})+3\varepsilon)} \end{aligned} \tag{11.37}$$

or, in other words,

$$p(\hat{\boldsymbol{x}}) \geq p(\hat{\boldsymbol{x}}|\boldsymbol{x})2^{-n(I(X;\hat{X})+3\varepsilon)} \tag{11.38}$$

To continue the encoding and decoding procedure should be specified. Firstly, let $p^*(\hat{x}|x)$ be a distribution that gives the rate-distortion function for distortion $\delta$, i.e.

$$p^*(\hat{x}|x) = \arg \min_{p(\hat{x}|x): E[d(X,\hat{X})] \leq \delta} I(X;\hat{X}) \tag{11.39}$$

From the source statistics $p(x)$ let

$$p^*(\hat{x}) = \sum_x p^*(\hat{x}|x)p(x) \tag{11.40}$$

Going back to Figure 11.1 a binary source vector $\boldsymbol{x}$ of length $n$ is encoded to a codeword $\boldsymbol{y}$. The decoder maps the codeword to a reconstructed binary vector $\hat{\boldsymbol{x}}$ of length $n$. When the rate is $R$, there are $2^{nR}$ codewords $\boldsymbol{y}$, and equally many reconstructed vectors $\hat{\boldsymbol{x}}$. To define a decoding rule, generate $2^{nR}$ reconstruction vectors using the distribution

$$p^*(\hat{\boldsymbol{x}}) = \prod_{i=1}^{n} p^*(\hat{x}_i) \tag{11.41}$$

and pair these with the codewords. Denote the decoding function $\hat{\boldsymbol{x}} = g(\boldsymbol{y})$. The encoding rule can be based on typical sequences. Given a vector $\boldsymbol{x}$, find a codeword $\boldsymbol{y}$ such that $(\boldsymbol{x}, g(\boldsymbol{y})) \in A_{\varepsilon,\delta}(X,\hat{X})$. If there are more than one possible codeword, choose one of them at random, and if there is no codeword forming a typical pair with $\boldsymbol{x}$ choose $\boldsymbol{y} = \mathbf{0}$. To see what this means for the average distortion first define the event that $\boldsymbol{x}$ and $\hat{\boldsymbol{x}} = g(\boldsymbol{y})$ are distortion typical sequences,

$$E_{\boldsymbol{y}|\boldsymbol{x}} \big\{ (\boldsymbol{x}, \hat{\boldsymbol{x}}) \in A_{\varepsilon,\delta}(X,\hat{X}) \big| \hat{\boldsymbol{x}} = g(\boldsymbol{y}) \big\} \tag{11.42}$$

Then the event that $\boldsymbol{x}$ does not have any matching codeword becomes

$$E_{e|\boldsymbol{x}} = \bigcap_{\boldsymbol{y}} E^c_{\boldsymbol{y}|\boldsymbol{x}} \tag{11.43}$$

Since the reconstructed vectors are generated i.i.d. the corresponding codewords are independent and

