


ITC 4/51 Information Technology and Control Vol. 51 / No. 4 / 2022 pp. 625-637 DOI 10.5755/j01.itc.51.4.30473	Performance Analysis of a 2-bit Dual-Mode Uniform Scalar Quantizer for Laplacian Source	
	Received 2022/01/09	Accepted after revision 2022/09/30
	 http://dx.doi.org/10.5755/j01.itc.51.4.30473	

HOW TO CITE: Perić, Z. H., Denić, B. D., Jovanović, A. Z., Milosavljević, S., Savić, M. S. (2022). Performance Analysis of a 2-bit Dual-Mode Uniform Scalar Quantizer for Laplacian Source. *Information Technology and Control*, 51(4), 625-637. <http://dx.doi.org/10.5755/j01.itc.51.4.30473>

Performance Analysis of a 2-bit Dual-Mode Uniform Scalar Quantizer for Laplacian Source

Zoran H. Perić, Bojan D. Denić, Aleksandra Z. Jovanović

University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia;
phone: +38118 529 367; e-mails: zoran.peric@elfak.ni.ac.rs, bojan.denic@elfak.ni.ac.rs,
aleksandra.jovanovic@elfak.ni.ac.rs

Srdjan Milosavljević

University of Priština–Kosovska Mitrovica, Faculty of Economics,
Kolasinska 156, 38220 Kosovska Mitrovica, Serbia; e-mail: srdjan.milosavljevic@pr.ac.rs

Milan S. Savić

University of Priština–Kosovska Mitrovica, Faculty of Sciences,
Ive Lole Ribara 29, 38220 Kosovska Mitrovica, Serbia; e-mail: milan.savic1@pr.ac.rs

Corresponding author: bojan.denic@elfak.ni.ac.rs

The main issue when dealing with the non-adaptive scalar quantizers is their sensitivity to variance-mismatch, the effect that occurs when the data variance differs from the one used for the quantizer design. In this paper, we consider the influence of that effect in low-rate (2-bit) uniform scalar quantization (USQ) of Laplacian source and also we propose adequate measure to suppress it. Particularly, the approach we propose represents the upgraded version of the previous approaches used to improve performance of the single quantizer. It is based on dual-mode quantization that combines two 2-bit USQs (with adequately chosen parameters) to process input data, selected by applying the special rule. Analysis conducted in theoretical domain has shown that the proposed approach is less sensitive to variance-mismatch, making the dual-mode USQ more efficient in terms of robustness than the single USQ. Also, a gain is achieved compared to other 2-bit quantizer solutions. Experimental results are also provided for quantization of weights of the multi-layer perceptron (MLP) neural network, where good matching with the theoretical results is observed. Due to these achievements, we believe that the solution we propose can be a good choice for compression of non-stationary data modeled by Laplacian distribution, such as neural network parameters.

KEYWORDS: Scalar quantization, Laplacian source, neural network compression, SQNR.

1. Introduction

Uniform scalar quantization (USQ) is extensively used in multiple data processing applications. The main attribute of USQ is the simplicity accompanied with the performance competitive to a more complex non-uniform quantization [10]. An extensive amount of research is dedicated to this topic, where different aspects were considered during the analysis and valuable conclusions were derived [9, 15–17]. That indeed helped to promote the USQ as an excellent candidate for various practical resource-constrained applications, such as speech coding [18] or neural network (NN) compression [1–3, 23].

Scalar quantizers are mostly designed for certain data source, whereby the Laplacian source is a frequently used one. This is because the Laplacian probability density function statistically models various real data such as speech [7, 10], image [10] or weights of a neural network [2]. Therefore, development of quantizers for the Laplacian source is significant from the practical point of view. In addition, quantizers are often designed for one variance value (usually the unit one) and applied to data with variance different from the designed one, which leads to mismatch in variance which further may cause a serious performance degradation [10, 13, 14]. Therefore, robust quantizers are preferred in a variance-mismatch scenario, that is, the ones that are able to suppress mismatch effect as high as possible and accordingly to provide satisfactory performance in the desired range of variances [10].

A typical practical example of the occurrence of variance-mismatch can be found in NN compression, as the statistical properties of NN parameters (e.g. weights) can change. Namely, NN compression is an active research area where scalar quantization (especially uniform scalar quantization) plays an important role. The application of quantization technique assumes using a lower bit-length representation for NN parameters (weights, activations, etc.) instead of commonly used 32-bit floating point format (full precision). In this way, quantized (compressed) NN is obtained. On the other hand, quantized NN should offer competitive performance with respect to full precision NN and should also be as simple as possible, which further enables easier implementation on the devices with limited power or memory, e.g. edge devices. Hence, it is of special importance to develop

an efficient quantization scheme, because it can significantly contribute to the performance of quantized NN. Most of the available quantization schemes are based on fixed-length coding, where different code-words length are employed including > 8 bits [23], 8 bits [1], 4 bits [3], 2 bits [4, 5, 20, 21, 26] or even 1 bit [8, 22, 27, 28]. It was reported that quantized NN provides a negligible decreasing of performance with respect to full precision NN when high bit lengths (≥ 8 bits) have been employed, where compression ratio ≤ 4 times was achieved. On the other hand, much higher compression ratios have been observed in the case of lower bit lengths (up to 16 times in case of 2 bits) at the cost of a certain performance degradation of quantized NN. Based on the previously achieved results, it is clear that low-rate (low-level) quantizers are very important in NN compression. However, most existing low-rate solutions are not properly designed in terms of NN parameter statistics (e.g. statistic of weights), which may be one of the main reasons for the degradation of quantized NN performance. Therefore, during quantizer design, it is preferred to take into account the actual statistical distribution of NN parameters, which is done in this paper.

This paper proposes a non-adaptive 2-bit ($N = 4$ levels) USQ intended for NN compression. Specifically, we investigate the variance-mismatch effect in case of 2-bit USQ and propose an adequate measure to deal with it. Note also that the practical significance of 2-bit quantizers is large, as they have already been successively applied in several data processing applications such as speech coding [18, 19], image coding [11, 29] or, as mentioned, NN compression [4, 5, 20, 21, 26]. Regarding the implementation in NN compression, it was shown in [4, 5, 26] that 2-bit USQ can enable a good trade-off between performance and size of NN. Namely, USQ models have been used in papers [4, 5, 26], which are designed using different approaches and further adapted to the statistics of input data (usually weights), enabling a high level of robustness. Recall that adaptation requires the estimation of some additional parameters (e.g. mean and variance) from the input data, which are used to fit the quantizer codebook. Robust non-adaptive quantizers are often used as alternative to adaptive quantizers, where aforementioned additional steps can be avoided at the expense of some performance degra-

dation. Note that such solutions may be adequate for NN compression, since NNs are more robust to change in data quality than e.g. speech, where adaptive solutions are more preferable, as shown in [18, 19]. A 2-bit quantizer model for NN weights, which is non-adaptive and non-uniform one (based on logarithmic compression function) with increased robustness property was proposed in [20].

The quantization approach presented here, used to increase the robustness (that is, to reduce the sensitivity to variance-mismatch) of the non-adaptive 2-bit USQ, actually upgrades the approach from [20] introduced to enhance performance of the single quantizer. Namely, the approach used in [20] is also related to Laplacian source and is based on scaling the quantizer key parameter, which resulted in increased performance in the variance range of interest. In this paper, we exploit the benefit offered by scaling, but in contrast to [20] we use two quantizers (the one with the scaled parameters and the one with the initial parameter settings) to cover the entire variance range. Since two 2-bit USQs (with different parameter settings) are at disposal, the resulting quantizer is named dual-mode USQ. The introduced dual-mode USQ divides input data into segments called blocks (block-by-block processing), and then to quantize the current block it selects one of two USQs depending on the estimated statistical characteristic of the block. In that case, 1 extra bit is needed per each block to identify USQ used for that block, slightly increasing the bit rate with respect to the single USQ. On the other hand, data are more accurately quantized using block-by-block data processing. The usefulness of mentioned data processing logic has been indicated in [6], where the gain in performance of quantized NN, obtained with block-by-block data quantization, was reported. Note that dual-mode scalar quantizers have already been proposed for Laplacian source, but for high bit rates and non-uniform quantizers [24, 25]. Namely, these solutions combined two adaptive quantizers with unequal support regions and equal number of quantization levels (restricted and unrestricted ones) for short data blocks (preferred for speech), favoring a more frequent utilization of the restricted quantizer to outperform single unrestricted quantizer. Instead of searching the content of a block as in [24, 25], in this paper's selection among the two USQs is done based on the estimated block variance. The idea originates from [21], but few progressive steps are performed in this paper. We change a heuristically de-

termined step size value (key parameter) of the initial 2-bit USQ by the one which is obtained by maximization of performance for the particular variance, further we determine the best value for scaling factor (used to scale the initial step size) and also we propose an iterative rule for determining the best switching threshold of the dual-mode USQ.

