IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 August 2023, accepted 1 November 2023, date of publication 9 November 2023, date of current version 16 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3331740

RESEARCH ARTICLE

Contextual Bandit-Based Amplifier IBO Optimization in Massive MIMO Network

MARCIN HOFFMANN, (Graduate Student Member, IEEE), AND PAWEL KRYSZKIEWICZ, (Senior Member, IEEE)
Institute of Radiocommunications, Poznań University of Technology, 60-965 Poznan, Poland
Corresponding author: Marcin Hoffmann (marcin.hoffmann@put.poznan.pl)

This work was supported by the Polish National Science Centre under Project 2021/41/B/ST7/00136.

ABSTRACT Massive Multiple-Input Multiple-Output (MMIMO) is one of the 5G key enablers. Though, most of the works consider MMIMO under assumptions of ideal hardware. It has been shown that Power Amplifiers (PAs) introduce nonlinear distortion while operating close to their saturation power. Moreover, these distortions are in some cases beamformed toward the user, preventing antenna array gain from solving this problem. One of the possible solutions is an adaptive adjustment of the PA operating point, measured by Input Back off (IBO), to find a balance between wanted signal power and nonlinear distortion power. This work proposes a Contextual Bandit-Based IBO Optimization (COBBIO) algorithm to find rate-maximizing IBO for a given user's radio conditions using learning through interaction. The proposed solution is tested in a realistic analog beamforming MMIMO cell simulator with multiple functional blocks, e.g., precoder, user scheduler, and utilizing an accurate 3D Ray-Tracing radio channel model. COBBIO provides throughput gains both over fixed-IBO schemes and state-of-the-art analytical IBO adjustment algorithms. The highest gains were observed for the so-called cell-edge users, where up to 83% improvement over the state-of-the-art algorithm was observed for the proposed COBBIO algorithm.

INDEX TERMS Massive MIMO, 5G, machine learning, nonlinear distortion, input back-off (IBO).

I. INTRODUCTION

The Massive Multiple-Input Multiple-Output (MMIMO) technology is a key enabler for achieving high user throughputs in 5G, and presumably 6G networks [1]. However, the phenomena related to hardware impairments in MMIMO still require attention. In [2] the influence of nonlinear PA on the out-of-band radiation of an MMIMO transmitter has been analyzed. Unlike previous studies, e.g., [3], it shows that the nonlinear distortion can achieve a similar array gain as the wanted signal. Therefore, this problem needs proper countermeasures, e.g., nonlinearity-minimizing precoders [4]. Moreover, the nonlinear distortion problem statement and some of its solutions, common for Orthogonal Frequency Division Multiplexing (OFDM) systems, can be applied directly to MMIMO OFDM systems, e.g., iterative, nonlinearity-aware reception [5], [6]. Another approach to decrease the impact of nonlinear distortion on the OFDM system is to reduce the Peak-to-Average Power Ratio (PAPR), e.g., with the use of the dedicated waveforms designed using Machine Learning (ML) models [7]. The drawback of this approach is that utilization of a new waveform requires redesigning network protocols for both Base Station (BS) and User Equipment (UE). However, this problem can be addressed from the transceiver control perspective, by adjustment of the PA operation point, measured by the Input-Back-Off (IBO) being the ratio between the input saturation power of the PA and the average power of the input signal. In state-of-the-art systems, the PA's IBO is fixed to make Error Vector Magnitude (EVM) or spectral emission mask at the transmitter output compliant with the standard. However, by changing IBO the relation between wanted signal power, distortion power, and the thermal noise at the receiver can be balanced, as such optimizing the network's performance. Most importantly, the adjustment of IBO does not require changes in the 5G network protocols enabling

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Kavak.

its adoption in the existing networks. Moreover, while the IBO modification can increase adjacent channel emission the coexistence ability can be restored by proper filtering [5].

In [8] the Signal-To-Noise-and-Distortion Ratio (SNDR) of an OFDM link is maximized by the PA IBO adjustment at BS which utilizes OFDM. While this solution can be adapted to some configurations of MMIMO systems, e.g., analog beamforming, the authors assumed a simplified system model, e.g., flat fading channel. This is not the case in real-world scenarios, where the radio channel is rich in reflections, and diffractions, making it frequency selective. Moreover, the authors of [8] do not consider layered signal processing in a real BS composed of, e.g., scheduling, and utilization of a fixed set of Modulation and Coding Schemes (MCS), that affect the throughput achievable by the network users.

While accurate mathematical modeling of a 5G MMIMO system may be difficult, its analytical optimization may be even harder. We propose to utilize ML, which is considered one of the key enablers for intelligent 6G networks [1]. In detail, we propose a COntextual Bandit-Based amplifier IBO Optimization (COBBIO) algorithm. Contextual bandit is a sub-class of the Reinforcement Learning (RL) algorithms, where the aim of the agent is to learn what actions should be taken within the current context in order to maximize reward [9], i.e., how to adjust the value of IBO dynamically, with respect to the radio channel conditions, so as to maximize user throughput. The proposed COBBIO algorithm utilizes a Deep Q Network (DQN) model, which unlike the state-of-the-art IBO optimization method [8] is trained directly on the network data making it aware of the frequency-selective radio channel, and multi-stage signal processing used inside the BS [10]. The proposed COBBIO algorithm is built on top of the contextual bandit framework that defines internal algorithms for data capture, model training, and providing a balance between exploitation and exploration. The superiority of the proposed solution is justified by an advanced, analog beamforming MMIMO BS simulation using a 3D Ray-Tracing radio channel model. One should notice that considered analog beamforming is the worst-case scenario from the perspective of nonlinear distortion, i.e., the same MMIMO array gain is applied to both the wanted signal and distortion term [6].

