Analysis of Nonlinear Fiber Interactions for Finite-Length Constant-Composition Sequences

Fehenberger, Tobias; Millar, David S.; Koike-Akino, Toshiaki; Kojima, Keisuke; Parsons, Kieran; Griesser, Helmut

TR2020-021   March 03, 2020

Abstract

In order to realize probabilistically shaped signaling within the probabilistic amplitude shaping (PAS) framework, a shaping device outputs sequences that follow a certain nonuniform distribution. In case of constant-composition (CC) distribution matching (CCDM), the sequences differ only in the ordering of their constituent symbols, whereas the number of occurrences of each symbol is constant in every output block. Recent results by Amari et al. have shown that the CCDM block length can have a considerable impact on the effective signal-to-noise ratio (SNR) after fiber transmission. So far, no explanation for this behavior has been presented. Furthermore, the block-length dependence of the SNR seems not to be fully aligned with previous results in the literature. This paper is devoted to a detailed analysis of the nonlinear fiber interactions for CC sequences. We confirm in fiber simulations the inverse proportionality of SNR with CCDM block length and present two explanations. The first one, which only holds in the short-length regime, is based on how two-dimensional symbols are generated from shaped amplitudes in the PAS framework. The second, more general explanation relates to an induced shuffling within a sequence, or equivalently a limited concentration of identical symbols, that is an inherent property for short CC blocks, yet not necessarily present in case of long blocks. This temporal property results in weaker nonlinear interactions, and thus higher SNR, for short CC sequences. For a typical multi-span fiber setup, the SNR difference is numerically demonstrated to be up to 0.7 dB. Finally, we evaluate a heuristic figure of merit that captures the number of runs of identical symbols in a concatenation of several CC sequences. For moderate block lengths up to approximately 100 symbols, this metric suggests that limiting the number identicalsymbol runs can be beneficial for reducing fiber nonlinearities and thus, for increasing SNR.

IEEE Journal of Lightwave Technology

Analysis of Nonlinear Fiber Interactions for Finite-Length Constant-Composition Sequences

Tobias Fehenberger, *Member, IEEE*, David S. Millar, *Member, IEEE*, Toshiaki Koike-Akino, *Senior Member, IEEE*, Keisuke Kojima, *Senior Member, IEEE*, Kieran Parsons, *Senior Member, IEEE*, and Helmut Griesser, *Member, IEEE*

Abstract—In order to realize probabilistically shaped signaling within the probabilistic amplitude shaping (PAS) framework, a shaping device outputs sequences that follow a certain nonuniform distribution. In case of constant-composition (CC) distribution matching (CCDM), the sequences differ only in the ordering of their constituent symbols, whereas the number of occurrences of each symbol is constant in every output block. Recent results by Amari *et al.* have shown that the CCDM block length can have a considerable impact on the effective signal-to-noise ratio (SNR) after fiber transmission. So far, no explanation for this behavior has been presented. Furthermore, the block-length dependence of the SNR seems not to be fully aligned with previous results in the literature. This paper is devoted to a detailed analysis of the nonlinear fiber interactions for CC sequences. We confirm in fiber simulations the inverse proportionality of SNR with CCDM block length and present two explanations. The first one, which only holds in the short-length regime, is based on how two-dimensional symbols are generated from shaped amplitudes in the PAS framework. The second, more general explanation relates to an induced shuffling within a sequence, or equivalently a limited concentration of identical symbols, that is an inherent property for short CC blocks, yet not necessarily present in case of long blocks. This temporal property results in weaker nonlinear interactions, and thus higher SNR, for short CC sequences. For a typical multi-span fiber setup, the SNR difference is numerically demonstrated to be up to 0.7 dB. Finally, we evaluate a heuristic figure of merit that captures the number of runs of identical symbols in a concatenation of several CC sequences. For moderate block lengths up to approximately 100 symbols, this metric suggests that limiting the number identical-symbol runs can be beneficial for reducing fiber nonlinearities and thus, for increasing SNR.

Index Terms—Fiber Nonlinearities, Constant-Composition Distribution Matching, Probabilistic Amplitude Shaping, Gaussian Noise Models

I. INTRODUCTION

Probabilistic amplitude shaping (PAS) [1] has become the de-facto standard coded modulation (CM) framework for realizing probabilistic shaping, offering about 1 dB of signal-to-noise ratio (SNR) improvement and rate adaptivity without major modifications to an existing CM architecture that was designed for unshaped, i.e., uniform, signaling. An important part of PAS is the shaper that transforms the uniform data

T. Fehenberger was with Mitsubishi Electric Research Laboratories. He is now with ADVA Optical Networking, Munich, Germany. E-mail: tfehenberger@advaoptical.com.

D. S. Millar, T. Koike-Akino, K. Kojima and K. Parsons are with Mitsubishi Electric Research Laboratories. E-mails: millar@merl.com; koike@merl.com; kojima@merl.com; parsons@merl.com.

H. Griesser is with ADVA Optical Networking, Munich, Germany. E-mail: hgriesser@advaoptical.com.

sequence to shaped amplitudes at the transmitter. The inverse operation at the receiver recovers the initially sent data from the shaped amplitudes, assuming that all transmission errors have been corrected by forward error correction (FEC), see Fig. 1 for a block diagram. The most prominent shaping system is constant-composition distribution matching (CCDM) [2], which is used in the initial PAS proposal [1, Sec. V]. As the name suggests, a CCDM of a certain length outputs sequences of constant composition (CC), i.e., which have a constant number of occurrences of each amplitude and are hence permuted version of each other. More advanced distribution matchers include multi-set partition distribution matching (MPDM) [3] and multi-composition codes [4]. Other fixed-length shapers that are not directly related to the concept of compositions include enumerative sphere shaping (ESS) [5] and shell mapping [6], both of which establishing an energy-efficient indexing of all sequences within a sphere of fixed radius, as well as prefix-free code distribution matching (PCDM) with framing [7].

Probabilistically shaped quadrature amplitude modulation (QAM) is well-studied for the nonlinear fiber channel in theory [8], transmission simulations [9]–[12] and experiments [12]–[15]. Temporal shaping of quaternary phase shift keying (QPSK) over a few time slots has been performed in [16]. In [11], which is the motivation for this manuscript, the authors investigate ESS and CCDM in multi-span wavelength-division multiplexing (WDM) simulations and find that the effective SNR after fiber transmission is inversely proportional to the block lengths of the employed CCDM and ESS. Unfortunately, no reasons for this relation are presented therein or, to the best of our knowledge, anywhere else. Upon re-analysis of the simulation data of our previous work [17] where we focused on the performance of shaped signaling with low-density parity-check FEC, we are able to confirm also in that paper that the average SNR varies with CCDM block length, which had not been mentioned therein. This finding of block-length dependent SNR poses questions on previously reported results, and also some challenges for future fiber models on which we will comment on briefly.

