Chapter 3

Shannon Information Theory

The information theory established by Shannon in 1948 is the foundation discipline for communication systems, showing the potentialities and fundamental bounds of coding. This includes channel coding (also called error-control coding) as well as source coding and secrecy coding (also called cryptography). However, in this chapter we will only consider the information theory of channel coding restricted to the discrete memoryless channel (DMC). Although block codes have not yet been discussed extensively, the results of Shannon's information theory will nevertheless become quite clear even with our present basic knowledge.

3.1 Channel Capacity of Discrete Memoryless Channels

First, we will need a more detailed probability description of the stochastic input and the stochastic discrete channel. Using the concepts of entropy and mutual information the channel capacity will be introduced as a fundamental property of the channel and calculated for the two standard examples of BSC and binary AWGN channel. In the next section the meaning of the capacity will become clear with the channel coding theorems.

3.1.1 The Joint Probability Distribution for Source and Channel

A detailed explanation of the necessary probability basics and the relations between joint, marginal and conditional distributions is given in Section A.3 of the appendix.

Assume an abstract discrete memoryless channel (DMC) which is defined by the triplet $(\mathcal{A}_{\text{in}}, \mathcal{A}_{\text{out}}, P_{y|x})$ as in Subsection 1.3.1, where $\mathcal{A}_{\text{in}}$ is the q-ary input alphabet and $\mathcal{A}_{\text{out}}$ is the output alphabet of the channel, and $P_{y|x}(\eta|\xi)$ or simply $P(y|x)$ is the transition probability (also called channel statistic) according to

Definition 1.1. The input to the abstract DC (which is the same as the output of the channel encoder) is a random variable x which is characterized by the a priori distribution $P_x(\xi)$ or simply $P(x)$. Recall that

$$\begin{aligned} P(y|x) &= P(y \text{ received} \mid x \text{ transmitted}), \\ P(x) &= P(x \text{ transmitted}). \end{aligned} \tag{3.1.1}$$

The probability description of the received signal y can be calculated from the probability description of the input and the channel. According to Bayes' theorem of total probability (A.3.1), the probability that y is received turns out to be

$$P(y) = \sum_{x \in \mathcal{A}_{\text{in}}} P(y|x) \cdot P(x). \tag{3.1.2}$$

For the joint probability distribution that x is transmitted and y is received,

$$P(x,y) = P(y|x) \cdot P(x) = P(x|y) \cdot P(y), \tag{3.1.3}$$

also introducing the conditional probability $P(x|y)$, i.e., x was transmitted assuming that y was received. The joint probability distribution $P(x,y)$ leads to $P(y)$ and $P(x)$ as marginal probability distributions:

$$P(y) = \sum_{x \in \mathcal{A}_{\text{in}}} P(x,y), \qquad P(x) = \sum_{y \in \mathcal{A}_{\text{out}}} P(x,y). \tag{3.1.4}$$

For completeness, note that

$$\sum_{x \in \mathcal{A}_{\text{in}}} P(x) = \sum_{y \in \mathcal{A}_{\text{out}}} P(y) = \sum_{y \in \mathcal{A}_{\text{out}}} P(y|x) = \sum_{x \in \mathcal{A}_{\text{in}}} P(x|y) = 1. \tag{3.1.5}$$

All previous statistical relations and the following definitions for entropy, mutual information and channel capacity can also be introduced for the inner DMC $(\mathcal{A}_{\text{mod}}, \mathcal{A}_{\text{dem}}, P_{\tilde{y}|\tilde{x}})$ with 2^M-ary input, in a similar way as for the abstract DMC $(\mathcal{A}_{\text{in}}, \mathcal{A}_{\text{out}}, P_{y|x})$ with q-ary input.

3.1.2 Entropy and Information

The amount of information (also called uncertainty) of a random variable with the q-ary range $\mathcal{A}_{\text{in}}$ is measured by the *entropy*

$$H(x) = -\sum_{x \in \mathcal{A}_{\text{in}}} P(x) \log_2 P(x) = -\sum_{i=1}^{q} p_i \log_2 p_i \tag{3.1.6}$$

where p_i denotes the probability that x attains the i-th value of the input alphabet. The concept of entropy, including all the proofs, is presented in detail in Appendix A.5. The entropy is measured in units of bits, because of the use of the logarithm to the base 2, and is generally bounded as $0 \le H(x) \le \log_2 q$.

The entropy $H(x)$ takes on the minimum value 0, if the random variable x is constant.

The entropy takes on its maximum, if all values are attained with the same probability of $p_i = 1/q$. Such a uniform distribution was presupposed for the derivation of the maximum-likelihood decoder in Section 1.6, i.e., all codewords and all components of the codeword occur with the same probability. Then the information symbols and the information words have the maximum possible entropy $H(x) = \log_2 q$ and $H(\boldsymbol{x}) = k \cdot \log_2 q$, respectively. For the entropy of the received values, $H(y) \leq \log_2 |\mathcal{A}_{\text{out}}|$ assuming that $|\mathcal{A}_{\text{out}}|$ is finite.

The *joint entropy* of two random variables x and y is a generalization of the entropy to the joint distribution:

$$H(x,y) = -\sum_{x \in \mathcal{A}_{\text{in}}, y \in \mathcal{A}_{\text{out}}} P(x,y) \log_2 P(x,y). \tag{3.1.7}$$

If x and y are two statistically independent random variables, then obviously $H(x,y) = H(x)+H(y)$. The definition of the entropy can be further generalized to the *conditional entropy* of x for a given y:

$$H(x|y) = -\sum_{x,y} P(x,y) \log_2 P(x|y), \tag{3.1.8}$$

which is lower and upper bounded by

$$0 \leq H(x|y) \leq H(x), \tag{3.1.9}$$

where equality occurs on the left side for $x \equiv y$ and equality on the right side for statistically independent random variables x and y. So if there is a dependence between x and y, the uncertainty of x is reduced by the increasing knowledge of y, i.e., conditioning on random variables can never increase uncertainty, but reduces uncertainty by the amount of $H(x) - H(x|y)$.

3.1.3 Mutual Information and Channel Capacity

We recall that the actual task of information transmission over a stochastic channel is to decide on the input x from the output y with a minimum of uncertainty. So the separated entropies of input and output are not the only relevant properties, but also the relation between input and output, which should be as close as possible. Now, we will introduce an important mathematical measure for this input-output relation, which is based on the joint distribution of input and output:

Definition 3.1. *Between two random variables and in particular between the input and the output of the DMC the* mutual information *indicates how much* *information one of the random variables reveals about the other:*

$$I(x;y) = \sum_{x\in\mathcal{A}_{\text{in}},y\in\mathcal{A}_{\text{out}}} P(x)P(y|x)\log_2 \frac{P(y|x)}{\sum\limits_{x'\in\mathcal{A}_{\text{in}}} P(x')P(y|x')} \tag{3.1.10}$$

$$= \sum_{x\in\mathcal{A}_{\text{in}},y\in\mathcal{A}_{\text{out}}} P(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)} \tag{3.1.11}$$

$$= \underbrace{\sum_{y\in\mathcal{A}_{\text{out}}} P(y)\log_2 \frac{1}{P(y)}}_{=H(y)} - \underbrace{\sum_{x\in\mathcal{A}_{\text{in}},y\in\mathcal{A}_{\text{out}}} P(x,y)\log_2 \frac{1}{P(y|x)}}_{=H(y|x)} \tag{3.1.12}$$

$$= \underbrace{\sum_{x\in\mathcal{A}_{\text{in}}} P(x)\log_2 \frac{1}{P(x)}}_{=H(x)} - \underbrace{\sum_{x\in\mathcal{A}_{\text{in}},y\in\mathcal{A}_{\text{out}}} P(x,y)\log_2 \frac{1}{P(x|y)}}_{=H(x|y)} \tag{3.1.13}$$

$$= -\underbrace{\sum_{x\in\mathcal{A}_{\text{in}},y\in\mathcal{A}_{\text{out}}} P(x,y)\log_2 \frac{1}{P(x,y)}}_{=H(x,y)} + H(x) \ + \ H(y). \tag{3.1.14}$$

The equivalence of these five terms follows from the elementary relations between the various probability functions previously mentioned in Subsection 3.1.1. In (3.1.10), $I(x;y)$ is characterized by the input distribution $P(x)$ and the channel properties represented by the transition probability $P(y|x)$. Equation (3.1.11) emerged from the joint distribution and the two marginal probability distributions. The last three equations are based on entropies. In (3.1.12) the mutual information is determined by the difference between the output entropy $H(y)$ and the conditional entropy $H(y|x)$ and in (3.1.13) by the difference between the input entropy $H(x)$ and the conditional entropy $H(x|y)$. These conditional entropies have specific names, $H(y|x)$ is called *noise entropy* or *prevarication*, and $H(x|y)$ is known as *equivocation*. In the last term (3.1.14), the mutual information is expressed by the joint entropy and the two marginal entropies.

The mutual information is obviously symmetrical in x and y. Appendix A.5 shows that all entropies, including the conditional entropies as well as the mutual information, are non-negative.

The relations between the entropies and the mutual information are shown in Figure 3.1, which implies the following interpretations.

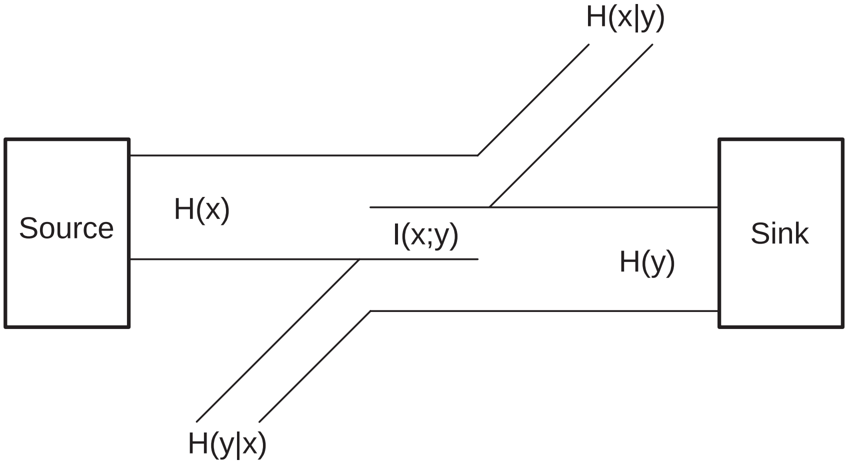


Figure 3.1. The DMC from the entropy's point of view

- Firstly, $H(x) = I(x;y) + H(x|y)$, so the transmitted information measured by $H(x)$ is reduced by $H(x|y)$. The uncertainty of the transmitted input, expressed by $H(x|y)$, should be as small as possible for a given received output.
- Secondly, $H(y) = I(x;y) + H(y|x)$, so a further entropy $H(y|x)$ is added to the channel, where the uncertainty of the output y for a given input x represents the stochastic behaviour of the channel.
- The information passed from the transmitter to the receiver is the actual mutual information $I(x;y)$, which is obviously bounded as $I(x;y) \leq \min\{H(x), H(y)\}$.

Example 3.1. Let us consider the two extreme cases, worst case and best case, for the DMC.

(1) The worst case is that there are no statistical dependencies between the input and the output of the DMC, i.e., no information can be transmitted. In this case x and y are statistically independent, and the joint distribution is factorized as $P(x,y) = P(x)P(y)$. Then in (3.1.11) $\log_2(\ldots) = 0$, and

$$I(x;y) = 0, \quad H(y|x) = H(y), \quad H(x|y) = H(x), \quad H(x,y) = H(x) + H(y). \tag{3.1.15}$$

The second equation, $H(y|x) = H(y)$, means that the information content of y is independent of whether or not x is known. For a good channel, y should have no content of information if x is known, since y should be determined by x. The third equation, $H(x|y) = H(x)$, can be interpreted accordingly. Furthermore, $I(x,y) = 0$ is equivalent to x and y being statistically independent.

(2) The best case is that the channel is transparent, i.e., $x \equiv y$. Then the channel is no longer stochastic but deterministic, thus obviously $P(y) = P(x)$

and

$$P(y|x) = \left\{ \begin{array}{ll} 1 & y = x \\ 0 & y \neq x \end{array} \right\} \quad , \quad P(x,y) = \left\{ \begin{array}{ll} P(x) & y = x \\ 0 & y \neq x \end{array} \right\}.$$

Considering $H(y|x)$ in (3.1.12) it turns out that for $y = x$, $\log_2(\ldots) = 0$, and $P(x,y) = 0$ otherwise. Thus $H(y|x) = 0$, i.e., for a given x there is no uncertainty of y. $H(x|y) = 0$ can be interpreted accordingly. Summarizing,

$$I(x;y) = H(y) = H(x) = H(x,y), \quad H(y|x) = 0, \quad H(x|y) = 0. \tag{3.1.16}$$

For uniformly distributed input symbols, $I(x;y) = H(x) = \log_2 q$. ■

Definition 3.2. *The* channel capacity C *of the DMC* $(\mathcal{A}_{\text{in}}, \mathcal{A}_{\text{out}}, P_{y|x})$ *is defined as the maximum of the mutual information over of all possible input statistics, i.e., over all a priori distributions:*

$$C = \max_{P_x} I(x;y), \tag{3.1.17}$$

where the capacity is measured in units of information bits per q-ary encoded symbol, in other words, per use of the abstract discrete channel.

The capacity is bounded as $0 \leq C \leq \log_2 q$ (with $q = 2$ for binary input). For $C = 0$ the output is statistically independent of the input and no information can be transmitted. For $C = \log_2 q$ the channel is transparent.

To achieve the maximum channel capacity, the input and thus the source probability distribution would have to be adapted according to the channel statistic. This is usually difficult to handle. However, for symmetric channels, the maximum of the mutual information will always occur for the uniform distribution of the input alphabet. This is particularly true for the binary channel. In this case the encoding is also binary and creates a uniform distribution of the codewords. The channel capacity does not only determine the design of a code, but more importantly it enables us to judge and to numerically evaluate a discrete channel or a modulation system.

3.1.4 Calculation of C for the BSC and the binary AWGN Channel

Example 3.2. Calculation of the channel capacity C for the BSC and the binary modulated baseband AWGN channel.

(1) BSC with the bit-error rate p_e. The maximum of $I(x;y)$ occurs because of the symmetry at $P_x(0) = P_x(1) = 0.5$. If x is uniformly distributed, then so is y with $P_y(0) = P_y(1) = 0.5$, so $H(y) = 1$. The equation

$$P(x,y) = P(x)P(y|x) = \left\{ \begin{array}{ll} (1-p_e)/2 & y = x \\ p_e/2 & y \neq x \end{array} \right\}$$

implies that

$$\begin{aligned} H(y|x) =& - P_{x,y}(0,0)\log_2 P_{y|x}(0|0) - P_{x,y}(0,1)\log_2 P_{y|x}(1|0) \\ & - P_{x,y}(1,0)\log_2 P_{y|x}(0|1) - P_{x,y}(1,1)\log_2 P_{y|x}(1|1) \\ =& - \frac{1-p_e}{2}\log_2(1-p_e) - \frac{p_e}{2}\log_2(p_e) \\ & - \frac{p_e}{2}\log_2(p_e) - \frac{1-p_e}{2}\log_2(1-p_e). \end{aligned}$$

Hence, the channel capacity of the BSC is

$$\begin{aligned} C &= 1 + p_e\log_2(p_e) + (1-p_e)\log_2(1-p_e) \\ &= 1 - H_2(p_e), \end{aligned} \qquad (3.1.18)$$

where $H_2(p_e)$ is the *binary entropy function* as defined in Appendix A.2. The channel capacity C is symmetrical with $C(p_e) = C(1-p_e)$ and we have $C(0) = 1$ and $C(0.5) = 0$. The capacity C is shown in Figure 3.2 together with the cutoff rate R_0 (see Definition 3.3).

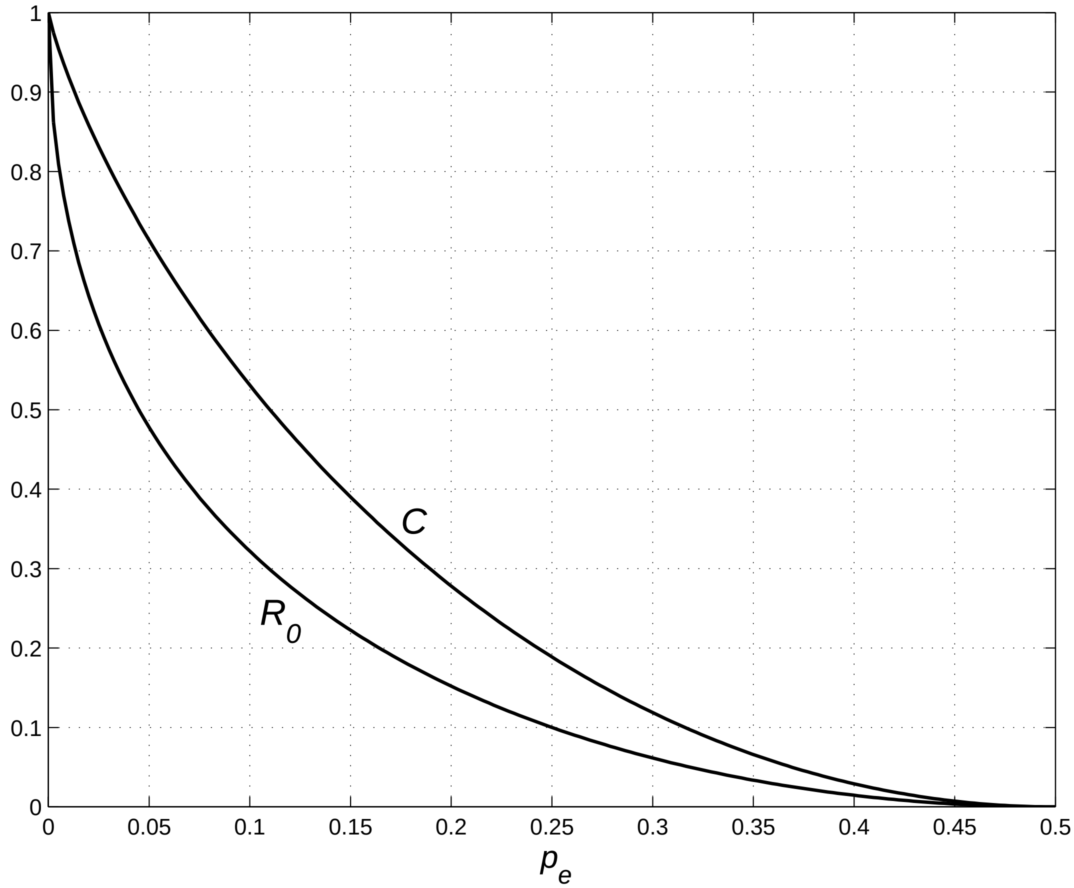


Figure 3.2. Channel capacity C and cutoff rate R_0 of the BSC

(2) AWGN channel. The maximum of $I(x;y)$ occurs because of the symmetry at $P_x(-\sqrt{E_c}) = P_x(+\sqrt{E_c}) = 0.5$. Then the probability density function (PDF) of y is given by

$$f_y(\eta) = \frac{1}{2}\left(f_{y|x}(\eta| - \sqrt{E_c}) + f_{y|x}(\eta| + \sqrt{E_c}) \right).$$

The sum over y in (3.1.10) is now transformed into an integral

$$C = \frac{1}{2}\int\limits_{-\infty}^{\infty} \left(f_{y|x}(\eta| + \sqrt{E_c}) \log_2 \frac{f_{y|x}(\eta| + \sqrt{E_c})}{f_y(\eta)} \right. \\ \left. + f_{y|x}(\eta| - \sqrt{E_c}) \log_2 \frac{f_{y|x}(\eta| - \sqrt{E_c})}{f_y(\eta)} \right) d\eta. \qquad (3.1.19)$$

Then with the PDF $f_{y|x}(\eta|\xi)$, as given in (1.3.11), this leads on to

$$C = \frac{1}{2\sqrt{\pi N_0}} \int\limits_{-\infty}^{\infty} \left(e^{-(\eta-\sqrt{E_c})^2/N_0} \log_2 \frac{2e^{-(\eta-\sqrt{E_c})^2/N_0}}{e^{-(\eta-\sqrt{E_c})^2/N_0} + e^{-(\eta+\sqrt{E_c})^2/N_0}} \right. \\ \left. + e^{-(\eta+\sqrt{E_c})^2/N_0} \log_2 \frac{2e^{-(\eta+\sqrt{E_c})^2/N_0}}{e^{-(\eta-\sqrt{E_c})^2/N_0} + e^{-(\eta+\sqrt{E_c})^2/N_0}} \right) d\eta.$$

The substitution $\alpha = \eta\sqrt{2/N_0}$ and the abbreviation $v = \sqrt{2E_c/N_0}$ simplify the result to

$$C = \frac{1}{2\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \left(e^{-(\alpha-v)^2/2} \log_2 \frac{2}{1+e^{-2\alpha v}} + e^{-(\alpha+v)^2/2} \log_2 \frac{2}{1+e^{2\alpha v}} \right) d\alpha. \qquad (3.1.20)$$

This integral can not be analytically calculated in a closed form. Numerical integration provides the curve C_{soft} of the channel capacity for the AWGN channel in Figure 3.3. The channel capacity increases continuously from 0 to 1, if E_c/N_0 runs from 0 to $+\infty$ (or from $-\infty$ to $+\infty$ if referring to decibels). For the R_0 curves see Subsection 3.2.5. ■

Although the channel models BSC and AWGN are of great practical and theoretical importance, there are also other relevant channels with asymmetric transition probabilities where the maximum of the mutual information does not occur for the uniform a priori distribution of the input alphabet.

