RIS-Assisted Integrated Sensing and Communication System With Physical Layer Security Enhancement by DRL Approach

Peng Jiang[1], Xiaowen Cao[1], Yejun He[1], Xianxin Song[2], and Zhonghao Lyu[2]

[1] State Key Laboratory of Radio Frequency Heterogeneous Integration,
Sino-British Antennas and Propagation Joint Laboratory of MOST,
Guangdong Engineering Research Center of Base Station Antennas,
Shenzhen Key Laboratory of Antennas and Propagation,
College of Electronics and Information Engineering, Shenzhen University, 518060, China

[2] FNii and SSE, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China

Email: 1044609737@qq.com, caoxwen@szu.edu.cn, heyejun@126.com,
xianxinsong@link.cuhk.edu.cn, zhonghaolyu@link.cuhk.edu.cn

2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring) | 979-8-3503-8741-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/VTC2024-SPRING62846.2024.10683048

***Abstract*—Reconfigurable intelligent surfaces (RIS) play a crucial role in enhancing the security of integrated sensing and communication (ISAC) systems. In this paper, RIS is explored to assist the secure transmission of user data in ISAC system. Through the joint design of the transmit beamforming and RIS discrete phase shifter, we aim to maximize user's secure rates while ensuring target sensing performance. Due to the coupling of optimization variables, conventional optimization methods are hard to address this formulated problem. Therefore, a deep reinforcement learning (DRL) scheme by utilizing the soft actor-critic (SAC) and alternating optimization (AO) algorithms is employed to design the transmit beamforming and the RIS discrete phase shifter, respectively. Simulation results indicate that the problem scheme could obtain a significant improvement in enhancing user secure rates compared to other benching schemes.**

I. Introduction

Integrated sensing and communication (ISAC) has been recently adopted as one of the six delineated usage scenarios for future sixth-generation 6G network [1]. Its potential property that allows for sharing the hardware and spectrum has motivated vast existing research works in both academia and industry, especially for emerging applications that require a seamless integration of high-performance sensing and communication functionalities [2]. Although ISAC could enjoy a significant performance gain by efficiently utilizing the spectrum, energy, and hardware, and then jointly designing the waveform and signal processing flow between communication and sensing, compromised performance is unavoidable in the presence of poor propagation conditions. One promising way to address this challenge is reconfigurable intelligent surfaces (RIS) technology, which is composed of a planar metasurface with cost-effective and intelligently reflecting elements [3]. To be specific, through intelligently manipulating the phase and amplitude of reflected waves, RIS elements can be reconfigured in real-time to reconstruct the propagation environments, thus making it possible to simultaneously modify the communication and sensing channel for ISAC.

Given the mentioned benefits, RIS-enabled ISAC has became as a hot topic and attracted many compelling technical research [4]–[6]. The work [4] considered a RIS-enabled multiuser ISAC system with its aim to minimize the Cramer-Rao bound (CRB) by jointly optimizing the transmit beamforming at the base station (BS) and the reflective beamforming at the RIS, while the authors in [5] first designed a ISAC protocol to enable uplink data transmission and multi-user localization under the assistance of a semi-passive RIS, and then proposed beamforming algorithms for ISAC and communication with known sensed location information. Besides, the authors in [6] considered a RIS assisted ISAC operating in a millimeter-wave (mmWave) network, where a transmission rate maximization problem is formulated under the constraint of the desired waveform.

The secure performance of RIS-assisted ISAC also has garnered significant attention due to its potential to mitigate eavesdropping and other security threats by controlling and manipulating the signal propagation environment via using RIS [7]. However, due to the dynamic and complex wireless environment in RIS-assisted ISAC network, deep reinforcement learning (DRL) is adopted to find the optimal strategy through agent-environment interactions. The most recent work in [8] considered an ISAC system where the BS needs to serve multiple users and tracks a target simultaneously only relying on the indirect link via RIS. The formulated secrecy rate maximization problem is well solved by DRL method. Moreover, the case with multiple eavesdropping targets is more general but is uncharted in the exiting work. Thus, it motivates our work.

