11th International Conference on Electrical, Electronics and Computer Engineering (IcETRAN), Nis, 3-6 June 2024

Performance Analysis of the 24-bits Floating-Point Format for a Gaussian Source

Zoran Perić
Department of Telecommunications
Faculty of Electronic Engineering,
University of Niš
Niš, Serbia
zoran.peric@elfak.ni.ac.rs
https://orcid.org/0000-0002-8267-9541

Bojan Denić
Department of Telecommunications
Faculty of Electronic Engineering,
University of Niš
Niš, Serbia
bojan.denic@elfak.ni.ac.rs
https://orcid.org/0000-0002-9003-2545

Milan Dinčić
Department of Measurements
Faculty of Electronic Engineering,
University of Niš
Niš, Serbia
milan.dincic@elfak.ni.ac.rs
https://orcid.org/0000-0001-7508-0277

***Abstract*—The research conducted in this paper targets to examine the efficiency of the 24-bits floating-point (FP24) format on data modelled by the Gaussian distribution. To accomplish this goal, we exploit a connection that exists between binary formats and quantization, so we apply the objective measures such as distortion and SQNR (signal-to-quantization noise ratio) to estimate the performance of the FP24 format. In particular, performance investigation of this format is performed in the theoretical domain for both mean-squared error (MSE) and absolute error (AE) measures. Results for both metrics are provided in the wide dynamic range of the input data variances and show that the SQNR is high (103.79 dB for MSE and 105.19 dB for AE) and varies negligibly. Therefore, the FP24 format can be considered as an excellent candidate for representing Gaussian data in practical variance-sensitive applications.**

Keywords—floating point format, Gaussian source, piecewise uniform quantizer, SQNR

I. Introduction

Various practical applications and systems support the floating-point (FP) format, although the implementation of the fixed-point (FXP) format is simpler. This is because the FP format, when confronted with the FXP format, can enable high performance (accuracy) in data representation in a much larger range of data variances. Codewords generated in the FP format, besides the sign bit, include bits that convey information about the exponent E and mantissa M of the real number. Different codeword lengths were proposed for this format so far, including 32 [1], 24 [2], 16 [3] or 8 bits [4]. Let us recall that 32-bits FP version (FP32) defined by IEEE 754 standard [1] applies 8 bits to encode E and 23 bits to encode M. It should also be noted that this standardized FP version is the first choice for applications with large available resources (such as memory and processing power). However, for resource-constrained applications it is recommended to use the FP variants with a lower number of bits; for example, the 8-bits FP format is a good choice for edge devices.

In [5−8], the analysis of FP and FXP formats was done using the equivalent quantization scheme and using SQNR as performance measure. That kind of analysis actually contributed in establishing a relationship between SQNR and accuracy in data representation. Namely, it was pointed out in [5−7] that the quantizer corresponding to the FP format is a piecewise uniform quantizer, while in [8] was indicated that quantizer equivalent to the FXP format is a uniform quantizer. In [5−7] are calculated performance of the FP32, FP24 (24-bits FP) and bfloat16 (16-bits FP) format, respectively, for the data with Laplacian distribution, where as a performance measure a MSE metric [9−11] is used. It was shown in a wide dynamic range of data variances that all considered FP formats provides high level of robustness as SQNR remains unchanged with the change of data variance, where the best SQNR performance is observed in the case of FP32 format, while FP24 is better that bfloat16. The study in [8] performs a performance analysis using the AE metric [9, 10, 12, 13], where FP32 and FXP32 (32-bits FXP) formats were observed for Laplacian data.

Besides the Laplacian source, the Gaussian source is also used as a statistical model for the data [9, 10]. Some important examples of data that follow the Gaussian distribution are neural network weights and measurement data collected from sensors [14]. Consequently, investigating the performance of binary formats for Gaussian data is significant from a practical point of view.

