# A DDPG-Based Procedure for Mitigating Pilot Contamination in Massive MIMO RSMA Systems

Felipe Augusto Dutra Bueno
*Dept. of Electrical and Computer Eng.*
*McMaster University*
Hamilton ON, Canada
buenof@mcmaster.ca

José Carlos Marinello Filho
*Dept. of Electrical Eng.*
*Federal University of Technology - UTFPR*
Cornelio Procopio PR, Brazil
jcmarinello@utfpr.edu.br

Telex M. N. Ngatched
*Dept. of Electrical and Computer Eng.*
*McMaster University*
Hamilton ON, Canada
ngatchet@mcmaster.ca

***Abstract*—The growing demand for improved spectral efficiency is one of the main challenges for the upcoming beyond fifth-generation wireless mobile communications networks. While massive multiple-input multiple-output (MIMO) technology has been demonstrating its potential in achieving higher spectral efficiency, the persistent problem of pilot contamination poses a significant hurdle for these systems. To address this issue, the Rate-Splitting Multiple Access (RSMA) framework has emerged as a potential solution. In this paper, we present a novel approach that leverages reinforcement learning (RL) with the Deep Deterministic Policy Gradient (DDPG) algorithm to maximize the sum spectral efficiency (SUM-SE) in a massive MIMO system implementing the RSMA framework with all users sharing a single pilot. The numerical results indicate that the proposed DDPG-based method is a competitive tool for optimizing the SUM-SE in massive MIMO scenarios employing the RSMA framework.**

***Index Terms*—Massive MIMO, Rate-Splitting Multiple Access, Reinforcement Learning, Deep Deterministic Policy Gradient (DDPG), Beyond Fifth-Generation (B5G)**

## I. Introduction

The number of wirelessly connected devices, including fifth-generation (5G) wireless communication devices, is increasing at unprecedented rates [1]. The upcoming sixth-generation (6G) of mobile communications and other beyond 5G (B5G) systems will require much higher performance compared to the current generation, which can be achieved by integrating different wireless services such as Wi-Fi, Bluetooth, THz, and Visible Light Communications [2]. As a result, there has been a strong drive to create new technologies that improve spectral efficiency (SE) in 5G, B5G, and wireless communications systems in general. Among these new technologies, multiple-input multiple-output (MIMO) has proved to be a great success. However, in massive MIMO systems, due to the scarcity of orthogonal sequences, pilot contamination is an obstacle for improving performance.

In [3], the authors demonstrated that with Minimum Mean Square Error (MMSE) precoding/combining schemes and a tiny amount of spatial channel correlation or large-scale fading variations over the array, the capacity increases without bound as the number of antennas increases, even under pilot contamination. Other typical solutions for pilot contamination involve the utilization of either grant-based random access protocols, as demonstrated in studies like [4]–[6], and [7], or grant-free approaches, as explored in works such as [8], [9], and [10]. Additionally, a more recent interference management approach called Rate Splitting Multiple Access (RSMA) [11], [12] has garnered attention for its ability to enhance SE within a wide range of interference scenarios.

What sets RSMA apart from its predecessors is its ambitious goal to serve as a comprehensive superset of other previous multiple access (MA) schemes, making it adaptable to fit most interference levels. This adaptability makes RSMA an extremely versatile tool that can optimize performance in a wide range of communication environments. Promising results have been obtained so far, showcasing RSMA's potential to enhance SE and energy efficiency (EE), Quality of Service, user-fairness, low latency requirements, among other characteristics that are essential for B5G communication systems.

In [13], the RSMA framework is applied to address the issue of intra-cell pilot contamination in massive MIMO systems. The researchers introduced a new downlink (DL) transmission approach of RSMA within Time Division Duplex (TDD) massive MIMO. The authors also developed a novel strategy for designing precoders and power allocation schemes to optimize various network utility functions such as SUM-SE, product of signal-to-interference-plus-noise ratio (SINR), and Minimum SE per user. The numerical findings demonstrated that RSMA outperforms the conventional linearly precoded massive MIMO transmission strategy consistently, exhibiting increased robustness against pilot contamination and achieving an equal or improved SE performance.

Deep learning techniques have been finding many applications in wireless systems. For example, in [14] and [15] a deep reinforcement learning (DRL) based algorithm named deep deterministic policy gradient (DDPG) has been used for controlling the phase shifts in a reconfigurable intelligent surface (RIS)-assisted wireless system with the aim of maximizing the DL sum-rate. In [16], the DDPG is also employed for phase shift control in RIS-assisted wireless system but seeking to maximize the received signal-to-noise ratio (SNR).

In this paper, we propose a DDPG-based method for allocating power in an RSMA framework operating within a massive MIMO system. Our goal is to maximize the sum spectral efficiency (SUM-SE) in a scenario where all user equipments (UEs) share the same pilot sequence for uplink (UL) training.

979-8-3503-8741-4/24/$31.00 ©2024 IEEE

The rest of the paper is organized as follows. In Section II, the system model is introduced. In Section III, SE expressions for the common and private streams, and the precoder are derived. In Section IV, the SUM-SE maximization problem is formulated and the proposed DDPG-based algorithm is presented. In Section V, numerical results are provided. Final remarks are presented in Section VI.

