面向多波束卫星系统的波束跳变与覆盖控制联合优化算法

许国良[1,2]，谭峰[1]，冉泳屹[1,2]，陈丰[1]
（1. 重庆邮电大学通信与信息工程学院，重庆 400065；2. 重庆邮电大学电子信息与网络工程研究院，重庆 400065）

摘 要：为了提升多波束卫星（MBS）系统的性能，提出了一种基于深度强化学习联合优化 MBS 的波束跳变和覆盖控制（BHCC）算法。首先，将多波束卫星中的资源分配问题转换为多目标优化问题，以最大化多波束卫星系统的系统吞吐量，最小化丢包率；其次，将多波束卫星环境表示为多维矩阵，并将目标问题建模为考虑随机通信需求的马尔可夫决策过程；最后，结合深度强化学习强大的特征提取能力和学习能力对目标问题进行求解。此外，提出了一种单智能体轮询复用机制以减少搜索空间，降低收敛难度，加速 BHCC 算法的训练。仿真结果表明，相对于遗传算法、贪婪算法及随机算法，BHCC 算法不仅能提高 MBS 的吞吐量，而且能降低系统的丢包率；相对于不考虑自适应波束覆盖范围的深度强化学习算法，BHCC 算法在不同通信场景下的性能更优异。

关键词：多波束卫星；深度强化学习；波束跳变技术；波束覆盖控制
中图分类号：TN92
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023076

Joint beam hopping and coverage control optimization algorithm for multibeam satellite system

XU Guoliang[1,2], TAN Feng[1], RAN Yongyi[1,2], CHEN Feng[1]
1. School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
2. Institute of Electronic Information and Network Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: To improve the performance of multibeam satellite (MBS) systems, a deep reinforcement learning-based algorithm to jointly optimize the beam hopping and coverage control (BHCC) algorithm for MBS was proposed. Firstly, the resource allocation problem in MBS was transformed to a multi-objective optimization problem with the objective maximizing the system throughput and minimizing the packet loss rate of the MBS. Secondly, the MBS environment was characterized as a multi-dimensional matrix, and the objective problem was modelled as a Markov decision process considering stochastic communication requirements. Finally, the objective problem was solved by combining the powerful feature extraction and learning capabilities of deep reinforcement learning. In addition, a single-intelligence polling multiplexing mechanism was proposed to reduce the search space and convergence difficulty and accelerate the training of BHCC. Compared with the genetic algorithm, the simulation results show that BHCC improves the throughput of MBS and reduces the packet loss rate of the system, greedy algorithm, and random algorithm. Besides, BHCC performs better in different communication scenarios compared with a deep reinforcement learning algorithm, which do not consider the adaptive beam coverage.

Keywords: multibeam satellite, deep reinforcement learning, beam hopping technology, beam coverage control

收稿日期：2022−11−09 ；**修回日期：**2023−02−01
通信作者：冉泳屹，ranyy@cqupt.edu.cn
基金项目：国家自然科学基金资助项目（No.62171072，No.62172064，No.62003067）；重庆市自然科学基金资助项目（No.cstc2021jcyj-msxmX0586）
Foundation Items: The National Natural Science Foundation of China (No.62171072, No.62172064, No.62003067), The Natural Science Foundation of Chongqing (No.cstc2021jcyj-msxmX0586)

0　引言

由于卫星通信资源的稀缺性和高成本，提高资源利用率成为满足日益增长的通信需求的关键[1]。已有研究表明，许多多波束卫星（MBS, multibeam satellite）系统中采用的波束跳变算法无法满足热点小区的需求，而非热点小区的波束容量被浪费。这导致卫星运营商和服务提供商的双重损失：一方面是未满足的需求所对应的收入损失，另一方面是未使用容量所对应的投资损失。所以，急需一种有效的波束跳变策略，用以提升卫星系统的资源利用率，优化卫星性能[2-3]。

为了有效地匹配有限的波束资源与非均匀的小区通信需求，研究人员应用启发式、凸优化和深度强化学习等优化方法对 MBS 中的波束跳变算法展开了大量研究。文献[4]指出波束跳变技术是匹配通信需求的有效解决方法，并且波束跳变可以处理不均匀的空间分布和时变的业务分布。基于此，文献[5]针对用户业务的突发性和时变性提出了一种基于遗传算法的波束调度方案，该方案可以实现智能调度波束，并且在一定程度上满足用户需求，但遗传算法复杂度较高，寻找最优解所需成本较大。文献[6]采用贪婪算法根据业务需求的分布灵活地分配星载资源，相较于文献[5]所采用的遗传算法，贪婪算法的复杂度更低，但是贪婪算法只基于业务分布情况提供服务，没有考虑服务公平性，所以仍然存在局限性。文献[7]提出了一种改进的布谷鸟算法来动态调度卫星波束，可以一定程度地提高用户的加权收益。文献[8]研究了多波束卫星系统中非正交多址接入（NOMA, non-orthogonal multiple access）和跳束的潜在协同效应，并将波束跳变和非正交多址技术联合功率分配问题定义为混合整数非凸规划，在多项式时间复杂度的时隙基础上提出了一种贪婪方案求解该问题。但是该工作并没有明确波束位置和系统性能的具体关系。针对这个问题，文献[9]发现并证明了波束位置数与排队时延呈负相关关系，并将波束位置划分问题转化为 P 中心问题，以最少的波束覆盖更多用户。

上述工作中应用传统启发式算法和凸优化算法虽然能在一定程度上提升系统性能，但是仍然存在一些不足。首先，上述工作大都未考虑地面小区在相邻时刻内的业务存在时间相关性，忽略了长期累积收益；其次，卫星性能的提升要综合考虑多个指标，上述工作大多偏向于单目标优化；最后，一旦卫星环境发生改变，启发式算法需要重新进行迭代才能找到次优解，算法的泛化性较弱。