In this paper, we consider a RIS-assisted multi-target single-user ISAC system, assuming that all targets are eavesdroppers. The objective is to maximize user's secure rates through the joint design of the transmit beamforming and RIS discrete phase shifter. The coupling of optimization variables in this problem makes traditional solving methods challenging. Therefore, we utilize a DRL algorithm, integrating soft actor-critic (SAC) with the alternating optimization (AO) algorithm,

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

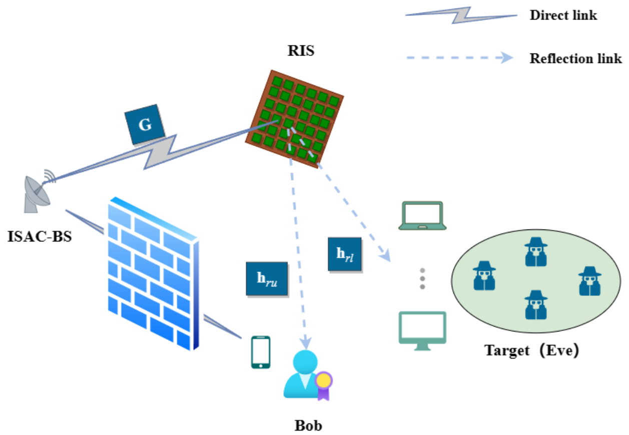

Figure 1: Illustration of RIS enabled ISAC system.

to jointly optimize transmit beamforming and RIS phase shifts. Simulation results validate the effectiveness of the SAC-AO algorithm in enhancing user security rates.

II. System Model

We consider a RIS-enabled ISAC system as shown in Fig.1, where a ISAC-BS is equipped with a uniform linear array (ULA) with M antennas, and the RIS is a uniform planar array (UPA) with N reflecting elements. The BS communicates simultaneously with a legitimate user (Bob) while also detecting L targets. Let $\mathcal{L} = \{1, \cdots, L\}$ denote the set of targets and $\mathcal{N} = \{1, \cdots, N\}$ denote the set of reflecting elements. We assume that all these targets are eavesdroppers attempting to intercept Bob's communication data, and both Bob and eavesdroppers are with a single antenna.

A. Transmitted Signal and Channel Model

Let $s_c(t)$ and $\mathbf{s}_r(t)$ represent the transmit communication signal for Bob and the sensing signal vector in each time slot t, respectively. We assume that radar signals are independent of communication symbols, indicating $\mathbb{E}\{s_c(t)\mathbf{s}_r(t)^H\} = \mathbf{0}$. Additionally, we assume unit power for the communication and radar signals, i.e., $\mathbb{E}\{s_c(t)s_c(t)^*\} = 1$ and $\mathbb{E}\{\mathbf{s}_r(t)\mathbf{s}_r(t)^H\} = \mathbf{I}_M$, respectively. $\mathbf{w}_c \in \mathbb{C}^{M\times 1}$ and $\mathbf{W}_r \in \mathbb{C}^{M\times M}$ denote the corresponding transmit beamforming vectors for communication and radar, respectively. Consequently, the ISAC-BS transmits a dual function signal $\mathbf{x}(t) \in \mathbb{C}^{M\times 1}$ in each time slot t given by

$$\mathbf{x}(t) = \mathbf{w}_c s_c(t) + \mathbf{W}_r \mathbf{s}_r(t). \tag{1}$$

In the time block T, we assume quasi-static flat fading for the channels. The channel between the BS and the RIS at time slot t is characterized as a Rician channel, which can be expressed as $\mathbf{G} = \sqrt{\lambda_{b,r}}\left(\sqrt{\frac{\kappa}{\kappa+1}}\mathbf{G}_{\text{LoS}} + \sqrt{\frac{1}{\kappa+1}}\mathbf{G}_{\text{NLoS}}\right) \in \mathbb{C}^{N\times M}$, where $\lambda_{b,r}$ represents the path loss depends on the distance, κ represents the Rician factor, $\mathbf{G}_{\text{LoS}}$ and $\mathbf{G}_{\text{NLoS}}$ represent the line-of-sight (LoS) and non line-of-sight (NLoS) components, respectively. For $\mathbf{G}_{\text{LoS}} = \mathbf{a}_r(\varphi_r, \zeta_r)\mathbf{a}_b^T(\varphi_b)$, where $\mathbf{a}_b \in \mathbb{C}^{M\times 1}$ and $\mathbf{a}_r \in \mathbb{C}^{N\times 1}$ represent the steering vectors at the BS and RIS, respectively, $\varphi_{r/b} \in (0, 2\pi]$ are azimuth angles, and $\zeta_r \in (-\pi/2, \pi/2]$ is the elevation angle. In accordance with [9], the terms $[\mathbf{a}_b(\varphi_b)]_n = e^{\frac{j2\pi(n-1)d_0\sin(\varphi_b)}{\lambda}}, \forall n$. $[\mathbf{a}_r(\varphi_r, \zeta_r)]_m = e^{j2\pi d_r(\lfloor \frac{m}{N_x} \rfloor \eta_1 + (m - \lfloor \frac{m}{N_x} \rfloor N_x)\eta_2)/\lambda}, \forall m$, where $\eta_1 = \sin(\varphi_r)\sin(\zeta_r)$, $\eta_2 = \sin(\varphi_r)\cos(\zeta_r)$, $\lfloor\cdot\rfloor$ denotes the floor function, N_x denotes the count of RIS elements in each row, d_0 signifies the separation between antennas, and d_r represents the gap between adjacent reflecting elements. The channel coefficients RIS-Bob and RIS-Eve are modeled as $\mathbf{h}_{ru} = \sqrt{\lambda_{ru}}\mathbf{g}_{\text{NLoS}} \in \mathbb{C}^{N\times 1}$ and $\mathbf{h}_{rl} = \sqrt{\lambda_{rl}}\mathbf{a}_r(\varphi_r, \zeta_r) \in \mathbb{C}^{N\times 1}$, respectively, where $\mathbf{g}_{\text{NLoS}}$ denotes the NLoS component with $[\mathbf{g}_{\text{NLoS}}]_j \sim \mathcal{CN}(0, 1)$.

