可重构智能表面赋能的无线信道均衡：自适应RIS均衡器与深度强化学习

Gal Ben-Itzhak，IEEE研究生会员；Ender Ayanoglu，IEEE会士

arXiv:2603.02489v1 [cs.IT] 3 Mar 2026

摘要

可重构智能表面能够重塑无线传播环境，前景广阔；然而，如何配置大型无源阵列以实现可靠的信号均衡，实用方法仍然有限。在宽带链路中，均衡对于抵消多径引起的脉冲失真至关重要，否则这种失真会降低符号恢复性能。本文研究可重构智能表面辅助的脉冲响应均衡和信号增强，采用的方法包括经典自适应滤波和无模型深度强化学习。我们提出一种最速下降法，利用基站-RIS-用户设备信道信息配置RIS系数，以抑制多径并提升信噪比；并指出，最速下降法与深度强化学习之间的权衡，主要源于使用无源RIS硬件实现精确均衡所需的大量信道估计。传统自适应滤波会在接收信号后更新延迟滤波器系数，而我们的方法利用部署在级联信道中的RIS，在用户设备接收信号之前完成均衡，无需延迟单元。在这一框架中，信道在均衡之前进行估计，这构成了我们所称的自适应RIS均衡（ARISE）的基础。为克服ARISE对信道估计的依赖，我们研究了多种深度强化学习算法——深度确定性策略梯度、双延迟深度确定性策略梯度和SAC——这些算法无需显式信道估计，直接根据接收脉冲响应优化RIS系数。通过在多种信道条件和不同可重构智能表面规模下进行大量仿真，我们表明，SAC能够快速、稳定地收敛，均衡性能与ARISE相当，同时实现复杂度显著更低。这些结果展现了深度强化学习作为未来无线系统中实时RIS控制的实用且可扩展方案的潜力。

索引词

可重构智能表面（RIS）、最速下降法（SD）、深度强化学习（DRL）、无线信道均衡、深度确定性策略梯度（DDPG）、双延迟DDPG（TD3）、软演员评论家（SAC）。

作者隶属于加利福尼亚大学欧文分校电气工程与计算机科学系 Nhu 系。此项工作部分由美国国家科学基金会拨款 2030029 资助。

I. 引言

可重构智能表面（RIS）已成为下一代无线网络的一项前景广阔的技术，可通过近乎无源、低功耗的硬件重塑传播环境 [1], [2]。通过调节入射信号的相位，可重构智能表面能够增强覆盖范围、缓解干扰，并提升基站（BS）与用户设备（UE）之间的链路可靠性，而无需采用有源射频（RF）链。然而，要在实践中实现这些优势，就需要在现实信道条件下高效配置大规模可重构智能表面阵列。一个核心挑战是，在不依赖完整信道状态信息（CSI）的情况下实现可靠的信号均衡与放大；由于基站—可重构智能表面—用户设备级联信道结构以及可重构智能表面辅助链路的高维特性，完整CSI难以获取 [3]。在动态或散射丰富的环境中，这些局限性更加突出，因为必须快速适应才能维持链路质量 [4]。这些挑战促使人们寻求无需依赖完整CSI获取或反复进行导频传输即可优化可重构智能表面配置的方法，尤其是在动态环境中。

大量研究已采用经典信号处理和基于模型的方法对可重构智能表面进行优化 [3], [5]–[11]。许多方法假设能够获取准确的信道状态信息，并着重于最大化接收功率、提高信噪比（SNR）、最大化和速率，或实现波束成形与聚焦 [12]–[17]。尽管这些方法在受控场景中有效，但随着可重构智能表面规模增大，它们会面临可扩展性问题，且信道估计开销可能变得难以承受。为解决这一问题，研究人员提出了低复杂度估计方案、压缩感知技术和解析优化框架 [18]。与此同时，基于学习的方法——尤其是深度强化学习（DRL）——作为一种无模型替代方案，日益受到关注 [19], [19]–[21]。深度强化学习能够根据观测到的性能度量指标直接配置可重构智能表面，无需显式信道状态信息，并能在解析模型不可靠的环境中保持鲁棒性。深度确定性策略梯度（DDPG） [19], 、双延迟深度确定性策略梯度（TD3） [22], 和软演员-评论家（SAC） [20] 等算法已应用于可重构智能表面的波束成形、资源分配和自适应控制，展现出很强的实时部署潜力，这得益于其可扩展性以及对连续动作空间的支持。此外，无模型特性使这些算法即便在可重构智能表面硬件存在非线性、耦合效应或频率相关行为（这些特性难以通过解析方式描述）时，也能有效运行。

尽管已有这些进展，基于可重构智能表面的信号均衡仍相对缺乏研究。在高速链路中，均衡对于缓解多径失真引起的码间干扰（ISI）仍然至关重要，这使得可重构智能表面辅助时域整形尤为相关。现有大多数可重构智能表面文献侧重于最大化接收功率或信噪比，而对接收信号的时域整形——这对缓解多径失真并确保可靠的符号恢复至关重要——所受到的关注仍然有限。

一些近期工作已开始填补这一空白。例如， [23] 表明可重构智能表面可被配置为在接收前抑制或对齐多径分量，从而有效执行空域均衡。同样， [24] 表明可重构智能表面能够以极少的接收端处理补偿信道失真，从而在受限环境中实现“空口均衡”。然而，这些研究仍处于早期阶段，而无源RIS硬件的实际约束带来了额外挑战：有效均衡可能需要大规模或精细调谐的系数集合，并且信号增益与均衡精度之间存在固有权衡。在本工作中，我们将RIS定位为“空口均衡器”，将其置于中间物理层节点——即基站-RIS-用户设备无线链路的中点——以便在信号到达接收机之前对整个信号进行均衡。RIS通过直接塑造电磁传播环境来补偿复合信道。这有别于传统均衡方法，后者在接收后利用数字基带处理方法或射频前端来校正失真。更多细节见第III节，我们在其中推导了一种非传统的最速下降（SD）算法 [25] ，该算法基于估计的级联基站-RIS-用户设备信道优化RIS反射系数，以均衡信号。尽管其本身已被证明富有成效，但这种SD实现依赖于准确的级联信道状态信息，这会在RIS优化开始前引入大量通信开销和时延。该算法实现所涉及的困难，以及上一段讨论的硬件相关挑战，使得一个问题仍待解答：在应用于RIS辅助均衡时，经典自适应滤波技术与现代深度强化学习方法相比表现如何。

为弥补这些不足，本文全面研究了采用经典方法和基于学习的方法进行RIS辅助信号均衡与放大。首先，我们提出一种基于SD的方法，利用估计得到的级联基站-RIS-用户设备信道来配置RIS系数，以实现脉冲响应均衡并提升信噪比。尽管我们提出的算法基于SD，但其实现方式不同于传统SD。因此，我们将其命名为自适应RIS均衡（ARISE）。我们分析了将ARISE应用于无源RIS时固有的权衡，包括为实现精确均衡而需要进行大量信道估计，以及由此带来的复杂度影响。随后，我们研究了无模型深度强化学习算法——DDPG、TD3和SAC；这些算法无需信道估计开销，直接根据接收脉冲响应优化RIS系数。通过针对多种信道条件和RIS规模开展大量仿真，我们证明SAC始终能够快速、稳定地收敛，性能与ARISE相当，同时实现复杂度显著更低。具体而言，SAC可从原始脉冲响应观测出发，立即开始优化RIS，从而避免ARISE固有的信道状态信息获取时延，并能在时变信道中实现更快的自适应。

本文的主要贡献如下：

- 基于可重构智能表面的均衡与放大框架：我们对可重构智能表面辅助的脉冲响应均衡和信号增强问题进行建模，并分析其在现实无源可重构智能表面约束下所面临的挑战。

- 基于最速下降算法的可重构智能表面优化：我们推导出一种最速下降算法，利用级联信道估计配置可重构智能表面系数，并分析其性能、复杂度和权衡。由于该算法与传统最速下降算法有所不同，我们将其命名为 ARISE。

- 用于可重构智能表面控制的无模型深度强化学习方法：我们开发并评估深度确定性策略梯度、双延迟深度确定性策略梯度和 SAC 算法，这些算法直接依据接收信号观测优化可重构智能表面系数，无需进行信道估计。

- 全面的比较分析：我们比较 ARISE 和深度强化学习方法在收敛行为、计算复杂度、通信复杂度、均衡性能以及不同信道类型和可重构智能表面规模下的鲁棒性。

- 将 SAC 确定为实用方案：我们表明，SAC 能够快速、稳定地收敛，并取得与 ARISE 相当的性能，因此有望成为未来无线系统中实时可重构智能表面控制的理想方案。

本文组织结构如下：第二节介绍系统和信道模型，并定义优化问题。第三节推导 ARISE 算法，并给出仿真结果，与基线情形进行比较。第四节介绍三种不同的深度强化学习算法，并讨论其复杂度。第五节展示 ARISE 算法与深度强化学习算法在多种环境模型下的对比仿真结果。

II. 系统模型与问题构建

A. 系统模型

可重构智能表面辅助的无线通信网络由一个单天线基站、一个包含 M 个单元的超表面可重构智能表面以及一个单天线用户设备组成。可重构智能表面单元为金属贴片，按周期性线性阵列结构排列，沿 x 方向以间距 d_x 分布；这些单元由变容二极管进行电控，并通过连续电压偏置调节反射相位。每个可重构智能表面单元 m 的反射系数表示为 $\Gamma_m = |\Gamma_m| e^{j\phi_m}$，其相移为 ϕ_m。为保证可重构智能表面的无源性，为简化起见，我们假设对于所有 $m = 1$、2、.....、M，均有 $|\Gamma_m| = 1$。

在图 1 所示的下行链路场景中，用户设备接收到两条主要信号射线：第一条直接来自基站，第二条是基站发出、经可重构智能表面反射后的信号。假设所有信道均为远场和频率选择性的。由于障碍物的存在，我们假设基站与用户设备之间不存在占主导地位的视距路径，因此信道冲激响应主要由瑞利系数表征

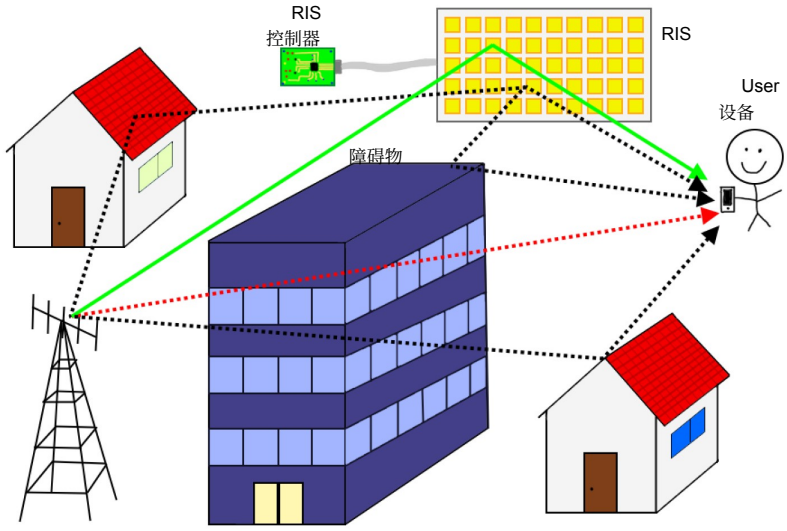


图1：下行链路可重构智能表面场景示意图。视距路径以实线标示，非视距路径以虚线标示。障碍物的反射会产生延迟路径和码间干扰分量。

基站与用户设备之间由于障碍物的存在不存在占主导地位的视距路径，因此信道冲激响应主要由瑞利系数表征

$$h_{\mathrm{BU}}(t) = (\beta_{\mathrm{BU}})^{1/2} h_{\mathrm{NLoS}}(t), \tag{1}$$

其中，β_{BU} 是与基站和用户设备之间距离相关的路径损耗，$\beta_{\mathrm{BU}} = Gd^{-\alpha}_{\mathrm{BU}}$。这里，$G$表示发射机增益、接收机增益和归一化路径损耗，$d_{\mathrm{BU}}$ 是基站与用户设备之间的距离，α 是路径损耗指数 [23], [26]。非视距分量服从标准圆对称复高斯随机变量 $h_{\mathrm{NLoS}}(t) \in \mathcal{CN}(0, \sigma^2(t))$，其方差 $\sigma^2(t)$取决于信号到达接收机的时间，这是由于多径效应产生的延迟信道分量所致，其中 $\sigma^2(0) = 1$ [27]。

表征基站与用户设备之间经由RIS单元的反射链路的信道，由冲激响应向量建模

$$\boldsymbol{h}_{\mathrm{BRU}}(t) = (\beta_{\mathrm{BRU}})^{1/2} \boldsymbol{R}^{1/2} \boldsymbol{\Gamma} \boldsymbol{h}(\theta_{\mathrm{BRU}}, t), \tag{2}$$

其中，$\beta_{\mathrm{BRU}} = G'(d_{\mathrm{BR}} d_{\mathrm{RU}})^{-\alpha}$ 表示与基站到可重构智能表面中心以及可重构智能表面中心到用户设备之间距离有关的路径损耗，分别记为 d_{BR} 和 d_{RU} [28], [29]。我们将可重构智能表面对视距信道的导向矢量定义为方位角 θ的函数：$\boldsymbol{h}_{\mathrm{LoS}}(\theta) = \left[1, e^{j\omega(\theta)}, \ldots, e^{j(M-1)\omega(\theta)}\right]^T$，其空间频率为 $\omega(\theta) = 2\pi d_x (f_c/c) \sin(\theta)$，其中 f_c为载波频率，c 为光速 [3]。结合上述内容，我们将衰落参数表示为

信道的莱斯系数

$$\boldsymbol{h}(\theta,t)=\sqrt{\frac{\kappa}{\kappa+1}}\boldsymbol{h}_{\text{LoS}}(\theta)+\sqrt{\frac{1}{\kappa+1}}\boldsymbol{h}_{\text{NLoS}}(t), \tag{3}$$

其中， κ, $0\leq\kappa<\infty$为莱斯因子。基站以垂直入射的方式照射可重构智能表面，基站、可重构智能表面与用户设备之间的方位角为 θ_{BRU}。我们假设视距信道分量仅在 $t=0$ 处的第一个信号采样时刻有效，此时接收到的信号能量最大。此后，延迟和失真的信号分量采用瑞利衰落建模，使用 $\boldsymbol{h}_{\text{NLoS}}(t)$。此外，信道由半正定空间相关矩阵 $\boldsymbol{R}$ 相关，其元素为

$$[\boldsymbol{R}]_{n,m}=\text{sinc}\left(\frac{2||u_n-u_m||}{c/f_c}\right)\quad n,m=1,\ldots,M, \tag{4}$$

其中， u_n 和 u_m 分别是RIS单元 n 和 m沿 x 的位置，而sinc(x) =sin(πx)/(πx)是sinc函数 [30], [31]。

对于一个发送信号 $s(t)$，用户设备处接收的基带信号表示为 s

$$y(t)=\left(h_{\text{BU}}(t)+\sum_{m=1}^{M}\Gamma_m h_{\text{BRU},m}(t)\right)*s(t)+n(t), \tag{5}$$

其中， $n(t)$ 为加性高斯白噪声（AWGN）， $*$ 为卷积算子 or.