Let us emphasize that the proposed 2-bit dual-mode USQ uses the same initial 2-bit USQ as the adaptive model in [26], whereby the adaptive model from [26] is equivalent to a switching scheme using 2^{32} different 2-bit USQs (as the data variance is quantized using 32 bits). Due to utilization of smaller number of USQs (2 vs. 2^{32} USQs), our proposal is expected to offer quality of quantized data being between that of single non-adaptive and adaptive 2-bit USQ. However, our proposal requires fewer extra bits per block compared to [26] (only 1 vs. 32 bits), making it more efficient in terms of overall bit rate. For the 2-bit adaptive quantizers in [4, 5] the same complexity and bit rate requirements can be observed as for the one in [26]. In this paper the quantizers from [4, 5] are used as the baselines for performance comparison. However, references [4, 5] did not provide the quantizer-related analysis from the viewpoint of signal processing that should reveal their actual performance, which is very important. From that reason, here, we calculate their performances in the case of Laplacian PDF.

In brief, this paper contributes with the following:

- We propose a scalar quantization solution based on dual-mode USQ, which combines two non-adaptive 2-bit USQs with adequately selected parameters. It upgrades existing approaches used to improve the performance of single non-adaptive quantizer.
- We design quantizer for NN weights compression by taking into consideration statistics of NN weights, in contrast to the available quantizer solutions that are suboptimal in that context.
- This type of quantizer, to the best of the author's knowledge, has not been proposed so far by other researchers for NN compression.
- The quantizer model we propose can be viewed as a compromise solution achieved between a non-adaptive 2-bit USQ whose robustness is not at the satisfactory level and a highly robust but more complex adaptive 2-bit USQ [26].

- We verify theoretical results with experimental ones, obtained by quantizing the weights of trained MLP NN intended for image classification.

2. Theoretical Model of Non-adaptive 2-bit USQ

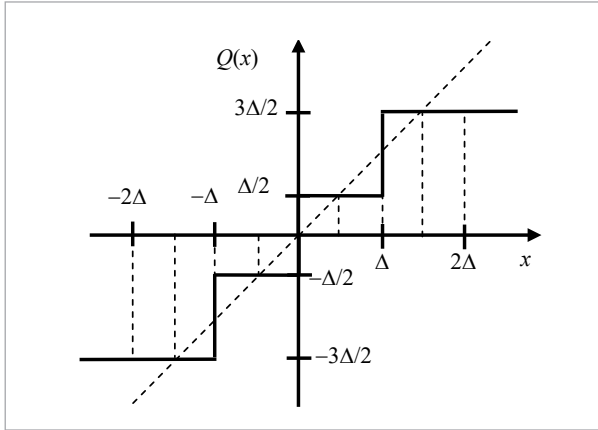
2.1. Description and Design of 2-bit USQ

The input-output characteristic of 2-bit USQ ($N = 4$ quantization levels) is depicted in Figure 1, where Δ denotes the step size. Actually, Δ is a key parameter of USQ, whereas the interval $[-2\Delta, 2\Delta]$ defines the support region.

If data to be quantized have probability density function (PDF) with infinite support, then, the quantiza-

Figure 1

Input-output characteristic of the 2-bit USQ



tion cells are defined in $(-\infty, -\Delta]$, $(-\Delta, 0]$, $(0, \Delta]$ and (Δ, ∞) , while quantization levels are uniformly distributed within the support region $Q(x) \in \{\pm\Delta/2, \pm3\Delta/2\}$. Input data value x is quantized by 2-bit USQ according to:

$$Q(x) = \begin{cases} \left(\left\lfloor \frac{|x|}{\Delta} \right\rfloor + \frac{1}{2} \right) \Delta \operatorname{sgn}(x), & |x| < 2\Delta \\ \frac{3\Delta}{2} \operatorname{sgn}(x), & |x| \geq 2\Delta \end{cases}, \quad (1)$$

where $\lfloor \cdot \rfloor$ denotes rounding to the nearest integer lower than x , while $\operatorname{sgn}(\cdot)$ is a sign function. In this

paper, we assume that x is generated by the memoryless Laplacian source of zero-mean and variance σ^2 , described by PDF [10]:

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

We design 2-bit USQ for Laplacian PDF with variance $\sigma^2 = \sigma_{\text{ref}}^2 = 1$ ($p(x, \sigma=1) \equiv p(x)$). Actually, we determine the key parameter $\Delta(\sigma_{\text{ref}}) = \sigma_{\text{ref}} \cdot \Delta = \Delta$ such that minimal mean-squared error (MSE) distortion or equivalently maximal signal to quantization noise ratio (SQNR) is achieved. MSE distortion of the 2-bit USQ is given by [21]:

$$D = 2 \left(\int_0^{\Delta} \left(x - \frac{\Delta}{2}\right)^2 p(x) dx + \int_{\Delta}^{+\infty} \left(x - \frac{3\Delta}{2}\right)^2 p(x) dx \right) = 1 + \frac{\Delta^2}{4} - \sqrt{2}\Delta \left(\frac{1}{2} + \exp\{-\sqrt{2}\Delta\} \right), \quad (3)$$

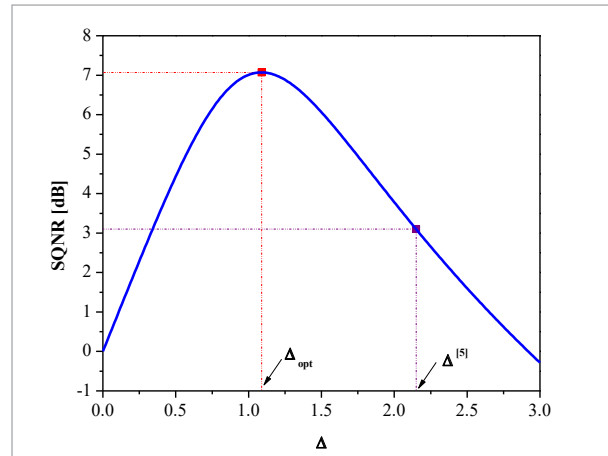
while SQNR is defined by:

$$\text{SQNR} = 10 \log_{10} \left(\frac{1}{D} \right) = -10 \log_{10} \left(1 + \frac{\Delta^2}{4} - \sqrt{2}\Delta \left(\frac{1}{2} + \exp\{-\sqrt{2}\Delta\} \right) \right). \quad (4)$$

In Figure 2, we depict SQNR as a function of Δ . It can be observed that $\Delta = \Delta_{\text{opt}} = 1.087$ maximizes SQNR ($\text{SQNR}(\Delta = \Delta_{\text{opt}}) = 7.07$ dB) [26]. For comparison pur-

Figure 2

SQNR vs. Δ of the 2-bit single USQ ($\sigma^2 = 1$)



poses, in Figure 2 we also indicate the step size (calculated for variance σ_{ref}^2) of 2-bit USQ developed in [5], $\Delta^{[5]}(\sigma_{\text{ref}}) = 2 \cdot c_1 \cdot \sigma_{\text{ref}} / 3$ (it is obtained as the result of setting the last quantization level to $c_1 \cdot \sigma_{\text{ref}}$ by assuming zero-mean of input data), where $c_1 = 3.2$. Given figure clearly demonstrates that step size value determined according to designing method from [5] is sub-optimal choice, as the loss in SQNR of about 4 dB is observed.

