Probabilistic Shaping: A Random Coding Experiment

Georg Böcherer
Optical Communications Technology Lab
Huawei France Technologies S.A.S.U.
Boulogne-Billancourt 92100, France
Email: georg.boecherer@ieee.org

Patrick Schulte and Fabian Steiner
Technical University of Munich
Institute for Communications Engineering
80333 München, Germany
Email: {patrick.schulte, fabian.steiner}@tum.de

***Abstract*—A layered probabilistic shaping (PS) ensemble is considered, which contains probabilistic amplitude shaping (PAS) as a practical instance. Layered PS consists of an inner layer for forward error correction (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, the probability of encoding failure and decoding failure is analyzed and it is shown that the layered PS architecture is capacity-achieving for a discrete input memoryless channel. Practical achievable spectral efficiencies of the layered PS architecture are discussed.**

I. Introduction

Probabilistic amplitude shaping (PAS) was proposed in [1] to integrate non-uniform channel input distributions with off-the-shelf linear forward error correction (FEC) codes. PAS quickly found industrial application in transceivers for fiber-optic transmission, e.g., [2]–[4]. Since PAS is not a sample of the classical random code ensemble (see Remarks 1, 2, and 3), the calculation of appropriate achievable rates for PAS is intricate, and several attempts were taken [2, Sec. III.C], [5], [6]. In [7] and [8, Chap. 10], achievable rates for PAS are derived using random sign coding and partially systematic FEC encoding. In this work, we discuss layered probabilistic shaping (PS), a random code ensemble that was developed in the line of work [8]–[11]. Layered PS contains PAS as a practical instance, but is more general, e.g., it also covers the probabilistic parity bit shaping proposed in [12].

In Sec. II, we define layered PS and derive a general channel coding theorem. In Sec. III, we show that layered PS achieves the capacity of discrete input memoryless channels and discuss practical matched and mismatched decoding metrics.

II. Layered Probabilistic Shaping

Consider a channel with finite input alphabet $\mathcal{X}$ and define

$$m = \log_2|\mathcal{X}|. \tag{1}$$

The channel output alphabet can be continuous or discrete.

A. Classical Random Code Ensemble

The classical random code ensemble [13, Ch. 5] for a channel with input alphabet $\mathcal{X}$ and codeword length n symbols in $\mathcal{X}$ is

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

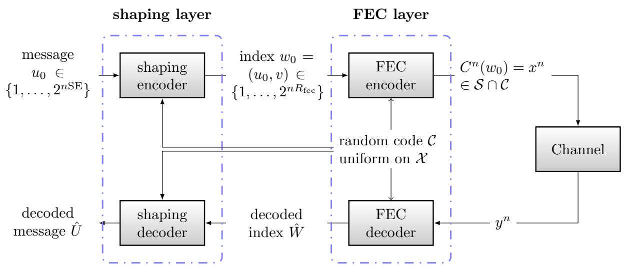

Fig. 1. The layered PS architecture discussed in Sec. II. In PAS [1], the FEC encoder is systematic and the shaping encoder is realized by a DM [14]–[19] that shapes the systematic symbols. The shaping encoder of PAS is zero error.

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 [13, Eq. (5.2.5)], the decoding rule for a memoryless channel with transition density $p_{Y|X}$ is

$$\hat{w} = \underset{w\in\{1,\dots,|\mathcal{C}|\}}{\operatorname{argmax}} \prod_{i=1}^{n} p_{Y|X}(y_i|c_i(w)) \tag{3}$$

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

$$\mathrm{SE}^* = \mathbb{I}(X;Y). \tag{4}$$

In particular, it achieves the capacity $\max_{P_X}\mathbb{I}(X;Y)$ when the optimal P_X is used.

B. Layered Random Code Ensemble

The layered PS architecture is displayed in Fig. 1. We consider the random code ensemble

$$\mathcal{C} = \left\{C^n(w), w = 1, 2, \dots, 2^{nmR_{\text{fec}}}\right\} \tag{5}$$

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 1. Note that the classical random code ensemble of Sec. II-A samples the codeword entries according to the desired channel input distribution P_X. In contrast, layered PS always uses the uniform distribution.

TABLE I
PS AND FEC OVERHEADS

	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 %	$100 \cdot \left(\frac{1}{R_{\text{ss}}+R_{\text{fec}}-1} - 1\right)$	

C. Encoding

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

$$R_{\text{ss}} = \frac{\log_2 |\mathcal{S}|}{nm}. \tag{6}$$

Note that by the definition of m in (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

$$\frac{2^{nmR_{\text{fec}}}}{2^{n\text{SE}}} = 2^{nm(R_{\text{fec}} - \frac{\text{SE}}{m})}. \tag{7}$$

The PS encoder maps message $u \in \{1, 2, \dots, 2^{n\text{SE}}\}$ 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 $v \in \{1, 2, \dots, 2^{nm(R_{\text{fec}} - \frac{\text{SE}}{m})}\}$. An encoding error occurs if the PS encoder cannot find such a codeword.