This paper addresses the FP24 format and aims to explore its performance in the case of a Gaussian source from the SQNR point of view, which is not done before. In other words, following the approach from [5−7], we use the adequate quantization model and then take into account the variance of the data when investigating the performance. Specifically, we consider both the MSE and AE metrics in order to provide a detailed performance analysis. For both metrics involved, the efficiency of the FP24 format in processing Gaussian data is shown to be high, as it maintains constancy in the accuracy of the representation when the variance of the data changes.

The rest of the paper is organized as follows. In Section II, after a short description of the FP24 format, we introduce the quantization scheme equivalent to this format and derive expressions for performance evaluation for both considered metrics. Section III presents and discusses theoretical results for the FP24 format obtained in processing of Gaussian data. Finally, Section IV provides concluding remarks.

II. FP24 Quantizer Model

In this Section, the FP24 format will be described and the performance of the quantizer analogous to this format (FP24 quantizer) will be evaluated for two commonly used metrics.

A. Description of the FP24 Format and Introduction of the FP24 Quantizer Model

Consider Fig. 1 where the binary representation of a real number x is given in FP24 format. We can see that FP24 codeword is composed of the sign bit 's', 8 bits $(e_1e_2 \dots e_8)_2$ for the exponent E and 15 bits $(m_1m_2 \dots m_{15})_2$ for the mantissa M.

This paper has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (grant number 451-03-65/2024-03/200102)

$$x:\ \underbrace{\underbrace{\underset{\text{sign bit}}{s}\ \underbrace{e_1e_2...e_8}_{\text{bits for } E}\ \underbrace{m_1m_2...m_{15}}_{\text{bits for } M}}_{24\,\text{bits in total}}}$$

Fig. 1. The binary representation of FP24 number.

It is worth noting that E and M are integers and can be obtained from their binary forms as:

$$E=\left(e_1e_2...e_8\right)_2=\sum_{i=1}^{8}e_i2^{8-i}\,, \tag{1}$$

$$M=\left(m_1m_2...m_{15}\right)_2=\sum_{i=1}^{15}m_i2^{15-i}\,. \tag{2}$$

From (1) it follows that $E_{\min}=0\leqslant E\leqslant E_{\max}=255$, while from (2) we have that $M_{\min}=0\leqslant M\leqslant M_{\max}=2^{15}-1$. Using E and M, the decimal form of the FP24 number can be determined according to [1]:

$$x=(-1)^s\,2^{E-127}\left(1+\frac{M}{2^m}\right)=(-1)^s\,2^{E^*}\left(1+\frac{M}{2^m}\right), \tag{3}$$

where E^* is the biased exponent, $E_{\min}^*=-127\leqslant E^*\leqslant E_{\max}^*=128$.

Observe that the FP24 format displays both positive and negative numbers. Since the FP24 format is zero-symmetrical, it follows that for each positive number there is a negative counterpart. The maximal positive FP24 value is $x_{\max}=2^{E\max^*}=2^{128}$, while $-x_{\max}=-2^{128}$ is the maximal negative FP24 value. Positive FP24 numbers are divided into 256 groups, where each group includes 2^{15} equidistant numbers. Note that the group is defined by the particular value of E^*; accordingly, numbers in the same group are located in the range $[2^{E^*},\,2^{E^*+1}]$, while the step size in the group is:

$$\Delta_{E^*}=2^{E^*}\left(1+\frac{M+1}{2^{15}}\right)-2^{E^*}\left(1+\frac{M}{2^{15}}\right)=2^{E^*-15}\,, \tag{4}$$

where E^* goes from -127 to 128.

As pointed out in [6], the FP24 quantizer is a symmetric $N=2^{24}$-levels piecewise uniform quantizer whose support region $[-x_{\max},\,x_{\max}]$ is partitioned into 256 unequal segments $[2^{E^*},\,2^{E^*+1}]$, $E^*=-127,\,\ldots,\,128$, where Δ_{E^*} (see (4)) is the step size within the segment containing 2^{15} quantization levels.