2.2. A 2-bit USQ in Variance-mismatch Scenario

Let us now consider the performance of the designed 2-bit USQ under variance-mismatch condition, that is, under condition that 2-bit USQ quantizes the Laplacian data having variance different from the designed one ($\sigma^2 \neq \sigma_{\text{ref}}^2 = 1$). Using PDF defined with (2), for distortion is obtained [21, 26]:

$$D(\sigma) = 2 \left(\int_0^{\Delta} \left(x - \frac{\Delta}{2}\right)^2 p(x, \sigma) dx + \int_{\Delta}^{+\infty} \left(x - \frac{3\Delta}{2}\right)^2 p(x, \sigma) dx \right) = \sigma^2 \left(1 + \frac{\Delta^2}{4\sigma^2} - \frac{\sqrt{2}\Delta}{\sigma} \left(\frac{1}{2} + \exp\left\{-\frac{\sqrt{2}\Delta}{\sigma}\right\} \right) \right) \quad (5)$$

and SQNR is given by:

$$\text{SQNR}(\sigma) = 10 \log_{10} \left(\frac{\sigma^2}{D(\sigma)} \right) = -10 \log_{10} \left(1 + \frac{\Delta^2}{4\sigma^2} - \frac{\sqrt{2}\Delta}{\sigma} \left(\frac{1}{2} + \exp\left\{-\frac{\sqrt{2}\Delta}{\sigma}\right\} \right) \right) \quad (6)$$

Since the variance σ^2 can change in a wide range around the reference variance ($\sigma_{\text{ref}}^2 = 1$) it is usual to express σ in logarithmic domain as $\sigma_{\text{dB}} = 20 \log_{10}(\sigma / \sigma_{\text{ref}})$, leading to:

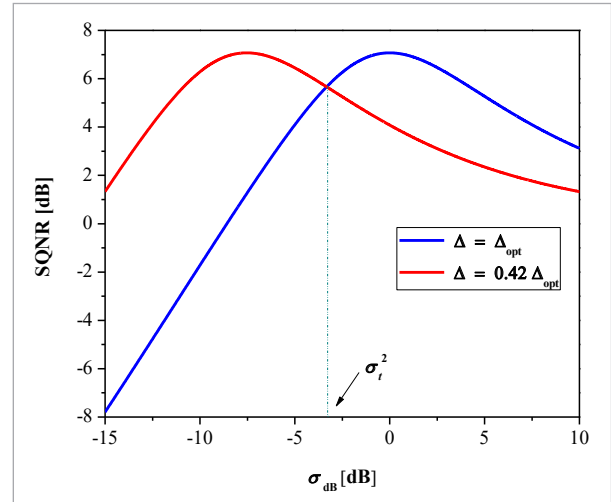
$$\sigma = 10^{\sigma_{\text{dB}}/20}. \quad (7)$$

Then, SQNR can be expressed as:

$$\text{SQNR}(\sigma_{\text{dB}}) = -10 \log_{10} \left(1 + \frac{1}{4} \frac{\Delta^2}{10^{\sigma_{\text{dB}}/10}} - \sqrt{2} \frac{\Delta}{10^{\sigma_{\text{dB}}/20}} \times \left(\frac{1}{2} + \exp\left\{-\sqrt{2} \frac{\Delta}{10^{\sigma_{\text{dB}}/20}}\right\} \right) \right). \quad (8)$$

Figure 3

SQNR of the 2-bit single USQ for $\Delta = \Delta_{\text{opt}}$ and $\Delta = c \cdot \Delta_{\text{opt}}$ ($c = 0.42$) in a wide dynamic range of the input data variances



In Figure 3, we plot the SQNR (Equation (8)) versus σ_{dB} for 2-bit USQ ($\Delta = \Delta_{\text{opt}}$), where we assume the range $[-15, 10]$ around the reference variance. It is obvious that a variance-mismatch significantly affects the performance, as SQNR is degraded. Namely, SQNR rapidly decreases when moving further away from the designed (i.e. reference) variance. This indicates that the robustness enhancement is required.

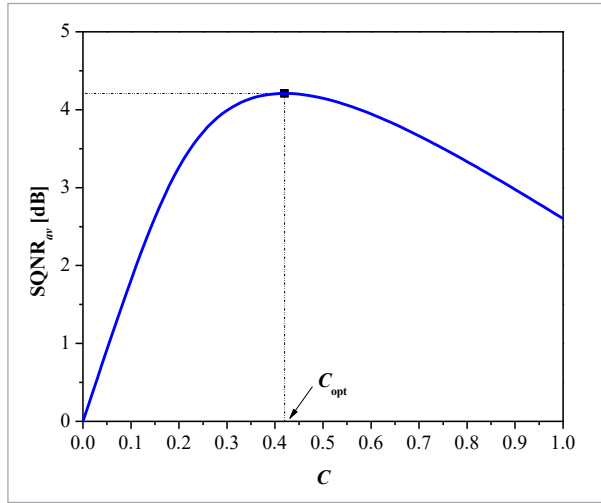
One efficient way to improve the performance of single USQ in the established variance range is to use the approach proposed in [20], which has already been successfully applied in case of non-uniform quantization. Particularly, it assumes scaling of the initial step size, $\Delta = c \cdot \Delta_{\text{opt}}$, where c is a real constant determined such that average SQNR (SQNR_{av}) in a given range of width 25 dB is maximal, where:

$$\text{SQNR}_{\text{av}} = \frac{1}{m} \sum_{i=1}^m \text{SQNR}(\sigma_i), \quad (9)$$

and where $m = 2500$ denotes the number of variances taken into account.

Figure 4 shows SQNR_{av} in dependence on c , where we can see that $c = c_{\text{opt}} = 0.42$ gives a maximum.

The SQNR curve in case of scaled step size value, $\Delta = c_{\text{opt}} \cdot \Delta_{\text{opt}}$ can be found in Figure 3, showing evident improvement compared to initial case, as negative SQNR values are avoided. In addition, observe that

Figure 4Dependence of SQNR_{av} on constant c 

scaling causes a shifting of the SQNR curve obtained for initial step size ($\Delta = \Delta_{\text{opt}}$). The amount of shifting can exactly be defined using the following lemma.

Lemma 1. The SQNR curves of the 2-bit USQ obtained for step sizes Δ and $c\Delta$ are shifted for c_{dB} , i.e. it holds:

$$\text{SQNR}(\sigma_{\text{dB}}, c \cdot \Delta_{\text{opt}}) = \text{SQNR}(\sigma_{\text{dB}} - c_{\text{dB}}, \Delta_{\text{opt}}). \quad (10)$$

Proof of Lemma 1. Based on (8), we have that:

$$\begin{aligned} \text{SQNR}(\sigma_{\text{dB}}, c \cdot \Delta_{\text{opt}}) &= -10 \log_{10} \left(1 + \frac{1}{4} \frac{(c \cdot \Delta_{\text{opt}})^2}{10^{\sigma_{\text{dB}}/10}} \right. \\ &\quad \left. - \sqrt{2} \frac{c \cdot \Delta_{\text{opt}}}{10^{\sigma_{\text{dB}}/20}} \left(\frac{1}{2} + \exp \left\{ -\sqrt{2} \frac{c \cdot \Delta_{\text{opt}}}{10^{\sigma_{\text{dB}}/20}} \right\} \right) \right) \end{aligned} \quad (11)$$

