FACTA UNIVERSITATIS
Series: **Automatic Control and Robotics** Vol. 21, Nº 1, 2022, pp. 37 - 46
https://doi.org/10.22190/FUACR220321004A

Regular Paper

ONE-BIT QUANTIZER PARAMETRIZATION FOR ARBITRARY LAPLACIAN SOURCES

UDC (621.391:517.988)

Danijela Aleksić, Zoran Perić

University of Niš, Faculty of Electronic Engineering,
Department of Telecommunications, Republic of Serbia

Abstract. *In this paper we suggest an exact formula for the total distortion of one-bit quantizer and for the arbitrary Laplacian probability density function (pdf). Suggested formula additionally extends normalized case of zero mean and unit variance, which is the most applied quantization case not only in traditional quantization rather in contemporary solutions that involve quantization. Additionally symmetrical quantizer's representation levels are calculated from minimal distortion criteria. Note that one-bit quantization is the most sensitive quantization from the standpoint of accuracy degradation and quantization error, thus increasing importance of the suggested parameterization of one-bit quantizer.*

Key words: *Laplacian source, one-bit quantization, symmetric quantizer design*

1. INTRODUCTION

Over the last few decades, numerous quantization methods have been suggested, which try to find a manner for minimizing the number of bits for real-valued presentations, striving for as higher as possible presentation accuracy [1]-[5]. The area of quantization application has constantly spread starting from traditional areas, i.e. information theory and digital signal processing, to come to the forefront in neural network (NN) field, primarily in resource-constrained environments [6]-[8]. While quantization in digital signal processing, as the signal compression method, fundamentally tries to minimize the difference between the quantized and the original signal, this minimization of inevitable quantization error is not recognized as the main target for quantized NN (QNN) models [6]-[8]. Instead, the main goal in indisputably attractive QNN area is to find the appropriate reduced-precision presentations, still generalizing well and attaining the high or satisfactory accuracy of the

Received March 21, 2022 / Accepted April 23, 2022
Corresponding author: Danijela Aleksić
University of Niš, Faculty of Electronic Engineering, Department of Telecommunications,
Aleksandra Medvedeva 14, 18000 Niš, Republic of Serbia
E-mail: danijelaal@telekom.rs

applied NN model. Application of quantization, individually or jointly with some other method that significantly shrink an NN model size, one can see as an important driving force for the NN deployment on edge devices, especially on IoT devices [9]. Generally, to benefit from quantization deployment in hardware-dependent NN solutions, it is important to select a preferable quantizer model and to leverage the knowledge about quantizers and their parameterization [1].

Moreover, here it is very important to emphasize that all data or datasets in many real-world scenarios are not accessible due to privacy, proprietary or security reasons. Zero shot quantization (ZSQ) anticipates the quantization scheme that does not require an access to the original data [6]. Recently, much attention has been given to automatic speech recognition (ASR) models [10]. In paper [10], to calibrate and finetune the quantization model, synthetic data are emerged or generated. These synthetic data directly accommodate to the internal data statistics by the minimization of the Kullback-Leibler (KL) divergence. Here synthetic data, that capture data pattern are leveraged to overcome the drawback of original data. If the access to the original data is infeasible, an obvious question arises - how to quantize this data if ultra low-bit presentations are required in hardware-constrained environments?

In the scope of this paper for ultra low-bit quantizer solution of particular interest are quantizer design and its optimization to the input data. We want to tackle the quantizer efficiency challenge, primarily important in extremely low-bit quantization, due to only two representations available. To address one-bit quantization, that poses a real challenge, we do not use the ACIQ method for the analytical clipping range determination [11], since clipping effect nullify the overload distortion. Instead, we indeed calculate the total distortion for the unrestricted Laplacian probability density function (pdf) of arbitrary mean and standard deviation. The reason why we deal with the Laplacian distribution with heavy tails is because it well describes many real-world scenarios or phenomena [12]. In particular, our contributions are as follows:

- We propose an exact formula for the total distortion when the one-bit quantizer is used, while the input data are well described with an unrestricted arbitrary Laplacian pdf, especially having in mind that most of the data or datasets do not necessarily tend to the zero mean and unit variance.
- Our framework supports the additional optimization of the symmetric one-bit quantizer for the arbitrary Laplacian distribution.