The paper is organized as follows: the system model is described in Sec. II. The proposed method of IBO optimization, i.e., the COBBIO algorithm based on the contextual bandit is described in Sec. III. The simulation environment is described in IV. The results are presented and discussed in Sec. V. Conclusions are formulated in Sec. VI.

II. SYSTEM MODEL

A downlink in a single MMIMO cell is considered utilizing M transmit antennas and N_{rb} Resource Blocks (RBs) with the block diagram depicted in Fig. 1. First, a user scheduler decides on the allocation of the radio resources. Its decisions are passed to the so-called 5G Distributed Unit (DU), which is responsible for the physical layer processing of the user's

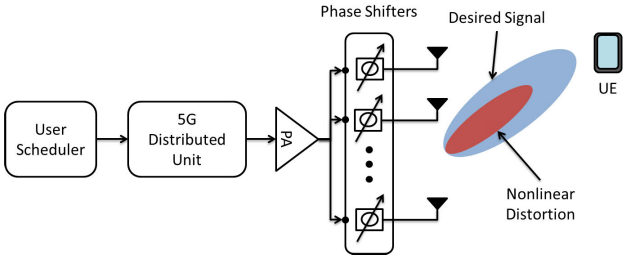

FIGURE 1. System model.

data, e.g., bit-symbol mapping and coding according to given Modulation and Coding Schemes (MCSs), channel estimation, and OFDM multiplexing. The output signal is then fed into the Power Amplifier (PA). The PA nonlinear effects are reflected by the soft limiter model where the output OFDM sample $\hat{s}(n)$ at n-th time instance is related to the input OFDM sample $s(n)$ by [8]:

$$\hat{s}(n) = \begin{cases} s(n), & \text{for } |s(n)| < A_{\text{sat}}, \\ A_{\text{sat}} \cdot e^{j \arg\{s(n)\}}, & \text{for } |s(n)| \geq A_{\text{sat}}, \end{cases} \tag{1}$$

where A_{sat} is the PA saturation voltage. The IBO (γ^2) is given by:

$$\gamma^2 = \frac{A_{\text{sat}}^2}{P_{\text{in}}} = \frac{A_{\text{sat}}^2}{\mathbb{E}\{|s(n)|^2\}}, \tag{2}$$

where P_{in} denotes the average power of input signal $s(n)$. It has been shown in [8] that the PA output can be decomposed as

$$\hat{s}(n) = \alpha s(n) + d(n), \tag{3}$$

where

$$\alpha = 1 - e^{-\gamma^2} + \frac{\sqrt{\pi}}{2}\gamma \cdot \text{erfc}(\gamma) \in \langle 0; 1 \rangle \tag{4}$$

is the wanted signal scaling factor and $d(n)$ is nonlinear distortion sample uncorrelated with signal $s(n)$ of power

$$\sigma_d^2 = \frac{A_{\text{sat}}^2}{\gamma^2}\left(1 - \alpha^2 - e^{-\gamma^2}\right). \tag{5}$$

If for a given PA the mean power of $s(n)$ is increased, decreasing γ, the higher wanted signal power at the receiver is expected at the cost of increased σ_d^2. Next, the signal from the PA output is equally divided between M antenna elements. Here, an analog beamformer is considered which utilizes M phase shifters to steer the beam in the direction of the user. Although such a solution allows simultaneously serving only one user, it has the advantage of low hardware and signal processing complexity [11]. While the same precoding is applied to both the wanted signal and distortion the same radiation pattern will be obtained for both signals, resulting in the worst-case scenario, i.e., the MMIMO array gain will not increase the signal-to-distortion ratio [12], independently from the wireless channel properties. The array-channel gain G_l of signal $\hat{s}(n)$ at the resource block l can be calculated

as $G_l = \sum_{m=0}^{M} h_{m,l} w_m$, where $h_{m,l}$ is a complex channel coefficient between the single antenna user, and m-th antenna of the BS at RB l, and $w_m = \frac{1}{\sqrt{M}} e^{j\varphi_m}$ is the beamforming coefficient for m-th antenna. The highest Signal-to-Noise Ratio (SNR) can be obtained using the considered PA for transmitting a single carrier of amplitude A_{sat} with perfect beamformer, i.e., $\varphi_m = -\arg\{h_{m,l}\}$, resulting in $G_l = \frac{1}{\sqrt{M}} \sum_{m=0}^{M} |h_{m,l}|$. In this case the SNR equals $G_l^2 A_{\text{sat}}^2 / \sigma_n^2$, where the σ_n^2 denotes the power of Additive White Gaussian Noise (AWGN). However, as there are multiple frequencies l to be used we average this metric over all RBs obtaining the saturation SNR

$$\text{SNR}_{\text{sat}} = \frac{A_{\text{sat}}^2}{\sigma_n^2} \cdot \frac{1}{N_{\text{rb}} M} \sum_{l=0}^{N_{\text{rb}}} \left(\sum_{m=0}^{M} |h_{m,l}| \right)^2. \tag{6}$$

This is an adaptation of SNR_{sat} metric used for Single Input Single Output (SISO) wireless channel description in [8] to an MMIMO system.

Authors of [8] assume the wireless channel is frequency flat over the whole OFDM band resulting in a constant gain in the whole band, i.e., $\forall l G_l = G$. In such a system signal-to-noise-plus-distortion ratio (SNDR) is given by:

$$\text{SNDR} = \frac{G\alpha^2 P_{in}}{G\sigma_d^2 + \sigma_n^2}. \tag{7}$$

While this assumption allowed to propose an analytical formula for the *optimal* IBO, it is suboptimal in a frequency-selective channel. Moreover, the practical 5G system has some limitations, e.g., due to the MCS selection mechanism at some point increasing SNDR would not provide further user-throughput improvement. This is not considered in [8].

