一种强化学习驱动的集中式MIMO雷达功率分配算法

黄洁瑜 谢军伟 张浩为* 冯为可 韩卫航
(空军工程大学防空反导学院 西安 710051)

摘 要：集中式多输入多输出(Multiple-input Multiple-output, MIMO)雷达资源分配算法可有效提升多目标跟踪精度，而传统优化算法仅优化下一时刻跟踪性能，难以实现全时域多目标累积跟踪精度提升；同时，传统优化算法的计算复杂度较高，难以满足资源分配过程的实时性要求。针对该问题，本文提出一种强化学习(Reinforcement Learning, RL)驱动的集中式 MIMO 雷达功率分配算法。首先，利用后验克拉美罗界(Posterior Cramér-Rao Lower Bound, PCRLB)构建RL模型的状态与奖励函数，将功率资源分配过程建模为马尔可夫决策过程(Markov Decision Process, MDP)，建立RL驱动的资源分配优化模型；其次，结合工程实际与算法收敛性要求，采用对决双深度 Q 网络(Dueling Double Deep Q-Network, D3QN)算法对模型进行求解。仿真结果表明，相较于固定功率分配与传统优化分配算法，所提算法可实现功率资源的自适应合理分配，显著提升全时域累积跟踪精度；并且训练后的网络可以根据状态进行实时决策输出资源分配结果，提升了算法实时性。

关键词：强化学习；集中式MIMO雷达；功率资源分配；对决双深度 Q 网络算法
中图分类号：TN959.6 **文献标识码**：A **文章编号**：1009-5896(2026)00-0001-10
DOI: 10.11999/JEIT260695 **CSTR**: 32379.14.JEIT260695

1 引言

随着战场环境日趋复杂，多目标跟踪精度面临更高要求。受限于雷达有限资源，传统固定分配策略难以在资源约束下实现系统性能最优化。动态优化分配雷达资源已成为破解资源与性能矛盾、提升作战效能的关键途径。集中式多输入多输出(Multiple-input Multiple-output, MIMO)雷达凭借正交波形发射机制，可实现波形与空间分集，灵活调控各通道发射功率与波束指向，在检测精度、参数估计及多目标跟踪容量等方面显著优于传统相控阵雷达，为精细化、自适应资源分配奠定了硬件基础[1]。

针对上述问题，国内外学者开展了广泛研究，现有研究主要沿两条路径展开。一是在满足跟踪性能约束下最小化资源消耗[2–4]。文献[2]以发射功耗最低为优化目标，在保证跟踪精度的前提下构建约束优化模型，能够显著降低雷达系统的能量消耗。文献[3]面向低截获作战场景，设计一种快速功率分配优化方法，通过构建机动目标跟踪精度与截获概率联合模型，同步改善多目标跟踪能力与雷达低截获水平。文献[4]针对组网雷达射频隐身需求，优化射频辐射资源分配方案，进一步强化组网雷达的隐身特性。二是在资源预算约束下最大化跟踪性能[5–7]。文献[5]依托集中式MIMO雷达组网架构，结合目标雷达散射截面积特征变化规律优化资源分配逻辑，改善隐身目标的跟踪精度。文献[6]依托服务质量约束框架构建鲁棒功率分配模型，可适配不同功率条件并优化跟踪效果。文献[7]设计了动态天线配置优化方案，该方案采用凸松弛联合局部搜索的混合求解策略处理NP难优化问题，能够在降低求解复杂度的同时提升多目标跟踪精度。然而，传统优化方法多聚焦单时刻性能寻优，缺乏全时域全局视野；且迭代求解计算复杂度高，难以满足复杂电磁环境下的实时分配需求。

近年来，强化学习(Reinforcement Learning, RL)凭借无模型依赖于长期收益优化特性，逐渐应用于雷达资源调控领域。文献[8]将强化学习引入频控阵MIMO雷达，优化发射功率分配方式，增强系统抗干扰能力。文献[9]提出数据驱动式一体化资源管理方案，依托动态干扰信息的在线感知能力，提升复杂电磁环境下多功能雷达的多目标跟踪性能。文献[10] 针对雷达系统多目标跟踪下的自适应资源管理问题，以有限资源预算约束下最小化全局跟踪代价为目标，构建基于深度 Q 网络(Deep Q-Network, DQN)的深度强化学习算法。文献[11]从多个方面梳理和总结了协同认知跟踪与资源调度算法近年来的研究进展，针对基于传统优化方法的资源分配进

收稿日期：2026-05-28；改回日期：2026-07-02；网络出版：2026-07-14
*通信作者： 张浩为 zhw_xhzf@163.com
基金项目：国家自然科学基金(62571544)，陕西省创新能力支撑计划项目(2025ZC-KJXX-81)，陕西省教育厅青年创新团队科研计划项目(24JP221)，西安市自然科学基金(26ZRKX00090)
Foundation Items: National Natural Science Foundation of China (62571544), Innovation Capability Support Program of Shaanxi (2025ZC-KJXX-81), Research Program Project of Youth Innovation Team of Shaanxi Provincial Education Department (24JP221), Natural Science Foundation of Xi'an (26ZRKX00090)

行了系统整理。但现有研究仍存在明显不足：一方面，针对集中式MIMO雷达体制的RL适配研究相对匮乏，现有模型多基于传统相控阵架构，直接迁移难以匹配模型；另一方面，现有RL模型在马尔可夫决策过程(Markov Decision Process, MDP)构建上存在理论缺陷，使得智能体极易形成极度短视的决策偏好，最终难以实现全时域累积跟踪精度提升。

为此，本文提出一种强化学习驱动的集中式MIMO雷达功率分配算法。该算法引入后验克拉美罗界(Posterior Cramér-Rao Lower Bound, PCRLB)量化跟踪误差下界，据此构建状态空间与奖励函数，将功率分配过程严格建模为MDP框架；并采用对决双深度Q网络(Dueling Double Deep Q-Network, D3QN)进行离线训练与在线实时求解。

2 系统模型

2.1 目标运动模型

假设有Q个目标在二维平面内运动，采用常速(Constant Velocity, CV)模型对目标运动进行描述，公式表示如下[2]：

$$\boldsymbol{x}_{q,k}=\boldsymbol{F}_q\boldsymbol{x}_{q,k-1}+\boldsymbol{w}_{q,k-1} \tag{1}$$

其中，$\boldsymbol{x}_{q,k}=[x_{q,k},\dot{x}_{q,k},y_{q,k},\dot{y}_{q,k}]^{\mathrm{T}}$ 表示k时刻时目标q的运动状态，$(x_{q,k},y_{q,k})$和$(\dot{x}_{q,k},\dot{y}_{q,k})$ 分别表示目标q的位置以及速度；$\boldsymbol{F}_q$为状态转移方程；$\boldsymbol{w}_{q,k-1}$表示协方差为$\boldsymbol{Q}_{q,k-1}$零均值高斯白噪声。$\boldsymbol{F}_q$和$\boldsymbol{Q}_{q,k-1}$分别表示如下：

$$\boldsymbol{F}_q=\boldsymbol{I}_2\otimes\begin{bmatrix}1 & T_{\mathrm{s}}\\ 0 & 1\end{bmatrix} \tag{2}$$

$$\boldsymbol{Q}_{q,k-1}=\mu\boldsymbol{I}_2\otimes\begin{bmatrix}T_{\mathrm{s}}^3/3 & T_{\mathrm{s}}^2/2\\ T_{\mathrm{s}}^2/2 & T_{\mathrm{s}}\end{bmatrix} \tag{3}$$

