Reinforcement Learning-Based Downlink Interference Control for Ultra-Dense Small Cells

Liang Xiao, *Senior Member, IEEE*, Hailu Zhang, Yilin Xiao, Xiaoyue Wan, Sicong Liu, *Member, IEEE*, Li-Chun Wang, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*— The dense deployment of small cells in 5G cellular networks raises the issue of controlling downlink inter-cell interference under time-varying channel states. In this paper, we propose a reinforcement learning based power control scheme to suppress downlink inter-cell interference and save energy for ultra-dense small cells. This scheme enables base stations to schedule the downlink transmit power without knowing the interference distribution and the channel states of the neighboring small cells. A deep reinforcement learning based interference control algorithm is designed to further accelerate learning for ultra-dense small cells with a large number of active users. Analytical convergence performance bounds including throughput, energy consumption, inter-cell interference, and the utility of base stations are provided and the computational complexity of our proposed scheme is discussed. Simulation results show that this scheme optimizes the downlink interference control performance after sufficient power control instances and significantly increases the network throughput with less energy consumption compared with a benchmark scheme.**

***Index Terms*— Ultra-dense small cells, interference control, power control, reinforcement learning.**

Manuscript received May 7, 2019; revised August 10, 2019; accepted September 25, 2019. Date of publication October 14, 2019; date of current version January 8, 2020. This work was supported in part by the Natural Science Foundation of China under Grant 61971366, Grant 61671396, and Grant 61901403, in part by the Fundamental Research Funds for the Central Universities of China under Grant 20720190034 and Grant 20720190029, in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University under Grant 2018D08, in part by the Natural Science Foundation of Fujian Province of China under Grant 2019J05001, and in part by the U.S. National Science Foundation under Grant CCF-0939370 and Grant CCF-1513915. The associate editor coordinating the review of this article and approving it for publication was J. Lee. *(Corresponding author: Liang Xiao.)*

L. Xiao is with the Department of Information and Communication Engineering, and the Key Laboratory of Digital Fujian on IoT Communication, Architecture and Security Technology, Xiamen University, Xiamen 361005, China, and also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China (e-mail: lxiao@xmu.edu.cn).

H. Zhang, Y. Xiao, X. Wan, and S. Liu are with the Department of Information and Communication Engineering, Xiamen University, Xiamen 361005, China, and also with the Key Laboratory of Digital Fujian on IoT Communication, Architecture and Security Technology, Xiamen University, Xiamen 361005, China.

L.-C. Wang is with the Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: lichun@g2.nctu.edu.tw).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2019.2945951

I. Introduction

ULTRA-DENSE small cell systems use low-cost and low-power cellular base stations (BSs) to enhance spatial multiplexing and significantly increase user capacity for fifth generation (5G) mobile networks. However, ultra-dense small cells are challenged by the unpredictable and increasing interference due to the spectrum scarcity and the uncoordinated network infrastructure. Interference control techniques, such as [1]–[4], have drawn significant attention for ultra-dense small cell systems. For instance, a big-data self-organizing network (Bi-SON) scheme as proposed in [4] statistically analyzes network data, especially the received signal power of the users in the neighboring cells, to determine the BS transmit power and improve the cellular communication efficiency for small cell networks. However, such an interference control scheme has to resolve unpredictable inter-cell interference and the inaccurate channel estimation caused by the over-deployment of small cells. A practical interference control scheme has to reduce the signaling and computation overhead of the BS due to the large number of BSs and wireless devices in ultra-dense small cells.

To address these issues, we propose a reinforcement learning (RL) based downlink interference control scheme to enable a BS to optimize the downlink transmit power without being aware of the channel states of the neighboring cells and their inter-cell interference distribution. This scheme depends on a state that consists of the estimated user density in the small cells, the downlink signal-to-interference-plus-noise ratio (SINR) for the users, and their downlink channel power gains. More specifically, the BS transmit power is chosen based on a Q-function, i.e., the expected accumulative reward for the BS to transmit with the power in the current state. The Q-function is updated via a Bellman iterative equation and initialized with transfer learning given by [5] to take advantage of the previous interference control experiences. This scheme can optimize the BS utility, i.e., increase the SINR, with less energy consumption and less inter-cell interference in the dynamic interference control process via trial-and-error.

We also propose a deep RL based interference control algorithm that applies deep learning to accelerate the optimization. By using a convolutional neural network (CNN) to estimate the Q values, this scheme compresses the dimension

1536-1276 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

of the state space and addresses the state quantization error for the ultra-dense small cells. We analyze the computational complexity and provide the convergence performance including throughput, energy consumption, interference, and utility, verified via simulations. We show that the proposed scheme significantly improves the throughput, reduces energy consumption, and increases the utility compared with Bi-SON [4].

This paper makes the following contributions:

1) We propose a downlink interference control scheme for ultra-dense small cell systems, in which a BS optimizes the transmit power without knowing the channel states of the neighboring cells and suppresses the inter-cell interference.
2) We introduce a deep RL based algorithm to further enhance the communication efficiency in the context of high cell density and analyze its computational complexity.
3) We analyze the convergence performance to support the design of the RL-based interference control scheme, confirming that the scheme can achieve a performance bound after sufficiently many power control interactions.

The remainder of this paper is organized as follows. Section II reviews related work, and Section III presents the system model. We propose the RL-based downlink interference control algorithm in Section IV and a deep RL-based algorithm in Section V. The convergence performance and the computational complexity are discussed in Section VI. We provide simulation results in Section VII. Concluding remarks are given in Section VIII.

II. Related Work

Power control serves as an efficient technique to mitigate the inter-cell interference in ultra-dense small cell networks. For example, a power control scheme as investigated in [6] uses Newton's method to improve both the network utility and the energy efficiency for millimeter-wave based ultra-dense small cell networks. An adaptive on-off power control method is presented in [7], which is capable of avoiding interference in a distributed pattern. A distributed target-SINR tracking based power control algorithm is presented in [8] by using the tracking power control and opportunistic power control method in a selective manner to improve the system throughput. An interference control algorithm as developed in [9] optimizes the power control to mitigate the cross-tier interference.

A power control algorithm as proposed in [10] combines the primal decomposition method in the coalition formation game to increase the system-wide utility and save energy for small cell systems. A δ_D-interference limited area control strategy as proposed in [11] combines with the conventional mechanism to improve the total capacity of cellular networks and device-to-device systems. An interference aware cell ranking scheme as proposed in [12] is combined with the switching on/off power control method to improve the communication performance for small cell systems in the dynamic environment. The data-driven resource management (DDRM) framework as proposed in [13] chooses the BS to transmit power and the channel according to the interference power from neighboring BSs to the local users and the channel characteristics of the other small cells. All this information can be obtained from the central controller home eNodeBs (HeNBs) management system, which is connected to each small cell. Moreover, the computational complexity and transmission latency of the proposed DDRM framework depends on the network size. Therefore, the proposed framework could not achieve high scalability when it is applied to the ultra-dense small cell networks with large coverage areas.

Recently, game theoretic interference control techniques are receiving considerable attention in ultra-dense small cell systems [14]–[16]. A dynamic pricing based power control scheme as proposed in [14] can derive the Nash equilibrium of the non-cooperative game for interference management. The joint power control and user scheduling scheme is proposed in [15] by formulating the interference control problem as a dynamic stochastic game between small cell BSs to improve the energy efficiency. Besides, a resource allocation algorithm based on graph theory as presented in [16] effectively manages the interference by a user-centric game.

As an emerging technology, RL has been applied in the interference control problem of wireless networks with more advanced features to improve the energy efficiency [17]–[29]. For instance, an RL-based decentralized power control strategy is proposed in [17], in which small cells jointly estimate the time-average performance and optimize the probability distribution for interference management in closed-access small cell networks. A centralized Q-learning algorithm with compact state representation is investigated in [18], where the BSs derive the optimal traffic offloading strategy based on the traffic observations to minimize the energy cost and maintain the quality of service (QoS). A dynamic Q-learning based interference coordination algorithm as proposed in [19] offloads traffic to the open-access picocells and then improves the system throughput. An online learning based traffic offloading strategy as proposed in [20] chooses the working mode for each small cell BS for heterogeneous cellular networks. Regret learning is applied in [21] to optimize the HeNBs in terms of energy efficiency following the QoS requirement for small-cell networks. The inter-cell interference coordination scheme as investigated in [22] applies fuzzy Q-learning to choose the BS transmit power and quantizes the continuous system state that consists of the physical resource block transmit power and the spectral efficiency, yielding state quantization noise in the interference control.

