IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 21, 2019, accepted February 1, 2019, date of publication February 7, 2019, date of current version March 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2898103

Probabilistic Shaping Using 5G New Radio Polar Codes

ONURCAN İŞCAN, RONALD BÖHNKE, AND WEN XU, (Senior Member, IEEE)
German Research Center, Huawei Technologies Düsseldorf GmbH, 80992 Munich, Germany

Corresponding author: Onurcan İşcan (onurcan.iscan@huawei.com)

This work has been partly performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union. The views expressed in this work are those of the authors and do not necessarily represent the project view.

ABSTRACT An extension to the 5G New Radio polar coding chain is proposed by introducing a shaping encoder prior to the polar encoder that improves the performance with higher order modulation via probabilistic shaping. The method is based on attaching shaping bits to the data bits that force the resulting channel input symbols (after rate matching, scrambling, and symbol mapping) to have a desired non-uniform distribution, such that a shaping gain is obtained. The shaping encoder is realized by using a polar decoder so that no additional hardware is required. The extension does not change the structure of the existing blocks in the transmission chain but only modifies the input parameters. The required changes at the receiver are also very small and can be realized by simple operations without increasing the required computational power. The simulation results validate the performance improvements on the additive white Gaussian noise channels, which can be more than 1 dB for typical scenarios.

INDEX TERMS 5G new radio, coded modulation, polar coding, probabilistic shaping.

I. INTRODUCTION

Polar codes [1] are error correction schemes that can provably achieve capacity of binary input discrete memoryless channels with low complexity successive cancellation (SC) decoding. Moreover, their finite length performance with cyclic redundancy check (CRC) aided SC-List (SCL) decoding [2] shows competitive performance compared to other modern channel codes (such as LDPC and turbo codes), which made them also attractive for practical applications. Recently, the first release of the fifth generation (5G) new radio (NR) standard has been completed by 3GPP, and polar codes were selected as the channel coding scheme for control data. Similar to legacy systems, 5G NR uses bit-interleaved coded modulation (BICM) with uniformly distributed quadrature amplitude modulation (QAM) symbols, which leads to a shaping loss and prevents the performance from approaching the Shannon limit for higher order modulation. In this work, we propose a practical modification to 5G NR polar codes such that the shaping loss is reduced and the performance with higher order modulation is improved.

Polar coded modulation is introduced in [3] and an efficient code design is considered in [4]. In these solutions, multi-level coding is preferred over BICM due to the loss introduced by the independent demapping of each bit-level for BICM. Moreover, these works did not consider signal shaping. In [5], a probabilistic shaping (PS) approach for polar codes based on [6] is studied, where a precoder is used for systematic encoding, and signal shaping is performed using a (non-binary) distribution matcher. This scheme requires multi-stage demapping and decoding, and it is based on codeword lengths that are an integer power of 2, i.e., rate-matching (e.g., puncturing or shortening) is not considered. Moreover, it is not possible to implement such a scheme, if a binary scrambler is used prior to symbol mapping (as in 5G NR), which can corrupt the desired symbol probability distribution. Another PS approach was presented in [7] by extending the ideas from [8] and [9] to higher order modulations. A precoder is used prior to polar encoding, which forces the codewords to have a desired probability distribution. The same method is applied to 5G NR polar codes in [10], but similar to [7], rate matching is not considered and the interleavers in the 5G NR chain need to be modified. Moreover, due to the special structure of this method, it is not straightforward to support 64-QAM modulation.

In this work, we extend the ideas of [7] and integrate probabilistic shaping into 5G NR polar codes in a way, such that it supports rate-matching (e.g. puncturing and shortening),

The associate editor coordinating the review of this manuscript and approving it for publication was Pietro Savazzi.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

and allows using any QAM modulation schemes of 5G NR (including 64-QAM). Moreover, the proposed scheme works with any binary scrambling operation prior to the symbol mapper, such that the desired QAM symbol distribution is obtained. Our proposal is related to the transmitter and allows to reuse existing hardware for signal shaping. We introduce a *polar shaping encoder* in the transmitter chain without altering the other elements in the chain. Moreover, the modifications at the receiver are small and can be realized by simple operations. This makes the proposed approach especially suitable for downlink communication, where the receivers of a user equipment (UE) have limited computational power.

A. ORGANIZATION AND ANNOTATIONS

The paper is organized as follows. We first introduce the system model and review the achievable rates in Sec. II. In Sec. III, we discuss how signal shaping can be combined with binary coded modulation schemes. In Sec. IV, we describe the polar coding steps of 5G NR, and our proposed modifications are presented in Sec. V. We validate our proposal with simulations in Sec. VI, and conclusions and discussions are given in Sec. VII.

In this work, $\mathbf{a}$ is a vector, A is a random variable representing the elements in $\mathbf{a}$, and P_{A} denotes their probability mass function. $\mathcal{D}$ denotes a sequence of indices, and $\mathbf{a}_{[\mathcal{D}]}$ is a vector that is built from the elements of $\mathbf{a}$ at the indices given by $\mathcal{D}$. $[\mathbf{a}\,\mathbf{b}]$ is the concatenation of two vectors $\mathbf{a}$ and $\mathbf{b}$. We follow the notation in [11] for variable names when possible.

