On Uniform Quantization for Successive Cancellation Decoder of Polar Codes

Zhengming Shi, Kai Niu
Key Laboratory of Universal Wireless Communication, Ministry of Education
Beijing University of Posts and Telecommunications, Beijing, China 100876
Email: {shzhengming, niukai}@bupt.edu.cn

Abstract—Polar codes have a regular-recursive structure, which can efficiently be mapped to hardware for practical applications. A good quantization scheme plays a great important role in hardware implementation. In this paper, three uniform quantizers are designed for successive cancellation (SC) decoding of polar codes based on either optimizing the equivalent channel capacity, cutoff rate or mean-squared error (MSE). Moreover, exploiting the cutoff rate maximizing criterion, a modified Gaussian approximation (GA) method is proposed to construct polar codes and estimate frame error rate (FER) performance under the quantized decoding algorithms. Simulation results have shown that a 6-bit uniform quantized SC decoder can achieve a near floating point performance and the upperbound of FER can be estimated precisely using the modified GA method under quantized decoding algorithms.

Index Terms—Polar Codes, SC Decoding, Quantization, Gaussian Approximation

I. INTRODUCTION

Polar codes, proposed by Arıkan [1], are the first proven capacity-achieving coding schemes to achieve the symmetric capacity of the binary-input discrete memoryless channels (B-DMCs) while having an explicit construction. It has a regular-recursive structure, which can be implemented in low-complexity encoding and decoding algorithms. However, although polar codes have astonishing asymptotic performance, the finite-length performance of polar codes under SC decoding is not satisfying. In addition, their low-complexity SC decoding algorithm which is serial in nature will lead to low decoding throughput [2].

Since the introduction of polar codes, a lot of work has been done to overcome the above drawbacks. The more sophisticated decoding algorithms, such as the list SC decoder [3],[4] and stack SC decoder [5], were introduced to improve its performance. For practical applications, hardware architectures for SC decoding of polar codes were also discussed in [2] and [6], and recent results even have started to address the issue of implementation of list SC decoding architectures [7]. However, there are few researches concentrating on the quantization of the SC decoding algorithm. As far as we know, the author in [8] only discusses the robustness of quantized decoding algorithm without any practical methods proposed for the quantization of polar codes. In this paper, we focus on the quantization of SC decoding algorithm. Due to the same recursive calculation manner, any improvement made to a quantized SC decoding can benefit its improved versions. In addition, a low-complexity quantized SC decoder can provide a good guidance for the hardware implementation of SC decoding and enable the practical applications of polar codes.

In this work, we focus on the quantization of polar codes and present three criteria for the design of quantized SC decoder by optimizing either the equivalent channel capacity, channel cutoff rate or mean-squared error (MSE). Based on the channel cutoff rate maximizing criterion, a modified GA method with respect to quantization is introduced to construct polar codes and estimate the upperbound of FER under the quantized SC decoding algorithms.

The remainder of this paper is organized as follows. Section II starts from the general framework proposed by Arıkan [1] and briefly reviews the encoding and decoding of polar codes. Sections III presents three quantization criteria for polar codes, and Section IV introduces the modified Gaussian approximation method for polar codes, then the simulation results are provided in Section V. Finally, Section VI concludes this paper.

II. PRELIMINARIES

A. Notation Conventions

Throughout this paper, we use \(a_{n}^{N-1}\) to denote a row vector \((a_0, \ldots, a_{N-1})\) and \(a_{j}^{i}\) denotes the subvector \((a_i, \ldots, a_j)\). Note here the index is assumed to start from 0 for reference. Moreover, \(a_{j, o}^{i}\) is used to denote a subvector with odd indices and \(a_{j, e}^{i}\) denotes a subvector with even indices. Given \(a_{N}^{N-1}\) and a set \(\mathcal{A} \subset \{0, \ldots, N-1\}\), \(a_{\mathcal{A}}\) refers to the subvector \((a_i : i \in \mathcal{A})\) and \(\mathcal{A}^c\) is the complement of \(\mathcal{A}\) in \(\{0, \ldots, N-1\}\).

