电讯技术
Telecommunication Engineering
ISSN 1001-893X,CN 51-1267/TN

《电讯技术》网络首发论文

题目：基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制方法
作者：刘骏，王永华，王磊，尹泽中
收稿日期：2022-04-28
网络首发日期：2022-08-04
引用格式：刘骏，王永华，王磊，尹泽中．基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制方法[J/OL]．电讯技术.
https://kns.cnki.net/kcms/detail/51.1267.tn.20220802.1104.002.html

网络首发时间：2022-08-04 08:55:28
网络首发地址：https://kns.cnki.net/kcms/detail/51.1267.tn.20220802.1104.002.html

基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制方法

刘骏，王永华**，王磊，尹泽中
(广东工业大学 自动化学院，广东广州 510006)

摘要：为了保证认知无线网络中次用户本身的通信服务质量，同时降低次用户因发射功率不合理而造成的功率损耗，提出一种基于 SumTree 采样结合深度双 Q 网络（Double Deep Q Network, Double DQN）的非合作式多用户动态功率控制方法。通过这种方法，次用户可以不断与辅助基站进行交互，在动态变化的环境下经过不断的学习，选择以较低的发射功率完成功率控制任务。其次，该方法可以解耦目标 Q 值动作的选择和目标 Q 值的计算，能够有效减少过度估计和算法的损失。并且，在抽取经验样本时考虑到不同样本之间重要性的差异，采用了结合优先级和随机抽样的 SumTree 采样方法，既能保证优先级转移也能保证最低优先级的非零概率采样。仿真结果表明，该方法收敛后的算法平均损失值能稳定在 0.04 以内，算法的收敛速度也至少快了 10 个训练回合，还能提高次用户总的吞吐量上限和次用户功率控制的成功率，并且将次用户的平均功耗降低了 0.5mW 以上。

关键词：认知无线网络；功率控制；SumTree 采样；深度强化学习

开放科学（资源服务）标识码（OSID）：

中图分类号：TN929.5　　文献标识码：A

A Non-cooperative Multi-user Dynamic Power Control Method Based on SumTree Sampling and Double DQN

LIU Jun, WANG Yonghua, WANG Lei, YIN Zezhong
(School of Automation, Guangdong University of Technology, Guangzhou 510006, China）

Abstract：To ensure the communication service quality of the secondary users in cognitive wireless networks and reduce the power loss caused by the unreasonable transmit power of the secondary users, so it proposes a non-cooperative multi-user dynamic power control method based on SumTree sampling and Double DQN(Double Deep Q Network). Through this method, the secondary users not only can continuously interact with the auxiliary base station and continuous learning in a dynamically changing environment, but also choose to complete the power control task by a lower transmit power. Moreover, this method can decouple the selection of the target Q-value action and the calculation of the target Q-value, which can effectively reduce overestimation and algorithm loss. In addition, this method takes into account in the importance of difference between samples when extracting empirical samples and adopts the SumTree sampling method combining priority and random sampling, which can ensure both priority transfer and non-zero probability sampling of the lowest priority. The simulation results show that the average loss value of the algorithm after the convergence of this method can be stabilized within 0.04. The convergence speed of the algorithm is at least 10 training rounds faster. It can also improve the total throughput upper limitation of the secondary users and the success rate of the power control in secondary users, reducing the average power consumption for secondary users by at least 0.5mW.

Key words：cognitive radio network; power control; SumTree sampling ; deep reinforcement learning

收稿日期：2022-04-28　修回日期：2022-06-30
基金项目：国家自然科学基金项目（61971147）；广东省研究生教育创新计划项目（2020JGXM040）
**通信作者：王永华 wangyonghua@gdut.edu.cn

0 引　言

伴随着无线通信技术的迅猛发展，5G 技术也迎来更加广泛的应用，不同的终端和设备可以通过无线技术接入到互联网，基于 5G 技术的万物互联也变成了可能[1]。无线通信业务快速增长的同时也导致对频谱需求的急剧增加，当前部分比较传统且固定的静态分配方法已无法满足频谱资源共享的需求[2]。认知无线电的目标是为了实现频谱资源的高效利用，而功率控制技术作为一项重要的认知无线电技术，自然也是频谱共享领域的重要研究方向[3-4]。

发射功率是一种重要的无线通信资源，对于认知无线网络（Cognitive Radio Network, CRN）的功率控制问题，目前已有很多相关的研究。文献[5]研究了由一个基站和多个用户组成的简单物联网系统，该论文使用深度 Q 网络(Deep Q Network, DQN)解决了该场景下的访问控制和连续控制问题，这样可使得对于功率的发射控制更加灵活，但该论文没有考虑发射功率的损耗问题，并且算法的复杂度也比较高。文献[6]提出了一种动态多目标方法用于认知无线电中的功率和频谱分配，该方法的算法复杂度不是很高，并且有利于释放认知无线电的潜能，但该方法没有考虑频谱分配的成功率以及发射功率的损耗问题。文献[7]设计了一种分布式深度强化学习的功率控制方案，在该方案中 CRN 采用分层 DQN 来实现动态频谱分配，这种方案的好处是算法复杂度较低，收敛速度较快，但却没有考虑到要最大化吞吐量上限和降低功率损耗等问题。文献[8]提出将用户社会关系考虑到频谱的功率控制中并验证了该方法的可行性，该方法创新性地提供了一个新的应用场景，但该方法的算法复杂度比较高，并且也没有考虑到发射功率的损耗问题。

