Data-Free Quantization of Neural Receivers: When 4-Bit Succeeds, Why 6-Bit Matters for 6G

SaiKrishna Saketh Yellapragada , Esa Ollila
Aalto University
Espoo, Finland
{saikrishna.yellapragada, esa.ollila}@aalto.fi

Mário Costa
Nokia Technologies
Amadora, Portugal
mario.costa@nokia.com

Abstract

As wireless systems evolve toward 6G, Artificial Intelligence (AI) and Deep Learning (DL) are poised to revolutionize physical-layer processing, offering superior performance over classical methods in throughput and Block Error Rate (BLER). Deploying DL-based receivers in resource-constrained environments requires balancing performance with inference latency, energy consumption, and computational overhead. We study data-free Post-Training Quantization (PTQ) of a neural receiver that processes frequency-domain baseband samples to generate Log-Likelihood Ratios (LLRs) for error-control decoding. Quantization parameters are derived directly from pretrained weights via symmetric per-channel uniform quantization, where each channel's scale captures the absolute-weight range—requiring no calibration data, synthetic data, or activation statistics. We reduce `float32` weights to `8`-, `6`-, and `4`-bit and evaluate radio performance across 3GPP Line-of-Sight (LoS)/Non-LoS (NLoS) channels and mobility scenarios. In NLoS, `8`- and `6`-bit achieve near-`float32` BLER, with gains up to 4.9 dB over baseline Least-Squares (LS) under high mobility. In LoS, `4`-bit remains robust, surpassing traditional receivers by 1.7–2.6 dB across mobilities, while yielding an $8\times$ smaller model. These findings inform hardware–software co-design for AI-native 6G air-interfaces, highlighting low-precision quantization as a key enabler for efficient edge, sensing, and cloud-radio deployments.

1 Introduction

Cellular systems have been shaped by transformative technologies. Orthogonal Frequency Division Multiplexing (OFDM) revolutionized 4G, while massive Multiple-Input-Multiple-Output (MIMO) became the cornerstone of 5G. As AI and DL achieve exceptional performance across domains from computer vision to natural language processing, an important question arises: how will these innovations shape 6G wireless communication systems? For decades, wireless communication challenges were addressed through statistical modeling, leading to linear methods in the physical-layer such as LS and Linear Minimum Mean Squared Error (LMMSE) channel estimation, linear equalizers, and Zero-Forcing (ZF) precoding, which have offered low complexity and near-optimal performance in 3G, 4G, and 5G. The emerging 6G environment challenges conventional frameworks, as recent work shows that DL applied to the physical layer outperforms model-based transceivers in throughput and BLER. To manage growing device and service heterogeneity, the 6G physical layer must prioritize simplicity and efficiency, necessitating an AI-native air interface to enhance radio access network functionalities [1]. Floating-point inference for DL models is computationally expensive and power-intensive, limiting its practicality on edge devices and in large-scale cloud-Radio Access Network (RAN) deployments with strict latency budgets, regardless of the hardware Central Processing Units (CPUs), Graphics Processing Units (GPUs), or Neural Processing Units (NPUs).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

Reducing inference complexity while preserving model performance is therefore essential to make deep learning receivers viable for operational 6G networks.

Two main approaches address this challenge: PTQ and Quantization-Aware Training (QAT). PTQ reduces the precision of a pre-trained floating-point model from 32-bit to lower-precision formats such as 16-bit floating-point or 8- and 4-bit integers without retraining.

2 Background and related work

Quantization This section provides an overview of PTQ and its relevance to the quantization of weight kernels in neural network receivers to be presented in Section 3.

Integer quantization centers on two fundamental operations that enable efficient model compression [2, 3, 4]. The first operation, quantization, transforms floating-point values into lower-precision integer representations such as 8-bit or 4-bit. The second operation, dequantization, converts these integer values back to approximate floating-point representations during inference.

