基于强化学习的深空探测器自主任务规划方法

毛维杨[1]，王 彬[1,2]，柳景兴[1]，熊 新[1]

（1. 昆明理工大学 信息工程与自动化学院，昆明 650500；2. 昆明理工大学 云南省人工智能重点实验室，昆明 650500）

摘 要：针对深空探测器自主任务规划多约束的需求，提出了基于动态奖励的强化学习深空探测器任务自主规划模型构建方法，建立了深空探测器智能体的交互环境，构建了策略网络和融合资源、时间以及时序约束的损失函数，并提出动态奖励机制对传统策略梯度学习方法进行了改进。仿真实验结果表明：该方法可实现自主任务规划，规划成功率和规划效率相比静态奖励策略梯度算法有明显的提升，并且能在任意状态下开始规划而无需改变模型结构，提高了对不确定规划任务的适应性。该方法为深空探测器自主任务规划与决策提供了一种新的解决方案。

关键词：深空探测；任务规划；策略梯度；强化学习；动态奖励

中图分类号：TP18 **文献标识码：**A **文章编号：**2096-9287(2023)02-0220-11

DOI:10.15982/j.issn.2096-9287.2023.20220049

引用格式：毛维杨，王彬，柳景兴，等. 基于强化学习的深空探测器自主任务规划方法[J]. 深空探测学报（中英文），2023，10（2）：220-230.

Reference format: MAO W Y，WANG B，LIU J X，et al. An autonomous planning method for deep space exploration tasks in reinforcement learning based on dynamic rewards[J]. Journal of Deep Space Exploration，2023，10（2）：220-230.

引 言

随着深空探测活动及手段的进步和发展，中国正在不断加快向更远深空迈进的步伐，从发展进入深空能力，到探索资源利用能力，再到拓展深空能力，深空探测的目标逐步深化[1]。近年来，对地外天体目标尤其是高科学价值区域开展原位探测和采样返回成为未来行星探测的重要发展方向[2]，环境更为复杂，任务也更加灵活多变，因此对深空探测器的自主性和智能性提出了更高要求。深空探测器自主任务规划技术是实现深空探测器自主智能的关键技术之一，深空探测任务的自主规划一般在多个子系统之间展开协同规划，不但要求根据任务实时决策输出可靠的动作序列，还必须满足时间、资源和同步等多种约束条件，因此常规任务规划方法和技术难以满足该领域的要求。

许多研究者将传统的规划算法应用于深空探测器任务规划中，赵凡宇等[3]提出了一种结合活动相关度选择机制的启发式规划算法，利用相关度来设计启发式信息，从而提高了规划效率。姜啸等[4]将深空探测规划问题转化为约束可满足问题，提出了一种以动作为中心的约束可满足技术用于深空探测规划问题，并设计了一种基于动态约束表的外延约束快速过滤算法，有效降低了约束处理中无效的约束检查次数[5]。金颢等[6]提出了一种适用于深空探测器的时间线转移路标启发式规划方法，引入了实现任务目标必需的状态集合，减少了规划过程中冗余节点扩展数量，提高了规划效率。赵宇庭等[7]提出基于动态智能体交互图的深空探测器任务规划方法，设计了一种智能体交互图来引导多个智能体协同规划。王晓晖等[8]提出了一种新的约束简化方法，通过在时间线规划模型中根据两两子系统间的实时状态关系定义启发式因子，并利用该因子在规划周期内的取值建立子系统间时间线临时从属关系，从而合理地降低规划过程中的约束复杂程度。

近年来很多机构和学者也围绕人工智能算法在深空探测器任务规划中的应用展开了理论和仿真研究。包含平台和载荷任务管理的两阶段航天器自主任务规划算法，实现了平载一体的航天器自主任务管理，其能够根据高级任务目标输出完整的指令序列[9]。王鑫等[10]针对多智能体深空探测器的知识表示问题，提出了一种基于知识图谱的新方法。李玉庆等[11]基于模糊神经网络提出了一种动态环境下航天器的重调度方法。贺东雷等[12]将遗传算法应用到深空探测任务规划问题，并用实数编码的方式对其进行了改进，提高了算法的

收稿日期：2022-05-18 修回日期：2022-07-11

基金项目：民用航天预研资助项目

计算效率。

强化学习（Reinforcement Learning， RL）模拟人脑可对多巴胺奖励产生连接模式及强度变化的神经元塑型机制，以“试错”的方式学习最优行为策略，通过与环境进行交互获得的奖惩不断改进行动方案以适应环境达到目标，从而获得最大回报[13]。史谦郡等[14]提出了一种基于深度强化学习的空间站任务重规划方法；郭林杰等[16]结合深度确定性算法设计了一种神经网络对深空探测器的跳跃过程进行规划；Furfaro等[17]设计了一组深度神经网络，实现了预测燃料最优控制动作以执行自主登月。由于强化学习在智能决策上的优势，近年来，有国内外学者将强化学习应用于深空探测领域并展开了相关研究[14-18]。

本文以深空探测器自主任务规划为目标，提出了一种强化学习深空探测器任务规划模型构建方法，重点研究了模型的架构、策略网络构建、融合资源约束与时间约束的损失函数建模以及动态奖励机制，并且以不确定的探测器初始态为输入对模型进行训练，提升了模型规划过程的鲁棒性。仿真结果表明，根据本文方法建立的强化学习模型可以实现深空探测器自动完成任务规划、可根据资源和约束自主选择最优决策结果输出；对比实验结果显示，本文方法在自主任务规划中具有较高的规划成功率，并且比静态奖励策略梯度算法有明显的提升，从而为深空探测器自主任务规划与决策提供了一种新的解决方案。

1　深空探测器任务规划强化学习参数

基于强化学习方法的深空探测器自主任务规划是根据对空间环境和探测器自身状态的感知，依据一段时间内的任务目标，对若干可供选择的动作、可用资源、约束关系等进行推理，自动自主地生成一组时间有序的动作序列。该规划一旦执行，便可将探测器状态成功转移到期望的目标状态[19]。

为建立本文算法的环境，本文在已有研究基础上给出深空探测器任务规划过程中强化学习参数定义。

定义1. 定义任务。

$$M=\{I_{\mathrm{M}},N_{\mathrm{M}},E_{\mathrm{ML}},E_{\mathrm{MP}},t_{\mathrm{MES}},t_{\mathrm{MLE}}\} \tag{1}$$

其中：M代表深空探测器可执行的科学任务；I_{M}代表任务的编号；N_{M}代表任务的名称；E_{ML}代表该项任务的价值；E_{MP}代表该项任务的优先级；t_{MES}代表该项任务的最早开始时间；t_{MLS}代表该项任务的最晚开始时间。

定义2. 资源集。资源集表示深空探测器在执行动作时会用到资源所组成的集合，用$\boldsymbol{R}$表示。

$$\boldsymbol{R}=\{r_1,r_2,\cdots,r_i,\cdots,r_{|\boldsymbol{R}|}\} \tag{2}$$

$$r_i=\{I_{r_i},N_{r_i},L_{r_i},C_{r_i}\} \tag{3}$$

其中：r_i表示第i种资源，式（3）中I_{r_i}代表资源的编号；N_{r_i}代表该资源的名字；L_{r_i}代表该资源的剩余量；C_{r_i}代表该资源的类型，即该资源为可再生资源还是为不可再生资源。