3.2 Channel Coding Theorems

We start with an overview of the section's content. First of all, the original Shannon coding theorem for noisy channels is introduced and discussed in Subsection

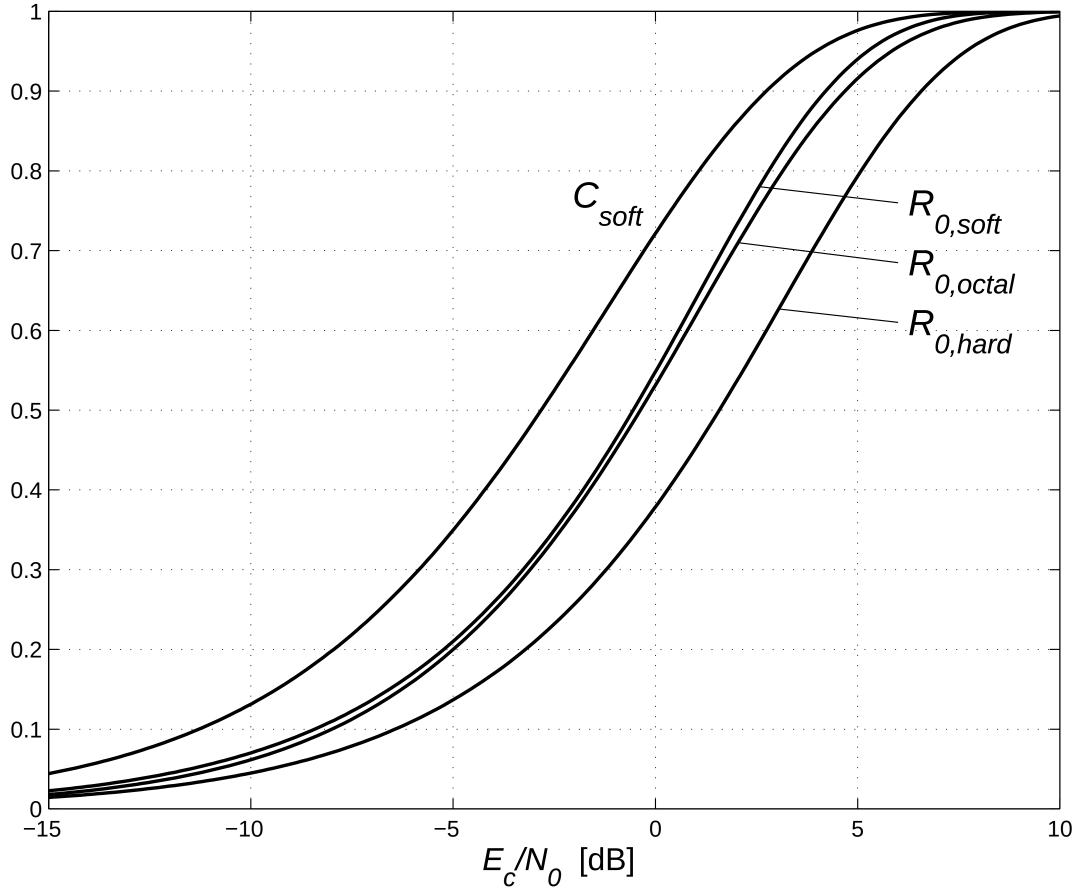


Figure 3.3. Channel capacity C and cutoff rate R_0 of the AWGN channel with binary input

3.2.1 for the abstract DC and in Subsection 3.2.2 in particular for an inner DC with high-level modulation. Furthermore, we will introduce an extension based on the error exponent in Subsection 3.2.3. More important in practice than the channel capacity C is the cutoff rate R_0 (also called R_0 criterion), as mentioned in Figures 3.2 and 3.3 and to be defined in Subsection 3.2.4, which is closely related to the Bhattacharyya bound as to be introduced in Subsection 3.2.5. After previously calculating C for the BSC and AWGN channel in Subsection 3.1.4, we will calculate R_0 correspondingly in Subsection 3.2.6.

3.2.1 Shannon's Noisy Channel Coding Theorem

The importance of the channel capacity becomes clear with the famous noisy channel coding theorem published by C.E.Shannon in the year 1948 (see [129, 130] for reprints of Shannon's classical papers), which forms the fundamental basis for digital communications.

Theorem 3.1 (Shannon's Noisy Channel Coding Theorem). *Let C be* *the channel capacity of the abstract DMC with the q-ary input alphabet $\mathcal{A}_{\text{in}}$. Then by using error-control coding and maximum-likelihood decoding, the word-error rate P_w can be made arbitrarily low, if the code rate $R_q = R \cdot \log_2 q$ is lower than C.*

A more precise statement is that for each $\varepsilon > 0$ and $\varepsilon' > 0$ there exists an $(n,k)_q$ block code with $R_q = k/n \cdot \log_2 q$ such that

$$C - \varepsilon' \le R_q < C \quad \textit{and} \quad P_w < \varepsilon.$$

The so-called converse to the coding theorem *is that for $R_q > C$, P_w can never fall below a certain bound.*

The proof is fairly simple for the special case of the BSC and is given in Section 3.6.

This result is certainly surprising: the channel properties only impose an upper bound on the transmission rate and on the throughput, but not on the quality of the transmission. So for higher quality requirements, we do not have to reduce the data rate, nor improve the channel, but merely increase the block length and thus the complexity. Important examples confirming these facts are shown in Figures 8.9 to 8.12 displaying the error probability of the BCH codes over E_b/N_0. Formally, the channel coding theorem can also be stated as: there exists a sequence of $(n_s, k_s)_q$ block codes $\mathcal{C}_s$ such that

$$\lim_{s\to\infty} P_w(\mathcal{C}_s) = 0 \quad \text{and} \quad \lim_{s\to\infty} \frac{k_s}{n_s} \cdot \log_2 q = C.$$

As already explained in Definition 1.4, $R_q = R \cdot \log_2 q$ is the code rate in units of information bits per q-ary input symbol of the abstract DC. Since the definition of the channel capacity is based on the binary logarithm and therefore refers to information bits per q-ary input symbol of the abstract DC, C has to be compared to R_q instead of R.

Example 3.3. We illustrate some applications of the channel coding theorem for the binary case, so $q = 2$ and $R_q = R$.

(1) Let C be the channel capacity of a binary DMC which can be used r_c times per second, hence the encoded bit rate is also r_c bit/s. Then we can choose a code rate R, which only has to be insignificantly lower than C, such that by using channel coding with an appropriately large block length, information bits can be transmitted at the rate of $r_b = R \cdot r_c$ for an arbitrarily low bit-error rate.

The channel capacity can also be expressed in information bits per second by $C^* = C \cdot r_c$, hence $R < C$ is equivalent to $r_b = R \cdot r_c < C \cdot r_c = C^*$ for the binary case (the general case with non-binary q will be examined closely in the next subsection).

(2) For example let $r_c = 1000$ encoded bit/s and $C = 0.60$ information bit/channel use, then $C^* = 600$ information bit/s. Thus almost $r_b = 600$ information bit/s can be transmitted with an arbitrarily low error rate for an appropriately large block length. Assume that a change in the modulation system creates another DMC with $r_c = 800$ encoded bit/s and $C = 0.75$ information bit/channel use, so that the same capacity $C^* = 600$ information bit/channel use results. Then r_b is again limited to 600 information bit/s. The questions of which DMC is actually more suitable and how C depends on r_c can generally not be answered definitively.

(3) Consider a BSC with $p_e = 0.01$ which can be used at $r_c = 1\ 000\ 000$ encoded bit/s. On average per second 990 000 bits are received correctly and 10 000 bits are wrong. Even information bit rates considerably lower than 900 000 bit/s can not be reliably transmitted without error-control coding. The channel capacity is

$$C = 1 + 0.01 \cdot \log_2 0.01 + 0.99 \cdot \log_2 0.99 = 0.919$$

information bit/channel use or $C^* = 919\ 000$ information bit/s. If $r_b = 900\ 000$ information bit/s is chosen with a code rate of $R = 0.9$, then by using coding, less than 1 error per second or an even lower error rate can be achieved. ■

The channel coding theorem is purely a theorem of existence and does not provide any instructions on how to construct the block codes. So Shannon did not create any clever codes, but chose block codes at random. Surprisingly, it is possible to prove the theorem for the average over of all block codes, and there is obviously at least one code which is as good as the mean. This technique is known as *random coding argument.* However, do not conclude that it is easy to find and to handle such codes. Actually, even now, half a century after Shannon's Theorem was published, no family of binary block codes is known (except for concatenated codes), where for a sequence of codes of increasing block length the error probability converges to zero. So we have to accept that *almost all codes are good except for the ones we already know.* The reason for this dilemma is that only a few codes of very large block length with their precise properties are really known in the algorithmic sense. Only codes with a very comprehensive and homogeneous mathematical structure are known or can be known in principle, but these only form a small subset of all codes. This subset hardly influences the mean of all block codes, thus the mean of all codes is far away from the properties of this subset.

A more extensive discussion of these observations is given, for example, in [155, 161] in a mathematical precise form. In Subsection 4.4.3 and in the following, random codes will be discussed again, however, under a slightly different aspect, i.e., referring to the minimum distance.

To conclude, the mathematical or algebraic structure of the codes is required for the analysis of the codes and their application in practice, and therefore forms the basis of error-control coding which will be discussed extensively in the following chapters. However, this structure prevents us from finding very powerful codes according to the channel coding theorem. In summary there are the following possibilities to improve the quality of the transmission:

(a) increase the channel capacity C by improving the channel (e.g., increase the transmit power or improve the link budget by larger antennas),

(b) reduce the code rate R to allow more redundancy, however, this requires an increase of discrete channel symbol rate and therefore an increase of the bandwidth,

(c) increase the block length n of the code, which is exactly the essence of the channel coding theorem.

3.2.2 Shannon's Noisy Channel Coding Theorem Restated for High-Level Modulation

In Subsections 1.3.3 and 1.4.2 we had explicitly distinguished between the inner discrete channel with 2^M-ary input and the abstract discrete channel with q-ary input, which is defined as the combination of the inner DC and the pair mapping-demapping. Based on this, we will state the Noisy Channel Coding Theorem once more and make clear which part q and M play.

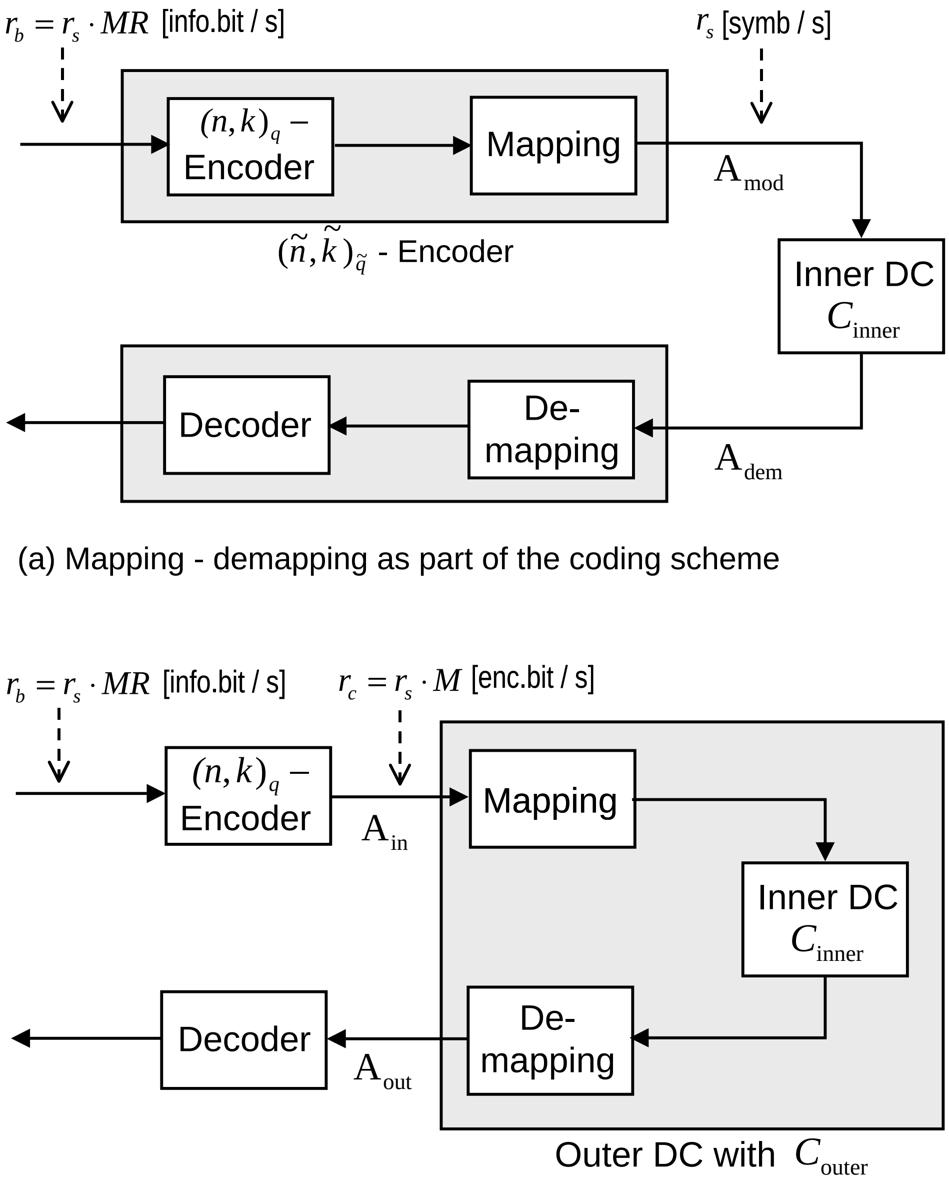
All previous considerations referred to the abstract DC, but for clarity, in this subsection, we use the term C_{abstract} to describe the corresponding channel capacity in units of information bits per q-ary input of the abstract DC. Correspondingly, C_{inner} in units of information bits per 2^M-ary input of the inner DC shall denote the channel capacity for the inner DC, which, in actual fact, has already been implicitly used in Subsection 3.1.4 for the example of the AWGN channel.

In Subsection 1.4.2, we considered the pair mapping-demapping as part of the coding scheme as well as part of the abstract channel, and we shall now discuss these two cases again but with the help of Figure 3.4. We start out from the inner DC with C_{inner} and the symbol rate $r_s = 1/T_s$ in units of 2^M-ary modulation symbols per second. We obtain an encoded symbol rate of $r_s \cdot M/\log_2 q$ and an encoded bit rate of $r_c = r_s \cdot M$ between encoder and mapping, since according to (1.2.6) the mapping causes an increase of the symbol rate by a factor of $(\log_2 q)/M$. Thus, the information bit rate is $r_b = r_c \cdot R = r_s \cdot RM$ in units of information bits per second.

The mapping as part of the coding scheme, as adopted in Figure 1.10 and in the upper half of Figure 3.4, leads to a $(\tilde{n}, \tilde{k})_{\tilde{q}}$ code, where $\tilde{n} = (n \log_2 q)/M$, $\tilde{k} = (k \log_2 q)/M$ and $\tilde{q} = 2^M$. According to Theorem 3.1, applied to the inner DC,

$$R_M = RM = \frac{\tilde{k}}{\tilde{n}} \cdot \log_2 \tilde{q} \overset{!}{<} C_{\text{inner}}. \tag{3.2.1}$$

The channel capacity C_{abstract} of the abstract DC can be obtained from C_{inner} by a simple consideration. Since the transition probabilities of the abstract DC for blocks of length n and of the inner DC for blocks of length $\tilde{n}$ are identical, $n \cdot C_{\text{abstract}} = \tilde{n} \cdot C_{\text{inner}}$ must be satisfied, thus $C_{\text{abstract}} = C_{\text{inner}} \cdot (\log_2 q)/M$.


