

# Deep Q-Network to Reduce PAPR in Communication Systems

Bianca S. de C. da Silva<sup>✉</sup>, Anderson R. R. Marins<sup>✉</sup>, Felipe A. P. de Figueiredo<sup>✉</sup>, and Luciano L. Mendes<sup>✉</sup>, *Member, IEEE*

**Abstract**—This paper presents a novel Reinforcement Learning (RL) strategy for reducing the Peak-to-Average Power Ratio (PAPR) in Orthogonal Frequency Division Multiplexing (OFDM) systems using only two pilot subcarriers and a single Inverse Fast Fourier Transform (IFFT) operation during inference. Unlike conventional techniques that require exhaustive searches, Partial Transmit Sequence (PTS), or fixed transforms, Discrete Fourier Transform (DFT), the proposed Deep Q-Network (DQN) learns an adaptive pilot mapping policy with negligible overhead. Simulation results show that the method reduces the PAPR of the original OFDM signal from 9.2 dB to 8.3 dB, surpassing DFT, and approaching the performance of Q-Learning, while maintaining a fraction of the complexity. Although PTS achieves lower PAPR as 4.5 dB, its computational cost is substantially higher, making it unsuitable for real-time deployment. The proposed DQN therefore offers a practical, low-complexity, learning-based alternative that provides meaningful PAPR reduction without the computational burden of traditional schemes.

**Index Terms**—BER, Machine Learning, OFDM, PAPR, Q-Learning, Reinforcement Learning.

## I. INTRODUCTION

THE rapid evolution of mobile communication systems, driven by the increased demand for high-speed services such as video streaming, videoconferencing, and real-time applications, has led to a significant increase in the number of connected users. This scenario, although positive from a connectivity perspective, poses a critical problem: the scarcity of available spectrum for accommodation of growing traffic, especially in the frequency bands traditionally used for mobile communications. The fixed and often inefficient allocation of these frequency bands contributes to the under-utilization of certain segments of the spectrum, further aggravating the difficulty of expanding network capacity [1].

To mitigate these constraints, regulatory agencies in several countries have encouraged the use of underutilized portions of the spectrum, known as TV White Space (TVWS). In Brazil, Anatel distributes guidelines that authorize the use of unoccupied television channels in the Very High Frequency (VHF) and Ultra High Frequency (UHF) bands [2] by secondary user systems. This not only aims to alleviate pressure on the traditional strategy but also to promote digital inclusion in less-served regions, such as rural areas and remote communities, taking advantage of the good propagation of these bands to expand Internet access. However, to ensure

non-interference with licensed primary services, strict limit standards are established, such as a maximum transmission power of 1 W for secondary devices operating in TVWS [3].

Although this approach offers advantages in terms of geographic coverage and spectral efficiency, power limitation poses a significant challenge to system performance, especially when using multi-carrier waveforms such as OFDM. Despite its recognized advantages, such as resistance to selective fading and high spectral efficiency, OFDM is characterized by a high PAPR, which may require highly linear power amplifiers. In practice, this requirement leads to the amplifier operating outside its ideal linearity zone, introducing distortions in the form of Inter-carrier Interference (ICI) and increasing the Out-of-Band Emissions (OOBE) level. These effects not only degrade the quality of the received signal but can also cause interference in adjacent systems [4], compromising harmonious coexistence in the spectrum.

PAPR reduction is also relevant in several other scenarios, such as spectrum sharing between heterogeneous mobile technologies, coexistence in unlicensed bands (Long-Term Evolution (LTE) and New Radio (NR)), compatibility with ISM bands, and satellite communication systems, where the cost and linearity of power amplifiers play a central role [5].

Several classical techniques have been proposed to mitigate PAPR in OFDM systems, including clipping and filtering, Selected Mapping (SLM), and PTS. However, many of these approaches suffer from high computational complexity, significant signaling overhead, or performance degradation [6, 7, 8]. In response to these limitations, modern approaches based on Machine Learning (ML), particularly Neural Network (NN), have gained attention as they can perform complex non-linear optimizations at reduced computational cost. Notable strategies include autoencoder-based PAPR suppression, deep unfolding, and adaptive modulation schemes [9, 10].

Recent advances include supervised neural networks trained using Monte Carlo Search-based Augmentation Memory-less Continuous Search Algorithm (MCSA), achieving significant PAPR reduction compared to conventional methods such as DFT spreading [11]. Although promising, these approaches rely on large datasets and extensive training procedures.

In this context, this work aims to reduce the PAPR of OFDM signals without relying on large training datasets or computationally intensive search procedures. Instead, we leverage RL to learn an adaptive pilot mapping policy that requires only two pilot subcarriers and a single IFFT operation during inference. This strategy provides a lightweight, scalable, and data efficient alternative to conventional PAPR

B. S. de C. da Silva, Anderson R. Reis, Felipe A. P. de Figueiredo and L. L. Mendes are from the National Institute of Telecommunications - INATEL, Santa Rita do Sapucaí, Brazil, e-mail: (bianca.silva@mtel.inatel.br, anderson.rufino@dtel.inatel.br, felipe.figueiredo@inatel.br, luciano@inatel.br).

reduction techniques.

