# Binary Quantization Analysis of Neural Networks Weights on MNIST Dataset

**Zoran H. Peric[1], Bojan D. Denic[1], Milan S. Savic[2], Nikola J. Vucic[1, *], Nikola B. Simic[3]**
*[1]Faculty of Electronic Engineering, University of Nis,
Aleksandra Medvedeva 14, 18000 Nis, Serbia
[2]Faculty of Sciences and Mathematics, University of Pristina,
Ive Lole Ribara 29, 38220 Kosovska Mitrovica, Serbia
[3]Faculty of Technical Sciences, University of Novi Sad,
Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia
nikola.vucic@elfak.ni.ac.rs*

***Abstract*—This paper considers the design of a binary scalar quantizer of Laplacian source and its application in compressed neural networks. The quantizer performance is investigated in a wide dynamic range of data variances, and for that purpose, we derive novel closed-form expressions. Moreover, we propose two selection criteria for the variance range of interest. Binary quantizers are further implemented for compressing neural network weights and its performance is analysed for a simple classification task. Good matching between theory and experiment is observed and a great possibility for implementation is indicated.**

***Index Terms*—Image classification; Multilayer perceptron; Neural network; Quantization; Source coding.**

## I. Introduction

Artificial neural networks (NNs) have become an attractive research field in recent decades for resolving different challenges due to the increasing availability of powerful hardware [1]. It is worth mentioning that the most significant achievements have been provided in tasks, such as image classification [2], object recognition [3], and speech processing [4]. However, the application in other fields has also been performed, where some promising results have been achieved [5]–[7].

Specifically, the improved performance (i.e., high prediction accuracy level) has often been provided using very complex NN architectures, with a large amount of parameters, computational and storage resources. This in turn can be a limiting factor for the application of NNs in portable and edge computing devices with limited memory and processing power, or in latency-critical services. Hence, the need for NN compression is evident and quantization is a widely used approach for that purpose. In that case, NN parameters (weights, activations, etc.), usually represented in 32-bits floating point format (full precision), are mapped to fixed-point representations using lower bit lengths.

The compression of NN parameters and various challenges have been observed in many research papers, where different codewords have been used, whose lengths are 8 bits [8], 4 bits [9], or 2 bits [10]. In addition, even lower representations using ternary [11] and binary [12]–[17] quantization have been considered, where a significant compression ratio accompanied with the competitive accuracy level have been offered by the quantized NN. Hence, binary quantization takes an important role in the compression of NN parameters and deserves to be explained in detail from the view of both signal processing and NN performance. This kind of analysis is supported in this paper and it has not been conducted in previous works [12]–[16]. Regarding the recently published paper [17], where a comprehensive analysis of the binary quantizer, including adaption and application, has been provided, here we deal with the design of a fixed (non-adaptive) quantizer for a wide dynamic range. This is important as the same quantizer can be used for different Laplacian inputs. Note also that Laplacian distribution can describe well the weights of NN [9], as well as speech [18]–[20]. In particular, we derive closed-form expressions for performance estimation and introduce two criteria for the selection of the quantizer for the variance range of interest. Theoretical results are further verified on real data using the weights of NN observed for the handwritten digit classification problem. The influence of binarized weights on prediction accuracy is also investigated and the relation between the weight quality (measured by SQNR) and accuracy is established, which has not been done so far.

The rest of the paper is organized as follows. In Section II, the design method for the optimal quantizer with respect to distortion is given. In Section III, the analysis in a wide dynamic range is provided in detail and criteria for selecting the quantizer for the defined variance range are proposed. In Section IV, the experimental results obtained by implementation in neural networks are summarized and discussed. Finally, we conclude the paper in Section V.

## II. Design of Binary Quantizer for the Reference Variance

Let us consider a symmetrical binary ($N$ = 2 levels) scalar

Manuscript received 31 January, 2021; accepted 7 May, 2021.

This work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under Grant No. TR32035 and by the Science Fund of the Republic of Serbia under Grant No. 6527104, AI-Com-in-AI.

quantizer presented in Fig. 1. With $\alpha$, the representational level in the positive range (the level in the negative range is simply reflection of the positive one) is denoted. Next, $x_{\max}$ denotes the maximal data limit, where $\alpha = x_{\max}/2$.

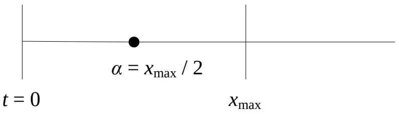

Fig. 1. The observed binary scalar quantizer.

