Journal of Electronic Research and Application, 2026, Volume 10, Issue 5
http://ojs.bbwpublisher.com/index.php/JERA
Online ISSN: 2208-3510
Print ISSN: 2208-3502

Progress on Probabilistic Shaping Techniques for Optical Fiber Communication Systems

Wenwen Bian*

School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China

Author to whom correspondence should be addressed.

Abstract: The continuous growth of global data transmission demands significantly higher channel capacity. According to Shannon's theorem, an upper bound exists for the capacity of additive white Gaussian noise (AWGN) channels. This limit can be closely approached by optimizing conventional modulation schemes. Probabilistic shaping (PS) represents a critical technique to achieve this goal. By employing PS, the signal-to-noise ratio (SNR) gap between practical modulation formats and the Shannon limit can be reduced by up to 1.53dB. PS methods are generally categorized into direct and indirect schemes. Direct PS features low hardware complexity and high processing speed, making it suitable for long-blocklength and linear systems. In contrast, indirect PS can approach the Shannon limit more closely and is better adapted to medium-to-short blocklength and nonlinear scenarios. Nevertheless, it suffers from high hardware complexity and low computational efficiency. Given that direct PS has been well developed and widely deployed, while indirect PS still exhibits considerable room for improvement, future research will concentrate on the enhancement and optimization of indirect PS for nonlinear channel environments.

Keywords: Direct PS; Indirect PS; Distribution matching (DM); Sphere shaping (SS)

Online publication: Jun 29, 2026

1. Introduction

In the digital society, the rapid growth of Internet services and network technologies has led to a sharp increase in network traffic. Channel capacity should be maximized, but it has an upper limit. According to Shannon's theorem, there exists a theoretical maximum information rate for error-free transmission over an additive white Gaussian noise (AWGN) channel. By optimizing the modulation at the transmitter, the actual information rate of the system can approach the Shannon limit with high spectral efficiency [1]. For high-speed long-haul coherent communication systems, constellation shaping is a key technique to improve spectral efficiency.

Constellation shaping includes two categories: geometric shaping (GS) and probabilistic shaping (PS). A Gaussian-like distribution is optimal in the AWGN channel [2]. Geometric shaping adjusts the positions

of constellation points to approximate a Gaussian distribution. PS changes the probability distribution of constellation points to follow a Gaussian-like distribution, thus improving spectral efficiency and system performance. From the perspective of implementation, PS is divided into direct PS and indirect PS.

This paper reviews PS techniques and summarizes the research progress of key methods in direct and indirect PS. The cores of direct and indirect PS are the Distribution Matching (DM) and Sphere Shaping (SS). The first proposed methods for each are Constant Composition Distribution Matching (CCDM) for direct PS and Enumerative Sphere Shaping (ESS) for indirect PS. They established the general framework and theoretical foundation for direct and indirect PS, and provided important references for technical comparison. When the block length approaches infinity, both CCDM and ESS are optimal shaping techniques. Their rate loss is close to 0. For short block lengths up to a few hundred symbols or smaller, indirect PS outperforms CCDM. The maximum energy of ESS is lower than that of CCDM [3].

Overall, indirect PS is better than direct PS. Indirect PS has higher computational and hardware complexity. However, it can approach the Shannon limit more closely. Therefore, it has become the main innovative area of current research.

2.Principle and classification of PS techniques

2.1. Principle of PS techniques

The main principle of PS is to change the occurrence probabilities of constellation points. It makes high-energy constellation points appear less frequently and low-energy constellation points appear more frequently. This allows the distribution of constellation points to approximate the Maxwell-Boltzmann (MB) distribution. As a result, the information rate in AWGN channels approaches the Shannon limit. The probability mass function of the MB distribution is:

$$P(x_i) = \frac{e^{-\lambda|x_i|^2}}{\Sigma_j e^{-\lambda|x_j|^2}}. \tag{1}$$

In **Equation (1)**, $P(x_i)$ represents the probability of the i^{th} constellation point，$|x_i|^2$ represents its energy，λ is the shaping parameter. When λ is larger low-energy constellation points have higher probabilities, high-energy points have lower probabilities, and the shaping gain is stronger. $\sum_j e^{-\lambda|x_j|^2}$ is the normalization factor, which ensures that the probabilities $\sum_i P(x_i)$ sum to 1.