The rest of paper is organized as follows: Section 2 describes the main motivation for accepting an arbitrary Laplacian pdf in our analysis, while Section 3 offers one-bit quantizer parameterization for the assumed Laplacian pdf. Section 4 provides the discussion on the performance achieved with the proposed quantizer. Section 5 summarizes and concludes on our research results.

2. WHY ARBITRARY LAPLACIAN PDF?

The Laplasian pdf, given by (1) is unimodal, log-concave pdf with a more pronounced peak and heavy tails [1]:

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

where μ is the mean and σ is the variance of input x.

Namely, the Laplacian distribution plays a prominent role in the probability theory, statistics and data modelling, since there is a widespread opinion that the Laplacian distribution fits many natural, economical and social phenomena [13]. Since in many real life situations, there is no prior information about data distribution, application of some well-known pdf becomes unavoidable. This explains our motivation to consider the arbitrary Laplacian distribution in the analysis presented in this paper.

Broadly speaking, many datasets exhibit some skewness and do not conform to the symmetry rules. Since we want to qualify our one-bit quantizer for special purposes where the low-complexity of hardware is one of the prerequisites, we will suggest the symmetric quantization model.

Recall that μ is a measure of the central tendency, determining where the values of x tend to be clustered, while σ points how x samples are spread out from μ to form the measure of dispersion [1]. As the pronounced peak is a specific feature of the Laplacian pdf, the largest number of samples x is concentrated around the mean value μ. Referring to our paper [14] we came to interesting conclusions for medium and high bit-rates when forming two granular regions for the restricted Laplacian pdf - Central Granular Region (CGR) and Peripheral Granular Region (PGR). We have shown that in general stands - the higher the bit rate, the higher the percentage of samples falls in the narrower CGR area. More precisely, for the amplitude dynamic defined by [1] and bit-rate 5bit/sample, 85.97% of the samples were concentrated in the CGR covering 31% of the granular region, while for the 8bit/sample, 92.37% of the samples falls in the CGR covering 25.89% of the granular region. As we want here to address the one-bit quantizer with only one pair of the quantization cells and symmetrically placed represents, we believe that we can expect that represents are close enough to the mean.

To the best of the authors' knowledge, an analysis of the influence of the low-bit quantizer's parameterization for the Laplacian pdf on Signal to Quantization Noise Ratio (SQNR) or NN model's accuracy has been addressed in numerous papers [15]-[18]. By applying two-bit and three-bit uniform quantization on the same NN model during the post-training phase, as in [15] and [16], the pure influence of the applied quantization to the NN model's accuracy is isolated. For the known Laplacian-like distribution of weights and MNIST dataset, in [15] and [16] we have proved that the quantizer design and relevant quantizer's representation levels had a stronger impact on the NN model's accuracy for the two-bit quantization case. Our anticipation is that the quantizer's representation levels determination is even more prominent in the one-bit quantization case.

Relying on a plethora of previous conclusions about uniform or nonuniform quantization [14], [19]-[21], further enhancements of one-bit quantizer parameterization are intuitively motivated by the better perceiving of the mean and variance for the arbitrary Laplacian pdf, particularly when the pdf of amplitudes being quantized was known in advance. Moreover, as a unique contribution of this paper we emphasize an analysis that outputs exact formulas for the simpler design and performance assessment of the one-bit quantizer.

3. One-bit Quantizer Parameterization

At the very beginning of this section, we recall briefly the basics of the quantization theory. An one-level quantizer Q_2 is defined by mapping $Q_2 : \mathbb{R} \rightarrow Y$ [1], where $\mathbb{R}$ is a set of real numbers, $Y = \{y_1, y_2\} \subset \mathbb{R}$ is the code book of size 2 containing representation levels y_i, (i = 1, 2). With the one-bit quantizer Q_2, $\mathbb{R}$ is partitioned into 2 one-side unbounded in width quantization cells $\Re_i$ (i = 1, 2), where y_i specifies the i-th codeword and is the only representative for all real values x from $\Re_i$. Note that these representation levels are symmetrically placed around 0, since we address symmetric one-bit quantizer.

Let us calculate for arbitrary $\mu \geq 0$ and σ the total distortion, composed of D_- and D_+, where D_- and D_+ are distortions in negative and positive axis parts of the assumed pdf, or in quantization cells $\Re_1$ and $\Re_2$, respectively. Foremost, we will give an exact formula for arbitrary chosen represents y_i, (i = 1, 2), later adapted to our case with the assumed symmetry of representation levels.

