Robust 2-bit Quantization of Weights in Neural Network Modeled by Laplacian Distribution

Zoran PERIC, Bojan DENIC, Milan DINCIC, Jelena NIKOLIC
Faculty of Electronic Engineering, University of Niš, Serbia
bojan.denic@elfak.ni.ac.rs

***Abstract*—Significant efforts are constantly involved in finding manners to decrease the number of bits required for quantization of neural network parameters. Although in addition to compression, in neural networks, the application of quantizer models that are robust to changes in the variance of input data is of great importance, to the best of authors knowledge, this topic has not been sufficiently researched so far. For that reason, in this paper we give preference to logarithmic companding scalar quantizer, which has shown the best robustness in high quality quantization of speech signals, modelled similarly as weights in neural networks, by Laplacian distribution. We explore its performance by performing the exact and asymptotic analysis for low resolution scenario with 2-bit quantization, where we draw firm conclusions about the usability of the exact performance analysis and design of our quantizer. Moreover, we provide a manner to increase the robustness of the quantizer we propose by involving additional adaptation of the key parameter. Theoretical and experimental results obtained by applying our quantizer in processing of neural network weights are very good matched, and, for that reason, we can expect that our proposal will find a way to practical implementation.**

***Index Terms*—image classification, neural networks, quantization, signal to noise ratio, source coding.**

I. INTRODUCTION

Deep neural networks (DNNs) are a very powerful tool [1-2], widely used in recent years to solve many complex problems, such as image classification [3-4], object recognition [5], text classification [6], speech processing [7] and vehicle routing [8]. Besides, some other applications of NNs in solving real-world challenges can be found in [9-12]. However, standard DNN implementation is based on full precision (32-bit) representation of DNN weights, requiring powerful and expensive hardware (specialized GPUs (Graphical Processor Units) and large memory space) for implementation. This prevents the application of DNNs on devices with limited hardware resources, such as edge devices [13]. Therefore, great research effort has been made in recent years to reduce the complexity of DNNs. One of the most efficient ways to achieve this goal is to perform quantization of DNN weights [14-16], reducing the number of bits for their representation from 32 bits to 16-bits [17], 8-bits [18], 4-bits [19] or 2-bits [20]. Moreover, ternary [21], and binary (1-bit) quantization [22-23] have also been taken into account.

Quantizers are usually designed for some referent variance [24]. However, the variance of DNN weights is usually not constant, that is, it varies in some range around the referent variance, producing a "variance-mismatch" scenario [25] that degrades performance of the quantizer. To deal with this problem, high-resolution quantizers are usually used, increasing complexity. In this paper, we came to an idea to use low-bit μ-law logarithmic companding quantizer to solve this "variance-mismatch" problem in a simple way. Our goal is to determine how successfully we can deal with this problem if we utilize μ-law logarithmic companding quantizer designed optimally, as described in the paper. Being the most used type of robust quantizers, the μ-law logarithmic companding quantizer [24], [26] has a number of applications in high-resolution cases (e.g. in speech [24] and audio [27] coding and in analog-to-digital converters of wireless receivers [28]), but it has not been considered in low-resolution cases so far.

This paper presents the design and performance analysis of a 2-bit μ-law logarithmic companding scalar quantizer for data with the Laplacian distribution, since it models well weights of NNs (neural networks) [15], [23]. Let us mention that Laplacian distribution is also widely used for modeling of speech [23-24], [29-30] and images [23-24]. The goal of our analysis is to maximize SQNR (signal-to-quantization noise ratio) by optimizing the support region threshold (as a key parameter) of the quantizer in question. In particular, we analyze two cases: designing the μ-law logarithmic companding quantizer for a referent variance, which is usually a unit variance, and designing it for a wide range of variances, that is, for a "variance-mismatch" scenario. These two cases have different design goals. In designing quantizer for the unit variance, the goal is to maximize SQNR for that variance. In designing quantizer for a wide range of variances, the goal is to maximize the average SQNR in the given variance range. These two cases require different optimization procedures, as will be shown in the paper. Of particular note, as an important contribution of this paper, is a new approach to optimization of the quantizer in question in a wide range of variances so that the average SQNR is maximized.

Companding quantizers are usually designed using approximate analysis [24], especially in the case of high-resolution quantizers where it achieves high accuracy, being much simpler compared to the exact analysis. However, in the case of 2-bit quantization, the accuracy of the approximate analysis is questionable, while the complexity of the exact and approximate analysis is of the same order. For that reason, we opted to perform both analyzes (exact and approximate) and to show which analysis is better and to what extent. This detailed comparison between the results of exact and approximate analysis represents another contribution of this paper, as, to the best of the authors

This work has been supported by the Science Fund of the Republic of Serbia (Grant No. 6527104, AI-Com-in-AI).

Digital Object Identifier 10.4316/AECE.2021.03001

knowledge, it has not been performed in the literature so far. Moreover, we provide a manner to increase the robustness of the quantizer in question by involving additional adaptation of the key parameter.

To test our 2-bit quantizer (designed based on exact formulas) we perform an experiment, where the weights of real NN are used as test data. The adopted NN model is Multi-Layer Perceptron (MLP) [1-2], developed for image classification. The paper shows a very good matching between the normalized histogram of NN weights and Laplacian PDF, proving the validity of the statistical modeling of NN weights with the Laplacian distribution. Also, a very good matching between theoretically and experimentally obtained SQNR values are demonstrated, as well as the ability of the proposed logarithmic quantizer to achieve better average SQNR than uniform quantization model, which is commonly used in NN compression (see e.g. [18-20]). In this way, we have indicated the suitability of using the proposed quantizer in NN compression.

II. Description of the 2-Bit Logarithmic Companding Scalar Quantizer

Let us consider a 2-bit logarithmic companding scalar quantizer (LCSQ) designed for data modeled with the zero-mean Laplacian PDF defined as [24]:

$$p(x,\sigma)=\frac{1}{\sigma\sqrt{2}}\exp\left(-\frac{\sqrt{2}|x|}{\sigma}\right), \tag{1}$$

where σ^2 denotes the variance of the input data. The quantizer is symmetric around zero, due to symmetry of the Laplacian PDF. The quantizer, we analyze in this paper (see Fig.1) is fully defined with its thresholds $x_{-2}, x_{-1}, x_0=0, x_1, x_2$ and representation levels y_{-2}, y_{-1}, y_1, y_2, whereas negative thresholds and representation levels are symmetric to the positive ones:

$$x_{-i}=-x_i, \quad y_{-i}=-y_i, \quad i=1,2. \tag{2}$$

Let $x_{\max}=x_2$ denote the support region threshold (also known as the maximal amplitude) of the quantizer.