Taking 16QAM as an example, **Figure 1** shows the uniformly distributed constellation diagram on the left and the PS-16QAM constellation on the right.

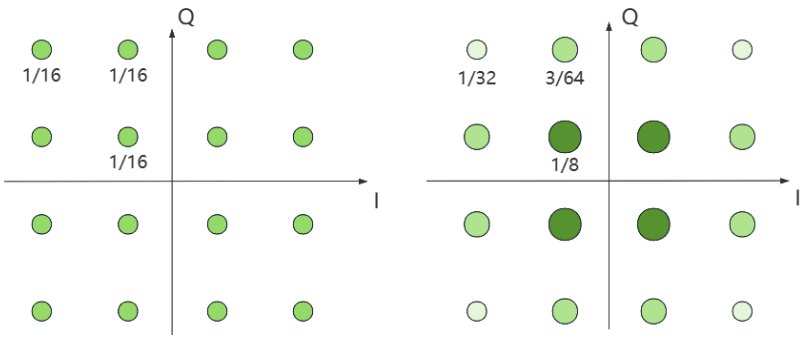

Figure 1. 16QAM vs. PS-16QAM.

In **Figure 1**, the size of constellation points reflects their occurrence probabilities. In uniformly distributed 16QAM, each constellation point has a probability of $\frac{1}{16}$. For a PS-16QAM constellation, symbols are partitioned into three concentric rings with different energies. The symbol energy rises from the inner to the outer ring, with their probabilities manually assigned as $\frac{1}{8}$, $\frac{3}{64}$, and $\frac{1}{32}$. The average power of each is computed individually:

$$P_{16QAM} = \frac{1}{16} \times 4 \times \left(1^2 + 1^2\right) + \frac{1}{16} \times 8 \times \left(1^2 + 3^2\right) + \frac{1}{16} \times 4 \times \left(3^2 + 3^2\right) = 10,$$

$$P_{PS-16QAM} = \frac{1}{8} \times 4 \times \left(1^2 + 1^2\right) + \frac{3}{64} \times 8 \times \left(1^2 + 3^2\right) + \frac{1}{32} \times 4 \times \left(3^2 + 3^2\right) = 7. \quad (2)$$

According to **Figure (2)**, the average output power of uniformly distributed 16QAM is 10, and that of PS-16QAM is 7. This shows that PS can effectively reduce the average output power. Outer constellation points exhibit poorer bit error rate (BER) performance than inner points. For this reason, PS can also enhance the overall BER performance of the system by reducing the probability of outer points.

2.2. Classification of PS techniques

PS techniques are divided into two categories: direct PS and indirect PS. Direct PS includes CCDM, Multiset-Partition Distribution Matching (MPDM), Hierarchical Distribution Matching (HiDM), and Prefix-free Code Distribution Matching (PCDM). Indirect PS includes ESS, PESS, and KESS. Fig**ure 2** shows the proposal time of each technique.

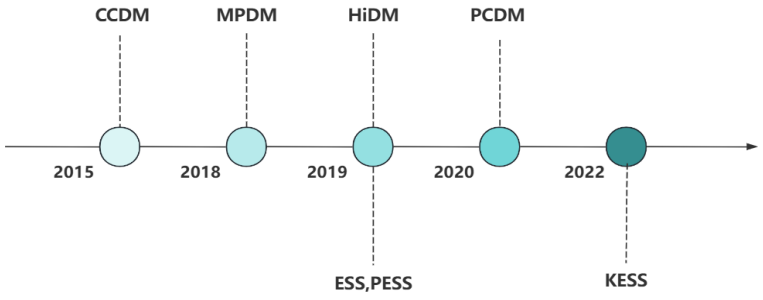

Figure 2. Development of PS techniques.

