Error Probability Analysis of Binary Asymmetric Channels

Intermediate Report of NSC Project

"Finite Blocklength Capacity"

Date:	31 May 2009
Project-Number:	NSC 97-2221-E-009-003-MY3
Project Duration:	1 August 2008 – 31 July 2011
Funded by:	National Science Council, Taiwan

Author:	Stefan M. Moser
Co-Authors:	Po-Ning Chen, Hsuan-Yin Lin
Organization:	Information Theory Laboratory Department of Communication Engineering National Chiao Tung University
Address:	Engineering Building IV, Office 727 1001 Ta Hsueh Rd. Hsinchu 30010, Taiwan
E-mail:	stefan.moser@ieee.org

Abstract

In his world-famous paper of 1948, Shannon defined *channel capacity* as the ultimate rate at which information can be transmitted over a communication channel with an error probability that will vanish if we allow the blocklength to get infinitely large. While this result is of tremendous theoretical importance, the reality of practical systems looks quite different: no communication system will tolerate an infinite delay caused by an extremely large blocklength. On the other hand, it is not necessary to have an error probability that is exactly zero either, a small, but finite value will suffice.

Therefore, the question arises what can be done in a practical scheme. In particular, what is the maximal rate at which information can be transmitted over a communication channel for a given fixed maximum blocklength (*i.e.*, a fixed maximum delay) if we allow a certain maximal probability of error? In this project, we have started to study these questions.

Keywords: Channel capacity, binary asymmetric channel (BAC), finite blocklengths, good codes, probability of error.

Contents

1 Introduction

Since the analytical study of communication over a channel is very difficult even if we restrict ourselves to discrete memoryless channels (DMCs). Most known results are derived using the mathematical trick of considering some limits, in particular, usually it is assumed that the blocklength tends to infinity. The insights that have been achieved are considerable, however, it still remains an open question how far these asymptotic results can be applied to the practical scenario where the blocklength usually is strongly restricted.

In this work we are trying to study the much more difficult setup with a communication channel of very short blocklength. Currently, we have restricted ourselves to binary memoryless channels. This report summarizes some of our first preliminary results.

In the subsequent section we will review some basic definitions needed in the context of reliable communication. This section is mainly based Shannon's first land-mark paper [1]. In Section 3 we give some preliminary definitions and results; Section 4 summarizes our results; and we conclude in Section 5.

2 Definitions

2.1 Discrete Memoryless Channel

The probably most fundamental model describing communication over a noisy channel is the so-called *discrete memoryless channel (DMC)*. A DMC consists of a

- a finite input alphabet $\mathcal{X}$;
- a finite output alphabet $\mathcal{Y}$; and

- a **conditional probability distribution** $P_{Y|X}(\cdot|x)$ for all $x \in \mathcal{X}$ such that

$$P_{Y_k|X_1,X_2,\ldots,X_k,Y_1,Y_2,\ldots,Y_{k-1}}(y_k|x_1,x_2,\ldots,x_k,y_1,y_2,\ldots,y_{k-1}) = P_{Y|X}(y_k|x_k) \qquad \forall\, k. \quad (1)$$

Note that a DMC is called *memoryless* because the current output Y_k depends only on the current input x_k. Moreover also note that the channel is *time-invariant* in the sense that for a particular input x_k, the distribution of the output Y_k does not change over time.

Definition 1. *We say a DMC is used* without feedback*, if*

$$P(x_k|x_1,\ldots,x_{k-1},y_1,\ldots,y_{k-1}) = P(x_k|x_1,\ldots,x_{k-1}) \quad \forall\, k, \quad (2)$$

i.e.*, X_k depends only on past inputs (by choice of the encoder), but not on past outputs! $\Longrightarrow$ There is no feedback link from the receiver back to the transmitter that would inform the transmitter about the last outputs!*

Note that even though we assume the channel to be memoryless, we do *not* restrict the encoder to be memoryless!

We now have the following theorem.

Theorem 2. *If a DMC is used without feedback, then*

$$P(y_1,\ldots,y_n|x_1,\ldots,x_n) = \prod_{k=1}^{n} P_{Y|X}(y_k|x_k) \quad \forall\, n \geq 1. \quad (3)$$

In the following we will concentrate on the special cases of *binary* DMCs, *i.e.*, we restrict our channel alphabets to be binary.

The most known example of a binary DMC is the *binary symmetric channel (BSC)* shown in Figure 1.

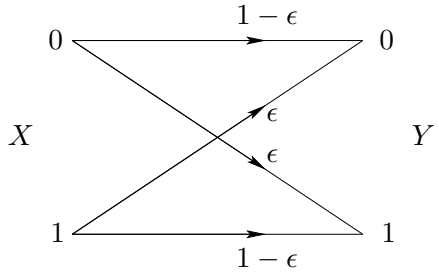

Figure 1: Binary symmetric channel (BSC).

Another important special case of a binary DMC is the Z-channel shown in Fig 2.

Both binary channel models BSC and Z-channel are both special cases of the most general binary DMC: the *binary asymmetric channel (BAC)* as shown in Figure 3.