$$\begin{aligned}
P(E_{e|\boldsymbol{x}}) &= P\big(\bigcap_{\boldsymbol{y}} E^c_{\boldsymbol{y}|\boldsymbol{x}}\big)\\
&= \prod_{\boldsymbol{y}} P(E^c_{\boldsymbol{y}|\boldsymbol{x}})\\
&= \prod_{\boldsymbol{y}} (1 - P(E_{\boldsymbol{y}|\boldsymbol{x}}))\\
&= \prod_{\boldsymbol{y}} \Big(1 - \sum_{\hat{\boldsymbol{x}}:(\boldsymbol{x},\hat{\boldsymbol{x}})\in A_{\varepsilon,\delta}} p(\hat{\boldsymbol{x}})\Big)\\
&\leq \prod_{\boldsymbol{y}} \Big(1 - \sum_{\hat{\boldsymbol{x}}:(\boldsymbol{x},\hat{\boldsymbol{x}})\in A_{\varepsilon,\delta}} p(\hat{\boldsymbol{x}}|\boldsymbol{x}) 2^{-n(I(X;\hat{X})+3\varepsilon)}\Big)\\
&= \Big(1 - \sum_{\hat{\boldsymbol{x}}:(\boldsymbol{x},\hat{\boldsymbol{x}})\in A_{\varepsilon,\delta}} p(\hat{\boldsymbol{x}}|\boldsymbol{x}) 2^{-n(I(X;\hat{X})+3\varepsilon)}\Big)^{2^{nR}}\\
&= \Big(1 - 2^{-n(I(X;\hat{X})+3\varepsilon)} \sum_{\hat{\boldsymbol{x}}:(\boldsymbol{x},\hat{\boldsymbol{x}})\in A_{\varepsilon,\delta}} p(\hat{\boldsymbol{x}}|\boldsymbol{x})\Big)^{2^{nR}} \qquad (11.44)
\end{aligned}$$

For $1 - \alpha x > 0$, the IT-inequality gives $\ln(1 - \alpha x) \leq -\alpha x$. Thus, $(1 - \alpha x)^M = e^{M \ln(1-\alpha x)} \leq e^{-M\alpha x}$. Furthermore, for $0 \leq x \leq 1$ it can be found that

$$e^{-M\alpha x} \leq 1 - x + e^{M\alpha} \qquad (11.45)$$

To see this, first notice that the bound is clearly fulfilled for the end points $x = 0$ and $x = 1$. In the considered interval the left hand side, $e^{-M\alpha x}$ is convex, while the right hand side is linearly decreasing with $x$, and, hence, the bound must be fulfilled in between the end points as well. So, for $0 \leq x \leq 1$, $0 \leq \alpha \leq 1$ and $M \geq 0$

$$(1 - \alpha x)^M \leq 1 - x + e^{M\alpha} \qquad (11.46)$$

Applying to (11.44) and identifying $M = 2^{nR}$, $x = \sum p(\hat{\boldsymbol{x}}|\boldsymbol{x})$ and $\alpha = 2^{-n(I(X;\hat{X})+3\varepsilon}$ gives

$$P(E_{e|\boldsymbol{x}}) \leq 1 - \sum_{\hat{\boldsymbol{x}}:(\boldsymbol{x},\hat{\boldsymbol{x}})\in A_{\varepsilon,\delta}} p(\hat{\boldsymbol{x}}|\boldsymbol{x}) + e^{2^{-n(I(X;\hat{X})+3\varepsilon)} 2^{nR}} \qquad (11.47)$$

Averaging over all $\boldsymbol{x}$ gives the total probability of no match as

$$\begin{aligned}
P(E_e) &= \sum_{\boldsymbol{x}} p(\boldsymbol{x}) P(E_{e|\boldsymbol{x}})\\
&\leq 1 - \sum_{(\boldsymbol{x},\hat{\boldsymbol{x}})\in A_{\varepsilon,\delta}} p(\boldsymbol{x}) p(\hat{\boldsymbol{x}}|\boldsymbol{x}) + e^{2^{n(R - I(X;\hat{X}) - 3\varepsilon)}}\\
&= P\big((\boldsymbol{x}, \hat{\boldsymbol{x}}) \notin A_{\varepsilon,\delta}\big) + e^{2^{n(R - R(\delta) - 3\varepsilon)}} \qquad (11.48)
\end{aligned}$$

where it is used that $p(\hat{x}|x) = p^*(\hat{x}|x)$ to get $I(X;\hat{X}) = R(\delta)$ in the last equality. From the definition of $A_{\varepsilon,\delta}(X,\hat{X})$, the term $P\big((\boldsymbol{x},\hat{\boldsymbol{x}}) \notin A_{\varepsilon,\delta}\big) \leq \varepsilon$, where $\varepsilon$ can be chosen arbitrarily small. For $R > R(\delta)$ and small enough $\varepsilon$, the exponent in the second term $R - R(\delta) - 3\varepsilon < 1$, and the term will decrease towards zero as $n$ grows. Thus, with $R > R(\delta)$ it is possible to find a code where $P(E_e) \to 0$ as $n \to \infty$.