深度强化学习（DRL, deep reinforcement learning）算法拥有较强的特征处理能力、学习能力、泛化能力和对抗稳健性[10-11]。因此，深度强化学习技术在多波束卫星资源配置领域的应用得到了相当多的关注[12]。Hu 等[12]将 DRL 算法应用于优化目标为降低服务阻塞概率的多波束卫星场景，提高了卫星系统的承载流量和频谱效率。针对系统资源利用问题，Liu 等[13]通过 DRL 技术对卫星进行动态波束调度，并且在阻塞率和吞吐量等方面较基准算法有一定提升。Luis 等[14]提出了一种基于深度强化学习的多波束卫星系统功率分配方法，利用近端策略优化算法来优化分配策略，最大限度地减少了未满足的系统需求和功耗。上述集中式强化学习算法会因为波束数量增加使可选空间增大，造成维度灾难问题，使算法难以收敛。针对此问题，Liao 等[15]采用了多智能体强化学习算法为多波束卫星系统动态功率分配问题提供了有效的解决方案，可以一定程度上避免“维度灾难”问题，同时使波束提供的负载容量与小区业务需求匹配度增加。

上述工作都致力于寻找最优的波束跳变策略，具有较大的局限性。由于每个小区的通信需求都是时变的，因此固定的波束覆盖半径可能会在非热点小区造成卫星传输资源的浪费，或对热点小区造成拥塞。因此，在设计卫星资源调度策略时，不仅需要考虑波束跳变策略及功率分配，还需要利用波束半径的灵活度，合理地调整波束覆盖半径。

本文提出了一种基于 DRL 的波束跳变和覆盖控制（BHCC, beam hopping and coverage control）联合优化算法，首先，将目标问题定义为提高系统吞吐量和降低丢包率的多目标优化问题；其次，将多波束卫星环境表示为多维矩阵，并且将目标问题建模为一个考虑随机通信需求的马尔可夫决策过程（MDP, Markov decision process）；最后，结合深度强化学习强大的特征提取能力和学习能力对目标问题进行求解。为了减小搜索空间，提出了“单智能体轮询复用”机制，即在训练过程中将波束作为一个独立的智能体，并对智能体进行训练，训练结束后，多个波束以轮询的方式共用一个智能体的算法模型，以实现同时控制多波束的目的。本文进行了大量实验，将 BHCC 算法与基于深度强化学习且固定波束半径的波束跳变算法，基于遗传算法、贪婪算法及随机算法的波束跳变及覆盖控制联合优化算法进行了性能对比。由仿真结果可知，

所提BHCC算法在系统吞吐量和丢包率方面相比其他算法具有更好的性能。本文的贡献总结如下。

1) 提出的 BHCC 算法从波束跳变和波束覆盖半径联合优化的角度求解多波束卫星的波束跳变问题，充分利用波束覆盖范围的灵活性，以匹配有限的波束传输容量与非均匀和时变的小区业务需求。

2) 为了减小搜索空间，降低收敛难度，提出“单智能体轮询复用”机制，将多波束问题转化为单波束问题，并且通过状态伪更新技术使各个波束在决策时考虑其他波束的行为从而使总体决策最优。

1 系统模型和目标优化问题

1.1 系统模型

本文针对卫星下行链路建立了多波束卫星系统模型，如图 1 所示。设地球同步轨道卫星工作于 Ka 频段，由于地球同步轨道卫星相对于地面静止，因此不考虑卫星的移动性[12]。卫星携载 K 个可转向且可动态调整覆盖范围的波束。此外，将卫星覆盖范围内的一个区域划分为 N 个小区（$K \ll N$）。卫星上携载一个缓存区，用于记录各个小区的通信需求。K 个波束以时分复用的方式覆盖 N 个小区，系统链路采用高斯白噪声信道。

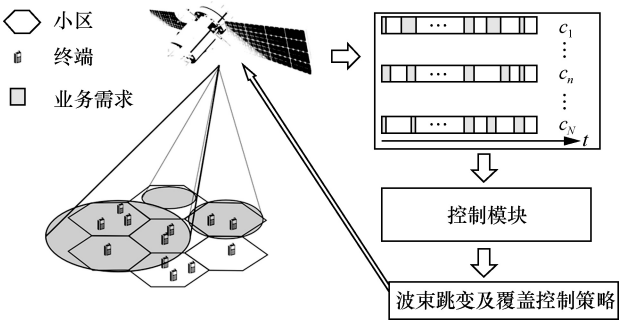


图 1 多波束卫星系统模型

波束集合表示为 $\mathcal{K}=\{k \mid k=1,2,\cdots,K\}$，小区集合表示为 $\mathcal{C}=\{c_n \mid n=1,2,\cdots,N\}$。小区 c_n 在 t 时刻的数据包请求数量 $\phi_{t,0}^{c_n}$ 服从泊松分布，被记录于缓存区的待服务数据包向量为 $\boldsymbol{E}_{t,n}^{\mathrm{T}}=[\phi_{t,l}^{c_n} \mid l=0,1,\cdots,\ l_{\mathrm{th}+1}]$，其中，$l$ 是数据包的排队时延，l_{th} 是数据包可容忍的最大排队时延。如果排队时延超过了可容忍的最大排队时延，即 $l=l_{\mathrm{th}+1}$，则该数据包将被丢弃。

小区 c_n 在 t 时刻待服务的数据包总数量为 $\lambda_t^{c_n}$，则相邻时刻小区 c_n 待服务数据包的数量可以表示为

$$\lambda_t^{c_n}=\lambda_{t-1}^{c_n}+\phi_{t,0}^{c_n}-\Theta_t^{c_n}-\phi_{t,l_{\mathrm{th}+1}}^{c_n} \tag{1}$$

其中，$\Theta_t^{c_n}$ 为 t 时刻小区 c_n 被传输的数据包数量。

根据 ITU-R S.672-4，离轴角 δ 和发射天线增益关系为

$$G_{\mathrm{tx}}(\delta)=\begin{cases} G_{\mathrm{m}},\ \delta<\delta_{\mathrm{b}} \\ G_{\mathrm{m}}-3\left(\dfrac{\delta}{\delta_{\mathrm{b}}}\right)^2,\ \delta_{\mathrm{b}}\leqslant\delta<a\delta_{\mathrm{b}} \\ G_{\mathrm{m}}+L_{\mathrm{s}},\ a\delta_{\mathrm{b}}\leqslant\delta<b\delta_{\mathrm{b}} \\ \max\left\{G_{\mathrm{m}}+L_{\mathrm{s}}+20-25\lg\left(\dfrac{\delta}{\delta_{\mathrm{b}}}\right),0\right\},\ \text{其他} \end{cases} \tag{2}$$