There are several studies that investigate scheduling based on deep RL for the cellular networks. For example, the downlink power control and rate adaption scheme as presented in [23] uses the Lagrange duality theory and artificial neural networks to enhance the cellular network throughput. A radio resource management approach as investigated in [24] uses the deep Q-network to select the on/off state for the cloud processor and the user equipment communication mode based on the cloud processor state, the current user communication mode, and the transmitter cache state to save power for fog radio access networks. A user scheduling and resource allocation

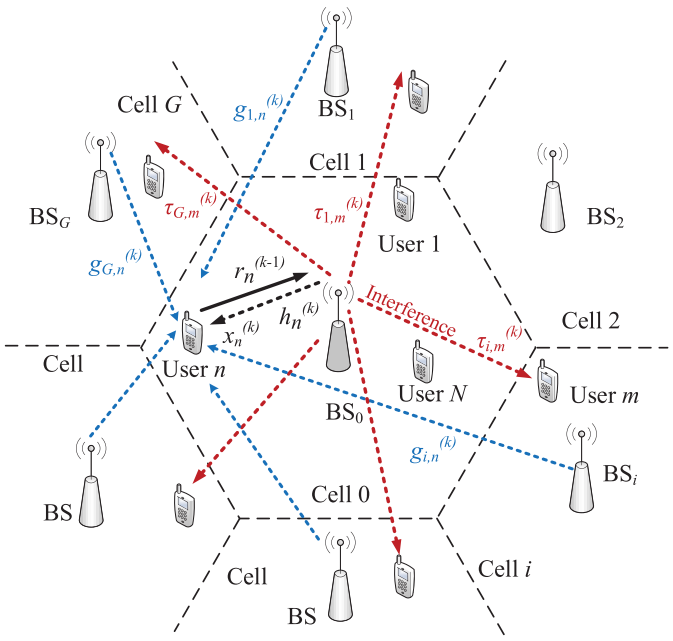


Fig. 1. Illustration of the interference mitigation in an ultra-dense small cell system, in which the target small cell BS_0 with N mobile users chooses its transmit power $x_n^{(k)}$ to mitigate the interference to the neighboring G small cells at time slot k, and the user n returns the estimated SINR $r_n^{(k)}$ to the target small cell BS_0.

scheme as proposed in [25] applies the policy gradient based actor-critic method to schedule the users, chooses the transmit power and the radio channel and uses the SINR of all the network users and the battery energy level of the BSs in the neighboring cells to improve the energy efficiency for heterogeneous networks.

In addition, an interference-aware path-planning scheme as investigated in [26] that uses the deep echo state network for the unmanned aerial vehicle (UAV) to select the moving direction, the transmission power and the cell association vector based on the UAV location and the BS location, is able to reduce the transmission latency and the interference on the ground network. A deep RL based scheduling algorithm as introduced in [27] determines the user for the resource block group based on the channel condition and the historic throughput information to improve the throughput for a single cell cellular network. In addition, the neighbor-agent actor critic scheduling algorithm as proposed in [28] that applies the deep RL to choose the resource block based on the channel information, the previous interference, and the previously selected resource block can reduce the outage probability and improve the sum rate in a single cell system. The distributed power allocation algorithm as proposed in [29] chooses the downlink transmit power based on the current channel state information, the last power set and the assisted feature to improve the average user sum-rate.

Furthermore, a power control scheme as presented in [30] uses the hotbooting Q-learning algorithm to determine the BS transmit power and mitigates the downlink inter-cell interference for ultra-dense small cells. Compared with the previous work in [30], we propose a deep RL based downlink interference control algorithm to accelerate the BS learning process with better communication efficiency. In addition, we provide convergence bounds on its interference control performance and discuss its computational complexity.

III. System Model

As depicted in Fig. 1, the BS of the target cell BS_0 that serves up to N mobile users is assumed to interfere with the users in the G neighboring cells. Equipped with multiple isotropic antennas, each BS in the system assigns orthogonal channel bandwidth B to each user without inter-cell interference in the downlink transmission. We index the time slot with k.

At time slot k, each BS obtains the downlink SINR denoted by $r_n^{(k-1)}$ from user n on the feedback channel. Let η be the cell density, i.e., the number of cells in a unit area. The large-scale fading gain from the BS to user m in cell i is denoted by $\xi_{i,m}^{(k)}$. The path loss $l_{i,m}^{(k)}$ between user m and the BS increases with distance $d_{i,m}^{(k)}$. According to [15] and [16], the interference factor denoted by $\tau_{i,m}^{(k)}$ represents the inter-cell interference from the BS to user m in the i-th neighboring cell, with $1 \le i \le G$, which is given by

$$\tau_{i,m}^{(k)} = \frac{\xi_{i,m}^{(k)}\sqrt{\eta}}{l_{i,m}^{(k)}\sqrt{|G|}}. \tag{1}$$

Given the maximum BS transmit power $\mathcal{P}_{\max}$, the BS transmit power to user n denoted by $x_n^{(k)}$ is quantized into $L+1$ levels, i.e. $x_n^{(k)} \in \mathbf{\Omega} = \{j\mathcal{P}_{\max}/L\}_{0\le j\le L}$, with $1 \le n \le N$. Note that the maximum transmit power $\mathcal{P}_{\max}$ will change with the type of the small cells (e.g., femtocells, microcells, and picocells) to meet the given coverage and service requirements. The downlink channel power gain in the target cell is denoted by $h_n^{(k)}$.

Cell i is assumed to have $M_i^{(k)}$ active users at time slot k. The average user density in the ultra-dense small cell system

TABLE I
SUMMARY OF SYMBOLS AND NOTATION

Symbol	Meaning
N	Maximum number of users in the cell
G	Number of neighboring cells
$\rho^{(k)}$	Average user density of the system
$r^{(k)}_{1\le n\le N}$	Downlink SINR
$\mathcal{P}_{\max}$	Maximum BS transmit power
$\tau^{(k)}_{i,m}$	Interference factor at time slot k
$M^{(k)}_i$	Number of users in the cell i
L	Feasible transmit power levels
$\mathbf{X}^{(k)}$	Transmit power for the N users
$\mathbf{\Omega}$	Transmit power set of the BS
$\Delta_{\mathbf{\Omega}}$	Action set of the BS
$u^{(k)}$	Utility of the BS
$\mathcal{C}_s$	Unit transmission cost
B	System bandwidth
$\theta^{(k)}$	CNN weights
σ	Receiver noise power
η	Cells Num. in unit area

denoted by $\rho^{(k)}$ is assumed to follow a two-dimensional Poisson process. According to [13], the number of users in cell i with area ϕ_i changes over time and its probability distribution is given by

$$\Pr\{M^{(k)}_i=\omega|\phi_i\}=\frac{(\rho^{(k)}\phi_i)^{\omega}}{\omega!}e^{-\rho^{(k)}\phi_i}. \tag{2}$$

The neighboring cell i interferes with user n in the target cell with channel power gain $g^{(k)}_{i,n}$. The common notations are summarized in Table I for later reference.

IV. RL-Based Downlink Interference Control

We propose an RL-based downlink interference control algorithm named RLIC to independently choose the downlink transmission power for the active users. This algorithm aims to suppress the inter-cell interference without relying on the knowledge of the channel state of the neighboring cells and the entire interference model. The state observed by the BS at time k is denoted by $\mathbf{s}^{(k)}$, which consists of the previous SINR for user n $r^{(k-1)}_n$ obtained from the feedback of user n, the estimated user density $\rho^{(k)}$, and the estimated channel state to user n denoted by $h^{(k)}_n$, from the channel estimation function according to [31], i.e., $\mathbf{s}^{(k)}=\{r^{(k-1)}_{1\le n\le N},\rho^{(k)},h^{(k)}_{1\le n\le N}\}\in\mathbf{S}$, where $\mathbf{S}$ denotes the state space.

A Q-function denoted by $Q(\mathbf{s},\mathbf{X})$ corresponds to the long-term discounted reward to the BS that takes action $\mathbf{X}$ at state $\mathbf{s}$, with $\mathbf{s}\in\mathbf{S}$ and $\mathbf{X}=[x_n]_{1\le n\le N}, x_n\in\mathbf{\Omega}$. The Q-values are initialized with the transfer learning method [32] by exploiting the previous similar interference control experiences to reduce the stochastic initial BS interference control explorations. More specifically, the BS learning parameters, such as the Q-values in Algorithm 1 are initialized before the dynamic interference control game, i.e, $\mathbf{Q}^*$ in Algorithm 1 based on λ interference control experiences sequences for similar scenarios. The learning parameter λ is chosen as a tradeoff between the fast interference mitigation and the over-fitting risk for the BS. Each of the λ experiences lasts F time slots, in which the BS uses Q-learning to choose the downlink transmit power for a given similar ultra-dense small cell system. In each of the F time slots, the Q-values are updated according to the Bellman iterative function.