The structure of this paper is organized as follows: Section II presents the background of the OFDM system, discussing its characteristics and the challenges related to PAPR. Section III describes the traditional methods to reduce PAPR in literature. Then, Section IV describes the principles of the proposed RL. Section V presents the simulations results, focusing on the reduction of PAPR, the Bit Error Rate (BER) and computational complexity. Finally, Section VI summarizes the main findings and specific points for future developments in this line of research.

## II. BACKGROUND

In this section, we present a concise overview of OFDM and discuss the PAPR problem inherent to these systems.

### A. Overview of OFDM Systems

At the OFDM transmitter, the incoming bit sequence  $\mathbf{b} = [b_0, b_1, \dots, b_{\mu K-1}]$ , where  $\mathbf{b} \in \mathbb{R}^{(\mu K) \times 1}$ , is converted into a complex data vector  $\mathbf{s} = [s_0, s_1, \dots, s_{K-1}]$ ,  $\mathbf{s} \in \mathbb{C}^{K \times 1}$ . This mapping is performed according to an in-phase/quadrature modulation scheme—most commonly Quadrature Amplitude Modulation (QAM). Here,  $\mu = \log_2 \mathcal{M}$  denotes the number of bits carried by each symbol,  $\mathcal{M}$  is the constellation size, and  $K$  corresponds to the number of orthogonal subcarriers. Every entry of  $\mathbf{s}$  thus encapsulates  $\mu$  information bits and modulates one specific subcarrier within the OFDM grid.

Once generated, the symbol vector is forwarded to the IFFT block, which transforms the frequency-domain representation of the  $\mathcal{M}$ -QAM symbols into a time-domain OFDM waveform. This operation is expressed as

$$\mathbf{x} = \mathbf{F}_K^H \mathbf{s}, \quad (1)$$

where  $(\cdot)^H$  denotes the Hermitian transpose and  $\mathbf{F}_K \in \mathbb{C}^{K \times K}$  is the normalized Fourier matrix.

A Cyclic Prefix (CP) of length  $K_{\text{CP}}$  samples is appended to the vector  $\mathbf{x}$ , producing the extended signal  $\bar{\mathbf{x}} \in \mathbb{C}^{(K+K_{\text{CP}}) \times 1}$ . The extended signal  $\bar{\mathbf{x}}$  is then propagated through a channel characterized by an impulse response  $\mathbf{h} \in \mathbb{C}^{N_t \times 1}$ , where  $\iota$  denotes the channel impulse response length. To recover the transmitted symbols, it is necessary to remove the CP. This means that the channel action can be expressed by the circular convolution between the OFDM symbol without CP with the channel impulse response [12].

Applying frequency-domain equalization under the zero-forcing criterion, the resulting estimate is

$$\hat{\mathbf{y}} = \mathcal{H}^{-1} \bar{\mathbf{y}} = \mathbf{s} + \mathcal{H}^{-1} \mathbf{w}, \quad (2)$$

where  $\mathbf{w} \in \mathbb{C}^{K \times 1}$  represents additive white noise in the frequency domain. The matrix  $\mathcal{H} = \mathbf{F}_K \mathbf{H} \mathbf{F}_K^H$ , with  $\mathcal{H} \in \mathbb{C}^{K \times K}$ , is diagonal and contains on its main diagonal the channel's frequency response [13]. After equalization, the symbols are demodulated to recover the estimated bit vector  $\hat{\mathbf{b}}$ .

A persistent issue in OFDM systems is their inherently high PAPR, as previously noted. The following section discusses this aspect in detail.

### B. Characterization of PAPR in OFDM Transmissions

The PAPR, denoted by  $\mathcal{P}$ , is defined as the ratio between the maximum instantaneous power of a signal and its average power:

$$\mathcal{P} = \frac{\max(|\mathbf{x}|^2)}{\mathbb{E}[|\mathbf{x}|^2]}, \quad (3)$$

where  $\mathbb{E}[\cdot]$  denotes the expectation operator.

In OFDM systems, the transmitted waveform is formed by the superposition of  $K$  independently modulated subcarriers, each carrying a symbol drawn from an  $\mathcal{M}$ -QAM constellation. This summation produces a time-domain envelope that is well approximated by a Gaussian random variable. As a consequence, there are instances where multiple subcarrier components combine constructively, generating large instantaneous amplitudes. Such constructive additions lead to pronounced peaks relative to the signal's Root Mean Square (RMS) level, which ultimately manifests as a high PAPR.

Large power peaks may force the transmitter's power amplifier to operate in its saturation region. In this regime, the amplifier clips the waveform and its input–output relationship ceases to be linear, thereby introducing nonlinear distortions into the transmitted signal. These distortions generate ICI and increase interference in neighboring channels due to the rise in OOB. Consequently, controlling and reducing the PAPR becomes essential in multicarrier modulation schemes.

The following section describes the method proposed for PAPR reduction.

## III. TRADITIONAL METHODS

Several techniques have been proposed in the literature to mitigate the high PAPR in multicarrier modulation systems such as OFDM. Among the most well-known are Clipping and Filtering, SLM, PTS, Tone Reservation (TR), and Active Constellation Extension (ACE) [14, 15, 16, 17]. Each of these methods provides a trade-off between PAPR reduction capability, implementation complexity, and impact on system performance metrics such as BER and spectral efficiency.