### A. Notation

In this paper, matrices are denoted by boldface uppercase letters, column vectors are denoted by boldface lowercase letters, sets are denoted by calligraphic letters, and scalars are denoted by standard letters. The trace of matrix $\mathbf{A}$ is denoted by $\mathrm{tr}(\mathbf{A})$. $\mathbf{A}^T$ and $\mathbf{A}^H$ denote the transpose and Hermitian operators on matrix $\mathbf{A}$, respectively. The Euclidean norm of vector $\mathbf{a}$ is denoted as $||\mathbf{a}||$ and the cardinality of a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. $\mathbb{E}_X\{Y\}$ represents the expectation of $Y$ with respect to the random variable $X$. $\mathbb{C}^{M\times N}$ and $\mathbb{R}^{M\times N}$ denote the sets of all $M \times N$-dimensional matrices with complex-valued and real-valued entries, respectively. A uniform distribution from $a$ to $b$ is denoted by $\mathcal{U}(a, b)$. A real Gaussian distribution with mean $\mu$ and variance $\sigma^2$ is denoted as $\mathcal{N}\left(\mu, \sigma^2\right)$, whereas a circularly symmetric complex Gaussian (CSCG) distribution with mean $\mu$ and variance $\sigma^2$ is denoted as $\mathcal{CN}\left(\mu, \sigma^2\right)$.

## II. System model

Consider a single-cell scenario where a base station (BS), equipped with $M$ antennas, is located at the center. The BS operates in TDD mode to serve $K$ single-antenna UEs in the same time-frequency resource block. The channel between each UE and the BS is represented by $\mathbf{g}_k \in \mathbb{C}^M$ and follows the standard Rayleigh fading model for each $k \in \mathcal{K}$, where $\mathcal{K} = \{1, 2, ..., K\}$ is the set of all UEs in the cell with cardinality $|\mathcal{K}| = K$.

The channel can be described as

$$\mathbf{g}_k = \sqrt{\beta_k}\mathbf{h}_k \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{R}_k\right), \tag{1}$$

where $\mathbf{R}_k \in \mathbb{C}^{M\times M}$ represents the spatial covariance matrix, which is assumed to be known by the BS. It is assumed that $\mathrm{tr}(\mathbf{R}_k) > 0$. $\beta_k = \frac{1}{M}\mathrm{tr}(\mathbf{R}_k)$ denotes the average large-scale fading coefficient, and $\mathbf{h}_k \in \mathbb{C}^M$ represents the small-scale fading coefficients between the $k$-th UE and the BS. The average large-scale fading parameter path loss $\beta_k$ for the $k$-th UE is expressed in decibels (dB) as follows: $\beta_k = \Gamma - 10\delta\log_{10}(\frac{d_k}{1\text{ km}}) + S_k$. In this equation, $d_k$ represents the distance of the $k$-th UE to the BS in kilometers, the path loss exponent $\delta$ determines the rate of signal power decay, and $\Gamma$ is the channel gain at a distance of 1 km. Additionally, $S_k \in \mathcal{N}(0, \sigma_s^2)$ denotes the shadow fading coefficient [17, 8, eq 2.3]. The small-scale fading variations, denoted by $\mathbf{h}_k$, are modeled using a CSCG distribution. On the other hand, the large-scale fading property of $\mathbf{R}_k$ incorporates the influences of both path loss and shadowing. It can be described by the following model

$$[\mathbf{R}_k]_{m_1,m_2} = \beta_k \times \frac{1}{S}\sum_{s=1}^{S} e^{i\pi(m_1-m_2)\sin(\varphi_{k,s})} \times e^{-\frac{\sigma_\varphi^2}{2}(\pi(m_1-m_2)\cos(\varphi_{k,s}))^2}, \tag{2}$$

where the value of $S$ is the number of clusters following the Gaussian scattering model in reference [13]. The geographical angle to the $k$-th UE from the BS is denoted as $\varphi_k$. Each cluster $s$ is characterized by a randomly generated angle-of-arrival, $\varphi_{k,s}$, distributed uniformly in the range $\mathcal{U}(\varphi_k - 40^\circ, \varphi_k + 40^\circ)$, while the angles of multipath components are distributed around their corresponding nominal angles with a standard deviation of $\sigma_\varphi$ [17, Sec. 2.6].

### A. Channel Estimation

We consider that each coherence block consists of $\tau_c = \tau_p + \tau_u + \tau_d$ channel uses, where $\tau_p$ are used for UL pilot transmission, $\tau_u$ for UL data transmission, and $\tau_d$ for DL data transmission. We assume that all UEs send the same pilot sequence $\boldsymbol{\phi} \in \mathbb{C}^{\tau_p}$, with $||\boldsymbol{\phi}|| = 1$, causing pilot contamination. As a result, the BS receives the matrix $\mathbf{Y}^p \in \mathbb{C}^{M\times\tau_p}$

$$\mathbf{Y}^p = \sum_{k=1}^{K}\sqrt{\rho^{tr}}\mathbf{g}_k\boldsymbol{\phi}^T + \mathbf{N}^p, \tag{3}$$