其中，$a=2.88$，$b=6.32$，常数 $L_{\mathrm{s}}=-25\ \mathrm{dB}$，$\delta$ 为离轴角，δ_{b} 为半波瓣增益角，G_{m} 为卫星发射天线的最大增益[16]。

$$G_{\mathrm{m}}=10\lg\left[4.93\left(\frac{70}{\delta_{\mathrm{b}}}\right)^2\right] \tag{3}$$

由式(3)可知，增大波束的半波瓣增益角会使发射天线的最大增益减小，波束半波瓣增益角决定了波束覆盖范围，所以波束覆盖范围和波束容量呈负相关。

信道增益可表示为 $h=G_{\mathrm{tx}}G_{\mathrm{r}}L_{\mathrm{f}}$，其中，$G_{\mathrm{r}}$ 为接收天线增益，L_{f} 为自由空间损耗。所以，波束 k 到小区 c_n 的信道容量可表示为

$$F_k^{c_n}=\varepsilon_k^{c_n}W_k\mathrm{lb}\left(1+\frac{h^{k,n}P_k}{W_kN_0+\sum\limits_{i\in\mathcal{K},i\neq k}h_t^{i,n}P_i}\right) \tag{4}$$

其中，P_k 为波束 k 的发射功率；N_0 为噪声功率谱密度；W_k 为波束 k 的带宽；$h^{k,n}$ 为波束 k 到小区 c_n 的信道增益；$\varepsilon_k^{c_n}$ 表示小区 c_n 是否在波束 k 的覆盖范围内，如果在，则 $\varepsilon_k^{c_n}=1$，否则 $\varepsilon_k^{c_n}=0$。小区 c_n 在时刻 t 内实际传输的数据量 $\Theta_t^{c_n}=\min\{F_k^{c_n},\lambda_t^{c_n}\}$。

1.2 目标优化问题

本文的目标是根据卫星覆盖范围内各个小区的通信需求动态地选择波束跳变和覆盖控制策略，使卫星缓存区中待服务的数据包更多地被波束传输，而不是因为超过最大容忍时延而被缓存区清除，从而最大限度地提高系统吞吐量，最小化丢包率。因此该优化问题可表示为

$$\max\quad \mathrm{P1}=\sum_{n=1}^{N}\Theta_t^{c_n}$$

$$\min\quad \mathrm{P2}=\frac{\sum\limits_{n=1}^{N}\phi_{t,l_{\mathrm{th}+1}}^{c_n}}{\sum\limits_{n=1}^{N}\Theta_t^{c_n}+\sum\limits_{n=1}^{N}\phi_{t,l_{\mathrm{th}+1}}^{c_n}}$$

$$\text{s.t.}\quad \begin{aligned} &\text{C1}:\sum_{k=1}^{K} P_k \leqslant P_{\text{tot}} \\ &\text{C2}: P_k \leqslant P_{\max} \\ &\text{C3}: v_t^k \leqslant R_{\max}, \forall c_n, t \end{aligned} \tag{5}$$

其中，P1为最大化系统吞吐量；P2为最小化丢包率；约束 C1 表示分配给所有波束的功率之和不能超过系统总功率；约束 C2 表示单个波束的发射功率不能超过单个波束能承载的最大功率；约束 C3 表示每个波束尺寸不能超过波束尺寸阈值（波束的最大宽度及最小宽度决定波束尺寸阈值），v_t^k 表示 t 时刻波束 k 的半径。

2　算法介绍

本节详细介绍了基于深度强化学习的波束跳变和波束覆盖控制联合优化算法。

2.1　MDP 模型

由式(1)可知，地面各个小区的业务需求在相邻时刻具有相关性，并且下一时刻的状态仅与上一时刻的状态和决策相关，所以多波束卫星系统的波束跳变问题可转换为一个顺序决策问题，也可以表示为离散时间马尔可夫决策过程。MDP 模型如图 2 所示。

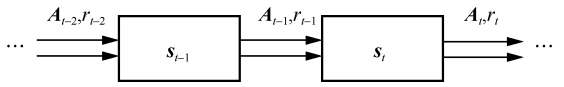


图 2　MDP 模型

MDP 可以用 $(\boldsymbol{s}_{t-1}, \boldsymbol{A}_{t-1}, r_{t-1}, \boldsymbol{s}_t)$ 来描述，其中，$\boldsymbol{s}$ 表示环境的状态空间，$\boldsymbol{A}$ 表示智能体采取的行动空间，r 表示奖励函数。智能体通过不断地试错学习，最终通过环境状态得到最佳行动，并将长期累积收益最大化。MDP 要素定义如下。

状态空间 $\boldsymbol{s}$。状态信息描述智能体所在的环境，智能体根据状态信息执行相应的动作。多波束卫星系统的状态信息由缓存信息来确定，卫星缓存区中记录的信息是各个小区随机产生的待服务的数据包数量。因此，状态空间定义为

$$\boldsymbol{s}_t = \{\boldsymbol{E}_{t,1}, \boldsymbol{E}_{t,2}, \cdots, \boldsymbol{E}_{t,N}\} = \begin{bmatrix} \phi_{t,l_{\text{th}+1}}^{c_1} & \phi_{t,l_{\text{th}}}^{c_1} & \cdots & \phi_{t,0}^{c_1} \\ \phi_{t,l_{\text{th}+1}}^{c_2} & \phi_{t,l_{\text{th}}}^{c_2} & \cdots & \phi_{t,0}^{c_2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{t,l_{\text{th}+1}}^{c_N} & \phi_{t,l_{\text{th}}}^{c_N} & \cdots & \phi_{t,0}^{c_N} \end{bmatrix} \tag{6}$$