In this work, we focus on two representative traditional methods: the DFT approach and the PTS technique [18, 19]. These methods were chosen due to their wide adoption, theoretical relevance, and their complementary nature, the former emphasizing signal orthogonality and transformation properties, and the latter leveraging phase optimization for PAPR minimization. The following subsections present an overview of each method and their main operating principles.

### A. Discrete Fourier Transform

The DFT approach is one of the most established strategies for lowering PAPR in multicarrier systems [20]. The method introduces an additional Fast Fourier Transform (FFT) stage applied independently to each of the  $D$  sub-blocks of the modulated symbol vector before the remaining steps of the traditional OFDM transmitter. By performing this transform, the symbol energy becomes more uniformly distributed across

the available subcarriers, which reduces fluctuations in the instantaneous amplitude and, consequently, decreases the PAPR.

At the receiver, the operations applied at the transmitter are reversed to fully restore the original symbol structure. The received signal first undergoes the standard FFT of the OFDM demodulation stage. Subsequently, an IFFT is applied to each sub-block, effectively inverting the FFT based spreading performed during transmission. This step ensures that the data symbols return to the appropriate time-domain representation prior to demodulation, thereby removing the impact of the DFT and preventing any distortion of the transmitted information [21].

### B. Partial Transmit Sequence

In the PTS scheme, the original data vector  $\mathbf{s}$  is divided into  $D$  distinct subblocks, each containing a contiguous portion of the input samples. Every subblock is constructed so that its active entries occupy a unique segment of  $\mathbf{s}$ , while all other positions are filled with zeros. For instance, the first subblock  $\mathbf{s}_1$  contains the initial  $K/D$  symbols of  $\mathbf{s}$ , with zeros appended in all remaining positions. The second subblock  $\mathbf{s}_2$  includes the subsequent  $K/D$  symbols, but with zero padding before and after this segment, and this pattern continues for all  $D$  partitions. This construction guarantees that the subblocks are mutually exclusive and non-overlapping.

Formally, the partitioning can be written as

$$\begin{aligned} s_m[i] &= s_i, & (m-1)\frac{K}{D} < i \leq m\frac{K}{D}; \\ s_m[j] &= 0, & j \neq i, \end{aligned} \quad (4)$$

where  $s_m[i]$  denotes the  $i$ -th entry of subblock  $\mathbf{s}_m$ , with  $m \in 1, \dots, D$  and  $K, D \in 2^v$  for  $v \in 1, 2, 3, \dots$  [22].

Each subblock is individually processed by an IFFT, producing time-domain vectors  $\mathbf{z}_m = [z_{m,1}, z_{m,2}, \dots, z_{m,K}]^T$ . These vectors are subsequently combined in a controlled manner so as to reduce the likelihood that their components add up coherently, which would otherwise generate large amplitude peaks. The goal of the method is therefore to select a set of phase factors that yields the smallest possible PAPR. The resulting transmitted signal can be expressed as [23]

$$\mathbf{z} = \sum_{m=1}^M p_m \mathbf{z}_m, \quad (5)$$

where each phase coefficient is typically defined by [23]

$$p_m = e^{j \frac{2\pi q_m}{u}}, \quad (6)$$

with  $0 \leq q_m \leq u-1$  and  $u$  denoting the number of allowed phase values. A block diagram illustrating the PTS-based OFDM transmitter can be found in [23].

At the receiver, the decoding stage mirrors the encoding procedure. Using the same phase factors employed during transmission, the receiver recovers each subblock, which are then recombined to reconstruct the original OFDM symbol.

## IV. PRINCIPLES OF REINFORCEMENT LEARNING

This section introduces the fundamental concepts of RL, focusing on the Q-learning algorithm and its deep extension, the DQN. First, an overview of both methods is presented to outline their main principles. Then, the parameters and configurations used in this work are described to show how each algorithm was implemented in the proposed scenario.

Before detailing the DQN, it is important to briefly introduce the concept of Q-learning, which serves as its foundation and will be used for comparison throughout this work.

### A. Q-Learning Overview

1) *In General:* Q-learning is a value-based RL algorithm that aims to learn an optimal action-value function  $Q(s, a)$ , representing the expected cumulative reward obtained by taking an action  $a$  in a given state  $s$  and following the optimal policy thereafter [24]. The agent iteratively updates this function using the Bellman equation

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)], \quad (7)$$

where  $\alpha$  is the learning rate,  $\gamma$  is the discount factor,  $r$  is the immediate reward, and  $s'$  is the next state. Although Q-learning is effective for discrete and small state-action spaces, it becomes computationally infeasible in high-dimensional environments, since it requires maintaining a lookup table for all possible state-action pairs.

2) *Parameters Considered in This Work:* In this work, each state corresponds to an OFDM symbol. A total of 1000 fixed OFDM symbols are used as states, while the allocation of  $K_p$  pilot subcarriers within these symbols defines the possible actions. These actions are selected from the sample space of the 4-QAM constellation, resulting in  $M^{K-K_p}$  possible combinations.

The reward function is defined as follows: the agent receives a reward of +1 if the resulting PAPR is lower than the previous value, and -1 otherwise.

3) *Computational Complexity:* Let  $S$  denote the set of all possible states and  $A$  the set of actions available to the agent, with cardinalities  $S$  and  $A$ , respectively. In tabular Q-Learning, the action-value function is represented explicitly as a table  $Q(s, a)$  containing one scalar for each state-action pair.