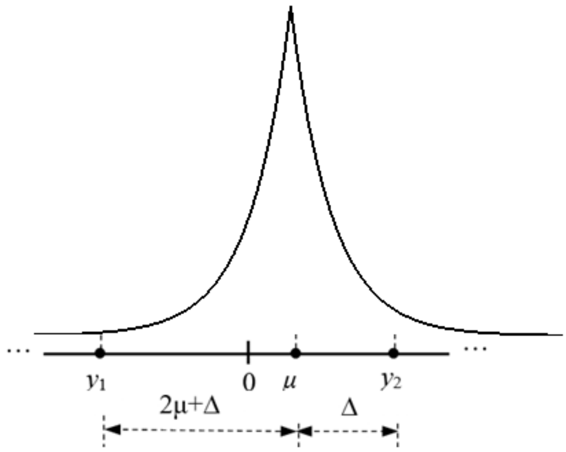

Fig. 1 Symmetrically placed representation levels for an arbitrary Laplacian pdf

$$D_- = \int_{-\infty}^{0} (x - y_1)^2 \, p(x|\mu,\sigma) \, dx, \; x < 0 \,, \tag{2}$$

$$D_+ = \int_{0}^{\mu} (x - y_2)^2 \, p(x|\mu,\sigma) \, dx + \int_{\mu}^{\infty} (x - y_2)^2 \, p(x|\mu,\sigma) \, dx, \; x \geq 0 \,, \tag{3}$$

Substituting (1) in (2) and (3) yields:

$$D_- = \frac{1}{\sqrt{2}\sigma} \int_{-\infty}^{0} (x - y_1)^2 \exp\left\{ -\frac{\sqrt{2}(x-\mu)}{\sigma} \right\} dx, \; x < 0 \,, \tag{4}$$

$$D_+ = \frac{1}{\sqrt{2}\sigma} \left[\int_{0}^{\mu} (x - y_2)^2 \exp\left\{ \frac{\sqrt{2}(x-\mu)}{\sigma} \right\} dx + \int_{\mu}^{\infty} (x - y_2)^2 \exp\left\{ \frac{-\sqrt{2}(x-\mu)}{\sigma} \right\} dx \right], \tag{5}$$

By further reorganization of formulas (4) and (5) we obtain:

$$D_- = \frac{1}{2} \exp\left\{ -\frac{\sqrt{2}\mu}{\sigma} \right\} (y_1^{\,2} + \sqrt{2}\sigma y_1 + \sigma^2) \,. \tag{6}$$

$$D_{+}=\frac{1}{2}\left[(y_2{}^2+\sigma^2)\left(2-\exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\}\right)-y_2\left(4\mu+\sqrt{2}\sigma\exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\}\right)+2\mu^2\right]. \quad (7)$$

Finally, we find out the total distortion as:

$$D=D_{-}+D_{+}=\mu^2+\sigma^2-2\mu y_2+y_2{}^2-\frac{1}{2}\exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\}[(y_2{}^2-y_1{}^2)+\sqrt{2}\sigma(y_2-y_1)]\,. \quad (8)$$

Further we define SQNR for the one-bit quantization case and the arbitrary unrestricted Laplacian pdf as:

$$\text{SQNR}=10\log_{10}\frac{\sigma^2}{D}\,. \quad (9)$$

$$\text{SQNR}=10\log_{10}\frac{\sigma^2}{\mu^2+\sigma^2-2\mu y_2+y_2{}^2-\frac{1}{2}\exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\}[(y_2{}^2-y_1{}^2)+\sqrt{2}\sigma(y_2-y_1)]}\,. \quad (10)$$

In the special case of zero mean and unit variance (μ = 0 and σ^2=1), or in normalized case, denoting distortion by D^{n} we calculate:

$$D^{\text{n}}=D_{-}+D_{+}\big|_{\mu=0,\sigma^2=1}=1+y_2{}^2-\frac{1}{2}[(y_2{}^2-y_1{}^2)+\sqrt{2}(y_2-y_1)]\,. \quad (11)$$

In our case with the symmetry of representation levels stands that y_2 = - y_1, distortion D^{s} for assumed symmetry condition becomes:

$$D^{\text{s}}=D_{-}+D_{+}\big|_{y_2=-y_1}=\mu^2+\sigma^2-2\mu y_2+y_2{}^2-\sqrt{2}\sigma y_2\exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\}. \quad (12)$$

