利用深度强化学习为智能无线电提供智能反射面控制

赵泽宇

摘要

摘要-智能反射面（Intelligent reflecting surface，IRS）将改变无线通信的模式，从” 适应无线信道” 转变为” 改变无线信道”。然而，目前的 IRS 配置方案包括子信道估计和依次进行的无源波束成形，符合传统的基于模型的设计理念，很难在复杂的无线电环境中实际实现。为了创建智能无线电环境，本文提出了一种独立于子信道信道状态信息（channel state information，CSI）的 IRS 控制无模型设计方案，它要求 IRS 与无线通信系统之间的交互最小。本文首先将 IRS 的控制建模为马尔可夫决策过程（Markov decisionprocess ，MDP），并应用深度强化学习（deep reinforcement learning，DRL）对 IRS 进行实时粗相控制。然后，我们应用极值寻优控制（extremum seeking control，ESC）作为 IRS 的精细相位控制。最后，通过更新帧结构，本文将 DRL 和 ESC 集成到 IRS 的无模型控制中，以提高其对不同信道动态的适应性。数值结果表明了本文提出的 DRL 和 ESC 联合方案的优越性，并验证了它在无子信道 CSI 的无模型 IRS 控制中的有效性。

关键词：智能反射表面，深度强化学习，极值寻优控制，免模型控制。

1 引言

超表面是一类由人工周期性或准周期性结构组成的新型功能材料，其特征在于亚波长的尺度 [1]，[2]。超表面具有负介电常数和负磁导率等非凡的电磁特性，使其在微波到可见光的宽频率范围内具备定制电磁波的潜力 [3]-[6]。智能反射表面（IRS），也称为可重构智能表面（RIS），是一类通过集成有源组件而实现电磁波电子调节的可编程超表面 [7]-[12]。IRS 的出现被视为无线通信等多个行业的潜在颠覆性技术。无线通信容易受到时变无线传播环境的影响，空间路径损耗、信号吸收、反射、折射及绕射效应共同导致信道高度动态化 [13]-[15]。由于 IRS 能够实时操控电磁波，使人们可以从“适应无线信道”转向“改变无线信道”[7]。因此，大量研究致力于获取 IRS 与无线收发器之间的子信道状态信息（CSI），以便设计 IRS 反射模式（即被动波束成形）[8]。然而，由于 IRS 的被动特性，其无法直接感知入射信号，导致信道估计过程远比传统无线通信复杂 [9]-[11]。尽管在 CSI 获取方面取得了一些进展，但 IRS 实际应用仍面临挑战，包括对现有无线通信协议的重新设计需求，实时优化计算成本高，和复杂传播环境下的建模困难 [12], [17], [18]。

2 相关工作

许多研究尝试提高 IRS 辅助无线通信的 CSI 获取技术。例如，文献 [9] 提出一种利用用户间信道相关性以降低训练开销的信道估计方案。文献 [10] 和 [11] 则提出了基于压缩感知的信道估计方法，适用于毫米波频段的基站与单天线用户之间的信道响应。然而，这些方法主要针对单天线用户，扩展到多用户场景会大幅增加训练开销。为此，文献 [16] 提出一种联合波束训练和定位方案，通过广播方式的随机波束成形在训练阶段实现独立于用户数量的训练开销。

尽管这些进展显著，但 IRS 的实际应用仍面临多重挑战。首先，IRS 辅助无线通信的信道估计需要根本性的协议更新，以协调发射器、接收器和 IRS 的操作。其次，即便获得完美的 CSI，利用凸或非凸优化技术实时优化 IRS 反射系数的计算成本依然极高 [12]。此外，当前 IRS 控制方案（CSI 获取及反射设计）依赖于精确的数学建模，然而由于 IRS 的反射系数与入射信号的载波频率相关，这种建模在动态传播环境中难以实现 [17]，[18]。为了应对这些挑战，已有一些研究尝试通过无模型控制方式来实现 IRS 的独立操作。文献 [27] 和 [28] 利用深度学习和深度强化学习方法，引导 IRS 与入射信号交互。然而，为了获取子信道的 CSI，这些方法通常需要在 IRS 上安装信道传感器，有违 IRS 作为被动设备的设计初衷。文献 [29] 提出一种深度学习方案，通过离线训练数据减少对子信道 CSI 的依赖，然而这限制了其在广泛场景中的适应性。

本文则提出一种无模型的 IRS 控制方案，通过深度强化学习（deep reinforcement learning，DRL）和极值寻优控制（extremum seeking control，ESC）优化 IRS 反射系数。不同于现有方法，本文的方案无需子信道 CSI，能够提升 IRS 的独立性并加速其实际部署。

本文的其余部分组织如下：在第 3 节，介绍系统模型。第 4 提出一种基于深度强化学习（DRL）的无模型 IRS 控制方法。第 5 节提出一种基于抖动的迭代方法，以增强 DRL 的控制效果。第 6 节展示数值结果。最后在第 7 节得出结论。

本文的符号定义：矩阵和列向量分别用黑体大小写字母表示，$\mathbf{x}[n]$ 表示向量 $\mathbf{x}$ 中的第 n 个元素，$\odot$ 表示哈达玛乘积，$(\cdot)^*$、$(\cdot)^T$ 和 $(\cdot)^H$ 分别表示共轭、转置和共轭转置运算。

3 本文方法（系统模型）

本节将介绍无模型 IRS 控制的系统模型。

3.1 IRS 的最优相移矢量

对于包含 N_R 个放射元件的 IRS 辅助无线通信，配备 N_B 根天线的基站（Base station，BS）与某一用户 k 之间的信道模型可表示为

$$\mathbf{h}_k = \mathbf{h}_{BU_k} + \mathbf{H}_{BR}\boldsymbol{\Theta}\mathbf{h}_{RU_k} \tag{1}$$

其中，用户 k 配备单天线，$\mathbf{h}_{BU_k} \in \mathbb{C}^{N_B\times 1}$ 是用户 k 与 BS 之间的信道响应向量，$\mathbf{H}_{BR} \in \mathbb{C}^{N_B\times N_R}$ 是 BS 与 IRS 之间的信道响应矩阵、$\mathbf{h}_{RU_k} \in \mathbb{C}^{N_R\times 1}$ 是用户 k 与 IRS 之间的信道响应向量，$\boldsymbol{\Theta} = diag\boldsymbol{\theta}$，其中 $\boldsymbol{\theta} \in \mathbb{C}^{N_R\times 1}$（$|\boldsymbol{\theta}[n]| = 1$）是 IRS 的相移向量。因此，多用户信道可写成

$$\mathbf{H} = \mathbf{H}_{BU} + \mathbf{H}_{BR}\boldsymbol{\Theta}\mathbf{H}_{RU} \tag{2}$$