To derive the average distortion consider first the vector pairs $(\boldsymbol{x},\hat{\boldsymbol{x}}) \in A_{\varepsilon,\delta}(X,\hat{X})$. Then the distortion is bounded by

$$\frac{1}{n}d(\boldsymbol{x},\hat{\boldsymbol{x}}) \leq E_{p^*}\big[d(X,\hat{X})\big] + \varepsilon \leq \delta + \varepsilon \tag{11.49}$$

For the vector pairs not included in the set of distortion typical sequences the distortion is bounded by $\frac{1}{n}d(\boldsymbol{x},\hat{\boldsymbol{x}}) \leq \hat{\delta}$, where $\hat{\delta} = \max_{(x,\hat{x})} d(x,\hat{x})$ is assumed to be finite. Then the average distortion is

$$\begin{aligned}\frac{1}{n}E\big[d(X,\hat{X})\big] &\leq (\delta+\varepsilon)P(E_e^c) + \hat{\delta}P(E_e)\\ &\leq \delta + \varepsilon + \hat{\delta}P(E_e) = \delta + \bar{\varepsilon}\end{aligned} \tag{11.50}$$

where $\bar{\varepsilon} = \varepsilon + \hat{\delta}P(E_e)$ can be chosen arbitrarily small, for large enough $n$. This completes the proof. ■

From the above rate-distortion theorem, the rate-distortion function plays the same role for lossy source coding as the entropy does for lossless source coding. It is the limit for when it is possible to find a code. It does not, however, say much on how to construct the code since the construction in the proof is not practically implementable. Especially in the area of image, video and voice coding there are active research ongoing. Another, closely related, topic is quantisation, which in its nature is both lossy and a compression. In Section 11.3 quantisation is treated more in detail, and in Section 11.4 transform coding is described, including overviews of JPEG and MPEC coding.

As in the case of the source coding theorem the rate-distortion theorem can be generalised to hold for stationary ergodic sources. The theory for this is out of the scope for this text.

## 11.2 Limit for fix $P_b$

In the previous section it was shown that the rate-distortion function has the same interpretation for lossy source coding as the entropy has for lossless source coding. In this section it will be seen that it also can be applied to the case of channel coding when a certain bit error rate can be acceptable. For this purpose the system model has to be expanded a bit to include the channel. In Figure 11.5 the source vector $\boldsymbol{X} = X_1 \dots X_k$ is of length $k$ and the code vector $\boldsymbol{Y} = Y_1 \dots Y_n$ of length $n$, which gives the encoding rate $R = \frac{k}{n}$. After transmission over the channel the received vector is $\hat{\boldsymbol{Y}} = \hat{Y}_1 \dots \hat{Y}_n$. Then the decoding outputs the estimated vector as $\hat{\boldsymbol{X}} = \hat{X}_1 \dots \hat{X}_k$.

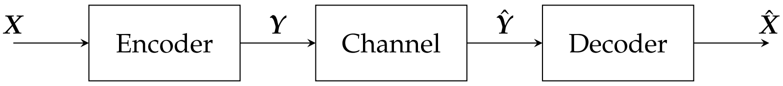

Figure 11.5: Block scheme with source and channel coding.

In Section 6.5 it was shown that reliable communication is possible if and only if the rate is bounded by the capacity,

$$R < C = \max_{p(y)} I(Y; \hat{Y}) \tag{11.51}$$

The term *reliable communication* refers to the case when the error probability after decoding can be made arbitrarily low. As with the case of lossless compared to lossy compression this puts some hard restrictions on the system. In a real system design a certain level of error probability can often be accepted. It is possible to treat this error level as an acceptable level of distortion at the decoder output. The next theorem shows the relation between the channel capacity and the rate-distortion function.