In the following, we will evaluate the performance (distortion or equivalently SQNR) of this quantizer for two different metrics (MSE and AE), when the input data is described by the Gaussian PDF [9−12]:

$$p(x,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{5}$$

B. Performance for the MSE Metric

Let us introduce the granular D_g and overload distortion D_{ov}, which represent the components of the total distortion D produced inside and outside the granular region, respectively. In the case of FP quantizers and MSE metric, the basic expressions for D_g and D_{ov} calculation are defined as follows [5−7]:

$$D_g=2\sum_{E^*=E_{\min}^*}^{E_{\max}^*-1}\frac{\Delta_{E^*}^2}{12}P_{E^*}\,, \tag{6}$$

$$D_{ov}=2\int_{x_{\max}}^{\infty}\left(x-x_{\max}\right)^2p(x,\sigma)\,dx\,, \tag{7}$$

where P_{E^*} denotes the segment probability. P_{E^*} for the Gaussian PDF can be calculated as:

$$P_{E^*}=\int_{2^{E^*}}^{2^{E^*+1}}p(x,\sigma)\,dx=\frac{1}{2}\left(erf\left(\frac{2^{E^*+1/2}}{\sigma}\right)-erf\left(\frac{2^{E^*-1/2}}{\sigma}\right)\right). \tag{8}$$

Applying (4) and (8) in (6), we derive the following expression for D_g of the FP24 quantizer:

$$D_g=\sigma^2\left(\sum_{E^*=-127}^{127}\frac{2^{2E^*-32}}{3\sigma^2}\left(erf\left(\frac{2^{E^*+1/2}}{\sigma}\right)-erf\left(\frac{2^{E^*-1/2}}{\sigma}\right)\right)\right), \tag{9}$$

while for D_{ov} of the FP24 quantizer it is obtained:

$$D_{ov}=\sigma^2\left(-\sqrt{\frac{2}{\pi}}\frac{x_{\max}}{\sigma}\exp\left\{-\frac{x_{\max}^2}{2\sigma^2}\right\}+\left(\frac{x_{\max}^2}{\sigma^2}+1\right)erfc\left(\frac{x_{\max}}{\sigma\sqrt{2}}\right)\right). \tag{10}$$

Using (9) and (10), we can define the expression for the SQNR of the FP24 quantizer for the Gaussian source:

$$\begin{aligned}\mathrm{SQNR}&=10\log_{10}\left(\frac{\sigma^2}{D_g+D_{ov}}\right)\\&=-10\log_{10}\left(\sum_{E^*=-127}^{127}\frac{2^{2E^*-32}}{3\sigma^2}\left(erf\left(\frac{2^{E^*+1/2}}{\sigma}\right)-erf\left(\frac{2^{E^*-1/2}}{\sigma}\right)\right)\right.\\&\left.-\sqrt{\frac{2}{\pi}}\frac{x_{\max}}{\sigma}\exp\left(-\frac{x_{\max}^2}{2\sigma^2}\right)+\left(\frac{x_{\max}^2}{\sigma^2}+1\right)erfc\left(\frac{x_{\max}}{\sigma\sqrt{2}}\right)\right)\end{aligned} \tag{11}$$

C. Performance for the AE Metric

For FP quantizers and AE metric, the following expressions can be applied for evaluation of the total distortion components [8]:

$$D_g = 2\sum_{E^*=E^*_{\min}}^{E^*_{\max}-1} \frac{\Delta_{E^*}}{4} P_{E^*}, \quad (12)$$

$$D_{ov} = 2\int_{x_{\max}}^{\infty} (x - x_{\max})\, p(x,\sigma)\, dx. \quad (13)$$