where $\mathbf{N}^p \in \mathbb{C}^{M\times\tau_p}$ is the receiver noise matrix, with all elements being independently distributed and following the distribution $\mathcal{CN}(0, \sigma_{ul}^2)$, and $\rho^{tr}$ is the transmit power. The received matrix $\mathbf{Y}^p$ is then correlated with the pilot sequence $\boldsymbol{\phi}$

$$\mathbf{Y}^p\boldsymbol{\phi}^* = \left(\sum_{k=1}^{K}\sqrt{\rho^{tr}}\mathbf{g}_k\boldsymbol{\phi}^T + \mathbf{N}^p\right)\boldsymbol{\phi}^*, \tag{4}$$

$$\mathbf{Y}^p\boldsymbol{\phi}^* = \sum_{k=1}^{K}\sqrt{\rho^{tr}}\mathbf{g}_k + \mathbf{n}_{t,k}, \tag{5}$$

where $\mathbf{n}_{t,k} \sim \mathcal{CN}(0, \sigma_{ul}^2\mathbf{I}_M)$. The channel $\mathbf{g}_k$ is then estimated, as computed in [17, Sec 3.2], through the MMSE estimator as

$$\widehat{\mathbf{g}}_k = \mathbf{R}_k\mathbf{Q}^{-1}\left(\mathbf{Y}^p\boldsymbol{\phi}^*\right) \sim \mathcal{CN}\left(\mathbf{0}, \boldsymbol{\Phi}_k\right), \tag{6}$$

where $\mathbf{Q} = \sum_{i\in\mathcal{K}}\mathbf{R}_i + \frac{\sigma_{ul}^2}{\rho^{tr}}\mathbf{I}_M$ and $\boldsymbol{\Phi}_k = \mathbf{R}_k\mathbf{Q}^{-1}\mathbf{R}_k$. To facilitate DL transmission, the 1-layer rate-splitting (RS) strategy is utilized. This strategy involves dividing the message $W_k$ intended for the $k$-th UE into two parts: a common part $W_{c,k}$ and a private part $W_{p,k}$. The common parts of all UEs are aggregated to create a unified common message $W_c$, which is encoded into a single common stream $s_c \in \mathbb{C}$, where $\mathbb{E}\{|s_c|^2\} = 1$, to be decoded by all UEs. Each UE's private message component is independently encoded into a private stream $s_k \in \mathbb{C}$ with $\mathbb{E}\{|s_k|^2\} = 1$, ensuring this property holds for all $k \in \mathcal{K}$. The purpose of this encoding is to enable the corresponding UE to decode its specific private stream exclusively. The resulting transmitting signal is given by

$$\mathbf{x} = \sqrt{\rho_c}\mathbf{w}_c s_c + \sum_{k=1}^{K} \sqrt{\rho_k}\mathbf{w}_k s_k, \tag{7}$$

where $\mathbf{w}_c \in \mathbb{C}^M$ is the common precoder, with $\mathbb{E}\{\|\mathbf{w}_c\|^2\} = 1$ and $\mathbf{w}_k \in \mathbb{C}^M$ is the private stream precoder, with $\mathbb{E}\{\|\mathbf{w}_k\|^2\} = 1, \forall k \in \mathcal{K}$. The power allocated to the common stream is denoted by $\rho_c$ and the power allocated to the private stream is denoted by $\rho_k$. The power allocated to the common stream and private streams must obey the constraint

$$\rho_c + \sum_{k=1}^{K} \rho_k \leq \rho_{\mathrm{dL}}, \tag{8}$$

where $\rho_{\mathrm{dL}}$ is the total amount of power reserved by the BS for DL transmission. The signal received by the $k$-th UE is given by

$$y_k = \sqrt{\rho_c}\mathbf{g}_k^H \mathbf{w}_c s_c + \sum_{i=1}^{K} \sqrt{\rho_i}\mathbf{g}_k^H \mathbf{w}_i s_i + n_k, \tag{9}$$

where $n_k \sim \mathcal{CN}(0, \sigma_{n,k}^2)$ is the noise perceived by the UE. The signal $y_k$ received by the $k$-th UE undergoes a decoding process that involves two main steps. Firstly, the common stream is decoded into $\widehat{W_c}$ by considering all private parts as noise. Subsequently, a successive interference cancellation technique is employed to remove the decoded common part from $y_k$. This allows the private part corresponding to the $k$-th UE to be decoded into $\widehat{W_{p,k}}$, treating the private parts of other UEs as noise during this decoding process. Finally, the $k$-th UE reconstructs its message by retrieving $\widehat{W_{c,k}}$ from the common stream $\widehat{W_c}$ and then combining it with $\widehat{W_{p,k}}$ to form $\widehat{W_k}$. Only the mean effective channel gains are known in such process, while the deviations around the mean are seen as interference.