**THEOREM 11.4** Given a source with probability distribution $p(\boldsymbol{x})$, that is encoded with a rate $R$ channel code before transmitted over a channel. If the acceptable distortion is $\delta$ for a distortion measure $d(x, \hat{x})$, such system can be designed if and only if

$$R \leq \frac{C}{R(\delta)} \tag{11.52}$$

where $C$ is the channel capacity for the channel and $R(\delta)$ the rate-distortion function. □

In this text the proof of the theorem is omitted. Instead refer to [47].

The above theorem gives a relation between the channel capacity and the rate-distortion function. In the next, the influence of the acceptable distortion on the fundamental limit in Section 9.3 is treated. The limit $E_b/N_0 \geq -1.59$ dB was derived by considering a binary equiprobable source where the bits are encoded by a rate $R$ channel code before the bits are transmitted over channel with signal to noise ratio $E_b/N_0$. For reliable communication the limit on the coding rate can be written as

$$R < C = \frac{1}{2}\log\Big(1 + 2R\frac{E_b}{N_0}\Big) \tag{11.53}$$

Assuming a binary source with equally distributed bits, and an acceptable bit error probability $P_b$ after decoding, the rate-distortion function is given by

$$R(P_b) = 1 - h(P_b) \tag{11.54}$$

Thus, the code rate bound becomes

$$R < \frac{C}{R(P_b)} = \frac{\frac{1}{2}\log\Big(1 + 2R\frac{E_b}{N_0}\Big)}{1 - h(P_b)} \tag{11.55}$$

Equivalently, rewritten as a bound on the signal to noise ratio for communication with a maximum bit error probability,

$$\frac{E_b}{N_0} > \frac{2^{2R(1-h(P_b))} - 1}{2R} \tag{11.56}$$

In Figure 11.6 the bound is plotted as the minimum signal to noise ratio for the bit error probability. In the figure there are four plots, one each for the coding rates $R = 1/4$, $R = 1/2$, $R = 3/4$, and the fourth, left most, curve is the case when the encoding rate tends to zero. This is the case when the fundamental limit is given, and the function becomes

$$\lim_{R\to 0}\frac{2^{2R(1-h(P_b))} - 1}{2R} = \ln(2)(1 - h(P_b)) \tag{11.57}$$

which describes the lowest achievable $E_b/N_0$ for an acceptable bit error probability of $P_b$ at the receiver. As the bit error rate becomes smaller the entropy function in the formula will close to zero and the curves fall down close to vertically below about $P_e = 10^{-3}$, where they equal the capacity limit for a given rate.

## 11.3 Quantisation

In the next two sections two examples of lossy compression are considered. Firstly it is quantisation that represents a continuous variable by a discrete, and thus introducing disturbance, and secondly transform decoding. A well known example of the latter is the image format JPEG that will be briefly described.

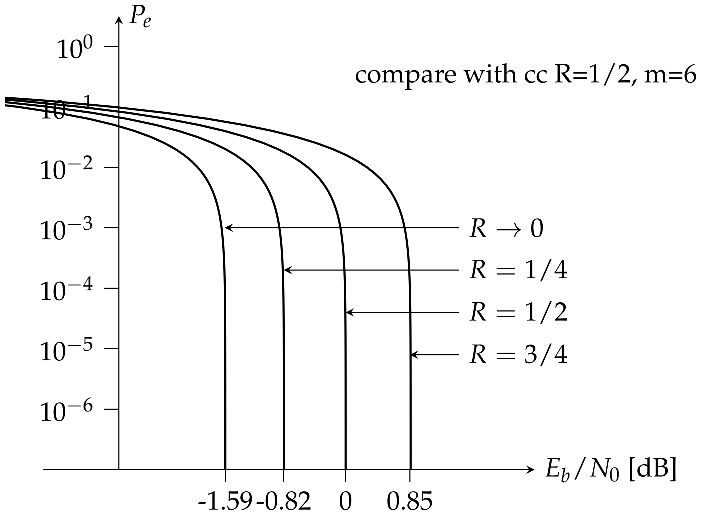

Figure 11.6: Plot of achievable SNR for certain bit error probability.

