A DYNAMIC Q-LEARNING BEAMFORMING METHOD FOR INTER-CELL INTERFERENCE MITIGATION IN 5G MASSIVE MIMO NETWORKS

Aidong Yang, Ph.D[1], Xinlang Yue[1], Mohan Wu, Ph.D[1], Ye Ouyang, Ph.D[1]
[1]Telecom Artificial Intelligence Lab, AsiaInfo Technologies, Beijing, China

NOTE: Corresponding author: Aidong Yang, Ph.D, yangad@asiainfo.com

Abstract – *Beamforming is an essential technology in 5G Massive Multiple-Input Multiple-Output (MMIMO) communications, which are subject to many impairments due to the nature of wireless transmission channel. The Inter-Cell Interference (ICI) is one of the main obstacles faced by 5G communications due to frequency-reuse technologies. However, finding the optimal beamforming parameter to minimize the ICI requires infeasible prior network or channel information. In this paper, we propose a dynamic Q-learning beamforming method for ICI mitigation in the 5G downlink that does not require prior network or channel knowledge. Compared with a traditional beamforming method and other industrial Reinforcement Learning (RL) methods, the proposed method has lower computational complexity and better convergence efficiency. Performance analysis shows the quality of service improvement in terms of Signal-to-Interference-plus-Noise-Ratio (SINR) and the robustness towards different environments.*

Keywords – 5G beamforming, inter-cell interference, massive MIMO, reinforcement learning

1. INTRODUCTION

Massive Multiple-Input Multiple-Output (MMIMO) technology in 5G is a competent solution that significantly improves system capacity, signal coverage and spectral-efficiency by configuring hundreds of Antenna Elements (AEs) at the Base Station (BS) to shape effective beamforming [1, 2]. However, the quality of MMIMO beamforming depends on accurate Channel State Information (CSI), pilot contamination and ICI estimation [3]. Moreover, the MMIMO beamforming complexity becomes a challenge as the number of AEs at the BS increases. Therefore, it is necessary to explore an effective and efficient beamforming method for ICI mitigation with low power and low complexity [4].

In recent years, the accurate MMIMO beamforming has attracted extensive research [3, 4, 5, 6, 7], which almost follow two main directions: with and without CSI. Hybrid beamforming [3, 4, 5] is the representative of the former. It aims to reduce the expense of Radio Frequency (RF) chains and decrease the complexity of beamforming compared to conventional methods [2], but it needs to update beams frequently when pilots are received continually at the BS. A smart pilot assignment scheme, which is effective in mitigating interference but is aimed at a single cell, is proposed in [5] to reduce pilot contamination by smartly assigning orthogonal pilots to users. The latter mainly includes the first Monte Carlo (MC) method, which searches the optimal beamforming parameters but suffers from increasing computational complexity, and second supervised Deep Learning (DL) methods. One of them is reported in [7] to research the characters of wireless spatial channels and explore preferable pilot assignments for better channel estimation and beamforming, but supervised methods require model training beforehand and time-consuming sample data collection.

In this paper, an RL-assisted full dynamic beamforming method is developed to efficiently acquire the optimal beamforming parameters in the MMIMO system to address ICI issues. We fully consider the microcell and macro-cell multi-path transmission channels which present radio features with high user density and traffic loads focusing on pedestrian and vehicular users (Dense Urban-eMBB) scenarios [1, 2, 5], such as buildings, mountains and rivers, where the distribution of User Equipments (UEs) changes infrequently; these factors significantly impact coverage. To get optimal beamforming, firstly, we utilize a Poisson Point distribution model to estimate the occurrences of UE in the target cells with a long-term data statistical analysis; secondly, we apply an algorithm to fast search through huge volumes of parameters and obtain optimal values. Lastly, we send the optimal parameters into the BS beamforming simulator for the best SINR.

In summary, the main contribution of this work includes:

- The proposed RL beamforming method for an MMIMO system is meant to get the optimal beamforming parameter, such a method with multi-cell ICI is rarely discussed in literature. Besides, it does not need any prior network or channel information and it works for different UE distribution.
- Compared with the traditional beamforming method and other industrial RL methods, the proposed dynamic Q-learning method shrinks the action space during its process, thus it requires less time and computational complexity to operate.
- As proven in many simulation results, the proposed method performs better than the other methods. Moreover, it is robust to various starting states and different environments.

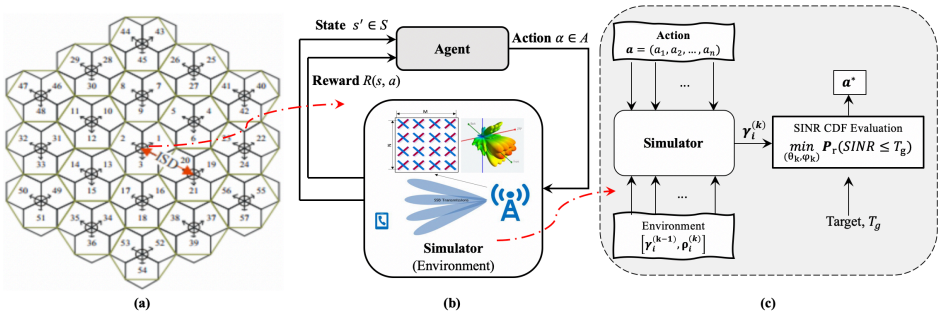


Fig. 1 – Illustration of the proposed RL-based beamforming (b) for MMIMO systems and the network layout (a) of 5G dense Urban-eMBB cells, in which the BS of target small cell #0 with $N_0(k)$ mobile users chooses the optimal beamforming $a_0^{*(k)}$ (c) to mitigate interference from the neighboring N_{cell} small cells at time slot k, and $N_0(k)$ users return estimated SINRs $\gamma_N(k)$ to the BS.

