基于学习的集成传感和通信系统资源管理

Ziyang Lu, M. Cenk Gursoy, Chilukuri K. Mohan, Pramod K. Varshney
Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse NY, 13066
{zlu112, mcgursoy, ckmohan, varshney}@syr.edu

摘要— 在本文中，我们解决了集成传感和通信系统（配备雷达和通信单元）中的自适应时间分配任务。双功能雷达-通信系统的任务包括为跟踪多个目标分配驻留时间，并利用剩余时间为估计的目标位置传输数据。我们介绍了一种新型的约束深度强化学习（CDRL）方法，旨在在时间预算限制下优化跟踪和通信之间的资源分配，从而提高目标通信质量。我们的数值结果证明了所提出的 CDRL 框架的有效性，证实了它能够在高度动态的环境中最大化通信质量的同时遵守时间约束。

I. 介绍

A. 背景

1) 认知雷达： 雷达技术，在环境感知、空间探索、导航和交通控制等各种应用中至关重要，随着自动驾驶汽车和无人机的出现而变得越来越重要。因此，在具有挑战性的环境中有效分配雷达资源已经成为研究的重点。传统的资源配置优化方法在 [1] 和 [2] 中进行了讨论，提供了问题的基础解决方案。

文献还探讨了雷达资源管理的博弈论方法。[3] 中的作者解决了多雷达系统中的功率分配问题，将其视为一个非合作博弈，并分析了纳什均衡及其收敛性。这种方法突出了雷达系统中的战略互动及其对资源管理的影响。

认知雷达，一个位于雷达技术和人工智能交叉领域的领域，正在迅速发展。特别是，认知雷达利用机器学习、博弈论和认知无线电技术来提高在动态和非平稳环境中的雷达性能。雷达技术的这一方面在 [4] 和 [5] 中进行了深入探讨，展示了认知雷达系统如何使用人工智能进行适应并作出知情决策。这些系统设计为能够持续更新对环境的理解，从而更有效地应对变化的条件。

[4] 的作者提供了对认知雷达系统的全面概述，重点介绍了信号处理、动态反馈和信息保存等方面。这一领域对于开发能够执行多目标跟踪和电子波束控制等功能的多功能雷达系统至关重要。文献强调了人工智能和认知技术对雷达技术的变革性影响，为能够在日益复杂的环境中操作的更加自适应和智能的系统铺平了道路。

2) 集成传感和通信（ISAC）： ISAC 代表无线网络中的范式转变，在单一平台上集成了感知和通信功能。这种集成在 5G 及以后的时代至关重要，为各种应用提供了增强的效率和能力。随着毫米波和大规模 MIMO 技术的出现，传感和通信技术的融合已经导致了 ISAC 研究和应用的重要进展，如 [6] 所示。

ISAC 也被预期在 6G 无线网络的演进中扮演关键角色。未来网络中的密集小区基础设施提供了一个独特的机会，构建感知网络，将传感直接集成到通信过程中。在未来蜂窝系统、WLAN 和 V2X 网络中引入 ISAC 预计将带来显著的整合与协调增益 [7]。

认知雷达是一个根据环境调整其操作参数的系统，并且与 ISAC 的原则紧密相关。它涉及智能决策和自适应感知，这些都是 ISAC 系统的关键组成部分。在 ISAC 的背景下，认知雷达可以从通信和传感资源的共享中获益，从而实现更高效、更响应迅速的系统，因此激发了本文的研究工作。

B. 相关工作

近期的研究工作集中在雷达系统的资源管理上。例如，研究 [8] 通过扩展卡尔曼滤波方法解决了在部分可观测马尔可夫决策过程和策略展开技术的背景下追踪多个目标的时间分配问题。另一条研究路线 [9] 利用模型预测控制（MPC）来为雷达系统分配时间。仿真结果显示，这两种策略展开技术和 MPC 都能有效地优化重访间隔和驻留时间，从而减少了估计方差。这些方法作为离线方法，其决策基于已有的知识，如关于测量噪声和机动性噪声的统计数据。然而，由于雷达环境的非平稳特性，这些离线方法可能难以应对快速变化的环境。

为应对此类动态场景，能够适应环境变化的在线策略较为理想。深度强化学习（DRL）在训练代理进行复杂决策任务方面获得了显著成就，在 [10] 中展示了令人印象深刻的结果。DRL 无需模型的特点及其固有特性使其适合于动态雷达应用挑战。事实上，DRL 已被越来越多地应用于与雷达相关的决策过程。例如，[11] 中的作者使用 DRL 进行多目标检测的频谱分配，证明了其提高了检测性能并减少了与其他无线系统的干扰。在 [12] 和 [13] 中还探讨了雷达环境下的其他 DRL 应用。