III. FRAMEWORK FOR CONTEXTUAL BANDIT-BASED IBO OPTIMIZATION

In this work, we propose to extend the MMIMO BS with the dedicated IBO optimization module, where the proposed COBBIO algorithm is deployed. Our objective is to adjust IBO (γ^2) for a currently scheduled user so as to maximize its throughput. The system model described in Sec. II consists of many functional blocks like user scheduler, analog precoder, MCS selection, and most importantly it is affected by the nonlinear distortion, that is steered toward UE together with a desired signal. Such a complex system is hard to be modeled analytically and optimized with the use of standard optimization methods. Instead, we propose to utilize ML techniques. The considered problem can be classified as the so-called contextual bandit problem [9], i.e., within the context of a currently scheduled user our objective is to select IBO, which will result in the highest throughput.

A. CONTEXTUAL BANDIT FRAMEWORK

The framework for the proposed COBBIO algorithm is depicted in Fig. 2. It is similar to a RL framework in that it involves an agent interacting with the environment by taking proper actions based on the observed states and received

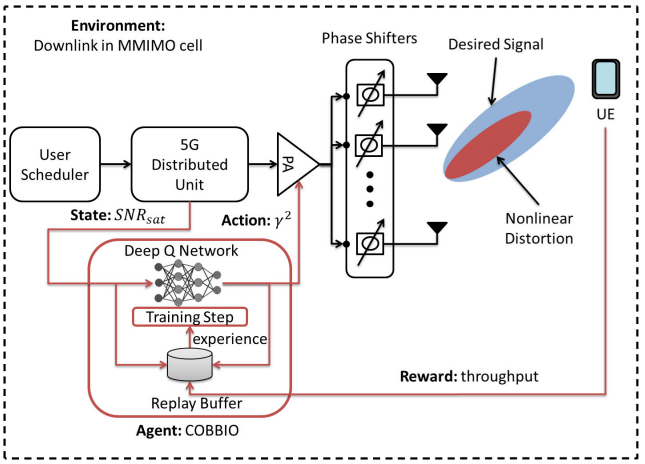

FIGURE 2. Contextual bandit based IBO optimization.

rewards. The difference is that there is no dependency between the consecutive states, i.e., the contextual bandit agent may focus only on the maximization of the reward in the current state. The components that constitute the contextual bandit framework for the proposed COBBIO algorithm are defined as follows:

- **Environment** is a downlink in the MMIMO cell, i.e., our system model described in Sec. II.
- **State** is defined as the SNR_{sat} computed according to (6). Due to the averaging over the RBs, this metric would be relatively stable, and good for the characterization of the user's radio conditions. It's important to note that SNR_{sat} is a continuous variable, making the RL state space continuous as well.
- **Action** is defined as the IBO (γ^2), and is also continuous. Both the wanted signal power and distortion power change monotonically in the function of γ^2. Thus the problem of continuous action space can be resolved by discretization as proposed in [13]. As a result, action is one of the fixed IBO levels ranging from γ^2_{min} to γ^2_{max} with the equal step of γ^2_{step}.
- **Reward** is defined as the throughput that was achieved by the currently scheduled UE. In Fig. 2 the throughput is reported by the UE to visualize the contextual bandit cycle. In practice, such a value is available at the 5G DU.
- **Agent** is the proposed COBBIO algorithm. It recognizes the state (SNR_{sat}) and performs an action, i.e., selects the value of IBO (γ^2). A detailed description of the agent's internal algorithms is provided in the following sections.

B. DEEP Q NETWORK

The aim of the agent is to select the IBO (γ^2) that provides the highest user throughput, based on the SNR_{sat}. In other words, the agent must approximate the so-called Q-function to determine the expected reward associated with each action (IBO value). For a problem that has a continuous state space and discrete action space, a common approach is to

utilize a dedicated artificial neural network, i.e., the so-called DQN [14]. The DQN takes the state (SNR_{sat}) as an input and outputs the Q-values. In the case of the contextual bandit, where the agent is focused only on the maximization of the reward in the current state, the Q-values are directly the expected reward (user throughput) associated with each action (γ^2). It has been proven that a 3-layer neural network can approximate any discontinuous function [15]. Thus we propose the DQN architecture to consist of an input layer of size 1, 3 hidden dense layers of size K, followed by the so-called rectified linear unit (ReLU), which introduces a following nonlinear function between input and output $g(x) = \max\{0, x\}$, and an output layer of size equal to the number of actions. As can be seen in Fig. 2, the training of DQN is incorporated at the end of the contextual bandit cycle. After the agent receives the reward, the experience sample that is defined as state, action, and reward tuple is put into the so-called Replay Buffer, a cyclic buffer data structure of size J. Then a batch of L experience samples is taken from the Replay Buffer and used to perform a single Stochastic Gradient Descent (SGD) step to update the weights of DQN [16]. The L samples are selected from the Replay Buffer according to the Combined Experience Replay (CER) [17]. CER is a low-complexity algorithm that takes the latest observed experience sample and randomly selects the remaining $L-1$ experience samples from the Replay Buffer. The SGD optimizes the DQN weights so as to minimize the Mean Absolute Error (MAE) between the received rewards, and DQN output, i.e., estimated Q-values.