II. SYSTEM MODEL

We consider a coded modulation scheme, where the binary message $\mathbf{a}$ (of length A) is encoded to a codeword $\mathbf{b}$ (of length B), which is then mapped to channel input symbols $\mathbf{x}$ using a symbol mapper. The encoder may contain the steps such as channel encoding, interleaving, scrambling, etc., that may be required prior to the generation of the channel input symbols. The symbol mapper maps M successive bits to a single complex valued 2^M-QAM symbol (with M bit-levels), where the real and imaginary parts are selected from the set $\{\pm 1, \pm 3, \ldots, \pm(2^{M/2}-1)\}$ according to the Gray labeling. We define $\mathcal{B}_j$ as the sequence containing the indices of the jth bit-level of the 2^M-QAM symbols, i.e.,

$$\mathcal{B}_j = \{0, M, 2M, \cdots, \left(\frac{B}{M} - 1\right) M\} + (j-1) \tag{1}$$

for $j \in \{1, \cdots, M\}$. The QAM symbols are transmitted over an additive white Gaussian noise (AWGN) channel and the receiver observes

$$\mathbf{r} = \mathbf{x} + \mathbf{w}, \tag{2}$$

where $\mathbf{w}$ contains complex Gaussian noise samples. The signal-to-noise ratio (SNR) is $\gamma = \mathrm{E}(|\mathsf{X}|^2)/\mathrm{E}(|\mathsf{W}|^2)$, where $\mathrm{E}(\cdot)$ denotes expectation. We assume a parallel symbol demapper at the receiver, which determines the log-likelihood ratios (L-values) for decoding without taking the dependencies of different bit-levels into account, as usually employed with Gray labeled BICM [12].

A. ACHIEVABLE RATES

For QAM transmission, the rates

$$R_{\mathrm{QAM}} = \mathrm{I}(\mathsf{X}; \mathsf{R}) \tag{3}$$

are achievable, where $\mathrm{I}(\cdot;\cdot)$ denotes mutual information. If binary codes are used to obtain X, the dependencies between different QAM bit-levels need to be taken into account during decoding to achieve R_{QAM}. Note that this rate depends on the probability distribution $\mathrm{P}_{\mathsf{X}}(x)$, and the maximum achievable rate (i.e. the QAM constrained capacity) can only be obtained if the symbol distribution is optimized for the channel. For AWGN channels with discrete alphabets, a Maxwell-Boltzmann (MB) distribution is nearly optimal [13]. However, many practical systems use uniformly distributed symbols, which lead to a shaping loss of up to 1.53dB.

For BICM, one usually considers parallel demapping, i.e. the dependencies between the QAM bit-levels are not considered. Accordingly, the rates

$$R_{\mathrm{BICM}} = \sum_{j=1}^{M} \mathrm{I}\left(\mathsf{B}_{[\mathcal{B}_j]}; \mathsf{R}\right) \tag{4}$$

are achievable for BICM [14, Sec. III], where $\mathsf{B}_{[\mathcal{B}_j]}$ represents the bits at the jth QAM bit-level. Similarly, a uniform distribution of the bits at each bit-level results in a sub-optimal rate on AWGN channels, and R_{BICM} can be maximized by optimizing the probability distributions of bits at each QAM bit-level.

III. SIGNAL SHAPING FOR BINARY CODED MODULATION SCHEMES

In general, binary channel encoders generate sequences with uniformly distributed bits, leading to uniformly distributed QAM symbols and thus to a shaping loss. One option to combat this loss is to explicitly use a shaping encoder, i.e., an additional (reversible) processing block in the transmitter chain. Shaping encoders force the channel input symbols to have the target probability distribution by introducing additional redundancy. By a proper design, the gain obtained by signal shaping is higher than the loss due to the additional redundancy. Note that signal shaping is beneficial especially for higher order modulation (i.e., $M > 2$) with non-binary symbol alphabets. Therefore, non-binary shaping encoders are required in general to obtain the optimal probability distribution.

A significant shaping gain can still be obtained if the optimal distribution is only roughly approximated. For example, the symbol distribution $\mathrm{P}_{\mathsf{X}}(x)$ can be approximated by a product of M binary distributions corresponding to each QAM bit-level (or $M/2$ distributions by combining the same bit-levels

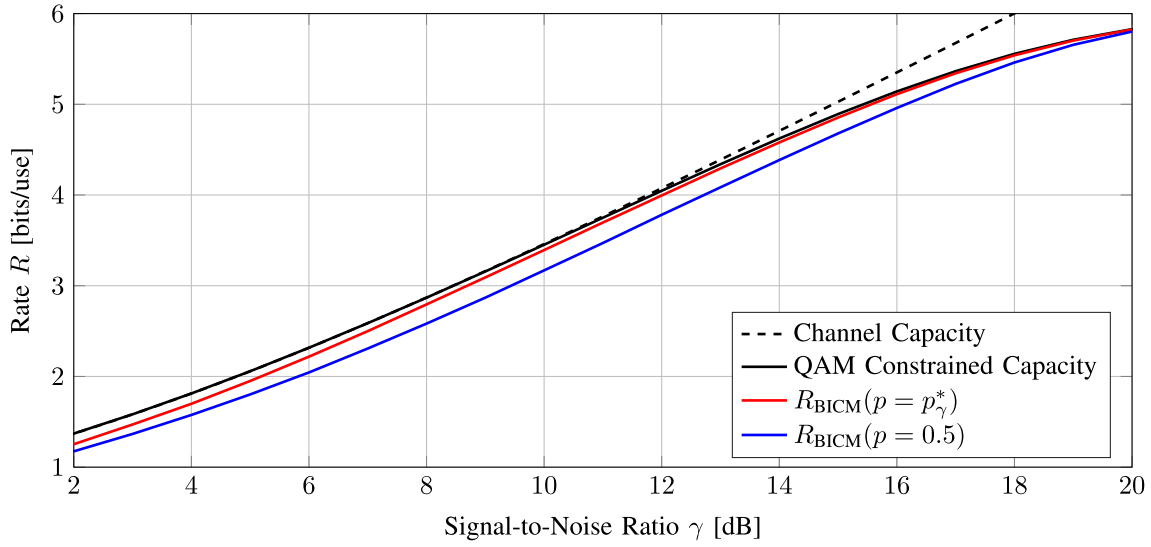

FIGURE 1. Achievable rates for 64-QAM on AWGN channels.

of the real and imaginary parts). Such a scheme can be realized by using an independent binary shaping encoder for each QAM bit-level [15]. This leads to a simple and parallelizable architecture which has implementation benefits compared to non-binary shaping encoders without significantly reducing the shaping gain.