In the dynamic interference control game, as shown in Algorithm 1, the BS reuses the prior knowledge, i.e., $\mathbf{Q}^*$ gained in the transfer learning and chooses the downlink transmit power for the N users served by the BS denoted by $\mathbf{X}^{(k)}=\left[x^{(k)}_n\right]_{1\le n\le N}$ based on the ϵ-greedy criterion similar to [33] as follows:

$$\Pr\left(\mathbf{X}^{(k)}=\mathbf{\Theta}\right)=\begin{cases}1-\epsilon, & \mathbf{\Theta}=\arg\max\limits_{\hat{\mathbf{X}}\in\Delta_{\mathbf{\Omega}}} Q(\mathbf{s}^{(k)},\hat{\mathbf{X}})\\ \frac{\epsilon}{|L+1|}, & \text{o.w.}\end{cases} \tag{3}$$

where $\hat{\mathbf{X}}$ is the power control policy that the BS tends to explore at state $\mathbf{s}^{(k)}$, and $\Delta_{\mathbf{\Omega}}$ is the BS action and can be given by $\Delta_{\mathbf{\Omega}}=\left\{\,[\hat{x}_n]_{1\le n\le N}\mid \hat{x}_n\in\mathbf{\Omega}\right\}$.

Algorithm 1 RL Based Downlink Interference Control

1: **Initialize**: α, β, $\mathbf{\Omega}$, $\mathbf{Q}=\mathbf{Q}^*$, $\mathbf{V}=\mathbf{0}$ and $r^{(0)}_{1\le n\le N}=0$
2: **For** $k=1,2,\ldots$
3: Estimate $\rho^{(k)}$ via (2)
4: Estimate the current channel state $h^{(k)}_{1\le n\le N}$
5: $\mathbf{s}^{(k)}=\left\{r^{(k-1)}_{1\le n\le N},\rho^{(k)},h^{(k)}_{1\le n\le N}\right\}$
6: Select the transmit power $\mathbf{X}^{(k)}$ via (3)
7: Evaluate the energy consumption of the target small cell and the inter-cell interference
8: Receive $r^{(k)}_{1\le n\le N}$ from the users feedback
9: Evaluate $u^{(k)}$ via (4)
10: Update $Q(\mathbf{s}^{(k)},\mathbf{X}^{(k)})$ via
$Q\left(\mathbf{s}^{(k)},\mathbf{X}^{(k)}\right)\leftarrow(1-\alpha)Q\left(\mathbf{s}^{(k)},\mathbf{X}^{(k)}\right)$
$+\alpha\Big(u^{(k)}+\beta\max_{\hat{\mathbf{X}}\in\Delta_{\mathbf{\Omega}}}Q\left(\mathbf{s}^{(k+1)},\hat{\mathbf{X}}\right)\Big)$
11: **End for**

Upon receiving the feedback from the N users, the BS evaluates the network user density $\rho^{(k)}$ and the SINR of the signal $r^{(k)}_{1\le n\le N}$ to determine the utility $u^{(k)}$ as follows:

$$u^{(k)}=\frac{B}{N}\sum_{n=1}^{N}\log_2\left(1+r^{(k)}_n\right)-\mathcal{C}_s\sum_{n=1}^{N}x^{(k)}_n-\mathcal{C}_s\sum_{i=1}^{G}\sum_{n=1}^{N}\sum_{m=1}^{M^{(k)}_i}x^{(k)}_n\tau^{(k)}_{i,m}. \tag{4}$$

The first term of $u^{(k)}$ in (4) represents the network throughput $R^{(k)}$ of the cell, the second term corresponds to the BS energy consumption $E^{(k)}$, and the third depends on the overall inter-cell interference $I^{(k)}$ according to [12], where $\mathcal{C}_s$ is the unit transmission cost. This utility function in (4) represents a tradeoff among the cellular throughput, the energy consumption and the interference suppression capability in the

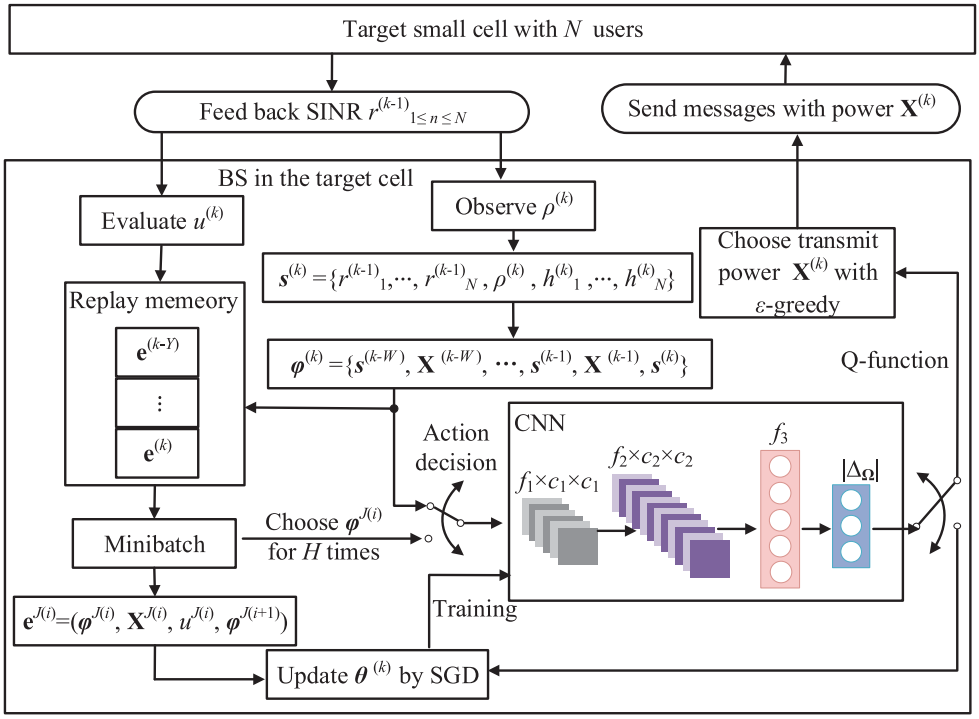


Fig. 2. Illustration of DRLIC for ultra-dense small cells, where $u^{(k)}$ and $\rho^{(k)}$ represent the observed network throughput and user density at time slot k respectively; and $\mathbf{X}^{(k)}$ is the suggested transmission power to N users at time k.

dynamic interference control game. This algorithm provides a dense reward signal for the BS to evaluate the impact of each interference control policy quickly.

The learning rate α is the weight of the current interference control experience and the discount factor β determines the importance of the future utility in the interference mitigation process. The Q-function $Q\left(\mathbf{s}^{(k)},\mathbf{X}^{(k)}\right)$ is updated according to the iterative Bellman equation as shown in Algorithm 1.

V. Deep RL-Based Interference Control

In this section, a deep RL-based interference control (DRLIC) algorithm is proposed by using a CNN to compress the BS state space, address the state quantization error and accelerate the dynamic optimization of the downlink transmit power. This algorithm depends on a current sequence of actions and states of the BS denoted by $\varphi^{(k)}$, i.e., $\varphi^{(k)}=\{\mathbf{s}^{(k-W)},\mathbf{X}^{(k-W)},\ldots,\mathbf{X}^{(k-1)},\mathbf{s}^{(k-1)},\mathbf{s}^{(k)}\}$ with $W(3N+1)+(2N+1)$ random variables. The sequence size W is set to achieve the tradeoff between the computation overhead and the interference control performance.

