引用格式：还婧文，杨少石，袁田浩，孟阔，毕嘉辉，唐玉蓉.一类基于定向Q-Learning的后5G无线网络上下行多业务并发功率分配方法[J].中国传媒大学学报（自然科学版），2022，29(02)：27-33.
文章编号：1673-4793（2022）02-0027-07

一类基于定向Q - Learning的后5G无线网络上下行多业务并发功率分配方法

还婧文[1]，杨少石[1,2*]，袁田浩[1]，孟阔[1]，毕嘉辉[1]，唐玉蓉[3]

（1.北京邮电大学 信息与通信工程学院，北京 100876；2.泛网无线通信教育部重点实验室，北京 100876；
3.中国移动研究院，北京 100053）

摘要：在后5G时代，基于动态时分双工技术的无线网络需要同时支持传输方向、速率、时延、可靠性等指标具有差异性的多类型业务共存及并发，这会导致复杂的跨小区交叉链路干扰问题。本文提出了一类基于定向Q-Learning的无线网络上下行多业务并发功率分配方法，利用平均意见分作为多业务的用户体验质量评价指标，对无线网络中的基站及用户发射功率进行分配。通过对新用户加入后Q-table的更新方式进行改进，提出了三种优化的Q-Learning算法。仿真结果表明，改进后的算法在用户数动态变化的场景下，在保证合理的平均意见分和拥塞率时，降低了迭代次数，提高了算法收敛性能。

关键词：无线网络；功率分配；机器学习；后5G；多业务
中图分类号：TN92 **文献标识码**：A

A family of directed Q-Learning based power allocation methods for uplink/downlink multi-service concurrency in beyond 5G wireless networks

HUAN Jingwen[1], YANG Shaoshi[1,2*], YUAN Tianhao[1], MENG Kuo[1], BI Jiahui[1], TANG Yurong[3]

*(1.School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing 100876, China;
2.Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing 100876, China;
3.China Mobile Research Institute, Beijing 100053, China)*

Abstract: In the beyond 5G era, the dynamic time-division duplexing (D-TDD) technique will be employed by wireless networks, in order to support the co-existence and concurrency of multiple services that have diverse requirements on the transmission direction, rate, latency and reliability, thus resulting in the complex inter-cell cross-link interference problem. In this paper we propose a family of directed Q-Learning based power allocation methods for uplink/downlink multi-service concurrency, where the mean opinion score (MOS) is invoked as a metric to characterize users′ quality of experience (QoE) for multiple services and to assist the transmission power allocation on the base station and the users. By improving the update mode of the Q-table after new users join the system, three optimized Q-Learning algorithms are proposed. Simulation results show that when the number of users changes, the improved algorithms maintain reasonable MOS values

基金项目：北京市自然科学基金-海淀原始创新联合基金前沿项目（L202012）；北京邮电大学-中国移动研究院联合创新中心资助项目

作者简介(*为通讯作者)：还婧文（1997-），女，硕士研究生，主要从事面向5G演进的异构网络混合业务资源分配技术研究。Email：jw_huan@bupt.edu.cn；杨少石（1983-），男，教授、博士生导师，主要从事B5G/6G和分布式感知-通信-计算-智能融合理论与技术研究。Email：shaoshi.yang@bupt.edu.cn

and congestion rate, while reducing the number of iterations and achieving improved convergence performance.

Key words: wireless networks; power allocation; machine learning; beyond 5G; multi-service

1 引言

为了满足未来无线通信系统面向多类型业务的灵活需求,文献[1][2]提出使用灵活双工技术来增强系统性能,这在赋能垂直行业的公网、专网联合部署环境中尤为重要。然而,当两个基于动态时分双工(Dynamic Time Division Duplexing, D-TDD)技术的相邻小区具有相反的传输方向并且共享相同时频资源时,可能会发生严重的小区间干扰,这种现象被称为交叉链路干扰(Cross-Link Interference, CLI)。它包括下行链路(Downlink, DL)到上行链路(Uplink, UL)的干扰和UL到DL的干扰。

5G赋能垂直行业的一个重要场景是工业物联网,其中多种不同类型的业务(如语音业务、数据业务和视频业务)在不同传输方向上并发成为常态。在后5G时代的无线网络中,使用D-TDD技术进行灵活的业务自适应传输有助于提高系统的传输资源利用率,但这也会导致复杂的CLI问题。如何进一步优化无线资源管理算法,有效缓解CLI问题带来的负面影响,是一个迫切需要研究的重要问题。

此外,对于不同的业务需求,优化的目标函数一般不同,这将增大无线资源分配问题的复杂性。为此,基于体验质量(Quality of Experience, QoE)对5G网络的资源管理技术进行性能评估[3]得到业界的广泛认可。平均意见分(Mean Opinion Score, MOS)是一种使用最广泛的QoE指标[4]。通过为不同类型的业务提供通用测量尺度,MOS使跨不同特征的业务进行综合业务管理和资源分配成为可能[5]。