A further simplification can be introduced by shaping only one bit-level per dimension (for real and imaginary part of the QAM symbols), such that those bit-levels contain ones with probability $p \neq 0.5$, resulting in a coarse approximation of the optimal distribution. This *single bit-level shaping* approach leads to a transmission scheme, where symbols with high energy are transmitted less frequently compared to symbols with low energy. For example, considering the QAM symbol mapper in 5G NR [16] the 3rd and 4th bit-levels distinguish between high and low energy symbols for modulation schemes with $M > 2$. Accordingly, if the probability of ones in $\mathbf{b}_{[\mathcal{B}_s]}$ is $p < 0.5$ with

$$\mathbf{b}_{[\mathcal{B}_s]} = [\mathbf{b}_{[\mathcal{B}_3]}\, \mathbf{b}_{[\mathcal{B}_4]}] \tag{5}$$

then the resulting $\mathrm{P}_X(x)$ is a coarse approximation of the MB distribution, as depicted in Fig. 2 for 64-QAM with $p = 0.25$. In this case, we can describe $\mathrm{P}_X(x)$ as

$$\mathrm{P}_X(x) = \begin{cases} \dfrac{(1-p)^2}{2^{M-2}}, & \text{if } \begin{cases} |\Re\mathfrak{e}(x)| < 2^{\frac{M}{2}-1} \\ |\Im\mathfrak{m}(x)| < 2^{\frac{M}{2}-1} \end{cases} \\ \dfrac{p^2}{2^{M-2}}, & \text{if } \begin{cases} |\Re\mathfrak{e}(x)| > 2^{\frac{M}{2}-1} \\ |\Im\mathfrak{m}(x)| > 2^{\frac{M}{2}-1} \end{cases} \\ \dfrac{p(1-p)}{2^{M-2}}, & \text{otherwise.} \end{cases} \tag{6}$$

The optimal value p^*_γ that maximizes (4) for a given SNR can be obtained numerically. Fig. 1 plots R_{BICM} for 64-QAM with $p = 0.5$ (uniform distribution) and with p^*_γ. As a reference, the figure also includes the channel capacity and

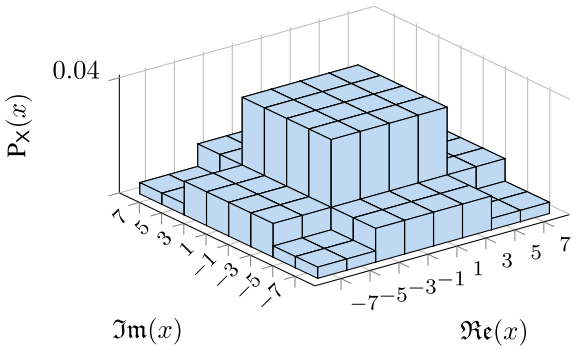

FIGURE 2. Resulting probability distribution of 64-QAM symbols, when mapped from binary sequences which contain ones with probability $p = 0.25$ at the 3rd and 4th bit levels.

the QAM constrained capacity. We observe that employing the distribution in (6) with p^*_γ already significantly improves the achievable rates compared to uniform distribution and closes most of the gap to the theoretical limits. In the next sections, we show how this gain can be obtained in practice using simple modifications of 5G NR polar codes.

IV. 5G NR POLAR CODING

Below we summarize how a message vector $\mathbf{a}$ (of length A) is encoded to the binary codeword $\mathbf{b}$ (of length B) and mapped to 2^M-QAM symbols in 5G NR. The required parameters for each step are defined in [11], where some of the parameters (such as activation flags for interleavers or CRC lengths) have fixed values based on the transmit direction (i.e., uplink or downlink). The rest of the internal parameters can be derived from A and B for uplink and downlink. For the sake of simplicity, we consider a single transport block

without code block segmentation, and describe the procedure in four steps, as depicted in Fig. 3.

A. PREPROCESSING

In this step, L bits of CRC (denoted as $\mathbf{p}$) are appended to the payload vector $\mathbf{a}$, resulting in the vector

$$\mathbf{c} = [\mathbf{a}\,\mathbf{p}] \tag{7}$$

of length $K = A + L$. For the downlink, $L = 24$ CRC bits are used, whereas for the uplink we have $L = 6$ if $A < 20$ and $L = 11$, otherwise. The vector $\mathbf{c}$ is then interleaved to produce

$$\mathbf{c}' = \Pi_{\mathrm{P}}(\mathbf{c}) \tag{8}$$

according to [11, Table 5.3.1.1-1]. This polar interleaver ensures that each CRC bit is placed after its relevant message bit, such that the decoder can perform early termination if an error is detected during decoding. Note that Π_{P} is only activated for the downlink, and for the uplink $\mathbf{c}'$ equals $\mathbf{c}$.

B. POLAR ENCODING

In this step, first a length-N vector $\mathbf{u}$ is generated, which is the input to the polar transform. Here, N is the mother codeword length that is a function of $K = A + L$ and B. Moreover, N is an integer power of 2 larger than 32. For the uplink, N can be up to 1024, whereas for the downlink the maximum value of N is limited to 512. $\mathbf{u}$ contains $\mathbf{c}'$ at the indices $\mathcal{Q}_I$, and zeros at the indices $\mathcal{Q}_F$, i.e.,

$$\mathbf{u}_{[\mathcal{Q}_I]} = \mathbf{c}' \quad \mathbf{u}_{[\mathcal{Q}_F]} = \mathbf{0}, \tag{9}$$

where $\mathcal{Q}_I$ and $\mathcal{Q}_F$ contain polar sub-channel indices that are used for transmission of data (message and CRC) bits and frozen bits, respectively. $\mathcal{Q}_I$ and $\mathcal{Q}_F$ are obtained from the polar sequence defined in [11, Table 5.3.1.2-1]. Note that $\mathcal{Q}_I$ is in ascending order of reliability. Moreover, $\mathcal{Q}_F$ is built from three sequences, $\mathcal{Q}_{F_p}$, $\mathcal{Q}_{F_s}$ and $\mathcal{Q}_{F_r}$, where $\mathcal{Q}_{F_p}$ and $\mathcal{Q}_{F_s}$ contain indices of the sub-channels that are frozen due to puncturing and shortening, respectively. $\mathcal{Q}_{F_r}$ includes indices with the lowest reliabilities that are selected from the remaining indices. Depending on the rate matching scheme, at least one of $\mathcal{Q}_{F_p}$ or $\mathcal{Q}_{F_s}$ is an empty sequence, i.e., puncturing and shortening are not applied at the same time.

