面向分层异构网络的资源分配：一种稳健分层博弈学习方案

邵鸿翔*①④ 赵杭生② 孙有铭③ 孙丰刚①
①(解放军理工大学通信工程学院 南京 210007)
②(南京电讯技术研究所 南京 210007)
③(解放军信息工程大学信息系统工程学院 郑州 450000)
④(洛阳理工学院 洛阳 471023)

摘 要：该文研究了信道状态不确定条件下分层异构微蜂窝网络中的无线资源分配优化问题。首先引入信道不确定模型描述无线信道的随机动态性，并将该问题建模为考虑信道不确定度的双层鲁棒斯坦伯格博弈；然后给出了该博弈的均衡点分析；最后提出了一种分布式改进型分层 Q 学习方案以实现宏基站和微基站的均衡策略搜索。理论分析和仿真表明，所提出的分层博弈模型可以有效抑制由于信道状态不确定引起的收益下降。所采用的学习方案较传统 Q 学习方案收敛速度明显加快，更加适用于短时快变的通信环境。

关键词：异构网络；斯坦伯格博弈；不完美信道信息；鲁棒决策；双层 Q 学习；离散策略
中图分类号： TN929.5 **文献标识码**： A **文章编号**：1009-5896(2017)01-0038-07
DOI: 10.11999/JEIT160285

Resource Allocation for Heterogeneous Wireless Networks: A Robust Layered Game Learning Solutions

SHAO Hongxiang①④ ZHAO Hangsheng② SUN Youming③ SUN Fenggang①
①(*College of Communications Engineering, PLA University of Science and Technology, Nanjing* 210007, *China*)
②(*Nanjing Telecommunication Technology Institute, Nanjing* 210007, *China*)
③(*Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou* 450000, *China*)
④(*Luoyang Institute of Science and Technology, Luoyang* 471023, *China*)

Abstract: This paper investigates a resource allocation scheme in heterogeneous wireless small cell networks with imperfect Channel State Information (CSI). In this work, the math expression for the stochastic dynamic uncertainty in CSI is proposed for model analysis and the robust Stackelberg game model with various interference power constraints is established firstly. Then, the Stackelberg game Equilibrium (SE) is obtained and analyzed. Lastly, an improved hierarchical Q-learning algorithm is also given to search the Stackelberg equilibrium strategies of macro-cell base station and small-cell base station. Both theoretical analysis and simulation results verify the proposed scheme can effectively restrain declining revenue due to incomplete CSI and the proposed algorithms can improves the convergence rate, especially applicable to the fast varying communication environment.

Key words: Heterogeneous wireless networks; Stackelberg game; Incomplete Channel State Information (CSI); Robust decision; Hierarchical Q-learning; Discrete strategy

1 引言

随着 5G 中新媒体数据应用需求的不断增长，密集组网技术将成为下一代通信的关键技术之一。通过在宏蜂窝基站(Macro-cell Base Station, MBS)周围布设小蜂窝基站(Small-cell Base Station, SBS)，能够扩展覆盖区域，改善能量效率，提高用户体验。异构分层蜂窝网大都采用分享复用的用频模式(shared-spectrum)，这种方式在增加频谱的空间重用效率的同时会引起小蜂窝与主蜂窝间的跨层干扰以及小蜂窝间的同层干扰，如果不进行适当的干扰协调，会造成基站间干扰的加剧和发射功率的巨大浪费[1]。

双层斯坦伯格博弈是一种处理不同等级理性参与者相互间利益决策的方法，已被广泛应用于分析和解决分层网络的资源分配问题[2]。文献[3,4]应用斯

收稿日期：2016-03-28；改回日期：2016-10-09；网络出版：2016-11-16
*通信作者：邵鸿翔 shaohongxiang2003@163.com
基金项目：国家自然科学基金(61471395, 61401508)，江苏省自然科学基金(BK20161125)
Foundation Items: The National Natural Science Foundation of China (61471395, 61401508), The Natural Science Foundation of Jiangsu Province, China (BK20161125)