As shown in **Figure 2**, CCDM, the core technique of direct PS, was proposed in 2015. Since then, various direct PS techniques have been proposed as improvements and extensions of CCDM. In 2019, indirect PS emerged. ESS, as the core of indirect PS, serves as a foundation for various subsequent indirect PS techniques.

3. Direct PS

Direct PS starts from the target distribution in a low-dimensional signal structure. It directly changes the occurrence probability of constellation points through a DM. It then generates this distribution directly using algorithms. The target distribution is typically the MB distribution. This is because it is the counterpart of the Gaussian distribution in discrete linear channels. It also maximizes the average information (entropy) under a

fixed average energy.

3.1. CCDM

CCDM was first proposed by Ludwig-Maximilians-Universität München in 2015 [4]. As a direct PS technique based on serial architecture, CCDM is a practical implementation of DM. Its core design goal is to precisely map uniformly distributed input bits into a symbol sequence following the MB distribution using fixed composition, meaning the proportion of symbols with different amplitudes is fixed in each output sequence [2]. This ensures the occurrence probability of constellation points meets the optimization requirements of PS. CCDM achieves optimal theoretical performance among similar PS techniques. It has low shaping loss, high amplitude resolution, and supports fixed-length reversible mapping. Since the composition of the output sequence is fixed, its energy remains constant. Therefore, CCDM has attracted extensive attention and has been widely applied in high-capacity, high-spectral-efficiency optical fiber transmission systems. The distribution matching process of CCDM is reversible. At the receiver, inverse DM only depends on the composition constraint within the block. Thus, a single symbol error will not spread to the entire data block. This feature improves its compatibility with Forward Error Correction (FEC) codes. It maintains stable decoding performance even at low signal-to-noise ratios and enhances the overall robustness of the system.

However, fixed-length-to-fixed-length DM requires a large block length to work properly [5]. It suffers from considerable rate loss at short block lengths. In addition, the serial encoding algorithm introduces certain latency. Therefore, CCDM is not suitable for high-throughput applications. Furthermore, CCDM is mostly employed in linear channels, where the performance of the DM is inferior to indirect PS methods such as SS. Moreover, CCDM has high hardware complexity. It cannot meet low-cost requirements. Thus, it is mainly used in long-distance optical transmission systems like backbone networks.

3.2. MPDM

CCDM requires all output sequences to have a fixed composition. Any finite-length DM causes rate loss inevitably, and this loss becomes larger when the block length is shorter. Therefore, CCDM needs a very long block length to achieve near-ideal performance. To solve this problem, MPDM was proposed in 2018 [6].

The basic principle of MPDM is as follows. First, it quantizes the target probability distribution into a discrete distribution and determines the typical composition accordingly. It breaks the limitation of CCDM, where all output sequences must strictly follow a single typical composition. It only requires the overall average composition of all output sequences to match the typical composition. Then, it approaches the target probability distribution based on the law of large numbers. Meanwhile, it maintains the fixed-length mapping between input and output, and achieves the reduction of block length.

Compared to the earliest CCDM, MPDM supports more output sequences. In all non-trivial scenarios, MPDM only requires that the overall average composition of all output sequences equals the target typical composition. It does not require every individual sequence to conform to this composition. Therefore, compared to CCDM, MPDM offers lower rate loss, reduced complexity, and lower latency. Additionally, in the medium to high signal-to-noise ratio (SNR) regime, to achieve a fixed gap to channel capacity, MPDM requires approximately 2.5 to 5 times less blocklength than CCDM. Therefore, MPDM can handle a larger set of sequences than CCDM [1].

