Probabilistic Amplitude Shaping

Georg Böcherer

Huawei Technologies, Germany; georg.boecherer@ieee.org

ABSTRACT

Probabilistic amplitude shaping (PAS) proposed in Böcherer, Steiner, Schulte [24] is a practical architecture for combining non-uniform distributions on higher-order constellations with off-the-shelf forward error correction (FEC) codes. PAS consists of a distribution matcher (DM) that imposes a desired distribution on the signal point amplitudes, followed by systematic FEC encoding, preserving the amplitude distribution. FEC encoding generates additional parity bits, which select the signs of the signal points. At the receiver, FEC decoding is followed by an inverse DM. PAS quickly had a large industrial impact, in particular in fiber-optic communications. This monograph details the practical considerations that led to the invention of PAS and provides an information-theoretic assessment of the PAS architecture. Because of the separation into a shaping layer and an FEC layer, the theoretic analysis of PAS requires new tools. On the shaping layer, the cost penalty and rate loss of finite length DMs is analyzed. On the FEC layer, achievable FEC rates are derived. Using mismatched decoding, achievable rates are studied for decoding metrics of practical importance. Combining the findings, it is shown that PAS with linear codes is capacity-achieving on a class of discrete input channels. Open questions for future study are discussed.

Georg Böcherer (2023), "Probabilistic Amplitude Shaping", Foundations and Trends® in Communications and Information Theory: Vol. 20, No. 4, pp 390–511. DOI: 10.1561/0100000111.

Preface

Almost 10 years ago, we simulated for the first time a communication system architecture that we later called Probabilistic Amplitude Shaping[1], the title of this monograph.

How to use this monograph All readers should read Section 1: it discusses the line of thoughts that led to the invention of PAS, outlines this monograph, and provides pointers to the literature.

The **theorist** may then read the discussion sections provided at the ends of Sections 2–5. The discussion sections summarize the sections, provide pointers to the literature, and outline open problems for future study. Also of interest to the theorist may be some of the proof techniques. For instance, the study of cost and rate scaling of distribution matchers[2] in Section 2.6, the layered probabilistic shaping (PS) random code ensemble in Section 3.1, the "any channel" achievable forward error correction (FEC) rate in Theorem 3.2, and the derivation of the PAS error exponent in Section 5.

The **practitioner** may implement the formulas provided throughout the monograph for numerical evaluation as guidance for designing PAS systems for industrial application. He/she may consult the PAS webpage (see below) to check for available implementations and may also consider contributing his/her implementations. For instance, one may use the

[1] We introduced the name Probabilistic Amplitude Shaping (PAS) in [24].
[2] We introduced the term Distribution Matching (DM) in [22].

formulas from Section 2 for choosing a DM class and dimensioning DM input and output lengths for trading rate and cost against latency and complexity. Also, the practitioner may implement the cross-equivocation formulas from Section 4 to compare the performance limits of binary and nonbinary codes, to choose between hard-decision and soft-decision, or to select the resolution for quantized soft-decision decoding. Similarily, one may implement the PAS achievable rate formulas from Section 5 for assessing the performance penalty caused by a constrained FEC rate, or for plotting PAS rate limits for finite length at a required reliability. Also of practical interest are the PAS system parameters FEC overhead, shaping set rate, and PS overhead as discussed in Sections 3.3 and 4.1.

For the **lecturer**, the cross-equivocation formalism from Section 4 may be of interest. Besides the basic decoding metrics discussed in this monograph, one can easily come up with many more variations, which according to my own teaching experience provide a rich source for homework and exam questions.

The **machine learning engineer** may find interest in the cross-equivocation formalism from Section 4. The underlying empirical cross-equivocation defined in Section 3 is identical to the cross-entropy loss frequently used in machine learning. Thus, the discussion in this monograph may provide the machine learning engineer with an interesting communication system perspective on the cross-entropy loss.

Webpage One shortcoming of this monograph is an insufficient number of plots with numerical evaluations for illustrating the developed concepts. I just did not have the time to add all the illustrations I would like to have. I have therefore set up a webpage[3] to accompany this monograph, for the following purposes:

- To host implementations of formulas and algorithms provided by the community.
- To share numerical plots of performance evaluations provided by the community.
- To publish the errata of this monograph.

[3] https://github.com/gbsha/PAS

I hope this provides an effective alternative to providing numerical evaluations in the monograph.

Acknowledgments Prof. Valdemar da Rocha and Prof. Cecilio Pimentel suggested to me as a master thesis topic the study of the discrete noiseless channel at their chair at the Federal University of Pernambuco. This triggered my interest in constrained coding and led to my study of variable length DM algorithms during my PhD at Prof. Rudolf Mathar's chair at the RWTH Aachen University. The work of Prof. David MacKay and his students (in particular the MacKay-Neal codes[4] and the sparse-dense codes[5]) inspired me to combine DM and FEC. Prof. Alex Alvarado brought my interest to the study of bit-interleaved coded modulation. Coded modulation in general was brought to my attention by Gottfried Ungerboeck when I served as his teaching assistant during the first months of my postdoc at Prof. Gerhard Kramer's chair at the Technical University of Munich.

The invention of PAS resulted in an exciting time with great people. Some memories are: Studying variable length DMs with Rana Ali Amjad and Sebastian Baur; Prof. Stephan ten Brink looking at an early PAS diagram and understanding it faster than anyone else before or after; a discussion with Irina Bocharova and Boris Kudryashov that led to the development of constant composition distribution matching (CCDM) by Patrick Schulte; the first implementation of PAS for a simulated optical transmission with Tobias Fehenberger; Gianluigi Liva asking whether one could change the DM distribution to adjust the PAS rate; the first PAS optical transmission experiment with Fred Buchali and Prof. Laurent Schmalen; the Bell Labs Prize 2015 together with Fabian Steiner and Patrick Schulte; Prof. Richard Wesel suggesting to change "rate-compatible" for "rate-matched" in the title of the PAS paper; working with Bernhard Geiger on quantization for distribution synthesis; Tobias Prinz developing polar coded PAS; the suggestion of Prof. Frans Willems to use sequences up to a maximum cost for DM, which led to the development of minimum cost distribution matching

[4] MacKay [54, Section VI].
[5] Ratzer [61, Chapter 5].

(MCDM) by several groups; the Johann-Philipp-Reis-Preis 2017; the collaboration with Prof. Neri Merhav on error exponents for layered PS; Huijian Zhang and Zhuhong Zhang appreciating the invention of PAS.

Prof. Frank Kschischang proposed this monograph to Prof. Alexander Barg, the editor in chief of this journal. Prof. Alexander Barg and publisher Mike Casey showed great patience during the making of this monograph. Two anonymous reviewers provided very valuable comments on a first version.

I thank you all.

Georg Böcherer
Munich, Germany
May 2023

1

Probabilistic Amplitude Shaping

In this section, we discuss the line of thoughts that led to the invention of probabilistic amplitude shaping (PAS). The key ingredients are three tools that have been available to the communications engineer already for some time. These three tools are: first, the additive white Gaussian noise (AWGN) capacity formula [71], second, powerful capacity-approaching binary low-density parity-check (LDPC) codes [37] and the possibility to simulate them on a personal computer [54], and third, the bit-interleaved coded modulation (BICM) architecture [27]. We briefly discuss the capacity formula in Section 1.1.1, binary forward error correction (FEC) in Section 1.1.2, word error rates (WERs) and bit error rates (BERs) in Section 1.1.3 and BICM in Section 1.2. With these tools at hand, the thought process that leads to PAS is rather of practical than theoretic nature. The steps consist in successive modifications of a practical system for simulating WERs of a binary FEC in AWGN. We discuss these modifications in Section 1.3. The PAS architecture raises several design questions, which we list in Section 1.4 and address in greater detail in the following sections of this monograph.

1.1 Preliminaries

1.1.1 AWGN Capacity

The real-valued discrete time AWGN channel is

$$Y_i = X_i + Z_i, \quad i = 1, 2, \ldots, n \tag{1.1}$$

where the Y_i, X_i, and Z_i are outputs, inputs, and noise, respectively. Inputs and noise are independent and the Z_i are independent zero mean Gaussian with variance σ^2, i.e.,

$$p_{Z_i}(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}}. \tag{1.2}$$

The input is subject to an average power constraint

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i^2) \leq \mathsf{P}. \tag{1.3}$$

The capacity of the AWGN channel is

$$\max_{P_X:\ \mathbb{E}(X^2)\leq \mathsf{P}} \mathbb{I}(X;Y) = \frac{1}{2}\log_2\left(1 + \frac{\mathsf{P}}{\sigma^2}\right) \tag{1.4}$$

where $\mathbb{I}(X;Y)$ denotes the mutual information of X and Y, see (A.3.5). The ratio P/σ^2 is called the signal-to-noise ratio (SNR). The capacity-achieving density of the AWGN is zero mean Gaussian with variance P.

1.1.2 Binary Linear FEC

Parity Check Matrix To protect a block $\boldsymbol{c} = c_1 \ldots c_n$ of n bits against errors, a linear FEC code imposes m_{fec} linear constraints on $\boldsymbol{c}$. Each constraint requires that a certain subset of the n bits in $\boldsymbol{c}$ add to an even number, i.e., zeros in the binary field. The constraints are therefore called *parity checks*. The ith parity check is compactly written as a length n row vector $\boldsymbol{h}_i = h_{i1} \ldots h_{in}$ and the vector $\boldsymbol{c}$ must fulfill $\boldsymbol{c}\boldsymbol{h}_i^{\mathrm{T}} = 0$. Arranging m_{fec} parity checks in a matrix results in the *parity check matrix* $\boldsymbol{H}$ with transpose

$$\boldsymbol{H}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{h}_1^{\mathrm{T}} & \boldsymbol{h}_2^{\mathrm{T}} & \cdots & \boldsymbol{h}_{m_{\text{fec}}}^{\mathrm{T}} \end{bmatrix} \tag{1.5}$$

and $\boldsymbol{c}$ is a codeword if and only if it fulfills all m_{fec} parity checks, i.e.,

$$\boldsymbol{c}\boldsymbol{H}^{\mathrm{T}} = \boldsymbol{0}. \tag{1.6}$$

This defines the linear FEC code

$$\mathcal{C} := \left\{ \boldsymbol{c} \in \{0,1\}^n \colon \boldsymbol{c}\boldsymbol{H}^{\mathrm{T}} = \boldsymbol{0} \right\}. \tag{1.7}$$

Systematic Encoding It is convenient for the last m_{fec} columns of $\boldsymbol{H}$ to be linearly independent, which can always be achieved, when $\boldsymbol{H}$ is full rank, by suitable rearrangement of columns. Then, the matrix is of the form

$$\boldsymbol{H} = [\boldsymbol{Q}|\boldsymbol{R}] \tag{1.8}$$

where the $m_{\text{fec}} \times m_{\text{fec}}$ matrix $\boldsymbol{R}$ is full rank and invertible. Systematic encoding of k bits $\boldsymbol{u}$ can now be done in two steps.

1. Calculate $\boldsymbol{s} = \boldsymbol{u}\boldsymbol{Q}^{\mathrm{T}}$.
2. Solve $\boldsymbol{p}\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{s} \Rightarrow \boldsymbol{p} = \boldsymbol{s}(\boldsymbol{R}^{\mathrm{T}})^{-1}$.

The vector $\boldsymbol{c} = [\boldsymbol{u}|\boldsymbol{p}]$ is then a codeword, i.e., it fulfills $\boldsymbol{c}\boldsymbol{H}^{\mathrm{T}} = \boldsymbol{0}$. A convenient way to represent systematic encoding is via a systematic generator matrix

$$\boldsymbol{G} = [\boldsymbol{I}|\boldsymbol{P}] \tag{1.9}$$

where $\boldsymbol{I}$ is a $k \times k$ identity matrix and $\boldsymbol{P} = \boldsymbol{Q}^{\mathrm{T}}(\boldsymbol{R}^{\mathrm{T}})^{-1}$. We can now compactly write systematic encoding by the multiplication of $\boldsymbol{u}$ with $\boldsymbol{G}$, i.e.,

$$\boldsymbol{u}\boldsymbol{G} = [\boldsymbol{u}|\boldsymbol{p}] = \boldsymbol{c}. \tag{1.10}$$

Since $\boldsymbol{p} = \boldsymbol{u}\boldsymbol{P}$, we call $\boldsymbol{P}$ the *parity forming part* of $\boldsymbol{G}$.

1.1.3 Word- and Bit Error Rate

The performance of FEC codes are usually characterized either by their WER or by their BER. While information theorists mainly use WER, e.g., for channel capacity, communications engineers mainly use BER.

In the remainder of this section, we consider WER, for the sake of simplicity. The obtained insights hold similarly for BER. We next define WER and BER formally and relate them to each other.

Consider a binary code with codeword length n. Suppose $\#\{\mathrm{W}\}$ codewords were transmitted and after decoding, $\#\{\mathrm{WE}\}$ word errors occured. The WER is then

$$\mathrm{WER} = \frac{\#\{\mathrm{WE}\}}{\#\{\mathrm{W}\}}. \tag{1.11}$$

The number of transmitted bits is $\#\{\mathrm{B}\} = n \cdot \#\{\mathrm{W}\}$. In each erroneous codeword, the number of erroneous bits is at least one and at most n. Thus, the number of bit errors is bounded as

$$\#\{\mathrm{WE}\} \leq \#\{\mathrm{BE}\} \leq n \cdot \#\{\mathrm{WE}\}. \tag{1.12}$$

The BER is

$$\mathrm{BER} = \frac{\#\{\mathrm{BE}\}}{\#\{B\}} \tag{1.13}$$

and bounded by

$$\frac{1}{n}\mathrm{WER} \leq \mathrm{BER} \leq \mathrm{WER}. \tag{1.14}$$

In particular, the BER is upper bounded by the WER, so if we design a communication link with low WER, we can guarantee that it has a low BER, too.

Another way to relate the two error rates is to consider error exponents. Suppose we have a family of FEC codes where we can choose the codeword length n as large as we want. Denote the corresponding error rates by $\mathrm{WER}(n)$ and $\mathrm{BER}(n)$. Then, the word and bit error exponents are respectively

$$E_W = \lim_{n\to\infty} -\frac{\log \mathrm{WER}(n)}{n} \tag{1.15}$$

$$E_B = \lim_{n\to\infty} -\frac{\log \mathrm{BER}(n)}{n}. \tag{1.16}$$

By (1.14), E_B is lower bounded by E_W and to bound E_B from above, consider

$$-\frac{\log \mathrm{BER}(n)}{n} \leq -\frac{\log\left(\frac{1}{n}\mathrm{WER}(n)\right)}{n} \tag{1.17}$$

$$= -\frac{\log \mathrm{WER}(n)}{n} + \frac{\log n}{n} \tag{1.18}$$

$$\xrightarrow{n\to\infty} -\frac{\log \mathrm{WER}(n)}{n} \tag{1.19}$$

which implies that E_B is also upper bounded by E_W. Consequently, the bit error exponent is equal to the word error exponent.

1.2 Bit-Interleaved Coded Modulation

1.2.1 BPSK in AWGN

Full System

This diagram lays out a coded transmission over the AWGN channel using a binary FEC code with a soft decision (SD) decoder. Let's go through the components from left to right.

Information bits b^k are encoded by a systematic encoder, which appends parity bits p^{n-k}. Together, information and parity bits form the codeword c^n. The coded bits are then mapped to binary phase shift keying (BPSK) symbols by the binary mapping

$$0 \mapsto x(0) = -1 \tag{1.20}$$

$$1 \mapsto x(1) = 1. \tag{1.21}$$

The BPSK symbols are transmitted over the channel and the channel output is

$$y_i = x_i + z_i, \quad i = 1, \ldots, n \tag{1.22}$$

where the z_i are independent zero mean Gaussians with variance σ^2. The demapper calculates the soft-decisions

$$\ell_i = \log \frac{p_{Y|B}(y_i|0)}{p_{Y|B}(y_i|1)} = \log \frac{p_{Y|X}(y_i|-1)}{p_{Y|X}(y_i|+1)}, \quad i = 1, \ldots, n \tag{1.23}$$

and the decoder outputs its decision $\hat{c}^n = \hat{b}^k \hat{p}^{n-k}$. For our discussion in this section, three ways to calculate the decision $\hat{c}^n$ from the soft-decision ℓ^n (or equivalently, from the likelihoods $p_{Y|B}(y_i|0)$ and $p_{Y|B}(y_i|1)$) are relevant.

1. To assess performance limits, we consider the mutual information $\mathbb{I}(B;Y)$ for uniformly distributed input bits. The achievability of mutual information is proven, e.g., in [39, Chapter 5], by considering a random code ensemble and the maximum-likelihood (ML) decision rule

$$\hat{c}^n = \arg\max_{c^n \in \mathcal{C}} \sum_{i=1}^{n} \ell_i (1 - 2c_i) \tag{1.24}$$

   which minimizes the WER. We discuss decision rules and achievable rates for non-uniformly distributed input in detail in Sections 3, 4, and 5.

2. A bitwise maximum a posteriori probability (MAP) decoder, see, e.g., [62, Section 2.5.1], uses the decision rule

$$\begin{aligned} \hat{c}_i &= \arg\max_{b \in \{0,1\}} P_{B_i|Y^n}(b|y^n) && (1.25)\\ &= \arg\max_{b \in \{0,1\}} \sum_{\substack{c^n \in \mathcal{C} \\ c_i = b}} P_{B^n|Y^n}(c^n|y^n) && (1.26)\\ &= \arg\max_{b \in \{0,1\}} \sum_{\substack{c^n \in \mathcal{C} \\ c_i = b}} \prod_{j=1}^{n} p_{Y|B}(y_j|c_j). && (1.27) \end{aligned}$$

3. A practical LDPC decoder approximates the bitwise MAP rule by message passing on a graph with cycles. All simulation results presented in this section were obtained by using the DVB-S2 rate 4/5 LDPC code with parameters specifed in Table 1.1.

We evaluate the peformance by estimating the WER

$$\text{WER} = \Pr(\hat{C}^n \neq C^n) \tag{1.28}$$

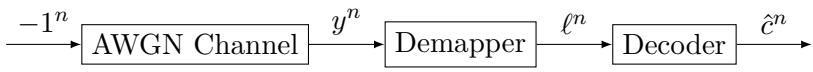
by Monte Carlo simulation. We display the WER curve in Figure 1.1, and we show the operating point at $\text{WER} = 1 \times 10^{-3}$ in Figure 1.2.

Table 1.1: Parameters of the DVB-S2 LDPC code.

R_{fec}	$4/5$
n	$64\,800$
k	$51\,840$
m_{fec}	$12\,960$
decoding algorithm	belief propagation
number of iterations	50

We note that the operating point is $\approx 0.6\,\text{dB}$ away from the BPSK limit $\mathbb{I}(B;Y)$ and $\approx 1.6\,\text{dB}$ away from capacity. Later in this section, we will use the $0.6\,\text{dB}$ gap to the BPSK limit as a rough estimate of the FEC penalty of the considered code.

All Zero Codeword



If we are only interested in evaluating the WER and don't need a fully functioning system, we can simplify our setup. For BPSK in AWGN, the two input symbols -1 and $+1$ are affected equally by noise. Therefore, since the FEC code is linear, the WER does not depend on the transmitted codeword. All linear codes have the all-zero vector as codeword, so that we can remove the encoder and the mapper and transmit the -1^n vector. The WER is then

$$\Pr(\hat{C}^n \neq 0^n) \tag{1.29}$$

which we can estimate by Monte Carlo simulation. As expected, in Figures 1.1 and 1.2, the all-zero codeword WER is on top of the full system WER. The all-zero codeword system has several advantages.

1. We need to write less code for implementing it.
2. The simulation runs faster since unnecessary calculations are skipped.
3. We can evaluate FEC codes for which we have a decoder but no encoder.

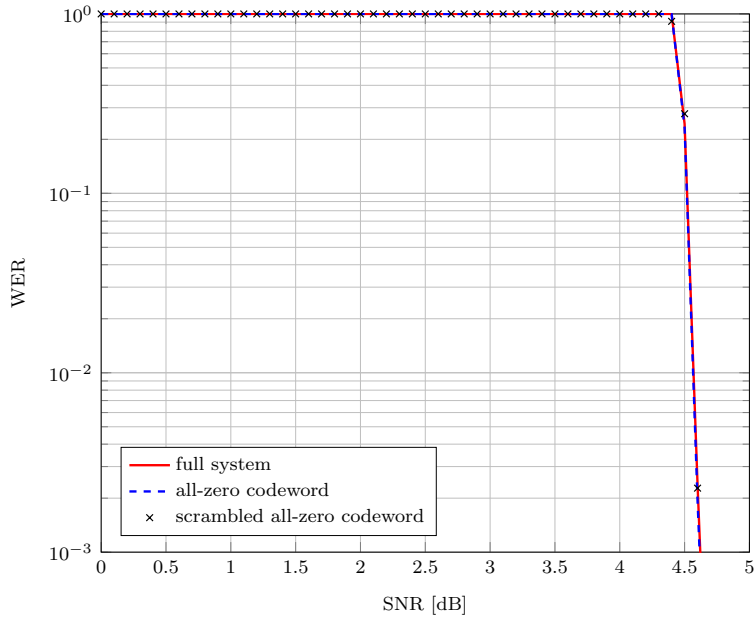


Figure 1.1: WER of BPSK.

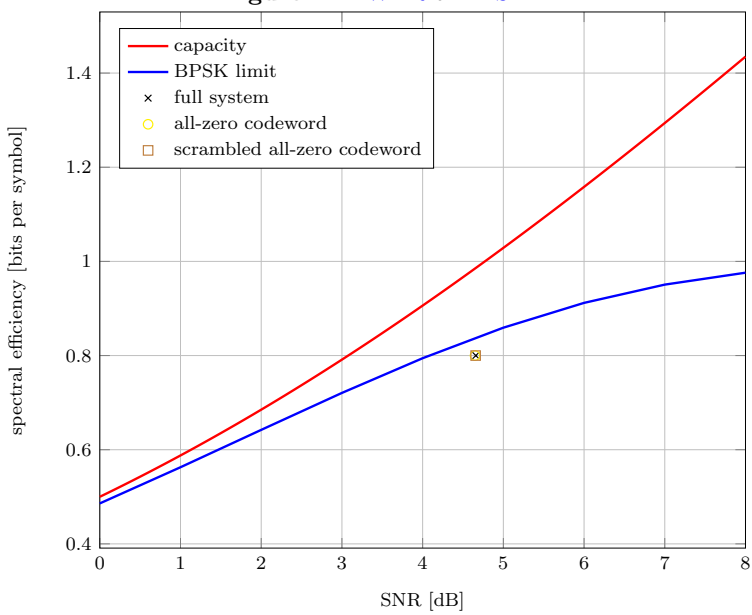
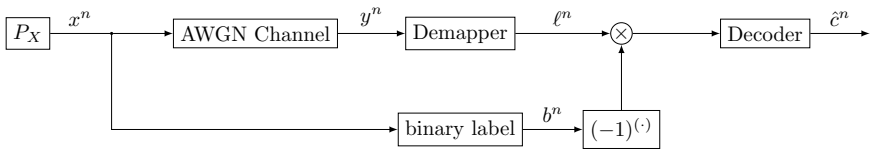


Figure 1.2: WER $= 1 \times 10^{-3}$ operating point of BPSK.

Scrambled All-Zero Codeword



Instead of transmitting always -1, we can also sample the BPSK symbols independently with distribution $P_X(-1) = P_X(1) = \frac{1}{2}$. The binary label b^n of the random sequence x^n is unlikely to be a codeword. Therefore, we interpret b^n as a scrambling sequence that was applied to the all-zero codeword. Accordingly, we must descramble the demapper output ℓ^n before we pass it to the decoder. The WER is now again

$$\Pr(\hat{C}^n \neq 0^n) \tag{1.30}$$

and we estimate it by Monte Carlo simulation. As expected, in Figures 1.1 and 1.2, the scrambled all-zero codeword WER is on top of the full system WER.

1.2.2 Bit-Interleaved Coded Modulation

As we can see in Figure 1.2, for sufficiently high SNR, the BPSK limit flattens out and the gap to capacity becomes arbitrarily large. We therefore need to use constellations larger than BPSK, which is called higher-order modulation. BICM [27] provides the appropriate framework for combining higher-order modulation with binary FEC. For specifying a BICM system, we first need some definitions.

Amplitude Shift Keying We use amplitude shift keying (ASK) constellations with M symbols

$$\mathcal{X} = \{\pm 1, \pm 3, \ldots, \pm(M-1)\} \tag{1.31}$$

where $M = 2^m$ for some integer m. Note that $M = 2$ recovers BPSK.

Bitwise Demapping We associate with each symbol $x \in \mathcal{X}$ a binary label $b^m = \phi(x) \in \{0,1\}^m$. The jth bit level is $b_j = \phi_j(x)$. Define the symbol sets

$$\mathcal{X}_b^j = \{x \in \mathcal{X} \colon \phi_j(x) = b\}, \quad j = 1, \ldots, m,\ b \in \{0,1\}. \tag{1.32}$$

Table 1.2: The BRGC for 8-ASK.

symbol x	label $\phi(x)$
-7	000
-5	001
-3	011
-1	010
1	110
3	111
5	101
7	100

For each bit level j, the constellation $\mathcal{X}$ is partitioned into $\mathcal{X}_0^j$ with symbols where bit level j is 0, and $\mathcal{X}_1^j$ where bit level j is 1. The demapper calculates

$$\ell_{ji} = \log \frac{P_{Y|B_j}(y_i|0)}{P_{Y|B_j}(y_i|1)} = \log \frac{\sum_{x\in\mathcal{X}_0^j} p_{Y|X}(y_i|x)}{\sum_{x\in\mathcal{X}_1^j} p_{Y|X}(y_i|x)}$$
$$j = 1,\ldots,m,\ i = 1,\ldots,n/m. \tag{1.33}$$

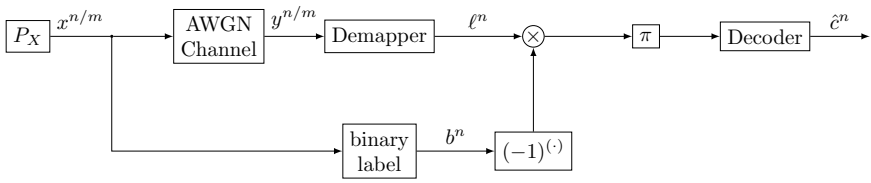
The ℓ_{ji} are reindexed to a length n vector ℓ^n and passed to the LDPC decoder, which outputs the decision $\hat{c}^n$. The internal processing of the LDPC decoder only depends on ℓ^n and not on whether ℓ^n was calculated for BPSK with one bit level or BICM with more than one bit level.

Gray Code BICM works best when the binary label ϕ is a Gray code, i.e., when any pair of neighboring symbols in $\mathcal{X}$ have labels that differ in only 1 bit level. For $M = 8$, a Gray code is listed in Table 1.2, specifically, a binary reflected Gray code (BRGC). We note that for bit level 1, we have one decision boundary, as all negative symbols have $b_1 = 0$ and all positive symbol have $b_1 = 1$. On the other hand, bit level 3 has three decision boundaries. This indicates that bit level 3 is affected more by noise than bit level 1.

Interleaver As the different bit levels have different reliability, their distribution over the codeword may affect performance. In BICM, an

interleaver takes care of how bit levels map to coded bits. Here, we simply use a bit interleaver π that we sample randomly once and then leave it fixed. We discuss interleaver design in more detail in Section 1.4.3.

BICM System with Scrambled All-Zero Codeword



This diagram shows a system for simulating the WER of BICM. We note that compared to the scrambled all-zero codeword BPSK system, not much has changed. The only differences are

- The demapper function (1.33) for calculating ℓ^n, which is more complex than before.
- The interleaver π, which distributes the different bit levels uniformly over the codeword.
- The source P_X, which now samples the x_i uniformly from a 2^m-ASK constellation.
- The length of the channel input sequence, which is reduced from n to n/m, as each symbol is labelled by m bits.

Note that for $m = 1$ bit levels, we recover the BPSK system we considered before. Note that the WER is always given by (1.28), independent of the number of bit levels. We show the operating point for WER $= 1 \times 10^{-3}$ in Figure 1.3 and we also plot the BICM limit

$$\sum_{i=1}^{m} \mathbb{I}(B_i; Y). \tag{1.34}$$

We provide a derivation of (1.34) in Section 4.4.2. We observe that the gap to the BICM achievable rate is ≈ 0.6 dB, similar to the FEC penalty we observed for BPSK. However, the BICM achievable rate

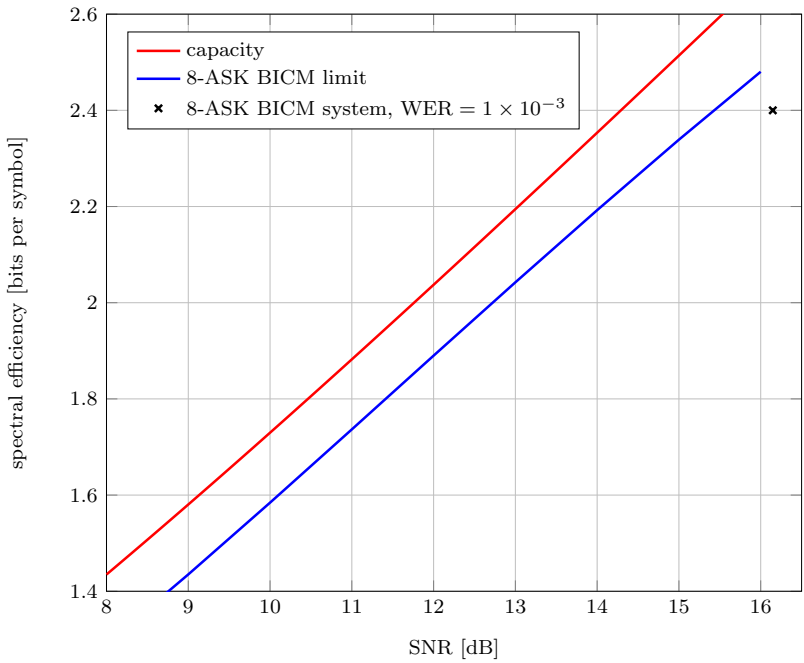


Figure 1.3: WER $= 1 \times 10^{-3}$ operating point of 8-ASK BICM.

itself has a gap of 1.2 dB to capacity. The achievable rate gap is pretty constant over the range of considered SNR values. Thus, it is unlikely to overcome this gap by using a larger constellation. Two alternative options for reducing the gap of the operating point to capacity are as follows.

1. Reduce the FEC penalty.
2. Use a non-uniform symbol distribution.

Because it is much simpler, let's focus on the second option.