Our LCSQ is implemented using the companding technique that consists of the following three steps (as authors illustrated in Fig. 2) [24]:

Step 1. Compress the input signal x by applying the compression function $c(\cdot)$. In this paper we use μ-law logarithmic compression function $c(x):[-x_{\max},x_{\max}]\rightarrow[-x_{\max},x_{\max}]$ defined as [24]:

$$c(x)=\frac{x_{\max}}{\ln(1+\mu)}\ln\left(1+\frac{\mu|x|}{x_{\max}}\right)\operatorname{sgn}(x), \tag{3}$$

where parameter μ denotes the compression factor.

Step 2. Apply the uniform quantizer on the compressed signal $c(x)$. The uniform quantizer divides the range $[-x_{\max},x_{\max}]$ into N uniform intervals with the same width $\Delta=2x_{\max}/N$, having positive thresholds and representation levels defined as:

$$x_{u,i}=i\cdot\Delta, \ i=0,1,2, \tag{4}$$

$$y_{u,i}=(2i-1)\Delta/2, \ i=1,2. \tag{5}$$

For $N=4$ we have that $\Delta=x_{\max}/2$, hence it follows that:

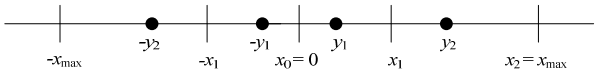

Figure 1. The appearance of a symmetric 2-bit non-uniform quantizer

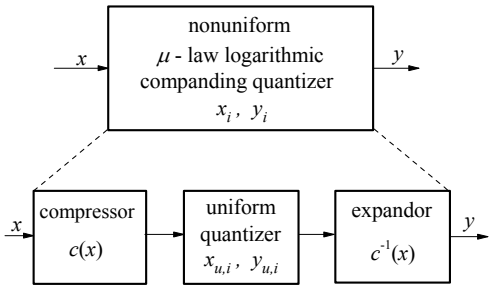

Figure 2. Structure of the companding quantizer [24]

$$x_{u,i}=i\cdot\frac{x_{\max}}{2}, \ i=0,1,2; \ y_{u,i}=\frac{(2i-1)\cdot x_{\max}}{4}, \ i=1,2. \tag{6}$$

Step 3. Expand the quantized version of the compressed signal using an inverse compressor function $c^{-1}(\cdot)$. For $c(x)$ defined with (3), the inverse function $c^{-1}(\cdot)$ that fulfills the condition $c^{-1}(c(x))=x$ is defined as:

$$c^{-1}(x)=\frac{x_{\max}}{\mu}\left((1+\mu)^{x/x_{\max}}-1\right)\operatorname{sgn}(x). \tag{7}$$

By applying the compression function $c(x)$, thresholds x_i and representation levels y_i of the μ-law logarithmic companding quantizer are mapped into thresholds $x_{u,i}$ and representation levels $y_{u,i}$ of the uniform quantizer:

$$c(x_i)=x_{u,i}, \ c(y_i)=y_{u,i}. \tag{8}$$

By applying the inverse compression function $c^{-1}(x)$ on equations defined with (8), we obtain the following expressions for x_i and y_i:

$$x_i=c^{-1}(x_{u,i})=\frac{x_{\max}}{\mu}\left((1+\mu)^{\frac{x_{u,i}}{x_{\max}}}-1\right), \tag{9}$$

$$y_i=c^{-1}(y_{u,i})=\frac{x_{\max}}{\mu}\left((1+\mu)^{\frac{y_{u,i}}{x_{\max}}}-1\right). \tag{10}$$

By substituting (6) into (9) and (10), we derive:

$$x_i=\frac{x_{\max}}{\mu}\left((1+\mu)^{\frac{i}{2}}-1\right), \ i=0,1,2, \tag{11}$$

$$y_i=\frac{x_{\max}}{\mu}\left((1+\mu)^{\frac{2i-1}{4}}-1\right), \ i=1,2. \tag{12}$$

Eventually, we come to the final expressions for positive thresholds and representation levels of 2-bit μ-law LCSQ:

$$x_0=0, \quad x_1=\frac{x_{\max}}{\mu}\left((1+\mu)^{\frac{1}{2}}-1\right), \quad x_2=x_{\max}, \tag{13}$$

$$y_1=\frac{x_{\max}}{\mu}\left((1+\mu)^{\frac{1}{4}}-1\right), \quad y_2=\frac{x_{\max}}{\mu}\left((1+\mu)^{\frac{3}{4}}-1\right). \tag{14}$$

The total distortion D is a sum of the distortion D_i in the inner region $[-x_{\max},x_{\max}]$ of the quantizer and the distortion D_o in the outer region $(-\infty,-x_{\max})\cup(x_{\max},+\infty)$ of the quantizer. We will firstly design 2-bit LCSQ for referent variance $\sigma_0^2=1$, that is a usual approach in the literature [24], so that the MSE (mean-square error) distortion is minimal. Then, the design of the quantizer in a wide range of variances will be performed.

III. Design of Optimal 2-bit LCSQ for the Unit Variance

For the unit variance $\sigma^2 = \sigma_0^2 = 1$, zero-mean Laplacian PDF becomes:

$$p(x) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|x|) . \quad (15)$$

To design 2-bit LCSQ, we will use two approaches: one based on exact formulas and the other based on approximate formulas.

A. Design of LCSQ by Using Exact Formulas

Distortions D_i and D_o are defined as [24]:

$$D_i = 2\int_{x_0=0}^{x_1} (x-y_1)^2 p(x)dx + 2\int_{x_1}^{x_{\max}} (x-y_2)^2 p(x)dx , \quad (16)$$

$$D_o = 2\int_{x_{\max}}^{+\infty} (x-y_2)^2 p(x)dx , \quad (17)$$

where x_i and y_i are specified by (13) and (14), respectively. The total distortion $D = D_i + D_o$ can be written as:

$$D = 2\int_{0}^{x_1} (x-y_1)^2 p(x)dx + 2\int_{x_1}^{+\infty} (x-y_2)^2 p(x)dx , \quad (18)$$

which can be expressed in closed-form:

$$D = 1 - \frac{\sqrt{2}x_{\max}}{\mu}\left((1+\mu)^{1/4} - 1\right) + \frac{x_{\max}^2}{\mu^2}\left((1+\mu)^{1/4} - 1\right)^2$$
$$+ \exp\left(-\frac{\sqrt{2}x_{\max}\left(-1+\sqrt{1+\mu}\right)}{\mu}\right)\frac{x_{\max}}{\mu}(1+\mu)^{1/4} \cdot \quad (19)$$
$$\cdot\left(\frac{x_{\max}}{\mu}\left(2\sqrt{1+\mu} - 2 + \mu\left((1+\mu)^{1/4} - 2\right)\right) - \sqrt{2}\left(\sqrt{1+\mu} - 1\right)\right)$$