总的来说，目前的功率控制方法往往没有很好地平衡次用户(Second User, SU)的功率损耗和次用户通信服务质量(Quality of Service, QoS)之间的性能关系。本文为了解决认知无线网络中多个用户的动态功率控制问题，既保证 SU 本身的 QoS，同时还要降低 SU 因发射功率不合理而造成的功率损耗，所做的主要贡献总结如下：

1) 提出一种基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制方法，该方可以解耦目标 Q 值动作的选择和目标 Q 值的计算，能够有效减少过度估计和算法的损失。在抽取经验样本时采用结合优先级和随机抽样的 SumTree 采样方法，既能保证优先级转移也能保证最低优先级的非零概率采样。

2) 建立 SU 重叠式接入到主用户(Primary User, PU)信道的动态频谱资源分配模型，主次用户为非合作式关系，SU 可以不断与辅助基站进行交互，在动态变化的环境下经过不断的学习，选择以较低的发射功率完成功率控制任务，达到减少能耗的目的。

3) 仿真结果还表明，本方法的算法损失函数更小，算法收敛速度更快，还能有效提高次用户功率控制的成功率和次用户总的吞吐量上限。

1 系统模型

如图1，本模型中有1个主基站（Primary Base Station, PBS），PBS位于本模型靠近中心的位置，它能够保障PU的正常通信。多个辅助基站(Auxiliary Base Station, ABS)随机散布于模型中的各个区域，ABS有采集主次用户接收信号强度信息（Receive signal strength information, RSSI）的作用[9]，同时ABS还能够与SU进行交互和通信，并把采集到的PU及SU的RSSI信息再传送给SU。M个PU和N个SU（N>M）随机散布于各个区域。同时采用自由路径损耗的信道模型，主次用户基于同一个信道进行通信，并且主次用户为非合作式的关系，SU采用重叠式接入PU信道。

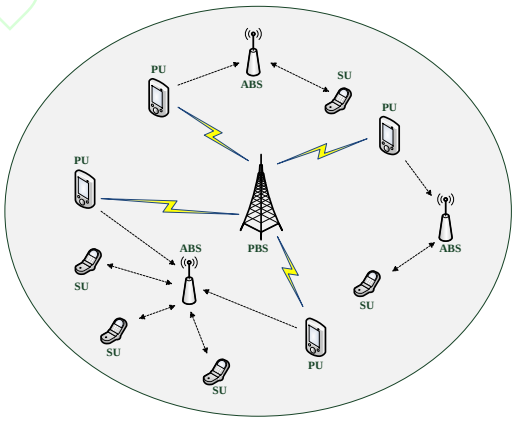

图1 系统模型

在本模型中，SU 采用重叠式接入 PU 信道，SU 接入到PU的频段是会对PU的通信造成一定干扰的，并且这个干扰会随着 SU 数量的增多而增大。而 PU 和 SU 进行频谱共享的一个重要要求是 SU 不能对 PU 的正常通信造成干扰，因此在进行功率控制时需要将SU的发射功率控制在一定的阈值之内。但是，SU 的发射功率过低会导致其无法满足自身的 QoS 要求，甚至可能会导致 SU 的通信发生中断。所以，对于功率控制技术而言，就是要处理好 SU 对 PU 的通信干扰问题，同时也要尽可能保证 SU 自身的 QoS。这就需要 SU 在进行功率控制时寻找到发射功

率的平衡点，通过学习主次用户的 RSSI 信息自适应地调整发射功率来完成通信任务。

信噪比（Signal to interference plus noise ratio, SINR）是衡量链路质量的重要指标。定义第i个 PU 的信噪比为[10]：

$$\gamma_i(t)=\frac{h_{ii}(t)P_i(t)}{\sum_{i\neq j}h_{ji}(t)P_j(t)+N_i(t)},i=1,2,...,M \tag{1}$$

同理，定义第j个 SU 的信噪比为：

$$\gamma_j(t)=\frac{h_{jj}(t)P_j(t)}{\sum_{j\neq k}h_{kj}(t)P_k(t)+h_{ij}(t)P_i(t)+N_j(t)},j=1,2,...,N \tag{2}$$

本模型的认知无线网络环境是动态变化的，PU 和 SU 都是通过信噪比来判断自身的 QoS。对于 PU 而言，设定任意主用户i正常工作的要求为：

$$\gamma_i\geq\mu_i \tag{3}$$

为了满足 QoS，PU 应根据自身需要智能调整发射功率。而一种高效的 PU 功率控制策略能够有效减少因频繁切换带来的能量损耗，从而可以在一定程度上节约能源。因此，本文引入一种较为高效且智能的 PU 功率控制策略[11]。

首先设置 PU 功率值集合为：

$$P_i(t)\triangleq\{P_1,\ P_2,\ ...,\ P_l\} \tag{4}$$

定义 PU 的功率控制策略如下：

$$P_i(t+1)=\begin{cases}P_{i+1}, & if\ P_i\leq\tau\leq P_{i+1}\ and\ i+1\leq l\\ P_{i-1}, & if\ \tau\leq P_{i-1}\ and\ i-1\geq 1\\ P_i, & otherwise\end{cases} \tag{5}$$