Regarding the latter, a block-length dependence of SNR after fiber transmission is not covered by the prominent enhanced Gaussian noise (EGN) model [18], [19] which considers only time-averaged statistics by assuming infinite-length blocks. The impact of the channel input on the fiber nonlinearities is described by the fourth and sixth moment of the QAM constellation, but finite-length temporal proper-
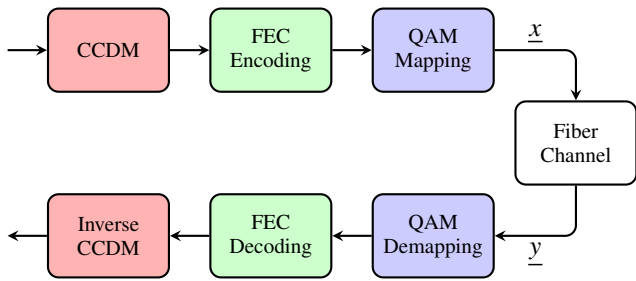
Fig. 1. Simplified block diagram of probabilistic amplitude shaping (PAS). This paper studies the impact of finite-length sequences $\underline{x}$, which are generated from the amplitude sequences output by the CCDM, on the fiber output $\underline{y}$, specifically on the effective SNR after fiber transmission. A complete block diagram for PAS can be found in [1, Fig. 3].

ties are not included. The time-domain model presented in [20] incorporates temporal effects, i.e., memory, by including the power of past and future symbols, yet not their phase information. The models proposed in [19, Sec. II], [21], [22] include all information of previous and past symbols. All of these models are arguably prohibitively complex to evaluate for many practical setups in which long memory is to be captured. Powerful models that are relatively simple to evaluate could thus be helpful in better understanding time-dependent nonlinear interactions, and in predicting the system performance particularly when short-length shaping is applied. We will set this open research problem aside and focus on understanding the origin of the SNR dependence on block length in the following.

Previous fiber simulations [9], [23], [24, Fig. 2] and experiments [14, Fig. 8], [25] found that the average SNR after transmission over the nonlinear fiber channel is smaller for shaped QAM than for uniform QAM. None of these papers, however, specifies the length of the employed DM, which suggests that either very long blocks were used where rate loss is negligibly small, or that a DM was emulated by drawing the QAM symbols randomly according to the target distribution. With the simulations results presented in [11] and the detailed analysis presented herein, the general finding that the SNR of shaped QAM is worse than that of uniform QAM must be further confined as this appears to be only true when sufficiently long shaping blocks are used. This manuscript is devoted to a detailed study into the origin of this block-length dependence.

We first analyze the generation of QAM symbols from CC amplitude sequences, as it is carried out for CCDM in the PAS framework. The main part of this work is then devoted to a study of temporal properties of a concatenation of CC sequences for varying DM block lengths. After a theoretical assessment, we perform extensive fiber simulations in order to analyze the dependence of the effective SNR after nonlinear fiber transmission, which we simply refer to as SNR, on the block length of these CC sequences. Throughout this work, we focus on CCDM of varying length but with a constant per-sequence distribution as to isolate the sequence properties (and thereby the nonlinear effects) that are present for different finite block-lengths. The study of low-rate-loss shapers, such

as ESS, MPDM or PCDM, is more complex since their outputs in general do not fulfill the CC property and is left for future work. The considered setting of having a fixed distribution at varying block length differs from [26] where the inverse case of SNR fluctuations for varying per-block composition at a fixed block length is considered.

The findings in this paper are as follows. We first demonstrate that the distribution of the QAM symbols, of each which is obtained from combining two amplitudes, can differ from the infinite-length expectation for very short block lengths. This is a direct result of how the one-dimensional shaped ASK signals are paired to QAM symbols in the PAS framework, and the first reason for an increased effective SNR at ultra-short blocks. The main reason for the SNR variations, however, is shown to be due to intra- and inter-block nonlinear effects, i.e., temporal nonlinear interactions within a shaping block and over adjacent blocks. By introducing an interleaver for each FEC codeword, we demonstrate that the SNR-dependence of the block length disappears almost entirely. When comparing short and long block lengths without interleaving, it becomes evident that some kind of temporally structured shuffling of the shaped amplitudes, which is implicitly generated with short-length CCDM, leads to reduced nonlinear effects and thus higher SNR than for long block-lengths. Using a heuristic metric that represents the number of identical-symbol runs in a compound sequence, we present strong arguments that, up to moderate block lengths, it is mainly the limited number of runs of identical symbols that gives the SNR improvement. To the best of our knowledge, this manuscript is the first detailed study of the characteristic effects that temporal structures in the transmit sequences have on the generation of nonlinear fiber interference.

II. QAM Symbols Generated From Constant-Composition Amplitude Sequences

A. Fundamentals of Probabilistic Amplitude Shaping

In PAS, the shaped amplitude output of the CCDM is combined with uniform bits, which are the FEC parity and possibly some data bits, into QAM symbols, see Fig. 1 for a coarse block diagram. More details on PAS can be found in [1, Sec. IV]. The output of a CCDM is fully described by its integer-valued composition, which describes exactly how often an amplitude occurs in every block. All possible CCDM output sequences for a fixed composition are permutations of each other, with the number of permutations given by the multinomial coefficient.

Example 1 (CC Sequences): The amplitude distribution $[0.4, 0.3, 0.2, 0.1]$ for the CCDM block length $n = 10$ gives the composition $[4, 3, 2, 1]$, meaning that the first amplitude occurs exactly four times in the output sequence of length 10, the second symbol three times and so forth. There are $\binom{10}{4} \cdot \binom{6}{3} \cdot \binom{3}{2} \cdot \binom{1}{1} \cdot = 12600$ different sequences with this composition. Increasing the block length to $n = 100$ gives the composition $[40, 30, 20, 10]$ and $\sim 4.9 \cdot 10^{52}$ different sequences.