B. 问题表述

我们的目标是在最大化接收信号功率的同时消除接收信号中的码间干扰。这可以通过将无源RIS同时用作空间均衡器和波束成形器来实现。与前馈均衡器（FFE）或判决反馈均衡器（DFE）等传统均衡器不同，RIS没有时延分量，但其相移可与环境中的多径传播所产生的时延配合使用，从而使产生码间干扰的信道分量正交化。

在数学上，我们将接收脉冲响应 $y(t)$ 的码间干扰 I 定义为以采样周期 T 的整数倍时延到达的信号内容。假设信号峰值在时刻 $t=0$ 到达接收机，则码间干扰表示为

$$I=\sum_{k\neq 0}y(t-kT)|_{t=0}, \tag{6}$$

其中，传输的离散时间基带信号是给定的脉冲:

$$s_k=\begin{cases}1 & k=0\\ 0 & k\neq 0.\end{cases} \tag{7}$$

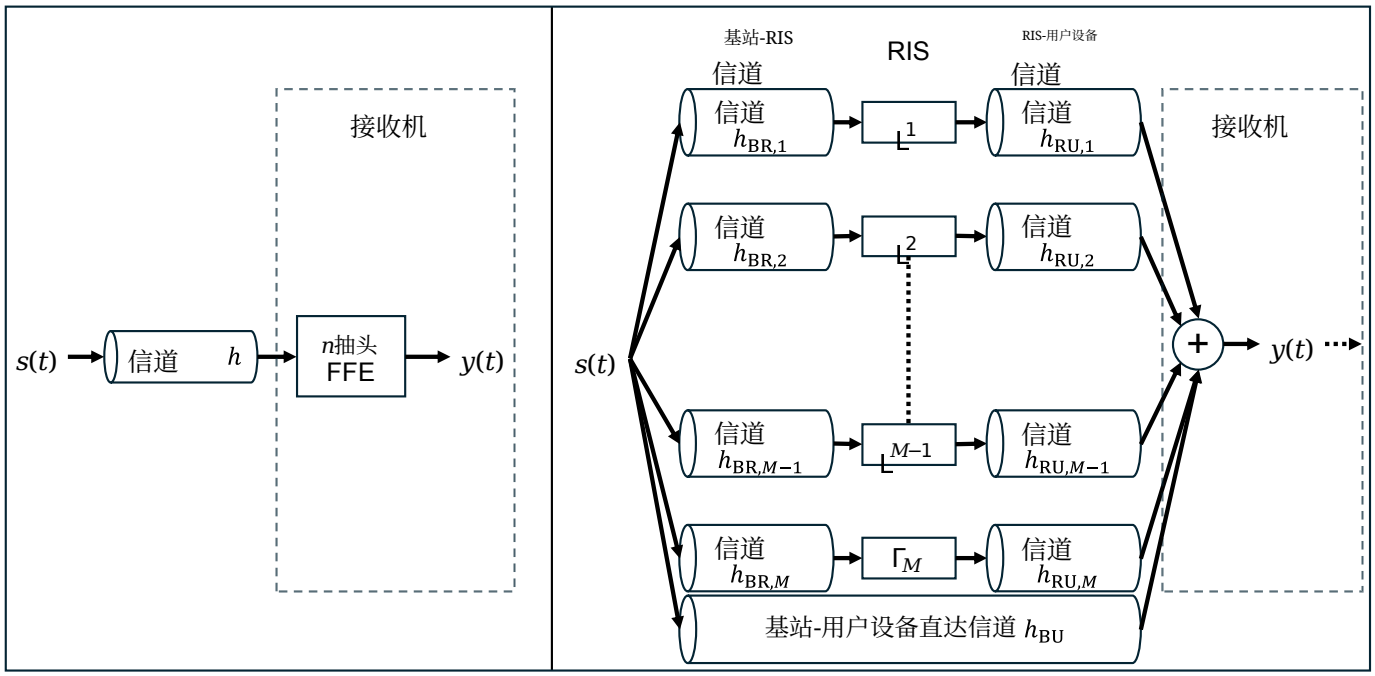


图 2：左：接收机处的传统前馈均衡器 (FFE)；右：我们提出的基于RIS的均衡器。传统均衡器在接收后使用 n 个延迟抽头处理基带信号，而基于RIS的均衡器位于级联的基站-RIS-用户设备链路中点，通过使用其每个 M 反射系数及其对应信道，在接收前对信号进行均衡。

该优化问题定义为

$$(P1): \qquad \max_{\boldsymbol{\theta}} \quad \eta, \tag{8}$$

其中

$$\eta = \operatorname{sgn}\big(\Re\{y(0)\}\big) \cdot \big[\Re\{y(0)\}\big]^2 - \sum_{k\neq 0} |y(t-kT)|^2 . \tag{9}$$

度量指标 η 旨在明确奖励接收脉冲响应中强且相位正确的主抽头，同时惩罚残余符号间干扰能量。第一项 $\operatorname{sgn}\big(\Re\{y(0)\}\big) \cdot \big[\Re\{y(0)\}\big]^2$ 鼓励在 $t=0$ 处形成较大的实值峰值（与实值发射脉冲一致），并通过符号因子确保相位对齐。否则，接收符号可能会在实轴和虚轴之间旋转，这一点将在图 7 和图 11 的部分曲线中看到。第二项 $-\sum_{k\neq 0} |y(t-kT)|^2$ 抑制所有符号间干扰分量的功率。

III. 自适应RIS均衡

A. ARISE算法

在通信系统中设计均衡器时，通常会采用一种常见方法来降低信号的码间干扰（ISI）：使用最速下降算法自适应地调整均衡器系数，以产生

期望信号 [32]。具体做法是发送导频序列（例如伪随机二进制序列（PRBS））, 或发送单比特脉冲, 以捕获给定数据速率 [33] 下的信道脉冲响应。对于基于RIS的均衡器, 均衡器系数就是RIS反射相位, 如图 2 所示。请注意图 2 中的一个重要区别：左侧的传统最速下降算法通过 n 抽头前馈均衡器（FFE）中的存储单元实现信道均衡；而我们在图 2 右侧提出的结构没有任何存储单元, 而是使用Γ_1、Γ_2、.....、Γ_M进行空域均衡。在传统均衡器中, 最速下降算法只需处理接收机处的输入信号 $x(t) = s(t) * h(t)$, 其中 $h(t)$ 表示从发射机到接收机的信道。假设已知 $s(t)$, 便可直接调整滤波器系数, 以最小化期望采样时刻 $t = kT$ 处误差信号 $e(t) = s(t) - y(t)$ 的平方幅度, 其中 $y(t) = x(t) * c(t)$, 而 $c(t)$ 是均衡器的冲激响应。

对于基于RIS的均衡器, RIS反射系数 $\mathbf{\Gamma}$ 就是均衡器系数；将其与级联信道 $\boldsymbol{h}_{\text{BRU}}$相乘后, 会得到发射信号 $s(t)$经过相位和幅度偏移的版本, 每个版本表示为 $x_m(t) = \Gamma_m h_{\text{BRU},\ m}(t) * s(t)$, 其中 $m = 1, 2, \ldots, M$。接收机处得到的信号由 RIS 辅助传播和从基站到用户设备的直达信道共同构成, 可写为 $y(t) = s(t) * h_{\text{BU}}(t) + \sum_{m=1}^{M} x_m(t)$, 与式（5）类似, 但省略了噪声项。通过正确调整系数并在接收机处合并所有这些信号, 我们可以消除（或至少尽量减小）由直达基站-用户设备链路以及基站-RIS-用户设备链路中的多径散射所导致的信号失真和码间干扰。由于有 M 个这样的信号, 对应于 M 个RIS单元, 因此我们需要估计 M 级联基站-RIS-用户设备信道, 以便正确计算梯度并最小化最终的误差信号, 正如本节推导并讨论的最速下降更新规则所示。这种信道估计会增加计算复杂度和通信复杂度, 因为这个名为ARISE的最速下降算法只有在获得 $\boldsymbol{h}_{\text{BRU}}$ 信道矩阵的全部信道状态信息后, 才能开始调整RIS。第四节E将进一步阐述这一问题。

我们假设接收机以周期 T 在整数时间点 k 对 $y(t)$ 进行采样。下文中, 对于所有 k, 我们使用离散时间信号表示法 y_k 作为 $y(kT)$ 的等价表示, 并使用数组表示法 $\boldsymbol{y}$, 即$[y_0, y_1, \ldots, y_L]$, 其中 L 是接收脉冲中码间干扰分量的数量。对于给定的接收脉冲响应 $\boldsymbol{y}$, ARISE算法需要将其与由 $1 \times (L+1)$ 数组 $\boldsymbol{s} = [1, 0, \ldots, 0]$ 给出的发射信号 $\boldsymbol{s}$ 相匹配, 以求解优化问题 (P1)。我们将代价函数 J 定义为误差幅度平方的平均值

$$J = E[\epsilon_k \epsilon_k^*], \tag{10}$$

其中，每个离散时间点 k 的误差信号 ϵ_k 定义为

$$\epsilon_k = s_k - y_k. \tag{11}$$

通过观察优化问题 (P1)，我们注意到，RIS既充当均衡器，也充当信号增强器，因为它能够进行波束成形、提高信号的信噪比，同时减少码间干扰。如果 s_0 的幅度远大于 $\boldsymbol{y}$的幅度，算法就会发现大部分误差贡献都来自主抽头，因此主要关注功率最大化，而几乎不关注符号间干扰最小化。相反，如果 s_0 过小，算法就会在保持主抽头较小的同时尝试最小化码间干扰，因为 ϵ_k、 $k>0$中的码间干扰误差幅度相对于 ϵ_0中的主抽头误差会变得更大。请注意，由于信道的影响，采样时刻 $t=0$ 处的接收脉冲会发生相移，并包含虚部。此外，其实部可能变为负值。基于上述原因，我们选择根据 $\boldsymbol{y}$ 在$k=0$ 处的幅度对其进行归一化，以匹配期望脉冲 $\boldsymbol{s}$。这样既考虑了信号通过信道时信号功率的损失，又确保符号间干扰分量相对于主抽头仍然较为显著。此外，为了利用RIS作为波束成形器来最大化信号功率，我们按RIS单元的数量对信号进行缩放，以计入使用无源RIS所能达到的最大增益。我们定义缩放变量 α_{s} 和缩放因子

$$y_s = \alpha_{\mathrm{s}} M |y_0|, \tag{12}$$

使用初始接收脉冲的主抽头 y_0 对 $\boldsymbol{y}$ 进行缩放，并在 ARISE 的迭代过程中保持该缩放。缩放方式可根据侧重消除码间干扰或放大主抽头进行调整，如后文所示，并通过调节 α_{s}加以控制。

为最小化代价函数 J，ARISE算法将直接修改 RIS反射系数Γ_m 以获得稳定梯度，而不是修改会以非线性方式影响接收信号的反射相位 θ_m 。因此，ARISE 的目标函数变为

$$\min_{\boldsymbol{\Gamma}} E[\epsilon_k \epsilon_k^*], \tag{13}$$

反射系数 $\boldsymbol{\Gamma}$ 的梯度下降更新规则为

$$\boldsymbol{\Gamma}_{t+1} = \boldsymbol{\Gamma}_t + \mu E\left[\sum_{l=0}^{L} \boldsymbol{h}_{\mathrm{BRU},l}^* s_{k-l}^* \epsilon_k\right], \tag{14}$$

按照附录A中的推导，我们对 $L+1$ 信号抽头取期望——这本质上相当于在脉冲响应上进行滑动窗口操作——以计算集合平均，从而得到“真实

梯度”，该梯度适用于每个时间步 [32]。在实践中，可以通过以下任一方式实现：

- 每一步都更新RIS配置，并从接收机获取新的脉冲响应。

- 假设RIS已更新，使用式（5）离线计算预期的脉冲响应。

请注意上述两种方法之间的权衡：第一种方法在优化RIS时需要与接收机实时通信，而第二种方法则需要对直接的 $\boldsymbol{h}_{\mathrm{BU}}$向量进行信道估计，才能准确计算预期的脉冲响应。

为确保RIS的无源性，我们在每次更新后使用以下公式对每个反射系数的幅度进行归一化

$$\Gamma_{t+1,m} \leftarrow \frac{\Gamma_{t+1,m}}{|\Gamma_{t+1,m}|} \tag{15}$$

对于 $m = 1$、2、.....、 M。

完整的ARISE算法如算法1所示。梯度下降的步长定义为

$$\mu = \frac{\alpha_\mu (L+1)}{\sum_k |y_k|}, \tag{16}$$

并采用任意缩放因子 α_μ 以及取自初始接收脉冲响应的信号分量 $|y_k|$ 。收敛的停止条件取决于阈值 η_{th} ，该阈值基于 η的归一化版本之间的绝对差，

$$\eta_{\mathrm{n}} = \frac{\eta}{\sum_k |y_k|^2}, \tag{17}$$

连续十步。作为额外的安全措施，我们定义$\eta_{\mathrm{n,max}}$ 以跟踪最优目标函数，并在算法过冲并发散时减小步长，同时使用另一个任意参数 η_{d} 来跟踪当前目标与最佳目标之间的差值。

尽管我们如(14)中的更新规则那样，通过应用期望算子采用最速下降法，但我们注意到，基于随机梯度下降的ARISE实现同样可行，类似于最小均方（LMS）算法 [34]。更新规则变为

$$\boldsymbol{\Gamma}_{t+1} = \boldsymbol{\Gamma}_t + \mu \sum_{l=0}^{L} \boldsymbol{h}^*_{\mathrm{BRU},l} s^*_{k-l} \epsilon_k. \tag{18}$$

然后，我们不再通过滑动窗口处理脉冲响应，并在每次更新可重构智能表面（RIS）之前对误差取平均，而是仅在每个时间步内通过一个窗口处理脉冲（或任何其他所需的导频信号），并使用由此得到的随机梯度。这种方法每次更新 RIS 的计算复杂度较低，但可能需要更多步才能收敛，并产生噪声

梯度。在本文讨论的仿真结果中，我们采用式（14）中带有期望算子的更新规则，以充分利用接收脉冲响应中的每个样本，从而使用平滑梯度对信号进行均衡。

算法 1 自适应RIS均衡（ARISE）

初始化 $\boldsymbol{\Gamma}, \boldsymbol{h}_{\mathrm{BRU}}, \boldsymbol{s}, \eta_{\mathrm{n,max}}$。
通过发送 $\boldsymbol{s}$ 并记录式（5）中的接收信号，初始化 $\boldsymbol{y}$。
初始化 $\mu \leftarrow \frac{\alpha(t+1)}{\sum_k |y_k|}$。
初始化 $y_s \leftarrow \alpha_{\mathrm{s}} M |y_0|$。
初始化 $i \leftarrow 0$。
重复
    $\boldsymbol{\epsilon} \leftarrow \boldsymbol{s} - \boldsymbol{y}/y_s$。
    使用 (9) 计算 η，并使用 (17) 计算 η_{n}。
    如果 $\eta_{\mathrm{n}} > \eta_{\mathrm{n,\ max}}$ 则
        $\eta_{\mathrm{n,max}} \leftarrow \eta_{\mathrm{n}}$。
    结束 if
    使用 (14) 更新 $\boldsymbol{\Gamma}$。
    使用 (15) 归一化 $\boldsymbol{\Gamma}$。
    根据 (5) 获取新的接收脉冲 $\boldsymbol{y}'$。
    使用 (9) 计算新的目标函数 η'，并使用 (17) 计算 η'_{n}。
    如果 $\eta'_{\mathrm{n}} < \eta_{\mathrm{n,max}} - \eta_{\mathrm{d}}$ 则
        $\mu \leftarrow \frac{\mu}{2}$.
        $\eta_{\mathrm{n,max}} \leftarrow -1.0$.
    结束如果
    如果 $|\eta'_{\mathrm{n}} - \eta_{\mathrm{n}}| < \eta_{\mathrm{th}}$ 则
        $i \leftarrow i + 1$.
    else
        $i \leftarrow 0$。
    结束条件判断
    如果 $i \geq 10$，则
        跳出
    结束条件判断
直到收敛

B. ARISE 仿真结果

RIS环境由单天线基站、单天线用户设备以及默认的 $M = 100$单元可重构智能表面组成。RIS单元之间的间距为 $d_x = 20$ mm。我们采用采样周期 $T = 16.3$ns，符合3GPP针对5G新空口（新空口） [35]的标准。归一化增益因子为$G = G' = -43$ dBm，如 [23], 中所述，加性高斯白噪声信号 $n(t)$ 中的噪声功率为 $P_n = -96$ dBm。我们定义一个二维坐标系，基站位于中心，坐标（单位：米）为(0, 0)。基站以垂直入射方式直接照射位于(10,0)的可重构智能表面。用户设备的起始位置为$(-20, -20)$，并允许在由点$(-100, -100)$和$(-10,100)$界定的矩形区域内移动。在每个信道相干块开始时，非视距信道

被完全更新，并且用户设备以随机游走方式移动，每次沿随机方向、以距前一点最多20米的随机距离移动，二者均服从均匀分布。考虑到基站与可重构智能表面之间的码间干扰路径也可能在RIS-用户设备链路中发生散射和延迟，我们将延迟路径数量 n_{r} 与码间干扰项数量 L 之间的关系定义为 $L = 2n_{\mathrm{r}}$。

我们定义两种基线RIS反射系数方案以供比较 :

- 随机相位——RIS反射相位在 $[-\pi, \pi]$之间均匀采样。
- 逆相位——RIS反射系数是 $\boldsymbol{h}_{\mathrm{BRU}\,k}$中 M 信道系数的复共轭，在采样实例 $k = 0$ 处求值并归一化为单位幅度，例如，

$$\boldsymbol{\Gamma} = \exp(-j\angle \boldsymbol{h}_{\mathrm{BRU},0}), \tag{19}$$

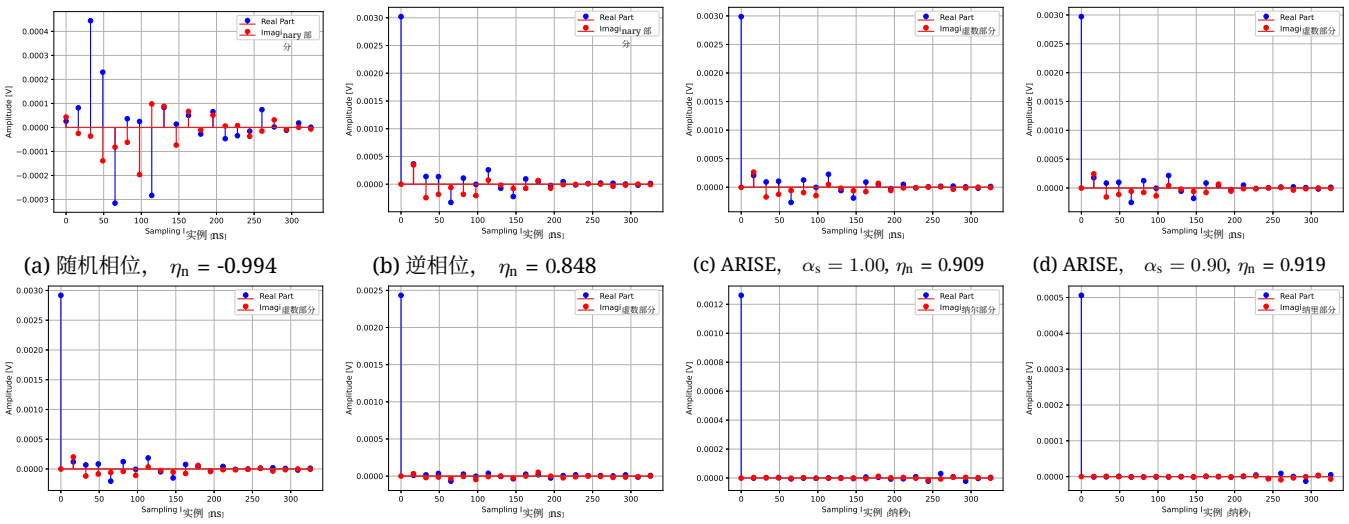
旨在针对 y_0 [11]在接收机处产生相干叠加。

对于ARISE算法，我们将阈值参数设为 $\eta_{\mathrm{th}} = 10^{-5}$ ，过冲参数设为$\eta_{\mathrm{d}} = 0.5$。我们将步长 μ 按 $\alpha_\mu = 10$ 缩放，以实现快速收敛。