1.3 Probabilistic Amplitude Shaping

Taking again a look at the BICM diagram, we note that we can change the probability distribution P_X and evaluate the WER, without affecting

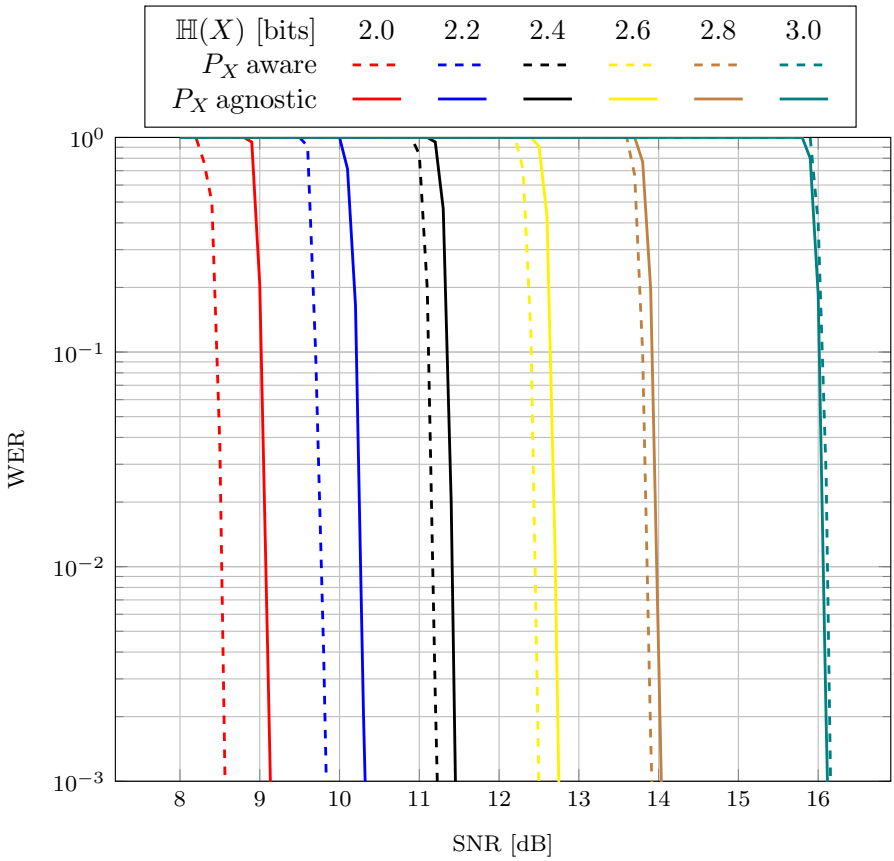


Figure 1.4: WER using demapper (1.33), agnostic of P_X, and demapper (1.36), aware of P_X.

any other part of the system. As the Gaussian density is capacity-achieving for AWGN, we choose a sampled Gaussian density, i.e.,

$$P_X(x) = \frac{e^{-\nu x^2}}{\sum_{a\in\mathcal{X}} e^{-\nu a^2}}, \quad x \in \mathcal{X}. \tag{1.35}$$

Following [51, Section IV.], we call (1.35) a Maxwell-Boltzmann (MB) distribution. The parameter $\nu \geq 0$ controls the shaping degree. For $\nu = 0$, P_X is uniform, and for $\nu \to \infty$, the probability mass concentrates on the two innermost points -1 and $+1$. We quantify the shaping degree by the entropy $\mathbb{H}(X)$ in bits. We now evaluate the WER curves for $\mathbb{H}(X) = 2.0, 2.1, \ldots, 3.0$ bits.

1.3.1 WER

We observe in Figure 1.4 (solid lines) that by lowering $\mathbb{H}(X)$, the SNR required for achieving WER $= 1 \times 10^{-3}$ is also lowered, using the same FEC code and decoder. The reason is that if we fix the noise variance and we decrease the entropy $\mathbb{H}(X)$, we also decrease the transmit power and thereby the SNR, while the distance between neighboring signal points remains unchanged. Equivalently, at the same SNR, lower entropy translates into larger distance.

We note that the demapper (1.33) is not aware of the input distribution P_X. To make the prior P_X available to the decoder, we modify the demapper to

$$\ell_{ji} = \log \frac{\sum_{x\in\mathcal{X}_0^j} P_X(x) p_{Y|X}(y_i|x)}{\sum_{x\in\mathcal{X}_1^j} P_X(x) p_{Y|X}(y_i|x)}$$
$$j = 1, \ldots, m, \; i = 1, \ldots, n/m. \tag{1.36}$$

We display the resulting WER curves in Figure 1.4 (dashed lines). We note that the WER curves are shifted to the left and the SNR required for WER $= 1 \times 10^{-3}$ is lowered further by up to 0.6 dB.

1.3.2 Spectral Efficiency

We now would like to display the WER $= 1 \times 10^{-3}$ operating point in the SNR versus spectral efficiency (SE) plot to evaluate the gap to capacity. However,

For $\mathbb{H}(X) < m$, what is the SE?

Note that as the FEC code is unchanged, the decoder still decodes against a code of rate R_{fec}, which corresponds to $R_{\text{fec}}m$ bits per symbol. For the unshaped case, the SE is mR_{fec}, which we can rewrite as

$$\text{SE} = m - m(1 - R_{\text{fec}}). \tag{1.37}$$

Here, m is the SE of an uncoded system, and $m(1 - R_{\text{fec}})$ is the FEC redundancy. For the shaped case, the uncoded SE is $\mathbb{H}(X)$ and in analogy to (1.37), we may guess the coded SE is

$$\text{SE} \stackrel{?}{=} \mathbb{H}(X) - m(1 - R_{\text{fec}}). \tag{1.38}$$

If entropy is very small, the right-hand side may become negative, which is not a meaningful value, so we modify our guess to

$$\text{SE} = [\mathbb{H}(X) - m(1 - R_{\text{fec}})]^+ . \tag{1.39}$$

In Figure 1.5 we plot required SNR versus SE assuming the correctness of (1.39). We note that below 14 dB of SNR, the curve is almost within the FEC penalty of capacity! This is an exciting observation!

1.3.3 Probabilistic Amplitude Shaping

We have to address two urgent questions:

1. How can we verify the spectral efficiency claim (1.39)?
2. How can we encode?

First, we observe that in our system, the decoder effectively decodes the binary labels of shaped symbols. Thus, we need to place the encoder between the shaped source and the channel. Using a systematic encoder, at least the information part is left unchanged by the encoder, so we may draw the following preliminary diagram.

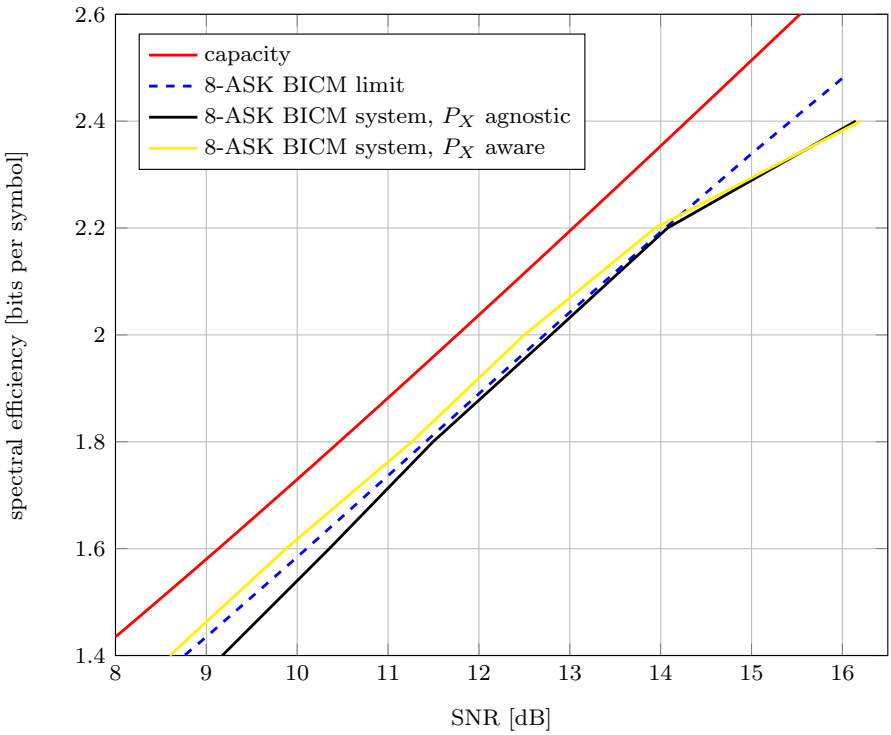
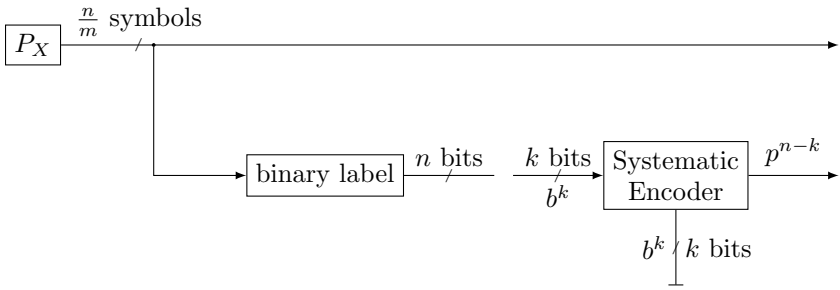
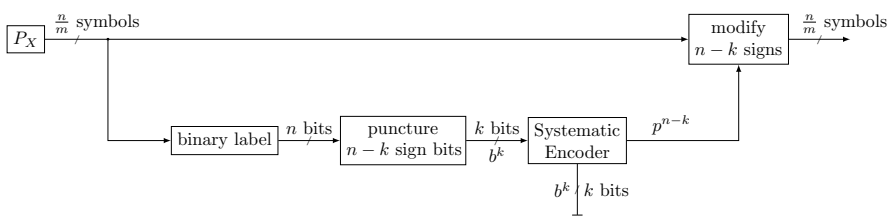


Figure 1.5: WER $= 1 \times 10^{-3}$ operating points. The P_X agnostic demapper uses (1.33) to calculate bitwise soft-decisions, while the P_X aware demapper uses (1.36).



The information bits b^k are left unchanged by systematic encoding and no further processing is required. We indicate this by the terminated encoder output in the diagram. In contrast, the parity bits p^{n-k} are newly generated by the encoder and do require further processing. This diagram has 2 issues. First, the encoder gets n bits at its input, while it only accepts k bits. Second, we must modulate the $n-k$ parity bits onto the transmitted signal somehow. The key observation is now that *we cannot impose any specific distribution onto the parity bits.* Looking at the Gray label in Table 1.2, we note that bit level 1 decides on the sign, and consequently, the transmitted power and thereby the received SNR does not depend on the distribution of bit level 1. A quick fix for the two issues is therefore as follows:

1. Mark $n-k$ sign bit positions.
2. Puncture these marked positions before the encoder, reducing the number of bits from n to k, as required.
3. Modify the $n-k$ signs corresponding to the marked positions according to the parity bits p^{n-k} output by the systematic encoder.



We are now in a position to calculate the SE. Note that for ASK constellations (1.31) the MB distribution P_X (1.35) can be factorized into amplitude A and sign S via

$$P_X(x) = P_A(|x|)P_S(\text{sign}(x)) \tag{1.40}$$

$$= P_A(|x|)\frac{1}{2}. \tag{1.41}$$

In terms of entropy, this corresponds to

$$\mathbb{H}(X) = \mathbb{H}(A) + \mathbb{H}(S) = \mathbb{H}(A) + 1. \tag{1.42}$$

The total amount of information per codeword is thus

$$\frac{n}{m}\,\mathbb{H}(A) + \left(\frac{n}{m} - (n-k)\right)\mathbb{H}(S) = \frac{n}{m}\left[\mathbb{H}(X) - \frac{(n-k)m}{n}\right] \tag{1.43}$$

$$= \frac{n}{m}\left[\mathbb{H}(X) - m(1 - R_{\text{fec}})\right] \tag{1.44}$$

which confirms the SE we postulated in (1.39). Note that on the right-hand side of (1.43), only the information bit carrying signs are counted, not the signs carrying parity bits. Thus, this SE calculation does not assume any specific distribution of the parity bits.

Having confirmed the SE, the complete PAS architecture is only a few steps away. We need to:

1. Separate the source into n/m amplitudes and $n/m - (n-k)$ signs.
2. Remove the sign puncturer.
3. Replace the sign polluter by a sign multiplexer.
4. Add an interleaver.

The diagram in Figure 1.6 shows the complete PAS architecture as we proposed it in [24].

1.4 PAS Components

At several points during the development of PAS in this section, we made design choices based on intuition, which require further study. In the following, we discuss some of them and if possible, we provide pointers to the parts of this monograph, where they are discussed in more detail.

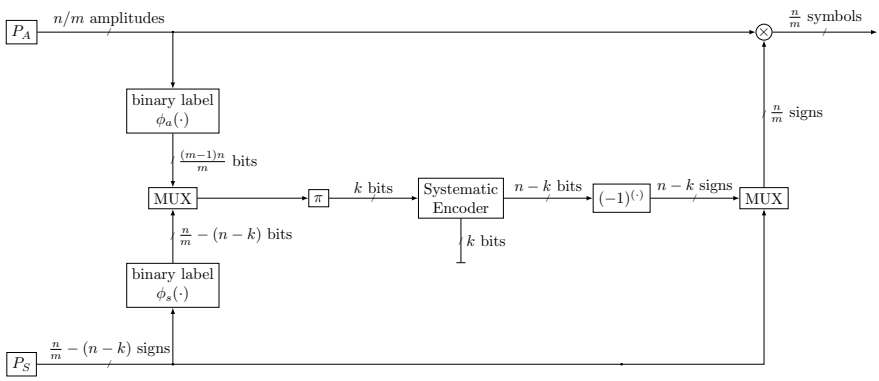


Figure 1.6: The PAS architecture as proposed in [24].

1.4.1 Distribution Matcher

A key ingredients of PAS is the amplitude source P_A, which generates amplitudes according to a desired distribution. To quantify the SE of PAS, we postulated the information content of this source to be $\mathbb{H}(A)$ bits per amplitude. In a practical system, we need to replace the amplitude source P_A by a distribution matcher (DM), which maps k uniformly distributed input bits to n amplitudes with distribution P_A. In Section 2, we study DMs in detail. The key result of Section 2 is that optimal DMs have an inherent rate loss $\mathbb{H}(A) - k/n$ that scales as $\frac{\log n}{n}$ and a cost penalty (e.g., increased average power) that also scales as $\frac{\log n}{n}$. On the downside, this requires the use of DMs that process sufficiently many amplitudes jointly. On the positive side, the rate $\mathbb{H}(A)$ can indeed be achieved, by a sufficiently long DM.

Remark 1.1. Because of the inherent rate loss, DMs operate at a rate that is *below* the entropy of the generated amplitude distribution P_A. This implies that a source decoder for a discrete memoryless source (DMS) P_A cannot be used as a DM, as it would operate at a rate *above* the entropy of P_A. We revisit this observation in Sections 2.4.2 and 2.5.6.

1.4.2 Achievable Spectral Efficiency

We postulated for PAS the SE

$$\mathrm{SE} = [\mathbb{H}(X) - m(1 - R_{\mathrm{fec}})]^+ \tag{1.45}$$

where $m = \log_2 |\mathcal{X}|$ is the logarithmic size of the channel input alphabet $\mathcal{X}$ and where $[\cdot]^+ = \max\{\cdot, 0\}$. In Section 3, we study what SEs are achievable by a PAS-like architecture that consists of two layers, namely the shaping layer and the FEC layer. The two layers are reflected in the achievable SEs, namely, it decomposes into two parts. The first part is the shaping set rate R_{ss}, which is bounded as

$$mR_{\mathrm{ss}} \leq \mathbb{H}(X) \tag{1.46}$$

and which can achieve this bound for sufficiently large n. The second part is the achievable FEC rate, which is given by

$$m(1 - R^*_{\mathrm{fec}}) = \mathbb{H}(X|Y) \tag{1.47}$$

that is, for $R_{\mathrm{fec}} < R^*_{\mathrm{fec}}$ and sufficiently large n, reliable communication is possible. The two parts together provide an achievable SE.

The use of a linear code is a key aspect of PAS. In Section 5, we derive an achievable SE for PAS using a random linear code. Again, this achievable SE consists of two parts. The shaping layer part is basically the rate of the employed DM (which, by Section 2, is asymptotically optimal). The FEC part recovers (1.47).

Both for the PAS-like architecture considered in Section 3 and the PAS architecture considered in Section 5, we find that $\mathbb{I}(X;Y)$ is an achievable SE, which shows that PAS is capacity-achieving for a certain class of discrete input channels.

1.4.3 Interleaver Design for Practical FEC

In our derivation of PAS we used an intra-codeword "random interleaver" (because we did not know better). In Section 3.4.3, we show that under ML-like decoding, the achievable FEC rate is invariant under intra-codeword interleaving and conclude that intra-codeword interleaver design should be considered part of practical FEC code design,

accounting for suboptimal decoding. In Section 4.4.3, we revisit the interleaver question and derive the optimal decoding metric for the case when the interleaver is not known to the decoder. The design of interleavers for PAS has been considered for different families of FEC codes.

LDPC Codes When using an already designed binary LDPC code with higher order modulation, one may optimize the interleaver separately as done, e.g., in [47]. This approach was used in [16, Section V.D], [6, Section V.B], and [24, Section VIII] for optimizing the interleaver for PAS with DVB-S2 LDPC codes. In [76], [77], the joint design of LDPC codes and interleavers for PAS is considered.

Product Codes The PAS interleaver design for product codes based on algebraic component codes is considered for hard-decision decoding in [72] and for soft-decision decoding in [20].

Spatially Coupled Codes (This family of codes is known under many different names, see, e.g., [79, Section I]). PAS is combined with spatially coupled LDPC codes in [14], [15]. The PAS interleaver design for staircase codes [75] under hard-decision decoding is considered in [73]. A similar design can be used for continuously interleaved algebraic component codes under hard-decision decoding [64] and under soft-decision decoding, e.g., the oFEC code [74]. Usually, the PAS interleaver design is simpler for spatially coupled codes than for product codes.

Polar Codes The work [59] designs a PAS interleaver for polar codes, a family of FEC codes proposed in [3], [78]. The work [48] points out that polar codes inherently allow for probabilistic shaping. Various strategies for polar coding with probabilistic shaping are evaluated in [63]. We note that [63] evaluates polar coded PAS for constant composition distribution matching (CCDM). As we detail in Section 2, minimum cost distribution matcher (MCDM) performs significantly better than CCDM for short output lengths, so the comparison of [48] and polar coded PAS with MCDM is an important topic for future work.

1.4.4 Decoding Metrics

We observed that switching the demapper from calculating $\log \frac{p_{Y|B}(y|0)}{p_{Y|B}(y|1)}$ to calculating $\log \frac{P_{B|Y}(0|y)}{P_{B|Y}(1|y)}$ improved the WER of PAS. In Section 4, we derive optimal decoding metrics for several practically relevant scenarios, including bitwise demapping and hard-decision decoding.

1.4.5 Optimal Input Distribution

By our findings in Section 3 and Section 5, PAS can achieve

$$\mathrm{SE} = \mathbb{I}(X;Y). \tag{1.48}$$

We may therefore assume that the optimal input distribution for PAS is

$$P_{X^*} = \arg\max_{P_X} \mathbb{I}(X;Y). \tag{1.49}$$

This, however, is only true if we can also choose the FEC rate freely. In practical applications, however, the FEC rate is often determined by the available FEC engine. In this case, the layered nature of the PAS architecture as reflected by the achievable SE expression needs to be taken into account. The optimization problem is then

$$\begin{aligned} \underset{P_X}{\text{maximize}} \quad & \mathbb{H}(X) && (1.50)\\ \text{subject to} \quad & \mathbb{H}(X|Y) \leq m(1-R_{\text{fec}}). && (1.51) \end{aligned}$$

This is a concave objective with a convex constraint. The Lagrangian to be maximized is

$$\mathbb{H}(X) - \lambda\, \mathbb{H}(X|Y) \tag{1.52}$$

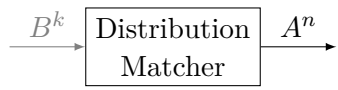
which is the sum of a concave and a convex function. For $\lambda = 1$, fortunately, we have

$$\mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{I}(X;Y) \tag{1.53}$$

which is known to be concave in P_X. However, for $\lambda > 1$, this may not be the case. Finding optimal distributions for FEC rate constrained PAS is interesting and of practical relevance, and we leave it for future research.

2

Distribution Matching



In this section, we consider the distribution matcher (DM), motivated by its central role in the PAS architecture. We next lay out a DM specification that is compliant with its application in the PAS architecture. Our focus in this section is on requirements, performance limits, and performance guarantees. We will not discuss efficient DM implementations, which are thoroughly treated in the existing literature, see, e.g., [66]. For further references, see Section 2.8.

2.1 Specification

2.1.1 Output Length n

We consider DMs that output symbol sequences A^n of fixed length n. The sequences $A^n(1), A^n(2), \ldots$ of successive calls are independent, it is therefore sufficient to study one such sequence A^n. The choice of the DM length n is to some degree independent of the FEC code length n_{fec}, though we may want to ensure that the DM sequence is not shared between different codewords, as illustrated by the following example.

Example 2.1. In our 8-ASK running example from Section 1, the FEC code length is $n_{\text{fec}} = 64\,800$ and the 8-ASK constellation corresponds to $m = 3$ bits per symbol. Thus, one codeword is mapped to $n_{\text{fec}}/m = 21\,600$ symbols, and in particular 21 600 amplitudes. By choosing a DM length n that divides 21 600, we can ensure that no DM sequence is shared among more than 1 codeword. For instance, for $n = 200$, the amplitude sequence passed to the FEC encoder corresponds to the outputs of $21\,600/200 = 108$ successive DM calls.

Suppose now that the DM length n divides n_{fec}/m, and furthermore, that the DM maps k input bits to n output symbols. Then, the PAS architecture maps a fixed number of bits $B^{\text{SE}\frac{n_{\text{fec}}}{m}}$ to coded bits $C^{n_{\text{fec}}}$ and the word error rate from PAS input to PAS output is no larger than the codeword error rate, i.e.,

$$\Pr(\hat{B}^{\text{SE}\frac{n_{\text{fec}}}{m}} \neq B^{\text{SE}\frac{n_{\text{fec}}}{m}}) \leq \Pr(\hat{C}^{n_{\text{fec}}} \neq C^{n_{\text{fec}}}) \tag{2.1}$$

where we have strict inequality if we occasionally have errors only in the parity bits, which are discarded before the information word error rate on the left-hand side is calculated. The relation (2.1) can not be guaranteed when DM sequences are shared between two or more codewords.

2.1.2 Rate

We define the DM rate as the average information per symbol in a DM sequence, i.e.,

$$R = \frac{\mathbb{H}(A^n)}{n}. \tag{2.2}$$

This definition is also meaningful when we leave the DM input unspecified, e.g., when we model the DM output as a random source. In case we specify the DM input, we assume the DM realizes a one-to-one mapping from k input bits to A^n. In case the input bits are independent and uniformly distributed, the rate simplifies to

$$R = \frac{\mathbb{H}(A^n)}{n} = \frac{k}{n}. \tag{2.3}$$

2.1.3 Cost

We associate with each symbol in the DM output alphabet $\mathcal{A}$ a cost w.

Example 2.2. In our running example, we use an 8-ASK constellation $\mathcal{X} = \{\pm 1, \pm 3 \pm 5 \pm 7\}$, with the amplitude set $\mathcal{A} = \{1, 3, 5, 7\}$. The symbol costs are the powers determined by the square and given in the following table

amplitude	1	3	5	7
symbol	± 1	± 3	± 5	± 7
cost w	1	9	25	49

The average cost of one symbol X is determined by its amplitude $A = |X|$ and given by

$$\mathbb{E}[w(A)] = \sum_{a \in \mathcal{A}} P_A(a) w(a). \tag{2.4}$$

For an arbitrary $\nu \neq 0$, define the probability distribution P by

$$P(a) = \frac{2^{-\nu w(a)}}{\sum_{a' \in \mathcal{A}} 2^{-\nu w(a')}}. \tag{2.5}$$

The average cost scaled by ν can now be written as a cross-entropy (A.23), up to normalization, i.e.,

$$\nu\, \mathbb{E}[w(A)] = \mathbb{X}(P_A \| P) - \log_2 \Big(\sum_{a' \in \mathcal{A}} 2^{-\nu w(a')} \Big) \tag{2.6}$$

where

$$\mathbb{X}(P_A \| P) = \mathbb{E}[-\log_2 P(A)]. \tag{2.7}$$

This relation will turn out to be useful at several occasions in the following.

2.1.4 Average Distribution

An important DM property is the average distribution $P_{\bar{A}}$, whose definition appears naturally when calculating the average DM sequence cost, as we will see next. We have

$$\frac{1}{n}\,\mathbb{E}[w(A^n)] = \frac{1}{n}\sum_{a^n\in\mathcal{A}^n} P_{A^n}(a^n)w(a^n) \tag{2.8}$$

$$= \sum_{a'\in\mathcal{A}} w(a') \sum_{a^n\in\mathcal{A}^n} P_{A^n}(a^n)\frac{N(a'|a^n)}{n} \tag{2.9}$$

$$= \sum_{a'\in\mathcal{A}} P_{\bar{A}}(a')w(a') \tag{2.10}$$

$$= \mathbb{E}[w(\bar{A})] \tag{2.11}$$

where we defined the average distribution as

$$P_{\bar{A}}(a') = \sum_{a^n\in\mathcal{A}^n} P_{A^n}(a^n)\frac{N(a'|a^n)}{n}, \quad a'\in\mathcal{A}. \tag{2.12}$$

The term $N(a'|a^n)$ is the number of occurences of letter a' in sequence a^n, see Appendix A.3.1. The average distribution is important for three reasons.

1. The average distribution determines the average cost via
$$\frac{1}{n}\,\mathbb{E}[w(A^n)] = \mathbb{E}[w(\bar{A})]. \tag{2.13}$$
For instance, knowing $P_{\bar{A}}$ allows us to calculate the average cost easily.

2. As we will see, the average distribution allows us to bound the DM rate from above via
$$R \le \mathbb{H}(\bar{A}). \tag{2.14}$$

3. The average distribution may be used by the demapper as prior, for instance, (1.36) may be realized by
$$\ell_{ji} = \log\frac{\sum_{x\in\mathcal{X}_0^j} P_{\bar{A}}(|x|)p_{Y|X}(y_i|x)}{\sum_{x\in\mathcal{X}_1^j} P_{\bar{A}}(|x|)p_{Y|X}(y_i|x)}$$
$$j = 1,\dots,m,\ i = 1,\dots,n/m. \tag{2.15}$$

2.2 Design Problem

We assume a required DM rate R and output length n. For instance, R may be determined by a required SE, the constellation size, and the

rate of the available FEC code. The DM design problem now consists in selecting from all available configurations the DM with rate at least R that has the minimum average cost, i.e.,

$$\text{minimize} \quad \frac{1}{n}\mathbb{E}[w(A^n)] \tag{2.16}$$

$$\text{subject to} \quad \frac{1}{n}\mathbb{H}(A^n) \geq R. \tag{2.17}$$

In the following, we will solve the DM design problem for the following three classes of DMs.

1. (MB Source) A source P_{A^n}.
2. (MCDM) A DM with $k = Rn$ uniformly distributed input bits and no further restrictions.
3. (CCDM) A DM with $k = Rn$ input bits and output sequences that all have the same composition.

We discuss the relation of MCDM and CCDM to source coding in the Sections 2.4.2 and 2.5.6, respectively.

2.3 Maxwell-Boltzmann Source

In this section, we aim to find the source P_{A^n} that minimizes the average cost, among all sources with rate R. By (2.11), we know that the average cost is determined by the average distribution $P_{\bar{A}}$. The rate of this source is bounded from above by

$$\frac{\mathbb{H}(P_{A^n})}{n} \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{H}(P_{A_i}) \tag{2.18}$$

$$\leq \mathbb{H}\Big(\sum_{i=1}^{n}\frac{1}{n}P_{A_i}\Big) \tag{2.19}$$

$$= \mathbb{H}(P_{\bar{A}}) \tag{2.20}$$

where the first inequality follows by the independence bound (A.19) and the second inequality follows by the concavity of entropy and Jensen's inequality (A.7). We have equality if $P_{A^n} = P_{\bar{A}}^n$. Thus, among all sources

with average distribution $P_{\bar{A}}$, the memoryless source $P_{\bar{A}}^n$ has highest rate. Consequently, the problem reduces to finding the distribution P_A that minimizes cost, among all distribution whose entropy is at least as large as the required rate.

The memoryless source with rate R and minimum cost is given by the solution of the optimization problem

$$\begin{aligned} \underset{P_A \in \mathcal{P}(\mathcal{A})}{\text{minimize}} \quad & \mathbb{E}[w(A)] \\ \text{subject to} \quad & \mathbb{H}(A) \geq R \end{aligned} \tag{2.21}$$

where $\mathcal{P}(\mathcal{A})$ is the set of all distributions on $\mathcal{A}$. Define now the MB source P_{A^*} as

$$P_{A^*} = \frac{2^{-\nu w(a)}}{\sum_{a'} 2^{-\nu w(a')}}, \quad \nu \colon \mathbb{H}(A^*) = R. \tag{2.22}$$

We claim that P_{A^*} solves (2.21). To show this, let P_A be an arbitrary distribution with the only restriction that $\mathbb{H}(A) = R$. By the information inequality (A.28), we have

$$\begin{aligned} 0 &\leq \mathbb{D}(P_A \| P_{A^*}) && (2.23)\\ &= \mathbb{X}(P_A \| P_{A^*}) - \mathbb{H}(A) && (2.24)\\ &= \mathbb{X}(P_A \| P_{A^*}) - \mathbb{H}(A^*) && (2.25)\\ &= \nu\, \mathbb{E}[w(A)] + \log_2\Big(\sum_a 2^{-\nu w(a)}\Big) \\ &\quad - \nu\, \mathbb{E}[w(A^*)] - \log_2\Big(\sum_a 2^{-\nu w(a)}\Big) && (2.26)\\ &= \nu\, \mathbb{E}[w(A)] - \nu\, \mathbb{E}[w(A^*)] && (2.27) \end{aligned}$$

where (2.25) follows because by assumption, $\mathbb{H}(A) = \mathbb{H}(A^*) = R$, and where we used (2.6) in (2.26). Thus,

$$\mathbb{E}[w(A)] \geq \mathbb{E}[w(A^*)] \tag{2.28}$$

with equality if and only if $P_A = P_{A^*}$.