Note that Δ_{E^*} and P_{E^*} in (12) are specified with (4) and (8) respectively. Accordingly, we derive the following for D_g and D_{ov} of the FP24 quantizer (AE metric):

$$D_g = \sigma \sum_{E^*=-127}^{127} \frac{2^{E^*-17}}{\sigma}\left(erf\left(\frac{2^{E^*+1/2}}{\sigma}\right) - erf\left(\frac{2^{E^*-1/2}}{\sigma}\right)\right), \quad (14)$$

$$D_{ov} = \sigma\left(\sqrt{\frac{2}{\pi}}\exp\left\{-\frac{x_{\max}^2}{2\sigma^2}\right\} - \frac{x_{\max}}{\sigma} erfc\left(\frac{x_{\max}}{\sigma\sqrt{2}}\right)\right). \quad (15)$$

For this performance metric, SQNR is defined with [8, 9]:

$$\mathrm{SQNR} = 20\log_{10}\left(\frac{2}{D}\int_0^{\infty} xp(x,\sigma)\, dx\right). \quad (16)$$

Substituting (5), (14) and (15) in (16), we obtain the SQNR for the FP24 quantizer:

$$\mathrm{SQNR} = 20\log_{10}\left(\sqrt{\frac{2}{\pi}}\frac{\sigma}{D}\right)$$
$$= -20\log_{10}\left(\sqrt{\frac{\pi}{2}}\sum_{E^*=-127}^{127}\frac{2^{E^*-17}}{\sigma}\left(erf\left(\frac{2^{E^*+1/2}}{\sigma}\right) - erf\left(\frac{2^{E^*-1/2}}{\sigma}\right)\right)\right.$$
$$\left. -\exp\left(-\frac{x_{\max}^2}{2\sigma^2}\right) + \sqrt{\frac{\pi}{2}}\frac{x_{\max}}{\sigma} erfc\left(\frac{x_{\max}}{\sigma\sqrt{2}}\right)\right) \quad (17)$$

III. Numerical Results and Discussion

The performance testing of the FP24 quantizer is done in the theoretical domain, considering the range of data variances [-30dB, 30dB] with respect to the reference variance $\sigma_{ref}^2 = 1$ (data follows the Gaussian PDF).

Let's first look at the structure of the FP24 quantizer and discover which segment (specified by E^* value) occurs more frequently for Gaussian data. Fig. 2 shows P_{E^*} versus E^*, calculated using (8) for three specific values of σ[dB]=$20\cdot\log_{10}\sigma$, i.e. for σ[dB] = -30 dB, 0 dB and 30 dB. From Fig.2, we can see that P_{E^*} is different from zero only in a small range of E^* values. It can also be seen that the width of that range is the same for each considered σ[dB], while the range shifts to the right as σ[dB] increases.

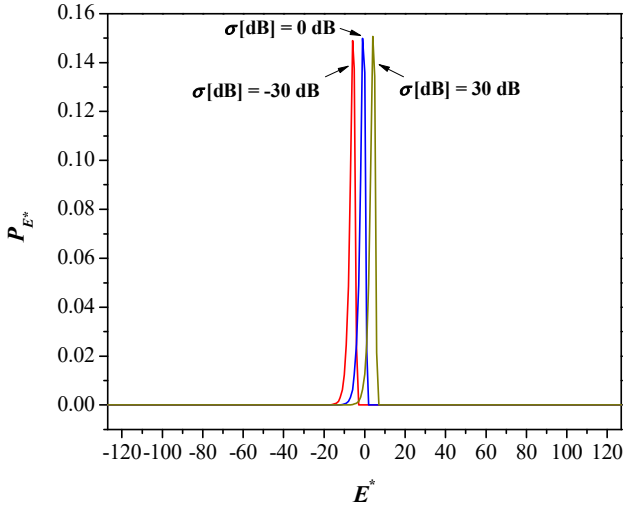

Fig. 2. Dependence of P_{E^*} on E^* for FP24 quantizer and Gaussian PDF.