The total number of stored parameters is therefore

$$P = SA, \quad (8)$$

where each entry corresponds to the Q-value associated with a specific  $(s, a)$  pair.

During training, each interaction step performs: (i) a lookup of  $Q(s, a)$ ; (ii) a maximization over all actions,  $\max_{a' \in A} Q(s', a')$ ; and (iii) a constant-time arithmetic update. Because the maximization operation dominates the per-step cost, the time complexity per update is

$$O(A). \quad (9)$$

Over  $N$  interaction steps, the overall training complexity becomes

$$\mathcal{O}(NA). \quad (10)$$

The memory complexity is proportional to the size of the Q-table

$$\mathcal{O}(P) = \mathcal{O}(SA). \quad (11)$$

### B. Deep Q-Network

1) *In General:* The DQN is an RL algorithm that integrates Q-learning with Deep Neural Network (DNN) to approximate the optimal action-value function. It was introduced to overcome the limitations of traditional tabular Q-learning, which becomes infeasible in large or continuous state space, but the actions continue to have discrete space [25]. By leveraging Deep Learning (DL), DQN generalizes across similar states, allowing agents to learn effective policies directly from high-dimensional inputs such as images or continuous sensor data [26].

The optimal Q-function satisfies the Bellman optimality equation [27]

$$Q^*(s, a) = \mathbb{E}_{s'} \left[ r + \gamma \max_{a'} Q^*(s', a') \right], \quad (12)$$

where  $\gamma \in [0, 1)$  is the discount factor balancing immediate and future rewards.

Instead of maintaining a lookup table for all state-action pairs, DQN employs a DNN  $Q(s, a; \theta)$ , parameterized by  $\theta$ , to approximate  $Q^*(s, a)$ . The network takes the current state as input and outputs an estimated Q-value for each possible action. The parameters  $\theta$  are optimized by minimizing the Temporal-Difference (TD) loss [27, 28]

$$L(\theta) = \mathbb{E}_{(s, a, r, s') \sim \mathcal{U}(\mathcal{D})} \left[ (r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta))^2 \right], \quad (13)$$

where  $\theta^-$  denotes the parameters of a target network, periodically updated to stabilize learning. The expectation  $\mathbb{E}_{(s, a, r, s') \sim \mathcal{U}(\mathcal{D})}$  is taken over transitions uniformly sampled from the replay memory  $\mathcal{D}$ .

The DQN relies on four key mechanisms to ensure stable and efficient learning. (i) Experience replay stores transitions  $(s_t, a_t, r_t, s_{t+1})$  in a buffer from which random minibatches are sampled, breaking temporal correlations and improving data efficiency [29, 30]; (ii) The target network  $Q(s, a; \theta^-)$  provides stable target values by updating its parameters less frequently ( $\theta^- \leftarrow \theta$ ) [31]; (iii) An  $\epsilon$ -greedy policy balances exploration and exploitation: with probability  $\epsilon$  the agent explores, otherwise it exploits the learned policy, with  $\epsilon$  decaying over time [32]; (iv) Finally, the neural architecture typically comprises fully connected layers with nonlinear activations such as Rectified Linear Unit (ReLU) [33], though convolutional or recurrent layers may be adopted to handle spatial or temporal dependencies.

Through repeated interactions with the environment, these mechanisms allow the DQN to iteratively refine its Q-function approximation and converge to near-optimal policies. Once

trained, the agent can select effective actions even in unseen states, making DQN suitable for high-dimensional and dynamic tasks such as vehicular routing, game playing, and autonomous control systems.

2) *Parameters Considered in This Work:* In this work, each state corresponds to an OFDM symbol. A total of 30,000 fixed OFDM symbols are used as states, while the allocation of  $K_p$  pilot subcarriers within these symbols defines the possible actions. These actions are selected from the sample space of the 4-QAM constellation, resulting in  $M^{K-K_p}$  possible combinations, where  $M$  denotes the modulation order.

Unlike the traditional tabular Q-learning, the DQN employs a NN to approximate the Q-function, as we can see in Fig. 1. The adopted architecture consists of a multilayer perceptron with one hidden layer, i.e.,  $L = 1$ , containing  $G_\ell = 256$  neurons, where  $\ell$  is the current layer, and a ReLU activation function. The input layer represents the environment state, in which the real and imaginary components must be separated, since the values are complex. The output layer, in turn, corresponds to the optimal action determined for that state. The network parameters are optimized using the following configuration: a learning rate of 0.0001, a replay buffer size of 1,000,000 transitions, and a batch size of 128. Training starts after 10,000 interactions, with a discount factor  $\gamma = 0.9252$ , target network updates every 100 steps, and a soft update coefficient  $\tau = 1.0$ .

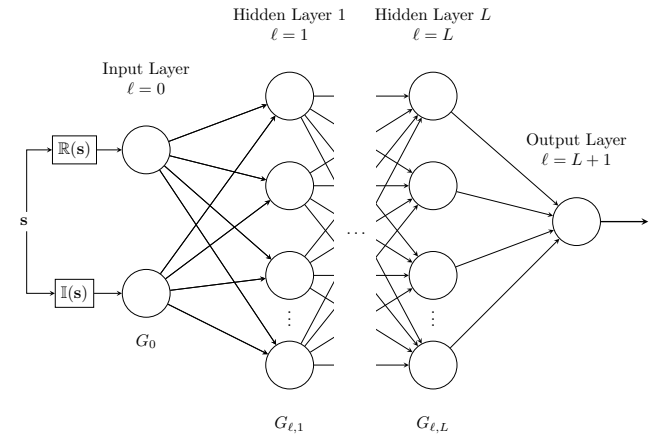


Fig. 1: Architecture of the Multilayer Perceptron

Exploration follows an  $\epsilon$ -greedy strategy, where  $\epsilon$  decays linearly from 1.0 to 0.05 over 80% of the total timesteps, ensuring a gradual transition from exploration to exploitation. The model was trained using the baseline DQN implementation [34].