2.2 Coding for DMC

Definition 3. *A* $(\mathcal{M}, n)$ code *for a DMC* $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$ *consists of*

- *the message set* $\mathcal{M} = \{1,\ldots,\mathcal{M}\}$ *of* $\mathcal{M}$ *equally likely random messages* M*;*

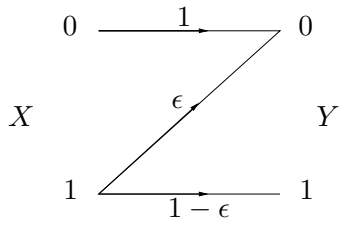

Figure 2: Z channel.

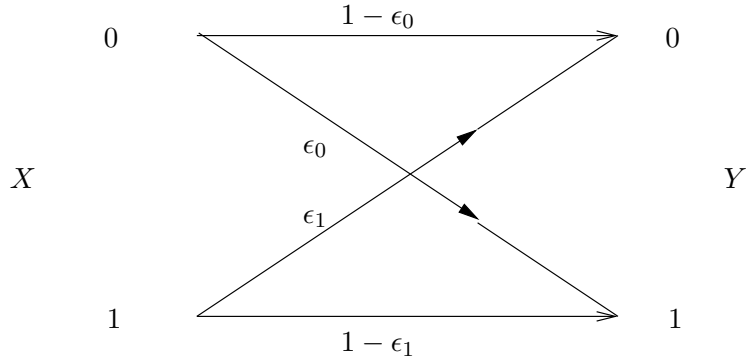

Figure 3: The binary asymmetric channel (BAC).

- *the codebook of* $\mathcal{M}$ *codewords which are channel input sequences of length* n*;*
- *an encoding function* ϕ *that assigns for every message* $m \in \mathcal{M}$ *a codeword* $\mathbf{x} = (x_1, \ldots, x_n)$*; and*
- *a decoding function* ψ *that maps the received channel output* n*-sequence* $\mathbf{y}$ *to a guess* $\hat{m} \in \mathcal{M}$*.*

Note that the two main parameters of interest of a code are the number of possible messages $\mathcal{M}$ (the larger, the more information is transmitted) and the blocklength n (the shorter, the less time is needed to transmit the message):

- we have $\mathcal{M}$ equally likely messages, *i.e.*, the entropy is $H(M) = \log_2 \mathcal{M}$ bits;
- we need n transmissions of a channel input symbol X_k over the channel in order to transmit the complete message.

Hence, we make the following definition:

Definition 4. *The* rate[1] *of a* $(\mathcal{M}, n)$ *code is defined as*

$$\mathcal{R} \triangleq \frac{\log_2 \mathcal{M}}{n} \text{ bits/transmission.} \tag{4}$$

However, this definition of a rate makes only sense if the message really arrives at the receiver, *i.e.*, if the receiver does not make a decoding error!

Definition 5. *Given that message* $M = m$ *has been sent, let* λ_m *be the* error probability*:*

$$\lambda_m \triangleq \Pr[\psi(Y_1^n) \neq m \mid X_1^n = \phi(m)] = \sum_{\mathbf{y} \in \mathcal{Y}^n} p(\mathbf{y}|\mathbf{x}(m)) I\{\psi(\mathbf{y}) \neq m\}, \tag{5}$$

[1] We define the rate here using a logarithm of base 2. However, we can use any logarithm as long as we adapt the units accordingly.

where $I\{\cdot\}$ is the indicator function

$$I\{\text{statement}\} = \begin{cases} 1 & \text{if statement is true,} \\ 0 & \text{if statement is wrong.} \end{cases}$$

The maximum error probability $\lambda^{(n)}$ *for an* $(\mathcal{M}, n)$ *code is defined as*

$$\lambda^{(n)} \triangleq \max_{m \in \mathcal{M}} \lambda_m. \tag{6}$$

The average error probability $P_{\mathrm{e}}^{(n)}$ *for an* $(\mathcal{M}, n)$ *code is defined as*

$$P_{\mathrm{e}}^{(n)} \triangleq \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} \lambda_m. \tag{7}$$

The most famous relation between code rate and error probability has been derived by Shannon in his landmark paper from 1948 [1].

Theorem 6 (The Channel Coding Theorem for a DMC). *Define*

$$\mathsf{C} \triangleq \max_{P_X(\cdot)} I(X;Y) \tag{8}$$

where X and Y have to be understood as input and output of a DMC and where the maximization is over all input distributions $P_X(\cdot)$.

Then for every $\mathsf{R} < \mathsf{C}$ there exists a sequence of $(2^{n\mathsf{R}}, n)$ codes with maximum error probability $\lambda^{(n)} \to 0$ as the blocklength n gets very large.

Conversely, any sequence of $(2^{n\mathsf{R}}, n)$ codes with maximum error probability $\lambda^{(n)} \to 0$ must have a rate $\mathsf{R} \leq \mathsf{C}$.

So we see that C denotes the maximum rate at which reliable communication is possible. Therefore C is called channel capacity*.*

Note that this theorem considers only the situation of n tending to infinity and thereby the error probability going to zero. However, in a practical system, we cannot allow the blocklength n to be too large because of delay and complexity. On the other hand it is neither necessary to have zero error probability.