2.3.1 Entropy-Cost Curve

For each entropy H, the MB source achieves the minimum cost W. This defines an entropy-cost curve $W(H)$, and similarly, a cost-entropy curve

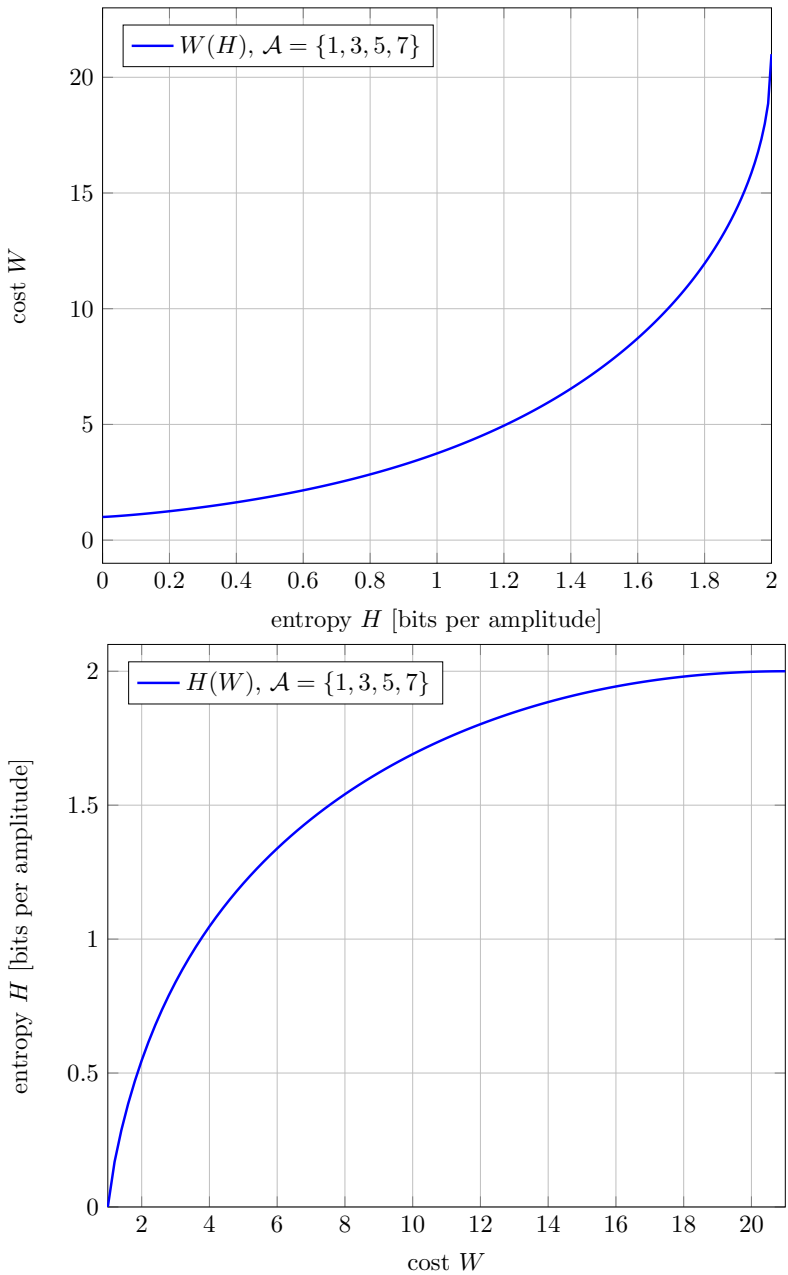


Figure 2.1: The entropy-cost curve $W(H)$ and the cost-entropy curve $H(W)$ for 8-ASK amplitudes $\mathcal{A} = \{1, 3, 5, 7\}$.

$H(W)$. Suppose now the MB distribution with parameter ν^* achieves the point $H^*, W(H^*)$ on the entropy-cost curve and the point $W^*, H(W^*)$ on the cost-entropy curve, respectively. Using convex optimization, we can then show that by [25, Section 5.6.3], the slopes of $W(H)$ and $H(W)$ at H^* and W^*, respectively, are given by

$$\frac{\partial W}{\partial H}(H^*) = \frac{1}{\nu^*} \tag{2.29}$$

$$\frac{\partial H}{\partial W}(W^*) = \nu^*. \tag{2.30}$$

Example 2.3. For our running 8-ASK example, we use the MB source P_{A^*} to generate amplitudes. We show in Figure 2.1 the cost $\mathbb{E}[w(A^*)]$ as function of rate $R = \mathbb{H}(A^*)$ and the rate as function of cost, respectively.

2.4 Minimum Cost Distribution Matching

We now consider a DM that maps a fixed number k of bits to 2^k distinct length n sequences in $\mathcal{A}^n$. The rate of such a DM is

$$R = \frac{k}{n}. \tag{2.31}$$

Independent of the distribution of the k input bits, to minimize the cost, the image must consist of the 2^k sequences in $\mathcal{A}^n$ of least cost, i.e., the optimization problem is

$$\underset{\mathcal{C}\subseteq\mathcal{A}^n}{\text{minimize}} \quad \sum_{a^n\in\mathcal{C}} \sum_{i=1}^{n} w(a_i) \tag{2.32}$$

$$\text{subject to} \quad \log_2 |\mathcal{C}| = k. \tag{2.33}$$

From now on, we assume the input bits are uniformly distributed, in which case the ordering of the sequences in $\mathcal{C}$ has no influence on the average cost. Furthermore, the output entropy is given by

$$\mathbb{H}(A^n) = k. \tag{2.34}$$

Remark 2.1. MCDMs have been analyzed under an informational divergence criterion in [69].

2.4.1 Calculating the Average Distribution

Of central interest is the average distribution $P_{\bar{A}}$, as it determines the cost and is also used by the demapper for calculating soft-decisions. Calculating $P_{\bar{A}}$ brute force is exponentially complex in n. In this section, we will devise a strategy for efficiently calculating $P_{\bar{A}}$. The key step is to partition $\mathcal{A}^n$ into type classes $\mathcal{T}^n(\hat{P})$. Within each type class, all sequences share by definition the same type $\hat{P}$ and thereby contribute equally to the average power. By [32, Section 2.1], we have the following properties.

1. The type class $\mathcal{T}^n(\hat{P})$ has size

$$|\mathcal{T}^n(\hat{P})| = \frac{n!}{n_1!n_2!\cdots n_M!} = \binom{n}{n_1, n_2, \ldots, n_M}. \tag{2.35}$$

2. The number of distinct types is

$$|\mathcal{P}_n(\mathcal{A})| = \binom{n+M-1}{M-1} \leq (n+1)^{M-1} \tag{2.36}$$

   in particular, the number of types is polynomial in n.

Example 2.4. For $M = 4$ and $n = 200$, the number of types is

$$\binom{n+M-1}{M-1} = 1\,373\,701 \tag{2.37}$$

that is, just a million and a little bit.

We can now proceed as follows:

1. Calculate size and weight of each type in $\mathcal{P}_n(\mathcal{A})$.
2. Sort the types by weight in increasing order. Denote the sorted types, sizes, and weights by $\hat{P}_1, \hat{P}_2, \ldots, N_1, N_2, \ldots$, and $w_1, w_2, \ldots$, respectively.
3. Determine $i_{\max}$ as

$$i_{\max} = \min\left\{ j \colon \sum_{i=1}^{j} N_i \geq 2^k \right\} \tag{2.38}$$

4. Calculate by how much the total number of sequences up to type $i_{\max}$ exceeds the required number 2^k by

$$d = \sum_{i=1}^{i_{\max}} N_i - 2^k. \tag{2.39}$$

5. Adjust the size of $N_{i_{\max}}$ to

$$N_{i_{\max}} \leftarrow N_{i_{\max}} - d \tag{2.40}$$

We can now calculate $P_{\bar{A}}$ by

$$P_{\bar{A}}(a) = \frac{\sum_{i=1}^{i_{\max}} N_i \hat{P}_i(a)}{\sum_{i=1}^{i_{\max}} N_i}, \quad a \in \mathcal{A}. \tag{2.41}$$

Example 2.5. For our 8-ASK running example, we compare in Figure 2.2 for $n = 200$ the MCDM cost to the MB source cost. Over the whole range of rates, the MCDM is within 0.1 dB of the MB source.

2.4.2 MCDM Decoder Is a Bad Source Encoder

Consider a rate $R = 1/2$ MCDM for the alphabet $\mathcal{A} = \{0, 1\}$ with weights $w(0) = 0, w(1) = 1$. The MCDM encoder maps k uniformly distributed bits to length $n = 2k$ sequences of low weight. The MCDM encoder output has an average distribution $P_{\bar{A}}$, which can be calculated following Section 2.4.1. The MCDM decoder maps length n sequences with average distribution $P_{\bar{A}}$ to k bits, and one may wonder if the MCDM decoder provides a useful source encoder. Specifically, consider a discrete memoryless source $P_{\bar{A}}$ outputting sequence $A^n = A_1 A_2 \dots A_n$, where the A_i are independent and identically-distributed (iid) $\sim P_{\bar{A}}$. We now source encode A^n to k bits by using an MCDM as follows.

- If A^n is one of the 2^k low weight sequences indexed by the MCDM encoder, we pass A^n to the MCDM decoder, which outputs the corresponding k bit index.
- If A^n is not one of the sequences indexed by the MCDM encoder, we throw an error.

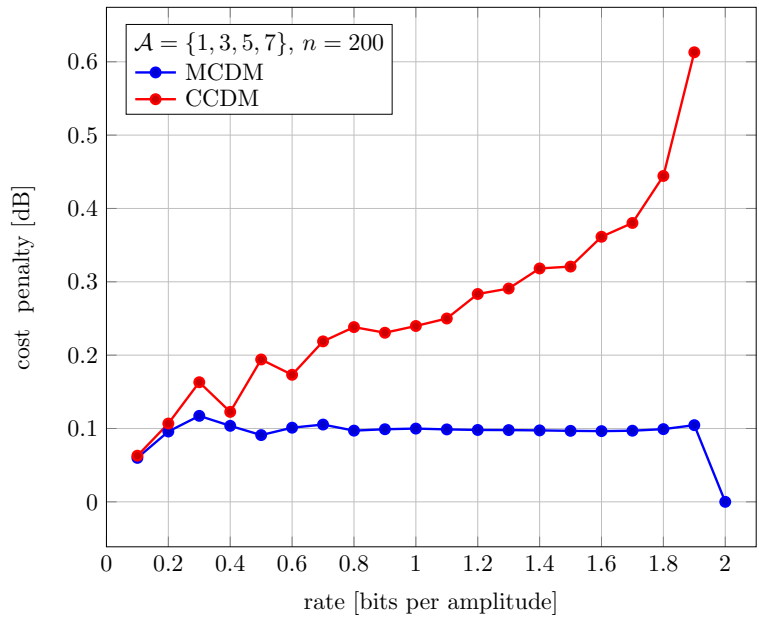


Figure 2.2: Comparison of MB source, CCDM, and MCDM.

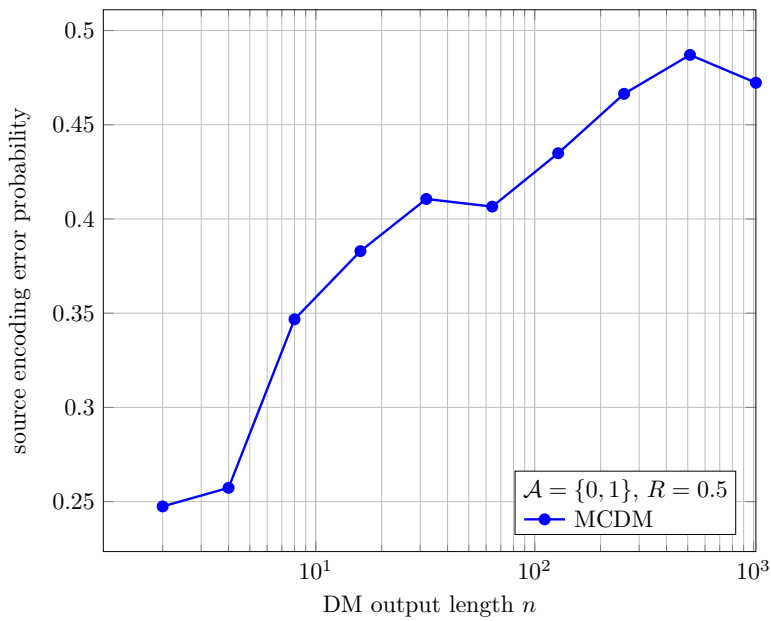


Figure 2.3: Error probability of using an MCDM decoder as source encoder.

This source encoder is useful, if the error probability is low. We evaluate the error probability by Monte Carlo simulation for increasing values of n. Using the tools developed in Section 2.4.1, we proceed as follows.

1. We calculate the index $i_{\max}$ of the type of largest weight sequences indexed by the MCDM. As we consider binary sequences with weights $\{0, 1\}$, the sequence weight uniquely identifies the type, i.e., $w_{i_{\max}} = i_{\max} - 1$.

2. We calculate the size excess d according to (2.39) and the excess fraction

$$p_{\max} = \frac{d}{N_{i_{\max}}}. \tag{2.42}$$

3. We repeatedly sample a^n from the discrete memoryless source $P_{\bar{A}}$.

   (a) If $w(a^n) > w_{i_{\max}}$, we increase the error counter by 1.

   (b) If $w(a^n) = w_{i_{\max}}$, we increase the error counter by $p_{\max}$, as this is the fraction of weight $w_{i_{\max}}$ sequences not indexed by the MCDM.

   We finally estimate the error probability by

$$\frac{\text{value of error counter}}{\#\{\text{number of sampled sequences}\}}. \tag{2.43}$$

In Figure 2.3, we plot the error probability estimates for $n = 2, 4, \ldots,$ 1024. For each value of n, the simulation was run until the error counter exceeded 10 000. The error probability starts at 0.25 at $n = 2$ and approaches 0.5 as n approaches 1024. The irregularity of the error curve is because the size excess d varies for different values of n. In summary, the error probability is unacceptably high and we conclude that for an MCDM encoder with average distribution $P_{\bar{A}}$, the corresponding MCDM decoder is a bad source encoder for a discrete memoryless source $P_{\bar{A}}$.

Remark 2.2. Suppose we have implemented an MCDM decoder. This implementation can very well be used as a source encoder: For a DMS P_A on $\mathcal{A}$, define the weight as $w(a) = -\log P_A(a)$. Then, for the desired

sequence length n, tune the compression ratio $R = k/n$ so that an acceptable probability of error is achieved. For n approaching infinity, such source encoder approaches the optimal compression ratio $\mathbb{H}(A)$ with vanishing probability of error.

2.5 Constant Composition Distribution Matching

We have seen that for any DM, the average distribution $P_{\bar{A}}$ plays a central role. We may therefore consider a DM with k input bits where all length n output sequences have the same empirical distribution $P_{\bar{A}}$. This requires that $P_{\bar{A}}$ is a type in $\mathcal{P}_n(\mathcal{A})$. Such DMs are called CCDM [65]. The optimization problem to find cost optimal CCDMs is

$$\underset{P_A \in \mathcal{P}_n(\mathcal{A})}{\text{minimize}} \quad \mathbb{E}[w(A)] \tag{2.44}$$

$$\text{subject to} \quad \frac{\log_2 |\mathcal{T}^n(P_A)|}{n} \geq \frac{k}{n}. \tag{2.45}$$

2.5.1 Calculating the Optimal Type

Problem (2.44)–(2.45) can be solved optimally by a strategy similar to the average distribution calculation that we described in Section 2.4.1:

1. Calculate type class size and weight of all types in $\mathcal{P}_n(\mathcal{A})$.
2. Index types, type class sizes, and weights as $\hat{P}_1, \hat{P}_2, \ldots, N_1, N_2, \ldots$ and $w_1, w_2, \ldots$, respectively.
3. Among all types whose type class is large enough, select the one with minimum weight, i.e.,

$$i^* = \underset{i\colon \log_2 N_i \geq k}{\arg\min} \; w_i. \tag{2.46}$$

The cost optimal type is then $\hat{P}_{i^*}$.

Example 2.6. For our 8-ASK running example, we compare in Figure 2.2 for $n = 200$ the CCDM cost to the MB source cost. The CCDM cost penalty compared to the MB source is around 0.2 dB for low rates and increases to 0.6 dB for high rates. The penalty of the MCDM is significantly lower, over the whole range of rates.

2.5.2 Distribution Quantization

Suppose we want to use the distribution P_A in our system by using a CCDM with output length n, however, P_A is not n-type. Thus, we must approximate P_A by an n-type distribution $P_{A'}$. Here, we will quantify how good $P_{A'}$ approximates P_A by the variational distance (VD), which is given by

$$\|P_A - P_{A'}\|_1 = \sum_{a \in \mathcal{A}} |P_A(a) - P_{A'}(a)|. \tag{2.47}$$

Other measures, e.g., the informational divergence, can be used; see for example [11]. We will first argue why the VD is a reasonable choice for our purposes and we will then state a simple algorithm to find an n-type approximation and bound the approximation error.

2.5.3 Cost and Entropy under Quantization

Two important parameters for system design are cost and rate. Suppose the VD of P_A and $P_{A'}$ is equal to δ. Let $w_{\max}$ be the largest cost of symbols in $\mathcal{A}$. The cost resulting from using $P_{A'}$ is then bounded above and below by

$$\mathbb{E}[w(A)] - \delta w_{\max} \leq \mathbb{E}[w(A')] \leq \mathbb{E}[w(A)] + \delta w_{\max}. \tag{2.48}$$

In particular, as the VD δ approaches zero, the cost $\mathbb{E}[w(A')]$ approaches the desired cost $\mathbb{E}[w(A)]$.

Next, consider entropy. By the continuity of entropy (A.22), if $\delta \leq \frac{1}{2}$, we have

$$\mathbb{H}(A) + \delta \log_2 \frac{\delta}{|\mathcal{A}|} \leq \mathbb{H}(A') \leq \mathbb{H}(A) - \delta \log_2 \frac{\delta}{|\mathcal{A}|}. \tag{2.49}$$

Again, as δ approaches zero, the entropy $\mathbb{H}(A')$ approaches the desired entropy $\mathbb{H}(A)$.

2.5.4 Variational Distance n-Type Approximation Algorithm

The following algorithm calculates an n-type approximation $P_{A'}$ for an arbitrary distribution P_A.

1. For each $a \in \mathcal{A}$, calculate

$$Q(a) = \frac{\lfloor nP_A(a) \rfloor}{n} \tag{2.50}$$

and define

$$L = n - \sum_{a \in \mathcal{A}} Q(a)n. \tag{2.51}$$

Note that by definition, L is an integer.

2. For L symbols with largest approximation error $P_A(a) - Q(a)$, assign $P_{A'}(a) = Q(a) + \frac{1}{n}$. For the remaining symbols, assign $P_{A'}(a) = Q(a)$.

The algorithm immediately implies

$$|P_{A'}(a) - P_A(a)| < \frac{1}{n}, \quad a \in \mathcal{A} \tag{2.52}$$

and

$$\|P_{A'} - P_A\|_1 < \frac{|\mathcal{A}|}{n}. \tag{2.53}$$

Example 2.7. Consider the distribution $P_A(0) = 1 - P_A(1) = 1/\pi$ and $n = 1 \times 10^3$. The rounding step of the algorithm yields

$$Q(0) = \frac{318}{1000}, \quad Q(1) = \frac{681}{1000} \tag{2.54}$$

and $L = 1000 - 318 - 681 = 1$. The approximation errors are

$$P_A(0) - Q(0) \approx 3.1 \times 10^{-4}, \quad P_A(1) - Q(1) \approx 6.9 \times 10^{-4} \tag{2.55}$$

so we increase $Q(1)$ by $1/n$ and leave $Q(0)$ unchanged. The resulting n-type approximation is

$$P_{A'}(0) = \frac{318}{1000}, \quad P_{A'}(1) = \frac{682}{1000}. \tag{2.56}$$

The VD is

$$\|P_{A'} - P_A\|_1 \approx 6.1977 \times 10^{-4}. \tag{2.57}$$

Remark 2.3. In [11], it is shown that the above algorithm is optimal in terms of VD, and furthermore, the bound (2.53) is tightened.

2.5.5 CCDM Rate Bound

Let $P_{\bar{A}}$ be an n-type and consider a CCDM that outputs sequences of type $P_{\bar{A}}$. The CCDM rate is

$$R_{\text{ccdm}}(P_{\bar{A}}, n) = \frac{k}{n} = \frac{\lfloor \log_2 \mathcal{T}^n(P_{\bar{A}}) \rfloor}{n} = \frac{\left\lfloor \log_2 \binom{n}{n_1, \ldots, n_M} \right\rfloor}{n} \tag{2.58}$$

where $n_1, n_2, \ldots, n_M$ specifies the type, $M = |\mathcal{A}|$. The average distribution of a CCDM with type $P_{\bar{A}}$ is $P_{\bar{A}}$, so that by (2.18)–(2.20), we know that the CCDM rate is bounded from above by

$$R_{\text{ccdm}}(P_{\bar{A}}, n) \leq \mathbb{H}(\bar{A}). \tag{2.59}$$

The following theorem quantifies the difference between the CCDM rate $R_{\text{ccdm}}(P_{\bar{A}}, n)$ and $\mathbb{H}(\bar{A})$ as a function of n.

Theorem 2.1. Let $P_{\bar{A}}$ be an n-type with letters in $\mathcal{A}$, $|\mathcal{A}| = M$. The rate of a CCDM outputting type $P_{\bar{A}}$ is bounded from above by

$$\begin{aligned} &R_{\text{ccdm}}(P_{\bar{A}}, n) \\ &\leq \mathbb{H}(\bar{A}) + \frac{\log_2 \frac{e}{(2\pi)^{\frac{M}{2}}}}{n} - \frac{\frac{1}{2} \log_2 \prod_{i=1}^{M} P_{\bar{A}}(i)}{n} - \frac{M-1}{2} \frac{\log_2 n}{n} \end{aligned} \tag{2.60}$$

and it is bounded from below by

$$\begin{aligned} &R_{\text{ccdm}}(P_{\bar{A}}, n) \\ &\geq \mathbb{H}(\bar{A}) - \frac{1}{n} + \frac{\log_2 \frac{\sqrt{2\pi}}{e^M}}{n} - \frac{\frac{1}{2} \log_2 \prod_{i=1}^{M} P_{\bar{A}}(i)}{n} - \frac{M-1}{2} \frac{\log_2 n}{n}. \end{aligned} \tag{2.61}$$

Proof. We provide the proof in Section 2.7. □

Remark 2.4. By appropriate reordering of the terms in (2.60) and (2.61), Theorem 2.1 provides bounds on the CCDM rate loss $\mathbb{H}(\bar{A}) - R_{\text{ccdm}}(P_{\bar{A}}, n)$.

Example 2.8. Consider the 4-ASK constellation $\mathcal{X} = \{\pm 1, \pm 3\}$ and the block length $n = 4$. The amplitudes are $\mathcal{A} = \{1, 3\}$. Suppose the desired amplitude distribution is

$$P_{\bar{A}}(1) = \frac{3}{4}, \quad P_{\bar{A}}(3) = \frac{1}{4}. \tag{2.62}$$

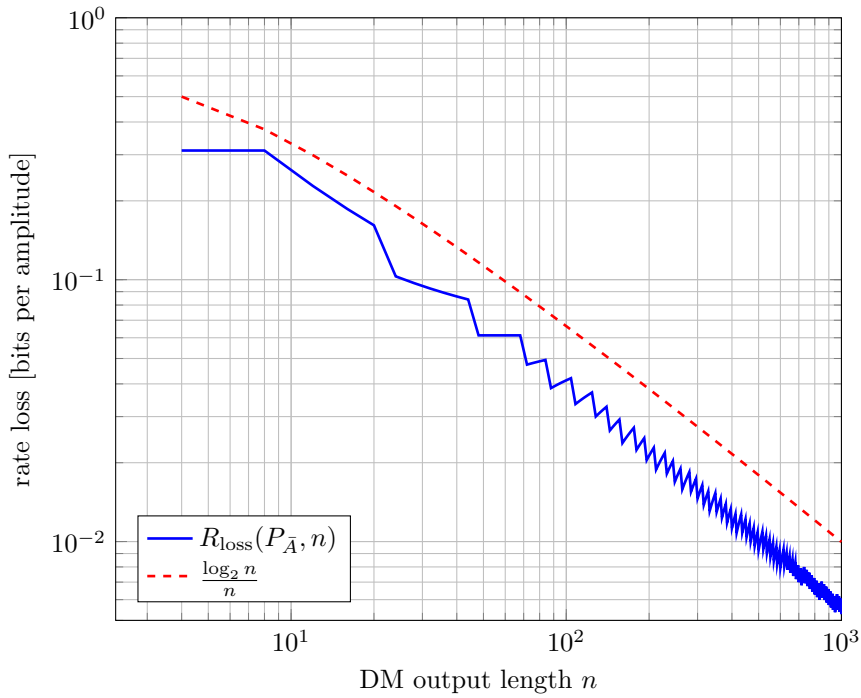


Figure 2.4: CCDM rate loss in Example 2.8.

The probabilities are integer multiples of $1/4$, so $P_{\bar{A}}$ is a 4-type. The corresponding 4-type class is

$$\mathcal{T}^4(P_{\bar{A}}) = \{(1,1,1,3), (1,1,3,1), (1,3,1,1), (3,1,1,1)\} \tag{2.63}$$

where in each sequence, the amplitudes 1 and 3 occur $n_1 = 3$ and $n_3 = 1$ times, respectively. A CCDM maps length k binary strings to sequences in $\mathcal{T}^4(P_{\bar{A}})$. There are 4 sequences in $\mathcal{T}^4(P_{\bar{A}})$, so we have

$$k = \log_2 |\mathcal{T}^4(P_{\bar{A}})| = 2. \tag{2.64}$$

The following lookup table (LUT) defines a CCDM.

$$\begin{aligned} 00 &\mapsto (1,1,1,3), \quad 01 \mapsto (1,1,3,1), \\ 10 &\mapsto (1,3,1,1), \quad 11 \mapsto (3,1,1,1). \end{aligned} \tag{2.65}$$

The mapping is one-to-one and therefore invertible on its image. The CCDM rate is

$$R_{\mathrm{ccdm}}(P_{\bar{A}}, 4) = \frac{k}{n} = \frac{\log_2 4}{4} = \frac{1}{2}. \tag{2.66}$$

The entropy of $P_{\bar{A}}$ is

$$\mathbb{H}(\bar{A}) = 0.8113. \tag{2.67}$$

Consequently, the rate loss is

$$R_{\mathrm{loss}}(P_{\bar{A}}, 4) = \mathbb{H}(\bar{A}) - R_{\mathrm{ccdm}}(P_{\bar{A}}, 4) \tag{2.68}$$

$$= 0.3113 \quad \left[\frac{\text{bit}}{\text{amplitude}}\right]. \tag{2.69}$$

For $n = 10\,000$, the CCDM rate is

$$R_{\mathrm{ccdm}}(P_{\bar{A}}, 10\,000) = \frac{8106}{10\,000} = 0.8106 \tag{2.70}$$

which is within 7×10^{-4} bits of $\mathbb{H}(\bar{A})$. In Figure 2.4, we display the rate loss for $n = 4, 8, \ldots, 1 \times 10^3$.

2.5.6 CCDM Decoder Is a Bad Source Encoder

Consider a CCDM with input length k, output length n, and n-type $P_{\bar{A}}$ on alphabet $\mathcal{A}$ of size $M = |\mathcal{A}| > 1$. In the following, we will consider the effect of letting n approach infinity. To limit the dependency of the n-type distribution $P_{\bar{A}}$ on n, we assume that the entries of $P_{\bar{A}}$ are bounded from below by a small positive value $p_{\min}$, independent of n, i.e.,

$$\forall a \in \mathcal{A}\colon\ P_{\bar{A}}(a) \geq p_{\min}. \tag{2.71}$$

The CCDM encoder f maps k bits to length n sequences in $\mathcal{T}^n(P_{\bar{A}})$, i.e.,

$$f\colon \{0,1\}^k \to \mathcal{T}^n(P_{\bar{A}}). \tag{2.72}$$

The encoder mapping f must be injective, which requires $2^k \leq |\mathcal{T}^n(P_{\bar{A}})|$. If $|\mathcal{T}^n(P_{\bar{A}})|$ is not an integer power of two, the image of encoder f must be a strict subset of $\mathcal{T}^n(P_{\bar{A}})$. The CCDM decoder g maps length n

sequences in $\mathcal{T}^n(P_{\bar{A}})$ to k bits, and one may wonder if the CCDM decoder g provides a useful source encoder. Specifically, consider a discrete memoryless source $P_{\bar{A}}$ outputting sequence $A^n = A_1 A_2 \dots A_n$, where the A_i are iid $\sim P_{\bar{A}}$. We now source encode A^n to k bits by using the CCDM as follows.

- If A^n is in the image of CCDM encoder f, we pass A^n to the CCDM decoder g and output the k bits $g(A^n)$.
- If A^n is not in the image of CCDM encoder f, we throw an error.

This source encoder is useful, if the error probability is low. We bound the probability of successful source encoding from above.

$$\begin{aligned}
\Pr\left[A^n \in f(\{0,1\}^k)\right] &\le \Pr\left[A^n \in \mathcal{T}^n(P_{\bar{A}})\right] && (2.73)\\
&= |\mathcal{T}^n(P_{\bar{A}})| 2^{-n\,\mathbb{H}(\bar{A})} && (2.74)\\
&\le \underbrace{\frac{e}{(2\pi)^{\frac{M}{2}}}}_{=:K} \cdot 2^{n\,\mathbb{H}(\bar{A})} \sqrt{\frac{n}{\prod_{a\in\mathcal{A}} n P_{\bar{A}}(a)}}\, 2^{-n\,\mathbb{H}(\bar{A})} && (2.75)\\
&= K \sqrt{\frac{n}{n^M} \frac{1}{\prod_{a\in\mathcal{A}} P_{\bar{A}}(a)}} && (2.76)\\
&\le K \sqrt{\frac{1}{n^{M-1} p_{\min}^M}} && (2.77)\\
&\xrightarrow{n\to\infty} 0 && (2.78)
\end{aligned}$$

where the inequality in (2.75) follows from (2.100)–(2.102) and where (2.77) follows from (2.71). Thus, with increasing output length, the probability of successful source encoding approaches 0, and we conclude that a CCDM decoder is a bad source encoder.

2.6 Cost and Rate Scaling

We observe in Figure 2.2 that for length $n = 200$ symbols, the MCDM is around 0.1 dB less energy efficient than the optimal MB source. In this section, we address the question how the cost penalty, and correspondingly, the rate loss, scale with the length n. What we will show is that both cost penalty and rate loss scale as $\frac{\log n}{n}$.

The main challenge in deriving this property is that when comparing a DM distribution P_{A^n} to the optimal distribution $P^n_{A^*}$, then these two distributions may differ both in cost and rate, while we are interested in comparing differences in cost at equal rate and differences in entropy at equal cost.

To bound the penalties from above, we proceed as follows. First, we restrict ourselves to CCDM, as MCDM is optimal and therefore has rate and cost scaling at least as good as CCDM. Second, we consider a (possibly suboptimal) CCDM by quantization, as this allows us to use the continuity of cost and entropy to bound the penalties. Finally, we will make use of a result from convex optimization to translate differences in cost into differences in entropy and vice-versa.

To bound the penalties from below, we use a result on divergence scaling [50].

2.6.1 Cost Scaling

Cost penalty upper bound We start with an MB distribution $P_{A'}$ with parameter ν' on the entropy-cost curve. We quantize $P_{A'}$ using the VD optimal n-type approximation from Section 2.5.4. We denote the resulting type by $P_{\bar{A}}$ and the distribution of the CCDM with composition $P_{\bar{A}}$ by P_{A^n}. Note that $P_{\bar{A}}$ is the average distribution of P_{A^n} and determines the cost. The resulting VD of $P_{A'}$ and its quantization is

$$\|P_{A'} - P_{\bar{A}}\| = \delta \leq \frac{M}{n} \tag{2.79}$$

where $M = |\mathcal{A}|$ is the size of the alphabet. Denote now by P_{A^*} the optimal distribution with $\mathbb{H}(A^*) = \frac{1}{n}\mathbb{H}(A^n)$. We illustrate the relation between P_{A^n}, $P_{A'}$, and P_{A^*} in Figure 2.5.