其中，μ表示过程噪声强度；$\boldsymbol{I}_2$表示二维单位矩阵；$\otimes$表示克罗内克积运算符；T_{s}表示采样时间间隔。

2.2 雷达观测模型

考虑一个集中式MIMO雷达，采用同时多波束的工作模式，即通过同时发射不同的窄波束进行多目标探测。然后在不同的通道内提取回波信息，进而得到目标量测值。k时刻从目标q反射回来的量测值可表示如下：

$$\boldsymbol{z}_{q,k}=g(\boldsymbol{x}_{q,k})+\boldsymbol{v}_{q,k} \tag{4}$$

其中，$\boldsymbol{z}_{q,k}=[R_{q,k},\theta_{q,k},f_{q,k}]^{\mathrm{T}}$表示量测信息；$g(\cdot)$表示雷达的非线性观测函数。函数$g(\boldsymbol{x}_{q,k})$表示如下：

$$\begin{cases}R_{q,k}=\sqrt{(x_{q,k}-x_{radar})^2+(y_{q,k}-y_{radar})^2}\\ \theta_{q,k}=\arctan[(y_{q,k}-y_{radar})/(x_{q,k}-x_{radar})]\\ f_{q,k}=-\dfrac{2}{\lambda R_{q,k}}[\dot{x}_{q,k}(x_{q,k}-x_{radar})\\ \qquad+\dot{y}_{q,k}(y_{q,k}-y_{radar})]\end{cases} \tag{5}$$

其中，$R_{q,k}$、$\theta_{q,k}$ 和$f_{q,k}$分别表示k时刻量测得到的目标q相对于雷达的距离、方位角和多普勒频率；$(x_{\mathrm{radar}},\ y_{\mathrm{radar}})$表示雷达的位置；$\lambda$ 表示雷达载波波长；$R_{q,k}$表示目标q与雷达之间的距离；$\boldsymbol{v}_{q,k}$表示协方差为$\boldsymbol{R}_{q,k}$零均值高斯白噪声。$\boldsymbol{R}_{q,k}$可以表示为：

$$\boldsymbol{R}_{q,k}(P_{q,k})=\mathrm{diag}(\sigma_{R_{q,k}}^2,\sigma_{\theta_{q,k}}^2,\sigma_{f_{q,k}}^2) \tag{6}$$

其中，$\mathrm{diag}(\cdot)$ 表示矩阵对角线元素；三个对角元素分别表示量测误差矩阵$\boldsymbol{R}_{q,k}$中的距离、方位角和多普勒频率误差协方差，满足以下条件：

$$\begin{cases}\sigma_{R_{q,k}}^2\propto[\alpha_{q,k}P_{q,k}|h_{q,k}|^2(\beta_{q,k}^n)^2]^{-1}\\ \sigma_{\theta_{q,k}}^2\propto[\alpha_{q,k}P_{q,k}|h_{q,k}|^2/B_{\mathrm{w}}]^{-1}\\ \sigma_{f_{q,k}}^2\propto[\alpha_{q,k}P_{q,k}|h_{q,k}|^2(T_{q,k}^n)^2]^{-1}\end{cases} \tag{7}$$

其中，$\beta_{q,k}^n$和$T_{q,k}^n$分别表示有效带宽和有效时宽；B_{w}表示半功率波束宽度；$P_{q,k}$表示雷达k时刻对目标q的发射功率；$\alpha_{q,k}\propto 1/(R_{q,k})^4$表示由路径损耗引起的传播衰减。

2.3 跟踪性能指标

传统资源分配依赖预测克拉美罗界进行优化，而强化学习采用数据驱动范式，策略优化依赖于动作执行后的真实环境反馈。PCRLB基于实际观测数据计算，能够客观表征资源分配后的真实估计性能。因此，本文中利用PCRLB作为跟踪性能的衡量标准，同时作为强化学习策略的反馈奖励。

PCRLB可以为无偏估计提供一个下界，并且在高信噪比的情况下接近实际跟踪误差。目标的状态估计满足以下不等式[3]：

$$\mathbb{E}_{\boldsymbol{x}_{q,k},\boldsymbol{z}_{q,1:\,k}}[(\breve{\boldsymbol{x}}_{q,k}-\boldsymbol{x}_{q,k})(\breve{\boldsymbol{x}}_{q,k}-\boldsymbol{x}_{q,k})^{\mathrm{T}}]\geq\boldsymbol{J}^{-1}(\boldsymbol{x}_{q,k}) \tag{8}$$

其中，$\boldsymbol{z}_{q,1:\,k}=\{\boldsymbol{z}_{q,1},\boldsymbol{z}_{q,2},\cdots,\boldsymbol{z}_{q,k}\}$表示目标$q$从开始时刻到当前时刻的量测信息集合；$\mathbb{E}_{\boldsymbol{x}_{q,k},\boldsymbol{z}_{q,1:\,k}}$表示对状态$\boldsymbol{x}_{q,k}$和量测集合$\boldsymbol{z}_{q,1:\,k}$的期望运算；$\boldsymbol{J}(\boldsymbol{x}_{q,k})$表示费舍尔信息矩阵(Fisher information matrix, FIM)，其逆为PCRLB。FIM表示如下：

$$\begin{aligned}\boldsymbol{J}(\boldsymbol{x}_{q,k})&=-\mathbb{E}_{\boldsymbol{x}_{q,k},\boldsymbol{z}_{q,1:\,k}}[\Delta_{\boldsymbol{x}_{q,k}}^{\boldsymbol{x}_{q,k}}\log p(\boldsymbol{x}_{q,k},\boldsymbol{z}_{q,1:\,k})]\\&=\boldsymbol{J}^{\mathrm{P}}(\boldsymbol{x}_{q,k})+\boldsymbol{J}^{\mathrm{Z}}(\boldsymbol{x}_{q,k})\end{aligned} \tag{9}$$

其中，$p(\boldsymbol{x}_{q,k},\boldsymbol{z}_{q,1:\,k})$表示目标状态及量测集合的联合概率密度函数；$\boldsymbol{J}^{\mathrm{P}}(\boldsymbol{x}_{q,k})$和$\boldsymbol{J}^{\mathrm{Z}}(\boldsymbol{x}_{q,k})$ 分别表示先验FIM和数据FIM，表示如下：

$$\begin{cases}\boldsymbol{J}^{\mathrm{P}}(\boldsymbol{x}_{q,k})=[\boldsymbol{Q}_{q,k}+\boldsymbol{F}_q[\boldsymbol{J}(\boldsymbol{x}_{q,k-1})]^{-1}\boldsymbol{F}_q^{\mathrm{T}}]^{-1}\\ \boldsymbol{J}^{\mathrm{Z}}(\boldsymbol{x}_{q,k})=[\boldsymbol{H}_{q,k}]^{\mathrm{T}}\boldsymbol{R}_{q,k}(P_{q,k})^{-1}\boldsymbol{H}_{q,k}\,|_{\boldsymbol{x}_{q,k}}\end{cases} \tag{10}$$