The detailed procedure for obtaining these sequences are described in [11, Sec. 5.3.1.2], which basically depends on the choice of A and B. Note that for uplink transmission with $K < 26$, $\mathcal{Q}_I$ may contain three additional indices for transmitting parity check (PC) bits that are generated from the data bits with a cyclic shift register. PC bits are not used for $K \geq 26$ or for the downlink.

After constructing $\mathbf{u}$, a polar transform is performed to obtain the mother codeword

$$\mathbf{d} = \mathbf{u}\mathbf{G} \tag{10}$$

where $\mathbf{G}$ is the $(\log_2 N)$-th Kronecker power of $\mathbf{G}_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$.

C. RATE MATCHING

This step consists of three parts. First, a sub-block interleaver reorders the vector $\mathbf{d}$ to the vector

$$\mathbf{y} = \Pi_{\mathrm{SB}}(\mathbf{d}) \tag{11}$$

by dividing $\mathbf{d}$ into 32 sub-blocks and interleaving the sub-blocks according to [11, Table 5.4.1.1-1].

Secondly, a bit selection procedure is performed as

$$\mathbf{e} = \mathbf{y}_{[\mathcal{E}]} \tag{12}$$

with

$$\mathcal{E} = \begin{cases} \{0, \cdots, N-1, 0, 1, \cdots\}, & \text{if } B > N \\ \{N-B, \cdots, N-1\}, & \text{else if } K/B \leq 7/16 \\ \{0, \cdots, B-1\}, & \text{otherwise,} \end{cases} \tag{13}$$

which outputs a vector of length B. For $B > N$ repetition is performed. Otherwise, depending on K/B, either the first $N-B$ bits are punctured, or the last $N-B$ bits are shortened. Note that the values of the punctured bits are unknown to the receiver, i.e., zeros are used as corresponding L-values. On the other hand, the shortened bits are always zero, which is known to the receiver.

Finally, $\mathbf{e}$ is interleaved to

$$\mathbf{f} = \Pi_{\mathrm{CB}}(\mathbf{e}) \tag{14}$$

with a code-bit interleaver. Π_{CB} is a length-B triangular interleaver defined in [11, Sec. 5.4.1.3], which is mainly considered to improve the performance for higher order modulation schemes. Note that Π_{CB} is activated only for the uplink, and $\mathbf{f}$ equals $\mathbf{e}$ for the downlink.

The rate matching step may be summarized as

$$\mathbf{f} = \mathbf{d}_{[\mathcal{D}]}, \tag{15}$$

where $\mathcal{D}$ is a sequence that can be constructed from Π_{SB}, Π_{CB} and $\mathcal{E}$.

D. MODULATION

In this step, $\mathbf{f}$ is scrambled with a user specific binary scrambling vector $\mathbf{v}$ (of length B) by an element-wise modulo-2 sum.

$$\mathbf{b} = \mathbf{f} \oplus \mathbf{v} \tag{16}$$

Then, a QAM symbol mapper is used to map $\mathbf{b}$ to channel input symbols

$$\mathbf{x} = g_M(\mathbf{b}) \tag{17}$$

according to 2^M Gray labeling, described in [16, Sec. 5.1]. For the downlink, $M = 2$ is used, but for the uplink higher order modulation is also supported if the physical uplink shared channel (PUSCH) is used.

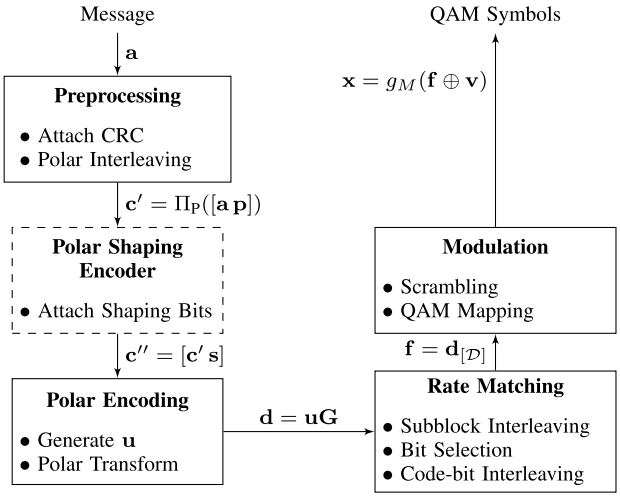

FIGURE 3. Block diagram of the 5G NR polar coding chain. The dashed box is the proposed addition to the chain, which appends shaping bits to its input.

V. SIGNAL SHAPING FOR 5G NR POLAR CODES

In this section, we propose an extension to the 5G NR polar transmission chain, such that it supports single bit-level shaping. For this aim, we introduce a *shaping encoder* between the preprocessing and polar encoding blocks, as depicted in Fig. 3. The shaping encoder appends additional shaping bits $\mathbf{s}$ (of length S) to the vector $\mathbf{c}'$, such that the polar encoding is performed on

$$\mathbf{c}'' = [\mathbf{c}'\,\mathbf{s}] \tag{18}$$

and the following blocks basically operate as if the input length is not K, but $K+S$, i.e., the shaping bits are treated as data bits that are transmitted in reliable polar sub-channels. For $S = 0$, the scheme is equivalent to the conventional 5G NR polar coding chain. The shaping bits are generated from $\mathbf{c}'$ and they do not carry new information, but they make $\mathbf{b}_{[\mathcal{B}_s]}$ (with $\mathcal{B}_s$ defined in (5)) contain ones with probability p, resulting in the symbol distribution $\mathrm{P}_X(x)$ given in (6).