In Fig. 2, the nesting of several DM codewords in one FEC block is shown. The overall considered simulation run (or experimental batch) contains several of these FEC blocks.
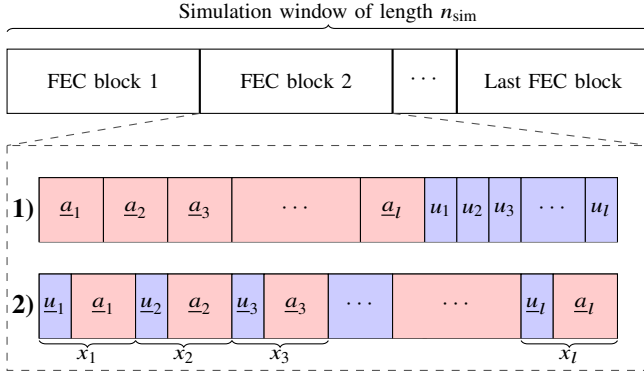
Fig. 2. Temporal structure of a typical simulation in the PAS framework: one simulation windows consists of several FEC blocks, which in turn contain one or more DM blocks. The block structures **1)** and **2)** are equivalent representations of the constituent symbols of a FEC word. In **1)**, the parity of the systematic encoder is at the end whereas **2)** has the parity information distributed over the codeword as prefix $\underline{u}_i$ of the shaped amplitudes $\underline{a}_1$, which is how QAM symbols $\underline{x}_1$ are generated in PAS.

The two rows **1)** and **2)** illustrate how the uniform sign bits $\underline{u}$ and the shaped amplitudes $\underline{a}$ are grouped within a FEC word.[1] The canonical view of a systematic code is **1)** where the parity bits go at the end of each FEC block. Row **2)** of Fig. 2 is closer to the structure that is actually transmitted over the channel: In PAS with QAM signaling, two uniform bits $\underline{u}_i$ that determine the quadrant are prefix to a pair of shaped amplitudes $\underline{a}_i$, which represents the shaped QAM payload, and combining these gives a shaped QAM symbol $\underline{x}_i$. It is important to realize that the generated QAM sequences is thus not CC (although the amplitudes are) because the quadrant identifiers are only approximately uniformly distributed. We further note that a pair of shaped amplitudes is required for each QAM symbol. These can be combined in two different ways, as we will discuss next.

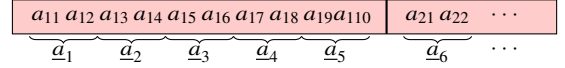### B. Two Strategies for Pairing Shaped Amplitudes

In the following, we study the QAM distributions that are obtained by combining the output amplitudes of varying-length CCDM. Intuitively, one would expect that the empirical distribution of the QAM symbols always follow the Cartesian product of the two-sided shaped amplitude shift keying (ASK) distributions. Firstly, note that this expectation only holds on average because the quadrant identifier, i.e., the first two bits of every QAM symbols as shown in Fig. 2 row **2)**, is in general random data.[2] This shows that the distribution of the QAM symbols is *not* constant-composition, even if the 1D amplitudes are generated with a CCDM.

We now set this aspect of random quadrant identifiers aside as it will average out over time, and analyze constraints in the 2D distributions that stem from short-length CCDM.

---

[1] Notation: Sequences are underlined with their elements in regular font, i.e., $\underline{x} = [x_1, \ldots, x_n]$. A discrete random variable X has the probability mass function (PMF) P_X on $\mathcal{X}$.

[2] In PAS, the quadrant identifier comes at first from the redundancy bits of the FEC codeword. If the FEC rate is chosen such that not enough parity bits are generated, uniformly distributed data bits are used in addition, see [1, Sec. IV-D].

Intra-DM pairing:
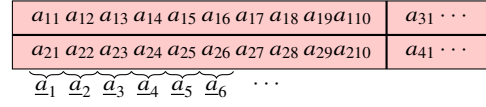

Inter-DM pairing:

Fig. 3. Two strategies for combining two shaped amplitudes into the payload of a QAM symbol, illustrated for $n = 10$. Top: Intra-DM pairing where two adjacent amplitudes are combined within a DM block. Bottom: Inter-DM pairing where the two amplitudes to be paired come from different DM outputs.

First, we investigate in detail how the QAM payload, i.e., the shaped data within each 2D symbol, is generated. In the PAS framework, as explained above, two ASK symbols are combined with a two-bit prefix, which determines the quadrant, into a QAM symbol. This means that the DMs must output twice as many shaped amplitudes as QAM symbols are transmitted. Note that this holds not only for CCDM but in general for any PAS shaper that generates amplitudes. An important aspect is how the amplitude symbols are paired into the QAM payload, which has, to the best our knowledge, not been studied yet. Figure 3 shows two different strategies. In the first approach, illustrated at the top of Fig. 3 and denoted as *intra-DM pairing*, the amplitudes sequences are simply concatenated and each QAM payload $\underline{a}$ is generated by combining consecutive symbols within a DM output. The pairing offset is fixed to zero, i.e., the first symbol is paired with the second symbol, the third with the fourth and so forth. A different strategy, referred to as *inter-DM pairing*, is shown at the bottom of Fig. 3. Two DM outputs are stacked on top of each other and two amplitudes, one from each DM output, are combined. Hence, the QAM payload $\underline{a}$ for this pairing strategy is the combination of two DM outputs that are independent in the sense that the CC constraint is fulfilled by each of them individually, yet not jointly. Note that the original PAS proposal does not specify which strategy to use but only states that QAM constellations can be generated "by mapping two real ASK symbols to one complex QAM symbol" [1, p. 4652]. We will illustrate in the following example that for intra-DM pairing, a different QAM distribution than expected can be obtained, in particular for very short blocks.

Example 2 (60QAM instead of 64QAM): We consider the amplitude distribution $[0.4, 0.3, 0.2, 0.1]$ and $n = 10$, which gives the composition $[4, 3, 2, 1]$. The expected probabilities of the shaped 64QAM, which are computed from the two-sided amplitude distribution as Cartesian products, are

$$\begin{bmatrix} 0.010 & 0.020 & 0.030 & 0.040 \\ 0.020 & 0.040 & 0.060 & 0.080 \\ 0.030 & 0.060 & 0.090 & 0.120 \\ 0.040 & 0.080 & 0.120 & 0.160 \end{bmatrix},$$

shown and normalized for the 16 constellation points in the top-left quadrant. With intra-DM pairing, which we recall to be
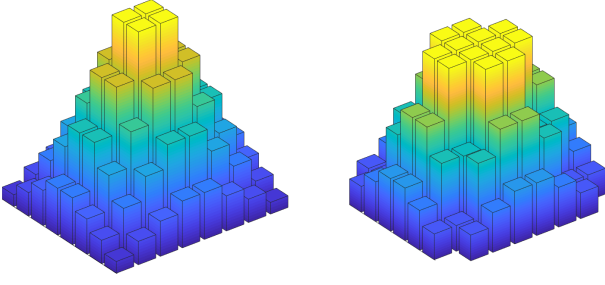
Fig. 4. 64QAM (left) expected for infinite-length sequences vs. 60QAM (right) obtained for the amplitude distribution $[0.4, 0.3, 0.2, 0.1]$ and $n = 10$, see Example 2 for details.
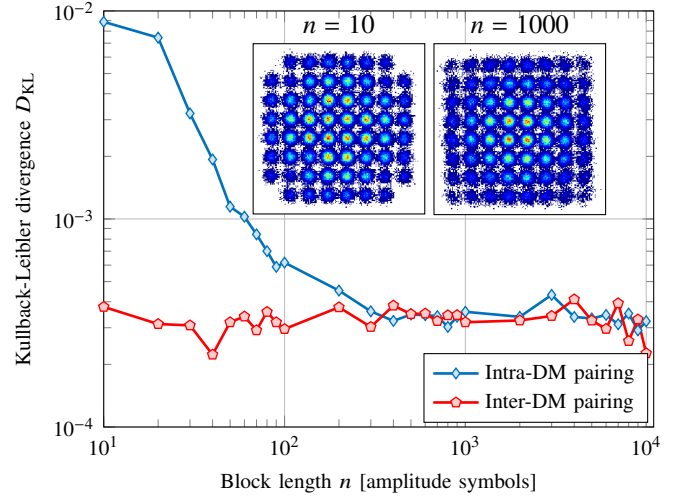