其中，$\boldsymbol{H}_{q,k}$表示雅可比矩阵。FIM的逆$\boldsymbol{J}^{-1}(\boldsymbol{x}_{q,k})$表示PCRLB，与发射功率$P_{q,k}$成反比关系。利用归一化思想，$k$时刻目标$q$的归一化PCRLB可以表示为：

$$\mathbb{F}^{\mathrm{PCRLB}}(P_{q,k})=\sqrt{\boldsymbol{\Lambda}\cdot\mathrm{tr}(\boldsymbol{J}^{-1}(\boldsymbol{x}_{q,k}))} \tag{11}$$

其中，$\boldsymbol{\Lambda}=\boldsymbol{I}_2\otimes[1,T_s]$；$\mathrm{tr}(\cdot)$ 表示矩阵求迹。$\mathbb{F}^{\mathrm{PCRLB}}(P_{q,k})$可以衡量每个目标在$k$时刻的跟踪精度，为后续模型建立中的奖励、状态提供基础。

3　强化学习驱动的功率分配算法

3.1　功率优化模型

从数学上来讲，资源分配问题可以看作是一个约束条件下的优化问题。针对多目标场景下的集中式MIMO雷达功率分配问题进行建模，将各个目标的跟踪精度权重和作为目标函数，利用最小和(Min-sum)准则进行优化，建立模型如下：

$$\begin{aligned}&\min_{\boldsymbol{P}_k}\sum_{q=1}^{Q}\omega_q\mathbb{F}^{\mathrm{PCRLB}}(P_{q,k})\\&\text{s.t.}\sum_{q=1}^{Q}P_{q,k}=P_{\mathrm{total}}\\&\quad P_{\min}\le P_{q,k}\le P_{\max}\end{aligned} \tag{12}$$

其中，$\boldsymbol{P}_k=[P_{1,k},P_{2,k},\cdots,P_{Q,k}]^{\mathrm{T}}$表示雷达发射功率向量；$\omega_q$表示目标权重，满足$\sum_{q=1}^{Q}\omega_q=1$，该参数受到目标威胁程度影响，可通过威胁评估方法进行计算；P_{total}表示雷达的发射总功率；$P_{\min}$和$P_{\max}$分别表示每个波束的最大和最小发射功率值。

3.2　强化学习基本模型

3.2.1　MDP定义

为利用强化学习算法对优化模型进行求解，需要建立强化学习驱动的资源分配框架。首先要定义一个MDP为$(\boldsymbol{S},\boldsymbol{A},p,r,\gamma)$[12]。$\boldsymbol{S}$表示状态空间，表示雷达目标跟踪场景下所有系统状态构成的集合；$\boldsymbol{A}$表示动作空间，代表雷达可执行的全部资源分配动作集合；p: $\boldsymbol{S}\times\boldsymbol{A}\times\boldsymbol{S}\to[0,1]$，表示转移概率函数，即表示系统在状态$\boldsymbol{s}_k$ 执行动作$\boldsymbol{a}_k$后转移至状态$\boldsymbol{s}_{k+1}$的转移概率；r: $\boldsymbol{S}\times\boldsymbol{A}\to\mathbb{R}$ ，表示奖励函数，即表征智能体在对应状态执行动作后获得的即时任务收益；γ表示折扣因子。基于MDP的定义，公式(12)中的优化模型可以转换如下：

$$\max_{\pi}\mathbb{E}\left[\sum_{m=0}^{\infty}\gamma^m r\left(\boldsymbol{s}_{k+m},\boldsymbol{a}_{k+m}\right)\right] \tag{13}$$

其中，目标函数表示在k时刻，最大化长期奖励。部分参数的定义在上文中给出，关于其余参数的具体含义在下文中定义。

3.2.2　模型基本要素

将雷达资源分配过程依据MDP进行定义，对应各元素的含义如下：

(1)状态：状态选择对于强化学习的收敛性至关重要，合适的状态选择可以促进强化学习收敛。将目标运动状态与归一化PCRLB结合构建联合状态，即可精准刻画目标运动态势，又能表征各目标理论跟踪估计精度下界。利用各个目标运动状态以及归一化PCRLB构成状态$\boldsymbol{s}_k$，表示如下：

$$\begin{cases}\boldsymbol{s}_k=[\boldsymbol{s}_{1,k},\boldsymbol{s}_{2,k},\cdots,\boldsymbol{s}_{Q,k}]\\\boldsymbol{s}_{q,k}=[\boldsymbol{x}_{q,k-1},\mathbb{F}^{\mathrm{PCRLB}}(P_{q,k-1})]\end{cases} \tag{14}$$

(2)动作：基于状态$\boldsymbol{s}_k$雷达依据策略选择动作$\boldsymbol{a}_k$，表示雷达的发射功率分配情况。为了提升策略收敛速度，同时满足约束条件，对动作进行离散化处理，定义如下：

$$\begin{cases}\boldsymbol{a}_k=[P_{1,k},P_{2,k},\cdots,P_{Q,k}]\\P_{q,k}=P_{\min}+\dfrac{(m-1)(P_{\max}-P_{\min})}{M-1},m=\{1,2,\cdots,M\}\\\sum_{q=1}^{Q}P_{q,k}=P_{\mathrm{total}}\end{cases} \tag{15}$$

其中，M表示单波束可用离散发射功率级数；同时将资源总量约束嵌入动作选取规则，即针对不同目标发射功率和满足$\sum_{q=1}^{Q}P_{q,k}=P_{\mathrm{total}}$，可以将动作空间$\boldsymbol{A}$进行降维，使智能体仅在合法可行域内完成决策搜索。

(3)奖励：利用强化学习进行功率分配时，利用归一化的PCRLB作为奖励，可以使得智能体感受到功率分配对于跟踪误差造成的影响，进而调整策略，实现最优资源分配。依据公式(12)，对应的奖励回报函数$r(\boldsymbol{s}_k,\boldsymbol{a}_k)$定义如下：

$$r(\boldsymbol{s}_k,\boldsymbol{a}_k)=-\zeta\sum_{q=1}^{Q}\mathbb{F}^{\mathrm{PCRLB}}(P_{q,k}) \tag{16}$$

其中，ζ表示放大因子，用来控制回报大小，以便于后续调试和回报收敛。

3.3　D3QN算法求解

3.3.1　强化学习驱动的资源分配框架

构建强化学习驱动的资源分配框架，如图1所示。雷达与目标跟踪的过程构成了环境，雷达内部的智能体与环境进行交互。k时刻，目标滤波后得到的运动状态估计值和归一化PCRLB共同构成状态$\boldsymbol{s}_k$。状态$\boldsymbol{s}_k$输入智能体后，对应智能体根据输入状态以及策略π进行决策，输出动作$\boldsymbol{a}_k$，调节雷达发射功率$\boldsymbol{P}_k$。之后，进一步滤波得到的运动状态估计值和归一化PCRLB共同构成状态$\boldsymbol{s}_{k+1}$，以及奖励$r(\boldsymbol{s}_k,\boldsymbol{a}_k)$等信息被反馈回智能体，用于训练及$k+1$时刻决策。上述过程构成了强化学习驱动的