Let the input data source be described by the Laplacian probability density function (PDF) given by [18]

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

where $\sigma^2$ is the variance of the data. If we adopt the unit variance as the reference one ($\sigma^2 = \sigma_0^2 = 1$), denoting the standard approach in scalar quantization [18], then PDF takes the following form

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

Given an input data source, the distortion in the case of symmetrical binary quantizer can be evaluated as

$$\mathrm{D} = 2\int_0^\infty (\mathrm{x}-\alpha)^2\, \mathrm{p}(\mathrm{x},\sigma=1)\,\mathrm{dx} = 1-\sqrt{2}\,\alpha+\alpha^2, \tag{3}$$

or in terms of $x_{\max}$

$$\mathrm{D} = 1 - \frac{\mathrm{x}_{\max}}{\sqrt{2}} + \frac{\mathrm{x}_{\max}{}^2}{4}. \tag{4}$$

Signal to quantization noise ratio (SQNR) is specified as

$$\mathrm{SQNR} = 10\log_{10}\left(\frac{\sigma_0{}^2}{D}\right) = 10\log_{10}\left(\frac{1}{1-\frac{x_{\max}}{\sqrt{2}}+\frac{x_{\max}{}^2}{4}}\right). \tag{5}$$

Figure 2 shows how $x_{\max}$ affects the SQNR. The commonly used criterion for the quantizer design is the maximal SQNR (or equivalently minimal distortion) [18]. Given results, the required criterion is accomplished for $x_{\max} = 1.4142$ ($\alpha = 0.7071$). This can also be verified using the following lemma.

*Lemma 1.* The value of $x_{\max}$ of Laplacian binary quantizer optimized in terms of distortion is specified as

$$x_{\max}^{opt} = \sqrt{2}. \tag{6}$$

*Proof.* Finding the first derivative of the distortion with respect to $x_{\max}$ and further equalling it to zero results in

$$\frac{\partial D}{\partial x_{\max}} = -\frac{1}{\sqrt{2}} + \frac{x_{\max}}{2} = 0. \tag{7}$$

From the last equation we obtain $x_{\max} = \sqrt{2}$, which concludes the proof. Based on Lemma 1 and relation among the quantizer parameters, it holds that $\alpha^{opt} = 1/\sqrt{2}$. Note also that for $x_{\max} = 2$ (i.e., $\alpha = 1$) we obtain the qunatizer widely used in NN applications [12]–[16], providing 0.7 dB lower SQNR than optimal one.

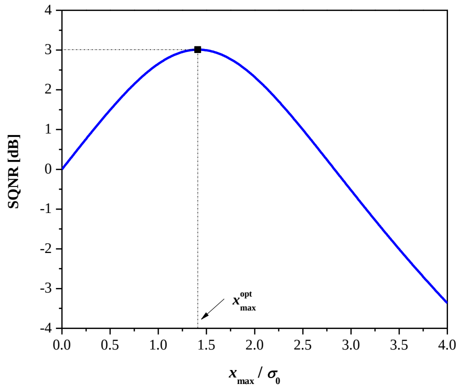

Fig. 2. Dependence of SQNR on $x_{\max}$ for the binary quantizer ($\sigma_0^2 = 1$)

## III. DESIGN OF BINARY SCALAR QUANTIZER FOR A WIDE DYNAMIC RANGE

In this section, we consider the situation when a binary quantizer (designed for the particular variance) is applied on the Laplacian inputs having a variance different from the designed one. This is known as variance-mismatched quantization [18], [21]. It is familiar that variance-mismatch effect reduces the efficiency of the quantization model over the broad variance range. Hence, robust quantization models are recommended for non-stationary data processing, as they can satisfy minimal quality requirements over the entire range. Here, we will analyse the binary quantizer in a wide dynamic range and derive expressions for performance evaluation. In addition, criteria for selection of a binary quantizer in the established variance range of interest will be proposed.

### *A. Derivation of Expressions for Performance Evaluation*

To evaluate the performance of the binary quantizer in a wide dynamic range of the input data variances, we use PDF defined with (1). Hence, we estimate the distortion as

$$\begin{aligned} D(\sigma) &= 2\int_0^\infty \left(x-\alpha(\sigma_0)\right)^2 p(x,\sigma)\,dx = \\ &= \sigma^2 - \sqrt{2}\,\alpha(\sigma_0)\sigma + \alpha^2(\sigma_0), \end{aligned} \tag{8}$$

or equivalently

$$D(\sigma) = \sigma^2 - \frac{x_{\max}(\sigma_0)}{\sqrt{2}}\sigma + \frac{x_{\max}{}^2(\sigma_0)}{4}, \tag{9}$$

where $\alpha(\sigma_0) = \alpha$ and $x_{\max}(\sigma_0) = x_{\max}$ denote the values of representation level and maximal data limit in the case of variance $\sigma_0^2$ (see Section II). Next, considering the previous expressions, SQNR is given by