As said, quantisation maps a continuous variable to a discrete version for the purpose of representing it with a finite length binary vector. An analog to digital converter (ADC) consist of sampling and quantisation, i.e. first mapping from continuous time to discrete time and then from continuous amplitude to discrete amplitude. This operation, as well as its inverse–digital to analog conversion (DAC), is a common component in circuits operating with signals from and to an outer unit, like a sensor of some kind.

In the sampling procedure, the optimal sampling frequency and the reconstruction formula is described by the sampling theorem used in Chapter 9. According to this, sampling and reconstruction does not introduce any distortion. However, to be able to represent the sample values in a computer using finite vectors they have to be quantised. This operation means representing a real value by a discrete variable, and it is inevitable that information is destroyed, and thus distortion introduced.

In this description a linear quantisation, as in Figure 11.7, is used. The input to the quantiser is the continuous variable $x$. The mapping to the quantiser output $x_Q$ is determined by a staircase function in the figure. In a linear quantiser the size of the steps is constant, say $\Delta$.

The term linear quantiser comes from the dashed line centered in the staircase function in the figure, which is a linear function. For a non-linear quantiser this center function can have more of an S-shape in either direction. This can

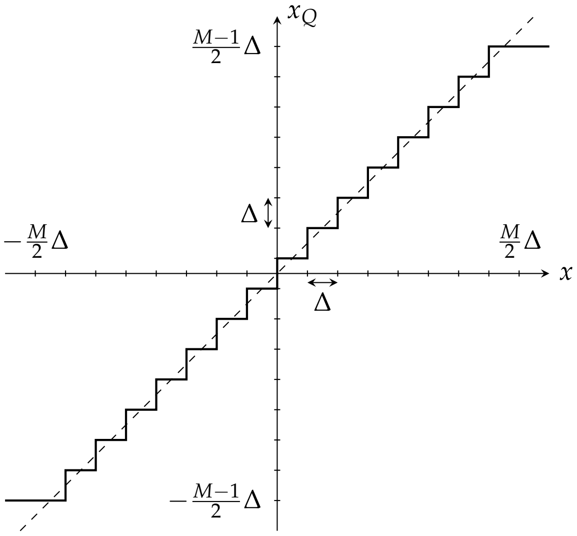

Figure 11.7: A linear quantisation function.

be used to form the quantiser for the statistics of the continuous source and to have different sizes of the quantisation intervals. However, in most practical implementations a linear quantiser is used. If the statistics differ much from the uniform distribution the quantiser can either be followed by a source code like the Huffman code or preceded by a compensation filter.

In the figure a quantiser with $M$ output levels is shown. Assuming a maximum level of the output mapping of $D = \frac{M-1}{2}\Delta$, the granularity of the quantisation becomes $\Delta = \frac{2D}{M-1}$. The $m$th output level then corresponds to the value

$$x_Q(m) = \Big(m\Delta - \frac{M-1}{2}\Delta\Big) = (2m - M + 1)\frac{\Delta}{2} \tag{11.58}$$

The mapping function shown in the figure is determined from finding an integer $m$ such that

$$x_Q(m) - \frac{\Delta}{2} \le x < x_Q(m) - \frac{\Delta}{2} \tag{11.59}$$

then the output index is $y = m$. If the input value $x$ can exceed the interval $x_Q(0) - \frac{\Delta}{2} \le x < x_Q(M-1) + \frac{\Delta}{2}$ the limits should be

$$y = \begin{cases} m, & x_Q(m) - \frac{\Delta}{2} \le x < x_Q(m) - \frac{\Delta}{2}, \quad 1 \le m \le M-2 \\ 0, & x < x_Q(0) + \frac{\Delta}{2} \\ M-1, & x \ge x_Q(M-1) - \frac{\Delta}{2} \end{cases} \tag{11.60}$$