The mapping between floating-point values and integer representations gives rise to two predominant quantization strategies: asymmetric and symmetric quantization. Asymmetric quantization operates through three critical parameters that govern the transformation process: the scale factor (s), the zero-point offset (z), and the bit-width (b). Symmetric quantization operates on the quantization grid by maintaining symmetry around the zero-point with $z = 0$. The mathematical formulation of the quantization function is given by:

$$\widehat{\mathbf{x}} := q\left(\mathbf{x};\, s, z, b\right) = s \cdot \left(\text{clip}\left(\left\lfloor \frac{\mathbf{x}}{s} \right\rceil + z; 0, 2^b - 1\right) - z\right), \tag{1}$$

where $\mathbf{x}$ denotes the quantizer input (i.e., network weights or activations). Moreover, $s \in \mathbb{R}_+$ denotes the scale factor, or step-size, $z \in \mathbb{Z}$ the zero point, and $b \in \mathbb{N}$ the bit-width. Finally, $\lfloor \cdot \rceil$ denotes the round-to-nearest-integer operator.

In this work, we consider symmetric quantization when performing per-channel PTQ [5, 6, 7, 8, 9, 10, 11]. We adopt symmetric per-channel quantization for two reasons: (i) **per-channel** scaling assigns each convolutional filter its own scale factor, preserving fine-grained dynamic-range information critical for resource-grid convolutions; (ii) **symmetric** quantization ($z = 0$) simplifies hardware by eliminating zero-point arithmetic, reducing latency and memory overhead on `8-bit/4-bit` accelerators. The scale factor s_c for channel c is computed as

$$s_c = \frac{\max(|\mathbf{w}_c|) - \min(|\mathbf{w}_c|)}{2^{b-1} - 1},$$

where $\mathbf{w}_c$ denotes the channel weights. This data-free approach derives quantization parameters solely from pre-trained floating-point model, eliminating calibration datasets or activation statistics.

Although there has been abundant research in quantization and efficient deep learning inference for computer vision and language modeling, its application to communication systems remains limited. In this paper, we focus on how quantization impacts radio performance when eural network receivers perform the majority of physical layer signal processing functions in an end-to-end learning fashion.

End-to-End Learning DL for the physical layer of wireless communication systems has been the focus of several scientific contributions [12, 13, 14]. Therein, it is shown that DL based receivers outperform model based receivers in terms of BLER and throughput. It has also been shown that uniform linear quantization of a Single-Input-Single-Output (SISO) OFDM neural network-based receiver reduces Floating Point Operations (FLOPs) by 50% with only a 0.25 dB degradation in radio performance [15]. Furthermore, 8-bit based PTQ of a DL receiver has been shown to perform similarly to that of a `float32` architecture [16]. In this paper, we investigate the impact of low-bit-width quantization to determine the practical limits of PTQ for DL-based physical-layer receivers in 6G systems.

- Evaluation of symmetric per-channel PTQ at `6-bit` and `4-bit` bit-widths, revealing trade-offs not explored in [16], such as `4-bit`'s enhanced robustness in LoS scenarios.
- Comprehensive analysis across both LoS and NLoS channel models. We consider NLoS Clustered Delay Line (CDL)-B for testing which usually models environments with more

clusters and a richer multipath structure than CDL-A and CDL-C. While CDL-D was used for LoS testing of the quantized neural receivers.

- We train the model using two distinct training scenarios, one is a hybrid of LoS and NLoS channels and another focusing solely on NLoS to investigate generalization and robustness to quantization.

3 System Model

We consider an uplink Single-Input-Multiple-Output (SIMO) OFDM wireless communication system with $N_{\text{Rx}} = 2$ receive antennas. The choice of N_{Rx} stems from isolating quantization effects from large-array combining gains.