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

$$=10\log_{10}\left(\frac{\sigma^2}{\sigma^2-\frac{x_{\max}(\sigma_0)}{\sqrt{2}}\sigma+\frac{x_{\max}{}^2(\sigma_0)}{4}}\right). \qquad (10)$$

Figure 3 plots SQNR (10) in the variance range (-10 dB, 25 dB) with respect to $\sigma_0^2 = 1$, when $x_{\max} = 1/2$, $x_{\max} = 1$, $x_{\max} = \sqrt{2}$, $x_{\max} = 2$, and $x_{\max} = 4$. It can be observed that all SQNR curves attain the same maximum (same as the optimal quantizer in Section II), but the SQNR does not retain the constant value in the rest of the variance range and it rapidly decreases. Accordingly, the quantizer robustness is low and the efficiency on non-stationary data is limited.

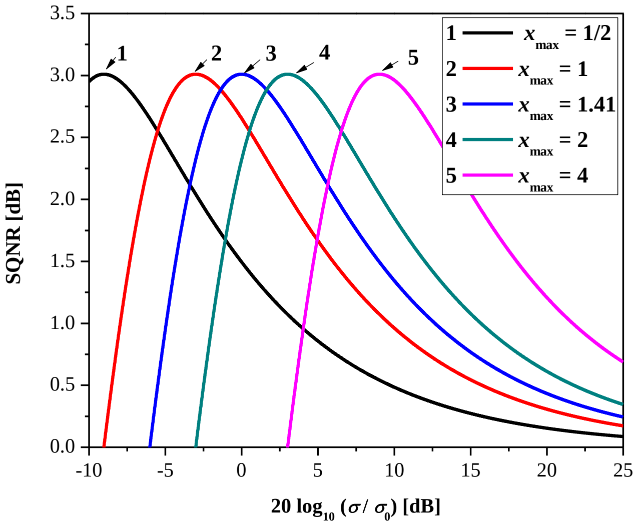

Fig. 3. SQNR in a wide dynamic range of input data variances for different values of parameter $x_{\max}$.

It is also important to discuss the impact of parameter $x_{\max}$ on the design approaches presented here (wide dynamic range) and in Section II (particular variance). While in the approach in Section II selection of non-optimal $x_{\max}$ value ($x_{\max} \neq \sqrt{2}$ ) causes the degradation in SQNR (see Fig. 2), here it causes shifting the curve left or right from the one with optimal value of $x_{\max}$ ($x_{\max} = \sqrt{2}$ ). Note also that each SQNR curve attains its maximum at different variance points, which is defined with the following lemma.

*Lemma 2.* Given variance range and parameter $x_{\max}$, the binary quantizer attains the maximum SQNR at the point specified as

$$\sigma = \frac{\mathrm{x}_{\max}(\sigma_0)}{\sqrt{2}}. \qquad (11)$$

*Proof.* Let us define the function $F$ as

$$F = \frac{\sigma^2}{D(\sigma)} = \frac{\sigma^2}{\sigma^2-\frac{x_{\max}}{\sqrt{2}}\sigma+\frac{x_{\max}{}^2}{4}}. \qquad (12)$$

Taking the first derivative of $F$ with respect to $\sigma$ results in

$$\frac{\partial F}{\partial\sigma} = \frac{2\sigma\left(\sigma^2-\frac{x_{\max}}{\sqrt{2}}\sigma+\frac{x_{\max}^2}{4}\right)-\left(2\sigma-\frac{x_{\max}}{\sqrt{2}}\right)\sigma^2}{\left(\sigma^2-\frac{x_{\max}}{\sqrt{2}}\sigma+\frac{x_{\max}^2}{4}\right)^2}. \qquad (13)$$

Furthermore, equalling (13) to zero, i.e., $\frac{\partial F}{\partial\sigma} = 0$ and solving with respect to $\sigma$, we obtain

$$\sigma = \frac{\mathrm{x}_{\max}}{\sqrt{2}}, \qquad (14)$$

or in terms of $\alpha$

$$\sigma = \frac{2\alpha}{\sqrt{2}}, \qquad (15)$$

which concludes the proof.

By replacing (14) in (12) and taking the logarithm (base 10), we obtain SQNR = 3.01 dB, the same as in Section II.

In addition, we will show that SQNR and distortion attain their extreme values (maximum or minimum) at different variance points.

Figure 4 shows the distortion (9) as a function of $\sigma$ (the values of $\sigma$ are given in the linear domain and the equivalent range in log-domain is [-15 dB, 10 dB]) for the same values of $x_{\max}$ as in the example of Fig. 3.