B. Radar Model

Assuming a sufficiently large value for T, we consider the sample covariance matrix of the transmitted signal $\mathbf{x}(t)$ as equivalent to its statistical covariance matrix. The covariance matrix of the transmitted signal can be formulated as

$$\begin{aligned}\mathbf{R}_x &\triangleq \frac{1}{T}\sum_{t\in\mathcal{T}}\mathbf{x}(t)\mathbf{x}^H(t) \approx \mathbb{E}(\mathbf{x}(t)\mathbf{x}^H(t)) \\ &= \mathbf{w}_c\mathbf{w}_c^H + \sum_{i=1}^{M}\mathbf{w}_i\mathbf{w}_i^H,\end{aligned} \tag{2}$$

where $\mathbf{w}_i$ denotes the i-th column of the matrix $\mathbf{W}_r$. For the l-th target, the power of the target illumination resulting from the signal transmitted by the BS is expressed as

$$\begin{aligned}p_l(\mathbf{w}_c, \mathbf{W}_r, \mathbf{\Phi}) &= \mathbb{E}\left[|\mathbf{h}_{rl}^H\mathbf{\Phi}\mathbf{G}\mathbf{x}|^2\right] = \mathbf{h}_{rl}^H\mathbf{\Phi}\mathbf{G}\mathbf{R_x}\mathbf{G}^H\mathbf{\Phi}^H\mathbf{h}_{rl} \\ &= \operatorname{Tr}\left(\mathbf{R_x}\mathbf{G}^H\mathbf{\Phi}^H\mathbf{h}_{rl}\mathbf{h}_{rl}^H\mathbf{\Phi}\mathbf{G}\right),\end{aligned} \tag{3}$$

where $\mathbf{\Phi} = \operatorname{diag}(\phi_1, \ldots, \phi_n, \ldots, \phi_N)$ represents the phase shift matrix of the RIS with $\phi_n = \beta^n e^{j\vartheta^n}$, $\beta^n \in [0, 1]$ represents the amplitude reflection coefficient, and $\vartheta^n \in [0, 2\pi]$ indicates the phase shift of the n reflecting element. For simplicity, here neglects the influence of amplitude, i.e., $\beta^n = 1, \forall n \in \mathcal{N}$ [10]. Due to hardware constraints, ϑ^n takes values from a discrete set, denoted as $\vartheta_t^m \in \mathcal{F} = \{0, \frac{2\pi\cdot 1}{2^b}, \ldots, \frac{2\pi\cdot(2^b-1)}{2^b}\}$, where b represents the quantized bits.

C. Communication Model

The received signal at Bob during time slot t can be expressed as $y_u(t) = \mathbf{h}_{ru}^H\mathbf{\Phi}\mathbf{G}\mathbf{x}(t) + n_u(t)$, where $n_u(t) \sim \mathcal{CN}(0, \sigma_u^2)$ represents the additive white Gaussian noise (AWGN) at the receiver. Similarly, the received information at Eve can be expressed as $y_l(t) = \mathbf{h}_{rl}^H\mathbf{\Phi}\mathbf{G}\mathbf{x}(t) + n_l(t)$, where $n_l(t) \sim \mathcal{CN}(0, \sigma_e^2)$ represents the AWGN at the Eve's receiver. For simplicity, we assume $\sigma_u^2 = \sigma_e^2 = \sigma^2$. As a result, signal-to-interference-noise ratio (SINR) at Bob side can be expressed as

$$\gamma_u = \frac{\left|\mathbf{h}_{ru}^H\mathbf{\Phi}\mathbf{G}\mathbf{w}_c\right|^2}{\sum_{i=1}^{M}\left|\mathbf{h}_{ru}^H\mathbf{\Phi}\mathbf{G}\mathbf{w}_i\right|^2 + \sigma^2}. \tag{4}$$