定义3. 深空探测器状态集。用$\boldsymbol{S}$来表示深空探测器的状态集，假设探测器一共有N个子系统。第i个子系统的状态集用$\boldsymbol{S}_i$表示，其状态有$|\boldsymbol{S}_i|$种，其中$1\leqslant i\leqslant N$可表示为

$$\boldsymbol{S}_i=\{s_{i1},s_{i2},\cdots,s_{ij},\cdots s_{i|S_i|}\} \tag{4}$$

则$\boldsymbol{S}$可表示为

$$\boldsymbol{S}=\boldsymbol{S}_1\cup\boldsymbol{S}_2\cup\cdots\boldsymbol{S}_i\cdots\cup\boldsymbol{S}_N \tag{5}$$

其中：s_{ij}表示第i个子系统的第j个状态；S_i表示第i个子系统的状态集。

定义4. 可执行动作。可执动作即为深空探测器可以执行的动作，用a_{ij}表示。

$$a_{ij}=\{I_{a_{ij}},D_{a_{ij}},R_{a_{ij}}\} \tag{6}$$

$$\boldsymbol{R}_{a_{ij}}\subset\boldsymbol{R} \tag{7}$$

其中：$I_{a_{ij}}$代表活动的编号；$D_{a_{ij}}$代表活动持续时长；$\boldsymbol{R}_{a_{ij}}$代表活动消耗的资源集合；$\boldsymbol{R}$代表深空探测器的资源集，$\boldsymbol{R}_{a_{ij}}$是$\boldsymbol{R}$的子集。

定义5. 深空探测器可执行动作集，用$\boldsymbol{A}$来表示深空探测器可执行动作集。第i个子系统的可执行动作集用$\boldsymbol{A}_i$表示，其种类有$|\boldsymbol{A}_i|$种，其中$1\leqslant i\leqslant N$。$\boldsymbol{A}_i$可表示为

$$\boldsymbol{A}_i=\{a_{i1},a_{i2},\cdots,a_{ij},\cdots a_{i|a_i|}\} \tag{8}$$

$$\boldsymbol{A}=\boldsymbol{A}_1\cup\boldsymbol{A}_2\cup\cdots\boldsymbol{A}_i\cdots\cup\boldsymbol{A}_N \tag{9}$$

其中：a_{ij}表示第i个子系统的第j个可执行动作；$\boldsymbol{A}_i$表示第i个子系统的可执行动作集。

定义6. 深空探测器状态转换集，在本文所研究问题中，构建三元组$\boldsymbol{G}=(\boldsymbol{S},\boldsymbol{A},\boldsymbol{H})$，其中$\boldsymbol{G}$表示深空探测器状态与动作的状态转换集合，$\boldsymbol{S}$表示深空探测器状态集，该集合中有$|\boldsymbol{S}|$种状态。$\boldsymbol{A}$表示深空探测器可执行动作集，该集合有$|\boldsymbol{A}|$种可执行动作。$\boldsymbol{H}\subseteq\boldsymbol{S}\times\boldsymbol{A}\times\boldsymbol{S}$表示三元组的集合，三元组的基本形式是$(s_{ij},a_{ij},s_{ij+1})$，其中$s_{ij}$表示第$i$个子系统的第$j$个状态，$a_{ij}$表示第$i$个子系统执行了该子系统可执行的第$j$个动作，$s_{ij+1}$表示第$i$个子系统执行了动作$a_{ij}$后到达的第$j+1$个状态。

2　基于强化学习的任务规划模型构建

强化学习的目标是通过与环境进行交互所获得的奖惩找到一个最优的行为策略从而获取最大的回报。强化学习的基本算法主要有基于值的算法（Value-

Based）和基于策略的算法（Policy-Based）[20]。其中基于策略梯度的强化学习方法在于直接对策略进行建模并优化，让神经网络直接输出状态s下应该执行何种动作的策略函数，而无需计算每一个动作的价值，减少了规划过程中的计算量，从而提高了规划的效率，能够有效解决离散动作序列的推理和决策[21-22]。

2.1 模型原理及架构

根据第1节中所定义的深空探测器任务规划强化学习参数，本文给出了一种基于策略梯度强化学习的深空探测器任务规划方法。该方法将执行任务的深空探测器看作一个智能体，根据包括动作规划、资源和时间约束在内的深空探测领域知识构建智能体的环境，深空探测器智能体与该环境进行交互，通过所获得的奖惩（回报）不断更新策略梯度，直至找到一个整体最优的行为策略并输出相应的行为序列。基于策略梯度强化学习深空探测器任务规划方法原理图如图1所示，主要包括智能体环境（Environment）和训练模型（Train）2大部分。

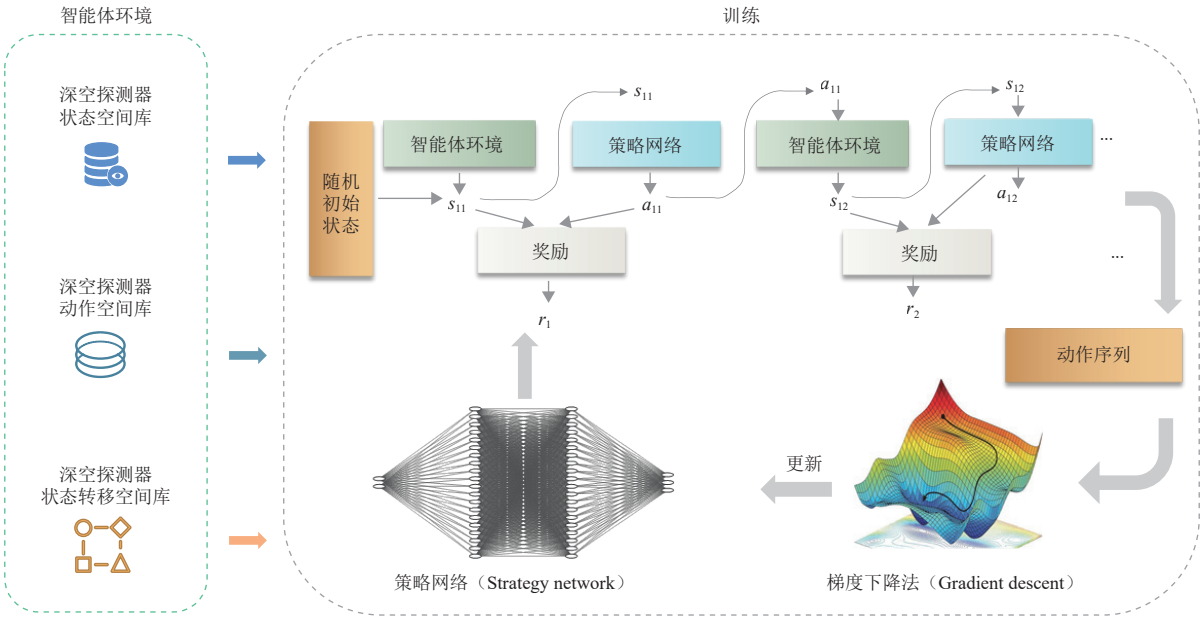

图 1 基于策略梯度强化学习深空探测器任务规划方法原理图

Fig. 1 Schematic diagram of task planning method for deep space detectors based on policy gradient reinforcement learning

智能体环境是智能体在训练和任务规划阶段所处的交互环境，深空探测器状态空间库包含深空探测器中各子系统可达的状态；深空探测器动作空间库包含深空探测器各子系统可执行的动作。深空探测器状态转移空间库包含深空探测器各子系统从当前状态执行某可执行动作后能到达的下一个状态的所有状态转移关系。