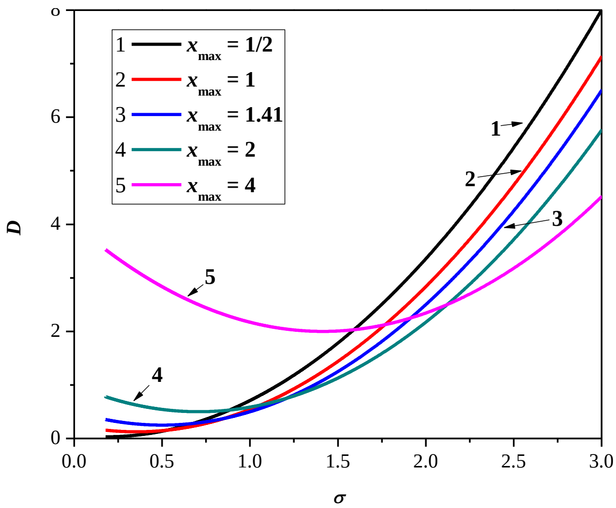

Fig. 4. Distortion versus $\sigma$ for the binary quantizer with different values of $x_{\max}$.

Note that each curve attains its minimum at different variance values; the point where the curve minimum is achieved can be determined as the solution $\partial D/\partial\sigma = 0$, which results in

$$\sigma = \frac{x_{\max}}{2\sqrt{2}} \qquad (16)$$

and is different from the one defined in (14). The corresponding values of $\sigma$ for both optimization functions, SQNR and $D$, in the case of different $x_{\max}$ are given in Table I.

In the following subsection, we provide the criteria for selecting the best quantizer (i.e., the appropriate $x_{\max}$ value) either for a particular variance and a range of variances having the width smaller than 35 dB.

TABLE I. THE VALUES OF $\sigma$ FOR WHICH DISTORTION IS MINIMIZED AND SQNR IS MAXIMIZED FOR DIFFERENT $x_{\max}$.

| $x_{\max}$ | $\alpha$ | $\sigma$ (SQNR) | $\sigma$ (Distortion) |
|---|---|---|---|
| 1/2 | 1/4 | 0.354 (-9.03 dB) | 0.178 |
| 1 | 1/2 | 0.707 (-3.01 dB) | 0.354 |
| $\sqrt{2}$ | $1/\sqrt{2}$ | 1 (0 dB) | 0.5 |
| 2 | 1 | 1.414 (3.01 dB) | 0.707 |
| 4 | 2 | 2.83 (9.03 dB) | 1.414 |

## B. Criteria for Selection of the Binary Quantizer

Firstly, we consider scenario when the best quantizer from the set of quantizers (i.e., the ones with different $x_{\max}$) needs to be selected for the particular variance in the defined variance range, observing SQNR as a performance criterion. Thus, by direct observing Fig. 3 for the variance defined in the point 0 dB ($\sigma$ = 1 in the linear domain), the best quantizer is the one with $x_{\max} = \sqrt{2}$ achieving the SQNR of 3.01 dB (as indicated in Section II). On the other hand, for the variance points, e.g., 15 dB ($\sigma$ = 5.62) and 20 dB ($\sigma$ = 10), the binary quantizer with $x_{\max}$ = 4 is the best since it provides SQNR of nearly 3 dB and 1.25 dB, respectively, and outperforms the other observed quantizers.

The following two criteria are proposed for selecting the best binary quantizer for the variance range of interest.

The first criterion proposes the selection of the quantizer such that the maximal average SQNR ($\text{SQNR}_{av}$) is achieved in a defined (fixed) variance range

$$\text{SQNR}_{av} = \frac{1}{m}\sum_{i=1}^{m}\text{SQNR}\left(\sigma_i\right), \tag{17}$$

where $m$ is the number of observed variances $\sigma_i$ taken from that fixed range.

With the second criterion, we want to emphasize the importance of robustness. Thus, besides taking into account $\text{SQNR}_{av}$, the best quantizer has to fulfil one additional condition in the given range

$$\text{SQNR} \geq \text{SQNR}_{\min} = 1\text{dB}, \tag{18}$$

where $\text{SQNR}_{\min}$ defines the minimal SQNR that should be achieved in the desired range. In other words, if in the defined range the quantizers achieve $\text{SQNR}_{av}$ values that are very close, then the best quantizer will be chosen the one providing the widest interval where criterion (18) is fulfilled.