At the transmitter, the input bit sequence is encoded with a Low-Density Parity-Check (LDPC) code. The encoded bits are mapped to complex modulation symbols and arranged into a Resource Grid (RG) of size $N_{\text{sym}} \times N_{\text{sc}}$, where N_{sym} is the number of OFDM symbols and N_{sc} is the number of subcarriers. Demodulation Reference Signals (DMRSs) are embedded at known time–frequency locations to facilitate channel estimation at the receiver. The grid is converted to a time-domain OFDM waveform via Inverse Fast Fourier Transform (IFFT) and cyclic prefix, and transmitted through a 3GPP CDL channel [17].

After Fast Fourier Transform (FFT) at the receiver, the system model is defined as follows

$$\mathbf{y}_{n,k} = \mathbf{h}_{n,k}\, x_{n,k} + \mathbf{n}_{n,k}, \qquad \mathbf{y}_{n,k} \in \mathbb{C}^{N_{\text{Rx}} \times 1}, \;\; \mathbf{h}_{n,k} \in \mathbb{C}^{N_{\text{Rx}} \times 1}, \tag{2}$$

where $x_{n,k}$ is the transmitted symbol at OFDM symbol n and subcarrier k. The noise vector satisfies $\mathbf{n}_{n,k} \sim \mathcal{CN}\big(\mathbf{0},\, \sigma^2 \mathbf{I}_{N_{\text{Rx}}}\big)$ with i.i.d. entries, and symbols are normalized as $\mathbb{E}[|x_{n,k}|^2] = 1$. DMRSs at pilot positions (n, k) enable channel estimation via nearest neighbor interpolation of $\mathbf{h}_{n,k}$ across the RG.

4 Neural Receiver: Training and Post-training Quantization

Traditionally, when the receiver is processing the OFDM waveform, after performing FFT, the post-FFT waveform is fed to perform channel estimation, equalization, and demapping. We consider a neural receiver architecture as shown in Table 1. This is designed to replace traditional signal processing operations where the input is the post-FFT sequence and the output is LLRs. The output of the network, which are the LLRs, are then used as input to LDPC decoding.

Layer	Channels	Kernel Size	Dilatation Rate
Input Conv2D	128	(3,3)	(1,1)
ResNet Blocks (1–8)	128	(3,3)	(1,1)
Output Conv2D	6	(3,3)	(1,1)

Table 1: Architecture Details of the Neural Receiver

Parameters	**Value**	**Randomization**
Carrier Frequency	3.5 GHz	*None*
RMS Delay Spread	10 ns – 100 ns	Uniform
UE Velocity	0 m/s – 50 m/s	Uniform
SNR	0 dB – 15 dB	Uniform
Subcarrier Spacing	30 kHz	*None*
Modulation Scheme	64-QAM	*None*
Code Rate	0.5	*None*
DMRS Configuration	3$^{\text{rd}}$ and 12$^{\text{th}}$ Symbol	*None*
Optimizer	Adam	*None*
Batch Size	128	*None*

Table 2: Training and Evaluation Parameters

We create two neural receiver models, where each model is trained on a combination of 3GPP channel models. The neural receiver and our experiments are performed using Sionna [18]:

- In Scenario I, we train the neural receiver with the parameters mentioned in Table 2 by using channel models CDL-A, C and E. This training scenario is a combination of NLoS and LoS models, while the testing is performed using CDL-B and CDL-D.
- In Scenario II, we train the neural receiver with the parameters using channel models CDL-A, B, and C. This training scenario is trained only on NLoS channels, and is tested on LoS channels CDL-D and CDL-E.

The neural receiver undergoes training with the objective of maximizing the bit-metric decoding performance. To achieve this optimization goal, Binary Cross-Entropy (BCE) loss serves as the cost function, quantifying the difference between the network's predicted LLRs and the ground truth transmitted coded bits distributed across the complete OFDM resource grid. The BCE loss-function is mathematically formulated as:

$$\mathcal{L}_{\text{BCE}}(B, \hat{L}) = -\mathbb{E}\left[B \log \sigma(\hat{L}) + (1 - B) \log(1 - \sigma(\hat{L}))\right], \tag{3}$$

where B denotes the ground-truth transmitted bits, $\hat{L}$ denotes the predicted LLRs and $\mathbb{E}[\cdot]$ expectation operator. The function $\sigma(\hat{L})$ represents the sigmoid activation, which maps LLRs to probabilities.

