Three Open Problems in Probabilistic Shaping

Georg Böcherer
Huawei Technologies

April 4, 2024
Munich Workshop on Shannon Coding Technologies

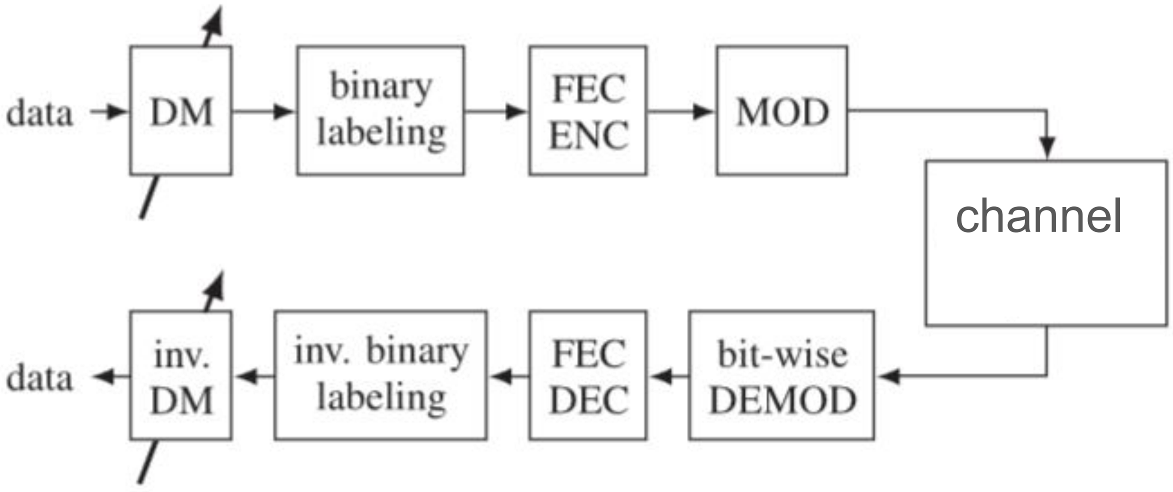
Outline

- Probabilistic Amplitude Shaping
- An Optimization Problem
- An Information Theoretic Problem
- A Machine Learning Problem

Outline

- **Probabilistic Amplitude Shaping**
- An Optimization Problem
- An Information Theoretic Problem
- A Machine Learning Problem