$r_b = r_s \cdot MR$ [info.bit / s]
r_s [symb / s]
$(n,k)_q -$ Encoder
Mapping
A_{mod}
$(\tilde{n},\tilde{k})_{\tilde{q}}$ - Encoder
Inner DC C_{inner}
Decoder
De-mapping
A_{dem}
$r_b = r_s \cdot MR$ [info.bit / s]
$r_c = r_s \cdot M$ [enc.bit / s]
$(n,k)_q -$ Encoder
A_{in}
Mapping
Inner DC C_{inner}
Decoder
A_{out}
De-mapping
Outer DC with C_{outer}

(a) Mapping - demapping as part of the coding scheme

(b) Mapping - demapping as part of the abstract DC

Figure 3.4. Capacities of abstract and inner discrete channel

According to Theorem 3.1, applied to the abstract DC, as shown in the lower half of Figure 3.4,

$$R_q = \frac{k}{n} \cdot \log_2 q \overset{!}{<} C_{\text{abstract}} = C_{\text{inner}} \cdot \frac{\log_2 q}{M}. \tag{3.2.2}$$

Since this is simply equivalent to (3.2.1), it turns out that both approaches lead to the same result. The term C^* is used to describe a different form of the channel capacity, which refers to information bits per second, and is defined as the product of the channel capacity in reference to symbols and the symbol rate:

$$C^* = C_{\text{abstract}} \cdot \frac{r_s M}{\log_2 q} = C_{\text{inner}} \cdot r_s \quad \text{in units of} \quad \left[\frac{\text{information bit}}{\text{second}}\right]. \tag{3.2.3}$$

The two conditions $R_M < C_{\text{inner}}$ and $R_q < C_{\text{abstract}}$ are equivalent to each other as well as to

$$R_M = RM < C_{\text{inner}}, \quad \text{or equivalently} \quad r_b < C^*. \tag{3.2.4}$$

Therefore the channel capacity C^* turns out to be an upper bound for the throughput r_b, if an arbitrarily small bit-error rate is to be achieved by channel coding. Obviously, the channel capacity can be solely described by the inner DC, since the mapping and the value of q are irrelevant.

A simple, direct trade-off between R and M seems to be possible at first glance, because C_{inner} is an upper bound for the product $R_M = RM$ and C_{inner} depends on the symbol rate (and therefore on the bandwidth of the waveform channel), which again only depends on $R_M = RM$ because of $r_s = r_b/(RM)$. In Section 3.4, we will examine this in detail for the example of the AWGN channel and practical modulation schemes, because this builds the foundation for trellis coded modulation (TCM), discussed in Chapter 11. We will realize that factoring a given value of R_M as

$$R = \frac{R_M}{R_M + 1} \quad \text{and} \quad M = R_M + 1 \tag{3.2.5}$$

delivers a reasonable result. For example, for $R_M = 2$, $R = 2/3$ and $M = 3$ is a sensible choice, whereas $R = 2/4$ and $M = 4$ is of no advantage. The results and curves for C and R_0 in Section 3.4 will help to make this become clear.

3.2.3 The Error Exponent

Theorem 3.1 is unsatisfactory in that the required block length can not be upper bounded. However, there is the following refinement [19, 42, 139]:

Theorem 3.2 (Channel Coding Theorem Based on Error Exponent). *Let C be the channel capacity of the abstract DMC with the q-ary input alphabet $|\mathcal{A}_{\text{in}}|$. Furthermore, let*

$$E_r(R_q) = \max_{0 \le s \le 1} \max_{P_x} \left[-sR_q - \log_2 \sum_{y \in \mathcal{A}_{\text{out}}} \left(\sum_{x \in \mathcal{A}_{\text{in}}} P(x) \cdot P(y|x)^{\frac{1}{1+s}} \right)^{1+s} \right] \tag{3.2.6}$$

be the error exponent *(also called* Gallager exponent*), which is solely defined by the DMC and the code rate. The behaviour of this function is described by*

$$\begin{aligned} E_r(R_q) &> 0 \quad \textit{for } R_q < C \\ E_r(R_q) &= 0 \quad \textit{for } R_q \ge C. \end{aligned} \tag{3.2.7}$$

Then there always exists an $(n,k)_q$ *block code with* $R_q = k/n \cdot \log_2 q < C$, *such that*

$$P_w < 2^{-n \cdot E_r(R_q)} \tag{3.2.8}$$

for the word-error probability P_w.

In (3.2.8), the word-error probability P_w, the block length n, the code rate R_q as well as the channel properties as represented by the error exponent $E_r(R_q)$, i.e., all important parameters of coded digital transmission, are condensed in one simple formula.

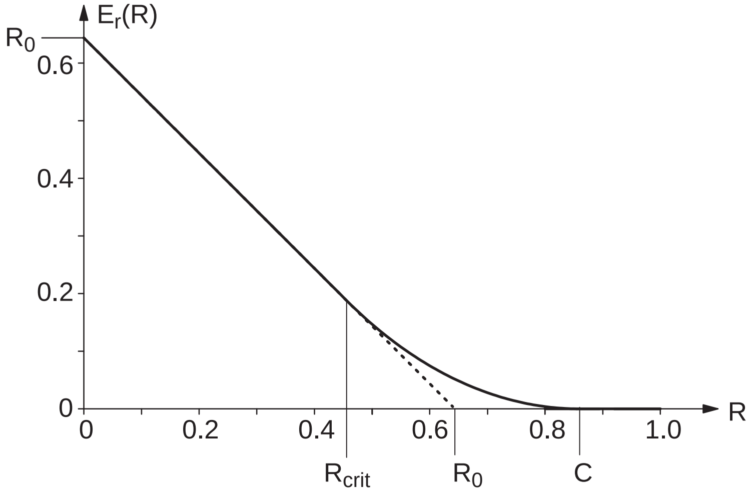


Figure 3.5. Typical behaviour of the error exponent (demonstrated by the BSC for $p_e = 0.02$, $C = 0.859$, $R_0 = 0.644$)

The error exponent is not a parameter only defined by the DMC, as is the channel capacity, but it is a function of both the code rate and the DMC. If the function $E_r(R_q)$ is given, then the required block length n can be determined explicitly for a given P_w. The typical curve of the error exponent is shown in Figure 3.5. We will not prove that $E_r(R_q)$ is a ∪-convex monotonic decreasing function. From $R_q = 0$ to a point $R_q = R_{\text{crit}}$ the gradient is -1 and the maximum is attained at $s = 1$. The value of the error exponent at $R_q = 0$ is important and therefore will be discussed separately in the following section.

3.2.4 The R_0 Theorem

The channel capacity C is a theoretical bound which the realizable codes or the codes used in practice are far from reaching. However, the cutoff rate R_0 [186, 204], defined below, is achievable with an acceptable amount of effort. This is not meant as a precise mathematical statement, but will become obvious from examples discussed in Section 12.?.

We will not prove that the maximum for $E_r(0)$ is attained at $s = 1$.

Definition 3.3. *The value of the error exponent at $R_q = 0$ is denoted R_0 and is called* cutoff rate *or R_0* criterion*:*

$$R_0 = E_r(0) = \max_{P_x} \left[-\log_2 \sum_{y \in \mathcal{A}_{\text{out}}} \left(\sum_{x \in \mathcal{A}_{\text{in}}} P(x) \sqrt{P(y|x)} \right)^2 \right]. \tag{3.2.9}$$

A third term often used for R_0 is *computational cutoff rate* or R_{comp}, however, this only makes sense in connection with the theory of sequential decoding which is not discussed here. If in the definition of the error exponent $s = 1$ is set, we obtain the lower bound (see also Figure 3.5)

$$\begin{aligned} E_r(R_q) &\geq \max_{P_x} \left[-R_q - \log_2 \sum_{y \in \mathcal{A}_{\text{out}}} \left(\sum_{x \in \mathcal{A}_{\text{in}}} P(x) \cdot P(y|x)^{\frac{1}{2}} \right)^2 \right] \\ &= R_0 - R_q. \end{aligned}$$

From (3.2.8) we obtain the following explicit estimate for the word-error probability P_w for code rates R_q between 0 and R_0:

Theorem 3.3 (*R_0* Theorem). *For an abstract DMC with q-ary input and with the cutoff rate R_0, there always exists an $(n, k)_q$ block code with the code rate $R_q = k/n \cdot \log_2 q < R_0$ such that when using maximum-likelihood decoding,*

$$P_w < 2^{-n(R_0 - R_q)} \tag{3.2.10}$$

for the word-error probability P_w. Similarly as for the channel coding theorem, a code rate R_q arbitrarily close to R_0 can be achieved.

The direct proof of the R_0 Theorem, without using Theorem 3.2, is fairly simple and will be given in Section 3.7 for the general DMC. However, we will use the random coding argument again, as we did for the channel coding theorem, so the proof will not provide any construction rules for good codes. So the closer R_q is to R_0, the larger the block length n has to be, to be able to guarantee the same error rate.

The different bounds on the word-error probability P_w depending on the various ranges of the code rate are listed below:

$0 < R_q < R_0$: P_w is explicitly bounded by the block length n and the difference $R_0 - R_q$. This bound can be numerically calculated.

$R_0 < R_q < C$: P_w is theoretically bounded by the block length n and the function $E_r(R_q)$. The bound can hardly be calculated except for some simple cases.

$C < R_q$: P_w can not become arbitrarily low, and a lower bound exists and can be calculated.

The upper bound for P_w is, of course, only valid for an optimally chosen code according to the R_0 Theorem or the channel coding theorem. For the codes applied in practice, which are designed with specific structures to save processing effort for encoding und decoding, P_w might be much larger. Other than for Theorem 3.3, the cutoff rate proves to be useful for evaluating a modulation scheme for channel coding, while other criteria are not suitable:

- the channel capacity is not suitable for evaluating since it is only a theoretical bound which requires codes with a large block length (and thus an accordingly long delay) and high complexity.
- the error probability is also not suitable since the quantization in hard-decision demodulators discards certain information which might have greatly improved the decoding. This loss of information is not captured by the error probability.

However, R_0 can be helpful for the following points. The error probability can be bounded for optimally chosen codes (see the R_0 Theorem 3.3) as well as for given codes (see Theorem 4.16 on the union bound). Furthermore, trade-offs between the block length, the code rate, the number of levels of the modulation scheme, the signal-to-noise ratio and the error probability can be calculated. Finally, R_0 enables us to determine the gain from soft-decision over hard-decision decoding as well as the design of the quantization operation for optimal soft decisions, this will become clear in Subsection 3.2.5.

3.2.5 The Bhattacharyya Bound

Assume a symmetric DMC with binary input $\mathcal{A}_{\text{in}} = \{0, 1\}$, then the defining expression (3.2.9) for R_0 can be greatly simplified to

$$\begin{aligned}
R_0 &= -\log_2 \left[\sum_{\eta \in \mathcal{A}_{\text{out}}} \left(\sum_{\xi \in \mathcal{A}_{\text{in}}} P_x(\xi) \sqrt{P_{y|x}(\eta|\xi)} \right)^2 \right] \\
&= -\log_2 \left[\frac{1}{4} \sum_{\eta \in \mathcal{A}_{\text{out}}} \left(\sqrt{P_{y|x}(\eta|0)} + \sqrt{P_{y|x}(\eta|1)} \right)^2 \right] \\
&= -\log_2 \left[\frac{1}{4} \sum_{\eta \in \mathcal{A}_{\text{out}}} P_{y|x}(\eta|0) + \frac{1}{4} \sum_{\eta \in \mathcal{A}_{\text{out}}} P_{y|x}(\eta|1) \right. \\
&\qquad \left. + \frac{1}{2} \sum_{\eta \in \mathcal{A}_{\text{out}}} \sqrt{P_{y|x}(\eta|0) P_{y|x}(\eta|1)} \right] \\
&= 1 - \log_2 \left[1 + \sum_{\eta \in \mathcal{A}_{\text{out}}} \sqrt{P_{y|x}(\eta|0) P_{y|x}(\eta|1)} \right] .
\end{aligned}$$

This leads to the following definition.

Definition 3.4. *For the symmetric DMC with binary input* $\mathcal{A}_{\text{in}} = \{0,1\}$ *the* Bhattacharyya bound *is defined as*

$$\gamma = \sum_{\eta \in \mathcal{A}_{\text{out}}} \sqrt{P_{y|x}(\eta|0) P_{y|x}(\eta|1)}. \tag{3.2.11}$$

The relation to R_0 *is*

$$R_0 = 1 - \log_2(1+\gamma) \quad , \quad \gamma = 2^{1-R_0} - 1. \tag{3.2.12}$$

It is obvious that $\gamma \geq 0$ and with *Schwarz's inequality* (A.1.9)

$$\begin{aligned} \gamma^2 &= \left(\sum_{\eta \in \mathcal{A}_{\text{out}}} \sqrt{P_{y|x}(\eta|0) P_{y|x}(\eta|1)} \right)^2 \\ &\leq \sum_{\eta \in \mathcal{A}_{\text{out}}} \sqrt{P_{y|x}(\eta|0)}^2 \cdot \sum_{\eta \in \mathcal{A}_{\text{out}}} \sqrt{P_{y|x}(\eta|1)}^2 = 1. \end{aligned}$$

Thus $0 \leq \gamma \leq 1$, so $0 \leq R_0 \leq 1$. The best case is $\gamma = 0, R_0 = 1$, and the worst case is $\gamma = 1, R_0 = 0$.

In Subsection 4.7.2 we will see that the error probability of a block code is characterized by a combination of code properties and channel properties. The channel properties are represented by γ. Therefore γ forms the basis for the calculation of P_w.

3.2.6 Calculation of R_0 for the BSC and the binary AWGN Channel

Example 3.4. Calculation of the Bhattacharyya bound γ and the cutoff rate R_0 for the BSC and the binary modulated baseband AWGN channel.

(1) BSC with the bit-error rate p_e. From (3.2.11) we can derive

$$\begin{aligned} \gamma &= \sqrt{P_{y|x}(0|0) P_{y|x}(0|1)} + \sqrt{P_{y|x}(1|0) P_{y|x}(1|1)} \\ &= \sqrt{(1-p_e)p_e} + \sqrt{p_e(1-p_e)} \\ &= \sqrt{4p_e(1-p_e)}, \end{aligned} \tag{3.2.13}$$

thus

$$R_0 = 1 - \log_2\left(1 + \sqrt{4p_e(1-p_e)}\right). \tag{3.2.14}$$

This result is shown in Figure 3.2 together with the channel capacity.

(2) AWGN channel. The summation in (3.2.11) is transformed into an integration:

$$\begin{aligned}\gamma &= \int\limits_{-\infty}^{\infty} \sqrt{f_{y|x}(\eta|\sqrt{E_c})f_{y|x}(\eta|-\sqrt{E_c})}\, d\eta \\ &= \int\limits_{-\infty}^{\infty} \sqrt{\frac{e^{-(\eta-\sqrt{E_c})^2/N_0}}{\sqrt{\pi N_0}} \cdot \frac{e^{-(\eta+\sqrt{E_c})^2/N_0}}{\sqrt{\pi N_0}}}\, d\eta \\ &= \frac{1}{\sqrt{\pi N_0}} \int\limits_{-\infty}^{\infty} e^{-(\eta^2+E_c)/N_0}\, d\eta \\ &= e^{-E_c/N_0} \cdot \frac{1}{\sqrt{\pi N_0}} \int\limits_{-\infty}^{\infty} e^{-\eta^2/N_0}\, d\eta \\ &= e^{-E_c/N_0}. \end{aligned} \tag{3.2.15}$$

This implies that

$$R_0 = 1 - \log_2\left(1 + e^{-E_c/N_0}\right) \tag{3.2.16}$$

which is shown in Figure 3.3, labeled $R_{0,\text{soft}}$, together with the channel capacity. For binary quantization at the output of the DMC we obtain a BSC with $p_e = Q(\sqrt{2E_c/N_0})$ and the curve $R_{0,\text{hard}}$. Obviously hard decisions imply a loss of about 2 dB over soft decisions in reference to R_0. So if the demodulator only provides the sign instead of the continuous-valued signal, then this loss of information has to be compensated for by increasing the transmit power by 2 dB. Of course, the distance of 3 dB for the asymptotic coding gain according to (1.7.13) only occurs as $E_c/N_0 \to \infty$ and only in reference to P_w.

Additionally, in Figure 3.3, R_0 is shown in the case of octal quantization as in Figures 1.6 and 1.7. It can be seen that in reference to R_0 the 3-bit quantization only causes a very small loss of information and is almost as good as ideal soft decisions. ■

3.3 Capacity Limits and Coding Gains for the Binary AWGN Channel

For the AWGN channel with baseband signaling and binary input, the channel capacity C and the cutoff rate R_0 have already been calculated in Examples 3.2(2) and 3.4(2) and are shown in Figure 3.3. In Subsection 3.3.1, we analyze the relation between E_b/N_0 and the code rate R for $R = C$ or $R = R_0$, assuming the error probability is arbitrarily close to zero. However, if we allow a certain value of the error probability, then we obtain a larger coding gain as examined in Subsection 3.3.2.

3.3.1 The Coding Gain as a Function of the Code Rate

In the following, we will derive a relation between the code rate R and the required E_b/N_0 for $R = R_0$ and $R = C$ for both soft and hard decisions, respectively. The result is four curves with E_b/N_0 as a function of R, shown in Figure 3.6. In all cases $E_b/N_0 \to \infty$ as $R \to 1$. Particularly the limits for $R \to 0$ are of great interest. For the relation between the energy per encoded bit and the energy per information bit, we generally have $E_c = RE_b$ according to (1.7.3).