第一次仿真将基线方案与采用不同缩放因子 α_{s}的ARISE收敛结果进行比较。在此情形下，基站-RIS-用户设备链路的莱斯因子为 $\kappa = 10$ ，而基站与用户设备之间的链路因路径中的障碍物而呈现瑞利衰落——用以建模预先存在的信号失真，并突出无线环境对可重构智能表面的依赖性。延迟路径数为 $n_{\mathrm{r}} = 10$，因此码间干扰项数为 $L = 20$。仿真采用随机游走方法，在1000个信道相干块上重复进行。收敛脉冲响应绘制于图 3，各算法对应的QPSK星座图在假设单位间隔（UI）为 $T = 16.3$ ns的条件下绘制于图 4，收敛的 η 与归一化的 η_{n} 的累积分布函数图见图 5。ARISE算法的优势首先体现在脉冲响应中。可以观察到，随着缩放因子 α_{s} 减小，均衡比放大得到更多强调，脉冲响应中的码间干扰项被最小化，直至几乎完全消除。这一点在图 7的QPSK图中得到进一步体现，显示出符号星座图中的信噪比提升和方差降低。图 5中的累积分布函数图进一步证实了这一行为，表明ARISE与逆相位方案取得近乎相等的 η 值，而ARISE具有严格更高的 η_{n} 值，展现出其更优的均衡能力。这些结果在更复杂的场景中得到进一步印证：所有信道完全为非视距（$\kappa = 0$），并且码间干扰项多得多， $n_{\mathrm{r}} = 50$ 见图 6、图 7和图 8。

ARISE的可行性在很大程度上取决于能否获得足够准确的信道状态信息（CSI），因为其在（14）中的更新规则需要明确知道级联信道矩阵 $\boldsymbol{h}_{\mathrm{BRU}}$，以抑制占主导地位的符号间干扰分量。随着可重构智能表面规模扩大，对这一高维信道矩阵进行估计会变得愈发繁重；原则上，要消除 L 码间干扰抽头，需要 $M(L+1)$

个与 M 可重构智能表面单元及（$L+1$）信号分量（主抽头和码间干扰）相关的系数。尽管实际估计器可能以少于$M(L+1)$种导频配置联合恢复主信道和延迟信道，但针对存在码间干扰的可重构智能表面辅助宽带信道设计此类估计器不在本文讨论范围内 [3], [36], [37]。重要的是，只有在完成信道估计后，ARISE才能开始优化可重构智能表面；由于需要针对多种可重构智能表面配置反复进行导频传输，这会带来额外的时延和通信开销。此外，ARISE隐含地假设可重构智能表面的响应是线性且频率平坦的，而实际的可重构智能表面硬件会呈现单元耦合、幅度损耗、相位误差以及频率相关反射系数等特性——这些因素会使准确建模变得复杂，尤其是在宽带场景下 [38]–[41]。在接下来的章节中，我们将采用基于深度强化学习的方法来解决上述问题；这些方法完全避免显式信道估计，并且本质上是无模型的。这些算法每个时间步只需传输一个导频符号，即可获得脉冲响应；智能体只要在其存储器中存入少量样本，就能立即开始优化可重构智能表面。



(a) 随机相位， η_{n} = -0.994　(b) 逆相位， η_{n} = 0.848　(c) ARISE， $\alpha_{\mathrm{s}}=1.00$, η_{n} = 0.909　(d) ARISE， $\alpha_{\mathrm{s}}=0.90$, η_{n} = 0.919

(e) ARISE， $\alpha_{\mathrm{s}}=0.75$， η_{n} = 0.943　(f) ARISE， $\alpha_{\mathrm{s}}=0.50$， η_{n} = 0.993　(g) ARISE， $\alpha_{\mathrm{s}}=0.25$， η_{n} = 0.997　(h) ARISE， $\alpha_{\mathrm{s}}=0.10$， η_{n} = 0.996

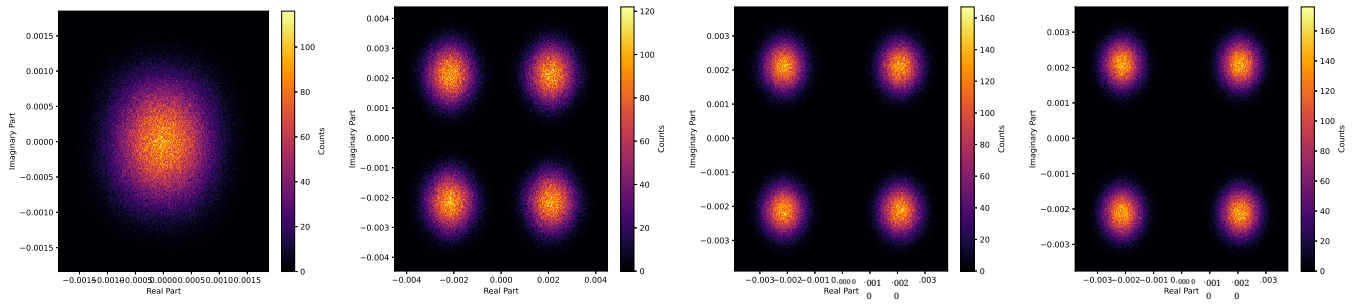
图3：用户设备位置（-20，-20）处收敛脉冲响应的实部和虚部， $M=100, \kappa=10, n_{\mathrm{r}}=10$。

IV. 深度强化学习

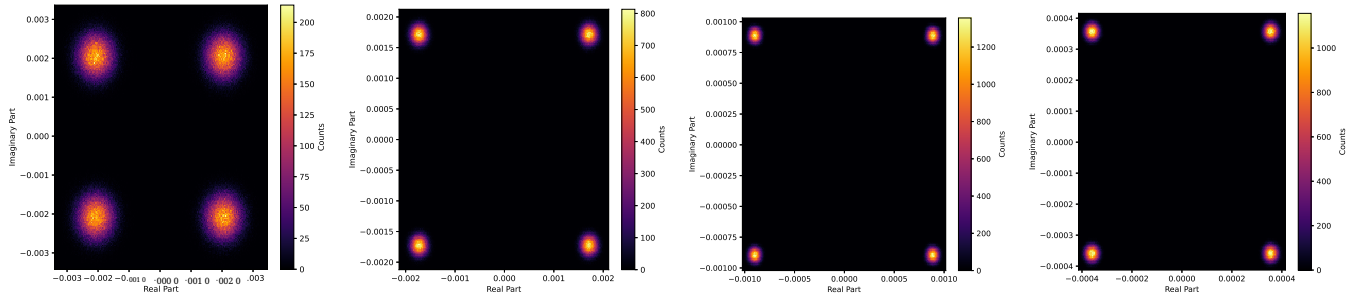
本节将回顾深度强化学习背后的一般概念，将其与可重构智能表面均衡器场景联系起来，并介绍三种不同的算法来求解优化问题（P1）。

A. 背景

从最广义的角度来看，强化学习（RL）涉及一个智能体在特定环境中采取动作，并因不同的动作而获得不同的奖励。其目标是学习一种策略，引导

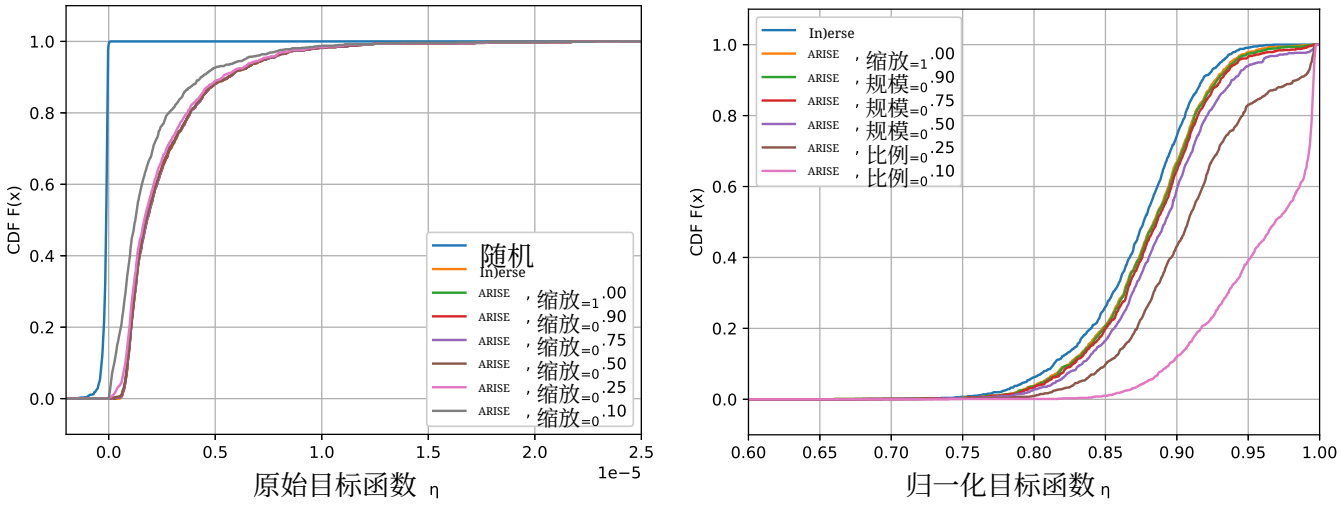


(a) 随机相位，信噪比 = 2.87 分贝　(b) 逆相位，信噪比 = 10.84 分贝　(c) ARISE，$\alpha_s = 1.00$，信噪比 = 13.22 分贝　(d) ARISE，$\alpha_s = 0.90$，信噪比 = 13.77 分贝



(e) ARISE，$\alpha_s = 0.75$，信噪比 = 15.32 分贝　(f) ARISE，$\alpha_s = 0.50$，信噪比 = 24.36 分贝　(g) ARISE，$\alpha_s = 0.25$，信噪比 = 27.57 分贝　(h) ARISE，$\alpha_s = 0.10$，信噪比 = 26.89 分贝

图 4：用户设备位置 (-20, -20) 处收敛的 QPSK星座图，$M = 100, \kappa = 10, n_r = 10$。

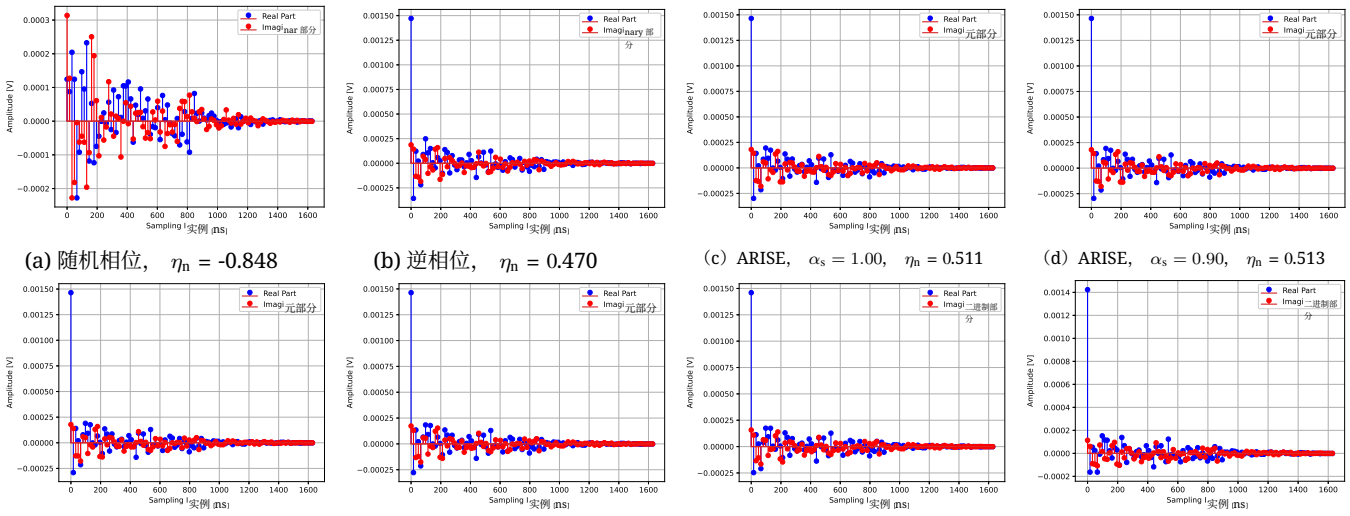


（a）收敛的 η 值　(b) 收敛的归一化 η_n 值

图 5：经过 1000 次仿真后收敛的累积分布函数，$M = 100, \kappa = 10, n_r = 10$。

智能体通过试错选择能够最大化长期奖励的动作。在深度强化学习（DRL）中，我们采用深度神经网络（DNN）对环境和智能体进行建模。应用于我们的可重构智能表面（RIS）场景时，强化学习中的关键要素如下所述 [19], [42]:

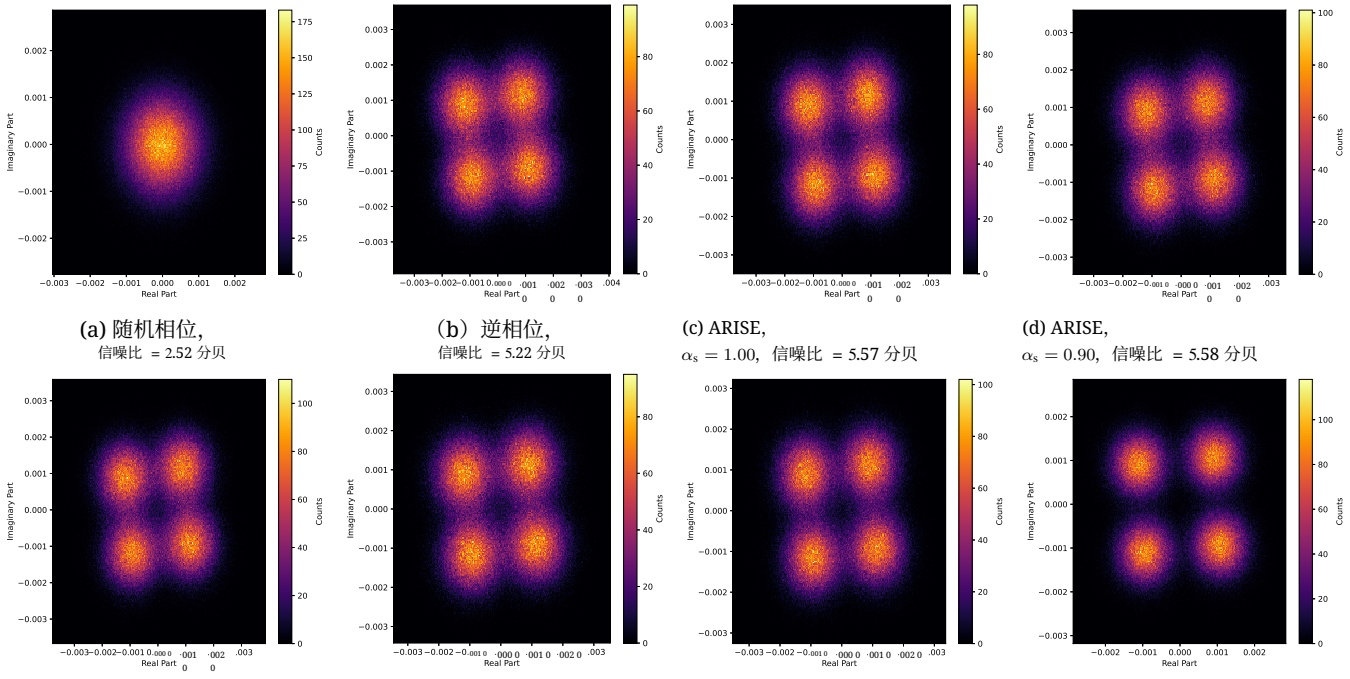
1) 状态：一组表征环境的参数。我们将时刻 t 的状态定义为时刻 t的接收脉冲响应 $\boldsymbol{y}$ 的函数 $s_t \in \mathcal{S}$，记作 $\boldsymbol{y}_t$，而 $\mathcal{S}$ 为连续状态空间。我们定义记号 $y_{t,\ k}$，用于标识接收脉冲响应$\boldsymbol{y}_t$中索引为 k 的样本。由于信道中的反射相移会使脉冲响应变为复数，我们将其展开为



(a) 随机相位， η_{n} = -0.848　(b) 逆相位， η_{n} = 0.470　(c) ARISE， $\alpha_{\mathrm{s}} = 1.00$， η_{n} = 0.511　(d) ARISE， $\alpha_{\mathrm{s}} = 0.90$， η_{n} = 0.513

(e) ARISE， $\alpha_{\mathrm{s}} = 0.75$， η_{n} = 0.518　(f) ARISE， $\alpha_{\mathrm{s}} = 0.50$， η_{n} = 0.532　(g) ARISE， $\alpha_{\mathrm{s}} = 0.25$, η_{n} = 0.568　(h) ARISE， $\alpha_{\mathrm{s}} = 0.10$, η_{n} = 0.680

图 6：收 敛后的脉冲响应在用户设备位置 (-20, -20) 处的实部和虚部， $M = 100, \kappa = 0, n_{\mathrm{r}} = 50$.