The rest of the paper is organized as follows. In Section 2, related work on RL-based ICI mitigation are presented. The system model and our proposed dynamic Q-learning scheme are presented in Section 3 and Section 4. In Section 5, simulation results are presented and the conclusion follows in Section 6.

2. RELATED WORK

ICI control is a key issue in 5G MMIMO systems, intensive research has been carried out to address this. Surveys have been carried out on ICI mitigation techniques in LTE downlink networks [8, 9], and research on ICI coordination techniques in 5G UFMC systems [10, 11].

RL-based approaches have been extensively applied in an ICI mitigation problem. For instance, a Q-learning-based power control scheme formulates the ICI coordination issue as a cooperative multi-agent control problem to improve the performance of the cellular systems is proposed in [12]. An RL-based power control scheme for ultra-dense small cells to improve network throughput and save energy consumption is presented in [13], in which the BS selects the downlink transmit power to manage interference. A dynamic RL-based ICI coordination algorithm as developed in [14] smartly offloads traffic to open access picocells and then improves the system throughput.

3. SYSTEM MODEL

ICI is caused by multiple sources transmitting signals with the same subcarrier and being received by a receiver. A user receives signals from the serving cell and neighboring cells but at different power levels due to the path loss.

3.1 AOA-based beamforming

The Angle-Of-Arrival(AOA)-based beamforming is usually used in 5G MMIMO systems, where the BS is configured with an antennas array composed of *W* AEs, and numbers of AEs are arranged as *M* per row and *L* per column [2].

In RF, the BS shapes beamforming for the k^{th} UE by configuring weights on AEs according to AOA $\langle \theta_k, \varphi_k \rangle$ [15], where θ_k is the azimuth and φ_k is the vertical angle of the k^{th} UE. The weights on the i^{th} AE in a row can be represented as

$$\omega_{ik} = e^{-j2\pi i d_h \sin\theta_k}, w_{ik} \in \mathcal{C}^{1*M} \quad (1)$$

where d_h is the row AE distance. And the l^{th} AE in the column can be obtained by

$$\xi_{lk} = e^{-j2\pi l d_v \cos\varphi_k}, \xi_{lk} \in \mathcal{C}^{L*1} \quad (2)$$

where d_v is the column AE distance. From *(1)* and *(2)*, the final beamforming weights for the k^{th} UE can be derived by

$$\Pi_k = \Psi_k \Omega_k \quad (3)$$

where

$$\begin{cases} \Omega_k = [\omega_{1k}, \omega_{2k}, \ldots, \omega_{Mk}], \\ \Psi_k = [\xi_{1k}, \xi_{2k}, \ldots, \xi_{Lk}]^T. \end{cases}$$

Since the final weights in *(3)* depend on $\langle \theta_k, \varphi_k \rangle$, the implementation complexity for $\langle \theta_k, \varphi_k \rangle$ estimation gets high as the perfect CSI needed, which is usually affected by ICI.

3.2 Search-based beamforming

To mitigate the ICI with a low complexity, a search-based beamforming algorithm is reported in [15], which uses MC to search the optimal weights rather than AOA estimation in *(3)*. In MC beamforming, the best weights are obtained by searching $\langle \theta_k, \varphi_k \rangle$ in all possible angles to minimize the ICI, i.e.

$$\langle \theta_k^*, \varphi_k^* \rangle \longleftarrow \arg \min_{\langle \theta_k, \varphi_k \rangle} P_r(SINR < T_g | h_j^{(k)}, \rho_j^{(k)}) \quad s.t. -\pi < \theta_k, \varphi_k \le \pi \quad (4)$$

where P_r is the probability of SINRs weaker than the target T_g given the channel $h_j^{(k)}$ and UE density $\rho_j^{(k)}$, and the SINR in (4) for the i^{th} UE located on the j^{th} cell can be expressed by [15]

$$SINR_{i,j} = \frac{p_{i,j} \varsigma_{i,j}^{-\nu}}{N_0 B + \sum_{k=1,k\neq j}^{N} p_k \varsigma_k^{-\nu}} \quad (5)$$

where ν is the path-loss exponent, $p_{.j}$ is the transmit power of the serving enode B_j, N is the number of neighboring enode B_s, p_k is the transmit power from B_s, $\varsigma_{.j}$ is the distance of the UE to the serving station, ς_k is the distance of the UE to each of the neighboring stations, and $N_0 B$ is the background noise with N_0 the thermal noise and B the system bandwidth.

According to [16], the UE density $\rho_0^{(k)}$ in *(4)* is assumed to follow the independently and identically distributed two-dimensional Poisson point process. The number of users $N_0^{(k)}$ of the target cell with area φ_0 is given by

$$Pr\{N_0^{(k)} = \lambda|\varphi_0\} = \frac{(\rho_0^{(k)}\varphi_0)^\lambda}{\lambda!} e^{-\rho_0^{(k)}\varphi_0} \qquad (6)$$

From *(4)* to *(6)*, the optimal parameters $\langle \theta^*, \varphi_k^* \rangle$ can be found, and the best weight Π_k can be derived by substituting *(4)-(6)* into *(1)-(3)*.