So the question arises what we can say about "capacity" for finite n, *i.e.*, if we allow a certain maximal probability of error, what is the smallest necessary blocklength n to achieve it? Or, vice versa, fixing a certain short blocklength n, what is the best average error probability that can be achieved?

3 Preliminaries

3.1 Capacity of BAC

Without loss of generality, we only consider BACs with $0 \leq \epsilon_0 \leq \epsilon_1 \leq 1$. The capacity of a BAC is given by

$$\mathsf{C}_{\mathrm{BAC}} = -\frac{1-\epsilon_0}{1-\epsilon_0-\epsilon_1} \cdot H_{\mathrm{b}}(\epsilon_1) - \frac{-\epsilon_1}{1-\epsilon_0-\epsilon_1} \cdot H_{\mathrm{b}}(\epsilon_0) - \log_2 \left(\frac{1}{1 + e^{\frac{H_{\mathrm{b}}(\epsilon_1) - H_{\mathrm{b}}(\epsilon_0)}{1-\epsilon_0-\epsilon_1}}} \right) \tag{9}$$

where $H_{\mathrm{b}}(\cdot)$ is the binary entropy function defined as

$$H_{\mathrm{b}}(p) \triangleq -p \log_2 p - (1-p) \log_2 (1-p).$$

The input distribution $P_X^*(\cdot)$ that achieves this capacity is given by

$$P_X^*(0) = \frac{r \cdot (1-\epsilon_1) - \epsilon_1}{(1-\epsilon_0-\epsilon_1) + r \cdot (1-\epsilon_0-\epsilon_1)}, \tag{10}$$

$$P_X^*(1) = 1 - P_X^*(0). \tag{11}$$

3.2 Empirical Distributions

In [2], Shamai and Verdú introduce the *empirical distribution of good codes*. To illustrate the basic idea, we will consider the special case of a BSC with $\mathsf{C}_{\mathrm{BSC}} = 1 - H_{\mathrm{b}}(\epsilon)$ and with a capacity-achieving input distribution

$$P_X^*(0) = P_X^*(1) = \frac{1}{2}.$$

The unique n-dimensional distribution that maximizes the n-block input-output mutual information of a BSC puts equal mass on all 2^n binary n-strings. This indicates that a good code for the BSC must contain an asymptotically equal proportion of 0's and 1's (note that is has not been proved!).

First we consider channels with input alphabets $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The random transformation operating on n-tuples (the channel) is denoted by $W^{(n)} : \mathcal{X} \to \mathcal{Y}$.

Definition 7. *A* good code-sequence *for a channel with capacity* C *is a sequence of codes with vanishing error probability whose rate satisfies*

$$\liminf_{n\to\infty} \frac{1}{n} \log_2 \mathcal{M} = \mathsf{C}. \tag{12}$$

Note that the existence of good code-sequence is guaranteed by the definition of channel capacity.

In general, a codebook is composed of $\mathcal{M}$ codewords of blocklength n

$$\{z_{i,m} \in \mathcal{X} : i = 1, \ldots, n;\ m = 1, \ldots, \mathcal{M}\},$$

i.e.,

$$\mathcal{C} = \begin{pmatrix} z_{11} & z_{21} & \cdots & z_{n1} \\ z_{12} & z_{22} & \cdots & z_{n2} \\ \vdots & & \ddots & \vdots \\ z_{1M} & z_{2M} & \cdots & z_{nM} \end{pmatrix}_{\mathcal{M}\times n}.$$

The 1st-order empirical distribution can be found by computing the fraction of symbols in the the codeword equal to each input letter.

The distribution of the n-tuple codeword into the channel is defined on $\mathcal{X}^{(n)}$ as

$$P_{\hat{X}^n}(x_1, x_2, \ldots, x_n) \triangleq \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} \prod_{i=1}^{n} I\{z_{im} = x_i\}.$$

In addition to the joint distribution of the whole n-tuple, it is interesting to deal with the joint distribution of subcodewords $(x_j, \ldots, x_l)$, $1 \le j \le l \le n$:

$$P_{\hat{X}_j^l}(x_j, \ldots, x_l) \triangleq \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} \prod_{i=j}^{l} I\{z_{im} = x_i\}.$$

Now we are ready for the definition of the kth-order empirical distribution: it is defined analogously to the first order empirical distribution by computing for each k-string $(\alpha_1, \ldots, \alpha_k) \in \mathcal{X}^k$ the fraction of k-strings anywhere within the codeword equal to $(\alpha_1, \ldots, \alpha_k)$, averaged over all equiprobable codewords.

Definition 8. *The* kth-order empirical distribution *($1 \leq k \leq n$) of the code*

$$\{z_{im} \in \mathcal{X},\ i = 1, \ldots, n,\ m = 1, \ldots, \mathcal{M}\}$$

is

$$Q^{(k)}_{\hat{X}^n}(\alpha_1, \ldots, \alpha_k) = \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} P_{\hat{X}_i^{i+k-1}}(\alpha_1, \ldots, \alpha_k). \tag{13}$$

When a code with empirical distribution $P_{\hat{X}_i^{i+k-1}}$ is used as input to a channel $W^{(n)} : \mathcal{X}^{(n)} \to \mathcal{Y}^{(n)}$, the output distribution induced on $\mathcal{Y}^{(n)}$ is denoted by $P_{\hat{Y}_i^{i+k-1}}$, and the corresponding k-th order empirical output distribution induced by the code is

$$Q^{(k)}_{\hat{Y}^n}(\alpha_1, \ldots, \alpha_k) = \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} P_{\hat{Y}_i^{i+k-1}}(\alpha_1, \ldots, \alpha_k).$$

Note that $P_{\hat{X}_j^l}$ (or $P_{\hat{Y}_j^l}$) can be obtained as *marginal distribution* by summing out all the components apart from $(\alpha_j, \ldots, \alpha_l)$.