## III. Spectral efficiency

Due to the limited knowledge of the channel at the UEs, accurately characterizing the DL SE for both the common and private streams becomes challenging [13]. Hence, the lower bound of the ergodic capacity of the common and private parts are calculated in (10) and (11), respectively as

$$\mathrm{SE}_{c,k} = \frac{\tau_d}{\tau_c} \log\left(1 + \gamma_{c,k}\right), \tag{10}$$

and

$$\mathrm{SE}_{p,k} = \frac{\tau_d}{\tau_c} \log\left(1 + \gamma_{p,k}\right), \tag{11}$$

where $\gamma_{c,k}$ and $\gamma_{p,k}$ are defined in (14) and in (15) at the top of the next page, respectively, and are the effective DL SINR lower bounds of the common and private streams perceived by the $k$-th UE. The achievable SE for the common stream, $\mathrm{SE_c}$, is defined as

$$\mathrm{SE}_c = \frac{\tau_d}{\tau_c} \log\left(1 + \gamma_c\right) = \sum_{k=1}^{K} C_k, \tag{12}$$

where $\gamma_c = \min_{k \in \mathcal{K}} \gamma_{c,k}$ and $C_k$ is the part of the common SE intended to the $k$-th UE. The total SE of the $k$-th UE is given as

$$\mathrm{SE}_k = \mathrm{SE}_{p,k} + C_k, \quad \forall k \in \mathcal{K}. \tag{13}$$

### A. Precoder design

Due to its computational tractability, the Maximum-Ratio (MR) precoder is employed in the private stream of the $k$-th UE, and is expressed as

$$\mathbf{w}_k = \frac{\widehat{\mathbf{g}}_k}{\sqrt{\mathbb{E}\left\{\|\widehat{\mathbf{g}}_k\|^2\right\}}} = \frac{\widehat{\mathbf{g}}_k}{\sqrt{\mathrm{tr}\left(\mathbf{\Phi}_k\right)}}. \tag{16}$$

The design of the common precoder $\mathbf{w}_c$ aims to solve the following max-min problem

$$\max_{\mathbf{w}_c} \quad \min_{k} \gamma_{c,k} \tag{17a}$$

$$\text{s.t.} \quad \mathbb{E}\{\|\mathbf{w}_c\|^2\} = 1. \tag{17b}$$

It can be transformed into a convex problem and solved by considering the weighted MR approach proposed in [13] as

$$\max_{\mathbf{c}, t > 0} \quad t \tag{18a}$$

$$\text{s.t.} \quad \mathbf{c}^T \mathbf{U}(:, k) \geq t, \forall k \in \mathcal{K},, \tag{18b}$$

$$\|\mathbf{c}\|^2 \leq 1. \tag{18c}$$

where $\mathbf{c} = [c_1, \ldots, c_K]^T$, and $\mathbf{U}(i, k) = \mathrm{tr}\left(\mathbf{R}_i \mathbf{Q}^{-1} \mathbf{R}_k\right)$.

The problem presented in (18) can be solved either using CVX or the CVXPY package from Python. In this work, we use the CVXPY with the solver ECOS. For the considered scenario where all UEs use the same pilot for UL channel estimation, the common precoder is expressed as

$$\mathbf{w}_c^* = \frac{\sum_{i=1}^{K} c_i^* \widehat{\mathbf{g}}_i}{\sqrt{\sum_{i=1}^{K} \sum_{j=1}^{K} c_i^* c_j^* \,\mathrm{tr}\left(\mathbf{R}_i \mathbf{Q}^{-1} \mathbf{R}_j\right)}}. \tag{19}$$

## IV. Maximizing The SUM-SE (MaxSum-SE)

For any given channel estimation technique and precoding scheme, the SUM-SE with the RS transmission strategy can be written as

$$\bar{\mathrm{S}} = \mathrm{SE}_c + \sum_{k=1}^{K} \mathrm{SE}_{p,k}. \tag{20}$$

Let $\boldsymbol{\rho} = [\rho_c, \rho_1 \ldots \rho_K]$, the MaxSum-SE problem can therefore be formulated as

$$\max_{\boldsymbol{\rho}} \quad \mathrm{SE}_c(\boldsymbol{\rho}) + \sum_{k=1}^{K} \mathrm{SE}_{p,k}(\boldsymbol{\rho}) \tag{21a}$$

$$\text{s.t.} \quad \rho_c + \sum_{i=1}^{K} \rho_i \leq \rho_{\mathrm{dL}}. \tag{21b}$$

To find the optimal power allocation between the common and private streams that maximizes the SUM-SE, the authors of [13] proposed a heuristic and low-complexity algorithm, referred to in the sequel as benchmark, for power allocation to maximize the SUM-SE.

$$\gamma_{c,k} = \frac{\rho_c \left|\mathbb{E}\left\{\mathbf{g}_k^H \mathbf{w}_c\right\}\right|^2}{\sum_{i=1}^K \rho_i \mathbb{E}\left\{\left|\mathbf{g}_k^H \mathbf{w}_i\right|^2\right\} + \rho_c \left(\mathbb{E}\left\{\left|\mathbf{g}_k^H \mathbf{w}_c\right|^2\right\} - \left|\mathbb{E}\left\{\mathbf{g}_k^H \mathbf{w}_c\right\}\right|^2\right) + \sigma_n^2}, \tag{14}$$

$$\gamma_{p,k} = \frac{\rho_k \left|\mathbb{E}\left\{\mathbf{g}_k^H \mathbf{w}_k\right\}\right|^2}{\sum_{i=1}^K \rho_i \mathbb{E}\left\{\left|\mathbf{g}_k^H \mathbf{w}_i\right|^2\right\} - \rho_k \left|\mathbb{E}\left\{\mathbf{g}_k^H \mathbf{w}_k\right\}\right|^2 + \rho_c \left(\mathbb{E}\left\{\left|\mathbf{g}_k^H \mathbf{w}_c\right|^2\right\} - \left|\mathbb{E}\left\{\mathbf{g}_k^H \mathbf{w}_c\right\}\right|^2\right) + \sigma_n^2}. \tag{15}$$