深空探测器强化学习训练模型的输入是根据深空探测器任务目标状态空间随机生成的初始状态，网络输出是深空探测器动作空间库包含的动作所组成的序列。本文将深空探测器智能体的行动策略转换为第1节中参数的非线性函数，使用策略网络对参数的寻优，计算每一个完整交互过程（Episode，用E表示）的损失函数值采用梯度下降法不断更新策略网络，使得网络参数向最优的方向逼近，直至找到函数的最值。在训练过程中随着策略网络的进化，深空探测器智能体的规划能力不断增强，训练结束后得到的最优策略网络即可用于深空探测器的自主任务规划模型。

2.2 智能体交互环境

深空探测器状态空间库。本文将深空探测器看成一个智能体，智能体的状态定义为深空探测器中各子系统的状态，即智能体状态来自由深空探测器各子系统状态组成的深空探测器状态空间库（***State***），其中$\boldsymbol{State}=\boldsymbol{S}$。

深空探测器动作空间库。深空探测器的每个子系统都有可执行动作集，第i个子系统的可执行动作集为$\boldsymbol{A}_i$，如式（8）所示，该集合中的可执行动作a_{ij}可以使该子系统从状态s_{ij}到达s_{ij+1}。当状态为s_{ij}的情况下智能体可执行动作集为$\boldsymbol{A}_i$的子集$\boldsymbol{A}_{ij}$，共有$|\boldsymbol{A}_{ij}|$种，如式（10）和式（11）所示，深空探测器动作空间库定义为***Action***，如式（12）所示

$$\boldsymbol{A}_{ij}\subset\boldsymbol{A}_i \tag{10}$$

$$\boldsymbol{A}_{ij}=\left\{a_{i1}^{j},a_{i2}^{j},\cdots,a_{k}^{j},\cdots,a_{i|A_j|}^{j}\right\} \tag{11}$$

$$\boldsymbol{Action}=\boldsymbol{A}_{11}\cup\boldsymbol{A}_{12}\cup\cdots\boldsymbol{A}_{ij}\cdots\cup\boldsymbol{A}_{N|S_N|} \tag{12}$$

式（10）中的$\boldsymbol{A}_{ij}$表示第i个子系统在第j个状态时可执行的动作集，$\boldsymbol{A}_i$表示第i个子系统可执行的动作集。式（11）中的a_{ik}^j表示第i个子系统在第j个状态时可执行的第k个动作。式（12）中的$\boldsymbol{A}_{ij}$表示第i个子系统的第j个状态下的可执行动作集，深空探测器一共分为N个子系统，其中第i个子系统有$|S_i|$种状态。

深空探测器状态转移空间库。本文用三元组来定义深空探测器状态转移空间库（$\boldsymbol{Transfer}$），该库中包含的状态转移是指智能体执行某个动作后会从当前状态转移到另一个状态，其用第1节的定义6中的三元组表示，$\boldsymbol{Transfer}$用式（13）表示

$$\boldsymbol{Transfer}=\{(s_{11},a_{11},s_{12}),\cdots,(s_{ij},a_{ij},s_{ij+1}),\cdots\} \tag{13}$$

2.3　策略网络构建

本文采用多层感知器（Multilayer Perceptron，MLP）神经网络结构[23]构建深空探测器的自主任务规划模型中的策略网络，采用策略梯度算法对深空探测器智能体在每种状态下要执行的动作进行选择。然后依据深空探测器状态转移空间的状态转移规则得到下一个状态和当前的即时奖励，多次重复以上过程直到智能体到达终止状态，此时得到由状态动作序列和期望奖励序列组成的；损失函数依据用策略梯度方法对策略网络进行更新，便完成了一次策略网络更新，通过多次更新，该策略网络就能拥有满足深空探测器任务规划问题要求的规划能力，训练过程如图2所示。

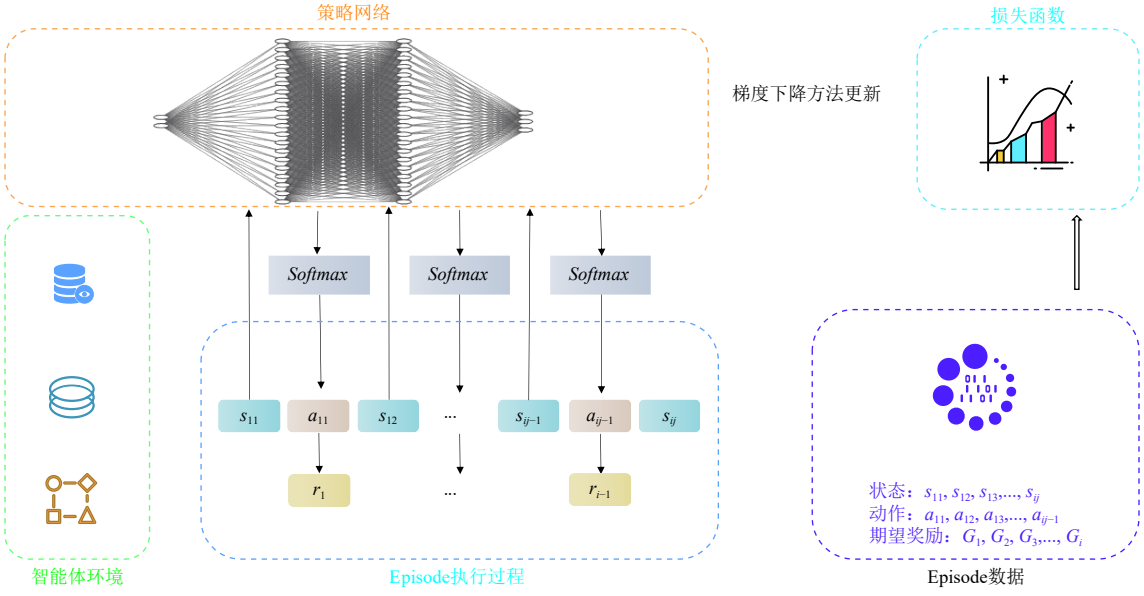

图 2　策略网络训练过程

Fig. 2　Policy network training process

结合2.2节智能体环境，定义一个4层的MLP策略网络结构如图3所示。其中网络的输入层是深空探测器智能体当前的状态，用2个参数表示，隐含层共2层，每层40个节点，输出是输入层状态下深空探测器智能体采取每个动作的概率。

MLP策略网络的连接权值计算方式[23]为

$$q_k^t=\sum_i^d w_{ik}x_i^t+b_k^t \tag{14}$$

其中：d为节点个数。

$$h_k^t=f(q_k^t) \tag{15}$$

在策略网络输出各动作的$Softmax$函数值之后，经过一个操作选取其中被执行概率最大的动作，其公式为

$$\pi_\theta(a_{ij}|s_{ij})=\frac{\mathrm{e}^{f_{a_{ij}}}}{\sum_{k=1}^{K}\mathrm{e}^{f_k}},\ j=1,2,\cdots,K \tag{16}$$

其中：$\pi_\theta(a_{ij}|s_{ij})$表示在状态下动作的被执行概率。此时根据状态转移空间里的状态转移规则到达下一个状态，一直到深空探测器到达不能转移的终止状态，一个完整的训练过程结束，得到所有被选动作组成的动作序列。

2.4　融合资源约束与时间约束的损失函数

强化学习算法通过构建损失函数用策略梯度方法对策略网络进行更新，与一般的强化学习问题不同，深空探测任务规划问题中存在着多种约束，包括时间、资源以及时序。在智能完成每一个完整交互过程的时候都需要考虑这3种约束。