Let us highlight here that eq. (19) has not been derived in the literature so far, since exact analysis of low-bit LCSQ has not been the topic of any previous research. From this worthy formula, we can notice that D is a function of two parameters: $x_{\max}$ and μ. For a given μ value, D depends only on $x_{\max}$. Accordingly, we determine $x_{\max}$ such that D is minimal, or equivalently, such that SQNR is maximal, whereas SQNR is defined as:

$$\text{SQNR}\,[\text{dB}] = 10 \cdot \log_{10}\left(\frac{\sigma^2}{D}_{|\sigma^2=\sigma_0^2=1}\right) = 10 \cdot \log_{10}\left(\frac{1}{D}\right) \quad (20)$$

In Fig. 3, we provide SQNR dependence on the parameter $x_{\max}$ for 2-bit LCSQ for several values of μ, that is for $\mu = 63$, $\mu = 127$ and $\mu = 255$ (note that μ is typically selected as 2^M-1, where M is an integer [24]). It can be observed that each SQNR curve (obtained for different values of μ) attains different maximal values at different values of $x_{\max}$. The maximal values of SQNR ($\text{SQNR}_{\max}$) and corresponding $x_{\max}^{opt}$ values are summarized in Table I.

TABLE I. $x_{\max}^{opt}$ AND CORRESPONDING $\text{SQNR}_{\max}$ VALUES FOR 2-BIT LCSQ DESIGNED USING EXACT FORMULAS

	$x_{\max}^{opt}$	$\text{SQNR}_{\max}$ [dB]
$\mu = 63$	3.707	5.21
$\mu = 127$	3.965	4.78
$\mu = 255$	4.318	4.44

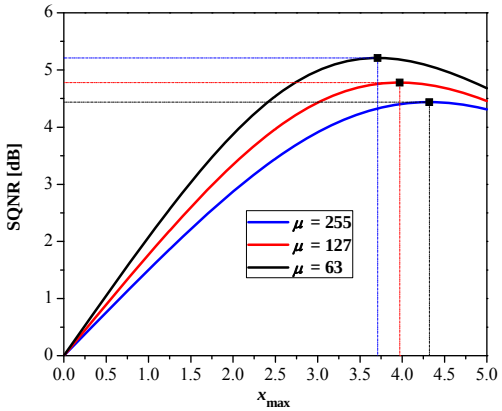

Figure 3. SQNR as a function of $x_{\max}$ for 2-bit LCSQ for different μ values

TABLE II. POSITIVE THRESHOLDS AND REPRESENTATION LEVELS OF 2-BIT LCSQ

	x_1	y_1	y_2
$\mu = 63$	0.412	0.108	1.274
$\mu = 127$	0.322	0.074	1.158
$\mu = 255$	0.254	0.051	1.067

We can notice that 2-bit LCSQ with a lower μ value (e.g. 63) provides a higher SQNR value due to a narrow support region. Eventually, we can conclude that corresponding pair of values (μ, $x_{\max}^{opt}$) given in Table I completely defines LCSQ, as values of thresholds and representation levels (given in Table II) can be easily calculated from them using (13) and (14).

B. Design of LCSQ by Using Approximated Formulas

This subsection also considers the design of 2-bit LCSQ optimized so that to minimize the MSE distortion for the case of using the approximate formulas. Specifically, for the inner distortion component of LCSQ, similarly as in [24], [27], we use the following expression, known as the Bennet's integral:

$$D_i^a = \frac{x_{\max}^2}{3N^2}\int_{-x_{\max}}^{x_{\max}} \left(\frac{dc(x)}{dx}\right)^{-2} p(x)dx , \quad (21)$$

while for the overload distortion component we use [24]:

$$D_o^a = 2\int_{x_{\max}}^{+\infty} (x - x_{\max})^2 p(x)dx . \quad (22)$$

Index a in the superscript refers to the approximate formulas. In case of the Laplacian PDF specified in (15) and compression function given in (3), the following expressions can be derived for components of the MSE distortion:

$$D_i^a = \frac{\ln^2(1+\mu)}{3N^2}\left(\frac{x_{\max}^2}{\mu^2} + \frac{\sqrt{2}x_{\max}}{\mu} + 1\right), \quad (23)$$

$$D_o^a = \exp(-\sqrt{2}x_{\max}) . \quad (24)$$

Finally, for the total distortion of LCSQ we have:

$$D^a = \frac{\ln^2(1+\mu)}{3N^2}\left(\frac{x_{\max}^2}{\mu^2} + \frac{\sqrt{2}x_{\max}}{\mu} + 1\right) + \exp(-\sqrt{2}x_{\max}) \quad (25)$$

TABLE III. $x_{\max}^{a,\,opt}$ AND CORRESPONDING $\text{SQNR}_{\max}^{a}$ VALUES FOR 2-BIT LCSQ DESIGNED USING APPROXIMATE FORMULAS

$N = 4$	**$\mu = 63$**	**$\mu = 127$**	**$\mu = 255$**
$x_{\max}^{a,\,opt}$	3.6	3.90	4.22
$\text{SQNR}_{\max}^{a}$ **[dB]**	4.01	2.87	1.82

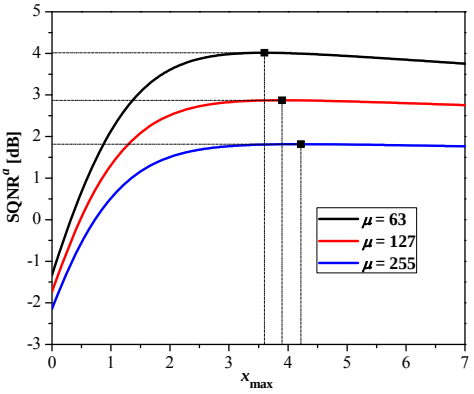

Figure 4. SQNR^a as a function of $x_{\max}$ for different μ values

From the last equation, we can conclude that D^a, for given N and μ values, is strongly dependent on $x_{\max}$. Accordingly, we optimize D with respect to $x_{\max}$. SQNR in this case is:

$$\mathrm{SQNR}^{a} = 10\log_{10}\left(\frac{1}{D^{a}}\right). \tag{26}$$

Fig. 4 plots SQNR^a versus $x_{\max}$ for 2-bit LCSQ using the same values of μ as in Fig. 3, while Table III summarizes the optimal values of $x_{\max}^{a,\,opt}$ and of the corresponding $\mathrm{SQNR}_{\max}^{a}$. From the results given in Table III, we can derive similar conclusions as from Table I. Additionally, for the same μ value, we can observe that the value of $x_{\max}^{a,\,opt}$ is slightly lower than the one from Table I, while the difference between the achieved SQNRs is notable, as will be discussed in the next subsection.