The SINR acquired by the l-th eavesdropper can be expressed as

$$\gamma_l = \frac{\left|\mathbf{h}_{rl}^H\mathbf{\Phi}\mathbf{G}\mathbf{w}_c\right|^2}{\sum_{i=1}^{M}\left|\mathbf{h}_{rl}^H\mathbf{\Phi}\mathbf{G}\mathbf{w}_i\right|^2 + \sigma^2}. \tag{5}$$

The achievable data rates for Bob and the l-th eavesdropper are denoted as $R_u = \log_2(1 + \gamma_u)$ and $R_l = \log_2(1 + \gamma_l)$, respectively. Thus, the system's secure rate can be formulated as

$$R_s = [R_u - \max_{l\in\mathcal{L}} R_l]^+, \tag{6}$$

where $[\cdot]^+ = \max\{\cdot, 0\}$.

D. Problem Formulation

In this paper, the objective is to maximize the secure rate by jointly designing $\mathbf{w}_c, \mathbf{W}_r, \mathbf{\Phi}$, and this optimization problem is formulated as

$$
\begin{aligned}
\text{(P1)}: \max_{\mathbf{w}_c, \mathbf{W}_r, \mathbf{\Phi}} \quad & R_s && \text{(7a)}\\
\text{s.t.} \quad & p_l(\mathbf{w}_c, \mathbf{W}_r, \mathbf{\Phi}) \geq \varepsilon, \forall l \in \mathcal{L}, && \text{(7b)}\\
& \|\mathbf{w}_c\|^2 + \mathrm{Tr}(\mathbf{W}_r \mathbf{W}_r^H) \leq P_{\max}, && \text{(7c)}\\
& \vartheta^n \in \mathcal{F}, \forall n \in \mathcal{N}. && \text{(7d)}
\end{aligned}
$$

Constraint (7b) is to guarantee satisfactory sensing performance by ensuring that the target illumination power remains above a certain threshold ε. Here, $P_{\max}$ represents the maximum transmit power of the BS, which is constrained by condition (7c). Furthermore, condition (7d) defines the permissible range for the phase shifts of all RIS elements.

III. Proposed Scheme

To ensure secure user data transmission, we formulate the transmit beamforming design as a Markov decision process (MDP) problem. The AO algorithm is applied for RIS phase shift, and the SAC-AO framework is utilized for problem resolution.

A. MDP Problem Formulation

The MDP consists of key elements: $\mathcal{S}$ (the set of states), $\mathcal{A}$ (the set of actions), $\mathcal{P}$ (the probabilities of transitioning to the next state from an action), $\mathcal{R}$ (immediate reward), and χ (discount factor). Our goal is to find the optimal policy, selecting the best actions in specific states to maximize the expected cumulative long-term reward. The states, actions, and rewards are defined as follows.

- Action $\mathbf{a}_t$: The beamforming matrices for communication and sensing signals are considerded as the key actions which are expressed as
$$\mathbf{a}_t = \left\{ \{\mathbf{w}_c^{(t)}\}, \{\mathbf{W}_r^{(t)}\} \right\}. \tag{8}$$
- State $\mathbf{s}_t$: States include the actions $\mathbf{a}^{t-1}$ taken at the $(t-1)$ time step, as well as the channel state information (CSI) among the BS-RIS, RIS-user, and RIS-eavesdroppers. Then, the state at the t step is defined as
$$\mathbf{s}_t = \left\{ \{\mathbf{a}_{t-1}\}, \{\mathbf{G}^{(t)}\}, \{\mathbf{h}_{ru}^{(t)}\}, \{\mathbf{h}_{rl}^{(t)}\}_{l \in \mathcal{L}} \right\}. \tag{9}$$
- Reward $\mathbf{r}_t$: The reward is composed of the security rate and can be expressed as
$$r_t = [R_u^{(t)} - \max_{l \in \mathcal{L}} R_l^{(t)}]^+ + \Delta_{penalty}. \tag{10}$$
where $\Delta_{penalty}$ is the penalty if (7b) is not satisfied.

B. The Proposed SAC-AO Algorithm

In this section, we introduce the SAC-AO algorithm, which employs the SAC algorithm to control the beamforming matrices for communication and sensing. Subsequently, the AO algorithm is utilized to determine the phase-shift matrix for the RIS.