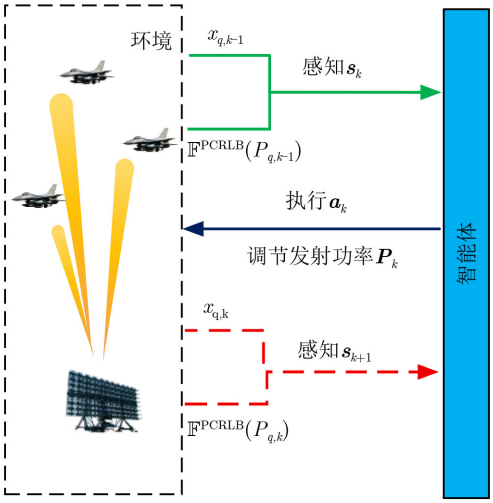


图 1 强化学习驱动的资源分配框架图

资源分配过程，为下一步的研究奠定理论及模型基础。

3.3.2 D3QN算法原理

在雷达多目标动态资源分配场景中，传统 Q-learning 算法受高维状态空间与离散动作组合爆炸的限制，难以直接应用；基础 DQN 算法虽借助神经网络实现高维状态拟合，但仍存在价值估计偏差、动作区分度不足、训练稳定性较弱等问题。为此，本文采用 D3QN 算法作为资源分配优化核心算法。该算法结合双DQN[13,14]、对决DQN两类改进思路，搭配异策略与经验回放实现离线高效训练，加速资源分配策略收敛。一方面，双网络实现动作选择与估值解耦，抑制价值高估；另一方面，网络采用价值、优势双流对决结构，双重优化共同提升训练稳定性。下面从奖励函数、双网络结构、对决网络结构、策略与探索机制四个方面，详细阐述 D3QN 算法的核心原理。

(1)奖励函数

强化学习通过智能体与环境的交互获得的奖励反馈来学习最优策略。智能体的目标是获得最大的折扣奖励，定义如下：

$$R=\sum_{m=0}^{\infty}\gamma^{m}r_{k+m} \tag{17}$$

其中，γ表示折扣因子，决定了优化过程中的远视程度；r_k表示k时刻对应的即时奖励。

(2)双网络结构

D3QN算法保留双网络结构，分别为决策Q网络与目标Q网络：决策Q网络用于实时拟合当前状态下的动作价值；目标Q网络采用延迟更新策略，用于生成稳定的标定Q值，抑制训练过程中的目标值震荡。构建的D3QN的损失函数如下：

$$\begin{cases}L(\theta)=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}[\hat{y}_i-Q(\boldsymbol{s}_i,\boldsymbol{a}_i,\boldsymbol{\theta})]^2\\ \hat{y}_i=r_i+\gamma Q^{\mathrm{target}}(\boldsymbol{s}_{i+1},\underset{\boldsymbol{a'}\in\boldsymbol{A}}{\mathrm{argmax}}Q(\boldsymbol{s}_{i+1},\boldsymbol{a'},\boldsymbol{\theta});\boldsymbol{\theta}^{-})\end{cases} \tag{18}$$

其中，$Q(\boldsymbol{s}_i,\boldsymbol{a}_i,\boldsymbol{\theta})$代表智能体在当前状态$\boldsymbol{s}_i$下执行动作$\boldsymbol{a}_i$的累积期望奖励；$\boldsymbol{\theta}$为决策Q网络的训练参数，负责最优动作选取；$\boldsymbol{\theta}^{-}$为目标Q网络训练参数，负责价值评估，实现动作选择与估值分离。完成批量样本采样后，采用梯度下降算法对决策Q网络参数进行迭代更新，参数更新公式为：

$$\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}-\alpha\nabla_{\theta}L(\boldsymbol{\theta}) \tag{19}$$

其中，α表示学习率；$\nabla_{\theta}L(\boldsymbol{\theta})$表示损失函数关于决策Q网络参数的梯度。

(3)对决网络结构

D3QN算法在保留双网络结构的基础上，将Q网络拆分为状态价值网络与动作优势网络，实现价值解耦表达。对决DQN由两个神经网络组成，通过归一化约束融合两支输出，如图2所示，最终动作价值表达式为：

$$\begin{aligned}Q(\boldsymbol{s},\boldsymbol{a},\boldsymbol{\theta})=&V(\boldsymbol{s};\boldsymbol{\theta}^{\mathrm{V}})+D(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}^{\mathrm{D}})\\&-\frac{1}{|\boldsymbol{A}|}\sum_{\boldsymbol{a}\in\boldsymbol{A}}D(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}^{\mathrm{D}})\end{aligned} \tag{20}$$

其中，$V(\boldsymbol{s};\boldsymbol{\theta}^{\mathrm{V}})$表征环境本身优劣程度且独立于动作；$D(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}^{\mathrm{D}})$为动作优势分支输出优势函数，用于刻画不同动作之间的相对选择优劣。

(4)策略方式

D3QN算法采用异策略方式，即目标策略与行为策略不同。异策略的优点是可以利用行为策略收集经验，之后用其训练目标策略，这种训练方式被称为经验回放。这使得智能体可以利用复用历史经验，搭建经验回放池缓存大量仿真轨迹，实现离线批量采样训练，数据复用率高、收敛速度快。其中

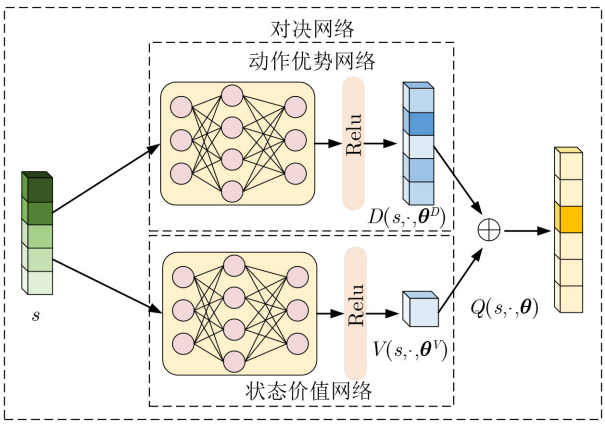


图 2 对决网络结构图

的经验来源可分为两种，一是来源于真实环境交互，二是来源于离线仿真训练，如图3所示。本文利用ε-greedy策略作为行为策略收集经验，表示如下：

$$\boldsymbol{a}_k=\begin{cases}\underset{\boldsymbol{a}\in\boldsymbol{A}}{\text{argmax}}Q(\boldsymbol{s}_k,\boldsymbol{a},\boldsymbol{\theta})，\text{以概率}(1-\varepsilon)\\ \boldsymbol{A}\text{中动作均匀采样，以概率}\varepsilon\end{cases}\tag{21}$$