行动空间 $\boldsymbol{A}$。由于本文的目标是共同优化波束跳变和覆盖控制策略，动作空间需要包括上述 2 个参数。因此，本文将 BHCC 算法输出的一维动作映射到实际的二维参数，表示为

$$\boldsymbol{a}_k = [n_t^k, v_t^k] \tag{7}$$

其中，n_t^k 表示波束 k 在 t 时刻选择的中心小区，v_t^k 表示 t 时刻波束 k 的半径。

奖励函数 r。在 BHCC 算法中，目标是最大化系统吞吐量和最小化丢包率。因此，可以用吞吐量和数据包丢失量来定义奖励函数，表示为

$$r = \sum_{n=1}^{N} \Theta_t^{c_n} - \sum_{n=1}^{N} \phi_{t,l_{\text{th}+1}}^{c_n} \tag{8}$$

2.2　单智能体轮询复用机制及状态伪更新过程

文献[12]中提出的以卫星作为智能体进行集中训练的 DRL 方法会随着波束数量和小区数量的增加而变得难以实现。一方面，从 N 个小区中选择 K 个波束进行服务，那么可供选择的方案有 $\text{C}_N^K = \dfrac{N!}{K!(N-K)!}$ 种。另一方面，假设卫星有 K 个波束，每个波束有 D 种可选半径，可用波束半径方案有 K^D 种。因此，整个行动空间大小为 $\text{C}_N^K \times K^D$，这对 DRL 算法的影响是灾难性的。

为了避免上述问题，本文提出了一种“单智能体轮询复用”机制，将多波束问题转化为单波束问题，即在训练过程中，只训练一个智能体，它在同一时刻仅可供一个波束使用。在决策过程中，多个波束以轮询的方式使用这个训练好的算法模型，得到其覆盖策略，这需要在每个波束做出决策后进行一次状态伪更新，即假设卫星已经执行了波束 k 的动作 a_k，并将其状态矩阵 $\boldsymbol{s}_k$ 更新为 $\boldsymbol{s}_{k+1}$，得到奖励 r_k，并将 $\boldsymbol{s}_{k+1}$ 作为波束 $k+1$ 的状态矩阵。如果不进行状态伪更新，多个波束将会输出同一种策略。对于单个波束而言，其可供选择的服务方案仅有 ND 种，这有效避免了以卫星为智能体带来的维度灾难问题，极大地缩小了算法的训练难度。

基于 BHCC 算法的多波束卫星系统模型如图 3 所示，该模型展示了状态重构过程、BHCC 算法的决策及训练过程和多波束卫星环境。首先，进行状态信息重构，将多波束卫星缓存区记录的随机产生的通信需求映射为多维矩阵作为状态信息 $\boldsymbol{s}_t$；其次，将状态矩阵 $\boldsymbol{s}_t$ 输入算法模型，得到动作向量 $\boldsymbol{A}_t = (a_1, a_2, \cdots, a_K)$；最后，将 $\boldsymbol{A}_t$ 返回多波束卫星环

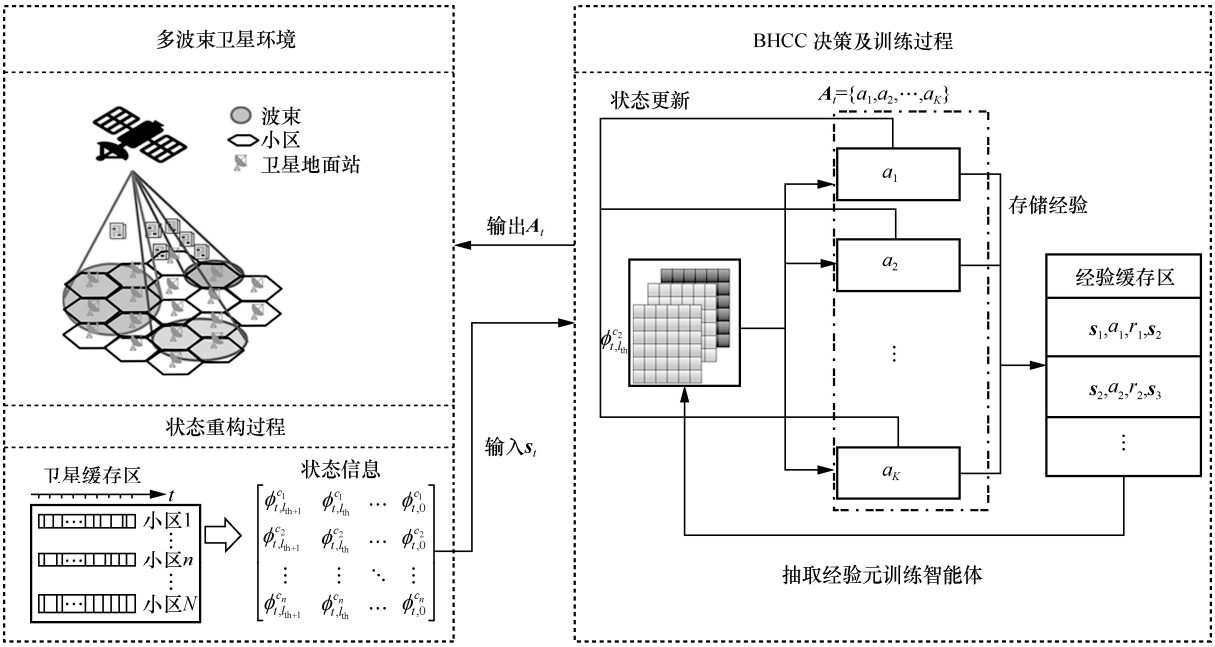


图 3 基于 BHCC 算法的多波束卫星系统模型

境，执行动作得到下一时刻的状态矩阵 $\boldsymbol{s}_{t+1}$，并更新神经网络参数。

2.3 网络的结构与训练

本节主要从深度 Q 网络结构和训练方面描述 BHCC 算法。

深度 Q 网络结构。由于式(7)中定义的动作空间是离散的，本文使用了深度 Q 网络学习方法[14]让智能体学习策略。深度 Q 网络是动作价值网络，它可以评估某个状态下进行某种行为的预期累积回报。与传统的强化学习（如 Q 学习）不同，深度强化学习的动作价值函数是通过神经网络而不是 Q 表来实现的。通过训练卷积神经网络得到最佳动作价值函数，并据此得到最优策略网络目标值。

$$Q^*(\boldsymbol{s},a)=\max_{\pi}\mathbb{E}\left[r_t+\gamma r_{t+1}+\cdots\mid \boldsymbol{s}_t=\boldsymbol{s},a_t=a,\pi\right] \quad (9)$$