By (2.48), the cost penalty is bounded by

$$\mathbb{E}[w(\bar{A})] - \mathbb{E}[w(A^*)] \leq \delta w_{\max} + \mathbb{E}[w(A')] - \mathbb{E}[w(A^*)]. \tag{2.80}$$

By (2.29), the slope of the entropy-cost curve in $H' = \mathbb{H}(A')$ is

$$\frac{\partial W}{\partial H}(H') = \frac{1}{\nu'}. \tag{2.81}$$

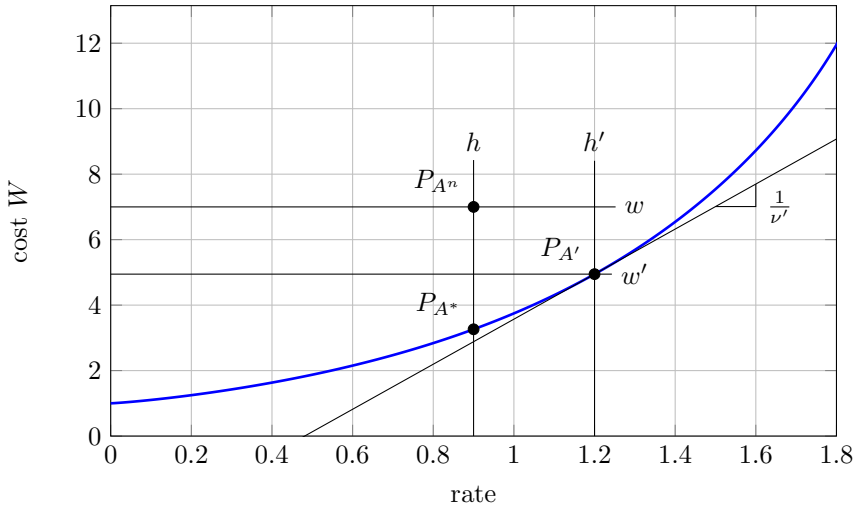


Figure 2.5: Illustration for deriving the cost penalty upper bound. In the figure, $w' = \mathbb{E}[w(A')]$, $w = \mathbb{E}[w(\bar{A})]$, $h = \mathbb{H}(A^n)/n$, $h' = \mathbb{H}(A')$. The goal is to bound $w - \mathbb{E}[w(A^*)]$.

We thus have

$$\mathbb{E}[w(A')] - \mathbb{E}[w(A^*)] \leq \frac{1}{\nu'} \left[\mathbb{H}(A') - \frac{\mathbb{H}(A^n)}{n} \right]. \tag{2.82}$$

We bound the entropy difference by

$$\mathbb{H}(A') - \frac{\mathbb{H}(A^n)}{n} \leq \mathbb{H}(A') - \mathbb{H}(\bar{A}) + \mathcal{O}\left(\frac{\log n}{n}\right) \tag{2.83}$$

$$\leq \delta \log_2 \frac{M}{\delta} + \mathcal{O}\left(\frac{\log n}{n}\right). \tag{2.84}$$

where the first inequality follows by (2.61) and the second inequality follows by (2.49). Thus, the cost penalty for MCDM and CCDM is bounded by

$$\mathbb{E}[w(\bar{A})] - \mathbb{E}[w(A^*)] \leq \delta w_{\max} + \frac{1}{\nu'} \left[\delta \log_2 \frac{M}{\delta} + \mathcal{O}\left(\frac{\log n}{n}\right) \right] \tag{2.85}$$

$$\leq \frac{M}{n} w_{\max} + \frac{1}{\nu'} \left[M \frac{\log_2 n}{n} + \mathcal{O}\left(\frac{\log n}{n}\right) \right] \tag{2.86}$$

$$= \mathcal{O}\left(\frac{\log n}{n}\right) \tag{2.87}$$

where the second inequality follows by (2.53).

Cost penalty lower bound Consider an MB distribution P_{A^*} on the entropy-cost curve and consider any DM with k uniformly distributed input bits and output distribution P_{A^n} with $\mathbb{H}(A^n)/n = \mathbb{H}(A^*)$. By [50, Theorem 3], we have

$$\frac{1}{n}\,\mathbb{D}(P_{A^n}\|P_{A^*}^n) = \Omega\left(\frac{\log n}{n}\right). \tag{2.88}$$

For the divergence, we have

$$\begin{aligned}
\frac{1}{n}\,\mathbb{D}(P_{A^n}\|P_{A^*}^n) &= \mathbb{X}(P_{\bar{A}}\|P_{A^*}) - \frac{1}{n}\,\mathbb{H}(A^n) && (2.89)\\
&= \nu\,\mathbb{E}[w(\bar{A})] + \log_2\left(\sum_{a'\in\mathcal{A}} 2^{-\nu w(a')}\right) - \frac{1}{n}\,\mathbb{H}(A^n) && (2.90)\\
&= \nu\,\mathbb{E}[w(\bar{A})] + \log_2\left(\sum_{a'\in\mathcal{A}} 2^{-\nu w(a')}\right) - \mathbb{H}(A^*) && (2.91)\\
&= \nu\,\mathbb{E}[w(\bar{A})] - \nu\,\mathbb{E}[w(A^*)] && (2.92)
\end{aligned}$$

where (2.90) follows by (2.6), (2.91) follows by assumption, and (2.92) again follows by (2.6). We conclude that for any DM with uniform input and rate $\mathbb{H}(A^*)$, the cost penalty is bounded from below by $\frac{1}{n}\log n$ asymptotically, i.e.,

$$\mathbb{E}[w(\bar{A})] - \mathbb{E}[w(A^*)] = \Omega\left(\frac{\log n}{n}\right). \tag{2.93}$$

2.6.2 Rate Scaling

The rate loss can be bounded using the same derivations as for the cost penalty.

Rate loss upper bound Consider an MB distribution $P_{A'}$ on the cost-entropy curve, a CCDM with uniform input, output distribution P_{A^n}, average distribution $P_{\bar{A}}$ that quantizes $P_{A'}$, and consider an MB distribution P_{A^*} on the entropy-cost curve with $\mathbb{E}[w(A^*)] = \mathbb{E}[w(\bar{A})]$. We illustrate the relation between $P_{A'}$, P_{A^n}, and P_{A^*} in Figure 2.6. Omitting the details, the rate loss is bounded from above by $\frac{1}{n}\log n$ asymptotically, i.e.,

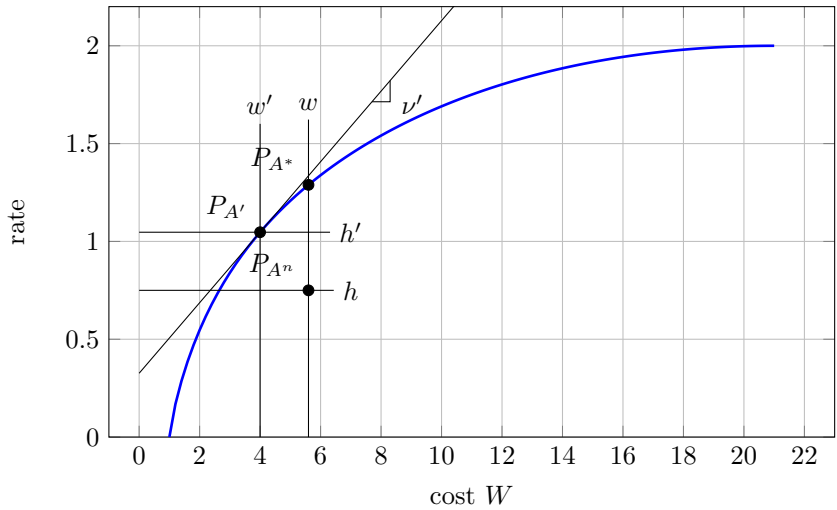


Figure 2.6: Illustration for deriving the rate loss upper bound. In the figure, $w' = \mathbb{E}[w(A')]$, $w = \mathbb{E}[w(\bar{A})]$, $h = \mathbb{H}(A^n)/n$, $h' = \mathbb{H}(A')$. The goal is to bound $\mathbb{H}(A^*) - h$.

$$\mathbb{H}(A^*) - \frac{\mathbb{H}(A^n)}{n} = \mathcal{O}\left(\frac{\log n}{n}\right). \tag{2.94}$$

As the MCDM has higher rate than the CCDM, this bound also holds for the MCDM.

Rate loss lower bound Consider an MB distribution P_{A^*} on the cost-entropy curve and consider any DM with uniform input, output distribution P_{A^n}, average distribution $P_{\bar{A}}$ and $\mathbb{E}[w(\bar{A})] = \mathbb{E}[w(A^*)]$. Then, by [50, Theorem 3] and omitting the details, the rate loss is bounded from below by $\frac{1}{n}\log n$, i.e.,

$$\mathbb{H}(A^*) - \frac{\mathbb{H}(A^n)}{n} = \Omega\left(\frac{\log n}{n}\right). \tag{2.95}$$

2.7 Proofs

We now prove Theorem 2.1. In the proof, we will use Stirling's formula (A.5), which provides the upper bound

$$n! < \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} e^{\frac{1}{12n}}. \tag{2.96}$$

To get rid of the $e^{\frac{1}{12n}}$ term, which depends on n, we use instead

$$n! \leq e \cdot n^{n+\frac{1}{2}} e^{-n}. \tag{2.97}$$

For $n = 1$, (2.97) holds with equality, and for $n \geq 2$, we have $\sqrt{2\pi} e^{\frac{1}{12n}} < e$, so (2.97) follows by (2.96). Stirling's formula (A.5) provides the lower bound

$$n! > \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} e^{\frac{1}{12n+1}} > \sqrt{2\pi} \cdot n^{n+\frac{1}{2}} e^{-n}. \tag{2.98}$$

Let $P_{\bar{A}}$ be an n-type on an alphabet of size M. We will need in our proof the identity

$$\frac{n^n}{n_1^{n_1} n_2^{n_2} \cdots n_M^{n_M}} = 2^{n\,\mathbb{H}(\bar{A})} \tag{2.99}$$

which is stated, e.g., in [32, Section 2.1].

For the size of the n-type class, we now have

$$|\mathcal{T}^n(P_{\bar{A}})| = \frac{n!}{n_1! \cdots n_M!} < \frac{e n^{n+\frac{1}{2}} e^{-n}}{\prod_{i=1}^M \sqrt{2\pi} n_i^{n_i+\frac{1}{2}} e^{-n_i}} \tag{2.100}$$

$$= \frac{e}{(2\pi)^{\frac{M}{2}}} \cdot \frac{n^{n+\frac{1}{2}}}{\prod_{i=1}^M n_i^{n_i+\frac{1}{2}}} \cdot \frac{e^{-n}}{e^{-(n_1+n_2+\cdots+n_M)}} \tag{2.101}$$

$$= \underbrace{\frac{e}{(2\pi)^{\frac{M}{2}}}}_{=:K_1} \cdot 2^{n\,\mathbb{H}(\bar{A})} \sqrt{\frac{n}{\prod_{i=1}^M n_i}} \tag{2.102}$$

where

- (2.100) follows by using (2.97) and (2.98),
- (2.102) follows by (2.99) and $n_1 + \cdots + n_M = n$.

Taking the logarithm and dividing by n, we get

$$\frac{\log_2 |\mathcal{T}^n(P_{\bar{A}})|}{n} \leq \mathbb{H}(\bar{A}) + \frac{\log_2 K_1}{n} + \frac{1}{2n} \log_2 \frac{n}{\prod_{i=1}^M P_{\bar{A}}(i) n} \tag{2.103}$$

$$= \mathbb{H}(\bar{A}) + \frac{\log_2 K_1 - \frac{1}{2} \log_2 \prod_{i=1}^M P_{\bar{A}}(i)}{n} - \frac{M-1}{2} \frac{\log_2 n}{n}. \tag{2.104}$$

Along the same lines, we get the lower bound

$$\frac{\log_2 |\mathcal{T}^n(P_{\bar{A}})|}{n} \geq \mathbb{H}(\bar{A}) + \frac{\log_2 K_2 - \frac{1}{2}\log_2 \prod_{i=1}^{M} P_{\bar{A}}(i)}{n} - \frac{M-1}{2}\frac{\log_2 n}{n}. \tag{2.105}$$

where

$$K_2 = \frac{\sqrt{2\pi}}{e^M}. \tag{2.106}$$

Finally, the CCDM rate is bounded below and above by

$$\frac{\log_2 |\mathcal{T}^n(P_{\bar{A}})|}{n} - \frac{1}{n} < \frac{\lfloor \log_2 |\mathcal{T}^n(P_{\bar{A}})| \rfloor}{n} \leq \frac{\log_2 |\mathcal{T}^n(P_{\bar{A}})|}{n} \tag{2.107}$$

The rate loss

$$R_{\text{loss}}(P_{\bar{A}}, n) = \mathbb{H}(\bar{A}) - \frac{\lfloor \log_2 |\mathcal{T}^n(P_{\bar{A}})| \rfloor}{n} \tag{2.108}$$

is now by (2.107) and (2.104) bounded above by $\frac{\log n}{n}$ asymptotically, i.e., $R_{\text{loss}}(P_{\bar{A}}, n) \in \mathcal{O}(\frac{\log n}{n})$, and by (2.107) and (2.105), it is bounded below by $\frac{\log n}{n}$ asymptotically, i.e., $R_{\text{loss}}(P_{\bar{A}}, n) \in \Omega(\frac{\log n}{n})$. This implies

$$R_{\text{loss}}(P_{\bar{A}}, n) \in \Theta\left(\frac{\log n}{n}\right). \tag{2.109}$$

2.8 Discussion

The term 'distribution matching' was first used in the title of [22]. The work [22] introduces geometric Huffman coding (GHC), an algorithm to construct variable-to-fixed length prefix-free codes for DM. GHC constructs codes different from Huffman codes, and the key insight in [22] is that DM by decoding a Huffman code as proposed, e.g., in [51, Section VIII.D], is suboptimal. The notes [17] discuss in detail optimal variable length source coding and DM, illustrating the differences. The observation we made in Section 2.4.2 ("MCDM decoders are bad source encoders") is empirical; future work may consider to study this observation theoretically, e.g., by bounding the probability of successful source encoding away from zero, similar to as we did for CCDMs in Section 2.5.6.

The work [68] states the $\frac{\log n}{n}$ normalized informational divergence (ID) scaling for binary DMs and [50] generalizes this result to arbitrary alphabet sizes. In [2], [4], [5], [22], it is shown that for variable length DMs, normalized ID scales as $1/n$. The relation between normalized ID and entropy rate is studied in [10]. Resolution coding discussed in [9] achieves a vanishing unnormalized ID, but is not invertible. Invertible low-divergence (ILD) coding proposed in [67] achieves vanishing unnormalized ID while being invertible, assuming access to additional random bits during encoding.

Algorithms for n-type approximations are discussed in [11] and [41].

Algorithms to implement variable length DMs are described in [2], [4], [5], [22], [52].

Fixed length DM algorithms are studied, e.g., in (alphabetically ordered) [12], [14], [29], [33]–[35], [45], [46], [57], [58], [60], [65], [66], [69], [70], [81].

PhD theses on DMs are [5], [42], [57], [66]. An overview paper on DM algorithms is [44].

The version of record is available at: http://dx.doi.org/10.1561/0100000111

3

Achievable Rates

In this section, we study a layered probabilistic shaping (PS) ensemble and its achievable rates. The motivation for this study lies in the (somewhat surprising) information-theoretic difficulties we encountered during our development of PAS in Section 1. In particular, in Section 1.3.2, we raised the question what the SE of the preliminary system without explicit encoding could be, and the natural follow-up question is what an *achievable* SE could be. As we will see, while the layered PS ensemble is more general than PAS, it appears to be an appropriate framework for answering the original questions, which we point out in the following for instance in Remark 3.1, Section 3.3, and Remark 3.2.

In Section 3.1, we define layered PS, which consists of an inner layer for FEC and an outer layer for PS. In the PS layer, message bits are mapped to FEC encoder inputs that map to channel input sequences in a shaping set. The shaping set specifies desired properties. For instance, it may consist of all sequences that have a capacity-achieving distribution for the considered channel. By random coding arguments, we analyze in Section 3.2 the probability of encoding failure, which results in a formula for the SE. In Section 3.4, we then analyze the decoding failure, which results in a formula for the achievable FEC rate. Finally, in

Section 3.5, we consider memoryless channels and we provide a formula for the achievable SE and we show that the layered PS architecture is capacity-achieving.

3.1 Layered Probabilistic Shaping

Consider a channel with finite input alphabet $\mathcal{X}$ and define its size in bits by

$$m = \log_2 |\mathcal{X}|. \tag{3.1}$$

The channel output alphabet can be continuous or discrete.

3.1.1 Classical Random Code Ensemble

The classical random code ensemble [39, Chapter 5] for a channel with input alphabet $\mathcal{X}$ and codeword length n symbols in $\mathcal{X}$ is

$$\mathcal{C} = \left\{ C^n(w), w = 1, 2, \ldots, 2^{nmR_{\text{fec}}} \right\} \tag{3.2}$$

where the entries of the $|\mathcal{C}| = 2^{nmR_{\text{fec}}}$ codewords are independently and identically distributed according to P_X on the constellation $\mathcal{X}$. We require $0 \leq R_{\text{fec}} \leq 1$ so that $mR_{\text{fec}} \leq \log_2 |\mathcal{X}|$. By [39, Equation (5.2.5)], the decoding rule for a memoryless channel with transition density $p_{Y|X}$ is the ML rule

$$\hat{w} = \underset{w \in \{1,\ldots,|\mathcal{C}|\}}{\arg\max} \prod_{i=1}^{n} p_{Y|X}(y_i | c_i(w)) \tag{3.3}$$

where y^n is the sequence observed at the channel output. The SE in bits per channel use is $\text{SE} = mR_{\text{fec}}$ and the classical random code ensemble achieves

$$\text{SE}^* = \mathbb{I}(X;Y). \tag{3.4}$$

In particular, it achieves the capacity $\max_{P_X} \mathbb{I}(X;Y)$ when the optimal P_X is used. When instead of $p_{Y|X}$ some other function q is used, the ML rule (3.3) becomes the mismatched rule

$$\hat{w} = \underset{w \in \{1,\ldots,|\mathcal{C}|\}}{\arg\max} \prod_{i=1}^{n} q(y_i, c_i(w)). \tag{3.5}$$

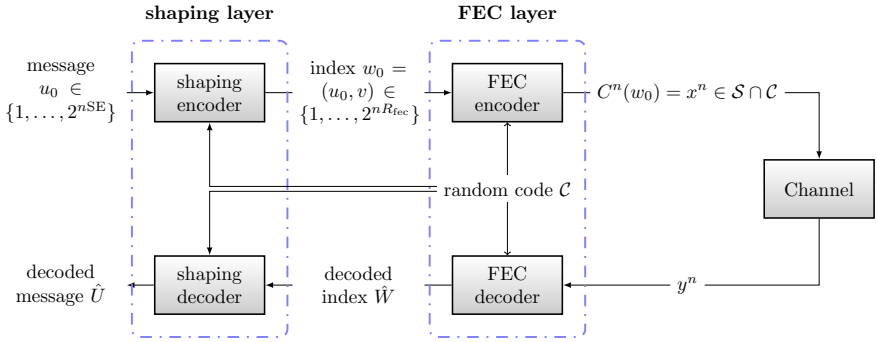


Figure 3.1: The layered PS architecture discussed in Section 3.1. In PAS, the FEC encoder is systematic and the shaping encoder is realized by a DM that shapes the systematic symbols. The shaping encoder of PAS is zero error.

3.1.2 Layered Random Code Ensemble

The layered PS architecture is displayed in Figure 3.1. We consider the random code ensemble

$$\mathcal{C} = \left\{ C^n(w), w = 1, 2, \ldots, 2^{nmR_{\text{fec}}} \right\} \tag{3.6}$$

where the entries of the $|\mathcal{C}| = 2^{nmR_{\text{fec}}}$ codewords are chosen independently and *uniformly* distributed on the constellation $\mathcal{X}$. As above, we require $0 \leq R_{\text{fec}} \leq 1$.

Remark 3.1. Note that the classical random code ensemble of Section 3.1.1 samples the codeword entries according to the desired channel input distribution P_X. In contrast, layered PS always uses the uniform distribution. Sampling the codeword entries uniform at random is compatible with linear codes. For instance, the codeword entries of the linear code ensemble considered in [39, Section 6.2] have uniform distributions. In our development of PAS in Section 1, we exclusively considered linear codes. Our layered random code ensemble accounts for that by sampling the codeword entries uniformly. See also Sections 5.1.2 and 5.2, where we derive PAS error exponents by uniform sampling and random linear coding, respectively.

3.2 Encoding

We consider a general shaping set $\mathcal{S} \subseteq \mathcal{X}^n$. Define the shaping set rate by

$$R_{\text{ss}} = \frac{\log_2 |\mathcal{S}|}{nm}. \tag{3.7}$$

Note that by the definition of m in (3.1), $0 \leq R_{\text{ss}} \leq 1$. We divide the FEC code into $2^{n\text{SE}}$ partitions, so that the number of codewords in each partition is

$$\frac{2^{nmR_{\text{fec}}}}{2^{n\text{SE}}} = 2^{nm(R_{\text{fec}} - \frac{\text{SE}}{m})}. \tag{3.8}$$

The PS encoder maps message $u \in \{1, 2, \ldots, 2^{n\text{SE}}\}$ to a codeword in the uth partition that is in $\mathcal{S}$. By double indexing $\mathcal{C}$, the chosen codeword has index $w = (u, v)$ for some $v \in \{1, 2, \ldots, 2^{nm(R_{\text{fec}} - \frac{\text{SE}}{m})}\}$. An encoding error occurs if the PS encoder cannot find such a codeword, i.e., if there is no index v so that codeword $c^n(u, v)$ is in the shaping set $\mathcal{S}$.

Theorem 3.1. The probability that the PS encoder cannot map its input to a codeword in $\mathcal{S} \cap \mathcal{C}$ is upper bounded by

$$\begin{aligned} &\Pr(\text{PS encoding failure}) \\ &\quad \leq \exp\left(-2^{nm\left[1-(1-R_{\text{ss}})-(1-R_{\text{fec}})-\frac{\text{SE}}{m}\right]}\right). \end{aligned} \tag{3.9}$$

Proof. We provide the proof in Section 3.6.1. □

3.3 Spectral Efficiency, Rate, Overhead

By the theorem, the SE is determined by two overheads, namely the PS overhead and the FEC overhead, which we summarize in Table 3.1. For a desired SE, the overhead allocation is a degree of freedom that can be exploited in the transceiver design, for example, a low FEC overhead may be desirable in practical systems for complexity reasons. Note that in the classical random coding experiment, the SE is always equal to mR_{fec}. The dependence of SE on both PS and FEC overhead is exactly what we encountered in our development of PAS in Section 1, Sections 1.3.2 and 1.3.3. For instance, reducing $\mathbb{H}(X)$ reduced the SE, for unchanged R_{fec}.

Table 3.1: PS and FEC overheads. For $R_{\text{ss}} = 1$ (no shaping), the total overhead is equal to the FEC overhead. For $R_{\text{ss}} = 1 - R_{\text{fec}}$ (maximum shaping), the total overhead is infinity, corresponding to $\text{SE} = 0$. For $R_{\text{ss}} < 1 - R_{\text{fec}}$, the total overhead becomes negative, which is meaningless. This is in accordance with Theorem 3.1: in this case, even for $\text{SE} = 0$, the bound on the probability of encoding failure is equal to 1, asymptotically in n.

	FEC	Shaping Set
Rate	R_{fec}	$R_{\text{ss}} = \frac{\log_2 \lvert\mathcal{S}\rvert}{nm}$
Redundancy	$1 - R_{\text{fec}}$	$1 - R_{\text{ss}}$
Overhead in %	$100 \cdot \left(\frac{1}{R_{\text{fec}}} - 1\right)$	$100 \cdot \left(\frac{1}{R_{\text{ss}}} - 1\right)$
Total overhead in %	$100 \cdot \left(\frac{1}{R_{\text{ss}}+R_{\text{fec}}-1} - 1\right)$	

3.4 Decoding

We consider a generic FEC decoder with a decoding metric q. For an observation y^n, the metric assigns to each sequence $x^n \in \mathcal{X}^n$ a non-negative score $q(x^n, y^n)$. The FEC encoder maps a message w to a codeword $c^n(w)$. For an observed output y^n, the decoder outputs as its decision the message that maps to the codeword with the maximum score, i.e,

$$\hat{w} = \underset{w \in \{1,\ldots,|\mathcal{C}|\}}{\arg\max}\ q\left(c^n(w), y^n\right). \tag{3.10}$$

Note that the decoder ignores the possible existence of a PS encoder in the sense that it *evaluates the metric over the entire codebook* $\mathcal{C}$ *and not over* $\mathcal{C} \cap \mathcal{S}$.

We explicitly allow the metric q to assign the score 0 to impossible input-output (I/O) pairs. For example, consider a noiseless binary symmetric channel (BSC) with input X, output Y, and crossover probability $\delta = 0$. We can describe it by the transition probabilities

$$P_{Y|X}(0|0) = P_{Y|X}(1|1) = 1 \tag{3.11}$$

$$P_{Y|X}(1|0) = P_{Y|X}(0|1) = 0. \tag{3.12}$$

A decoding metric for this channel is $q(x, y) = P_{Y|X}(y|x)$, which assigns score 0 to the impossible I/O pairs $(0, 1)$ and $(1, 0)$ and score 1 to

the possible I/O pairs $(0,0)$ and $(1,1)$. In the proof of the following Theorem 3.2 (see Section 3.6.2), we will need to divide by $q(x^n, y^n)$ for a *possible* I/O pair. For this technical reason, we define a non-negative decoding metric as a metric that assigns non-negative scores (including score 0) to impossible I/O pairs and strictly positive scores to possible I/O pairs. This definition is consistent with using channel transition probabilities as decoding metric.

Theorem 3.2. Suppose the codeword $C^n(w_0) = x^n$ is transmitted, let y^n be a channel output sequence, and let q be a non-negative decoding metric. Define the empirical cross-equivocation

$$\mathsf{x}(q, x^n, y^n) = -\frac{1}{n} \log_2 \frac{q(x^n, y^n)}{\sum_{a^n \in \mathcal{X}^n} q(a^n, y^n)}. \tag{3.13}$$

The probability that the decoder (3.10) does not recover the index w_0 from the sequence y^n is bounded from above by

$$\begin{aligned} &\Pr(\hat{W} \neq w_0 | C^n(w_0) = x^n, Y^n = y^n) \\ &\quad \leq 2^{-nm\left(1 - R_{\text{fec}} - \frac{\mathsf{x}(q, x^n, y^n)}{m}\right)}. \end{aligned} \tag{3.14}$$

Proof. We provide the proof in Section 3.6.2. □

Note that in Figure 3.1, if the index decision $\hat{W}$ is correct, then the shaping decoder can error-free recover the message u_0 from $\hat{W}$. That is, $\Pr(\hat{W} \neq w_0)$ upper bounds $\Pr(\hat{U} \neq u_0)$.

3.4.1 Achievable FEC Rate

Note that in Theorem 3.2, the transmitted codeword x^n and the noisy observation y^n are *deterministic* values. The code whose decoder attempts to recover x^n from y^n is *random*, i.e., the theorem bounds the probability of picking a bad code. This corresponds to what we observe when designing FEC codes in practice: with a small R_{fec}, it is easy to find a practical rate R_{fec} code that decodes correctly, corresponding to a vanishing probability of picking a bad code. The larger R_{fec} is (i.e., the smaller the backoff becomes), the more difficult it is to find a

practical rate R_{fec} code that decodes correctly. The error exponent in (3.14) suggest to define an achievable FEC rate by

$$R^*_{\text{fec}} = \left[1 - \frac{\mathsf{x}(q, x^n, y^n)}{m}\right]^+ \tag{3.15}$$

where $[\cdot]^+ = \max\{0, \cdot\}$ ensures non-negativity. For $R_{\text{fec}} < R^*_{\text{fec}}$, the error probability bound in (3.14) vanishes, asymptotically in n, and for $R_{\text{fec}} \geq R^*_{\text{fec}}$, it is equal to or exceeds the trivial bound 1.

By making the achievable FEC rate $R^*_{\text{fec}}(q, x^n, y^n)$ dependent on the input sequence x^n and the output sequence y^n, Theorem 3.2 holds in great generality and attaches a precise operational meaning to the empirical cross-equivocation for arbitrary channels. Consider now a binary channel and the decoding metric

$$q(x^n, y^n) = 2^{-\mathrm{d_H}(x^n, y^n)} \tag{3.16}$$

where $\mathrm{d_H}$ is the Hamming distance, which counts the number of entries in which x^n and y^n differ. We calculate the empirical cross-equivocation and the achievable FEC rate for various sequence pairs:

$$\mathsf{x}(q, 000, 000) = 0.585, \qquad R^*_{\text{fec}}(q, 000, 000) = 0.415 \tag{3.17}$$

$$\mathsf{x}(q, 000, 100) = 0.918, \qquad R^*_{\text{fec}}(q, 000, 100) = 0.082 \tag{3.18}$$

$$\mathsf{x}(q, 000, 110) = 1.25, \qquad R^*_{\text{fec}}(q, 000, 110) = 0. \tag{3.19}$$

We observe that, depending on the received output sequence, the achievable FEC rate may be 0.415, 0.08, or 0. Thus, we need to choose an FEC rate that is achievable for a sufficient fraction of sequence pairs that occur during transmission and this FEC rate then provides an achievable SE. We derive an achievable SE for memoryless channels in Section 3.5.

3.4.2 Tightened Empirical Cross-Equivocation

For $s > 0$, the non-negative metric q and the metric q^s implement exactly the same decision rule (3.10). Consequently, their error probability is the same. This allows us to tighten the error bound in Theorem 3.2 and

thereby improve upon the achievable FEC rate (3.15). The tightened empirical cross-equivocation is

$$\mathsf{x}^*(q, x^n, y^n) = \min_{s>0} \mathsf{x}(q^s, x^n, y^n). \tag{3.20}$$

3.4.3 Interleaving

In practical systems, we often use memoryless decoding metrics of the form

$$q(x^n, y^n) = \prod_{i=1}^{n} q(x_i, y_i). \tag{3.21}$$

For instance, the bitwise metrics (1.33) and (1.36) that we considered in our development of PAS in Section 1 are examples of memoryless metrics.

From (3.21), we can see that the value of the metric does not change when we use a length n block interleaver between the channel and the FEC components. Consequently, the empirical cross-equivocation does not change either, so the achievable FEC code rate cannot be improved by interleaving. The reason for this is twofold. First, the empirical cross-equivocation is calculated assuming the ML-like decoder (3.10). Second, the achievable FEC code rate is stated in terms of the probability of picking a code uniformly at random that achieves it. The selection of a code uniformly at random has the effect of a random interleaver between the FEC components and the channel, and consequently, any effect of choosing a specific interleaver is undone by the random coding experiment. When we design practical FEC codes with practical decoding, e.g., LDPC codes with iterative decoding, then the interleaver can indeed improve the performance, see, e.g., [47], [76]. Practical FEC codes and block interleavers should therefore be designed jointly.