4. THE PROPOSED REINFORCEMENT LEARNING ASSISTED BEAMFORMING

Due to the lack of prior knowledge that is required to find the theoretical optimal solution of *(4)*, some research has been conducted over related surrogate RL optimization problems. Generally, apart from the MC method, Sarsa [17] and Q-learning [13] are attempted. Those methods lack convergence efficiency in practice even though their convergence can be guaranteed [18].

In this section, a dynamic Q-learning beamforming method is proposed to mitigate ICI and enhance convergence efficiency. Each BS exploits the user SINRs in a dense Urban-eMBB transmission environment and estimates the Probability Density Function (PDF) of users' occurences to achieve an optimal beamforming solution via trial without knowledge of the network and transmission channel.

4.1 RL-based beamforming

In the RL-based beamforming process as shown in Fig. 1(a)(b), the BS in the target cell estimates the probability density $\rho_0^{(t)}$ of users' occurrences in the target small cell $\#0$ by a long-term data statistical analysis in *(6)* at time slot t. Once all served users send SINRs $\gamma^{(t-1)}$ at the time slot $(t-1)$ to the BS, the state $[\rho_0^{(t)}, \gamma^{(t-1)}]$ observed by the BS at the time slot t is obtained, and then an RL-based beamforming algorithm is applied for searching the optimal parameters for the ICI mitigation and coverage optimization. We formulate the beamforming optimization problem under the MMIMO system context as an RL problem and therefore provide a dynamic Q-learning scheme to address the issue.

First, we define the agent to be the MMIMO system, the set of states $S \triangleq \{s_l\}_{l=0}^{m-1}$ to be the levels of average regional SINR, the set of actions $A \triangleq \{a_j\}_{j=0}^{n-1}$ to be the possible combinations of antenna parameters. More precisely, each s_l is an interval of the SINR value, s_0 is the optimal SINR value interval, i.e. the highest achievable SINR value derived from expert experiences in the current environment. Similarly, s_{m-1} is the lowest SINR range. And as l increases, the boundary values of s_l decreases, thus a higher l implies poorer signal performance state s_l. Each action a_j is an antenna parameter's choice made by the MMIMO system, and consists of azimuth, vertical angle and beam width. The environment is a signal simulator, see Section 5.1 for more detail. The objective is to approach the optimal target SINR state s_0 to achieve the best signal performance. It covers the probability in *(4)* of average regional SINR given by the simulator and guided by selected action a. The environment (Fig. 1(c)) grants the agent a reward $r_{s,a}$ after the latter takes an $a \in A$ when it is in $s \in S$.

Formally, we denote the state-action value function, the expected discounted reward, as $Q(s, a)$. In the table $\mathbf{Q} \in R^{m \times n}$, we use notation [19] $Q(s, a) \triangleq [\mathbf{Q}]_{s,a}$ and update entries by:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r_{s,a} + \delta \max_{a'} Q(s', a') - Q(s,a)] \qquad (7)$$

where $\alpha : 0 < \alpha < 1$ is the learning rate and $\delta : 0 < \delta < 1$ is the *discount* factor and determines the importance of future rewards. s' and a' are the next state and action, respectively.

An *episode* is a period of time in which an interaction between the environment and the agent takes place. Here, an episode is of (at most) τ transitional discrete time step t. During an episode $i : i \in \{0, 1, \ldots, \zeta\}$, the agent makes a decision to maximize the effects of actions decided by itself. To achieve this goal, we apply the ϵ- greedy learning strategy to balance exploration and exploitation, where $1 - \epsilon : 0 < \epsilon < 1$ is the exploration rate and serves as the threshold probability to select a random $a \in A$, as opposed to selecting an action based on exploitation. To add randomness, the ϵ increases in every episode from ϵ_{min} until it reaches a preset upper bound.

The S space is constructed by partitioning the range of the process Cumulative Distribution Function (CDF), which is the probability of users with SINR under the given T_g in *(4)*. The components of A space is shown in Table 1. Through a finite series of $a \in A' := A - \mathbf{C}$ (will be discussed later), the agent attempts to approach s_0 in response to simulated s_i at step t within an episode.

Table 1 – Learning Parameters

Parameter	Value
Learning rate α	0.01
Reward decay rate δ	0.9
Minimum exploration rate ϵ_{min}	0.9
Number of episodes ζ	22
Number of steps in each episode τ	40
Number of states	30
Number of actions	855

4.2 Reward signals

4.2.1 Reward design with Q-initialization

As discussed in [20], reward signals in our simulation environment are crucial to the RL Markov Decision Process (MDP) since agents are expected to learn the optimal policy under industrial criteria.

Since adding additional rewards follows the policy invariance [20], the reward function $r(s,a)$ within our problem setting consists of two main parts:

$$r(s,a) = r(s_0,a)_{goal} + r(s,a)_{inter} \tag{8}$$

$r(s_0,a)_{goal}$ is given to the agent if s_0 is approached and $r(s,a)_{inter}$ works as intermediate reward in the training process when $s \neq s_0$.

We aim to construct reward shaping for $r(s,a)_{inter}$ using the potential-based method to help guide the agent in MDP; the potential-based shaping function is defined as [20]:

Definition 1 *Let any S, A, δ and any shaping reward function $F: S \times A \times S \to \mathbb{R}$ in MDP be given. F is **potential-based** if there exists a real-valued function $\Phi: S \to \mathbb{R}$ s.t.*

$$F(s,a,s') = \delta\Phi(s') - \Phi(s) \tag{9}$$

for all $s \neq s_0, s' \in S, a \in A$.