Finally, we calculate the total distortion for the normalized case with symmetrically placed representation levels as:

$$D^{\text{ns}}=D_{-}+D_{+}\big|_{\mu=0,\sigma^2=1,y_2=-y_1}=1-\sqrt{2}y_2+y_2{}^2\,. \quad (13)$$

By minimizing the distortion, that is, by setting the first derivative of so obtained distortion D^{ns} with respect to y_2 equal to zero:

$$\frac{\partial D^{\text{ns}}}{\partial y_2}=2y_2-\sqrt{2}=0\,. \quad (14)$$

we can find $y_2{}^{ns}=\frac{\sqrt{2}}{2}$, while for D^{s} we derive the minimal distortion condition as:

$$\frac{\partial D^{\text{s}}}{\partial y_2}=-2\mu+2y_2-\sqrt{2}\sigma\exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\}=0\,. \quad (15)$$

Here the positive representation level y_2 is determined as:

$$y_2 = \mu + \frac{\sqrt{2}}{2}\sigma \exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\} = \mu + \Delta . \quad (16)$$

whereas Δ shows how far the representation level y_2 is from mean μ, while y_1 is distant from mean for $2\mu+\Delta$ (see Fig.1).

$$\Delta = \frac{\sqrt{2}}{2}\sigma \exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\} . \quad (17)$$

By further reorganization of (16) we have:

$$y_2 = \mu\left(1 + \frac{\sqrt{2}\sigma}{2\mu}\exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\}\right) = \mu\left(1 + \frac{\Delta}{\mu}\right). \quad (18)$$

Substituting (16) in (10) yields:

$$\mathrm{SQNR}^s = 10\log_{10}\frac{1}{1 - \frac{1}{2}\exp\left\{\frac{-2\sqrt{2}\mu}{\sigma}\right\} - \frac{\sqrt{2}\mu}{\sigma}\exp\left\{\frac{-\sqrt{2}\mu}{\sigma}\right\}} . \quad (19)$$

4. Numerical Results

In this section, we discuss achieved results as the outcomes of the parameterization for symmetrical one-bit quantizer. Let us remind that in previous section we have given an exact formula for SQNR calculation (19) when the symmetrical one-bit quantizer is used, while the input data are well described with an unrestricted arbitrary Laplacian pdf. Therefore, we find an interest to analyze how quantization of data having arbitrary Laplacian pdf affects data representation levels for the assumed symmetry of quantizer design.

Paying attention to the Eq. (19), one can note that SQNR^s for symmetrically placed representation levels depends on μ/σ (see Fig.2)

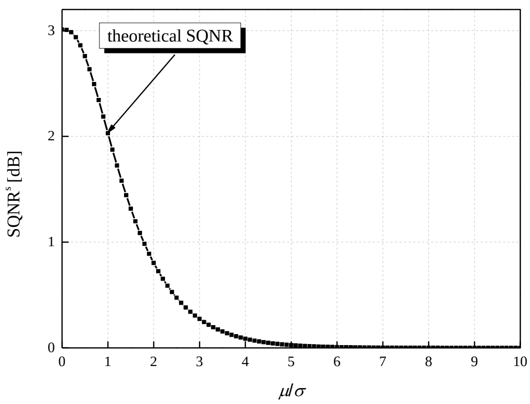

Fig. 2 SQNR^s dependence on μ/σ for symmetrical quantizer design

Table 1 $SQNR^s$ for some specific μ and σ.

μ	σ	SQNR [db]
0	arbitrary	3.0103
1	1	2.0299
1	2	2.7590
1	5	2.9856
2	1	0.8036
2	2	2.0299
2	5	2.8616
5	1	0.0262
5	2	0.4743
5	5	2.0299

Table 1 shows $SQNR^s$ for some specific μ and σ values. The highest $SQNR^s$ is achieved for zero mean and arbitrary variance, while $SQNR^s$ rapidly decreases with an increase of mean and for unit variance. If μ and σ are the same, $SQNR^s$ amounts 2.0299.