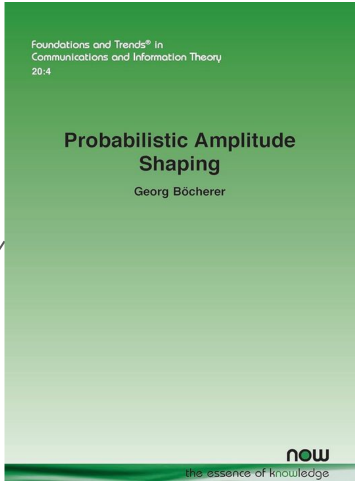
Probabilistic Amplitude Shaping

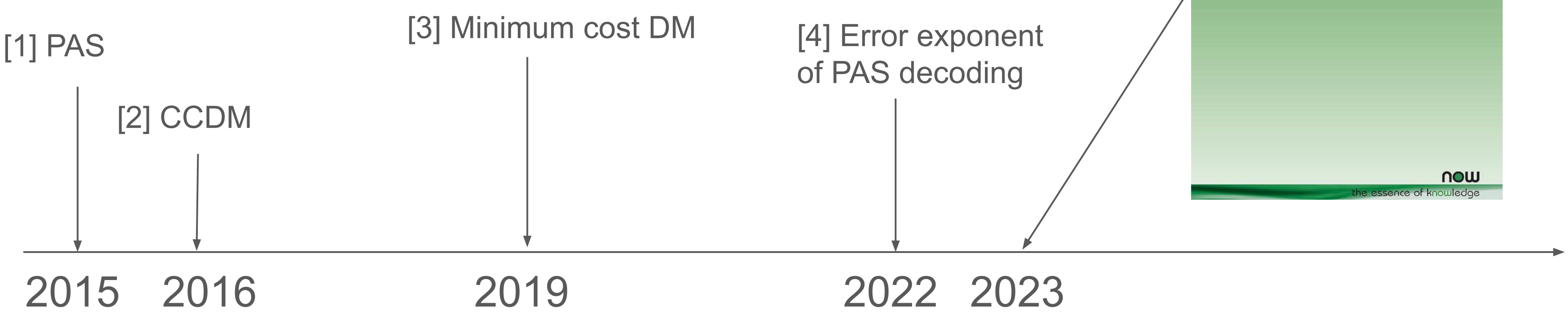


- **PAS encoding:** systematic linear FEC encoder preserves amplitude distribution **imposed by DM**.
- **PAS decoding:** linear FEC decoder on full linear code, **agnostic of DM**.

This talk focuses on theoretic aspects of PAS

Probabilistic Amplitude Shaping
Theoretical advances





[1] 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, pp. 4651–4665, Dec. 2015

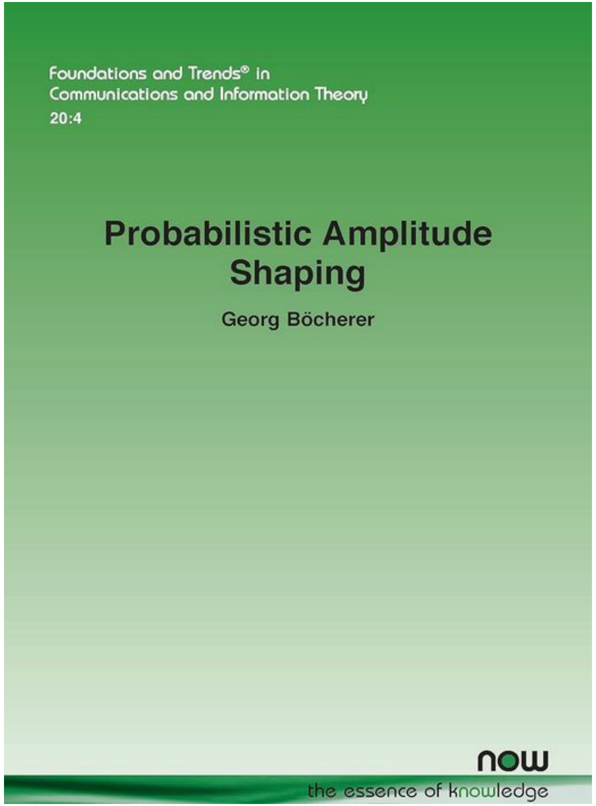
[2] P. Schulte and G. Böcherer, "Constant composition distribution matching," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 430–434, Jan. 2016

[3] P. Schulte and F. Steiner, "Divergence-optimal fixed-to-fixed length distribution matching with shell mapping," *IEEE Wireless Commun. Letters*, vol. 8, no. 2, pp. 620–623, Apr. 2019

[4] N. Merhav and G. Böcherer, "Codebook mismatch can be fully compensated by mismatched decoding," *IEEE Trans. Inf. Theory*, vol. 69, no. 4, pp. 2152–2164, Apr. 2023

Probabilistic Amplitude Shaping

https://github.com/gbsha/PAS



Outline

- Probabilistic Amplitude Shaping
- **An Optimization Problem**
- An Information Theoretic Problem
- A Machine Learning Problem