(a) 随机相位，
信噪比 = 2.52 分贝

(b) 逆相位，
信噪比 = 5.22 分贝

(c) ARISE，
$\alpha_{\mathrm{s}} = 1.00$， 信噪比 = 5.57 分贝

(d) ARISE，
$\alpha_{\mathrm{s}} = 0.90$， 信噪比 = 5.58 分贝

(e) ARISE，
$\alpha_{\mathrm{s}} = 0.75$， 信噪比 = 5.62 分贝

(f) ARISE，
$\alpha_{\mathrm{s}} = 0.50$， 信噪比 = 5.75 分贝

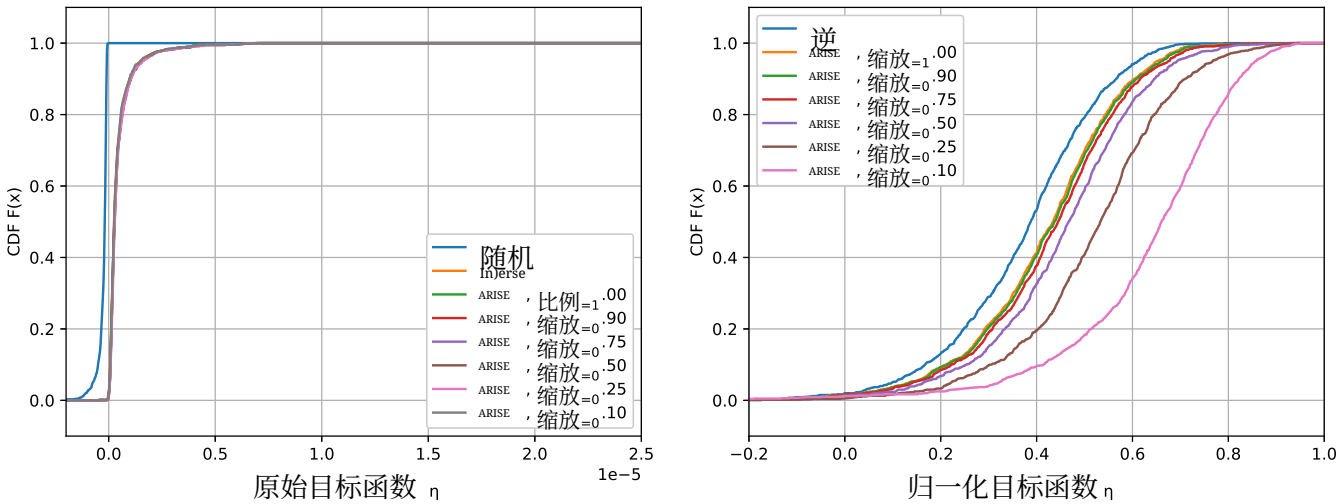
(g) ARISE，
$\alpha_{\mathrm{s}} = 0.25$， 信噪比 = 6.09 分贝

(h) ARISE，
$\alpha_{\mathrm{s}} = 0.10$， 信噪比 = 7.42 分贝

图 7：用户设备位置 (-20, -20) 处收敛的 QPSK星座图， $M = 100$、 $\kappa = 0$、 $n_{\mathrm{r}} = 50$。

长度为 $L + 1$ 的脉冲转换为长度为 $2(L + 1)$ 的数组，其中包含脉冲中每个采样点的实部和虚部，作为深度神经网络的输入。状态定义为所得数组 $s_t = \boldsymbol{y}'_t$，其中 $\boldsymbol{y}'_t = [\Re\{y_{t,0}\}, \Im\{y_{t,0}\}, \ldots, \Re\{y_{t,L}\}, \Im\{y_{t,L}\}]$。

2) 动作：智能体在每个时间步所做的选择 $a_t \in \mathcal{A}$，用于从环境的当前状态 s_t 转移到下一个状态 s_{t+1}，该选择从连续动作空间 $\mathcal{A}$中采样。我们将动作定义为时间 t 的 RIS反射系数 $\boldsymbol{\Gamma}$，其目标是改善脉冲的质量



（a）收敛的 η 值　　(b) 收敛的归一化 η_{n} 值

图8：经过1000次仿真后收敛的累积分布函数，$M=100, \kappa=0, n_{\mathrm{r}}=50$。

响应，即 $a_t=\mathbf{\Gamma}_t'$。这里，我们还将 M维反射系数向量 $\mathbf{\Gamma}_t$的实部和虚部拆分为长度为 $2M$ 的数组，即 $\mathbf{\Gamma}_t'=[\Re\{\Gamma_{t,1}\},\Im\{\Gamma_{t,1}\},\ldots,\Re\{\Gamma_{t,M}\},\Im\{\Gamma_{t,M}\}]^T$。3) 奖励：环境提供的瞬时性能指标，用于告知智能体其在状态 s_t 下采取的动作 a_t 的质量，以及智能体向下一个状态 s_{t+1}转移时的情况。这里，奖励 r_{t+1} 根据式 (17) 定义为 η 的归一化版本，即 η_{n}，表示为

$$r_{t+1}=\frac{\eta}{\sum_k|y_{t+1,k}|^2}. \tag{20}$$

这种归一化可确保奖励始终介于最差情况下的 -1 和最佳情况下的 $+1$ 之间，从而在训练深度神经网络时提供稳定梯度。请注意，根据(9)中 η 的定义，并且由于此处进行了归一化，当接收信号 $\boldsymbol{y}_t$ 受码间干扰主导时、y_t的实部$_{,0}$ 较小时，以及y_t的虚部$_{,0}$ 较大时，奖励函数会受到惩罚；这是通过计算 $\boldsymbol{y}_t$每个采样索引处的功率来实现的。虽然这种方法不会明确确定利用可重构智能表面进行波束成形所获得的增益大小，但通常较高的奖励值意味着信号的码间干扰较少、主抽头的实部相较于信号其余部分更强，或两者兼有。这也表明可重构智能表面配置中会存在不同的局部最优，类似于ARISE算法的性能取决于接收信号的缩放方式。这个问题涉及深度强化学习算法在探索与利用之间的权衡，后文将对此进行讨论。

4) 经验：智能体与环境先前的观察和交互记录，以（$s_t, a_t, r_{t+1}, s_{t+1}$）的形式存储在内存中。

5) 策略：根据当前状态确定选择特定动作的概率，记为

$\pi(a_t|s_t)$，并满足 $\sum_{a_t\in\mathcal{A}}\pi(a_t|s_t)=1$。我们的目标是确定一种策略，以快速找到一组稳定的RIS反射系数，从而减少码间干扰并提高接收无线信号的信噪比。

6) 状态-动作价值函数：在状态 s_t 中采取动作 a_t 的质量，基于潜在的未来奖励。

我们将 V_t 定义为未来累积折扣奖励

$$V_t=\sum_{\tau=0}^{\infty}\gamma^{\tau}r_{t+\tau+1}, \tag{21}$$

其中 τ 取非负整数值，$0<\gamma<1$ 为折扣率。状态-动作价值函数由Q函数给出，该函数满足贝尔曼方程

$$\begin{aligned}Q_{\pi}(s_t,a_t)&=E_{\pi}[V_t|s_t=s,a_t=a]\\&=E_{\pi}[r_{t+1}|s_t=s,a_t=a]\\&+\gamma\sum_{s'\in\mathcal{S}}P_{ss'}^{a}\left(\sum_{a'\in\mathcal{A}}\pi(a'|s')Q_{\pi}(s',a')\right)\end{aligned} \tag{22}$$

并可利用Q学习算法 [43]递归学习，以找到最优策略。在上文中，$P_{ss'}^{a}$ 是马尔可夫决策过程（MDP）中给定动作 a 时从状态 s 转移到状态 s' 的转移概率。目标Q函数先被初始化，然后通过学习以达到最优Q函数，其期望是对所有可能的下一状态 s_{t+1}取的，

$$Q^{*}(s_t,a_t)=E\left[r_{t+1}+\gamma\max_{a_{t+1}}Q^{*}(s_{t+1},a_{t+1})\right], \tag{23}$$

使用更新规则

$$\begin{aligned}Q^{*}(s_t,a_t)\leftarrow&(1-\alpha)Q^{*}(s_t,a_t)\\&+\alpha\left(r_{t+1}+\gamma\max_{a_{t+1}}Q_{\pi}(s_{t+1},a_{t+1})\right),\end{aligned} \tag{24}$$

其中 α 为学习率。根据定义，Q函数根据给定策略、当前状态以及动作选择来确定未来累积折扣奖励。最优Q函数——即为当前状态和动作产生最高未来奖励的那个——将遵循期望的最优策略 π^{*}。

鉴于本工作的背景，动作空间是高维的，并且会随可重构智能表面的规模扩大，因此需要使用机器学习，通过两个深度神经网络估计Q函数和最优策略。评论家深度神经网络根据智能体在环境中执行动作时的状态、动作和奖励来估计Q函数。使用符号 θ 表示评论家深度神经网络的参数（神经元权重、

激活函数、层数等），我们有 $Q(s_t, a_t) = Q_\theta(s_t, a_t)$。演员深度神经网络以当前状态作为输入，并在输出端估计一个动作，以寻求最优策略。根据算法的选择，这两个深度神经网络中的每一个还可能进一步分为训练深度神经网络和目标深度神经网络 [44]:

- 训练评论家网络 $Q_\theta(s_t, a_t)$ 通过梯度下降学习 Q 函数。
- 目标评论家网络 $Q_{\theta_{目标}}(s_t, a_t)$ 提供评论家网络的缓慢更新副本，用于生成稳定的时间差分（TD）目标。
- 训练演员网络 $\mu_\phi(s_t)$ 是策略网络，并使用确定性策略梯度进行优化。
- 目标行动者网络 $\mu_{\phi_{目标}}(s_t)$ 是演员网络的缓慢更新副本，用于在训练期间为损失函数计算生成稳定的下一状态动作。

请注意，现在可以通过使用随机优化算法和反向传播更新权重来训练深度神经网络，具体如下

$$\theta_{t+1} = \theta_t - \mu \nabla_\theta l(\theta), \tag{25}$$

其中，μ 为学习率，$\nabla_\theta l(\theta)$ 为损失函数 $l(\theta)$ [45] 的梯度。诸如 TensorFlow [46] 之类的现代机器学习框架会通过反向传播自动计算梯度，因此我们无需手动推导这些表达式。

深度强化学习的另一个关键组成部分是经验回放缓冲区，它用于存储智能体与环境之间的交互数据。与使用连续状态训练不同，从该缓冲区随机采样批次来训练深度神经网络，可以减弱数据中存在的强时间相关性。随机性增强有助于防止网络对近期经验过拟合并陷入局部最优，从而提升整体学习性能。这一机制也与探索和利用这一更广泛的概念相互作用：探索通过带噪声或随机化的动作来寻找新状态，而利用则借助先前收集的经验（例如存储在经验回放缓冲区中的经验）来优化策略。

本小节讨论的原则为接下来介绍的深度强化学习算法奠定了基础：深度确定性策略梯度、双延迟深度确定性策略梯度和 SAC；这些思想将在具体的算法框架中得到应用。

B. 深度确定性策略梯度

深度确定性策略梯度（DDPG）是一种离策略（从由不同行为策略 [47]生成的数据中学习目标策略）、面向连续控制任务的无模型演员-评论家算法。它将确定性策略梯度框架与深度强化学习技术相结合，使用独立的演员与评论家网络以及目标网络和经验回放缓冲区。评论家网络学习动作价值函数，而演员则通过反向传播评论家网络相对于动作的梯度进行更新，从而能够在高维连续动作空间中学习确定性策略

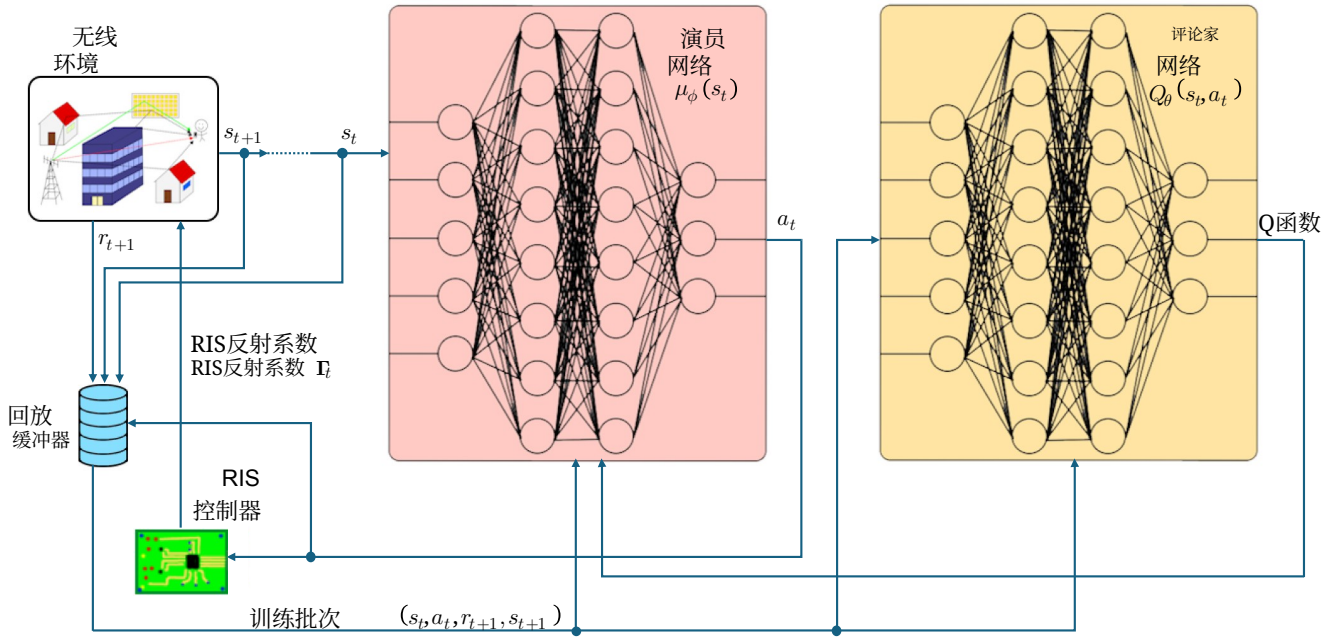


图 9：基于深度强化学习的可重构智能表面均衡器实现示意图。演员（策略）深度神经网络接收来自环境的状态，并输出对应于可改善信号完整性的可重构智能表面反射系数的动作。可重构智能表面控制器会更新可重构智能表面反射系数，环境也随之变化。环境状态（脉冲响应）、动作（可重构智能表面反射系数）和奖励（归一化的 η_n 度量指标）会存储在经验回放缓冲区中，并从中提取训练批次以优化深度神经网络超参数。演员也会利用评论家网络（Q网络）的输出进行训练。

[44], [47]。演员深度神经网络以状态作为输入，并根据其学习到的策略输出连续动作；评论家网络接收状态和动作，并根据（22） [19]输出估计的Q函数 $Q(s_t, a_t)$。参照（24），演员的作用是找到能使$\max_{a_{t+1}} Q_\pi(s_{t+1}, a_{t+1})$项取得最大值的最佳动作，从而无需穷举搜索和非凸优化。DDPG使用确定性策略 $\mu_\phi(s_t)$，其作用与随机策略 $\pi(a_t|s_t)$相同，但输出的是单一动作，而非动作上的分布。

顶层算法概述如下。我们采用上一节所述的RIS辅助环境，其中当前状态 s_t 定义为时刻 t接收到的采样脉冲响应， $\boldsymbol{y}_t'$；当前动作 $a_t = \mu_\phi$ （s_t）是将 $\boldsymbol{y}_t'$ 输入策略深度神经网络后得到的RIS反射系数更新量，输出为 $\mathbf{\Gamma}_t'$。我们将 $\boldsymbol{y}_t'$ 数组归一化至范围 $[-1, 1]$ ，以便在使用反向传播 [45]训练深度神经网络时获得稳定梯度。我们将环境在当前RIS配置下的状态定义为当前“步”，将无线信道保持相干的时间段定义为当前“回合”。在每个状态下，我们通过将随机噪声向量 $\boldsymbol{\epsilon}$ 加到RIS反射系数上来进行探索，将反射系数裁剪至 $[-1, 1]$，并将 $\mathbf{\Gamma}_t'$ 中分离出的实部和虚部合并为 $\mathbf{\Gamma}_t$ 后，依据式（15）赋予单位幅度。之后

更新RIS反射系数后，我们根据式（20）得到下一个（归一化）状态 $s_{t+1} = \boldsymbol{y}'_{t+1}$ 和下一个奖励 r_{t+1}，并将（s_t， a_t， r_{t+1}， s_{t+1}）存入经验回放缓冲区 $\mathcal{D}$。