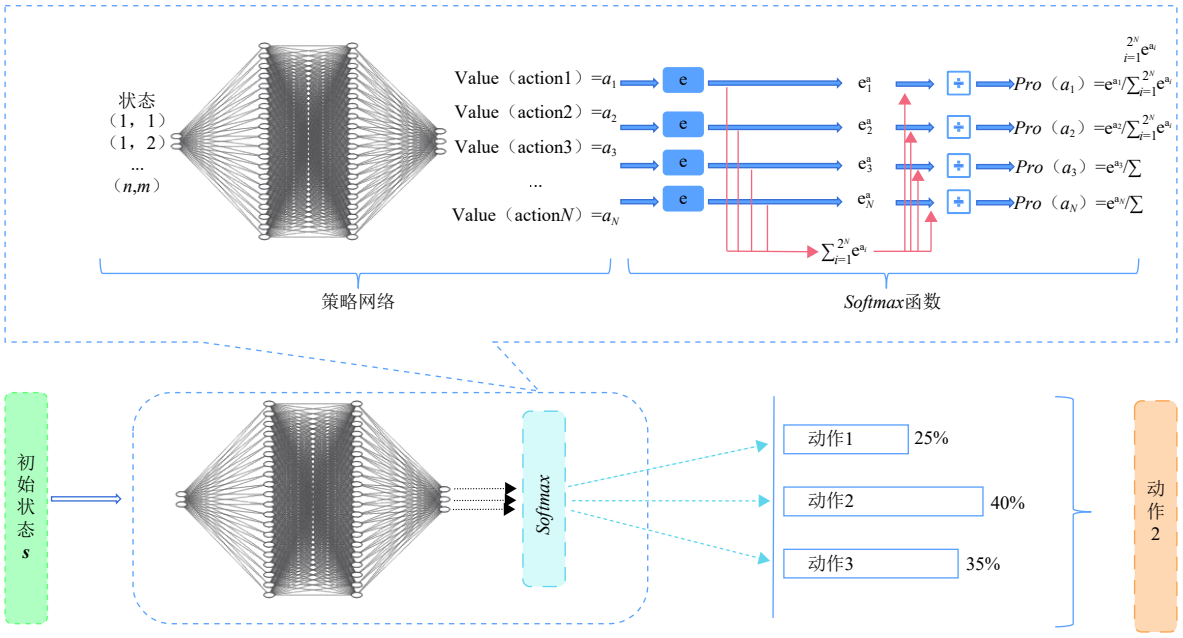

图 3 基于策略梯度强化学习方法动作输出原理图

Fig. 3 Principle diagram of action output based on policy gradient reinforcement learning method

1）时间约束。为描述深空探测器各子系统的可执行动作所需要的时间情况，建立动作时间消耗矩阵$\boldsymbol{T}$，如式（18）所示，该矩阵保存了各个子系统的可执行动作的时间消耗情况，以满足规划过程中的时间约束。

$$M = \max\{|A_1|,|A_2|,\cdots,|A_N|\} \tag{17}$$

$$\boldsymbol{T} = [t_{ij}]_{N\times M} = \begin{bmatrix} t_{11} & \cdots & t_{1j} & \cdots & t_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ t_{i1} & \cdots & t_{ij} & \cdots & t_{iM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ t_{N1} & \cdots & t_{Nj} & \cdots & t_{NM} \end{bmatrix} \tag{18}$$

其中：$|A_i|$为深空探测器的第i个子系统的可执行动作数；M为深空探测器的N个子系统对应的可执行动作数中最大的动作数；$\boldsymbol{T}$为动作时间消耗矩阵；t_{ij}为第i个子系统的第j个可执行动作所需要消耗的时间。

2）资源约束。为描述深空探测器各子系统的可执行动作所需要的资源，建立动作资源消耗矩阵$\boldsymbol{R}$，式（19）所示，该矩阵保存了各子系统可执行动作的资源消耗的情况，以满足规划过程中的资源约束。

$$\boldsymbol{R} = [r_{ij}]_{N\times M} = \begin{bmatrix} r_{11} & \cdots & r_{1j} & \cdots & r_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{i1} & \cdots & r_{ij} & \cdots & r_{iM} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{N1} & \cdots & r_{Nj} & \cdots & r_{NM} \end{bmatrix} \tag{19}$$

其中：$\boldsymbol{R}$为动作资源消耗矩阵；r_{ij}为第i个子系统的第j个可执行动作所需要消耗的资源。

3）时序约束。对于深空探测器任务规划中存在时序约束问题定义动作a_{ij}的前序动作集合$\boldsymbol{P}_{a_{ij}}$，前序动作集合包含了该动作被执行以前必须先执行的动作，如式（20）所示

$$\boldsymbol{P}_{a_{ij}} = \left\{a_1^{ij},\cdots,a_i^{ij},\cdots,a_{|\boldsymbol{P}_{a_{ij}}|}^{ij}\right\} \tag{20}$$

综合以上资源、时间和时序约束，采用动态奖励策略，即时奖励的计算公式为

$$\boldsymbol{E} = \{a_{11},a_{12},\cdots,a_{ij}\} \tag{21}$$

$$r_m = \begin{cases} \dfrac{1}{a\_t} + \dfrac{1}{a\_r}, & \boldsymbol{P}_{a_{ij}} \subset \boldsymbol{E} \\ 0, & P_{a_{ij}} \not\subset \boldsymbol{E} \end{cases} \tag{22}$$

其中：$\boldsymbol{E}$表示深空探测智能体在完成一个完整交互过程后已规划出的动作序列；式（22）中当动作a_{ij}的前序动作集合$\boldsymbol{P}_{a_{ij}}$为$\boldsymbol{E}$的子集的时候，取即时奖励r_m为当前动作所消耗的时间a_t以及所消耗的资源a_r的倒数的和，其中$a_t \in \boldsymbol{T}$，$a\_r \in \boldsymbol{R}$；否则取即时奖励r_m为0。采用这样的动态奖励机制把深空探测任务规划中的时间束、资源和时序约束融合在强化学习的训练过程中。

策略网络以得出最好状态动作序列为目标，状态动作序列是策略网络多次动作预测得到的结果，所以用来评估策略网络的损失函数需考虑策略网络在状态动作序列中每个动作预测上的表现，且依据每个动作潜在的价值来设置损失函数在该动作的关注度，取当前动作的期望回报G_m作为其在策略网络被评估过程中需要被关注的程度。G_m为当前动作即时奖励r_m与状态动作序列中剩下动作对应的即时奖励的衰减和，其定

义如下

$$G_m = r_m + \sum_{z=m+1}^{anm} \gamma^{z-m} r_z = r_m + \gamma G_{m+1} \tag{23}$$

其中：γ为衰减因子，用于惩罚未来回报中的不确定性，$0 < \gamma \leqslant 1$。

根据策略梯度下降方法构造式（24）所示的损失函数对其进行更新。

$$L = \frac{\sum_{1}^{enm} \left\{ \sum_{m=1}^{anm} -a'_{ij} G_m \log[\pi_\theta(a_{ij}|s_{ij})] \right\}}{enm} \tag{24}$$

其中：anm为当前$Episode$所包含的动作数量；enm为训练迭代总次数；a'_{ij}为当前选取的动作；$\pi_\theta(a_{ij}|s_{ij})$为策略网络的可执行动作预测结果。