Therefore, based on the results in [20], such an F can guarantee consistency with the optimal policy that the agent learned. Luckily, there is no need to construct the shaping function from scratch [21], since the design of F is equivalent to the initialization of $[\mathbf{Q}]_{s,a}$.

Suppose the optimal policies learnt in our model with and without potential-based F are π' and π, respectively. Let initial Q function of π be $Q(s,a) = Q_0(s,a)$ with shaping rewards $\delta\Phi(s') - \Phi(s)$, and initial Q function of π' be $Q'(s,a) = Q_0(s,a) + \Phi(s)$ with no shaping rewards.

By *(7)*, we have the update error:

$$\begin{cases} Q_{error} = r_{s,a} + \delta\Phi(s') - \Phi(s) + \delta \max\limits_{a'} Q(s',a') - Q(s,a) \\ Q'_{error} = r_{s,a} + \delta \max\limits_{a'} Q'(s',a') - Q'(s,a) \end{cases} \tag{10}$$

and now insert ΔQ and $\Delta Q'$, the difference between current and initial values of Q and Q' respectively, into the update error:

$$\begin{cases} \Delta Q(s,a) = Q(s,a) - Q_0(s,a) \\ \Delta Q'(s,a) = Q'(s,a) - Q_0(s,a) - \Phi(s) \end{cases} \tag{11}$$

we have

$$\begin{aligned} Q_{error} &= r_{s,a} + \delta\Phi(s') - \Phi(s) + \delta \max_{a'}(Q_0(s',a') \\ &\quad + \Delta Q(s',a')) - Q_0(s,a) - \Delta Q(s,a) \\ &= r_{s,a} + \delta \max_{a'}(\Phi(s') + Q_0(s',a') + \Delta Q(s',a')) \\ &\quad - Q_0(s,a) - \Delta Q(s,a) - \Phi(s) \\ &= r(s,a) + \delta \max_{a'} Q'(s',a') - Q'(s',a') \\ &= Q'_{error} \end{aligned} \tag{12}$$

Therefore, we investigate the relationship between r_{inter} and $Q_0(s,a)$ to decide the form of r_{inter}. In the MDP problem setting [18], the discounted return from time step t is $G_t = \sum_{k=0}^{\infty} \delta^k r_{t+k+1}$, and since $\delta \in (0,1)$, if r_{inter} is formed as a bounded series based on the distance from s_0 to s_l: $r(s,a)_{inter} \leq r_{bound}$, where $r_{bound} \leq 1$, we have

$$\begin{aligned} G_t &= \sum_{k=0}^{\infty} \delta^k r_{t+k+1} \\ &\leq \sum_{k=0}^{\infty} \delta^k r_{bound} \\ &\leq r_{bound} \sum_{k=0}^{\infty} \delta^k \\ &= \frac{r_{bound}}{1-\delta} \end{aligned} \tag{13}$$

then for optimal policy π' [18]

$$\max_a Q^{\pi'}(s,a) = E[G_t] \leq r_{goal} \tag{14}$$

we know r_{inter} and r_{goal} satisfy:

$$\frac{r_{bound}}{1-\delta} \leq r_{goal} \tag{15}$$

(15) gives an explicit gap between the two parts of $r(s,a)$ and also directly influences the following initialization of $Q(s,a)$.

4.2.2 Q-initialization setting

We rewrite the initial Q table of policy π' and the final converged table as $Q_0^{\pi'}$ and $Q_{final}^{\pi'}$ respectively. By (*(7)*),

$$\begin{aligned} Q^{\pi'}(s,a) &\leftarrow Q_0^{\pi'} + \alpha(r_{bound} + \delta Q_0^{\pi'} - Q_0^{\pi'}) \\ &= Q_0^{\pi'} + \alpha(1-\delta)(Q_{final}^{\pi'} - Q_0^{\pi'}) \end{aligned} \tag{16}$$

we can derive that

$$Q_0^{\pi'} > Q_{final}^{\pi'} = \frac{r_{bound}}{1-\delta} \tag{17}$$

to guarantee the convergence of the model update. And under the Q-learning scheme, *(17)* always provide chances of exploration for actions that have not been attempted.

In this end, we give reward signals for $r(s,a)$ as follows:

$$r_{s_l,a} := \begin{cases} -\frac{e^{0.1\cdot(l-2)}}{e^{2.8}+1} & s_l \geq s_2 \\ -\frac{0.01}{e^{2.8}+1} & s_l = s_1 \\ \frac{e^{0.2\cdot(30-i)}}{e^{2.8}+1} & s_l = s_0 \end{cases} \quad (18)$$

here $s_l \geq s_2$ means $l \geq 2$ and set $r_{bound} = -\frac{0.01}{e^{2.8}+1}$ from *(18)* to follow the conditions we derived in *(13)*, *(15)*. Therefore, we can initialize the Q function as $[\mathbf{Q}]_{s,a} := \mathbf{0}_{|S|\times|A|}$ to satisfy *(17)*.

Algorithm 1 Optimal Action Selection Control

Input: Initial CDF state s_{init} and target state s_0.
Output: Optimal a to approach s_0 during episode i.