3) *Computational Complexity:* As in the tabular case, let  $S$  denote the set of possible states and  $A$  the set of available actions. Unlike Q-Learning, the DQN algorithm does not store a table of size  $|S| \times |A|$ . Instead, it approximates the action-value function using a neural network parameterized by  $\theta$

$$Q(s, a; \theta) \approx Q^*(s, a). \quad (14)$$

Let the input dimension be  $d_{\text{in}}$  and let the network architecture be defined by layers with  $G_\ell$  neurons, where the final layer

satisfies  $G_L = |A|$ . The total number of trainable parameters is

$$P = \sum_{\ell=1}^L (G_{\ell-1}G_{\ell} + G_{\ell}), \quad (15)$$

where each term corresponds to the weights and biases of layer  $\ell$ .

During training, each gradient update requires a forward pass and a backward pass for a mini-batch of size  $B$ . The computational cost of one such update is proportional to the number of parameters:

$$\mathcal{O}(BP). \quad (16)$$

Thus, the total training complexity over  $N$  updates is

$$\mathcal{O}(NBP). \quad (17)$$

## V. NUMERICAL RESULTS

This section presents the performance results in terms of PAPR and BER, considering a Q-Phase Shift Keying (PSK) modulation scheme. Two pilot subcarriers were allocated in the first and last positions of each OFDM symbol, totaling 16 subcarriers, of which 14 transmit useful data. Only two pilots were used because with just two we were able to reduce the PAPR as desired, and the idea is to reduce the PAPR without significantly impacting the system's spectral efficiency, since the greater the number of pilots, the greater the loss of spectral efficiency. This issue will be detailed further when the BER results are presented. The BER performance was evaluated over an Additive White Gaussian Noise (AWGN) channel. The following subsections provide a detailed analysis of the obtained PAPR and BER results<sup>1</sup>.

### A. PAPR Performance

Fig. 2 shows the PAPR performance of Q-Learning, DQN, DFT, and PTS compared to the original signal, without any reduction technique. Both Q-Learning and DQN outperform DFT, demonstrating that learning based approaches can more effectively exploit the structure of the OFDM symbol to suppress peaks.

Among the evaluated techniques, PTS achieves the lowest PAPR, as expected. This method explicitly searches through a set of candidate phase vectors each representing a combination of rotation factors to approximate the optimal signal configuration. Because each candidate requires a full IFFT evaluation, PTS attains strong performance at the cost of significantly higher complexity, which grows with the number of phase combinations. As a result, despite its effectiveness, PTS becomes impractical for systems with large  $M$ , strict latency requirements, or high throughput operation.

Within this scenario, DQN stands out as the most balanced solution. Although its PAPR reduction does not reach the same level as PTS, it offers a substantially better performance and complexity trade-off. Once training is complete, DQN

requires only a single forward pass to determine the adjustment applied to the symbol, no phase-vector search is performed, and only one IFFT is needed per transmission. This results in low runtime cost, high scalability, and feasibility for real-time deployments, advantages that PTS cannot provide.

It is also observed that Q-Learning reduces PAPR only for OFDM symbols seen during training, since the tabular representation cannot generalize to new states. In contrast, DQN incorporates a NN that learns a mapping from states to actions, enabling generalization to unseen symbols and providing consistent performance across the entire dataset.

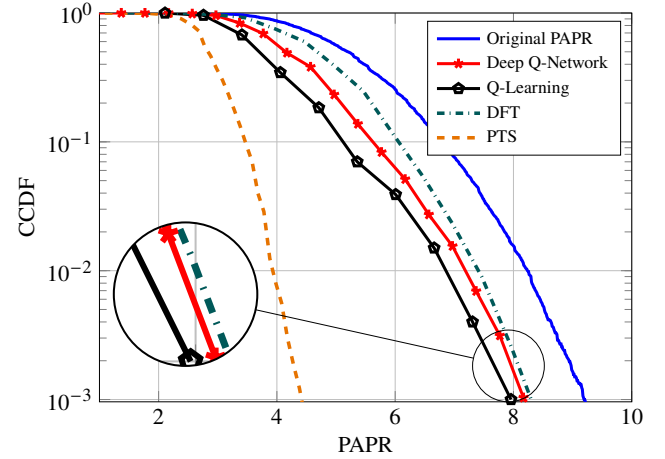


Fig. 2: Performance of PAPR for DFT, PTS, Q-Learning and DQN for 2 pilots and 16 subcarriers.