Being developed for the interference control of ultra-dense small cells, the CNN has 2 convolutional (Conv) layers to make a tradeoff between the accurate feature exaction from the sequence and the overfitting risk in the interference control according to [34] and [35]. As shown in Fig. 2, the sequence $\varphi^{(k)}$ is shaped into a $c_0\times c_0$ matrix and input to the first Conv layer of the CNN. Conv layer l that convolves f_l filters, each with dimension $c_l\times c_l$ and stride s_l, is followed by a rectified linear unit (ReLU) that has f_l feature maps as the output, with $1\le l\le 2$. The feature maps in Conv 2 are sent to two full connected (FC) layers. The first FC layer consists of v ReLUs. The second FC layer outputs the Q-values for the feasible transmit power levels. The CNN provides the $(L+1)^N$ estimated Q-values for the power control policies at the current sequence $\varphi^{(k)}$, i.e., $\mathbf{Q}\left(\varphi^{(k)},\mathbf{X},\boldsymbol{\theta}^{(k)}\right)$. The CNN weight vector denoted by $\boldsymbol{\theta}^{(k)}$ contains the weights of the four layers.

The ϵ-greedy algorithm in (3) is applied to determine the transmit power $\mathbf{X}^{(k)}$ based on the Q-values estimated to compromise between exploitation and exploration. After the downlink signal transmission and the feedback on the uplink control channel at time k, the BS evaluates the downlink SINR vector that is measured by the N users and applies (4) to calculate the utility $u^{(k)}$. The current interference management experience $\mathbf{e}^{(k)}=(\varphi^{(k)},\mathbf{X}^{(k)},u^{(k)},\varphi^{(k+1)})$ is stored together with the last Y experiences to realize the experience reuse in the replay memory, i.e., $\mathcal{D}=\{\mathbf{e}^{(k-Y)},\ldots,\mathbf{e}^{(k)}\}$ at each time slot.

Similar to the hotbooting process in [36], the BS exploits the previous interference mitigation experience to initialize the CNN weights $\overline{\boldsymbol{\theta}}$ for faster initial learning. By performing the experience replay at time slot k with H interference control experiences, i.e., $\{\mathbf{e}^{J(i)}\}_{1\le i\le H}=\{\varphi^{J(i)},\mathbf{X}^{J(i)},u^{J(i)},\varphi^{J(i+1)}\}_{1\le i\le H}$, with $J(\cdot)\sim U(k-Y,k)$, which are randomly sampled from the stored memory pool $\mathcal{D}$. The resulting minibatch is used to update the CNN weights $\boldsymbol{\theta}^{(k)}$ at each time slot. The experience replay randomizes over the previous BS interference control experiences, and thereby removes the correlation in the observation sequences, avoiding oscillations or divergence in the CNN parameters for better data efficiency [37].

According to the previous CNN weights $\boldsymbol{\theta}^{(k-1)}$, the target Q-value function Q' is given by

$$Q' = \left[u^{J(i)} + \beta \max_{\mathbf{X}' \in \Delta_{\mathbf{\Omega}}} Q\left(\boldsymbol{\varphi}^{J(i+1)}, \mathbf{X}'; \boldsymbol{\theta}^{(k-1)}\right) \right]_{1 \le i \le H}, \tag{5}$$

where $\mathbf{X}'$ is the next action to maximize the Q-value at state $\boldsymbol{\varphi}^{J(i+1)}$. Based on the H sampled experiences, the CNN weights $\boldsymbol{\theta}^{(k)}$ are updated by minimizing the mean square error between the estimated Q-value and the target Q-value according to the stochastic gradient decent (SGD) method at time slot k as shown in Algorithm 2, i.e.,

$$\boldsymbol{\theta}^{(k+1)} = \arg\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{\varphi}^{J(i)}, \mathbf{X}^{J(i)}, u^{J(i)}, \boldsymbol{\varphi}^{J(i+1)}} \left[\left(Q' - Q\left(\boldsymbol{\varphi}^{J(i)}, \mathbf{X}^{J(i)}, \hat{\boldsymbol{\theta}}\right) \right)^2 \right]_{1 \le i \le H}. \tag{6}$$

Algorithm 2 Deep RL Based Interference Control

Initialize: β, W, $\mathbf{\Omega}$, ϵ, and $r^{(0)}_{1 \le n \le N} = 0$
Set $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$, $\mathcal{D} = \emptyset$
For $k = 1, 2, \ldots$
  Estimate the average user density $\rho^{(k)}$
  Estimate the local channel state $h^{(k)}_{1 \le n \le N}$
  $\mathbf{s}^{(k)} = \left\{ r^{(k-1)}_{1 \le n \le N}, \rho^{(k)}, h^{(k)}_{1 \le n \le N} \right\}$
  If $k < W$ **then**
    Select $\mathbf{X}^{(k)}$ randomly
  Else
    Set $\boldsymbol{\varphi}^{(k)}$ as input of the CNN
    Set $\mathbf{Q}\left(\boldsymbol{\varphi}^{(k)}, \mathbf{X}, \boldsymbol{\theta}^{(k)}\right)$ as the CNN output
    Select $\mathbf{X}^{(k)}$ via (3) for the local users
  End if
  Send messages to the users with power $\mathbf{X}^{(k)}$
  Receive the SINR $r^{(k)}_{1 \le n \le N}$ from the feedback
  Evaluate $u^{(k)}$ via (4)
  $\boldsymbol{\varphi}^{(k+1)} = \left\{ \mathbf{s}^{(k-W+1)}, \mathbf{X}^{(k-W+1)}, \ldots, \mathbf{s}^{(k)}, \mathbf{X}^{(k)}, \mathbf{s}^{(k+1)} \right\}$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \left(\boldsymbol{\varphi}^{(k)}, \mathbf{X}^{(k)}, u^{(k)}, \boldsymbol{\varphi}^{(k+1)} \right)$
  $\mathcal{D} = \left\{ \mathbf{e}^{(k-Y)}, \ldots, \mathbf{e}^{(k)} \right\}$
  For $J(i) = 1, 2, \ldots, H$
    Select $\mathbf{e}^{J(i)} \in \mathcal{D}$ randomly
    Calculate Q' via (5)
  End for
  Update $\boldsymbol{\theta}^{(k)}$ by (6)
End for

VI. Performance Evaluation

In this section, we analyze the performance limit of the proposed interference control schemes, including the performance of energy consumption, data throughput, inter-cell interference, and the utility of the BS. $P^{(k)}_{i,n}$ denotes the power of the interference signal sent from BS_i and received by user n at time slot k. σ is the receiver noise power. According to [12], the SINR of user n (denoted by r_n) and the throughput of the target cell (denoted by R) are represented respectively as

$$r_n = \frac{x_n h_n}{\sigma + \sum_{i=1}^{G} P_{i,n} g_{i,n}} \tag{7}$$

and

$$R = \frac{B}{N} \sum_{n=1}^{N} \log_2 \left(1 + \frac{x_n h_n}{\sigma + \sum_{i=1}^{G} P_{i,n} g_{i,n}} \right). \tag{8}$$

To simplify the notation, the superscript of k denoting time slot k is omitted in the following content. Also, this algorithm ignores the circuit power consumption and the power consumption of the power amplifier in the small cell, as these factors have negligible impact on the utility of the BS in the learning process. Thus the energy consumption E and the overall interference I are modeled respectively as

$$E = \mathcal{C}_s \sum_{n=1}^{N} x_n \tag{9}$$

and

$$I = \mathcal{C}_s \sum_{i=1}^{G} \sum_{n=1}^{N} \sum_{m=1}^{M_i} x_n \tau_{i,m}. \tag{10}$$

The interference control process in the repetitive game on multiple interactions can be viewed as a Markov decision process since the future state of the BS is independent of the previous states for a given interference control policy at the current state. Therefore, the convergence performance of the RLIC algorithm (i.e., Algorithm 1) can be evaluated as follows.