1) SAC for Transmit Beamforming Optimization: Given the discrete phase shifts $\mathbf{\Phi}$ of RIS, This optimization problem for transmit beamforming can be formulated as follows

$$
\begin{aligned}
\text{(P1.1)}: \max_{\mathbf{w}_c, \mathbf{W}_r} \quad & R_s && \text{(11)}\\
\text{s.t.} \quad & \text{(7b)}, \text{(7c)}. &&
\end{aligned}
$$

Fundament of SAC Algorithm: The SAC algorithm excels in handling continuous action spaces. The introduction of entropy maximization improves exploration, stabilizes training, and reduces sensitivity to hyperparameters [11]. It aims to find the optimal action strategy by maximizing the entropy of the action distribution in the current state while optimizing future cumulative rewards. Therefore, optimal action strategy π^* can be expressed as

$$\pi^* = \arg\max_{\pi} \sum_t \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \xi \mathcal{H}(\pi(\cdot|\mathbf{s}_t)) \right], \tag{12}$$

where $\pi(\mathbf{a}_t|\mathbf{s}_t)$ is a policy that a mapping from state $\mathbf{s}_t$ to action $\mathbf{a}_t$. and ρ_π denotes the state-action trajectory distribution formed by the policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$. $\mathcal{H}(\pi(\cdot|\mathbf{s}_t)) \triangleq -\log(\pi(\cdot|\mathbf{s}_t))$ represents the entropy objective of the policy π. ξ represents the temperature parameter, which can be adjusted to modify the trade-off between entropy and reward. An auto-adjustment method for ξ is introduced in [12], denoted by

$$J(\xi) = \mathbb{E}_{a_t \sim \pi_{\boldsymbol{\psi}}} \left[-\xi \log \pi_{\boldsymbol{\psi}}(\mathbf{a}_t|\mathbf{s}_t) - \xi \bar{\mathcal{H}} \right], \tag{13}$$

where $\bar{\mathcal{H}}$ is the minimum expected entropy of target, and $\pi_{\boldsymbol{\psi}}$ represents the approximation of the policy distribution using a neural network with parameters $\boldsymbol{\psi}$.

During the policy evaluation stage, for a given policy π, the soft-Q function is iteratively computed using the bellman iteration equation

$$Q(\mathbf{s}_t, \mathbf{a}_t) = r_t + \chi \mathbb{E}_{\mathbf{s}_{t+1} \sim p}[V(\mathbf{s}_{t+1})], \tag{14}$$

where $V(\mathbf{s}_{t+1})$ is the soft state value function and can be expressed as

$$V(\mathbf{s}_{t+1}) = \mathbb{E}_{\mathbf{a}_{t+1} \sim \pi} \left[Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - \xi \log \pi(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}) \right]. \tag{15}$$

Similarly, deep neural network (DNN) is employed to approximate the soft Q-function. The network is trained through stochastic gradient descent, parameterizing the soft Q-function with θ and it is trained to minimize squared residual error

$$
\begin{aligned}
J_Q(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}} \Bigg[\frac{1}{2} \big(Q_{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}_t) - (\mathbf{r}(\mathbf{s}_t, \mathbf{a}_t) \\
+ \chi \mathbb{E}_{\mathbf{s}_{t+1} \sim p}[V_{\bar{\boldsymbol{\theta}}}(\mathbf{s}_{t+1})]) \big)^2 \Bigg],
\end{aligned} \tag{16}
$$

where $\mathcal{D}$ represents replay buffer and $V_{\bar{\boldsymbol{\theta}}}(\mathbf{s}_{t+1})$ represents the estimated soft state value obtained through the use of a target neural network.

In the policy improvement phase, the new policy is updated based on the exponential gradient of the Q-function, guided by the Kullback-Leibler (KL) divergence. Therefore, the actor network is updated by minimizing the KL divergence and can be defined as

$$J_\pi(\boldsymbol{\psi}) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[D_{\mathrm{KL}} \left(\pi_{\boldsymbol{\psi}}(\cdot \mid \mathbf{s}_t) \middle\| \frac{\exp(Q_{\boldsymbol{\theta}}(\mathbf{s}_t, \cdot))}{Z_{\boldsymbol{\theta}}(\mathbf{s}_t)} \right) \right], \tag{17}$$

where $Z_{\boldsymbol{\theta}}$ is employed to normalize the distribution.