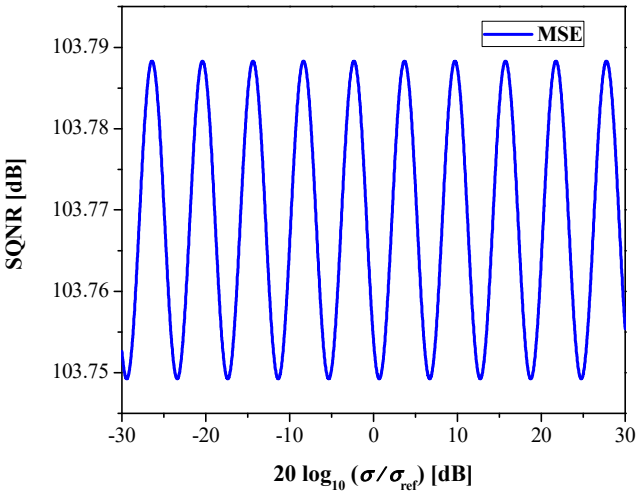

Fig. 3. Performance of the FP24 quantizer over a wide dynamic range of variances for MSE metric.

To compute the SQNR of the FP24 quantizer in the previously determined range of variances we used (11), and the results are presented in Fig. 3. Note that the achieved SQNR is high and changes negligibly in this range as the SQNR dynamics ΔSQNR (i.e. the difference between the maximal (SQNR_{max}) and minimal SQNR (SQNR_{min}) values) is only 0.04 dB; so, it follows that the FP24 quantizer is robust. Based on this fact, we report that, in the case of Gaussian data, change in variance causes a negligible impact on the accuracy of the FP24 format.

In Fig. 4 is provided the SQNR of the FP24 quantizer for the AE metric, that is determined according to (17). It can be noted that the results for this metric differ from those for the MSE metric presented in Fig. 3. From Fig. 4, the robustness of the FP24 quantizer can be easily ascertained, as SQNR preserves stability in the whole range (the SQNR dynamics is ΔSQNR = 0.016 dB). In this way, we reconfirmed that the performance of the FP24 format is independent on data variance.

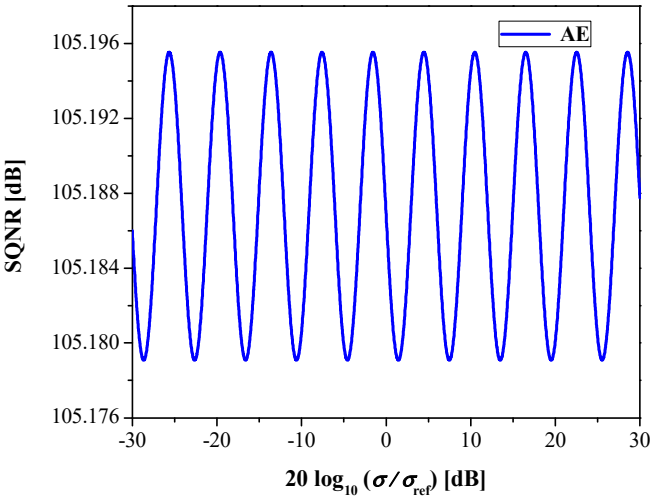

Fig. 4. Performance of the FP24 quantizer across a wide dynamic range of variances for AE metric.

TABLE I. THE OBTAINED RESULTS FOR FP24 FORMAT FOR DIFFERENT MEASURES AND GAUSSIAN SOURCE

Metric	$SQNR_{max}$ [dB]	$SQNR_{min}$ [dB]	ΔSQNR [dB]
MSE	103.7883	103.7492	0.0391
AE	105.1955	105.1791	0.0165

In Table I are provided performance details of the FP24 format for both observed measures.

To summarize, based on the SQNR analysis, a high efficiency of the FP24 quantizer in Gaussian data processing is observed. Hence, it is worth implementing FP24 format in applications where Gaussian data occurs and where the variance of the data is a variable parameter.