Theorem 1 ([11, Theorem 1])**.** *The probability that the PS encoder cannot map its input to a codeword in $\mathcal{S} \cap \mathcal{C}$ is upper bounded by*

$$\begin{aligned}&\Pr(\text{PS encoding failure})\\ &\quad\leq \exp\left(-2^{nm\left[1-(1-R_{\text{ss}})-(1-R_{\text{fec}})-\frac{\text{SE}}{m}\right]}\right).\end{aligned} \tag{8}$$

Remark 2. By the theorem, the SE is determined by two overheads (see Table I), namely the PS overhead and the FEC overhead. 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 for complexity reasons. Note that in the classical random coding experiment, the SE is always equal to mR_{fec}.

D. 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 $q(x^n, y^n)$ (see [11, Sec. V.A] for the definition and detailed discussion of non-negative scores). 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,

$$\hat{w} = \underset{w \in \{1,\dots,|\mathcal{C}|\}}{\operatorname{argmax}} \; q\left(c^n(w), y^n\right). \tag{9}$$

Theorem 2 ([11, Theorem 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-entropy*

$$\mathsf{x}(q, x^n, y^n) = -\frac{1}{n} \log_2 \frac{q(x^n, y^n)}{\sum_{a^n \in \mathcal{X}^n} q(a^n, y^n)}. \tag{10}$$

The probability that the decoder (9) *does not recover the index w_0 from the sequence y^n is bounded from above by*

$$\begin{aligned}&\Pr(\hat{W} \neq w_0 | C^n(w_0) = x^n, Y^n = y^n)\\ &\qquad\leq 2^{-nm\left(1 - R_{\text{fec}} - \frac{\mathsf{x}(q, x^n, y^n)}{m}\right)}.\end{aligned} \tag{11}$$

Note that in Fig. 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 $\Pr(\hat{U} \neq u_0)$.

E. Channel Coding Theorem

We now consider a memoryless channel

$$p_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i) \tag{12}$$

and memoryless decoding metrics

$$q(x^n, y^n) = \prod_{i=1}^{n} q(x_i, y_i). \tag{13}$$

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

$$\mathsf{x}(q, X^n, Y^n) \approx \mathbb{E}\left[\mathsf{x}(q, X, Y)\right] =: \mathbb{X}(q, X, Y) \tag{14}$$

where $\mathbb{X}(q, X, Y)$ is a cross-entropy. By Theorems 1 and 2, following the line of arguments in [20] (leaving out the ϵs and δs) we arrive at the following channel coding theorem.

Corollary 1. *For a shaping set with distribution P_X, an achievable spectral efficiency allowing for successful encoding and decoding with high probability is*

$$\text{SE}^* = \left[mR_{\text{ss}} - \mathbb{X}(q, X, Y)\right]^+ \tag{15}$$

where $[\cdot]^+ = \max\{0, \cdot\}$ ensures non-negativity.

Note that (15) is the same as [11, Eq. (1)] with slightly different notation.

III. DECODING METRICS

We now instantiate the achievable SE in (15) for various shaping sets and decoding metrics. See Table II for an overview.

A. Capacity-Achieving Symbol-Metric

We use as shaping set $\mathcal{S}$ all sequences with distribution P_X. For sufficiently large n, we have $R_{\text{ss}} m \approx \mathbb{H}(X)$. With the decoding metric $P_{X|Y}$, 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. Note that the classical random code ensemble achieves capacity with a maximum likelihood (ML) rule on a codebook of size $2^{n\text{SE}}$ while layered PS achieves capacity with a maximum a posteriori (MAP) rule on a codebook of size $2^{n(\text{SE}+m(1-R_{\text{ss}}))}$, which is larger.

TABLE II
IMPORTANT DECODING METRICS

	mR_{ss}	q	$\mathbb{X}^*(q,X,Y)$	SE*
symbol-metric	mR_{ss}	$P_{X\|Y}$	$\mathbb{H}(X\|Y)$	$[mR_{\text{ss}} - \mathbb{H}(X\|Y)]^+$
+ capacity-achieving	$\mathbb{H}(X)$	$P_{X\|Y}$	$\mathbb{H}(X\|Y)$	$\mathbb{I}(X;Y)$
bit-metric	mR_{ss}	$\prod_{i=1}^m P_{B_i\|Y}$	$\sum_{i=1}^m \mathbb{H}(B_i\|Y)$	$\left[mR_{\text{ss}} - \sum_{i=1}^m \mathbb{H}(B_i\|Y)\right]^+$
	$\mathbb{H}(X)$	$\prod_{i=1}^m P_{B_i\|Y}$	$\sum_{i=1}^m \mathbb{H}(B_i\|Y)$	$\left[\mathbb{H}(X) - \sum_{i=1}^m \mathbb{H}(B_i\|Y)\right]^+$
mismatched metric	mR_{ss}	q	$\min_{s>0} \mathbb{X}(q^s,X,Y)$	$\max_{s>0} [mR_{\text{ss}} - \mathbb{X}(q^s,X,Y)]^+$