### *A. Proposed Solution*

Although the benchmark approach is a simple solution, it is worth noting that it always allocate the same private power share for each individual UE. Therefore, we propose to employ a DRL-based method to find the optimal power allocation policy.

In the proposed DRL-based solution, at each time step $t$, an agent observes the current state $\mathbf{x}_t$, takes an action $\mathbf{a}_t$ based on the policy $\pi$, receives a reward $r_t$, and transitions to a new state $\mathbf{x}_{t+1}$.

We define the state-space at time step $t$ by the vector

$$\mathbf{x}_t = [r_{t-1}, \mathbf{a}_{t-1}, \beta_1 \ldots \beta_K, \varphi_1 \ldots \varphi_K] + \kappa, \tag{22}$$

where $\mathbf{a}_{t-1}$ and $r_{t-1}$ are the action and reward at step $t-1$, respectively, $\varphi_k$ is the angle of the $k$-th UE in relation to the BS, and $\kappa \in \mathbb{R}^{3K+2}$, which follows the distribution $\mathcal{N}(0, \vartheta)$, is a random exploratory noise added to the input state. The action space at time step $t$ is expressed as $\mathbf{a}_t = [\rho_c, \rho_1, \rho_2, \ldots, \rho_K]$, and the reward at time step $t$ is given by $\bar{\mathrm{S}}$.

Since our action space is continuous, we employ a policy gradient based algorithm [15]. The goal of the reinforcement learning agent is to learn a policy that maximizes the expected cumulative discounted reward from the start state. The proposed algorithm uses the actor-critic approach with two evaluation deep neural network (DNN) models: actor $\mu\left(\mathbf{x}_t \mid \boldsymbol{\theta}_\mu\right)$, and critic, $Q\left(\mathbf{x}_t, \mathbf{a}_t \mid \boldsymbol{\theta}_q\right)$, where $\boldsymbol{\theta}$ represents the DNN parameters. The actor receives the state as input and produces the action $\mathbf{a}_t$ based on the state using the parameterized function $\mu(\mathbf{x}_t|\theta_\mu)$, then a random exploratory epsilon-greedy ($\epsilon$-greedy) process $\xi \in \mathbb{R}^{K+1}$ is added to the action $\mathbf{a}_t$ to encourage action exploration. The critic, on the other hand, takes both the state $\mathbf{x}_t$ and action $\mathbf{a}_t$ as input and outputs the Q-value, which represents the evaluation network's estimation.

The $\epsilon$-greedy exploratory process is defined as follows

$$\xi = \begin{cases} x \sim \mathcal{N}(0, \sigma_\xi^2), & \text{if } p < \epsilon, \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

where $p \sim \mathcal{U}(0,1)$ and $\epsilon$ is the probability of adding $\xi$ to the action $\mathbf{a}_t$. The $\epsilon$ is initialized as 1 and is updated at each time step $t$ by the decay factor $\eta$

$$\epsilon_t = \eta \epsilon_{t-1}, 0 \leq \eta \leq 1. \tag{24}$$

At the start of the DRL algorithm, four networks are created: the target actor network ($\mu'$), the target critic network ($q'$), and their corresponding evaluation networks. These target networks are initialized by making exact copies of the actor and critic evaluation networks, $\mu'\left(\mathbf{x}_t \mid \boldsymbol{\theta}_{\mu'}\right)$ and $Q'\left(\mathbf{x}_t, \mathbf{a}_t \mid \boldsymbol{\theta}_{q'}\right)$. Additionally, an experience replay memory $\mathcal{D}$ is constructed to minimize correlation among the training samples.

During each episode, the agent obtains the channel state information. Using the actor network, the agent generates an action $\mathbf{a}_t$, calculates the reward $r_t$, and transitions to the next state $\mathbf{x}_{t+1}$. The experience tuple $e_t$, defined as $(\mathbf{x}_t, \mathbf{a}_t, r_t, \mathbf{x}_{t+1})$, is stored in the experience replay memory $\mathcal{D}$.

The critic evaluation network samples a minibatch of transitions $\mathcal{N}_\mathcal{B} \subseteq \mathcal{D}$ with cardinality $N_B = |\mathcal{N}_\mathcal{B}|$ to calculate the target value $\Upsilon_j$ for each transition. The target value is computed as

$$\Upsilon_j = r_j + \lambda Q'(\mathbf{x}_{j+1}, \mu'(\mathbf{x}_{j+1} \mid \boldsymbol{\theta}_{\mu'}) \mid \boldsymbol{\theta}_{q'}), \tag{25}$$

where $\lambda$ is the discount factor.