### B. BER Performance

As illustrated in Fig. 3, a slight deterioration in BER performance is observed for both Q-Learning and DQN. This degradation is associated with a reduction in spectral efficiency, a direct consequence of the insertion of pilot subcarriers. The presence of these subcarriers reduces the number of carriers available for data transmission, which in turn impacts the energy efficiency of the transmission. Part of the total energy is directed to the pilot subcarriers used in PAPR mitigation, instead of being used for the transmission of useful information. However, this degradation is relatively small when compared to the substantial benefits achieved in terms of PAPR reduction, making the proposed approach a highly effective trade-off between performance and efficiency. The equation 18 can show this with details [35]:

$$p_e = \bar{\mu}Q \left( \sqrt{\xi \frac{\epsilon E}{N_0}} \right), \quad (18)$$

where  $E$  denotes the average energy of the QAM constellation,  $\bar{\mu} = 4(\sqrt{M} - 1)/\sqrt{M}$  represents the average number of nearest symbols, and  $\xi = 3/(M - 1)$  is the constellation scaling factor. The term  $\epsilon = (K - K_p)/K$  accounts for the spectral efficiency adjustment in the Q-Learning and DQN, compensating for the presence of  $K_p$  pilot subcarriers, which do not convey data.

<sup>1</sup>The code used to generate the results presented in this paper is available at: [https://github.com/BiaSabrina/Bianca\\_TP558\\_ML\\_Avancado/tree/main/Projeto\\_Final](https://github.com/BiaSabrina/Bianca_TP558_ML_Avancado/tree/main/Projeto_Final)

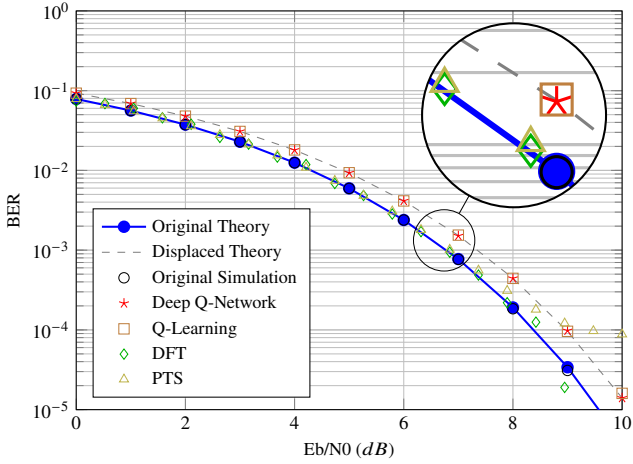


Fig. 3: Performance of BER for DFT, PTS, Q-Learning and DQN for 2 pilots and 16 subcarriers.

### C. Computational Complexity

The exhaustive search for optimal phase factors in PTS methods is computationally expensive and impractical for real-time applications. Even in its simplified form, the PTS scheme requires evaluating  $D$  candidate phase combinations, each demanding an IFFT computation. Since IFFT has complexity  $\mathcal{O}(K \log_2 K)$ , the total cost becomes  $\mathcal{O}(DK \log_2 K)$ , which grows linearly with the number of phase factors. This iterative processing is the main reason PTS achieves strong PAPR reduction, but it also makes the method unsuitable for latency constrained systems. Suboptimal PTS variants reduce  $D$ , but at the expense of performance, as also reported in [36]. The DFT method has fixed complexity  $\mathcal{O}(K \log_2 K + K \log_2(K/D))$ , but its achievable PAPR reduction is inherently limited.

In contrast, RL approaches such as Q-Learning and DQN avoid combinatorial searches entirely. Once trained, both methods require only a single action selection per OFDM symbol. Q-Learning performs a table lookup and a maximization over actions, with complexity  $\mathcal{O}(A)$ , while DQN performs a single forward pass with cost  $\mathcal{O}(P)$ , where  $P$  is the number of network parameters. Crucially, both methods require only one IFFT per transmission, adding a cost of  $\mathcal{O}(K \log_2 K)$ , regardless of the number of phase candidates representing a drastic reduction compared to the  $D$  IFFTs required by PTS.

Although PTS achieves stronger absolute PAPR reduction in our simulations, its runtime complexity increases proportionally with the number of candidate phase factors. In contrast, the proposed DQN achieves competitive PAPR reduction while maintaining fixed and significantly lower computational cost during inference, besides offering generalization capability and robustness to signal variations.

Table I summarizes the computational complexity of all algorithms using the specific parameters of our simulation.

## VI. CONCLUSION

This article presented a new approach for reducing PAPR in OFDM systems based on the DQN algorithm. The proposed method, which uses two pilot subcarriers, demonstrated

TABLE I: Computational Complexities

Method	Complexity
<b>Q-Learning</b>	$\mathcal{O}(A + K \log_2 K)$
<b>Deep Q-Network</b>	$\mathcal{O}(P + K \log_2 K)$
<b>DFT</b>	$\mathcal{O}(K \log_2 K + K \log_2(K/D))$
<b>PTS</b>	$\mathcal{O}(DK \log_2 K)$

a significant reduction in PAPR compared to conventional techniques such as DFT and PTS. Among the methods evaluated, DQN showed promising results: although Q-Learning achieved slightly superior performance, DQN demonstrated better generalization capabilities and effective mitigation of power peaks. A small degradation in BER performance was observed due to the insertion of the pilot subcarriers; however, this impact is minimal compared to the significant gains in PAPR reduction. As a future perspective, we intend to extend the proposed structure to scenarios with a larger number of subcarriers in order to evaluate its scalability and robustness.