However, this advantage is significant only in the medium to high SNR regime. In low SNR scenarios,

MPDM's benefits in blocklength saving and rate loss reduction are greatly diminished. Furthermore, its approach of approximating the target distribution through an average of multiple compositions may perform worse due to noise interference, making it less robust and stable than CCDM's single typical composition mapping.

3.3. HiDM

Although CCDM is a DM for reverse concatenated PS, it still has room for simplification in high-speed optical fiber communication systems. The DM and inverse distribution matching (invDM) consume considerable circuit resources. In practice, at high throughput, imperfect signal processing in deployable hardware leads to a practical performance penalty relative to achievable rates that assume ideal DM and FEC.

To reduce the complexity of PS, HiDM was proposed in 2019 [7]. HiDM is a low-complexity distribution matching scheme for PS. By hierarchically cascading small look-up tables (LUTs), it enables the output symbol sequence of the DM to follow a MB distribution. Each LUT layer generates output bits based on the "constraint bits" from the previous layer and the input "information bits", achieving a fixed-length input-output mapping. The architecture adopts a fully parallel input-output interface and a pipeline structure, avoiding complex operations such as arithmetic coding used in serial-structured CCDM. Compared to MPDM, HiDM's fixed-length conversion and hierarchical LUT design greatly reduce hardware storage requirements. It also meets the high throughput demands of high-speed optical communication systems.

However, to achieve a low-complexity and high-throughput architecture, HiDM incurs a moderate performance penalty. Compared to CCDM, HiDM cannot perfectly approximate the MB distribution due to the approximation nature of its cascaded LUT layers. As a result, it has a slightly higher rate loss and requires a higher SNR to achieve the same post-FEC BER. In simulations, under the same BER condition, its required SNR is 0.13dB higher than that of CCDM. Its constellation gain is 0.39dB lower than that of the ideal MB distribution. Its performance is consistently worse than a single-LUT scheme that can store optimal low-energy sequences. However, these performance penalties do not negate its core advantages in hardware complexity, memory requirements, and throughput in practical applications.

3.4. PCDM

Traditional HiDM is based on a modified, non-prefix-free Huffman coding. This leads to several issues. Decoding can be ambiguous. Channel errors can easily cause synchronization loss and error propagation. To address this, PCDM was proposed in 2020.

PCDM uses a naturally prefix-free code set, enabling instantaneous symbol-by-symbol decoding. An error affects only the current codeword. As a result, decoding robustness is significantly improved. To handle the variable-length output of prefix-free codes, it uses a fixed-frame encapsulation algorithm based on dual code set switching and minimum-energy symbol padding. This ensures perfect compatibility with the fixed-frame transmission structure of communication systems, reducing hardware buffering and synchronization overhead. The code sets are constructed using a tree-based structure, so decoding only requires simple LUTs. This results in low storage complexity, making the scheme cost-effective and easy to implement in hardware. Furthermore, in PS applications over AWGN channels, PCDM achieves significant shaping gains ranging from 0.21dB to 0.98dB compared to uniform QAM formats (4, 8, 16, and 32). Its performance is closer to that of CCDM, with only a small gap of ≤ 0.2dB [8].

However, PCDM still has technical limitations despite its strong performance and practical engineering value, it requires additional processing logic for code set switching and symbol padding. This increases the complexity of frame encapsulation and synchronization compared to fixed-length distribution matching. The padding symbols also introduce a small amount of energy overhead. In short blocklength transmission scenarios, the variable-length nature of its prefix-free codes leads to fluctuations in the effective information rate. It also makes it difficult to fully achieve the energy gain. Compared to distribution matching techniques optimized for short blocklengths, PCDM experiences a more significant rate loss. As a result, its shaping gain is also reduced.

In contrast to direct PS techniques, indirect PS approaches offer distinct advantages in short blocklength and nonlinear scenarios, as discussed in the following section.