2.5　强化学习深空探测任务规划算法实现

本文是基于策略梯度强化学习方法，在其框架基础上，根据深空探测器任务规划问题的特点，按照上文方法构建了训练模型，经过训练后得到了最优策略网络参数，将其用于深空探测器的自动任务规划，算法实现伪代码如表1所示。

表 1　算法实现

Table 1　Algorithm implementation

Algorithm 1 An Autonomous Planning Method For Deep Space Exploration Tasks In Reinforcement Learning Based On Dynamic Rewards

Input: Tasks Γ, Learning Rate α, Step Size β, End State Collection S_e, Policy Network Parameters θ Model Training Times N;

Output: al;

Start training the model
For 1 to N **do**:
randomly initialize state $s \leftarrow s_i$
For s_i not in S_e **do**:
get action a_i from Strategy network f_θ
$s_i \leftarrow s_{i+1}$
get state-action sequence al and reward rl
End For
$\theta \leftarrow \theta - \beta \nabla_\theta Loss(f_\theta)$
End For
get the model f_θ
initial state s'_i: $s'_i \leftarrow \Gamma$
get action sequence al: $al \leftarrow f_\theta$

3　仿真验证

本文针对深空探测器执行拍照任务，采用基于动态奖励的策略梯度强化学习深空探测器自主任务规划方法对其进行规划，分别设计完成了自主任务规划的有效性实验、规划成功率统计，以及动态奖励与非动态奖励策略梯度算法结果对比等3个实验，验证了本文方法的有效性。

3.1　实验场景

本文算法的实验环境如表2所示。本文实验场景是规划深空探测器拍照任务，目标是在尽可能短的时间内使用较少的资源完成任务，涉及4个子系统，分别是电源子系统（power subsystem）、姿态调整子系统（attitude adjustment subsystem）、相机子系统（camera subsystem）、通信子系统（communication subsystem），共包含16个状态以及25个可执行动作，如表3所示，其介绍了各个子系统对应的状态以及状态对应的可执行动作。

表 2　实验环境

Table 2　Experimental environment

参数	配置
操作系统	Windows 1 064位
编程语言	Python3.8.8
CPU	Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz
内存	16GB
强化学习框架	PARL[24]

表 3　所用深空探测器拍照案例

Table 3　Examples used in deep space probes

子系统名称	状态	可执行动作	子系统名称	状态	可执行动作
电源子系统	休眠	启动	拍照子系统	待机	接收姿态调整信号
	运行	发送低于安全阈值信号			关闭镜头
		发送高于安全阈值信号		就绪	执行拍照
		读取目标坐标		拍照	写入储存器
		打开镜头		保存	发送失败信号
		打开天线			发送成功信号
	充电	收起太阳能板			
	返回				
姿态调整子系统	就绪	转向目标	通信子系统	待机	写入缓存
	运行	发送指向完成信号			清空缓存并删除图像
	等待	接收任务完成信号		就绪	建立连接
		接收任务失败信号			发送连接异常信号
	阻塞	接收新任务信号		发送	断开
	指向完成	发送准备信号			发送传输完成信号

3.2　模型有效性验证

基于2.2节建立的深空探测器自主任务规划智能体交互环境对模型进行6次训练，模型训练过程中，损失值（Loss，用L表示）随着E的变化情况如图4所示，可知6次实验中，在E达到2 000时模型即可达到收敛状

态，说明模型收敛速度较快，本文构建的深空探测器智能体交互环境是有效的。

模型在训练过程中，步长（Step，用S表示）随着E的变化情况如图5所示，可知模型进行深空探测器任务规划的步长随着训练次数的增多而逐渐下降，在E达到1 500，初始状态不确定的情况下，步长值稳定在50以内，说明模型在不确定初始状态下能快速完成深空探测器规划任务。