Setup

- 2^m-ASK constellation

$$X \in \{\pm 1, \pm 3, \ldots, \pm(2^m - 1)\}$$

- AWGN channel

$$Y = X + \sigma Z, \quad Z \sim \mathcal{N}(0, 1)$$

- Signal-to-noise ratio

$$\mathsf{SNR} = \frac{\mathbb{E}(X^2)}{\sigma^2}$$

- FEC rate

$$0 \leq R_{\text{fec}} \leq 1$$

PAS Achievable Rate

- Maxwell-Boltzmann (MB) distribution

$$P_X(x) \propto \exp(-\nu|x|^2)$$

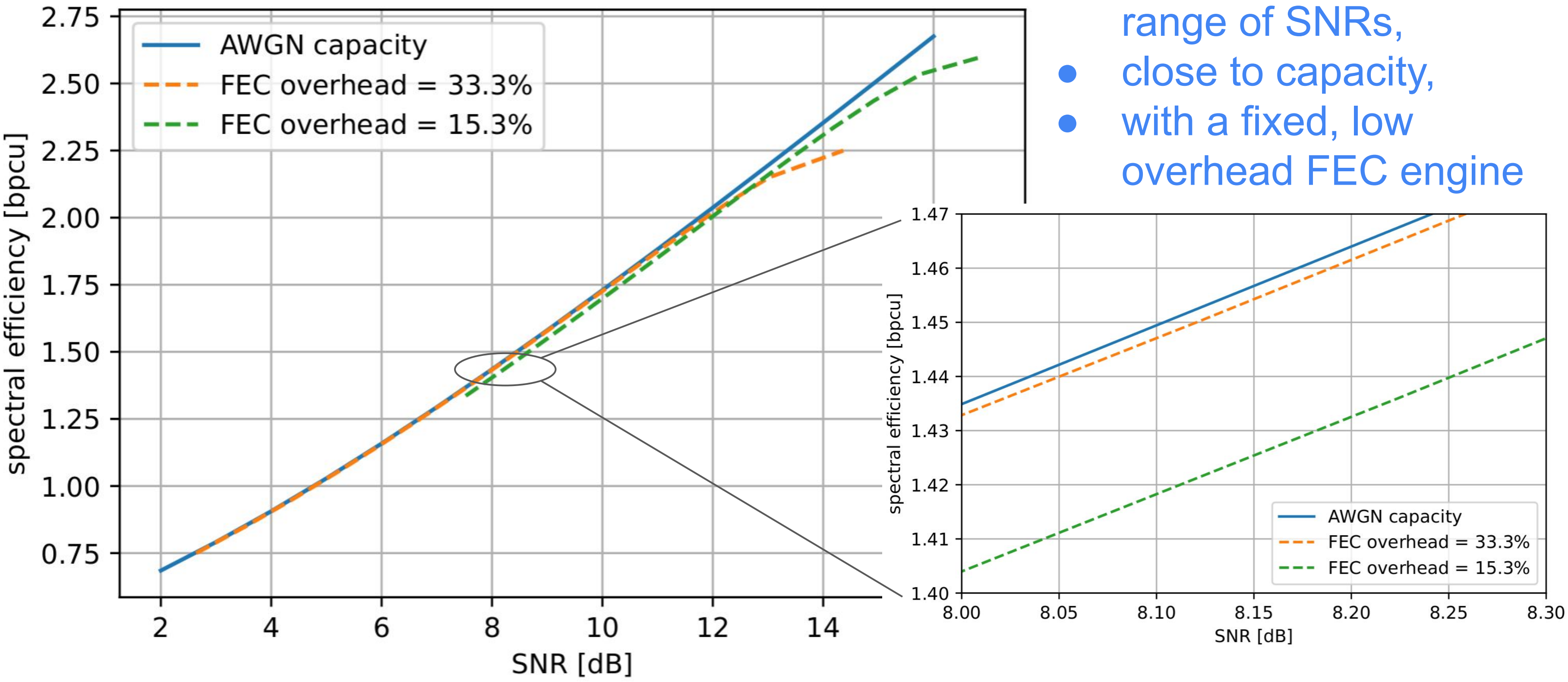
- Spectral efficiency

$$\mathrm{SE} = \mathbb{H}(X) - m(1 - R_{\mathrm{fec}})$$

- By [5, Example 5.5], an achievable noise level is

$$\sigma \colon \mathbb{H}(X|X + \sigma Z) = m(1 - R_{\mathrm{fec}})$$

PAS Achievable Rate



Killer application in optical communications:

- Operate over a large range of SNRs,
- close to capacity,
- with a fixed, low overhead FEC engine

[5, Sec. 1.4.5]

What if we want to optimize P_X?