坦伯格博弈模型研究了双层网络中功率分配和干扰控制的问题。然而这些文献的研究都是假设所有用户和基站间信道状态信息(Channel State Information, CSI)已知，并据此做相应的决策。但是在实际情况下，由于无线信道的动态随机特性，现有模型中不同层级间的基站用户完美获取相互间信道信息并不实际。如果使用以往在完美信道信息条件下得到的资源分配策略很可能使实际系统的性能恶化。在优化领域有两种方法用来处理信息的不确定，分别是基于概率分布的贝叶斯模型[5]和考虑极端情况的鲁棒优化模型[6]。在已知不确定信息发生概率分布的前提下，贝叶斯模型是用其期望值来表示不确定信息，但现实中信息的分布却难以得到；由于现实中环境参数的不确定度往往是有界的，鲁棒控制理论通过假定不确定度在一定范围内变化来进行建模[7]。另外，现有的工作大都是考虑连续数值的资源分配问题。相比连续的资源分配策略，离散策略的资源分配方式可简化传输设计和数据处理，降低基站之间的信息交换开销，如在 3GPP LTE 蜂窝网络中就只支持离散功率控制的下行传输[8]。

本文将基于频谱复用的下行异构蜂窝网络模型，研究在不完美信道信息条件下双层网络的分布式离散策略资源分配问题。通过引入干扰付费机制，建立鲁棒离散策略的斯坦伯格博弈模型。针对常用离散策略决策中使用的强化 Q 学习方法收敛速度慢的问题，提出一种改进的分布式双层 Q 学习高效算法寻找稳定解，并探讨不确定因素对参与者的决策的影响。

2　系统模型和问题描述

下行链路的 OFDM 双层蜂窝网络模型如图 1 所示。MBS 和 SBSs 分享复用网络频谱资源。每个基站间通过数字用户线链接，作为控制信道用来交换信息。为便于分析，假设每个基站在一个时隙只服务一个用户。因为 SBS 与 MBS 复用相同的频段，就不可避免地发生不同基站间的跨层和同层干扰。为了保护 MBS 内宏用户的通信质量，我们使用干扰价格对下层 SBS 的发射功率加以约束，并限定 SBS 对 MBS 的累积干扰必须小于门限值 Z。这样，SBS 需要优化自己的功率策略来获取干扰代价和自身通信收益的平衡。而上层 MBS 希望在干扰满足限定约束的条件下，尽可能提高对下层 SBS 干扰收费的总收益。斯坦伯格博弈是一种存在双层结构的非合作博弈，可用于本文去联合优化上下层用户的效用。上层博弈参与者作为 leader，具有强势地位，首先做出决策并向下层广播；下层参与者 follower 是跟随关系，根据上层的决策从可能的策略集中选择对

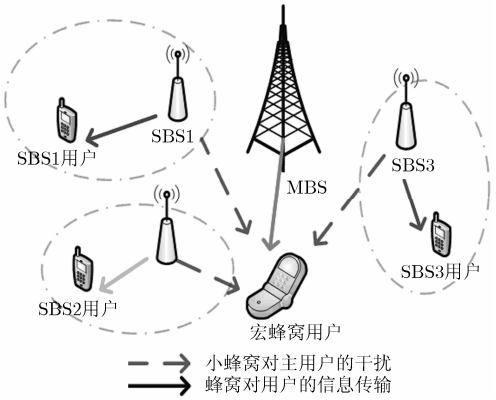


图 1 异构双层网络模型

自己最有利的策略。本文采用单 leader 多 follower 形式。MBS 作为 leader 首先行动，发布单位干扰定价；SBSs 作为 follower，根据上层 MBS 的定价，选择最优功率分配策略来最大化其效用收益。

对于下层小蜂窝，SBS i 接收到的信干噪比可写为

$$\gamma_i(p_i,p_{-i})=\frac{p_i h_{ii}}{\sum_{j\neq i}p_j h_{ji}+\sigma_0},\ \forall i\in\{1,2,\cdots,N\} \quad (1)$$