### ACKNOWLEDGMENT

This work has received partial funding from the project XGM-AFCCT-2024-2-15-1, supported by xGMobile – EM-BRAP II - Intel Competence Center on 5G and 6G Networks, with financial resources from the PPI IoT/Manufacturing 4.0 program of MCTI grant number 052/2023, signed with EM-BRAP II. Additionally, this work was partially supported by the Ciência por Elas project (APQ-04523-23 funded by Fapemig), the SEMEAR project (22/09319-9 funded by FAPESP), the Brasil 6G project (01245.010604/2020-14 funded by RNP and MCTI), and CNPq-Brasil.

### REFERENCES

- [1] Federal Communications Commission. Spectrum policy task force report. <https://www.fcc.gov/>, 2010.
- [2] Agência Nacional de Telecomunicações (Anatel). Anatel aprova regulamento para uso de espectro ocioso – tv white spaces. <https://www.gov.br/anatel/>, 2021.
- [3] M. G. Vieira et al. Impact of regulation on tv white space implementation in brazil: Laboratory and field analyses using the 5g-range system. *Sensors*, 25(8):2469, 2024. doi: 10.3390/s25082469.
- [4] Yasir Rahmatallah and Seshadri Mohan. Peak-to-average power ratio reduction in ofdm systems: A survey and taxonomy. *IEEE Communications Surveys & Tutorials*, 15(4):1567–1592, 2013. doi: 10.1109/SURV.2013.021313.00164.
- [5] Y.-C. Wang and Chintha Tellambura. A simplified clipping and filtering technique for par reduction in ofdm systems. *IEEE Signal Processing Letters*, 12(6):453–456, 2005. doi: 10.1109/LSP.2005.849697.
- [6] Kee-Hoon Kim, Hyun-Seung Joo, Jong-Seon No, and Dong-Joon Shin. Adaptive generation method of ofdm signals in slm schemes for low-complexity. *arXiv preprint*, 2012.
- [7] Kee-Hoon Kim, Hyun-Bae Jeon, Jong-Seon No, and Dong-Joon Shin. A new low-complexity selected mapping scheme using cyclic shifted ifft. *arXiv preprint*, 2012.

- [8] N. Arora et al. Partial transmit sequence (pts) – papr reduction scheme. In *Proc. International Conference on Emerging Trends*. Atlantis Press, 2013.
- [9] Bianca S. C. da Silva et al. A survey of papr techniques based on machine learning. *Sensors*, 24(6), 2024. doi: 10.3390/s24061918.
- [10] Zainab Alnaseri, Ahmed Al-Saedi, Yassine Himeur, and Kuan Li. Papr reduction based on deep learning autoencoder in coherent optical ofdm systems. *arXiv preprint*, 2024.
- [11] Chenglong Bao, Yi Fang, Yunjin Chen, and Ting-Zhu Jiang. Perturbation-assisted papr reduction for mimo-ofdm via admm. *arXiv preprint*, 2016.
- [12] J.-J. van de Beek, O. Edfors, M. Sandell, S.K. Wilson, and P.O. Borjesson. On channel estimation in ofdm systems. In *1995 IEEE 45th Vehicular Technology Conference. Countdown to the Wireless Twenty-First Century*, volume 2, pages 815–819 vol.2, 1995. doi: 10.1109/VETEC.1995.504981.
- [13] S. Coleri, M. Ergen, A. Puri, and A. Bahai. Channel estimation techniques based on pilot arrangement in ofdm systems. *IEEE Transactions on Broadcasting*, 48(3):223–229, 2002. doi: 10.1109/TBC.2002.804034.
- [14] Y.-C. Wang and Z.-Q. Luo. Optimized iterative clipping and filtering for papr reduction of ofdm signals. *IEEE Transactions on Communications*, 59(1):33–37, 2011. doi: 10.1109/TCOMM.2010.102910.090040.
- [15] Seung Hee Han and Jae Hong Lee. Modified selected mapping technique for papr reduction of coded ofdm signal. *IEEE Transactions on Broadcasting*, 50(3):335–341, 2004. doi: 10.1109/TBC.2004.834200.
- [16] B.S. Krongold and D.L. Jones. An active-set approach for ofdm par reduction via tone reservation. *IEEE Transactions on Signal Processing*, 52(2):495–509, 2004. doi: 10.1109/TSP.2003.821110.
- [17] B.S. Krongold and D.L. Jones. Par reduction in ofdm via active constellation extension. *IEEE Transactions on Broadcasting*, 49(3):258–268, 2003. doi: 10.1109/TBC.2003.817088.
- [18] Mohamed A. Aboul-Dahab, Mohamed M. Fouad, and Radwa A. Roshdy. Generalized discrete fourier transform for fbmc peak to average power ratio reduction. *IEEE Access*, 7:81730–81740, 2019. doi: 10.1109/ACCESS.2019.2921447.
- [19] Robert J. Baxley and G. Tong Zhou. Comparing selected mapping and partial transmit sequence for par reduction. *IEEE Transactions on Broadcasting*, 53(4):797–803, 2007. doi: 10.1109/TBC.2007.908335.
- [20] Gilberto Berardinelli. Generalized dft-s-ofdm waveforms without cyclic prefix. *IEEE Access*, 6:4677–4689, 2018. doi: 10.1109/ACCESS.2017.2781122.
- [21] Koteswara Rao Gudimitla, M. Sibgath Ali Khan, Said-hiraj Amuru, and Kiran Kuchi. Pre-dft multiplexing of reference signals and data in dft-s-ofdm systems. *IEEE Open Journal of the Communications Society*, 5:514–525, 2024. doi: 10.1109/OJCOMS.2023.3348190.
- [22] Dae-Woon Lim, Seok-Joong Heo, Jong-Seon No, and Habong Chung. A new pts ofdm scheme with low complexity for PAPR reduction. *IEEE Transactions on Broadcasting*, pages 77–82, 2006. doi: 10.1109/TBC.2005.861605.
- [23] Yasir Rahmatallah and Seshadri Mohan. Peak-to-average power ratio reduction in ofdm systems: A survey and taxonomy. *IEEE Communications Surveys & Tutorials*, 15(4):1567–1592, 2013. doi: 10.1109/SURV.2013.021313.00164.
- [24] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.
- [25] Xu Wang, Sen Wang, Xingxing Liang, Dawei Zhao, Jincui Huang, Xin Xu, Bin Dai, and Qiguang Miao. Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5064–5078, 2024. doi: 10.1109/TNNLS.2022.3207346.
- [26] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Transactions on Cybernetics*, 50(9):3826–3839, 2020. doi: 10.1109/TCYB.2020.2977374.
- [27] Sean Meyn. The projected bellman equation in reinforcement learning. *IEEE Transactions on Automatic Control*, 69(12):8323–8337, 2024. doi: 10.1109/TAC.2024.3409647.
- [28] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- [29] Beakcheol Jang, Myeonghwi Kim, Gaspard Harerimana, and Jong Wook Kim. Q-learning algorithms: A comprehensive classification and applications. *IEEE Access*, 7:133653–133667, 2019. doi: 10.1109/ACCESS.2019.2941229.
- [30] Pingli Lv, Xuesong Wang, Yuhu Cheng, and Ziming Duan. Stochastic double deep q-network. *IEEE Access*, 7:79446–79454, 2019. doi: 10.1109/ACCESS.2019.2922706.
- [31] Xiaoyu Tan, Chao Qu, Junwu Xiong, James Zhang, Xihe Qiu, and Yaochu Jin. Model-based off-policy deep reinforcement learning with model-embedding. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(4):2974–2986, 2024. doi: 10.1109/TETCI.2024.3369636.
- [32] Christopher Painter-Wakefield and Ronald Parr. Greedy algorithms for sparse reinforcement learning. *arXiv preprint arXiv:1206.6485*, 2012.
- [33] Digvijay Boob, Santanu S Dey, and Guanghui Lan. Complexity of training relu neural network. *Discrete Optimization*, 44:100620, 2022.
- [34] Antonin Raffin, Ashley Hill, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, and Noah Dormann. *Stable-Baselines3: DQN Documentation*. Stable-Baselines3 Contributors, 2024. Available at: <https://stable-baselines3.readthedocs.io/en/master/modules/dqn.html>.
- [35] P.A. Humblet and M. Azizoglu. On the bit error rate of lightwave systems with optical amplifiers. *Journal*