Let us express c in the logarithmic domain as $c_{\text{dB}} = 20 \log_{10} c$. Using the following relations:

$$\frac{c}{10^{\sigma_{\text{dB}}/20}} = \frac{10^{c_{\text{dB}}/20}}{10^{\sigma_{\text{dB}}/20}} = \frac{1}{10^{(\sigma_{\text{dB}} - c_{\text{dB}})/20}}, \quad (12)$$

$$\frac{c^2}{10^{\sigma_{\text{dB}}/10}} = \frac{10^{c_{\text{dB}}/10}}{10^{\sigma_{\text{dB}}/10}} = \frac{1}{10^{(\sigma_{\text{dB}} - c_{\text{dB}})/10}}. \quad (13)$$

and substituting them into (11) results in:

$$\begin{aligned} \text{SQNR}(\sigma_{\text{dB}}, c \cdot \Delta_{\text{opt}}) &= -10 \log_{10} \left(1 + \frac{1}{4} \frac{\Delta_{\text{opt}}^2}{10^{(\sigma_{\text{dB}} - c_{\text{dB}})/10}} - \sqrt{2} \frac{\Delta_{\text{opt}}}{10^{(\sigma_{\text{dB}} - c_{\text{dB}})/20}} \right. \\ &\quad \left. \times \left(\frac{1}{2} - \exp \left\{ -\sqrt{2} \frac{\Delta_{\text{opt}}}{10^{(\sigma_{\text{dB}} - c_{\text{dB}})/20}} \right\} \right) \right) = \text{SQNR}(\sigma_{\text{dB}} - c_{\text{dB}}, \Delta_{\text{opt}}) \end{aligned} \quad (14)$$

which proves the lemma

Alternatively, the shifting can be determined as follows. First, it is necessary to find, in case of both Δ and $c\Delta$, the variance in which the SQNR has maximum. Then, the shifting in log domain can simply be found as the difference between these two variances. The following lemma specifies these variances (in the linear domain).

Lemma 2. The variance ($\sigma^2 = \sigma_d^2$) in which the 2-bit USQ defined by Δ reaches a maximum of the SQNR can be determined according to the following iterative rule:

$$\sigma_d^{(i)} = \frac{1}{\sqrt{2}} \left(\Delta - 2 \left(\sqrt{2} \sigma_d^{(i-1)} - 2\Delta \right) \exp \left\{ -\frac{\sqrt{2}\Delta}{\sigma_d^{(i-1)}} \right\} \right). \quad (15)$$

Proof of Lemma 2. Let us introduce the function S :

$$S = \frac{\sigma^2}{D(\sigma)} = 1 + \frac{\Delta^2}{4\sigma^2} - \frac{\sqrt{2}\Delta}{\sigma} \left(\frac{1}{2} - \exp \left\{ -\frac{\sqrt{2}\Delta}{\sigma} \right\} \right). \quad (16)$$

By differentiating S with respect to σ we obtain:

$$\frac{\partial S}{\partial \sigma} = \frac{\Delta - \sqrt{2}\sigma - 2 \left(\sqrt{2}\sigma - 2\Delta \right) \exp \left\{ -\frac{\sqrt{2}\Delta}{\sigma} \right\}}{\left(4\sigma^2 + \Delta^2 - 4\sqrt{2}\sigma\Delta \left(\frac{1}{2} - \exp \left\{ -\frac{\sqrt{2}\Delta}{\sigma} \right\} \right) \right)^2}. \quad (17)$$

From the condition $\frac{\partial S}{\partial \sigma} \Big|_{\sigma=\sigma_d} = 0$, we arrive to the following:

$$\frac{\Delta - \sqrt{2}\sigma_d}{2 \left(\sqrt{2}\sigma_d - 2\Delta \right)} = \exp \left\{ -\frac{\sqrt{2}\Delta}{\sigma_d} \right\}. \quad (18)$$

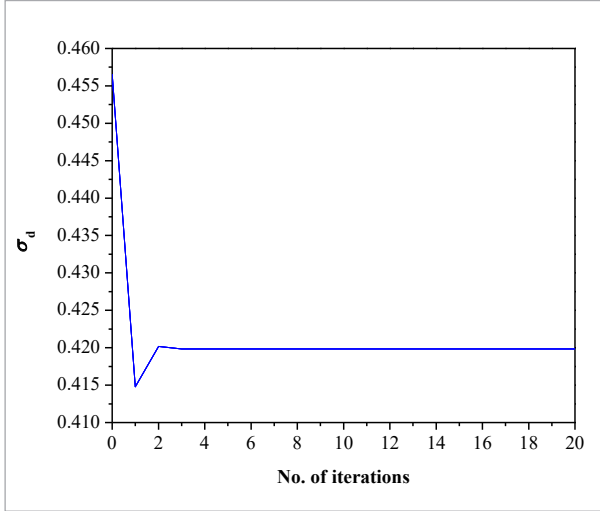
Based on (18), we can express σ_d as:

$$\sigma_d = \frac{1}{\sqrt{2}} \left(\Delta - 2 \left(\sqrt{2}\sigma_d - 2\Delta \right) \exp \left\{ -\frac{\sqrt{2}\Delta}{\sigma_d} \right\} \right). \quad (19)$$

The equation (19) can be solved iteratively, thus concluding the proof.

Figure 5

The iterative process for determination of σ_d defined by (15) in case when $\Delta = 0.42 \cdot \Delta_{opt}$



In order to show the efficiency as well as the correctness of the iterative process defined with (15), in Figure 5 we show σ_d over different iterations for one specific step size value $\Delta = 0.42 \cdot \Delta_{opt}$ ($\Delta_{opt} = 1.087$), where the initial value was set to $\sigma_d^{(0)} = \Delta$. We can see that σ_d after only few iterations takes the value 0.42 and then saturates, showing the efficiency of the iterative process (15). The outcome of proces (15), $\sigma_d = 0.42 = c_{opt}$ ($\sigma_d^2 = 0.1764$ (-7.53 dB)) is in accordance with Figure 3 and Lemma 1, proving its correctness.

To summarize, by scaling the initial step size ($\Delta = \Delta_{opt}$) with properly chosen constant value we indeed improve the performance of single 2-bit USQ in the observed variance range. However, in this paper, we want to upgrade the approach from [20], with a goal to provide further performance enhancement. Particularly, we propose quantization based on switching between two 2-bit-USQs with unequal support regions (the one with the scaled parameters and the one with the initial parameter settings), resulting in the dual-mode USQ. In this way, the complexity of the method is slightly increased compared to single USQ, and it is much simpler compared to the adaptive models reported in [4, 5, 26]. A detailed description is given in the next section.

3. Theoretical Model of a 2-bit Dual-Mode USQ

In this section, we propose a dual-mode USQ, which improves the performance of the single USQ in a wide dynamic range of data variances. Dual-mode USQ is composed of two 2-bit USQs denoted with Q_1 (defined with $\Delta_1 = c \cdot \Delta_{opt}$, $c = 0.42$) and Q_2 (defined with $\Delta_2 = \Delta_{opt}$) having unequal support regions ($2\Delta_1 < 2\Delta_2$), whereby the switching among them, based on data variance classification, is adopted from [21]. Namely, data variance of block is classified into one of two possible variance ranges (an adequate USQ is associated to each range) by comparing data variance with the threshold between ranges, denoted with σ_t^2 . Based on the starting assumption that $\Delta_2 = \Delta_{opt}$, it is evident from Figure 3 that Q_2 is more appropriate (due to better SQNR scores) for data with the variance greater than σ_t^2 . The following lemma defines σ_t^2 .