4. Indirect PS

Indirect PS techniques alter the boundary geometry of the signal space. This indirectly induces a non-uniform distribution in low dimensions. Given a fixed volume, the N-dimensional sphere is the most energy-efficient geometric shape. Therefore, the signal space is generally considered as an N-dimensional sphere. SS can utilize all energy sequences within the sphere. Therefore, it achieves the smallest possible rate loss for a given target rate.

Figure 3 shows the energy shell model of indirect PS under uniform signaling. From the inner layer to the outer layer, the energy increases progressively. In an N-dimensional sphere, the outermost layer with the highest energy is removed. Only the inner layers with lower energy are kept for transmission. This is done without directly changing the probability of each amplitude signal. This approach ensures sufficient information throughput while significantly reducing the average transmit power.

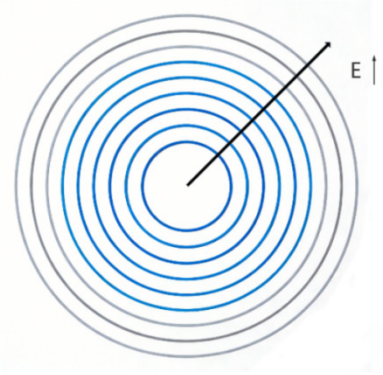

Figure. 3. Energy shell structure under uniform signalling.

4.1. ESS

CCDM suffers from high rate loss at short blocklengths. Therefore, ESS is proposed as an alternative. ESS can minimize rate loss at any blocklength.

ESS is a probability amplitude shaping (PAS) technique designed for short-packet wireless communications. It minimizes rate loss at any blocklength by utilizing all sequences within an N-dimensional sphere that satisfy the energy constraint. Especially in short blocklength scenarios, its rate loss is only one-

fifth that of CCDM. ESS achieves lower average symbol energy and more significant shaping gain (reaching 1.11dB for 8ASK with a blocklength of 96). In AWGN channels, it provides up to 1.6dB energy efficiency improvement over uniform signaling. In frequency-selective channels, it can achieve a gain of up to 0.7dB. It also supports flexible rate adaptation with a granularity of up to $\frac{1}{N}$ bit/symbol by adjusting the maximum energy constraint. Even at very short blocklengths (i.e., N=6), it maintains an energy gain of 0.59dB [3].

However, due to technical principles and engineering implementation requirements, several inherent limitations still exist. For example, ESS relies on a bounded-energy lattice structure as its core, which leads to the storage requirement growing cubically with block length N in full-precision implementations. Although its computational complexity is linear, it is still much higher than the minimal storage overhead of CCDM. This issue of hardware resource occupation is particularly severe in resource-constrained scenarios. Furthermore, since ESS sorts sequences lexicographically, the unused sequences in actual transmission are not necessarily the outer-layer sequences with the highest energy. This may cause the actual average energy to be higher than the theoretical value. Meanwhile, although ESS outperforms CCDM for short block lengths, its shaping gain at extremely short block lengths (i.e., N = 6) is significantly lower than that for medium and long block lengths. Moreover, a LUT implementation at such lengths still incurs high storage costs.

4.2. Partial enumerative sphere shaping (PESS)

The PESS scheme only shapes part of the amplitude bits. It uses a smaller set of amplitude values to run ESS and lowers the storage and computation complexity. Shaping bits combine with uniform data bits to form a partially shaped constellation. This constellation has performance very close to a fully shaped constellation.

For 16ASK modulation with a shaping rate of 2.67 bits per amplitude and a block length shorter than 300 symbols, PESS achieves lower rate loss than CCDM. Simulation results on the AWGN channel show that shaping two amplitude bits of 16ASK signals achieves similar performance gain to shaping three bits. Twobit and onebit PESS provide 1.27dB and 0.95dB gains respectively, while CCDM provides 0.45dB. Two-bit PESS performs very close to three-bit ESS with only a 0.08dB difference, while its required memory and computational complexity are greatly reduced.