C. ACTION SELECTION

One of the challenges in solving the Contextual Bandit problem is the balance between exploration and exploitation, i.e., how much time an agent should spend on exploring new actions, and when it should act greedy by selecting the best-known action (the one associated with the highest Q-value). In our previous work, we have shown that Upper Confidence Bound (UCB) provides good exploration-exploitation balance, and fast convergence [18]. However, UCB is not proper for dealing with continuous state space as it requires storing a number of visits in each state. We propose to utilize a well-known ϵ-greedy strategy [9], i.e., with the probability of ϵ agent selects the random action, and with the probability of $1-\epsilon$ selects the greedy (best-known) action, associated with the highest Q-value. It is expected that the agent would spend more time on exploration during the first phases of training, and after getting enough experience would turn into the exploitation of current knowledge. We propose to start with an $\epsilon = 1.0$ and decay it according to the following rule:

$$\epsilon \leftarrow \max\left(\epsilon_{\min}, \epsilon - \frac{\delta_\epsilon}{N_a}\right), \tag{8}$$

where $\epsilon_{\min}$ is the minimal arbitrary chosen probability of exploration, δ_ϵ is the decay step, and N_a is the total number of actions that the agent has already taken.

IV. SIMULATION ENVIRONMENT

To evaluate the proposed IBO optimization module based on the deep contextual bandit framework, we have developed an advanced simulator of the MMIMO 5G cell. In this section, the utilized 5G network simulator is described together with its parameters. Moreover, the utilized 3D Ray-Tracing channel model is presented that has been used to generate accurate and realistic radio channel coefficients.

We are considering a downlink in a single MMIMO cell, which operates at the center frequency of 3.6 GHz, i.e., within the 3GPP n78 band [19]. The available bandwidth is equal to the 25 MHz and is divided between $N_{\text{rb}} = 69$ resource blocks, including a guardband. The MMIMO BS is equipped with a rectangular antenna array of $M = 128$ elements (8 vertical $\times$ 16 horizontal). The saturation power of PA A^2_{sat} is equal to the 38 dBm, which corresponds to the 3GPP *Medium Range BS*. The transmit power is divided equally between the resource blocks. The power of thermal noise is -174 dBm/Hz. The MMIMO cell utilizes the following algorithms for the purpose of downlink transmission:

- **User Scheduler:** we utilize a Round Robin user scheduler. This ensures the same sequence of scheduled users during each simulation in order to provide a fair comparison between the proposed IBO optimization algorithm and baseline solutions.
- **Precoder:** we utilize the so-called Equal Gain Transmission (EGT) precoder [20]. The EGT is a phase-only precoder, proper for analog systems, and ensures that equal power is being allocated per antenna.
- **MCS Selection:** we consider MCSs selection algorithm that is based on the SNR estimates obtained at the stage of user scheduling and precoding, i.e., one of the 15 MCSs is assigned to the scheduled user based on the Exponential Effective SNR (EES) mapping, as defined in [21] and [22]. The minimal required EES is -6.28 dB, while the highest, 15th MCS is assigned when the estimated EES is above 20.13 dB.

While evaluating the algorithms oriented on the optimization of the MMIMO network it is of high importance to utilize realistic radio channel models. Measurement studies show that the commonly used i.i.d. Rayleigh channel model significantly differs from the real propagation environment [23]. Thus, to obtain radio channel coefficients between the BS and users, we utilize the realistic Wireless InSite$^{\text{TM}}$ 3D Ray-Tracer. It is configured to consider 15 reflections and 1 diffraction between the MMIMO BS's antennas and each of the single-antenna users. We have defined the 3D urban scenario that follows the well-established Madrid Grid test environment [24]. The deployment of a MMIMO BS, and example placement of users is depicted in Fig. 3. The MMIMO BS is deployed 2.5 m above the rooftop of the central building, i.e., at a height of 45 m, with a 5 deg downtilt. The users are uniformly distributed over the cell area to create a heterogenous radio environment that includes both Line of Sight (LOS) conditions in a park area (green square), and Non-Line of Sight (NLOS) in the narrow streets between

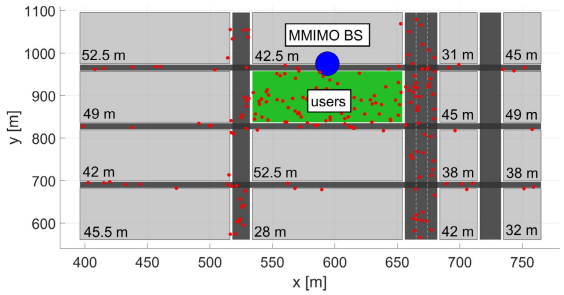

FIGURE 3. Deployment of the MMIMO BS (blue dot) and example placement of the users (red dots). Values on grey rectangles, e.g., 52.5 m, denote building heights.

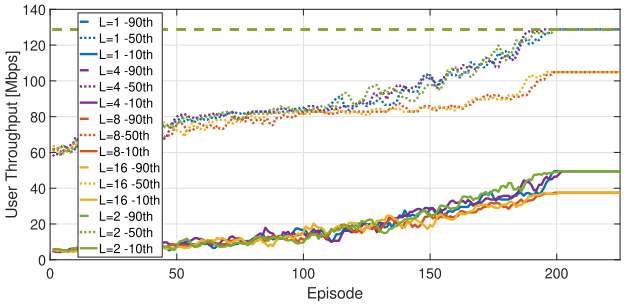

FIGURE 4. 90th, 50th, and 10th percentiles of user throughput distribution in the function of training epoch, for number of neurons per layer $K = 128$, and varying experience samples L.

TABLE 1. Simulation parameters.