SAC Algorithm: First, the actor network $\pi_{\boldsymbol{\psi}}$ interacts with the environment, generating an action distribution ($\mathbf{a}_t \sim \pi_{\boldsymbol{\phi}}(\mathbf{a}_t|\mathbf{s}_t)$) based on the current state $\mathbf{s}_t$. Then, the action $\mathbf{a}_t$ and state $\mathbf{s}_t$ are input into the critic network for evaluation to adjust the action policy of the actor network. The critic network consists of two Q networks $Q_{\boldsymbol{\theta}_i}$ and two target

Table I: HYPER-PARAMETERS DESCRIPTIONS

Description	Value	Description	Value
Discount rate(χ)	0.99	Learning rate for actor network	1×10^{-3}
Target smoothing efficient(τ)	5×10^{-4}	Learning rate for critic network	5×10^{-4}
Replay buffer size($\mathcal{D}$)	10^6	Batch size($\mathcal{B}$)	64
Activation function	Relu	Layer hidden units	256
Number of steps for updating the target network	1	Panalty value $(\Delta_{penalty})$	-0.5

networks $Q_{\bar{\boldsymbol{\theta}}_i}, i \in (1,2)$. The minimum Q-value is chosen as the estimated action-state value to prevent overestimation. During the parameter update phase, a minimal batch is selected from the replay buffer using prioritized experience replay (PER) technique for updating the parameters of network.

PER Technique: During the parameter update phase, uniformly random sampling from the replay buffer may lead to training instability. Therefore, the PER [13] method has been proposed to expedite the convergence of the algorithm and stabilize the training process of SAC. Performing priority sampling in the replay buffer, where the sampling probability is proportional to the temporal difference (TD) error. Since higher TD errors can provide more valuable information, the sampling probability can be computed as follows $P(i) = \frac{p_i^{\varrho}}{\sum_j p_j^{\varrho}}$, where $p_i = |\delta_i| + \Delta$ represents the relationship between priority and TD error δ_i, with Δ being a small positive constant. ϱ is a hyperparameter employed to regulate the impact of priority on the sampling probability.

2) AO Algorithm for Designing Phase Shifts of RIS: Given the transmit beamforming matrix according to the SAC algorithm, the optimization problem for the phase shifts of the RIS can be formulated as follow

$$(\text{P2}): \quad \max_{\boldsymbol{\Phi}} \quad R_s \tag{18a}$$

$$\text{s.t.} \quad (7\text{b}), (7\text{d}). \tag{18b}$$

After obtaining the transmit beamforming matrix, we employ a simple AO algorithm to solve for the phase shifts of the RIS. Initially, phase shifts for all RIS elements are randomly generated from the set $\mathcal{F}$. Subsequently, during each element's phase shift update process, while keeping the phase shifts of other elements unchanged, we calculate the illumination power and secure rate for each feasible phase shift setting. We choose the phase shift that satisfies constraint (7b) and maximizes the secure rate as the value for that particular element's phase shift. After implementing a maximum number of iterations I_{max}, the optimized phase matrix for the RIS is obtained.

The SAC-AO algorithm proposed in this paper begins by initializing the relevant network parameters to obtain the initial state of the environment $\mathbf{s}_0$. Then in each iteration t, the agent observes the state $\mathbf{s}_t$ to select the action $\mathbf{a}_t$ based on the action strategy, and then the phase shift of RIS is updated according to AO algorithm. Next, the agent calculates the corresponding reward r_t, and updated the next state $\mathbf{s}_{t+1}$, while $\{(\mathbf{s}_t, \mathbf{a}_t, \mathbf{r}_t, \mathbf{s}_{t+1})\}$ is stored in the replay buffer as a tuple. Finally, the PER technique is used for batch sampling to update network parameters. The detailed algorithm is summarized in Algorithm 1.

IV. NUMERICAL RESULTS

In this section, we present numerical simulation results to demonstrate the performance of the algorithm. We consider the distance-dependent path loss model, i.e., $\beta = \beta_0 (d/d_0)^{-\bar{\alpha}}$, where $\beta_0 = -30$ dB represents the path loss at the reference distance of $d_0 = 1$m, and $\bar{\alpha}$ is the path loss exponent. We set $\bar{\alpha}$ as 2.2 and 2.8 for the BS-RIS and RIS-user link, respectively. Throughout the entire simulation, we set the positions of the BS, RIS, and user as (0, 0, 2.5m), (20m, 0, 2.5m), and (20m, 5m, 0) respectively. Simultaneously, we consider the target located at a random position within a distance of 5 meters from the user. The settings for other relevant parameters are as follows $\sigma^2 = -90$ dBm, $M = 4$, $\kappa = 10$. Besides, we show the network-related hyperparameters in the Table I.