3.5 Channel Coding Theorem

We now consider a memoryless channel

$$p_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i) \tag{3.22}$$

and memoryless decoding metrics

$$q(x^n, y^n) = \prod_{i=1}^{n} q(x_i, y_i). \tag{3.23}$$

Furthermore, we require that most sequences in the shaping set $\mathcal{S}$ have the distribution P_X, so that with high probability

$$\mathsf{x}(q, X^n, Y^n) \approx \mathbb{E}\left[\mathsf{x}(q, X, Y)\right] =: \mathsf{X}(q, X, Y) \tag{3.24}$$

where $\mathsf{X}(q, X, Y)$ is a cross-equivocation (A.26). This can be seen as follows. We have

$$\mathsf{X}(q, X, Y) = \mathbb{E}\Big[-\log_2 \underbrace{\frac{q(X, Y)}{\sum_{a\in\mathcal{X}} q(a, Y)}}_{=:Q_{X|Y}(X|Y)}\Big] \tag{3.25}$$

$$= \mathbb{E}\left[-\log_2 Q_{X|Y}(X|Y)\right] \tag{3.26}$$

$$= \mathbb{X}(P_{X|Y} \| Q_{X|Y} | p_Y) \tag{3.27}$$

where because of normalization, $Q_{X|Y}(\cdot|b)$ is a distribution on $\mathcal{X}$, for each b in the channel output alphabet.

By Theorems 3.1 and 3.2, following the line of arguments in [53] (leaving out the ϵs and δs) we arrive at the following channel coding theorem.

Theorem 3.3. For a shaping set with distribution P_X, an achievable spectral efficiency allowing for successful encoding and decoding with high probability is

$$\mathrm{SE}^* = \left[mR_{\mathrm{ss}} - \mathsf{X}(q, X, Y)\right]^+. \tag{3.28}$$

We can improve upon (3.28) by tightening the cross-equivocation following Section 3.4.2, i.e., define

$$\mathsf{X}^*(q, X, Y) = \min_{s>0} \mathsf{X}(q^s, X, Y). \tag{3.29}$$

This provides the following corollary, which improves upon Theorem 3.3.

Corollary 3.4. For a shaping set with distribution P_X, an achievable spectral efficiency allowing for successful encoding and decoding with high probability is

$$\mathrm{SE}^* = \left[mR_{\mathrm{ss}} - \mathsf{X}^*(q, X, Y)\right]^+. \tag{3.30}$$

We discuss in detail optimization over q and over s in Section 4.

3.5.1 Capacity-Achieving Symbol-Metric

We use as shaping set $\mathcal{S}$ all sequences with distribution P_X. For sufficiently large n, we have $R_{\rm ss}m \approx \mathbb{H}(X)$. With the decoding metric $P_{X|Y}$, we have $\mathsf{X}(q, X, Y) = \mathbb{H}(X|Y)$ and the achievable SE becomes equal to the mutual information $\mathbb{I}(X;Y)$, which shows that the layered PS architecture is capacity-achieving.

Remark 3.2. Note that the classical random code ensemble achieves capacity with an ML rule on a codebook of size $2^{n\mathrm{SE}}$ while layered PS achieves capacity with a MAP rule on a codebook of size $2^{n(\mathrm{SE}+m(1-R_{\rm ss}))}$, which is larger. Note that this is inline with what we observed in our development of PAS in Section 1, where switching from the bitwise ML metric (1.33) to the bitwise MAP metric (1.36) led to significant gains.

3.6 Proofs

3.6.1 Proof of Theorem 3.1

We consider a general shaping set $\mathcal{S} \subseteq \mathcal{X}^n$ and we consider the random code ensemble that we defined in Section 3.1.2. Note that we do not condition on a specific realization of the code $\mathcal{C}$, i.e., for each index $w = 1, 2, \ldots, 2^{nmR_{\rm fec}}$, a codeword is picked uniformly at random from $\mathcal{X}^n$, independent of the other codewords. We divide the FEC code into $2^{n\mathrm{SE}}$ partitions, so that the number of codewords in each partition is

$$\frac{2^{nmR_{\rm fec}}}{2^{n\mathrm{SE}}} = 2^{nm(R_{\rm fec}-\frac{\mathrm{SE}}{m})}. \tag{3.31}$$

The PS encoder maps message $u \in \{1, 2, \ldots, 2^{n\mathrm{SE}}\}$ to a codeword in the uth partition that is in $\mathcal{S}$. In our random coding experiment, each codeword is chosen uniformly at random from $\mathcal{X}^n$ and it is not in $\mathcal{S}$ with probability

$$\frac{|\mathcal{X}|^n - |\mathcal{S}|}{|\mathcal{X}|^n} = 1 - \frac{|\mathcal{S}|}{|\mathcal{X}|^n}. \tag{3.32}$$

For a considered message u, the probability that of the $2^{nm(R_{\text{fec}}-\frac{\text{SE}}{m})}$ codewords in the corresponding uth partition, none is in $\mathcal{S}$ is therefore

$$\left(1-\frac{|\mathcal{S}|}{|\mathcal{X}|^n}\right)^{2^{nm(R_{\text{fec}}-\frac{\text{SE}}{m})}} \leq \exp\left(-\frac{|\mathcal{S}|}{|\mathcal{X}|^n}2^{nm(R_{\text{fec}}-\frac{\text{SE}}{m})}\right)$$

where we used $1-s \leq \exp(-s)$. Finally, the exponent is

$$\frac{|\mathcal{S}|}{|\mathcal{X}|^n}2^{nm(R_{\text{fec}}-\frac{\text{SE}}{m})} = 2^{nm\left[R_{\text{fec}}-\frac{\text{SE}}{m}+\frac{1}{nm}\log_2|\mathcal{S}|-1\right]} \tag{3.33}$$

$$= 2^{nm\left[1-(1-R_{\text{ss}})-(1-R_{\text{fec}})-\frac{\text{SE}}{m}\right]}. \tag{3.34}$$

3.6.2 Proof of Theorem 3.2

Our proof is based on Markov's inequality and is similar in style to the proof of the channel coding theorem in [53].

Recall the setup from Section 3.4:

- Finite constellation $\mathcal{X}$ with cardinality $|\mathcal{X}| = 2^m$.
- Random code $\mathcal{C}$ with codewords in $\mathcal{X}^n$, cardinality $|\mathcal{C}| = 2^{nmR_{\text{fec}}}$ and FEC rate $R_{\text{fec}} = \frac{\log_2|\mathcal{C}|}{mn}$.
- The $2^{nmR_{\text{fec}}} \cdot n$ entries of the $2^{nmR_{\text{fec}}}$ codewords $C^n(w)$, $w = 1, 2, \ldots, 2^{nmR_{\text{fec}}}$, are generated independently and uniformly distributed on $\mathcal{X}$.
- For the observation y^n, the decoder decides

$$\hat{W} = \underset{w\in\{1,\ldots,|\mathcal{C}|\}}{\arg\max}\ q(C^n(w), y^n) \tag{3.35}$$

  where $q\colon \mathcal{X}^n \times \mathcal{Y}^n \to \mathbf{R}_{\geq 0}$ is a non-negative metric defined on $\mathcal{X}^n \times \mathcal{Y}^n$, and where $\mathcal{Y}$ is a possibly continuous and/or multi-dimensional output alphabet. As defined in Section 3.4, the metric is restricted to assign strictly positive scores to input-output pairs that are possible for the considered channel.

We condition on the event that index W was encoded to $C^n(W) = x^n$ and the sequence y^n was output by the channel. For notational

convenience, we assume without loss of generality $W = 1$. A decoding error $\hat{W} \neq 1$ implies the existence of a $w' \neq 1$ such that

$$L(w') := \frac{q(C^n(w'), y^n)}{q(x^n, y^n)} \geq 1 \Rightarrow \sum_{w=2}^{|\mathcal{C}|} L(w) \geq 1. \tag{3.36}$$

If event $\mathcal{A}$ implies event $\mathcal{B}$, then $\Pr(\mathcal{A}) \leq \Pr(\mathcal{B})$. Therefore, we have

$$\begin{aligned}
&\Pr(\hat{W} \neq 1 | X^n = x^n, Y^n = y^n) \\
&\leq \Pr\left[\sum_{w=2}^{|\mathcal{C}|} L(w) \geq 1 \,\middle|\, X^n = x^n, Y^n = y^n\right] && (3.37)\\
&\leq \mathbb{E}\left[\sum_{w=2}^{|\mathcal{C}|} L(w) \,\middle|\, X^n = x^n, Y^n = y^n\right] && (3.38)\\
&= q(x^n, y^n)^{-1}\, \mathbb{E}\left[\sum_{w=2}^{|\mathcal{C}|} q(C^n(w), y^n)\right] && (3.39)
\end{aligned}$$

where

- Inequality in (3.38) follows by Markov's inequality.
- Equality in (3.39) follows because for $w \neq 1$, the codeword $C^n(w)$ and the transmitted codeword $C^n(1)$ were generated independently so that $C^n(w)$ and $[C^n(1), Y^n]$ are independent and the conditioning on $Y^n = y^n$ can be dropped.

The right-hand side of (3.39) is further bounded as

$$\begin{aligned}
(3.39) &= (|\mathcal{C}| - 1) q(x^n, y^n)^{-1}\, \mathbb{E}\left[q(C^n, y^n)\right] && (3.40)\\
&\leq |\mathcal{C}| q(x^n, y^n)^{-1}\, \mathbb{E}\left[q(C^n, y^n)\right] && (3.41)\\
&= 2^{nmR_{\text{fec}}} q(x^n, y^n)^{-1} \sum_{x^n \in \mathcal{X}^n} |\mathcal{X}|^{-n} q(x^n, y^n) && (3.42)\\
&= 2^{-nm\left(1 + \frac{1}{nm} \log_2 \frac{q(x^n, y^n)}{\sum_{x^n \in \mathcal{X}^n} q(x^n, y^n)} - R_{\text{fec}}\right)} && (3.43)
\end{aligned}$$

where

- C^n in (3.40) is a random variable with entries that are independent and uniformly distributed on $\mathcal{X}$.

- Equality in (3.40) holds because in our random coding experiment, for each index w, we generated the codeword entries $C_1(w), C_2(w), \ldots, C_n(w)$ independent and uniformly distributed on $\mathcal{X}$.

3.7 Discussion

Since PAS is not a sample of the classical random code ensemble (see Remark 3.1, Section 3.3, and Remark 3.2), the calculation of appropriate achievable rates for PAS is intricate, and several attempts were taken [26, Section III.C], [28], [80]. In this section, we discussed layered PS, a random code ensemble that was developed in the line of work [7], [8], [13], [18], [19], [23]. Theorems 3.1 and 3.2 were first stated in [13, Theorem 1] and [13, Theorem 2], respectively. Layered PS is more general than PAS, e.g., it also covers probabilistic parity bit shaping as proposed in [21].

In [56], linear layered PS is studied for discrete memoryless channels, and it is shown that it achieves capacity and the error exponent [31, Theorem 10.2]. This is an interesting result, as it shows that linear codes are capacity-achieving for discrete memoryless channels without the need for alphabet extension [39, Page 208], and without the PAS requirement, namely that the channel input alphabet can be partitioned into two halfs that are selected equally likely by the capacity-achieving distribution. An interesting problem for future work is to generalize the result of [56] to discrete-time, continuous-value output channels.

Another interesting future work is the study of finite length error exponents for layered PS, accounting for the distribution spectrum of the sequences in the shaping set.

Another interesting problem is practical encoding for layered PS, for making the practical layered PS proposed in [21] feasible for longer codeword lengths.

4

Calculating Practical Achievable Rates

In this section, we combine our results on DM rates in Section 2 with the achievable SE for memoryless channels that we stated in Theorem 3.3 and Corollary 3.4 in Section 3. Our aim is to calculate achievable rates for practical configurations. Throughout this section, we will call the SE "rate" and use the letter R, so that (3.30) becomes

$$R^*(R_{\mathrm{ss}}, q) = [mR_{\mathrm{ss}} - \mathsf{X}^*(q, X, Y)]^+ . \tag{4.1}$$

This expression provides a framework for analyzing achievable rates for shaping set rates R_{ss} and decoding metrics q. We note that (4.1) splits $R^*(R_{\mathrm{ss}}, q)$ into two term, which we can analyze independently. For instance, in case we do not want to account for suboptimal shaping set rates, we can assume an optimal shaping set rate $R_{\mathrm{ss}}m = \mathbb{H}(X)$, and (4.1) becomes

$$R^*(q) = [\mathbb{H}(X) - \mathsf{X}^*(q, X, Y)]^+ . \tag{4.2}$$

Similarily, in case we do not shape at all and use uniformly distributed input X, (4.2) becomes

$$R^*(q) = [\log_2 |\mathcal{X}| - \mathsf{X}^*(q, X, Y)]^+ . \tag{4.3}$$

Table 4.1: Important decoding metrics.

		q	$\mathsf{X}^*(q,X,Y)$	Section
symbolwise	SMD	$P_{X\|Y}$	$\mathbb{H}(X\|Y)$	4.4.1
bitwise	BMD	$\prod_{i=1}^{m} P_{B_i\|Y}$	$\sum_{i=1}^{m} \mathbb{H}(B_i\|Y)$	4.4.2
interleaver-agnostic	IACM	(4.51)–(4.52)	$m_s\, \mathbb{H}(X\|Y)$	4.4.3
bit-interleaver-agnostic	BIACM	(4.51)–(4.52)	$m_s\, \mathbb{H}(B\|Y)$	4.4.3
symbolwise hard-decision		$\exp[\mathbb{1}(x,\omega(y))]$	$\mathbb{H}_2(\epsilon) + \epsilon \log_2(\|\mathcal{X}\|-1)$	4.5.3
bitwise hard-decision		$\exp[\sum_{j=1}^{m} \mathbb{1}(x_j,\omega_j(y))]$	$m\, \mathbb{H}_2(\epsilon)$	4.5.4

This section's results on decoding metrics q and cross-equivocation $\mathsf{X}(q,X,Y)$ apply in the general R_{ss} case (4.1), the optimal R_{ss} case (4.2), and the uniform case (4.3). The remainder of this section is organized as follows.

1. In Section 4.1, we show how to explicitly calculate shaping set rates R_{ss} for PAS.

2. In Section 4.2, we review how the FEC rate and overhead from Section 3.3 relate to the cross-equivocation $\mathsf{X}(q,X,Y)$.

3. In Section 4.3, we review the concept of mismatched probabilistic models, which we will use in the following sections for analyzing the cross-equivocation $\mathsf{X}(q,X,Y)$ for mismatched decoding metrics q.

4. In Section 4.4, we consider metric design, where we restrict the permitted metrics to a certain class $\mathcal{Q}$, and then ask the question which metric q within that class is optimal. Our task is to solve

$$q^* = \arg\min_{q\in\mathcal{Q}} \mathsf{X}^*(q,X,Y) \tag{4.4}$$

$$= \arg\min_{q\in\mathcal{Q}} \left[\min_{s>0} \mathsf{X}(q^s,X,Y)\right]. \tag{4.5}$$

   All considered metric classes are closed under $(\cdot)^s$, i.e., $q \in \mathcal{Q}$ implies $q^s \in \mathcal{Q}$ and the optimization problem simplifies to

$$q^* = \arg\min_{q\in\mathcal{Q}} \mathsf{X}(q,X,Y). \tag{4.6}$$

5. In Section 4.5, we consider metric assessment, i.e., for a fixed and given metric q, we calculate the maximum achievable rate for that metric. Our task is to calculate $\mathsf{X}^*(q,X,Y)$.

The results of Section 4.4 and Section 4.5 are summarized in Table 4.1.

4.1 PAS Shaping Set Rate and Overhead

The shaping layer contributes to the achievable rate (4.1) via the shaping set rate R_{ss}. In practical designs, rates are often expressed as overheads. If needed, we can translate the shaping set rates discussed in the following into shaping overheads via Table 4.3. In this section, we consider PAS, in which case we can calculate the shaping set rate R_{ss} explicitly.

Remark 4.1. In Section 3, we derived the achievable rate (4.1) for layered PS, which captures the essence of PAS, but is not identical to PAS. We assume for now that (4.1) also holds for PAS. In Section 5, we revisit this assumption, discuss the differences in the theoretical analysis of layered PS and PAS, and we prove in Theorem 5.3 the achievability of (4.1) by PAS for the special case when the DM is a CCDM. We leave the theoretical analysis of PAS with a general DM (including the important case of an MCDM) for future work.

We consider a channel with input alphabet $\mathcal{X}$ and start from the general definition of R_{ss} that we gave in Section 3.2, i.e.,

$$R_{\text{ss}} = \frac{\log_2 |\mathcal{S}|}{nm} \tag{4.7}$$

where

- $\mathcal{S} \subseteq \mathcal{X}^n$ is the shaping set of permitted sequences,
- $m = \log_2 |\mathcal{X}|$ is the alphabet size in bits,
- nm is the codeword length in bits.

For PAS, the channel input alphabet decomposes into $\mathcal{X} = \mathcal{A} \times \mathcal{U}$ and the shaping set $\mathcal{S}$ decomposes into $\mathcal{S}_A \times \mathcal{U}^n$, where $\mathcal{U}^n$ is unconstrained and used for the parity bits generated by the systematic FEC encoder, and where $\mathcal{S}_A \subseteq \mathcal{A}^n$ contains the sequences output by the DM. The shaping set rate now becomes

$$mR_{\text{ss}} = \frac{\log_2 |\mathcal{S}_A||\mathcal{U}|^n}{n} = \frac{\log_2 |\mathcal{S}_A|}{n} + \log_2 |\mathcal{U}|. \tag{4.8}$$

In practical systems, the DM output length n_{dm} is often shorter than the FEC codeword length n. We therefore assume that $n = \ell \cdot n_{\mathrm{dm}}$, i.e., the FEC codeword length is a multiple of the DM output length. The shaping set rate for PAS now becomes

$$mR_{\mathrm{ss}} = \frac{\log_2 |\mathcal{S}_A|^{\ell}}{\ell n_{\mathrm{dm}}} + \log_2 |\mathcal{U}| \tag{4.9}$$

$$= \frac{\log_2 |\mathcal{S}_A|}{n_{\mathrm{dm}}} + \log_2 |\mathcal{U}| \tag{4.10}$$

$$= \frac{k_{\mathrm{dm}}}{n_{\mathrm{dm}}} + \log_2 |\mathcal{U}|. \tag{4.11}$$

To account for a specific DM in our achievable rate calculation, we proceed as follows:

1. For DM input length k_{dm} and output length n_{dm}, calculate the shaping set rate R_{ss} via (4.11).
2. Calculate the average distribution $P_{\bar{A}}$ of the DM, and the resulting channel input distribution
$$P_X = P_{\bar{A}} P_S, \quad P_S \text{ uniform on } \mathcal{U}. \tag{4.12}$$
3. Calculate the cross-equivocation term $\mathsf{X}^*(q, X, Y)$ for $X \sim P_X$.
4. The achievable PAS rate is
$$R^*_{\mathrm{pas}} = \left[\frac{k_{\mathrm{dm}}}{n_{\mathrm{dm}}} + \log_2 |\mathcal{U}| - \mathsf{X}^*(q, X, Y) \right]^+. \tag{4.13}$$

The shaping set rate calculation for PAS is summarized in Table 4.2. We next provide examples for the calculation of (4.11) and $P_{\bar{A}}$, using the techniques we developed in Section 2.

4.1.1 Average Power Constrained 8-ASK PAS

Let's revisit the 8-ASK running example from Section 2. The input alphabet is

$$\mathcal{X} = \{\pm 1, \pm 3, \pm 5, \pm 7\} \tag{4.14}$$

Table 4.2: Shaping set rates for PAS.

$mR_{\mathrm{ss}} = \frac{\log_2 \lvert\mathcal{S}_A\rvert}{n} + \log_2 \lvert\mathcal{U}\rvert$ $P_X(as) = P_{\bar{A}}(a) \cdot \frac{1}{\lvert\mathcal{U}\rvert}, \quad a \in \mathcal{A}, s \in \mathcal{U}, \mathcal{X} = \mathcal{A} \times \mathcal{U}$		
	$\frac{\log_2 \lvert\mathcal{S}_A\rvert}{n}$	$P_{\bar{A}}$
MB source	$\mathbb{H}(A)$	P_A
MCDM	$\frac{k_{\mathrm{mcdm}}}{n_{\mathrm{mcdm}}}$	$P_{\bar{A}}(k_{\mathrm{mcdm}}, n_{\mathrm{mcdm}})$ by Sec. 2.4.1
CCDM	$\frac{k_{\mathrm{ccdm}}}{n_{\mathrm{ccdm}}}$	$P_{\bar{A}}(k_{\mathrm{ccdm}}, n_{\mathrm{ccdm}})$ by Sec. 2.5.1

and as symbol cost, we consider the power $w(x) = x^2$. The input alphabet decomposes into amplitude and sign via

$$\mathcal{A} = \{1, 3, 6, 7\} \tag{4.15}$$

$$\mathcal{U} = \{-1, 1\} \tag{4.16}$$

$$\mathcal{X} = \mathcal{A} \otimes \mathcal{U} \tag{4.17}$$

where $\otimes$ denotes the Kronecker product. The task of the DM is to minimize the average cost, and as the average cost is invariant under the choice of the sign, the sign set is the unconstrained set $\mathcal{U}$. Suppose now input length k_{dm} and output length n_{dm} are fixed and given, for instance, because the DM rate $k_{\mathrm{dm}}/n_{\mathrm{dm}}$ is required for achieving a specific SE, and because the output length is restricted to n_{dm} for complexity reasons or because of latency constraints. Thus, independent of the employed DM, the shaping set rate is

$$mR_{\mathrm{ss}} = \frac{k_{\mathrm{dm}}}{n_{\mathrm{dm}}} + \log_2 |\mathcal{U}| = \frac{k_{\mathrm{dm}}}{n_{\mathrm{dm}}} + 1, \qquad m = \log_2 |\mathcal{X}| = 3. \tag{4.18}$$

The resulting average amplitude distribution $P_{\bar{A}}$ *does* depend on the employed DM. The DM selects $2^{k_{\mathrm{dm}}}$ amplitude sequences from $\mathcal{A}^{n_{\mathrm{dm}}}$, with the aim to minimize the average cost $\mathbb{E}[w(\bar{A})]$. The resulting average amplitude distribution $P_{\bar{A}}$ can now be approximated by Monte Carlo simulation or calculated explicitly, depending on the employed DM. For instance, for MCDMs, we can calculate $P_{\bar{A}}$ by Section 2.4.1, and for CCDMs by Section 2.5.1.

4.1.2 Target Distribution PAS

Suppose now we have a desired distribution P_X on a channel input alphabet $\mathcal{X}$ that factorizes into $P_A P_S$, where P_A is some distribution on $\mathcal{A}$, P_S is uniform on $\mathcal{U}$, and $\mathcal{X} = \mathcal{A} \times \mathcal{U}$. Our task is to approximate P_X sufficiently well using PAS.

One possible approach is to choose a DM who's output distribution approximates P_A best possible, e.g., one may use a CCDM whose type is the result of quantizing P_A, following Section 2.5.2. In practice, this approach has a downside, as we loose control of the DM rate: While it approaches the entropy of $P_{\bar{A}}$ for large output length n_{dm}, the DM rate can become arbitrarily small for small output lengths. Also, it is not clear why to approximate P_X, we should approximate P_A but not P_S.

For practical system design, we therefore advocate to follow [69] and translate the desired input distribution P_X into a cost via the self-information

$$w(x) = -\log_2 P_X(x), \qquad x \in \mathcal{X}. \tag{4.19}$$

Writing $x = as$, we have

$$\begin{aligned} w(x) = w(as) &= -\log_2[P_A(a)P_S(s)] && (4.20)\\ &= -\log_2 P_A(a) - \log_2 P_S(s) && (4.21)\\ &= -\log_2 P_A(a) + \log_2 |\mathcal{U}|. && (4.22) \end{aligned}$$

We observe that the cost of $x = as$ depends on a but not on s, so under the self-information cost, $\mathcal{U}$ is the unconstrained set. Consequently, to approximate $P_X = P_A P_S$ with P_S uniform on $\mathcal{U}$, we use PAS with a DM that selects $2^{k_{\mathrm{dm}}}$ sequences from $\mathcal{A}^{n_{\mathrm{dm}}}$. We choose k_{dm} and n_{dm} according to the required DM rate, complexity, and latency, and the DM selects the sequences with the aim to minimize the average cost

$$\mathbb{E}[w(\bar{A})] = \mathbb{E}[-\log_2 P_A(\bar{A})] = \mathbb{X}(P_{\bar{A}} \| P_A). \tag{4.23}$$

By using the self-information (4.19) as cost, we can again calculate the average distribution $P_{\bar{A}}$ explicitly for MCDMs by Section 2.4.1 and for CCDMs by Section 2.5.1.

Table 4.3: Achievable rates and overheads.

FEC Rate	
Rate	$R_{\text{fec}} \leq 1 - \frac{1}{m}\mathsf{X}^*(q, X, Y)$
Redundancy	$1 - R_{\text{fec}} \geq \mathsf{X}^*(q, X, Y)$
Overhead in %	$100 \cdot \frac{1-R_{\text{fec}}}{R_{\text{fec}}} \geq 100 \cdot \frac{\mathsf{X}^*(q,X,Y)}{m-\mathsf{X}^*(q,X,Y)}$
Shaping Set Rate	
Rate	$R_{\text{ss}} = \frac{\log_2 \lvert\mathcal{S}\rvert}{nm}$
Redundancy	$1 - R_{\text{ss}}$
Overhead in %	$100 \cdot \left(\frac{1}{R_{\text{ss}}} - 1\right)$
Total overhead in %	$100 \cdot \left(\frac{1}{R_{\text{ss}}+R_{\text{fec}}-1} - 1\right)$

4.2 Achievable FEC Rate and Overhead

The FEC layer contributes via the cross-equivocation $\mathsf{X}^*(q, X, Y)$ to the achievable rate (4.1). The normalized cross-equivocation $\mathsf{X}^*(q, X, Y)/m$ directly provides a benchmark for the FEC rate and overhead of practical systems, which we summarize in Table 4.3. This benchmark accounts for the employed decoding metric q and the employed channel input distribution P_X, however, as we discussed in Section 3, it considers asymptotically long codewords and is calculated for non-constructive random code ensembles. Consequently, the practical FEC rate R_{fec} at which the employed FEC code achieves the desired reliability is usually smaller than the benchmark. The gap

$$R^*_{\text{fec}} - R_{\text{fec}} = 1 - \frac{1}{m}\mathsf{X}^*(q, X, Y) - R_{\text{fec}} \tag{4.24}$$

between the achievable FEC rate R^*_{fec} and the actual FEC rate R_{fec} is a good measure to quantify the quality of a practical FEC code.

4.3 Mismatch

Let X be a random variable with distribution P_X on the finite alphabet $\mathcal{X}$ and let Q_X be some other distribution on $\mathcal{X}$, possibly different from P_X. By the information inequality (A.24), we know that

$$\mathbb{X}(P_X \| Q_X) = \mathbb{E}[-\log_2 Q_X(X)] \tag{4.25}$$
$$\geq \mathbb{E}[-\log_2 P_X(X)] = \mathbb{H}(X) \tag{4.26}$$

with equality if and only if $Q_X = P_X$. If Q_X models P_X and $Q_X \neq P_X$, we say that Q_X is mismatched. The information inequality provides us an objective function for matching our model Q_X to P_X, as

$$\arg\min_{Q_X} \mathbb{E}[-\log_2 Q_X(X)] = P_X. \tag{4.27}$$

In the following, we use this property to identify optimal metrics.

4.4 Decoding Metric Design

4.4.1 Symbol-Metric Decoding

Suppose we have no restriction on the decoding metric q. To maximize the achievable rate, we need to minimize the cross-equivocation term in (4.2). We have

$$\mathbb{E}\left[-\log_2 \frac{q(X,Y)}{\sum_{a\in\mathcal{X}} q(a,Y)}\right] = \mathbb{E}\left[\mathbb{E}\left[-\log_2 \frac{q(X,Y)}{\sum_{a\in\mathcal{X}} q(a,Y)}\middle| Y\right]\right] \tag{4.28}$$
$$\overset{(4.26)}{\geq} \mathbb{E}\left[\mathbb{E}\left[-\log_2 P_{X|Y}(X|Y)\middle| Y\right]\right] \tag{4.29}$$
$$= \mathbb{H}(X|Y) \tag{4.30}$$

with equality if we use the posterior probability distribution as metric, i.e.,

$$q(a,b) = P_{X|Y}(a|b), \quad a \in \mathcal{X}, b \in \mathcal{Y}. \tag{4.31}$$

Note that this choice of q is not unique, in particular,

$$q(a,b) = P_{X|Y}(a|b)P_Y(b) \tag{4.32}$$

is also optimal, since the factor $P_Y(b)$ cancels out. For the optimal metric, the achievable rate is

$$R^{\mathsf{opt}} = [\mathbb{H}(X) - \mathbb{H}(X|Y)]^+ = \mathbb{I}(X;Y) \tag{4.33}$$

where we dropped the $(\cdot)^+$ operator because by the information inequality, mutual information is non-negative.

4.4.2 Bit-Metric Decoding

Suppose the channel input is a binary vector $\boldsymbol{B} = B_1 \cdots B_m$ and the receiver uses a bit-metric, i.e.,

$$q(\boldsymbol{a}, y) = \prod_{j=1}^{m} q_j(a_j, y). \tag{4.34}$$

In this case, we have for the cross-equivocation term in (4.2)

$$\begin{aligned}
&\mathbb{E}\left[-\log_2 \frac{q(\boldsymbol{B}, Y)}{\sum_{\boldsymbol{a}\in\{0,1\}^m} q(\boldsymbol{a}, Y)}\right] \\
&\qquad = \mathbb{E}\left[-\log_2 \frac{\prod_{j=1}^{m} q_j(B_j, Y)}{\sum_{\boldsymbol{a}\in\{0,1\}^m} \prod_{j=1}^{m} q_j(a_j, Y)}\right] && (4.35)\\
&\qquad = \mathbb{E}\left[-\log_2 \frac{\prod_{j=1}^{m} q_j(B_j, Y)}{\prod_{j=1}^{m} \sum_{a\in\{0,1\}} q_j(a, Y)}\right] && (4.36)\\
&\qquad = \mathbb{E}\left[-\sum_{j=1}^{m} \log_2 \frac{q_j(B_j, Y)}{\sum_{a\in\{0,1\}} q_j(a, Y)}\right] && (4.37)\\
&\qquad = \sum_{j=1}^{m} \mathbb{E}\left[-\log_2 \frac{q_j(B_j, Y)}{\sum_{a\in\{0,1\}} q_j(a, Y)}\right] && (4.38)
\end{aligned}$$

where equality in (4.36) follows by (A.9). For each $j = 1, \ldots, m$, we now have

$$\begin{aligned}
&\mathbb{E}\left[-\log_2 \frac{q_j(B_j, Y)}{\sum_{a\in\{0,1\}} q_j(a, Y)}\right] \\
&\qquad = \mathbb{E}\left[\mathbb{E}\left[-\log_2 \frac{q_j(B_j, Y)}{\sum_{a\in\{0,1\}} q_j(a, Y)}\,\middle|\, Y\right]\right] && (4.39)\\
&\qquad \geq \mathbb{H}(B_j|Y) && (4.40)
\end{aligned}$$

with equality if

$$q_j(a, b) = P_{B_j|Y}(a|b), \quad a \in \{0, 1\}, b \in \mathcal{Y}. \tag{4.41}$$

The achievable rate becomes the bit-metric decoding (BMD) rate

$$R^{\mathsf{bmd}} = \Big[\mathbb{H}(\boldsymbol{B}) - \sum_{j=1}^{m} \mathbb{H}(B_j|Y)\Big]^+. \tag{4.42}$$

By defining the L-value

$$L_i = \log \frac{P_{B_i|Y}(0|Y)}{P_{B_i|Y}(1|Y)}$$

the equivocation sum can also be written as

$$\sum_{i=1}^{m} \mathbb{H}(B_i|Y) = \sum_{i=1}^{m} \mathbb{E}\left[\log_2\{1+\exp\left[-(1-2B_i)L_i\right]\}\right]. \tag{4.43}$$

For independent bit levels $B_1, B_2, \ldots, B_m$, the BMD rate can be also written in the form [55]

$$R^{\mathsf{bmd,ind}} = \sum_{j=1}^{m} \mathbb{I}(B_j;Y). \tag{4.44}$$

4.4.3 Interleaver-Agnostic Coded Modulation

Suppose we have a vector channel with input $\boldsymbol{X} = X_1 \cdots X_{m_s}$ with distribution $P_{\boldsymbol{X}}$ on the input alphabet $\mathcal{X}^{m_s}$ and output $\boldsymbol{Y} = Y_1 \cdots Y_{m_s}$ with distributions $P_{\boldsymbol{Y}|\boldsymbol{X}}(\cdot|\boldsymbol{a})$, $\boldsymbol{a} \in \mathcal{X}^{m_s}$, on the output alphabet $\mathcal{Y}^{m_s}$. We consider the following situation:

- The Y_i are potentially correlated, in particular, we may have $Y_1 = Y_2 = \cdots = Y_{m_s}$.
- Despite the potential correlation, the receiver uses a memoryless metric q defined on $\mathcal{X} \times \mathcal{Y}$, i.e., a vector input $\boldsymbol{x}$ and a vector output $\boldsymbol{y}$ are scored by

$$q^{m_s}(\boldsymbol{x},\boldsymbol{y}) = \prod_{i=1}^{m_s} q(x_i,y_i). \tag{4.45}$$

  The reason for this decoding strategy may be an interleaver between encoder output and channel input that is reverted at the receiver but not known to the decoder. We therefore call this scenario interleaver-agnostic coded modulation (IACM).