C. Analysis of the Theoretical Results

In this subsection, we analyze the accuracy of the approximate formula for SQNR (see (25) and (26)) with respect to the exact one (see (19) and (20)). Specifically, we use the relative error as the comparison measure:

$$\delta_{\mathrm{SQNR}}[\%] = \frac{\left|\mathrm{SQNR} - \mathrm{SQNR}^{a}\right|}{\mathrm{SQNR}} \cdot 100 \tag{27}$$

Table IV lists the values of the relative error we have calculated. It is obvious that the employed approximated formula is very inaccurate, as the relative error ranges up to 60 % for some values of μ. Hence, considering the design of our LCSQ for the particular variance, we can conclude that one should better use the exact formulas, defined with (19) and (20). Let us highlight that this is intuitively expected result, as the accuracy of approximate formulas decreases as N decreases [24]. However, to the best of the authors knowledge, a detailed comparative analysis of the results presented in this paper including two design approaches for LCSQ, approximate and exact one, has not been reported in the literature so far.

TABLE IV. RELATIVE ERRORS IN PERCENTS FOR SQNR

$N = 4$	$\mu = 63$	$\mu = 127$	$\mu = 255$
$\delta_{\mathrm{SQNR}}[\%]$	23.03	39.96	59.01

IV. A 2-BIT LCSQ IN A WIDE RANGE OF VARIANCES

This section is devoted to the performance analysis of 2-bit LCSQ optimally designed for variance σ_0^2, that is, applied for quantization of data from Laplacian source with a variance σ^2 different from the variance used for design of the quantizer ($\sigma^2 \neq \sigma_0^2$). In the literature, this case is known as mismatched quantization case [24-25]. In particular, it is of importance to analyze mismatched quantization from practical reasons, as it provides information about the robustness of non-adaptive quantizer models in non-stationary data processing, where robust quantizers are greatly required. The research conducted so far, for instance in [25], has shown that variance-mismatch effect has negative impact on the quantizer performance (degrades SQNR). In what follows, beside a detailed theoretical analysis, we also provide a manner for performance improvement of LCSQ in a wide dynamic range of input data variances.

A. Robustness Analysis of LCSQ by Using Exact Formulas

To evaluate performance of 2-bit LCSQ in the variance-mismatched case, we use the PDF defined with (1). Thus, for the distortion components we have:

$$D_i(\sigma) = 2\int_0^{x_1(\sigma_0)} (x - y_1(\sigma_0))^2\, p(x,\sigma)dx + 2\int_{x_1(\sigma_0)}^{x_{\max}(\sigma_0)} (x - y_2(\sigma_0))^2\, p(x,\sigma)dx, \tag{28}$$

$$D_o(\sigma) = 2\int_{x_{\max}(\sigma_0)}^{+\infty} (x - y_2(\sigma_0))^2\, p(x,\sigma)dx, \tag{29}$$

where $x_{\max}(\sigma_0) = x_{\max}$, $x_i(\sigma_0) = x_i$ and $y_i(\sigma_0) = y_i$ denote thresholds and representation levels of our 2-bit LCSQ calculated for the unit variance, as in Tables I and II. The total distortion $D(\sigma) = D_i(\sigma) + D_o(\sigma)$ can be written as:

$$D(\sigma) = 2\int_0^{x_1} (x - y_1)^2\, p(x,\sigma)dx + 2\int_{x_1}^{+\infty} (x - y_2)^2\, p(x,\sigma)dx \tag{30}$$

and it can be expressed in closed-form:

$$D(\sigma) = \sigma^2\left[1 - \frac{\sqrt{2}x_{\max}\left((1+\mu)^{1/4} - 1\right)}{\mu\cdot\sigma} + \frac{x_{\max}^2\left((1+\mu)^{1/4} - 1\right)^2}{\mu^2\sigma^2} + \exp\left(-\frac{\sqrt{2}x_{\max}\left(-1+\sqrt{1+\mu}\right)}{\mu\sigma}\right)\frac{x_{\max}}{\mu\cdot\sigma}(1+\mu)^{1/4}\cdot \left(\frac{x_{\max}}{\mu\cdot\sigma}\left(2\sqrt{1+\mu} - 2 + \mu\left((1+\mu)^{1/4} - 2\right)\right) - \sqrt{2}\left(\sqrt{1+\mu} - 1\right)\right)\right] \tag{31}$$

In this case, we estimate SQNR as:

$$\mathrm{SQNR}(\sigma) = 10\log_{10}\left(\frac{\sigma^2}{D(\sigma)}\right). \tag{32}$$

Since the variance σ^2 can vary in a wide range, it is usual to express σ in logarithmic domain as $\sigma_{dB} = 20\log_{10}(\sigma/\sigma_0)$. For $\sigma_0 = 1$, we derive:

$$\sigma_{dB} = 20\log_{10}(\sigma);\ \sigma = 10^{(\sigma_{dB}/20)} \tag{33}$$

SQNR can be expressed as:

$$\mathrm{SQNR}(\sigma_{dB}) \equiv \mathrm{SQNR} = -10\cdot\log_{10}(D(\sigma)/\sigma^2) = -10\cdot\log_{10}\left[1 - \frac{\sqrt{2}x_{\max}\left((1+\mu)^{1/4} - 1\right)}{\mu\cdot 10^{\sigma_{dB}/20}} + \frac{x_{\max}^2\left((1+\mu)^{1/4} - 1\right)^2}{\mu^2\cdot 10^{\sigma_{dB}/10}} + \exp\left(-\frac{\sqrt{2}x_{\max}\left(-1+\sqrt{1+\mu}\right)}{\mu\cdot 10^{\sigma_{dB}/20}}\right)\frac{x_{\max}}{\mu\cdot 10^{\sigma_{dB}/20}}(1+\mu)^{1/4}\cdot \left(\frac{x_{\max}\left(2\sqrt{1+\mu} - 2 + \mu\left((1+\mu)^{1/4} - 2\right)\right)}{\mu\cdot 10^{\sigma_{dB}/20}} - \sqrt{2}\left(\sqrt{1+\mu} - 1\right)\right)\right] \tag{34}$$

Fig. 5 illustrates SQNR of 2-bit LCSQ designed optimally for variance σ_0^2 and various values of μ in accordance with the exact analysis and applied to quantize data from Laplacian source with variance σ^2, where wide range of σ^2 values is assumed. In particular, eq. (34) is utilized for obtaining curves shown in Fig. 5. One can notice from Fig. 5 that each curve, beside the global maximum obtained at the point 0 dB ($\sigma^2 = \sigma_0^2$), also provides the local maximum. The occurrence of this local maximum can be explained as follows. For some specific variance value, the levels $-y_1$ and y_1 more dominantly contribute to D compared to $-y_2$ and y_2, and, for that reason, at this particular variance, our 2-bit LCSQ actually behaves as 1-bit LCSQ. We can notice that maximum SQNR values are equal to ones from Table I. Fig. 5 also reveals that the robustness of this quantizer is low, since even negative SQNR values are calculated and presented.