Parameter	Value
Center Frequency	3.6 GHz
Bandwidth	25 MHz
Number of Resource Blocks N_{rb}	69
Number of Antennas M	128 (8 vertical, 16 horizontal)
Saturation Power A^2_{sat}	38 dBm
Thermal noise	-174 dBm/Hz
Analog precoder	Equal Gain Transmission [20]
User Scheduler	Round Robin
Urban model	Madrid Grid [24]
Radio channel model	Wireless InSite$^{\mathrm{TM}}$ Ray-Tracer

the relatively high buildings of tens of meters. The simulation parameters are summarized in Table 1.

We compare the proposed COBBIO algorithm against the baseline algorithm (“Reference”), that maximizes the SNDR given by (7). The baseline algorithm approximates optimal IBO $\hat{\gamma}^2$ based on the SNR_{sat} according to the following equation [8]:

$$\hat{\gamma}^2 = 5.975 \cdot e^{0.00943 \cdot \mathrm{SNR}_{\mathrm{sat}}} - 12.79 \cdot e^{-0.0775 \cdot \mathrm{SNR}_{\mathrm{sat}}} [\mathrm{dB}]. \tag{9}$$

Besides the baseline algorithm, we also consider two schemes of constant IBO (“Fixed IBO”): $\gamma^2 = 0$ dB, and $\gamma^2 = 6$ dB respectively.

V. RESULTS

The simulation environment described in the previous section is utilized to evaluate the proposed COBBIO algorithm in terms of computer simulations. Regarding the RL terminology that is also valid for the contextual bandit, the simulation experiments consist of episodes. Each episode is a sequence of steps, i.e., a sequence of contexts (states) that the agent recognizes to take proper action and observe the reward. In this simulation, the experiment step is a single time slot. Within the time slot, IBO is adjusted based on the SNR_{sat}, and user throughput is observed as a reward. However, firstly, COBBIO’s hyperparameters must be obtained. Some of them can be selected based on state-of-the-art knowledge about the RL and 5G MMIMO networks, while others must be obtained through simulation studies. The action space (discretized values of IBO) ranges from $\gamma^2_{\mathrm{min}} = 0$ dB, to $\gamma^2_{\mathrm{min}} = 9$ dB with a step of $\gamma^2_{\mathrm{step}} = 1$ dB. The motivation for such a range is that low values of IBO, e.g., γ^2_{min} below 0 dB produce large nonlinear distortion that causes significant QoS degradation. On the other hand high values of IBO γ^2_{min}, e.g., above 9 dB can result in the poor energy efficiency of a PA, without additional QoS improvement [25]. The studies on CER have shown that the size of a Replay Buffer equal to $J = 1000$ is a good choice [17]. Our target is to train the agent to select greedy actions that will maximize the throughput of a currently scheduled user, thus the minimal probability of exploration is $\epsilon_{\mathrm{min}} = 0$. While implementing this solution in a real 5G network, one may consider setting the probability of exploration ϵ_{min} to a non-zero value, to deal with changes in a radio environment. The training step takes only a one-time slot (0.5 ms for a 5G OFDM network under the assumption of 30 kHz subcarrier spacing). The convergence time is not crucial in this case, i.e., collecting 1000 data samples takes only 5 seconds. Thus we have set a relatively large epsilon decay step of $\delta_\epsilon = 1000$.

To adjust the number of experience samples L to be taken from Replay Buffer for a single SGD step, we have set a large number of neurons in hidden layers: $K = 128$ and tested different values of L for 200 pedestrian users randomly placed over the cell area, and moving with the speed of 1.5 m/s. We have conducted 225 episodes of online training, within every episode each user was scheduled once so there were 200 steps taken by the agent . The 90th, 50th, and 10th percentiles of user throughput distribution in the function of training episode, for $K = 128$ neurons per layer, and varying number of experience samples L are depicted in Fig. 4. The first observation is that users with the best radio conditions (90th percentile) reach the maximum throughput all the time. The throughput achieved by 50th and 10th percentile users stabilize after about 225 episodes. It can be seen that for $L > 4$ performance of COBBIO starts to degrade for the 50th and 10th percentile of users. One of the hypothesis for such a behavior is that during SGD the loss and related gradient are computed over a bigger set of samples. This reduces the noise, and increases the stability of learning, but for arbitrary non-convex functions, SGD with large batch size can stuck in local optimum [26]. In such a case some instability related

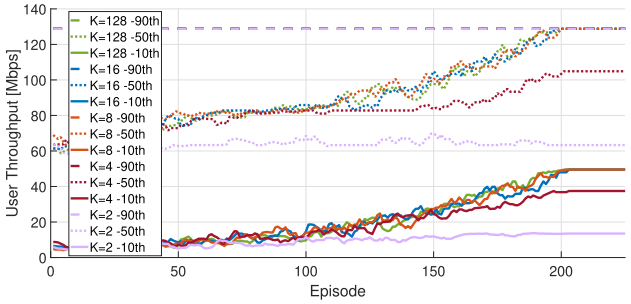

FIGURE 5. 90th, 50th, and 10th percentiles of user throughput distribution in the function of training episode, for experience samples $L = 2$, and a varying number of neurons per layer K.

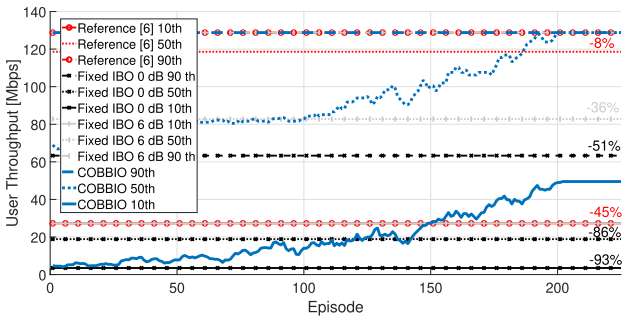

FIGURE 6. Comparison of 90th, 50th and 10th percentiles of user throughput distribution between the proposed COBBIO algorithm ($L = 2$, $K = 8$), reference algorithm and fixed IBO schemes ($\gamma^2 = 0$ dB, and $\gamma^2 = 6$ dB).