Algorithm 1 SAC-AO algorithm for solving problem (P1)

```
Initialize network parameters $\boldsymbol{\theta_i}$, $\boldsymbol{\psi}$, $\bar{\boldsymbol{\theta}}_i$,$i \in \{1, 2\}$.
Initialize experience replay buffer $\mathcal{D}$
for each iteration do
    Initialize state $\mathbf{s}_0$, $\mathbf{s} \leftarrow \mathbf{s}_0$;
    for each step do
        $\mathbf{a}_t \sim \pi_{\boldsymbol{\psi}}(\mathbf{a}_t|\mathbf{s}_t)$;
        Obtain the phase shifts of the RIS with AO algorithm.
        Obtain $\mathbf{s}_{t+1}$ and the reward $r_t$.
        Updata replay buffer $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, \mathbf{r}_t, \mathbf{s}_{t+1})\}$.
    end for
    for each gradient step do
        Sample batch of experiences $\mathcal{B}$ by PER technique;
        Updata network parameters based on (13), (16), (17);
        $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \chi \nabla_{\boldsymbol{\theta}_i} J_Q(\boldsymbol{\theta}_i)$ for $i \in \{1, 2\}$;
        $\boldsymbol{\psi} \leftarrow \boldsymbol{\psi} - \chi \nabla_{\boldsymbol{\psi}} J_\pi(\boldsymbol{\psi})$;
        $\xi \leftarrow \xi - \chi \nabla_\xi J_\xi(\xi)$;
        Update the parameters of the target network
            $\bar{\boldsymbol{\theta}}_i \leftarrow \tau \boldsymbol{\theta}_i + (1 - \tau) \bar{\boldsymbol{\theta}}_i$ for $i \in \{1, 2\}$,
        where $\tau$ represents the target smoothing coefficient.
    end for
end for
```

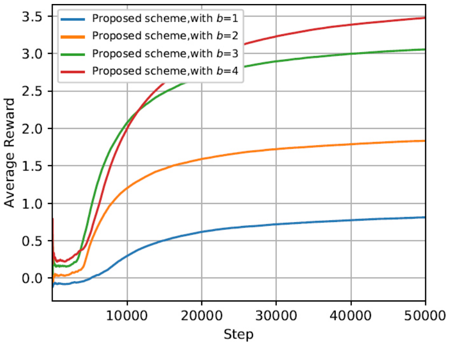

Figure 2: Convergence performance with $T = 4$, $N = 20$, $P_{max} = 10$W, $\varepsilon = 5\mu$W.

First, we validate the convergence of the algorithm, where the average reward value versus the learning step is shown in Fig.2. Simulation results indicate that under different quantization bit settings, the average reward converges to a stable value. Moreover, with an increase in the number of quantization bits, although the convergence rate of the algorithm decreases, higher average reward values can be achieved.

In the proposed scheme, we set $b = 3$ and introduce three benchmarks as comparative schemes.

i) Fixed RIS: Fixed RIS phase shift matrix and optimized beamforming matrix.

ii) Separate Beamforming: The optimization of RIS phase shift is separated from the transmit beamformers. It involves maximizing the norm of the RIS's channel towards the desired sensing target, expressed as $\max_{\mathbf{\Phi}} \|\mathbf{h}_{rl}^{H}\mathbf{\Phi}\mathbf{G}\|$. Subsequently, the transmit beamformers are optimized based on the given phase shift.

iii) Random RIS: The phase shift is randomly obtained from a uniform distribution in the range (0,2π].

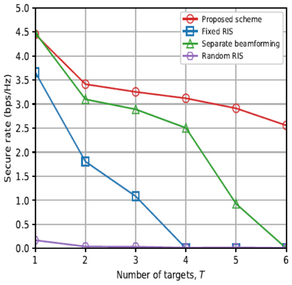

Figure 3: Secure rate versus T with $N = 20$, $P_{max} =$ 10W, $\varepsilon = 5\mu$W

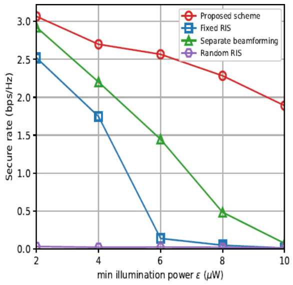

Figure 4: Secure rate versus ε with $T = 4$, $N = 20$, $P_{max} = $ 10W.

Figure 3 depicts the secure rate versus the number of target. It can be observed that as the number of targets increases, and consequently more eavesdroppers are present, the secure rate tends to decrease. In comparison to the benchmarks, our proposed algorithm exhibits better robustness. Figure 4 illustrates the secure rate versus the illumination power threshold. It can be seen that as the ε increases, the secure rate undergoes a decrease. This is due to a trade-off between enhanced sensing performance and a certain degradation in communication performance.