The meaning of $P_{\hat{X}_j^l}(x_1, x_2, \ldots, x_n)$ can be seen as putting mass $\frac{m}{\mathcal{M}}$ on $(x_j, \ldots, x_l)$, if there are m subcodewords equal to $(x_j, \ldots, x_l)$. In [2] it is shown that the input empirical distribution converges to the capacity-achieving input distribution in divergence sense as n gets very large. In the following we will compare the capacity-achieving distribution to the empirical distribution of *good codes* for finite blocklength n.

3.3 The n-Dependent Critical Rate

In [3] [4], Shannon's coding theorem has been further refined by giving a an exponential bound on the decay of the error probability in the blocklength n.

Theorem 9 (Gallager's Random Coding Bound for Block Codes). *Over any ensemble of block codes with $\mathcal{M} = e^{n\mathcal{R}}$ codewords of length n for use on a given DMC, wherein each code is assigned a probability in such a way that the codewords are pairwise independent and that the digits in the i-th codeword are statistically independent and each is distributed according to a given probability distribution Q over the channel input alphabet, then the average block error probability of these codes for ML decoding satisfies*

$$\mathsf{E}[P_{\mathrm{e}}] \leq 2^{-n(E_0(\rho,Q)-\rho\mathcal{R})} \tag{14}$$

for all ρ, such that $0 \leq \rho \leq 1$, where

$$E_0(\rho, Q) \triangleq -\log_2 \sum_{y \in \mathcal{Y}} \left(\sum_{x \in \mathcal{X}} Q_X(x) \cdot P_{Y|X}(y|x)^{\frac{1}{1+\rho}} \right)^{1+\rho}. \tag{15}$$

In particular,

$$\mathsf{E}[P_{\mathrm{e}}] \leq 2^{-nE_{\mathrm{G}}(\mathcal{R})} \tag{16}$$

where

$$E_{\mathrm{G}}(\mathcal{R}) \triangleq \max_{Q} \max_{0 \leq \rho \leq 1} \{E_0(\rho, Q) - \rho\mathcal{R}\}. \tag{17}$$

Note that in (17) for $0 \leq \mathcal{R} \leq \mathcal{R}_{\text{crit}}$, the maximum is achieved for $\rho = 1$ where the *critical rate* $\mathcal{R}_{\text{crit}}$ is defined as

$$\mathcal{R}_{\text{crit}} \triangleq \sup_Q \left.\frac{\partial E_0(\rho, Q)}{\partial \rho}\right|_{\rho=1} = \inf_{0\leq\rho\leq 1} \frac{E_0(1) - E_0(\rho)}{1-\rho} = \lim_{\rho\to 1} \frac{\partial E_0(\rho)}{\partial \rho}$$

with

$$E_0(\rho) \triangleq \sup_Q E_0(\rho, Q).$$

The Gallager exponent E_{G} and the critical rate $\mathcal{R}_{\text{crit}}$ do not depend on n.

Note that the derivation of Theorem 9 is based on the following bound on the average probability of error:

$$\mathsf{E}[P_{\text{e}}] \leq (\mathcal{M}-1)^{\rho} 2^{-nE_0(\rho,Q)}.$$

In this expression the first factor is then further weakened to

$$(\mathcal{M}-1)^{\rho} \leq \mathcal{M}^{\rho} = 2^{\rho n \mathcal{R}}$$

in order to get to the bound (14). If we do not follow this step we get an improved bound:

$$\mathsf{E}[P_{\text{e}}] \leq 2^{-n\left(E_0(\rho,Q) - \rho\frac{\log_2\left(2^{n\mathcal{R}}-1\right)}{n}\right)}. \tag{18}$$

Comparing this to (14), we define

$$\mathcal{R}'(n) \triangleq \frac{1}{n}\log_2\left(2^{n\mathcal{R}}-1\right). \tag{19}$$

Similarly, we define a new version of the critical rate as

$$\mathcal{R}'_{\text{crit}} \triangleq \frac{1}{n}\log_2\left(2^{n\mathcal{R}_{\text{crit}}}-1\right). \tag{20}$$

Note that this new version of the critical rate depends on n, but will converge to the original critical rate when n tends to infinity.

4 Main Results

4.1 Error Probability of BSC

Consider the situation of a BSC and assume that we transmit the m-th codeword $\mathbf{x}_m$, $1 \leq m \leq \mathcal{M}$, and that we receive $\mathbf{y}$. The ML decision is then

$$g(\mathbf{y}) \triangleq \arg\max_{1\leq i\leq\mathcal{M}} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}_i).$$

The average probability of error can be computed as

$$P_{\text{e}} = \frac{1}{\mathcal{M}}(1-\epsilon)^n \sum_{\mathbf{y}} \sum_{\substack{i=1 \\ i\neq g(\mathbf{y})}}^{\mathcal{M}} \left(\frac{\epsilon}{1-\epsilon}\right)^{d_{\text{H}}(\mathbf{x}_i,\mathbf{y})} \tag{21}$$

where $d_{\text{H}}(\cdot,\cdot)$ is the Hamming distance.