Fig. 5. KL divergence between the expected QAM PMF and the PMF that is obtained from constant-composition amplitude sequences, as a function of the block length n. Higher KL divergence means larger difference in distributions. Insets: Received constellations at approximately 19 dB SNR for the intra-DM amplitude pairing and two different block lengths. The top left-most marker and the left inset correspond to Example 2.

the pairing of adjacent symbols within a DM output, the outermost QAM symbol is only obtained when the highest-energy amplitude occurs as two consecutive symbols within one CC sequence. In this short-length example, this is impossible to happen because the highest-energy sequence occurs only once per DM output. Therefore, the outermost QAM symbols have zero probability of occurrence instead of the expected 0.01. Rather than the expected 64QAM, a 60QAM distribution is generated with the per-quadrant PMF

$$\begin{bmatrix} 0 & 0.022 & 0.034 & 0.045 \\ 0.022 & 0.022 & 0.067 & 0.090 \\ 0.034 & 0.067 & 0.067 & 0.133 \\ 0.045 & 0.090 & 0.133 & 0.133 \end{bmatrix}.$$

The full PMFs of the expected 64QAM and the obtained 60QAM are shown in Fig. 4. As noted above, the outer-most, i.e., highest-energy QAM symbol, cannot occur at all, while some of inner QAM points are more likely to appear than expected.

C. Kullback-Leibler Divergence

The above Example 2 is an extreme case of ultra-short block length which leads to certain QAM symbols not occurring at all. In Fig. 5, we generalize this analysis to a wider range of block lengths and to both intra-DM and inter-DM pairing that were introduced in the context of Fig. 3. The Kullback-Leibler (KL) divergence between the expected 2D QAM distribution and the actually obtained PMF is used as measure of difference in distributions [27, Sec. 2.3]. We generated $n_{QAM} = 100000$ QAM symbols x_i for $i = \{1, \ldots, n_{QAM}\}$ from CC sequences of different length n and computed the KL divergence D_{KL} as an average over all symbols as

$$D_{KL} = \sum_{x' \in X} P_X(x') \log_2 \frac{P_X(x')}{\frac{1}{n_{QAM}} \sum_{i=1}^{n_{QAM}} \delta(x_i, x')}, \quad (1)$$

where $P_X(\cdot)$ is the expected QAM distribution obtained from the Cartesian product of the shaped ASK PMF, and the denominator is the empirical distribution obtained over n_{QAM} symbols. The Kronecker delta function is defined as

$$\delta(a, b) = \begin{cases} 0 & \text{if } a \neq b, \\ 1 & \text{if } a = b. \end{cases} \quad (2)$$

For the results of Fig. 5, the amplitude distribution was fixed to $[0.4, 0.3, 0.2, 0.1]$, corresponding to the QAM PMF used in Example 2. Using intra-DM amplitude-pairing, a clear difference between the expected and the empirical QAM distribution is observed. When the block length is increased to a few hundred symbols, the KL divergence saturates at approximately 4×10^{-4}. The reason for it not reaching zero is the randomness in the QAM quadrant bits that leads to non-CC QAM sequences although the 1D amplitudes are indeed CC. Inter-DM pairing, which combines the outputs of two independent DMs, gives approximately constant KL divergence and most notably, does not lead to the large KL divergence at short block lengths. This can be explained by the fact that independent DM outputs are combined and thus, no constraints are imposed on which amplitude pairs can be generated.

D. Interpretation within the EGN Model

This finding of infeasible and rare QAM symbols is now applied to fiber transmission. The EGN model predicts an SNR decrease with increase in kurtosis [9], which is the standardized fourth moment of the QAM input X and is defined as

$$\mu_4 = \frac{\mathbb{E}\left(|X - \mu|^4\right)}{\mathbb{E}\left(|X - \mu|^2\right)^2}, \quad (3)$$

where X has mean μ and $\mathbb{E}(\cdot)$ denotes expected value. Figure 6 shows the kurtosis μ_4 over block length n for the above setting where KL divergence was analyzed. We observe that in the length regime of up to approximately $n = 100$ symbols where intra-DM pairing caused a large KL divergence (see Fig. 5), the kurtosis is reduced, whereas it is constant around 1.65 for inter-DM pairing. Hence, some of the SNR improvement for short block lengths that will be presented in Sec. IV-B can
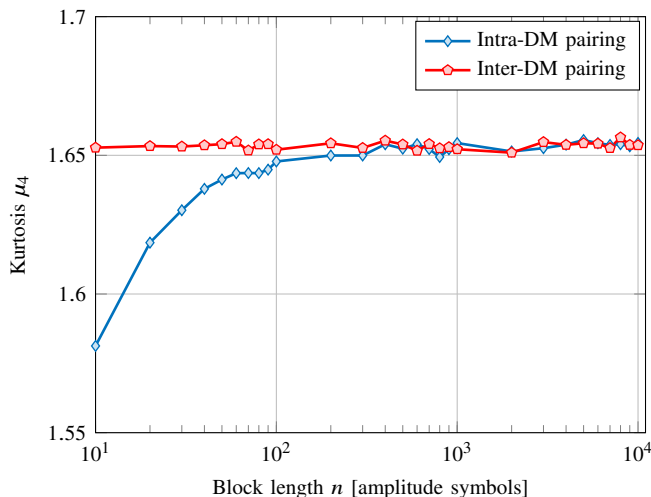
Fig. 6. 2D kurtosis versus CCDM block length for the two pairing strategies introduced in Fig. 3. The considered amplitude distribution is $[0.4, 0.3, 0.2, 0.1]$, which gives shaped 64QAM.

be attributed to the empirically generated QAM distribution having smaller kurtosis than for long blocks. However, as we will see in Sec. IV, this only partly explains the SNR dependence on shaping block length that extends into length regimes where the kurtosis has reached its asymptotic value.