5 Experiments

In this section, we evaluate the neural receiver under various scenarios as mentioned in Section 4. The trained `float32` neural receiver models were post-training quantized to low bit-width representations (`8-bit`, `6-bit`, and `4-bit`) before being used for inference to analyze the correspondig radio performance. Model training was performed on devices equipped with NVIDIA A40 GPU.

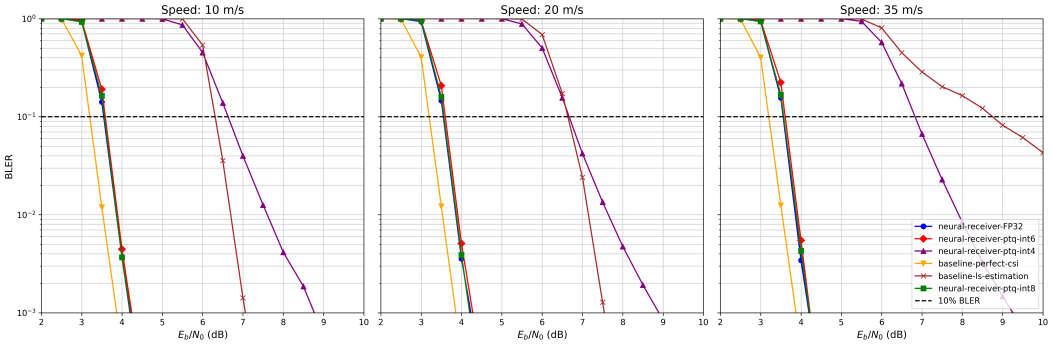

Figure 1: Performance Analysis for Scenario I on CDL-D in terms of BLER vs E_b/N_0

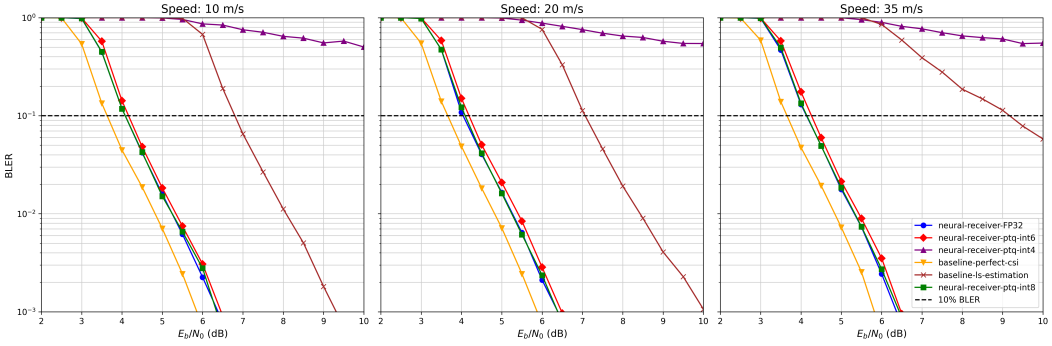

Figure 2: Performance Analysis for Scenario I on CDL-B in terms of BLER vs E_b/N_0

The BLER performance was evaluated for several receiver configurations, including a neural receiver `float32`, PTQ'd neural receivers `8-bit, 6-bit, 4-bit`, as well as baseline methods employing