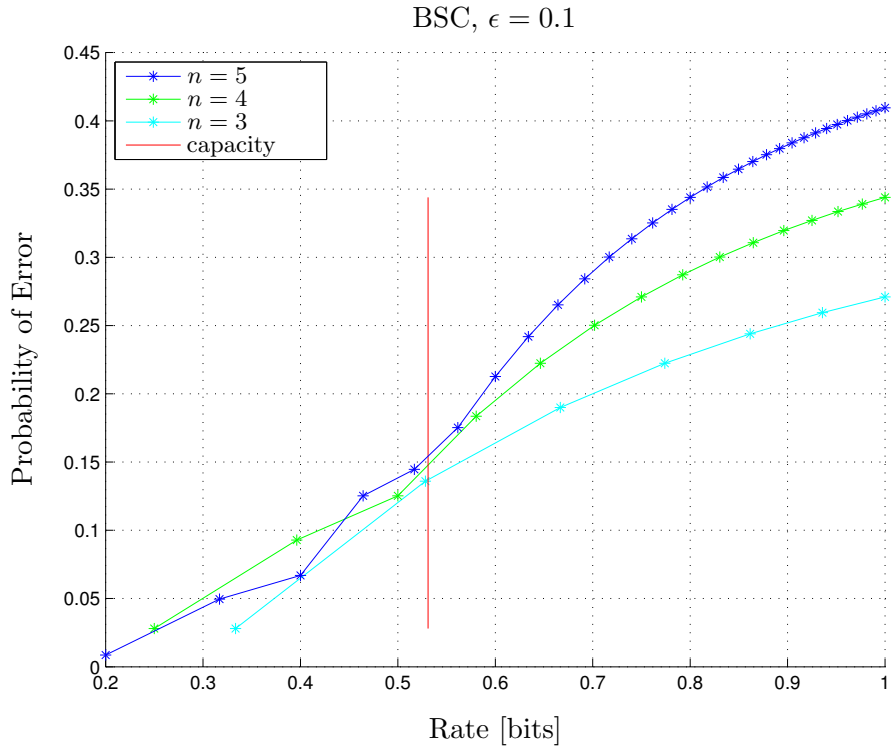

Figure 4: The average error probability of the best code versus the rate for a BSC with cross-over probability $\epsilon = 0.1$.

In Figure 4, we plot the *best* average error probability versus the rate for block-length $n = 3$, $n = 4$, and $n = 5$. Note that we find this *best error probability* by checking through **all possible** codes (including both linear and nonlinear codes).

It is interesting to note that there seems to be a clear separation between the region above capacity where the error probability increases fast as a function of n and also as a function of $\mathcal{R}$, and a region below capacity where for a fixed rate the probability of error is not monotonically increasing in n.

Also note that the complexity of this computation grows very fast in n: for $\mathcal{M} = 4$ and

- for $n = 3$ there are $\binom{8}{4} = 70$ different codes;
- for $n = 4$ there are $\binom{16}{4} = 1820$ different codes;
- for $n = 5$ there are $\binom{32}{4} = 35960$ different codes, etc.

Since the BSC is **strongly symmetric**, it is not surprising that there are a huge number of codes that can achieve the best error probability. Among them many are linear.

4.2 Error Probability of BAC

To simplify our notation we introduce $d_{\alpha\beta}$ to be the number of positions j where where $x_j = \alpha$ and $y_j = \beta$, where as usual $\mathbf{x}$ is the sent codeword and $\mathbf{y}$ is the received sequence.

The conditional probability can then be written as

$$P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = (1-\epsilon_0)^{d_{00}} \cdot \epsilon_0^{d_{01}} \cdot \epsilon_1^{d_{10}} \cdot (1-\epsilon_1)^{d_{11}}.$$

Note that we can express these different d's as follows:

$$\begin{aligned} d_{11} &= \frac{1}{2}\text{sum}(\mathbf{x}+\mathbf{y}-|\mathbf{x}-\mathbf{y}|), \\ d_{10} &= \text{sum}(I\{\mathbf{x}-\mathbf{y}<0\}), \\ d_{01} &= \text{sum}(I\{\mathbf{y}-\mathbf{x}<0\}), \\ d_{00} &= n-d_{11}-d_{10}-d_{01}, \end{aligned}$$

where

$$\text{sum}(\mathbf{x}) \triangleq \sum_{j=1}^{n} x_j.$$

The error probability of a BAC can now be expressed as

$$P_{\text{e}} = \frac{1}{\mathcal{M}}(1-\epsilon_0)^n \sum_{\mathbf{y}} \sum_{\substack{i=1 \\ i\neq g(\mathbf{y})}}^{\mathcal{M}} \left(\frac{\epsilon_0}{1-\epsilon_0}\right)^{d_{01}} \left(\frac{\epsilon_1}{1-\epsilon_0}\right)^{d_{10}} \left(\frac{1-\epsilon_1}{1-\epsilon_0}\right)^{d_{11}}, \qquad (22)$$

where $g(\mathbf{y})$ is the ML decision for the observation $\mathbf{y}$.