C. 贡献

本工作解决了 ISAC 系统中的时间分配问题，旨在优化跟踪与通信之间的平衡，以提高与目标的整体通信质量。我们的主要贡献如下：

- *时间分配问题的形式化作为约束优化问题*：这涉及为 ISAC 系统开发一种策略，以平衡跟踪和通信阶段，从而最大化目标通信质量。
- *设计一个约束深度强化学习（CDRL）框架来解决优化问题*：我们详细介绍了同时学习深度 Q 网络（DQN）参数和对偶变量的方法。
- *数值分析证明了 CDRL 框架的有效性*：结果验证了我们的方法成功地学习了一种有效的时长分配策略，同时遵守了可用时间预算的限制。

II. 雷达跟踪模型

A. 目标运动模型

在时间槽 t，目标的状态被捕获为 $\boldsymbol{x_t} = [x_t, y_t, \dot{x}_t, \dot{y}_t]^T$，其中 (x_t, y_t) 表示目标的当前位置，而 ($\dot{x}_t$ 和 $\dot{y}_t$) 分别表示其水平和垂直速度。在重访期间的恒定速度模型下，目标状态从 $\boldsymbol{x_t}$ 过渡到

$$\boldsymbol{x_{t+1}} = \boldsymbol{F_t x_t} + \boldsymbol{w_t}, \tag{1}$$

本工作部分得到了国家科学基金会 CNS-2221875 拨款的支持。

，其中 $\boldsymbol{F_t} \in \mathbb{R}^{4\times4}$ 是状态转移矩阵，表示为

$$\boldsymbol{F_t} = \begin{bmatrix} 1 & 0 & T_t & 0 \\ 0 & 1 & 0 & T_t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

，其中 T_t 是在时间 t 时雷达系统的跟踪时间间隔。这里，$\boldsymbol{w_t}$ 表示机动噪声，建模为一个多元零均值高斯噪声，协方差矩阵为

$$\boldsymbol{Q_t} = \begin{bmatrix} T_t^4/4 & 0 & T_t^3/2 & 0 \\ 0 & T_t^4/4 & 0 & T_t^3/2 \\ T_t^3/2 & 0 & T_t^2 & 0 \\ 0 & T_t^3/2 & 0 & T_t^2 \end{bmatrix} \sigma_w^2 \tag{3}$$

其中，σ_w^2 表示时间 t 处的机动噪声方差。

B. 测量模型

雷达系统获取目标的距离 r 和方位角 θ 的测量值以进行位置估计。测量向量，用 $\boldsymbol{z_t}$ 表示，以及从 $\boldsymbol{x_t}$ 到 $\boldsymbol{z_t}$ 的非线性映射函数，用 $h(\cdot)$ 表示，建立了状态和测量向量之间的关系如下：

$$\boldsymbol{z_t} = h(\boldsymbol{x_t}) + \boldsymbol{v_t} = \left[\sqrt{x_t^2 + y_t^2}, \quad \tan^{-1}\left(\frac{y_t}{x_t}\right)\right]^T + \boldsymbol{v_t}, \tag{4}$$

其中，$\boldsymbol{v_t}$ 为时间 t 处的测量噪声向量，包括距离测量噪声 $v_{r,t}$ 和角度测量噪声 $v_{\theta,t}$，均被建模为零均值高斯噪声，其方差分别为 $\sigma_{r,t}^2$ 和 $\sigma_{\theta,t}^2$。雷达的假设位置在笛卡尔坐标系的原点。

测量噪声方差与目标反射雷达信号在时间 t 的信噪比 SNR_t 动态相关。SNR_t 受雷达驻留时间 τ_t、目标-雷达距离 r_t 影响，如公式 [8] 和 [14] 所示，以及波束对准损失 L_{bm}^T：

$$\mathrm{SNR}_t(\tau_t, r_t, \Theta, \hat{\Theta}) = \mathrm{SNR}_0 \left(\frac{\tau_t}{\tau_0}\right) \left(\frac{r_t}{r_0}\right)^{-4} L_{bm}^T \tag{5}$$

其中，SNR_0、τ_0 和 r_0 是信噪比、驻留时间和目标到雷达距离的参考值。L_{bm}^T 表示由于雷达波束方向与追踪目标的真实方位角不一致而造成的功率损失。在这项工作中，我们用方位上的余弦响应来建模雷达波束的天线图样。例如，如果我们把目标估计的方位角记作 $\hat{\Theta}$，把目标的真实方位角记作 Θ。然后跟踪光束对准损失 L_{bm}^T 可以表示为

$$L_{bm}^T = \begin{cases} \cos^j(|\Theta - \hat{\Theta}|) & \text{if } |\Theta - \hat{\Theta}| \le \frac{\pi}{2}, \\ 0 & \text{if } |\Theta - \hat{\Theta}| > \frac{\pi}{2} \end{cases} \tag{6}$$