In order to improve the performance of our initial quantizer model in a wide dynamic range of variances, we propose the following manner for determining the crucial design parameter:

$$x_{\max}^{new} = k \cdot x_{\max}, k \in R, \tag{35}$$

where k is a real constant chosen to maximize the average SQNR (SQNR$_{av}$) in the variance range of particular importance for our analysis, where:

$$\text{SQNR}_{av} = \frac{1}{v}\sum_{i=1}^{v} \text{SQNR}(\sigma_i), \tag{36}$$

and where v represents the number of the particular variances σ_i^2 in the observed range. We consider the variance range [-30 dB, 30 dB] around σ_0^2, averaging in v = 1200 points.

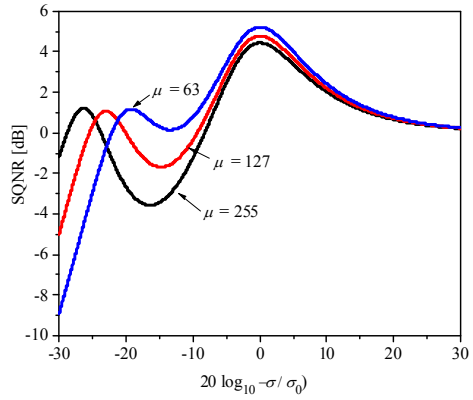

Figure 5. SQNR of 2-bit LCSQ in a wide dynamic range of input data variances, for various μ values

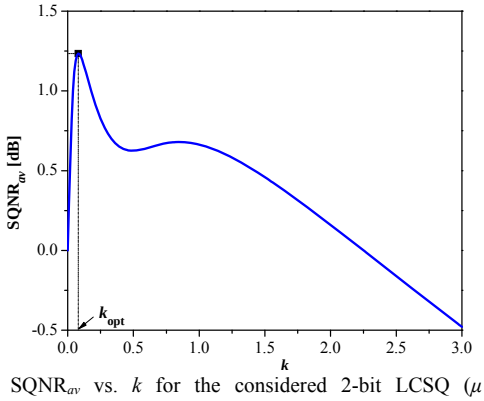

Figure 6. SQNR$_{av}$ vs. k for the considered 2-bit LCSQ (μ = 255)

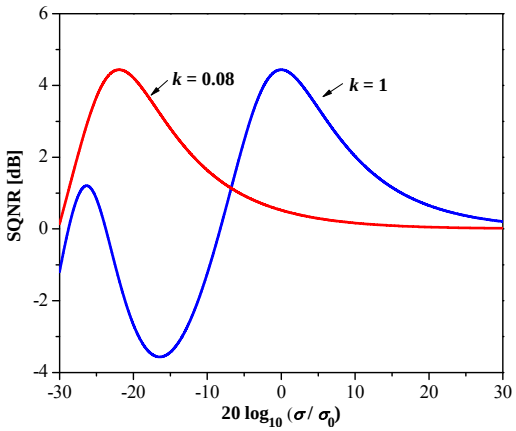

Figure 7. SQNR of 2-bit LCSQ (μ = 255) for k = 1 and k = 0.08

TABLE V. THE OPTIMAL VALUES OF K AND CORRESPONDING AVERAGE SQNR FOR 2-BIT LCSQ DESIGNED USING EXACT FORMULAS

N = 4	μ = 63	μ = 127	μ = 255
k_{opt}	0.4	0.09	0.08
SQNR$_{av}^{k=k_{opt}}$ [dB]	1.67	1.37	1.23
SQNR$_{av}^{k=1}$ [dB]	1.09	1.03	0.66

Fig. 6 shows SQNR$_{av}$ as a function of k, for 2-bit LCSQ and μ = 255. From Fig. 6 we can conclude that the most appropriate value for the parameter k is $k = k_{opt} = 0.08$, as in that case SQNR$_{av}$ is maximized. Fig. 7 provides comparison of the corresponding SQNR curves for 2-bit LCSQ (μ = 255) when k is set to 1 and also to 0.08, showing the evident benefit in terms of the robustness of the introduced design approach. In addition, the values of k_{opt} in case of some other values of μ are presented in Table V, along with the corresponding SQNR$_{av}$ values. For comparison purposes, in Table V we also provide SQNR$_{av}$ for k = 1 ($x_{\max}^{new} = x_{\max}$). From Table V we can notice that improvements in average SQNR amount up to 0.6 dB in the predefined variance range having width of 60 dB.

B. Robustness Analysis of LCSQ Using Approximated Formulas

For this specific case, the corresponding expressions for distortion components can be obtained by substituting $p(x)$ (eq. (15)) with $p(x,\sigma)$ (eq. (1)) into (21) and (22), which gives:

$$D_i^a(\sigma) = \sigma^2 \frac{\ln^2(1+\mu)}{3N^2}\left(\frac{x_{\max}^2}{\mu^2\sigma^2} + \frac{\sqrt{2}}{\mu}\frac{x_{\max}}{\sigma} + 1\right), \tag{37}$$

$$D_o^a(\sigma) = \sigma^2 \exp\left(-\frac{\sqrt{2}x_{\max}(\sigma_0)}{\sigma}\right), \tag{38}$$

where $x_{\max}(\sigma_0) \equiv x_{\max}$ refers to the $x_{\max}^{a,\,opt}$ value optimized for variance σ_0^2, as given in Table II. In brief, it holds:

$$D^a(\sigma) = \sigma^2\left[\frac{\ln^2(1+\mu)}{3N^2}\left(\frac{x_{\max}^2}{\mu^2\sigma^2} + \frac{\sqrt{2}}{\mu}\frac{x_{\max}}{\sigma} + 1\right) + \exp\left(-\frac{\sqrt{2}x_{\max}}{\sigma}\right)\right] \tag{39}$$

and

$$\text{SQNR}^a = 10\log_{10}\left(\frac{\sigma^2}{D^a(\sigma)}\right). \tag{40}$$

Fig. 8 demonstrates SQNR^a (eq. (40)) of 2-bit LCSQ in the predefined wide dynamic range of variances, for different μ values. We can observe quite different shapes of SQNR curves with respect to former case (see Fig. 5) and, as expected, we can notice lower maximal SQNR values.

By using the same design approach as in Section IV-A defined with (35), we can improve the performance of LCSQ in a wide dynamic range of variances. The optimal value of the parameter k can be found in the same way as in Fig. 6. Table VI summarizes the optimal values of k and corresponding SQNR$_{av}$ values for different values of μ. However, quite small improvements are achieved in this case (up to 0.1 dB), much smaller than in the case with exact formulas. Fig. 9 shows SQNR curves in a wide range of variances for μ = 255, for two values of k (k = 1 and k = 1.18).