Theorem 1: The proposed RL-based interference control algorithm (Algorithm 1) can converge to the optimal transmit power control policy after a sufficient number of interactions. Specifically, when

$$\begin{aligned} &\frac{\ln 2 \cdot \mathcal{C}_s N}{B} \left(\sigma + \sum_{i=1}^{G} P_{i,n} g_{i,n} \right) \left(\sum_{i=1}^{G} \sum_{m=1}^{M_i} \tau_{i,m} + 1 \right) \le h_n \\ &\le \frac{\ln 2 \cdot \mathcal{C}_s N}{B} \left[\left(\sigma + \sum_{i=1}^{G} P_{i,n} g_{i,n} \right) \left(\sum_{i=1}^{G} \sum_{m=1}^{M_i} \tau_{i,m} + 1 \right) + \mathcal{P}_{\max} \right] \end{aligned} \tag{11}$$

holds for $\forall 1 \le n \le N$, the performance bounds of throughput R, energy consumption E, interference level I, and utility u of Algorithm 1 can be represented as

$$\begin{aligned} R = {} & B \log_2 \frac{B}{\ln 2 \cdot \mathcal{C}_s N \left(\sum_{i=1}^{G} \sum_{m=1}^{M_i} \tau_{i,m} + 1 \right)} \\ & + \frac{B}{N} \sum_{n=1}^{N} \log_2 \frac{\sigma + \sum_{i=1}^{G} P_{i,n} g_{i,n}}{h_n}; \end{aligned} \tag{12}$$

$$\begin{aligned} I = {} & \sum_{i=1}^{G} \sum_{m=1}^{M_i} \frac{B \tau_{i,m}}{\ln 2 \cdot \left(\sum_{i=1}^{G} \sum_{m=1}^{M_i} \tau_{i,m} + 1 \right)} \\ & - \mathcal{C}_s \sum_{i=1}^{G} \sum_{m=1}^{M_i} \tau_{i,m} \sum_{n=1}^{N} \frac{\sigma + \sum_{i=1}^{G} P_{i,n} g_{i,n}}{h_n}; \end{aligned} \tag{13}$$

TABLE II
ENVIRONMENT PARAMETERS OF THE ULTRA-DENSE SMALL CELL SYSTEM

Parameters	Value/Mode
System bandwidth B	10 MHz
Density of small cells η	12 cells/km^2
Small cell radius	25 m
Maximum Num. of users per cell N	6
Received noisy power σ	-174 dBm/Hz
Distribution of small cell path loss $l_{i,m}^{(k)}$	$140.7 + 36.7\log_{10}(d_{i,m}^{(k)})$ dB ($d_{i,m}^{(k)}$ [km])

TABLE III
CNN PARAMETERS FOR DEEP RL-BASED INTERFERENCE CONTROL

Layer	Conv1	Conv2	FC1	FC2
Input	9×9	$8 \times 8 \times 20$	360	180
Filter size	2×2	2×2	/	/
Stride	1	1	/	/
Filter Num.	30	40	180	$\lvert\Delta_{\mathbf{\Omega}}\rvert$
Output	$8 \times 8 \times 30$	$7 \times 7 \times 40$	180	$\lvert\Delta_{\mathbf{\Omega}}\rvert$

$$E = \frac{B}{\ln 2 \cdot \left(\sum_{i=1}^{G}\sum_{m=1}^{M_i}\tau_{i,m}+1\right)} - \mathcal{C}_s \sum_{n=1}^{N}\frac{\sigma+\sum_{i=1}^{G}P_{i,n}g_{i,n}}{h_n}; \tag{14}$$

and

$$u = B\log_2\frac{B}{\ln 2\cdot \mathcal{C}_s N\left(\sum_{i=1}^{G}\sum_{m=1}^{M_i}\tau_{i,m}+1\right)} + \frac{B}{N}\sum_{n=1}^{N}\log_2\frac{\sigma+\sum_{i=1}^{G}P_{i,n}g_{i,n}}{h_n} - \frac{B}{\ln 2} + \mathcal{C}_s\sum_{n=1}^{N}\frac{1}{h_n}\left(\sum_{i=1}^{G}\sum_{m=1}^{M_i}\tau_{i,m}+1\right)\left(\sigma+\sum_{i=1}^{G}P_{i,n}g_{i,n}\right). \tag{15}$$

Proof: See Appendix A. ∎

Remark 1: The condition (11) means that the channel power gains of the N users $h_{1\le n\le N}$ exceed the bound based on the BS transmission cost $\mathcal{C}_s$ and the downlink bandwidth B, and are smaller than the bound given by the maximum transmission power $\mathcal{P}_{\max}$. In this case, the BS optimizes the transmit power for user n after a sufficient number of interactions, i. e.,

$$x_n^* = \frac{B}{\ln 2\cdot N\mathcal{C}_s\left(\sum_{i=1}^{G}\sum_{m=1}^{M_i}\tau_{i,m}+1\right)} - \frac{\sigma+\sum_{i=1}^{G}P_{i,n}g_{i,n}}{h_n}. \tag{16}$$

The convergence throughput given by (12) increases with the bandwidth B and decreases with the transmission cost $\mathcal{C}_s$. The converged inter-cell interference level given by (13) increases with the cell density η and the number of users in the neighboring cells. The converged energy consumption given by (14) decreases with the channel gains between the BSs in neighboring cells and the users in the target cell.

Let K be the number of steps per episode to converge, and Z be the number of the episodes. According to [38], the computational complexity of Algorithm 1 (denoted by $\mathcal{T}_1$) depends on the total number of the convergence steps to the optimal policy.

Theorem 2: The computational complexity of RLIC in Algorithm 1 is given by

$$\mathcal{T}_1 = \mathcal{O}(KZ), \tag{17}$$

if $KZ \ge \text{ploy}(|\boldsymbol{S}|, |\Delta_{\mathbf{\Omega}}|, K)$.

Proof: See Appendix B. ∎

Remark 2: The computational complexity of RLIC grows with the total number of learning samples. The random exploration of Q-learning at the initial interference control process in RLIC requires more interactions to converge than the optimal policy [5]. The transfer leaning in RLIC exploits the interference control experiences in similar networks to initialize $\mathbf{Q}^*$ to reduce the random initial explorations and thus reduces the sample size.

The computational complexity of the deep RL-based interference control algorithm depends on the CNN computational complexity in Algorithm 2, denoted by $\mathcal{T}_2$. According to [39], the CNN computational complexity depends on the number of filters, the filter size and the filter stride in each convolutional layer.

Theorem 3: The computational complexity of DRLIC in Algorithm 2 is given by

$$\mathcal{T}_2 = \mathcal{O}\left(f_1 c_2^2 f_2\left(\frac{c_0-c_1}{s_1 s_2} - \frac{c_2-1}{s_2}+1\right)^2\right). \tag{18}$$

Proof: See Appendix C. ∎

Remark 3: The communication performance improves with the number of filters that represents the inter-cell communication features. A shorter Conv stride that captures more interference details of the small cell systems has more