Define customized S, A, ϵ and δ.
Initialize $\mathbf{C} := \{\}$, $\boldsymbol{Q} := \mathbf{0}_{|S|\times|A|}$, $i := 0$
Initialize $s := s_{init}$, $t := 0$
repeat
  while $t < \tau$ **do**
    $\epsilon := \max_{a'}(\epsilon_{min}, \epsilon_{min} + \delta \cdot t/(\tau \cdot \zeta))$
    Sample $k_1, k_2 \sim \mathcal{U}(0,1)$
    if $k_1 \leq \epsilon$ **then**
      if $k_2 > \delta(1-\epsilon)$ **then**
        Select $a \in A - \mathbf{C}, a = \arg\max_{a'} Q(s,a')$
      else
        Select $a \in A, a = \arg\max_{a'} Q(s,a')$
      end
    else
      Select $a \in A - \mathbf{C}$ randomly
    end
    Perform a in the simulator obtain $s', r(s,a)$
    Update the entry $Q(s,a)$ as in *(7)*
    $s \leftarrow s', t \leftarrow t+1$
    if $s \neq s_0$ **then**
      Append a in $\mathbf{C}$
    else
      Early stopping
      return a
    end
  end while
until $s = s_0$ otherwise proceed to episode $i+1$

4.3 Dynamic Q-Learning algorithm

Considering the computational and equipment cost in an MMIMO system, the delaying effect of reward should be minimized. Then after each step t, we use twice ϵ-greedy strategy, the controller to help avoid the action that is unrelated to s_0 to dynamically shrink the A space in order to make up for the delay in *(18)*. Therefore, the controller plays a highly efficient role as the penalty signal in our reward and serves as a reinforced mechanism to assist the selection. The upper bound of the time complexity for the dynamic Q-learning method is in $\mathcal{O}(mn)$ [22].

For a total of at most n trials in ζ episodes with a fixed initial environment setting, algorithm 1 will stop training the agent once s_0 is approached rather than continuing the process due to the reward signals design in our model:

Controller **C**: As shown in Algorithm 1, controller **C** will shrink the action space related to s in every step t based on the double ϵ- greedy principle. This operation enables the optimal action selection with higher and higher probability as t goes on.

Reward $r_{s,a}$: *(18)* guarantees the agent learns a global optimum, our target action, instead of continuously jumping on some local optimum for meaningless rewarding [23].

Reward signals and controller **C** attempt to guide the agent by avoiding redundant scoring and long term penalties. The agent itself continuously updates the learning policy under the guidance of both of them.

4.4 Other existing methods

For the not too large $S \times A$ space defined in Section 4.1, the MC exhaustion algorithm often serves as a baseline solution for the problem in Section 3. It requires testing on all possible $a \in A$ to ensure the best action among space A.

Therefore, we apply classical model-free RL methods: Q-learning (off-policy) and Sarsa (on-policy) [18] in this problem setting. They differ mainly in the Q function updating style, while Q-learning holds (*(7)*), Sarsa follows the update below:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r_{s,a} + \delta Q(s',a') - Q(s,a)] \quad (19)$$

Parameters for these models are set the same as in Table 1. Unsurprisingly, off-policy based methods are superior to on-policy methods [18] in the experiment discussed later.

With the experience gained from Algorithm 1, Algorithm 2 is proposed to test the trained agent's policy with any randomized given s_{init}.

5. SIMULATIONS AND DISCUSSIONS

To thoroughly investigate the performance of the proposed RL-assisted full dynamic beamforming method and validate the effectiveness of the theoretical analysis previously, we present statistical results of SINRs and computational complexity of the proposed algorithm compared to other industrial methods. We implement Algorithm 1 within the environment below with preset parameters shown in both Table 1 and Table 2.

Table 2 – Environment components

Simulation Parameter	Value	Simulation Parameter	Value
Antenna 3dB-Bandwidth in Azimuth (°)	$15 \sim 110$	Number of Observed UEs K_0	100
Antenna 3dB-Bandwidth in Elevation (°)	$0 \sim 30$	Receiver Bandwidth (MHz)	20
Antenna Tilt Angle (°)	$-3 \sim 15$	Receiver Height (m)	1.5
Carrier Frequency (GHz)	3.5	MM Array Type	URA
Height of BS (m)	25	MM Array Size	8×8
Transmit Power (dBm)	44	MM Mechanical Downtilt	15

Algorithm 2 Evaluation Algorithm

Input: Target state s_0 and $\boldsymbol{Q}$ from Algorithm 1.
Output: Optimal a, target s with rewards for Z episodes.

Load the experienced $\boldsymbol{Q} := [\mathbf{Q}]_{s,a}$, $z := 0$
repeat
  Randomize s_{init}
  Choose $a = \arg\max_{a'} Q(s_{init}, a')$
  Perform a in the simulator and obtain $s', r_{s_{init}}, t$
  Update $(s := s', a, r_{s_{init}}, t)_z$
  $z \leftarrow z + 1$
until $z = Z$

5.1 Environment setting

We use the three following metrics to set up the simulation environment and help compare:

- The simulation is based on the guidelines deined in [24] for evaluating 5G radio technologies in an urban macro-cell test environment which presents a radio channel with high user density and trafic loads focusing on pedestrian and vehicular users (Dense Urban-eMBB) [25].
- As shown in Fig. 1(a), the environment layout consists of 19 sites placed in a hexagonal layout, each with 3 cells, and the Inter-Site Distance (ISD) is 200 m.
- To visualize SINR for the simulation scenario we use the Close-In (CI) propagation path loss model [26], which calculates the path loss of transmitted power in 5G urban microcell and macro-cell scenarios. This model produces an RSRP (Reference Signal Receiving Power) map and a SINR map that shows reduced interference effects compared to other beamforming methods.