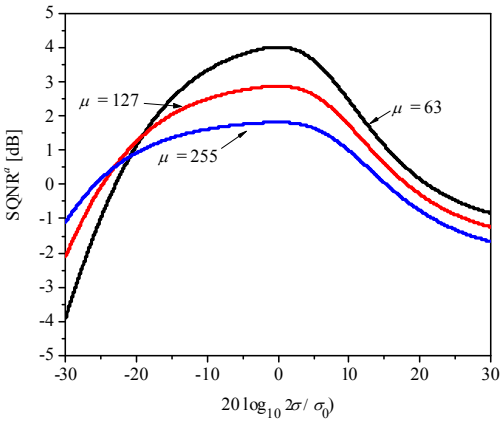

Figure 8. SQNR of 2-bit LCSQ designed by using the approximated formula in a wide dynamic range of variances, for various μ values

TABLE VI. THE OPTIMAL VALUES OF K AND CORRESPONDING AVERAGE SQNR FOR 2-BIT LCSQ DESIGNED USING APPROXIMATED FORMULAS

N = 4	μ = 63	μ = 127	μ = 255
k_{opt}^{a}	0.57	0.82	1.18
SQNR$_{av}^{k=k_{opt}}$ [dB]	1.67	1.15	0.59
SQNR$_{av}^{k=1}$ [dB]	1.55	1.14	0.58

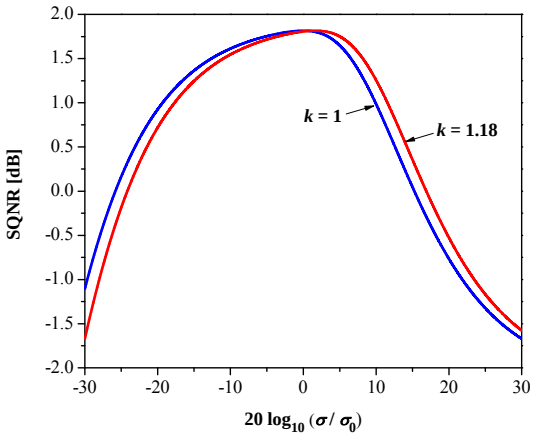

Figure 9. SQNR of 2-bit LCSQ (μ = 255) designed using approximated formula for k = 1 and k = 1.18

C. Performance Comparison in a Wide Dynamic Range

In Fig. 10, we give SQNR curves for 2-bit LCSQ in case of using exact and approximated formulas, for μ = 255. A large deviation between the exact and approximated SQNR values is apparent in a wide dynamic range of variances, which again indicates that proper analysis and design of the considered quantizer is enabled only by the exact formulas.

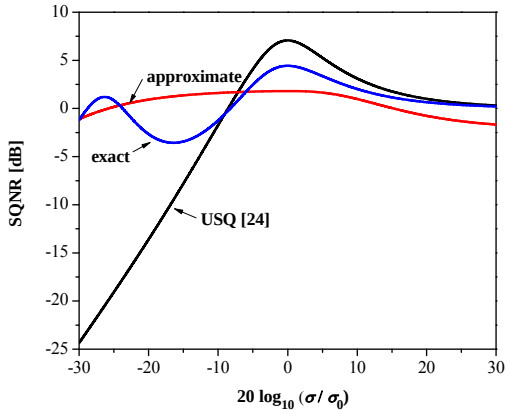

Figure 10. Performance comparison of different 2-bit scalar quantizers in a wide dynamic range of variances (LCSQ and USQ [24])

In addition, for comparison purposes, we include in Fig. 10 the 2-bit uniform scalar quantizer (USQ), due to the fact that uniform scalar quantization is dominantly used compression method in NN [13-16], [18-20]. Thus, for design of 2-bit USQ we use the optimal step size Δ = 1.0874 [24]. The provided results demonstrate that the proposed 2-bit LCSQ is a better candidate for a wide range of variances than the employed baseline USQ (although it can achieve higher maximal SQNR). Moreover, this can be proved by observing the average SQNR values of the proposed LCSQ (see Table V for k = 1 and μ = 255) and baseline quantizer (SQNR$_{av}^{USQ}$ = -2.57 dB), where the gain of more than 3 dB is provided.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section we will apply the proposed 2-bit LCSQ (designed using exact formulas) for quantization of weights of a trained neural network, with the aim to investigate matching with the theoretical results in terms of SQNR and the possibility of applying LCSQ for NN compression. As the neural network model we will use MLP (Multilayer Perceptron) [1-2] composed of an input and an output layer, that is developed for image classification task.

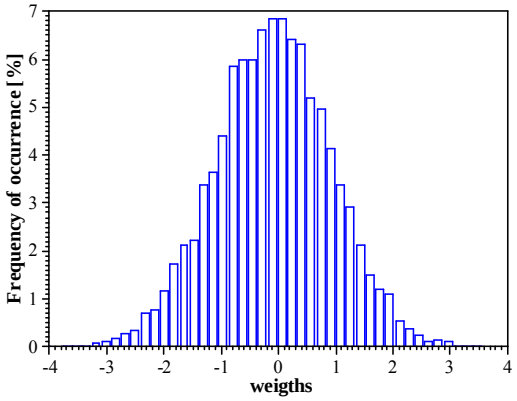

Figure 11. The histogram of weights of trained MLP network.

Training, validation and test data are used from the MNIST database [30], which consists of 60000 monochrome images of hand-written single digits with dimension 28 × 28 = 784 pixels, among which 50000 images are used for training and 10000 images are used for testing. Therefore, MLP uses input layer with 784 nodes, while the output layer has 10 nodes (due to 10 different digits). In addition, *softmax* function is used as activation

function at the output layer. For other relevant settings of MLP, such as the learning rate and batch size we adopt the values 0.5 and 250, respectively.

The weights to be used for our analysis are obtained as the result of MLP training, performed in 20 epochs. Fig. 11 shows the histogram of weights of trained MLP network. We can see that the histogram is very close to the zero-mean Laplacian PDF, proving its appropriateness for statistical modeling of NN weights and also forming the base for implementation of the 2-bit LCSQ.