为了更新深度神经网络，我们从经验回放缓冲区中随机采样一个大小为 B 的批次 $\mathcal{B}$，计算梯度并将其应用于所有深度神经网络，如算法2所示。用于以梯度下降训练评论家网络的损失函数由均方贝尔曼误差（MSBE）函数给出，用以确定 $Q_\theta(s,a)$与满足(23)中贝尔曼方程的接近程度 [48],

$$l(\theta) = E\Big[\big(Q_\theta(s,a) - (r' + \gamma Q_{\theta_{\text{targ}}}(s', \mu_{\phi_{\text{targ}}}(s')))\big)^2\Big], \tag{26}$$

其中，期望是在从经验回放缓冲区 $\mathcal{D}$中随机选取的(s, a, r', s')批次上取得的。演员使用梯度上升，以损失函数进行训练

$$l(\phi) = E\Big[Q_\theta\big(s, \mu_\phi(s)\big)\Big], \tag{27}$$

以学习使Q函数最大化的策略。目标深度神经网络使用Polyak平均[49] 以及加权因子 τ进行更新，

$$\theta_{\text{targ}} \leftarrow \tau\theta + (1-\tau)\theta_{\text{targ}}. \tag{28}$$

注意，在每个时间步 t内，可通过从经验回放缓冲区中采样随机批次，将演员和评论家深度神经网络训练 N_{train} 次。然而，只有当经验回放缓冲区包含足够多的样本以构成一个批次时，它们才会开始训练。

C. 双延迟深度确定性策略梯度

尽管深度确定性策略梯度能够取得优异性能，但其学习到的Q函数常常存在高估偏差，这会使训练不稳定，并提高对超参数的敏感性 [50]。双延迟深度确定性策略梯度（双延迟深度确定性策略梯度）是一种建立在深度确定性策略梯度基础之上、旨在解决上述问题的算法。它引入了截断双Q学习的概念，即学习两个“孪生”评论家网络（每个都有各自的训练网络和目标网络），Q_{θ_1} 和Q_{θ_2}。在梯度更新过程中，我们将（26）中的损失函数修改为取两个目标评论家网络Q函数估计器中的最小值

$$\begin{aligned} l(\theta_i) = E\Bigg[\Big(&Q_{\theta_i}(s,a) \\ &- (r' + \gamma \min_{i=1,2} Q_{\theta_{\text{targ},i}}(s', \mu_{\phi_{\text{targ}}}(s')))\Big)^2\Bigg], \end{aligned} \tag{29}$$

从而考虑了过估计偏差。

算法2：用于可重构智能表面均衡器的深度确定性策略梯度

1：初始化演员和评论家网络的训练深度神经网络与目标深度神经网络，赋值 $\phi_{\text{targ}} \leftarrow \phi$，$\theta_{\text{targ}} \leftarrow \theta$。

2：对于 N_{ep} 回合，执行

3：　对于 t，在 N_{steps} 中执行

4：　观测状态 $s_t \leftarrow \boldsymbol{y}'_t$ 并设置可重构智能表面系数

$$\boldsymbol{\Gamma}'_t \leftarrow \text{clip}(\mu_{\phi_{\text{targ}}}(\boldsymbol{y}'_t) + \boldsymbol{\epsilon}, -1, 1), \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_{\text{a}}^2)^{2M\times 1}.$$

5：　设置 $q_t \leftarrow \boldsymbol{\Gamma}'_t$.

6：　将 RIS反射系数 更新为

$$\boldsymbol{\Gamma}_t \leftarrow [\Gamma'_{t,1} + j\Gamma'_{t,2}, \ldots, \Gamma'_{t,2M-1} + j\Gamma'_{t,2M}]^T$$

并使用 (15) 进行归一化。

7：　观察新状态 $s_{t+1} \leftarrow \boldsymbol{y}'_{t+1}$。

8：　使用 (20) 计算奖励 r_{t+1}。

9：　将 $(s_t, a_t, r_{t+1}, s_{t+1})$ 存入经验回放缓冲区 $\mathcal{D}$。

10：　如果 $|\mathcal{D}| \geq B$，则

11：　对于 N_{train} 中的 i **do**

12：　从 $\mathcal{D}$ 中采样批次 $\mathcal{B} = \{(s, a, r', s')\}$。

13：　使用 (26) 计算 $l(\theta)$。

14：　更新评论家网络：

$$\theta \leftarrow \theta - \mu_{\text{c}} \nabla_\theta l(\theta)。$$

15：　使用 (27) 计算 $l(\phi)$。

16：　更新演员：

$$\phi \leftarrow \phi - \mu_{\text{a}} \nabla_\phi l(\phi).$$

17：　软更新目标评论家网络：

$$\theta_{\text{targ}} \leftarrow \tau\theta + (1-\tau)\theta_{\text{targ}}.$$

18：　软更新目标行动者网络：

$$\phi_{\text{targ}} \leftarrow \tau\phi + (1-\tau)\phi_{\text{targ}}.$$

19：　结束循环

20：　结束条件判断

21：　结束循环

22：结束循环

此外，双延迟深度确定性策略梯度通过每隔 d_{TD3} 步更新演员网络和目标网络，引入延迟策略更新，以减少价值函数估计中的误差并降低发散风险。这样一来，评论家网络便能在策略更新之间取得更大进展，从而减少评论家估计误差向策略的传播。

最后，双延迟深度确定性策略梯度在训练过程中使用目标策略平滑正则化。在计算用于训练评论家网络的目标值时，双延迟深度确定性策略梯度会向目标行动者网络 $\mu_{\phi_{\text{targ}}}(s)$ 的输出添加截断噪声，促使评论家网络学习更平滑的 Q函数，并防止利用尖锐的价值峰值。

TD3算法的伪代码见附录B中的算法 3。

D. SAC

软演员-评论家（SAC）通过引入最大熵强化学习扩展了演员-评论家框架；该方法在标准目标函数中加入熵项，以鼓励探索并增强鲁棒性 [51]。与依赖确定性策略的深度确定性策略梯度和双延迟深度确定性策略梯度不同，SAC 学习一种随机策略 $\pi_\phi(a|s)$，同时最大化期望回报和策略熵。由此得到的目标函数为

$$J(\pi) = E_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + \alpha H(\pi(\cdot|s_t)) \right) \right], \tag{30}$$

其中 α 是控制奖励最大化与熵 $H(\cdot)$之间权衡的温度参数， γ 是来自(21)的折扣率。请注意，深度确定性策略梯度和双延迟深度确定性策略梯度中的探索完全依赖于向从策略深度神经网络中选出的RIS反射系数添加噪声。对于SAC，随机策略通过对每个RIS反射系数生成高斯分布并从中随机采样来产生不同的动作。

SAC采用三个主要深度神经网络：随机策略网络 π_ϕ 作为演员，以及两个Q函数近似器Q_{θ_1} 和 Q_{θ_2} 作为评论家网络（及其训练网络和目标网络）。这两个评论家网络类似于双延迟深度确定性策略梯度中引入的双Q学习，通过在形成损失函数的目标时取其输出的最小值来减轻过估计偏差。然而，SAC的不同之处在于，损失函数同时包含下一状态价值和策略的熵。因此，用于训练评论家网络的新均方贝尔曼误差损失函数变为

$$\begin{aligned} l(\theta_i) = E\Bigg[\Big(& Q_{\theta_i}(s,a) - \\ & \Big(r' + \gamma \big(\min_{i=1,2} Q_{\theta_{\text{targ},i}}(s', \mu_{\phi_{\text{targ}}}(s')) - \alpha \log \pi_\phi(a'|s') \big) \Big) \Big)^2 \Bigg]. \end{aligned} \tag{31}$$

演员旨在最大化预期未来奖励和预期未来熵。演员损失函数为

$$l(\phi) = E_{a\sim\pi}[\min_{i=1,2} Q_{\theta_i}(s,a) - \alpha \log \pi_\phi(a|s)]. \tag{32}$$

SAC 的一个关键特性是自动温度调节机制，该机制会在训练期间调整 α，以维持目标熵水平 [52]。这样，算法无需手动调整熵权重，即可自适应地平衡探索与利用，并使用损失函数进行梯度下降

$$l(\log \alpha) = E_{a\sim\pi}[-\log \alpha (\log \pi_\phi(a|s) - \mathcal{H})], \tag{33}$$

其中， $\mathcal{H} = -\dim\ \mathcal{A} = -2M$ 是目标熵， $\mathcal{A}$ 是对应于可重构智能表面反射系数 $2M$实部和虚部的动作空间。我们更新 α 的对数值，以避免意外出现

负 α 值，并确保梯度平滑稳定。SAC的伪代码见附录B中的算法4。

总体而言，与深度确定性策略梯度和双延迟深度确定性策略梯度等确定性方法相比，SAC 具有多项优势，这主要归因于随机性带来的探索能力提升、对超参数的敏感性降低，以及熵正则化和双Q学习带来的稳定性增强。上述所有算法都适用于高维连续控制任务，例如本文讨论的 RIS 均衡问题。在下一小节中，我们将讨论这些算法各自的计算与通信成本，以便与 ARISE 算法进行公平比较。

E. 复杂度分析

根据深度确定性策略梯度、双延迟深度确定性策略梯度和 SAC 的更新规则结构，我们通过计算演员与评论家网络的前向/反向传播次数，推导每次更新的计算成本。据我们所知，尚无先前研究对这些算法进行正式的渐近复杂度分析。我们将单个深度神经网络每批次进行一次前向/反向传播的成本定义为 C，该成本取决于各深度神经网络的架构和超参数，以及批次大小。我们假设所有批次大小均相同，且这三种算法中的深度神经网络架构相似，因此 C 在这些算法之间不会有显著差异。请注意，各算法收敛前的总体计算复杂度会因其在 RIS 均衡问题上的表现而异——例如，是否需要长时间训练，或者能否快速优化反射系数。所有算法的单步复杂度汇总于表I。

对于深度确定性策略梯度，每步更新演员和评论家深度神经网络 N_{train} 次，得到 $T_{\text{DDPG}} = \mathcal{O}(N_{\text{train}}(C_{\text{critic}} + C_{\text{actor}}))$，其中 C_{critic} 是每次评论家网络更新的成本，C_{actor} 是每次演员更新的成本。我们在此表达式中不计入目标深度神经网络的更新，因为与通过深度神经网络进行一次前向/反向传播的复杂度相比，Polyak平均的复杂度可以忽略不计。

对于双延迟深度确定性策略梯度，我们有两个评论家深度神经网络和一个演员。演员每隔 d_{TD3} 次迭代更新一次，得到 $T_{\text{TD3}} = \mathcal{O}(N_{\text{train}}(2C_{\text{critic}} + \frac{1}{d_{\text{TD3}}} C_{\text{actor}}))$。SAC 同样有两个评论家深度神经网络和一个演员，并且每次迭代都会更新演员。假设与深度神经网络训练所需的反向传播相比，温度和目标更新的开销可以忽略不计，则得到 $T_{\text{SAC}} = \mathcal{O}(N_{\text{train}}(2C_{\text{critic}} + C_{\text{actor}}))$。

为推导 ARISE 的每次更新复杂度，我们参考式 (14) 中的更新规则。由于捕获的脉冲响应长度为 $L+1$，我们对 M 维向量执行 $L+1$ 次向量乘法。假设每个时间步都会引入一个新的脉冲响应，其中包含 $L+1$ 个待均衡的抽头，我们对 $L+1$ 个信号抽头取期望以计算集合平均，最终每步的开销为 $T_{\text{ARISE}} = \mathcal{O}(M(L+1)^2)$ [32]。请注意，对于 ARISE 的随机梯度下降（SGD）版本

在式 (18) 中，开销变为 $T_{\text{ARISE-SGD}} = \mathcal{O}(M(L+1))$，因为每次可重构智能表面更新时，脉冲只处理一次，不进行集合平均。

算法	每步复杂度
ARISE	$\mathcal{O}(M(L+1)^2)$
ARISE-SGD	$\mathcal{O}(M(L+1))$
DDPG	$\mathcal{O}(N_{\text{train}}(C_{\text{critic}} + C_{\text{actor}}))$
TD3	$\mathcal{O}(N_{\text{train}}(2C_{\text{critic}} + \frac{1}{d}C_{\text{actor}}))$
SAC	$\mathcal{O}(N_{\text{train}}(2C_{\text{critic}} + C_{\text{actor}}))$

表I：所提算法每步的计算复杂度。

每步的复杂度最终取决于深度神经网络的规模及其激活函数，这些因素决定了 C_{critic} 和 C_{actor}。根据可重构智能表面控制器所配备的硬件，深度神经网络的训练可以加速，从而减少每步所需的时间。

此外，ARISE与深度强化学习所涉及的通信复杂度差异显著。如前所述，我们的ARISE算法能否成功，主要取决于是否拥有 $\boldsymbol{h}_{\text{BRU}}$ 级联信道矩阵的信道状态信息，以便梯度下降规则能够正确运行。如果有 M 个信道，且脉冲响应中有 $L+1$ 个样本，则需要估计 $M(L+1)$个信道系数；这需要高度复杂的信道估计流程，其复杂度随可重构智能表面的规模以及所需均衡的程度而增加，近期研究（如 [18], [53]–[56]）对此已有记载。这会造成可重构智能表面优化的瓶颈，因为必须先估计信道，ARISE才能开始运行——这一过程通常需要将可重构智能表面配置为足够多种状态，才能准确确定信道矩阵。该估计过程需要时间，并可能在运行期间降低信号质量。此外，为了在实践中获取每次可重构智能表面更新后的脉冲响应，我们要么离线计算预期脉冲（这需要估计基站-用户设备链路 $\boldsymbol{h}_{\text{BU}}$的信道），要么重新配置可重构智能表面并从接收机获取脉冲响应。对于ARISE-SGD，由于梯度存在噪声，需要更多时间步才能收敛。相比之下，上文分析的深度强化学习算法无需显式信道估计，可直接处理接收脉冲响应。只要在每次可重构智能表面更新后提供脉冲响应，深度强化学习便可随时开始优化可重构智能表面，以提升信号质量。脉冲响应信息可通过带外通信可靠传输，既不会中断实时信号流量，又具有低时延，且不依赖于可重构智能表面的配置 [57]。此外，深度强化学习所支持的无模型运行能够自然地适应可重构智能表面的非线性、耦合效应和频率相关行为，因此是实现宽带可重构智能表面辅助均衡的一种更具可扩展性和鲁棒性的方法。

总的来说，尽管 ARISE 的计算开销可能较低，但其实际可行性主要受制于

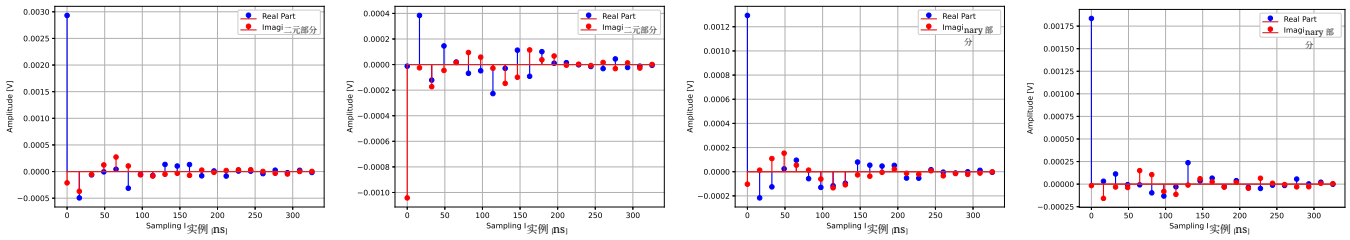
其所依赖的信道估计流程的复杂度显著更高，以及可重构智能表面的规模和所需的均衡量。相比之下，上文分析的深度强化学习算法每次更新的计算成本主要来自演员与评论家网络中的前向/反向传播，但它们无需显式进行信道估计，并且可以直接根据接收脉冲响应立即优化可重构智能表面。

V. 对比仿真结果

变量名称	描述	值
N_{train}	初始演员和评论家网络每次更新step	4
B	批次大小	4
$D_{\max}$	Max 经验回放缓冲区大小	400
d_{TD3}	SAC 演员更新周期	2
σ_{a}	深度确定性策略梯度和双延迟深度确定性策略梯度初始动作噪声标准差	0.3
τ_{d}	噪声衰减率	0.999
σ_{t}^2	双延迟深度确定性策略梯度训练动作标准差viation	0.3
c_{e}	双延迟深度确定性策略梯度训练动作噪声界限	0.5
μ_{c}	评论家网络学习rate	10^{-3}
μ_{a}	演员学习rate	深度确定性策略梯度: 10^{-3}; 双延迟深度确定性策略梯度和SAC: 2×10^{-3}
μ_{α}	SAC α学习rate	10^{-3}
τ	目标网络Polyak平均因子	0.001

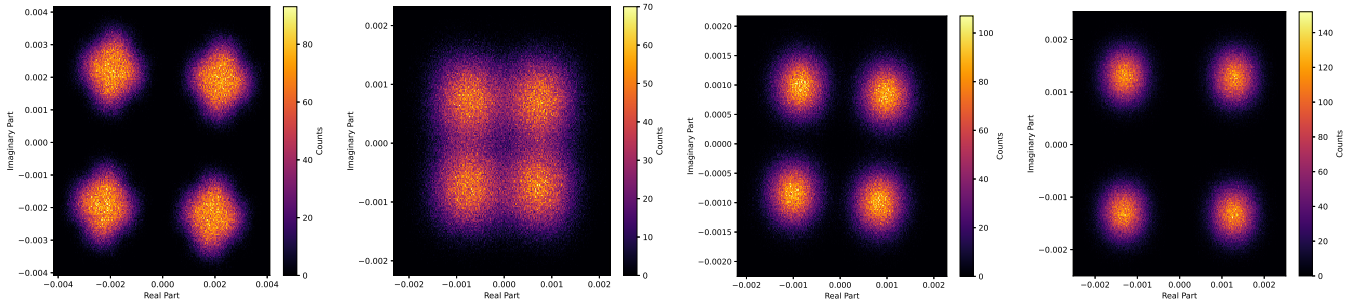
表II：深度强化学习实现参数。

我们采用与第III-B节相同的仿真设置，将深度强化学习算法的性能与 ARISE 进行比较。根据第III-B节的结果，我们为 ARISE算法使用缩放因子 $\alpha_{\text{s}} = 0.10$，以增强其均衡能力，并提供一个强有力的基线，用于与深度强化学习算法进行比较。这里，我们将一个“回合”定义为一个信道相干块。在每个回合开始时，