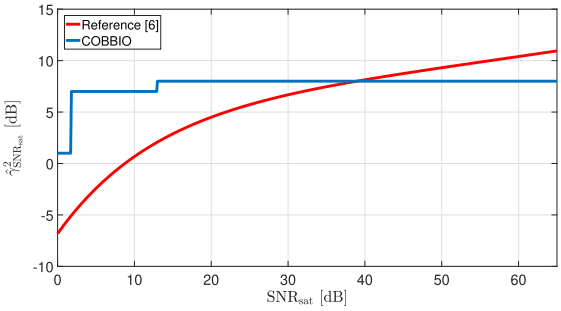

FIGURE 7. Comparison of selected IBO in the function of SNR_{sat} between the reference algorithm, and the proposed COBBIO algorithm.

to small batch size can potentially help to search for global optimum. Based on the observation of the results we have selected the $L = 2$, because of the best convergence for 10th percentile users, i.e., the most challenging group of users that suffer the worst radio conditions.

After tuning the number of experience samples to $L = 2$ our target is to tune the number of neurons in the hidden layers K. For this purpose, we have utilized the same setup of online training as for adjusting the number of experience samples L. The results in terms of 90th, 50th, and 10th percentiles of user throughput distribution are depicted in Fig. 5. It can be seen that for the number of neurons in hidden layer K equal to 2 and 4 the performance of COBBIO is significantly degraded. However, the number of neurons per hidden layer can be lowered to $K = 8$ without decreasing the COBBIO's performance in terms of user throughput. On the other hand, a lower number of DQN parameters reduces the prediction time and required memory. The final architecture of the DQN can be summarized as follows: input layer of size 1, followed by the three hidden layers of size $K = 8$, and an output layer of size 9, i.e., the number of actions that the agent can take. This DQN has a total number of 241 trainable parameters.

After tuning the hyperparameters we have compared the COBBIO algorithm against the two fixed IBO schemes of $\gamma^2 = 0$ dB, and $\gamma^2 = 6$ dB, and reference algorithm based on [8]. The scenario was the same as for the adjustment of L and K. The results in terms of 90th, 50th, and 10th percentiles of user throughput distribution are depicted in Fig. 6. It can be seen that the 90th percentile is deteriorated by 51% for fixed IBO of $\gamma^2 = 0$ dB, compared to the remaining algorithms. This is caused by the high nonlinear distortion. In the case of the 50th percentile the proposed COBBIO algorithm has the best performance, i.e., the reference algorithm, fixed IBO of $\gamma^2 = 6$ dB fixed IBO of $\gamma^2 = 0$ dB are characterized by the median user throughput decreased by 8%, 36%, and 86% respectively, in relation to the COBBIO algorithm. A similar tendency is observed for the 10th percentile user throughput, where compared to COBBIO the 45% degradation is observed for the reference algorithm and fixed IBO of $\gamma^2 = 6$ dB, and 93% for $\gamma^2 = 0$ dB. This shows the benefits of learning through interaction that allows one to select proper IBO after approximately 200 training episodes.

The comparison of IBO selected by the reference algorithm [8] and the proposed COBBIO algorithm in a function of $\mathrm{SNR_{sat}}$ is presented in Fig. 7. We can see that the theoretical IBO values computed according to [8] do not align with the actions taken by the COBBIO. Most importantly for the low values of $\mathrm{SNR_{sat}}$ (e.g., below 20 dB), the theoretical optimum is up to several dBs lower than the IBO values that are selected by the COBBIO. Within the next few paragraphs, we will show that this is the result of the simplified system model assumed by the authors of [8], i.e., mostly by the flat radio channel. On the other hand for $\mathrm{SNR_{sat}} > 40$ dB the theoretical optimal IBO values are much bigger than the ones selected by the COBBIO. It is because the practical system has an upper bound of the reasonable SNDR to be achieved related to the maximal possible MCS. In such a case it is not necessary to increase further IBO in order to reduce nonlinear distortion power.

For further investigations we compared the previously trained COBBIO ($K = 8$, $L = 2$) against the reference algorithm, and fixed IBO schemes under the new set of states (contexts, related to a newly generated set of UEs), independent from those used previously to tune the hyperparameters. From Fig. 7 it can be seen that the biggest difference between the value of IBO indicated by the reference algorithm, and the proposed COBBIO algorithm was observed for the

relatively low values of $\mathrm{SNR}_{\mathrm{sat}} < 40$. These values of IBO correspond mainly to the 10th and 50th percentile of user throughput distribution depicted in Fig. 6, for which the highest benefits of utilizing COBBIO were observed. To focus on these challenging cases, 1620 pedestrian users are randomly placed over the cell area such that their path loss is at least 100 dB. Later on they are moving with the speed of 1.5 m/s. In such a scenario the proposed prediction of IBO is the most beneficial, i.e., its results are significantly different than the results of the reference algorithm. Following the Round Robin scheduling strategy, each user receives in total 6 time slots of 0.5 ms duration, with the first allocation being neglected while obtaining statistics, i.e., average user rate, Capacity Effective SINR Mapping (CESM) [27], and wideband SNDR calculated according to the (7). The statistics are aggregated over the 81 simulation runs. As a result, there are 1620 values of each metric taken for statistical analysis.