其中：

$$\tau\triangleq\frac{\mu_i}{\gamma_i(t)}P_i(t) \tag{6}$$

当前时刻对下一时刻的 PU 信噪比预测值为：

$$\hat{\delta}\triangleq\frac{P_i(t+1)}{P_i(t)}\gamma_i(t) \tag{7}$$

上式中，l为 PU 动作空间的长度。本策略通过判断当前时刻 PU 信噪比是否达到阈值要求，并预测下一时刻是否能够满足要求，只通过一次切换即可完成功率发射。具体的，若当t时刻 PU 信噪比低于要求阈值，但预测到下一时刻提高自身发射功率就能够满足要求，此时 PU 会增加发射功率；若t时刻 PU 已满足阈值要求，且预测下一时刻下调功率仍然满足要求，此时 PU 会减小发射功率以降低能耗。其余情况维持当前功率不变。

在实际情况中，SU 接入到 PU 的频段是会对 PU 的通信造成一定干扰的，而 PU 和 SU 进行频谱共享的一个重要前提是 SU 不能对 PU 的正常通信造成干扰。所以对于 SU 的功率控制策略，本文将采用深度强化学习的方法，使 SU 通过不断的学习，能够智能选择和更新自身的发射功率，达到提高频谱资源利用率的目的。

为了模拟更加复杂的动态环境，本模型中的信道增益每隔一段时间会更新一次。根据香农定理，第j个 SU 的吞吐量上限与信噪比间的函数可表示为[12]：

$$T_j(t)=W\log_2(1+\rho\gamma_j(t)) \tag{8}$$

在该动态变化的系统中，要保证系统的功率分配效果最佳，就是既要满足 PU 的信噪比高于预设阈值，还要保证 SU 能够通过不断学习来调整自身发射功率，从而让整个系统中 SU 总的吞吐量上限最大化。本节涉及到的系统参数，如表 1 所示。

表 1 系统参数

参数	含义	单位
$P_i(t)$	第i个 PU 在t时刻的发射功率	mW
$h_{ji}(t)$	t时刻，发射方第j个 SU 到接收方第i个 PU 的信道增益	-
$P_j(t)$	第j个 SU 在t时刻的发射功率	mW
$N_i(t)$	第i个 PU 在t时刻接收到的噪声功率	mW
$N_j(t)$	第j个 SU 在t时刻接收到的噪声功率	mW
$h_{ii}(t)$	t时刻，第i个 PU 的发射方到接收方的信道增益	-
$h_{jj}(t)$	t时刻，第j个 SU 的发射方到接收方的信道增益	-
$h_{kj}(t)$	t时刻，第k个 SU 的发射方到第j个 SU 接收方的信道增益	-
$h_{ij}(t)$	t时刻，第i个 PU 的发射方到第j个 SU 接收方的信道增益	-
μ_i	第i个 PU 预设的阈值	dB
μ_j	第j个 SU 预设的阈值	dB
$T_j(t)$	第j个 SU 在t时刻的吞吐量上限	b/s
W	第j个 SU 的可用带宽	b/s
ρ	一个常系数	-

2 算法设计

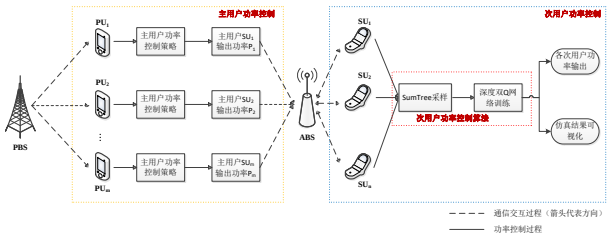

图2 基于SumTree采样结合Double DQN的非合作式多用户动态功率控制方法框图

如图 2 所示，为了解决认知无线网络中多个用户的动态功率控制问题，既保证 SU 本身的 QoS，同时还要降低 SU 因发射功率不合理而造成的功率

损耗，本文采用一种基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制方法（a non-cooperative multi-user dynamic power control method based on SumTree sampling and double DQN，简称 ST_double DQN 方法）来进行求解。使用 Double DQN 算法来解耦目标 Q 值动作的选择和目标 Q 值的计算，这样可以有效减少过度估计。而普通的深度强化学习使用等概率随机采样进行经验回放，这样会存在重要经验利用率不足和收敛速度较慢等问题，本文的 SumTree 采样方法能够对经验样本赋予不同优先级并在采样时还加入随机性，可以防止系统的过拟合，提高认知无线网络的性能。本方法能够在动态变化的环境下使 SU 经过不断的学习，选择以较低的发射功率完成功率控制任务，达到减少能耗的目的。

2.1 SumTree 采样

如图 3 所示，本文使用一种二叉树结构的存储单元作为记忆库的存储结构。SumTree 存储示意图中从上往下一共有四层节点结构，最顶部的那个节点称之为根节点，最底层一行称之为叶子节点，中间两行称之为内部节点。所有经验样本的数据都是储存在叶子节点，不仅如此，叶子节点还会存储样本的优先级。除叶子节点外的所有节点都是不存储数据的，但是会保存下级的左右子节点优先级之和，并且把子节点优先级之和用数字显示出来。

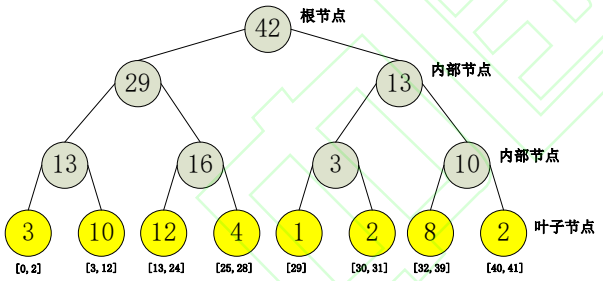

图 3 SumTree 存储单元结构示意

SumTree 采样主要是根据优先级来对样本进行训练，优先级取决于时序差分(Temporal-Difference Learning, TD)误差的大小，TD 误差的值越大说明神经网络的反向传播作用越强，样本被学习的重要性就越高，相应的优先级也越高，这些样本就会优先被训练[13]。