B. Polar Coding

Polar codes are based on a universal phenomenon called channel polarization. Let \(W : \mathcal{X} \rightarrow \mathcal{Y}\) denote an arbitrary B-DMC with input alphabet \(\mathcal{X} = \{0, 1\}\) and output alphabet \(\mathcal{Y}\) which is characterized by the transition probabilities \(W(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\). Channel polarization refers to the fact that we can obtain a second set of \(N\) binary-input synthesized channels \(\{W_n^{(j)}\}, 0 \leq j < N, n = \log_2 N\), by recursively applying a polarizing transform to multiple independent copies of the given B-DMC \(W\) \(n\) times, where \(W_n^{(j)}\) is the \(j\)th subchannel on stage \(n\) [3],[8]. As \(N\) goes to infinity, these polarized channels tend to be either purely noisy or noise-free, where the fraction of these noise-free channels...
approaches the symmetric capacity of the original memoryless channel \( W \). Polar coding method takes advantage of the polarization effect and suggests to transmit information bits through these noiseless channels and send frozen bits through the remaining, which can achieve the symmetric capacity under an SC decoder.

The polar coding method can be described as follows: A binary source block \( u_{0}^{N-1} \in \{0,1\}^{N} \) which consists of \( K \) information bits and \( N - K \) frozen bits is mapped to a codeword \( x_{0}^{N-1} \) via \( x_{0}^{N-1} = u_{0}^{N-1}G_{N} \). The matrix \( G_{N} = B_{N}F_{2}^{\otimes n} \) is the generation matrix where \( B_{N} \) is the bit-reversal permutation matrix and \( F_{2}^{\otimes n} \) denotes the \( n \)-th Kronecker power of \( F_{2} \). Then transmitted over the underlying channel \( K \) with the stopping condition \( L = 2^{n} \).

A binary source block \( W \) approaches the symmetric capacity of the original memoryless \( W \). Polar coding method takes advantage of the polarization effect and suggests to transmit information bits through these noiseless channels and send frozen bits through the remaining, which can achieve the symmetric capacity under an SC decoder.

**C. SC Decoding**

As mentioned in [1], given the noisy codeword \( y_{0}^{N-1} \), polar codes can be decoded by an SC decoding algorithm [1]. Let \( \hat{u}_{0}^{N-1} \) denote the estimate of source block \( u_{0}^{N-1} \), the bits \( \hat{u}_{j} \) are determined successively with index \( j = 0, 1, \ldots, N - 1 \). If \( j \in A_{j} \), \( A_{j} \) is known to both encoder and decoder and \( \hat{u}_{j} \) is simply set to \( u_{j} \). Otherwise, the SC decoder estimates \( \hat{u}_{j} \) according to the sign of its corresponding computed log-likelihood-ratio (LLR) \( L_{n}^{(j)}(y_{0}^{N-1}, \hat{u}_{0}^{j-1}) = \log \frac{W_{y_{0}^{j}}(y_{0}^{N-1}, \hat{u}_{0}^{j-1})}{W_{y_{0}^{j}}(y_{0}^{N-1}, \hat{u}_{0}^{j-1})} \). If the LLR is nonnegative, \( \hat{u}_{i} = 0 \), otherwise, \( \hat{u}_{i} = 1 \). From [3], the LLR \( L_{n}^{(j)}(y_{0}^{N-1}, \hat{u}_{0}^{j-1}) \) can be calculated recursively by

\[
L_{n}^{(2j)}(y_{0}^{N-1}, \hat{u}_{0}^{2j-1}) = 2\tanh^{-1} \left( \tanh \left( \frac{L_{n-1}^{(j)}(y_{0}^{N-1}, \hat{u}_{0}^{2j-1})}{2} \right) \right)
\]

and

\[
L_{n}^{(2j+1)}(y_{0}^{N-1}, \hat{u}_{0}^{2j}) = L_{n}^{(0)}(y_{0}^{N-1}, \hat{u}_{0}^{2j}) + (-1)^{u_{j}}L_{n-1}^{(j)}(y_{0}^{N-1}, \hat{u}_{0}^{2j})
\]

with the stopping condition \( L_{n}^{(0)}(y_{0}) = \log \frac{W_{y}(y_{0})}{W_{y}(y_{0})} \). Note here \( W_{y}(y_{0}) = \log \frac{W_{y}(y_{0})}{W_{y}(y_{0})} \).

**III. QUANTIZED SC DECODER**

Consider a quantization function \( Q(x) \) with quantization step \( \Delta \) and \( q \)-bit representing the quantization levels, with one of the \( q \) bits denoting the sign. \( Q(x) \) is defined as follows

\[
Q(x) = \begin{cases} 
\left( \frac{\lfloor x \rfloor}{2} + \frac{1}{2} \right) \Delta, & x \in [-M, M], \\
\text{sign}(x) \left( M - \frac{x}{\Delta} \right), & \text{otherwise.}
\end{cases}
\]

where the quantization levels \( L = 2^{q} \) and the truncation threshold \( M = \frac{\Delta}{2} \). Note here the quantization function is symmetric, i.e., \( Q(x) = -Q(-x) \). The input to \( Q(x) \) is limited to the range \([M, -M]\) and the values out of this range are all clipped.