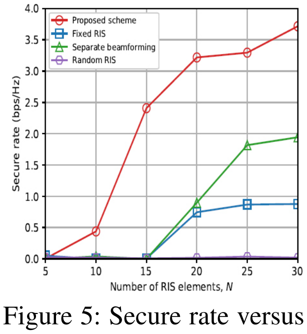

Figure 5: Secure rate versus N with $T = 4$, $P_{max} =$ 10W, $\varepsilon = 5\mu$W

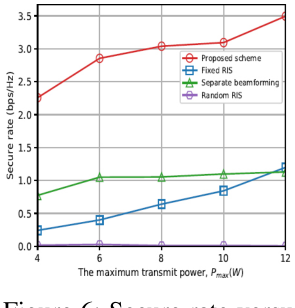

Figure 6: Secure rate versus P_{max} with $T = 4$, $N = 20$, $\varepsilon = 5\mu$W

In Fig. 5, the secure rate exhibits a positive correlation with the number of RIS elements, showcasing an increase in secure communication performance. This is due to the increased number of reflective elements, which facilitates resource allocation and further promotes beamforming gains. Figure 6 illustrates the relationship between the secure rate and $P_{\max}$. As power increases, there is a corresponding positive impact on the secure rate, and the proposed scheme outperforms the other three schemes.

V. Conclusions

In this paper, a RIS-assisted multi-objective single-user secure ISAC system was considered, where a secure rate maximization problem was formulated while ensuring target sensing performance. The SAC-AO algorithm was used to solve this non-convex optimization problem. Simulation results show that the algorithm was effective. In future studies, the focus will lie on improving the sensing performance and considering the synergistic effect between eavesdroppers.

Acknowledgment

This work is supported by National Key Research and Development Program of China under Grant 2023YFE0107900, in part by the National Natural Science Foundation of China under Grant 62071306, and in part by Shenzhen Science and Technology Program under Grants JSGG20210802154203011, JCYJ20200109113601723 and JSGG20210420091805014.

References

[1] ITU-R, Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond, *International Telecommunication Union Radiocommunication Sector (ITU-R)*, Internal Document, September 2023. [Online]. Available: https://www.itu.int/en/ITU- R/study- groups/rsg5/rwp5d/imt- 2030/Pages/default.aspx

[2] F. Liu, Y. Cui, C. Masouros, J. Xu, T. Han, Y. Eldar, and S. Buzzi, "Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728-1767, Jun. 2022.

[3] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent Reflecting Surface-Aided Wireless Communications: A Tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313-3351, May 2021.

[4] X. Song, T. X. Han, and J. Xu, "Cramer-Rao Bound Minimization for IRS-Enabled Multiuser Integrated Sensing and Communication with Extended Target," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Rome, Italy, May 2023, pp. 5725-5730.

[5] Z. Yu, X. Hu, C. Liu, M. Peng, and C. Zhong, "Location Sensing and Beamforming Design for IRS-Enabled Multi-User ISAC Systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 5178-5193, Nov. 2022.

[6] Z. Zhu, Z. Li, Z. Chu, G. Sun, W. Hao, P. Xiao, and I. Lee, "Resource Allocation for IRS Assisted mmWave Integrated Sensing and Communication Systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Seoul, Korea, May 2022, pp. 2333-2338.

[7] J. Chu, Z. Lu, R. Liu, M. Li, and Q. Liu, "Joint Beamforming and Reflection Design for Secure RIS-ISAC Systems," *IEEE Trans. Veh. Technol*, vol. 73, no. 3, pp. 4471-4475, Mar. 2024.

[8] Q. Liu, Y. Zhu, M. Li, R. Liu, Y. Liu, and Z. Lu, "DRL-Based Secrecy Rate Optimization for RIS-Assisted Secure ISAC Systems," *IEEE Trans. Veh. Technol*, vol. 72, no. 12, pp. 16871-16875, Dec. 2023.

[9] S. Zhang and R. Zhang, "Capacity Characterization for Intelligent Reflecting Surface Aided MIMO Communication," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1823-1838, Aug. 2020.

[10] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838-1851, Mar. 2020.

[11] T. Haarnoja, A. Zhou, P . Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. IEEE Int. Conf. Machine Learning (ICML)*, Stockholm Sweden, Jul. 2018, pp. 2976-2989.

[12] T. Haarnoja et al., "Soft actor-critic algorithms and applications," 2018, arXiv: 1812.05905.

[13] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. Int. Conf. Learn. Represent.*, San Juan, Puerto Rico, 2016, pp. 1-9.