From (11.58) this can equivalently be written as

$$y = \begin{cases} m, & (2m-M)\frac{\Delta}{2} \leq x < (2m-M)\frac{\Delta}{2} + \Delta, \quad \text{for } 1 \leq m \leq M-2 \\ 0, & x < (2-M)\frac{\Delta}{2} \\ M-1, & x \geq (M-2)\frac{\Delta}{2} \end{cases} \tag{11.61}$$

The output values from the quantiser can be represented by a finite length binary vector. The price for representing a real value with finite levels is an error introduced in the signal. In Figure 11.8 this error, defined as the difference $x - x_Q$, is shown in the upper plot, and the corresponding distortion, $d(x, x_Q) = (x - x_Q)^2$ in the lower plot.

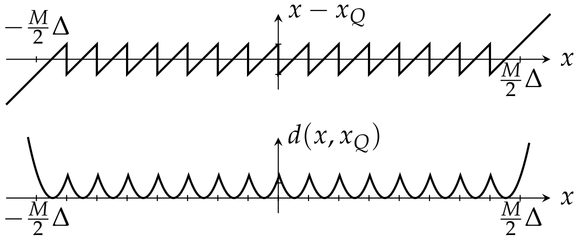

Figure 11.8: The quantisation error, $x - x_Q$, and distortion, $d(x, x_Q) = (x - x_Q)^2$, for the linear quantisation function in Figure 11.7.

An estimate of the distortion introduced can be made by considering a uniformly distributed input signal, $X \sim \mathrm{U}(-M\frac{\Delta}{2}, M\frac{\Delta}{2})$. Then all quantisation levels will have uniformly distributed input with $f(x) = \frac{1}{\Delta}$, and deriving the average distortion can be made with normalised $x_Q = 0$,

$$E\big[(X - X_Q(m))^2 | Y = m\big] = \int_{-\Delta/2}^{\Delta/2} x^2 \frac{1}{\Delta} dx = \frac{\Delta^2}{12} \tag{11.62}$$

From the uniform assumption $P(Y = m) = \frac{1}{M}$, and hence

$$E\big[(X - X_Q)^2\big] = \sum_{m=0}^{M-1} \frac{1}{M} \frac{\Delta^2}{12} = \frac{\Delta^2}{12} \tag{11.63}$$

When the quantised value is mapped to a vector of length $k$ and $M = 2^k$ this is equivalent to, $E\big[(X - X_Q(m))^2\big] \approx 2^{2(k-1)} \frac{D^2}{12}$, where the approximation $M - 1 \approx M$ is used.

Viewing the distortion as a noise, it is convenient to consider the signal to quantisation noise ratio, SQNR. Since the signal has zero mean its variance is

$$E\big[X^2\big] = \int_{-M\frac{\Delta}{2}}^{M\frac{\Delta}{2}} x^2 \frac{1}{M\Delta} dx = \frac{(M\Delta)^2}{12} \tag{11.64}$$

Hence the signal to quantisation noise is

$$\text{SQNR} = \frac{E\big[X^2\big]}{E\big[(X - X_Q)^2\big]} = M^2 = 2^{2k} \tag{11.65}$$

Expressed in dB this means

$$\text{SQNR}_{\text{dB}} = 2k \cdot 10 \log_{10} 2 \approx k \cdot 6 \text{ dB} \tag{11.66}$$

i.e. the SQNR increases with 6 dB with each bit in the quantisation.