Lemma 3. For a dual-mode USQ composed of two 2-bit USQs defined with the corresponding step sizes Δ_1 and Δ_2 ($\Delta_1 < \Delta_2$), σ_t can be determined iteratively using:

$$\sigma_t^{(i)} = \frac{\Delta_2^2 - \Delta_1^2}{2\sqrt{2} \left(\Delta_2 - \Delta_1 + 2 \left(\Delta_2 \exp \left\{ -\frac{\sqrt{2}\Delta_2}{\sigma_t^{(i-1)}} \right\} - \Delta_1 \exp \left\{ -\frac{\sqrt{2}\Delta_1}{\sigma_t^{(i-1)}} \right\} \right) \right)} \tag{20}$$

Proof of Lemma 3. In order to provide the highest possible SQNR of the dual-mode USQ composed of Q_1 (defined with Δ_1) and Q_2 (defined with Δ_2), σ_t^2 has to be determined as the variance where the SQNR curves of Q_1 and Q_2 intersect, that is, by equaling corresponding SQNRs (see Figure 3). SQNRs of Q_1 and Q_2 can be found using (6). From the condition:

$$\text{SQNR}(\sigma = \sigma_t, \Delta_1) = \text{SQNR}(\sigma = \sigma_t, \Delta_2), \tag{21}$$

we have:

$$\begin{aligned} & \frac{1}{4} \frac{\Delta_1^2}{\sigma_t} - \sqrt{2} \frac{\Delta_1}{\sigma_t} \left(\frac{1}{2} - \exp \left\{ -\frac{\sqrt{2}\Delta_1}{\sigma_t} \right\} \right) \\ & = \frac{1}{4} \frac{\Delta_2^2}{\sigma_t} - \sqrt{2} \frac{\Delta_2}{\sigma_t} \left(\frac{1}{2} - \exp \left\{ -\frac{\sqrt{2}\Delta_2}{\sigma_t} \right\} \right) \end{aligned} \tag{22}$$

From (22), after some mathematical manipulations, σ_t can be expressed as:

$$\sigma_t = \frac{\Delta_2^2 - \Delta_1^2}{2\sqrt{2} \left(\Delta_1 - \Delta_0 + 2 \left(\Delta_2 \exp \left\{ -\frac{\sqrt{2}\Delta_2}{\sigma_t} \right\} - \Delta_1 \exp \left\{ -\frac{\sqrt{2}\Delta_1}{\sigma_t} \right\} \right) \right)} \quad (23)$$

For $\Delta_1 = 0.42 \cdot \Delta_{opt}$ and $\Delta_2 = \Delta_{opt}$, when the iterative method (20) is initialized with $\sigma_t^{(0)} = (\Delta_1 + \Delta_2)/2$, the results presented in Figure 6 show fast convergence of the itera-

Figure 6

The iterative process for determination of the switching threshold σ_t^2 defined by (20) in the case when $\Delta_1 = 0.42 \cdot \Delta_{opt}$ and $\Delta_2 = \Delta_{opt}$

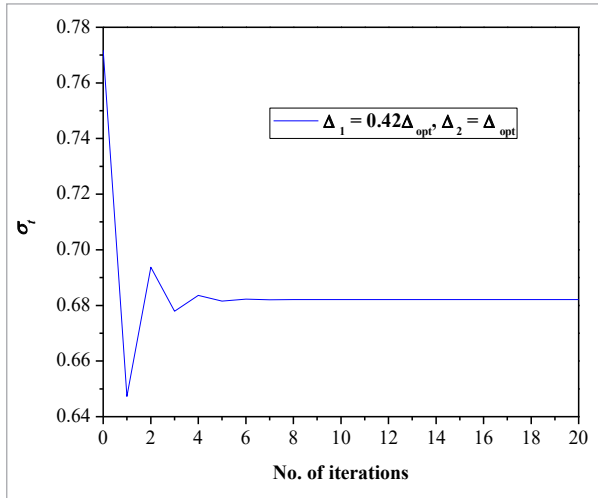
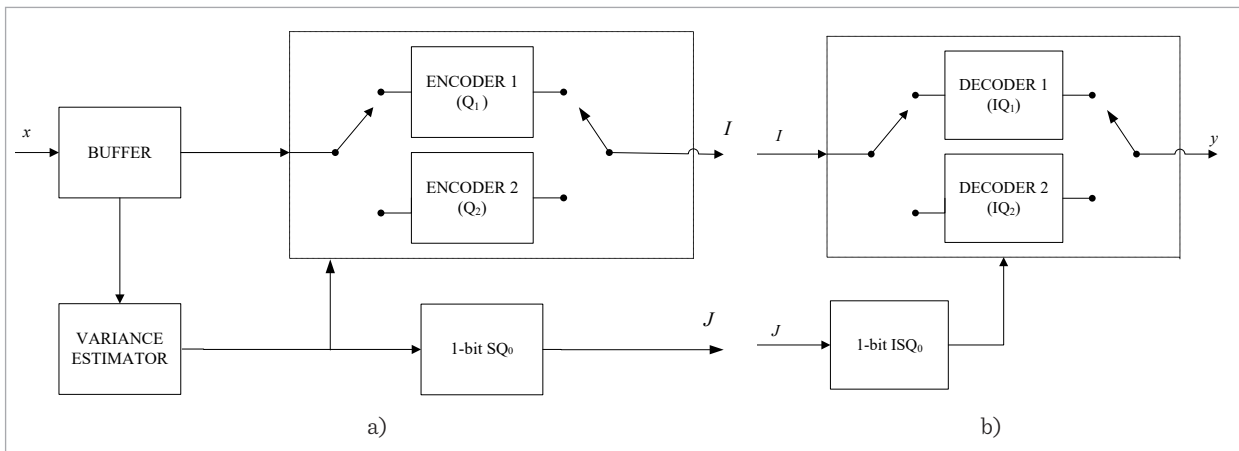


Figure 7

Block diagram of the 2-bit dual-mode USQ: a) encoder; b) decoder



tive method. The obtained value $\sigma_t = 0.6821$ ($\sigma_t^2 = 0.4653$ (-3.32 dB)) matches the result in Figure 3 (see σ_t^2).

Figure 7 shows a block diagram of the proposed dual-mode USQ operating in a block-by-block manner. The following steps describe the operating principle of encoder (Figure 7(a)):

- 1 **Buffering.** The input data is stored in the buffer forming the block, with a capacity of M samples.
- 2 **Variance estimation.** The variance of the buffered block, σ^2 , is estimated as follows [10, 21, 26]:

$$\sigma^2 = \frac{1}{M} \sum_{n=1}^M x^2(n). \quad (24)$$

- 3 **Switching.** For the current block, the choice of one of two disposable USQs is based on the following comparison: if $\sigma^2 \leq \sigma_t^2$ switch to Q_1 , otherwise to Q_2 . Information about the employed quantizer is represented with 1 bit, and it denotes the side information that should be stored per each block. As Figure 7(a) shows, the side information can be obtained by quantizing estimated variance of the block using the 1-bit asymmetric scalar quantizer (1-bit SQ_0) as follows:

$$\hat{\sigma} = Q(\sigma) = \begin{cases} 1, & \sigma^2 \leq \sigma_t^2 \\ 0, & \sigma^2 > \sigma_t^2 \end{cases}. \quad (25)$$

- 5 **Encoding (quantization).** Each data sample within the block $x(n)$, $n = 1, \dots, M$, is encoded (quantized) with the chosen USQ, resulting in sequence of M codewords of length 2 bits (see index I).

The decoder is shown in Figure 7(b). Index J is decoded using the inverse 1-bit SQ_0 (1-bit ISQ_0) and used to select inverse USQ for decoding of sequence I . After that, the sequence (or index) I is decoded by using the selected inverse USQ providing in that manner the reconstruction of the samples within the current data block, $x^a(n) = y(n)$, $n = 1, \dots, M$.

Finally, let us define bit rate R and SQNR of the considered dual-mode USQ. Thus, the bit rate is given by:

$$R = 2 + \frac{1}{M} \text{ [bits/ data sample]}, \quad (26)$$

while SQNR can be evaluated by using [21]:

$$\text{SQNR} = \begin{cases} -10 \log_{10} \left(1 + \frac{\Delta_1^2}{4\sigma^2} - \frac{\sqrt{2}\Delta_1}{\sigma} \left(\frac{1}{2} - \exp \left\{ -\frac{\sqrt{2}\Delta_1}{\sigma} \right\} \right) \right), & \sigma^2 \leq \sigma_i^2 \\ -10 \log_{10} \left(1 + \frac{\Delta_2^2}{4\sigma^2} - \frac{\sqrt{2}\Delta_2}{\sigma} \left(\frac{1}{2} - \exp \left\{ -\frac{\sqrt{2}\Delta_2}{\sigma} \right\} \right) \right), & \sigma^2 > \sigma_i^2 \end{cases}, \quad (27)$$

In Figure 8, we show SQNR for the proposed dual-mode USQ and single 2-bit USQ ($\Delta = c \cdot \Delta_{\text{opt}}$, $c = 0.42$) over the same variance range as in Figure 3. We can notice evident improvement in performance (SQNR) compared to scaling-based approach. Figure 8 also shows that SQNR of dual-mode USQ for two values of the input variance reaches a maximum.