为了保证采样能够按照优先级进行，并且所有样本都有被抽到的可能，本文采用这种结合优先级和随机抽样的方法，既能保证优先级转移也能保证最低优先级的非零概率采样。采样转移概率为：

$$p(j)=\frac{p_j^{\alpha}}{\sum_k p_k^{\alpha}} \tag{9}$$

p_j 和 p_k 分别表示样本 j 和任意样本 k 的优先级，对于 p_j 有：

$$p_j=|TD_{error}(j)|+\epsilon \tag{10}$$

上面式子中，ϵ 是一个非常小的正常数，这样可保证 $p_j>0$，而 α 为优先级指数，$\alpha=0$ 时为随机均匀采样，k 代表采样的批量数。上面的采样机制会带来偏差，会使得系统不稳定，于是根据样本重要性权重来纠正偏差：

$$\omega_j=\left(\frac{1}{N}\cdot\frac{1}{p(j)}\right)^{\beta} \tag{11}$$

上式中，ω_j 表示权重系数，N 代表经验池大小，β 表示非均匀概率补偿系数，当 $\beta=1$ 时就完全补偿了 $p(j)$。

2.2 基于 Double DQN 的动态功率控制

强化学习属于机器学习的一个大的分支，它是多个学科交叉的领域。强化学习的数据和环境是动态的，它通过不断与环境进行交互，在试错的过程中找到有效的策略，所以它主要用来解决智能决策问题。

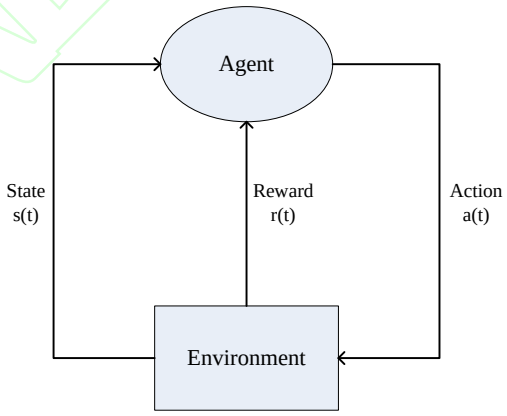

图 4 强化学习过程

图 4 为强化学习的学习过程，该过程是伴随着智能体(Agent)和环境(Environment)的交互而产生[14]。其中，在 t 时刻的 Agent 能够感知 Environment 所处的状态 $s(t)$，并且 Agent 可以选择一个动作 $a(t)$ 作用于 Environment，Environment 在受到动作的作用后对 Agent 产生一个奖励 $r(t)$，此时 Environment 转移进入到下一个状态 $s(t+1)$。

本文研究的动态功率控制问题，本质上是一个马尔科夫决策过程[15]。设置一个四元组结构，其中 S 表示环境的状态值集合，A 表示Agent动作值集合，R 表示在状态下采用动作所得到的奖励值，π 为策略函数。Agent在状态下选择动作获得的奖励期望回报为：

$$G_t = \sum_{t=0}^{+\infty} \gamma^t R_{t+1} \tag{12}$$

上式中，折扣因子$\gamma \in [0,1]$，R_{t+1}为t时间步所获环境奖励值。在状态s_t下采取策略π的状态价值函数为：

$$V_\pi(s) = E_\pi[G_t | s_t = s] \tag{13}$$

基于策略π，采用动作a_t，则s_t状态下的动作价值函数为：

$$Q_\pi(s,a) = E_\pi[G_t | s_t = s, a_t = a] \tag{14}$$

最优策略和最优价值函数可表示为：

$$\begin{cases} V_*(s) = \max\limits_{\pi} V_\pi(s) \\ Q_*(s,a) = \max\limits_{\pi} Q_\pi(s,a) \\ \pi_*(s) = arg \ \max\limits_{a} Q_*(s,a) \end{cases} \tag{15}$$

用Bellman最优方程求解最优价值函数可得：

$$\begin{cases} V_*(s) = \max\limits_{a} \left[R(s,a) + \gamma \sum\limits_{s'} p(s'|s,a) V_*(s') \right] \\ Q_*(s,a) = R(s,a) + \gamma \sum\limits_{s'} p(s'|s,a) \max\limits_{a'} Q_*(s',a') \end{cases} \tag{16}$$

其中，p表示状态转移概率，s'表示在s下一时刻的状态。Q-learning和上述马尔科夫决策过程一样，也是通过迭代更新值函数，其更新公式为：

$$Q(s,a) = Q(s,a) + \alpha \left[R + \gamma \max_{a'} Q(s',a') - Q(s,a) \right] \tag{17}$$

对于要解决的动态功率控制问题，SU 的任务就是通过不断的学习得到最优的功率控制策略$\pi_*(s)$。SU 的发射功率是不断变化的，传统的强化学习方法很难对其进行求解，而 DQN 就能很好解决这一问题。DQN 既有深度学习强大的感知能力，又具备强化学习的决策能力，还能够克服传统强化学习不能处理高维连续状态和动作空间的问题。