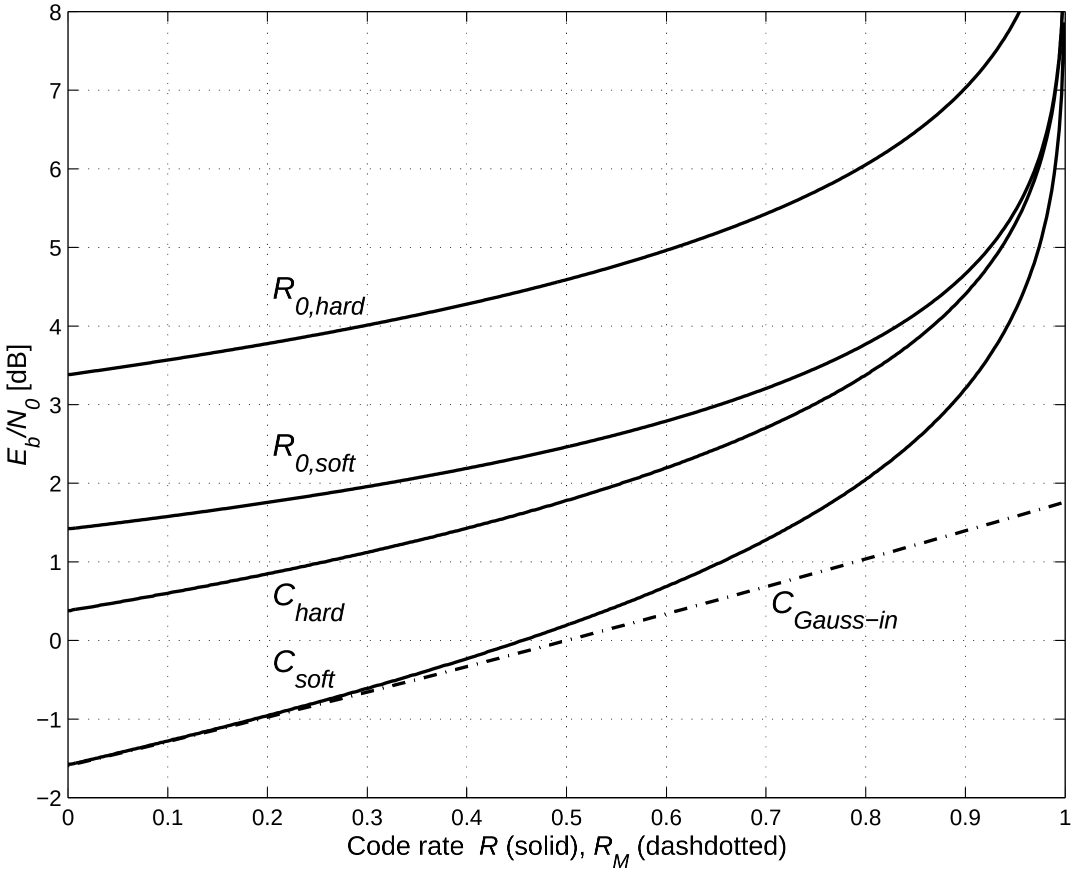


Figure 3.6. Required E_b/N_0 to achieve $R = R_0$ or $R = C$ for the AWGN channel with binary input (solid lines) and non-binary input (dashdotted lines)

Case 1. First $R = R_0$ is assumed to be the maximum possible code rate in *practice.* For soft decisions, according to (3.2.16),

$$R = R_0 = 1 - \log_2\left(1 + e^{-R \cdot E_b/N_0}\right). \tag{3.3.1}$$

This equation relates R to E_b/N_0 with the solution

$$\frac{E_b}{N_0} = -\frac{\ln(2^{1-R} - 1)}{R}. \tag{3.3.2}$$

The resulting curve labeled $R_{0,\text{soft}}$ is shown in Figure 3.6. As $R \to 1$, $E_b/N_0 \to \infty$. As $R \to 0$, (3.3.1) only provides a trivial statement. In (3.3.2) the quotient has the form 0/0, thus we can apply l'Hôspital's rule (A.1.3):

$$\lim_{R\to 0} \frac{E_b}{N_0} = \lim_{R\to 0} -\frac{\dfrac{1}{2^{1-R}-1} \cdot 2^{1-R} \ln(2) \cdot (-1)}{1}$$

$$= 2 \cdot \ln 2 \cong 1.42 \text{ dB} \qquad (\text{curve } R_{0,\text{soft}}). \tag{3.3.3}$$

This limit can also be taken from Figure 3.6. As a conclusion, for E_b/N_0 smaller than 1.42 dB a transmission with $R = R_0$ is impossible. An explanation could be given as follows: a very low code rate R requires a very large bandwidth (this is quite unrealistic) and therefore there is very little energy per encoded bit in contrast to the noise power spectral density. Then every single encoded bit is mostly superimposed by noise, and R_0 is very low. After a certain limit, R_0 is lower than R so that $R = R_0$ can no longer be achieved.

Case 2. The curve $R_{0,\text{hard}}$ in Figure 3.6 corresponds to hard decisions with binary quantization and is obtained according to (3.2.14) by

$$R = R_0 = 1 - \log_2\left(1 + \sqrt{4p_e(1-p_e)}\right) \quad \text{with} \quad p_e = Q\left(\sqrt{\frac{2RE_b}{N_0}}\right). \tag{3.3.4}$$

The distance between soft and hard decisions for $R = R_0$ is about 2 dB throughout the whole range as shown in Figure 3.3. This again emphasizes the meaning of soft-decision demodulation. As $R \to 0$, (A.4.23) at first implies that $2p_e \approx 1 - \sqrt{4RE_b/(\pi N_0)}$ and $2(1-p_e) \approx 1 + \sqrt{4RE_b/(\pi N_0)}$, thus

$$\begin{aligned} R = R_0 &\approx 1 - \log_2\left(1 + \sqrt{\left(1 - \sqrt{\frac{4RE_b}{\pi N_0}}\right)\left(1 + \sqrt{\frac{4RE_b}{\pi N_0}}\right)}\right) \\ &= 1 - \log_2\left(1 + \sqrt{1 - \frac{4RE_b}{\pi N_0}}\right) \\ &\approx 1 - \log_2\left(2 - \frac{2RE_b}{\pi N_0}\right) \quad \text{according to (A.1.8)} \\ &\approx \frac{R}{\pi \ln 2} \cdot \frac{E_b}{N_0} \quad \text{according to (A.1.6).} \end{aligned}$$

Therefore

$$\lim_{R\to 0} \frac{E_b}{N_0} = \pi \ln 2 \cong 3.38 \text{ dB} \qquad (\text{curve } R_{0,\text{hard}}). \tag{3.3.5}$$

Case 3. Next, $R = C$ is presupposed as the *theoretically* maximum possible code rate. The capacity for soft decisions is the same as in (3.1.20) with $v = \sqrt{2RE_b/N_0}$. The resulting curve is labeled C_{soft} in Figure 3.6. The calculation of the limit as $R \to 0$ is fairly time-consuming, since using linear approximations would be too inaccurate. The integral in (3.1.20) is at first written as $C = \int f(\alpha, v) d\alpha$. As $R = C \to 0$,

$$\begin{aligned} 1 &= \lim_{R\to 0} \frac{1}{R} \int\limits_{-\infty}^{\infty} f(\alpha, v) d\alpha \\ &= \lim_{R\to 0} \frac{d}{dR} \int\limits_{-\infty}^{\infty} f(\alpha, v) d\alpha \quad \text{with (A.1.3)} \\ &= \lim_{v\to 0} \frac{E_b}{N_0} \cdot \int\limits_{-\infty}^{\infty} \frac{\frac{d}{dv} f(\alpha, v)}{v} d\alpha \quad \text{since} \quad \frac{dv}{dR} = \frac{E_b}{N_0} \frac{1}{v}. \end{aligned}$$

A lengthy calculation leads to the *Shannon limit*

$$\lim_{R\to 0} \frac{E_b}{N_0} = \ln 2 \cong -1.59 \text{ dB} \qquad (\text{curve } C_{\text{soft}}), \tag{3.3.6}$$

i.e., for E_b/N_0 smaller than -1.59 dB a transmission with $R = C$ is not possible.

Case 4. The curve C_{hard} in Figure 3.6 corresponds to hard decisions with binary quantization and is taken from (3.1.18):

$$\begin{aligned} R = C &= 1 - H_2(p_e) \qquad\qquad (3.3.7) \\ &= 1 + p_e \log_2 p_e + (1 - p_e) \log_2 (1 - p_e) \quad \text{with} \quad p_e = Q\left(\sqrt{\frac{2RE_b}{N_0}}\right). \end{aligned}$$

For the limit as $R \to 0$,

$$\begin{aligned} R = C &= 1 - H_2\left(Q\left(\sqrt{2R\frac{E_b}{N_0}}\right)\right) \\ &\approx 1 - H_2\left(\frac{1}{2} - \sqrt{\frac{RE_b}{\pi N_0}}\right) \quad \text{according to (A.4.23)} \\ &\approx \frac{2}{\ln 2} \cdot \frac{RE_b}{\pi N_0} \quad \text{according to (A.2.5).} \end{aligned}$$

Thus

$$\lim_{R\to 0} \frac{E_b}{N_0} = \frac{\pi \ln 2}{2} \approx 1.09 \cong 0.37 \text{ dB} \qquad (\text{curve } C_{\text{hard}}). \tag{3.3.8}$$

The distance between C_{soft} and C_{hard} does not remain more or less constant, as it does for the corresponding R_0 curves, but decreases from about 2 dB at $R = 0$ to about 1 dB at $R = 1$. As $R \to 1$, the curves C_{hard} and $R_{0,\text{soft}}$ draw closer together.

However, the most important implication of Figure 3.6 is that it hardly makes sense to use code rates lower than 1/2 in practice, since the gain of the transition from $R = 1/2$ to $R \to 0$ is only a maximum of about 1 dB in reference to R_0. However, recall that these observations are based on the presumption of limited transmit power and unlimited bandwidth.

The fifth graph in Figure 3.6 labeled $C_{\text{Gauss-in}}$ refers to the AWGN channel with non-binary input. We will see in Subsection 3.4.1 that the optimum for non-binary inputs is given by a continuous-valued input signal with Gaussian distribution. The exact details will be defined in Theorem 3.4.

Example 3.5. For a bit-error probability of $P_b = 10^{-5}$ without coding, a channel with $E_b/N_0 = 9.59$ dB is required according to Table 1.1. If a rate-1/2 code is used, then for $R = R_0$ an arbitrarily low error rate can be achieved for 2.45 dB, and as $R \to C$ further 2 dB can be saved. Thus the coding gain is 7.14 dB at $P_b = 10^{-5}$. The coding gain can be increased by a lower code rate R (or by lower error rates). However, in practice these coding gains can only be achieved by using very complex codes. ■

3.3.2 The Coding Gain as a Function of the Bit-Error Rate

When we derived the coding gains in the previous subsections, we assumed a required error probability arbitrarily close to zero. If we are willing to tolerate a certain given value of the error probability, we can obtain a larger coding gain as shown in Figure 3.7. It follows from Shannon's rate distortion theory [4, 19, 27], that if we tolerate an error rate P_b, then a k-tuple of source bits can be shortened to a k'-tuple

$$k' = k \cdot (1 - H_2(P_b)) = 1 + P_b \log_2 P_b + (1 - P_b) \log_2(1 - P_b), \tag{3.3.9}$$

where $H_2(P_b)$ denotes the binary entropy function and recalling that $1 - H_2(P_b) = C_{\text{hard}}(P_b)$ according to (3.1.18). First, the k source bits are compressed to k' bits and are then expanded to n encoded bits with error-control coding. So for $R = k/n$ and $R' = k'/n$,

$$R \cdot (1 - H_2(P_b)) = R' = C_{\text{soft}} \left(\text{at } \frac{E_c}{N_0} = R' \cdot \frac{E_b}{N_0} \right). \tag{3.3.10}$$

So P_b and R lead to E_b/N_0 being

$$\frac{E_b}{N_0} = \frac{C_{\text{soft}}^{-1}(R(1 - H_2(P_b)))}{R(1 - H_2(P_b))}. \tag{3.3.11}$$

In Figure 3.7, these coding limits are illustrated for some values of R by a graph of P_b over E_b/N_0. A similar plot of graphs can be found in [157]. For small P_b, the graphs are almost vertical and the value of E_b/N_0 in Figure 3.7 is equal to the values of the curve C_{soft} in Figure 3.6. So Figure 3.7 only yields additional

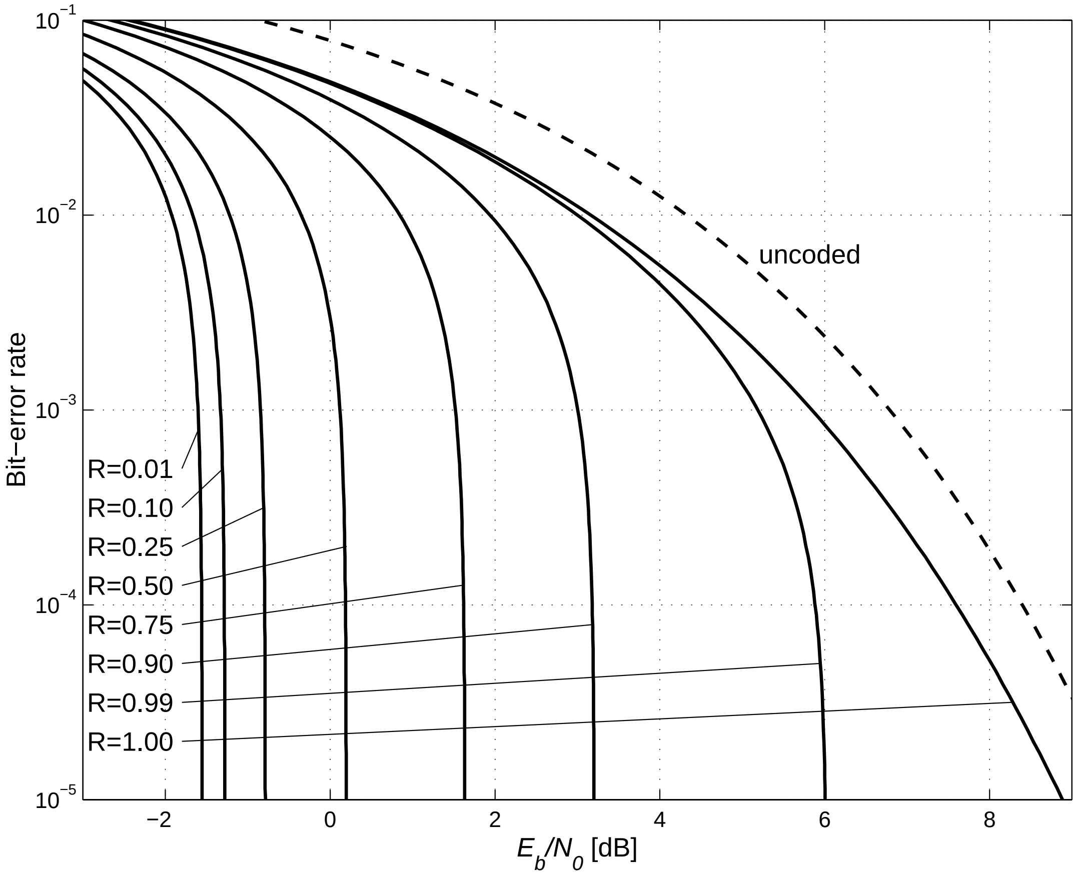


Figure 3.7. Required E_b/N_0 to achieve a bit-error rate P_b under the condition of $R = C_{\mathrm{soft}}$ for the AWGN channel with binary input

information above $P_b > 0.001$. The Shannon limit (3.3.6) can be found at $R = 0.01$.

The maximum possible coding gain (presupposing $R = C_{\mathrm{soft}}$, i.e. for extremely high coding complexity) for a particular given bit-error rate is the horizontal gap to the uncoded BPSK curve. For example, according to Figure 10.10, a rate-1/2 convolutional code with $m = 6$ (we refer to Chapters 9 and 10 for details) requires $E_b/N_0 = 1.6$ dB for a bit-error rate of 0.01, whereas according to Figure 3.7, -0.4 dB would suffice (however, requiring the aforementioned, enormous effort), so the gap to the theoretical Shannon boundary is 2 dB.

3.4 C and R_0 for AWGN Channels with High-Level Modulation

So far we have only considered the discrete-time AWGN channel with baseband signaling and binary input $\mathcal{A}_{\mathrm{mod}} = \{+\sqrt{E_c}, -\sqrt{E_c}\}$, so a very simple modulation system with only two different signals was presupposed. Now, we will ignore this restriction to be able to use the waveform channel in the best possible way. In the following subsections, we consider the AWGN channel with continuous-valued Gaussian-distributed input (as the theoretical optimum) as well as with ASK-, PSK- and QAM-distributed input (as the most important practical modulation schemes), where the output is always assumed with soft decisions. Occasionally, we have to take into account the differences between baseband and passband AWGN channels. In Section 3.5 we will extend the AWGN channel to a continuous-time model.

3.4.1 Gaussian-Distributed Continuous-Valued Input

Before considering high-level modulation schemes in the following subsections and comparing these to the binary modulation used so far, we will first consider the extreme case of the modulation alphabet comprising the whole set of real numbers, $\mathcal{A}_{\text{mod}} = \mathbb{R}$, thus $M = \infty$ formally. This means that continuous-valued input signals to the inner discrete-time channel will be considered, the objective being the joint optimization of coding and modulation, i.e., the optimization of the whole transmitter and receiver in Figure 1.1. This will enable us to determine by how much a high-level modulation scheme deviates from the best case of an extremely high-level modulation with an optimum a priori probability distribution of the individual symbol levels.

Furthermore, we will presume the discrete-time baseband AWGN channel as given in Definition 1.3, however, this time with continuous-valued input x. For an arbitrary input x the output $y = x + \nu$ is always continuous-valued because of the continuous-valued noise ν. For the transition from the discrete-valued to the continuous-valued system the channel capacity is still defined as

$$C = \max_{P_x} \Big(H(y) - H(y|x) \Big), \tag{3.4.1}$$

where the *differential entropy* of a continuous-valued random variable y with the probability density function f_y is defined as

$$H(y) = -\int_{-\infty}^{\infty} f_y(\eta) \cdot \log_2 f_y(\eta)\, d\eta. \tag{3.4.2}$$

However, $H(y)$ can not be interpreted as in the case of discrete values, since a continuous-valued random variable attains infinitely many values and therefore might have an infinite entropy. The function $H(y)$ may even attain negative values. Yet, C can be defined as in (3.4.1), see [19] for a more detailed discussion. We will not prove that the maximum of the channel capacity is achieved by a Gaussian distributed input x. Recall that according to Subsection 1.6.1 for a q-ary discrete-valued input a uniform distribution was presupposed.

Since the input x and the noise ν are both Gaussian distributed then so is the output $y = x + \nu$, thus an analytically closed calculation of C is possible.