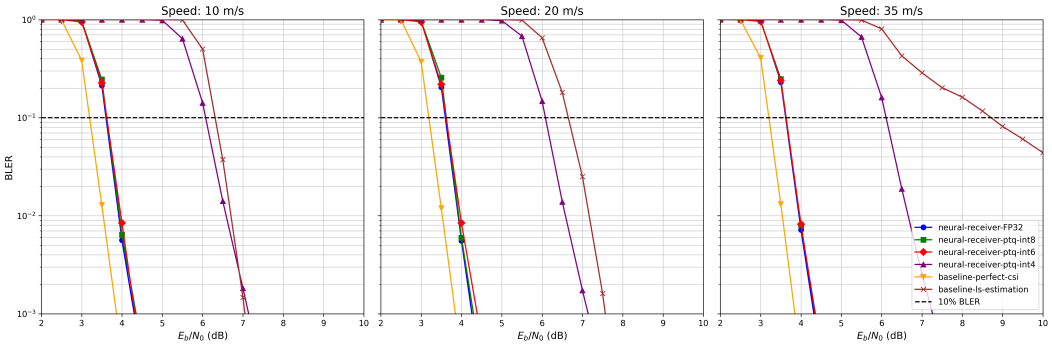

Figure 3: Performance Analysis for Scenario II on CDL-D in terms of BLER vs E_b/N_0

perfect Channel State Information (CSI) and LS channel estimation. LS channel estimation is performed with nearest neighbour interpolation followed by LMMSE equalization. Experiments were conducted under three mobility scenarios: low speed (10 m/s), medium speed (20 m/s), and high speed (35 m/s).

In Figure 1, the neural receiver and its `8-bit/6-bit` quantized models consistently achieved low BLER across the entire Signal-to-Noise-Ratio (SNR) range, demonstrating robustness to both quantization and increased User Equipment (UE) speed. In contrast, the neural receiver quantized to `4-bit` exhibits noticeable performance degradation compared to higher-precision neural receivers but consistently outperforms the baseline LS estimation method in high-mobility scenarios. At higher speeds (e.g., 35 m/s), the `4-bit` neural receiver achieves significantly lower BLER with up to a gain of 1.7 dB compared to the LS receiver.

Figure 2 illustrates that `4-bit` quantization of the neural receiver leads to performance degradation on a CDL-B NLoS channel with 10% BLER as a target metric. The CDL-B channel may represent urban macrocell environments with large delay spreads. The BLER remains high across the entire SNR range at various mobility conditions. In this case, LS estimation achieved significantly lower BLER at moderate and high SNRs, outperforming the `4-bit` neural receiver and highlighting its limitations under challenging channel conditions. These findings suggest that `4-bit` symmetric uniform per-channel quantization is not suitable for deployment in such environments, whereas `8-bit` and `6-bit` variants are seen as a good compromise that retain robust performance and consistently outperform LS estimation by 4.7 dB. Specifically, being within 0.1 dB at 10% BLER makes the `6-bit` a desirable choice.

Finally, Figure 3 shows that in Scenario II, the `4-bit` quantized neural receiver surpasses LS estimation by approximately 2.5 dB under high UE speed, highlighting its robustness in LoS conditions. Additionally, across the entire SNR range and mobility conditions, the `8-bit` and `6-bit` models closely match the `float32` baseline, indicating that `6-bit` quantization offers an effective trade-off between computational efficiency and 10% BLER performance.

6 Discussion

Benefits of Quantized Neural Receivers The presented study demonstrates the trade-offs of adopting sub-8-bit representations in neural receiver architectures by applying PTQ with `6-bit` quantization to neural receivers, and evaluating the performance across diverse scenarios. Figure 4 illustrates the E_b/N_0 required to achieve 10% BLER at high mobility (35 m/s) for Scenario I (CDL-B, NLoS) and Scenario II (CDL-D, LoS) . The `6-bit` quantized receiver achieves near-`float32` performance, requiring 4.26 dB (vs. 4.13 dB for `float32`) on CDL-B and 3.63 dB (vs. 3.62 dB) on CDL-D, with a $5.33\times$ model size reduction. This robustness, along with gains of 4.92 dB over LS estimation in high-mobility NLoS conditions (CDL-B) and 5.09 dB in LoS (CDL-D), positions `6-bit` as an excellent choice for 6G's AI-native air interfaces, balancing efficiency and radio performance [1].