For zero mean pdfs, representation level $y_2 = \sigma / \sqrt{2}$ depends only on σ. In case of non-zero mean pdfs, $y_2 = \mu\ (1+ \Delta/\mu)$ and according to Δ/μ dependence on μ/σ (see Fig. 3) we can conclude that μ predominantly influences y_2, as well as its symmetrical pair y_1.

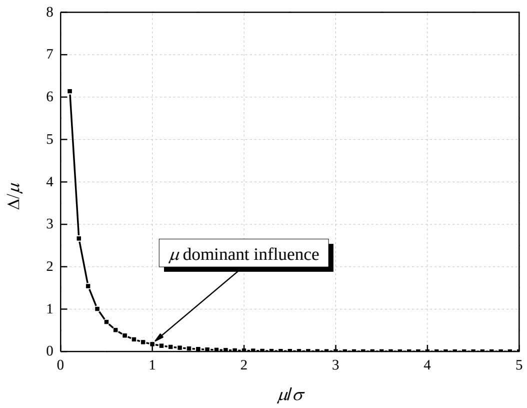

Fig. 3 Δ/μ dependence on μ/σ

For pre-trained QNN, customized for some specific task, further reused as the foundation for an another task, in case of aggressive one-bit quantization, QNN performance can be tremendously degraded. To alleviate this inevitable degradation, more precise mean and standard deviation assessment one can set as an indispensable prerequisite. In most of the performance analysis, zero mean and unit variance are accepted as the most common conditions of the ground-up quantizer design. Here we want to analyse an influence of two main parameters, μ and σ, that well describe Laplacian pdf, since their determination holds the key to improving the quantization's efficiency.

In doing so, we will analyse the Kullback-Leibler (KL) divergence for two arbitrary Laplacian pdfs [22], [23] (see Table 2 Case 1), while in general stands that:

$$\mathrm{KL}(\,p(x|\mu_1,\sigma_1)\|\,p(x|\mu_2,\sigma_2)) \neq \mathrm{KL}(\,p(x|\mu_2,\sigma_2)\|\,p(x|\mu_1,\sigma_1))\,. \tag{20}$$

Case 1 shows the measure or divergence of $p(x|\mu_1,\sigma_1)$ from $p(x|\mu_2,\sigma_2)$, while Eq. (20) underlines the inequality of reciprocal divergences of the same pair of functions $\{p(x|\mu_1,\sigma_1), p(x|\mu_2,\sigma_2)\}$. For two identical pdfs the KL divergence is zero, while for a large deviation of μ_1 from μ_2 and σ_1 from σ_2, KL divergence is large.

Kullback-Leibler divergence, as a similarity measure, is also used to specify extra bits required to describe a pair of distributions. If pdfs differ significantly, there is a need to provide additional bits that is not desirable in low-bit presentations.

If one of the pdfs implies a normalized case of zero mean and unit variance, we can find a KL divergence in cases where this normalization characterizes the first or the second pdf (Case 2 and Case 3, respectively). Note that the KL divergence given in Case 5, for a specific case of $\sigma_1 = \sqrt{2}$ does not depend on μ_1. In contrast, applying similar condition $\sigma_2 = \sqrt{2}$ in Case 4, one can notice that the KL divergence indeed depends on μ_2, that additionally confirms an inequality statement given by (20).

Table 2 KL divergence for some specific cases.

Case	μ_1	σ_1	μ_2	σ_2	KL divergence
1	μ_1	σ_1	μ_2	σ_2	$\dfrac{\sigma_1 \exp\left\{\dfrac{-\sqrt{2}\|\mu_1-\mu_2\|}{\sigma_1} + \|\mu_1-\mu_2\|\right\}}{\sigma_2} + \log\dfrac{\sigma_2}{\sigma_1} - 1$
2	0	1	μ_2	σ_2	$\dfrac{\exp\{(1-\sqrt{2})\|\mu_2\|\}}{\sigma_2} + \log\sigma_2 - 1$
3	μ_1	σ_1	0	1	$\sigma_1 \exp\left\{\dfrac{-\sqrt{2}\|\mu_1\|}{\sigma_1} + \|\mu_1\|\right\} - \log\sigma_1 - 1$
4	0	1	μ_2	$\sqrt{2}$	$\dfrac{1}{\sqrt{2}}\exp\{(1-\sqrt{2})\|\mu_2\|\} + \log\sqrt{2} - 1$
5	μ_1	$\sqrt{2}$	0	1	$\sqrt{2} - \log(\sqrt{2}) - 1$
6	μ_1	σ	μ_2	σ	$\exp\left\{\dfrac{(\sigma-\sqrt{2})\|\mu_1-\mu_2\|}{\sigma}\right\} - 1$
7	μ	σ_1	μ	σ_2	$\dfrac{\sigma_1}{\sigma_2} + \log\dfrac{\sigma_2}{\sigma_1} - 1$