(a) ARISE, η_{h} = 0.848　(b) 深度确定性策略梯度, η_{h} = -0.258　(c) 双延迟深度确定性策略梯度, η_{h} = 0.770　(d) SAC, η_{h} = 0.883

图10：用户设备位置(-20, -20)处收敛脉冲响应的实部和虚部，$M = 100, \kappa = 10, n_{\text{r}} = 10$。

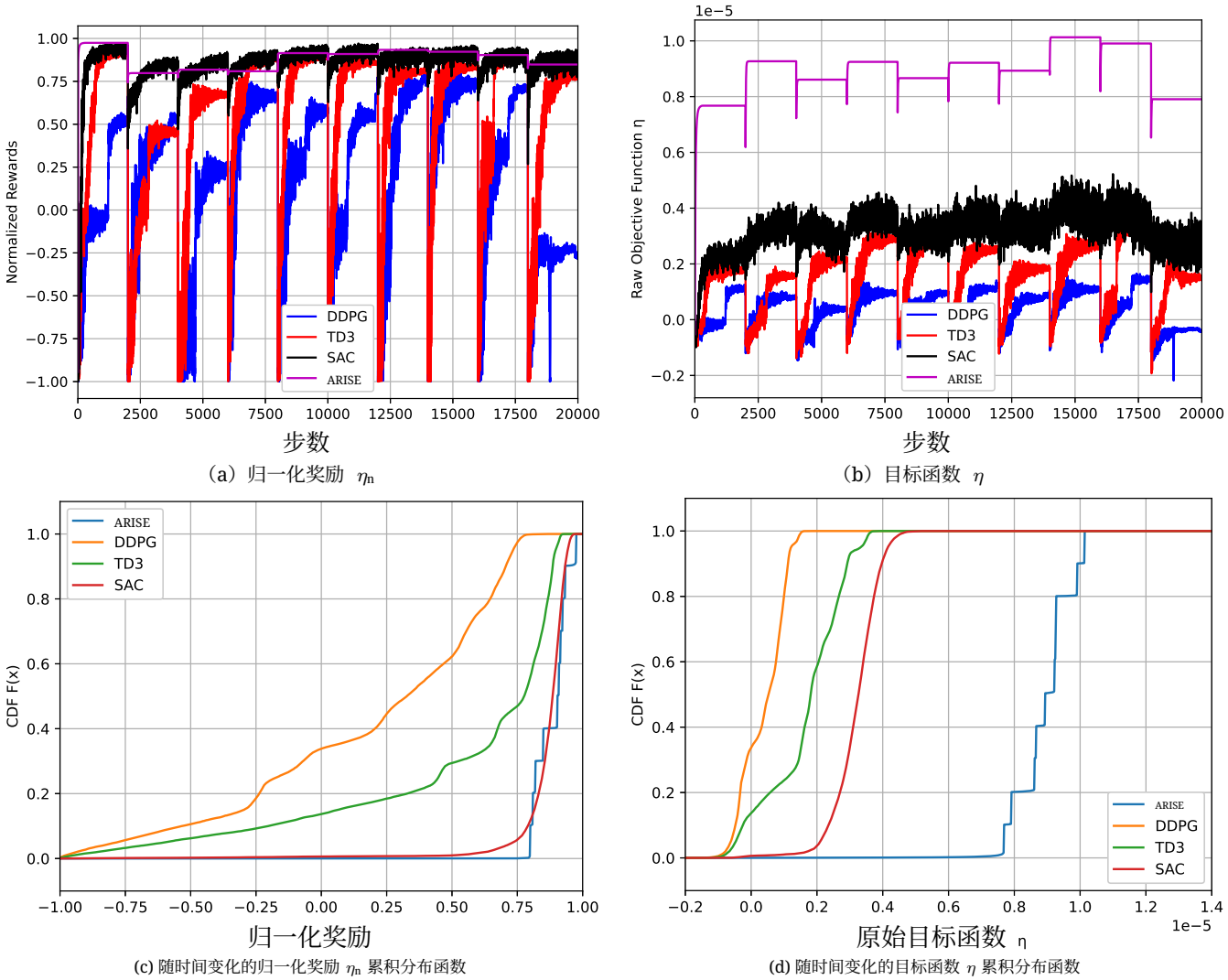


(a) ARISE，信噪比 = 11.01 分贝　(b) 深度确定性策略梯度，信噪比 = 5.20 分贝　(c) 双延迟深度确定性策略梯度，信噪比 = 9.00 分贝　(d) SAC，信噪比 = 12.11 分贝

图11：用户设备位置(-20, -20)处收敛的QPSK星座图，$M = 100, \kappa = 10, n_{\text{r}} = 10$。

用户设备按照随机游走方式移动。“时间步”对应于可重构智能表面反射系数的一次变化，随后获得相应的脉冲响应和奖励。每个回合包含 2000 个时间步。我们通过在将每个 r_{t+1} 存入经验回放缓冲区之前乘以 100 来进行奖励缩放，以提高网络的学习率。经验回放缓冲区按照先进先出（FIFO）策略更新，最大缓冲区大小为 $D_{\max}$。在每个回合开始时，我们会重置所有深度神经网络的经验回放缓冲区和学习率，以及 SAC温度系数 α。仅对于深度确定性策略梯度，我们还会随机化演员与评论家网络的权重，以避免陷入局部最优。

DDPG 和双延迟深度确定性策略梯度的动作噪声在每个时间步按 $10^{-\tau_{\text{d}}t}$ 进行缩放，从而提供衰减噪声调度，在每个回合开始时促进探索，并在每个回合结束时增加利用，以促进收敛稳定性，其中 t 为当前时间步索引，$t = 0$、1、……、1999，并在每个回合开始时重置。此外，我们根据每种算法获得的最近连续五个奖励的平均值，减少每个时间步的训练次数和演员学习率。例如，我们从 $\mu_{\text{a}} = 10^{-3}$ 和 $N_{\text{train}} = 4$开始。当连续 5 个奖励的平均值超过 0.75 时，我们设置 $\mu_{\text{a}} \leftarrow 0.8\mu_{\text{a}}$、$N_{\text{train}} \leftarrow N_{\text{train}} - 1$，并将奖励目标增加 0.05。对于奖励超过 0.8、然后超过 0.85 等情况，我们重复执行相同操作，同时将 N_{train} 的最小界限设为 1，以确保网络继续学习，同时促进稳定性，并在奖励值较高时减少训练迭代次数。所有其他相关的实现



（a）归一化奖励 η_{n}

（b）目标函数 η

(c) 随时间变化的归一化奖励 η_{n} 累积分布函数

(d) 随时间变化的目标函数 η 累积分布函数

图 12：目标函数随时间的变化， $M = 100, \kappa = 10, n_{\mathrm{r}} = 10$。

参数如表II所示。

所有演员和评论家深度神经网络均按照图 9 所示进行实现，包含两个隐藏层，每个隐藏层有 512 个节点，并使用 ReLU 激活函数 [58]。具体而言，对于深度确定性策略梯度，我们在演员与评论家网络的每个隐藏层之后应用层归一化，以提高算法的训练稳定性。所有层均使用随机正态初始化器进行初始化，均值为零，标准差为 0.1 [59]。这些网络使用 TensorFlow Keras [46] 在 Python 中进行训练，并使用 Adam 优化器 [60]。

我们首先分析对于位于$(-20, -20)$的静止用户设备，深度强化学习算法相较于ARISE算法的性能。假设基站-用户设备链路服从瑞利衰落，基站-RIS-用户设备链路的莱斯因子为 $\kappa = 10$ ，且存在 $n_{\mathrm{r}} = 10$ 条延迟路径（因此有 $L = 20$ 个符号间干扰分量）。仿真场景包含10个回合，每个回合开始时，非视距信道都会随机变化。为确保公平，我们假设ARISE在每个回合开始时都能获取完整的级联信道状态信息，

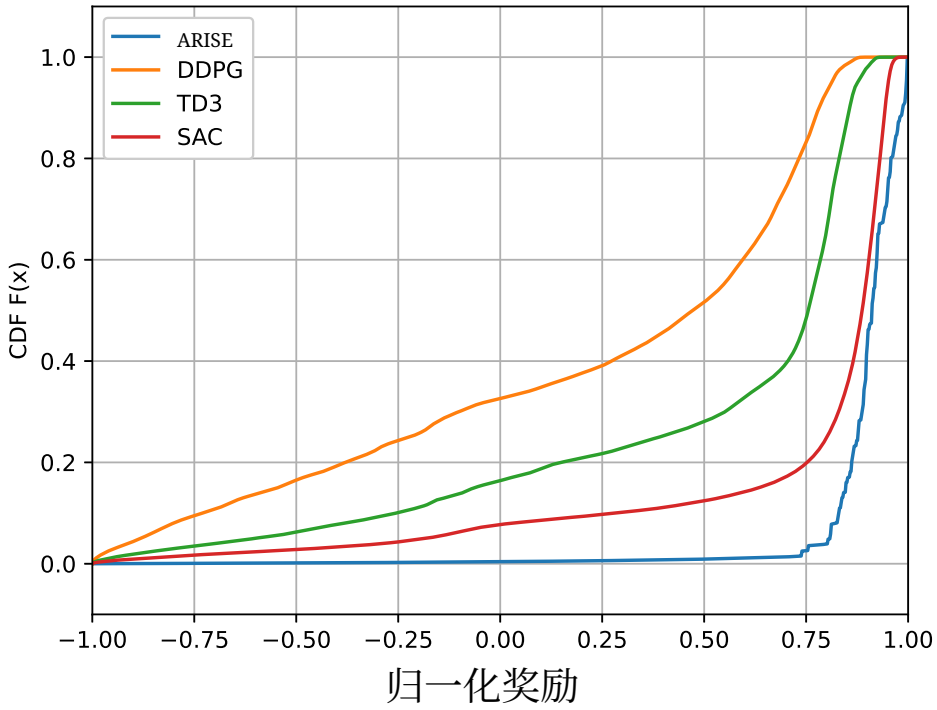


图13：归一化奖励 η_{n} 在100个回合持续期间的每一步， $M = 100$， $\kappa = 10$， $n_{\mathrm{r}} = 10$。

尽管在实际应用中，这需要在开始优化之前进行一个专门的信道估计阶段，其中涉及多种可重构智能表面配置。图10绘出了经过10个回合后收敛的脉冲响应，图11则展示了相应的QPSK星座图。此外，图12展示了归一化奖励 r_{t+1} 和目标函数 η。我们观察到，各深度强化学习算法的表现差异显著。对于相同的信道系数，SAC的均衡性能与ARISE相当，有时甚至更优；其收敛速度非常快，在初始阶段不到1000步便能收敛至较高的奖励值，此后即使信道发生变化，也始终保持较高的奖励。其次，双延迟深度确定性策略梯度（TD3）的收敛时间比SAC更长，但似乎在几个回合后便能收敛至较高的奖励值，并保持稳定。深度确定性策略梯度（DDPG）则表现落后，且陷入局部最优，往往无法达到TD3的性能。比较各算法实现的目标函数 η 可见，ARISE始终取得最高的目标值，这主要归功于其放大能力：它会尝试将衰减信号与其放大并均衡后的版本相匹配。深度强化学习算法的奖励定义并未明确要求信号放大，因为奖励是相对于脉冲响应采样值绝对值之和进行归一化的。平均而言，使用ARISE优化的可重构智能表面比使用SAC优化的可重构智能表面多出约3分贝的功率增益，尽管在给定场景下，两者的均衡能力几乎相同。值得特别注意的是，SAC在无需ARISE所依赖的大量导频传输和完整级联信道估计的情况下，就达到了与ARISE相当的均衡性能，从而降低了

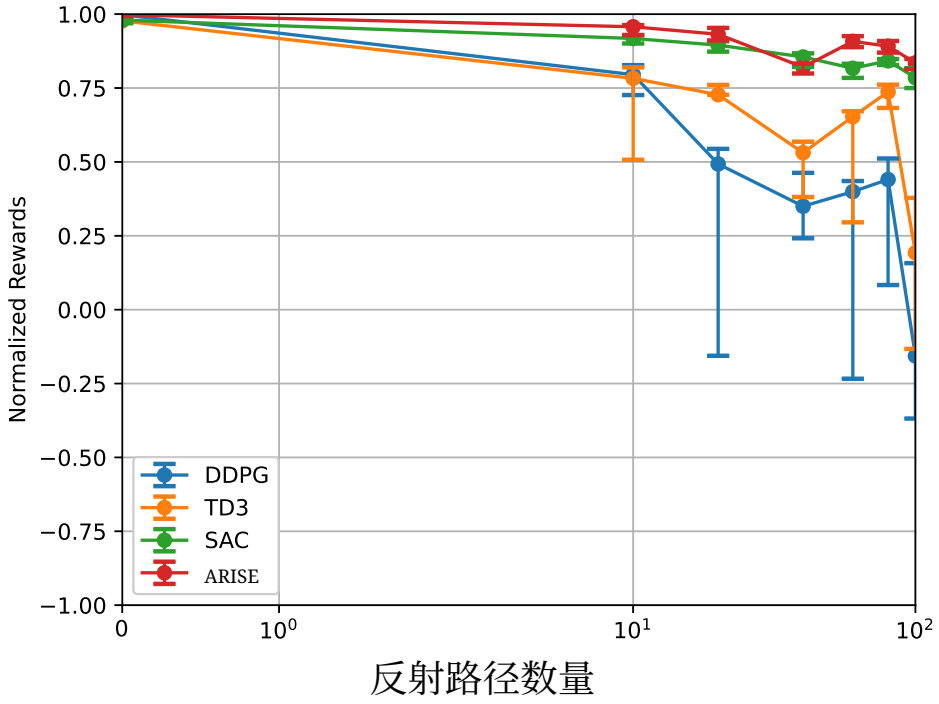


图14：收敛的归一化奖励 η_{n} 与延迟路径数量 n_{r}, $M = 100$, $\kappa = 10$的关系。

与可重构智能表面优化相关的通信开销。

接下来，我们采用随机游走模型，并在更具挑战性的情况下比较各算法的性能，以评估它们对不同可重构智能表面场景的适用性。在图13中，我们使用用户设备持续100个回合的随机游走模型，并比较归一化奖励 η_{n}随时间的变化；可重构智能表面规模和环境仍与前一种情况相同。尽管视距分量和非视距分量在每个回合开始时都会完全改变，SAC仍始终保持快速且出色的均衡能力，超过80%的时间归一化奖励高于0.75，性能接近ARISE。与ARISE不同，后者在用户设备移动时需要反复进行信道重新估计，而SAC可直接根据接收脉冲响应进行适应，从而避免了额外的通信负担。双延迟深度确定性策略梯度的收敛速度比SAC慢，但大多数时间仍能保持较高奖励，而深度确定性策略梯度则表现较差。

接下来，我们在图 14 中比较收敛的归一化奖励与信号所经历的延迟路径数量 n_{r} 之间的关系，其中码间干扰分量总数为 $L = 2n_{\mathrm{r}}$ 。每个数据点表示连续 10 个回合中最后 10 步的平均归一化奖励，误差条显示均值上下的标准差。在这里，SAC 和 ARISE 仍表现出相近且稳定性较高的性能，这从较小的标准差和较高的平均归一化奖励可以看出；而 TD3 和 SAC 的奖励较低，其中 SAC 在不同回合间的方差更大。重要的是，随着码间干扰分量数量的增加，SAC 无需像 ARISE 那样估计不断增加的