Furthermore, `4-bit` quantization significantly enhances LoS performance, requiring only 6.12 dB on CDL-D (compared with 8.72 dB for LS, thus a 2.60 dB gain) in Scenario II, surpassing LS by up to 2.5

dB in high-mobility LoS conditions. This enables ultra-low-power inference with an $8\times$ model size reduction, facilitating latency and energy demands compared to `8-bit`, making `4-bit` a compelling choice for site-specific 6G deployments like fixed wireless access or Vehicle-to-Everything (V2X) use-cases, where LoS channels predominate.

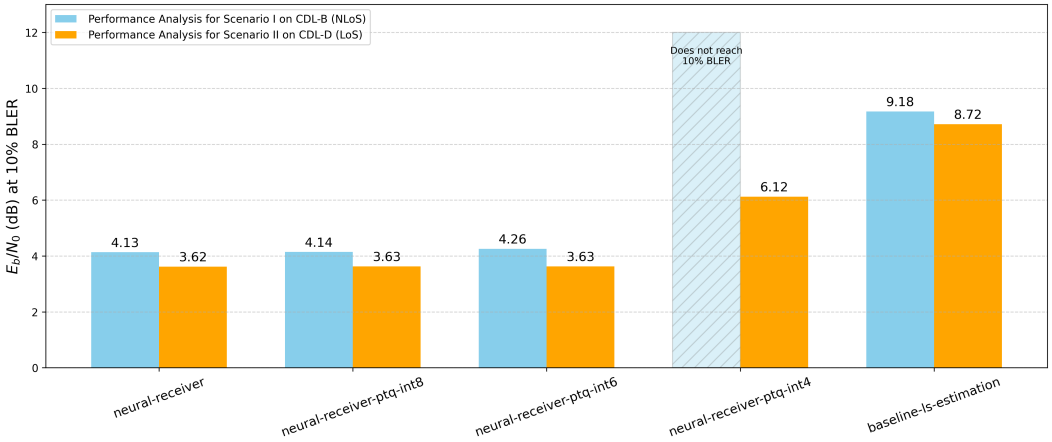

Figure 4: E_b/N_0 required to reach 10% BLER at high mobility (35 m/s) for various receiver architectures

Limitations The significant performance gap between `4-bit` and `8-bit` quantization, particularly in NLoS channels, e.g., CDL-B, where `4-bit` fails to achieve 10% BLER while `8-bit` requires 4.14 dB, indicates that `4-bit` quantization-induced errors outweigh the effects of challenging channel conditions. This highlights the need for advanced PTQ techniques, such as adaptive rounding, to mitigate these errors [11] and QAT.

While `6-bit` weight quantization achieves performance comparable to `8-bit` and `float32`, and offers gains in model compression, hardware efficiency as well as memory access, its adoption is limited by the lack of native hardware support. Current GPUs and accelerators mainly support `float16`, `8-bit`, and `4-bit`, leaving `6-bit` operations reliant on inefficient emulation. Thus, despite promising results, practical deployment of `6-bit` remains uncertain.

7 Conclusions

We trained multiple neural receivers across diverse channel profiles and applied symmetric per-channel PTQ with varying bit-widths. The `4-bit` neural receiver emerges as a viable low-power alternative for LoS deployments, outperforming `float32`-based LS receivers at medium to high UE velocities, particularly in site-specific 6G scenarios. However, in NLoS settings, the `4-bit` receiver fails to achieve 10% BLER. Our analysis can guide hardware–software co-design toward accelerators supporting `6-bit` and `4-bit`, with results highlighting the potential of dynamic precision switching between `6-bit`/`8-bit` and `4-bit` to balance power efficiency and reliability. The robust LoS performance of `4-bit` also makes it a promising candidate for integrated communications and sensing, a cornerstone of 6G. Future work can explore emerging `FP6` and `FP4` formats to combine efficiency with competitive performance, while native `6-bit` support in accelerators and mixed-precision PTQ remain promising directions for on-device optimization.