其中 $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_K]$，$\mathbf{H}_{BU} = [\mathbf{h}_{BU_1}, \mathbf{h}_{BU_2}, ..., \mathbf{h}_{BU_K}]$ 以及 $\mathbf{H}_{RU} = [\mathbf{h}_{RU_1}, \mathbf{h}_{RU_2}, ..., \mathbf{h}_{RU_K}]$。聚合等效信道 $\mathbf{H}$ 与子信道之间的关系如图 1 所示。

基于强化学习的 IRS 配置的目标是开发一种广泛兼容的方法，可以在各种无线通信场景中部署，而无需了解无线系统的内部工作机制。研究问题表述为

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & P_m \\ \text{s.t.} \quad & \boldsymbol{\theta}[n] = e^{-j\boldsymbol{\varphi}[n]},\ \forall n \in \{1, 2, ..., N_R\} \\ & \boldsymbol{\varphi}[n] \in \mathcal{B},\ \forall n \in \{1, 2, ..., N_R\} \end{aligned} \tag{3}$$

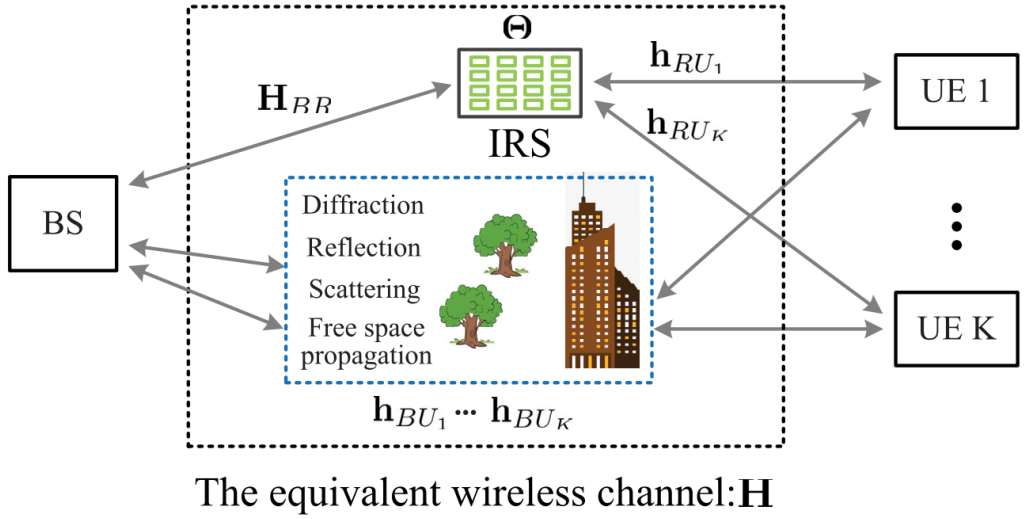

图 1: IRS 辅助无线通信。

其中，P_m 是要优化的无线系统的性能指标。P_m 取决于无线信道 $\mathbf{H}$，而 $\mathbf{H}$ 取决于反射模式 $\mathbf{\Theta}$。$\varphi[n]$ 是从有限集合 $\mathcal{B} = \{-\pi, \frac{-2^r+2}{2^r}\pi, \frac{-2^r+4}{2^r}\pi, ..., \pi\}$ 中选择的量化相位，共有 $2^r + 1$ 个可能值。

3.2 经典场景：TDD 多用户 MIMO

在不失一般性的前提下，本文使用无线通信中的一个典型场景，即 TDD 多用户多输入多输出（Multi input multi output，MIMO）。在 TDD 中，通过利用信道互易性，BS 可以根据上行信道的先导估计下行信道。因此，TDD 多用户 MIMO 包括两个阶段（参见图 2），即上行链路先导传输和下行链路数据传输 [30]，[31]。

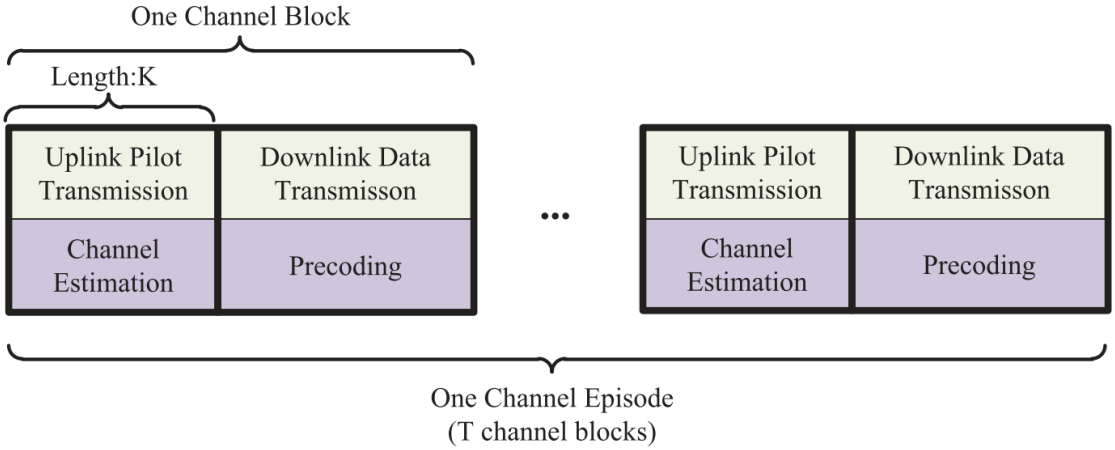

图 2: TDD 多用户 MIMO 的帧结构。

在上行链路阶段，多个用户同时向 BS 发射先导信号。接收到的先导信号表示为

$$\mathbf{Y}_U = \mathbf{HS} + \mathbf{N} \tag{4}$$

其中 $\mathbf{S} \in \mathbb{C}^{K\times K}$ 是导频模式，$\mathbf{N}$ 是加性高斯白噪声。接收到先导信号后，BS 对信道矩阵进行最小均方误差（MMSE）估计，即 $\hat{\mathbf{H}} = \mathbf{Y}_U\mathbf{S}^H(\mathbf{S}\mathbf{S}^H + \sigma_U^2\mathbf{I})^{-1}$，当 $\mathbf{S}$ 为酉矩阵时，其可以化简为

$$\hat{\mathbf{H}} = \frac{\mathbf{Y}_U\mathbf{S}^H}{1+\sigma_U^2}. \tag{5}$$

在下行链路阶段，采用迫零（zero-forcing，ZF）预编码进行数据传输，预编码矩阵表示为

$$\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_K] = \mathbf{D}_p(\hat{\mathbf{H}}^H\hat{\mathbf{H}})^{-1}\hat{\mathbf{H}}^H \tag{6}$$

其中，$\mathbf{D}_p = diag([\frac{1}{|\mathbf{m}_1|_2}, \frac{1}{|\mathbf{m}_2|_2}, ..., \frac{1}{|\mathbf{m}_K|_2}])$ 用于功率归一化。用户 k 的接收信号如下

$$y_{D,k} = \mathbf{m}_k^H \mathbf{h}_k x_k + \sum_{l \neq k}^{K} \mathbf{m}_l^H \mathbf{h}_k x_l + n_k \tag{7}$$