Q-Learning可以通过与环境交互获得的即时回报生成接近最优的解决方案。通过优化当前奖励实现长期优化目标对于动态变化的复杂无线网络的资源管理至关重要。在用户数动态变化时,新用户加入后如何更有效地分配基站及用户的发射功率是后5G研究中的一个难点。针对此问题,有研究者提出了认知学习的思想,该思想允许新用户从提前接入小区的用户那里学习,以改进学习过程[6]。

总的来说,现有的很多针对多小区场景下的功率分配方法仅仅围绕干扰消除展开,并没有考虑用户业务类型需求不同的情况。因此,本文对多小区无线网络中上下行多业务并发场景下的功率分配方法进行研究。首先,给出了宏小区用户和微小区用户的语音、数据以及视频业务的系统模型、业务模型以及评价指标。其次,基于Q-Learning对多业务并发时的基站和用户发射功率进行分配,基于Q-table的更新方式提出了三种定向学习方法。最后,将设计的三种定向学习方法与无定向学习能力的原始Q-Learning算法进行比较分析。仿真结果显示本文提出的方法在保证系统合理的MOS值和拥塞率时,降低了算法收敛所需的迭代次数,提升了算法收敛性能。

2 系统模型

本文所考虑的系统模型如图1所示,包含两个小区(宏小区和微小区)。宏小区的传输方向为DL,信号由宏基站(Macro Base Station, MBS)发送给宏小区用户(Macro-cell User Equipment, MUE)。微小区的传输方向为UL,微小区用户(Small-cell User Equipment, SUE)将信号上传至微基站(Small Base Station, SBS)。宏小区和微小区的用户数分别为K和L。

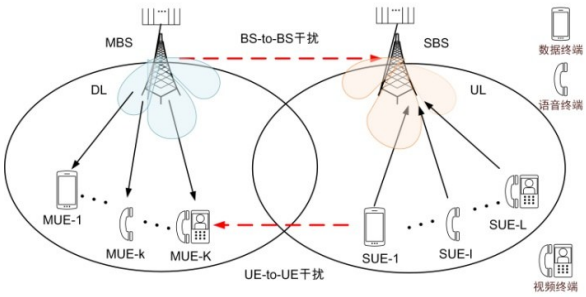

图1 系统模型示意图

DL中MBS天线数为M;UL中SBS天线数为N;MUE和SUE均为单天线。第k个MUE的DL接收信号为:

$$y_k^{\mathrm{DL}} = \mathbf{h}_k^{\mathrm{DL}}\mathbf{w}_k^{\mathrm{DL}}s_k^{\mathrm{DL}} + \mathbf{h}_k^{\mathrm{DL}}\sum_{i\in\Phi_k\setminus k}\mathbf{w}_i^{\mathrm{DL}}s_i^{\mathrm{DL}} + \sum_{l\in\Psi_k}h_{l,k}\sqrt{p_l^{\mathrm{UL}}}\,s_l^{\mathrm{UL}} + n_k \tag{1}$$

$\mathbf{h}_k^{\mathrm{DL}}\in\mathbb{C}^{1\times M}$是第$k$个MUE的DL信道状态信息(Channel State Information, CSI),$\mathbf{w}_i^{\mathrm{DL}}\in\mathbb{R}^{M\times 1}$是功率分配矢量,其公式为:

$$\mathbf{w}_i^{\mathrm{DL}} = \left[\sqrt{p_{1,i}^{\mathrm{DL}}},\cdots,\sqrt{p_{m,i}^{\mathrm{DL}}},\cdots,\sqrt{p_{M,i}^{\mathrm{DL}}}\right]^{\mathrm{T}}, i\in\Phi_k \tag{2}$$

$s_k^{\mathrm{DL}}\in\mathbb{C}$是MBS向第$k$个MUE发送的符号,

$s_l^{\mathrm{UL}} \in \mathbb{C}$是第$l$个SUE向SBS发送的符号，$h_{l,k} \in \mathbb{C}$是第$l$个SUE到第$k$个MUE之间的CSI。$\Phi_k$是与第$k$个MUE占用同一时频资源块（Resource Block, RB）的MUE的集合，因此$\left|\Phi_k\right| = K$。Ψ_k是与第k个MUE占用相同RB的SUE的集合，因此$\left|\Psi_k\right| = L$。$p_{m,i}^{\mathrm{DL}} \in \mathbb{R}$是MBS的第$m$个天线分配给符号$s_i^{\mathrm{DL}}$的功率，假设$\left|s_k^{\mathrm{DL}}\right|^2 = 1$且$\left|s_l^{\mathrm{UL}}\right|^2 = 1$。$n_k \sim CN\left(0,\sigma^2\right)$代表第$k$个MUE接收到的加性高斯白噪声（Addictive White Gaussian Noise, AWGN）。