DQN 采用的是两个结构完全相同但是参数不同的神经网络，其中一个是作为实时更新参数的神经网络结构，另一个是用于更新目标 Q 值的神经网络结构。目标 Q 值的更新公式为：

$$Q_{target} = R + \gamma \max_{a'} Q(s',a';\theta') \tag{18}$$

将目标 Q 值Q_{target}和当前网络 Q 值之间的均方差称之为损失函数,则损失函数可表示为：

$$L(\theta) = E[(Q_{target} - Q(s,a;\theta))^2] \tag{19}$$

然而，普通的 DQN 算法是从目标 Q 网络中寻找各个动作最大的 Q 值，这样会大概率选择过估计的 value，进而导致 value 的过乐观估计。

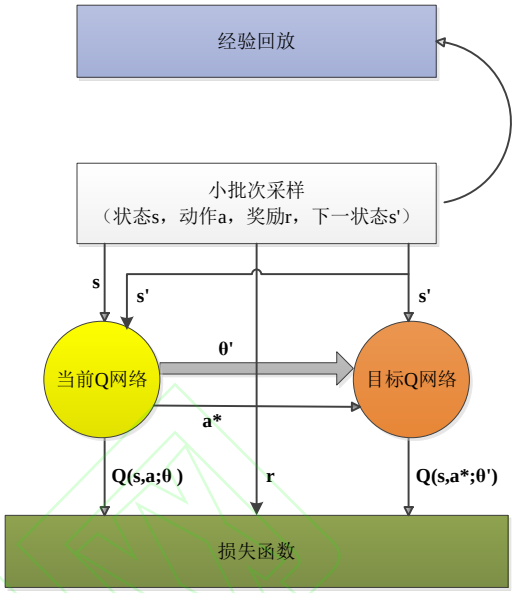

图 5 Double DQN 结构图

为了解决这个问题，本文采用 Double DQN 算法来进行动态功率控制。如图 5 所示，Double DQN 算法具有与普通 DQN 相同的网络结构，即 Double DQN 也有一个环境、两个结构相同但参数不同的神经网络、一个回放记忆单元以及误差函数。但 Double DQN 在从目标 Q 网络中寻找最大的 Q 值时能够解耦合：

$$Q_{target}^{Double} = R + \gamma Q\left(s', arg \max_{a} Q(s',a;\theta);\theta'\right) \tag{20}$$

Q_{target}^{Double}不再是直接在目标 Q 网络里面找各个动作中最大 Q 值，而是先在当前 Q 网络中先找出最大 Q 值对应的动作，然后再利用这个选择出来的动作在目标网络里面去计算目标 Q 值[16]。

采用 Double DQN 的动态功率控制方法，ABS 可将采集到的主次用户 RSSI 信息传送给 SU，SU 根据这些信息进行学习，并智能选择和更新自身的发射功率，使得在满足 PU 正常通信不受影响的情况下还能以一定的发射功率进行通信，从而实现频谱资源的动态共享。

（1）状态

本模型下的主次用户基于同一个信道进行通信，且主次用户为非合作式的关系，SU 采用重叠式通信。因为主次用户不能直接感知对方的功率发射方法，所以需要通过 ABS 来进行状态感知，ABS 可不间断采集主次用户的 RSSI 信息并再传送给 SU。若系统中有x个 ABS，那么状态值可表示为：

$$S(t) = [s_1(t), s_2(t), ..., s_k(t), ..., s_x(t)] \tag{21}$$

定义第k个 ABS 采集到的 RSSI 信息为：

$$s_k(t)=\sum_{i=1}^{m}P_i(t)\left[\frac{l_{ik}(t)}{l_0(t)}\right]^{-\tau}+\sum_{j=1}^{n}P_j(t)\left[\frac{l_{jk}(t)}{l_0(t)}\right]^{-\tau}+\sigma(t) \tag{22}$$

式中，在t时刻，PU 与 ABS 的距离用$l_{ik}(t)$表示，SU 与 ABS 的距离用$l_{jk}(t)$表示，基准距离用$l_0(t)$表示，τ表示路径损耗指数，该环境下的平均噪声功率用$\sigma(t)$表示。

（2）动作

SU 通过不断接收 ABS 的 RSSI 信息，能够根据获得的状态值选择到一个动作进行输出，输出的动作为发射功率。若每个 SU 有 H 种发射功率值，那么对于有 N 个 SU 的系统，动作空间的大小为H^n。定义动作空间为：

$$A(t)=[P_1(t),P_2(t),...,P_n(t)] \tag{23}$$

（3）奖励

SU 想要通过不断的学习，完成动态功率控制，那么一个重要的环节就是奖励函数的设计。本文依据 SU 从 ABS 处接收观测到的不同频谱接入情况，首先设置从S_1到S_4的四种不同频谱接入条件：

$$\begin{cases}S_1: & \forall\gamma_i\geqslant\mu_i\\ S_2: & \forall P_i\geqslant\sum P_j\\ S_3: & \exists\gamma_j\geqslant\mu_j\\ S_4: & \forall\gamma_i<\mu_i\end{cases} \tag{24}$$

满足任意PU信噪比都大于等于PU预设阈值记为S_1，满足任意 PU 发射功率都大于等于所有 SU 发射功率的和记为S_2，若存在 SU 信噪比大于等于 SU 预设阈值则记为S_3，若无 PU 信噪比高于预设阈值则记为S_4。

根据以上接入情况，定义奖励函数为：

$$R(t)=\begin{cases}-a_1\left(\sum_i\gamma_i+\sum_j\gamma_j\right), & if\ S_4=True\\ a_2\left(\sum_i\gamma_i+\sum_j\gamma_j\right), & if\ S_1、S_2、S_3都为True\\ -a_3\left(\sum_i\gamma_i\right), & otherwise\end{cases} \tag{25}$$