However, PESS still has inherent limitations. As it only shapes part of the amplitude bits, its rate loss cannot fully drop to zero. The rate loss converges to 0.071 bit per amplitude for 1-bit shaping and 0.015 for 2-bit shaping. Fewer shaping bits lead to higher average symbol energy and lower shaping gain. The gain of 1-bit shaping is only 0.81dB, lower than the 1.29dB of 3-bit full shaping [9].

4.3. Kurtosis-limited enumerative sphere shaping (KESS)

Although ESS can achieve high energy efficiency through energy constraints, the Gaussian-like distribution sequences it generates have high kurtosis. High kurtosis inputs increase nonlinear interference (NLI) in fiber channels and reduces the effective SNR. With a small number of wavelength division multiplexing (WDM) channels, the high kurtosis limits the transmission distance either. These effects are especially significant in single-span fiber transmission.

KESS is an improved algorithm of ESS. It adds a kurtosis constraint to conventional spherical shaping, which only limits sequence energy. By limiting both energy and the sum of fourth-order moments, it constructs a signal set with constrained kurtosis. It extends the energy accumulation lattice of ESS to a two-dimensional energy-kurtosis accumulation lattice. By counting the number of sequences at each lattice point,

it realizes a reversible mapping from binary indices to amplitude sequences that satisfy both energy and kurtosis constraints. This generates low-kurtosis shaping inputs and reduces nonlinear interference at the source, solving the key incompatibility between ESS and the nonlinear nature of fiber channels. In addition, KESS can revert to traditional ESS by relaxing the kurtosis constraint [10]. KESS provides irreplaceable advantages in key scenarios with few channels, such as dedicated fiber links [11].

KESS only achieves significant gains in short-distance, single-span, few-channel transmission. In multi-span long-distance or multi-channel systems, its performance converges to traditional ESS. The advantage of kurtosis constraint disappears, and the optimal scheme becomes the same as ESS without kurtosis constraint. When the shaping block length is close to the channel optimum, the performance gap between KESS and ESS becomes small. If the block length is too small, the difference disappears completely, and the room for optimization is greatly reduced. To make up for the reduced number of sequences caused by the kurtosis constraint, KESS requires higher average energy to maintain the rate. This is the necessary cost for its kurtosis optimization.

5. Comprehensive comparison of PS techniques

This section focuses on a technical comparison between direct and indirect PS techniques. **Table 1** offers a high-level overview, which compares the overall characteristics of the two techniques. **Table 2** delivers an in-depth analysis, examining the advantages and disadvantages of specific methods under both categories.

5.1. Comparison of direct and indirect PS techniques

Table 1 shows that direct PS uses a large blocklength. It's hardware complexity is low, making it easier to implement. It is often used in serial linear environments and long-haul optical transmission systems. Indirect PS is suitable for medium to short blocklengths. Its performance outperforms direct PS. However, its implementation complexity is higher. It is suitable for parallel nonlinear environments and single-span short-distance scenarios.

Table 1. Direct PS vs. indirect PS

Types of PS techniques	Block length	Shaping gain	Computational complexity	Channel regime	Application scope
Direct PS	Suitable for long blocklengths.	Approach Shannon capacity, closing the 1.53dB SNR gap.	Lower	Linear	Long-haul serial optical transmission system.
Indirect PS	Suitable for medium to short blocklengths.	Provides nonlinear shaping gain at a block length of 256. Has lower rate loss than CCDM.	Higher	Nonlinear	Single-span short-distance parallel scenario.

5.2. Comprehensive comparison of various PS techniques

As shown in **Table 2**, each PS technique has its unique technical characteristics. In general, direct PS offers fast implementation and low hardware complexity, making it widely applicable. Indirect PS offers better technical performance. However, it has slower implementation speed and higher hardware complexity. It is an important aspect of innovation and development in PS technology.