其中雷达波束的天线图样由 j 确定。较大的 j 对应于较窄的波束图样。

然后，测量噪声的方差与信噪比之间的关系可以确定为 [15]

$$\sigma_{\bullet,t}^2 = \frac{\sigma_{\bullet,0}^2}{\mathrm{SNR}_t(\tau_t, r_t, \Theta, \hat{\Theta})} \tag{7}$$

其中 $\bullet \in (r, \theta)$. $\sigma_{\bullet,0}^2$ 表示相应的测量噪声方差的参考值。值得注意的是，根据公式 (5) 和 (7)，当分配给目标的驻留时间更长或目标向雷达靠近时，测量噪声的方差会减小。

请注意，测量与状态之间的映射函数 $h(\cdot)$ 是非线性的，因此本工作中采用了扩展卡尔曼滤波器（EKF）。在使用 EKF 时，引入了一个观测矩阵 $\boldsymbol{H_t} \in \mathbb{R}^{2\times4}$ 来线性化 $\boldsymbol{z_t}$ 和 $\boldsymbol{x_t}$ 之间的关系。$\boldsymbol{H_t}$ 被定义为测量函数 $h(\cdot)$ 的雅可比矩阵：

$$\boldsymbol{H_t} = \frac{\partial h(\cdot)}{\partial \boldsymbol{x}}|_{\boldsymbol{x_t}} = \begin{bmatrix} \frac{x_t}{\sqrt{x_t^2+y_t^2}} & \frac{y_t}{\sqrt{x_t^2+y_t^2}} & 0 & 0 \\ \frac{-y_t}{x_t^2+y_t^2} & \frac{x_t}{x_t^2+y_t^2} & 0 & 0 \end{bmatrix}. \tag{8}$$

考虑独立测量，测量的协方差矩阵由

$$\boldsymbol{R_t} = \begin{bmatrix} \sigma_{r,t}^2 & 0 \\ 0 & \sigma_{\theta,t} \end{bmatrix}. \tag{9}$$

给出

C. 扩展卡尔曼滤波器

卡尔曼滤波器是一种著名的递归算法，用于估计过程 [16] 的状态，在扩展卡尔曼滤波器（EKF）中得到了进一步的应用。EKF 擅长处理非线性测量场景，如雷达目标跟踪，其操作主要分为两个阶段——预测和更新。

1) 预测阶段*:* 在此阶段，EKF 使用以下内容预测目标的未来状态：

$$\hat{\boldsymbol{x}}_{\boldsymbol{t|t-1}} = \boldsymbol{F_t}\boldsymbol{x}_{\boldsymbol{t-1|t-1}}, \tag{10}$$

$$\hat{\boldsymbol{P}}_{\boldsymbol{t|t-1}} = \boldsymbol{F_t}\boldsymbol{P}_{\boldsymbol{t-1|t-1}}\boldsymbol{F_t}^T + \boldsymbol{Q_t}, \tag{11}$$

其中 $\hat{\boldsymbol{x}}_{\boldsymbol{t|t-1}}$ 和 $\hat{\boldsymbol{P}}_{\boldsymbol{t|t-1}}$ 是预测的状态及其协方差。

2) 更新阶段*:* 在此，EKF 根据新的测量值改进其预测：

$$\boldsymbol{K_t} = \hat{\boldsymbol{P}}_{\boldsymbol{t|t-1}}\boldsymbol{H_t}^T(\boldsymbol{H_t}\hat{\boldsymbol{P}}_{\boldsymbol{t|t-1}}\boldsymbol{H_t}^T + \boldsymbol{R_t})^{-1}, \tag{12}$$

$$\boldsymbol{x_{t|t}} = \hat{\boldsymbol{x}}_{\boldsymbol{t|t-1}} + \boldsymbol{K_t}(\boldsymbol{z_t} - h(\hat{\boldsymbol{x}}_{\boldsymbol{t|t-1}})), \tag{13}$$

$$\boldsymbol{P_{t|t}} = (\boldsymbol{I} - \boldsymbol{K_t}\boldsymbol{H_t})\hat{\boldsymbol{P}}_{\boldsymbol{t|t-1}}, \tag{14}$$

最终得到更新的状态 $\boldsymbol{x_{t|t}}$ 和协方差 $\boldsymbol{P_{t|t}}$。

EKF 初始化 $\boldsymbol{x_{t|t}}$ 和 $\boldsymbol{P_{t|t}}$ 分别为零向量和单位矩阵，迭代以最小化 $\boldsymbol{P_{t|t}}$ 的迹。