In Fig. 8 there is a comparison between the Cumulative Distribution Function (CDF) of the wideband SNDR distribution among users for all tested IBO adjustment solutions. It can be seen that the reference algorithm provides the best wideband SNDR. Such a result could be expected because wideband SNDR is exactly what has been optimized by the authors of [8]. However, the real radio environment is not characterized by a flat wideband channel. One of the statistics that includes channel frequency-selectivity is CESM which relies on per-RB Shannon capacity. The CDFs of CESM are shown in Fig. 9. The fixed IBO scheme of $\gamma^2 = 0$ dB is characterized by the worst CESM resulting from too high distortion power. The second fixed IBO scheme of $\gamma^2 = 6$ dB outperforms the previous one, the reference algorithm, and sometimes even slightly the proposed COBBIO algorithm (recall the COBBIO maximizes rate, not CESM). Recalling Fig. 7 the IBO of 6 dB is good for users that suffer poor radio conditions. However, this does not allow users under better channel conditions to achieve CESM higher than around 27 dB, limiting potentially their rate. On the other hand, the reference algorithm, due to the assumption of a flat radio channel, obtains in many cases CESM lower than the other solutions. This shows that the wideband SNDR optimization is not optimal in a frequency-selective channel. Finally, the COBBIO algorithm is designed to maximize each user rate that is reflected by relatively high CESM values. The main advantage of this approach is that it is trained through interaction on real-network data considering, e.g., a limited set of MCS and frequency-selective radio channels. It is visible that the reference solution obtains higher CESM for around 10% of best channel users. This is caused by the limited MCS set, i.e., the COBBIO algorithm achieves for these users maximal MCS and does not need to increase CESM any further.

Finally, in Fig. 10 the users' throughput is shown. All considered schemes are compared in terms of the 10th percentile (cell-edge users throughput), median, and 90th percentile of user throughput distribution. The COBBIO

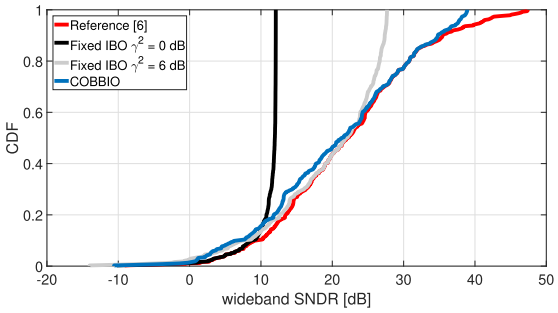

FIGURE 8. CDF of the wideband SNDR calculated using (7) for the fixed IBO schemes ($\gamma^2 = 0$ dB, and $\gamma^2 = 6$ dB), reference algorithm, and the proposed COBBIO algorithm.

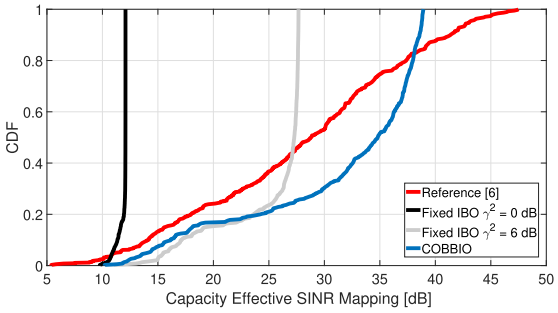

FIGURE 9. CDF of CESM obtained for the fixed IBO schemes ($\gamma^2 = 0$ dB, and $\gamma^2 = 6$ dB), reference algorithm, and the proposed COBBIO algorithm.

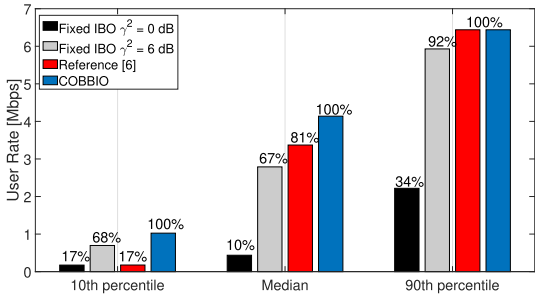

FIGURE 10. Statistics of user-rates: 10th percentile, median, and 90th percentile, computed for the reference algorithm, COBBIO algorithm, and fixed IBO schemes, of $\gamma^2 = 0$ dB, and $\gamma^2 = 6$ dB.

algorithm provides the best user throughputs for all considered percentiles. While the reference solution is the closest in terms of performance for the median and 90th percentile, it is significantly outperformed by the 10th percentile achieving only 17% of the COBBIO's throughput. For the worst-case users (10th percentile) the fixed IBO of $\gamma^2 = 6$ dB achieves the user's rate closest to the ML-based solution. Still, the achievable throughput is lower by around 32%.

VI. CONCLUSION

The management of contemporary 5G and future 6G networks should take into account the nonlinear distortion generated by the PAs. We have shown utilizing an accurate 3D Ray-tracing radio channel model that adjustment of PA IBO based on the proposed COBBIO algorithm can

significantly improve throughput in an MMIMO 5G network. This is not achievable with state-of-the-art analytical IBO adjustment solutions nor with the fixed IBO solutions, that are contemporarily used to guarantee a given EVM at the transmitter output.