The quantized SC decoder, denoted by \( QSC \), is defined as a version of the SC decoder in which the quantization function \( Q(x) \) is applied to all the output LLRs that the SC decoder computes. However, during SC decoding process, the LLRs in each stage will undergo remarkable changes due to channel polarization, so a series of quantization functions should be designed for the \( QSC \) decoder, and these functions are denoted as \( Q_{i}^{(j)}(x) \) which represents the quantization function for the \( j \)-th subchannel in the \( i \)-th stage in the decoding process.\( (0 \leq i < n, 0 \leq j < 2^{i}) \). Precisely speaking, the decoder computes LLRs of the received symbols from channel \( W \) and applies function \( Q_{0}(x) \) to them, these quantized LLRs are then fed into the SC decoder to estimate messages further. Namely, the function \( Q_{i}^{(j)}(x) \) is applied to the message update equations (1) and (2) in each step, then these new quantized message is employed for further computations.

Intuitively, two factors may impact the error-correction performance of \( QSC \): the number of quantization bits used to represent LLRs and the maximum message magnitude allowed before saturation occurs. Obviously, the more quantization bits, the better performance will be got, and we aim to find good schemes and use as less quantization bits as possible to reach near-floating point performance.

As for the standard SC decoder, the soft messages in each stage have different ranges due to channel polarization, the dynamic changes of LLRs in the decoding process should be taken into account. Namely, given the quantization bit \( q \), the quantization space for each subchannel may be different in the \( QSC \). Fortunately, GA can help to obtain the approximate probability density function (pdf) of LLRs in each subchannel, now we just need to design a good quantizer which can find the optimal \( \Delta \) for a given message distribution in each subchannel. The following three criteria are presented to solve this problem.
A. MMSE Criterion

As established in [9], q-bit quantization of the output of the binary input additive white Gaussian noise (BIAWGN) channel can be regarded as an equivalent channel with binary inputs and L-ary outputs, where \( L = 2^q \). A demonstration of this channel quantizations is shown in Fig. 1. The output symbols are characterized by the following transition probabilities:

\[
P_{ij} \equiv \Pr [y_k \in V_j | x_k = i] = \int_{V_j} p(y_k | x_k = i) dy_k, \tag{4}
\]

where \( i \in \{0, 1\}, j \in \{0, 1, \ldots, L - 1\}, x_k \) and \( y_k \) are respectively the transmitted value and unquantized received value at the \( k \)th time instant, \( V_j = (a_j, a_{j+1}) \) is the \( j \)th quantization interval with \( a_0 = -\infty \) and \( a_L = \infty \) and \( p(y_k | x_k = i) \) is the conditional pdf for the received unquantized value \( y_k \), given that \( x_k = i \in \{0, 1\} \) is transmitted.

Assuming that the binary input to the BIAWGN channel is equiprobable with binary phase-shift keying (BPSK) modulation and the quantization is symmetric around zero, all the \( L = 2^q \) quantization boundaries which correspond to the probability distribution of the L quantized outputs should be decided. Considering the symmetry of the problem under these assumptions, i.e., \( a_j = -a_{L-j}, j \in \{0, 1, \ldots, L\} \), we just need to determine \( L/2 - 1 \) region boundaries \( a_1, a_2, \ldots, a_{L/2-1} \), instead of \( L - 1 \). However, we focus more on the uniform quantizer, in which all the boundaries \( a_j \) are evenly spaced by \( \Delta \) and all reconstruction values \( v_j \) are the midpoints of the intervals \( V_j = (a_j, a_{j+1}), j \in \{0, 1, \ldots, L - 1\} \).

Generally speaking, as for the minimum mean-squared error (MMSE) quantization, the quantizer is not only determined by its region boundaries but also by the quantized values. However, for a general uniform quantizer, the quantization distortion \( D \) is defined as

\[
D(\Delta) = 2 \sum_{i=1}^{M-1} \int_{-(i-1)\Delta}^{i\Delta} f(x) - \frac{(2i-1)\Delta p(x)}{2} \cdot dx \\
+ 2 \sum_{i=M-1}^{L} \int_{-(M-1)\Delta}^{(M-1)\Delta} f(x) - \frac{(2M-2i)\Delta p(x)}{2} \cdot dx, \tag{5}
\]

where \( M = L/2 = 2^{(q-1)} \), \( f(x) \) is the distortion function, \( p(x) \) is the input amplitude probability density, and for SC decoder \( p(x) \) represent the pdf of LLR in a certain subchannel. For a minimum of MSE, the objective function is

\[
\min_{\Delta} D(\Delta). \tag{6}
\]

we set \( f(x) = x^2 \) and require \( \frac{\partial D(\Delta)}{\partial \Delta} = 0 \), then the optimal quantization space can be determined by the MMSE criterion.