其中，x_k 是发送给用户 $k \in \{1, ..., K\}$ 的信号（$\mathbb{E}(x_k) = 0$ 且 $\mathbb{E}(|x_k|^2) = 1$），$n_k \sim \mathcal{CN}(0, \sigma_k^2)$ 为加性白高斯噪声。因此，第 k 个用户的信噪比为

$$SINR_k = \frac{|\mathbf{m}_k^H \mathbf{h}_k|^2}{\sum_{l \neq K}^{K} |\mathbf{m}_l^H \mathbf{h}_k|^2 + \sigma_k^2} \tag{8}$$

对于通信系统来说，性能指标可以是信噪比（SINR）、数据速率、帧误码率（FER）等。在不失一般性的前提下，本文采用总数据率作为性能指标，即

$$P_m = \sum_{k=1}^{K} r_k = \sum_{k=1}^{K} \log_2(1 + SINR_k) \tag{9}$$

3.3 信道建模

本文假定 $\mathbf{h}_{BU_k}$、$\mathbf{H}_{BR}$ 和 $\mathbf{h}_{RU_k}$ 采用瑞利信道模型。以 $\mathbf{H}_{BR}$ 为例，它表示为

$$\mathbf{H}_{BR} = \sqrt{\frac{K}{K+1}} \mathbf{H}_{BR,LoS} + \sqrt{\frac{1}{K+1}} \mathbf{H}_{BR,NLoS} \tag{10}$$

其中，$\mathbf{H}_{BR,LoS}$ 表示确定性视线（Line of sight，LoS）分量，$\mathbf{H}_{BR,NLoS}$ 表示快速衰减非视线（None line of sight，NLoS）分量，其分量为独立同分布（i.i.d.）的圆周对称复高斯随机变量，均值为零，方差为一，K 为 LoS 路径功率与 NLoS 路径功率之比 [32]。

LoS 分量与位置有关，因此是慢性时变的；NLoS 分量由多径效应引起，因此是快速时变的 [33]。结合无线信道的特点和强化学习的设置，本文引入了以下两个概念：

1）信道 block：一个信道 block 由上行链路先导传输阶段和下行链路数据传输阶段组成（如图 2所示），信道矩阵在信道 block 期间保持不变。**2）信道 episode：**一个信道 episode 由 T 个信道 block 组成（如图 2所示）。一个信道 episode 内的 LoS 分量保持不变；NLoS 分量随时间变化，不同信道 block 的 NLoS 分量均为 i.i.d。

4 本文方法（基于深度强化学习的无模型 IRS 控制）

4.1 深度强化学习基础知识

为便于介绍本文的设计，本小节将简要介绍 DRL 的一些关键概念。

强化学习的目标: 一个 MDP 由四元组 $< \mathcal{S}, \mathcal{A}, P, R >$ 指定，其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，P 是状态转移概率，R 是代理获得的直接奖励。当处于状态 $s \in \mathcal{S}$ 的智能体采取了动作 $a \in \mathcal{A}$ 时，环境将以概率 $P(s'|s, a) = \Pr(S_{t+1} = s'|S_t = s, A_t = a)$ 演化到下一个状态 $s' \in \mathcal{S}$，与此同时，智能体将获得即时奖励 $R^a_{s \to s'}$。在 S、A、R 中加入时间下标后，MDP 的演化可以用下面的轨迹来表示

$$< S_0, A_0, R_1, S_1, A_1, R_2, ..., S_{T-1}, A_{T-1}, R_T, S_T, ... > \tag{11}$$

智能体的动作由策略函数

$$\pi(a|s) = \Pr(A_t = a|S_t = s) \tag{12}$$

所指导，即当当前状态为 s 时，智能体采取动作 a 的概率。强化学习任务的目的是找到一种能获得长

期良好回报的策略，回报被定义为未来的累积折扣奖励，即

$$U_t = \sum_{\tau=0}^{\infty} \gamma^{\tau} R_{t+\tau+1} \tag{13}$$

$\gamma \in [0,1]$ 是未来回报的折扣因子。由于状态转换（由动态环境引起）和动作选择的随机性，回报 U_t 是一个随机变量。从数学上讲，智能体在强化学习中的目标是找到一个能使预期收益最大化的好策略，即

$$\max_{\pi} \mathbb{E}(U_t) \tag{14}$$

动作值函数和最优策略：强化学习中动作选择的一个关键指标是动作值函数，即

$$Q_{\pi}(s,a) = \mathbb{E}[U_t | S_t = s, A_t = a], \tag{15}$$

其表示在策略 π 下，从状态 s 中选择动作 a 的预期收益。对于任意策略 π 和任意状态 $s \in \mathcal{S}$，动作值函数满足以下递归关系

$$Q_{\pi}(s,a) = \mathbb{E}_{s'}\left[R^{a}_{s\to s'} + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|s') Q_{\pi}(a', s') \middle| s', a'\right] = \sum_{s' \in \mathcal{S}} P(s'|s,a)\left(R^{a}_{s\to s'} + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|s') Q_{\pi}(a', s')\right), \tag{16}$$

其中 $R^{a}_{s\to s'}$ 是当采取动作 a 后，环境从状态 s 转到状态 s' 时的直接奖励，公式 (16) 是著名的动作值函数的贝尔曼方程 [34]。

如果一个策略在所有状态和所有动作下的预期收益都高于另一个策略，那么这个策略就被定义为优于另一策略。因此，最优动作值函数为

$$Q^{*}(s,a) = Q_{\pi^{*}}(s,a) = \max_{\pi} Q_{\pi}(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{17}$$

有了最优动作值函数，最优政策定义为

$$\pi^{*}(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a \in \mathcal{A}} Q^{*}(s,a), \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

将 (17)、(18) 与 (16) 相结合，$Q^{*}(s,a)$ 的贝尔曼最优方程为

$$Q^{*}(s,a) = \sum_{s'} P(s'|s,a)\left(R^{a}_{s\to s'} + \gamma \max_{a' \in \mathcal{A}} Q^{*}(s', a')\right) = \mathbb{E}_{s'}\left[R^{a}_{s\to s'} + \gamma \max_{a' \in \mathcal{A}} Q^{*}(s', a') \middle| s_t = s, a_t = a\right]. \tag{19}$$

利用贝尔曼最优方程，可以通过迭代法（即基于策略的迭代法和基于值的迭代法）得到最优策略 $\pi^{*}(a|s)$ 或最优动作值函数 $Q^{*}(s,a)$ [34]。本文将主要使用基于值迭代的方法。

时序差分：上述迭代法需要完整的环境知识，即状态转移概率 $P(s'|s,a)$、奖励函数 $R^{a}_{s\to s'}$ 等。然而，在实践中通常无法获得环境动态的明确知识。(19) 中的条件期望可以通过对与环境实际交互的状态、动作和奖励的样本序列进行数值平均来实现，例如时序差分法或蒙特卡罗法。