From the theoretical SQNR curves in Fig. 3, it can be shown that the width of the range where condition (18) is fulfilled is equal and amounts to approximately 17.6 dB. Furthermore, the borders of that range denoted as ($\sigma_{\min}$, $\sigma_{\max}$) for each curve (obtained for different $x_{\max}$) can be calculated as solutions of the following equation

$$10\log_{10}\left(\frac{\sigma^2}{\sigma^2 - \frac{x_{\max}}{\sqrt{2}}\sigma + \frac{x_{\max}{}^2}{4}}\right) \geq 1, \tag{19}$$

and are provided in Table II.

TABLE II. THE BORDER VALUES OF THE RANGE WHERE SQNR ≥ 1 dB FOR BINARY QUANTIZER WITH DIFFERENT $x_{\max}$.

| $x_{\max}$ | $\sigma_{\min}$ | $\sigma_{\max}$ | $20\log_{10}\sigma_{\min}$ [dB] | $20\log_{10}\sigma_{\max}$ [dB] |
|---|---|---|---|---|
| 1/2 | 0.200 | 1.519 | -13.98 | 3.63 |
| 1 | 0.400 | 3.038 | -7.96 | 9.65 |
| $\sqrt{2}$ | 0.566 | 4.296 | -4.95 | 12.66 |
| 2 | 0.800 | 6.076 | -1.94 | 15.67 |
| 4 | 1.600 | 12.152 | 4.08 | 21.53 |

Table III summarizes the calculated values of $\text{SQNR}_{av}$ of the binary quantizer with different $x_{\max}$ for three arbitrary selected variance ranges. According to the first criterion (maximal $\text{SQNR}_{av}$ in the observed range), we conclude that the binary quantizer with $x_{\max}$ = 2 is the best for the ranges of [-3 dB, 15 dB] and [0 dB, 18 dB], while the binary quantizer with $x_{\max}$ = 4 is the best for the range of [5 dB, 25 dB].

Table IV includes the width of the interval (in decibels) within the chosen variance ranges, where condition (18) is fulfilled. Using Tables III and IV, the basis for the application of the second criterion is provided. Namely, the results show matching with the first criterion, as the same quantizers are chosen for the established variance ranges.

TABLE III. $\text{SQNR}_{av}$ [dB] FOR DIFFERENT BINARY QUANTIZERS AND VARIANCE RANGES.

| $x_{\max}$ | [-3 dB, 15 dB] | [0 dB, 18 dB] | [5 dB, 25 dB] |
|---|---|---|---|
| 1/2 | 0.896 | 0.642 | 0.337 |
| 1 | 1.638 | 1.235 | 0.667 |
| $\sqrt{2}$ | 1.996 | 1.639 | 0.929 |
| 2 | 2.060 | 1.997 | 1.270 |
| 4 | 0.414 | 1.571 | 1.937 |

TABLE IV. WIDTH OF THE INTERVAL WITHIN THE CORRESPONDING VARIANCE RANGE WHERE SQNR ≥ 1 dB FOR BINARY QUANTIZER WITH DIFFERENT $x_{\max}$.

| $x_{\max}$ | [-3 dB, 15 dB] | [0 dB, 18 dB] | [5 dB, 25 dB] |
|---|---|---|---|
| 1/2 | 6.63 | 3.63 | 0 |
| 1 | 12.65 | 9.65 | 4.65 |
| $\sqrt{2}$ | 15.66 | 12.66 | 7.66 |
| 2 | 16.94 | 15.67 | 10.67 |
| 4 | 10.92 | 13.92 | 16.53 |

# IV. EXPERIMENTAL RESULTS AND DISCUSSION

The goal of the section is to verify the theoretical analysis provided in previous Section III by applying a binary quantizer in processing the weights of NN. In addition, we will investigate the influence of binarized weights on NN performance, measured by prediction accuracy [1].

Our experiment is focused on the feedforward neural network named the “multilayer Perceptron” (MLP) [1]. This is a classical network and it is composed of input, hidden, and output layers. We use MNIST database [22] as input, having 60.000 monochrome images of handwritten single digits of dimension 28×28 pixels, where for training and testing purposes, 50000 and 10000 images, respectively, are used. Accordingly, MLP is used for classification tasks; the number of nodes in the input, hidden, and output layers is 784 (28×28), 128, and 10 (the number of classes), respectively. Rectified Linear Unit (*ReLU*) and *softmax* activation functions are used in the hidden and output layer, respectively. In addition, the regularization rate, learning rate, number of iterations per epoch, and batch size are set to 0.01, 0.0005, 468, and 128, respectively.

The MLP NN is trained for 20 epochs achieving the prediction accuracy score of 96.7 % (in this case, the weights are represented using 32-bit floating point format (full precision)). The histogram of the learned weights is depicted in Fig. 5. Given the figure, one can note that the weights can be approximated with Laplacian PDF of variance $\sigma_w^2$ and mean $\mu_w$ (in our case, $\mu_w$ tends to zero), providing the basis for implementation of the considered binary quantizer (post-training quantization will be performed).