The expected values of x, ν and y are all zero. The parameter $E_s = E(x^2) = \sigma_x^2$ represents the *energy per encoded modulation symbol.* The variance of $y = x + \nu$ is

$$\sigma_y^2 = E(y^2) = E(x^2) + E(\nu^2) = E_s + N_0/2,$$

since the transmit signal and the noise are statistically independent. Therefore the entropy of y is

$$\begin{aligned} H(y) &= -\int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\eta^2}{2\sigma_y^2}\right) \cdot \log_2\left(\frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\eta^2}{2\sigma_y^2}\right)\right) d\eta \\ &= -\log_2\left(\frac{1}{\sqrt{2\pi\sigma_y^2}}\right) \cdot \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\eta^2}{2\sigma_y^2}\right) d\eta \\ &\quad - \log_2(e) \cdot \left(-\frac{1}{2\sigma_y^2}\right) \cdot \int\limits_{-\infty}^{\infty} \frac{\eta^2}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\eta^2}{2\sigma_y^2}\right) d\eta. \end{aligned}$$

The first part of the integral covers the probability density function and delivers the value 1. The second integral gives us the variance σ_y^2, thus

$$H(y) = -\log_2\left(\frac{1}{\sqrt{2\pi\sigma_y^2}}\right) + \frac{1}{2}\log_2(e) = \frac{1}{2}\log_2(2\pi e\sigma_y^2),$$

and therefore

$$H(y|x) = \frac{1}{2}\log_2\left(2\pi e\frac{N_0}{2}\right).$$

The channel capacity is the difference $H(y) - H(y|x) = \dfrac{1}{2}\log_2\left(\dfrac{2\sigma_y^2}{N_0}\right)$ and thus we have proved the following theorem on the channel capacity:

Theorem 3.4. *For the discrete-time baseband AWGN channel with Gaussian-distributed continuous-valued input, the channel capacity is*

$$C_{1-\mathrm{dim}} = \frac{1}{2} \cdot \log_2\left(1 + \frac{2E_s}{N_0}\right) = \frac{1}{2} \cdot \log_2\left(1 + \frac{E(x^2)}{E(\nu^2)}\right) \qquad (3.4.3)$$

in units of information bits per discrete channel use. For the transition from the baseband to the passband AWGN channel both the noise power, according to (2.2.28), as well as the capacity are doubled:

$$C_{2-\mathrm{dim}}\left(\frac{E_s}{N_0}\right) = 2 \cdot C_{1-\mathrm{dim}}\left(\frac{E_s}{2N_0}\right) = \log_2\left(1 + \frac{E_s}{N_0}\right). \qquad (3.4.4)$$

The capacity bounds from Theorem 3.4 are sketched in Figures 3.7 and 3.9. As in Section 3.3, we will again consider the case of $R_M = C$ as the theoretically maximum possible code rate. Instead of $E_c = RE_b$, we will use $E_s = R_M \cdot E_b$. Theorem 3.4 implies that

$$R_M = \frac{1}{2}\log_2\left(1 + 2R_M\frac{E_b}{N_0}\right), \quad \text{or} \quad \frac{E_b}{N_0} = \frac{2^{2R_M}-1}{2R_M} \tag{3.4.5}$$

for baseband channels, and

$$R_M = \log_2\left(1 + R_M\frac{E_b}{N_0}\right), \quad \text{or} \quad \frac{E_b}{N_0} = \frac{2^{R_M}-1}{R_M} \tag{3.4.6}$$

for passband channels. If $R_M = 1$, then $E_b/N_0 = 1.76$ dB (baseband) and $E_b/N_0 = 0$ dB (passband), and if R_M is higher, E_b/N_0 also becomes higher. As $R_M \to 0$, according to (A.1.3),

$$\lim_{R_M\to 0}\frac{E_b}{N_0} = \lim_{R_M\to 0}\frac{2^{2R_M}\cdot\ln(2)\cdot 2}{2} = \ln 2 \cong -1.59 \text{ dB}. \tag{3.4.7}$$

This *Shannon limit* applies both for baseband and passband channels with continuous-valued input and, furthermore, is identical to the case of binary input (3.3.6). The explanation for the existence of such a limit is similar to that for the curves in Figure 2.4.

Example 3.6. According to Theorem 3.4, for $E_s/N_0 = -5$ dB the channel capacity is $C = \frac{1}{2}\log_2(1 + 2\cdot 0.316) \approx 0.35$ information bit per modulation symbol. Coding with $R_M = 0.3$ information bit per modulation symbol makes an almost error-free transmission possible. However, then $E_b/N_0 = E_s/(R_M N_0) = 0.316/0.3 \cong 0.2$ dB, but this value lies above the Shannon limit. For $E_b/N_0 < -1.59$ dB there is no R_M such that for $E_s/N_0 = R_M E_b/N_0$ we obtain a channel with a capacity of $C > R_M$. ■

The cutoff rate R_0 can also be generalized to a continuous-valued Gaussian distributed input alphabet. To obtain mathematically and physically significant results a restriction of the maximum energy per symbol is introduced. For more details and a derivation we refer to [42, 219]. For baseband signaling, with the abbreviation $v = E_s/N_0$,

$$R_{0,1-\text{dim}} = \frac{1 + v - \sqrt{1+v^2}}{2\ln 2} + \frac{\log_2\left(1+\sqrt{1+v^2}\right) - 1}{2}. \tag{3.4.8}$$

Similarly as in Theorem 3.4, the transition to passband signaling, gives the following result with the abbreviation $v = E_s/(2N_0)$,

$$R_{0,2-\text{dim}} = \frac{1 + v - \sqrt{1+v^2}}{\ln 2} + \log_2\left(1+\sqrt{1+v^2}\right) - 1. \tag{3.4.9}$$

These theoretical R_0 limits are shown in Figures 3.10 and 3.12.

3.4.2 The Coding Gain as a Function of the Bit-Error Rate

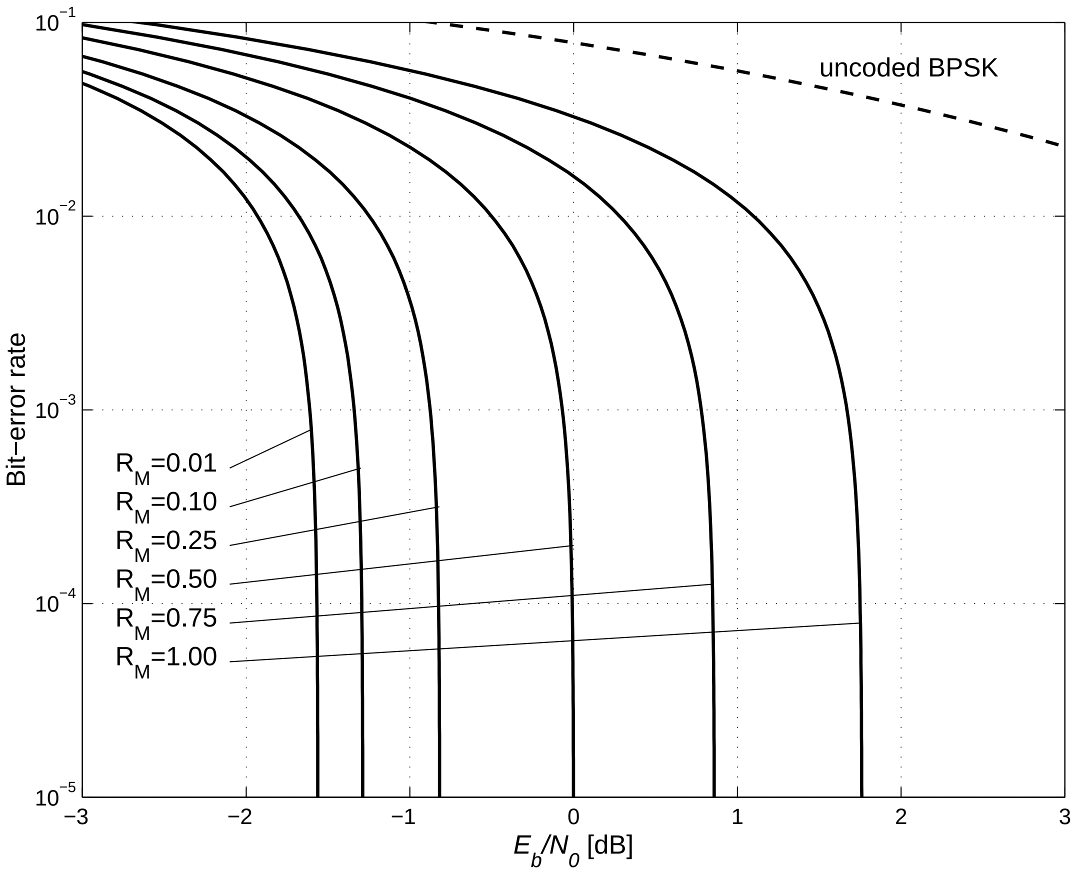


Figure 3.8. Required E_b/N_0 to achieve a bit-error rate P_b under the condition of $R = C_{\text{Gauss-in}}$ for the AWGN channel with non-binary input

The heading to this subsection is identical to that of Subsection 3.2.2, and in both subsections we consider the relationship between the bit-error rate P_b after decoding and the compulsory E_b/N_0 on the condition that $R = C_{\text{soft}}$ for the AWGN channel. The only difference is that now we consider the AWGN channel with Gaussian input instead of binary input. For the output we presume soft decisions in both cases.

Similar to Subsection 3.3.2, according to (3.4.5),

$$\begin{aligned} R'_M = R_M \cdot (1 - H_2(P_b)) &= C_{\text{Gauss-in}} \left(\text{ at } \frac{E_s}{N_0} = R'_M \cdot \frac{E_b}{N_0} \right) \\ &= \frac{1}{2} \log_2 \left(1 + 2R'_M \frac{E_b}{N_0} \right) \qquad (3.4.10) \end{aligned}$$

for the baseband AWGN channel and therefore

$$\frac{E_b}{N_0} = \frac{2^{2R'_M(1-H_2(P_b))}}{2R'_M}. \tag{3.4.11}$$

In Figure 3.8, these coding limits are illustrated for some values of R_M by a graph of P_b over E_b/N_0 (again we use the term R_M instead of R'_M). Of course, values of R_M greater than 1 are possible. As in Figure 3.7, the graphs are almost vertical for small P_b and the value of E_b/N_0 in Figure 3.8 is equal to the values of the curve $C_{\text{Gauss-in}}$ in Figure 3.6. So Figure 3.8 only yields additional information above $P_b > 0.001$. The Shannon limit (3.3.6) can be found at $R = 0.01$ again.

The comparison of Figures 3.7 (binary input) and 3.8 (Gaussian input) as well as the comparison of $C_{\text{Gauss-in}}$ with C_{soft} in Figure 3.6 show that the transition from binary to non-binary input signals is only advantageous for code rates greater than about 1/2.

Table 3.1. Overview of capacities for the AWGN channel

Parameter	Input	Output
C_{hard}	binary ($M = 1$)	binary
C_{soft}	binary ($M = 1$)	soft
$C_{\text{Gauss-in}}$	Gaussian ($M = \infty$)	soft
C_{ASK}, C_{PSK}, C_{QAM}	2^M-ary modulated	soft

To be able to keep track of the channel capacities used here, Table 3.1 gives an overview of all channel capacities mentioned so far as well as those to come for the AWGN channel. All these capacities are given in units of information bits per symbol (i.e., per channel use), whereas C^* denotes the channel capacity in units of information bits per second.

3.4.3 Amplitude Shift Keying (ASK)

In practice a Gaussian distributed input is of course unrealistic, not only because this would require unlimited peak values. Therefore in the following we will consider the AWGN channel with 2^M-ary input where the 2^M amplitudes $\xi_i \in \mathcal{A}_{\text{mod}}$ in the modulation alphabet are equidistant and occur with the same a priori probabilities $P_x(\xi_i) = 2^{-M}$. In this subsection on baseband signaling we consider 2^M-ASK (Amplitude Shift Keying) as the modulation scheme with the input alphabet (2.3.3). Even without optimization of the input distribution P_x as required in Definition 3.2, the mutual information is still called channel capacity for a uniform input distribution. According to (3.1.10), the channel capacity for 2^M-ASK can be written as

$$\begin{aligned} C_{2^M-\mathrm{ASK}} &= \sum_i \int\limits_{-\infty}^{\infty} \frac{1}{2^M} f_{y|x}(\eta|\xi_i) \log_2 \frac{f_{y|x}(\eta|\xi_i)}{2^{-M}\sum_l f_{y|x}(\eta|\xi_l)}\, d\eta \\ &= M - \frac{1}{2^M\sqrt{\pi N_0}} \sum_i \int\limits_{-\infty}^{\infty} e^{-(\eta-\xi_i)^2/N_0} \log_2\left(\sum_l e^{-(\eta-\xi_l)^2/N_0+(\eta-\xi_i)^2/N_0} \right) d\eta. \end{aligned}$$

With the substitution $u = \eta\sqrt{2/N_0}$ and the abbreviation $a_i = \xi_i\sqrt{2/N_0}$ we obtain

$$\begin{aligned} C_{2^M-\mathrm{ASK}} &= M - \frac{1}{2^M\sqrt{2\pi}} \sum_i \int\limits_{-\infty}^{\infty} e^{-(u-a_i)^2/2} \log_2\left(\sum_l e^{-(u-a_l)^2/2+(u-a_i)^2/2} \right) du \\ &= M - \frac{1}{2^M\sqrt{2\pi}} \sum_i \int\limits_{-\infty}^{\infty} e^{-u^2/2} \log_2\left(\sum_l e^{-(a_i-a_l)^2/2-u(a_i-a_l)} \right) du. \end{aligned} \tag{3.4.12}$$

This integral can only be numerically evaluated as in the binary case. The resulting curves are shown in Figure 3.7. The curve of 2-ASK is identical to the curve C_{soft} in Figure 3.3 and to C in (3.1.20). The curve $C_{\mathrm{Gauss-In}} = \log_2(1 + 2E_s/N_0)/2$ is obtained according to (3.4.3). Obviously,

$$C_{2^M-\mathrm{ASK}} \approx \left\{ \begin{array}{ll} M & \text{for large } E_s/N_0 \\ C_{\mathrm{Gauss-In}} & \text{for small } E_s/N_0 \end{array} \right\}. \tag{3.4.13}$$

The large dots in Figures 3.7 and 3.8 refer to uncoded transmission with a symbol-error rate of $P_s = 10^{-5}$ (see for example [114, 112] or Section 2.4 for the calculation of P_s). For 4-ASK and $E_s/N_0 = 16.8$ dB, $P_s = 10^{-5}$ for the uncoded transmission and $C_{4-\mathrm{ASK}} = 2$. So, coding with a code rate of $R_M = 2$ information bit per modulation symbol could at least theoretically make an arbitrarily small error probability possible.

If the modulation scheme is changed from 4-ASK to 8-ASK, then $C_{8-\mathrm{ASK}} = 2$ already for $E_s/N_0 = 9.6$ dB, hence we obtain a coding gain of 7.2 dB as shown in Figure 2.5. But for a coding gain of only 4 dB, i.e., for $E_s/N_0 = 12.8$ dB, $C_{8-\mathrm{ASK}} = 2.455$ is clearly larger than $R_M = 2$, so coding can enable a very small error probability in practice. If 16-ASK is used, then $C_{16-\mathrm{ASK}} = 2$ at $E_s/N_0 = 9.5$ dB in contrast to $C_{8-\mathrm{ASK}} = 2$ at $E_s/N_0 = 9.6$ dB, therefore the coding gain increases for the transition from 8-ASK to 16-ASK by only 0.1 dB.

Thus a doubling of the alphabet is more than sufficient. This result forms the basis of trellis coded modulation (TCM), introduced in Chapter 10, where the error-control coding scheme and the higher-level modulation scheme are jointly optimized. Even if a continuous-valued Gaussian distributed input is considered, instead of a doubling of the alphabet, then according to Figure 3.7

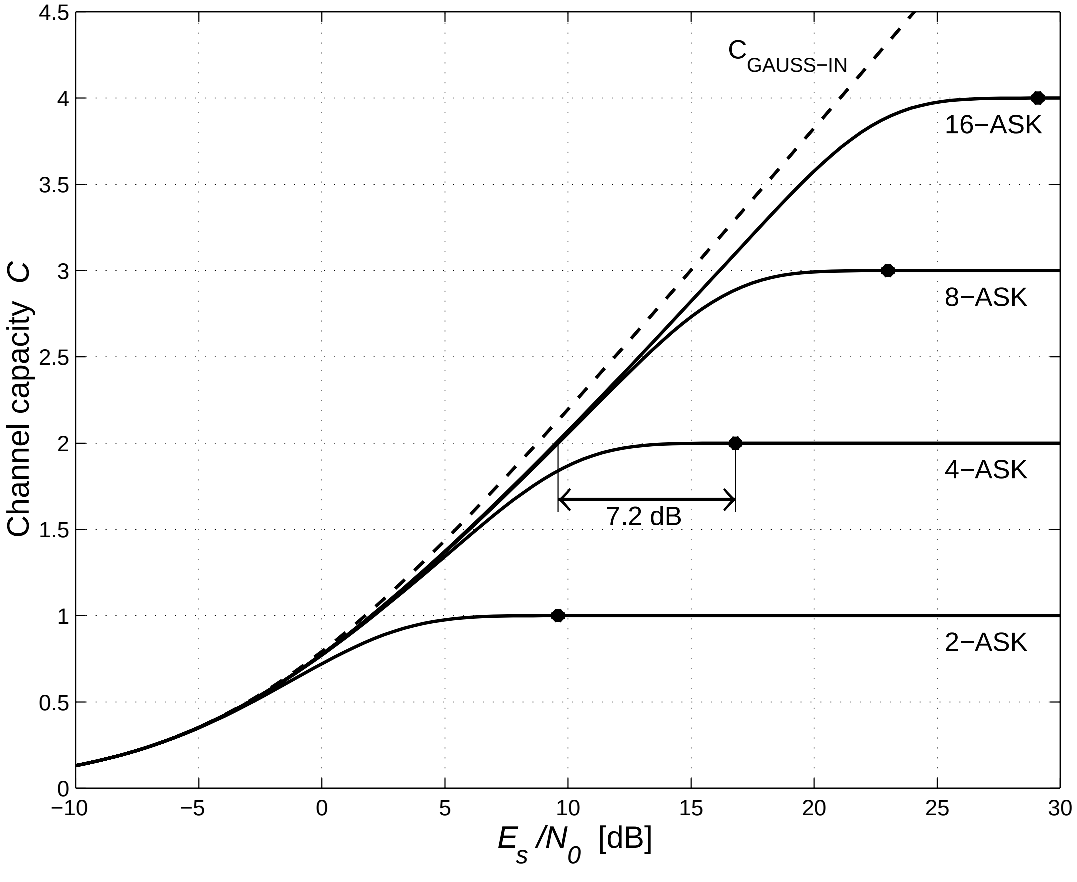


Figure 3.9. Channel capacity C of ASK for the AWGN channel with soft decisions

the additional gain is less than 1 dB. For $R_M = 2$ and 8-ASK the code rate, according to (1.4.3) or (2.6.1), is

$$R = \frac{R_M}{M} = \frac{2}{3} \quad \left[\frac{\text{info symbol}}{\text{encoded symbol}} = \frac{\text{info bit/modulation symbol}}{\text{encoded bit/modulation symbol}}\right].$$

The capacity curves in Figure 3.7 are compared to the corresponding cutoff rate curves in Figure 3.8. The boundary curves $C_{\text{Gauss-In}}$, as in (3.4.3), are identical in Figures 3.7 and 3.8. The boundary curve $R_{0,\text{Gauss-In}}$, as in (3.4.6), is shown in Figure 3.8.