III. PROPERTIES OF CONSTANT-COMPOSITION SEQUENCES FOR VARYING BLOCK LENGTHS

In the following, we analyze some important properties of compound CC sequences that are generated by varying the constituent CCDM length while keeping the per-block distribution fixed. This setting allows to isolate some properties that will be relevant for the study of the nonlinear fiber channel in Sec. IV.

A. Inherent Shuffling

An important realization when comparing block lengths for the same distribution is that all sequences that are generated by concatenating several short CCDM outputs can also come from a single long DM. The inverse statement, however, is not true as there are sequences of a long DM that cannot be the result of several short DMs. When looking at the temporal properties of these two cases, an obvious difference is that for many short CCDMs, long runs of consecutive symbols are prohibited since the CC principle must be fulfilled within every constituent sequence. This implies that using multiple short-DM blocks imposes some kind of shuffling of the amplitudes and simultaneously prevent a concentration of identical symbols, which does not necessarily have to occur for long blocks. We illustrate this property in the following example.

Example 3 (Limited Concentration of Identical Symbols): Consider two CCDMs with $n = 10$ and $n = 1000$, both targeting the distribution $[0.4, 0.3, 0.2, 0.1]$ for the amplitudes $[\alpha, \beta, \gamma, \delta]$. The overall length of the compound sequence to be generated is set to 1000 symbols, which means that 100

CCDM outputs are concatenated for $n = 10$, whereas for the long DM a single output is sufficient. Considering the short-DM case, the rarest amplitude δ must occur exactly once in every ten-symbol block. Over the entire sequence of length 1000, this gives a maximum temporal spacing of two occurrences of δ of 19 when one CCDM output begins with δ and the next one ends with this symbol. For the most likely symbol α, at most eight consecutive occurrences are feasible, which would be the case if a DM output ends with four times α and the next one begins with four times α. This shuffling, or equivalently limited concentration of symbols, holds over all 100 CCDM blocks and thus over the entire sequence of length 1000 as long as no interleaving is performed. In contrast, much longer runs of the same amplitude are feasible for the single CCDM output with $n = 1000$, and rare symbols such as δ can be spaced further apart. An extreme case of a compound sequence that is impossible to be generated by several short DMs is when all amplitudes are lumped together, i.e., the output is $\underline{a} = [\underbrace{\alpha, \ldots, \alpha}_{400\times}, \underbrace{\beta, \ldots, \beta}_{300\times}, \underbrace{\gamma, \ldots, \gamma}_{200\times}, \underbrace{\delta, \ldots, \delta}_{100\times}]$.

We emphasize that these temporal properties are the only difference between using many short DMs and a few (or just one) long DM when the same distribution and the same compound sequence length is used. Thus, the observation that the effective SNR after fiber transmission varies with CCDM block length (while keeping all other parameters unchanged) can only be attributed to these time-dependent properties. Further note that, for a purely linear channel with additive white Gaussian noise only, such length aspects would not have any impact on the channel and thus, the SNR measured after transmission would remain fixed.

B. Run Ratio

In the following, we introduce a heuristic for compound CC sequences that is intended to describe, to some extend, these time-dependent properties. The idea is that the shuffling of a shaped CC sequence can be partially represented by how often a transmit symbol differs from its precursor, or equivalently by the number of *runs* within that sequence. The concept of runs is known from run tests used for determining randomness, see for example [28, Ch. 12] and [29]. We normalize the number of runs by the number of simulated QAM symbols n_{sim} in the overall sequence in order to obtain a figure of merit that allows to compare different compound sequence lengths with respect to their concentration of symbols. We refer to this metric as *run ratio* and define it as

$$\texttt{runratio} = \frac{1}{n_{\text{sim}}} \left(1 + \sum_{i=2}^{n_{\text{sim}}} \bar{\delta}\left(x_{i-1}, x_i\right) \right), \qquad (4)$$

where $\bar{\delta}(a, b)$ is the inverse of the Kronecker delta function of (2),

$$\bar{\delta}(a, b) = \begin{cases} 1 & \text{if } a \neq b, \\ 0 & \text{if } a = b. \end{cases} \qquad (5)$$

The factor $\frac{1}{n_{\text{sim}}}$ in (4) is the aforementioned normalization and the addend 1 in front of the sum ensures that the run beginning with the first symbol x_1 is also counted. A large

`runratio` means that there are few consecutive occurrences of identical symbols, whereas a small `runratio` indicates that a sequence has many runs of identical symbols and thus, little shuffling.

Example 4 (Run Ratios): Consider a sequence of $n_{sim} = 10$ QPSK symbols with the four unit-energy constellation points denoted as $[\alpha, \beta, \gamma, \delta]$. The sequences $\underline{a}_1 = [\alpha\beta\gamma\delta\alpha\beta\gamma\alpha\beta\alpha]$ and $\underline{a}_2 = [\alpha\beta\alpha\gamma\alpha\delta\gamma\beta\alpha\beta]$ each have 10 runs (of length 1), and thus a run ratio of `runratio` $= 10/10 = 1$. The sequence $\underline{a}_3 = [\alpha\alpha\alpha\alpha\beta\beta\beta\gamma\gamma\delta]$ has 4 runs, giving a run ratio of $4/10 = 0.4$.

We further introduce an absolute-value version of the run ratio of (4) and define it as

$$\texttt{runratio}_{||} = \frac{1}{n_{sim}}\left(1 + \sum_{i=2}^{n_{sim}} \bar{\delta}\left(|x_{i-1}|, |x_i|\right)\right). \quad (6)$$

A further extension is to consider only the argument (i.e., phase) of the complex symbols,

$$\texttt{runratio}_{\angle} = \frac{1}{n_{sim}}\left(1 + \sum_{i=2}^{n_{sim}} \bar{\delta}\left(\angle x_{i-1}, \angle x_i\right)\right). \quad (7)$$

Example 5 (Run Ratios: Absolute Value and Argument): Consider the QPSK sequences of Example 4. In this trivial example, `runratio`$_{||}$ is always 0.1 because all symbols have the same amplitude, while `runratio`$_{\angle}$ is identical to the regular `runratio` of (4) because no two QPSK symbols share the same phase. This would obviously be different for higher-order QAM.

Note that the objective of the heuristic run ratios of Eqs. (4), (6), and (7) is not to exactly determine the absolute SNR variation after nonlinear fiber transmission, but to give some insight into whether the SNR variation with block length is in fact due to the temporal properties that is observed in a compound CC sequence. We will use these heuristics in the numerical analysis of the subsequent section to present strong arguments that it is indeed the induced structure when concatenating short CCDMs that leads to high effective SNR.