这一增强的描述提供了对 EKF 机制更清晰的理解，特别是在雷达目标跟踪应用中。

III. 问题表述

A. 约束马尔可夫决策过程（*CMDP*）

如 [17] 所定义的约束马尔可夫决策过程（CMDP）由元组 $(S, A, C, \Theta, T, \mu, \gamma)$ 表征。它由状态 S、动作 A、一个成本函数 $C : S \times A \times S \to \mathbb{R}$、一个预算函数 $\Theta : S \times A \times S \to \mathbb{R}$ 和一个转移概率 $T : S \times A \times S \to [0, 1]$ 组成。初始状态分布表示为 μ，$\gamma \in [0, 1)$ 表示未来奖励和成本的折扣因子。目标是制定一个最优策略 $\pi : S \times A \to [0, 1]$，以最小化未来成本的折现和（或最

大化奖励的折现和），同时遵守预算约束。数学上，这表示为

$$\begin{aligned} \min_{\pi} \quad & \sum_{m=0}^{\infty} \gamma^m c_{t+m} \\ \text{s.t.} \quad & \sum_{m=0}^{\infty} \gamma^m (\Theta_{t+m} - \Theta_{max}) \leq 0. \end{aligned} \tag{15}$$

B. 时间分配在集成传感和通信中

ISAC 系统将诸如雷达跟踪等传感功能与通信能力在统一框架内协同工作。本研究专注于 ISAC 系统的时隙分配问题，旨在优化传感和通信效率的综合指标。这包括将总的雷达重访时间，记为 T_0，划分为目标跟踪的驻留时间 $\{\tau_t^n\}_{n=1}^N$，以及目标通信的时间 $\tau_t^c = T_0 - \sum_{n=1}^N \tau_t^n$。

ISAC 操作分为两个阶段：目标跟踪和通信。在跟踪阶段，雷达使用先前的目标位置估计值，向目标的估计方位角发射波束，并捕获嵌入噪声测量值的回波信号。每个目标 n 分配一个特定的驻留时间 τ_t^n。然后雷达采用扩展卡尔曼滤波器来预测目标的后续位置。

在通信阶段，雷达系统首先根据前一帧的目标位置估计确定其通信波束的方向。考虑波束对准损失，记为 L_{bm}^C，以及路径损耗，当前时间帧内所有目标的总数据传输速率量化如下：

$$R(\{\tau_t^n\}_{n=1}^N) = \tau_t^c B \sum_{n=1}^{N} \log_2 \left(1 + \frac{P_t L L_{bm}^C}{\sigma^2}\right) \tag{16}$$

其中，B 表示带宽，P_t 是发射功率，L 是路径损耗，定义为：

$$L = \left(\frac{d_0}{d}\right)^{\eta/2} \tag{17}$$

其中，d_0 是一个参考距离，d 是从雷达到目标的实际距离，η 是路径损耗因子。

类似于在 (6) 中定义的波束失配损耗，通信波束失配损耗 L_{bm}^C 被定义为

$$L_{bm}^C = \begin{cases} \cos^i(|\Theta - \hat{\Theta}|) & \text{if } |\Theta - \hat{\Theta}| \leq \frac{\pi}{2} \\ 0 & \text{if } |\Theta - \hat{\Theta}| > \frac{\pi}{2} \end{cases}. \tag{18}$$

目标是找到一个时间分配策略 π，能够有效地在跟踪和通信之间分配可用时间，以最大化接收到的目标的总速率。雷达必须分配足够的时间来准确估计目标位置，从而提高 L_{bm}^T 和 L_{bm}^C。同时，应该为通信分配充足的时间 τ_t^c，因为目标函数 R 与 τ_t^c 成线性比例。

根据 [17] 和 [18]，本文考虑的问题可以形式化为以下约束优化问题：

$$\begin{aligned} \underset{\pi}{\text{maximize}} \quad & \sum_{m=0}^{\infty} \gamma^m R(\{\tau_{t+m}^n\}_{n=1}^N) \\ \text{subject to} \quad & \sum_{m=0}^{\infty} \gamma^m \left(\sum_{n=1}^{N} \tau_{t+m}^n - T_0\right) \leq 0. \end{aligned} \tag{19}$$

利用拉格朗日松弛法，可以在目标函数中整合在 (19) 中指定的预算约束。这是通过引入一个非负对偶变量实现的，该变量表示为 λ_t。因此，问题 (19) 被转化为一个无约束优化问题，具体如下：

$$\min_{\lambda_t \geq 0} \max_{\pi} \sum_{m=0}^{\infty} \gamma^m \left[R(\{\tau_{t+m}^n\}_{n=1}^N) - \lambda_t \left(\sum_{n=1}^{N} \tau_{t+m}^n - T_0\right)\right]. \tag{20}$$