其中，γ 是折扣因子。

深度Q网络训练。当采用 CNN 来近似 Q 值函数时，深度强化学习算法的训练结果被高估或者不稳定甚至发散。为了避免高估问题，本文的 BHCC 算法中采用了双网络技术（策略网络 Q^* 和目标网络 Q^-）[7]。为了使训练结果趋于稳定，本文在 BHCC 算法中加入了记忆重放技术。在记忆池 $\mathfrak{R}$ 中记录训练经验。训练时，从缓存区中随机抽取一批经验元基于贝尔曼方程计算出目标值[13]。目标值计算如下

$$y_t=r_t+\gamma\max Q^-\left(\boldsymbol{s}_{t+1},a;\theta^-\right) \quad (10)$$

其中，θ^- 是目标网络的参数，每 $\mathfrak{I}$ 步随策略网络参数 θ 更新一次。根据 Q^* 中的目标值，t 时刻网络的损失值 $L_t(\theta_t)$ 为[15]

$$L_t(\theta_t)=\mathbb{E}_{(\boldsymbol{s}_t,a_t,r_t,\boldsymbol{s}_{t+1})\sim U(\mathfrak{R})}\left(y_t-Q^*(\boldsymbol{s},a;\theta_t)\right)^2 \quad (11)$$

其中，θ_t 是 t 时刻的策略网络参数。

2.4 BHCC 算法步骤和流程

本文提出的BHCC算法基本步骤如算法1所示。

算法 1 BHCC 算法

输入 状态信息 $\boldsymbol{s}_t$

输出 动作 $\boldsymbol{A}_t$

1) 参数初始化。初始化目标网络参数 θ^-、策略网络参数 θ、记忆池 $\mathfrak{R}$、探索参数 ε、采样批量数 ℓ。

2) 接收初始化状态 $\boldsymbol{s}$，由各小区生成通信需求。

3) for epsiode=1 to max_epsiode do

4) 根据目前的状态选择动作，执行动作并进入下一状态。

5) 根据式(8)计算奖励。

6) 将当前状态、动作、奖励以及下一状态四要素存入记忆池中。

7) 更新卫星缓存区。

8) 更新网络参数 θ，每 $\mathfrak{I}$ 步更新一次 θ^-

9) end for

3　仿真分析

3.1　场景和参数设置

仿真基于 Python3.6 平台进行，参数如下：Intel® Core(TM) i5-1135G7，8 GB RAM，Intel® UHD Graphics630（应用于深度强化学习训练阶段）。仿真场景为工作于 Ka 频段、运行于地球同步轨道的多波束卫星，主要参数依据地球同步轨道卫星移动无线接口规范设置。系统内设置 30 个小区，卫星上携载 7 个指向和覆盖范围可变的波束，系统和算法仿真参数如表 1 所示。每个时刻地面小区的通信需求量都遵循泊松分布。由于每个小区中用户量存在差异，各个小区的通信需求量也有所不同。为了反映通信需求的时变特性，每个小区请求的通信需求在不同时刻也不完全相同。具体地，单个小区的业务需求为 0～150 Mbit/s，但为了业务需求不同的通信场景，仿真时会控制各小区总业务需求。本文以 500 个测试结果的统计平均值作为评价指标。

表 1　　系统和算法仿真参数

仿真参数	取值
卫星高度/km	35 786
Ka 波段频率/GHz	20
总带宽 B_{tot} /MHz	500
波束个数 K /个	7
卫星总功率 P_{tot} /dBW	34.5
波束发射功率 P_k /dBW	26
小区数 N /个	30
最大发射天线增益 G_{m} /dBi	40.3
自由空间损耗 L_{f} /dB	209.6
最大接收天线增益 G_{r} /dBi	31.6
最小波束覆盖半径/m	500
最大波束覆盖半径/m	1 000
时隙持续时间/ms	2
重放缓存区大小	100 000
采样批量	128
折扣因子 γ	0.9

3.2　性能指标

为了评估 BHCC 算法的性能，本文定义了以下性能评估指标。

系统吞吐量：某时刻内由系统传输的数据包总数。

丢包率：由于排队时间超过最大容忍时延而丢失的数据包数量占总数据包数量的百分比。

3.3　不同算法性能分析

3.3.1　BHCC 算法收敛性

图 4 给出了 BHCC 算法在总业务需求为 3 000 Mbit/s 时目标奖励值的收敛效果。从图 4 可以看出，算法在训练过程初始时刻，奖励值比较小，因为此时神经网络的参数是随机初始化的，不能准确地预测累积收益和选择能得到最大奖励的动作，当训练约 30 000 次后，奖励值逐渐收敛到最大值，并趋于稳定。

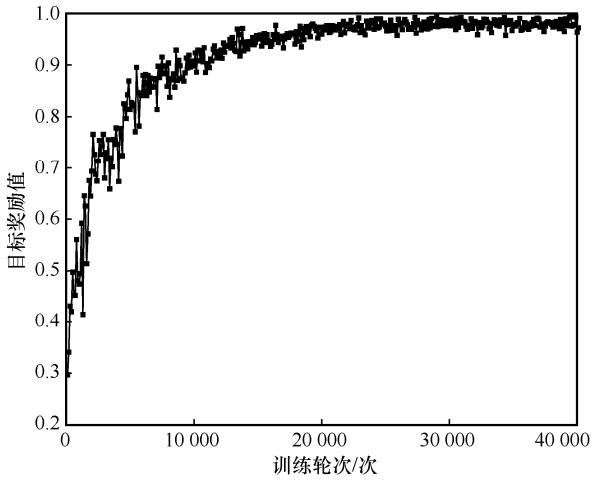


图 4　BHCC 算法收敛趋势

3.3.2　性能分析

为了验证所提出的 BHCC 算法在多波束卫星系统中的性能，本文将基于 BHCC 算法与随机算法、贪婪算法和遗传算法的波束跳变及覆盖联合优化算法进行了对比，具体介绍如下。