Obtaining polar codewords with a non-uniform distribution of bits (by transmitting shaping bits at reliable sub-channels) was initially discussed in [8] and [9]. This idea is further extended in [7] and [10] to higher order modulation, exploiting the fact that each polar codeword can be seen as a combination of two polar codewords that are further polarized. Moreover, it is shown how the shaping bits can be obtained by using a polar decoder as a precoder. However, the approach in [7] and [10] assumes that the codeword length is an integer power of two, i.e., it does not support rate-matching operations and 2^M-QAM modulations where M is not an integer power of two (such as 64-QAM). Moreover, [10] requires the interleavers in the transmission chain to be modified.

In the following, we describe a polar shaping encoder that overcomes these problems and has the flexibility to support all codeword lengths and modulation orders included in 5G NR without changing the existing processing blocks.

A. POLAR SHAPING ENCODER

Let us define the sequence $\mathcal{D}_s$ such that

$$\mathbf{d}_{[\mathcal{D}_s]} = \mathbf{f}_{[\mathcal{B}_s]}, \tag{19}$$

i.e., $\mathcal{D}_s$ includes the bit indices of the polar mother codeword that will be mapped to the 3rd and 4th QAM bit-level after rate matching and scrambling. Similarly, let us define $\mathcal{D}_d$ as the sequence containing the indices of the mother codeword that are mapped to the remaining bit-levels. The task of the precoder is to generate and attach shaping bits $\mathbf{s}$ (of length S) to $\mathbf{c}'$ (of length K), such that

$$\mathbf{b}_{[\mathcal{B}_s]} = \mathbf{d}_{[\mathcal{D}_s]} \oplus \mathbf{v}_{[\mathcal{B}_s]} \tag{20}$$

contains ones with probability p, resulting in $\mathrm{P}_X(x)$ in (6). Concretely, we use the following steps to generate $\mathbf{s}$.

- Determine the sequences $\mathcal{Q}_I$ and $\mathcal{Q}_F$ (including $\mathcal{Q}_{F_r}$, $\mathcal{Q}_{F_s}$ and $\mathcal{Q}_{F_p}$) according to [11] for a data length $K+S$, i.e., $\mathcal{Q}_I$ contains $K+S$ indices, and $\mathcal{Q}_F$ contains $N-K-S$ indices.
- Define the first K and the last S elements of $\mathcal{Q}_I$ as $\mathcal{Q}_{I_d}$ and $\mathcal{Q}_{I_s}$, respectively.
- Use a polar decoder, where $\mathbf{\Lambda}$ (defined below) is used as the noisy observation vector containing the L-values, and $\mathbf{c}'$ and $\mathbf{0}$ are treated as frozen bits at the indices $\mathcal{Q}_{I_d}$ and $\mathcal{Q}_F$, respectively. The decoder output $\mathbf{s}$ corresponds to the bits of the sub-channels with indices $\mathcal{Q}_{I_s}$.
- Concatenate $\mathbf{c}'$ and $\mathbf{s}$ to obtain $\mathbf{c}''$.

Here, $\mathbf{\Lambda}$ is a length-N vector with the following elements:

$$\mathbf{\Lambda}_{[\mathcal{Q}_{F_p}]} = \mathbf{0} \quad \mathbf{\Lambda}_{[\mathcal{Q}_{F_s}]} = +\infty$$
$$\mathbf{\Lambda}_{[\mathcal{D}_d]} = \mathbf{0} \quad \mathbf{\Lambda}_{[\mathcal{D}_s]} = (-2 \cdot \mathbf{v}_{[\mathcal{B}_s]} + 1) \log \frac{1-p}{p}$$

The decoder basically looks for a vector $\mathbf{s}$ that leads to a codeword $\mathbf{d}$ with the following properties:

- Shortened bits at $\mathcal{Q}_{F_s}$ are always zero (corresponding to an L-value of $+\infty$).
- Bits that correspond to non-shaped bit-levels or punctured bits are one or zero with equal probability (corresponding to an L-value of 0).
- Bits at the indices $\mathcal{D}_s$ are 1 with probability p, if the corresponding bit in the scrambling vector is 0, and $1-p$ otherwise.

Accordingly, if $\mathbf{c}'' = [\mathbf{c}'\,\mathbf{s}]$ is used as the input for the polar encoder, then the resulting vector $\mathbf{b}$ after rate matching and scrambling should ideally contain ones with probability p at the indices $\mathcal{B}_s$, and the QAM symbols are distributed according to (6). This procedure can be seen as an extension of the method presented in [7], such that it supports rate-matching, code-bit interleaving and binary scrambling.

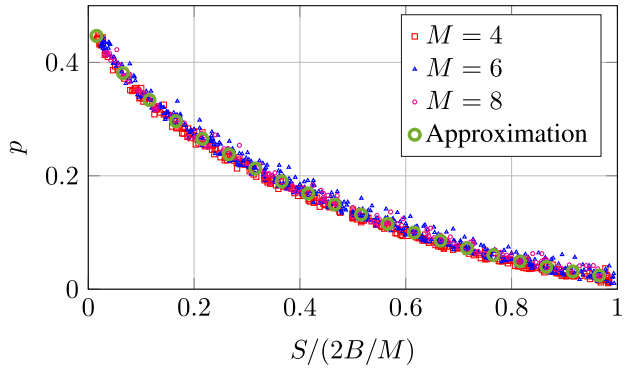

FIGURE 4. The resulting p for different choices of S, B and M. A polynomial approximation is also given.

TABLE 1. Parameters of the simulated setups with $B = 768$.

A	S	B	M	B/M	R	Δ_γ
384	192	768	4	192	2	0.51
480	115	768	4	192	2.5	0.57
576	38	768	4	192	3	0.44
672	10	768	4	192	3.5	0.35
384	192	768	6	128	3	0.80
480	115	768	6	128	3.75	0.86
576	58	768	6	128	4.5	0.68
672	10	768	6	128	5.25	0.45
384	77	768	8	96	4	1.06
480	115	768	8	96	5	1.10
576	58	768	8	96	6	0.87
672	29	768	8	96	7	0.62

TABLE 2. Parameters of the simulated setups with $B/M = 120$.