Using the same approach as for BMD, we have

$$\frac{1}{m_s}\mathbb{E}\left[-\log_2\frac{\prod_{i=1}^{m_s}q(X_i,Y_i)}{\sum_{\boldsymbol{a}\in\mathcal{X}^{m_s}}\prod_{i=1}^{m_s}q(a_i,Y_i)}\right]$$

$$=\frac{1}{m_s}\mathbb{E}\left[-\log_2\frac{\prod_{i=1}^{m_s}q(X_i,Y_i)}{\prod_{i=1}^{m_s}\sum_{a\in\mathcal{X}}q(a,Y_i)}\right] \tag{4.46}$$

$$=\frac{1}{m_s}\sum_{i=1}^{m_s}\mathbb{E}\left[-\log_2\frac{q(X_i,Y_i)}{\sum_{a\in\mathcal{X}}q(a,Y_i)}\right] \tag{4.47}$$

where equality in (4.46) follows by (A.9). Expression (4.47) is not very insightful. We could optimize q for, say, the ith term, which would be

$$q(a,b)=P_{X_i|Y_i}(a|b),\quad a\in\mathcal{X},b\in\mathcal{Y} \tag{4.48}$$

but this would not be optimal for the other terms. We therefore choose a different approach. Let I be a random variable uniformly distributed on $\mathcal{I}=\{1,2,\ldots,m_s\}$ and define $X=X_I$, $Y=Y_I$. Then, we have

$$\frac{1}{m_s}\sum_{i=1}^{m_s}\mathbb{E}\left[-\log_2\frac{q(X_i,Y_i)}{\sum_{a\in\mathcal{X}}q(a,Y_i)}\right]=\mathbb{E}\left[-\log_2\frac{q(X_I,Y_I)}{\sum_{a\in\mathcal{X}}q(a,Y_I)}\right] \tag{4.49}$$

$$=\mathbb{E}\left[-\log_2\frac{q(X,Y)}{\sum_{a\in\mathcal{X}}q(a,Y)}\right]. \tag{4.50}$$

Thus, the optimal metric for interleaving is

$$q(a,b)=P_{X|Y}(a|b) \tag{4.51}$$

which can be calculated from

$$P_X(a)p_{Y|X}(b|a)=\sum_{j=1}^{m_s}\frac{1}{m_s}P_{X_j}(a)p_{Y_j|X_j}(b|a). \tag{4.52}$$

The achievable rate becomes

$$R^{\mathsf{iacm}}=\left[\mathbb{H}(\boldsymbol{X})-m_s\,\mathbb{H}(X|Y)\right]^+. \tag{4.53}$$

When the input is a binary vector $\boldsymbol{B}=B_1\cdots B_m$, we get the bit-interleaver-agnostic coded modulation (BIACM) rate

$$R^{\mathsf{biacm}}=\left[\mathbb{H}(\boldsymbol{B})-m\,\mathbb{H}(B|Y)\right]^+. \tag{4.54}$$

4.5 Decoding Metric Assessment

Suppose a decoder is constrained to use a specific metric q. In this case, our task is to assess the metric performance by calculating a rate that can be achieved. If q is a non-negative metric, an achievable rate is given by (4.2), which evaluates the cross-equivocation in q^s and minimizes over s. Elaborating on Section 3.4.2, the reason for this is as follows: suppose we have another metric $\tilde{q}$ that scores the codewords in the same order as metric q, i.e., we have

$$\tilde{q}(a_1, b) > \tilde{q}(a_2, b) \Leftrightarrow q(a_1, b) > q(a_2, b), \quad a_1, a_2 \in \mathcal{X}, b \in \mathcal{Y}. \tag{4.55}$$

Then, $R(\tilde{q})$ is also achievable by q. An example for an order preserving transformation is $\tilde{q}(a, b) = e^{q(a,b)}$. For a non-negative metric q, another order preserving transformation is $\tilde{q}(a, b) = q(a, b)^s$ for $s > 0$. We may now find a better achievable rate for metric q by calculating for instance

$$\max_{s>0} R(q^s). \tag{4.56}$$

In the following, we will say that two metrics q and $\tilde{q}$ are equivalent if and only if the order-preserving condition (4.55) is fulfilled.

Example 4.1 (AWGN Channel with BPSK)**.** Consider a BPSK constellation $\mathcal{X} = \{-1, 1\}$ and uniformly distributed input, i.e., $P_X(-1) = P_X(1) = \frac{1}{2}$. The channel output is

$$Y = |h| \cdot X + Z$$

where $|h|$ is a constant positive real number and where Z is zero mean Gaussian with variance σ^2. At the receiver, the decoder uses the metric

$$q(b, a) = b \cdot a, \quad b \in \mathbf{R}, a \in \mathcal{X}.$$

Note that the decoder does not make use of the channel parameters $|h|, \sigma^2$. We transform the metric into the equivalent non-negative metric $e^{sq(b,a)} =: \tilde{q}(b, a)$ with $s > 0$ and calculate the cross-equivocation term for $\tilde{q}$. We have

$$\mathbb{E}\left[-\log_2 \frac{\tilde{q}(Y, X)}{\sum_{a\in\{-1,1\}} \tilde{q}(Y, a)}\right] = \mathbb{E}\left[\mathbb{E}\left[-\log_2 \frac{\tilde{q}(Y, X)}{\sum_{a\in\{-1,1\}} \tilde{q}(Y, a)}\middle| Y\right]\right]$$

For $Y = b$, the inner expectation is

$$\mathbb{E}\left[-\log_2 \frac{\tilde{q}(Y,X)}{\sum_{a\in\{-1,1\}} \tilde{q}(Y,a)} \middle| Y = b \right]$$

and it is calculated for $a' \in \{-1, 1\}$ according to

$$\begin{aligned} P_{X|Y}(a'|b) &= \frac{P_X(a')p_{Y|X}(b|a')}{p_Y(b)} \\ &= \frac{\frac{1}{2}p_Z(b-|h|a')}{\frac{1}{2}p_Z(b+|h|)+\frac{1}{2}p_Z(b-|h|)} \\ &= \frac{p_Z(b-|h|a')}{p_Z(b+|h|)+p_Z(b-|h|)} \\ &= \frac{e^{-\frac{(b-|h|a')^2}{2\sigma^2}}}{e^{-\frac{(b+|h|)^2}{2\sigma^2}}+e^{-\frac{(b-|h|)^2}{2\sigma^2}}} \\ &= \frac{e^{\frac{|h|a'b}{\sigma^2}}}{e^{-\frac{|h|b}{\sigma^2}}+e^{\frac{|h|b}{\sigma^2}}}. \end{aligned}$$

Note that for $s = |h|/\sigma^2$, we also have

$$\frac{\tilde{q}(b,a')}{\sum_{a\in\{-1,1\}} \tilde{q}(b,a)} = \frac{e^{\frac{|h|a'b}{\sigma^2}}}{e^{-\frac{|h|b}{\sigma^2}}+e^{\frac{|h|b}{\sigma^2}}}.$$

Thus, we have

$$\begin{aligned} \mathbb{E}\left[-\log_2 \frac{\tilde{q}(Y,X)}{\sum_{a\in\{-1,1\}} \tilde{q}(Y,a)} \middle| Y = b \right] &\overset{(a)}{\geq} \mathbb{E}\left[-\log_2 P_{X|Y}(X|b) \middle| Y = b \right] \\ &= \mathbb{H}(X|Y=b) \end{aligned}$$

with equality in (a) if $s = |h|/\sigma^2$. Thus, for this choice of s, the cross-equivocation term is equal to the equivocation $\mathbb{H}(X|Y)$ and the achievable rate is $\mathbb{I}(X;Y)$. Observations:

- Although the original metric $q(b,a) = a \cdot b$ does not take the channel parameters $|h|, \sigma^2$ into account, it is optimal because it achieves the mutual information $\mathbb{I}(X;Y)$.
- The optimal value $s^* = |h|/\sigma^2$ recovers the channel parameters.

4.5.1 Generalized Mutual Information

Suppose the input distribution is uniform, i.e., $P_X(a) = 1/|\mathcal{X}|, a \in \mathcal{X}$. In this case, we have

$$\max_{s>0} R(q^s) = \max_{s>0} \left[\mathbb{H}(X) - \mathbb{E}\left[-\log_2 \frac{q(X,Y)^s}{\sum_{a\in\mathcal{X}} q(a,Y)^s} \right] \right]^+ \tag{4.57}$$

$$= \max_{s>0} \left[\mathbb{E}\left[\log_2 \frac{q(X,Y)^s \frac{1}{P_X(X)}}{\sum_{a\in\mathcal{X}} q(a,Y)^s} \right] \right]^+ \tag{4.58}$$

$$= \max_{s>0} \mathbb{E}\left[\log_2 \frac{q(X,Y)^s}{\sum_{a\in\mathcal{X}} P_X(a) q(a,Y)^s} \right] \tag{4.59}$$

where we could move $P_X(a)$ under the sum, because P_X is by assumption uniform, and where we could drop the $[\cdot]^+$ operator because for $s = 0$, the expectation is zero. The expression in (4.59) is called generalized mutual information (GMI) in [49] and was shown to be an achievable rate for the classical transceiver. This is as expected, as for uniform input, layered PS is equivalent to the classical transceiver, see Section 3.1. For non-uniform input, the GMI and (4.58) may differ, i.e., we may not have equality in (4.59).

Discussion

Suppose for a non-uniform input distribution P_X and a metric q, the GMI evaluates to R, implying that a classical transceiver can achieve R. Can layered PS also achieve R, possibly by using a different metric? The answer is yes. Define

$$\tilde{q}(a,b) = q(a,b) P_X(a)^{\frac{1}{s}}, \quad a \in \mathcal{X}, b \in \mathcal{Y} \tag{4.60}$$

where s is the optimal value maximizing the GMI. We calculate a PS achievable rate for $\tilde{q}$ by analyzing the equivalent metric $\tilde{q}^s$. We have

$$R(\tilde{q}^s) = \left[\mathbb{H}(X) - \mathbb{E}\left[-\log_2 \frac{\tilde{q}^s(X,Y)}{\sum_{a\in\mathcal{X}} \tilde{q}^s(a,Y)}\right]\right]^+ \tag{4.61}$$

$$= \left[\mathbb{E}\left[\log_2 \frac{\tilde{q}^s(X,Y)\frac{1}{P_X(X)}}{\sum_{a\in\mathcal{X}} \tilde{q}^s(a,Y)}\right]\right]^+ \tag{4.62}$$

$$= \left[\mathbb{E}\left[\log_2 \frac{q^s(X,Y)}{\sum_{a\in\mathcal{X}} P_X(a) q^s(a,Y)}\right]\right]^+ \tag{4.63}$$

$$= R \tag{4.64}$$

which shows that R can also be achieved by layered PS. It is important to stress that this requires a change of the metric: for example, suppose q is the Hamming metric of a hard-decision decoder (see Section 4.5.3). In general, this does *not* imply that $\tilde{q}$ defined by (4.60) is also a Hamming metric.

4.5.2 LM-Rate

For the classical transceiver of Section 3.1.1, the work [40] shows that the so-called LM-Rate defined as

$$R_{\mathrm{LM}}(s,r) = \left[\mathbb{E}\left[\log_2 \frac{q(X,Y)^s r(X)}{\sum_{a\in\operatorname{supp} P_X} P_X(a) q(a,Y)^s r(a)}\right]\right]^+ \tag{4.65}$$

is achievable, where $s > 0$ and where r is a function on $\mathcal{X}$. By choosing $s = 1$ and $r(a) = 1/P_X(a)$, we have

$$R_{\mathrm{LM}}(1, 1/P_X) = \left[\mathbb{E}\left[\log_2 \frac{q(X,Y)\frac{1}{P_X(X)}}{\sum_{a\in\operatorname{supp} P_X} q(a,Y)}\right]\right]^+ \tag{4.66}$$

$$\geq \left[\mathbb{E}\left[\log_2 \frac{q(X,Y)\frac{1}{P_X(X)}}{\sum_{a\in\mathcal{X}} q(a,Y)}\right]\right]^+ \tag{4.67}$$

$$= \left[\mathbb{H}(X) - \mathbb{E}\left[-\log_2 \frac{q(X,Y)}{\sum_{a\in\mathcal{X}} q(a,Y)}\right]\right]^+ \tag{4.68}$$

$$= R \tag{4.69}$$

with equality in (4.67) if $\operatorname{supp} P_X = \mathcal{X}$. Thus, formally, our achievable rate can be recovered from the LM-Rate. We emphasize that [40] shows the achievability of the LM-Rate for the classical transceiver of Section 3.1.1, and consequently, R_{LM} and R have different operational meanings, corresponding to achievable rates of two different transceiver setups, with different random coding experiments, and different encoding and decoding strategies.

4.5.3 Hard-Decision Decoding

Hard-decision decoding consists of two steps. First, the channel output alphabet is partitioned into disjoint decision regions

$$\mathcal{Y} = \bigcup_{a\in\mathcal{X}} \mathcal{Y}_a, \quad \mathcal{Y}_a \cap \mathcal{Y}_b = \emptyset \text{ if } a \neq b \tag{4.70}$$

and a quantizer ω maps the channel output to the channel input alphabet according to the decision regions, i.e.,

$$\omega\colon \mathcal{Y} \to \mathcal{X}, \quad \omega(b) = a \Leftrightarrow b \in \mathcal{Y}_a. \tag{4.71}$$

Second, the receiver uses the Hamming metric on $\mathcal{X}$ for decoding, i.e., we have

$$q(a,\omega(y)) = \mathbb{1}(a,\omega(y)) = \begin{cases} 1, & \text{if } a = \omega(y) \\ 0, & \text{otherwise.} \end{cases} \tag{4.72}$$

We next derive an achievable rate by analyzing the equivalent metric $e^{s\mathbb{1}(\cdot,\cdot)}$, $s > 0$. For the cross-equivocation term, we have

$$\mathbb{E}\left[-\log_2 \frac{e^{s\mathbb{1}[X,\omega(Y)]}}{\sum_{a\in\mathcal{X}} e^{s\mathbb{1}[a,\omega(Y)]}}\right] = \mathbb{E}\left[-\log_2 \frac{e^{s\mathbb{1}[X,\omega(Y)]}}{|\mathcal{X}|-1+e^s}\right] \tag{4.73}$$

$$= -\Pr[X=\omega(Y)]\log_2 \frac{e^s}{|\mathcal{X}|-1+e^s} - \Pr[X\neq\omega(Y)]\log_2 \frac{1}{|\mathcal{X}|-1+e^s} \tag{4.74}$$

$$= -(1-\epsilon)\log_2 \frac{e^s}{|\mathcal{X}|-1+e^s} - \epsilon \log_2 \frac{1}{|\mathcal{X}|-1+e^s} \tag{4.75}$$

$$= -(1-\epsilon)\log_2 \frac{e^s}{|\mathcal{X}|-1+e^s} - \sum_{\ell=1}^{|\mathcal{X}|-1} \frac{\epsilon}{|\mathcal{X}|-1} \log_2 \frac{1}{|\mathcal{X}|-1+e^s} \tag{4.76}$$

where we defined $\epsilon = \Pr(X \neq \omega(Y))$. By (4.26), the last line is minimized by choosing

$$s\colon 1-\epsilon = \frac{e^s}{|\mathcal{X}|-1+e^s} \text{ and } \frac{\epsilon}{|\mathcal{X}|-1} = \frac{1}{|\mathcal{X}|-1+e^s} \tag{4.77}$$

which is achieved by

$$e^s = \frac{(|X|-1)(1-\epsilon)}{\epsilon}. \tag{4.78}$$

With this choice for s, we have

$$\begin{aligned} &-(1-\epsilon)\log_2(1-\epsilon) - \sum_{\ell=1}^{|\mathcal{X}|-1} \frac{\epsilon}{|\mathcal{X}|-1} \log_2 \frac{\epsilon}{|\mathcal{X}|-1} \\ &\quad = \underbrace{-(1-\epsilon)\log_2(1-\epsilon) - \epsilon\log_2\epsilon}_{=:\mathbb{H}_2(\epsilon)} + \epsilon\log_2(|\mathcal{X}|-1) \qquad (4.79) \\ &\quad = \mathbb{H}_2(\epsilon) + \epsilon\log_2(|\mathcal{X}|-1) \qquad (4.80) \end{aligned}$$

where $\mathbb{H}_2(\cdot)$ is the binary entropy function. The term (4.80) corresponds to the equivocation of a $|\mathcal{X}|$-ary symmetric channel with uniform input, see Figure 4.1 for an illustration. We conclude that by hard-decision decoding, we can achieve

$$R^{\mathsf{hd}} = [\mathbb{H}(X) - [\mathbb{H}_2(\epsilon) + \epsilon\log_2(|\mathcal{X}|-1)]]^+ \tag{4.81}$$

where

$$\begin{aligned} \epsilon &= 1 - \Pr[X = \omega(Y)] \qquad (4.82) \\ &= 1 - \sum_{a\in\mathcal{X}} P_X(a) \int_{\mathcal{Y}_a} p_{Y|X}(\tau|a)\,\mathrm{d}\tau. \qquad (4.83) \end{aligned}$$

Remark 4.2. Our derivations for hard-decision decoding can be related to rate distortion theory, see, e.g., [30, Chapter 10].

4.5.4 Binary Hard-Decision Decoding

Suppose the channel input is the binary vector $\boldsymbol{B} = B_1 \cdots B_m$ and the decoder uses m binary quantizers, i.e., we have

$$\mathcal{Y} = \mathcal{Y}_{0j} \cup \mathcal{Y}_{1j}, \quad \mathcal{Y}_{1j} = \mathcal{Y} \setminus \mathcal{Y}_{0j} \tag{4.84}$$

$$\omega_j\colon \mathcal{Y} \to \{0,1\}, \quad \omega_j(b) = a \Leftrightarrow b \in \mathcal{Y}_{aj}. \tag{4.85}$$

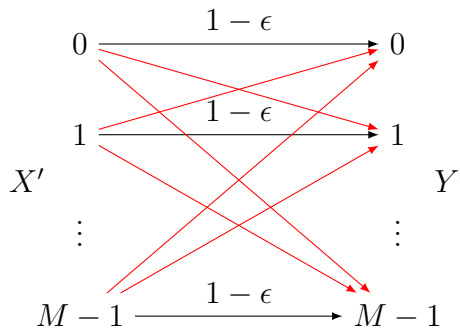


Figure 4.1: The M-ary symmetric channel. Each red transition has probability $\frac{\epsilon}{M-1}$. Note that for $M = 2$, the channel is the binary symmetric channel. For uniformly distributed input X', we have $\mathbb{H}(X'|Y') = \mathbb{H}_2(\epsilon) + \epsilon \log_2(M-1)$.

The receiver uses a binary Hamming metric, i.e., we have

$$q(a,b) = \mathbb{1}(a,b), \quad a,b \in \{0,1\} \tag{4.86}$$

$$q^m(\boldsymbol{a},\boldsymbol{b}) = \sum_{j=1}^{m} \mathbb{1}(a_j,b_j) \tag{4.87}$$

and we analyze the equivalent metric

$$e^{sq^m(\boldsymbol{a},\boldsymbol{b})} = \prod_{j=1}^{m} e^{s\mathbb{1}(a_j,b_j)}, \quad s > 0. \tag{4.88}$$

Since the decoder uses the same metric for each bit level $j = 1,2,\ldots,m$, binary hard-decision decoding is an instance of interleaved coded modulation, which we discussed in Section 4.4.3. Thus, defining the auxiliary random variable I uniformly distributed on $\{1,2,\ldots,m\}$ and

$$B = B_I, \quad \hat{B} = \omega_I(Y) \tag{4.89}$$

we can use the interleaved coded modulation result (4.50). We have for the normalized cross-equivocation term

$$\begin{aligned} &\frac{1}{m}\mathbb{E}\left[-\log_2 \frac{\prod_{j=1}^{m} e^{s\mathbb{1}[B_j,\omega_j(Y)]}}{\sum_{\boldsymbol{a}\in\{0,1\}^m}\prod_{j=1}^{m} e^{s\mathbb{1}[a_j,\omega_j(Y)]}}\right] \\ &\quad \overset{(4.50),(4.89)}{=} \mathbb{E}\left[-\log_2 \frac{e^{s\mathbb{1}(B,\hat{B})}}{\sum_{a\in\{0,1\}} e^{s\mathbb{1}(a,\hat{B})}}\right] \end{aligned} \tag{4.90}$$

$$= -\Pr(B=\hat{B})\log_2\frac{e^s}{e^s+1} - \underbrace{\Pr(B\neq\hat{B})}_{=:\epsilon}\log_2\frac{1}{e^s+1} \tag{4.91}$$

$$\overset{(4.26)}{\geq} \mathbb{H}_2(\epsilon) \tag{4.92}$$

with equality if

$$s\colon \frac{1}{e^s+1} = \epsilon. \tag{4.93}$$

Thus, with a hard decision decoder, we can achieve

$$R^{\mathsf{hd,bin}} = [\mathbb{H}(\boldsymbol{B}) - m\,\mathbb{H}_2(\epsilon)]^+ \tag{4.94}$$

where

$$1-\epsilon = \sum_{j=1}^{m}\frac{1}{m}\sum_{a\in\{0,1\}} P_{B_j}(a)\int_{\mathcal{Y}_{aj}} p_{Y|B_j}(\tau|a)\,\mathrm{d}\tau. \tag{4.95}$$

For uniform input, the rate becomes

$$R^{\mathsf{hd,bin}}_{\mathsf{uni}} = m - m\,\mathbb{H}_2(\epsilon) = m[1-\mathbb{H}_2(\epsilon)]. \tag{4.96}$$

4.6 Discussion

We state the BMD rate (4.42) for the first time in [16] and we discuss it in detail in [24, Section VI.]. In [7], we prove the achievability of (4.42) for discrete memoryless channels.

In this section, we derived optimal decoding metrics for several special cases, and we can imagine there are many more variations that lead to interesting optimal metrics. An interesting direction of study is to optimize for specific metrics the achievable SE over the input distribution, possibly with a constraint on the FEC rate. According to our discussion in Section 1.4.5, this could lead to interesting non-convex optimization problems.

The version of record is available at: http://dx.doi.org/10.1561/0100000111

5

PAS Error Exponent

In this section, we show that on memoryless channels, PAS can achieve the spectral efficiency

$$\mathrm{SE} = \mathbb{I}(X;Y) \tag{5.1}$$

under certain conditions, namely

- The input alphabet $\mathcal{X}$ decomposes into $\mathcal{X} = \mathcal{A} \times \mathcal{U}$.
- The distribution P_X decomposes into $P_A \times P_S$ where P_S is uniform on $\mathcal{U}$.
- The FEC rate is sufficiently large, i.e.,

$$R_{\mathrm{fec}} \geq 1 - \frac{\log_2 |\mathcal{U}|}{\log_2 |\mathcal{X}|}. \tag{5.2}$$

In particular, when the capacity and the capacity-achieving distribution fulfill these conditions, and the FEC rate is chosen appropriately (recall our discussion in Section 1.4.5), then PAS is capacity-achieving. This result holds for random codes and for random linear codes.

To prove this result, we derive a PAS error exponent using similar techniques as in Section 3. Our setup differs from the layered PS ensemble that we analyzed in Section 3 in the following aspects.

- The non-constructive shaping layer is replaced by a constructive shaping layer using the CCDM developed in Section 2.5.
- The encoding is partially systematic, i.e., systematic on $\mathcal{A}$ and non-systematic on $\mathcal{U}$.
- As consequence of the partially systematic encoding, the code ensemble is deterministic on $\mathcal{A}$ and random on $\mathcal{U}$.
- The channel is memoryless.
- The decoding metric is memoryless.

This setup is more restrictive than in Section 3. The downside is that the results are less general than in Section 3. The advantage is that the error exponent is also valid for finite length, without the need to invoke an asymptotic measure concentration argument. Another advantage is that important practical aspects of PAS are accounted for, namely encoding by an invertible DM followed by systematic FEC encoding and the use of linear FEC.

5.1 FEC Layer

We consider the following transceiver setup:

- (As in Section 3) The channel is discrete-time with finite input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. We derive our results assuming a continuous-valued output. Our results also apply for discrete output alphabets.
- (**Different** from Section 3) Random coding: For indices $w = 1, 2, \ldots, |\mathcal{C}|$, we generate codewords $C^n(w)$ according to a general set of distributions $P_{C^n(w)}$. In particular, we permit different distributions for different indices and we also allow dependence among the codeword entries of the same codeword. The code is

$$\mathcal{C} = \{C^n(1), C^n(2), \ldots, C^n(|\mathcal{C}|)\}. \tag{5.3}$$

  This code ensemble is more general than needed and we will restrict it later to the partially systematic random code ensemble in our

derivation. We do so as starting with a more general ensemble and specializing later simplifies notation.

- (As in Section 3) The FEC rate is

$$R_{\text{fec}} = \frac{\log_2 |\mathcal{C}|}{nm} \tag{5.4}$$

where

$$m = \log_2 |\mathcal{X}|. \tag{5.5}$$

Equivalently, we have $|\mathcal{C}| = 2^{nmR_{\text{fec}}}$ codewords.

- (As in Section 3.5) We consider a memoryless non-negative decoding metric q on $\mathcal{X} \times \mathcal{Y}$ defined by

$$q^n(x^n, y^n) := \prod_{i=1}^{n} q(x_i, y_i), \quad x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n. \tag{5.6}$$

For the channel output y^n, we let the receiver decode with the rule

$$\hat{w} = \underset{w \in \{1,\ldots,2^{nmR_{\text{fec}}}\}}{\arg\max} \prod_{i=1}^{n} q(c_i(w), y_i). \tag{5.7}$$

- (As in Theorem 3.2) We consider the decoding error probability

$$P_e = \Pr(\hat{W} \neq w_0 | C^n(w_0) = x^n, Y^n = y^n) \tag{5.8}$$

where w_0 is the index of the transmitted codeword, $C^n(w_0) = x^n$ is the transmitted codeword, y^n is the channel output sequence, and $\hat{W}$ is the decoded index at the receiver. Note that the codewords $C^n(w)$, $w \neq w_0$ against which the decoder attempts to decode are random and the transmitted codeword $C^n(w_0) = x^n$ and the channel output y^n are deterministic.

5.1.1 FEC Rate Error Exponent

We consider the setting in Figure 5.1, i.e., we condition on that index w_0 was encoded to $C^n(w_0) = x^n$. For notational convenience, we assume without loss of generality $w_0 = 1$. We analyze the error probability

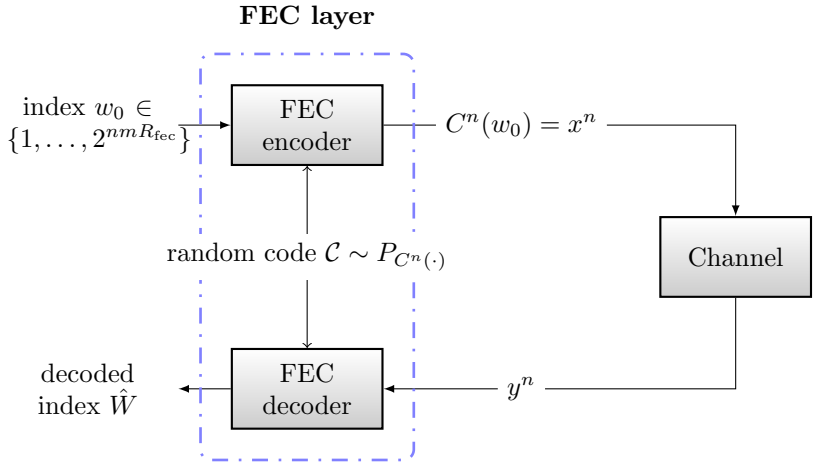


Figure 5.1: Random coding experiment for FEC rate error exponent. The codewords are generated for each index w according to an individual distribution $P_{C^n(w)}$.

$\Pr(\hat{W} \neq 1 | C^n(1) = x^n, Y^n = y^n)$, averaged over the random code. Note that we have $C^n(1) = x^n$ and for $w = 2, 3, \ldots, |\mathcal{C}|$, we have $C^n(w) \sim P_{C^n(w)}$. We have the implications

$$\hat{W} \neq 1 \Rightarrow \hat{W} = w' \neq 1 \tag{5.9}$$

$$\Rightarrow L(w') := \frac{q^n(C^n(w'), y^n)}{q^n(x^n, y^n)} \geq 1 \tag{5.10}$$

$$\Rightarrow \sum_{w=2}^{|\mathcal{C}|} L(w) \geq 1 \tag{5.11}$$

$$\Rightarrow \Big[\sum_{w=2}^{|\mathcal{C}|} L(w)\Big]^{\rho} \geq 1, \qquad \rho \geq 0. \tag{5.12}$$

We will use ρ for the same purposes as in [39, Chapter 5]: We will optimize over ρ to maximize the error exponent and to identify achievable rates. If event $\mathcal{A}$ implies event $\mathcal{B}$, then $\Pr[\mathcal{A}] \leq \Pr[\mathcal{B}]$. Therefore, we have

$$\Pr(\hat{W} \neq 1 | C^n(1) = x^n, Y^n = y^n)$$
$$\leq \Pr\left\{ \Big[\sum_{w=2}^{|\mathcal{C}|} L(w)\Big]^{\rho} \geq 1 \,\middle|\, C^n(1) = x^n, Y^n = y^n \right\} \tag{5.13}$$

$$\leq \mathbb{E}\left\{ \Big[\sum_{w=2}^{|\mathcal{C}|} L(w)\Big]^{\rho} \,\middle|\, C^n(1) = x^n, Y^n = y^n \right\} \tag{5.14}$$

$$= q^n(x^n, y^n)^{-\rho}\, \mathbb{E}\left[\Big[\sum_{w=2}^{|\mathcal{C}|} q^n[C^n(w), y^n]\Big]^{\rho} \right] \tag{5.15}$$

where

- the inequality in (5.14) follows by Markov's inequality (A.12),
- equality in (5.15) follows because for $w \neq 1$, the codeword $C^n(w)$ and the transmitted codeword $C^n(1)$ were generated independently so that $C^n(w)$ and $[C^n(1), Y^n]$ are independent.