Let us perform quantization of NN weights w_i with the proposed 2-bit LCSQ with μ = 255, obtaining quantized weights w_i^q, ($i = 1,...,W$), where W denotes the total number of weights. Let $\sigma_w^2 = (1/W)\cdot\sum_{i=1}^{W} w_i^2$ denote the variance of MLP weights and $D_w = (1/W)\cdot\sum_{i=1}^{W}(w_i - w_i^q)^2$ denote the distortion of quantization of weights. Then, we can calculate experimental SQNR (SQNR^{exp}) for weights quantization as:

$$\mathrm{SQNR}^{\exp} = 10\log_{10}\left(\frac{\sigma_w^2}{D_w}\right) = 10\log_{10}\left(\frac{\sum_{i=1}^{W} w_i^2}{\sum_{i=1}^{W}(w_i - w_i^q)^2}\right) \quad (41)$$

The 2-bit LCSQ with μ = 255 is designed for the variance σ_w^2 (called referent variance), whereas we consider two versions of 2-bit LCSQ, for k = 1 and k = 0.08. Thresholds and representation levels for the 2-bit LCSQ designed for the referent variance σ_w^2 are obtained by multiplying corresponding thresholds and representation levels from Table II by $k\cdot\sigma_w$. Experimentally obtained SQNR^{exp}, defined with (41), is shown in Fig.12 in the wide range [-30 dB, 30 dB] of variances of the MLP weights relative to σ_w^2. SQNR^{exp} for some weight variance σ^2 (expressed in the logarithmic domain as σ^2[dB] = $10\cdot\log_{10}(\sigma^2/\sigma_w^2)$ belonging to the range [-30 dB, 30 dB] relative to σ_w^2) is calculated by feeding the quantizer input with weights obtained by multiplying all original MLP weights w_i by σ. We can see that experimentally obtained SQNR values are consistent with the theoretical ones shown in Fig. 7.

Eventually, the performance of the proposed 2-bit LCSQ (k = 1, μ = 255) and the baseline 2-bit USQ [24] are compared in real data processing scenario using the same input data (MLP weights), and the results obtained using (41) are shown in Fig. 13. Note that our quantizer provides better robustness in a wide range of variance of weights than

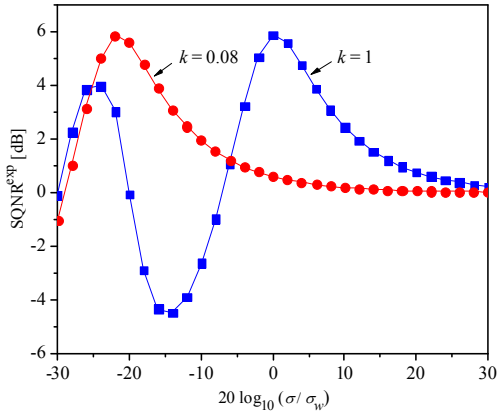

Figure 12. SQNR^{exp} of 2-bit LCSQ in a wide range of weights variance

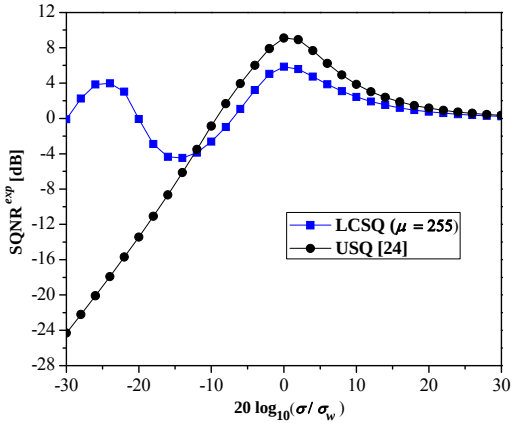

Figure 13. Performance of the proposed 2-bit LCSQ (μ = 255) and baseline USQ [24] in a wide dynamic range of weights variance

the baseline USQ [24], as it is a case with the theoretical results given in Fig. 10. Based on the conducted theoretical and experimental analysis, we can anticipate that the proposed 2-bit LCSQ is a very good solution for NN quantization and compression.

VI. Conclusion

In this paper, we have described in detail two approaches to design of optimized 2-bit LCSQ, where optimization has been performed in terms of MSE distortion, derived as a result of the exact and approximate analysis of LCSQ. In particular, for both approaches, we have derived novel formulas for the distortion of LCSQ that show which parameters affect the distortion of the considered 2-bit LCSQ. Specifically, we have shown that corresponding pair of values (μ, x_{max}^{opt}) completely defines LCSQ performance as well as that other quantizer design parameters are straightforwardly determined from it. In other words, by optimizing the distortion of the proposed quantizer with respect to x_{max}, for the given bit rate and parameter μ, we have ended up with the optimal support region threshold value x_{max}^{opt} for both design approaches. We have uncovered that employed approximated formulas are very inaccurate, and, for that reason, we have highlighted that considering designing LCSQ for the particular variance, one should better use the exact formulas for the performance studying of our LCSQ. Specifically, a large deviation between the exact and approximate SQNR values has been determined in a wide dynamic range of input data variances indicating that accurate analysis and design of the considered 2-bit LCSQ is enabled only by using the exact formulas. As an additional result of this paper, we have provided a manner for performance improvement of LCSQ in a wide dynamic range of input data variances. Comparison of theoretical results has shown that, for processing data from Laplacian source in a wide range of input data variances, the proposed 2-bit LCSQ is a better candidate than the employed baseline USQ, especially if the robustness of the quantizer performance is the most important feature required. Moreover, the performance of the proposed 2-bit LCSQ and the baseline 2-bit USQ has been compared in real data processing scenario, and the results have shown that our quantizer provides better performance robustness in a wide dynamic range of weight variances than the baseline USQ, as it is a case with the theoretical results. Based on the

presented analysis and both, theoretical and experimental results of our LCSQ model, we have anticipated that there is a great possibility to implement the proposed 2-bit LCSQ in NN compression.

In brief, the new ideas presented in the paper are: 1) the description of the 2-bit μ-law logarithmic quantizer opposed to the usually used high resolution μ-law logarithmic quantizers and its design in according to the exact analysis opposed to the usually used asymptotic analysis; 2) the comparison of exact and asymptotic analysis showing the dominance of the exact analysis for the 2-bit case; 3) optimization of the maximal amplitude of the quantizer in order to increase its robustness; 4) the application of the designed μ-law logarithmic quantizer for quantization and compression of weights of neural networks opposed to commonly applied uniform quantizer; 5) the analysis of quantization of NN weights in the wide range of weights variances, using SQNR as an objective measure.

ACKNOWLEDGMENT

This work has been supported by the Science Fund of the Republic of Serbia (Grant No. 6527104, AI- Com-in-AI).