Table 2. Technical characteristics of direct and indirect PS

Types of PS techniques	Technical name	Advantages	Disadvantages	Reference
Direct PS	CCDM	Low processing loss, high amplitude resolution, constant energy.	Larger block length, high hardware complexity.	[4]
	MPDM	Shorter block length, lower rate loss, complexity and latency than CCDM.	Low robustness under low SNR.	[6]
	HiDM	Lower hardware storage, supports high-speed transmission.	Slight rate loss, slightly inferior performance to CCDM.	[7]
Indirect PS	PCDM	No decoding ambiguity, strong practicality, performance close to CCDM.	Higher frame encapsulation complexity, significant rate loss at short block lengths.	[8]
	ESS	Much lower rate loss than CCDM, low computational complexity.	Low shaping gain at ultra-short block lengths, significant gain degradation in dynamic channels.	[3]
	PESS	Low storage and computational complexity.	Rate loss exists, lower gain with fewer shaping bits.	[9]
	KESS	Better single-span short-haul performance.	Advantage over ESS vanishes in multi-span, multi-channel scenarios, requires higher average energy.	[10]

6. Conclusion

PS is a key technology to improve spectral efficiency and energy efficiency in communication systems. Direct methods focus on distribution matching to balance performance and complexity. Indirect methods focus on geometric design of signal space to fit different transmission scenarios. Each method has its own advantages and disadvantages. CCDM performs well in long-distance scenarios, and KESS is suitable for single-span short-distance transmission. MPDM, HiDM, PCDM, ESS, and PESS solve specific problems in targeted situations. Future research should further balance performance and complexity. Since PS for linear channels is nearly mature in theory and practice, nonlinear channels are the core focus for current and future breakthroughs. Sequence Selection is a recently proposed method for investigating the ultimate potential of PS in nonlinear channels.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Liu X, Zhang J, Zhu M, et al., 2022, Research Status and Progress of Probabilistic Shaping Technology in Optical Communications. Adv. Laser Optoelectron. 2022(59): 62–82.

[2] Shu C, Zhang Q, Shatin N, 2020, Probabilistic Shaping: A Step Closer to the Shannon Limit in Channel Capacity for Optical Communications. Electron. Lett., 2020(56): 1162–1163.

[3] Ritter F, Rode A, Schmalen L, 2025, An Extension of Enumerative Sphere Shaping for Arbitrary Channel Input Distributions, arXiv, https://doi.org/10.48550/arXiv.2512.16808

[4] Böcherer G, Steiner F, Schulte P, 2015, Bandwidth Efficient and Rate-Matched Low-Density Parity-Check Coded Modulation. IEEE Transactions on Communications, 63(12): 4651–4665.

[5] Schulte P, Böcherer G, 2016, Constant Composition Distribution Matching. IEEE Transactions on Information Theory, 62(1): 430–434.

[6] Fehenberger T, Millar D, Koike-Akino T, et al., 2019, Multiset-Partition Distribution Matching. IEEE Transactions on Communications, 67(3): 1885–1893.

[7] Yoshida T, Karlsson M, Agrell E, 2019, Hierarchical Distribution Matching for Probabilistically Shaped Coded Modulation. Journal of Lightwave Technology, 37(6): 1579–1589.

[8] Cho J, 2020, Prefix-Free Code Distribution Matching for Probabilistic Constellation Shaping. IEEE Transactions on Communications, 68(2): 670–682.

[9] Gultekin Y, van Houtum W, Koppelaar A, et al., 2019, Partial Enumerative Sphere Shaping, 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), 1–6.

[10] Askari M, Lampe L, 2025, Probabilistic Shaping for Nonlinearity Tolerance. Journal of Lightwave Technology, 43(4): 1565–1580.

[11] Civelli S, Parente E, Forestieri E, et al., 2023, On the Nonlinear Shaping Gain with Probabilistic Shaping and Carrier Phase Recovery. Journal of Lightwave Technology, 41(10): 3046–3056.