In Figures 5 and 6, we plot the *best* average error probability versus the rate for blocklength $n=3$, $n=4$, and $n=5$. In Figure 5 we have chosen $\epsilon_0 = 0.1$

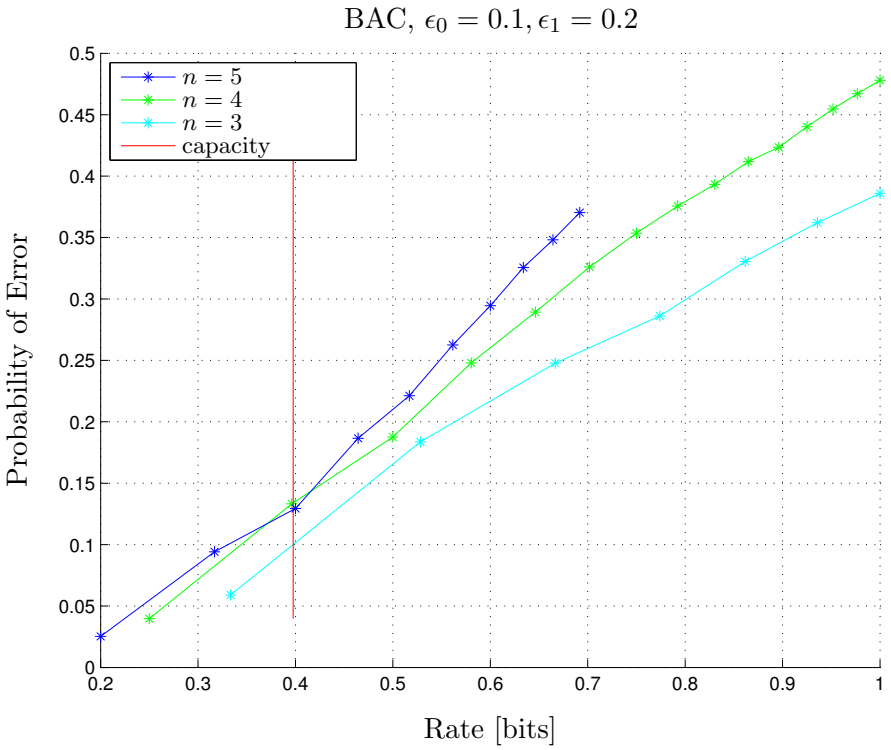

Figure 5: The average error probability of the best code versus the rate for a BAC with cross-over probabilities $\epsilon_0 = 0.1$ and $\epsilon_1 = 0.2$.

and $\epsilon_1 = 0.2$ (*i.e.*, $|\epsilon_0 - \epsilon_1| = 0.1$ is not very big), and in Figure 6 we have chosen $\epsilon_0 = 0.01$ and $\epsilon_1 = 0.4$ (*i.e.*, an extreme case with $|\epsilon_0 - \epsilon_1| = 0.39$ relatively large).

Now we can observe that the number of codes that achieves the best error probability is reduced drastically. This can be seen in Figure 7. Here, the vertical red line denotes the capacity of the channel, and the vertical dashed red line depicts the

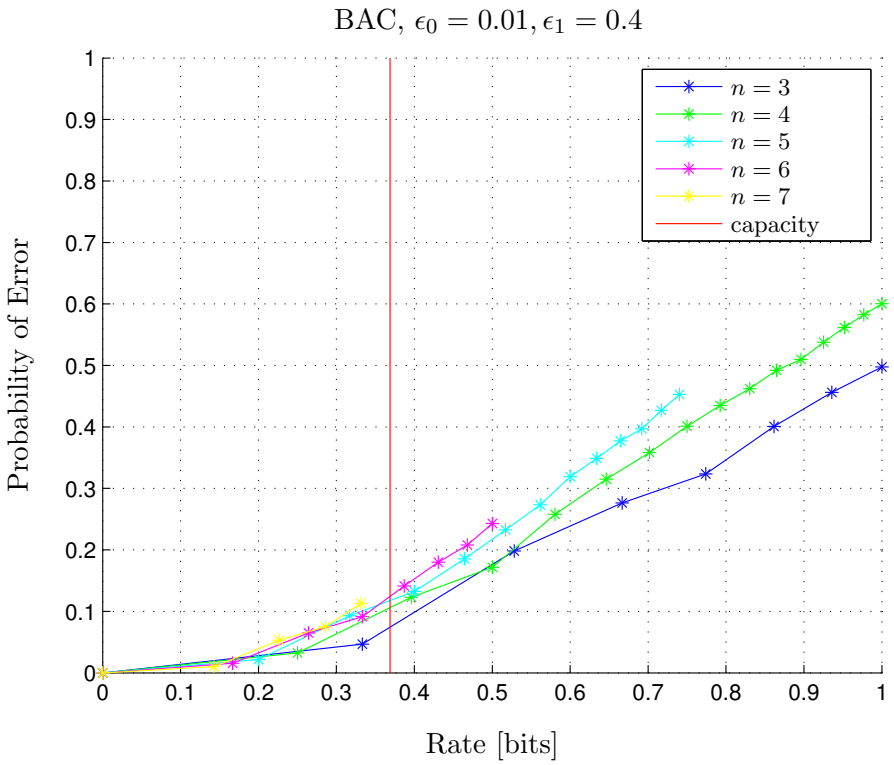

Figure 6: The average error probability of the best code versus the rate for a BAC with cross-over probabilities $\epsilon_0 = 0.01$ and $\epsilon_1 = 0.4$.