Figure 8

SQNR of the 2-bit dual-mode USQ ($\Delta_1 = 0.42 \cdot \Delta_{\text{opt}}$ and $\Delta_2 = \Delta_{\text{opt}}$) and single 2-bit USQ ($\Delta = 0.42 \cdot \Delta_{\text{opt}}$) in a wide dynamic range of input data variances

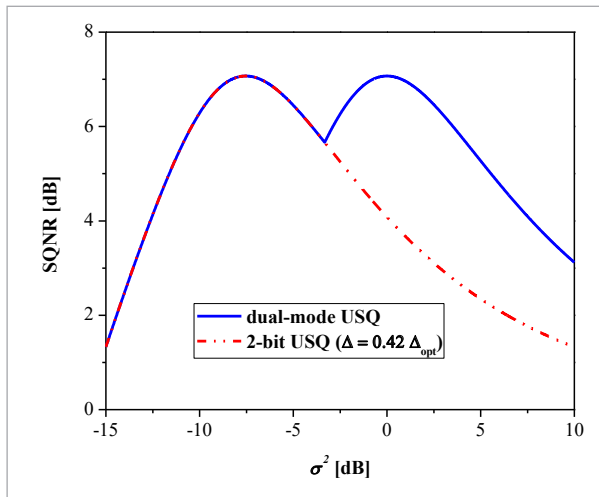


Table 1

SQNR_{av} of the proposed 2-bit dual-mode USQ ($\Delta_1 = 0.42 \cdot \Delta_{\text{opt}}$ and $\Delta_2 = \Delta_{\text{opt}}$) and single 2-bit USQs with $\Delta = 0.42 \cdot \Delta_{\text{opt}}$ and $\Delta = \Delta_{\text{opt}}$

Dual-mode USQ ($\Delta_1 = 0.42 \cdot \Delta_{\text{opt}}$, $\Delta_2 = \Delta_{\text{opt}}$)	USQ ($\Delta = 0.42 \cdot \Delta_{\text{opt}}$)	USQ ($\Delta = \Delta_{\text{opt}}$)
5.55 dB	4.21 dB	2.6 dB

To measure the achieved performance gain of the dual-mode USQ over the single USQ in the observed range of variances, we calculated the average SQNR as given in Table 1. It can be noted that the proposed dual-mode USQ improves average SQNR for 1.34 dB with respect to USQ with $\Delta = c \cdot \Delta_{\text{opt}}$ ($c = 0.42$), while the gain of 2.8 dB is observed with respect to initial USQ ($\Delta = \Delta_{\text{opt}}$). Based on these achievements, we can conclude that the dual-mode USQ proposed in this paper is a better candidate for applications where data variance tends to change (e.g. quantization of the weights of NN) than the single 2-bit USQ. Let us emphasize that the performance gain is achieved with slightly increased bit rate of $1/M$ bits/data sample (see eq. (26)). Typically, selection of M depends on concrete application.

4. Results and Discussion

In this Section, we present, discuss and compare the performances (in both the theoretical and experimental domain) of the proposed 2-bit dual-mode USQ and some other known 2-bit quantizers used as baselines. Obviously, the goal is to point out the advantages that can be achieved in quantization of non-stationary data by using this simple quantization approach we propose in this paper.

As baseline quantization models, we consider two adaptive USQs developed in [4, 5], as well as non-adaptive 2-bit non-uniform logarithmic scalar quantizer introduced in [20]. These models are selected as they have already proved the efficiency in processing NN weights. In this manner, we will demonstrate that our quantization model is suitable for application in NN compression.

Note that relation between the step size and the designed-for variance of the quantizer given in [5] has already been mentioned in Section 1 (see Figure 2). As that quantizer is adaptive, it holds that $\Delta^{[5]}(\sigma) = \sigma \cdot \Delta^{[5]}$

$(\sigma_{\text{ref}}) = 2 \cdot c_1 \cdot \sigma / 3$. For its performance evaluation (in the theoretical domain) expression (6) holds, which gives:

$$\begin{aligned} \text{SQNR}(\sigma) &= -10 \log_{10} \left(1 + \frac{c_1^2}{9} - \frac{2\sqrt{2}}{3} c_1 \left(\frac{1}{2} + \exp \left\{ -\frac{2\sqrt{2}}{3} c_1 \right\} \right) \right) \\ &= 3.17 \text{ dB} \end{aligned} \quad (28)$$

pointing out that SQNR is independent on σ . On the other hand, quantizer from [4] is adaptive uniform quantizer having step size $\Delta^{[4]}(\sigma) = \sigma \beta$, $\beta = c_1 = 3.2$ (quantization levels are defined as $\{-\sigma\beta, 0, 0, \sigma\beta\}$). Distortion in this case is specified by:

$$\begin{aligned} D(\sigma) &= 2 \int_0^{\Delta^{[4]}(\sigma)/2} x^2 p(x, \sigma) dx \\ &+ 2 \int_{\Delta^{[4]}(\sigma)/2}^{\infty} (x - \sigma \cdot \beta)^2 p(x, \sigma) dx = \cdot \\ &= \sigma^2 \left(1 - \sqrt{2} \beta \exp \left\{ -\frac{\beta}{\sqrt{2}} \right\} \right) \end{aligned} \quad (29)$$

whereas SQNR is defined as:

$$\begin{aligned} \text{SQNR}(\sigma) &= 10 \log_{10} \left(\frac{\sigma^2}{D(\sigma)} \right) \\ &= -10 \log_{10} \left(1 - \sqrt{2} \beta \exp \left\{ -\frac{\beta}{\sqrt{2}} \right\} \right) \\ &= 2.77 \text{ dB} \end{aligned} \quad (30)$$

also pointing out that SQNR is independent on σ . Regarding the 2-bit logarithmic quantizer from [20], it was designed using the following parameters: $\mu = 255$ and $x_{\text{max}} = 4.318$.

The SQNR dependence on data variance for the proposed dual-mode USQ and the considered baseline quantizers [4, 5, 20] is provided in Figure 9, while their average SQNRs are listed in Table 2. By observing Figure 9, we can see that our proposal is better than the 2-bit quantizer from [20], as the SQNR curve corresponding to the quantizer from [20] is overreached in the entire variance range. The same conclusion can be drawn for quantizers from [4, 5]. Furthermore, in Table 2 we report the significant gains in average SQNR of about 2.8 dB, 2.4 dB, and 3.85 dB in comparison to the invoked 2-bit baselines [4, 5, 20], respectively.

Figure 9

SQNR of the 2-bit dual-mode USQ ($\Delta_1 = 0.42 \cdot \Delta_{\text{opt}}$ and $\Delta_2 = \Delta_{\text{opt}}$) and baseline quantizers in a wide dynamic range of input data variances

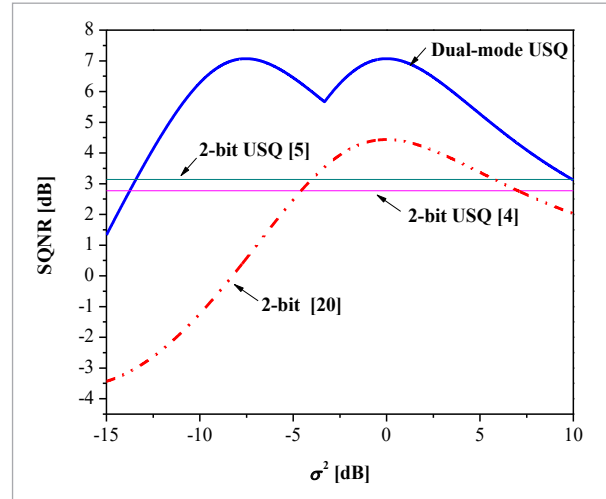


Table 2

SQNR_{av} of the proposed 2-bit dual-mode USQ ($\Delta_1 = 0.42 \cdot \Delta_{\text{opt}}$ and $\Delta_2 = \Delta_{\text{opt}}$) and baseline 2-bit quantizers

Dual-mode USQ ($\Delta_1 = 0.42 \cdot \Delta_{\text{opt}}$, $\Delta_2 = \Delta_{\text{opt}}$)	2-bit USQ [4]	2-bit USQ [5]	2-bit [20]
5.55 dB	2.77 dB	3.17 dB	1.70 dB

The theoretical results are supported by experimental ones, obtained in quantization of weights of the MLP neural network with one hidden layer [30], being used for image classification task.