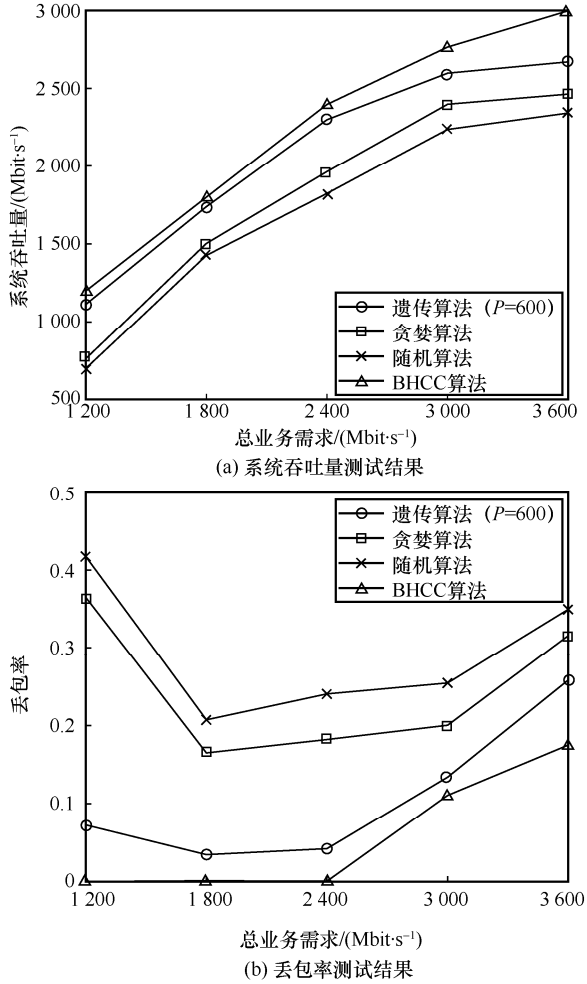
随机算法。每个决策时刻随机选择波束覆盖的中心小区，并且随机确定波束尺寸[4]。

贪婪算法。每个决策时刻计算一次各个小区待服务业务的总需求量，选择业务需求最大的 7 个小区作为中心小区，在波束尺寸阈值内随机选择波束尺寸[6]。

遗传算法。优化目标设为最大化传输数据量，最小化传输失败的数据量。算法的父代个数为 60，变异概率为 0.001，交叉概率为 0.06，迭代次数 P=600 。在每个决策时刻，根据各小区的通信需求，通过迭代选择每个波束覆盖的中心小区，并确定波束尺寸[5]。

不同总业务需求下的测试结果如图 5 所示。从图 5(a)可以看出，随着总业务需求从 1 200 Mbit/s 增长至 3 600 Mbit/s，几种算法的系统吞吐量均

增加。当总业务需求较小时，BHCC 算法和遗传算法的性能比较接近，但当总业务需求较大时，BHCC 算法的性能明显优于遗传算法。本文提出的基于强化学习的 BHCC 算法的决策是基于长期累积收益所做出的，而其他几种算法都是根据当前时刻状态做出的最优解或者次优解。



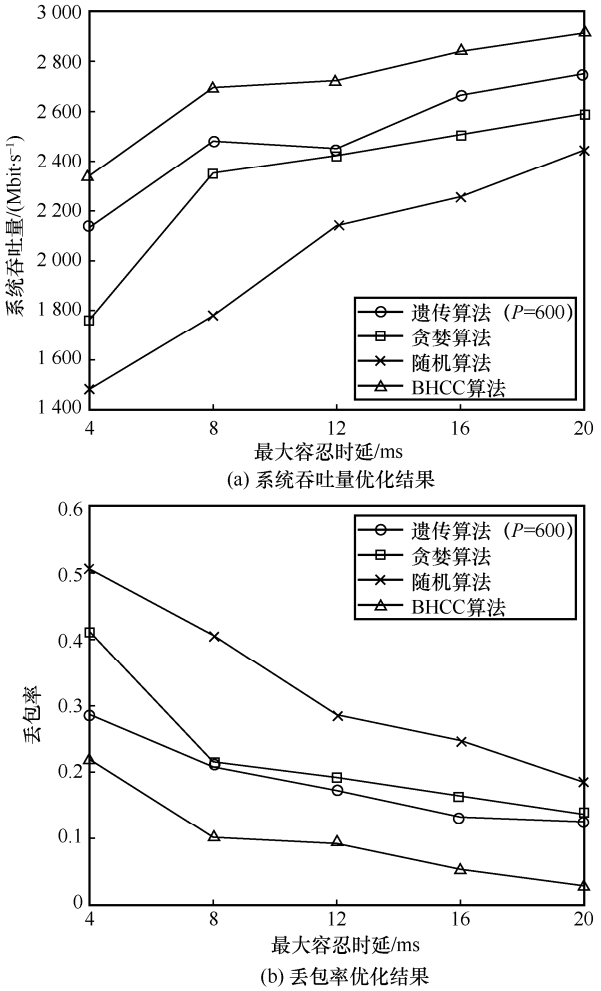
(a) 系统吞吐量测试结果

(b) 丢包率测试结果

图 5　不同总业务需求下的测试结果

从图 5(b)可以看出，当总业务需求为 1 200 Mbit/s 时，贪婪算法仅服务业务需求相对较大的小区，但由于总业务需求较小，导致业务需求较大的小区和其他小区需求差值并不大，因此丢包率较高。随机算法具有较大随机性，在小业务需求时的性能较差。遗传算法在进行迭代之后可以获得较好的性能，但一旦小区需求发生变化，遗传算法需要重新迭代才能得到最优解，算法复杂度较高。

训练结束后，当总业务需求为 3 000 Mbit/s 时，不同最大容忍时延下的测试结果如图 6 所示。数据包最大容忍时延表示数据包能够接受在队列中排队的最长时间，如果数据包在队列中的等待时延超过最大容忍时延，那么该数据包将会传输失败。从图 6 可以看出，随着数据包最大容忍时延的增大，基于不同算法的波束跳变和波束覆盖控制方案的系统吞吐量总体呈上升趋势，丢包率呈下降趋势。但在最大容忍时延较小的场景下，几种算法性能差异较大。基于深度强化学习的 BHCC 算法在训练过程中积累了足够的经验，它可以在各种场景下做出更适应的决策。