B. Capacity-maximizing Criterion

Following the notation of [9], the capacity \( C \) of the equivalent channel of Fig. 1 with equiprobable and independent binary inputs is

\[
C = I(X;Y) = H(Y) - H(Y|X) \\
= - \sum_{j=0}^{L-1} p_j \log_2 (p_j) + \sum_{j=0}^{L-1} P_{0j} \log_2 (P_{0j}), \tag{7}
\]

where \( p_j = \Pr [y_k \in T_j] \) and considering the fact \( P_{0j} = P_{1(L-1-j)} \) due to the quantizer symmetry. Since \( \Pr [y_k = i] = 0.5, (i = 0, 1) \) and \( p_j = 0.5 (P_{0j} + P_{1j}) = 0.5 (P_{0j} + P_{1(L-1-j)}) \), (7) can be written as

\[
C(\Delta) = 1 - \sum_{j=0}^{L-1} P_{0j} \log_2 \frac{P_{0j} + P_{1(L-1-j)}}{P_{0j}}. \tag{8}
\]

To maximize the equivalent channel capacity \( C(\Delta) \) with respect to the \( P_{0j} \), the problem can be written as

\[
\max_{\Delta} C(\Delta). \tag{9}
\]

We just set \( \frac{dC(\Delta)}{d\Delta} = 0 \), then the optimal quantization space \( \Delta \) can also be obtained by this method.

C. Cutoff rate maximizing Criterion

The cutoff rate for the equivalent equiprobable binary input discrete memoryless channel is

\[
R(\Delta) = \max_{p(x)} \left\{ -\log_2 \left( \sum_{y \in Y} \left( \sum_{x \in X} p(x) \sqrt{p(y|x)^2} \right) \right) \right\} \\
= -\log_2 \left[ \frac{1}{2^q} \sum_{y \in Y} \left( \sum_{x \in X} \sqrt{p(y|x)^2} \right) \right] \\
= 2 \log_2 2 - \log_2 \left[ \sum_{j=0}^{L-1} \left( \sqrt{P_{0j} + P_{1j}} \right)^2 \right] \\
= 1 - \log_2 \left[ 1 + \sum_{j=0}^{L-1} \sqrt{P_{0j} P_{1j}} \right], \tag{10}
\]

where \( p(x) = \frac{1}{2} \) and \( P_{0j} \) are defined above, \( i \in \{0, 1\}, j \in \{0, 1, \ldots, L - 1\} \). Similarly, taking into account the symmetry of the quantizer, i.e., \( P_{1j} = P_{0(L-1-j)} \), (10) can be simplified as follows:

\[
R(\Delta) = 1 - \log_2 \left( 1 + 2 \sum_{j=0}^{L/2-1} \sqrt{P_{0j} P_{0(L-1-j)}} \right). \tag{11}
\]

The problem of maximizing the cutoff rate \( R(\Delta) \) can be written as

\[
\max_{\Delta} R(\Delta). \tag{12}
\]

It is the same way as the capacity-maximizing criterion, the optimal quantization space can be determined by \( \frac{dC(\Delta)}{d\Delta} = 0 \) which maximizes the equivalent channel cutoff rate.

As described above, when the quantization bit \( q \) is fixed, the mean-squared error, channel capacity and cutoff rate can be all considered as a unary function of the quantization space \( \Delta \). In order to solve the three optimization problems above, bisection method is employed numerically to search for the optimal quantization space \( \Delta \) in the local region, where the calculation of transition probability can be simplified by the function \( Q(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \).
IV. MODIFIED GAUSSIAN APPROXIMATION

As we all known, GA is a useful approach for the construction of polar codes in AWGN channel [10]. However, when quantization is taken into account, the polarized channels are deteriorated and the traditional GA method will be inaccurate, which can result in bad performance under quantized SC decoder. In this section, a modified GA method is proposed for the construction of polar codes which is a combination of the traditional GA method and the cutoff rate maximizing criterion for quantization, this method can be used to construct polar codes and estimate the upperbound of FER approximately under the quantized SC decoder.