在 (20) 中表达的优化问题作为在 (19) 中概述的主要问题的对偶问题。在凸优化场景中，这两个问题之间的对偶间隙消失。重要的是要认识到 λ_t 随时间动态变化，随意设置其值可能会导致次优解。接下来的部分介绍了一种约束深度强化学习方法来有效解决这一优化挑战。

IV. 约束深度强化学习

在这项工作中，我们提出了一种约束深度强化学习（CDRL）框架来解决 ISAC 系统中多目标跟踪和通信的时间分配问题。目标是在总时间预算限制在某一阈值以下的情况下最大化基站与目标之间的总体通信质量。我们利用深度 Q 学习（DQL）来实现这一目标。

提出的框架中包含一个单独的 DQN 和 N 个跟踪任务。在每个时间槽 t，每个任务 n 需要决定其用于跟踪目标 n 的驻留时间 τ_t^n，使用共同的 DQN。确定驻留时间 $\{\tau_t^n\}_{n=1}^N$ 后，计算通信时间 τ_t^c，并根据 (16) 计算通信的总速率 R。

A. 状态

在此研究中，任务 n 在时间槽 t 的状态，表示为 s_t^n，定义如下：

$$s_t^n = [\{\sigma^2_{\theta_{t-1}^n}\}_{n=1}^N, \{\tau_{t-1}^n\}_{n=1}^N, \lambda_{t-1}]. \tag{21}$$

这里，$\{\sigma^2_{\theta_{t-1}^n}\}_{n=1}^N$ 表示从前一个时间槽中目标估计方位角的方差。需要注意的是，这些值并非直接提供给代理；而是通过蒙特卡洛方法推导出来的。这涉及根据扩展卡尔曼滤波器中的协方差矩阵生成大量的假设目标位置 (x, y)，该矩阵是每个时间槽中 $P_{t|t}$ 的一部分。通过对这些位置计算方位角 θ 并分析这些值的分布，我们得到方差 σ_θ^2。在训练阶段，目标的实际 x 和 y 值取为它们的均值，而在测试阶段，则使用估计的 x 和 y。

第二个组件，$\{\tau_{t-1}^n\}_{n=1}^N$，是一个表示上一时隙中选定的停留时间的数组。最终项，λ_{t-1}，表示来自上一时隙的对偶变量。状态向量的总大小为 $2N+1$，包含这些元素。

B. 动作

动作 a_t^n 是为在时间槽 t 中跟踪目标 n 所选的驻留时间。每个任务 n 可以选择一个驻留时间 $\frac{\tau_t^n}{T_0} \in [0, 1]$。我们将范围量化为十个等级，因此 $a_t^n = \frac{\tau_t^n}{T_0} \in \{0, 0.1, ..., 1\}$。动作由 ϵ-贪婪方法选择。动作要么根据 DQN 的输出以概率 $(1-\epsilon)$ 确定，要么以概率 ϵ 从动作空间中随机采样。

C. 奖励

所有任务联合最大化时间段 t 内的全局奖励 r_t。并且 r_t 被定义为

$$r_t = R(\{\tau_t^n\}_{n=1}^N) - \lambda_t \left(\sum_{n=1}^{N} \tau_t^n - T_0 \right). \quad (22)$$

r_t 中的第一项表示 CDRL 算法旨在最大化的目标函数，第二项则处理约束条件。

我们观察到，基于瞬时通信质量 $R(\{\tau_t^n\}_{n=1}^N)$ 的奖励函数由于通信过程中的内在随机性和噪声而表现出不稳定性。为了解决这一挑战，我们在奖励函数中实施了关键修改。我们没有使用最初在 (18) 中定义的瞬时光束对准角度 $|\Theta - \hat{\Theta}|$，而是转向采用方位角的标准差 σ_θ。因此，通信质量 $R(\{\tau_t^n\}_{n=1}^N)$ 按照 (16) 进行重新计算。

对奖励函数的这一修改非常重要。通过依赖方位角的标准差，奖励指标变得更加稳健，能够应对过程中固有的波动和不确定性。这种变化增强了奖励函数的稳定性，使其成为衡量所选动作在通信性能方面有效性的一个更可靠的指示器。它使我们能够更一致且可靠地评估动作对通信质量的影响，这对于我们的框架的成功至关重要。