If both pdfs have the same $\sigma_1 = \sigma_2 = \sigma$ (Case 6) the KL divergence depends on σ and difference of means, while for $\mu_1 = \mu_2 = \mu$, the KL divergence depends only on σ_1 / σ_2 (Case 7).

Since pre-trained model can be accommodated to the normalized case, by calculating the KL divergence similar as in paper [10], one can estimate μ, σ or μ/σ and specify symmetrical representation levels of one-bit quantization model according to the suggested formula (16).

As the total bit rate R in signal processing area according to formula $R = r_{fixed} + r_{add} /M$ [24], shows a strong dependence on r_{add} and M, where r_{add} – is the bit rate for additional information (here for μ and σ) and M – denotes the frame size, we should also pay attention to the analysis of R for the suggested one-bit design. (r_{fixed} = 1, see Table 3). Our NN model architecture specified in [15,16] consists of three FC (dense) layers. MNIST training and testing datasets are loaded and then flattened into one-dimensional vectors of 784 (28*28) elements. First two FC layers consist of 512 nodes, where the first layer accepts an input shape of (784,), so the overall number of bits for that level can be calculated as M = 784x512, and the total bit rate is R = 1+32/784x512 = 1,00008 ≈ 1. We can notice that neural networks are less demanding in terms of additional bits for side information transmitting, as we concluded - the larger the NN model, the closer the bit rate is to r_{fixed} or to R = 1, in one-bit quantizer case.

For the most common used presentations with M = 32 and M^* = 16 bits, required total bit-rates are shown in Table 3. As for low-bit presentations, it is desirable that additional information is displayed with as few bits as possible, so that $M \geq 120$ frame sizes should be preferred.

Table 3 Total bit rate R for one-bit quantizer in signal processing area.

r_{fixed}	r_{add}	M	R	r_{add}^*	M^*	R^*
1	32	20	2.6	16	20	1.8
1	32	40	1.8	16	40	1.4
1	32	120	1.267	16	120	1.133
1	32	240	1.133	16	240	1.067

5. Conclusion

In this paper, an importance of determining of the exact formula for one-bit quantizer's total distortion is highlighted in the case when the data are well described by an arbitrary Laplacian pdf. Aiming for the simplicity of quantization design, symmetrical quantizer is proposed, as well as the symmetry of representational levels. Note that precise determination of the mean and variance is a prerequisite to achieve the high quantization model's accuracy, since quantization accuracy can be tremendously degraded in the case when the mean and variance are not well adjusted. We have shown that the mean has a predominantly important influence on the representation level determination in the non-normalized case, since one of the representation levels is close to the mean. One-bit quantizers are especially convenient in highly proliferated resource-constrained devices thus increasing the need for the simple but accurate low-bit quantizer design.

Acknowledgement: *This research has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia and by the Science Fund of the Republic of Serbia, 6527104, AI-Com-in-AI.*