MLP is trained and tested on MNIST database [12] with 60000 gray-scaled images of handwritten digits used for network training and 10000 images for testing with the size 28 x 28 pixels. The input network layer is composed of 784 nodes that correspond to 28x28 pixels image size, 128 nodes in the hidden layer and 10 nodes in the output layer that correspond to 10 classes (digits). We use the following hyperparameter settings: regularization rate = 0.01, learning rate = 0.0005 and batch size = 128.

As indicated in [21], the weights of such trained MLP have the distribution that is roughly Laplacian, with parameters: zero-mean and variance $\sigma_w^2 = 0.01$ ($\sigma_{w,\text{dB}} = 10 \cdot \log_{10}(\sigma_w^2) = -20$ dB). It is also worth emphasizing that the uncompressed weights are available in a matrix form of dimensions 784x128. To obtain a

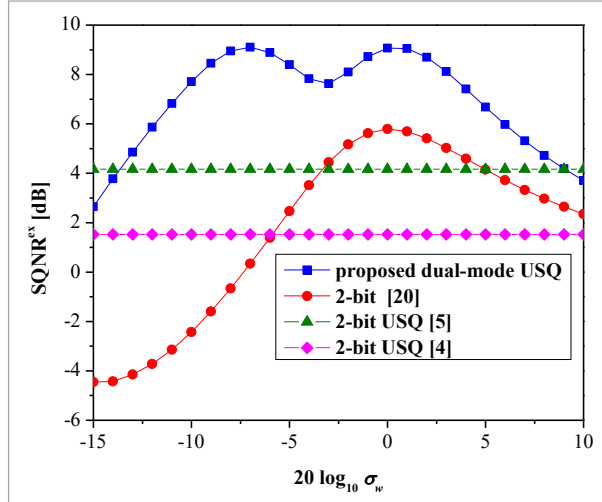
quantized matrix, we fully apply the steps described in the previous Section 3, whereas the processing is done row-wise (that is, channel-wise), implying that $M = 128$. For baseline quantizers [4, 5], for each row (block) the variance is estimated (see (24)) and further used to adapt their parameters. In the case of non-adaptive quantizer form [20], the matrix of weights is processed sample by sample. To measure the quantizer efficiency on real data the experimental value of SQNR is evaluated as [21, 26]:

$$\text{SQNR}^{\text{ex}} = 10 \log_{10} \left(\frac{\sum_{i=1}^W w_i^2}{\sum_{i=1}^W (w_i - w_i^q)^2} \right), \quad (31)$$

where w_i are the original and w_i^q are the quantized network weights and W is total number of weights.

Figure 10

SQNR^{ex} of the 2-bit dual-mode USQ ($\Delta_1 = 0.42 \cdot \Delta_{\text{opt}}$ and $\Delta_2 = \Delta_{\text{opt}}$) and considered baselines in a wide dynamic range of weights variances



Experimental results are presented in Figure 10. They show the dependence of SQNR^{ex} on the weights' variance for the dual-mode USQ and the corresponding baselines (same settings as in Figure 9). By multiplying the matrix of trained weights (the original, uncompressed weight matrix) with the appropriate constant k we obtain the matrix of weights with variances $\sigma^2 = k^2 \cdot \sigma_w^2$ ($\sigma_{\text{dB}} = 20 \cdot \log_{10}(k \cdot \sigma_w)$), which allows us to analyze

the performance in the wide variance range. We can see that experimental SQNR values (obtained using (31)) verify the theoretical results shown in Figure 9. Table 3 summarizes the overall bit rate required for all considered quantizers. It can be seen that our model requires a slightly increased bit rate for 1/128 bits (1 bit per block size) compared to the quantizer in [20], but the rate is lower than that required for adaptive baselines [4, 5] that use 32 bits to quantize side information (i.e. block variance) per each block. To summarize, in addition to the increased SQNR value within the wide variance range, our solution is less demanding in terms of bit rate compared with adaptive baselines. This makes the proposed dual-mode USQ adequate for weights processing and accordingly adequate for NN compression.

Table 3

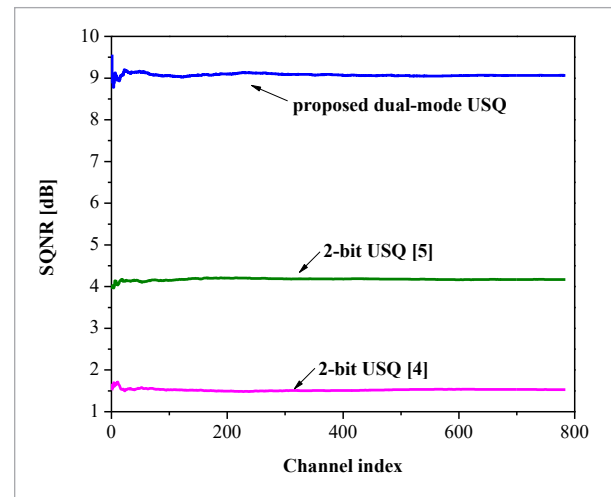
Bit rate in bits/data sample of the proposed 2-bit dual-mode USQ ($\Delta_1 = 0.42 \cdot \Delta_{\text{opt}}$, $\Delta_2 = \Delta_{\text{opt}}$) and baseline 2-bit quantizers

Dual-mode USQ ($\Delta_1=0.42 \cdot \Delta_{\text{opt}}, \Delta_2 = \Delta_{\text{opt}}$)	2-bit USQ [4]	2-bit USQ [5]	2-bit [20]
2.00775 bits	2.25 bits	2.25 bits	2 bits

Figure 11 depicts the SQNR values per each channel (784 channels in total) achieved by the proposed and baseline quantizers, in the case when the weights matrix has variance 0 dB ($k = 10$).

Figure 11

The SQNR values per channels (784 channels in total) attained by the proposed dual-mode USQ and the considered baselines, for the case when the weights variance is 0 dB



5. Conclusion

Quantization and compression of non-stationary data with the Laplacian distribution by using low rate USQ are considered in this paper. We propose a 2-bit dual-mode USQ, which operates by switching between two non-adaptive 2-bit USQs depending on the estimated variance of the buffered data. The idea originates from [21], but few progressive steps are performed in this paper. We change a heuristically determined step value of the initial 2-bit USQ by the one which is obtained by maximizing the performance for the particular variance. Further, we determine the best value for scaling factor (used to scale the initial step size) and finally we propose an iterative rule for determining the optimal variance threshold used to compare variances during USQ selection. Apart this, we determine the relation between SQNR dependences on variance for USQs whose step sizes are linearly proportional. Specifically, we prove that by multiplying the step size of USQ with a constant c , the dependence of the SQNR on the variance is shifted for $20 \cdot \log_{10} c$. Owing to that the performance analysis of the dual-mode USQ is performed in a more convenient manner. The proposed dual-mode USQ improves average SQNR of the single 2-bit USQ with

slightly increased bit rate for $1/M$ bits/data samples. Moreover, we record the significant gain in average SQNR of about 2.8 dB and 2.4 dB with respect to the adaptive 2-bit USQs given in [4] and [5]. Finally, the quantization model we propose is characterized by a reduced complexity and lower bit rate in comparison with the corresponding adaptive quantizers. Based on these achievements, we can conclude that the quantization model we propose is suitable for quantization of data modeled with Laplacian distribution whose variance tends to change. In particular, it can be suitable for applications when the compression rate is of greater interest than the high data quality. The theoretical results are verified by experimental ones, obtained in quantization of weights of a real NN. In that way, possibility of dual-mode 2-bit USQ implementation in NN compression is indicated. In our further research the influence of 2-bit quantization on NN performance will be studied.