Acknowledgements

The authors thank Sebastian Cammerer from NVIDIA for his insights during the inception of this work. This work has been supported in parts by Research Council of Finland (grant no:359848) and by the 6GARROW project which has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101192194 and from the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (No. RS-2024-00435652).

References

[1] Jakob Hoydis, Fayçal Ait Aoudia, Alvaro Valcarce, and Harish Viswanathan. Toward a 6G AI-native air interface. *IEEE Communications Magazine*, 59(5):76–81, 2021. doi: 10.1109/MCOM.001.2001187.

[2] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Blankevoort Tijmen. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

[3] Andrey Kuzmin, Markus Nagel, Mart Van Baalen, Arash Behboodi, and Tijmen Blankevoort. Pruning vs quantization: Which is better? *Advances in neural information processing systems*, 36:62414–62427, 2023.

[4] Hao Wu, Patrick Judd, Xiaoxia Zhang, Mikhail Isaev, and Paul Micikevicius. Integer quantization for deep learning inference: Principles and empirical evaluation. *arXiv preprint arXiv:2004.09602*, 2020.

[5] Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Post-training 4-bit quantization of convolution networks for rapid-deployment. *arXiv preprint arXiv:1810.05723*, 2018.

[6] Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *ICCV Workshops*, pages 3009–3018, 2019.

[7] Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Improving post training neural quantization: Layer-wise calibration and integer programming. *arXiv preprint arXiv:2006.10518*, 2020.

[8] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.

[9] Eldad Meller, Alexander Finkelstein, Uri Almog, and Mark Grobman. Same, same but different: Recovering neural network quantization error through weight factorization. In *International Conference on Machine Learning*, pages 4486–4495. PMLR, 2019.

[10] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1325–1334, 2019.

[11] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *International Conference on Machine Learning (ICML)*, 2020.

[12] Mikko Honkala, Dani Korpi, and Janne M. J. Huttunen. DeepRx: Fully convolutional deep learning receiver. *IEEE Transactions on Wireless Communications*, 20(6):3925–3940, 2021. doi: 10.1109/TWC.2021.3054520.

[13] Fayçal Ait Aoudia and Jakob Hoydis. End-to-end learning for ofdm: From neural receivers to pilotless communication. *IEEE Transactions on Wireless Communications*, 21(2):1049–1063, 2022. doi: 10.1109/TWC.2021.3101364.

[14] K. Pavan Srinath and Jakob Hoydis. Bit-metric decoding rate in multi-user MIMO systems: Theory. *IEEE Transactions on Wireless Communications*, 22(11):7961–7974, 2023. doi: 10.1109/TWC.2023.3257196.

[15] Moritz Benedikt Fischer, Sebastian Dörner, Takayuki Shimizu, Chinmay Mahabal, Hongsheng Lu, and Stephan ten Brink. On the implementation of neural network-based OFDM receivers. In *2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, pages 1–6, 2024. doi: 10.1109/VTC2024-Spring62846.2024.10683574.

[16] SaiKrishna Saketh Yellapragada, Esa Ollila, and Mario Costa. Efficient deep neural receiver with post-training quantization. *arXiv preprint arXiv:2508.06275*, 2025.

[17] 3GPP. Study on channel model for frequencies from 0.5 to 100 GHz. Technical Report TR 38.901, 3rd Generation Partnership Project (3GPP), 2020. version 16.1.0.

[18] Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Merlin Nimier-David, Lorenzo Maggi, Guillermo Marcus, Avinash Vem, and Alexander Keller. Sionna, 2022. https://nvlabs.github.io/sionna/.