当观察到 (11) 中新的轨迹段时，即 $< S_t = s, A_t = a, R_{t+1} = R^{a}_{s\to s'}, S_{t+1} = s' >$，动作值函数 $Q(s,a)$ 更新如下：

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\left(R^{a}_{s\to s'} + \gamma \max_{a' \in \mathcal{A}} Q_t(s', a') - Q_t(s,a)\right), \tag{20}$$

其中 $\alpha \in (0,1]$ 是学习率，括号内的项是估计动作值函数与目标值之间的误差。这意味着值函数会沿着误差的方向更新，当误差变得无穷小时，迭代就会终止。

双重深度 Q 网络（Double Deep Q-Network，DDQN）：当状态 s 和动作 a 都是离散的时候，最佳状态-动作函数 $Q^*(s,a)$ 可以通过查找表（也称为 Q-表格 [36]）的形式，按照 (20) 中的迭代过程得到。然而，状态（或动作）空间的大小可能相当之大，状态（或行动）甚至可能是连续的。在这种情况下，用查找表来表示 $Q(s,a)$ 是不切实际的。幸运的是，可以采用深度神经网络（Deep neural network，DNN）将 Q-表格近似为 $Q(s,a) \approx \tilde{Q}(s,a;\mathbf{w})$，从而使强化学习能够扩展到更广泛的决策问题。$\tilde{Q}(s,a;\mathbf{w})$ 的系数 **w** 是 DNN 的权重，这里的 DNN 又被称为深度 Q 网络（Deep Q-network，DQN）[36], [37]。

(11) 中的轨迹段 $< S_t = s, A_t = a, R_{t+1} = R^a_{s\to s'}, S_{t+1} = s' >$ 构成一个" 经验样本"，将用于训练 DQN，根据（20），DQN 训练过程中采用的损失函数为

$$Loss = \Big(\underbrace{R^a_{s\to s'} + \gamma \max_{a'\in\mathcal{A}} \tilde{Q}(s', a'; \mathbf{w})}_{T_{DQN}} - \tilde{Q}(s,a;\mathbf{w}) \Big)^2 \tag{21}$$

其中，T_{DQN} 是 Q 网络的目标值。

目标 T_{DQN} 取决于即时奖励 $R^a_{s\to s'}$，以及 DQN 的输出 $\tilde{Q}(s,a;\mathbf{w})$。这种结构不可避免地会导致在训练过程中过高的估计动作状态值（又称 Q 值），从而大大降低 DRL 的性能。为了减少高估，本文将采用 DDQN 结构 [39]。DDQN 的基本思想是应用单独的目标网络 $\tilde{Q}(s', a';\mathbf{w})$ 来估计目标值 [39]，DDQN 中目标的表达式为

$$T_{DQN} = R^a_{s\to s'} + \gamma \tilde{Q}\left(s', \arg\max_{a'\in\mathcal{A}} \tilde{Q}(s', a'; \mathbf{w}); \mathbf{w}^- \right) \tag{22}$$

总而言之，DDQN 与 DQN 的区别在于以下两个方面，即：(1) 使用权重为 **w** 的 Q-网络来选择最优行动；(2) 目标值 Q-值取自权重为 $\mathbf{w}^-$ 的目标 Q-网络。

4.2 利用深度强化学习实现 IRS 的无模型控制

为了将 DRL 应用于无模型的 IRS 配置，本文首先将 IRS 辅助无线通信建模为 MDP。

1）智能体：智能体是 IRS 控制器，能够通过 IRS 自主与环境互动，以实现设计目标。

2）环境：环境是指与智能体交互的事物，包括 BS、无线信道、IRS 和移动用户。

3）状态：为了便于准确预测给定行动的预期奖励和下一个状态，定义状态为 $\{\mathbf{H}, \boldsymbol{\theta}\}$，它由两个子状态组成，即等效无线信道 **H** 和 IRS 的反射向量 $\boldsymbol{\theta}$。

4）动作：动作被定义为当前反射模式的增量相移，即

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} \odot \triangle\boldsymbol{\theta}^{(t)} \tag{23}$$

其中，$\boldsymbol{\theta}^{(t)}$ 是第 t 个信道 block 的反射模式，$\triangle\boldsymbol{\theta}^{(t)}$ 是 $\boldsymbol{\theta}^{(t)}$ 的增量相移。本文使用离散傅立叶变换（DFT）矢量的子集（或全集）作为作用集。例如，当动作空间大小为 5 时，我们设置

$$\mathcal{A} = \{\mathrm{v}(-\frac{6}{N_R}), \mathrm{v}(-\frac{2}{N_R}), \mathrm{v}(0), \mathrm{v}(\frac{2}{N_R}), \mathrm{v}(\frac{6}{N_R}), \}$$

其中 $\mathrm{v}(\Psi_R)$ 为转向向量（在不失一般性的前提下，本文假设 IRS 的反射器阵列是一个均匀线性阵列。），即 $\mathrm{v}(\Psi_R) = \left[1, e^{j\pi\Psi_R}, ..., e^{j(N_R-1)\pi\Psi_R}\right]^T$。当 $\triangle\boldsymbol{\theta}^{(t)} = \mathbf{v}(0)$ 时，子状态 $\boldsymbol{\theta}$ 保持不变，而子状态 **H** 只是由于信道 NLoS 成分的变化而发生变化；$\triangle\boldsymbol{\theta}^{(t)} = \mathbf{v}(-\frac{2}{N_R})$ 和 $\triangle\boldsymbol{\theta}^{(t)} = \mathbf{v}(\frac{2}{N_R})$ 方向相反，这使得智能体能够迅速从负面动作中纠正过来；$\triangle\boldsymbol{\theta}^{(t)} = \mathbf{v}(-\frac{6}{N_R})$ 和 $\triangle\boldsymbol{\theta}^{(t)} = \mathbf{v}(\frac{6}{N_R})$ 用于加快反射模式的转换。使用增量相移而非绝对相移作为动作的原因有两个。一方面由马尔可夫的状态转移特性决定，另一方面，可以缩小动作空间的大小，加快收敛速度。

5）奖励：在状态 s 下执行动作 a 到达下一步状态 s' 所得到的奖励其函数定义如下：

$$R = \begin{cases} P_m, & \text{when } P_m \geqslant P_{th}, \\ P_m - 100, & \text{when } P_m < P_{th}, \end{cases} \tag{24}$$