REFERENCES

[1] N. S. Jayant, P. Noll, *Digital Coding of Waveforms*, Englewood Cliffs, NJ: Prentice-Hall, 1984.
[2] D. Hui, D. L. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization," *IEEE Transaction on Information Theory,* vol. 47, no. 3, pp. 957–977, 2001, doi: 10.1109/18.915652.
[3] S. Na, D. Neuhoff, "On the convexity of the MSE distortion of symmetric uniform scalar quantization," *IEEE Transaction on Information Theory*, vol. 64, no.4, pp. 2626−2638, 2018, doi: 10.1109/TIT.2017.2775615.
[4] S. Na, D. Neuhoff, "Monotonicity of step sizes of MSE-optimal symmetric uniform scalar quantizers," *IEEE Transaction on Information Theory*, vol. 65, no. 3, pp. 1782−1792, 2019, doi: 10.1109/TIT.2018.2867182.
[5] J. Lee, S. Na, "A rigorous revisit to the partial distortion theorem in the case of a Laplacian source," *IEEE Commun. Lett*., vol. 21, pp. 2554–2557, 2017, doi: 10.1109/LCOMM.2017.2749218.
[6] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, K. Keutzer, "A survey of quantization methods for efficient neural network inference," arXiv:2103.13630
[7] Y. Guo, "A survey on methods and theories of quantized neural networks," arXiv:1808.04752
[8] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, Y. Bengio, "Quantized neural networks: training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, vol. 18, pp. 6869–6898, 2017.
[9] D. Liu, H. Kong, X. Luo, W. Liu, R. Subramaniam, "Bringing AI to edge: from deep learning's perspective," arXiv:2011.14808.
[10] S. Kim, A. Gholami, Z. Yao, N. Lee, P. Wang, A. Nrusimha, B. Zhai, T. Gao, M. W. Mahoney, K. Keutzer, "Integer-only zero-shot quantization for efficient speech recognition," arXiv:2103.16827v2
[11] R. Banner, Y. Nahshan and D. Soudry, "Postraining 4-bit quantization of convolutional networks for rapid-deployment," *33rd Conference on Neural Information Processing Systems* (*NeurIPS 2019*), *Vancouver*, *Canada*, pp. 7948-7956, 2019.
[12] S. Gazor and W. Zhang, "Speech probability distribution," IEEE Signal Processing Letters, vol. 10, no. 7, pp. 204-207, 2003, doi: 10.1109/LSP.2003.813679.
[13] S. Kotz, T. Kozubowski and K. Podgorski, "The Laplace distribution and generalizations," *Birkhäuser*, Boston, 2001.
[14] J. Nikolić, D. Aleksić, Z. Perić and M. Dinčić, "Iterative algorithm for parameterization of two-region piecewise uniform quantizer for the Laplacian source," *Mathematics*, vol. 9, pp. 3091, 2021, doi: 10.3390/math9233091.
[15] Stefan Tomić, Jelena Nikolić, Zoran Perić, Danijela Aleksić, "Performance of Post-Training Two-Bits Uniform and Layer-Wise Uniform Quantization for MNIST Dataset from the Perspective of Support Region Choice,",*Mathematical Problems in Engineering*, Volume 2022, Article ID 1463094, 2022.J.
[16] Nikolić, Z. Perić, D. Aleksić, S. Tomić, A. Jovanović, "Whether the support region of three-bit uniform quantizer has a strong impact on post-training quantization for MNIST dataset?," *Entropy*, vol. 23, pp. 1699, 2021, doi: 10.3390/e23121699.
[17] W. Zhao, M. Teli, X. Gong, B. Zhang, D. Doermann, "A review of recent advances of binary neural networks for edge computing," *IEEE J. Miniat. Air Space Syst.*, vol. 2, pp. 25–35. 2021, doi: 10.1109/JMASS.2020.3034205.
[18] P. E. Novac, G. B. Hacene, A. Pegatoquet, B. Miramond, V. Gripon, "Quantization and deployment of deep neural networks on microcontrollers," *Sensors*, vol. 21, pp. 2984, 2021, doi: 10.3390/s21092984.
[19] D. Aleksić, Z. Perić, "Analysis and design of robust quasilogarithmic quantizer for the purpose of traffic optimisation," *Inf. Technol. Control*, vol. 47, pp. 615-622, 2018, doi: http://dx.doi.org/10.5755/j01.itc.47.4.20668.
[20] Z. Perić, J. Nikolić, D. Aleksić, A. Perić, "Symmetric quantile quantizer parameterization for the Laplacian source: qualification for contemporary quantization solutions, " *Math. Probl. Eng.*, vol. 2021, Article ID 6647135, 2021, doi: 10.1155/2021/6647135
[21] Z. Perić, D. Aleksić, "Quasilogarithmic quantizer for Laplacian source: support region ubiquitous optimization task," *Rev. Roum. Sci. Techn.*, vol. 64, pp. 403−408, 2019.
[22] G. Dehaune, "A deterministic and computable Bernstein-von Mises theorem," arXiv:1904.02505.
[23] S. Kullback, R. Leibler, "On information and sufficiency. Annals of mathematical statistics," vol. 22, pp. 79–86, 1951.
[24] S. Tomić, Z. Perić, J. Nikolić, "Modified BTC algorithm for audio signal coding," *Advances in Electrical and Computer Engineering*, vol. 16, 2016.