The set of experience tuples $\mathcal{N}_\mathcal{B}$ is sampled randomly from $\mathcal{D}$ where each entry $e_t \in \mathcal{D}$ is chosen with probability

$$p_t = \begin{cases} \frac{P_t}{\sum_{j \in D} P_j}, & \text{if } p < 0.5 \\ \frac{1}{D}, & \text{otherwise} \end{cases}, \tag{26}$$

where $p \sim \mathcal{U}(0,1)$, $D = |\mathcal{D}|$ is the cardinality of the set $\mathcal{D}$, and $P_t$ is calculated as

$$P_t = u_g(t-1) + u_g/2, \tag{27}$$

where $u_g$ is an unbalancing gap which holds the property $0 < u_g \leq 1$ and can be adjusted to increase the probability of choosing the most recently stored experience tuples when its value is near 1 or towards a uniform distribution when its value is near 0. In summary, when $p > 0.5$ the most recent stored experience tuples are prioritized and when $p \leq 0.5$, the choice is done uniformly among all the stored entries.

The parameters of the critic and actor evaluation networks, $\boldsymbol{\theta}_q$ and $\boldsymbol{\theta}_\mu$, are updated using stochastic gradient descent and policy gradient, with learning rates $c_{lr}$ and $a_{lr}$, respectively.

The critic and actor networks are updated using stochastic gradient descent (SGD) and policy gradient, respectively. The critic network is updated by minimizing the loss function given in the following equation

$$L(\boldsymbol{\theta}_q) = \frac{1}{N_B} \sum_j \left(\Upsilon_j - Q\left(\mathbf{x}_j, \mathbf{a}_j \mid \boldsymbol{\theta}_q\right)\right)^2. \tag{28}$$

TABLE I
NUMERICAL PARAMETERS.

| Parameter | Value | Description |
|---|---|---|
| $M$ | [20,100] | Number of BS antennas |
| $K$ | 8 | Number of UEs |
| $\rho^{tr}$ | 10 dBm | Transmit power of the UEs |
| $\rho_{\mathrm{dl}}$ | 20 dBm | Downlink power |
| $c_r$ | 125 m | Cell radius |
| $\sigma_n^2$ | -94 dBm | Noise power |
| $\sigma_{ul}^2$ | -94 dBm | Noise power |
| $\Gamma$ | -148.1 | Channel gain at 1 km |
| $\delta$ | 3.76 | Path loss exponent |
| $\sigma_s^2$ | 16 | Shadow fading variance |
| $\tau$ | 0.002 | Target network learning rate |
| $c_{lr}$ | 0.002 | Critical network learning rate |
| $a_{lr}$ | 0.001 | Actor network learning rate |
| $\tau_c$ | 200 | Number of resource blocks |
| $\tau_p$ | 20 | Length of pilot sequence |
| $\tau_d$ | 180 | Number of resource blocks reserved for data |
| $\vartheta$ | 0.005 | Variance of $\kappa$ |
| $\sigma_\xi^2$ | 0.3 | Variance of $\xi$ |
| $\eta$ | 0.9981 | Decay of $\epsilon$ |
| $\lambda$ | 0.99 | Discount factor |
| $u_g$ | 0.8 | Unbalacing gap |
| $N$ | 400 | Number of Episodes |
| $T$ | 500 | Number of Steps per episode |
| $N_B$ | 100 | Batch size |
| $D$ | 50000 | Replay buffer size |

The actor network is updated using the following policy gradient:

$$\nabla_{\boldsymbol{\theta}_\mu} = \frac{1}{N_B} \sum_j \nabla_a Q\left(s, a \mid \boldsymbol{\theta}_q\right) \nabla_{\boldsymbol{\theta}_\mu} \mu\left(s \mid \boldsymbol{\theta}_\mu\right) |_{s=\mathbf{x}_j, a=\mu(\mathbf{x}_j)} \tag{29}$$

Finally, the target network parameters are updated using a soft update coefficient $\tau$ in the following manner

$$\begin{aligned} \boldsymbol{\theta}_{q'} &\longleftarrow \tau \boldsymbol{\theta}_q + (1-\tau)\boldsymbol{\theta}_{q'}, \\ \boldsymbol{\theta}_{\mu'} &\longleftarrow \tau \boldsymbol{\theta}_{\mu'} + (1-\tau)\boldsymbol{\theta}_{\mu'}. \end{aligned} \tag{30}$$

This process is repeated for $N$ episodes and $T$ steps until convergence is achieved. At the beginning of each episode, the environment is reset to its initial state. The structure of the proposed DRL algorithm is summarized in Algorithm 1.

The proposed DNN models are feedforward fully connected networks with input, hidden, and output layers. The actor network consists of an input layer, two hidden layers that utilize the ReLU activation function, and an output layer with softmax activation. The critic network has two input layers, one for the state-space and the other for the action. There are two hidden layers after each input layer with ReLU activation, and a concatenation layer that combines the result of the last hidden layer of each input. The concatenated input then goes through two hidden layers before reaching the output layer with linear activation.