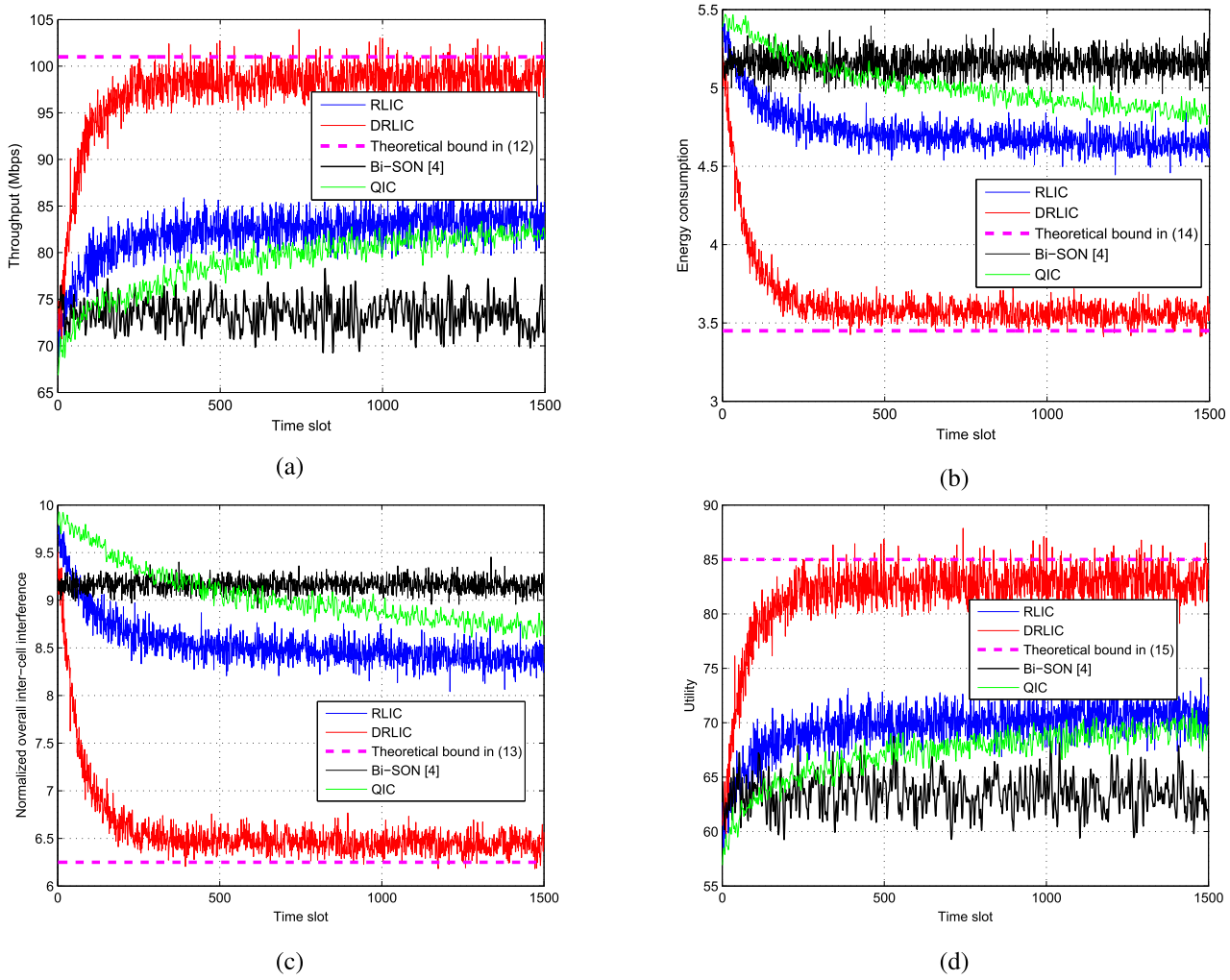


Fig. 3. Performance of the target BS interference control algorithms in the ultra-dense small cell system, with $G = 6$, $\alpha = 0.7$, $\beta = 0.5$, $\epsilon = 0.1$ and $B = 10$ MHz. (a) Throughput of the target cell. (b) Normalized energy consumption of the target cell. (c) Normalized overall inter-cell interference. (d) Utility of the BS.

computation overhead. The selection of the learning parameters in Algorithm 2 has to make a tradeoff between the communication performance and the computational complexity.

VII. Simulation Results

Simulations were performed to evaluate the proposed RL-based interference control schemes for an ultra-dense small cell system with 50 small cells. The corresponding cell density is 12 cells/km^2. For simplicity, the BS applies the quadrature phase-shift keying modulation and the first-tier inter-cell interference $G = 6$. The BS sends messages to six (i.e., $N = 6$) users [15] and receives feedback at each time slot for 8 ms. The system parameters for the small cells summarized in Table II are set according to [13].

The proposed interference control schemes take empirically effective learning parameters $\alpha = 0.7$, $\beta = 0.5$, $\epsilon = 0.1$ and $W = 3$ to compromise between communication efficiency and CNN memory overhead. The other CNN parameters in Table III are chosen to increase the communication efficiency according to the simulation results that are not shown here. For example, the minibatch size H is selected as a tradeoff between fast interference mitigation and the over-fitting risk for the BS. The feature size c_l captures the number of the cellular features at the cost of slow optimization. Similar to [35], the other learning parameters, i.e., the number of the filters f_l and the stride of filters s_l, are selected based on an informal search for the interference control scheme and the CNN structure that can optimize interference mitigation in ultra-dense small cells. For instance, the CNN of 30 filters with size 2×2 in Conv layer 1 can catch more features about the small cell system, compared with 20 filters with size 3×3 in Conv layer 1 given in [40].

Figure 3 provides the comparative performance of the proposed RLIC and DRLIC schemes with Bi-SON [4], which uses a data-driven based power control algorithm to suppress downlink interference, and a Q-learning based interference control (QIC) scheme according to the utility given by (4) and the same state signal with Algorithm 1.

As shown in Fig. 3, DRLIC converges to the performance bounds given by (12)–(15) after 300 time slots. Both RLIC and QIC improve the power control policy in the dynamic

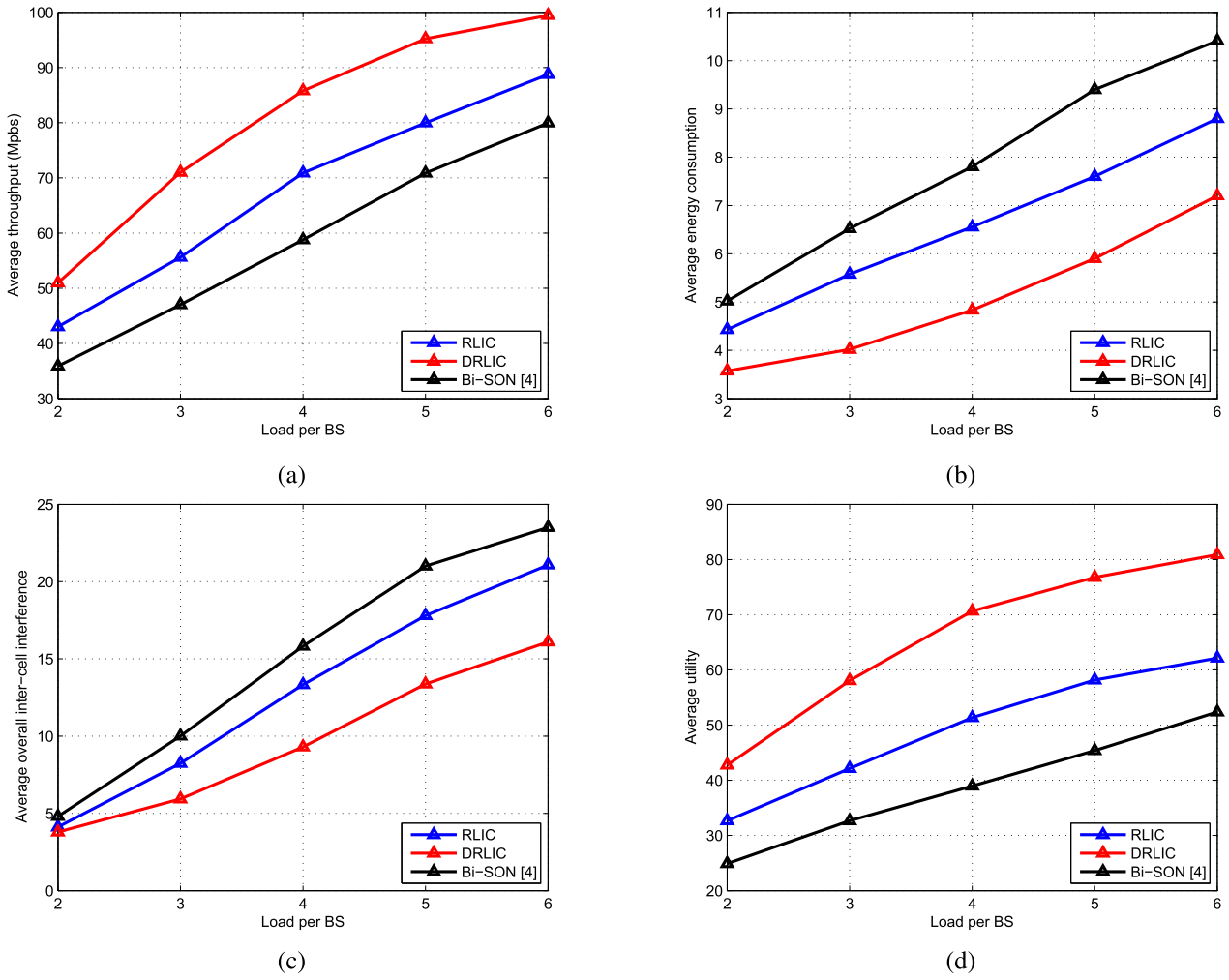


Fig. 4. Average performance of the target small cell for different schemes versus the load per BS over 100 learning processes and 400 time slots. (a) Average throughput of the target cell. (b) Average energy consumption of the target cell. (c) Average overall inter-cell interference. (d) Average utility of the BS.

game until reaching the performance bounds given by Theorem 1 after sufficient interactions with the users and the other BSs. Nevertheless, both algorithms can reduce the energy consumption, and the overall interference, and increase the system throughput and the utility of the BS compared with Bi-SON in [4] and QIC. For instance, RLIC converges to its optimal interference control policy after 300 time slots and saves the convergence time by 80% compared with QIC, and increases the throughput by 18.3% compared with Bi-SON. DRLIC can further improve the throughput of RLIC at the 300-th time slot by 20.2%. DRLIC has the lowest energy consumption, which is followed by RLIC. For instance, RLIC consumes 13.2% less energy consumption, yields 8.7% inter-cell interference, and results in 13.6% more BS utility than Bi-SON at the 300-th time slot. DRLIC further improves the performance of energy consumption, interference and the utility by 23.9%, 22.6% and 19.7% respectively.