其中，P_{th} 是性能阈值。当 P_m 小于 P_{th} 时，我们会增加 -100 的惩罚，以鼓励 IRS 最大限度地提高性能，同时将性能维持在阈值之上。

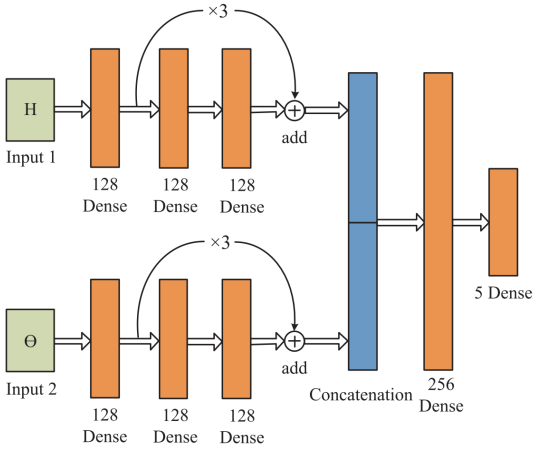

图 3: 网络结构。

根据建模的 MDP 和 4.1 小节介绍的 DRL 基本原理，本文设计了算法 1 来最大化在 (13) 中的累积折扣奖励 U_t。算法 1 中应用的一些关键技术解释如下:

(1) $DDQN$：与简单的 DQN 方法不同，在 DQN 中使用带有权重 $\mathbf{w}$ 的 $\tilde{Q}(s,a;\mathbf{w})$ 来生成目标值，而在 DDQN 中使用一个独立的目标网络 $\tilde{Q}(s,a;\mathbf{w}^-)$（带有权重 $\mathbf{w}^-$）来生成目标值，并且目标网络的权重每 N_{TNet} 个时间间隔通过 $\mathbf{w}^- = \mathbf{w}$ 进行更新。网络的结构如图 3所示。具体来说，本文应用深度残差网络（ResNet）[40] 来处理两个子状态（即 $\mathbf{H}$ 和 $\boldsymbol{\theta}$），然后使用两层密集网络融合这两个子状态。本文使用的激活函数是 Swish 函数 [41]。

(2) $\varepsilon - Greedy$: 给定最优的 $\tilde{Q}(s,a;\mathbf{w})$，最优策略是选择产生最大状态-动作值的动作。然而，最优的 $\tilde{Q}(s,a;\mathbf{w})$ 需要无限量的经验，这在动态无线环境中是不切实际且不可行的。因此，为了防止陷入次优策略，智能体需要持续探索。为此，本文采用 $\varepsilon - Greedy$ 策略。在 $\varepsilon - Greedy$ 策略中，ε 指的是选择探索的概率，即从所有可能的动作中随机选择，而 1−ε 则是根据已获得的 DQN 做出决策时选择利用的概率。因此，$\varepsilon - Greedy$ 策略可以表示为：

$$\pi^{\varepsilon} = \begin{cases} \pi^*(a|s), & w.p.\ 1-\varepsilon \\ P(a) = \dfrac{1}{|\mathcal{A}|}, & w.p.\ \varepsilon \end{cases} \tag{25}$$

其中，$\pi^*(a|s)$ 为 (18) 中的基于 Q-网络的策略。在本文的设计中，ε 的初始值为 1，并以以 θ 的速率指数递减（$0 < \theta < 1$），直至达到下限 ε_{min}。

(3) $Experience\ Replay$：本文不使用最新的经验来训练 DDQN 而是以“先进先出”（FIFO）的方式在经验回放池 $\mathcal{M}$ 中存储最近的经验，然后从中随机获取一个批量大小为 N_e 的小批量经验进行训练。

Procedure 1 基于 DDQN 的无模型 IRS 控制。

Initialize parameters s_0, ε; Initialize the FIFO memory $\mathcal{M}$ with the size N_m;
Initialize the weights of the DQN $\mathbf{w}$ and set the targetnetwork as $\mathbf{w}^- = \mathbf{w}$.
for $t = 0, 1, 2, \cdots$ **do**
  Input s_t to the DQN and obtain the state-action values $\tilde{Q}(s_t, a; \mathbf{w}), a \in \mathcal{A}$;
  With $\tilde{Q}(s_t, a; \mathbf{w}), a \in \mathcal{A}$, select an action a_t using ε-greedy policy;
  Receive the reward r_{t+1} and the estimated channel response $\hat{\mathbf{H}}_{t+1}$, and compute the next state s_{t+1} from $\hat{\mathbf{H}}_{t+1}$, s_t and a_t;
  Store the experience tuple $< s_t, a_t, r_{t+1}, s_{t+1} >$ to the FIFO memory $\mathcal{M}$.
  if $|\mathcal{M}| > N_e$ **then**
    Randomly select a mini batch of N_e experience tuples $< s_i, a_i, r_{i+1}, s_{i+1} >$ from $\mathcal{M}$.
    Calculate the target values $T_{DQN,i}$ for the mini batch according to (22).
    With the input $\{s_i\}$ and the output $\{T_{DQN,i}\}$, train the DQN, and update its weights $\mathbf{w}$.
    if t mod $N_{TNet} = 0$ **then**
      update the weights of the target network, i.e., set $\mathbf{w}^- = \mathbf{w}$.
    end
  end
end

4.3 无模型 IRS 的工作流程

根据图 4，在特定的循环中，IRS 根据 $\varepsilon - Greedy$ 策略配置反射模式（$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} \odot \triangle\boldsymbol{\theta}^{(t)}$），然后用户设备和 BS 依次执行上行导频传输和下行数据传输，就像不存在 IRS 一样。之后，BS 将估计的信道矩阵 $\hat{\mathbf{H}}^{(t+1)}$ 发送给 IRS 控制器，这可以通过有线通信实现，同时用户设备将它们的性能指标反馈给 IRS 控制器。最后，IRS 控制器根据接收到的信道估计 $\hat{\mathbf{H}}^{(t+1)}$ 和性能反馈 P_m^{t+1}，推导出经验元组 $< \{\hat{\mathbf{H}}^{(t)}, \boldsymbol{\theta}^{(t)}\}, \triangle\boldsymbol{\theta}^{(t)}, R^{t+1}, \{\hat{\mathbf{H}}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}\} >$，并将其存储在先进先出队列中作为 DDQN 的训练数据。

与传统的时分双工多用户 MIMO 相比，整合 IRS 所需的额外努力仅在于 $\hat{\mathbf{H}}^{(t+1)}$ 和 P_m^{t+1} 的反馈。前者可以通过 BS 与 IRS 之间的有线通信轻松实现，后者则消耗移动用户设备几乎可以忽略不计的无线通信资源。值得注意的是，IRS 控制器不了解基站和用户设备的工作机制，也不需要子信道的信道状态信息。

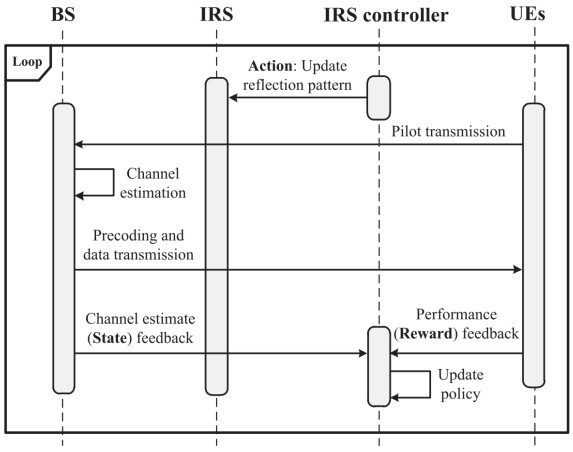

图 4: 无模型 IRS 配置流程

5 本文方法（利用极值寻优控制加强 IRS 控制）

在深度强化学习中，为了加快收敛速度，动作空间受到限制，这限制了 IRS 的相位自由度。为了增强对 IRS 的控制，另一种无模型实时优化方法极值搜索控制（Extremum seeking control，ESC），被用来设计 IRS 的精细相位控制。