**EXAMPLE 11.3** In the 4G mobile standard LTE, the downstream signals are constructed with an OFDM modulation scheme. The modulation carries 2, 4 or 6 bits per tone and transmission. To get the maximum data rate of the system a reasonable lower requirement on the signal to noise ratio is 30 dB. Then the quality of the total channel, both quantisation and air channel, will not constrain the modulation due to the quantisation. If the air channel is good enough for full speed, so will the combination with quantisation. From the approximation of 6 dB per bit, this corresponds to $k = 5$ b/sample.

There are six possible bandwidths for the communication link,

$$W \in \{1.4, 3, 5, 10, 15, 20\} \text{ [MHz]} \tag{11.67}$$

Following the Nyquist sampling rate $F_S \geq 2W$, and since the samples are complex the total required bit rate is $R_b = 2F_s k \geq 2W \cdot 2 \cdot 5 = W \cdot 20$. In the next table the resulting minimum bit rates for the LTE bands are shown. The calculations are based on uniformly distributed amplitude of the samples, which is not the case in reality. So, the result is a bit optimistic and a real signal would require some extra bits per real sample.

| $W$ [MHz] | $R_{b,\min}$ [Mbps] | $CPRI$ [Mbps] |
|---|---|---|
| 1.4 | 28 | 614.4/8 |
| 3 | 60 | 614.4/3 |
| 5 | 100 | 614.4 |
| 10 | 200 | 1228.8 |
| 15 | 300 | 1228.8 |
| 20 | 400 | 2457.6 |

As a comparison, for each bit rate, the rates used by the fronthaul protocol CPRI is shown. This is a standard developed for transporting samples within the base-station, but also often considered for transporting LTE samples over fibre connections further distances. The relatively high bit rates comes from the requirement of 15 b/real sample. In CPRI the specified bit rates in Mbps are 614.4, 1228.8, 2457.6, 3072, 4915.2, 6144, 9830.4. Then for the 1.4 MHz band there can be 8 signals in one 614.4 Mbps stream and for the 3 MHz band 3 signals in a 614.4 Mbps stream. For the others it is one signal per stream.

In the case of uniform distribution it is natural to set the reconstructed value to the centre in the quantisation interval. In the general case, for a given intervall $\Delta_m \leq x < \Delta_{m+1}$ and a reconstruction value $x_m$ in the interval $m$ with the distribution $f(x|m)$, the average distortion is

$$d_m = E_{X|m}\big[(X - x_m)^2\big] = E_{X|m}\big[X^2\big] - 2x_m E_{X|m}\big[X\big] + x_m^2 \tag{11.68}$$

Thus, to find the reconstruction value that minimises the distortion take the derivative with respect to $x_m$ to get

$$\frac{\partial d_m}{\partial x_m} = -2E_{X|m}\big[X\big] + 2x_m = 0 \tag{11.69}$$

and hence the optimal reconstruction value is $x_m = E_{X|m}\big[X\big]$. For the uniform distribution this is indeed the centre in the intervall as used above. For other distributions the value can change. In the next example the reconstruction value for a Gaussian source when using a 1-bit quantiser is derived.

**EXAMPLE 11.4** Assume a Gaussian source where $X \sim \mathrm{N}(0, \sigma)$ and a 1 bit quantiser. The natural intervalls are divided buy the value $x = 0$, i.e. for $x < 0, y = 0$ and for $x \geq 0$, $y = 1$. Since the two sides are symmetric it is only needed to derive the optimal reconstruction level for the positive side. There, the distribution is given by

$$f(x|y = 1) = \frac{2}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} \tag{11.70}$$

and the reconstruction value is

$$x_1 = \int_0^\infty x \frac{2}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} dx = \sqrt{\frac{2}{\pi}}\sigma \tag{11.71}$$

Consequently, the reconstruction value for the negative side

$$x_0 = -\sqrt{\frac{2}{\pi}}\sigma \tag{11.72}$$

With these levels the average quantisation distortion becomes $E\big[d(X, X_Q)\big] = \frac{\sigma^2}{\pi}(\pi - 2)$.