Acknowledgement

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

References

1. Banner, R., Hubara, I., Hoffer, E., Soudry, D. Scalable Methods for 8-bit Training of Neural Networks. Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada, December 2-8, 2018.
2. Banner, R., Nahshan, Y., Hoffer, E., Soudry, D. ACIQ: Analytical Clipping for Integer Quantization of Neural Networks. arXiv preprint arXiv: 1810.05723, 2018.
3. Banner, R., Nahshan, Y., Soudry, D. Post Training 4-bit Quantization of Convolutional Networks for Rapid-Deployment. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, December 8-10, 2019.
4. Baskin, C., Zheltonozhkii, E., Rozen, T., Liss, N., Chai, Y., Schwartz, E., Giryes, R., Bronstein, A., Mendelson, A. NICE: Noise Injection and Clamping Estimation for Neural Network Quantization. Mathematics, 2021, 9(17), 2144. <https://doi.org/10.3390/math9172144>
5. Choi, J., Venkataramani, S., Srinivasan, V., Gopalakrishnan, K., Wang, Z., Chuang, P. Accurate and Efficient 2-Bit Quantized Neural Networks. Proceedings of the 2nd SysML Conference, Palo Alto, CA, USA, March 31-April 2, 2019.
6. Dai, S., Venkatesan, R., Ren, H., Zimmer, B., Dally, W. J., Khailany, B. VS-QUANT: Per-Vector Scaled Quantization for Accurate Low-Precision Neural Network Inference. Proceedings of the 4th MLSys Conference, San Jose, CA, USA, 2021.
7. Gazor, S., Zhang, W. Speech Probability Distribution. IEEE Signal Processing Letters, 2003, 10(7), 204-207. <https://doi.org/10.1109/LSP.2003.813679>
8. Hirtzlin, T., Penkovsky, B., Bocquet, M., Klein, J.-O., Portal, J.-M., Querlioz, D. Stochastic Computing for Hardware Implementation of Binarized Neural Networks. IEEE Access, 2019, 76394-76403. <https://doi.org/10.1109/ACCESS.2019.2921104>

9. Hui, D., Neuhoff, D. L. Asymptotic Analysis of Optimal Fixed-Rate Uniform Scalar Quantization. *IEEE Transactions on Information Theory*, 2001, 47(3), 957-977. <https://doi.org/10.1109/18.915652>
10. Jayant, N. C., Noll, P. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, New Jersey, 1984. [https://doi.org/10.1016/0165-1684\(85\)90053-2](https://doi.org/10.1016/0165-1684(85)90053-2)
11. Kim, S., Lee, H-J. RGBW Image Compression by Low-Complexity Adaptive multi-Level Block Truncation Coding. *IEEE Transaction on Consumer Electronics*, 2016, 62(4), 412-419. <https://doi.org/10.1109/TCE.2016.7838094>
12. LeCun, Y., Cortez, C., Burges, C. The MNIST Handwritten Digit Database. Available online: yann.lecun.co.
13. Na, S. Asymptotic Formulas for Mismatched Fixed-Rate Minimum MSE Laplacian Quantizers. *IEEE Signal Processing Letters*, 2008, 15, 13-16. <https://doi.org/10.1109/LSP.2007.910240>
14. Na, S. Variance-mismatched fixed-rate scalar quantization of Laplacian sources. *IEEE Transactions on Information Theory*, 2011, 57 (7), 4561-4572. <https://doi.org/10.1109/TIT.2011.2146390>
15. Na, S., Neuhoff, D. L. Asymptotic MSE Distortion of Mismatched Uniform Scalar Quantization. *IEEE Transactions on Information Theory*, 2012, 58(5), 3169-3181. <https://doi.org/10.1109/TIT.2011.2179843>
16. Na, S., Neuhoff, D. L. Monotonicity of Step Sizes of MSE-Optimal Symmetric Uniform Scalar Quantizers. *IEEE Transactions on Information Theory*, 2018, 65(3), 1782-1792. <https://doi.org/10.1109/TIT.2018.2867182>
17. Na, S., Neuhoff, D. L. On the Convexity of the MSE Distortion of Symmetric Uniform Scalar Quantization. *IEEE Transactions on Information Theory*, 2017, 64(4), 2626-2638. <https://doi.org/10.1109/TIT.2017.2775615>
18. Perić, Z., Denić, B., Despotović, V. Algorithm Based on 2-Bit Adaptive Delta Modulation and Fractional Linear Prediction for Gaussian Source Coding. *IET Signal Processing*, 2021, 15 (6), 410-423. <https://doi.org/10.1049/sil2.12040>
19. Perić, Z., Denić, B., Despotović, V. Novel Two-Bit Adaptive Delta Modulation Algorithms. *Informatika*, 2019, 30(1), 117-134. <https://doi.org/10.15388/Informatika.2019.200>
20. Perić, Z., Denić, B., Dinčić, M., Nikolić, J. Robust 2-Bit Quantization of Weights in Neural Network Modeled by Laplacian Distribution. *Advances in Electrical and Computer Engineering*, 2021, 21(3), pp. 3-10. <https://doi.org/10.4316/AECE.2021.03001>
21. Perić, Z., Denić, B., Jovanović, A., Savić, M., Vučić, N., Nikolić, A. A Dual-Mode 2-bit Uniform Scalar Quantizer of Laplacian Source. *Proceedings of the 29th Telecommunications forum TELFOR 2021, Belgrade, Serbia, November 23-24, 2021*. <https://doi.org/10.1109/TELFOR52709.2021.9653253>
22. Perić, Z., Denić, B., Savić, M., Despotović, V. Design and Analysis of Binary Scalar Quantizer of Laplacian Source with Applications. *Information*, 2020, 11, Article ID: 501. <https://doi.org/10.3390/info11110501>
23. Perić, Z., Denić, B., Savić, M., Dinčić, M., Mihajlov, D. Quantization of Weights of Neural Networks with Negligible Decreasing of Prediction Accuracy. *Information Technology and Control*, 2021, 50(3), 558-569. <https://doi.org/10.5755/j01.itc.50.3.28468>
24. Perić, Z., Nikolić, J. An Adaptive Waveform Coding Algorithm and its Application in Speech Coding. *Digital Signal Processing*, 2012, 22(1), 199-209. <https://doi.org/10.1016/j.dsp.2011.09.001>
25. Perić, Z., Nikolić, J., Denić, B., Despotović, V. Forward Adaptive Dual-Mode Quantizer Based on the First-Degree Spline Approximation and Embedded G.711 Codec. *Radioengineering*, 2019, 28(4), 729-739. <https://doi.org/10.13164/re.2019.0729>
26. Perić, Z., Savić, M., Simić, N., Denić, B., Despotović V. Design of a 2-Bit Neural Network Quantizer for Laplacian Source. *Entropy*, 2021, 23(8), Article ID: 933. <https://doi.org/10.3390/e23080933>
27. Simons, T., Lee, D-J. A Review of Binarized Neural Networks. *Electronics*, 2019, 8 (6), 661. <https://doi.org/10.3390/electronics8060661>
28. Wu, Q., Chen, C., Wang, C., Wu, Y., Zhao, Y., Wu, X. Improving the Accuracy of Binarized Neural Networks and Application on Remote Sensing Data. *IEEE Geoscience and Remote Sensing Letters*, 2019, 17(7), 1278-1282. <https://doi.org/10.1109/LGRS.2019.2942348>
29. Yang, C-H., Chou, Y-C., Chang, T-K., Kim, C. An Enhanced Adaptive Block Truncation Coding with Edge Quantization Scheme. *Applied Sciences*, 2020, 10(20), Article ID: 7340. <https://doi.org/10.3390/app10207340>
30. Zhang, A., Lipton, Z. C., Li, M., Smola, A. J. *Dive into Deep Learning*. Amazon Science, 2020