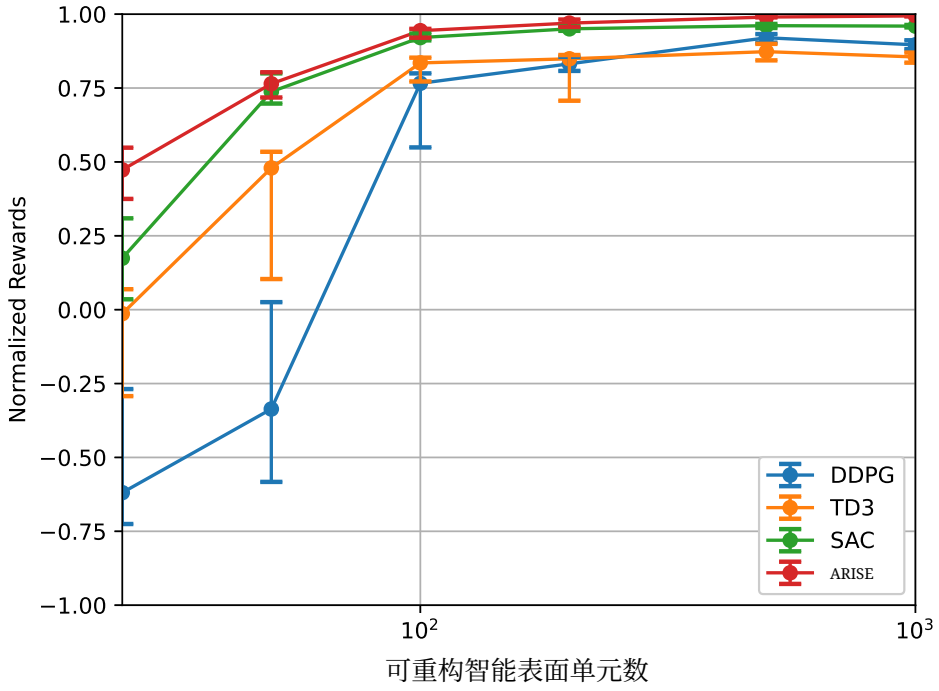


图 15：收敛的归一化奖励 η_{n} 与可重构智能表面单元数的关系 M, $n_{\mathrm{r}}=10$, $\kappa=10$。

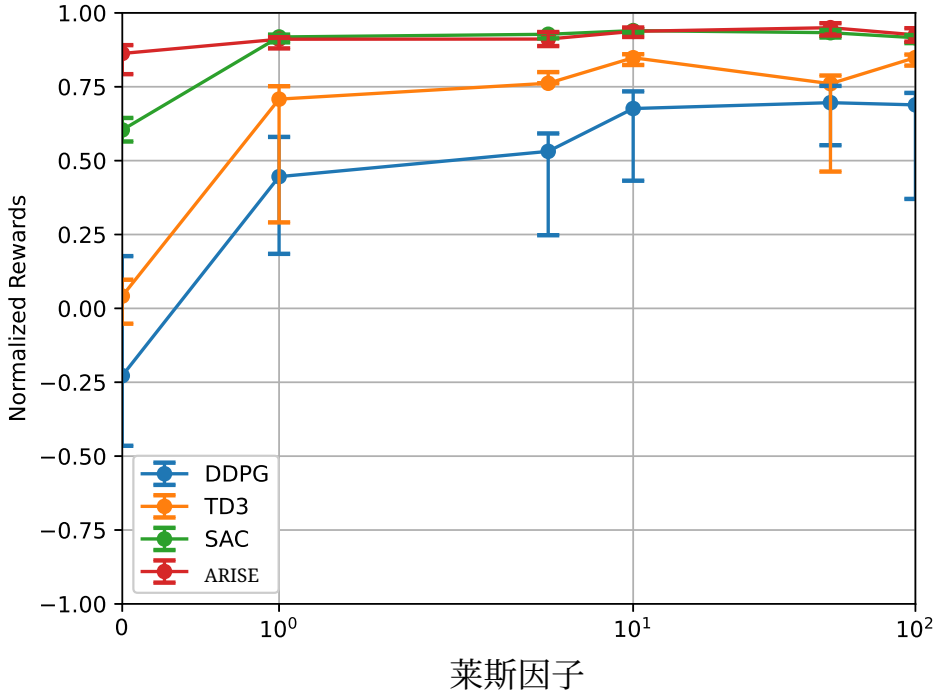


图 16：收敛的归一化奖励 η_{n} 与莱斯因子的关系 κ, $M=100$, $n_{\mathrm{r}}=10$。

级联信道的　　信道系数。

随着向信号中引入更多码间干扰项，100单元可重构智能表面作为均衡器的整体性能仍保持一致。我们在图15中评估了随着可重构智能表面规模扩大，深度强化学习算法的均衡能力。我们注意到，修改RIS单元数量 M 会改变输出维度

演员深度神经网络的输出维度，而改变延迟路径数量 n_{r} 可能会改变演员和评论家深度神经网络的输入维度（具体取决于所选择的可重构智能表面算子，以均衡特定数量的码间干扰项）。在实验中，我们会根据每个变量的变化，相应地修改演员和评论家深度神经网络的输入和输出维度，同时保持其隐藏层不变。在图15所示的场景中，我们观察到，当可重构智能表面规模较小时，其均衡能力明显低于之前的仿真案例，并且这一现象适用于所有算法，其中ARISE表现最佳。随着可重构智能表面规模增大，各算法的性能系统性提升，稳定性也不断增强。此外，当RIS单元数量超过200后，ARISE仍在小幅持续改进，而深度强化学习算法似乎趋于饱和。这很可能是由于深度神经网络的隐藏层保持相同规模，在需要适应不断增加的随机信道的同时，还要具备足够的泛化能力以消除大部分码间干扰，因此难以满足这一要求。

最后，我们在图16中评估了各算法在基站-RIS-用户设备链路 κ的莱斯因子不同时的性能。对于瑞利情形，ARISE的性能显著优于 $\eta_{\mathrm{n}} \approx 0.85$的深度强化学习算法：DDPG完全无法均衡信号，TD3略好一些，而SAC始终达到 $\eta_{\mathrm{n}} \approx 0.6$的平均值。随着莱斯因子增大，深度神经网络能够学习无线信道中视距分量的几何结构，其性能显著提升，SAC的均衡性能也达到与ARISE相同的水平。尽管SAC无需显式进行CSI获取，也无需使用可重构智能表面模型，仍能达到同等性能，这进一步凸显了其在通信效率方面的优势。

总体而言，我们证明了深度强化学习算法能够成功将可重构智能表面配置为均衡器：在大多数情况下，SAC的性能与ARISE相同，在其他情况下也不相上下。尽管其单步计算成本高于TD3和DDPG，但它提供了一种快速高效地提升链路信号完整性的方法，无需进行过多的信道估计。综合来看，这些结果表明，SAC在实现ARISE级别的均衡精度的同时，通信复杂度显著更低，计算开销也仅适中，因此是更适合实际可重构智能表面部署的可扩展解决方案。

VI. 结论

在本研究中，我们展示了可重构智能表面技术在各种信道条件和无线环境下，既可用作信号增强器，也可用作均衡器。我们开发了一种基于最速下降法的方法——ARISE，该方法利用估计得到的级联基站-RIS-用户设备信道，并展示了其配置RIS系数以均衡接收脉冲响应、增强信噪比的能力。我们还探讨了将ARISE应用于无源可重构智能表面时面临的挑战，尤其是实现准确均衡所需的大量信道

估计，以及由此产生的信号增益与均衡性能之间的权衡——这种影响可能需要迭代优化或更深入的信噪比分析。

作为替代方案，我们探索了几种深度强化学习方法——深度确定性策略梯度、双延迟深度确定性策略梯度和SAC——这些方法无需进行信道估计，仅依靠接收脉冲响应来优化RIS系数。通过比较计算复杂度和可实现的性能，我们发现SAC最为有效：它能在不同信道和可重构智能表面规模下快速收敛至稳定解，且性能可与ARISE相媲美。更重要的是，SAC无需像ARISE那样进行繁重的信道估计即可达到这一性能，从而显著降低通信复杂度，使整个优化流程更具实用性。SAC的无模型特性使其能够天然适应RIS非线性、单元耦合以及频率相关反射特性，而这些因素难以通过解析信道模型加以刻画。此外，与必须等到完成完整CSI获取后才能开始优化的ARISE不同，SAC可以根据原始脉冲响应观测立即调整RIS配置，从而在动态环境中实现更快的适应并降低时延。总的来说，凭借实现简单、鲁棒性强和性能出色等优势，深度强化学习有望成为未来RIS辅助无线系统中一个重要且不可或缺的组成部分。

基于本研究的贡献，未来工作应分析在多用户、多天线基站的情况下 ARISE 和深度强化学习算法的性能，并将可重构智能表面从均匀线性阵列（ULA）扩展为均匀平面阵列（UPA），同时纳入具有硬件限制和非线性的物理RIS模型。这些工作将把可重构智能表面用作均衡器的概念推广到更广泛的场景，并提升其在未来无线多输入多输出（MIMO）系统中的实用性。

附录A

ARISE算法推导

我们希望最小化期望信号 $\boldsymbol{s}$ 与接收信号$\boldsymbol{y}$之间的均方误差（MSE）

$$\boldsymbol{y} = \left(\boldsymbol{h}_{\text{BU}} + \sum_{m=1}^{M} \Gamma_m \boldsymbol{h}_{\text{BRU},m} \right) * \boldsymbol{s} \tag{34}$$

通过对 RIS反射系数 Γ_m进行梯度下降，其中 $m = 1$、2、.....、M。代价函数定义为 $J = E[\epsilon_k \epsilon_k^*]$，误差信号为 $\boldsymbol{\epsilon} = \boldsymbol{s} - \boldsymbol{y}$。这里， $\boldsymbol{s} = [1, 0, \ldots, 0]$是传输的离散时间脉冲。我们可以将 Γ_m 表示为其实部和虚部的组合

$$\Gamma_m = a_m + j b_m, \tag{35}$$

从而将 $\boldsymbol{y}$ 表示为

$$\boldsymbol{y}=\left(\boldsymbol{h}_{\text{BU}}+\sum_{m=1}^{M}(a_m+jb_m)\boldsymbol{h}_{\text{BRU},m}\right)*\boldsymbol{s}. \tag{36}$$

相对于 Γ_m 的误差梯度定义为

$$\frac{\partial\boldsymbol{\epsilon}}{\partial\Gamma_m}=\frac{\partial\boldsymbol{\epsilon}}{\partial a_m}+j\frac{\partial\boldsymbol{\epsilon}}{\partial b_m} \tag{37}$$

并且与 $\boldsymbol{y}$ 的关系为

$$\frac{\partial\boldsymbol{\epsilon}}{\partial a_m}=-\frac{\partial\boldsymbol{y}}{\partial a_m};\quad \frac{\partial\boldsymbol{\epsilon}}{\partial b_m}=-\frac{\partial\boldsymbol{y}}{\partial b_m}. \tag{38}$$

展开卷积算子，并考虑 $\boldsymbol{y}$、 y_k中的各项，可得

$$\frac{\partial y_k}{\partial a_m}=\sum_{l=0}^{L}h_{\text{BRU},m,l}s_{k-l} \tag{39}$$

and

$$\frac{\partial y_k}{\partial b_m}=j\sum_{l=0}^{L}h_{\text{BRU},m,l}s_{k-l}. \tag{40}$$

对代价函数求导可得

$$\begin{aligned}\frac{\partial J}{\partial a_m}&=E\left[\frac{\partial\epsilon_k}{\partial a_m}\epsilon_k^*+\frac{\partial\epsilon_k^*}{\partial a_m}\epsilon_k\right]\\&=E\left[-\sum_{l=0}^{L}h_{\text{BRU},m,l}s_{k-l}\epsilon_k^*-\sum_{l=0}^{L}h_{\text{BRU},m,l}^*s_{k-l}^*\epsilon_k\right]\end{aligned} \tag{41}$$

and

$$\begin{aligned}\frac{\partial J}{\partial b_m}&=E\left[\frac{\partial\epsilon_k}{\partial b_m}\epsilon_k^*+\frac{\partial\epsilon_k^*}{\partial b_m}\epsilon_k\right]\\&=E\left[-j\sum_{l=0}^{L}h_{\text{BRU},m,l}s_{k-l}\epsilon_k^*+j\sum_{l=0}^{L}h_{\text{BRU},m,l}^*s_{k-l}^*\epsilon_k\right].\end{aligned} \tag{42}$$

合并后可得

$$\frac{\partial J}{\partial\Gamma_m}=E\left[-2\sum_{l=0}^{L}h_{\text{BRU},m,l}^*s_{k-l}^*\epsilon_k\right] \tag{43}$$

最终梯度为

$$\nabla_{\boldsymbol{\Gamma}}=-2E\begin{bmatrix}\sum_{l=0}^{L}h_{\text{BRU},1,l}^*s_{k-l}^*\\\sum_{l=0}^{L}h_{\text{BRU},2,l}^*s_{k-l}^*\\\vdots\\\sum_{l=0}^{L}h_{\text{BRU},M,l}^*s_{k-l}^*\end{bmatrix}\epsilon_k. \tag{44}$$

最后，我们得到 $\mathbf{\Gamma}$ 的更新规则，该规则由梯度下降给出

$$\begin{aligned}\mathbf{\Gamma}_{t+1} &= \mathbf{\Gamma}_t - \frac{1}{2}\mu\nabla_{\mathbf{\Gamma}_t} \\ &= \mathbf{\Gamma}_t + \mu E\left[\sum_{l=0}^{L}\boldsymbol{h}_{\mathrm{BRU},l}^{*}s_{k-l}^{*}\epsilon_k\right],\end{aligned} \tag{45}$$

其中， μ 为步长。为满足可重构智能表面的无源性要求，我们在每次更新后将反射系数的幅度归一化为

$$\Gamma_{t+1,m} \leftarrow \frac{\Gamma_{t+1,m}}{|\Gamma_{t+1,m}|} \tag{46}$$

对所有 $m=1$、2、.....、M 均成立。

APPENDIX B
TD3 AND SAC ALGORITHMS PSEUDOCODES

附录中的算法 3 给出了双延迟深度确定性策略梯度算法的伪代码，算法4 则给出了 SAC 算法的伪代码。

参考文献

[1] C. Pan、H. Ren、K. Wang、J. F. Kolb、M. Elkashlan、M. Chen、M. Di Renzo、Y. Hao、J. Wang、A. Swindlehurst、X. You 和 L. Hanzo，“面向 6G系统的可重构智能表面：原理、应用与研究方向”，《IEEE通信杂志》，第 59 卷，第 6 期，第 14–20 页，2021。[2] M. A. ElMossallamy、H. Zhang、L. Song、K. G. Seddik、Z. Han 和 G. Y. Li，“面向无线通信的可重构智能表面：原理、挑战与机遇”，《IEEE认知通信与网络汇刊》，第 6 卷，第 3 期，第 990–1002 页，2020。[3] C. Pan、G. Zhou、K. Zhi、S. Hong、T. Wu、Y. Pan、H. Ren、M. D. Renzo、A. LeeSwindlehurst、R. Zhang 和 A. Y. Zhang，“RIS/IRS辅助无线系统的信号处理技术概述”，IEEE 信号处理专题期刊，第 16 卷，第 5 期，第 883–917 页，2022。[4] X. Wei、D. Shen 和 L. Dai，“RIS辅助无线通信的信道估计——部分 I：基础、解决方案与未来机遇”，《IEEE通信快报》，第 25 卷，第 5 期，第 1398–1402 页，2021。[5] H. Zhou、M. Erol-Kantarci、Y. Liu 和 H. V. Poor，“关于基于模型、启发式和机器学习优化方法在RIS辅助无线网络中的综述”，2023年。 [在线]。网址：https://arxiv.org/abs/2303.14320[6] C. Huang、A. Zappone、G. C. Alexandropoulos、M. Debbah 和 C. Yuen，“可重构智能表面在无线通信能源效率方面的应用”，IEEE无线通信汇刊，第 18 卷，第 8 期，第 4157–4170 页，2019。[7] M. Di Renzo、A. Zappone、M. Debbah、M.-S. Alouini、C. Yuen、J. de Rosny 和 S. Tretyakov，“可重构智能表面赋能的智能无线电环境：工作原理、研究现状与未来展望”，IEEE通信选域期刊，第 38 卷，第 11 期，第 2450–2525 页，2020。[8] G. Zhou、C. Pan、H. Ren、K. Wang 和 A. Nallanathan，“面向具有不完美级联信道的IRS辅助MISO通信的鲁棒传输设计框架”，IEEE信号处理汇刊，第 68 卷，第 5092–5106 页，2020。[9] X. Pei、H. Yin、L. Tan、L. Cao、Z. Li、K. Wang、K. Zhang 和 E. Bj¨ornson，“RIS辅助无线通信：原型开发、自适应波束成形及室内/室外现场试验”，IEEE通信汇刊，第 69 卷，第 12 期，第 8627–8640 页，2021。

算法 3 用于RIS均衡器的双延迟深度确定性策略梯度

初始化演员和评论家网络的训练深度神经网络与目标深度神经网络，并赋值 $\phi_{\text{targ}} \leftarrow \phi$, $\theta_{\text{targ}} \leftarrow \theta$。

对于 N_{ep} 中的每个回合，执行

  对于 N_{steps} 中的每个 t，执行

    观测状态 $s_t \leftarrow \boldsymbol{y}'_t$ 并设置RIS系数

$$\boldsymbol{\Gamma}'_t \leftarrow \text{clip}(\mu_{\phi_{\text{targ}}}(\boldsymbol{y}'_t) + \boldsymbol{\epsilon}, -1, 1), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_{\text{a}}^2)^{2M\times 1}.$$

    设置 $q \leftarrow \boldsymbol{\Gamma}'_t$.