Assuming that the all-zero codeword is transmitted, the probability density function of LLRs of the received bits is $L_0(y_i) \sim \mathcal{N}\left(\frac{2}{\sigma^2}, \frac{1}{\sigma^2}\right)$, where $\sigma^2$ is the channel noise variance of $W_0$, i.e., $W$. In each step of SC decoding, the LLRs will be quantized using $q$ bits, which should be taken into account in the construction of polar codes. In each step of GA construction, given the pdf of LLRs in a certain channel, we can compute the optimal quantization space $\Delta_{ij}(q)$ using the cutoff rate maximizing criterion by (12), then the maximum cutoff rate, denoted as $R_{ij}(q)$, can be obtained simultaneously by (11). However, the quantized channel is still considered as a Gaussian channel in floating point and the equivalent channel noise variance $\sigma^2$ can be obtained from

$$\sigma^2 = -\frac{1}{2\ln(2^{1-R_f} - 1)}.$$  \hspace{1cm} (13)

where $R_f$ is the cutoff rate. Then the mean and variance of LLRs in this subchannel are respectively modified by $E_{ij}(q) = \frac{2}{\sigma^2}$ and $D_{ij}(q) = 2E_{ij}(q)$. Similar to the traditional GA [10], the means of LLRs of the polarized channels can be updated using equations

$$E_{i+1}(q) = \phi^{-1}\left(1 - \left(1 - \phi\left(E_{i}(q)\right)\right)^2\right)$$  \hspace{1cm} (14)

and

$$E_{i+1}^{(2j-1)} = 2E_{i}(q),$$  \hspace{1cm} (15)

where

$$\phi(x) = \begin{cases} 1 - \frac{1}{\sqrt{\pi}} \int_{-\infty}^{x} \tanh \frac{u}{2} e^{-\frac{(u-x)^2}{4}} \, du, & x > 1 \\ 1, & x = 0 \end{cases}$$

The process above is repeated until the means of LLRs of the subchannels in last stage are obtained. Finally, the error probability of each subchannel can be determined by

$$P_e\left(W^{(i)}_n\right) \approx Q\left(\sqrt{E_{n}(q)/2}\right)$$  \hspace{1cm} (16)

where $Q(x) = \frac{1}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2/2} \, dt$.

After the reliability vector $P_e(W^{(i)}_n)$ is obtained, the $K$ most reliable polarized channels are selected to carry the information bits and the FER of an $(N, K, A)$ polar code under quantized SC decoder is upper bounded by

$$P_B(N, K, A) \leq \sum_{i \in A} P_e(W^{(i)}_n).$$  \hspace{1cm} (17)

V. SIMULATION RESULTS

In this section, we simulate an $N = 1024, R = 0.5$ polar code under quantized SC decoder based on the three proposed criteria and evaluate their performances. In Fig. 2, Fig. 3 and Fig. 4, the traditional GA in floating point is used for the construction of polar codes. From the three figures, we can see how 4-bit 5-bit and 6-bit quantization with different criteria affect the decoding performance of the $N = 1024, R = 0.5$ polar code.

![Fig. 2. FER performance of an $N = 1024, R = 0.5$ polar code under QSC using MMSE criterion.](image)

First, we can see that with 6-bit quantization the quantized SC decoder using the three proposed criteria can all achieve a near floating point performance. However, when the number of quantization bit is less than 6, the performance of polar codes using the MMSE criterion deteriorates the most comparing with the other two criteria. Furthermore, the cutoff rate maximization criterion works slightly better than capacity maximization criterion in high SNR and both are significantly better than MSE minimization criterion. In addition, Fig. 4 shows the loss from the unquantized performance can be as low as 0.1 dB when using 5 quantization bits for quantized SC decoder with cutoff rate maximization criterion. Therefore, it is clear that MSE minimization is not such a good criterion for the quantizer design for polar codes and cutoff rate maximization may be the best choice in this setup.

From the simulation results, we can see that the performance of polar codes using capacity maximization criterion is not so good as the cutoff rate criterion in high SNR and this is because the numerical calculation is instable. This is one of the reason why cutoff rate maximization criterion is applied to the modified GA method. On the other hand, the calculation of capacity in floating point is also a bit difficult, especially in the case of modifying GA. It will lead to great error in the calculation of the equivalent channel noise variance due to its high complexity and instability in the process of numerical calculation. In contrast, the channel cutoff rate has an explicit analytic expression in floating point which can
MSE minimization is not a good criterion for quantizer design and cutoff rate maximization yields slightly better designs in terms of decoding performance than capacity maximization and may be the best choice for the quantization of polar codes. Furthermore, we presented a reformulation of Gaussian approximation which enables us to construct polar codes and estimate the upperbound of FER under the quantized SC decoding algorithms.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 61171099), the National Science and Technology Major Project of China (No. 2013ZX03003012) and Qualcomm Corporation.

REFERENCES