另一方面，SBS以第l个SUE为目标用户时的接收信号为：

$$\mathbf{y}_l^{\mathrm{UL}} = \mathbf{h}_l^{\mathrm{UL}}\sqrt{p_l^{\mathrm{UL}}}\,s_l^{\mathrm{UL}} + \sum_{i\in\Psi_l\setminus l}\mathbf{h}_i^{\mathrm{UL}}\sqrt{p_i^{\mathrm{UL}}}\,s_i^{\mathrm{UL}} + \mathbf{H}^{\mathrm{BS}}\mathbf{W}^{\mathrm{DL}}\mathbf{s}^{\mathrm{DL}} + \mathbf{n} \tag{3}$$

其中，$\mathbf{h}_l^{\mathrm{UL}} \in \mathbb{C}^{N\times 1}$是第$l$个SUE到SBS的信道矢量，$\mathbf{H}^{\mathrm{BS}} \in \mathbb{C}^{N\times M}$是MBS到SBS的信道矩阵，$W^{DL} = \left[\mathbf{w}_1^{\mathrm{DL}},\mathbf{w}_2^{\mathrm{DL}},\cdots,\mathbf{w}_K^{\mathrm{DL}}\right] \in \mathbb{R}^{M\times K}$，$\mathbf{s}^{\mathrm{DL}} = \left[s_1^{\mathrm{DL}},s_2^{\mathrm{DL}},\cdots,s_K^{\mathrm{DL}}\right]^{\mathrm{T}} \in \mathbb{C}^{K\times 1}$。方便起见，本文定义$\mathbf{w}^{\mathrm{UL}} \triangleq \left[p_1^{\mathrm{UL}},p_2^{\mathrm{UL}},\cdots,p_L^{\mathrm{UL}}\right]$。

下面根据接收信号模型对信干噪比（Signal to Interference plus Noise Ratio, SINR）进行推导。第k个MUE的SINR为：

$$\mathrm{SINR}_k^{\mathrm{DL}} = \frac{\left\|\mathbf{h}_k^{\mathrm{DL}}\mathbf{w}_k^{\mathrm{DL}}\right\|^2}{\varphi_k + \psi_k + \sigma^2} \tag{4}$$

其中，

$$\varphi_k = \sum_{i\in\Phi_k\setminus k}\left\|\mathbf{h}_k^{\mathrm{DL}}\mathbf{w}_i^{\mathrm{DL}}\right\|^2 \tag{5}$$

$$\psi_k = \sum_{l\in\Psi_k}\left|h_{l,k}\right|^2 p_l^{\mathrm{UL}} \tag{6}$$

相似地，SBS以第l个SUE为目标用户时的SINR为：

$$\mathrm{SINR}_l^{\mathrm{UL}} = \frac{\left\|\mathbf{h}_l^{\mathrm{UL}}\right\|^2 p_l^{\mathrm{UL}}}{\underline{\psi_l} + \underline{\varphi_l} + \sigma^2} \tag{7}$$

其中，

$$\underline{\psi_l} = \sum_{i\in\Psi_l\setminus l}\left\|\mathbf{h}_i^{\mathrm{UL}}\right\|^2 p_i^{\mathrm{UL}} \tag{8}$$

$$\underline{\varphi_l} = \left\|\mathbf{H}^{\mathrm{BS}}\mathbf{W}^{\mathrm{DL}}\mathbf{s}^{\mathrm{DL}}\right\|^2 \tag{9}$$

根据上述推导结果，第k个MUE的速率可以表示为：

$$R_k^{\mathrm{DL}} = B_k\log_2\left(1 + \frac{\left\|\mathbf{h}_k^{\mathrm{DL}}\mathbf{w}_k^{\mathrm{DL}}\right\|^2}{\varphi_k + \psi_k + \sigma^2}\right) \tag{10}$$

其中，B_k是第k个MUE的带宽。第l个SUE的速率表示为：

$$R_l^{\mathrm{UL}} = B_l\log_2\left(1 + \frac{\left\|\mathbf{h}_l^{\mathrm{UL}}\right\|^2 p_l^{\mathrm{UL}}}{\underline{\psi_l} + \underline{\varphi_l} + \sigma^2}\right) \tag{11}$$

其中，B_l是第l个SUE的带宽。

3　基于Q-Learning的混合多业务功率分配

3.1　业务模型

本文的目标函数要求针对不同的业务将包括数据速率、误包率（Packet Error Probability, PEP）、峰值信噪比（Peak Signal-to-Noise Ratio, PSNR）等在内的评价参数映射到MOS。语音业务、数据业务和视频业务的具体映射关系如下所述。