其中，ε为概率值。为增加探索性，在实验初期设置的ε值较大；在训练过程中逐渐衰减，并在训练后期保持一个较小值，以便于收敛。

3.3.3 D3QN算法流程

采用D3QN算法对上述强化学习驱动的资源分配优化模型进行离线训练，先积累相应的经验储存到经验池中，之后批量采样经验用于决策Q网络和目标Q网络的更新。首先，利用损失函数计算梯度，用于更新决策Q网络参数$\boldsymbol{\theta}$；之后，每隔N_{update}步将决策Q网络的参数复制到目标Q网络。经过离线训练之后，网络参数不断更新。在线时固定网络参数，训练好的网络就能够根据状态输入进行实时资源分配决策，算法的整体流程如算法1所示，算法结构图如图3所示。

4　仿真分析

4.1 基本参数设定

假设目标数量$Q=3$；采样时间间隔$T_{\text{s}}=2$ s；仿真步数为30帧；最小和最大发射功率分别为$P_{\min}=0.1P_{\text{total}}$和$P_{\max}=0.8P_{\text{total}}$；离散发射功率级数$M=8$，则合法动作数量$|\boldsymbol{A}|=36$。D3QN算法参数设置如下：状态维数为15；动作维数为36；状态价值网络和动作优势网络的结构相同，包含两个隐藏层，每层有128个神经元，利用Relu函数对隐藏层进行激活；学习率为0.01；网络更新步数$N_{\text{update}}=50$；经验库容量为5 000，批量采样数$N=128$；放大因子$\zeta=100$。为了验证D3QN算法在不同目标运动场景下的有效性，本文设计了对应仿真场景：目标间距离差距较大，信噪比差异较大。目标初始运动状态设置如表1所示，对应雷达与目标位置如图4所示，此时目标与雷达之间的距离如图5所示。定义目标跟踪过程中的目标均方根误差和(Root Mean Square Error sum, RMSEs)为：

$$\text{RMSEs}_k=\sum_{q=1}^{Q}\sqrt{\frac{1}{N_{sim}}\sum_{\text{i}=1}^{N_{\text{sim}}}\left[(x_{q,k}-\tilde{x}_{q,k}^{i})^2+(y_{q,k}-\tilde{y}_{q,k}^{i})^2\right]}\tag{22}$$

算法 1　D3QN算法

1 初始化雷达、目标、仿真参数
2 初始化智能体 Agent，环境 Env
3 初始化经验回放池 Replaybuffer
4 For epieode $=1:E$ do
5　初始化状态s_1
6　For $k=1:T$ do
7　　将状态 Ak 输入智能体Agent
8　　依据s - greedy 策略选取动作a_k
9　　与环境Env交互
10　　得到奖励η以及下一时刻状态$\Delta+1$
11　　保存经验(s_k,a_k,η_k,s_{k+1})到经验回放池
12　　If aize (Replaybuffer) $\geq N$ then
13　　从经验回放池中批量采样N个样本
14　　更新决策Q网络参数：$\theta\leftarrow\theta-\alpha\nabla_\theta L(\theta)$
15　　End If
16　　If $[F+T\times(\text{epiaode}-1)]\text{mod}\ N_{\text{upulate}}==0$ then
17　　更新目标Q网络参数：$\theta\leftarrow\theta$
18　　End If
19　End For
20 End For

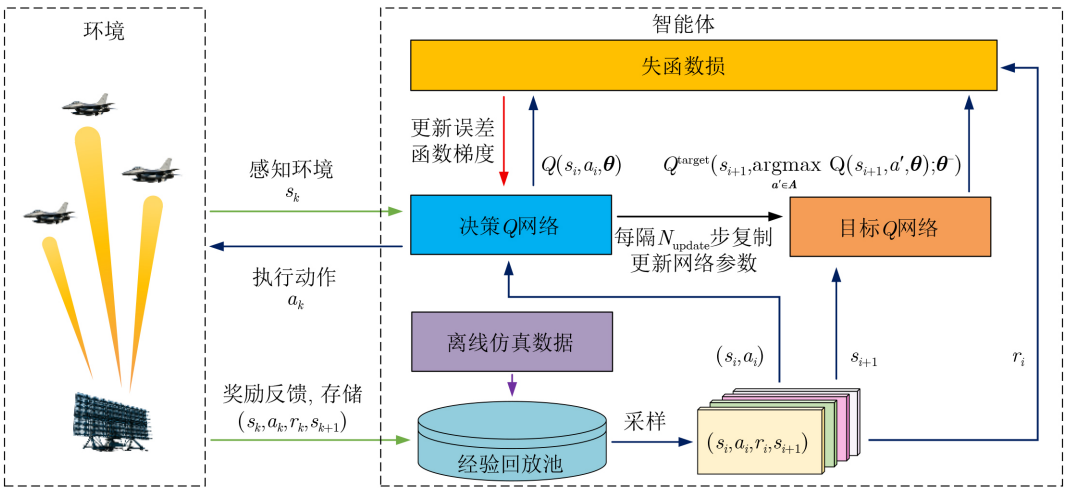


图 3 D3QN算法结构图

其中，$(x_{q,k}, y_{q,k})$表示目标q的真实位置；$(\tilde{x}^i_{q,k}, \tilde{y}^i_{q,k})$表示第$i$次仿真中目标$q$状态估计位置。

4.2 算法性能验证

4.2.1 算法有效性及实时性验证

所有仿真实验均在搭载英特尔酷睿 Ultra 9 275HX 处理器(主频 2.70 GHz)、128GB 内存的计算机上完成，文中所提算法基于 Python 3.13 实现。传统优化算法采用 Python 的 scipy.optimize.minimize 库实现，采用内点法进行求解，最大迭代次数设为1 000，收敛阈值设为$1e^{-6}$，同时对传统优化算法的输出连续结果进行了公平化处理，映射到强化学习动作空间中最近的离散动作。表2中展示了不同算法的运行时间，可以看出基于D3QN算法的资源分配方法的实时性优于传统算法。这是因为传统算法需在线迭代求解优化问题，计算开销大；而 D3QN算法离线完成训练后，在线仅需一次网络前向推理即可输出调度方案，计算耗时更低。

图6中分别展示了基于功率平均分配、传统优化算法、以及不同γ取值下D3QN算法的资源分配方法在多目标跟踪过程中的RMSEs和PCRLBs。结果表明，基于功率平均分配的资源分配方法具有最低的跟踪精度，此时的资源利用效率较低；基于传统优化算法的资源分配方法依托单时刻的优化求解，其跟踪精度高于功率平均分配方法；$\gamma=0$条件下，基于D3QN算法的资源分配方法与基于传统优化算法的方法具有相近的精度，这是因为该条件下强化学习仅优化即时收益，本质上与基于传统优化算法的方法一致，即仅关注当前时刻的优化；$\gamma=0.99$的条件下，基于D3QN算法的资源分配方法早期的跟踪精度低于其他方法，但是在后期的跟踪精度高于其他方法。这一现象表明折扣因子γ可以控制未来跟踪精度对于当前资源分配的重要性，即关注优化过程的远视程度。γ越高，当前优化过程就更关注未来时刻的跟踪精度，更关注全时域累积跟踪精度提升。产生上述性能差距核心原因在

表 1 目标初始运动状态

目标	初始位置(km)	速度(m/s)
1	(100, 24)	(–220, –250)
2	(–20, 20)	(200, –200)
3	(15, 10)	(–200, –100)