D. CDRL 的更新

Algorithm 1 CDRL 算法

1: Initialize the DQN parameters Φ_0^π with random values.
2: Initialize states $\{\boldsymbol{s_0^n}\}_{n=1}^N$ as zero vectors and λ_t as λ_0.
3: **for** time slot $t = 0, 1, ..., T_{max}$ **do**
4: **for** each agent $n = 1, 2, ..., N$ **do**
5: Select an action a_t^n based on the current state $\boldsymbol{s_t^n}$ with the DQN $\Phi_t^{\pi,n}$ and ϵ-greedy method.
6: **end for**
7: Compute reward r_t according to (22).
8: **for** each agent $n = 1, 2, ..., N$ **do**
9: Store the experience $(\boldsymbol{s_t^n}, a_t^n, r_t, \boldsymbol{s_{t+1}^n})$ to the DQN experience buffer.
10: Update Φ_t^π to Φ_{t+1}^π with experience replay and back-propagation.
11: **end for**
12: $\lambda_{t+1} = \max(0,\ \lambda_t - \alpha(\sum_{n=1}^N \tau_t^n - T_0))$.
13: **end for**

提出的 CDRL 算法如上所述的算法 1中所示。我们同时更新 DQN 参数 Φ_t^π 和对偶变量 λ_t。

所提出的 CDRL 算法的目标是找到问题 (20) 的解决方案。通过将奖励函数设置为 (22)，DQN 将学习最大化折扣奖励 r_t，即

$$\max_\pi \sum_{m=0}^{\infty} \gamma^m \left[R(\{\tau_{t+m}^n\}_{n=1}^N) - \lambda_t \left(\sum_{n=1}^{N} \tau_{t+m}^n - T_0 \right) \right]. \quad (23)$$

我们将 (23) 中的目标函数表示为 $\mathcal{L}$。然后，对偶变量 λ_t 通过最小化 $\mathcal{L}$ 关于 λ_t 进行更新，即

$$\begin{aligned} \lambda_{t+1} &= \max(0,\ \lambda_t - \alpha \bigtriangledown_{\lambda_t} \mathcal{L}) \\ &= \max\left(0,\ \lambda_t + \alpha \sum_{m=0}^{\infty} \gamma^m \left(\sum_{n=1}^{N} \tau_{t+m}^n - T_0 \right)\right) \end{aligned} \quad (24)$$

其中 α 是对偶变量的学习率。梯度 $\bigtriangledown_{\lambda_t} \mathcal{L}$ 可以通过一个额外的神经网络进行估计，但我们简化使用的是

$$\lambda_{t+1} = \max\left(0,\ \lambda_t + \alpha \left(\sum_{n=1}^{N} \tau_t^n - T_0 \right)\right). \quad (25)$$

E. 收敛性分析

所提出的 CDRL 算法迭代地确定 DQN 参数和对偶变量（Φ_t，λ_t）。基于在 [18] 和 [19] 的第六章中呈现的理论基础，这种迭代方法被识别为多时间尺度随机逼近过程，最终收敛于一个稳定点（Φ_t^*, λ_t^*）。

V. 数值结果

表 I
模拟参数

$\sigma_{r,0}^2$ (m^2)	10
$\sigma_{\theta,0}^2$ (rad^2)	1e-4
σ_w $((m/s^2)^2)$	5
Reference distance r_0 (m)	800
Reference dwell time τ_0 (s)	2
Revisit interval T_0 (s)	3
Transmit Power P_t (W)	1
Noise Level σ	0.1
Reference Distance in Path Loss d_0 (m)	500
Bandwidth B (Hz)	500
DRL discount factor γ	0.9
DRL mini-batch size	32
Replay memory buffer size	50000
Exploring probability ϵ	0.1
Initial dual variable (λ_0)	100
Step size of dual variable (α)	10

A. 实验设置

本节概述了实验框架，包括环境和 CDRL 算法所使用的具体超参数。这些参数在表 I中详细列出。我们在研究中采用的 DQN 架构设计了两个隐藏层，每个层包含 64 个神经元。为了促进层间有效的神经传输和非线性建模，采用了 ReLU（修正线性单元）激活函数。

我们的实验模型假设一个最多存在四个目标的场景。在 CDRL 算法的训练阶段，这些目标被随机生成的位置和时间槽引入系统中，从而模拟了一个动态且不可预测的环境，这与实际情况非常接近。