上式中，a_1、a_2、a_3为三个不同常系数。当同时满足条件S_1、S_2、S_3时，表明 SU 观测到所有 PU 正常工作，SU 顺利进行功率控制且达到了频谱资源共享的目的，此时把所有主次用户的信噪比求和并将一个正的系数a_2与之相乘作为奖励值。当满足条件S_4时，所有 PU 信噪比均低于阈值，若触发此条件，表明 SU 观测到所有 PU 的通信都受到了影响，此时将所有主次用户的信噪比求和并将一个负的a_1系数与之相乘作为奖励值，这个奖励值实际上是一个惩罚，说明选择到的该动作是错误的。其余所有情况，把 PU 信噪比求和并乘以一个负的a_3系数作为奖励值。

2.3 基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制方法

本文采用基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制方法进行求解，该算法的伪代码如算法 1 所示。

算法 1 基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制算法

固定经验池 D 的容量为 O，确定主次用户发射功率分别为$P_i(t)$和$P_j(t)$，初始化两个神经网络的参数，初始化优先级指数α和β，权重差值△＝0。

1) 对于每个回合 For m = 1 to M do

2) 根据初始状态$S_0(t)$，选择出动作$A_0(t)$，到达下一个状态

3) 对于回合中的每一步 For t = 1 to T do

4) PU 自适应控制自身发射功率

5) SU 基于贪婪算法选择动作，以ε概率随机选择动作A_t，或者以$1-\varepsilon$概率选择动作$A_t=\max_a Q(s_t,a;\theta,\alpha,\beta)$；

6) 通过选择动作A_t，得到奖励R_t，到达下一个状态S_{t+1}

7) 将样本数据$(S_t,A_t,R_t,r_{t+1},S_{t+1})$存储到经验池中的叶子节点中，并根据各样本的时序差分误差确定优先级

8) If $t>O/2$ then

9) 从经验池通过 SumTree 采样结合随机采样的方法进行采样，遵循$p(j)=\dfrac{p_j^{\alpha}}{\sum_k p_k^{\alpha}}$

10) 计算采样重要性权重w_j

11) 计算深度双 Q 网络的算法损失函数：$L(\theta)=E[(Q_{target}^{Double}-Q(s,a;\theta))^2]$，更新两个神经网络的权重参数$\theta$

12) 更新样本的优先级

13) 基于梯度下降法更新梯度

14) 更新深度双 Q 网络的两个神经网络的权重参数$\theta=\theta+\eta\cdot\triangle$，重置△＝0

15) 更新目标 Q 网络的权重参数θ'

16) End If

17) End For

18) 每隔一定步数随机更新环境的参数

19) End For

3 仿真实验与结果分析

本节通过 Python 平台进行仿真实验，在相同环境中，比较不同深度强化学习算法下各项指标性能的差异。本实验的场景为半径 200 米范围的圆形区域，以 PBS 为正中心，1 个 PU、2 个 SU 和 10 个 ABS 随机分布在区域内，PBS 负责 PU 的正常通信，ABS 可收集主次用户通信信息并把收集到的数据通过专用信道再发送给 SU。设置 PU 发射功率 $P_i(t)$ 为 0~30 范围内以 5 为间隔的 7 个离散值，SU 发射功率 $P_j(t)$ 为 0~4.5 范围内以 0.5 为间隔的 10 个离散值，发射功率单位均为 mW。参照现有文献以及梯度进行仿真实验的结果，本实验规定 PU 的预设信噪比阈值不低于 1.0dB，SU 预设信噪比阈值不低于 0.5 dB，环境噪声均为 0.1mW。

本实验的当前 Q 网络和目标 Q 网络均使用含 4 个隐含层的全连接层，其中前三层隐含层神经元个数分别为 256、128 和 256，输出层神经元个数为动作空间的大小。前两层使用 ReLUs 函数作为激活函数，第三层使用 tanh 函数作为激活函数，神经网络权重的更新方法为 Adam 算法。经验池的容量为 1000，当经验池容量超过 500 时才开始训练，批量采样数目为 128，目标 Q 网络的更新频率为 250。在动作的选择上采用贪婪算法($\varepsilon-greedy$)，ε 的初始值为 0.8，随着训练次数线性迭代至 0。

本文提出的基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制方法，因核心算法为 SumTree 采样结合 Double DQN，故后文及仿真实验中将该算法简称为 ST_double DQN。将相同仿真环境和奖励函数下的 natural DQN 算法、double DQN 算法、dueling DQN 算法按照其算法名称直接命名并进行实验对比和分析。同时将文献 8 中的算法(Dueling deep Q-networks for social awareness-aided spectrum sharing, 简称 SAA_dueling DQN 算法)也进行仿真实验与分析。

图 6 为相同环境下五种算法的损失函数对比图，从图中可以看到，五种深度强化学习算法在经过一定回合的学习之后均能够达到收敛。natural DQN 算法、double DQN 算法和 SAA_dueling DQN 算法的收敛速度都比较慢，都是直到第 30 回合左右才逐渐趋于稳定，并且这三种算法在收敛之前的波动也比较大。dueling DQN 算法较之上述三种算法，在收敛速度上稍好一些，但还是存在较大波动。而本文提出的 ST_double DQN 算法能够在 5 个回合的训练内将损失值迅速降至 0.1 以内，在第 15 回合左右即可达到收敛，并且收敛的平均损失值在 0.04 以内，说明本方法具有更好的适应性和学习能力。