5.2 Computational complexity

The agent learns from the environment for 1000 epochs of all randomized s_{init} and stores policy experience in the Q-table described in Algorithm 1. In this stage, our model performs faster and is more stable than other methods mentioned above. We utilize the three following metrics to help compare:

Normalized Iteration Expectation $\mathcal{I}_E$: it indicates the scaled steps expectation to approach s_0 in 1000 epochs of training.

Computational Eficienc y (CE) $\mathcal{E}_*$: we deine the ratio below to relect computational cost saving:

$$\mathcal{E}_* \triangleq \frac{\mathcal{I}_{\mathrm{E}} \text{ for Baseline MC}}{\mathcal{I}_{\mathrm{E}} \text{ for method } i} \tag{20}$$

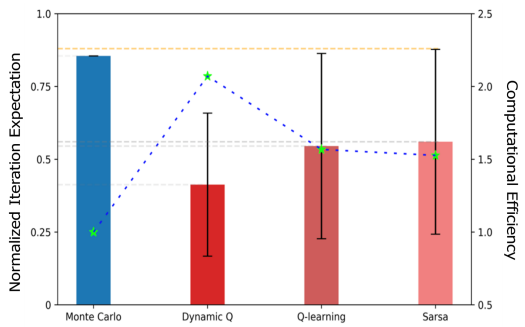
where $i \in \{\text{Dynamic Q, Q-learning, Sarsa}\}$.

Reward Scoring: This metric indicates how the dynamic Q method is different from other methods in speed and convergence when achieving reward.

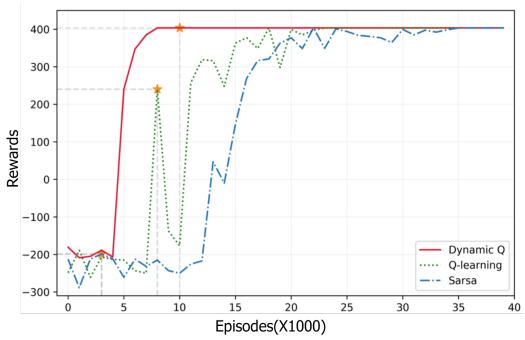
Fig. 2(a) displays the $\mathcal{I}_E$ with standard deviation, which implies stability in 1000 epochs, of how the dynamic Q model acts differently from Q-learning, Sarsa and MC. It takes the lowest normalized $\mathcal{I}_E$ needed to meet s_0 with the highest computational eficiency $\mathcal{E}_*$ (highlighted stars) and even doubles $\mathcal{E}_*$ compared to the baseline MC. (b) indicates the agility of our model in adapting to the environment. Given randomized s_{init}, the 95% conidence interval shadow indicates within 1000 epochs of training, the range and convergence rate of reward scoring for the dynamic Q model differs from other RL methods. Our model is able to fully train its agent in 10 episodes (without early stopping) with robustness and obtain the highest reward while other methods are still unstable under the two criteria.

5.3 SINR performance

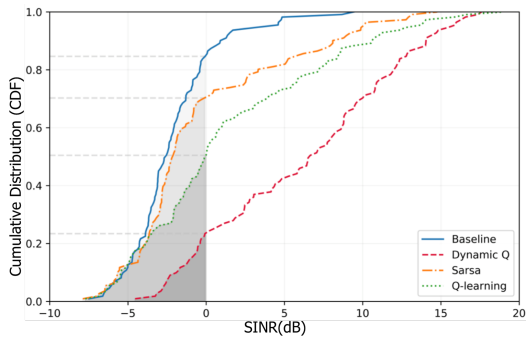
We show our model's shifting effect on SINR coverage in Fig. 2(c)(d) compared to other methods. With the optimal parameters derived from four models, respectively, within 10 episodes in (b) and sent into the simulator, (c) indicates ours is of the smallest weak SINR coverage that is lower than 0dB. The dynamic Q model suficientl y shifts the SINR coverage towards a strong SINR direction, and it enlarges the SINR coverage larger than 0 dB to over 50% of the total population in the Region of Interest (ROI). Specifically, (d) discloses that our model has the smallest probability density of users with weak SINR, for example, when SINR $\in (-5, 0]$, the probability is 23% with the dynamic Q model while it is 74%, 58%, 38% with the rest of the methods, respectively.



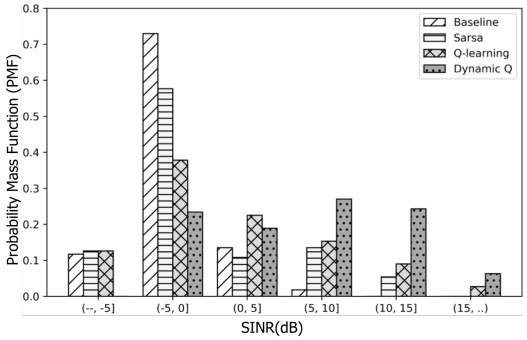
(a) The complexity of iteration expectation