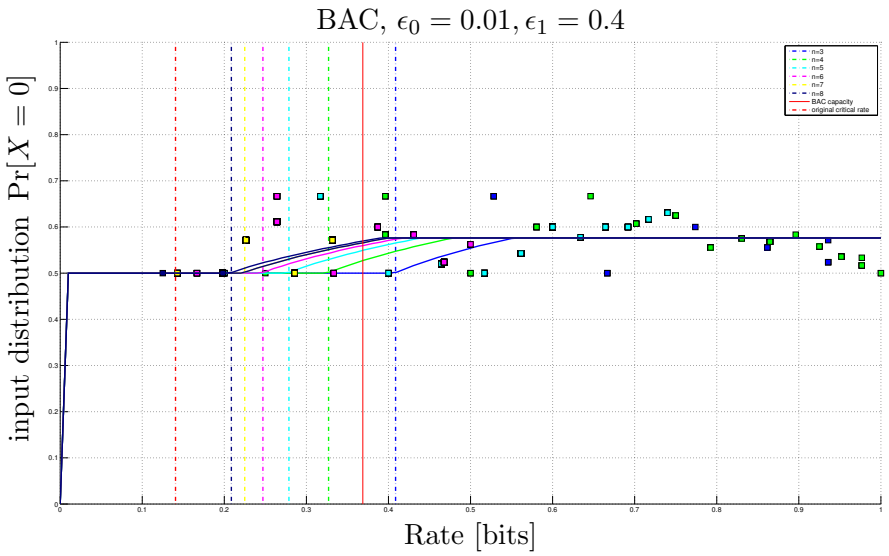

Figure 7: Optimum empirical input distributions of a BAC.

(original) critical rate. The other vertical dashed lines show the (n-dependent) new critical rate (20) for various values of n. Note that as n increases th n-dependent critical rate approaches the original one.

The horizontal lines describe the input distribution[2] Q that achieves the maximum in the Gallager exponent (17). We see that for small rates, a uniform distribution is optimal, while for larger rates (in particular for any rates above capacity if n is large) the optimum distribution is the capacity-achieving distribution, which for this BAC with $\epsilon_0 = 0.01, \epsilon_1 = 0.4$ is given by

$$\Pr[X = 0] = 0.5762.$$

The colored square boxes denote the empirical distribution of the best code, *i.e.*, the code that actually achieves the smallest possible probability of error for a given blocklength on this channel. Note that for certain rates the optimal code is unique, but for other rates, there are several optimal solutions.

Note that a closer investigation shows that some of these optimal codes are linear, but some are not.

We see that in the given example for each n and for any rate smaller than the n-dependent critical rate, the optimal code has a uniform empirical input distribution.

4.3 Error Probability of Z-channel

A special case of the BAC is the Z-channel where we have $\epsilon_1 = 0$ and $\epsilon_0 = \epsilon$. The probability of error in this case is

$$P_{\mathrm{e}} = \frac{1}{\mathcal{M}} \sum_{\mathbf{y}} \sum_{\substack{i=1 \\ i \neq g(\mathbf{y})}}^{\mathcal{M}} (1-\epsilon_0)^{z(\mathbf{x})} \left(\frac{\epsilon_0}{1-\epsilon_0} \right)^{d_{01}} \cdot I\{\text{if } x_j = 1 \Rightarrow y_j = 1, \ \forall\, j\} \tag{23}$$

where

$$z(\mathbf{x}) \triangleq \text{the number of 0's in } \mathbf{x}.$$

Similarly to before, in Figure 8 we plot the best error probability versus rate.

In Figure 9 we again depict the empirical distribution of the optimum code for various n, including the n-dependent critical rates and the Gallager-bound achieving input distribution. Note that the capacity-achieving distribution is

$$\Pr[X = 0] = \frac{2}{5}.$$

Again we see that for each n and for any rate smaller than the n-dependent critical rate, the optimal code has a uniform empirical input distribution.

5 Discussion & Conclusion

We have seen that the behavior of the optimal code for finite, small blocklength n is much less clear than what the asymptotic results predict. While asymptotically, we have a huge number of possible codes that achieve the optimal performance, many of which are linear, this is not anymore the case for finite blocklength and a BAC (for BSC it still holds because of the many symmetries in the channel itself). It depends

[2] Actually, since we are in a binary situation, it is sufficient to depict the probability of a zero input $\Pr[X = 0]$.