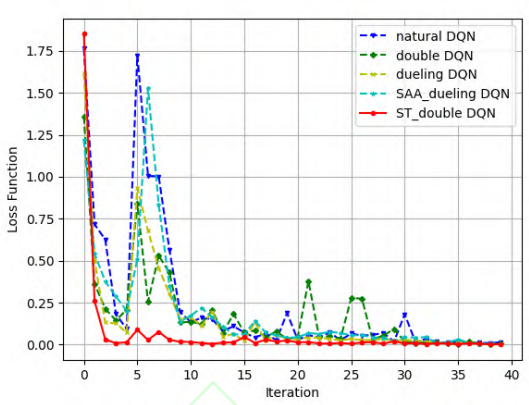

图 6 损失函数

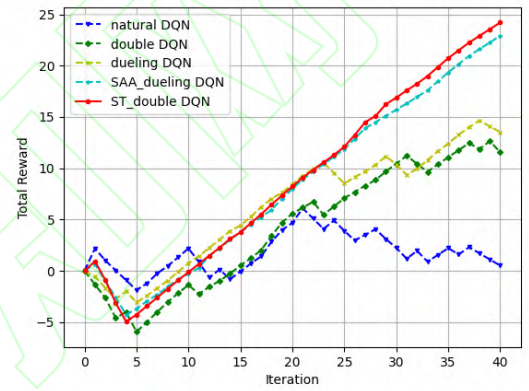

图 7 累计奖励

图 7 为五种不同算法累计奖励对比曲线。从图中可以看出：natural DQN 算法、double DQN 算法和 dueling DQN 算法的累计奖励一直在震荡，说明这三种算法在探索过程中，一直在接入成功的动作和接入不成功的动作之间切换，最终都没有确定合适的接入动作。SAA_dueling DQN 算法和 ST_double DQN 算法在前面 4 个回合都没有采取正确的动作，从第 5 回合开始就能探索出 SU 接入成功的动作，开始获得正奖励，累计奖励也持续保持上升，表明这两种算法都能够快速探索出 SU 接入成功的动作，找到合适的功率控制策略。

图 8 为 SU 总的吞吐量上限图像。本模型下的动态功率控制问题，就是在满足 PU 通信质量的前提下，SU 自适应调整自身发射功率，达到最大化 SU 总吞吐量的目的。初始阶段，五种算法的吞吐量上限都几乎没有变化，随着不断学习和交互，从第 5 回合开始，natural DQN 算法、double DQN 算法和 dueling DQN 算法的总吞吐量上限开始持续增加，但这三种算法都在第 35 回合左右总吞吐量上限开始几乎保持不再增加。而 SAA_dueling DQN 算法和

ST_double DQN 算法在训练 4 回合后，SU 的总吞吐量上限一直保持持续上升，没有出现再次下降或者不变的情况，表明这两种算法都可以使 SU 总的吞吐量上限持续提高，但 ST_double DQN 算法的上限值略高于 SAA_dueling DQN 算法。

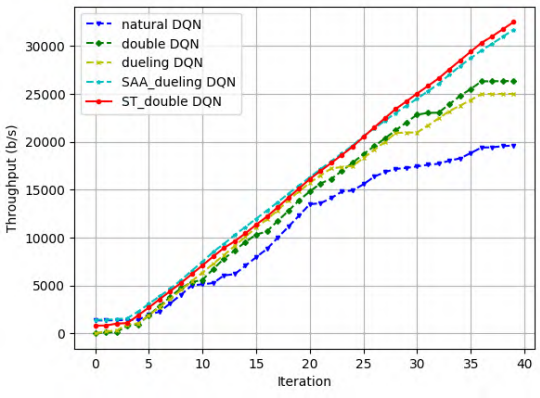

图 8 次用户总的吞吐量上限

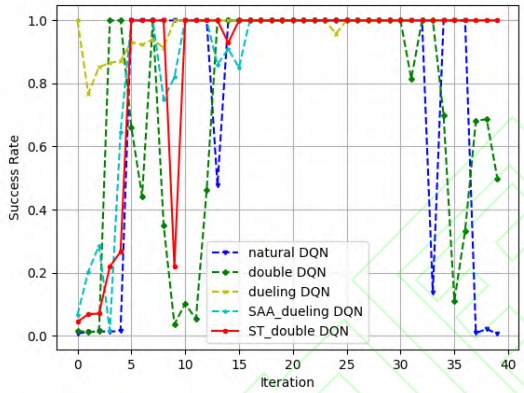

图 9 功率控制成功率

图 9 中，本实验训练的总次数为 40000 次，每 1000 次定义为图像所显示的一个回合，在每个回合内选取 50 次训练进行测试，若测试中 SU 能够选择成功的接入动作则视为成功完成传输任务，成功的次数与测试总次数的比值定义为功率控制的成功率。仿真结果中，natural DQN 算法、double DQN 算法和 dueling DQN 算法的成功率波动幅度很大，并且始终没有收敛，很不稳定。而 SAA_dueling DQN 算法和 ST_double DQN 算法，虽然初始阶段波动也比较大，但这两种算法分别能够在第 16 回合和第 15 回合后就收敛并达到 100%的测试成功率。

图 10 显示的是五种算法 SU 的平均发射功率。总的来看，natural DQN 算法的平均发射功率是最高的，double DQN 算法、dueling DQN 算法和 SAA_dueling DQN 算法的平均发射功率大部分在 2.5mW 以上。而 ST_double DQN 算法的平均发射功率是最低的，绝大部分处于 2.0mW 和 2.5mW 之间，只有极少数在区间范围之外，这表明该算法能够有效降低能耗，这也是本方法的一个优势。