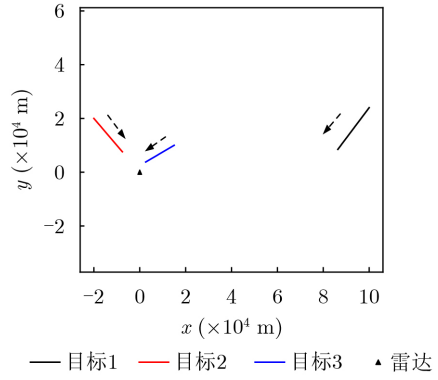


图 4 雷达与目标位置示意

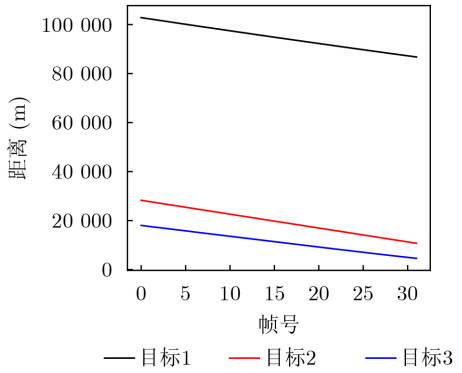
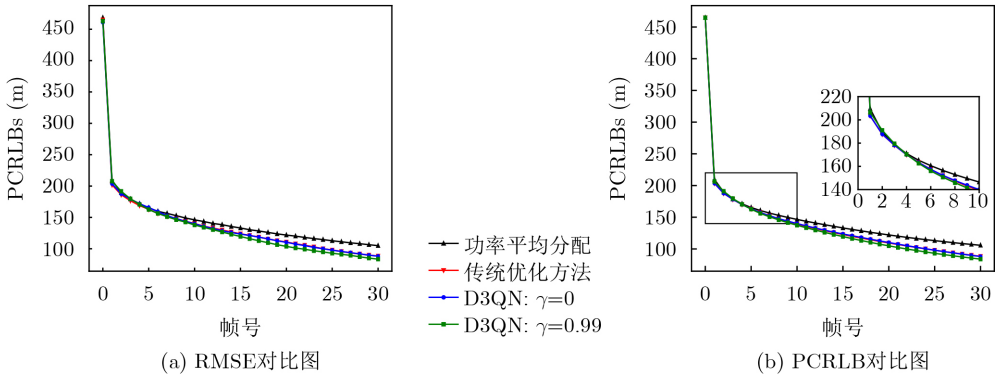


图 5 雷达与目标间距离

表 2 算法实时性对比

算法名称	时间(s)
传统优化算法	0.4
D3QN算法($\gamma=0$)	0.044
D3QN算法($\gamma=0.99$)	0.043



(a) RMSE对比图 (b) PCRLB对比图

图 6 不同策略下RMSEs和PCRLBs对比图

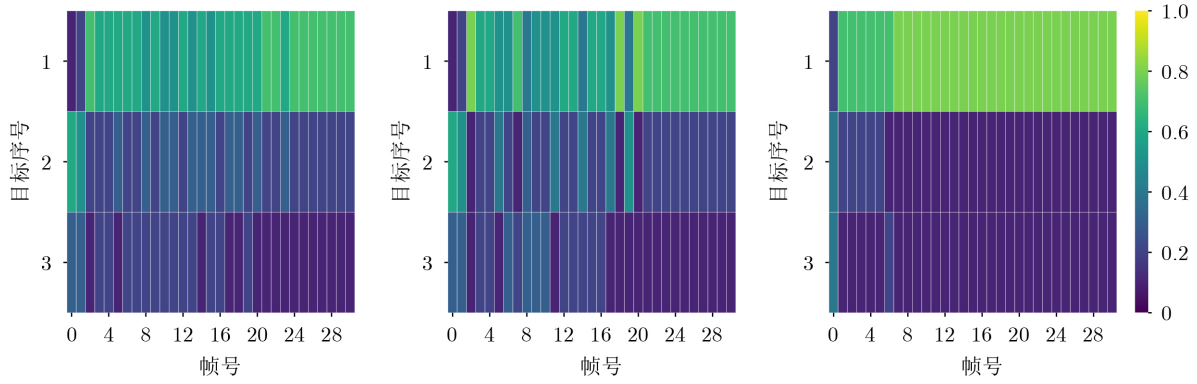
于，本文所提强化学习调度框架依托图1 确立的长期奖励优化机制，结合目标运动状态及归一化PCRLB完成动态决策，突破传统静态优化与时序滞后决策缺陷，能够根据不同目标信噪比需求精准调配资源，进而提升全时域累积跟踪精度。

图7中分别展示了基于传统优化算法、$\gamma=0$和$\gamma=0.99$下D3QN算法的资源分配结果。不同方法的资源分配结果整体趋势是一致的，即倾向于给远距离目标分配更多的资源。从图5中可以看出，目标1距离雷达最远，因此更多的资源被分配给目标1。从图7(a)和图7(b)中可以看出，基于传统优化算法和$\gamma=0$下D3QN算法的两种方法资源分配结果相近，但是两种方法仅关注当前时刻，缺乏远期规划。图7(c)表示$\gamma=0.99$下基于D3QN算法的资源分配方法能提前向远距离低信噪比目标倾斜更多发射功率资源，同时合理缩减近距离高信噪比目标的冗余资源分配量，依据长期奖励优化机制，提升全时域累积跟踪精度。

4.2.2 算法收敛性验证

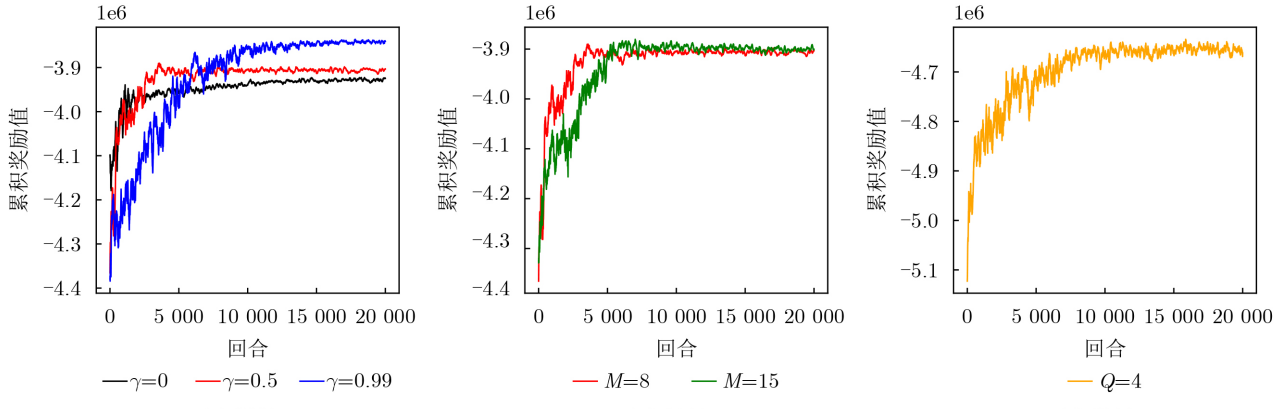
为对比不同情况下的算法收敛性与鲁棒性，本文设置了不同情况下的对比实验：场景1：$Q=3$，$M=8$，γ取值不同；场景2：$Q=3$，$\gamma=0.5$，M取值不同；场景3：$\gamma=0.5$，$M=8$，Q取值不同。其中，Q代表目标数；M代表离散功率档数；γ代表折扣因子。