A	S	B	M	B/M	R	Δ_γ
180	120	480	4	120	1.5	0.38
240	96	480	4	120	2	0.41
300	36	480	4	120	2.5	0.38
360	36	480	4	120	3	0.39
420	6	480	4	120	3.5	0.20
360	180	720	6	120	3	0.78
450	81	720	6	120	3.75	0.81
540	54	720	6	120	4.5	0.56
630	18	720	6	120	5.25	0.35
480	144	960	8	120	4	0.90
600	144	960	8	120	5	0.92
720	72	960	8	120	6	0.83
840	36	960	8	120	7	0.52

Note that hardware efficient implementations of an L-value based polar decoder usually make use of the so-called min-sum approximation, which simplifies the check-node operations during decoding [17]. This approximation makes the decoder output independent of the scaling of $\mathbf{\Lambda}$. Assuming $p < 0.5$, $\mathbf{\Lambda}_{[\mathcal{D}_s]}$ can be therefore further simplified to

$$\mathbf{\Lambda}_{[\mathcal{D}_s]} = -2 \cdot \mathbf{v}_{[\mathcal{B}_s]} + 1, \tag{21}$$

such that the shaping encoder does not need to know the value of p.

The polar shaping encoder is a systematic encoder that generates and appends the shaping bits at the end of its input. As $\mathcal{Q}_I$ is in ascending order of reliability, the shaping bits are transmitted in the most reliable sub-channels, similar to [10], which correspond to sub-channels with (mostly) large indices. Due to the structure of the polar transform and the successive nature of the decoding process, this allows the shaping bits to be chosen depending on the data and frozen bits, which are transmitted in sub-channels with smaller indices.

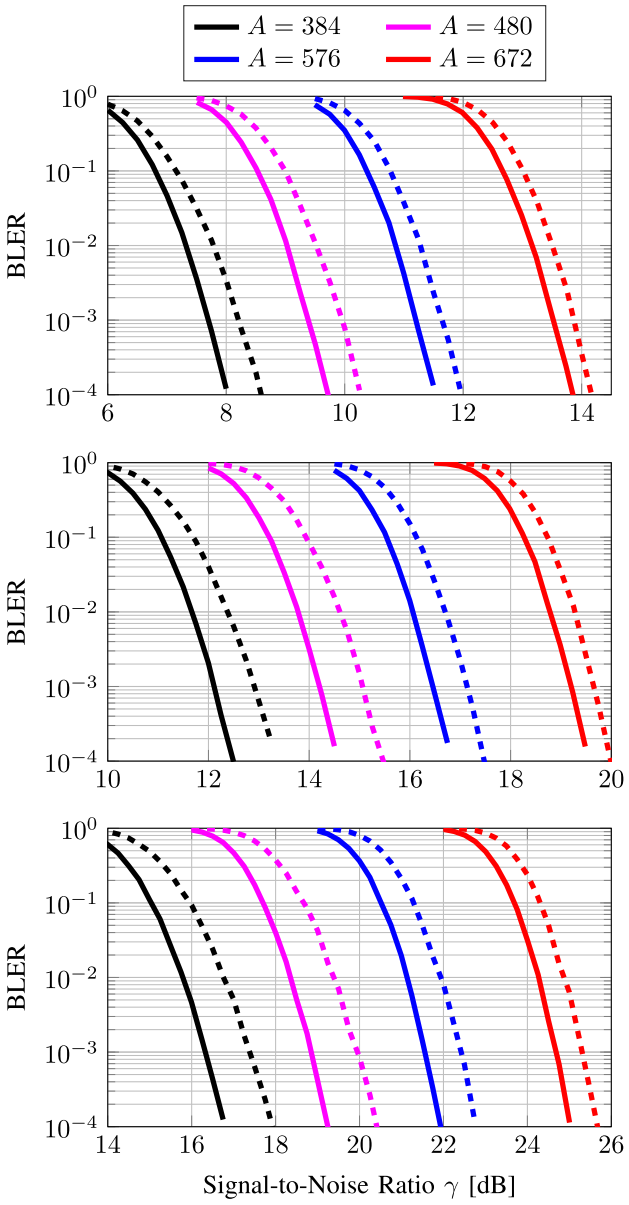

FIGURE 5. BLER performance of 5G NR polar codes for 16-QAM, 64-QAM and 256-QAM (from top to bottom) with the proposed probabilistic shaping approach (solid lines), where $A \in \{384, 480, 576, 672\}$ message bits (from left to right) are encoded to $B = 768$ codeword bits. The dashed curves show the performance without signal shaping ($S = 0$) for the same rate.

B. RELATION BETWEEN THE NUMBER OF SHAPING BITS AND THE OUTPUT PROBABILITY

The probability of ones p in $\mathbf{d}_{[\mathcal{D}_s]}$ is directly related to the amount of shaping bits S that are used by the polar shaping