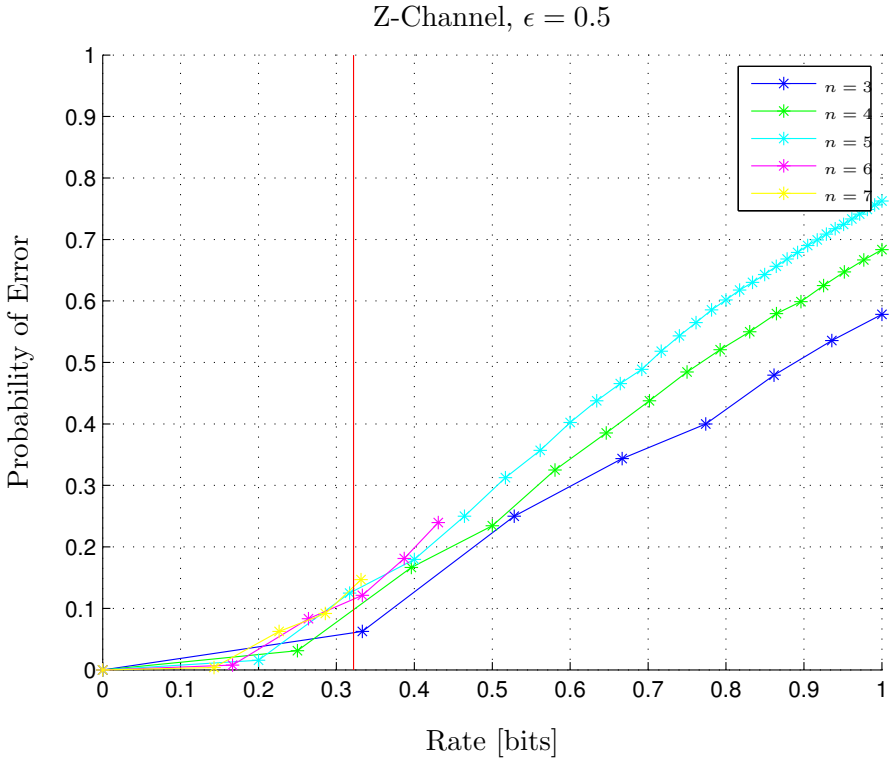

Figure 8: The average error probability of the best code versus the rate for a Z-channel with cross-over probability $\epsilon = 0.5$.

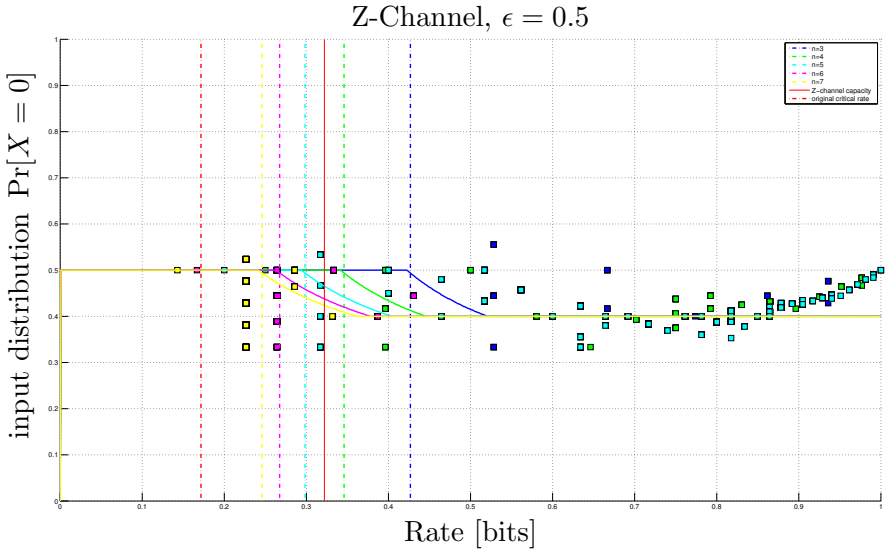

Figure 9: Optimum empirical input distributions of a Z-channel.

very much on the exact channel parameters and the exact value of the blocklength n whether a linear code is optimal or not.

Interestingly, for the most asymmetric channel, the Z-channel with $\epsilon = 0.5$, there are many values of the blocklength for which a linear code is optimal.

We also note that the empirical distribution of a good code for finite n is often not close to the capacity-achieving input distribution. We believe that the techniques that lead to good codes for rates close to capacity and large blocklengths might be very suboptimal in the situation of codes of small rates and small blocklength.

In the situation of a BSC we see a very clear separation between the behavior of codes with rates above and below capacity. Above capacity the probability of error of the best code increases quickly as a function of n, while for codes with a rate below capacity this is not true. However, once we consider less symmetric channels this behavior is less obvious.

Finally, we point out once more that the behavior of codes of finite n is rather different from the asymptotic behavior once we let n tend to infinity. It has been shown in [2] that for rates close to capacity and large n, the empirical distributions of good codes are close to the capacity-achieving input distribution. This is not anymore true once we consider small n and rates much smaller than capacity.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

[2] S. Shamai (Shitz) and S. Verdú, "The empirical distribution of good codes," *IEEE Transactions on Information Theory*, vol. 43, no. 3, pp. 836–846, May 1997.

[3] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels," *Information and Control*, pp. 65–103, December 1966, part I.

[4] ——, "Lower bounds to error probability for coding on discrete memoryless channels," *Information and Control*, pp. 522–552, May 1967, part II.