The efficiency of the quantizer on the real data is measured using the SQNR$^{ex}$ that (assuming the zero-mean) is defined as

$$\mathrm{SQNR}^{ex} = 10\log_{10}\left(\frac{\sigma_w^2}{D_w}\right) = 10\log_{10}\left(\frac{\frac{1}{W}\sum_{i=1}^{W} w_i^2}{\frac{1}{W}\sum_{i=1}^{W}\left(w_i - w_i^q\right)^2}\right), \quad (21)$$

where $D_w$ is the distortion obtained by binarization of weights, $W$ is the total number of weights, $w_i$ is the original, and $w_i^q$ is the quantized value of weights.

The correctness of the theoretical results (in terms of SQNR) is investigated on the range of [0 dB, 18 dB] in relation to the reference variance that is set to be one of the original weights ($\sigma_w^2$). Thus, in Table V, the SQNR$^{ex}$ values for some selected points from the observed range are summarized, considering the quantizer with different values of $x_{\max}$. For illustration purposes, we plotted in Fig. 6 the SQNR$^{ex}$ versus the variance of the data (weights).

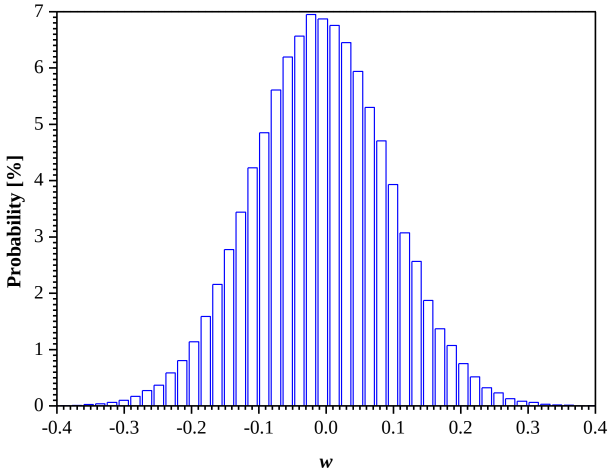

Fig. 5. Histogram of trained weights.

The selection of the best quantizer based on experimental results will be done using the criteria proposed in Section III. In Table VI, the values of the average SQNR$^{ex}$ (SQNR$_{av}^{ex}$) and the width of the interval, where SQNR$^{ex}$ is higher than 1 dB (denoted by Δ), achieved in the range under question by different binary quantizers, are listed.

TABLE V. EXPERIMENTAL RESULTS: THE VALUES OF SQNR$^{EX}$ AND PREDICTION ACCURACY (PA) IN CASE OF BINARY QUANTIZER (DIFFERENT $x_{\max}$) AND DIFFERENT VARIANCES OF WEIGHTS.

| $\sigma_w^2$ [dB] | $x_{\max} = 1/2$ | | $x_{\max} = 1$ | | $x_{\max} = \sqrt{2}$ | | $x_{\max} = 2$ | | $x_{\max} = 4$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SQNR [dB] | PA [%] | SQNR [dB] | PA [%] | SQNR [dB] | PA [%] | SQNR [dB] | PA [%] | SQNR [dB] | PA [%] |
| 0 | 1.779 | 85.07 | 3.441 | 90.82 | 4.282 | 91.55 | 3.918 | 91.93 | -2.580 | 92.24 |
| 3.52 | 1.180 | | 2.368 | | 3.264 | | 4.184 | | 1.858 | |
| 6.02 | 0.881 | | 1.779 | | 2.500 | | 3.441 | | 3.918 | |
| 7.96 | 0.703 | | 1.420 | | 2.007 | | 2.821 | | 4.383 | |
| 9.54 | 0.584 | | 1.180 | | 1.671 | | 2.368 | | 4.184 | |
| 10.88 | 0.500 | | 1.009 | | 1.430 | | 2.033 | | 3.815 | |
| 12.04 | 0.437 | | 0.881 | | 1.249 | | 1.779 | | 3.441 | |
| 13.06 | 0.388 | | 0.782 | | 1.108 | | 1.580 | | 3.107 | |
| 13.98 | 0.349 | | 0.703 | | 0.995 | | 1.420 | | 2.821 | |
| 14.81 | 0.317 | | 0.638 | | 0.904 | | 1.289 | | 2.577 | |
| 15.56 | 0.290 | | 0.584 | | 0.827 | | 1.180 | | 2.368 | |
| 16.26 | 0.268 | | 0.539 | | 0.763 | | 1.088 | | 2.189 | |
| 16.90 | 0.249 | | 0.500 | | 0.707 | | 1.009 | | 2.033 | |
| 17.5 | 0.232 | | 0.466 | | 0.660 | | 0.941 | | 1.898 | |
| 18 | 0.217 | | 0.437 | | 0.618 | | 0.881 | | 1.779 | |

It can be perceived that the first criterion proposes a binary quantizer with parameter $x_{\max} = 4$, while according to the second criterion, the best quantizer is designed using $x_{\max} = 2$. Note also that the quantizer designed using $x_{\max} = 2$ has been the theoretical choice for both criteria for the considered range (see Tables III and IV) and matching of the theoretical and experimental results is observed in that case.