(b) The best reward of ix ed iteration $N_{it} = 10$

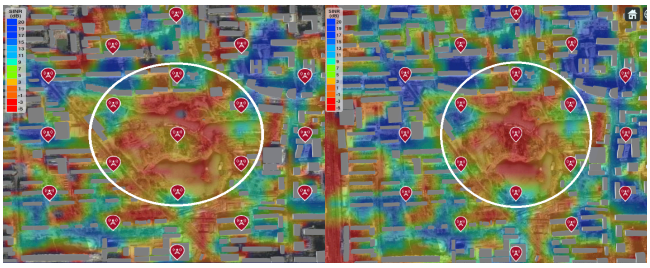


(c) The CDF of SINRs



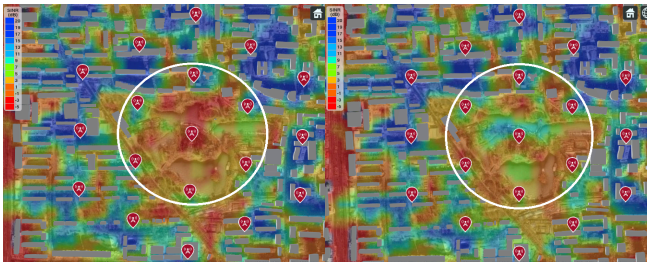
(d) The PMF of SINRs

Fig. 2 – Comparison of computational complexity and SINR improvements of the proposed dynamic Q algorithm with other industrial methods: MC (Baseline), Q-Learning and Sarsa. (a) shows the total iteration number for the optimal parameters; (b) displays their best reward when the iterate number is ix ed to $N_{\text{it}} = 10$, each point on the mean curve of rewards is averaged across 1000 epochs with random s_{init}, the shadow is the 95% conidence interval across 40 episodes of three models setting; (c) and (d) give the CDF and PMF of their SINRs.



(a) Baseline: $\bar{\gamma} = -2.036dB$

(b) SARSA: $\bar{\gamma} = -1.724dB$



(c) Q-learning: $\bar{\gamma} = 1.488dB$

(d) Dynamic Q $\bar{\gamma} = 6.319dB$

Fig. 3 – The average SINRs of different RL-based ICI mitigation algorithms in 5G MMIMO system, with parameters fed from Fig. 2(b). White circles are ROI. $W = 64$, $N_{cell} = 57$, $K_0 = 100$.

Fig. 3 displays the application when the optimal action is sent into the simulator of different models in 10 training episodes. The dynamic Q model is of the best average SINR of $\overline{\gamma} = 6.319$ dB in the ROI among all models.

In Table 3, we compare the average SINRs, across 6 different scenarios, for the dynamic Q model against MC, SARSA, and Q-Learning with parameters fed from Fig. 2(b). It is clear that the dynamic Q model improves the UE SINRs across 6 different environments, particularly in comparison with MC, where we achieve the average SINR improvements of around 8.3 dB, 10.4 dB, 12.2 dB 11.2 dB and 11.8 dB, respectively.

6. CONCLUSION

In this paper, we propose an RL (i.e. dynamic Q-learning) assisted full dynamic beamforming algorithm for the ICI mitigation in 5G MMIMO systems. This algorithm mitigates the ICI and reduces the computational complexity of the BS without knowledge of the network and transmission channel. Simulation results show the implementation complexity is lower and UE SINRs are signiicantly improved compared to other industrial methods. For example, in the dense Urban-eMBB scenario, the probability of weak SINRs in the target cell is about 60% lower and computational complexity is reduced by more than 50% compared to the benchmark.

Table 3 – Application scenarios

BS location		RL-based ICI mitigation algorithms			
Longitude	**Latitude**	**MC**	**Sarsa**	**Q-learning**	**Dynamic Q**
116.395659	39.959522	–2.036	–1.724	1.488	6.319
117.212147	39.161901	–4.732	–2.543	0.563	5.783
111.713038	40.832723	–7.931	–3.472	–1.239	4.374
111.710787	40.832027	–6.293	–2.174	0.897	4.978
111.709219	40.837586	–6.517	–2.573	1.296	5.381

REFERENCES

[1] Emil Björnson, Erik G Larsson, and Thomas L Marzetta. "Massive MIMO: Ten myths and one critical question". In: *IEEE Communications Magazine* 54.2 (2016), pp. 114–123.

[2] Jakob Hoydis, Stephan Ten Brink, and Mérouane Debbah. "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" In: *IEEE Journal on selected Areas in Communications* 31.2 (2013), pp. 160–171.

[3] Xudong Zhu, Zhaocheng Wang, Linglong Dai, and Chen Qian. "Smart pilot assignment for massive MIMO". In: *IEEE Communications Letters* 19.9 (2015), pp. 1644–1647.

[4] Vishnu V Ratnam, Andreas F Molisch, Ozgun Y Bursalioglu, and Haralabos C Papadopoulos. "Hybrid beamforming with selection for multiuser massive MIMO systems". In: *IEEE Transactions on Signal Processing* 66.15 (2018), pp. 4105–4120.

[5] Xiaoguang Zhao, Elena Lukashova, Florian Kaltenberger, and Sebastian Wagner. "Practical hybrid beamforming schemes in massive mimo 5g NR systems". In: *WSA 2019; 23rd International ITG Workshop on Smart Antennas*. VDE. 2019, pp. 1–8.

[6] Deepak Mishra and Håkan Johansson. "Optimal channel estimation for hybrid energy beamforming under phase shifter impairments". In: *IEEE Transactions on Communications* 67.6 (2019), pp. 4309–4325.

[7] Kwihoon Kim, Joohyung Lee, and Junkyun Choi. "Deep learning based pilot allocation scheme (DL-PAS) for 5G massive MIMO system". In: *IEEE Communications Letters* 22.4 (2018), pp. 828–831.

[8] A Daeinabi, K Sandrasegaran, and X Zhu. "Survey of intercell interference mitigation techniques in LTE downlink networks". In: *Australasian Telecommunication Networks and Applications Conference (ATNAC) 2012*. IEEE. 2012, pp. 1–6.