IV. NUMERICAL ANALYSIS

A. Simulation Setup

An idealized multi-span WDM fiber system is simulated to investigate the nonlinear fiber effects for varying block lengths. The simulation parameters are given in Table I. We focus on 64QAM and consider for illustration purposes only CC sequences with the fixed amplitude distribution $[0.4, 0.3, 0.2, 0.1]$, which was chosen as a coarse approximation to a Maxwell-Boltzmann distribution that allows to generate valid, i.e., integer-valued, composition for all block lengths that are multiplies of ten. We note that the presented findings are expected to be similar for other high-order QAM formats and shaped distributions. The shaped symbols in both polarizations are generated from independent random data, and each WDM channel is modulated separately. When explicitly stated, an ideal interleaver that randomly shuffles the QAM symbols within each FEC block is included. The length per FEC word is 10800 64QAM symbols (64800 bits).

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Modulation	64QAM
Pol.-mux.	yes (independent pol's)
Symbol rate	32 GBd
WDM spacing	50 GHz
WDM channels	5
Pulse shape	root-raised cosine
Roll-off	10%
Fiber length	80 km
No. of spans	10
Fiber loss	0.2 dB/km
Dispersion	17 ps/nm/km
Nonlinearity γ	1.37 1/W/km
Power per ch.	-0.5 dBm
EDFA noise figure	6 dB
Oversampling	8×
QAM symbols per run	~500,000
No. of simulation runs	10
Amplitude PMF	[0.4, 0.3, 0.2, 0.1]
Block length n	$10, \ldots, 10000$

After signal generation with root-raised cosine pulse shaping, the dual-polarization WDM signal is transmitted over 10 spans of standard single-mode fiber without polarization mode dispersion. Each span is followed by an Erbium-doped fiber amplifier (EDFA) with 6 dB noise figure. The propagation is simulated with the symmetrized split-step Fourier method with an adaptive step size such that the maximum nonlinear phase shift is 10^{-3} rad. At the receiver, the center channel is ideally filtered with a matched filter, chromatic dispersion is fully compensated, and any constant phase offset is removed in a genie-aided approach (see below). The effective SNR, averaged over both polarizations, is estimated from all pairs of the transmitted symbols $\underline{x}$ (whose constellation has unit average energy) and received symbols $\underline{y}$ as

$$\text{SNR}_{\text{eff}} = 1/\text{var}(h\underline{y} - \underline{x}), \quad (8)$$

where $\text{var}(\cdot)$ denotes variance and h is a complex number that removes any constant phase offset by applying a single phase rotation and that scales the received symbols such that their conditional mean values are aligned with those of the transmit symbols, which is a similar approach to [13, Sec. III-B]. The effective SNR includes all additive noise from the in-line amplifiers as well as nonlinear interference. A simulation run with approximately 500,000 QAM symbols is repeated ten times and the averaged effective SNR over all simulations is used as figure of merit in the following. We also show a 95% confidence interval in the results via error bars.

B. Results

1) Block Length Analysis: Figure 7 shows the effective SNR as defined in (8) versus the CCDM block length n for the above fiber system. We consider the two different amplitude-to-QAM pairing strategies introduced in the context of Fig. 3 as well as interleaving within each FEC codeword. As shown previously in [11] for a different fiber setup, a clear dependence of SNR on the block length is observed
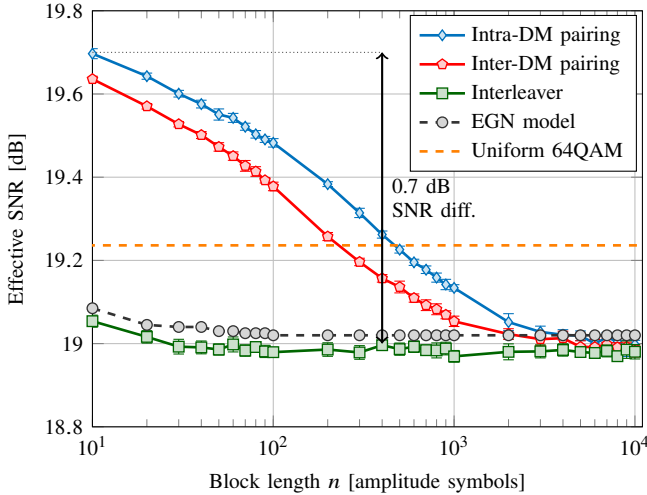
Fig. 7. Effective SNR after transmission in dB versus CCDM block length for 64QAM with one-sided amplitude distribution [0.4, 0.3, 0.2, 0.1]. Uniform 64QAM simulations and EGN model results for intra-DM pairing are included as references.

when no interleaver is present and for the conventional intra-DM pairing (blue curve). The gap between ultra-short CCDM ($n = 10$) and long CCDM ($n = 10000$) amounts to 0.7 dB SNR. The effective SNR of shaped 64QAM with CCDM is approximately equal to that of uniform signaling at $n = 500$. For longer block lengths, the SNR of shaped QAM is worse than for uniform, which is precisely what has been observed in the literature, see the introduction for a list of references. When the inter-DM pairing strategy is used (red curve in Fig. 5), SNR is also inversely proportional to n, but with approximately 0.06 dB offset compared to intra-DM pairing up to a block length of about 1000. We believe that the reason for this difference is that the inherent shuffling discussed in Sec. III-A is not as distinct when the QAM symbols are combined from two independent DMs, as it is done for inter-DM pairing. When an interleaver is included (green curve), the block-length dependence of the SNR has completely disappeared except for a very small difference of 0.05 dB SNR at very short lengths where the distribution is different from the one at larger n, see Fig. 5. This means that the SNR decrease with increasing n reported in Fig. 7 is due to some nonlinear temporal interactions that occur within each DM block and also between adjacent blocks.

Further included in Fig. 7 are the results of the EGN model [18], evaluated with 100,000 Monte Carlo samples and including self- and cross-channel interference, for the intra-DM pairing strategy, for which the kurtosis was analyzed in Fig. 6. As expected, the EGN curve is mostly flat as the EGN model is by design unable to predict time-dependent NLI effects.

We stress that we only show effective SNR in Fig. 7, which does not take into account the DM rate loss. The best operating point in terms of maximizing the achievable information rate (AIR) for finite-length CCDM is not at ultra-short lengths because the CCDM rate loss is prohibitively large in this regime, see, e.g., [3, Fig. 6]. In further simulations (not shown

herein) we find that the optimal block length for the considered system is at approximately 500 symbols, with longer block lengths having slightly lower AIR, which is in good agreement with [11, Fig. 7].