According to (3.2.9), for 2^M-ary ASK,

$$\begin{aligned} R_{0,2^M\text{-ASK}} &= -\log_2 \int_{-\infty}^{\infty} \left(\sum_i \frac{1}{2^M} \sqrt{\frac{1}{\sqrt{\pi N_0}} e^{-(\eta-\xi_i)^2/N_0}} \right)^2 d\eta \\ &= -\log_2 \frac{1}{4^M \sqrt{\pi N_0}} \int_{-\infty}^{\infty} \sum_{i,l} e^{-(\eta-\xi_i)^2/2N_0-(\eta-\xi_l)^2/2N_0} d\eta \end{aligned}$$

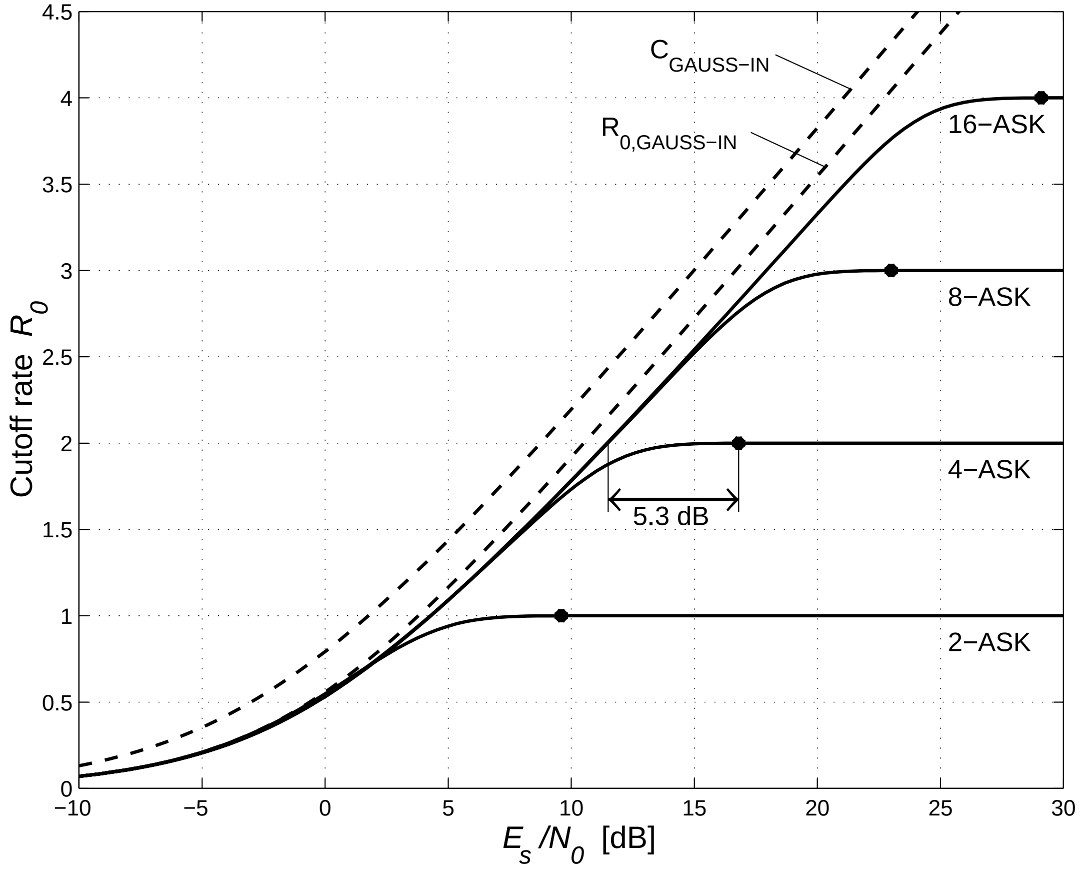


Figure 3.10. Cutoff rate R_0 of ASK for the AWGN channel with soft decisions

$$= -\log_2 \frac{1}{4^M\sqrt{\pi N_0}} \int_{-\infty}^{\infty} \sum_{i,l} e^{-(\eta-(\xi_i+\xi_l)/2)^2/N_0} e^{-(\xi_i-\xi_l)^2/4N_0} d\eta$$

$$= -\log_2 \frac{1}{4^M} \sum_{i,l} \exp\left(-(i-l)^2 \frac{3}{4(4^M-1)} \frac{E_s}{N_0}\right) \tag{3.4.14}$$

$$= -\log_2 \frac{1}{4^M} \left(2^M + 2\sum_{v=1}^{2^M-1} (2^M - v) \exp\left(-v^2 \frac{3}{4^M-1} \frac{E_s}{N_0}\right)\right). \tag{3.4.15}$$

The R_0-curves in Figure 3.8 show the same behaviour as the C-curves in Figure 3.7, only with a shift of 1.5 to 2 dB. The coding gains are smaller accordingly. Again an increase of M by 1 proves to be sufficient. The curve $R_{0,2-\text{ASK}}$ is identical to the curve $R_{0,\text{soft}}$ in Figure 2.2.

3.4.4 Phase Shift Keying (PSK) and Quadrature Amplitude Modulation (QAM)

In this subsection we consider PSK and QAM as the most important passband modulation schemes. Their complex-valued modulation alphabets $\mathcal{A}_{\text{mod}}$ are defined in (2.3.5) and (2.3.7). The channel output can also be interpreted as a complex number $y = y_I + jy_Q \in \mathbb{C}$ as described in Section 2.1. The 2-dimensional Gaussian probability density function (PDF) is given in (2.1.6) or in the more compact form

$$f_{y|x}(\eta|\xi_i) = \frac{1}{2\pi\sigma^2} \cdot \exp\left(-\frac{|\eta - \xi_i|^2}{2\sigma^2}\right) \tag{3.4.16}$$

with $|\eta - \xi_i|^2 = (\eta_I - \xi_{i,I})^2 + (\eta_Q - \xi_{i,Q})^2$ and $2\sigma^2 = N_0$. The 2-dimensional Gaussian PDF is sketched in Figure 2.1 and a comprehensive discussion of the multi-dimensional Gaussian distribution can be found in Subsection A.4.3.

The channel capacity C for arbitrary 2-dimensional signal constellations $\mathcal{A}_{\text{mod}} = \{\xi_i \mid i = 0, \ldots, 2^M - 1\}$ can be calculated from equation (3.1.10) as follows:

$$\begin{aligned}
C &= \int\limits_{-\infty}^{\infty} \sum_i \frac{1}{2^M} f_{y|x}(\eta|\xi_i) \cdot \log_2 \frac{f_{y|x}(\eta|\xi_i)}{2^{-M} \sum_l f_{y|x}(\eta|\xi_l)} \, d\eta \\
&= -\frac{1}{2^M} \int\limits_{-\infty}^{\infty} \sum_i f_0(\eta - \xi_i) \cdot \log_2 \frac{\sum_l \exp(-|\eta - \xi_l|^2/N_0)}{2^M \cdot \exp(-|\eta - \xi_i|^2/N_0)} \, d\eta \\
&= -\frac{1}{2^M} \int\limits_{-\infty}^{\infty} f_0(\eta) \underbrace{\sum_i \log_2\left(2^{-M} \sum_l \exp\left(-\frac{|\eta + \xi_i - \xi_l|^2 - |\eta|^2}{N_0}\right)\right)}_{= g(\eta)} d\eta,
\end{aligned} \tag{3.4.17}$$

where $f_0(\eta) = \exp(-|\eta|^2/N_0)/(\pi N_0)$ is the probability density function of the $N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_2)$ 2-dimensional Gaussian distribution. The 2-dimensional integral in (3.4.13) can not be analytically calculated and the numerical integration is also very difficult. An easier evaluation can be performed by using the *Monte-Carlo technique* [225], for this (3.4.13) is interpreted as the mean of the random variable $g(y)$, where y is Gaussian distributed with $y \sim N(0, \sigma^2 \boldsymbol{I}_2)$. This is equivalent to $y = (y_I, y_Q)$ with the two statistically independent random variables $y_I, y_Q \sim N(0, \sigma^2)$. The mean is approximated by N random samples η_n with the PDF f_0 as follows:

$$C = E(g(\eta)) = \int\limits_{-\infty}^{\infty} f_0(\eta) \cdot g(\eta) \, d\eta \approx \frac{1}{N} \sum_{n=1}^{N} g(\eta_n). \tag{3.4.18}$$

The curves in Figure 3.9 were calculated with $N = 1000$ samples and 80 points per curve.

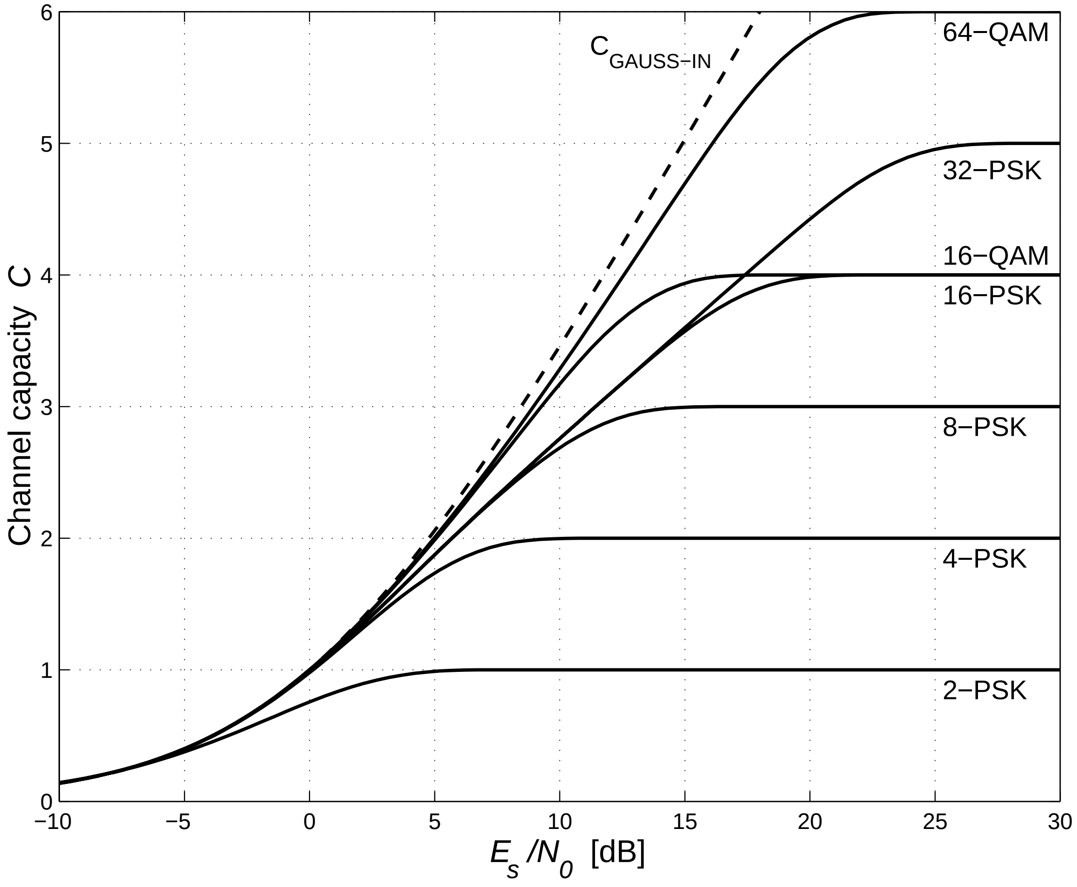


Figure 3.11. Channel capacity C of PSK and QAM for the AWGN channel with soft decisions

Obviously QAM is much closer to the theoretical limit than PSK, as can be seen from the comparison of 16-QAM to 16-PSK in the range of 5 to 20 dB. The large deviation of PSK from the capacity boundary, compared to QAM or compared to the 1-dimensional ASK, can be easily explained by the signal constellations: PSK with uniform distributed signal points on a circle deviates even more from the optimum Gaussian distribution over the complex plane than QAM with uniform distributed signal points within a square.

The effects of doubling the size of the symbol alphabet for PSK and QAM are identical to those for ASK.

For completeness, the cutoff rate R_0 for PSK is given in Figure 3.10 although we do not obtain any profoundly new results. For 2^M-PSK with the signal alphabet (2.3.5) and $2\sigma^2 = N_0$ according to (3.2.9)

$$R_{0,2^M-\mathrm{PSK}} = -\log_2 \int_{-\infty}^{\infty} \left(\sum_i \frac{1}{2^M} \sqrt{\frac{1}{\pi N_0} e^{-|\eta-\xi_i|^2/N_0}} \right)^2 d\eta$$

$$
\begin{aligned}
&= -\log_2 \frac{1}{4^M \pi N_0} \int\limits_{-\infty}^{\infty} \sum_{i,l} \exp\left(-\frac{|\eta-\xi_i|^2+|\eta-\xi_l|^2}{2N_0}\right) d\eta \\
&= -\log_2 \frac{1}{4^M \pi N_0} \sum_{i,l} \exp\left(-\frac{2|\xi_i|^2+2|\xi_l|^2-|\xi_i+\xi_l|^2}{4N_0}\right) \\
&\qquad\qquad \cdot \underbrace{\int\limits_{-\infty}^{\infty} \exp\left(-\frac{|\eta-(\xi_i+\xi_l)^*/2|^2}{N_0}\right) d\eta}_{=\pi N_0} \\
&= -\log_2 \frac{1}{4^M 2} \sum_{i,l} \exp\Bigg(-\frac{E_s}{N_0}\Bigg(1-\frac{1}{4}\underbrace{\left|e^{j2\pi\xi_i/2^M}+e^{j2\pi\xi_l/2^M}\right|^2}_{=2+2\cos(2\pi(i-l)/2^M)}\Bigg)\Bigg) \\
&= -\log_2 \frac{1}{4^M} \sum_{i,l} \exp\left(-\frac{E_s}{2N_0}(1-\cos(2\pi(i-l)/2^M))\right) \\
&= -\log_2 \frac{1}{2^M} \sum_{v=0}^{2^M-1} \exp\left(-\frac{E_s}{2N_0}(1-\cos(2\pi v/2^M))\right) \\
&= -\log_2 \frac{1}{2^M} \sum_{v=0}^{2^M-1} \exp\left(-\frac{E_s}{N_0}\sin^2(\pi v/2^M)\right). \qquad (3.4.19)
\end{aligned}
$$

Further curves of C and R_0 for PSK and QAM can be found for example in [19, 186, 225].

3.5 Band-Limited AWGN Channels

So far we only assumed the transmit power to be limited. The code rates were in no way restricted, in other words the AWGN channel was assumed to be useable as often as necessary and therefore to have unrestricted bandwidth. Many applications require an efficient use of the available spectrum. Therefore we will now presuppose a continuous-time AWGN channel with the bandwidth W, i.e., the channel is limited to frequencies f with

$$|f| < W \text{ (baseband)} \quad \text{and} \quad |\pm f_c - f| < W/2 \text{ (passband)}. \qquad (3.5.1)$$

As discussed in Subsection 2.2.2, we further assume W as the minimum Nyquist bandwidth with a rolloff factor of $\alpha = 0$. According to (2.2.19) or (2.6.3), the maximum symbol rate (baud rate) is limited to

$$r_s = 2W \text{ (baseband)} \quad \text{and} \quad r_s = W \text{ (passband)} \qquad (3.5.2)$$

modulation symbols per second, given that no degradations due to intersymbol interferences are permitted.

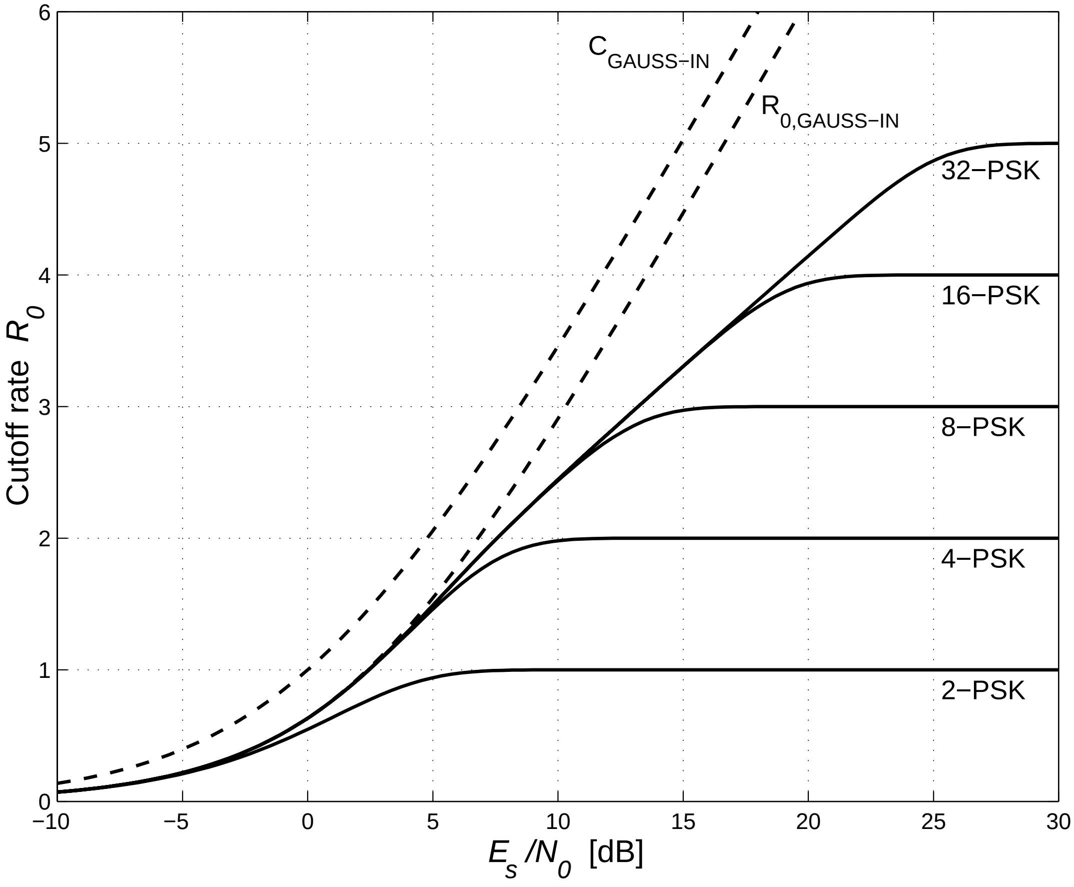


Figure 3.12. Cutoff rate R_0 of PSK for the AWGN channel with soft decisions

3.5.1 The Shannon-Hartley Theorem

The capacity of the continuous-time AWGN channel with a continuous-valued input as in Subsection 3.4.1 can be derived from Theorem 3.4 for the discrete-time channel. With (2.6.8),

$$C = \left\{ \begin{array}{ll} \frac{1}{2} \cdot \log_2 \left(1 + \frac{S}{N}\right) & \text{baseband channel} \\ \log_2 \left(1 + \frac{S}{N}\right) & \text{passband channel} \end{array} \right\} \tag{3.5.3}$$

information bits per discrete channel symbol can be transmitted. So a maximum of $C^*_{1-\text{dim}} = 2W \cdot C_{1-\text{dim}}$ or $C^*_{2-\text{dim}} = W \cdot C_{2-\text{dim}}$ information bits per second can be transmitted. The form C^* of the channel capacity was already introduced in (3.2.3) and again we emphasize the important difference that C refers to symbols and C^* to seconds.

Thus we have proved the following Shannon-Hartley Theorem for the capacity of a bandwidth-limited AWGN channel for an optimally distributed, i.e., Gaussian distributed, input signal. The right-hand side of the formula in the

Theorem follows from $S = E_b \cdot r_b$ and $N = N_0 \cdot W$ according to (2.6.6) and (2.6.7), respectively. Recall that r_b denotes the throughput in units of information bit per second.