[9] Beatriz Soret, Klaus I Pedersen, Niels TK Jørgensen, and Víctor Fernández-López. "Interference coordination for dense wireless networks". In: *IEEE Communications Magazine* 53.1 (2015), pp. 102–109.

[10] Shendi Wang, John S. Thompson, and Peter M. Grant. "Closed-Form Expressions for ICI/ISI in Filtered OFDM Systems for Asynchronous 5G Uplink". In: *IEEE Transactions on Communications* 65.11 (2017), pp. 4886–4898. DOI: 10.1109/TCOMM.2017.2698478.

[11] Reshma Ravindran and Abhishek Viswakumar. "Performance evaluation of 5G waveforms: UFMC and FBMC-OQAM with Cyclic Prefix-OFDM System". In: *2019 9th International Conference on Advances in Computing and Communication (ICACC)*. 2019, pp. 6–10. DOI: 10.1109/ICACC48162.2019.8986195.

[12] Mariana Dirani and Zwi Altman. "A cooperative reinforcement learning approach for inter-cell interference coordination in OFDMA cellular networks". In: *8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*. IEEE. 2010, pp. 170–176.

[13] Hailu Zhang, Minghui Min, Liang Xiao, Sicong Liu, Peng Cheng, and Mugen Peng. "Reinforcement learning-based interference control for ultra-dense small cells". In: *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2018, pp. 1–6.

[14] Meryem Simsek, Mehdi Bennis, and Ismail Güvenç. "Learning based frequency-and time-domain inter-cell interference coordination in HetNets". In: *IEEE Transactions on Vehicular Technology* 64.10 (2014), pp. 4589–4602.

[15] Joy Iong-Zong Chen, Bo Hueng Lee, and Wen Bin Wu. "Performance evaluation of BER for an Massive-MIMO with M-ary PSK scheme over Three-Dimension correlated channel". In: *Computers & Electrical Engineering* 65 (2018), pp. 196–206.

[16] Vladimir Poulkov, Pavlina Koleva, Oleg Asenov, and Georgi Iliev. "Combined power and inter-cell interference control for LTE based on role game approach". In: *Telecommunication Systems* 55.4 (2014), pp. 481–489.

[17] Aya Mostafa Ahmed, Alaa Alameer Ahmad, Stefano Fortunati, Aydin Sezgin, Maria Greco, and Fulvio Gini. "A Reinforcement Learning based approach for Multi-target Detection in Massive MIMO radar". In: *IEEE Transactions on Aerospace and Electronic Systems* (2021), pp. 1–1.

[18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[19] Raghavendra V Kulkarni, Anna F rster, and Ganesh Kumar Venayagamoorthy. "Computational intel- ligence in wireless sensor networks: A survey". In: *IEEE communications surveys & tutorials* 13.1 (2010), pp. 68–96.

[20] Andrew Y Ng, Daishi Harada, and Stuart Russell. "Policy invariance under reward transformations: Theory and application to reward shaping". In: *ICML*. Vol. 99. 1999, pp. 278–287.

[21] Eric Wiewiora. "Potential-based shaping and Q-value initialization are equivalent". In: *Journal of Artiicial Intelligence Research* 19 (2003), pp. 205–208.

[22] Sven Koenig and Reid G Simmons. "Complexity analysis of real-time reinforcement learning". In: *AAAI*. 1993, pp. 99–107.

[23] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. "Inverse reward design". In: *Advances in neural information processing systems*. 2017, pp. 6765–6774.

[24] M Series. "Guidelines for evaluation of radio interface technologies for IMT-2020". In: (2017).

[25] M Series. "Guidelines for evaluation of radio interface technologies for IMT-Advanced". In: *Report ITU* 638 (2009), pp. 1–72.

[26] Shu Sun, Theodore S Rappaport, Timothy A Thomas, Amitava Ghosh, Huan C Nguyen, István Z Kovács, Ignacio Rodriguez, Ozge Koymen, and Andrzej Partyka. "Investigation of prediction accuracy, sensitivity, and parameter stability of large-scale propagation path loss models for 5G wireless communications". In: *IEEE Transactions on Vehicular Technology* 65.5 (2016), pp. 2843–2860.

AUTHORS

Aidong Yang received a Ph.D. degree in wireless communications from the Dalhousie University, Halifax, NS, Canada, in 2017. His research interests include wireless techniques, 5G communications, machine learning and its applications. He has authored and co-authored over 20 journal and conference papers.

Xinlang Yue received an M.S. degree in applied mathematics from Columbia University, New York, NY, US, in 2020. His research interests mainly focus on deep reinforcement learning algorithms with real-world applications.

Mohan Wu received a Ph.D. degree in numerical mathematics from University of Pittsburgh, Pittsburgh, PA, US, in 2019. His research interests mainly focus on the init e element method and machine learning algorithm.

Ye Ouyang, Ph.D., is CTO & Senior Vice President of AsiaInfo Technologies. Dr. Ouyang has distinguished experience in R&D and management in the telecommunications industry. Prior to AsiaInfo, Dr. Ouyang has been Verizon Fellow and Senior Manager in Verizon. His research is in the interdisciplinary area of wireless communications, data science, and AI. Dr. Ouyang has authored more than 30 academic papers, 40 patents, 10 international standards, and 8 books. Dr. Ouyang obtained a Ph.D. from Stevens Institute of Technology, a Master of Science from Tufts University, another Master of Science from Columbia University, and a Bachelor's degree of from Southeast University.