1.语音业务。确定语音业务质量的传统方法是进行主观测试。这些测试的结果进行平均得出MOS，但此类测试成本高昂，不适用于在线语音业务质量评估。因此，ITU-T提出了一个标准化模型，即语音业务质量的感知评估（Perceptual Evaluation of Speech Quality, PESQ）[7]，这是一种能够以高度相关性预测典型主观测试中给出的质量分数的算法。然而PESQ算法在计算上过于昂贵，无法用于实时场景。为了解决这个问题，Giupponi等人提出了一个模型来估计MOS与传输速率R和PEP的函数[8, Fig. 2]。本文中，我们以$\mathrm{MOS_u}$来表示语音业务的MOS值，其具体数值及与传输速率R的对应关系，可由文献[8 Fig.2]中给出的PEP值确定。

2.数据业务。为了估计数据业务的用户满意度，本文使用对数形式的MOS与传输速率的关系[9]，它是传输速率R的递增严格凹连续可微函数[8, Fig. 3]。

基于系统提供给用户的R和PEP来估计MOS，具体计算公式如下：

$$\mathrm{MOS_w} = a*\log_{10}\left[b*R*(1 - PEP)\right] \tag{12}$$

式中，a和b由用户感知质量确定。通过R和PEP来计算MOS。

3.视频业务。对于视频业务质量进行评估的现有技术很多，其中ITU对多媒体业务质量进行了主观评估。PSNR作为一种视频业务质量的评价指标，被普遍用来客观地衡量视频的编码性能。然而，PSNR不能准确反映人类对视频质量的主观感知。PSNR和MOS之间具有线性映射关系[8, Fig. 2]，它为40 dB及以上的PSNR分配的MOS值为4.5，为20 dB及以下的PSNR分配的MOS值为1。上限来自这样一个事实，即PSNR为40 dB的重建视频序列几乎无法与原始视

频序列区分,低于20 dB的视频序列会因严重的退化而失真[8]。因此如果使用客观指标(例如PSNR)测量图像失真,可以使用以下逻辑函数来表征MOS和PSNR之间的关系[10]:

$$\mathrm{MOS}_v = \frac{c}{1+\exp\left[d*(PSNR-e)\right]} \tag{13}$$

其中,c、d和e是函数的参数,取c=6.6431,d=−0.1344和e=30.4264。本文选择log函数来评估视频的质量。为了表征重建视频的PSNR随传输速率的变化,得到关系如下:

$$PSNR = k\log_{10}R + p \tag{14}$$

其中k和p是常数。

3.2 优化目标

由于MOS作为所有类型业务的通用质量评估指标[11],允许以集成方式用于为所有类型的业务分配传输资源,因此本文将语音、视频和数据业务评价指标统一化为:

$$\frac{1}{U+V+W}\left(\begin{array}{l}\sum_{u=1}^{U}\mathrm{MOS}_u\\+\sum_{v=U+1}^{U+V}\mathrm{MOS}_v\\+\sum_{w=U+V+1}^{U+V+W}\mathrm{MOS}_w\end{array}\right) \tag{15}$$

其中,U是语音业务用户的数量,V是视频业务用户的数量,W是数据业务用户的数量。在这种情况下,即使系统性能最大化,也有可能无法满足给定用户的需求,这可能是因为其SINR过低,导致传输资源被分配给其他用户,这与试图为用户提供的公平性相矛盾。为了解决这个问题,本文根据估计的MOS历史值选择公平系数λ_{ui}、λ_{vi}和λ_{wi}。假设当前处于分配步骤Z中,系统中某业务有Q个用户,用户的最大MOS值:

$$\mathrm{MaxMOS}_Z = \frac{1}{Z-1}\max\left\{\begin{array}{l}\sum_{n=1}^{Z-1}\mathrm{MOS}_{1n},\sum_{n=1}^{Z-1}\mathrm{MOS}_{2n},\\\cdots,\sum_{n=1}^{Z-1}\mathrm{MOS}_{Qn}\end{array}\right\} \tag{16}$$

用户的公平系数的计算式如下:

$$\lambda_{qZ} = \frac{\mathrm{MaxMOS}_Z}{\left(1/(Z-1)\right)\sum_{n=1}^{Z-1}\mathrm{MOS}_{qn}},\quad q=1\cdots Q \tag{17}$$