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

The Lagrangian is

$$L(P_X, \lambda) = \mathbb{H}(X) - \lambda \mathbb{H}(X|Y)$$

- For $\lambda = 1$: $L(P_X, \lambda) = \mathbb{I}(X; Y) \Rightarrow$ convex $\cap$ in P_X.
- For large enough $\lambda > 1$, $L(P_X, \lambda)$ becomes **non-convex**

Problem 1: Find algorithm to calculate optimal P_X

Remarks on Problem 1

- The solution may assign probability 0 to some signal points.
- Enhancing the peak power constrained problem in [6] by a constraint on the equivocation may be a viable approach.

[6] J. G. Smith, "The information capacity of amplitude-and variance-constrained scalar Gaussian channels," *Inf. control*, vol. 18, no. 3, pp. 203–219, 1971

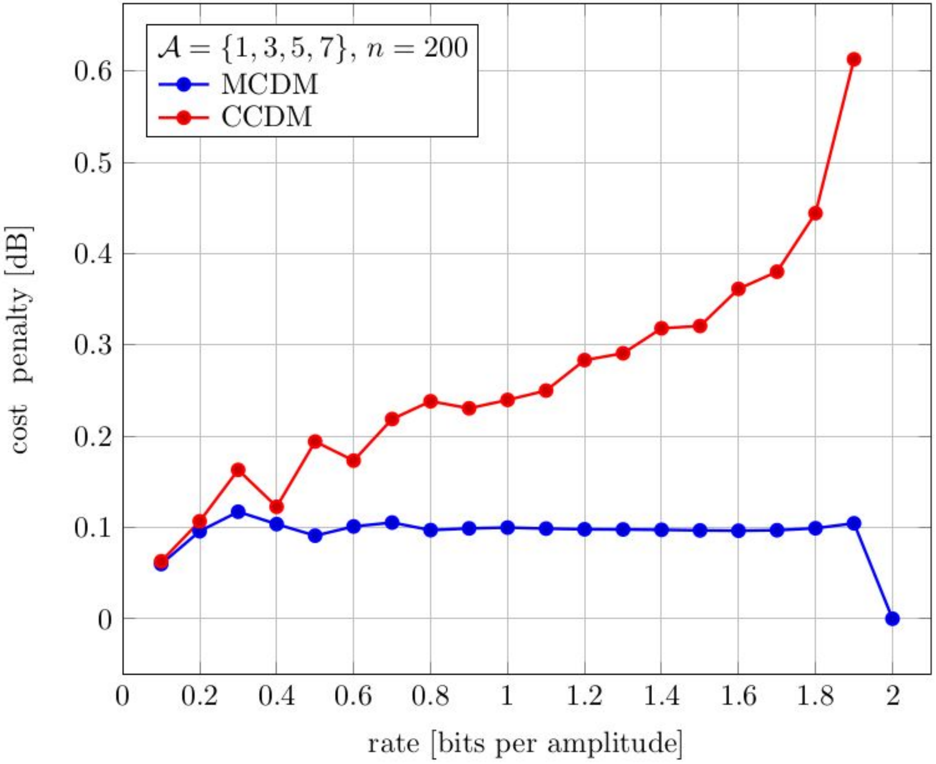
Outline

- Probabilistic Amplitude Shaping
- An Optimization Problem
- **An Information Theoretic Problem**
- A Machine Learning Problem

Distribution Matching [5, Chapter 2]