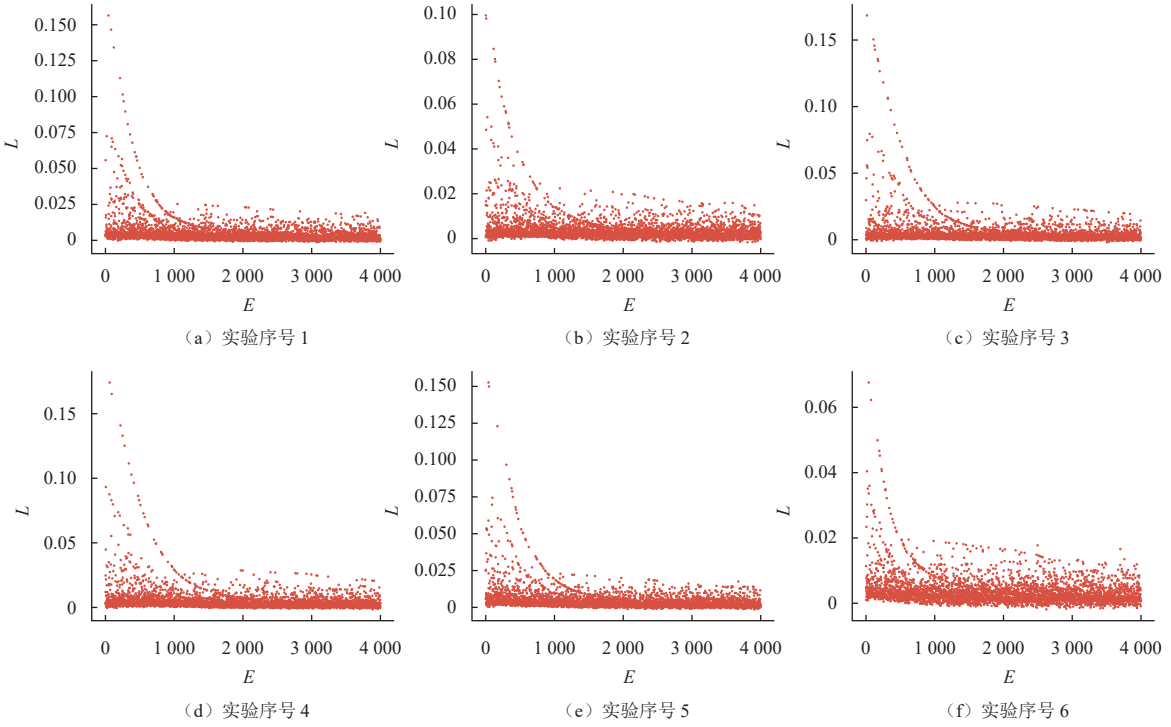

图 4　6次实验的损失变化情况

Fig. 4　Loss changes in six experiments

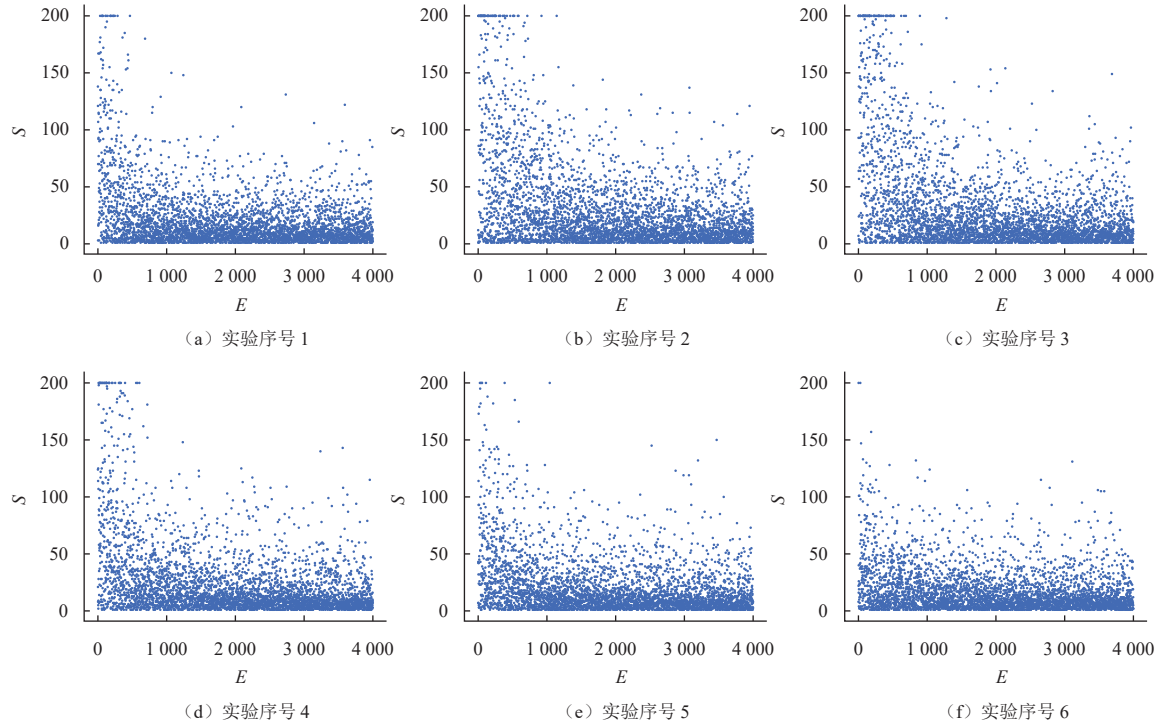

图 5　6次实验的步长变化情况

Fig. 5　Step changes in six experiments

采用本文方法对训练完成后的深空探测自主任务规划算法模型进行10次实验，每次实验进行1 000次规划，按照式（24）计算每一次规划结果的奖励（Reward，用R_{Episode}表示），其中anm为规划结果所包含的动作数量，r_m为规划结果中的第m个动作应的即时奖励。将10次实验得的奖励绘制成箱形图如图6所示，奖励值平均水平在2 000～2 500，箱宽较小，说明波动较小，稳定性较好。

$$R_{\text{Episode}} = \sum_{m=1}^{anm} r_m \tag{25}$$

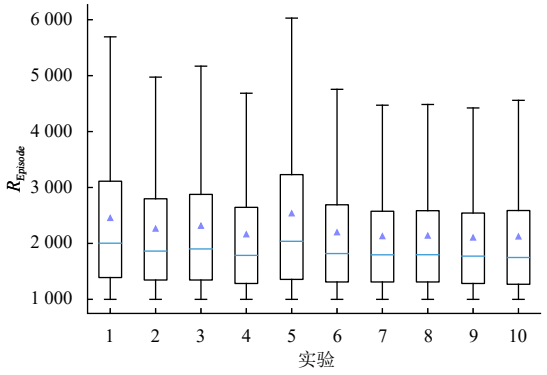

图 6　10次实验奖励值箱形图

Fig. 6　Reward box plot for ten experiments

对上述10次实验中R_{Episode}极大值最大的实验5对应的1 000次规划输出结果进行分析，1 000次规划的R_{Episode}值如图7所示。

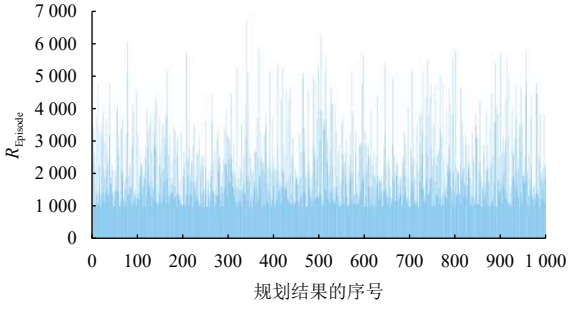

图 7　1 000次规划的奖励结果图

Fig. 7　Reward result of 1,000 times of planning

因为初始状态不确定，本文以规划步长较长的规划任务，即初始状态为电源子系统“待机”状态规划为例，选取其中奖励值最高的规划结果，即图7中第343次奖励值为6 757的结果，模型输出为一条从初始状态到任务终止状态的状态动作序列，其中状态用数值对表达，数值对的第1个数值代表子系统的序号，第2个数值代表子系统内的状态序号，动作用一个数值表示，将所得状态动作序列按照状态的第1个数值分别分给对应的子系统，相对顺序保持不变，得到输出结果如图8所示，以电源子系统为例，“（1，1）”代表“休眠”状态，“（1，2）”代表“运行”状态，“（1，3）”代表“充电”状态，“（1，4）”代表“完成”状态，其它子系统同理。为更清楚地展现深空探测器在该规划结果下的运行过程，根据图9绘制深空探测器各子系统的任务甘特图如图10所示。

```
This is a main thread@<ipython-input-8-3d4c6513e175>:3θ
Episode 343, reward sum 6 757.θ
输出结果:
电源子系统: (1,1), θ, (1, 2), 1, (1, 3), θ, (1,4)
姿态子系统: (2,1), θ, (2, 2), θ, (2, 5), θ, (2,3), θ
拍照子系统: (3,2), θ, (3, 3), θ, (3, 4), 2
通信子系统: (4,1), θ, (4, 2), θ, (4, 3)2
```

图 8　实验输出规划结果

Fig. 8　Experimental output planning results

图 9　各子系统的任务甘特图

Fig. 9　Task Gantt chart of each subsystem

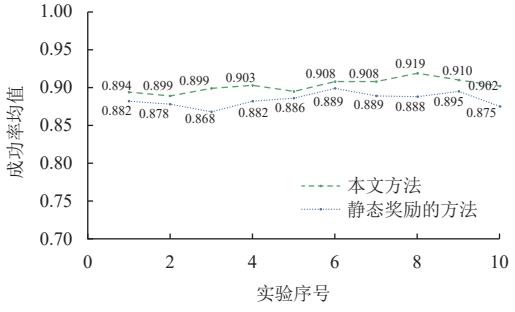

图 10　动态奖励和非动态奖励成功率结果对比

Fig. 10　Comparison of accuracy results of dynamic rewards and non-dynamic rewards

可以看出采用本文算法能实现深空探测器自主规划任务，并可根据R_{Episode}值自动决策输出其中较为合理的规划，从而证明了本文方法的有效性和适用性。

3.3　自主任务规划成功率

利用本文提出的方法进行10次实验，每次实验用该算法进行1 000次规划，同时为了全面评价该算法，将每次实验的1 000次规划分为10个小组，即每个小组100次，对每个小组利用式（26）所示的成功率（Success *rate*，用*Su*表示）对模型进行评价，取10个小组的A值