5.1 利用 ESC 对 IRS 进行无模型控制

极值搜索控制（ESC）是一种无模型的方法，用于实现基于学习的自适应控制器，以最大化或最小化某些系统性能指标 [42], [43]。ESC 的基本思想是在系统输入中添加抖动信号（例如，正弦波信号和随机噪声），并通过观察其对输出的影响来获取未知系统的非线性静态映射的近似隐梯度。根据 ESC 的设计理念，本文提出了基于抖动的 IRS 无模型控制方法，其包括三个部分，即抖动信号生成模块、上升方向估计模块和参数更新模块。抖动信号生成模块生成随机抖动/扰动信号以探测系统的响应 $P_m(\cdot)$；梯度估计模块根据系统性能 $P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi})$ 确定系统输入的更新方向，以保证性能的单调增加，同时指导随机抖动信号生成的开关；参数更新模块根据估计的方向更新系统输入 $\boldsymbol{\varphi}$。具体来说，迭代过程如下：

（1）**步骤一**，抖动信号生成：

通过均匀随机分布生成一个小的随机抖动信号，即

$$\Delta\boldsymbol{\varphi}[n]=\frac{a}{2^{r-1}},\ \ a\in\mathcal{U}\{-\frac{2^{r-1}}{N_R},\frac{2^{r-1}}{N_R}\} \tag{26}$$

然后，将抖动信号 $\triangle\boldsymbol{\varphi}$ 加到参数 $\boldsymbol{\varphi}$ 上，使用 $\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi}$ 作为系统的输入，并接收性能指标 $P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi})$ 的反馈。

（2）**步骤二**，方向估计和参数更新：

条件一。如果 $P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi})\geqslant P_m(\boldsymbol{\varphi})$，则采用 $\Delta\boldsymbol{\varphi}$ 作为方向。参数更新为

$$\boldsymbol{\varphi}\leftarrow\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi} \tag{27}$$

并更新性能指标为

$$P_m(\boldsymbol{\varphi})\leftarrow P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi}) \tag{28}$$

然后，跳转到步骤一开始下一次迭代；

条件二。如果 $P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi})<P_m(\boldsymbol{\varphi})$，则采用 $-\Delta\boldsymbol{\varphi}$ 作为方向。更新参数为

$$\boldsymbol{\varphi}\leftarrow\boldsymbol{\varphi}-\Delta\boldsymbol{\varphi} \tag{29}$$

关闭抖动信号，仅使用 $\boldsymbol{\varphi}$ 作为系统输入，并测量系统性能 $P_m(\boldsymbol{\varphi})$。然后，跳转到步骤一开始下一次迭代。

命题 1：在基于 ESC 的迭代过程中，每次迭代都可以保证性能指标 P_m 的单调增加，前提是随机抖动信号的范数，即 $|\Delta\boldsymbol{\varphi}|$，足够小。

证明：要证明命题 1，关键是要验证步骤二中的条件二的操作 (29) 能够保证 P_m 的增加，即当 $P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi})<P_m(\boldsymbol{\varphi})$ 时，下列不等式 $P_m(\boldsymbol{\varphi}-\Delta\boldsymbol{\varphi})>P_m(\boldsymbol{\varphi})$ 成立。为此，将 $P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi})$ 在 $\boldsymbol{\varphi}$ 点进行泰勒展开可得

$$P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi})=P_m(\boldsymbol{\varphi})+\frac{\partial P_m(\boldsymbol{\varphi})}{\partial\boldsymbol{\varphi}^H}\Delta\boldsymbol{\varphi}+\mathcal{O}(|\Delta\boldsymbol{\varphi}|^2),\ \ \Delta\boldsymbol{\varphi}\to 0.$$

由于 $|\Delta\boldsymbol{\varphi}|$ 足够小，即 $|\Delta\boldsymbol{\varphi}|\to 0$，因此可以取其一阶近似：

$$P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi})\approx P_m(\boldsymbol{\varphi})+\frac{\partial P_m(\boldsymbol{\varphi})}{\partial\boldsymbol{\varphi}^H}\Delta\boldsymbol{\varphi}$$

根据条件二可知 $P_m(\boldsymbol{\varphi}+\Delta\boldsymbol{\varphi}) < P_m(\boldsymbol{\varphi})$，于是可以得出 $\dfrac{\partial P_m(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}^H}\Delta\boldsymbol{\varphi} < \mathbf{0}$ 然后，很容易验证

$$P_m(\boldsymbol{\varphi}-\Delta\boldsymbol{\varphi}) \approx P_m(\boldsymbol{\varphi}) - \frac{\partial P_m(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}^H}\Delta\boldsymbol{\varphi} > P_m(\boldsymbol{\varphi})$$

成立。

因为 $\boldsymbol{\theta}[n] = e^{-j\pi\boldsymbol{\varphi}[n]}$，所以操作 (27) 和 (29) 可以写为如下形式

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} \odot \Delta\boldsymbol{\theta}, \text{ 和 } \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} \odot \Delta\boldsymbol{\theta}^* \tag{30}$$

其中，$\Delta\boldsymbol{\theta} = e^{-j\pi\Delta\boldsymbol{\varphi}}$，$\Delta\boldsymbol{\theta}^*$ 表示 $\Delta\boldsymbol{\theta}$ 的共轭。

5.2 将 ESC 集成到 DRL 中

为了使 DRL 达到快速收敛的目的，在本文中有意限制了 DRL 的动作空间。而基于抖动的迭代法则依赖于小尺度相移。因此基于抖动的迭代法是 DRL 动作的补充，可用于增强 DRL 的动作。DRL 的增强动作定义如下。

步骤一，粗相移（Coarse phase shift，CPS）：当 $l=0$ 时，设置

$$\boldsymbol{\theta}_{temp}^{(t+1)} = \boldsymbol{\theta}^{(t)} \odot \Delta\boldsymbol{\theta}_c^{(t)}, \tag{31}$$