(a) 系统吞吐量优化结果

(b) 丢包率优化结果

图 6　不同最大容忍时延下的测试结果

3.4　不同方案性能分析

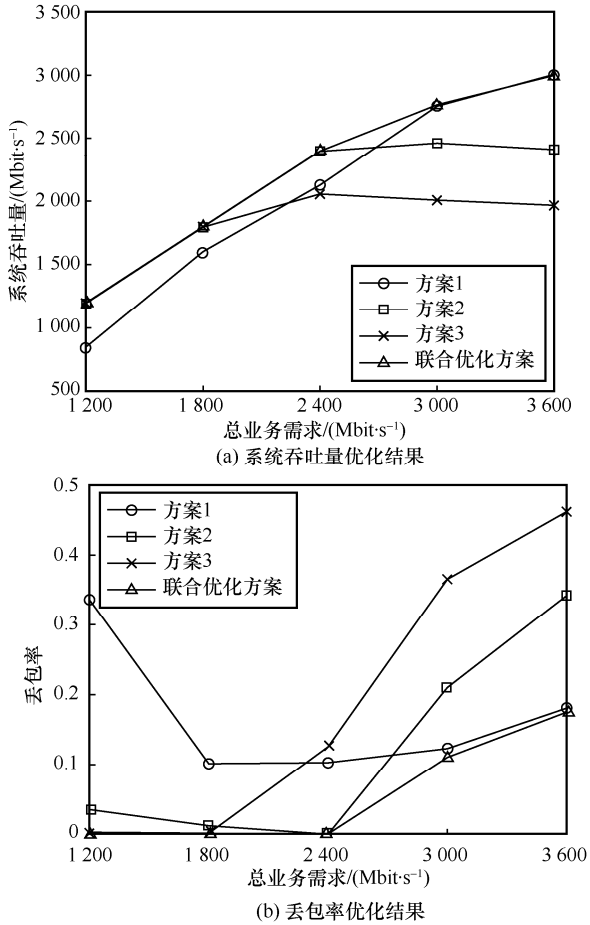
为了验证本文提出的 BHCC 算法在多波束卫星系统中的性能，本文将基于 BHCC 算法的联合优化方案与 3 种固定波束覆盖方案进行了对比，其区别在于 BHCC 算法不仅可以控制波束跳变策略，还能自适应调节波束的覆盖范围。而另外 3 种波束覆盖方案仅对波束跳变策略进行优化，其波束尺寸固定。

由式(3)可知，波束的宽度会影响波束的传输容量，$F(1)$ 表示波束半径 $v_t^k=1$ 时波束的传输容量，如果增大波束的覆盖范围，使一个波束同时覆盖 n 个小区，则波束的传输总容量 $F(n)<F(1)$，为了简化实验，波束传输容量与波束覆盖小区数的关系如表 2 所示。

表 2　波束传输容量与波束覆盖小区数的关系

方案	波束传输容量	波束覆盖小区数/个
方案 1	$F(1)$	1
方案 2	$0.85F(1)$	3
方案 3	$0.65F(1)$	5

训练结束后，不同总业务需求下各方案的优化结果如图 7 所示。从图 7(a)可以看出，随着总业务需求从 1 200 Mbit/s 增加到 3 600 Mbit/s，基于 BHCC 算法的联合优化方案都接近最优策略，并且不同方案的系统吞吐量随着总业务需求的增加总体呈上升趋势，方案 3 和方案 2 覆盖范围较大，导致其传输容量上限较小，会更快地达到吞吐量上限。



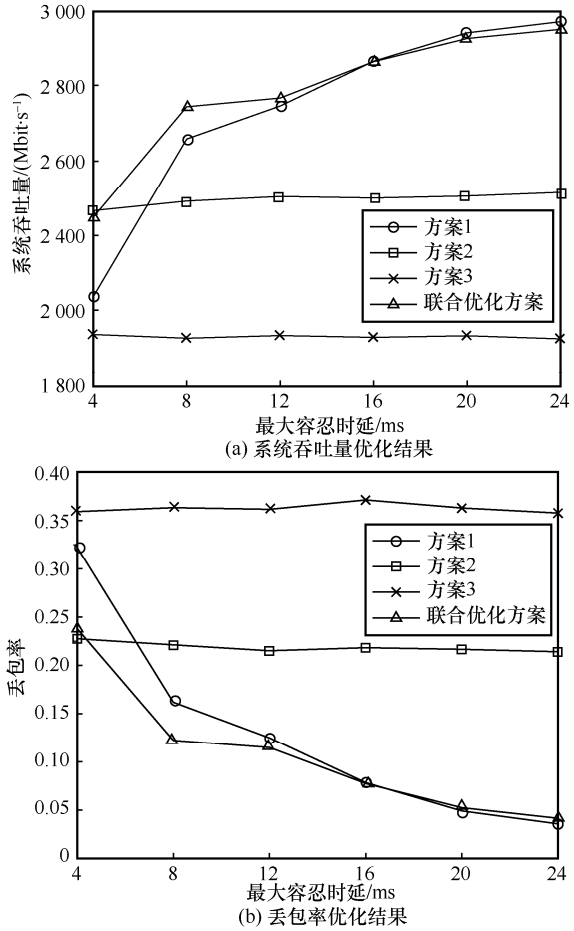
(a) 系统吞吐量优化结果

(b) 丢包率优化结果

图 7　不同总业务需求下各方案的优化结果

从图 7(b)可以看出，随着总业务需求从 1 200 Mbit/s 增加到 3 600 Mbit/s，基于 BHCC 算法的联合优化方案都能接近最优策略，方案 3 在低业务需求时表现出较好的性能，因为其覆盖范围较大，可以支持更多的小区传输，但是在高业务需求时，因为其传输容量小，导致需求不能满足，丢包率较高。相反，方案 1 因为传输容量大，在高业务需求时可以传输更多的数据流，丢包率更小。