Observe that $z \mapsto z^\rho$ is for $0 \leq \rho \leq 1$ a concave function and for $1 \leq \rho$ a convex function, see Figure 5.2. We therefore restrict the parameter ρ to

$$0 \leq \rho \leq 1 \tag{5.16}$$

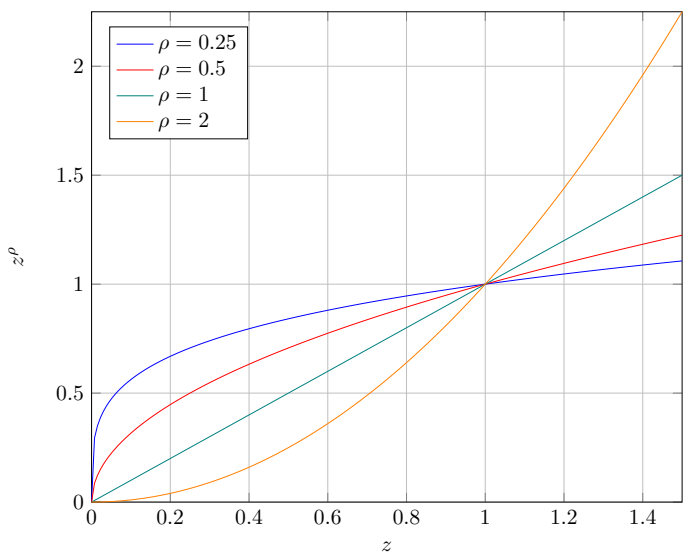


Figure 5.2: The function $z \mapsto z^\rho$ is concave for $0 \leq \rho \leq 1$ and it is convex for $1 \leq \rho$.

so that $(\cdot)^\rho$ is concave and by Jensen's inequality (A.7), we have $\mathbb{E}(Z^\rho) \leq \mathbb{E}(Z)^\rho$. We use this to bound (5.15) further:

$$(5.15) \leq q^n(x^n, y^n)^{-\rho}\, \mathbb{E}\left[\sum_{w=2}^{|\mathcal{C}|} q^n[C^n(w), y^n]\right]^\rho \tag{5.17}$$

$$= |\mathcal{C}|^\rho q^n(x^n, y^n)^{-\rho}\, \mathbb{E}\left[\frac{1}{|\mathcal{C}|}\sum_{w=2}^{|\mathcal{C}|} q^n[C^n(w), y^n]\right]^\rho. \tag{5.18}$$

We have pulled out the factor $|\mathcal{C}| = 2^{nmR_{\text{fec}}}$, as we want to bound the error probability in terms of the FEC rate R_{fec}.

5.1.2 PAS FEC Rate Error Exponent

So far, we have analyzed the error probability of decoding a transmitted codeword $C^n(w_0) = x^n$ against the random code $\mathcal{C} \sim P_{C^n(\cdot)}$. We next consider a specific instance of the random coding experiment $P_{C^n(\cdot)}$, which reflects the PAS that we developed in Section 1. This will allow us to explicitly state an encoder and to quantify into how many distinct codewords we can encode.

We now make the assumption that the input alphabet decomposes into two parts $\mathcal{X} = \mathcal{A} \times \mathcal{U}$. We represent the codeword index by $w = a^n s^{\gamma n}$, i.e., code size and FEC rate are respectively given by

$$|\mathcal{C}| = |\mathcal{A}|^n |\mathcal{U}|^{\gamma n} \tag{5.19}$$

$$mR_{\text{fec}} = \log_2 |\mathcal{A}| + \gamma \log_2 |\mathcal{U}| \tag{5.20}$$

$$= \log_2 |\mathcal{X}| - (1-\gamma)\log_2 |\mathcal{U}| \tag{5.21}$$

We now have

$$(1-\gamma)\log_2 |\mathcal{U}| = m(1 - R_{\text{fec}}) \tag{5.22}$$

where the right-hand side is the FEC redundancy per symbol.

Example 5.1 (Amplitude Shift Keying)**.** The 8-ASK constellation

$$\mathcal{X} = \{\pm 1, \pm 3, \pm 5, \pm 7\} \tag{5.23}$$

decomposes into an amplitude set and a sign set via

$$\mathcal{A} = \{1, 3, 5, 7\}, \quad \mathcal{U} = \{-1, 1\} \tag{5.24}$$

$$\mathcal{A} \times \mathcal{U} \to \mathcal{X} \colon (a, s) \mapsto a \cdot s \tag{5.25}$$

$$\mathcal{X} \to \mathcal{A} \times \mathcal{U} \colon x \mapsto (|x|, \operatorname{sign}(x)). \tag{5.26}$$

For a 2^m-ASK constellation, the FEC rate is

$$mR_{\text{fec}} = \log_2 |\mathcal{A}| + \gamma \log_2 |\mathcal{U}| = m - 1 + \gamma. \tag{5.27}$$

Example 5.2 (Quadrature Amplitude Modulation)**.** The 16-quadrature amplitude modulation (QAM) constellation

$$\mathcal{X} = \{\pm 1 \pm j, \pm 1 \pm 3j, \pm 3 \pm j, \pm 3 \pm 3j\} \tag{5.28}$$

decomposes into

$$\mathcal{A} = \{1, 3\} \times \{1, 3\}, \quad \mathcal{U} = \{(\pm 1, \pm 1)\} \tag{5.29}$$

$$\mathcal{A} \times \mathcal{U} \to \mathcal{X} \colon (a, s) \mapsto a_1 s_1 + j a_2 s_2 \tag{5.30}$$

$$\mathcal{X} \to \mathcal{A} \times \mathcal{U} \colon$$

$$x \mapsto [(|\operatorname{Re}(x)|, |\operatorname{Im}(x)|), (\operatorname{sign}[\operatorname{Re}(x)], \operatorname{sign}[\operatorname{Im}(x)])]. \tag{5.31}$$

Note that this is equivalent to interpreting 16-QAM as the Cartesian product of two 4-ASK constellations.

Example 5.3 (PAS is not restricted to amplitude shaping)**.** We revisit the 8-ASK constellation of Example 5.1

$$\mathcal{X} = \{\pm 1, \pm 3, \pm 5, \pm 7\} \tag{5.32}$$

and different from Example 5.1, we decompose it into sets $\mathcal{A}$ and $\mathcal{U}$ via

$$\mathcal{A} = \{-7, -3, 1, 5\}, \quad \mathcal{U} = \{-1, 1\} \tag{5.33}$$

$$\mathcal{A} \times \mathcal{U} \to \mathcal{X} \colon (a, s) \mapsto a \cdot s = x. \tag{5.34}$$

Note that in this example, if $\operatorname{sign}(X)$ is uniformly distributed on $\{-1, 1\}$, then S as defined by (5.34) is uniformly distributed on $\{-1, 1\}$, although S is not the sign of X and A as defined by (5.34) is not the amplitude of X. In that sense, 'amplitude' in "probabilistic amplitude shaping" does not necessarily need to be an amplitude.

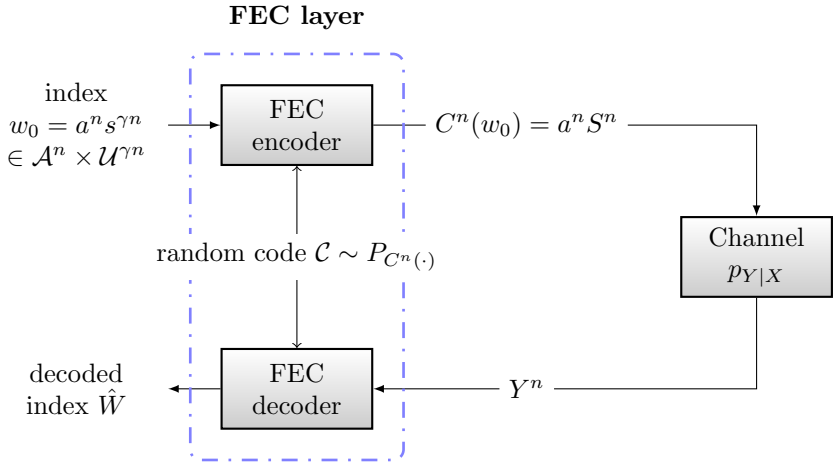


Figure 5.3: PAS random coding experiment.

Our random coding experiment is as follows. We encode the index $w \in \mathcal{A}^n \times \mathcal{U}^{\gamma n}$ to

$$w = a^n s^{\gamma n} \mapsto C^n(w) = a^n S^n(w) \tag{5.35}$$

where the $S_i(w)$, $i = 1, 2, \ldots, n$ are independent and uniformly distributed on $\mathcal{U}$. Note that this corresponds to partially systematic encoding: the part a^n of the codeword index appears in the codeword. Next, we condition on an index $w_0 = a^n s^{\gamma n}$. This index selects a codeword $a^n S^n$, which consists of the deterministic part a^n and a random part S^n, which is stochastically independent of w_0. In terms of the random coding distribution $P_{C^n(\cdot)}$, we have

$$P_{C^n(a^n s^{\gamma n})}(\alpha^n \sigma^n) = \begin{cases} \frac{1}{|\mathcal{U}|^n}, & \text{if } \alpha^n = a^n \\ 0, & \text{otherwise} \end{cases}, \quad a^n, \alpha^n \in \mathcal{A}^n, s^{\gamma n} \in \mathcal{U}^{\gamma n}, \sigma^n \in \mathcal{U}^n \tag{5.36}$$

where $P_{C^n(w_0)}$ indeed depends on w_0. The PAS coding experiment is displayed in Figure 5.3. We have the following differences to the coding experiment in Figure 5.1:

1. For the transmitted codeword, we condition on $C^n(w_0) = a^n S^n$, which is random.

2. Since the transmitted codeword is random, also the channel output Y^n is random. We therefore analyze the average error probability, where the average is over S^n and Y^n.

3. We assume the channel is memoryless with channel law $p_{Y|X}$.

We now have

$$\Pr(\hat{W} \neq w_0 | A^n = a^n) = \mathbb{E}\left[\Pr(\hat{W} \neq w_0 | X^n = a^n S^n, Y^n)\right] \tag{5.37}$$

$$\overset{(5.18)}{\leq} |\mathcal{C}|^\rho \, \mathbb{E}\Bigg\{ q^n(a^n S^n, Y^n)^{-\rho} \cdot \mathbb{E}\left[\frac{1}{|\mathcal{C}|} \sum_{\substack{w \in \mathcal{A}^n \times \mathcal{U}^{\gamma n} \\ w \neq w_0}} q^n[C^n(w), Y^n] \,\middle|\, Y^n \right]^\rho \,\middle|\, A^n = a^n \Bigg\}. \tag{5.38}$$

We develop the innermost expectation in (5.38). We have

$$\mathbb{E}\left[\frac{1}{|\mathcal{C}|} \sum_{\substack{w \in \mathcal{A}^n \times \mathcal{U}^{\gamma n} \\ w \neq w_0}} q^n[C^n(w), Y^n] \,\middle|\, Y^n \right]$$

$$= \mathbb{E}\left[\frac{1}{|\mathcal{C}|} \sum_{\substack{w \in \mathcal{A}^n \times \mathcal{U}^{\gamma n} \\ w \neq w_0}} q^n[a^n(w) S^n(w), Y^n] \,\middle|\, Y^n \right] \tag{5.39}$$

$$= \frac{1}{|\mathcal{A}|^n |\mathcal{U}|^{\gamma n}} \sum_{\substack{w \in \mathcal{A}^n \times \mathcal{U}^{\gamma n} \\ w \neq w_0}} \sum_{t^n \in \mathcal{U}^n} \frac{1}{|\mathcal{U}|^n} q^n[a^n(w) t^n, Y^n] \tag{5.40}$$

$$= \frac{1}{|\mathcal{U}|^{\gamma n}} \sum_{\substack{w \in \mathcal{A}^n \times \mathcal{U}^{\gamma n} \\ w \neq w_0}} \sum_{t^n \in \mathcal{U}^n} \frac{1}{|\mathcal{X}|^n} q^n[a^n(w) t^n, Y^n] \tag{5.41}$$

$$= \frac{1}{|\mathcal{U}|^{\gamma n}} \sum_{\substack{w \in \mathcal{A}^n \times \mathcal{U}^{\gamma n} \\ w \neq w_0}} \sum_{t^n \in \mathcal{U}^n} \prod_{i=1}^n \frac{1}{|\mathcal{X}|} q[a_i(w) t_i, Y_i] \tag{5.42}$$

$$\leq \frac{1}{|\mathcal{U}|^{\gamma n}} \sum_{a^n \in \mathcal{A}^n} \sum_{s^{\gamma n} \in \mathcal{U}^{\gamma n}} \sum_{t^n \in \mathcal{U}^n} \prod_{i=1}^n \frac{1}{|\mathcal{X}|} q(a_i t_i, Y_i) \tag{5.43}$$

$$= \sum_{a^n \in \mathcal{A}^n} \sum_{t^n \in \mathcal{U}^n} \prod_{i=1}^n \frac{1}{|\mathcal{X}|} q(a_i t_i, Y_i) \tag{5.44}$$

$$= \prod_{i=1}^{n} \sum_{a\in\mathcal{A}} \sum_{t\in\mathcal{U}} \frac{1}{|\mathcal{X}|} q(at, Y_i) \tag{5.45}$$

$$= \prod_{i=1}^{n} \frac{1}{|\mathcal{X}|} \sum_{c\in\mathcal{X}} q(c, Y_i) \tag{5.46}$$

where

- (5.39) follows by (5.35),
- (5.40) follows because for each $w \neq w_0$, $S^n(w)$ is independent of Y^n and because the entries of the $S^n(w)$ are independent and uniformly distributed on $\mathcal{U}$,
- (5.41) and (5.46) use $\mathcal{X} = \mathcal{A} \times \mathcal{U}$,
- (5.42) follows because the metric q^n is memoryless,
- the inequality in (5.43) follows since the sum now includes index w_0.
- for (5.46), $a^n S^n \longrightarrow \boxed{p_{Y|X}} \longrightarrow Y^n$.

Remark 5.1. Note that we would get (5.46) also for a random code with its $n|\mathcal{C}|$ entries independent and uniformly distributed on $\mathcal{X}$. Our PAS random code can thus be interpreted as a partially systematic version of a uniformly distributed random code.

Remark 5.2. Note that we have not yet made use of the assumption that the channel is memoryless, in particular, (5.46) also holds for channels with memory.

We now make use of our assumption that the channel is memoryless. By inserting (5.46) in the outer expectation in (5.38) we have

$$\mathbb{E}\left\{ \prod_{i=1}^{n} q(a_i S_i, Y_i)^{-\rho} \left[\frac{1}{|\mathcal{X}|} \sum_{c\in\mathcal{X}} q(c, Y_i) \right]^{\rho} \middle| A^n = a^n \right\}$$

$$= \prod_{i=1}^{n} \mathbb{E}\left\{ q(a_i S_i, Y_i)^{-\rho} \left[\frac{1}{|\mathcal{X}|} \sum_{c\in\mathcal{X}} q(c, Y_i) \right]^{\rho} \middle| A_i = a_i \right\} \tag{5.47}$$

$$= \prod_{i=1}^{n} \mathbb{E}\left\{ q(a_i S, Y)^{-\rho} \left[\frac{1}{|\mathcal{X}|} \sum_{c \in \mathcal{X}} q(c, Y) \right]^{\rho} \middle| A = a_i \right\} \tag{5.48}$$

$$= \prod_{\alpha \in \mathcal{A}} \mathbb{E}\left\{ q(\alpha S, Y)^{-\rho} \left[\frac{1}{|\mathcal{X}|} \sum_{c \in \mathcal{X}} q(c, Y) \right]^{\rho} \middle| A = \alpha \right\}^{N(\alpha|a^n)} \tag{5.49}$$

where

- equality in (5.47) follows because by assumption, the channel is memoryless, so conditioned on $A^n = a^n$, $S_i Y_i$ and $S_j Y_j$ are independent for $i \neq j$,
- for the ith factor in (5.48), $a_i S \longrightarrow \boxed{p_{Y|X}} \longrightarrow Y$, where we replace $A_i S_i Y_i$ by ASY. We can do so, since the S_i are iid $\sim P_S$ and independent of A_i, and the channel is memoryless, in other words, the statistics of $A_i S_i Y_i | A_i = a$ does not depend on the index i.
- for the αth factor in (5.49), $\alpha S \longrightarrow \boxed{p_{Y|X}} \longrightarrow Y$,
- in (5.49),

$$N(\alpha|a_n) = |\{i \colon a^n = \alpha\}| \tag{5.50}$$

$$= \text{number of occurrences of letter } \alpha \text{ in } a^n. \tag{5.51}$$

We denote the empirical distribution of a^n by P_A and write $N(\alpha|a^n) = nP_A(\alpha)$. Evaluating $-\frac{1}{n}\log_2(\cdot)$ in (5.49) gives

$$\tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho, P_A, q) = \sum_{a \in \mathcal{A}} P_A(a) \log_2 \mathbb{E}\left\{ \left[|\mathcal{X}| \frac{q(aS, Y)}{\sum_{c \in \mathcal{X}} q(c, Y)} \right]^{\rho} \middle| A = a \right\}. \tag{5.52}$$

Finally, by inserting (5.49) in (5.38) and then evaluating $-\frac{1}{n}\log_2(\cdot)$ in (5.38) yields the PAS error exponent and the bound on decoding error probability, respectively:

$$\mathrm{E}_{\mathrm{PAS}}(R_{\mathrm{fec}}, \rho, P_A, q) = \tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho, P_A, q) - \rho m R_{\mathrm{fec}} \tag{5.53}$$

$$\Pr(\hat{W} \neq W | A^n = a^n) \leq 2^{-n\, \mathrm{E}_{\mathrm{PAS}}(R_{\mathrm{fec}}, \rho, P_A, q)}. \tag{5.54}$$

5.2 PAS with Random Linear Code

Suppose we have $|\mathcal{X}| = 2^m$, $|\mathcal{A}| = 2^{m-1}$, and $|\mathcal{U}| = 2$, for instance, $\mathcal{X}$ may be a 2^m-ASK constellation. We represent $\mathcal{A}$ by a binary label $\boldsymbol{b} = b_1 \cdots b_{m-1} \in \{0,1\}^{m-1}$ and $\mathcal{U}$ by $s \in \{0,1\}$. For this scenario, we next show that our results for PAS also hold for random linear coding. This can be generalized to other scenarios in a straightforward manner. For the considered scenario, the PAS random coding (5.35) is

$$\boldsymbol{w} = \boldsymbol{b}^n s^{\gamma n} \mapsto \boldsymbol{b}^n S^n(\boldsymbol{w}), \quad \boldsymbol{w} \in \{0,1\}^{(m-1+\gamma)n} \tag{5.55}$$

where the entries of $S^n(\boldsymbol{w})$, $\boldsymbol{w} \in \{0,1\}^{(m-1+\gamma)n}$ are independent and uniformly distributed on $\{0,1\}$. We used two properties in our derivation:

1. In (5.15) and (5.40), we used that for $\boldsymbol{w} \neq \boldsymbol{w}'$, $S^n(\boldsymbol{w})$ and $S^n(\boldsymbol{w}')$ are independent.

2. In (5.40), we used that for each $\boldsymbol{w} \in \{0,1\}^{(m-1+\gamma)n}$, the n entries of $S^n(\boldsymbol{w})$ are independent and uniformly distributed on $\{0,1\}$.

The results we derived apply for any coding experiment that fulfills properties 1. and 2. In particular, consider the random coset coding

$$\boldsymbol{w} = \boldsymbol{b}^n s^{\gamma n} \mapsto (\boldsymbol{b}^n, \boldsymbol{w}\boldsymbol{P} \oplus \boldsymbol{p}) \tag{5.56}$$

where $\boldsymbol{P}$ is an $(m-1+\gamma)n \times n$ matrix and $\boldsymbol{p}$ is a length n vector, both with their entries independent and uniformly distributed on $\{0,1\}$. For this random coset code (5.56), Property 2. holds for all messages $\boldsymbol{w}$, in particular, it holds for $\boldsymbol{w} = \boldsymbol{0}$ because of $\boldsymbol{p}$. Furthermore, Property 1. holds because of the independent and uniformly distributed rows of $\boldsymbol{P}$. Thus, the PAS error exponent also holds for the random coset coding (5.56).

Remark 5.3. In the original PAS paper [24], it was argued that the parity bits after systematic encoding are approximately uniformly distributed (uniform check bit assumption), and empirical evidence was provided. In our introductory Section 1 on practical PAS, we do not make this assumption anymore, as it is not needed: it is sufficient to observe that the signal power does not depend on the sign. For the random coset

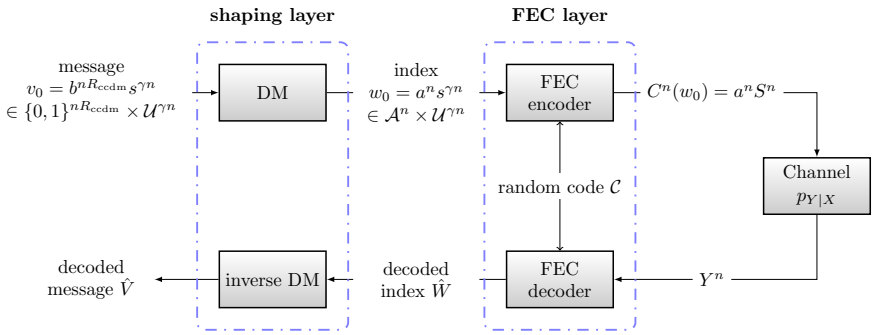


Figure 5.4: The PAS ensemble. The shaping layer is deterministic. The DM maps nR_{ccdm} input bits to n symbols in $\mathcal{A}$ using a CCDM and it copies γn input symbols in $\mathcal{U}$ unchanged to its output.

coding (5.56), the check bits are exactly independent and uniformly distributed. The reason is that we are considering a random code ensemble, not a specific code. Accordingly, the PAS error exponent holds for the error probability averaged over the random code ensemble, which implies the existence of a code in the ensemble that achieves the error exponent, but an explicit construction of such code is not provided.

5.3 Shaping Layer

We display the full PAS scheme including the shaping layer in Figure 5.4. Our goal is to associate an SE with the PAS error exponent. Recall that the FEC rate is

$$mR_{\text{fec}} = \log_2 |\mathcal{A}| + \gamma \log_2 |\mathcal{U}|. \tag{5.57}$$

The PAS error exponent $\mathrm{E}_{\text{PAS}}(R_{\text{fec}}, \rho, P_A, q)$ depends on the type of the amplitude sequence a^n. Thus, the number of codewords for which it holds is

$$(\text{number of sequences } a^n \text{ of type } P_A) \times |\mathcal{U}|^{\gamma n}. \tag{5.58}$$

Using the CCDM we developed in Section 2.5, we can encode

$$R_{\text{ccdm}}(P_A, n) \cdot n \text{ bits}$$

to length n sequences of type P_A. In total, we can encode

$$[R_{\text{ccdm}}(P_A, n) + \gamma \log_2 |\mathcal{U}|] \cdot n \text{ bits} \tag{5.59}$$

and the SE is

$$\begin{aligned} \text{SE} &= R_{\text{ccdm}}(P_A, n) + \gamma \log_2 |\mathcal{U}| && (5.60)\\ &= R_{\text{ccdm}}(P_A, n) + \log_2 |\mathcal{U}| - (1-\gamma) \log_2 |\mathcal{U}| && (5.61)\\ &= R_{\text{ccdm}}(P_A, n) + \log_2 |\mathcal{U}| - m(1 - R_{\text{fec}}) \quad \left[\frac{\text{bits}}{\text{symbol}}\right] && (5.62) \end{aligned}$$

where the last line follows by (5.22). The last line has the following interpretation:

- $R_{\text{ccdm}}(P_A, n) + \log_2 |\mathcal{U}|$ is the uncoded SE.
- $m(1 - R_{\text{fec}})$ is the FEC redundancy per symbol.
- The constraint $\gamma \geq 0$ translates into $R_{\text{fec}} \geq 1 - \frac{\log_2 |\mathcal{U}|}{m}$.

Remark 5.4. We can express the FEC rate in terms of the SE by

$$mR_{\text{fec}} = \text{SE} + [\log_2 |\mathcal{A}| - R_{\text{ccdm}}(P_A, n)]\,. \tag{5.63}$$

Since $R_{\text{ccdm}}(P_A, n) < \log_2 |\mathcal{A}|$, mR_{fec} is larger than the SE. This makes sense, since of all codewords, we only transmit those that have amplitude type P_A.

For a given decoding metric q, we can improve the error bound by evaluating the error exponent in q^s and maximizing over s, following the arguments of Sections 3.4.2 and 4.5. The following theorem summarizes our findings.

Theorem 5.1. For $R_{\text{fec}} \geq 1 - \frac{\log_2 |\mathcal{U}|}{m}$, PAS can operate at the spectral efficiency

$$\text{SE} = R_{\text{ccdm}}(P_A, n) + \log_2 |\mathcal{U}| - m(1 - R_{\text{fec}}) \quad \left[\frac{\text{bits}}{\text{symbol}}\right] \tag{5.64}$$

with the decoding error probability bounded from above by

$$2^{-n\, \text{E}^*_{\text{PAS}}(R_{\text{fec}}, \rho, P_A, q)} \tag{5.65}$$

where

$$\text{E}^*_{\text{PAS}}(R_{\text{fec}}, \rho, P_A, q) = \max_{s>0} \text{E}_{\text{PAS}}(R_{\text{fec}}, \rho, P_A, q^s). \tag{5.66}$$

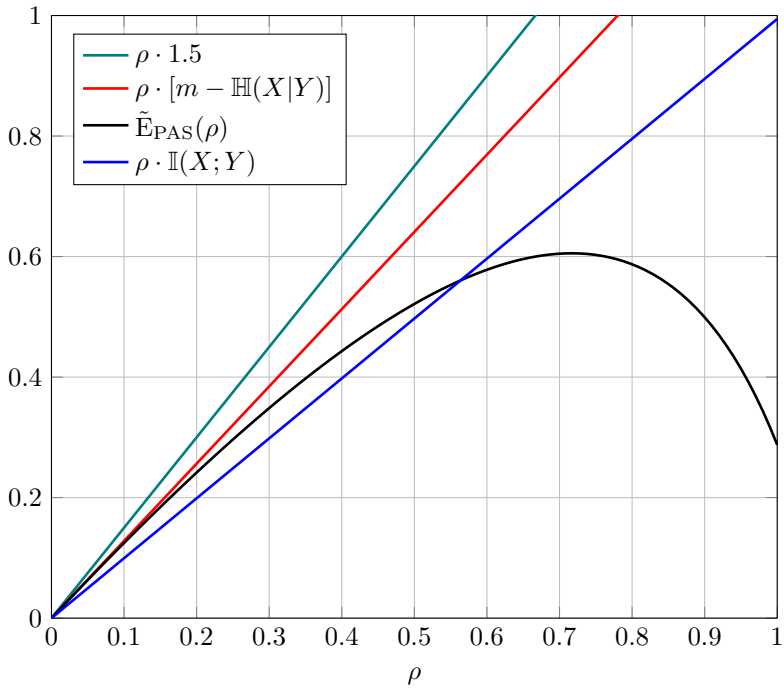


Figure 5.5: Visualization of error exponent. By increasing the FEC rate R_{fec}, the slope of $\rho m R_{\mathrm{fec}}$ becomes steeper, until it exceeds $\tilde{\mathrm{E}}_{\mathrm{PAS}}$ everywhere. The phase transition occurs in $\rho = 0$ for $mR_{\mathrm{fec}} = m - \mathbb{H}(X|Y)$, which corresponds to the FEC redundancy $m(1 - R_{\mathrm{fec}}) = \mathbb{H}(X|Y)$.

5.4 PAS Achieves the AWGN Capacity

By making n large, the error probability bound approaches zero, as long as the error exponent is positive, which is the case when $\tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho, P_A, q)$ is larger than $\rho m R_{\mathrm{fec}}$.

Example 5.4. Consider an AWGN channel

$$Y = X + Z \tag{5.67}$$

where $X = A \cdot S$ is 4-ASK input with alphabet $\mathcal{X} = \{\pm 1, \pm 3\}$, amplitude distribution $P_A(1) = \frac{4}{5}$, $P_A(3) = \frac{1}{5}$, uniform sign distribution $P_S(1) = P_S(-1) = \frac{1}{2}$ and where Z is zero mean Gaussian with variance $\sigma^2 =$

$\mathbb{E}(X^2)/3$, i.e., the SNR is 3 and the AWGN channel capacity is $\frac{1}{2}\log_2(1+\mathsf{snr}) = 1$. For the decoding metric

$$q(a,b) = P_{X|Y}(a|b) \tag{5.68}$$

we plot $\tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho)$ and $\rho m R_{\mathrm{fec}}$ in Figure 5.5. We make the following two observations:

- The function $\tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho)$ is concave in ρ.
- $\tilde{\mathrm{E}}_{\mathrm{PAS}}(0) = 0$.

This implies that the largest achievable FEC rate mR_{fec} is given by the slope of $\tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho)$ in $\rho = 0$.

We now show that our observations in Example 5.4 are true in general. First,

$$\tilde{\mathrm{E}}_{\mathrm{PAS}}(0) = 0 \tag{5.69}$$

follows directly by (5.52).

Theorem 5.2. $\tilde{\mathrm{E}}_{\mathrm{PAS}}(0)$ is concave for $\rho \in \mathbf{R}$.