- Rate $R = k/n$
- Cost $\mathbb{E}[w(\bar{A})]$
- MB source A^* with $\mathbb{H}(A^*) = R$

$$\text{cost penalty} = 10\log_{10}\frac{\mathbb{E}[w(\bar{A})]}{\mathbb{E}[w(A^*)]}$$



Remark: The CCDM distribution (“type”) is chosen in [5] according to the minimum cost for which the required rate is achieved. This results in lower cost than starting from an MB distribution and quantizing it to an n-type. The minimum cost DM (MCDM) indexes the 2^k least cost length n sequences. Thus, both for CCDM and MCDM, the distribution (averaged both over the sequences and the entries of the sequences) has only a subordinate role. Further, in practice, the cost penalty has proven to be a more relevant metric for DM comparison than rate loss or relative entropy, for two reasons: (1) the MB distribution itself results from minimizing cost; so directly minimizing cost is more meaningful than approximating an MB distribution. (2) rate is a system parameter, which usually has to be realized exactly.

PAS Channel Coding Theorem

- By [5, Theorem 5.3], asymptotically in FEC length n, PAS can achieve

$$\mathsf{SE}^* = R_{\mathsf{ccdm}}(P_{|X|}, n_{\mathsf{dm}}) + 1 - \mathbb{H}(X|Y)$$

  with sign(X) uniform on $\{-1, 1\}$.

- For $n_{\mathsf{dm}} \to \infty$,

$$\mathsf{SE}^* = \mathbb{H}(X) - \mathbb{H}(X|Y)$$

  which is the best we can do.

- However, for short DM length, we know empirically that MCDM is much better than CCDM. **Example:** 800G: FEC decoding window $> 100\,000$, $n_{\mathsf{dm}} = 12$.

Problem 2: Prove PAS channel coding theorem for MCDM

Comments on Problem 2

- [4] proves that PAS decoding of linear codes achieves the best known error exponent for constant composition codes [7, Theorem 10.2].
- [8] proves a PAS channel coding theorem where the DM is replaced by a discrete memoryless source.

[4] N. Merhav and G. Böcherer, "Codebook mismatch can be fully compensated by mismatched decoding," *IEEE Trans. Inf. Theory*, vol. 69, no. 4, pp. 2152–2164, Apr. 2023

[7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011

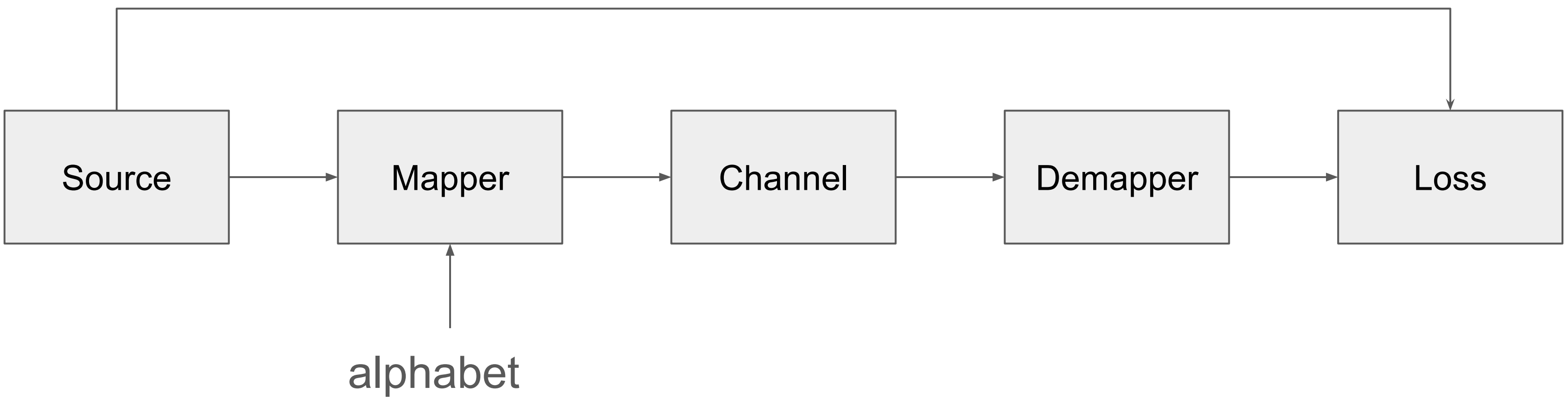
[8] R. A. Amjad, "Information rates and error exponents for probabilistic amplitude shaping," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Guangzhou, China, Nov. 2018