IV. CONCLUSION

In this paper, the analysis of the FP24 format in the presence of data described by the Gaussian distribution was provided. Specifically, the performance of this format was investigated through an equivalent quantization scheme (called the FP24 quantizer which is actually a piecewise uniform quantizer) and two different performance measures (MSE and AE), where adequate expressions were derived in both cases. It was shown that the FP24 quantizer is robust, as a high value of SQNR along with its negligible change in a wide dynamic range of data variances was observed for both MSE and AE metrics. This achievement makes the FP24 format highly efficient in representation of different Gaussian data, since the SQNR and the accuracy of the binary formats are directly related. Note that the robustness property is very important and can be a critical factor from the angle of practical implementation, especially in applications which are sensitive to changes in data accuracy and have large resources (memory, processing power, etc.) at disposal. Future work will include the implementaion of this format in neural networks, as neural network parameters (e.g., weights) can be statistically modelled by a Gaussian distribution.

ACKNOWLEDGMENT

This paper has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (grant number 451-03-65/2024-03/200102)

REFERENCES

[1] IEEE Standard for Floating-Point Arithmetic, 754, 2019.

[2] M. Junaid, S. Arslan, T. Lee and H. Kim, "Optimal architecture of floating-point arithmetic for neural network training processor," Sensors, vol. 22, Article ID:1230, 2022.

[3] A. Agrawal, S. M. Mueller, B. M. Fleischer, X. Sun, N. Wang, J. Choi, et al., "DLFloat: A 16-bfloating point format designed for deep learning training and inference," 26th IEEE Symp. on Computer Arithmetic (ARITH), pp. 92−95, 2019.

[4] N. Wang, J. Choi, D. Brand, C.Y. Chen and K. Gopalakrishnan, "Training deep neural networks with 8-bit floating point numbers," 32nd Conf. on Neural Information Processing Systems (NeurIPS 2018), pp. 1−11, 2018.

[5] N. Vučić, Z. Perić and A. Jovanović, "Model of improved floating point 32-bits quantizer," 57th Int. Sci. Conf. on Information, Communication and Energy Systems and Technologies (ICEST), 2022.

[6] M. Dinčić, Z. Perić, M. Savić, M. Milojković and N. Vučić, "SQNR analysis and classification accuracy of the 24-bit floating point representation of the Laplacian data source applied for quantization of weights of a multilayer perceptron," 15th Int. Conference SAUM, 2020.

[7] Z. Perić, B. Denić and M. Dinčić, "Improvement of the Bfloat16 Floating-point for the Laplacian Source," 13th International Symposium on Advanced Topics in Electrical Engineering (ATEE), 2023.

[8] Z. Perić, B. Denić, N. Vučić, A. Jovanović and J. Nikolić, "Performance analysis of fixed-point and floating-point 32-bit quantizers for L1 norm and Laplacian source," 58th Int. Sci. Conf. on Information, Communication and Energy Systems and Technologies (ICEST), 2023.

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

[10] A. Gersho and R. Gray, Vector Quantization and Signal Compression. New York, USA: Kluwer Academic Publishers, 1992.

[11] D. Marco and D. L. Neuhoff, "Low resolution scalar quantization for Gaussian sources and squared error," IEEE Trans. Inf. Theory, vol. 52, issue 4, pp. 1689-1697, 2006.

[12] D. Marco and D. L. Neuhoff, "Low-resolution scalar quantization for Gaussian sources and absolute error," IEEE Transactions on Information Theory, vol. 53, issue 3, pp. 1177-1179, 2007.

[13] R. W. Farebrother, L1-Norm and L∞- Estimation: An Introduction to the Least Absolute Residuals the Minimax Absolute Residual and Related Fitting Procedures. London, UK: Springer Science & Business Media, 2013.

[14] M. Dinčić, Z. Perić, B. Denić and B. Denić, "Optimization of the fixed-point representation of measurement data for intelligent measurement systems," Measurement, vol. 217, Article ID: 113037, 2023.