因此,具有最大MOS值的用户公平系数为1。由于分母在区间[1,4.5]内变化,所以其他用户的公平系数在[1,4.5]范围内,给在当前优化步骤之前MOS较低的用户提供更多的资源来确保公平性。通过使用相应算法求解以下优化问题,来获得最佳功率分配方案,以使目标函数最大化,即最大化系统平均MOS性能:

$$\max_{\mathbf{w}^{\mathrm{UL}};\{\mathbf{w}_k^{\mathrm{DL}}\},k=1,2,\cdots,K}\frac{1}{K+L}\left(\begin{array}{l}\sum_{u=1}^{U}\lambda_u\mathrm{MOS}_u\\+\sum_{v=U+1}^{U+V}\lambda_v\mathrm{MOS}_v\\+\sum_{w=U+V+1}^{U+V+W}\lambda_w\mathrm{MOS}_w\end{array}\right) \tag{18}$$

$$\sum_{k=1}^{K}\mathrm{Tr}\left(\mathbf{w}_k\mathbf{w}_k^{\mathrm{T}}\right)\le P_{\max}^{\mathrm{DL}} \tag{19}$$

$$p_l^{\mathrm{UL}}\le P_{l,\max}^{\mathrm{UL}},\quad l=1,2,\cdots,L \tag{20}$$

$$R_u\ge R_{u,\min} \tag{21}$$

$$R_v\ge R_{v,\min} \tag{22}$$

$$R_w\ge R_{w,\min} \tag{23}$$

其中,λ表示三种业务的公平系数。公式(19)表示MBS的M根天线对于K个MUE的发送功率之和小于基站的最大发送功率;公式(20)表示SBS给每个SUE分配的发送功率要小于其本身最大的发送功率;公式(21)表示语音业务的用户需满足文献[8 Fig.2]所示的四种语音编码器其中一个的速率,即$R_{u,\min}$的值为6.4 kbit/s、15.2k bit/s、24.6 kbit/s、64 kbit/s;公式(22)(23)分别表示视频业务和数据业务的用户速率需满足的传输速率。

3.3 算法设计

Q-Learning定义一组状态(State)S、一组动作(Action)A和奖励函数(Reward)R,奖励函数表示所选动作对环境的影响。每个代理(Agent)将从A中选择下一个Action。在本文的系统模型中,Agent对应于MBS,它的每根天线均给多个MUE分配下行功率,并告知SBS如何为每个SUE分配上行功率,这些功率的分配对应A。本文的目标函数对应奖励函数,反映了系统的QoE;约束条件对应状态S。MBS在可行域中对功率的有限离散空间进行搜索。本文选择Q-Learning强化学习方法来解决前述功率分配问题。MBS获取环境当前状态S,并相应地在特定策略π下采取行动a,也就是$\pi(s)\to a$,即时奖励为$R(a,s)$。然后,使用折扣(discount)因子$\gamma(0<\gamma<1)$来最大化未来奖励,该因子代表未来奖励的重要性。在满足约束的情况下,MBS将寻求一个最优分配,以最大化目标值。

此外,定义系统拥塞率来表示系统学习性能:

$$\eta = 1-\frac{count\_new}{count} \tag{24}$$

其中,$count_new$表示新用户加入后成功学习(当Q-table值保持不变时)的次数,$count$表示新用户加入前原有用户成功学习的次数。

第t次选择的动作$a_t\in A=\left\{\cdots,A_i^{\mathrm{DL}}\cup A_j^{\mathrm{UL}},\cdots\right\}$,其中

$A_i^{\mathrm{DL}}=\left\{p_{1,1}^{\mathrm{DL}},\cdots,p_{M,1}^{\mathrm{DL}},\cdots,p_{1,K}^{\mathrm{DL}},\cdots,p_{M,K}^{\mathrm{DL}}\right\}$，$A_j^{\mathrm{UL}}=\left\{p_1^{\mathrm{UL}},\cdots,p_L^{\mathrm{UL}}\right\}$。第 t 次的状态为 $s_t\in S=\left\{\cdots,S_i^{\mathrm{DL}}\cup S_j^{\mathrm{UL}},\cdots\right\}$，其中 $S_i^{\mathrm{DL}}=\left(S_{i1}^{\mathrm{DL}},S_{i2}^{\mathrm{DL}}\right)$，$S_j^{\mathrm{UL}}=\left(S_{j1}^{\mathrm{UL}},S_{j2}^{\mathrm{UL}}\right)$，两者均由约束条件决定。对于 MUE 以下关系成立：

$$S_{i1}^{\mathrm{DL}}=\begin{cases}0 & \sum_{k=1}^{K}\mathrm{Tr}\left(\mathbf{w}_k\mathbf{w}_k^{\mathrm{T}}\right)\le P_{\max}^{\mathrm{DL}}\\ 1 & \text{其他}\end{cases}\tag{25}$$

$$S_{i2}^{\mathrm{DL}}=\begin{cases}0 & R_k\ge R_{k,\min}\left(\forall k\in\{1,\cdots,K\}\right)\\ 1 & \text{其他}\end{cases}\tag{26}$$