ACKNOWLEDGMENT

REFERENCES

[1] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133995–134030, 2020.
[2] E. G. Larsson and L. Van Der Perre, "Out-of-band radiation from antenna arrays clarified," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 610–613, Aug. 2018.
[3] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.
[4] F. Rottenberg, G. Callebaut, and L. Van der Perre, "The Z3RO family of precoders cancelling nonlinear power amplification distortion in large array systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 2036–2047, Mar. 2023.
[5] Y. Sun and H. Ochiai, "Performance analysis and comparison of clipped and filtered OFDM systems with iterative distortion recovery techniques," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7389–7403, Nov. 2021.
[6] M. Wachowiak and P. Kryszkiewicz, "Clipping noise cancellation receiver for the downlink of massive MIMO OFDM system," *IEEE Trans. Commun.*, vol. 71, no. 10, pp. 6061–6073, Oct. 2023.
[7] Y. Huleihel, E. Ben-Dror, and H. H. Permuter, "Low PAPR waveform design for OFDM systems based on convolutional autoencoder," in *Proc. IEEE Int. Conf. Adv. Netw. Telecommun. Syst. (ANTS)*, Dec. 2020, pp. 1–6.
[8] C. H. A. Tavares, J. C. M. Filho, C. M. Panazio, and T. Abrão, "Input back-off optimization in OFDM systems under ideal pre-distorters," *IEEE Wireless Commun. Lett.*, vol. 5, no. 5, pp. 464–467, Oct. 2016.
[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.
[10] W. Yu, F. Sohrabi, and T. Jiang, "Role of deep learning in wireless communications," *IEEE BITS Inf. Theory Mag.*, vol. 2, no. 2, pp. 56–72, Nov. 2022.
[11] E. Bjornson, L. Van der Perre, S. Buzzi, and E. G. Larsson, "Massive MIMO in sub-6 GHz and mmWave: Physical, practical, and use-case differences," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 100–108, Apr. 2019.
[12] C. Mollen, E. G. Larsson, U. Gustavsson, T. Eriksson, and R. W. Heath, "Out-of-band radiation from large antenna arrays," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 196–203, Apr. 2018.
[13] Y. Tang and S. Agrawal, "Discretizing continuous action space for on-policy optimization," in *Proc. 34th AAAI Conf. Artif. Intell., (AAAI), 32nd Innov. Appl. Artif. Intell. Conf., (IAAI), 10th AAAI Symp. Educ. Adv. Artif. Intell., (EAAI)*, New York, NY, USA: AAAI Press, Feb. 2020, pp. 5981–5988, doi: 10.1609/aaai.v34i04.6059.
[14] F. R. Yu and Y. He, *Reinforcement Learning and Deep Reinforcement Learning*. Cham, Switzerland: Springer, 2019, pp. 15–19.
[15] V. E. Ismailov, "A three layer neural network can represent any multivariate function," *J. Math. Anal. Appl.*, vol. 523, no. 1, Jul. 2023, Art. no. 127096. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022247X23000999
[16] N. Ketkar, *Stochastic Gradient Descent*. Berkeley, CA, USA: Apress, 2017, pp. 113–132.
[17] S. Zhang and R. S. Sutton, "A deeper look at experience replay," in *Proc. Deep Reinforcement Learn. Symp. (NIPS)*, 2017. [Online]. Available: https://sites.google.com/view/deeprl-symposium-nips2017/home and https://arxiv.org/abs/1712.01275
[18] M. Hoffmann and P. Kryszkiewicz, "Reinforcement learning for energy-efficient 5G massive MIMO: Intelligent antenna switching," *IEEE Access*, vol. 9, pp. 130329–130339, 2021.
[19] *Base Station (BS) Radio Transmission and Reception (Release 18)*, Standard 3GPP, TS 38.104 v.18.2.0, 3GPP, Jun. 2023.
[20] S. Zhang, R. Zhang, and T. J. Lim, "Massive MIMO with per-antenna power constraint," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2014, pp. 642–646.
[21] Z. Hanzaz and H. D. Schotten, "Performance evaluation of link to system interface for long term evolution system," in *Proc. 7th Int. Wireless Commun. Mobile Comput. Conf.*, Jul. 2011, pp. 2168–2173.
[22] B. Bossy, P. Kryszkiewicz, and H. Bogucka, "Optimization of energy efficiency in the downlink LTE transmission," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[23] S. Willhammar, J. Flordelis, L. Van Der Perre, and F. Tufvesson, "Channel hardening in massive MIMO: Model parameters and experimental assessment," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 501–512, 2020.
[24] *Deliverable D6.1, Simulation Guidelines v1.0*, document ICT-317669-METIS/D6.1, METIS, Mobile and Wireless Communications Enablers for the Twenty-Twenty Information, 2013.
[25] P. Kryszkiewicz, "Efficiency maximization for battery-powered OFDM transmitter via amplifier operating point adjustment," *Sensors*, vol. 23, no. 1, p. 474, Jan. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/1/474
[26] T. Takase, S. Oyama, and M. Kurihara, "Why does large batch training result in poor generalization? A comprehensive explanation and a better strategy from the viewpoint of stochastic optimization," *Neural Comput.*, vol. 30, no. 7, pp. 2005–2023, Jul. 2018, doi: 10.1162/neco_a_01089.
[27] Z. Hanzaz and H. D. Schotten, "Analysis of effective SINR mapping models for MIMO OFDM in LTE system," in *Proc. 9th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jul. 2013, pp. 1509–1515.

MARCIN HOFFMANN (Graduate Student Member, IEEE) received the M.Sc. degree (Hons.) in electronics and telecommunication from the Poznań University of Technology, in 2019, where he is currently pursuing the Ph.D. degree with the Institute of Radiocommunications. He is also gaining scientific experience by being involved in both national and international research projects. His research interests include the utilization of machine learning and location-dependent information for the purpose of network management.

PAWEL KRYSZKIEWICZ (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees (Hons.) in telecommunications from the Poznań University of Technology (PUT), Poland, in 2010 and 2015, respectively. He is currently an Associate Professor with the Institute of Radiocommunications, PUT. He was involved in a number of international research projects. His research interests include multicarrier signal design for green communications and problems related to the practical implementation of massive MIMO systems.