B. Bit-Metric

Bit metric decoding uses an m-bit label $\boldsymbol{B} = B_1B_2\dots B_m$ of the channel input alphabet and a bit-metric

$$q(\boldsymbol{b},y) = \prod_{i=1}^{m} q_i(b_i,y). \tag{16}$$

Table II shows achievable SEs when $q_i = P_{B_i|Y}$. By defining the L-value $L_i = \log P_{B_i|Y}(0|Y)/P_{B_i|Y}(1|Y)$, the conditional entropy sum can also be written as

$$\sum_{i=1}^{m} \mathbb{H}(B_i|Y) = \sum_{i=1}^{m} \mathbb{E}\left[\log_2\{1+\exp\left[-(1-2B_i)L_i\right]\}\right]. \tag{17}$$

C. Mismatched Metrics

For $s > 0$, the non-negative metric q and the metric q^s implement exactly the same decision rule. Consequently, their error probability is the same. This allows us to tighten the error bound in Theorem 2 and thereby the achievable SE in Corollary 1. The tightened cross-entropy is

$$\mathbb{X}^*(q,X,Y) = \min_{s>0} \mathbb{X}(q^s,X,Y). \tag{18}$$

For uniform distributions P_X, the mismatched achievable SE recovers the generalized mutual information (GMI) in [21]. For non-uniform P_X, it is different from the GMI, because in [21], the classical random code ensemble of Sec. II-A is considered.

IV. CONCLUSIONS

We defined layered probabilistic shaping (PS) and derived achievable rates. In particular, we showed that layered PS is capacity-achieving for a particular shaping sets and decoding metrics. Several differences between layered PS and the classical random code ensemble were pointed out. The achievable rates of layered PS are directly applicable for probabilistic amplitude shaping (PAS). An 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.

REFERENCES

[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] 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, Apr. 2016.

[3] A. Ghazisaeidi, I. F. de Jauregui Ruiz, R. Rios-Müller, L. Schmalen, P. Tran, P. Brindel, A. C. Meseguer, Q. Hu, F. Buchali, G. Charlet, *et al.*, "Advanced C+L-band transoceanic transmission systems based on probabilistically shaped PDM-64QAM," *J. Lightw. Technol.*, vol. 35, no. 7, pp. 1291–1299, Apr. 2017.

[4] J. Li, A. Zhang, C. Zhang, X. Huo, Q. Yang, J. Wang, J. Wang, W. Qu, Y. Wang, J. Zhang, *et al.*, "Field trial of probabilistic-shaping-programmable real-time 200-Gb/s coherent transceivers in an intelligent core optical network," in *Asia Commun. Photonics Conf. (ACP)*, 2018.

[5] 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)*, Paper M.2.D.2, Gothenburg, Sweden, Sep. 2017.

[6] 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, pp. 2111–2114, Dec. 2017.

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

[8] G. Böcherer, *Principles of coded modulation*, Habilitation thesis, Technical University of Munich, 2018. [Online]. Available: http://www.georg-boecherer.de/bocherer2018principles.pdf.

[9] ——, "Achievable rates for shaped bit-metric decoding," *arXiv preprint*, 2016. [Online]. Available: http://arxiv.org/abs/1410.8075.

[10] G. Böcherer, "Achievable rates for probabilistic shaping," *arXiv preprint*, [Online]. Available: https://arxiv.org/abs/1707.01134v5.

[11] 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, pp. 230–244, Jan. 2019.

[12] G. Böcherer, D. Lentner, A. Cirino, and F. Steiner, "Probabilistic parity shaping for linear codes," *arXiv preprint*, 2019. [Online]. Available: https://arxiv.org/abs/1902.10648.

[13] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., 1968.

[14] G. Böcherer and R. Mathar, "Matching dyadic distributions to channels," in *Proc. Data Compression Conf. (DCC)*, 2011, pp. 23–32.

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

[16] Y. C. Gültekin, F. M. Willems, W. van Houtum, and S. Serbetli, "Approximate enumerative sphere shaping," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, Colorado, USA, 2018, pp. 676–680.

[17] 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.

[18] T. Yoshida, M. Karlsson, and E. Agrell, "Hierarchical distribution matching for probabilistically shaped coded modulation," *J. Lightw. Technol.*, vol. 37, no. 6, pp. 1579–1589, 2019.

[19] J. Cho, "Prefix-free code distribution matching for probabilistic constellation shaping," *IEEE Trans. Commun.*, 2019, accepted for publication, DOI: 10.1109/TCOMM.2019.2924896.

[20] Y. Lomnitz and M. Feder, "A simpler derivation of the coding theorem," *arXiv preprint*, 2012. [Online]. Available: https://arxiv.org/abs/1205.1389.

[21] 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, pp. 228–239, 1993.