的均值作为该次实验的最终结果，结果如表4所示。取每个小组的100次规划总耗时的平均值为该小组平均单次规划耗时，取每次实验中10个小组的平均单次规划耗时的均值为该次实验的最终结果，如表5所示。

$$Su = \frac{success\_num}{total\_num} \tag{26}$$

其中：$success\_num$ 表示本组规划中成功的次数，$total\_num$ 表示本组规划的总次数。

表 4　10次实验的成功率结果

Table 4　Success rate results of 10 experiments

组号	1	2	3	4	5	6	7	8	9	10	均值
实验1	0.91	0.86	0.85	0.87	0.91	0.87	0.92	0.93	0.92	0.90	0.89
实验2	0.90	0.92	0.82	0.89	0.88	0.91	0.90	0.91	0.90	0.86	0.88
实验3	0.90	0.92	0.82	0.89	0.92	0.91	0.90	0.91	0.93	0.89	0.89
实验4	0.91	0.87	0.88	0.89	0.88	0.91	0.94	0.92	0.93	0.90	0.90
实验5	0.91	0.87	0.88	0.92	0.87	0.91	0.90	0.89	0.90	0.90	0.89
实验6	0.82	0.86	0.86	0.93	0.93	0.94	0.91	0.94	0.94	0.95	0.90
实验7	0.91	0.90	0.91	0.93	0.92	0.89	0.90	0.91	0.92	0.89	0.90
实验8	0.95	0.89	0.91	0.94	0.92	0.93	0.93	0.92	0.89	0.91	0.91
实验9	0.91	0.92	0.91	0.90	0.92	0.90	0.91	0.92	0.89	0.92	0.91
实验10	0.91	0.89	0.83	0.93	0.89	0.89	0.94	0.88	0.94	0.92	0.90

表 5　10次实验的单次规划用时情况

Table 5　Single planning time of 10 experiments

单位：s

组号	1	2	3	4	5	6	7	8	9	10	均值
实验1	0.109	0.111	0.122	0.121	0.117	0.113	0.112	0.132	0.110	0.114	**0.116 1**
实验2	0.116	0.112	0.115	0.110	0.107	0.112	0.109	0.108	0.107	0.112	0.110 8
实验3	0.112	0.111	0.111	0.112	0.110	0.111	0.109	0.113	0.109	0.113	0.111 1
实验4	0.110	0.112	0.113	0.113	0.113	0.106	0.107	0.106	0.111	0.111	0.110 2
实验5	0.112	0.113	0.114	0.116	0.116	0.119	0.114	0.121	0.116	0.12	**0.116 1**
实验6	0.116	0.111	0.113	0.110	0.112	0.107	0.110	0.108	0.112	0.112	0.111 1
实验7	0.113	0.113	0.112	0.111	0.113	0.110	0.114	0.108	0.112	0.113	0.111 9
实验8	0.110	0.109	0.111	0.114	0.110	0.110	0.112	0.110	0.106	0.110	0.110 2
实验9	0.109	0.107	0.111	0.110	0.112	0.112	0.109	0.109	0.112	0.110	0.110 1
实验10	0.112	0.111	0.111	0.111	0.110	0.112	0.115	0.112	0.112	0.112	0.111 8

通过表4的结果可以看出，本文所设计的基于动态奖励的策略梯度强化学习深空探测器自主任务规划方法在实验中规划的成功率均值在88%以上，并且在每次实验中各组的值波动均不大，说明基于本文方法的深空探测器任务规划具有较大的成功率与较好的稳定性。由表5可知，单次规划耗时在0.116 1 s（表5中加黑部分）内，说明规划效率较高，当规划失败时，立即重新规划生成新的规划结果，以保证深空探测器自主规划任务顺利完成。

3.4　动态奖励与非动态奖励策略梯度强化学习结果对比

本文针对深空探测器任务规划问题的时间、资源和时序约束通过引入动态奖励机制来对传统的策略梯度强化学习方法进行改进，为验证本文算法引入的动态奖励机制合理性和优越性，用本文所提的基于动态奖励策略梯度强化学习方法和改进前的基于非动态奖励的策略梯度强化学习方法展开了对比实验，动态奖励与非动态奖励实验分别进行10组，每组1 000次规划，对每组用式（26）中的成功率算数平均值对2组实验结果进行对比分析，如图10所示，可以看出改进后的动态奖励算法规划成功率平均在88%以上，并且高于静态奖励策略算法的平均结果，说明本文提出的动态奖励算法在深空探测器任务规划问题表现更佳。

4　结　论

本文以深空探测器自主任务规划为目标，提出了

一种动态奖励梯度策略方法，并以此为基础实现了深空探测器任务规划强化学习模型，动态奖励算法将资源约束、时间约束和时序约束融合在一起构建了即时奖励模型，并据此定义损失函数更新策略梯度，在保证多约束条件的前提下使模型能够从任意初始状态快速获得规划解。

仿真实验结果验证了本文算法可根据资源和约束动态调整任务规划结果的评价指标和奖励值，从而确保规划过程中算法向最优方向自动调整策略网络参数，使得决策模型的规划质量得到了提高。此外本文方法中构建的模型可以将任何一个状态作为输入起点，即该任务规划方法能在任意状态下开始规划而无需改变模型结构，因此增强了对不确定规划任务的适应性。对比实验结果显示本文提出的动态奖励策略与一般的静态奖励策略梯度算法相比规划成功率有明显的提升。本文方法为深空探测器自主任务规划与决策提供了一种新的解决方案，但对于动态奖励机制中的复杂资源约束以及进一步提高规划效率等问题仍需深入研究。

参　考　文　献

[1] 崔平远. 深空探测：空间拓展的战略制高点[J]. 人民论坛·学术前沿，2017(5)：13-18.
CUI P Y. Deep space exploration：strategic height of space expansion[J]. People's Forum. Academic Frontier，2017(5)：13-18.

[2] 于登云，张兴旺，张明，等. 小天体采样探测技术发展现状及展望[J]. 航天器工程，2020，29(2)：1-10.
YU D Y，ZHANG X W，ZHANG M，et al. Current status and prospects of small object sampling and detection technology[J]. Spacecraft Engineering，2020，29(2)：1-10.

[3] 赵凡宇，徐瑞，崔平远. 启发式深空探测器任务规划方法[J]. 宇航学报，2015，36(5)：496-503.
ZHAO F Y，XU R，CUI P Y. Heuristic mission planning method for deep space probes[J]. Journal of Astronautics，2015，36(5)：496-503.

[4] 姜啸，徐瑞，朱圣英. 基于约束可满足的深空探测任务规划方法研究[J]. 深空探测学报(中英文)，2018，5(3)：262-268.
JIANG X，XU R，ZHU S Y. Research on constrained satisfiable deep space mission planning method[J]. Journal of Deep Space Exploration，2018，5(3)：262-268.

[5] 姜啸，徐瑞，陈俐均. 深空探测器动态约束规划中的外延约束过滤方法研究[J]. 深空探测学报(中英文)，2019，6(6)：586-594.
JIANG X，XU R，CHEN L J. Study on extensive constraint filtering method for dynamic constraint planning of deep space detector[J]. Journal of Deep Space Exploration，2019，6(6)：586-594.

[6] 金颢，徐瑞，朱圣英，等. 适用于深空探测器的时间线转移路标启发式规划方法[J]. 宇航学报，2021，42(7)：862-872.
JIN B，XU R，ZHU S Y，et al. Time line transfer landmark heuristic planning method for deep space detector[J]. Journal of Astronautics，2021，42(7)：862-872.

[7] 赵宇庭，徐瑞，李朝玉，等. 基于动态智能体交互图的深空探测器任务规划方法[J]. 深空探测学报(中英文)，2021，8(5)：519-527.
ZHAO Y T，XU R，LI C Y，et al. Mission planning method for deep space probe based on dynamic agent interaction diagram[J]. Journal of Deep Space Exploration ，2021，8(5)：519-527.