of *Lightwave Technology*, 9(11):1576–1582, 1991. doi: 10.1109/50.97649.

- [36] Yasir Rahmatallah and Seshadri Mohan. Peak-to-average power ratio reduction in OFDM systems: A survey and taxonomy. *IEEE communications surveys and tutorials*, pages 1567–1592, 2013. doi: 10.1109/SURV.2013.021313.00164.



**Bianca S. de C. da Silva** was born in Santa Rita do Sapucaí, Minas Gerais, Brazil, in 1998. She received the B.S. degree in Control and Automation Engineering in 2020 and the M.Sc. degree in Telecommunications Engineering in 2025, both from the National Institute of Telecommunications (INATEL), Santa Rita do Sapucaí, where she is currently pursuing a Ph.D. degree in Telecommunications Engineering. In 2023, she supported field technicians with remote site integration at Ericsson-INATEL.



**Anderson R. R. Marins** He received his B.Sc. and M.Sc. degrees in Electrical Engineering from Inatel, Brazil, in 2007 and 2025, respectively. Since 2013, he has been working as a researcher and systems specialist at Inatel, Brazil. During this period, he has actively participated in various research projects and collaborations with industry partners. His areas of expertise include artificial intelligence, machine learning, digital signal processing, and digital communication systems. He is currently pursuing a Ph.D. degree in Telecommunications Engineering, with

research focused on PAPR reduction for TV White Spaces applications and future mobile communication systems.



**Luciano L. Mendes** received the B.Sc. and M.Sc. degrees from INATEL, Brazil, in 2001 and 2003, respectively, and the Ph.D. degree from Unicamp, Brazil, in 2007, all in electrical engineering. Since 2001, he has been a Professor with INATEL, where he has acted as the Technical Manager of the Hardware Development Laboratory, from 2006 to 2012. From 2013 to 2015, he was a Visiting Researcher with Vodafone Chair Mobile Communications Systems, Technical University of Dresden, where he had developed his postdoctoral training. In 2017, he was

elected as the Research Coordinator of the 5G Brazil Project, an association involving industries, telecom operators, and academia, which aims for funding and build an ecosystem toward 5G in Brazil. He is the Technical Coordinator of Brazil 6G Project and general coordinator of the XGMobile - Competence Center.