其中，$\boldsymbol{\theta}^{(t)}$ 是第 t 个信道 block 的反射模式，$\Delta\boldsymbol{\theta}_c^{(t)}$ 是第 t 个信道 block 的粗增量相移其由 DRL 决定，而 $\boldsymbol{\theta}_{temp}^{(t+1)}$ 是第 $t+1$ 个信道 block 的临时反射模式。按照第 4.2 节中的例子，增量相移 $\Delta\boldsymbol{\theta}_c^{(t)}$ 的动作集为 $\mathcal{A} = \left\{\mathbf{v}\left(-\dfrac{6}{N_R}\right), \mathbf{v}\left(-\dfrac{2}{N_R}\right), \mathbf{v}(0), \mathbf{v}\left(\dfrac{2}{N_R}\right), \mathbf{v}\left(\dfrac{6}{N_R}\right)\right\}$。

步骤二，细相移（Fine phase shift，FPS）：对于从 1 到 L，执行以下操作：

$$\boldsymbol{\theta}_{temp}^{(t+1)} = \boldsymbol{\theta}_{temp}^{(t+1)} \odot \Delta\boldsymbol{\theta}_f, \tag{32}$$

其中，$\Delta\boldsymbol{\theta}_f = \Delta\boldsymbol{\theta}$ 或者 $\Delta\boldsymbol{\theta}_f = \Delta\boldsymbol{\theta}^*$，而 $\Delta\boldsymbol{\theta}$ 和 $\Delta\boldsymbol{\theta}^*$ 是 (30) 中的随机抖动信号。例如，当量化等级 $r=8$，且 $N_R = 32$ 时，粗相移的步长为 $\dfrac{2\pi}{32}$，细相移的步长为 $\dfrac{2\pi}{256}$，范围为 $[-\dfrac{\pi}{32}, \dfrac{\pi}{32}]$。为了兼容增强动作，帧结构需要进行更新，如图 5所示。在前 K 个时隙中，用户设备传输导频信号，BS 使用 (31) 中的反射模式执行信道估计，随后在接下来的 $(L-1)K$ 个时隙中，用户设备重复传输导频信号，BS 执行信道估计，同时反射模式按照 (32) 式进行更新。值得注意的是，由于性能反馈每个信道 block 只进行一次，因此基于抖动的方法中使用的性能指标 P_m 是通过用信道估计 $\hat{\mathbf{H}}$ 替代 (2) 中的真实信道响应 $\mathbf{H}$ 得出的一个近似值。需要注意的是，参数 L 可以根据信道动态自适应设置。对于一个实际的无线通信系统，不同的 L 值对应于不同的模式。

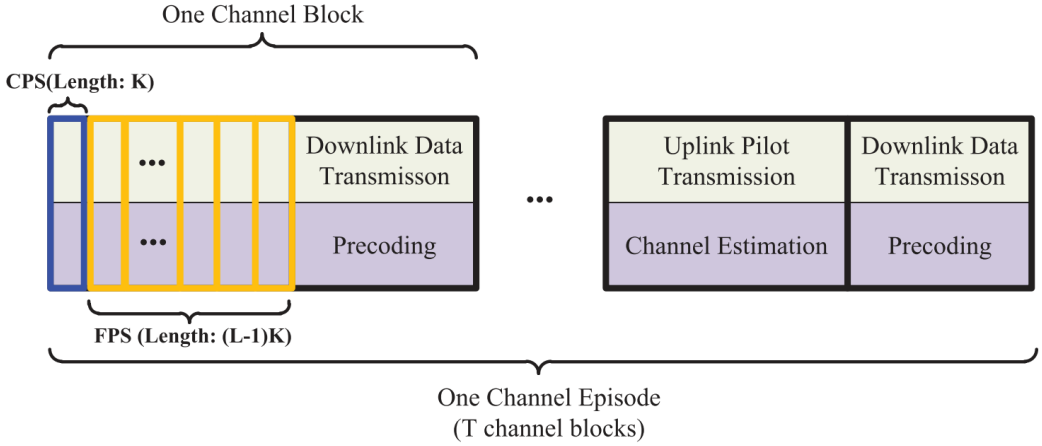

图 5: 将 ESC 集成到 DRL 中的帧结构

6 复现细节

6.1 与已有开源代码对比

本篇文章的代码已经公开，但是其所用的框架为 TensorFlow，且运行时会报错。因此论文原作者提供的代码只供参考并不能直接运行。为此我用 Pytorch 框架重新对本篇文章进行复现，并将本文所用到的 DDQN 算法用 SAC 算法 [44] 所替代。其优势在于 SAC 算法引入了信息熵，相比于 DDQN 算法它能使智能体更充分的探索环境进而避免陷入局部最优。对于 SAC 算法的细节描述可以参考文献 [44]。

6.2 实验环境搭建

本次复现的开发环境为 Python 3.11.7, 共有 5 个文件。本次复现的主要工作在于对本文 DDQN 算法的实现，以及对已有方法的改进。在文件 Main.py 中，我们对系统进行初始化以及相关参数的初始化。在文件 MuMIMOClass.py 中是对 TDD 多用户 MIMO 系统的环境建模。文件 DQN.py 和 SAC.py 则包含了对 DDQN 算法和 SAC 算法的实现。文件 plot_result.py 为复现结果展示。

7 实验结果分析

本次实验进行了两组对比。左图为对原文的复现，分别为 DDQN 算法、随机相移和 DDQN+Dither 算法。实验结果表面，单独使用 DDQN 算法对于性能的提升并不大，而加入了随机抖动方法后性能有较大提升。右图为对原文方法的改进，分别为 SAC 算法、SAC+Dither 算法、DDQN 算法和 DDQN+Dither 算法。从右图中可以发现我们用 SAC 算法替换 DDQN 算法后性能进一步得到了提升。

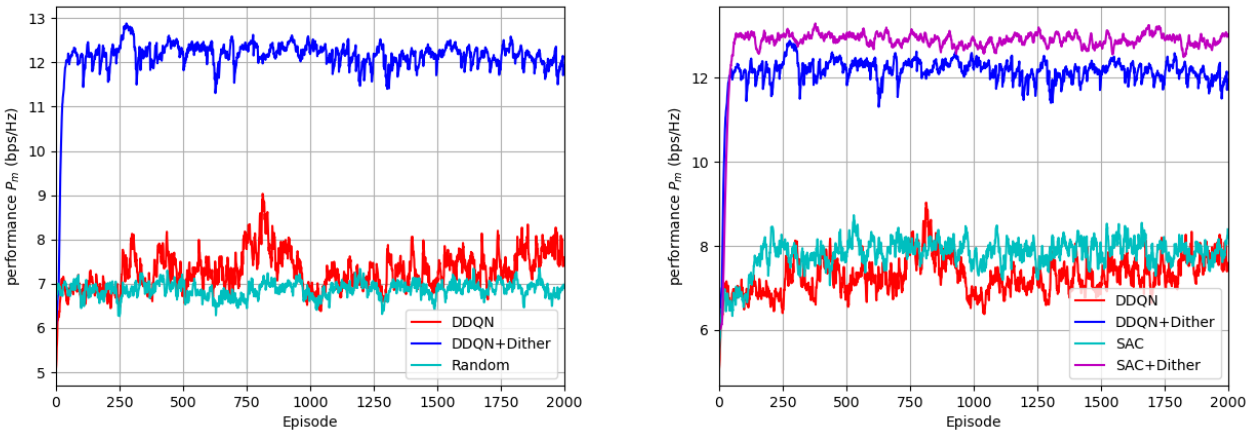

图 6: 实验结果示意

8 总结与展望

在本次课程设计中，我学习到了有关 IRS 的相关知识，这对我未来的研究有极大的帮助，此外还加强了我对深度强化学习的理解。对于本篇文章而言，通过对实验结果的分析，我发现本文所提出的深度强化学习对于性能的提升比较有限，主要还是靠随机抖动信号生成才能使性能有较大提升。而这需要用户向基站发送更多的导频信号用于信道估计，对于一些实时性比较高的环境而言这将不切实际，因此如何设计一个只需要少量导频传输且还能提升系统性能的方法是未来研究的重点。

9 参考文献

[1] N. Engheta and R. W. Ziolkowski, Metamaterials: Physics and Engineering Explorations. Hoboken, NJ, USA: Wiley, 2006.