Figure 4 provides the simulation results of the average performance over 100 learning processes with each containing 400 time slots in terms of different number of active users per cell, i.e., the BS load changes from 2 to 6 according to [15]. The average throughput increases logarithmically with the BS load. For instance, if the BS load changes from 2 to 6, the average throughput increases from 51 Mbps to 98 Mbps for the BS applying DRLIC. It is shown that, the average energy consumption increases from 3.6 to 7.1, and the average overall inter-cell interference increases from 4 to 16 if the BS load changes from 2 to 6. In addition, RLIC improves the average throughput by 12.6%, saves the average energy consumption by 14.4%, reduces the average overall interference by 12.5%, and increases the average utility by 19.2% compared with Bi-SON, if the BS load is 6. DRLIC further improves the performance of average throughput, average energy consumption, average interference and the average utility by 12.4%, 19.3%, 24.9% and 29.8% respectively, if the BS load is 6.

VIII. Conclusion

In this paper, we have proposed an RL based downlink interference control scheme for ultra-dense small cell systems, which enables a BS to optimize its transmit power without being aware of the inter-cell interference distribution and the channel state of the neighboring cells. A deep RL based framework has been presented to further enhance the communication efficiency of dynamic ultra-dense small cells with an acceptable computational complexity. A performance bound of

the proposed interference control schemes has been provided in terms of the downlink throughput, inter-cell interference, overall energy consumption and utility of the BS. Simulation results show that the proposed schemes significantly improve the throughput, reduce the overall inter-cell interference, and reduce the BS energy consumption. For example, after convergence the RL-based interference control algorithm reduces the BS energy consumption by 13.2% and increases the downlink throughput by 18.3% compared with Bi-SON. The deep RL-based interference control algorithm further reduces the energy consumption by 23.9% and increases the throughput by 20.2%.

Appendix A
Proof of Theorem 1

By (4) and (7), if (11) holds, we have

$$\begin{aligned}\frac{\partial u}{\partial x_n} &= \frac{\partial\left[\frac{B}{N}\sum_{n=1}^{N}\log_2\left(1+\frac{x_n h_n}{\sigma+\sum_{i=1}^{G}P_{i,n}g_{i,n}}\right)\right]}{\partial x_n} \\ &\quad - \frac{\partial\left[\mathcal{C}_s\sum_{n=1}^{N}x_n\left(\sum_{i=1}^{G}\sum_{m=1}^{M_i}\tau_{i,m}+1\right)\right]}{\partial x_n} \\ &= \frac{Bh_n}{\ln 2\cdot N\left(x_n h_n+\sigma+\sum_{i=1}^{G}P_{i,n}g_{i,n}\right)} \\ &\quad - \mathcal{C}_s\left(\sum_{i=1}^{G}\sum_{m=1}^{M_i}\tau_{i,m}+1\right),\end{aligned} \tag{19}$$

and

$$\frac{\partial^2 u}{\partial x_n^2} = -\frac{Bh_n^2}{\ln 2\cdot N\left(x_n h_n+\sigma+\sum_{i=1}^{G}P_{i,n}g_{i,n}\right)^2} < 0. \tag{20}$$

From (19), we have

$$\left.\frac{\partial u}{\partial x_n}\right|_{x_n=x_n^*} = 0, \tag{21}$$

where

$$\begin{aligned}x_n^* = &\frac{B}{\ln 2\cdot N\mathcal{C}_s\left(\sum_{i=1}^{G}\sum_{m=1}^{M_i}\tau_{i,m}+1\right)} \\ &- \frac{\sigma+\sum_{i=1}^{G}P_{i,n}g_{i,n}}{h_n},\end{aligned} \tag{22}$$

If (11) and (20) hold, we have $0 \le x_n \le \mathcal{P}_{\max}$ and

$$u\left(x_n^*\right) \ge u\left(x_n\right). \tag{23}$$

According to [41], the RL-based scheme can achieve the policy x_n^* in the MDP after a sufficient long time. Therefore, this algorithm can achieve x_n^* in (22). By integrating (22) into (4) and (8) - (10), we have (12) - (15).

Appendix B
Proof of Theorem 2

According to [38], the computational complexity of RL algorithms on episodic MDP is $\mathcal{O}(T)$, if $T \ge \text{ploy}(\mathbb{S}, \mathbb{A}, K)$, in which T is total number of steps, $\mathbb{S}$ denotes the number of states, and $\mathbb{A}$ represents the number of actions. In the RLIC scheme, the number of the system state is $|\mathbf{S}|$, the number of the BS action $\mathbf{X}$ is $|\Delta_{\mathbf{\Omega}}|$, and $T = KZ$. Therefore, we have (17).

Appendix C
Proof of Theorem 3

Similar to the analysis in [39], the computational complexity of the DRLIC scheme in Algorithm 2 is $\mathcal{O}(\sum_{l=1}^{2} f_{l-1}c_l^2 \ f_l m_l^2)$, in which f_{l-1} denotes the number of input channels of the l-th Conv layer, and m_l denotes the spatial size of the output feature map in the l-th Conv layer. Conv layer 1 involves f_1 filters, each with size $c_1 \times c_1$ and stride s_1. The first Conv layer has f_1 output feature maps. The second Conv layer consists of f_2 filters with size $c_2 \times c_2$, stride s_2, and f_2 output feature maps. According to [42], the output sizes of both convolutional layers are $m_1 = (c_0 - c_1)/s_1 + 1$ and $m_2 = (c_0 - c_1)/(s_1\ s_2) - (c_2 - 1)/s_2 + 1$, respectively. Therefore, we have

$$\begin{aligned}\mathcal{T}_2 = \mathcal{O}\Bigg(&c_1^2 f_1\left(\frac{c_0-c_1}{s_1}+1\right)^2 \\ &+ f_1 c_2^2 f_2\left(\frac{c_0-c_1}{s_1 s_2}-\frac{c_2-1}{s_2}+1\right)^2\Bigg).\end{aligned} \tag{24}$$

According to the CNN architecture in [43], we have

$$c_1^2 f_1\left(\frac{c_0-c_1}{s_1}+1\right)^2 \ll f_1 c_2^2 f_2\left(\frac{c_0-c_1}{s_1 s_2}-\frac{c_2-1}{s_2}+1\right)^2. \tag{25}$$

Thus by (24) and (25), we have (18).

References

[1] V. N. Ha and L. B. Le, "Distributed base station association and power control for heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 282–296, Jan. 2014.

[2] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 118–127, Jun. 2014.

[3] K. I. Pedersen, Y. Wang, S. Strzyz, and F. Frederiksen, "Enhanced inter-cell interference coordination in co-channel multi-layer LTE-advanced networks," *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 120–127, Jun. 2013.

[4] L.-C. Wang, S.-H. Cheng, and A.-H. Tsai, "Bi-SON: Big-data self organizing network for energy efficient ultra-dense small cells," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Montréal, QC, Canada, Sep. 2016, pp. 1–5.

[5] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.

[6] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. Leung, and H. V. Poor, "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936–1947, Sep. 2017.

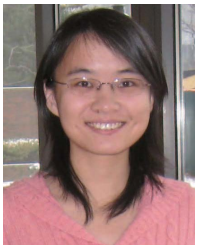
[7] J. Liu, M. Sheng, L. Liu, and J. Li, "Interference management in ultra-dense networks: Challenges and approaches," *IEEE Netw.*, vol. 31, no. 6, pp. 70–77, Dec. 2017.

[8] M. Rasti, A. R. Sharafat, and J. Zander, "A distributed dynamic target-SIR-tracking power control algorithm for wireless cellular networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 906–916, Feb. 2010.

[9] H. Zhang, Y. Nie, J. Cheng, V. C. M. Leung, and A. Nallanathan, "Sensing time optimization and power control for energy efficient cognitive small cell with imperfect hybrid spectrum sensing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 730–743, Feb. 2017.

[10] L. P. Qian, Y. Wu, H. Zhou, and X. Shen, "Joint uplink base station association and power control for small-cell networks with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5567–5582, Sep. 2017.