**Algorithm 1** DRL-based solution

```
Initialize:
    $\boldsymbol{\theta}_\mu$ and $\boldsymbol{\theta}_q$ with random weights, $\mathcal{D}$, $\lambda$, $\tau$, $\eta$,
    $\vartheta$, $\sigma_\epsilon$, $\epsilon$, $c_{lr}$, $a_{lr}$, $u_g$, $\boldsymbol{\theta}_{\mu'} \leftarrow \boldsymbol{\theta}_\mu$ and
    $\boldsymbol{\theta}_{q'} \leftarrow \boldsymbol{\theta}_q$.
    Initialize $\kappa$.
    Randomly select UEs location.
    Observe the initial state $\mathbf{x}_1$ as expressed in (22).
Iterate for N episodes:
    Initialize $\rho_c = \frac{\rho_{\mathrm{dL}}}{2}$ and $\rho_k = \frac{\rho_{\mathrm{dL}-\rho_c}}{K}, \forall k \in \mathcal{K}$.
    Initialize $\xi$ according to (23).
  Iterate for T steps:
      Extract $\mathbf{a}'_t = \mu(\mathbf{x}_t|\boldsymbol{\theta}_\mu) + \xi$ from the actor network.
      To meet the constraints in (21), make $\mathbf{a}_t = \frac{\mathbf{a}'_t}{\sum_j a_t^j}\rho_{\mathrm{dl}}$,
      where $a_t^j$ is the $j$-th element of $\mathbf{a}'_t$ .
      Observe new state, $\mathbf{x}_{t+1}$, given $\mathbf{a}_t$.
      Keep storing ($\mathbf{x}_t$, $\mathbf{a}_t$, $r_t$, $\mathbf{x}_{t+1}$) in $\mathcal{D}$ until it has $N_B$
      elements.
      If $\mathcal{D}$ has at least $N_B$ elements, sample a minibatch
      of transitions $(\mathbf{x}_j, \mathbf{a}_j, r_j, \mathbf{x}_{j+1})$ from it according to
      (26) and do:
        Compute the target value using (25).
        Update the critic with by minimizing the loss
        function in (28).
        Update the actor using the policy gradient in (29).
        Update the target NNs using (30).
Output: Optimal power allocation policy for the given
        scenario.
```

## V. NUMERICAL RESULTS

In this section, we present the results of our proposed Deep Deterministic Policy Gradient (DDPG)-based SUM-SE maximization method applied in a massive MIMO system, implementing the RSMA framework, where all users share a single pilot. The parameters used in the system simulation and the DDPG algorithm are presented in Table I.

The DDPG-based framework is trained for a set of $K$ UEs fixed at the same positions during the entire training process. The $k$-th UE has its distance from the BS initialized randomly following the uniform distribution $\mathcal{U}(6.25, 125)$, and its angle $\varphi_k$ is also initialized from a uniform distribution $\mathcal{U}(0, \pi/4)$.

Both the actor evaluation network and the critic evaluation network utilize the Adam optimizer to update their parameters. The actor network consists of an input layer with $3K + 2$ neurons and an output layer with $K + 1$ neurons. Between them, there are two hidden layers with 800 and 600 neurons, respectively, followed by a ReLU activation function. The output layer of the actor network employs the softmax function. The critic network consists of an input layer with $4K + 3$ neurons and an output layer with 1 neuron with linear activation. The input layer receives the state vector $\mathbf{x}_t$ and the action vector $\mathbf{a}_t$ as inputs. Before concatenating them, both vectors pass through independent layers. The state vector $\mathbf{x}_t$ passes through two layers with 32 and 128 neurons, respectively, both followed by ReLU activation. The action vector $\mathbf{a}_t$ passes through two layers with 128 neurons each, also followed by ReLU activation. Then, the results are concatenated and pass through two layers with 800 and 600 neurons, respectively, followed by ReLU activation.

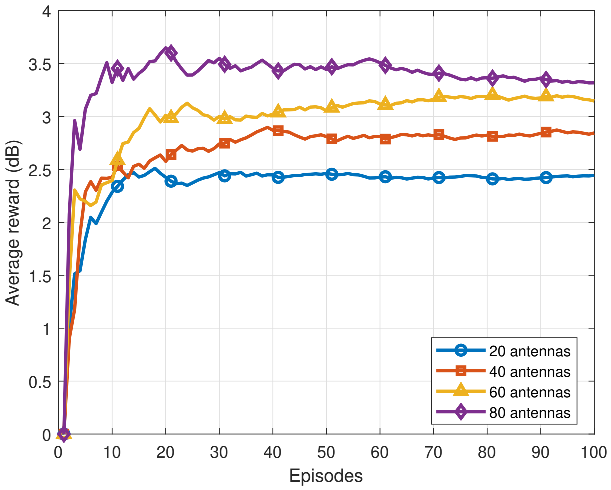

Fig. 1. SUM-SE moving average for $K = 8$.