公式(25)表示满足功率约束下的状态，满足条件判断为0；公式(26)表示满足多业务用户速率下的状态，只要有某个用户没达到速率要求，S_{i2}^{DL} 就被判断为1。同样，对于 SUE 以下关系成立：

$$S_{j1}^{\mathrm{UL}}=\begin{cases}0 & p_l^{\mathrm{UL}}\le P_{l,\max}^{\mathrm{UL}}\left(\forall l\in\{1,\cdots,L\}\right)\\ 1 & \text{其他}\end{cases}\tag{27}$$

$$S_{j2}^{\mathrm{UL}}=\begin{cases}0 & R_l\ge R_{l,\min}\left(\forall l\in\{1,\cdots,L\}\right)\\ 1 & \text{其他}\end{cases}\tag{28}$$

第 t 次的即时奖励表示为：

$$R_t\left(s_t,a_t\right)=\begin{cases}\beta & S_{i1}^{\mathrm{DL}}+S_{i2}^{\mathrm{DL}}+S_{j1}^{\mathrm{UL}}+S_{j2}^{\mathrm{UL}}>0\left(\text{所有用户}\right)\\ \dfrac{1}{K+L}\left(\sum_{k=1}^{K}\mathrm{MOS}_k+\sum_{l=1}^{L}\mathrm{MOS}_l\right) & \text{其他}\end{cases}\tag{29}$$

其中，β 是一个小于任何其他策略奖励的常数，取0.01表示采取了违反约束的不成功操作。当满足约束时，式中的值为语音业务、数据业务和视频业务的平均 MOS 值。Q-table 更新函数表示为：

$$\begin{aligned}&Q_{t+1}\left(s_t,a_t\right)\leftarrow(1-\alpha)\cdot Q_t\left(s_t,a_t\right)\\&+\alpha\cdot\left[R_t\left(s_t,a_t\right)+\gamma\max_{a_{t+1}}Q_{t+1}\left(s_{t+1},a_{t+1}\right)\right]\end{aligned}\tag{30}$$

式中 α 是学习效率，$0<\alpha<1$。公式(30)中出现的最大化表示在所有可能的 Action 中选择使 Q_{t+1} 最大的 a_{t+1}。基于 Q-Learning 的无线网络资源分配具体流程如图2所示。

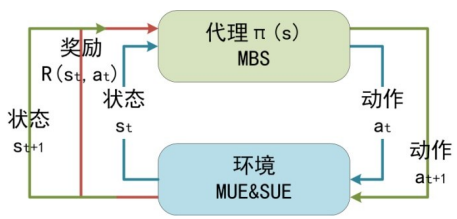

图2 基于Q-Learning的无线网络资源分配流程

本文的算法旨在对系统中新加入的多业务用户进行功率分配，。为使 Q-Learning 算法满足环境变化，本文引入定向学习能力，也就是说，对新加入用户的三种业务进行针对性的学习方式设计。每个新加入的多业务用户首先了解其周围环境，然后由 MBS 继续选择与最大奖励相关的 Action，通过运行 Q-Learning 获得所选行动的奖励，最后根据收到的即时奖励更新 Q-table。对于解决特定系统问题，具有更多“经验”的节点将教授能力较差的节点，以减少学习时间，同时提高学习性能的思想被称作 docitive[12]。本文所提三种学习方式如下所述。

第一种为相同业务类型定向学习，取相同业务类型用户的 Q-table 均值作为新用户的 Q-table；第二种为最近用户定向学习，选取距离新用户最近的用户的 Q-table 作为新用户的 Q-table；第三种为随机选择定向学习，在原始用户中随机选择某个用户的 Q-table 作为新用户的 Q-table。已经存在于网络中的用户使用 Q-Learning 算法学习 Q-table。新用户加入后，利用上述三种方案，获取新用户的 Q-table。算法详细过程描述如下：

步骤一：初始化学习效率 α、discount 因子 γ、Q-table；初始化带宽 B，MBS、SBS 位置；MUE、SUE 随机撒点；初始化信道矩阵。为当前所有用户随机分配业务类型。

步骤二：随机选择一个状态作为起点 $s_0\in S=\left\{\cdots,S_i^{\mathrm{DL}}\cup S_j^{\mathrm{UL}},\cdots\right\}$。

步骤三：根据 $\pi\left(s_t\right)$，在当前状态 s_t 的所有可选行动中选择一个作为 a_t。

步骤四：计算给定参数下生成的速率、MOS 值集合。

步骤五：系统移动到下一状态 s_{t+1}，反馈即时奖励值 $R\left(s_t,a_t\right)$。

步骤六：在新状态上选择使 Q-table 值最大的行动 a_{t+1}。

步骤七：更新 Q-table。

步骤八：新用户加入，方法一是取与新用户相同业务类型用户的 Q-table 取均值赋给新用户；方法二是取与新用户最近用户的 Q-table 赋给新用户；方法三是从原有用户中随机选择一个 Q-table 赋给新用户。