REFERENCES

[1] A. Zhang, Z. C. Lipton, M. Li, A. J. Smola, Dive into Deep Learning. Amazon Science, 2020

[2] Z. Nagy, Artificial Intelligence and Machine Learning Fundamentals: Develop Real-World Applications Powered by the Latest AI Advances. Packt Publishing, 2018

[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. of the International Conference on Neural Information Processing Systems, Harrahs and Harveys, Lake Tahoe, NV, USA, 2012, pp. 1097–1105

[4] G. Mukhtar, S. Farhan, "Convolutional neural network based prediction of conversion from mild cognitive impairment to Alzheimer's disease: A technique using hippocampus extracted from MRI," Advances in Electrical and Computer Engineering, vol. 20, no. 2, pp. 113–122, 2020. doi:10.4316/AECE.2020.02013

[5] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Proc. of the Conference on Advances in neural information processing systems (NeurIPS), Montreal, Canada, 2015, pp. 91–99

[6] A. Conneau, H. Schwenk, L. Barrault, Y. Lecun, "Very deep convolutional networks for text classification," arXiv preprint arXiv:1606.01781, 2016

[7] V. Delić, Z. Perić, M. Sečujski, N. Jakovljević, J. Nikolić, et. al., "Speech technology progress based on new machine learning paradigm," Computational Intelligence and Neuroscience, vol. 2019, Article ID 4368036, 19 pages, 2019. doi:10.1155/2019/4368036

[8] A. Albu, R. E. Precup, T. A. Teban, "Results and challenges of artificial neural networks used for decision-making and control in medical applications," Facta Universitatis Series: Mechanical Engineering, vol. 17, no. 3, pp. 285–308, 2019. doi:10.22190/FUME190327035A

[9] M. U. Ahmed, S. Brickman, A. Dengg,N. Fasth, M. Mihajlovic, et. al., "A machine learning approach to classify pedestrians' events based on IMU and GPS," International Journal of Artificial Intelligence, vol. 17, no 2, pp. 154-167, 2019

[10] U. L. Yuhana, N. Z. Fanani, E. M. Yuniarno, S. Rochimah, L. T. Koczy, et. al., "Combining fuzzy signature and rough sets approach for predicting the minimum passing level of competency achievement," International Journal of Artificial Intelligence, vol. 18, no. 1, pp. 237–249, 2020

[11] Y. Sheng, H. Ma, W. Xia, "A Pointer neural network, for the vehicle routing problem with task priority and limited resources," Information Technology and Control, vol. 49, no.2, pp. 237−248, 2020. doi:10.5755/j01.itc.49.2.24613

[12] Y. Li, Y. Bao, W. Chen, "Fixed-sign binary neural network: An efficient design of neural network for Internet-of-Things devices," IEEE Access, vol. 8, pp. 164858–164863, 2018. doi:10.1109/ACCESS.2020.3022902

[13] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," Journal of Machine Learning Research, vol. 18, no. 1, pp. 6869–6898, 2017. doi:10.5555/3122009.3242044

[14] D. Lin, S. Talathi, S. Annapureddy, "Fixed point quantization of deep convolutional networks," in Proc. of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016

[15] R. Banner, Y. Nahshan, E. Hoffer, D. Soudry, "ACIQ: Analytical clipping for integer quantization of neural networks," arXiv preprint arXiv:1810.05723, 2018

[16] L. Enderich, F. Timm, W. Burgard, "SYMOG: Learning symmetric mixture of Gaussian modes for improved fixed-point quantization," Neurocomputing, vol. 416, pp. 310−315, 2020. doi:10.1016/j.neucom.2019.11.114

[17] A. Nannarelli, "Variable precision 16-bit floating-point vector unit for embedded processors," in Proc. of IEEE 27th Symposium on Computer Arithmetic, (ARITH 2020), Portland, OR, USA, 2020

[18] R. Banner, I. Hubara, E. Hoffer, D. Soudry, "Scalable methods for 8-bit training of neural networks," in Proc. of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada, 2018

[19] R. Banner, Y. Nahshan, D. Soudry, "Post training 4-bit quantization of convolutional networks for rapid-deployment," in Proc. of the 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2019

[20] J. Choi, S. Venkataramani, V. Srinivasan, K. Gopalakrishnan, Z. Wang, et. al., "Accurate and efficient 2-bit quantized neural networks," in Proc. of the 2nd SysML Conference, Stanford, CA, USA, 2019

[21] L. Deng, P. Jiao, J. Pei, Z. Wu, G. Li, "GXNOR-Net: Training deep neural networks with ternary weights and activations without full-precision memory under a unified discretization framework," Neural Networks, vol. 100, pp. 49-58, 2018. doi:10.1016/j.neunet.2018.01.010

[22] H. Qina, R. Gonga, X. Liu, X. Baie, J. Songc, et. al, "Binary neural networks: A survey," Pattern Recognition, vol. 105, Article ID: 107281, 2020. doi:10.1016/j.patcog.2020.107281

[23] Z. Peric, B. Denic, M. Savic, V. Despotovic, "Design and analysis of binary scalar quantizer of Laplacian source with applications," Information, vol. 11, 18 pages, 2020. doi:10.3390/info11110501

[24] N. S. Jayant, P. Noll, "Digital coding of waveforms: Principles and applications to speech and video,", New Jersey, Prentice Hall, Chapter 4, pp. 115−188, 1984

[25] S. Na, "Asymptotic formulas for mismatched fixed-rate minimum MSE Laplacian Quantizers," IEEE Signal Processing Letters, vol. 15, pp. 13−16, 2008. doi:10.1109/LSP.2007.910240

[26] Z. Peric, G. Petkovic, B. Denic, A. Stanimirovic, V. Despotovic, et. al., "Gaussian source coding using a simple switched quantization algorithm and variable length codewords," Advances in Electrical and Computer Engineering, vol. 20, no. 4, pp. 11−18, 2020. doi:10.4316/AECE.2020.04002

[27] S. Tomić, Z. Perić, M. Tančić, J. Nikolić, "Backward adaptive and quasi-logarithmic quantizer for sub-band coding of audio," Information Technology and Control, vol. 47, no. 1, pp. 131−139, 2018. doi:10.5755/j01.itc.47.1.16190

[28] M. Dincic, Z. Peric, D. Denic, Z. Stamenkovic, "Design of robust quantizers for low-bit analog-to-digital converters for Gaussian source," Journal of Circuits, Systems and Computers, vol. 28, no. supp01, 1940002, 2019. doi:10.1142/S0218126619400024

[29] Z. Perić, J. Nikolić, D. Aleksić, A. Perić, "Symmetric quantile quantizer parameterization for the Laplacian source: Qualification for contemporary quantization solutions," Mathematical Problems in Engineering, vol. 2021, Article ID 6647135, 12 pages, 2021. doi:10.1155/2021/6647135

[30] S. Gazor, W. Zhang, "Speech probability distribution," IEEE Signal Processing Letters, vol. 10, pp. 204−207, 2003. doi:10.1109/LSP.2003.813679

[31] Y. LeCun, C. Cortez, C. Burges, "The MNIST Handwritten Digit Database," available online: yann.lecun.co