Outline

- Probabilistic Amplitude Shaping
- An Optimization Problem
- An Information Theoretic Problem
- **A Machine Learning Problem**

Data Driven ("Machine Learning") Geometric Shaping

- Geometric shaping is the vanilla case of **data driven** transmitter optimization.
- Loss function: crossequivocation ("crossentropy") $\mathbb{E}[-\log_2 Q(X|Y)] \geq \mathbb{H}(X|Y)$.
- Differential channel simulator.
- Backpropagate loss gradient to alphabet.



Data Driven Probabilistic Shaping

- Loss function: crossequivocation is minimized by Dirac delta P_X
  $\Rightarrow$ regularize, e.g., by $\mathbb{H}(X)$.
- How to define a differentiable estimator of $\mathbb{H}(X)$?
- How to define a differentiable source?

Problem 3: Data driven optimization of input distribution P_X

Remarks on Problem 3

- For differential entropy, a differentiable estimator exists [9].
  $\Rightarrow$ estimate density in x_i by distance to closest neighbor.
  - I use it in [10, Chapter 10] for stochastic channel modelling.
- There may be no solution, but why?

[9] J. Jiao, W. Gao, and Y. Han, "The nearest neighbor information estimator is adaptively near minimax rate-optimal," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 3160–3171

[10] G. Böcherer, *Lecture notes on machine learning for communications*, [Online]. Available: `http://georg-boecherer.de/mlcomm`

Summary

Practical PAS is well understood, theoretical understanding is still incomplete.

We lack

- Algorithm for optimizing input distribution **under constrained FEC overhead**.
- Channel coding theorem for **minimum cost DM** (optimal DM?)
- Better understanding difficulty of data-driven input distribution optimization.

References I

[1] 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, pp. 4651–4665, Dec. 2015.

[2] P. Schulte and G. Böcherer, "Constant composition distribution matching," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 430–434, Jan. 2016.

[3] P. Schulte and F. Steiner, "Divergence-optimal fixed-to-fixed length distribution matching with shell mapping," *IEEE Wireless Commun. Letters*, vol. 8, no. 2, pp. 620–623, Apr. 2019.

[4] N. Merhav and G. Böcherer, "Codebook mismatch can be fully compensated by mismatched decoding," *IEEE Trans. Inf. Theory*, vol. 69, no. 4, pp. 2152–2164, Apr. 2023.

References II

[5] G. Böcherer, "Probabilistic amplitude shaping," *Foundations and Trends® in Communications and Information Theory*, vol. 20, no. 4, pp. 390–511, 2023, ISSN: 1567-2190. DOI: 10.1561/0100000111. [Online]. Available: http://dx.doi.org/10.1561/0100000111.

[6] J. G. Smith, "The information capacity of amplitude-and variance-constrained scalar Gaussian channels," *Inf. control*, vol. 18, no. 3, pp. 203–219, 1971.

[7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[8] R. A. Amjad, "Information rates and error exponents for probabilistic amplitude shaping," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Guangzhou, China, Nov. 2018.

[9] J. Jiao, W. Gao, and Y. Han, "The nearest neighbor information estimator is adaptively near minimax rate-optimal," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 3160–3171.

References III

[10] G. Böcherer, *Lecture notes on machine learning for communications*, [Online]. Available: http://georg-boecherer.de/mlcomm.