B. 测试结果

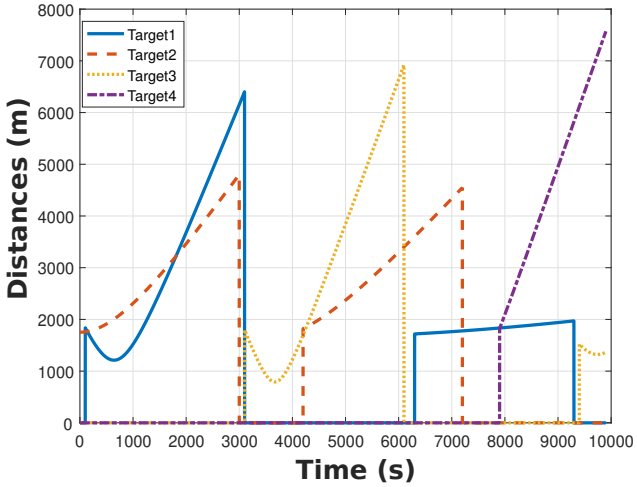

图 1. 雷达的目标距离

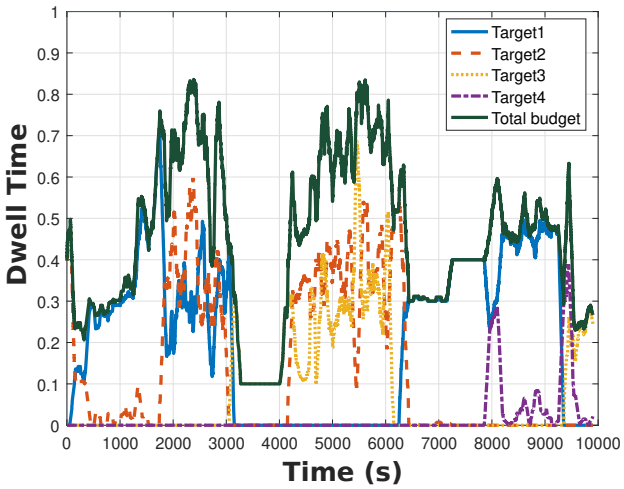

图 2. 驻留时间分配策略由 CDRL 学习得到

测试阶段的目标移动模式如图 1所示，具体展示了我们设计的一个测试场景中目标之间的不同距离。为了增强测试环境的复杂性和多样性，遵循了一项特定协议：任何在环境中停留超过三千个时隙的目标都会被系统移除。这种方法确保了测试环境持续变化，挑战算法适应频繁变化条件的能力，从而对其性能进行全面评估。

图 2展示了由我们提出的框架学习到的驻留时间分配策略。从该图中可以注意到，目标距离与跟踪时长分配之间存在相关性。具体而言，当所有目标距离雷达更远时，框架倾向于分配更高的比例的时间用于跟踪。这

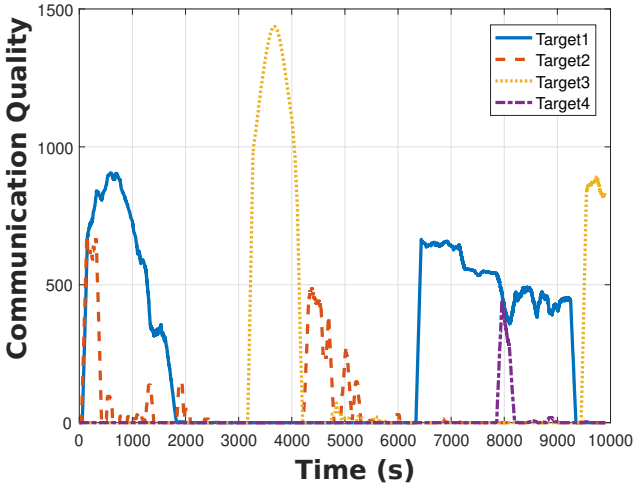

图 3. 通过 CDRL 实现的通信质量

一现象在区间 [2000, 3000] 和 [6000, 6500] 尤为明显。这种策略背后的原理是直观的：由于目标距离增加导致跟踪精度变得至关重要，必须优先考虑跟踪而非通信，以避免因跟踪误差而失去通信质量的风险。

反之，如图 2所示，在区间 [1000, 1200]、[3700, 4300] 和 [6700, 7000] 内，当目标靠近雷达时，我们观察到策略发生了转变。在这些情况下，算法通过分配更多时间给通信阶段来适应变化。这种战略转变基于增强的通信时间能够直接且线性地提高数据传输总速率的理解。在这种情形下，进一步增加跟踪的时间对于改善通信质量带来的回报逐渐减少。这种自适应的时间分配展示了该框架根据实时动态环境灵活平衡跟踪和通信所分配时间的能力。图 3展示了在此次测试案例中 CDRL 实现的每个目标的个体通信速率。