图8(a)展示了在$Q=3$和$M=8$的前提下，不同γ设置下的算法训练对比曲线。随着γ的增大，智能体的策略表现出更强的“远视”特性，倾向于最大化长期累积收益，因此最终收敛至更高的奖励平台，实现全时域多目标累积跟踪精度提升。然而，γ增大的同时也使前期收敛速度显著放缓，并伴随更剧烈的震荡。原因在于算法所需考虑的步数显著增多，这种对长序列状态的依赖直接加剧了价值评估的难度从而干扰了前期的探索效率。图8(b)反映了在$Q=3$和$\gamma=0.5$的前提下，不同M设置下的算法收敛趋势。$M=8$时，动作空间较为紧凑，初始阶段学习效率相对较高；而$M=15$时，虽然动作空间的扩张增加了前期的探索难度，使得初期上升速率相对延缓，但由于发射功率的选择区间变得更加细致，算法在稳态期的精度也因此获得了略微提升。该结果体现了M的设定，需要在初期探索效率与最终调控精度之间取得合理的折中。图8(c)



(a) 传统优化算法功率分配图　(b) γ=0下D3QN算法功率分配图　(c) γ=0.99下D3QN算法功率分配图

图 7 资源分配结果图



(a) 场景1　(b) 场景2　(c) 场景3

图 8 不同场景下算法累积奖励值对比图

展示了在$\gamma=0.5$和$M=8$的前提下，$Q=4$时算法的收敛趋势。随着Q的增加，系统状态空间与任务复杂度提升，这导致算法的前期收敛速度较$Q=3$出现下降。然而，算法在经过较长周期的迭代探索后仍有效收敛。这一趋势证明了所提方法在面对多目标高维复杂场景时，仍具备必要的泛化拓展能力。

5 结论

本文针对集中式MIMO雷达多目标跟踪中传统优化算法全时域累积跟踪精度优化能力不足、实时性较差的问题，提出一种强化学习驱动的集中式MIMO雷达功率分配算法。通过理论分析与仿真实验验证，本文采用的D3QN算法能够自适应完成雷达功率动态分配，有效提升跟踪精度且算法实时性明显优于传统迭代优化方案。本文研究结果为复杂战场环境下的雷达智能资源管理提供了新思路，对智能化雷达调度体系构建具有参考价值。然而，本文算法仅基于理想化数学模型开展设计，未考虑复杂电磁干扰[15]等实战扰动因素。未来将进一步优化网络结构，提升算法泛化能力与运算效率，推动该智能分配算法在实战雷达系统中的工程落地。

参 考 文 献

[1] 何子述, 程子扬, 李军, 等. 集中式MIMO雷达研究综述[J]. 雷达学报, 2022, 11(5): 805–829. doi: 10.12000/JR22128.
HE Zishu, CHENG Ziyang, LI Jun, *et al*. A survey of collocated MIMO radar[J]. *Journal of Radars*, 2022, 11(5): 805–829. doi: 10.12000/JR22128.

[2] 严俊坤, 陈林, 刘宏伟, 等. 基于机会约束的MIMO雷达多波束稳健功率分配算法[J]. 电子学报, 2019, 47(6): 1230–1235. doi: 10.3969/j.issn.0372-2112.2019.06.007.
YAN Junkun, CHEN Lin, LIU Hongwei, *et al*. Chance constrained based robust multibeam power allocation algorithm for MIMO radar[J]. *Acta Electronica Sinica*, 2019, 47(6): 1230–1235. doi: 10.3969/j.issn.0372-2112.2019.06.007.

[3] 李正杰, 谢军伟, 张浩为, 等. 一种低截获背景下的集中式MIMO雷达快速功率分配算法[J]. 雷达学报, 2023, 12(3): 602–615. doi: 10.12000/JR22203.
LI Zhengjie, XIE Junwei, ZHANG Haowei, *et al*. A fast power allocation algorithm in a collocated MIMO radar under low interception backgrounds[J]. *Journal of Radars*, 2023, 12(3): 602–615. doi: 10.12000/JR22203.

[4] 时晨光, 丁琳涛, 汪飞, 等. 面向射频隐身的组网雷达多目标跟踪下射频辐射资源优化分配算法[J]. 电子与信息学报, 2021, 43(3): 539–546. doi: 10.11999/JEIT200636.
SHI Chenguang, DING Lintao, WANG Fei, *et al*. Radio frequency stealth-based optimal radio frequency resource allocation algorithm for multiple-target tracking in radar network[J]. *Journal of Electronics & Information Technology*, 2021, 43(3): 539–546. doi: 10.11999/JEIT200636.

[5] 黄洁瑜, 张浩为, 谢军伟, 等. 一种面向隐身目标跟踪的雷达组网系统资源优化分配算法[J]. 北京航空航天大学学报, 2026, 52(2): 470–481. doi: 10.13700/j.bh.1001-5965.2023.0782.
HUANG Jieyu, ZHANG Haowei, XIE Junwei, *et al*. A resource optimization allocation algorithm for radar networked system for stealth target tracking[J]. *Journal of Beijing University of Aeronautics and Astronautics*, 2026, 52(2): 470–481. doi: 10.13700/j.bh.1001-5965.2023.0782.

[6] YUAN Ye, YI Wei, HOSEINNEZHAD R, *et al*. Robust power allocation for resource-aware multi-target tracking with colocated MIMO radars[J]. *IEEE Transactions on Signal Processing*, 2021, 69: 443–458. doi: 10.1109/TSP.2020.3047519.

[7] ZHANG Haowei, SHI Junpeng, ZHANG Qiliang, *et al*. Antenna selection for target tracking in collocated MIMO radar[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2021, 57(1): 423–436. doi: 10.1109/TAES.2020.3031767.

[8] 丁梓航, 谢军伟, 齐铖. 基于强化学习的频控阵-多输入多输出雷达发射功率分配方法[J]. 电子与信息学报, 2023, 45(2): 550–557. doi: 10.11999/JEIT211555.
DING Zihang, XIE Junwei, and QI Cheng. Transmit power allocation method of frequency diverse array-multi input and multi output radar based on reinforcement learning[J]. *Journal of Electronics & Information Technology*, 2023, 45(2): 550–557. doi: 10.11999/JEIT211555.

[9] 张鹏, 严俊坤, 高畅, 等. 动态电磁环境下多功能雷达一体化发射资源管理方案[J]. 雷达学报, 2025, 14(2): 456–469. doi: 10.12000/JR24230.
ZHANG Peng, YAN Junkun, GAO Chang, *et al*. Integrated transmission resource management scheme for multifunctional radars in dynamic electromagnetic environments[J]. *Journal of Radars*, 2025, 14(2): 456–469. doi: 10.12000/JR24230.