In addition, the weights (original weights, as well as the weights having variance different from the original one) quantized using a binary quantizer with various $x_{\max}$ are then separately implemented to MLP for classification purposes on test data (10000 images from MNIST database [22]) and the prediction accuracy is examined. This corresponds to the situation when the same (non-adaptive) binary quantizer is used for different MLP networks (as the set of weights is different in each case). The accuracy scores are provided in Table V, where some interesting conclusions can be derived. Observe that for a given binary quantizer defined with $x_{\max}$, each MLP achieves the same prediction accuracy score, although different SQNRs are provided. This is because the same quantized weights are obtained regardless the variance of the weights (the weights are quantized to the values -$x_{\max}(\sigma_w)/2$, $x_{\max}(\sigma_w)/2$) and thus the quantized MLP is the same. Accordingly, in that case, the relationship between the SQNR and prediction accuracy cannot be uniquely defined (i.e., SQNR does not dominantly contribute to neural network performance).

On the other hand, in Table V, we can see that the accuracy of the quantized MLP increases as the binary quantizer uses higher values of $x_{\max}$. This can be explained as follows. As $x_{\max}$ increases, the distance among the representational levels increases, enabling better classification and higher accuracy scores. The highest performance of the quantized MLP is achieved when quantizer with $x_{\max} = 4$ is applied (92.24 %) and slightly lower when the quantizer with $x_{\max} = 2$ is applied (91.93 %). Note that these two quantizers are already proposed as the most appropriate based on SQNR analysis performed above. Further increasing of the parameter $x_{\max}$ ($x_{\max} > 4$) will result in negligible increasing of MLP performance.

Finally, one can notice that MLP with binarized weights provides a lower accuracy score for 4.46 % ($x_{\max} = 2$) or for 4.84 % ($x_{\max} = 4$) than that achieved with full precision weights, at the same time reducing the network size by a factor 32.

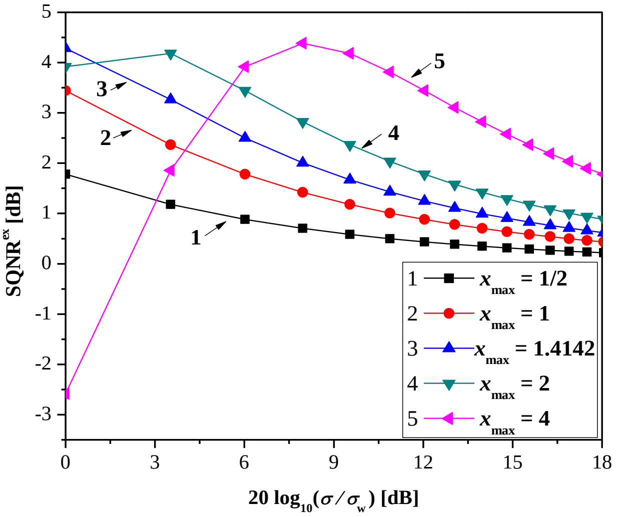

Fig. 6. Experimental results: SQNR$^{ex}$ as the function of weight variance.

TABLE VI. EXPERIMENTAL RESULTS: THE VALUES OF SQNR$_{av}^{ex}$ AND Δ IN THE RANGE OF [0 dB, 18 dB] FOR BINARY QUANTIZER WITH DIFFERENT $x_{\max}$.

| $x_{\max}$ | 1/2 | 1 | $\sqrt{2}$ | 2 | 4 |
|---|---|---|---|---|---|
| **SQNR$_{av}^{ex}$ [dB]** | 0.516 | 1,031 | 1.418 | 1.856 | 2.41 |
| **Δ** | 4.5 | 10.5 | 13.5 | 16.90 | 15 |