Proof. In Section 5.6, we show the non-negativity of the second derivative of a more general expression, from which the concavity of $\tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho)$ follows. □

By (5.69) and the concavity of $\tilde{\mathrm{E}}_{\mathrm{PAS}}$, it follows that for $0 \le \rho \le 1$, the largest slope of $\tilde{\mathrm{E}}_{\mathrm{PAS}}$ is in $\rho = 0$. To calculate the slope, define

$$Z_a = \left\{ \log |\mathcal{X}| \frac{q(Y, aS)}{\sum_{c\in\mathcal{X}} q(Y,c)} \middle| A = a \right\}. \tag{5.70}$$

We now have

$$\tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho, P_A, q) = \sum_{a\in\mathcal{A}} P_A(a) \log_2 \mathbb{E}(e^{\rho Z_a}) \tag{5.71}$$

$$= \sum_{a\in\mathcal{A}} P_A(a) \frac{\log[\mathsf{mgf}_{Z_a}(\rho)]}{\log 2} \tag{5.72}$$

where mgf is the moment generating function (MGF) (A.13) and by (A.14), we have

$$\frac{\partial}{\partial\rho}\tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho,P_A,q)\Big|_{\rho=0} = \sum_{a\in\mathcal{A}} P_A(a)\frac{\mathbb{E}(Z_a)}{\log 2} \tag{5.73}$$

$$= \mathbb{E}\left[\log_2|\mathcal{X}|\frac{q(X,Y)}{\sum_{c\in\mathcal{X}}q(c,Y)}\right] \tag{5.74}$$

$$= m - \mathbb{E}\left[-\log_2\frac{q(X,Y)}{\sum_{c\in\mathcal{X}}q(c,Y)}\right] \tag{5.75}$$

$$= m - \mathsf{X}(q,X,Y). \tag{5.76}$$

Example 5.5. We continue with Example 5.4. For the decoding metric $q(c,b) = P_{X|Y}(b|c)$, an achievable FEC rate is

$$mR_{\mathrm{fec}} = \frac{\partial}{\partial\rho}\tilde{\mathrm{E}}_{\mathrm{PAS}}(\rho,P_A,q)\Big|_{\rho=0} \tag{5.77}$$

$$= m - \mathbb{E}\left[-\log_2\frac{P_{X|Y}(X|Y)}{\sum_{c\in\mathcal{X}}P_{X|Y}(c|Y)}\right] \tag{5.78}$$

$$= m - \mathbb{H}(X|Y) \tag{5.79}$$

By Theorem 2.1, for large n a CCDM generates type P_A sequences with rate $R_{\mathrm{ccdm}}(P_A,\infty) = \mathbb{H}(A)$. The SE is

$$\mathrm{SE} = \mathbb{H}(A) + \log_2|\mathcal{U}| - m(1-R_{\mathrm{fec}}) = \mathbb{H}(X) - m(1-R_{\mathrm{fec}}) \tag{5.80}$$

and the achievable SE is

$$\mathrm{SE}^* = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{I}(X;Y) \tag{5.81}$$

i.e., PAS asymptotically achieves $\mathbb{I}(X;Y)$.

5.5 PAS with Finite Length CCDM

Consider a CCDM with output length n_{ccdm} amplitudes and rate $R_{\mathrm{ccdm}}(P_{\bar{A}},n_{\mathrm{ccdm}})$. We now use the FEC code over ℓ CCDM outputs, so that the codeword length $n = \ell\cdot n_{\mathrm{ccdm}}$ is a multiple of n_{ccdm}. As all CCDM outputs are of type $n_a = n_{\mathrm{ccdm}}\cdot P_{\bar{A}}(a)$, $a\in\mathcal{A}$, all sequences

consisting of ℓ concatenated CCDM outputs belong to the same type class as well, i.e., the number of occurences of amplitude $a \in \mathcal{A}$ is always

$$\ell \cdot n_a = \ell \cdot n_{\text{ccdm}} \cdot P_{\bar{A}}(a). \tag{5.82}$$

Consequently, by letting ℓ approach infinity, the achievable FEC rate is (5.76). By the arguments of Sections 3.4.2 and 4.5, we can replace the cross-equivocation in (5.76) by the tightened cross-equivocation

$$\mathsf{X}^*(q, X, Y) = \min_{s>0} \mathsf{X}(q^s, X, Y) \tag{5.83}$$

and we arrive at the following theorem.

Theorem 5.3. For a DM that concatenates ℓ outputs of a CCDM with (short) output length n_{ccdm}, asymptotically in ℓ, an achievable rate is

$$\begin{aligned}&[R_{\text{ccdm}}(P_{\bar{A}}, n_{\text{ccdm}}) + \log_2 |\mathcal{U}| - \mathsf{X}^*(q, X, Y)]^+ \\ &\quad = [mR_{\text{ss}} - \mathsf{X}^*(q, X, Y)]^+\end{aligned} \tag{5.84}$$

where R_{ss} is the PAS shaping set rate (4.11).

5.6 Proofs

We now prove the concavity of $\tilde{\text{E}}_{\text{PAS}}$. Consider the function

$$g(x) = \log\left(\sum_{i=1}^{n} a_i b_i^x\right) = \log\left(\sum_{i=1}^{n} a_i e^{x \log b_i}\right) \tag{5.85}$$

where the a_i and b_i are non-negative. We verify that $g(x)$ is convex on $\mathbf{R}$ by showing that the second derivative is non-negative. We have

$$\frac{\partial}{\partial x} g(x) = \frac{\sum_{i=1}^{n} \log(b_i) a_i e^{x \log b_i}}{\sum_{i=1}^{n} a_i e^{x \log b_i}} \tag{5.86}$$

$$\begin{aligned}&\frac{\partial^2}{\partial x^2} g(x) \\ &= \frac{(\sum_{i=1}^{n} \log(b_i)^2 a_i e^{x \log b_i})(\sum_{i=1}^{n} a_i e^{x \log b_i}) - (\sum_{i=1}^{n} \log(b_i) a_i e^{x \log b_i})^2}{(\sum_{i=1}^{n} a_i e^{x \log b_i})^2}\end{aligned} \tag{5.87}$$

For $i = 1, \ldots, n$, define

$$u_i = \log(b_i)\sqrt{a_i e^{x \log b_i}} \tag{5.88}$$

$$v_i = \sqrt{a_i e^{x \log b_i}}. \tag{5.89}$$

The numerator of the second derivative is now

$$\boldsymbol{u}\boldsymbol{u}^T \boldsymbol{v}\boldsymbol{v}^T - (\boldsymbol{u}\boldsymbol{v}^T)^2 \tag{5.90}$$

which is non-negative, by the Cauchy-Schwarz inequality (A.1). The derivation above also holds if the sum over i is replaced by an integral over some variable τ.

5.7 Discussion

In [1], achievable rates are derived for PAS assuming a random source, in place of a DM. The work [43] analyzes PAS using typicality.

The PAS error exponent that we derived in this section has several appealing properties, for instance, it holds for linear codes, it explicitly uses a DM, and it provides an error bound for finite length. Somewhat unsatisfactory is that we had to use a CCDM so that all amplitudes have equal composition. As we have seen in Section 2, CCDM loses significantly compared to MCDM for finite length. Thus, a finite length analysis that allows for the use of an MCDM is interesting to study.

The version of record is available at: http://dx.doi.org/10.1561/0100000111

Appendices

A

Preliminaries

A.1 Mathematics

Cauchy-Schwarz Inequality

For two row vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^M$, the Cauchy-Schwarz inequality is

$$\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{v}\boldsymbol{v}^T - (\boldsymbol{u}\boldsymbol{v}^T) \geq 0 \tag{A.1}$$

with equality if and only if $\boldsymbol{u}$ and $\boldsymbol{v}$ are linearly dependent.

Big O Notation

- f is bounded below by g asymptotically:

$$f \in \Omega(g) \Leftrightarrow \liminf_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| > 0. \tag{A.2}$$

- f is bounded above by g asymptotically:

$$f \in \mathcal{O}(g) \Leftrightarrow \liminf_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| < \infty. \tag{A.3}$$

- f is bounded above and below by g asymptotically:

$$f \in \Theta(g) \Leftrightarrow f \in \Omega(g) \text{ and } f \in \mathcal{O}(g). \tag{A.4}$$

Stirling's Formula

By [36, Section II.9],

$$\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n}}. \tag{A.5}$$

Convexity

- A real-valued function f is *convex* on the interval $[A, B] \subseteq \mathbf{R}$ if for each $x_1, x_2 \in [A, B]$ and $0 \leq \lambda \leq 1$, we have

$$f[\lambda x_1 + (1-\lambda)x_2] \leq \lambda f(x_1) + (1-\lambda)f(x_2).$$

- The function f is *concave* on $[A, B]$ if $-f$ is convex on $[A, B]$.

- Let X be a random variable with support $[A, B]$. Jensen's inequality states that for f convex on $[A, B]$, we have

$$f[\mathbb{E}(X)] \leq \mathbb{E}[f(X)]. \tag{A.6}$$

  For f concave on $[A, B]$, Jensen's inequality states that

$$f[\mathbb{E}(X)] \geq \mathbb{E}[f(X)]. \tag{A.7}$$

Sum-of-Products and Product-of-Sums

Consider m sets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_m$. The Cartesian product of the m sets is the set of ordered m tuples

$$\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_m = \{\boldsymbol{a} = (a_1, a_2, \ldots, a_m) | a_i \in \mathcal{X}_i, i = 1, 2, \ldots, m\}. \tag{A.8}$$

We now have the following sum-of-products as product-of-sums identity:

$$\sum_{\boldsymbol{a} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m} \prod_{j=1}^{m} a_j = \prod_{j=1}^{m} \sum_{a \in \mathcal{X}_j} a. \tag{A.9}$$

Example A.1. Consider

$$m = 2, \quad \mathcal{X}_1 = \{b, c\}, \quad \mathcal{X}_2 = \{d, e, f\}.$$

We have

$$\sum_{\boldsymbol{a}\in\mathcal{X}_1\times\mathcal{X}_2}\prod_{j=1}^{2}a_j = bd+be+bf+cd+ce+cf$$

$$\prod_{j=1}^{2}\sum_{a\in\mathcal{X}_j} = (b+c)(d+e+f) = bd+be+bf+cd+ce+cf.$$

Example A.2. We often encounter the case when $\mathcal{X}_j$ is the set of probabilities defined by a distribution P_{X_j} on an alphabet $\mathcal{X}$, i.e.,

$$\mathcal{X}_j = \{P_{X_j}(a)|a\in\mathcal{X}\}.$$

In particular, the sets $\mathcal{X}_j$ are all of the same size, i.e., $|\mathcal{X}_1| = |\mathcal{X}_2| = \cdots = |\mathcal{X}_m| = |\mathcal{X}|$. The Cartesian product of m copies of $\mathcal{X}$ is

$$\mathcal{X}^m = \underbrace{\mathcal{X}\times\mathcal{X}\times\cdots\times\mathcal{X}}_{m \text{ times}}$$

The sum-of-products as product-of-sums identity can now be written as

$$\begin{aligned}\sum_{\boldsymbol{p}\in\mathcal{X}_1\times\cdots\mathcal{X}_m}\prod_{j=1}^{m}p_j &= \sum_{\boldsymbol{a}\in\mathcal{X}^m}\prod_{j=1}^{m}P_{X_j}(a_j)\\ &= \prod_{j=1}^{m}\sum_{a\in\mathcal{X}}P_{X_j}(a).\end{aligned}$$

A.2 Probability

- **Probability distribution** P_X on discrete set $\mathcal{X}$:

$$\forall x\in\mathcal{X}\colon\ \Pr(X=x) = P_X(x). \tag{A.10}$$

- **Probability density function** (pdf) p_X on real numbers $\mathbf{R}$:

$$\forall x\in\mathbf{R}\colon\ \Pr(X\le x) = \int_{-\infty}^{x}p_X(\tau)d\tau. \tag{A.11}$$

- **Markov's inequality, [38, Section 1.6.1]:** Let X be a non-negative random variable, i.e., $\Pr(X<0)=0$. Then for $a>0$

$$\Pr(X\ge a)\le\frac{\mathbb{E}(X)}{a}. \tag{A.12}$$

- **Moments:** Real-valued random variable X, positive integer k.

$$\mathsf{mgf}_X(r) = \mathbb{E}(e^{rX}) \tag{A.13}$$

$$\frac{\partial^k}{\partial r^k}\mathsf{mgf}_X(r)\bigg|_{r=0} = \mathbb{E}(X^k). \tag{A.14}$$

  $\mathsf{mgf}_X(r)$ is the moment generating function (MGF) of X and $\mathbb{E}(X^k)$ is the kth *moment* of X.

A.3 Information Theory

A.3.1 Types and Typical Sequences

Types Consider a sequence $x^n = x_1 x_2 \cdots x_n$ with entries in a finite alphabet $\mathcal{X}$. Let $N(a|x^n)$ be the number of times letter $a \in \mathcal{X}$ occurs in x^n, i.e.,

$$N(a|x^n) = \Big|\Big\{i \in \{1, 2, \ldots, n\} \colon x_i = a\Big\}\Big|, \quad a \in \mathcal{X}. \tag{A.15}$$

The empirical distribution of x^n is

$$P_{x^n}(a) = \frac{N(a|x^n)}{n}, \quad a \in \mathcal{X}. \tag{A.16}$$

Since every permutation of x^n has the same empirical distribution, we define $n_a = N(a|x^n)$ and write

$$P_X(a) = \frac{n_a}{n}, \quad a \in \mathcal{X}. \tag{A.17}$$

Note that every probability $P_X(a)$, $a \in \mathcal{X}$, is an integer multiple of $1/n$. The distribution P_X is therefore called an n-type. The set of all length n sequences with empirical distribution P_X is called the type class of the n-type P_X and denoted by $\mathcal{T}^n(P_X)$.

A.3.2 Differential Entropy

- **Differential entropy:**

$$\mathrm{h}(X) := \mathbb{E}[-\log_2 p_X(X)]. \tag{A.18}$$

- **Independence bound:**

$$\mathrm{h}(X, Y) \leq \mathrm{h}(X) + \mathrm{h}(Y). \tag{A.19}$$

A.3.3 Entropy

Random variable X with distribution P_X on finite set $\mathcal{X}$.

- **Entropy:**

$$\mathbb{H}(P_X) = \mathbb{H}(X) := \mathbb{E}[-\log_2 P_X(X)]. \tag{A.20}$$

- **Conditional Entropy, Equivocation:**

$$\mathbb{H}(P_{X|Y}|P_Y) = \mathbb{H}(X|Y) := \mathbb{E}[-\log_2 P_{X|Y}(X|Y)]. \tag{A.21}$$

- **Relation to differential entropy:** Properties (A.19) also hold for entropy.

- **Continuity:** Distributions P_X, $P_{X'}$ on finite set $\mathcal{X}$. Suppose $\|P_X - P_{X'}\|_1 = \delta \leq \frac{1}{2}$. Then

$$|\,\mathbb{H}(P_X) - \mathbb{H}(P_{X'})| \leq -\delta \log_2 \frac{\delta}{|\mathcal{X}|}. \tag{A.22}$$

- **Cross-Entropy:** P_X, Q_X distributions on $\mathcal{X}$.

$$\mathbb{X}(P_X\|Q_X) = \mathbb{E}[-\log_2 Q_X(X)]. \tag{A.23}$$

- **Information inequality:**

$$\mathbb{X}(P_X\|Q_X) \geq \mathbb{H}(P_X) \tag{A.24}$$

  with equality if and only if $Q_X = P_X$.

- **Cross-Equivocation:** $P_{X|Y}(\cdot|b)$ distribution on $\mathcal{X}$ for each $b \in \mathcal{Y}$. $Y \sim p_Y$.

  - $Q_{X|Y}(\cdot|b)$ distribution on $\mathcal{X}$ for each $b \in \mathcal{Y}$.

$$\mathbb{X}(P_{X|Y}\|Q_{X|Y}|p_Y) = \mathbb{E}[-\log_2 Q_{X|Y}(X|Y)]. \tag{A.25}$$

  - $q(\cdot,\cdot)$ non-negative function on $\mathcal{X} \times \mathcal{Y}$.

$$\mathsf{X}(q, X, Y) = \mathbb{E}\left[-\log_2 \frac{q(X,Y)}{\sum_{a\in\mathcal{X}} q(a,Y)}\right]. \tag{A.26}$$

A.3.4 Informational Divergence

- **Informational divergence:**

$$\mathbb{D}(p_X \| p_Y) := \mathbb{E}\left[\log_2 \frac{p_X(X)}{p_Y(X)}\right] \tag{A.27}$$

- **Information inequality:**

$$\mathbb{D}(p_X \| p_Y) \geq 0 \tag{A.28}$$

with equality if and only if $p_X = p_Y$.

A.3.5 Mutual Information

- **Mutual Information:**
  - X, Y continuous:

$$\begin{aligned}
\mathbb{I}(X;Y) &:= \mathbb{D}(p_{XY} \| p_X p_Y) && \text{(A.29)}\\
&= \mathbb{D}(p_{Y|X} \| p_Y | p_X) && \text{(A.30)}\\
&= \mathbb{D}(p_{X|Y} \| p_X | p_Y) && \text{(A.31)}\\
&= \mathrm{h}(Y) - \mathrm{h}(Y|X) && \text{(A.32)}\\
&= \mathrm{h}(X) - \mathrm{h}(X|Y). && \text{(A.33)}
\end{aligned}$$

  - X discrete, Y continuous:

$$\begin{aligned}
\mathbb{I}(X;Y) &:= \mathbb{D}(P_X p_{Y|X} \| P_X p_Y) && \text{(A.34)}\\
&= \mathbb{D}(P_{X|Y} \| P_X | p_Y) && \text{(A.35)}\\
&= \mathbb{D}(p_{Y|X} \| p_Y | P_X) && \text{(A.36)}\\
&= \mathrm{h}(Y) - \mathrm{h}(Y|X) && \text{(A.37)}\\
&= \mathbb{H}(X) - \mathbb{H}(X|Y). && \text{(A.38)}
\end{aligned}$$

  - Other combinations of discrete/continuous accordingly.

B

Acronyms

ASK amplitude shift keying

AWGN additive white Gaussian noise

BER bit error rate

BIACM bit-interleaver-agnostic coded modulation

BICM bit-interleaved coded modulation

BMD bit-metric decoding

BPSK binary phase shift keying

BRGC binary reflected Gray code

BSC binary symmetric channel

CCDM constant composition distribution matching

DM distribution matcher

DMS discrete memoryless source

FEC forward error correction

GHC geometric Huffman coding

GMI generalized mutual information

IACM interleaver-agnostic coded modulation

ID informational divergence

iid independent and identically-distributed

ILD invertible low-divergence

LDPC low-density parity-check

LUT lookup table

MAP maximum a posteriori probability

MB Maxwell-Boltzmann

MCDM minimum cost distribution matcher

MGF moment generating function

ML maximum-likelihood

PAS probabilistic amplitude shaping

PS probabilistic shaping

QAM quadrature amplitude modulation

SD soft decision

SE spectral efficiency

SNR signal-to-noise ratio

VD variational distance

WER word error rate

References

[1] R. A. Amjad, "Information rates and error exponents for probabilistic amplitude shaping," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Guangzhou, China, 2018.

[2] R. A. Amjad and G. Böcherer, "Fixed-to-variable length distribution matching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1511–1515, Istanbul, Turkey, 2013.

[3] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, 2009, pp. 3051–3073.

[4] S. Baur and G. Böcherer, "Arithmetic distribution matching," in *Proc. Int. ITG Conf. Syst. Commun. Coding (SCC)*, pp. 1–6, Hamburg, Germany, 2015.

[5] G. Böcherer, "Capacity-achieving probabilistic shaping for noisy and noiseless channels," Ph.D. dissertation, RWTH Aachen University, 2012. URL: http://www.georg-boecherer.de/capacityAchievingShaping.pdf.

[6] G. Böcherer, "Labeling non-square QAM constellations for one-dimensional bit-metric decoding," *IEEE Commun. Lett.*, vol. 18, no. 9, 2014, pp. 1515–1518.

[7] G. Böcherer, "Achievable rates for shaped bit-metric decoding," *arXiv preprint*, 2016. URL: http://arxiv.org/abs/1410.8075.

[8] G. Böcherer, "Principles of coded modulation," Habilitation thesis, Technical University of Munich, 2018. URL: http://www.georg-boecherer.de/bocherer2018principles.pdf.

[9] G. Böcherer and R. A. Amjad, "Fixed-to-variable length resolution coding for target distributions," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Seville, Spain, 2013.

[10] G. Böcherer and R. A. Amjad, "Informational divergence and entropy rate on rooted trees with probabilities," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 176–180, Honolulu, HI, USA, 2014.

[11] G. Böcherer and B. C. Geiger, "Optimal quantization for distribution synthesis," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, 2016, pp. 6162–6172.

[12] G. Böcherer, P. Schulte, and F. Steiner, "High throughput probabilistic shaping with product distribution matching," *arXiv preprint*, 2017. URL: http://arxiv.org/abs/1702.07510.

[13] G. Böcherer, P. Schulte, and F. Steiner, "Probabilistic shaping and forward error correction for fiber-optic communication systems," *J. Lightw. Technol.*, vol. 37, no. 2, 2019, pp. 230–244.

[14] G. Böcherer, F. Steiner, and P. Schulte, "Fast probabilistic shaping implementation for long-haul fiber-optic communication systems," in *Proc. Eur. Conf. Optical Commun. (ECOC)*, Gothenburg, Sweden, 2017.

[15] G. Böcherer, F. Steiner, and P. Schulte, "Fast probabilistic shaping implementation for long-haul fiber-optic communication systems," in *Proc. Eur. Conf. Optical Commun. (ECOC)*, Gothenburg, Sweden, 2017. URL: http://www.georg-boecherer.de/bocherer2017fast_slides.pdf.

[16] G. Böcherer, "Probabilistic signal shaping for bit-metric decoding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 431–435, Honolulu, HI, USA, 2014.

[17] G. Böcherer, "Lecture notes on variable length coding," 2016. URL: http://www.georg-boecherer.de/bocherer2016variable.pdf.

[18] G. Böcherer, "Integration of probabilistic shaping and forward error correction: Spectral efficiency, rate, overhead," in *CNRS/GdR ISIS Workshop on Coding, Modulation, and Signal Processing for Optical Communications, Telecom Paris Tech*, Paris, France, 2019. URL: http://www.georg-boecherer.de/bocherer2019integration_slides.pdf.

[19] G. Böcherer, "Achievable rates for probabilistic shaping," *arXiv preprint*, URL: https://arxiv.org/abs/1707.01134v5.

[20] G. Böcherer, F. Diedolo, and F. Pittala, "Label extension for 32QAM: The extra bit for a better FEC performance-complexity tradeoff," in *Proc. Eur. Conf. Optical Commun. (ECOC)*, Brussels, Belgium, 2020.

[21] G. Böcherer, D. Lentner, A. Cirino, and F. Steiner, "Probabilistic parity shaping for linear codes," *arXiv preprint*, 2019. URL: https://arxiv.org/abs/1902.10648.

[22] G. Böcherer and R. Mathar, "Matching dyadic distributions to channels," in *Proc. Data Compression Conf. (DCC)*, pp. 23–32, Snowbird, UT, USA, 2011.

[23] G. Böcherer, P. Schulte, and F. Steiner, "Probabilistic shaping: A random coding experiment," in *Proc. Int. Zurich Seminar Commun.*, ETH Zurich, pp. 12–14, 2020.

[24] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Trans. Commun.*, vol. 63, no. 12, 2015, pp. 4651–4665.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[26] F. Buchali, F. Steiner, G. Böcherer, L. Schmalen, P. Schulte, and W. Idler, "Rate adaptation and reach increase by probabilistically shaped 64-QAM: An experimental demonstration," *J. Lightw. Technol.*, vol. 34, no. 8, 2016.

[27] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, 1998, pp. 927–946.

[28] J. Cho, L. Schmalen, and P. J. Winzer, "Normalized generalized mutual information as a forward error correction threshold for probabilistically shaped QAM," in *Proc. Eur. Conf. Optical Commun. (ECOC)*, Gothenburg, Sweden, 2017.

[29] J. Cho, "Prefix-free code distribution matching for probabilistic constellation shaping," *IEEE Trans. Commun.*, vol. 68, no. 2, 2020, pp. 670–682.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., 2006.

[31] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[32] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Found. Trends Comm. Inf. Theory*, vol. 1, no. 4, 2004, pp. 417–528.

[33] M. Dia, V. Aref, and L. Schmalen, "A compressed sensing approach for distribution matching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1266–1270, Vail, Colorado, USA, 2018.

[34] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, and K. Parsons, "Multiset-partition distribution matching," *IEEE Trans. Commun.*, vol. 67, no. 3, 2019, pp. 1885–1893.

[35] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, and K. Parsons, "Parallel-amplitude architecture and subset ranking for fast distribution matching," *IEEE Trans. Commun.*, vol. 68, no. 4, 2020, pp. 1981–1990.

[36] W. Feller, *An Introduction to Probability Theory and Its Applications, Volume I*. John Wiley & Sons, Inc, 1968.

[37] R. G. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, 1962, pp. 21–28.

[38] R. G. Gallager, *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

[39] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., 1968.

[40] A. Ganti, A. Lapidoth, and E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, 2000, pp. 2315–2328.

[41] B. C. Geiger and G. Böcherer, "Greedy algorithms for optimal distribution approximation," *Entropy*, vol. 18, no. 7, 2016, pp. 1–10. DOI: 10.3390/e18070262.

[42] Y. C. Gültekin, "Enumerative sphere shaping techniques for short blocklength wireless communications," Ph.D. dissertation, Technische Universiteit Eindhoven, 2020.

[43] Y. C. Gültekin, A. Alvarado, and F. M. Willems, "Achievable information rates for probabilistic amplitude shaping: An alternative approach via random sign-coding arguments," *Entropy*, vol. 22, no. 7, 2020.

[44] Y. C. Gültekin, T. Fehenberger, A. Alvarado, and F. M. Willems, "Probabilistic shaping for finite blocklengths: Distribution matching and sphere shaping," *Entropy*, vol. 22, no. 5, 2020, p. 581.

[45] Y. C. Gültekin, W. J. van Houtum, S. Şerbetli, and F. M. Willems, "Constellation shaping for IEEE 802.11," in *Proc. IEEE Int. Symp. Personal, Indoor, Mobile Radio Commun. (PIMRC)*, Montreal, Quebec, Canada, 2017.

[46] Y. C. Gültekin, F. M. Willems, W. van Houtum, and S. Serbetli, "Approximate enumerative sphere shaping," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 676–680, Vail, Colorado, USA, 2018.

[47] C. Häger, A. Graell i Amat, F. Brännström, A. Alvarado, and E. Agrell, "Improving soft FEC performance for higher-order modulations via optimized bit channel mappings," *Optics Express*, vol. 22, no. 12, 2014, pp. 14 544–14 558.

[48] J. Honda and H. Yamamoto, "Polar coding without alphabet extension for asymmetric models," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, 2013, pp. 7829–7838.

[49] G. Kaplan and S. Shamai (Shitz), "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *AEÜ*, vol. 47, no. 4, 1993, pp. 228–239.

[50] G. Kramer, "Divergence scaling for distribution matching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1159–1163, Melbourne, Australia, 2021.

[51] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, 1993, pp. 913–929.

[52] A. Lempel, S. Even, and M. Cohn, "An algorithm for optimal prefix parsing of a noiseless and memoryless channel," *IEEE Trans. Inf. Theory*, vol. 19, no. 2, 1973, pp. 208–214.

[53] Y. Lomnitz and M. Feder, "A simpler derivation of the coding theorem," *arXiv preprint arXiv:1205.1389*, 2012.

[54] D. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, 1999, pp. 399–431.

[55] A. Martinez, A. Guillén i Fàbregas, G. Caire, and F. Willems, "Bit-interleaved coded modulation revisited: A mismatched decoding perspective," *IEEE Trans. Inf. Theory*, vol. 55, no. 6, 2009, pp. 2756–2765.

[56] N. Merhav and G. Böcherer, "Codebook mismatch can be fully compensated by mismatched decoding," *IEEE Trans. Inf. Theory*, vol. 69, no. 4, 2023, pp. 2152–2164.

[57] M. Pikus, "Finite-precision and multi-stream distribution matching," Ph.D. dissertation, Technical University of Munich, 2019.

[58] M. Pikus and W. Xu, "Bit-level probabilistically shaped coded modulation," *IEEE Commun. Lett.*, vol. 21, no. 9, 2017, pp. 1929–1932.

[59] T. Prinz, P. Yuan, G. Böcherer, F. Steiner, O. Iscan, R. Böhnke, and W. Xu, "Polar coded probabilistic amplitude shaping for short packets," in *IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, Sapporo, Japan, 2017.

[60] T. V. Ramabadran, "A coding scheme for m-out-of-n codes," *IEEE Trans. Commun.*, vol. 38, no. 8, 1990, pp. 1156–1163.

[61] E. A. Ratzer, "Error-correction on non-standard communication channels," Ph.D. dissertation, University of Cambridge, 2003.

[62] T. J. Richardson, M. A. Shokrollahi, and R. L. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, 2001, pp. 619–637.

[63] C. Runge, T. Wiegart, and D. Lentner, "Improved list decoding for polar-coded probabilistic shaping," *arXiv preprint*, 2023. URL: https://arxiv.org/abs/2305.07962v1.

[64] M. Scholten, T. Coe, and J. Dillard, "Continuously-interleaved BCH (CI-BCH) FEC delivers best in class NECG for 40G and 100G metro applications," in *National Fiber Optic Engineers Conference*, NTuB3, 2010.

[65] P. Schulte and G. Böcherer, "Constant composition distribution matching," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, 2016, pp. 430–434.

[66] P. Schulte, "Algorithms for distribution matching," Ph.D. dissertation, Technische Universität München, 2020.

[67] P. Schulte, R. A. Amjad, T. Wiegart, and G. Kramer, "Invertible low-divergence coding," *IEEE Trans. Inf. Theory*, vol. 68, no. 1, 2021, pp. 178–192.

[68] P. Schulte and B. C. Geiger, "Divergence scaling of fixed-length, binary-output, one-to-one distribution matching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 3075–3079, Aachen, Germany, 2017.

[69] P. Schulte and F. Steiner, "Divergence-optimal fixed-to-fixed length distribution matching with shell mapping," *IEEE Wireless Commun. Letters*, vol. 8, no. 2, 2019, pp. 620–623.

[70] P. Schulte, F. Steiner, and G. Böcherer, "Four dimensional probabilistic shaping for fiber-optic communication," in *Proc. Signal Process. Photonic Commun. (SPPCOM)*, New Orleans, Louisiana, USA, 2017.

[71] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, 1948, 379–423 and 623–656.

[72] A. Sheikh, A. Graell i Amat, and A. Alvarado, "On product codes with probabilistic amplitude shaping for high-throughput fiber-optic systems," *IEEE Commun. Lett.*, vol. 24, no. 11, 2020, pp. 2406–2410.

[73] A. Sheikh, A. Graell i Amat, G. Liva, and F. Steiner, "Probabilistic amplitude shaping with hard decision decoding and staircase codes," *J. Lightw. Technol.*, vol. 36, no. 9, 2018, pp. 1689–1697.

[74] M. A. Sluyski, *Open ROADM MSA 3.01 W-port digital specification (200G-400G)*, 2019. URL: https://tinyurl.com/openroadm.

[75] B. P. Smith, A. Farhood, A. Hunt, F. R. Kschischang, and J. Lodge, "Staircase codes: FEC for 100 Gb/s OTN," *J. Lightw. Technol.*, vol. 30, no. 1, 2012, pp. 110–117.

[76] F. Steiner, G. Böcherer, and G. Liva, "Protograph-based LDPC code design for shaped bit-metric decoding," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 2, 2016, pp. 397–407.

[77] F. Steiner, "Coding for higher-order modulation and probabilistic shaping," Ph.D. dissertation, Technical University of Munich, 2020.

[78] N. Stolte, "Recursive codes with the Plotkin construction and their decoding," Ph.D. dissertation, Technical University of Darmstadt, 2002.

[79] A. Y. Sukmadji, U. Martínez-Peñas, and F. R. Kschischang, "Zipper codes," *J. Lightw. Technol.*, vol. 40, no. 19, 2022, pp. 6397–6407.

[80] T. Yoshida, M. Karlsson, and E. Agrell, "Performance Metrics for Systems With Soft-Decision FEC and Probabilistic Shaping," *IEEE Photon. Technol. Lett.*, vol. 29, no. 23, 2017, pp. 2111–2114.

[81] T. Yoshida, M. Karlsson, and E. Agrell, "Hierarchical distribution matching for probabilistically shaped coded modulation," *J. Lightw. Technol.*, vol. 37, no. 6, 2019, pp. 1579–1589.