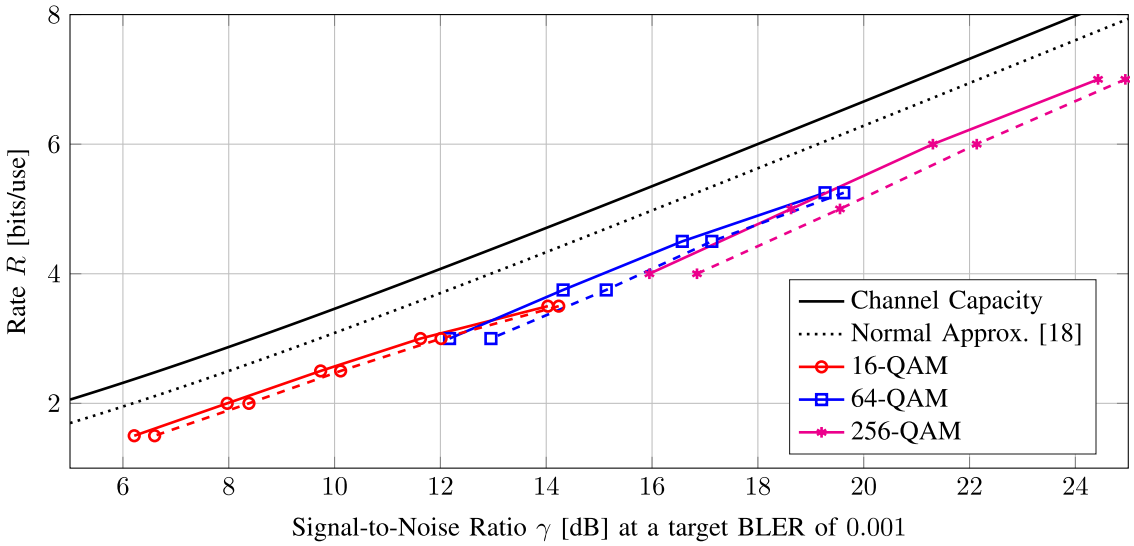

FIGURE 6. AWGN channel performance for $B/M = 120$ symbols with (solid lines) and without shaping (dashed lines).

encoder. Obviously, for $S = 0$ no shaping is performed and $p = 0.5$. The probability p diverges from 0.5 with increasing S.

If the whole codeword is shaped, the relation between S and p can be expressed asymptotically for large N as $S/N = 1 - h_2(p)$ [7], where $h_2(\cdot)$ denotes the binary entropy function. As the scheme proposed here only performs shaping on a single bit-level per real dimension, we can expect a similar relation if we replace N with $2B/M$, i.e., with the total number of bits in the third and fourth bit-levels. However, as this relation is valid asymptotically, it does not necessarily represent the finite length performance accurately. Therefore, we evaluated the resulting p for $S \in \{8, 16, \cdots, 2B/M\}$ and for different choices of B and M by using Monte-Carlo simulations. We randomly generated data bits and appended S shaping bits as described above. Then we obtained the corresponding codewords and calculated the average p. As a polar shaping encoder, we used an SCL decoder with list size 8 that employs the min-sum approximation. Results are plotted in Fig. 4, where each dot corresponds to another combination of S, B and M. We observe a similar trend for the relation between p and S normalized by $2\,B/M$. Note that this (finite length) relation can be well approximated with an analytical function, e.g. with a polynomial, as also depicted in Fig. 4.

C. MODIFICATIONS AT THE RECEIVER

The receiver requires only two small modifications. Firstly, $\mathrm{P}_X(x)$ needs to be considered in the demapping of the QAM symbols. This can be done easily by first obtaining the L-values corresponding to $\mathbf{b}$ as usual with an independent symbol demapper, and then adding a constant value of $\log((1-p)/p)$ to each element at the indices $\mathcal{B}_s$. This allows the decoder to utilize the a-priori information about the symbol distribution.

Secondly, after descrambling and rate dematching, polar decoding should be performed assuming that there are $K + S$ bits to recover (instead of K). Afterwards, the shaping bits can be discarded before performing deinterleaving and CRC-check. The decoder can also perform an early termination after the first K bits are extracted. We remark that both modifications at the receiver are simple and only lead to a negligible increase of the receiver complexity. Note that other probabilistic shaping methods may require additional shaping decoders (such as distribution dematchers) that need additional hardware at the receiver.

VI. NUMERICAL EVALUATION

The performance of the proposed scheme for AWGN channels was evaluated via Monte-Carlo simulations. We considered A message bits and $L = 24$ CRC bits (with polynomial g_{crc24c} from [11]) without polar interleaving prior channel encoding. We appended S shaping bits that are generated by using an SCL decoder (employing min-sum approximation) with list size 8. Rate matching was employed such that codewords of length B are obtained, which are mapped to 2^M-QAM symbols ($M \in \{4, 6, 8\}$) after code-bit interleaving. We allowed N to be up to 1024, and chose the parameters according to [11], assuming the input of the polar encoding step is $A + L + S$ bits. For each choice of A, B and M, we performed simulations with different number of shaping bits to obtain the optimal value of S that gives the best block error rate (BLER) performance, if an SCL decoder (employing min-sum approximation) with list size 8 is used at the receiver. As a reference, we also evaluated the performance without shaping ($S = 0$), and estimate the shaping gain Δ_γ, which reflects the SNR gain in dB at a target BLER of 10^{-3}.

We first evaluate the performance with different modulation orders and code rates for a fixed B. Fig. 5 shows the

BLER performance for the parameters given in Table 1 with $B = 768$. We observe gains between 0.35dB and 1.1dB compared to the conventional scheme without shaping.

Similarly, we also performed simulations for a fixed total amount of QAM symbols $B/M = 120$ (with parameters given in Table 2). Fig. 6 plots the performance in terms of the required SNR to achieve a target BLER of 10^{-3} with and without shaping for different modulation orders. As a reference, the figure also depicts the finite length normal approximation according to [18] for the given target BLER and the number of symbols. The proposed scheme again shows significant gains compared to the conventional 5G NR polar coding scheme.

VII. DISCUSSIONS AND CONCLUSION

In this work we proposed a modification to the 5G NR polar code transmission chain by introducing a shaping encoder. This allows the generated QAM symbols to have a non-uniform distribution for reducing the shaping loss. The proposed scheme is attractive from several perspectives:

- Except for the introduction of the shaping encoder, the transmitter chain remains almost unchanged. It can operate as described in the standard, only with different input parameters, i.e., the polar encoder and rate matcher process $A + L + S$ data bits (instead of $A + L$) to obtain B bits that are mapped to QAM symbols. Note that the overall transmission rate is not influenced by this operation, as it remains $R = A/(B/M)$ for both cases.
- The polar shaping encoder can be realized by using a polar decoder, which is already included in the transceiver chain of bidirectional communication systems. Hence, no additional hardware is required.
- The whole modification at the transmitter can be described by a single parameter S, which simplifies signaling: only S needs to be signaled in order to revert the operation at the receiver (p can be obtained from S e.g. by look-up tables). Moreover, for $S = 0$ the scheme simplifies to the existing 5G NR transmission chain.
- Only minor modifications are required at the receiver and the additional complexity is negligible. This makes the scheme especially suitable for downlink transmission.
- In contrast to other probabilistic shaping methods, the proposed scheme considers binary scrambling prior to the symbol mapper (as it is done in 5G NR). Moreover, it supports all codeword lengths and 2^M-QAM modulation schemes specified in 5G NR, which was not the case in [10].