[10] LU Ziyang and GURSOY M C. Resource allocation for multi-target radar tracking via constrained deep reinforcement learning[J]. *IEEE Transactions on Cognitive Communications and Networking*, 2023, 9(6): 1677–1690. doi: 10.1109/TCCN.2023.3304634.

[11] 易伟, 袁野, 刘光宏, 等. 多雷达协同探测技术研究进展: 认知跟踪与资源调度算法[J]. 雷达学报, 2023, 12(3): 471–499. doi: 10.12000/JR23036.
YI Wei, YUAN Ye, LIU Guanghong, *et al*. Recent advances in multi-radar collaborative surveillance: Cognitive tracking and resource scheduling algorithms[J]. *Journal of Radars*, 2023, 12(3): 471–499. doi: 10.12000/JR23036.

[12] ZHANG Peng, YAN Junkun, PU Wenqiang, *et al*. Collaborative strategy learning based transmit resource management scheme for multiple target tracking in multi-cooperative jamming environment[J]. *IEEE Transactions on Signal Processing*, 2026, 74: 2053–2068. doi: 10.1109/TSP.2026.3693184.

[13] 陈佳美, 孙慧雯, 李玉峰, 等. 基于双深度Q网络算法的无人机辅助密集网络资源优化策略[J]. 电子与信息学报, 2025, 47(8): 2621–2629. doi: 10.11999/JEIT250021.
CHEN Jiamei, SUN Huiwen, LI Yufeng, *et al*. Double deep Q network algorithm-based unmanned aerial vehicle-assisted dense network resource optimization strategy[J]. *Journal of Electronics & Information Technology*, 2025, 47(8): 2621–2629. doi: 10.11999/JEIT250021.

[14] 周黎鸣, 余汐, 范明虎, 等. 基于双深度Q网络的多目标遥感产品生产任务调度算法[J]. 电子与信息学报, 2025, 47(8): 2819–2829. doi: 10.11999/JEIT250089.
ZHOU Liming, YU Xi, FAN Minghu, *et al*. Multi-objective remote sensing product production task scheduling algorithm based on double deep Q-network[J]. *Journal of Electronics & Information Technology*, 2025, 47(8): 2819–2829. doi: 10.11999/JEIT250089.

[15] 李一兵, 孙柳晴, 戚昌龙. 基于改进秘书鸟算法的协同干扰资源分配方法[J]. 电子与信息学报, 2025, 47(5): 1494–1504. doi: 10.11999/JEIT240709.
LI Yibing, SUN Liuqing, and QI Changlong. Collaborative interference resource allocation method based on improved secretary bird algorithm[J]. *Journal of Electronics & Information Technology*, 2025, 47(5): 1494–1504. doi: 10.11999/JEIT240709.

黄洁瑜：男，博士生，研究方向为雷达资源管理、电子对抗技术，邮箱：hjy_ed@163.com.
谢军伟：男，教授，研究方向为新体制雷达、干扰与抗干扰技术.
张浩为：男，副教授，研究方向为雷达资源管理，邮箱：zhw_xhzf@163.com.
冯为可：男，副教授，研究方向为雷达信号处理、雷达成像及目标识别.
韩卫航：男，博士生，研究方向为频控阵雷达、电子对抗技术.

A Reinforcement Learning Driven Power Allocation Algorithm for Collocated MIMO Radar

HUANG Jieyu　XIE Junwei　ZHANG Haowei　FENG Weike　HAN Weihang

(*Air Force Engineering University Air and Missile Defense College*, *Xi'an* 710051, *China*)

Abstract:

Objective Traditional optimization based on power allocation algorithm for collocated MIMO radar has two fundamental limitations. First, it optimizes tracking performance only for the next time step, lacking a global view across the entire time horizon. This myopic strategy cannot achieve optimal multiple target tracking accuracy over a long duration, especially when target trajectories vary significantly. Second, their iterative solving process involves nonlinear constrained optimization, which incurs high computational complexity. In dynamic battlefield environments where target states change rapidly, such algorithms fail to meet real-time requirements. To address these issues, this paper proposes a reinforcement learning (RL) driven power allocation algorithm. Unlike traditional methods, the proposed approach formulates the problem as a Markov decision process that maximizes long-term cumulative rewards. The algorithm adaptively allocates limited power resources among multiple beams based on the current system state, balancing immediate tracking performance and future gains.

Methods The posterior Cramér-Rao lower bound (PCRLB) is employed to quantify the theoretical tracking error lower bound for each target. The state space is constructed by combining the motion states (position and velocity) of all targets and normalized PCRLB from the previous allocation. The action space consists of discrete transmit power levels for each beam, subject to the total power budget and beam power limits. All feasible power allocation vectors are enumerated and encoded to reduce dimensionality. The reward function is defined as the negative weighted sum of normalized PCRLB, encouraging the agent to minimize tracking errors. The power allocation process is formulated as a Markov decision process (MDP). The Dueling Double Deep Q-Network (D3QN) algorithm is adopted to solve this MDP. D3QN integrates three key enhancements: (1) a double network training framework (decision Q-network and target Q-network) to stabilize learning; (2) a

dueling architecture that decomposes the Q-value into state value and action advantage functions, improving action discrimination; (3) off-policy learning with experience replay, enabling efficient use of historical trajectories. The ε-greedystrategy is used for exploration, with epsilon decaying over episodes. After offline training, the learned network outputs power allocation decisions in real-time given the current system state, without any iterative optimization.

Results and Discussions Simulations are conducted with three targets following constant velocity models. Fixed power allocation yields the lowest tracking accuracy due to inefficient resource utilization. The traditional optimization method, which minimizes the immediate tracking error, achieves moderate accuracy but remains myopic. For D3QN with discount factor $\gamma = 0$, the performance is nearly identical to the traditional method. In contrast, D3QN with $\gamma = 0.99$ achieves significantly better full time horizon accuracy. Power allocation patterns reveal that the D3QN with $\gamma = 0.99$ preemptively allocates more power to distant and low signal to noise ratio (SNR) targets earlier, while reducing redundant power to close and high SNR targets. The training curves show that $\gamma = 0.99$ achieves a higher steady state cumulative reward, although with more oscillations due to the complexity of estimating future returns. Moreover, the trained D3QN network outputs decisions instantaneously, whereas traditional optimization requires solving a constrained optimization problem at each time step, providing a clear real-time advantage.

Conclusions This paper proposes a RL driven power allocation algorithm for collocated MIMO radar multiple target tracking that overcomes the myopic and computationally intensive limitations of traditional optimization methods. The algorithm uses PCRLB to construct state and reward functions, models the allocation process as an MDP, and solves it using the D3QN algorithm. Simulation results demonstrate that the approach based on D3QN with a suitable discount factor ($\gamma = 0.99$) significantly improves full time horizon target tracking accuracy. The improvement stems from the agent's ability to learn a long-term optimal policy that preemptively allocates resources to future challenging targets. Furthermore, the trained network enables real-time decision, substantially reducing computational latency from iterative solving to instantaneous forward propagation. This work provides a new approach for intelligent radar resource management in complex battlefield environments.

Key words: Reinforcement learning (RL); Collocated MIMO radar; Power resource allocation; Dueling Double Deep Q-Network (D3QN) algorithm