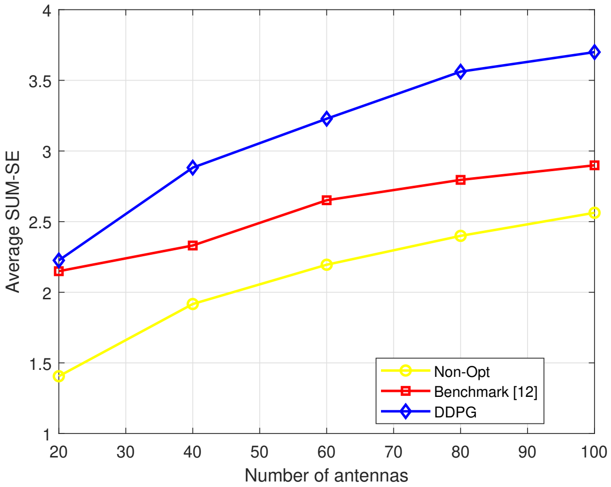

Fig. 2. Average SUM-SE vs. $M$

Figure 2 shows the comparative performance using the proposed DDPG-based approach, the benchmark, and without applying any optimization technique (Non-Opt). For all three cases, the common precoder is optimized through the weighted MR approach [13]. It can be observed that, for 20 antennas, the achieved average SUM-SE is $1.247$ bits/s/Hz without optimization, $3.054$ bits/s/Hz with the benchmark, and $3.271$ bits/s/Hz using the DDPG. The average relative gain of the DDPG-based method compared to the benchmark [13] is about $6\%$, and it is $124\%$ in relation to the Non-Opt case. These results were obtained by averaging the outcomes of 30 setups, while keeping the UEs at the same positions for both the proposed DDPG-based algorithm and the benchmark. These results indicate that the DDPG can be a valuable tool for optimizing the SUM-SE in massive MIMO RSMA scenarios as it considers each UE individually when optimizing the private power allocation.

## VI. Final Remarks

In this work, we have shown that the proposed DDPG-based method can effectively maximize the SUM-SE in RSMA scenarios with shared pilots among all UEs. Our results are competitive with the benchmark method, achieving a remarkable $124\%$ increase in the SUM-SE compared to non-optimized scenarios and a $6\%$ improvement over the benchmark method. Future research will consider multicell scenarios with dynamic UE position, and explore different DRL-based approaches.

## References

[1] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, pp. 615–637, Mar. 2021.

[2] M. Moussaoui, E. Bertin, and N. Crespi, "5G shortcomings and beyond-5G/6G requirements," in *2022 1st International Conference on 6G Networking (6GNet)*, pp. 1–8, Jul. 2022.

[3] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 574–590, Jan. 2018.

[4] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 2220–2234, Apr. 2017.

[5] J. C. Marinello, T. Abrão, R. D. Souza, E. de Carvalho, and P. Popovski, "Achieving fair random access performance in massive MIMO crowded machine-type networks," *IEEE Wireless Commun. Lett.*, vol. 9, pp. 503–507, Apr. 2020.

[6] J. C. Marinello and T. Abrão, "Collision resolution protocol via soft decision retransmission criterion," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 4094–4097, Apr. 2019.

[7] F. A. D. Bueno, C. F. Yamamura, A. Goedtel, and J. C. Marinello Filho, "A random access protocol for crowded massive MIMO systems based on a bayesian classifier," *IEEE Wireless Commun. Lett.*, vol. 11, pp. 2455–2459, Nov. 2022.

[8] L. M. Bello, P. Mitchell, and D. Grace, "Application of Q-learning for RACH access to support M2M traffic over a cellular network," in *European Wireless 2014; 20th European Wireless Conference*, pp. 1–6, May 2014.

[9] L. Bai, J. Liu, Q. Yu, J. Choi, and W. Zhang, "A collision resolution protocol for random access in massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 39, pp. 686–699, Mar. 2021.

[10] J. Ding, D. Qu, H. Jiang, and T. Jiang, "Success probability of grant-free random access with massive MIMO," *IEEE Internet Things J.* , vol. 6, pp. 506–516, Feb. 2019.

[11] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: a promising PHY-layer strategy for LTE evolution," *IEEE Commun. Mag.*, vol. 54, pp. 98–105, May 2016.

[12] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Trans. Wireless Commun.*, vol. 64, pp. 4847–4861, Nov. 2016.

[13] A. Mishra, Y. Mao, C. K. Thomas, L. Sanguinetti, and B. Clerckx, "Mitigating intra-cell pilot contamination in massive MIMO: A rate splitting approach," *IEEE Trans. Wireless Commun.*, vol. 22, pp. 3472–3487, May 2023.

[14] A. Faisal, I. Al-Nahhal, O. A. Dobre, and T. M. N. Ngatched, "Deep reinforcement learning for optimizing RIS-assisted HD-FD wireless systems," *IEEE Commun. Lett.*, vol. 25, pp. 3893–3897, Dec 2021.

[15] A. Faisal, I. Al-Nahhal, O. A. Dobre, and T. M. N. Ngatched, "Deep reinforcement learning for RIS-assisted FD systems: single or distributed RIS?," *IEEE Commun. Lett.*, vol. 26, pp. 1563–1567, Jul. 2022.

[16] K. Feng, Q. Wang, X. Li, and C.-K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, pp. 745–749, May 2020.

[17] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, pp. 154–655, Jan. 2017.