Recall that there is an optimal number of shaping bits S for each choice of message length A, codeword length B and modulation order M. The optimal S can be obtained by numerical simulations and stored in a look-up table. However, we observed in our simulations (not included in this work) a certain correlation between S and A/B, i.e., it may be possible to derive an analytical function for the calculation of S from A and B, similar to other parameters in the chain. This would further simplify signaling, as one bit (i.e., shaping on/off) would be sufficient.

Note that while the probability p is not necessarily required at the transmitter side, it is needed at the receiver before decoding. The relation between S and p depends on B and M, but also on the precoder parameters such as list size. This relation may also be stored in a look-up table, or an analytical function (like the polynomial approximation depicted in Fig. 4) can be derived to avoid look-up tables.

We would like to point out that in the first release of 5G NR, polar coding is only specified for transmitting control information, where higher order modulation is used only in PUSCH. However, the presented scheme demonstrates significant improvements for higher order modulation, which makes it promising for the next releases of 5G NR, where ultra reliable low latency communications (URLLC) and machine type communications (MTC) will be supported.

REFERENCES

[1] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, Jul. 2009.
[2] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2213–2226, May 2015.
[3] M. Seidl, A. Schenk, C. Stierstorfer, and J. B. Huber, "Polar-coded modulation," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4108–4119, Oct. 2013.
[4] G. Böcherer, T. Prinz, P. Yuan, and F. Steiner, "Efficient polar code construction for higher-order modulation," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2017, pp. 1–6.
[5] T. Prinz *et al.*, "Polar coded probabilistic amplitude shaping for short packets," in *Proc. IEEE Int. Workshop Signal Proc. Adv. Wireless Commun.*, Jul. 2017, pp. 83–87.
[6] 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.
[7] O. İşcan, R. Böhnke, and W. Xu, "Shaped polar codes for higher order modulation," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 252–255, Feb. 2018.
[8] J. Honda and H. Yamamoto, "Polar coding without alphabet extension for asymmetric models," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7829–7838, Dec. 2013.
[9] M. Mondelli, S. Hassani, and R. Urbanke, "How to achieve the capacity of asymmetric channels," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, Sep. 2014, pp. 789–796.
[10] O. İşcan and W. Xu, "Polar codes with integrated probabilistic shaping for 5G new radio," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Aug. 2018, pp. 1–5.
[11] *Technical Specification Group Radio Access Network, NR, Multiplexing and Channel Coding*, document 3GPP TS 38.212, 2017.
[12] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 927–946, May 1998.
[13] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 913–929, May 1993.
[14] A. G. I. Fàbregas and A. Martinez, "Bit-interleaved coded modulation with shaping," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Aug./Sep. 2010, pp. 1–5.
[15] M. Pikus and W. Xu, "Bit-level probabilistically shaped coded modulation," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 1929–1932, Sep. 2017.
[16] *Technical Specification Group Radio Access Network, NR, Physical Channels Modulation*, document 3GPP TS 38.211 2017.
[17] A. Balatsoukas-Stimming, M. B. Parizi, and A. Burg, "LLR-based successive cancellation list decoding of polar codes," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5165–5179, Oct. 2015.
[18] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

ONURCAN İŞCAN received the B.Sc. degree (Hons.) from Istanbul Technical University, Turkey, in 2006, and the M.Sc. and Dr. Ing. degrees from the Technische Universität München, Germany, in 2008 and 2014, respectively, where he was a Research and Teaching Assistant with the Institute for Communications Engineering, from 2009 to 2015. In 2014, he was a Visiting Researcher with the University of South Australia. Since 2015, he has been with the Huawei German Research Center, Munich, Germany. He has been involved in several national and EU funded research projects. His main research interests include physical layer communication, including relaying, coding, modulation, and security.

RONALD BÖHNKE received the Dipl.Ing. and Dr. Ing. degrees in electrical engineering from the University of Bremen, Germany, in 2002 and 2014, respectively, where he was a Research and Teaching Assistant with the Department of Communications Engineering, from 2002 to 2010. In 2002, he was a Visiting Researcher with the Fraunhofer Heinrich Hertz Institute, Berlin. From 2010 to 2014, he was with the Institute for Communications and Navigation and the Institute for Communications Engineering, Technische Universität München, Germany. In 2014, he joined the Huawei German Research Center, Munich, Germany. He has been involved in several national and EU funded research projects. His main research interests include efficient detection algorithms and adaptive transmission for MIMO systems, non-orthogonal multiple access, channel coding, and modulation.

WEN XU (SM'03) received the B.Sc. and M.Sc. degrees from the Dalian University of Technology, Dalian, China, in 1982 and 1985, respectively, and the Dr. Ing. degree from the Technische Universität München, Munich, Germany, in 1996, all in electrical engineering. From 1995 to 2006, he was with Siemens Mobile, Munich, where he was the Head of the Algorithms and Standardization Laboratory. As a competence center, the laboratory was responsible for physical layer and multimedia signal processing, and partly protocol stack aspects of 2G, 3G, and 4G mobile terminals, and actively involved in the standardization activities of ETSI, 3GPP, DVB, and ITU. From 2007 to 2014, he was with Infineon Technologies AG, Neubiberg, Germany, focusing on wireline and wireless system concepts/architectures and software/hardware implementations. In 2014, he joined the European Research Center, Huawei Technologies Düsseldorf GmbH, Munich, where he is currently the Head of Radio Access Technologies Department. His research interests include signal processing, source/channel coding, and wireless communication systems. He is a member of the Verband der Elektrotechnik Elektronik Informationstechnik, Germany.