## V. CONCLUSIONS

In this paper, a detailed analysis of binary scalar quantization of Laplacian source has been carried out along with an application for compression of NN parameters. Closed-form expressions in terms of SQNR and distortion have been derived for analysis in a wide dynamic range, and two criteria have been proposed to select the best quantizer. Verification of the theoretical results in terms of SQNR achieved in a wide dynamic range and quantizer selection has been done on real data using NN weights. Furthermore, the selected fixed (non-adaptive) binary quantizer has been applied to compress different MLP networks (whose coefficients follow the Laplacian PDF, but have different variances) with a goal to establish relationship among the SQNR and prediction accuracy. It has been shown that each quantized MLP is the same regardless of the weight variance (i.e., the same prediction accuracy has been achieved), although for different weights different SQNRs have been observed. Therefore, the uniquely defined relationship has not been established as SQNR does not dominantly contribute to NN performance. In addition, the relatively high prediction accuracy has been reported (over 92 %), that is only 4.46 % lower than the full-precision model, along with a compression gain of 32 times.

## CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

## REFERENCES

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

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks”, *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. DOI: 10.1145/3065386.

[3] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks”, in *Proc. of the 28th International Conference on Neural Information Processing Systems*, 2015, vol. 1, pp. 91–99.

[4] A. Conneau, H. Schwenk, L. Barrault, and Y. Le Cun, “Very deep convolutional networks for text classification”, in *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, vol. 1, pp. 1107–1116. arXiv: 1606.01781v2.

[5] M. Togacar, B. Ergen, and M. E. Sertkaya, “Subclass separation of white blood cell images using convolutional neural network models”, *Elektronika ir Elektrotechnika*, vol. 25, no. 5, pp. 63−68, 2019. DOI: 10.5755/j01.eie.25.5.24358.

[6] Y. Liu, C. Hu, and Y. Hong, “Electric energy substitution potential prediction based on logistic curve fitting and improved BP neural network algorithm”, *Elektronika ir Elektrotechnika*, vol. 25, no. 3, pp. 18–24, 2019. DOI: 10.5755/j01.eie.25.3.23671.

[7] R. Yayla and B. Sen, “A new classification approach with deep mask r-cnn for synthetic aperture radar image segmentation”, *Elektronika ir Elektrotechnika*, vol. 26, no. 6, pp. 52−57, 2020. DOI: 10.5755/j01.eie.26.6.25849.

[8] R. Banner, I. Hubara, E. Hoffer, and 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, 2018, pp. 5151–5159.

[9] R. Banner, Y. Nahshan, and 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, 2019.

[10] J. Choi, S. Venkataramani, V. Srinivasan, K. Gopalakrishnan, Z. Wang, and P. Chuang, “Accurate and efficient 2-bit quantized neural networks”, *in Proc. of the 2nd MLSys Conference*, Stanford, 2019.

[11] C. Zhu, S. Han, H. Mao, and W. J. Dally, “Trained ternary quantization”, *in Proc. of the 5th International Conference on Learning Representations (ICLR 2017),* Toulon, 2017.

[12] H. Qin, R. Gong, X. Liu, X. Bai, J. Song, and N. Sebe, “Binary neural networks: A survey”, *Pattern Recognition*, vol. 105, art. 107281, 2020. DOI: 10.1016/j.patcog.2020.107281.

[13] T. Simons and D.-J. Lee, “A review of binarized neural networks”, *Electronics*, vol. 8, no. 6, p. 661, 2019. DOI: 10.3390/electronics8060661.

[14] Y. Wang, J. Lin, and Z. Wang, “An energy-efficient architecture for binary weight convolutional neural networks”, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 2, pp. 280–293, 2018. DOI: 10.1109/TVLSI.2017.2767624.

[15] Y. Li, Y. Bao, and 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.

[16] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized neural networks”, in *Proc. of the 30th Conference on Neural Information Processing Systems (NIPS)*, Barcelona, 2016, pp. 4114–4122.

[17] Z. Peric, B. Denic, M. Savic, and V. Despotovic, “Design and analysis of binary scalar quantizer of Laplacian source with applications”, *Information*, vol. 11, no. 11, p. 501, 2020. DOI: 10.3390/info11110501.

[18] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall: New Jersey, USA, 1984.

[19] S. Gazor and W. Zhang, “Speech probability distribution”, *IEEE Signal Processing Letters*”, vol. 10, pp. 204–207, 2003. DOI: 10.1109/LSP.2003.813679.

[20] Z. Peric, B. Denic, and V. Despotovic, “Multilevel delta modulation with switched first-order prediction for wideband speech coding”, *Elektronika ir Elektrotechnika*, vol. 24, no. 1, pp. 46−51, 2018. DOI: 10.5755/j01.eie.24.1.20156.

[21] 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

[22] Y. LeCun, C. Cortez, and C. Burges, “The MNIST handwritten digit database”. [Online]. Available: http://yann.lecun.com/