[8] 王晓晖，李爽. 深空探测器约束简化与任务规划方法研究[J]. 宇航学报，2016，37(7)：768-774.
WANG X H，LI S. Research on constraint simplification and task planning method for deep space detector[J]. Journal of Astronautics，2016，37(7)：768-774.

[9] 冯小恩，李玉庆，杨晨，等. 面向自主运行的深空探测航天器体系结构设计及自主任务规划方法[J]. 控制理论与应用，2019，36(12)：2035-2041.
FENG X E，LI Y Q，YANG C，et al. Architecture design and autonomous mission planning for autonomous deep space exploration spacecraft[J]. Control Theory and Application，2019，36(12)：2035-2041.

[10] 王鑫，赵清杰，徐瑞. 基于知识图谱的深空探测器任务规划建模[J]. 深空探测学报(中英文)，2021，8(3)：315-323.
WANG X，ZHAO Q J，XU R. Modeling of deep space probe mission planning based on knowledge map[J]. Journal of Deep Space Exploration ，2021，8(3)：315-323.

[11] 李玉庆，徐敏强，王日新. 航天器观测重调度问题中的模糊性不确定因素及其处理[J]. 宇航学报，2009，30(3)：1106-1111.
Li Y Q，XU M Q，WANG R X. Fuzzy uncertainty factors in spacecraft observation rescheduling problem and their processing[J]. Journal of Astronautics，2009，30(3)：1106-1111.

[12] 贺东雷，冯小恩，雷明佳，等. 面向深空探测任务的实数遗传编码多星任务规划算法[J]. 控制理论与应用，2019，36(12)：2055-2064.
HE D L，FENG X E，LEI M J，et al. Real-number genetic encoding multistar mission planning algorithm for deep space mission[J]. Control Theory and Application，2019，36(12)：2055-2064.

[13] SUTTON R S，BARTO AG. Reinforcement learning：an introduction[J]. IEEE Transactions on Neural Networks，1998，9(5)：1054.

[14] 史兼郡，张进，罗亚中，等. 基于深度强化学习算法的空间站任务重规划方法[J]. 载人航天，2020，26(4)：469-476.
SHI J，ZHANG J，LUO Y Z，et al. Space station task replanning method based on deep enhanced learning algorithm[J]. Manned Space，2020，26(4)：469-476.

[15] 伍国威，崔本杰，曲耀斌，等. 基于深度强化学习的卫星实时引导任务规划方法及系统：中国，CN111950873A[P]. 2022-11-15.
WU G W，CUI B J，QU Y B，et al. Satellite real-time guidance mission planning method and system based on deep reinforcement learning：China，CN111950873A[P]. 2022-11-15.

[16] 郭林杰. 基于深度强化学习的跳跃式小行星探测器规划策略研究[D]. 哈尔滨：哈尔滨工业大学，2019.
GUO L J. Study on planning strategy of skip asteroid detector based on deep reinforcement learning [D]. Harbin：Harbin University of Technology，2019.

[17] FURFARO R，LINARES R. Deep learning for autonomous lunar landing[C]// Proceedings of AAS/AIAA Astrodynamics Specialist Conference. [S. l.]：AIAA，2018.

[18] HECKE K V，DE CROON G C H E，HENNES D，et al. Self-supervised learning as an enabling technology for future space

exploration robots：ISS experiments on monocular distance learning[J]. Acta Astronautica，2017：S0094576517302862.

[19] 徐瑞，李朝玉，朱圣英，等. 深空探测器自主规划技术研究进展[J]. 深空探测学报（中英文），2021，8(2)：111-123.
XU R，LI C Y，ZHU S Y，et al. Progress in deep space explorer autonomous planning[J]. Journal of Deep Space Exploration ，2021，8(2)：111-123.

[20] 刘志荣，姜树海. 基于强化学习的移动机器人路径规划研究综述[J]. 制造业自动化，2019，41(3)：90-92.
LIU Z R，JIANG S H. A review of path planning for mobile robots based on reinforcement learning[J]. Manufacturing Automation，2019，41(3)：90-92.

[21] 俞胜平，韩忻辰，袁志明，等. 基于策略梯度强化学习的高铁列车动态调度方法[J]. 控制与决策，2022(9)：2407-2417.
YU S P，HAN X C，YUAN Z M，et al. Dynamic scheduling method of high-speed train based on policy gradient reinforcement learning [J]. Control and Decision, 2022(9)：2407-2417.

[22] 张淼，张琦，刘文韬，等. 一种基于策略梯度强化学习的列车智能控制方法[J]. 铁道学报，2020，42(1)：69-75.
ZHANG B，ZHANG Q，LIU W T，et al. A train intelligent control method based on strategic gradient enhanced learning[J]. Journal of Railways，2020，42(1)：69-75.

[23] 周飞燕，金林鹏，董军. 卷积神经网络研究综述[J]. 计算机学报，2017，40(6)：1229-1251.
ZHOU F Y，JIN L P，DONG J. A review of convolution neural networks[J]. Journal of Computer Science，2017，40(6)：1229-1251.

[24] 李高杨，吕晓鹏，张星. 基于强化学习的交通信号控制及深度学习应用[J]. 人工智能，2020(3)：84-9.
LI G Y，LV X P，ZHANG X. Application of traffic signal control and in-depth learning based on reinforcement learning[J]. Artificial Intelligence，2020(3)：84-9.

作者简介：
毛维杨(1997–)，硕士研究生，主要研究方向：人工智能，深空探测器自主任务规划。
通信地址：昆明理工大学信息工程与自动化学院(650500)
E-mail：1518887260@qq.com
王彬(1977–)，博士，副教授，主要研究方向：实时控制、智能控制、深空探测、智能信息处理。**本文通信作者**。
通信地址：昆明理工大学信息工程与自动化学院(650500)
E-mail：wangbin@kust.edu.cn

An Autonomous Planning Method for Deep Space Exploration Tasks in Reinforcement Learning Based on Dynamic Rewards

MAO Weiyang[1]，WANG Bin[1,2]，LIU Jingxing[1]，XIONG Xin[1]
（1. Faculty of Information Engineering & Automation, Kunming University of Science & Technology, Kunming 650500, China；
2. Yunnan key Laboratory of Artificial Intelligence, Kunming University of Science & Technology, Kunming 650500, China）

Abstract：Aiming at the characteristics of multi-system parallelism and the need to meet various constraints in the proceAiming at the characteristics of multi-system parallelism and the need to meet various constraints in the process of autonomous mission planning of deep space detectors, a reinforcement learning task autonomous planning model construction method for deep space detectors was proposed based on dynamic rewards, and a deep space detector agent was established. In the interactive environment, a policy network and a loss function integrating resource constraints, time constraints and timing constraints were constructed, and a dynamic reward mechanism was proposed to improve the traditional policy gradient learning method. The simulation results show that the method in this paper could realize autonomous task planning. Compared with the static reward policy gradient algorithm, the planning success rate and planning efficiency were significantly improved, and the method could start planning in any state without changing the model structure, which improved the accuracy of the algorithm. This method provides a new solution for autonomous mission planning and decision-making of deep space probes.

Keywords：deep space exploration；task planning；policy gradient；reinforcement learning；dynamic reward

Highlights：

- A reinforcement learning interactive environment for deep space probe agents is built.
- The traditional policy gradient reinforcement learning method is improved by constructing a loss function which integrates resource, time and timing constraints for task planning of deep space detectors.
- A dynamic reward mechanism is proposed.
- A deep space exploration task planning model with random initial states is presented.

[责任编辑：宋宏，英文审校：宋利辉]