我们提出的算法与各种基准策略的和速率性能比较分析详见表 II。例如，标记为“固定 0.1”的策略是指一个固定的驻留时间分配协议。在此策略中，每当环境中存在目标时，都会为其分配一致的驻留时间 $0.1T_0$ 以对其进行跟踪。随后，在将指定的跟踪时间分配给每个检测到的目标后，该时间段内剩余的时间专用于通信阶段。比较结果表明，提出的 CDRL 算法在通信效率方面始终优于预设的固定分配方案。值得注意的是，“固定 0.2”策略在此特定测试场景中表现出几乎与 CDRL 相当的性能指标，但重要的是要认识到这种固定方案固有的局限性。具体来说，“固定 0.2”策略缺乏优化动态和不可预测环境中的性能所需的关键适应能力，而这是我们的 CDRL 算法内在的优势。

	平均和速率	**百分比**
CDRL	466.37	100%
Fixed 0.1	345.94	74.18%
Fixed 0.2	449.42	96.37%
Fixed 0.3	383.02	82.13%

表 II
和速率比较

VI. 结论

本研究介绍了一种专为 ISAC 系统高效时间分配设计的新型 CDRL 框架。该 CDRL 框架创新性地集成了神经网络参数和对偶变量的同时更新，以在预定义的预算约束内找到优化通信性能的战略平衡。我们的数值结果证明了 CDRL 算法具备智能学习并实施均衡时间分配策略的能力。这种策略能够熟练处理跟踪精度与通信效率之间的权衡。我们发现的关键亮点是，所提出的 CDRL 框架相较于几种任意设计的固定分配策略具有更优的总和速率性能。这种增强的性能突显了 CDRL 在适应动态环境条件以及使资源分配有利于最大化通信吞吐量方面的高效性。

参考文献

[1] A. Orman, C. N. Potts, A. Shahani, and A. Moore, "Scheduling for a multifunction phased array radar system," *European Journal of Operational Research*, vol. 90, no. 1, pp. 13–25, 1996.

[2] J. Butler, A. Moore, and H. Griffiths, "Resource management for a rotating multi-function radar," in *Radar 97 (Conf. Publ. No. 449)*. IET, 1997, pp. 568–572.

[3] A. Deligiannis, A. Panoui, S. Lambotharan, and J. A. Chambers, "Game-theoretic power allocation and the nash equilibrium analysis for a multistatic MIMO radar network," *IEEE Transactions on Signal Processing*, vol. 65, no. 24, pp. 6397–6408, 2017.

[4] A. Charlish, F. Hoffmann, C. Degen, and I. Schlangen, "The development from adaptive to cognitive radar resource management," *IEEE Aerospace and Electronic Systems Magazine*, vol. 35, no. 6, pp. 8–19, 2020.

[5] S. Haykin, "Cognitive radar: a way of the future," *IEEE Signal Processing Magazine*, vol. 23, no. 1, pp. 30–40, 2006.

[6] A. Liu, Z. Huang, M. Li, Y. Wan, W. Li, T. X. Han, C. Liu, R. Du, D. K. P. Tan, J. Lu *et al.*, "A survey on fundamental limits of integrated sensing and communication," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 994–1034, 2022.

[7] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728–1767, 2022.

[8] M. Schöpe, H. Driessen, and A. Yarovoy, "A constrained POMDP formulation and algorithmic solution for radar resource management in multi-target tracking," *ISIF Journal of Advances in Information Fusion*, vol. 16, no. 1, p. 31, 2021.

[9] T. de Boer, M. I. Schöpe, and H. Driessen, "Radar resource management for multi-target tracking using model predictive control," in *2021 IEEE 24th International Conference on Information Fusion (FUSION)*. IEEE, 2021, pp. 1–8.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[11] C. E. Thornton, M. A. Kozy, R. M. Buehrer, A. F. Martone, and K. D. Sherbondy, "Deep reinforcement learning control for radar detection and tracking in congested spectral environments," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 4, pp. 1335–1349, 2020.

[12] E. Selvi, R. M. Buehrer, A. Martone, and K. Sherbondy, "Reinforcement learning for adaptable bandwidth tracking radars," *IEEE Transactions on Aerospace Electronic Systems*, vol. 56, no. 5, pp. 3904–3921, 2020.

[13] F. Meng, K. Tian, and C. Wu, "Deep reinforcement learning-based radar network target assignment," *IEEE Sensors Journal*, vol. 21, no. 14, pp. 16 315–16 327, 2021.

[14] W. Koch, "Adaptive parameter control for phased-array tracking," in *Signal and Data Processing of Small Targets 1999*, vol. 3809. SPIE, 1999, pp. 444–455.

[15] H. Meikle, *Modern radar systems*. Artech House, 2008.

[16] G. Welch, G. Bishop *et al.*, "An introduction to the Kalman filter," 1995.

[17] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999, vol. 7.

[18] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *arXiv preprint arXiv:1805.11074*, 2018.

[19] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.