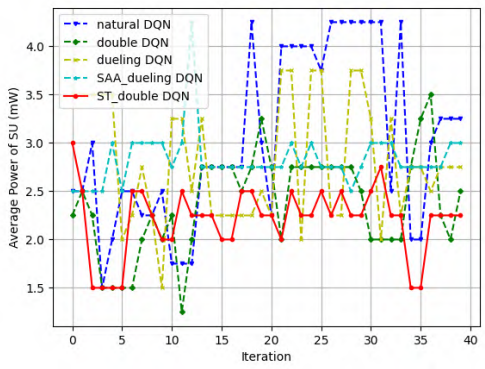

图 10 次用户平均功率

4 结束语

为了解决认知无线网络中多个用户的动态功率控制问题，既保证次用户本身的通信服务质量 QoS，同时还要降低次用户因发射功率不合理而造成的功率损耗，本文提出一种基于 SumTree 采样结合 Double DQN 的非合作式多用户动态功率控制方法。仿真结果也表明本方法可以保证次用户的 QoS,并且有效降低了次用户的功率损耗，达到了节约能耗的目的。然而，本文使用的是离散化的功率空间，未来可在研究中使用连续化的功率空间，这样可使得对于功率的发射控制更加灵活。同时，在后续研究中也可以基于更加复杂化的场景来进行建模，例如基于车联网、工业物联网等场景来进行建模。最后，在后续的研究中可采用不同信道模型来研究其对算法性能的影响。

参考文献

[1] ADESHINA B S, SHAHID M, SABA A R, et al. 5G Millimeter-Wave Mobile Broadband: Performance and Challenges[J]. IEEE Communications Magazine, 2018, 56(6):137-143.

[2] 盘小娜,陈哲,李金泽,等. 一种利用优先经验回放深度 Q-Learning 的频谱接入算法[J]. 电讯技术, 2020, 60(5):489-495.

[3] KUMAR P, HASAN N, IBRAHEE M, et al. Sub-optimal automatic generation control of interconnected power system using constrained feedback control strategy[J]. Electric Machines & Power Systems, 2012, 40(9):977-994.

[4] LEE W, LEE K. Deep Learning-Aided Distributed Transmit Power Control for Underlay Cognitive Radio Network[J]. IEEE Transactions on Vehicular Technology, 2021, 70(4):3990-3994.

[5] CHU M, LIAO X, LI H, et al. Power Control in Energy Harvesting

Multiple Access System With Reinforcement Learning[J]. IEEE Internet of Things Journal, 2019, 6(5): 9175-9186.

[6] CHUANG C L, CHIU W Y, CHUANG Y C. Dynamic Multiobjective Approach for Power and Spectrum Allocation in Cognitive Radio Networks[J]. IEEE Systems Journal, 2021, 15(4): 5417-5428.

[7] CHEN X, XIE X, SHI Z, et al. Dynamic Spectrum Access Scheme of Joint Power Control in Underlay Mode Based on Deep Reinforcement Learning[C]//2020 IEEE/CIC International Conference on Communications in China (ICCC). China: IEEE, 2020: 536-541.

[8] WANG Y, LI X, WAN P, et al. Dueling deep Q-networks for social awareness-aided spectrum sharing[J]. Complex & Intelligent Systems, 2021, 1: 1-12.

[9] 叶梓峰,王永华,万频,等.基于优先记忆库结合竞争深度 Q 网络的动态功率控制[J]. 电讯技术, 2019, 59(10):1132-1139.

[10] LI F, TAN X, WANG L. A new game algorithm for power control in cognitive radio networks[J]. IEEE Transactions on Vehicular Technology, 2011, 60(9): 4384-4391.

[11] LI X, FANG J, CHENG W, et al. Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach[J]. IEEE Access, 2018, 6: 25463-25473.

[12] KASGARI A T Z, MAHAM B, KEBRIAEI H, et al. Dynamic learning for distributed power control in underlaid cognitive radio networks[C]//2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC). Cyprus: IEEE, 2018: 213-218.

[13] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized Experience Replay[C]//ICLR 2016: International Conference on Learning Representations. San Juan: [s.n,], 2016:1-21.

[14] BAN T W. An Autonomous Transmission Scheme Using Dueling DQN for D2D Communication Networks[J]. IEEE Transactions on Vehicular Technology, 2020, 69(12):16348-16352.

[15] VIKRAM K. Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing[J]. IEEE Control Systems Magazine, 2019, 39(4) : 76-79.

[16] PAN J, WANG X, CHENG Y, et al. Multisource Transfer Double DQN Based on Actor Learning[J]. IEEE Transactions on Neural Networks and Learning Systems,2018,29(6) : 2227-2238.

作者简介：

刘骏　男，1996 年生于湖北鄂州，2019 年获学士学位，现为广东工业大学控制工程专业硕士研究生，主要研究方向为认知无线网络和深度强化学习。

王永华　男，1979 年生于河北石家庄，2009 年获博士学位，现为广东工业大学自动化学院副教授，主要研究方向：认知无线网络、机器学习。

王磊　男，1993 年生于陕西汉中，2017 年获学士学位，现为广东工业大学控制科学与工程专业硕士研究生，主要研究方向为认知无线网络。

尹泽中　男，1997 年生于湖南郴州，2019 年获得学士学位，现为广东工业大学控制工程专业硕士研究生，主要研究方向为深度强化学习。