Theorem 3.5 (Shannon-Hartley). *For the baseband and passband band-limited AWGN channel with continuous-valued input the capacity is*

$$C^* = W \cdot \log_2\left(1 + \frac{S}{N}\right) = W \cdot \log_2\left(1 + \frac{E_b}{N_0} \cdot \frac{r_b}{W}\right) \tag{3.5.4}$$

in units of information bit per second. For $r_b < C^*$ an almost error-free transmission is possible with very high effort. The parameters S/N and W set an elementary bound for the throughput, but not for the quality of the transmission.

Of particular importance is that the three key parameters of digital communications, the channel bandwidth W, the signal-to-noise ratio S/N or E_b/N_0 and the throughput r_b, can be summarized in one single formula. Obviously, an exchange between the bandwidth and the signal-to-noise ratio is possible, for example, a small S/N can be compensated for by a larger W. However, there is a fundamental restriction, since $E_b/N_0 \to 0$ can not be compensated for by $W \to \infty$, as can be easily seen:

$$\begin{aligned}\lim_{W\to\infty} C^* &= \lim_{W\to\infty} W \cdot \log_2\left(1 + \frac{S}{N_0 \cdot W}\right)\\ &= \lim_{W\to\infty} \log_2\left(\left(1 + \frac{S}{N_0 \cdot W}\right)^W\right) = \log_2 \exp\left(\frac{S}{N_0}\right)\\ &= \frac{1}{\ln 2} \cdot \frac{S}{N_0} = 1.44 \cdot \frac{r_b \cdot E_b}{N_0}.\end{aligned} \tag{3.5.5}$$

Since $r_b < C^*$, the implication is the *Shannon limit* for E_b/N_0 again, which can not be violated for a reliable transmission:

$$\frac{E_b}{N_0} > \ln 2 \cong -1.59 \text{ dB}. \tag{3.5.6}$$

Once more we note that this is only a theoretical bound, which can not be reached in practice, since (i) the code rate must converge to zero or equivalently the bandwidth must approach infinity, (ii) the input of the channel must be continuous-valued and Gaussian distributed and (iii) the block length and the complexity of the code must exceed each bound.

Of course, there is no lower bound for S/N. Obviously, $W \to 0$ implies that $C^* \to 0$, hence it is trivial that there can not be a transmission without any bandwidth. A noise-free channel is physically impossible, but mathematically for $N_0 = 0$ the channel capacity is infinite.

Example 3.7. We consider some applications of the Shannon-Hartley Theorem.

(1) For the standard voice-band telephone channel (in the passband domain) over switched lines the bandwidth is typically $W = 3000$ Hz and the signal-to-noise ratio is $S/N = 30$ dB. Thus the capacity is

$$C^* = 3000 \cdot \log_2(1 + 1000) \approx 29.9 \text{ kbit/s}.$$

Under different conditions the channel capacity can also be larger or smaller. However, the telephone channel is not a pure AWGN channel, since in addition to the noise other deteriorations and imperfections also have to be considered. Since 1994 modems with up to 28.8 kbit/s according to the ITU-T standard V.34, have been introduced to the market. This was later improved to 33.6 kbit/s with the standard V.34bis. An overview of modem technology can be found in Section 16.2?.

(2) We will now compare baseband 2-PSK and passband 4-PSK, presupposing a bit-error rate of 10^{-8}. For uncoded 2-PSK, about $E_b/N_0 = E_s/N_0 = 12.0$ dB are required according to Table 1.1, Figure 2.10 or Figure 2.11. For uncoded 4-PSK about $E_s/N_0 = 15.0$ dB are required according to Figure 2.10, and $E_b/N_0 = 12.0$ dB according to Figure 2.11. In each case, $S/N = 15.0$ dB according to (2.6.9). A bandwidth of $W = 4000$ Hz allows a maximum baud rate of $r_s = 8000$ symbol/s for 2-PSK and $r_s = 4000$ symbol/s for 4-PSK, according to (2.6.3). In each case, $r_b = 8000$ bit/s for the throughput. According to Theorem 3.5,

$$C^* = 4000 \cdot \log_2(1 + 10^{15/10}) \approx 20.1 \text{ kbit/s}$$

for baseband as well as for passband signaling. Thus, in contrast to uncoded signaling with a bit-error rate of 10^{-8}, error-control coding can enable a throughput that is about 2.5 times higher for arbitrarily small bit-error rate.

(3) Now, we consider baseband 256-ASK, again with $W = 4000$ Hz and $r_s = 8000$ symbol/s. Obviously, $r_b = 64000$ bit/s for uncoded signaling. According to Figure 2.14, in contrast to binary modulation, high-level ASK requires an additional asymptotic amount of 43.4 dB with respect to E_s/N_0. So all in all about $E_s/N_0 = 12.0 + 43.4 = 55.4$ dB or $S/N = 58.4$ dB or $E_b/N_0 = 46.4$ dB are required. According to Theorem 3.5,

$$C^* = 4000 \cdot \log_2(1 + 10^{58.4/10}) \approx 77.6 \text{ kbit/s} .$$

Here the difference between uncoded signaling with $P_b = 10^{-8}$ and the channel capacity is smaller than for (2). This effect will become even clearer with Figure 3.13, where the vertical (logarithmic) difference (in reference to the throughput) between ASK or QAM and the capacity boundary decreases as $M \to \infty$, whereas the horizontal gap (in reference to E_b/N_0) remains almost the same.

However, the more important implication here is that such a high throughput in combination with such a small bandwidth is only possible with extremely high signal-to-noise ratio. The constellation considered in this example is not just theoretical but is similar to the situation for ITU-T V.90 modems (also known as 56K modems), which have been widely-used for internet access since 1998. However, this is only the case for the downstream direction from the internet service provider (ISP) to the subscriber modem. Furthermore, the link from the switching station to the modem, which is only a few kilometers long, should be the only analog circuit in the whole connection, whereas the ISP is connected via a digital line to the digital public switched telephone network (PSTN). The V.90 standard is also discussed in detail in Section 16.2?. ■

3.5.2 Spectral Efficiency and the Bandwidth-Efficiency Diagram

Definition 3.5. *The value r_b/W is called the* spectral efficiency *or* spectral bit rate *or* bandwidth efficiency *and*

$$\frac{C^*}{W} = \log_2\left(1 + \frac{S}{N}\right) = \log_2\left(1 + \frac{E_b}{N_0}\cdot\frac{r_b}{W}\right) \tag{3.5.7}$$

is called the normalized channel capacity*. Both r_b/W and C^*/W are in units of information bits per second per Hertz, i.e., simply in units of information bits.*

The Shannon-Hartley Theorem presumes that $r_b/W < C^*/W$. For the limit case of $r_b = C^*$, we have a relation between C^*/W and E_b/N_0

$$\frac{C^*}{W} = \log_2\left(1 + \frac{E_b}{N_0}\cdot\frac{C^*}{W}\right) \quad \text{or} \quad \frac{E_b}{N_0} = \frac{2^{C^*/W} - 1}{C^*/W} \tag{3.5.8}$$

and once more the Shannon limit (3.5.4)

$$\lim_{C^*/W\to 0}\frac{E_b}{N_0} = \ln 2 \cong -1.59 \text{ dB}, \tag{3.5.9}$$

which applies both for baseband and passband signaling.

Figure 3.11 shows the normalized channel capacity C^*/W over E_b/N_0 in comparison to various digital modulation schemes without coding. Above the capacity bound ($r_b > C^*$) the error probability can never be brought below a certain value. Below the channel capacity bound ($r_b < C^*$) an almost error-free transmission is possible with an appropriate effort. The modulation schemes ASK (Amplitude Shift Keying), PSK (Phase Shift Keying), QAM (Quadrature Amplitude Modulation) and FSK (Frequency Shift Keying see [57, 128, 151]) are meant to be without coding, with coherent demodulation (i.e., ideal carrier and clock recovery) and refer to a symbol-error rate of 10^{-5} (see Figure 2.11 for PSK and QAM). Modulation schemes of the same type are connected by lines

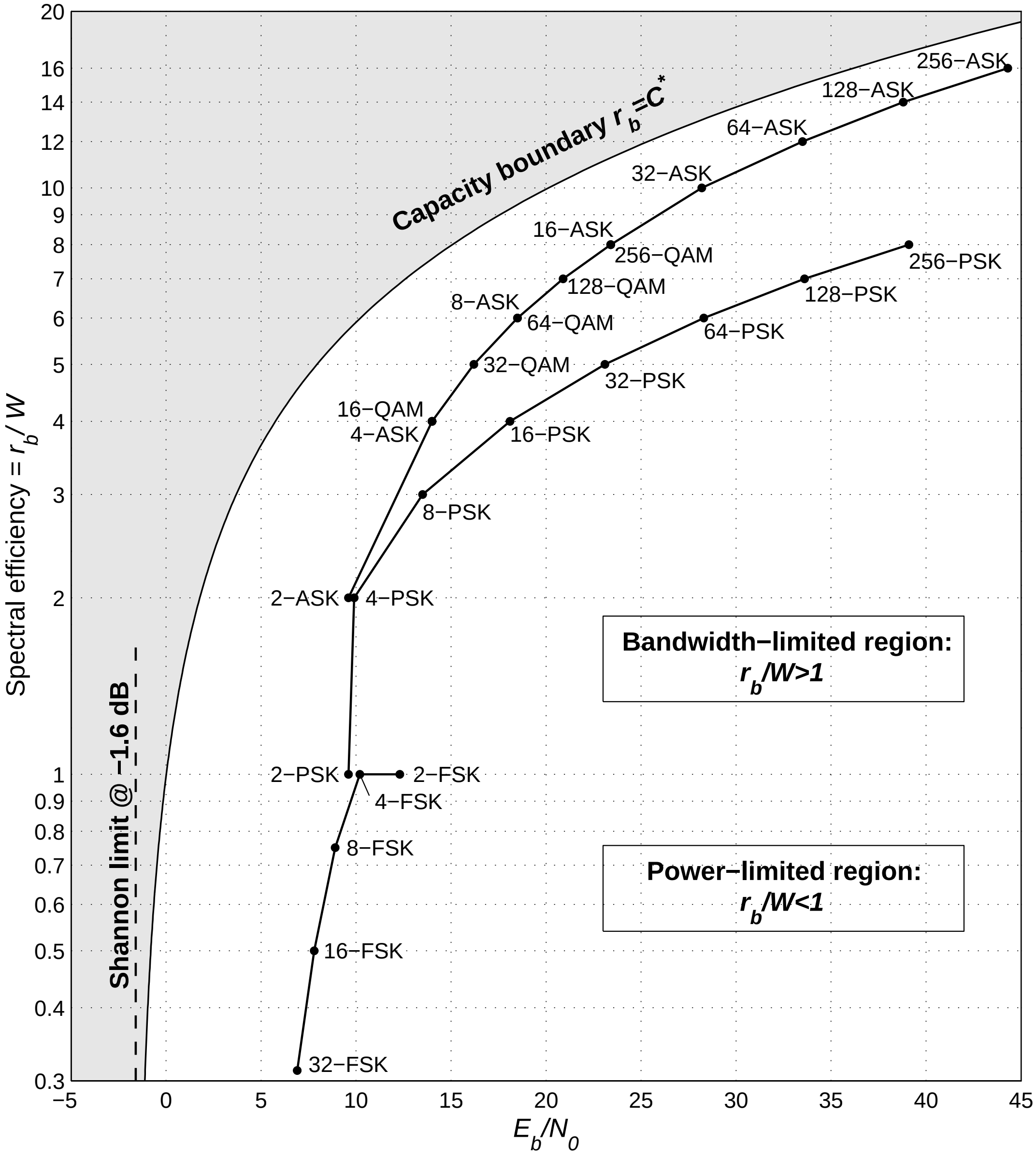


Figure 3.13. Bandwidth-efficiency diagram
(uncoded modulation schemes at $P_s = 10^{-5}$, ordinate logarithmically scaled)

in Figure 3.11. The spectral efficiency is

$$\frac{r_b}{W} = \left\{ \begin{array}{ll} 2M & 2^M\text{-ASK} \\ M & 2^M\text{-PSK, } 2^M\text{-QAM} \\ M/2^{M-1} & 2^M\text{-FSK (coherent)} \end{array} \right\}. \tag{3.5.10}$$

According to (2.4.11) and (2.4.12), 2^M-ASK and 2^{2M}-QAM have almost the same bit-error rate for equal E_b/N_0, thus ASK and QAM lie on a mutual curve in Figure 3.13, which was obvious anyway.

We insert also some remarks on FSK: The 2^M tones are spaced at $\Delta f = 1/(2T_s) = r_s/2$ Hz for coherent FSK or at $\Delta f = 1/T_s = r_s$ Hz for non-coherent FSK. So the bandwidth is approximately $W = 2^M \cdot \Delta f$. The result $r_b/W = M/2^{M-1}$ (coherent) or $r_b/W = M/2^M$ (non-coherent) follows from $r_b = Mr_s$. For more details on FSK we refer to textbooks on digital communications, e.g., [128].

All modulation schemes in Figure 3.11 are about 10 dB away from the channel capacity bound in reference to E_b/N_0. For a smaller error probability this distance becomes even larger. Therefore error-control coding can lead to considerable improvements of digital transmission systems, which actually justifies the big efforts required in theory and implementation. Also, note that, as $M \to \infty$, the vertical logarithmic gap between the capacity boundary and ASK and QAM continuingly decreases.

In Figure 3.11 one can roughly distinguish between two main areas (which are not exactly separated, but overlap):

Power-limited region ($r_b/W < 1$) is characterized by a large or medium bandwidth compared to the information bit rate. The minimum of E_b/N_0 is -1.59 dB given by the Shannon limit. FSK is a bandwidth-intensive modulation scheme where for higher levels the spectral efficiency as well as the necessary E_b/N_0 decrease. Smaller reductions of E_b/N_0 require a considerably larger bandwidth for FSK. For the classic channel coding with 2-ASK or 4-PSK (Chapters 4 to 10) saving transmit power can be compensated for by larger bandwidth and higher complexity. Deep-space satellite communication is one typical application (see Section 12.1).

Bandwidth-limited region ($r_b/W > 1$) is characterized by a small or medium bandwidth and a large E_b/N_0. High-level modulation schemes, for example QAM and PSK or a combination of both, are used. In contrast to FSK, higher levels for QAM and PSK increase the spectral efficiency as well as the necessary E_b/N_0. For example, there are line-of-sight microwave radio systems with 1024-QAM providing a spectral efficiency of 10 bit/sec/Hz. However, a further improvement of the spectral efficiency causes an immense increase of the required E_b/N_0. For instance, an improvement of the spectral efficiency from 2 to 10 bit/sec/Hz implies an increase of the required E_b/N_0 from 1.76 dB to 20.10 dB, or in terms of $E_s/N_0 = M \cdot E_b/N_0$ from 4.77 dB to 30.10 dB.

Error-control coding should not raise the required bandwidth for most bandwidth-limited applications. As an attractive alternative to classic error-control coding there are bandwidth-efficient coding schemes where channel coding and modulation schemes are jointly optimized. An essential technique is that the number of levels of the modulation scheme is increased while the modulation symbol rate remains constant. These methods are known as trellis coded modulation (TCM) and will be introduced in Chapter 10.

> Modems (see Section 16.2?) and line-of-sight microwave radio (see Section 16.5?) are typical applications of bandwidth-limited communication. For mobile radio systems (see Sections 16.3? and 16.4?) both power- as well as bandwidth-efficient transmission methods are required.

In summary, the designer of a coded communication system has the following possibilities for exchanges between the most important parameters of digital signaling, i.e., the signal-to-noise ratio E_b/N_0, the spectral efficiency r_b/W and the symbol-error probability P_s:

- exchange between P_s and E_b/N_0 for a constant r_b/W.
- exchange between P_s and r_b/W for a constant E_b/N_0.
- exchange between r_b/W and E_b/N_0 for a constant P_s.

However, only the first exchange does not require a change of the modulation scheme.

3.6 Appendix: Proof of Shannon's Noisy Channel Coding Theorem for the BSC

For the special case of the binary symmetric channel (BSC) with $C = 1 - H_2(p_e)$ the channel coding Theorem 3.1 can be proved fairly easily, also the random coding argument becomes quite clear.

Let $p_e < 0.5$ and let $\varepsilon > 0$ and $\varepsilon' > 0$ be arbitrary. We are to prove the existence of an $(n,k)_2$ code with

$$R = \frac{k}{n} \geq C - \varepsilon' \quad \text{and} \quad P_w < \varepsilon. \tag{3.6.1}$$

The decoding is performed with the maximum-likelihood (ML) rule or a worse method. The proof is carried out in several steps:

Step 1. Since $H_2(p_e)$ is a strict monotonic increasing function for $p_e < 0.5$ (see Figure A.1), there exists a $\beta > 0$ with

$$H_2(p_e + \beta) - H_2(p_e) < \frac{\varepsilon'}{4} \quad \text{and} \quad \beta < \frac{1}{2} - p_e. \tag{3.6.2}$$

Then the block length n is chosen large enough such that

$$\frac{p_e(1-p_e)}{n\beta^2} < \frac{\varepsilon}{2} \quad \text{and} \quad 2^{-n\varepsilon'/2} < \frac{\varepsilon}{2} \tag{3.6.3}$$

and furthermore a threshold is defined as

$$t = n(p_e + \beta). \tag{3.6.4}$$

Step 2. Let an arbitrary $(n,k)_2$ code be given as

$$\mathcal{C} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{2^k}\} \subseteq \{0,1\}^n \quad \text{with} \quad C - \frac{\varepsilon'}{2} > R \geq C - \varepsilon' \tag{3.6.5}$$

in an enumerating description, where codewords may be identical.

Step 3. We will now knowingly degrade the ML rule of decoding, to make the proof easier. In reference to the defined threshold t, the received word $\boldsymbol{y}$ is decoded as follows:

(a) decision for $\boldsymbol{a}_i$, if $\boldsymbol{a}_i$ is the only codeword within a distance $\leq t$ of $\boldsymbol{y}$, i.e.: $d_H(\boldsymbol{y}, \boldsymbol{a}_i) \leq t$ and $d_H(\boldsymbol{y}, \boldsymbol{a}_j) > t$ for all j with $j \neq i$.

(b) no decision, if (a) is not fulfilled, i.e., either there is no codeword or there is more than one codeword within a distance $\leq t$ of $\boldsymbol{y}$.

In the case of (a), the ML rule is obviously fulfilled according to Theorem 1.3. Since there is no decoding in the case of (b), the defined rule is worse than the ML rule.