When we recall that it was argued in Sec. III-A that few long DMs can generate the same sequence as many short DMs, the question might arise why the SNR in Fig. 7 is worse for large n than for small n. A possible answer is that the probability of a particular CCDM sequence being generated is $\mathcal{A}^{-n}$ where $\mathcal{A}$ is the number of shaped amplitudes, which is 4 in the case of 64QAM. Even for a moderate length of $n = 200$, this number is $3.88 \cdot 10^{-121}$, which shows how unlikely it is that such a CCDM happens to output a sequence that has the same temporal properties (e.g., number of runs) as an ultra-short CCDM.

The implications of these findings are twofold. Firstly, if an interleaver is included in the system, all blocks lengths perform approximately the same; there is no block-length-dependent SNR but only an approximately constant one. Secondly, as outlined in Sec. III-A, any group of short-length CCDM blocks can, in principle, be generated by a single long CCDM. Therefore, it should be possible for a fixed large n to modify which CCDM sequences are selected such that an SNR gain is achieved. It is an open research problem to find algorithms that enable such a nonlinearity-aware mapping and also to identify a priori which sequences are good and which ones are bad. Note that this can only be realized without additional rate loss when the DM mapping function is not bijective, i.e., when there are output sequences that cannot be reached regardless of the binary input word. As indicated in Fig. 5 by the SNR results for interleaving, no structure, i.e., random shuffling, seems not to be the way forward. This means that some structure in the output sequence (as it is introduced by concatenating many short CCDMs) is beneficial for increasing the SNR after fiber transmission. A first demonstration of nonlinearity-tailored distribution matching is presented in [26] where the construction of an MPDM is modified to increase the SNR of the worst shaping block as well as the average SNR.

2) Run Ratio Analysis: In the following, we use the run ratios defined in (4), (6), and (7) as proxies to study the structure within the DM codewords. Figure 8 shows, for the same simulation data that was used in Fig. 7, run ratio as a function of block length. Focusing on the regular run ratio `runratio` of (4) first, we observe that short blocks (which resulted in high SNR) indeed have a higher run ratio. We recall that for a fixed sequence length, an increase in run ratio means a larger number of runs and thus, some shuffling in the sequence. In combination with the results presented in Fig. 7, this suggests for the considered setup that when long runs of identical symbols are prohibited, or equivalently when the concentration of symbols is limited, a high effective SNR is achieved. Note that Fig. 8 shows a block-length-dependent run ratio only up to approximately $n = 100$, and simulations with the average run length as figure of merit found qualitatively the same trend as `runratio`. This indicates that more refined metrics are required to fully capture the time-dependent nonlinear interactions of longer
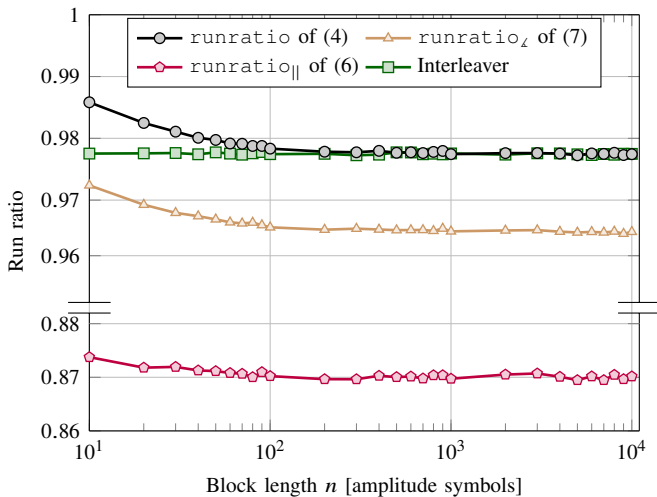
Fig. 8. Run ratios versus block length n. Only intra-DM pairing is considered.

blocks. For example, any metric that is based only on the number of runs does not differentiate between whether a high- or low-energy symbol constitutes a run, which could, however, have a significant impact on the magnitude of the induced nonlinear interactions. Using the notation of Example 4, the two sequences $[\alpha\alpha\alpha\alpha\beta\gamma\beta\gamma\delta\beta]$ and $[\alpha\alpha\beta\alpha\alpha\beta\beta\gamma\delta\gamma]$ have the same run ratio of 0.7, but there is no reason to assume that they induce an identical amount of nonlinear interference. Furthermore, we note that when the entire QAM sequences is sorted prior to transmission, the run ratio does not give a clear relationship with effective SNR.[3]

When an interleaver is inserted into the system, the run ratio is independent of the block length, as expected from the approximately flat SNR-curve of Fig. 7. The run ratio for inter-DM pairing (not shown in Fig. 8) follows roughly the same behavior as intra-DM pairing, indicating that the heuristic run ratio is able to qualitatively describe the trend of SNR dependence, but is unable to predict the exact SNR value. Further included in Fig. 8 for intra-DM pairing are the absolute-value and phase run ratios defined in (6) and (7), respectively. We observe that runratio$_\parallel$ has very little variation for the considered block lengths, whereas runratio$_\angle$ follows the same behavior as runratio. This could suggest that amplitude jumps are not the main reason for the SNR variation with block length, but phase variations have a stronger impact. A thorough analysis of this aspect is part of our ongoing work.

V. CONCLUSIONS

We have investigated constant-composition sequences of varying lengths and have confirmed in extensive simulations of a multi-span WDM system that the effective SNR after transmission is indeed decreasing with increasing block length, as previously noted in [11]. The reasons for this effect are

twofold. Firstly, for very short blocks, combining two shaped amplitudes into the shaped QAM payload, as it is typically done in the PAS framework, can lead to certain high-energy symbols occurring more rarely than for longer blocks. Hence, varying the CCDM block length for a fixed amplitude distribution can impact the QAM distribution, which in turn impacts the fiber nonlinearities and thus the SNR. Secondly, and more importantly, we have studied properties of long compound sequences that consists of many short CCDM blocks, or just a few long blocks. We explain that a key characteristic of short sequences is an inherent shuffled structure which prohibits long runs of identical symbols. To further study this temporal property, a heuristic called run ratio is introduced which describes the number of consecutive-symbol runs in a sequence. A numerical study of this metric for compound sequences of short-to-moderate length suggests that such a limited symbol concentration is beneficial as weaker nonlinear fiber effects occur. We believe that this finding could motivate and facilitate the search for signaling schemes that suppress the generation of nonlinear interference.

REFERENCES

[1] G. Böcherer, P. Schulte, and F. Steiner, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4651–4665, Dec. 2015.

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

[3] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, and K. Parsons, "Multiset-Partition Distribution Matching," *IEEE Trans. Commun.*, vol. 67, no. 3, Mar. 2019.