    更新 RIS反射系数，使其为

$$\boldsymbol{\Gamma}_t \leftarrow [\Gamma'_{t,1} + j\Gamma'_{t,2}, \ldots, \Gamma'_{t,2M-1} + j\Gamma'_{t,2M}]^T$$

  并使用 (15) 进行归一化。

    观察新状态 $s_{t+1} \leftarrow \boldsymbol{y}'_{t+1}$。

    使用 (20) 计算奖励 r_{t+1}

    将 $(s_t, a_t, r_{t+1}, s_{t+1})$ 存储到经验回放缓冲区 $\mathcal{D}$ 中。

    如果 $|\mathcal{D}| \geq B$ 那么

      对于 i，遍历 N_{train} **do**

        从 $\mathcal{D}$中采样样本批次 $\mathcal{B} = \{(s, a, r', s')\}$。

        向 a添加噪声：

$$a \leftarrow \text{clip}(a + \text{clip}(\epsilon, -c_e, c_e), -1, 1) \quad \epsilon \sim \mathcal{N}(0, \sigma_{\text{t}}^2).$$

        使用公式 (29) 计算 $l(\theta_i)$，其中 $i = 1$ 为 2。

        更新评论家网络：

$$\theta_i \leftarrow \theta_i - \mu_{\text{c}} \nabla_{\theta_i} l(\theta_i).$$

        使用（27）计算 $l(\phi)$。

        如果 $i \bmod d_{\text{p}} == 0$ 则

          更新演员：

$$\phi \leftarrow \phi - \mu_{\text{a}} \nabla_{\phi} l(\phi).$$

          软更新目标评论家：

$$\theta_{\text{targ},i} \leftarrow \tau\theta_i + (1-\tau)\theta_{\text{targ},i}, \quad i = 1, 2.$$

          软更新目标演员：

$$\phi_{\text{targ}} \leftarrow \tau\phi + (1-\tau)\phi_{\text{targ}}.$$

        结束条件判断

      结束循环

    结束条件判断

  结束循环

结束循环

算法4 可重构智能表面均衡器的软演员-评论家（SAC）

1：初始化演员深度神经网络超参数 ϕ、训练评论家深度神经网络和目标评论家深度神经网络，并赋值 $\theta_{\text{targ}} \leftarrow \theta$。

2：初始化温度参数 α。

3：对于 N_{ep} 中的每个回合，执行

4：　对于 N_{steps} 中的每个 t，执行

5：　观测状态 $s_t \leftarrow \boldsymbol{y}'_t$。

6：　从随机策略中采样动作：

$$a_t \sim \pi_\phi(a_t \mid s_t).$$

7：　裁剪RIS系数： $\boldsymbol{\Gamma}'_t \leftarrow \text{clip}(a_t, -1, 1)$。

8：　将RIS反射系数更新为

$$\boldsymbol{\Gamma}_t \leftarrow [\Gamma'_{t,1} + j\Gamma'^{\,2}_{t,}, \cdots, {}^{\Gamma'_t}{}_{,2M-1} + j\Gamma'_{t,2M}]^T$$

并用(15)进行归一化。

9：　观察新状态 $s_{t+1} \leftarrow \boldsymbol{y}'_{t+1}$。

10：　使用公式 (20)计算奖励 r_{t+1}。

11：　将 $(s_t, a_t, r_{t+1}, s_{t+1})$ 存储到经验回放缓冲区 $\mathcal{D}$ 中。

12：　如果 $|\mathcal{D}| \geq B$ 则

13：　对于 i 中的每个 N_{train}，执行

14：　从 $\mathcal{D}$. 中采样批次 $\mathcal{B} = \{(s, a, r', s')\}$

15：　采样下一个动作和对数概率：

$$a' \sim \pi_\phi(a' \mid s'), \quad \log \pi_\phi(a' \mid s').$$

16：　使用 (31) 计算 $l(\theta_i)$，其中 $i = 1$ 为 2。

17：　更新评论家网络：

$$\theta_i \leftarrow \theta_i - \mu_{\text{c}} \nabla_{\theta_i} l(\theta_i).$$

18：　为演员更新采样动作：

$$a \sim \pi_\phi(a \mid s), \quad \log \pi_\phi(a \mid s).$$

19：　使用 (32) 计算 $l(\phi)$。

20：　更新演员：

$$\phi \leftarrow \phi - \mu_{\text{a}} \nabla_\phi l(\phi).$$

21：　使用 (33) 计算 $l(\log \alpha)$。

22：　更新温度：

$$\log \alpha \leftarrow \log \alpha - \mu_\alpha \nabla_{\log \alpha} l(\log \alpha).$$

23：　$\alpha \leftarrow \exp(\log \alpha)$。

24：　软更新目标评论家：

$$\theta_{\text{targ},i} \leftarrow \tau\theta_i + (1-\tau)\theta_{\text{targ},i}, \qquad i = 1, 2。$$

25：　结束循环

26：　结束条件判断

27：　结束循环

28：结束循环

[10] R. Liu, M. Li, Y. Liu, Q. Wu, and Q. Liu, 《RIS 辅助双功能雷达通信系统的发射波形与无源波束成形联合设计》，IEEE 信号处理专题期刊，第 16 卷，第 5 期，第 995–1010 页，2022 年。[11] G. Ben-Itzhak, M. Saavedra-Melo, B. Bradshaw, E. Ayanoglu, F. Capolino, and A. Lee Swindlehurst, 《波控可重构智能表面的设计与运行原理》，IEEE 通信学会开放期刊，第 5 卷，第 7730–7751 页，2024 年。[12] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, 《可重构智能表面：原理与机遇》，IEEE Communications Surveys & Tutorials，第 23 卷，第 3 期，第 1546–1577 页，2021 年。[13] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, 《可重构智能表面辅助无线网络的加权和速率最大化》，IEEE无线通信汇刊，第 19 卷，第 5 期，第 3064–3076 页，2020 年。[14] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, 《同时透射与反射（STAR）RIS 辅助无线通信》，IEEE无线通信汇刊，第 21 卷，第 5 期，第 3083–3098 页，2022 年。[15] Z. Yu, H. Ren, C. Pan, G. Zhou, B. Wang, M. Dong, and J. Wang, 《有源 RIS 辅助通感一体化系统：波束成形设计与性能分析》，IEEE通信汇刊，第 72 卷，第 3 期，第 1578–1595 页，2024 年。[16] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, 《面向可重构智能表面多用户通信的混合波束成形：有限离散相移下的可达速率》，IEEE通信选域期刊，第 38 卷，第 8 期，第 1809–1822 页，2020 年。[17] J. He, H. Wymeersch, and M. Juntti, 《基于原子范数最小化的 RIS 辅助毫米波 MIMO 系统信道估计》，IEEE无线通信汇刊，第 20 卷，第 9 期，第 5786–5797 页，2021 年。[18] X. Wei, D. Shen, and L. Dai, 《RIS辅助无线通信的信道估计——第二部分：基于双结构稀疏性的改进方案》，《IEEE通信快报》，第 25 卷，第 5 期，第 1403–1407 页，2021 年。[19] C. Huang, R. Mo, and C. Yuen, 《利用深度强化学习的可重构智能表面辅助多用户 MISO 系统》，IEEE通信选域期刊，第 38 卷，第 8 期，第 1839–1850 页，2020 年。[20] B. Saglam, D. Gurgunoglu, and S. S. Kozat, 《存在硬件损伤和不完美信道状态信息时，RIS 辅助 MU-MISO 系统中基于深度强化学习的下行波束成形与 RIS 配置联合优化》，载于 2023 年 IEEE 国际通信会议研讨会（ICC Workshops），2023 年，第 66–72 页。[21] K. Kim, Y. K. Tun, M. S. Munir, W. Saad, and C. S. Hong, 《用于 RIS 辅助无线网络信道估计的深度强化学习》，《IEEE通信快报》，第 27 卷，第 8 期，第 2053–2057 页，2023 年。[22] M.-L. Tham, Y. J. Wong, A. Iqbal, N. B. Ramli, Y. Zhu, and T. Dagiuklas, 《利用可重构智能表面的深度强化学习保密能效无人机通信》，载于 2023 年 IEEE 无线通信与网络会议（WCNC），2023 年，第 1–6 页。[23] H. Zhang, L. Song, Z. Han, and H. V. Poor, 《接收前的空域均衡：用于多径抑制的可重构智能表面》，载于 ICASSP 2021——2021 年 IEEE 声学、语音与信号处理国际会议（ICASSP），2021 年，第 8062–8066 页。[24] H. Prod’homme, M. F. Imani, S. Abadal, and P. del Hougne, 《资源受限无线网络中基于 RIS 的空口信道均衡》，载于 2024 年第 18 届欧洲天线与传播会议（EuCAP），2024 年，第 1–5 页。[25] B. Widrow and J. McCool, 《基于最速下降法和随机搜索法的自适应算法比较》，IEEE 天线与传播汇刊，第 24 卷，第 5 期，第 615–637 页，1976 年。[26] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, 《面向未来无线通信系统设计的宽带毫米波传播测量与信道模型》，IEEE通信汇刊，第 63 卷，第 9 期，第 3029–3056 页，2015 年。[27] A. Saleh and R. Valenzuela, 《室内多径传播的统计模型》，IEEE通信选域期刊，第 5 卷，第 2 期，第 128–137 页，1987 年。[28] S. Zeng, H. Zhang, B. Di, Z. Han, and L. Song, 《可重构智能表面（RIS）辅助无线覆盖扩展：RIS 朝向与位置优化》，《IEEE通信快报》，第 25 卷，第 1 期，第 269–273 页，2021 年。

[29] H. Zhang, B. Di, L. Song, and Z. Han, 《有限相移下的可重构智能表面辅助通信：多少个相移才够？》，IEEE Vehicular Technology 汇刊，第 69 卷，第 4 期，第 4498–4502 页，2020 年。[30] E. Bj¨ornson and L. Sanguinetti, 《可重构智能表面的瑞利衰落建模与信道硬化》，IEEE 无线通信快报，第 10 卷，第 4 期，第 830–834 页，2021 年。[31] Z. Abdullah, A. Papazafeiropoulos, S. Kisseleff, S. Chatzinotas, and B. Ottersten, 《相位噪声与空间相关性对双 RIS 辅助多用户 MISO 网络的影响》，IEEE 无线通信快报，第 11 卷，第 7 期，第 1473–1477 页，2022 年。[32] A. H. Sayed, 《自适应滤波器》。John Wiley & Sons，2011 年。[33] S. Qureshi, 《自适应均衡》，IEEE 论文集，第 73 卷，第 9 期，第 1349–1387 页，1985 年。[34] B. Widrow, J. McCool, M. Larimore, and C. Johnson, 《LMS 自适应滤波器的平稳与非平稳学习特性》，IEEE 论文集，第 64 卷，第 8 期，第 1151–1162 页，1976 年。[35] 3第 3 代合作伙伴计划（3GPP），新空口；物理信道与调制，3GPP 技术规范 TS 38.211，2024 年，第 17 版，版本 17.6.0。[36] M. Jian, G. C. Alexandropoulos, E. Basar, C. Huang, R. Liu, Y. Liu, and C. Yuen, 《面向无线通信的可重构智能表面：硬件设计、信道模型与估计技术综述》，《智能与融合网络》，第 3 卷，第 1 期，第 1–32 页，2022 年。[37] A. L. Swindlehurst, G. Zhou, R. Liu, C. Pan, and M. Li, 《可重构智能表面的信道估计——通用框架》，IEEE 论文集，第 110 卷，第 9 期，第 1312–1338 页，2022 年。[38] G. Ben-Itzhak, M. Saavedra-Melo, E. Ayanoglu, F. Capolino, and A. L. Swindlehurst, 《波控可重构智能表面的人工智能驱动优化》，IEEE 通信学会开放期刊，第 6 卷，第 6650–6665 页，2025 年。[39] K. Keykhosravi, M. F. Keskin, G. Seco-Granados, P. Popovski, and H. Wymeersch, 《用户移动性与空间宽带效应下 RIS 辅助 SISO 定位》，IEEE 信号处理专题期刊，第 16 卷，第 5 期，第 1125–1140 页，2022 年。[40] R. Wang, Y. Yang, B. Makki, and A. Shamim, 《面向 5G 毫米波应用的宽带可重构智能表面》，IEEE 天线与传播汇刊，第 72 卷，第 3 期，第 2399–2410 页，2024 年。[41] X. Zhao, M. Jian, Y. Chen, Y. Zhao, and L. Mu, 《面向 6G 的可重构智能表面：工程挑战与未来发展方向》，《智能与融合网络》，第 6 卷，第 1 期，第 53–81 页，2025 年。[42] Y. Li, 《深度强化学习：综述》，2018 年。[43] R. Bellman, 《动态规划》。美国：普林斯顿大学出版社，2010 年。[44] D. Silver, G. Lever, N. M. O. Heess, T. Degris, D. Wierstra, and M. A. Riedmiller, 《确定性策略梯度算法》，载于国际机器学习会议，2014 年。[45] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, 《通过误差反向传播学习表征》，《自然》，第 323 卷，第 6088 期，第 533–536 页，1986 年2月。[46]M. Abadi 等, 《TensorFlow：异构系统上的大规模机器学习》，2015 年。 [在线]。链接：https://www.tensorflow.org/[47] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, 《基于深度强化学习的连续控制》，2019 年。 [在线]。链接：https://arxiv.org/abs/1509.02971[48] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, 《通过深度强化学习实现人类水平控制》，《自然》，第 518 卷，第 7540 期，第 529–533 页，2015年2月。 [在线]。链接：https://doi.org/10.1038/nature14236[49] B. T. Polyak and A. B. Juditsky, 《通过平均加速随机逼近》，SIAM 控制与优化期刊，第 30 卷，第 4 期，第 838–855 页，1992 年。[50] S. Fujimoto, H. van Hoof, and D. Meger, 《解决演员-评论家方法中的函数逼近误差》，2018 年。 [在线]。链接：https://arxiv.org/abs/1802.09477

[51] T. Haarnoja、A. Zhou、P. Abbeel 和 S. Levine，《软演员-评论家：采用随机演员的离策略最大熵深度强化学习》，2018年。[在线]。获取地址：https://arxiv.org/abs/1801.01290[52] T. Haarnoja、A. Zhou、K. Hartikainen、G. Tucker、S. Ha、J. Tan、V. Kumar、H. Zhu、A. Gupta、P. Abbeel 和 S. Levine，《软演员-评论家算法及其应用》，2019年。[在线]。获取地址：https://arxiv.org/abs/1812.05905[53] H. Liu、X. Yuan 和 Y.-J. A. Zhang，《面向可重构智能表面辅助多用户MIMO的基于矩阵校准的级联信道估计》，IEEE通信选域期刊，第38卷，第11期，第2621–2636页，2020年。[54] N. K. Kundu 和 M. R. McKay，《可重构智能表面辅助MISO通信的信道估计：从线性最小均方误差到深度学习解决方案》，IEEE 通信学会开放期刊，第2卷，第471–487页，2021年。[55] J. An、C. Xu、L. Gan 和 L. Hanzo，《基于离散相移的RIS辅助MIMO系统低复杂度信道估计与无源波束成形》，IEEE通信汇刊，第70卷，第2期，第1245–1260页，2022年。[56] H. Guo 和 V. K. N. Lau，《智能反射面辅助多用户MISO系统的上行级联信道估计》，IEEE信号处理汇刊，第70卷，第3964–3977页，2022年。[57] A. Ali、N. Gonz´alez-Prelcic 和 R. W. Heath，《利用带外空间信息进行毫米波波束选择》，IEEE无线通信汇刊，第17卷，第2期，第1038–1052页，2018年。[58] B. Ding、H. Qian 和 J. Zhou，《深度神经网络中的激活函数及其特性》，载于《中国控制与决策会议（CCDC）论文集》，2018年，第1836–1841页。[59] M. V. Narkhede、P. P. Bartakke 和 M. S. Sutaone，《神经网络权重初始化策略综述》，人工智能评论，第55卷，第1期，第291–322页，2022年1月。[在线]。获取地址：https://doi.org/10.1007/s10462-021-10033-z[60] D. P. Kingma 和 J. Ba，《Adam：一种随机优化方法》，2017年。[在线]。获取地址：https://arxiv.org/abs/1412.6980



GAL BEN-ITZHAK（IEEE研究生学生会员）分别于2023年和2024年获得加利福尼亚大学欧文分校（UCI，美国加利福尼亚州欧文）电气工程理学学士和理学硕士学位。他目前正在UCI攻读电气工程博士学位。其当前研究与专业兴趣包括高速通信、优化、数字信号处理、高速电路设计和信号完整性。



ENDER AYANOGLU（IEEE会士）于1986年在美国加利福尼亚州斯坦福的斯坦福大学获得电气工程博士学位。他曾在美国新泽西州霍尔姆德尔的贝尔实验室通信系统研究实验室工作。1999年至2002年，他在加利福尼亚州圣何塞的思科系统公司担任系统架构师。自2002年以来，他一直担任加利福尼亚州欧文的加利福尼亚大学欧文分校电气工程与计算机科学系教授；2002年至2010年期间，他曾担任普适通信与计算中心主任及Conexant-Broadcom捐赠讲席教授。他曾获IEEE通信学会Stephen O.

他于1995年获得赖斯奖论文奖，1997年获得IEEE通信学会最佳教程论文奖，2014年获得IEEE通信学会通信理论技术委员会杰出服务奖。2000年至2001年，他担任IEEE-ISTO宽带无线互联网论坛的创始主席；该论坛是一个行业标准组织。1990年至2002年，他在IEEE通信学会通信理论委员会执行委员会任职，并于1999年至2002年担任该委员会主席。1993年至2014年，他担任《IEEE通信汇刊》编辑。2004年至2008年，他担任《IEEE通信汇刊》主编；2014年至2016年，他担任《IEEE通信选域期刊》绿色通信与网络系列主编。2016年至2020年，他担任《IEEE绿色通信与网络汇刊》的创刊主编。2022—2023年和2024—2025年，他担任IEEE通信学会杰出讲师。2023年，他因作为编辑、主编和创刊主编为IEEE通信学会期刊作出的杰出贡献，获得IEEE通信学会Joseph L. LoCicero奖。