Step 4. Let $P_{w|i}$ be the word-error probability assuming that $\boldsymbol{a}_i$ was transmitted, then

$$\begin{aligned} P_{w|i} &= P(\text{decoding error} \mid \boldsymbol{a}_i \text{ transmitted}) \\ &= P(d_H(\boldsymbol{y}, \boldsymbol{a}_i) > t \text{ or there exists } j \neq i \text{ with } d_H(\boldsymbol{y}, \boldsymbol{a}_j) \leq t \mid \boldsymbol{a}_i) \\ &\leq \underbrace{P(d_H(\boldsymbol{y}, \boldsymbol{a}_i) > t \mid \boldsymbol{a}_i)}_{= P_{1|i}} + \underbrace{P(\text{there exists } j \neq i \text{ with } d_H(\boldsymbol{y}, \boldsymbol{a}_j) \leq t \mid \boldsymbol{a}_i)}_{= P_{2|i}}. \end{aligned}$$

Step 5. The case of $d_H(\boldsymbol{y}, \boldsymbol{a}_i) > t$ means that at least $t+1$ errors occur. The number of errors $d_H(\boldsymbol{y}, \boldsymbol{a}_i)$ is binomially distributed, according to (1.3.9), with the expected value np_e and the variance $np_e(1-p_e)$:

$$\begin{aligned} P_{1|i} &= P(d_H(\boldsymbol{y}, \boldsymbol{a}_i) > t) \\ &= P(d_H(\boldsymbol{y}, \boldsymbol{a}_i) - np_e > n(p_e + \beta) - np_e) \\ &\leq P(|d_H(\boldsymbol{y}, \boldsymbol{a}_i) - np_e| > n\beta) \\ &\leq \frac{np_e(1-p_e)}{n^2\beta^2} \qquad \text{according to Theorem A.2} \\ &= \frac{p_e(1-p_e)}{n\beta^2} \\ &< \frac{\varepsilon}{2} \qquad \text{according to (3.6.3).} \end{aligned}$$

So $P_{1|i}$ is independent of i and was also upper bounded independent of the properties of the code.

Step 6. For $P_{2|i}$ the situation is more complicated:

$$\begin{aligned} P_{2|i} &= P(\text{there exists } j \neq i \text{ with } d_H(\boldsymbol{y}, \boldsymbol{a}_j) \leq t \mid \boldsymbol{a}_i) \\ &\leq \sum_{\substack{j=1 \\ j \neq i}}^{2^k} P(d_H(\boldsymbol{y}, \boldsymbol{a}_j) \leq t \mid \boldsymbol{a}_i) = \sum_{\substack{j=1 \\ j \neq i}}^{2^k} P(\boldsymbol{a}_j \in K_t(\boldsymbol{y}) \mid \boldsymbol{a}_i). \end{aligned}$$

The set $K_t(\boldsymbol{y})$ denotes the sphere of all words around $\boldsymbol{y}$ of radius t (see also Definition 4.5). For the probability $P_{2|i}$ it is difficult to find an upper bound or $P_{2|i}$ may also be very large. However, an upper bound for $P_{2|i}$ is easy to derive, if the average over all randomly chosen codes is considered. Then $\boldsymbol{a}_j$ is uniformly distributed in $\{0,1\}^n$, i.e., for each arbitrary set $\mathcal{M}$,

$$P(\boldsymbol{a}_j \in \mathcal{M}) = 2^{-n} \cdot |\mathcal{M}|. \tag{3.6.6}$$

The spheres of radius t contain $\displaystyle\sum_{r=0}^{t} \binom{n}{r}$ words, according to (4.2.3), thus

$$P_{2|i} \le \sum_{\substack{j=1\\ j\neq i}}^{2^k} 2^{-n}|K_t(\boldsymbol{y})| \le 2^{k-n}|K_t(\boldsymbol{y})| = 2^{k-n}\sum_{r=0}^{t}\binom{n}{r}. \tag{3.6.7}$$

Let $\lambda = \dfrac{t}{n} = p_e + \beta$. Because of (3.6.2), $\lambda \le \dfrac{1}{2}$ and according to Theorem A.1,

$$\begin{aligned}
P_{2|i} &\le 2^{k-n} \cdot 2^{nH_2(\lambda)} \\
&= 2^{n(R-1+H_2(\lambda))} \\
&\le 2^{n(C-\varepsilon'/2-1+H_2(p_e+\beta))} \qquad \text{according to (3.6.5)} \\
&= 2^{n(H_2(p_e+\beta)-H_2(p_e)-\varepsilon'/2)} \\
&\le 2^{n(\varepsilon'/4-\varepsilon'/2)} \qquad \text{according to (3.6.2)} \\
&= 2^{-n\varepsilon'/2} \\
&\le \frac{\varepsilon}{2} \qquad \text{according to (3.6.3).}
\end{aligned}$$

Step 7. In summary, the results of steps 5 and 6 imply that

$$P_{w|i} \le P_{1|i} + P_{2|i} \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

The rest is quite simple:

$$P_w = \sum_{i=1}^{2^n} \underbrace{P(\text{decoding error} \mid \boldsymbol{a}_i \text{ transmitted})}_{=\,P_{w|i} \,\le\, \varepsilon} \cdot \underbrace{P(\boldsymbol{a}_i \text{ transmitted})}_{=\,2^{-n}} \;\le\; \varepsilon.$$

So this is valid for the average over all randomly chosen codes. Trivially, there must be at least one code which is as good as the mean. The larger n is, the stronger the code properties are concentrated around the expected value. Therefore, as already mentioned in Subsection 3.2.1, *almost all codes are good.* The converse theorem requires an independent proof, which we will leave out here. ■

To help with the interpretation of the main ideas we review the proof. On average the received word $\boldsymbol{y}$ contains np_e errors. For large n this mean is approximately identical to the actual number of errors. A decoding only takes place, if there is only one codeword in $K_t(\boldsymbol{y})$. So there are two possible wrong decodings:

$P_{1|i}$: the transmitted codeword is not in $K_t(\boldsymbol{y})$.

$P_{2|i}$: a different codeword than the transmitted codeword is in $K_t(\boldsymbol{y})$.

Depending on the choice of the threshold $t = np_e + n\beta$, two scenarios are possible:

- as $t \to np_e$ or as $\beta \to 0$, (3.6.3) is not fulfilled. The sphere becomes so small that the transmitted codeword is often outside of it, hence $P_{1|i}$ is large and $P_{2|i}$ is small.
- for $t \gg np_e$ or $\beta \gg 0$, (3.6.2) is not fulfilled. The sphere becomes so large that a wrong codeword is contained in it too often, hence $P_{1|i}$ is small and $P_{2|i}$ is large.

The suitable choice of β for a fixed n (or for a given $\beta \approx 0$ and an accordingly large n) makes $P_{1|i}$ as well as $P_{2|i}$ small. The principle introduced in step 3 is also called *bounded-distance decoder* (BDD). A comparison to the maximum-likelihood method and a further decoding method will be given in Subsection 4.2.4.

A very simplified approach without the random coding argument also leads to the channel capacity: on average the received word $\boldsymbol{y}$ contains np_e errors. The number of error patterns of weight np_e is $\binom{n}{np_e} \approx 2^{nH_2(p_e)}$ according to Theorem A.1. Per channel input there are $2^{nH_2(p_e)}$ outputs which are highly likely, and a total of 2^n possible outputs. For the number of inputs with different outputs,

$$2^k = \frac{2^n}{2^{nH_2(p_e)}} = 2^{n(1-H_2(p_e))} = 2^{nC},$$

thus $R = k/n = C$. A similar conclusion also leads to Theorem 3.4 for the baseband AWGN channel. The received value $y = x + \nu$ has the variance $\sigma_y^2 = E_s + N_0/2$, therefore $E(\|\boldsymbol{y}\|^2) = n\sigma_y^2$, hence $\boldsymbol{y} \in K_{\sqrt{n\sigma_y^2}}(\boldsymbol{0})$ for random $\boldsymbol{x}$. Correspondingly, $\boldsymbol{y} \in K_{\sqrt{nN_0/2}}(\boldsymbol{0})$ for a given $\boldsymbol{x}$. As for the BSC, the implication is that

$$2^k = \frac{|K_{\sqrt{n\sigma_y^2}}(\boldsymbol{0})|}{|K_{\sqrt{nN_0/2}}(\boldsymbol{0})|} = \frac{c_n(n\sigma_y^2)^{n/2}}{c_n(nN_0/2)^{n/2}} = \left(1 + 2\frac{E_s}{N_0}\right)^{n/2},$$

where $c_n t^n$ is the content of the n-dimensional sphere of radius t. Thus $k = n/2 \cdot \log_2(1 + 2E_s/N_0)$ which is the statement of Theorem 3.4.

3.7 Appendix: Proof of the R_0 Theorem for the DMC

The R_0 Theorem 3.3 can be fairly easily proved for the general DMC, again by using the random coding argument. However, the proof is quite different from that of the channel coding theorem. With the same methods, as used here for Theorem 3.3, we will later prove the union bound in Theorem 4.16. First we will formulate a lemma:

Lemma 3.1. *Let $h(v,w)$ be an arbitrary function of the variables $v \in \mathcal{A}$ and $w \in \mathcal{B}$. Then for each $\boldsymbol{w} = (w_0, \dots, w_{n-1}) \in \mathcal{B}^n$,*

$$\sum_{(v_0,\dots,v_{n-1})\in\mathcal{A}^n} \prod_{r=0}^{n-1} h(v_r, w_r) = \prod_{r=0}^{n-1} \sum_{v\in\mathcal{A}} h(v, w_r). \tag{3.7.1}$$

Proof. Obviously, the following is valid

$$\begin{aligned} &\sum_{v_0\in\mathcal{A}} \cdots \sum_{v_{n-1}\in\mathcal{A}} h(v_0,w_0)\cdots h(v_{n-1},w_{n-1}) \\ &= \left(\sum_{v_0\in\mathcal{A}} h(v_0,w_0)\right)\cdots\left(\sum_{v_{n-1}\in\mathcal{A}} h(v_{n-1},w_{n-1})\right). \end{aligned}$$

The upper term corresponds to the left side and the lower term corresponds to the right side of (3.7.1). ■

Proof of Theorem 3.3 in several steps:

Step 1. Let $\mathcal{C} = \{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ be an arbitrary code with 2 codewords, denoted $\boldsymbol{a}_i = (a_{i,0}, \dots, a_{i,n-1}) \in \mathcal{A}_{\text{in}}^n$. Maximum-likelihood (ML) decoding, according to Theorem 1.2, leads to

$$\begin{array}{lcllcl} P(\boldsymbol{y}|\boldsymbol{a}_1) & > & P(\boldsymbol{y}|\boldsymbol{a}_2) & \Longrightarrow & \hat{\boldsymbol{a}} = \boldsymbol{a}_1 \\ P(\boldsymbol{y}|\boldsymbol{a}_1) & < & P(\boldsymbol{y}|\boldsymbol{a}_2) & \Longrightarrow & \hat{\boldsymbol{a}} = \boldsymbol{a}_2. \end{array}$$

Then for $P_i = P(\text{decoding error} \mid \boldsymbol{a}_i \text{ transmitted})$,

$$\begin{aligned} P_1 &= P\Big(P(\boldsymbol{y}|\boldsymbol{a}_1) < P(\boldsymbol{y}|\boldsymbol{a}_2)\Big|\boldsymbol{a}_1\Big) \\ &= \sum_{\substack{\boldsymbol{y}\in\mathcal{A}_{\text{out}}^n \\ P(\boldsymbol{y}|\boldsymbol{a}_1)<P(\boldsymbol{y}|\boldsymbol{a}_2)}} P(\boldsymbol{y}|\boldsymbol{a}_1) \\ &\le \sum_{\substack{\boldsymbol{y}\in\mathcal{A}_{\text{out}}^n \\ P(\boldsymbol{y}|\boldsymbol{a}_1)<P(\boldsymbol{y}|\boldsymbol{a}_2)}} P(\boldsymbol{y}|\boldsymbol{a}_1)\sqrt{\frac{P(\boldsymbol{y}|\boldsymbol{a}_2)}{P(\boldsymbol{y}|\boldsymbol{a}_1)}} \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{\boldsymbol{y}\in\mathcal{A}_{\text{out}}^n} \sqrt{P(\boldsymbol{y}|\boldsymbol{a}_1)P(\boldsymbol{y}|\boldsymbol{a}_2)} \\
&= \sum_{\boldsymbol{y}\in\mathcal{A}_{\text{out}}^n} \prod_{r=0}^{n-1} \sqrt{P(y_r|a_{1,r})P(y_r|a_{2,r})} \quad \text{with (1.3.2).}
\end{aligned}
$$

In Lemma 3.1, we set $v = y \in \mathcal{A}_{\text{out}}$ and $\boldsymbol{w} = (a_1, a_2) \in \mathcal{A}_{\text{in}}^2$, then

$$
P_1 \leq \prod_{r=0}^{n-1} \underbrace{\sum_{y\in\mathcal{A}_{\text{out}}} \sqrt{P(y|a_{1,r})P(y|a_{2,r})}}_{= J(a_{1,r},a_{2,r})} .
$$

Of course this result is also valid for P_2.

Step 2. For the code consisting of 2 codewords,

$$
\begin{aligned}
P_w &= P(\text{decoding error}) \\
&= P_1 \cdot P(\boldsymbol{a}_1 \text{ transmitted}) + P_2 \cdot P(\boldsymbol{a}_2 \text{ transmitted}) \\
&\leq \prod_{r=0}^{n-1} J(a_{1,r}, a_{2,r}).
\end{aligned}
$$

The code $\mathcal{C} = \{\boldsymbol{a}_1, \boldsymbol{a}_2\}$ is randomly chosen such that the total of $2n$ code symbols are statistically independent with the distribution P_x, which maximizes R_0. For the expected value of P_w,

$$
P_w \leq \sum_{(\boldsymbol{a}_1,\boldsymbol{a}_2)\in\mathcal{A}_{\text{in}}^{2n}} \prod_{r=0}^{n-1} P_x(a_{1,r})P_x(a_{2,r})J(a_{1,r}, a_{2,r}).
$$

In Lemma 3.1, we set $\boldsymbol{v} = (a_1, a_2) \in \mathcal{A}_{\text{in}}^2$ (w is omitted), then

$$
\begin{aligned}
P_w &\leq \prod_{r=0}^{n-1} \sum_{(a_1,a_2)\in\mathcal{A}_{\text{in}}^2} P_x(a_1)P_x(a_2)J(a_1, a_2) \\
&= \left[\sum_{(a_1,a_2)\in\mathcal{A}_{\text{in}}^2} P_x(a_1)P_x(a_2)J(a_1, a_2) \right]^n \\
&= \left[\sum_{y\in\mathcal{A}_{\text{out}}} \sum_{(a_1,a_2)\in\mathcal{A}_{\text{in}}^2} P_x(a_1)P_x(a_2)\sqrt{P(y|a_1)P(y|a_2)} \right]^n \\
&= \left[\sum_{y\in\mathcal{A}_{\text{out}}} \left(\sum_{a\in\mathcal{A}_{\text{in}}} P_x(a)\sqrt{P(y|a)} \right)^2 \right]^n = 2^{-nR_0} \quad \text{with (3.2.9).}
\end{aligned}
$$

Step 3. Let $\mathcal{C} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{q^k}\}$ be an arbitrary $(n,k)_q$ code with the code rate $R_q = k/n \cdot \log_2 q$, then

$$\begin{aligned} P_w &= \sum_{i=1}^{q^k} P(\text{decoding error} \mid \boldsymbol{a}_i \text{ transmitted}) \cdot P(\boldsymbol{a}_i \text{ transmitted}) \\ &\le \sum_{i=1}^{q^k} P(\boldsymbol{a}_i \text{ transmitted}) \cdot \sum_{\substack{j=1 \\ j\neq i}}^{q^k} \underbrace{P(\text{decision for } \boldsymbol{a}_j | \boldsymbol{a}_i \text{ transmitted})}_{\le 2^{-nR_0}} \\ &\le \sum_{i=1}^{q^k} P(\boldsymbol{a}_i \text{ transmitted}) \cdot (q^k - 1) \cdot 2^{-nR_0} \\ &\le q^k \cdot 2^{-nR_0} \;=\; 2^{-n(R_0 - R_q)}. \end{aligned}$$

So this is valid for the mean of all randomly chosen codes. Trivially, there must be at least one code which is as good as the mean. ■

3.8 Problems

3.1. Assume a BSC with the bit-error probability p_e and a binary input with the distribution described by $P(x=0) = \alpha$ and $P(x=1) = 1-\alpha$. The output is denoted y. Calculate the entropies $H(x)$, $H(y)$, $H(y|x)$, the mutual information $I(x;y)$ as well as the channel capacity C. Draw $I(x;y)$ as a function of α for various parameters p_e.

3.2. Determine R_0 for the erasure channel BEC (see (1.3.10)).

3.3. Determine the channel capacity for the binary DMC with the transition probability

$$P_{y|x}(\eta|\xi) = \left\{ \begin{array}{ll} p_e & \eta = 0 \\ 1-p_e & \eta = 1 \end{array} \right\}.$$

3.4. Calculate the channel capacity for the binary channel with the transition probability

$$P_{y|x}(\eta|\xi) = \left\{ \begin{array}{ll} 1 & \xi = 0, \eta = 0 \\ 0 & \xi = 0, \eta = 1 \\ 1/2 & \xi = 1 \end{array} \right\}.$$

This model is called *Z-channel* since the diagram of the transition probability (as in Figure 1.3) looks like the letter Z.

3.5. For the q-ary symmetric DMC according to (1.3.3), prove that the channel capacity is $C = \log_2 q - H_q(p_e)$, where

$$H_q(p_e) = p_e \log_2(q-1) + H_2(p_e)$$

denotes an extension of the binary entropy function.

3.6. Prove for a binomially distributed random variable x with the distribution

$$P(x = l) = \binom{n}{l} p_e^l (1 - p_e)^{n-l} \quad , \quad l = 0, \ldots, n$$

the upper bound of the entropy $H(x) \leq nH_2(p_e)$.

3.7. Calculate the differential entropy of the two following distributions which are represented by their probability density functions and interpret the result:

$$f_x(\xi) = \left\{ \begin{array}{ll} 1/a & 0 < \xi < a \\ 0 & \text{otherwise} \end{array} \right\}, \quad f_x(\xi) = \left\{ \begin{array}{ll} 1/(\xi \ln^2 \xi) & \xi > e \\ 0 & \xi < e \end{array} \right\}.$$

3.8. Over a BSC, 10000 encoded bits per second can be transmitted with an error rate of 0.09. By using coding, can 4800 information bits per second be transmitted with an error rate of 10^{-5}? What is the situation for an error rate of 10^{-12}?

3.9. Assume a bit-error probability of $P_b = 10^{-5}$. (1) How large is the maximum theoretically possible gain for a binary transmission over the power-limited AWGN channel? (2) ... with a restriction to $R = R_0$? (3) ... with a restriction to $R = R_0$ and 4-times bandwidth expansion?

3.10. Prove (similar as for Theorem 3.1) that the average number of codewords in a sphere of radius t about an arbitrary word $\boldsymbol{y}$ is

$$q^{k-n} \cdot |K_t(\boldsymbol{y})| \;=\; q^{k-n} \cdot \sum_{r=0}^{t} \binom{n}{r} (q-1)^r.$$

How large is the average number of codewords on the surface of the sphere?