[2] T. J. Cui, M. Q. Qi, X. Wan, J. Zhao, and Q. Cheng, “Coding metamaterials, digital metamaterials and programmable metamaterials,” Light,Sci. Appl., vol. 3, no. 10, p. e218, Oct. 2014.

[3] D. Schurig et al., “Metamaterial electromagnetic cloak at microwave frequencies,” Science, vol. 314, no. 5801, pp. 977–980, Oct. 2006.

[4] L. Liang et al., “Anomalous terahertz reflection and scattering by flexible and conformal coding metamaterials,” Adv. Opt. Mater., vol. 3, no. 10,pp. 1374–1380, Oct. 2015.

[5] N. Yu et al., “Light propagation with phase discontinuities: Generalized laws of reflection and refraction,” Science, vol. 334, no. 6054,pp. 333–337, Oct. 2011.

[6] C. Zhang, W. Chen, Q. Chen, and C. He, “Distributed intelligent reflecting surfaces-aided device-to-device communications system,” J. Commun. Inform. Netw., vol. 6, no. 3, pp. 197–207, 2021.

[7] M. Di Renzo et al., “Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come,” EURASIPJ. Wireless Commun. Netw., vol. 2019, no. 1, pp. 1–20, 2019.

[8] 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.

[9] Z. Wang, L. Liu, and S. Cui, “Channel estimation for intelligent reflecting surface assisted multiuser communications,” in Proc. IEEE WirelessCommun. Netw. Conf. (WCNC), May 2020, pp. 1–6.

[10] P. Wang, J. Fang, W. Zhang, and H. Li, “Fast beam training and alignment for IRS-assisted millimeter wave/terahertz systems,” 2021,arXiv:2103.05812.

[11] P. Wang, J. Fang, H. Duan, and H. Li, “Compressed channel estimation for intelligent reflecting surface-assisted millimeter wave systems,” IEEESignal Process. Lett., vol. 27, pp. 905–909, 2020.

[12] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming,” IEEE Trans. WirelessCommun., vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[13] D. Tse and P. Viswanath, Fundamentals of Wireless Communication. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[14] C. Qi, P. Dong, W. Ma, H. Zhang, Z. Zhang, and G. Y. Li, “Acquisition of channel state information for mmWave massive MIMO: Traditionaland machine learning-based approaches,” Sci. China Inf. Sci., vol. 64,no. 8, pp. 1–16, Aug. 2021.

[15] F. Liu, J. Pan, X. Zhou, and G. Y. Li, “Atmospheric ducting effect in wireless communications: Challenges and opportunities,” J. Commun.Inform. Netw., vol. 6, no. 2, pp. 101–109, 2021.

[16] W. Wang and W. Zhang, “Joint beam training and positioning for intelligent reflecting surfaces assisted millimeter wave communications,” IEEETrans. Wireless Commun., vol. 20, no. 10, pp. 6282–6297, Oct. 2021.

[17] H. Yang et al., “A programmable metasurface with dynamic polarization, scattering and focusing control,”

Sci. Rep., vol. 6, no. 1, pp. 1–11,Oct. 2016.

[18] W. Tang et al., “Wireless communications with programmable metasurface: Transceiver design and experimental results,” China Commun.,vol. 16, no. 5, pp. 46–61, May 2019.

[27] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb, “Deep reinforcement learning for intelligent reflecting surfaces: Towards standaloneoperation,” in Proc. IEEE 21st Int. Workshop Signal Process. Adv.Wireless Commun. (SPAWC), May 2020, pp. 1–5.

[28] A. Taha, M. Alrabeiah, and A. Alkhateeb, “Enabling large intelligent surfaces with compressive sensing and deep learning,” IEEE Access,vol. 9, pp. 44304–44321, 2021.

[29] B. Sheen, J. Yang, X. Feng, and M. M. U. Chowdhury, “A deep learning based modeling of reconfigurable intelligent surface assisted wirelesscommunications for phase shift configuration,” IEEE Open J. Commun.Soc., vol. 2, pp. 262–272, 2021.

[30] Y. Kim, G. Miao, and T. Hwang, “Energy efficient pilot and link adaptation for mobile users in TDD multi-user MIMO systems,” IEEETrans. Wireless Commun., vol. 13, no. 1, pp. 382–393, Jan. 2014.

[31] W. Zhang et al., “Large-scale antenna systems with UL/DL hardware mismatch: Achievable rates analysis and calibration,” IEEE Trans. Commun., vol. 63, no. 4, pp. 1216–1229, Apr. 2015.

[32] A. Goldsmith, Wireless Communications. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[33] A. K. Samingan, I. Suleiman, A. A. A. Rahman, and Z. M. Yusof, “LTF-based vs. pilot-based MIMO-OFDM channel estimation algorithms: An experimental study in 5.2 GHz wireless channel,” inProc. IEEE 9th Malaysia Int. Conf. Commun. (MICC), Dec. 2009,pp. 794–800.

[34] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA, USA: MIT Press, 2018.

[36] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” IEEE Signal Process.Mag., vol. 34, no. 6, pp. 26–38, Nov. 2017.

[37] V. Mnih et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[39] H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double Q-learning,” in Proc. AAAI Conf. Artif. Intell., 2016, vol. 30,no. 1, pp. 1–7.

[40] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR), Jun. 2016, pp. 770–778.

[41] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for activation functions,” 2017, arXiv:1710.05941.

[42] K. B. Ariyur and M. Krsti´c, Real-Time Optimization by Extremum Seeking Control. Hoboken, NJ, USA: Wiley, 2003.

[43] D. Nesic, A. Mohammadi, and C. Manzie, “A framework for extremum seeking control of systems with parameter uncertainties,” IEEE Trans.Autom. Control, vol. 58, no. 2, pp. 435–448, Feb. 2013.

[44] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Offpolicy maximum entropy deep reinforcement learning with a stochasticactor,” in International conference on machine learning. PMLR, 2018,pp.

1861–1870.