步骤九：将新状态更新为当前状态，重复步骤三到步骤七，直到 Q-table 值保持不变。并判断为成功学习。

4　仿真结果分析

仿真参数如表1所示，在仿真过程中宏、微小区的

原有用户数保持不变,原有用户的业务类型和新加入系统的用户的业务类型均随机分配,新加入用户数为1。

表1 仿真参数

参数名	参数值
噪声功率	174 dBm
宏小区半径	600 m
微小区半径	100 m
MBS天线数	4
SBS天线数	4
MBS最大功率	10 W
SUE最大功率	3 W
用户带宽	4 MHz
学习效率	0.1
discount因子	0.4

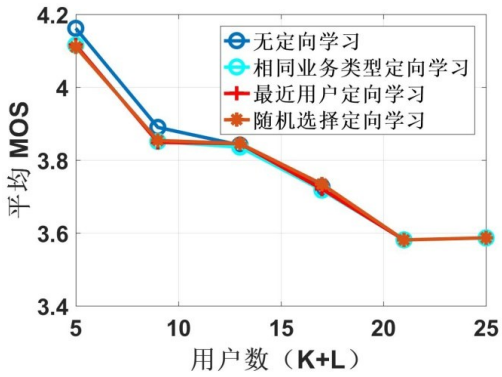

图3 各用户数下的平均MOS

图3所示为当其他系统参数保持不变,令系统原有用户数分别从4,8,12,16,20,24变化(新加入1个用户后总用户数为5,9,13,17,21,25),系统新加入用户业务类型随机分配时,分别以无定向学习、相同业务类型定向学习、最近用户定向学习以及随机选择定向学习四种算法进行平均MOS变化仿真。从图中可以观察到:

1)随着用户数的增加,四种不同学习方式所实现的系统平均MOS值均在减小,这意味着新用户无论是无定向学习,还是选择相似业务类型的用户进行定向学习,选择最近的用户进行定向学习,随机选择用户进行定向学习,用户数的增加均会造成系统性能的降低。

2)随着用户数的增加,无定向学习能力的Q-Learing算法在用户数较少时(图中看少于13个)可达到的系统性能较引入定向学习能力的算法略高。但是当用户数增加时,几种方式区别不大。考虑到当用户数增加时,无定向学习能力的算法复杂度比所提出的定向学习方法显著增加(如图5所示),这与其系统性能上所取得的微弱优势几乎抵消,而且定义的不同业务公平系数的引入会增加定向学习算法的最终平均MOS值,最终导致用户数增加时各算法性能差别不大。

3)三种不同定向学习能力的Q-Learning算法在系统平均MOS值性能上差别不大,这主要是因为三种算法在新用户刚加入时对Q-table的更新上具有差别,在后续的Q-table值更新上是没有区别的,最终达到的系统性能也不会有明显差别。

4)该仿真结果还表明,我们设计的三种基于Q-Learning资源分配算法均获得了较高的MOS值(始终高于可接受的MOS水平,包括仿真中存在最多的用户数的情况)。

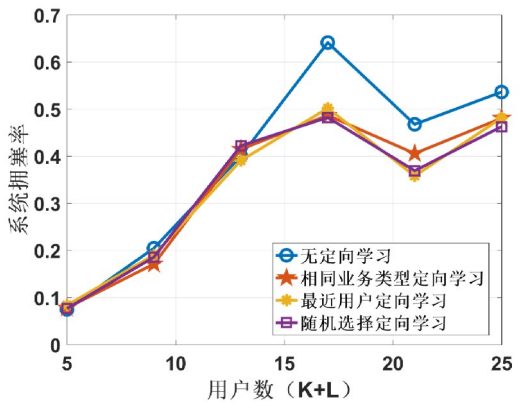

图4 各用户数下的系统拥塞率

图4所示为当系统用户数逐渐增加时,所定义的系统拥塞率的变化。可以看出:

1)该结果可以用于确定对系统拥塞率有要求时的用户数目的选择范围。基于此图所示结果,如果要求网络以预定义的拥塞率运行,则引入定向学习的解决方案始终能接受更多的用户数。

2)无定向学习的拥塞率较定向学习更高些。这是由于无定向学习算法中当新用户加入时,原有用户的Q-table会清空,继而随着新加入的用户重新分配资源,会加大算法的复杂度,导致拥塞率增加。

3)三种定向学习Q-Learning算法的系统拥塞率差别不大。这是因为,虽然实际拥塞率在不同类型的业务之间可能有所不同,但对于不同的s_t和a_t所获得的$R_t(s_t,a_t)$之间的关系是保持不变的,并且始终是通过目标函数MOS度量来计算的。

图5所示为用户数增加的过程中几种算法的效率(或计算复杂度)对比,分别展示了新用户加入后不同算法的迭代次数。可以得到如下结论:

1）本文所提三种定向学习算法能够将算法的平均收敛迭代次数减少约 2/3。

2）本文所提三种定向学习算法的复杂度几乎相同，只是选择学习的方式不同。这是由于定向学习通过有经验的用户将对周围环境的感知准确地转化为新用户的 Q-table，并减少实现收敛所需的迭代次数。可以看出，与加入定向学习能力前的无学习能力算法相比，实现收敛所需的迭代次数减少达 65%。

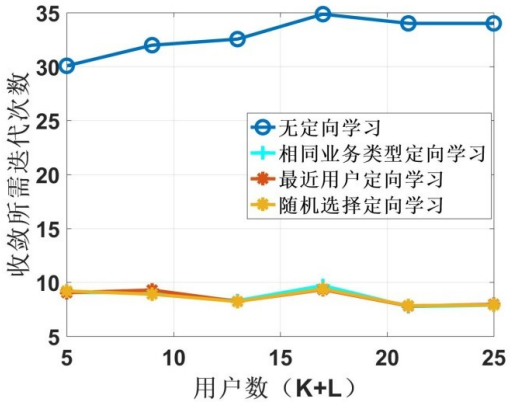

图 5 各用户数下的迭代次数

5 结论

本文针对后 5G 无线网络对上下行多业务的并发需求，利用 MOS 作为多业务用户的 QoE 评价指标，将多业务资源分配的优化目标统一化，给出了解决上下行多业务并发系统性能优化的目标函数，并提出了一类具有定向学习能力的 Q-Learning 方法对多业务用户基站侧和用户侧发射功率分配进行优化。由于当系统中加入新用户时发射功率需要重新分配，本文在原始 Q-Learning 算法的 Q-table 更新方式上进行了三种改进，分别取与新用户相同业务类型用户的 Q-table 取均值赋给新用户；取与新用户最近用户的 Q-table 赋给新用户；从原有用户中随机选择一个 Q-table 赋给新用户。将所设计的三种定向学习算法与无定向学习能力的 Q-Learing 算法进行了比较分析，可知在平均 MOS 值方面，所有用户进行定向学习与无定向学习算法系统性能差别不大；在算法的系统拥塞率方面，定向学习算法低于无定向学习；在算法所需迭代次数方面，定向学习算法可将迭代次数大大降低。综上所述，改进后的算法在用户数动态变化的场景下，在保证合理的系统 MOS 值和拥塞率的同时降低了迭代次数，提高了算法收敛性能。

参考文献（References）：

[1] Elbamby M S, Bennis M, Saad W, et al. Dynamic uplink-downlink optimization in TDD-based small cell networks [C]. 11th International Symposium on Wireless Communications Systems (ISWCS), 2014.

[2] Yu B, Yang L, Ishii H, et al. Dynamic TDD support in macro cell assisted small cell architecture[J]. IEEE Journal on Selected Areas in Communications, 2015, 33(6): 1201-1213.

[3] Wu Y, Hu F, Kumar S, et al. A learning-based QoE-driven spectrum handoff scheme for multimedia transmissions over cognitive radio networks[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(11): 2134-2148.

[4] Chen Y, Wu K, Zhang Q. From QoS to QoE: A tutorial on video quality assessment[J]. IEEE Communications Surveys & Tutorials, 2015, 17(2):1126-1165.

[5] Dobrijevic O, Kassler A J, Skorinkapov L, et al. Q-point: QoE-driven path optimization model for multimedia services [C]. International Conference on Wired/Wireless Internet Communications, 2014:134-147.

[6] Giupponi L, Galindo-Serrano A, Blasco P, et al. Docitive networks: an emerging paradigm for dynamic spectrum management[J]. IEEE Wireless Communications, 2010, 17(4): 47-54.

[7] ITU-T Recommendation P. 862. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs [EB/OL]. (2001) [2022-02-16]. https://www.itu.int/rec/T-REC-P.862.

[8] Khan S, Duhovnikov S, Steinbach E, et al. MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication[J]. Advances in Multimedia, 2007(01): 6.

[9] Kelly F P. Charging and rate control for elastic traffic[J]. European Transactions on Telecommunications, 1997, 8(1): 33-37.

[10] Hanhart P, Ebrahimi T. Calculation of average coding efficiency based on subjective quality scores[J]. Journal of Visual Communication & Image Representation, 2014, 25(3): 555-564.

[11] Piamrat F K, Viho C, Bonnin J M, et al. Quality of experience measurements for video streaming over wireless networks[C]. Sixth International Conference on Information Technology, 2009：1184-1189.

[12] Zhao Q, Grace D, Clarke T. Transfer learning and cooperation management: balancing the quality of service and information exchange overhead in cognitive radio networks[J]. Transactions on Emerging Telecommunications Technologies, 2015, 26(2):290-301.

编辑：王谦，李树锋