[4] M. Pikus and W. Xu, "Arithmetic Coding Based Multi-Composition Codes for Bit-Level Distribution Matching," *arXiv*, Apr. 2019. [Online]. Available: https://arxiv.org/abs/1904.01819

[5] Y. C. Gültekin, W. van Houtum, A. Koppelaar, and F. M. Willems, "Enumerative sphere shaping for wireless communications with short packets," *arXiv*, Mar. 2019. [Online]. Available: https://arxiv.org/abs/1903.10244

[6] R. Laroia, N. Farvardin, and S. A. Tretter, "On optimal shaping of multidimensional constellations," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1044–1056, Jul. 1994.

[7] J. Cho, "Prefix-free code distribution matching for probabilistic constellation shaping," *IEEE Transactions on Communications*, 2019, pre-print.

[8] R. Dar, M. Feder, A. Mecozzi, and M. Shtaif, "On shaping gain in the nonlinear fiber-optic channel," in *Proc. IEEE International Symposium on Information Theory (ISIT)*. Honolulu, HI, USA: pp. 2794–2798, Jun. 2014.

[9] T. Fehenberger, A. Alvarado, G. Böcherer, and N. Hanik, "On probabilistic shaping of quadrature amplitude modulation for the nonlinear fiber channel," *Journal of Lightwave Technology*, vol. 34, no. 22, pp. 5063–5073, Nov. 2016.

[10] E. Sillekens, D. Semrau, G. Liga, N. A. Shevchenko, Z. Li, A. Alvarado, P. Bayvel, R. I. Killey, and D. Lavery, "A simple nonlinearity-tailored probabilistic shaping distribution for square QAM," *Optical Fiber Communication Conference*, 2018.

[11] A. Amari, S. Goossens, Y. C. Gultekin, O. Vassilieva, I. Kim, T. Ikeuchi, C. Okonkwo, F. M. J. Willems, and A. Alvarado, "Introducing Enumerative Sphere Shaping for Optical Communication Systems with Short Blocklengths," *arXiv:1904.06601 [cs, math]*, Apr. 2019.

---

[3]Albeit theoretically interesting, we consider this case somewhat unrealistic since by sorting every sequence, the amount of information that is conveyed over the channel becomes vanishingly small.

[12] J. Cho, S. Chandrasekhar, R. Dar, and P. J. Winzer, "Low-complexity shaping for enhanced nonlinearity tolerance," in *Proc. European Conference on Optical Communications (ECOC)*. Düsseldorf, Germany: Paper W.1.C.2, Sep. 2016.

[13] 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," *Journal of Lightwave Technology*, vol. 34, no. 7, pp. 1599–1609, Apr. 2016.

[14] A. Ghazisaeidi, I. F. d. J. Ruiz, R. Rios-Müller, L. Schmalen, P. Tran, P. Brindel, A. C. Meseguer, Q. Hu, F. Buchali, G. Charlet, and J. Renaudier, "Advanced C+L-band transoceanic transmission systems based on probabilistically shaped PDM-64QAM," *Journal of Lightwave Technology*, vol. 35, no. 7, pp. 1291–1299, Apr. 2017.

[15] R. Maher, K. Croussore, M. Lauermann, R. Going, X. Xu, and J. Rahn, "Constellation shaped 66 GBd DP-1024QAM transceiver with 400 km transmission over standard SMF," in *Proc. European Conference on Optical Communications (ECOC)*, Gothenburg, Sweden, Sep. 2017.

[16] M. P. Yankov, K. J. Larsen, and S. Forchhammer, "Temporal probabilistic shaping for mitigation of nonlinearities in optical fiber Systems," *Journal of Lightwave Technology*, vol. 35, no. 10, pp. 1803–1810, May 2017.

[17] T. Fehenberger, D. S. Millar, T. Koike-Akino, K. Kojima, and K. Parsons, "Partition-based probabilistic shaping for fiber-optic communication systems," in *Proc. Optical Fiber Communication Conference (OFC)*, San Diego, CA, USA, Mar. 2019.

[18] A. Carena, G. Bosco, V. Curri, Y. Jiang, P. Poggiolini, and F. Forghieri, "EGN model of non-linear fiber propagation," *Optics Express*, vol. 22, no. 13, pp. 16 335–16 362, Jun. 2014.

[19] R. Dar, M. Feder, A. Mecozzi, and M. Shtaif, "Properties of nonlinear noise in long, dispersion-uncompensated fiber links," *Optics Express*, vol. 21, no. 22, pp. 25 685–25 699, Oct. 2013.

[20] E. Agrell, A. Alvarado, G. Durisi, and M. Karlsson, "Capacity of a nonlinear optical channel with finite memory," *Journal of Lightwave Technology*, vol. 32, no. 16, pp. 2862–2876, Aug 2014.

[21] R. Dar, M. Feder, A. Mecozzi, and M. Shtaif, "Pulse collision picture of inter-channel nonlinear interference in fiber-optic communications," *Journal of Lightwave Technology*, vol. 34, no. 2, pp. 593–607, Jan. 2016.

[22] O. Golani, R. Dar, M. Feder, A. Mecozzi, and M. Shtaif, "Modeling the bit-error-rate performance of nonlinear fiber-optic systems," *Journal of Lightwave Technology*, vol. 34, no. 15, pp. 3482–3489, Aug. 2016.

[23] T. Fehenberger, A. Alvarado, G. Böcherer, and N. Hanik, "On the impact of probabilistic shaping on SNR and information rates in multi-span WDM systems," in *Proc. Optical Fiber Communication Conference (OFC)*. Los Angeles, CA, USA: Paper M3C.4, Mar. 2017.

[24] A. S. Kashi, A. I. A. El-Rahman, J. C. Cartledge, and S. A. Etemad, "Extending a nonlinear SNR estimator to include shaping distribution identification for probabilistically shaped 64-QAM signals," *Journal of Lightwave Technology*, vol. 37, no. 13, pp. 3252–3260, Jul. 2019.

[25] J. Renner, T. Fehenberger, M. P. Yankov, F. Da Ros, S. Forchhammer, G. Böcherer, and N. Hanik, "Experimental comparison of probabilistic shaping methods for unrepeated fiber transmission," *Journal of Lightwave Technology*, vol. 35, no. 22, pp. 4871–4879, Nov. 2017.

[26] T. Fehenberger and A. Alvarado, "Analysis and optimisation of distribution matching for the nonlinear fibre channel," in *Proc. European Conference on Optical Communications (ECOC)*, Dublin, Ireland, Sep. 2019, available on arXiv: https://arxiv.org/abs/1907.02846.

[27] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. New York, NY, USA: John Wiley & Sons, 2006.

[28] J. Bradley, *Distribution-free statistical tests*. Prentice-Hall, 1968.

[29] A. Wald and J. Wolfowitz, "On a test whether two samples are from the same population," *The Annals of Mathematical Statistics*, vol. 11, no. 2, pp. 147–162, 1940.