训练结束后，当总业务需求为 3 000 Mbit/s 时，不同业务最大容忍时延下各方案的优化结果如图 8 所示。



(a) 系统吞吐量优化结果

(b) 丢包率优化结果

图 8　不同业务最大容忍时延下各方案的优化结果

从图 8 可以看出，基于 BHCC 算法的联合优化方案在不同的最大容忍时延的情况下都是最优策略，当最大容忍时延为 4 ms 时，自适应范围的方案和方案 2 的性能接近，因为方案 2 相比方案 1 可以覆盖更多的小区，但是随着最大容忍时延增加，自适应覆盖范围方案和方案 1 的性能更接近，因为方案 2 的传输容量相比方案 1 较小，当最大容忍时延增加时，每个小区队列中待传输的数据量增加，选择具有更大传输容量的方案 1 能得到更好的性能。

4 结束语

本文研究了多波束卫星系统中的波束跳变和覆盖控制联合优化问题，以匹配有限的卫星资源与非均匀和时变的通信需求。首先，将目标优化问题描述为一个马尔可夫过程。其次，提出了“单智能体轮询复用”机制，避免联合优化导致的“维度灾难”问题，即在训练过程中，将波束作为一个独立的智能体，并对智能体进行训练，训练结束后，多个波束以轮询的方式共用一个智能体的算法模型，以实现在卫星上同时控制多波束的目的。最后，仿真结果表明，相对于遗传算法、贪婪算法及随机算法，BHCC 算法不仅能提高 MBS 的吞吐量，而且能降低系统的丢包率；相对于不考虑自适应波束覆盖范围的深度强化学习算法，BHCC 算法在不同通信场景下的性能更优异。

参考文献：

[1] 张晨, 张更新, 王显煜. 基于跳波束的新一代高通量卫星通信系统设计[J]. 通信学报, 2020, 41(7): 59-72.
ZHANG C, ZHANG G X, WANG X Y. Design of next generation high throughput satellite communication system based on beam-hopping[J]. Journal on Communications, 2020, 41(7): 59-72.

[2] LEI L, LAGUNAS E, YUAN Y X, et al. Deep learning for beam hopping in multibeam satellite systems[C]//Proceedings of 2020 IEEE 91st Vehicular Technology Conference. Piscataway: IEEE Press, 2020: 1-5.

[3] LEI J, VÁZQUEZ-CASTRO M Á. Multibeam satellite frequency/time duality study and capacity optimization[J]. Journal of Communications and Networks, 2011, 13(5): 472-480.

[4] CHOI J P, CHAN V W S. Optimum power and beam allocation based on traffic demands and channel conditions over satellite downlinks[J]. IEEE Transactions on Wireless Communications, 2005, 4(6): 2983-2993.

[5] WANG L B, HU X, MA S J, et al. Dynamic beam hopping of multi-beam satellite based on genetic algorithm[C]//Proceedings of 2020 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking. Piscataway: IEEE Press, 2021: 1364-1370.

[6] TIAN F, HUANG L L, LIANG G, et al. An efficient resource allocation mechanism for beam-hopping based LEO satellite communication system[C]//Proceedings of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting. Piscataway: IEEE Press, 2019: 1-5.

[7] SHI D Y, LIU F, ZHANG T. Resource allocation in beam hopping communication satellite system[C]//Proceedings of International Wireless Communications and Mobile Computing. Piscataway: IEEE Press, 2020: 280-284.

[8] WANG A Y, LEI L, LAGUNAS E, et al. Joint beam-hopping scheduling and power allocation in NOMA-assisted satellite systems[C]//Proceedings of IEEE Wireless Communications and Networking Conference. Piscataway: IEEE Press, 2021: 1-6.

[9] TANG J Y, BIAN D M, LI G X, et al. Optimization method of dynamic beam position for LEO beam-hopping satellite communication systems[J]. IEEE Access, 2021, 9: 57578-57588.

[10] RAN Y Y, ZHOU X, HU H, et al. Optimizing data centre energy efficiency via event driven deep reinforcement learning[C]//Proceedings of IEEE World Congress on Services. Piscataway: IEEE Press, 2022: 20.

[11] RAN Y Y, HU H, WEN Y G, et al. Optimizing energy efficiency for data center via parameterized deep reinforcement learning[J]. IEEE Transactions on Services Computing, 2022, PP(99): 1-14.

[12] HU X, LIU S J, WANG Y P, et al. Deep reinforcement learning-based beam hopping algorithm in multibeam satellite systems[J]. IET Communications, 2019, 13(16): 2485-2491.

[13] LIU S J, HU X, WANG W D. Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems[J]. IEEE Access, 2018, 6: 15733-15742.

[14] LUIS J J G, GUERSTER M, DEL PORTILLO I, et al. Deep reinforcement learning for continuous power allocation in flexible high throughput satellites[C]//Proceedings of IEEE Cognitive Communications for Aerospace Applications Workshop. Piscataway: IEEE Press, 2019: 1-4.

[15] LIAO X L, HU X, LIU Z J, et al. Distributed intelligence: a verification for multi-agent DRL-based multibeam satellite resource allocation[J]. IEEE Communications Letters, 2020, 24(12): 2785-2789.

[16] LIN Z Y, NI Z Y, KUANG L L, et al. Dynamic beam pattern and bandwidth allocation based on multi-agent deep reinforcement learning for beam hopping satellite systems[J]. IEEE Transactions on Vehicular Technology, 2022, 71(4): 3917-3930.

[作者简介]

许国良（1973– ），男，浙江金华人，博士，重庆邮电大学教授，主要研究方向为大数据信息挖掘、智能分析、传感检测、感知系统和传感网络等领域的新技术和应用。

谭峰（1998– ），男，重庆人，重庆邮电大学硕士生，主要研究方向为卫星通信、资源分配、路由算法等。

冉泳屹（1986– ），男，重庆人，重庆邮电大学讲师、硕士生导师，主要研究方向为深度强化学习、深度学习、随机优化等方法在计算机系统和通信网络的应用。

陈丰（1997– ），男，安徽池州人，重庆邮电大学硕士生，主要研究方向为卫星通信、路由算法等。