[11] H. Min, J. Lee, S. Park, and D. Hong, "Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 3995–4000, Dec. 2011.
[12] L.-C. Wang and S.-H. Cheng, "Self-organizing ultra-dense small cells in dynamic environments: A data-driven approach," *IEEE Syst. J.*, vol. 13, no. 2, pp. 1397–1408, Jun. 2019.
[13] L.-C. Wang and S.-H. Cheng, "Data-driven resource management for ultra-dense small cells: An affinity propagation clustering approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 3, pp. 267–279, Jul./Sep. 2019.
[14] J. Zheng, Y. Wu, N. Zhang, H. Zhou, Y. Cai, and X. Shen, "Optimal power control in ultra-dense small cell networks: A game-Theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4139–4150, Jul. 2017.
[15] S. Samarakoon, M. Bennis, W. Saad, M. Debbah, and M. Latva-Aho, "Ultra dense small cell networks: Turning density into energy efficiency," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1267–1280, May 2016.
[16] J. Cao, T. Peng, Z. Qi, R. Duan, Y. Yuan, and W. Wang, "Interference management in ultradense networks: A user-centric coalition formation game approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 5188–5202, Jun. 2018.
[17] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. V. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3202–3212, Jul. 2013.
[18] X. Chen, C. Wu, Y. Zhou, and H. Zhang, "A learning approach for traffic offloading in stochastic heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K, Sep. 2015, pp. 3347–3351.
[19] M. Simsek, M. Bennis, and I. Güvenç, "Learning based frequency- and time-domain inter-cell interference coordination in HetNets," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4589–4602, Oct. 2015.
[20] X. Chen, J. Wu, Y. Cai, H. Zhang, and T. Chen, "Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 627–640, Apr. 2015.
[21] Y. Wang, X. Dai, J. M. Wang, and B. Bensaou, "A reinforcement learning approach to energy efficiency and QoS in 5G wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1413–1423, Jun. 2019.
[22] M. Dirani and Z. Altman, "A cooperative reinforcement learning approach for inter-cell interference coordination in OFDMA cellular networks," in *Proc. 8th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, Avignon, France, Jun. 2010, pp. 170–176.
[23] E. Ghadimi, F. D. Calabrese, G. Peters, and P. Soldati, "A reinforcement learning approach to power control and rate adaptation in cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–7.
[24] Y. Sun, M. Peng, and S. Mao, "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1960–1971, Apr. 2019.
[25] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 680–692, Jan. 2018.
[26] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2125–2140, Apr. 2019.
[27] J. Wang, C. Xu, Y. Huangfu, R. Li, Y. Ge, and J. Wang, "Deep reinforcement learning for scheduling in cellular networks," May 2019, *arXiv:1905.05914*. Online Available: https://arxiv.org/abs/1905.05914
[28] Z. Li, C. Guo, and Y. Xuan, "A multi-agent deep reinforcement learning based spectrum allocation framework for D2D communications," Apr. 2019, *arXiv:1904.06615*. [Online]. Available: https://arxiv.org/abs/1904.06615
[29] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," Jan. 2019, *arXiv:1901.07159*. [Online]. Available: https://arxiv.org/abs/1901.07159
[30] H. Zhang, M. Min, L. Xiao, S. Liu, M. Peng, and P. Cheng, "Reinforcement learning-based interference control for ultra-dense small cells," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.
[31] S. Coleri, M. Ergen, A. Puri, and A. Bahai, "Channel estimation techniques based on pilot arrangement in OFDM systems," *IEEE Trans. Broadcast.*, vol. 48, no. 3, pp. 223–229, Nov. 2002.
[32] L. Xiao, Y. Li, G. Han, H. Dai, and H. V. Poor, "A secure mobile crowdsensing game with deep reinforcement learning," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 35–47, Jan. 2018.
[33] E. R. Gomes and R. Kowalczyk, "Dynamic analysis of multiagent Q-learning with ε-greedy exploration," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, Montreal, QC, Canada, Jun. 2009, pp. 369–376.
[34] I. Hadji and R. P. Wildes, "What do we understand about convolutional networks?" Mar. 2018, *arXiv:1803.08834*. [Online]. Available: https://arxiv.org/abs/1803.08834
[35] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
[36] M. Min, L. Xiao, C. Xie, M. Hajimirsadeghi, and N. B. Mandayam, "Defense against advanced persistent threats in dynamic cloud storage: A Colonel Blotto game approach," vol. 5, no. 6, pp. 4250–4261, Dec. 2018.
[37] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," Dec. 2013, *arXiv:1312.5602*. [Online]. Available: https://arxiv.org/abs/1312.5602
[38] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montréal, QC, Canada, Dec. 2018, pp. 4863–4873.
[39] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Reconit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 5353–5360.
[40] L. Xiao, D. Jiang, D. Xu, H. Zhu, Y. Zhang, and H. V. Poor, "Two-dimensional antijamming mobile communication based on reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9499–9512, Oct. 2018.
[41] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
[42] C. C. T. Mendes, V. Frémont, and D. F. Wolf, "Exploiting fully convolutional neural networks for fast road detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, Stockholm, Sweden, May 2016, pp. 3174–3179.
[43] A. Pritzel *et al.*, "Neural episodic control," Mar. 2017, *arXiv:1703.01988*. [Online]. Available: https://arxiv.org/abs/1703.01988



Liang Xiao (M'09–SM'13) received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications, China, in 2000, the M.S. degree in electrical engineering from Tsinghua University, China, in 2003, and the Ph.D. degree in electrical engineering from Rutgers University, NJ, USA, in 2009. She was a Visiting Professor with Princeton University, Virginia Tech, and the University of Maryland, College Park. She is currently a Professor with the Department of Communication Engineering, Xiamen University, China. She was a recipient of the Best Paper Award for 2016 INFOCOM Big Security WS and 2017 ICC. She has served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING.

Hailu Zhang received the B.S. degree in information and communication engineering from Xiamen University, Xiamen, China, in 2018, where she is currently pursuing the M.S. degree with the Department of Information and Communication Engineering. Her current research interests include network security and wireless communication.

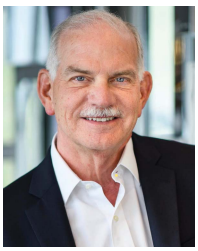
Yilin Xiao received the B.S. degree in automation and the M.S. degree in pattern recognition and intelligent system from the Hefei University of Technology, Hefei, China, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Department of Information and Communication Engineering, Xiamen University, Xiamen, China. His research interests include network security and privacy protection.

Xiaoyue Wan received the B.S. and M.S. degrees in information and communication engineering from Xiamen University, Xiamen, China, in 2016 and 2019, respectively.

Sicong Liu (S'15–M'17) received the B.S.E. and Ph.D. degrees (Hons.) in electronic engineering from Tsinghua University, Beijing, China, in 2012 and 2017, respectively. He was a Visiting Scholar with the City University of Hong Kong in 2010. He served as a Senior Research Engineer for Huawei Technologies. He joined Xiamen University in 2018, where he is currently an Assistant Professor. His current research interests include sparse signal processing, wireless communications, communications security, and machine learning. He has served as an editor, the track chair or a TPC member for several IEEE and other academic journals and conferences.

Li-Chun Wang (M'96–SM'06–F'11) received the Ph.D. degree from the Georgia Institute of Technology, Atlanta, in 1996.

From 1996 to 2000, he was with AT&T Laboratories, where he was a Senior Technical Staff Member with the Wireless Communications Research Department. Since August 2000, he has been with the Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan, where he is jointly appointed to the Department of Computer Science and Information Engineering. He has published more than 300 journal and conference articles and holds 23 U.S. patents, and have co-edited a book *Key Technologies for 5G Wireless Systems* (Cambridge University Press, 2017). His current research interests are in the areas of software-defined mobile networks, heterogeneous networks, and data-driven intelligent wireless communications. He was elected to the IEEE Fellow for his contributions to cellular architectures and radio resource management in wireless networks in 2011. He was a recipient of the Distinguished Research Award from the National Science Council, Taiwan, in 2012. He was a co-recipient of the IEEE Communications Society Asia–Pacific Board Best Award in 2015, the Y. Z. Hsu Scientific Paper Award in 2013, and the IEEE Jack Neubauer Best Paper Award in 1997.



H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in electrical engineering and computer science from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton University, where he is currently the Michael Henry Strater University Professor of Electrical Engineering. From 2006 to 2016, he served as the Dean of the Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory and signal processing, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the recent book *Multiple Access Techniques for 5G Wireless Networks and Beyond* (Springer, 2019).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the Marconi and Armstrong Awards of the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, the 2019 ASEE Benjamin Garver Lamme Award, a D.Sc. *honoris causa* from Syracuse University in 2017, and a D.Eng. *honoris causa* from the University of Waterloo in 2019.