式(1)中，σ_0 代表高斯噪声功率，p_i 表示下层 SBS i 的发射功率，p_{-i} 表示除了 SBS i 外的其他 SBS 的功率策略，h_{ji} 表示 SBS j 对 SBS i 用户干扰的信道增益，$i,j\in\{1,2,\cdots,N\}$，N 为 SBS 的总数，则 $\sum_{j\neq i}p_j h_{ji}$ 代表使用同频信道的其他基站对 SBS i 用户带来的干扰。所以下层 SBS i 的效用函数可以定义为

$$\begin{aligned}\boldsymbol{U}_i(p_i,p_{-i},u_i,\lambda_0)&=W\log_2(1+\gamma_i(p_i,p_{-i}))\\&\quad-u_i p_i-\lambda_0 g_{i0}p_i\end{aligned} \quad (2)$$

式(2)由 3 部分组成，分别表示 SBS 的容量收益，功耗代价和 SBS 对 MBS 带来的干扰，其中 W 表示带宽，g_{i0} 表示 SBS i 对 MBS 宏用户的信道增益，u_i 表示单位能耗定价，λ_0 表示单位干扰定价。下层 SBS 必须选择合适的功率策略最大化自己的效用。对于每个 SBS 而言，优化问题可建模为式(3)所示的问题 1：

$$\max_{P_i}\ \boldsymbol{U}_i(p_i,p_{-i},u_i,\lambda_0) \quad (3)$$

对于上层 MBS，其目标是在其干扰可承受的范围内，最大化下层 SBS 对其干扰的累加付费收益。所以上层的优化目标可建立为带约束优化问题，如式(4)所示的问题 2：

$$\left.\begin{aligned}&\max_{\lambda_0}\boldsymbol{U}_0(\lambda_0,p_i)=\sum_i^N\lambda_0 g_{i0}p_i\\&\text{s.t.}\quad\sum_i^N g_{i0}p_i\le Z\end{aligned}\right\} \quad (4)$$

信道不确定条件下的鲁棒斯坦伯格博弈

假设基站只知道自身的信道增益 h_{ii}，但并不确切知道同层干扰的信道增益 h_{ji} 和跨层干扰的信道增益 g_{i0}。我们把信道增益表示为标称值和不确定值的求和形式，即 $g_{i0}=\overline{g}_{i0}+\Delta g_{i0}$，$h_{ji}=\overline{h}_{ji}+\Delta h_{ji}$。本文从信道信息不确定引起的最差情况出发，将斯坦伯格博弈问题转化为双层的最大最小化问题。

下层 SBS 的效用函数可转化为

$$\begin{aligned}&\max\min\ \boldsymbol{\mathcal{U}}_i\left(p_i,p_{-i},u_i,\lambda_0\right)\\&=\ W\log_2\left(1+\frac{p_ih_{ii}}{\sum\limits_{j\neq i}p_j\left(\overline{h}_{ji}+\Delta h_{ji}\right)+\sigma_0}\right)\\&\quad-u_ip_i-\lambda_0(\overline{g}_{i0}+\Delta g_{i0})p_i\end{aligned}\tag{5}$$

类似地，上层 MBS 的效用函数转化为

$$\left.\begin{aligned}&\max\min\ \boldsymbol{\mathcal{U}}_0(\lambda_0,p_i)=\sum_i^N\lambda_0\left(\overline{g}_{i0}+\Delta g_{i0}\right)p_i\\&\text{s.t.}\quad\sum_i^N\left(\overline{g}_{i0}+\Delta g_{i0}\right)p_i\le Z\end{aligned}\right\}\tag{6}$$

利用柱形模型[9]和柯西不等式，信道增益不确定分量的上界及由不确定所带来的最大干扰可分别表征为

$$|\Delta g_{i0}|\le\varepsilon_{i0},$$

$$\sum_{j\neq i}p_j\Delta h_{ji}\le\left[\sum_{j\neq i}\left|p_j\right|^2\sum_{j\neq i}\left|\Delta h_{ji}\right|^2\right]^{\frac{1}{2}}\le\varepsilon_{ji}\sqrt{\sum_{j\neq i}p_j^2}\tag{7}$$

其中，ε 表示不确定上界。利用式(7)，式(5)和式(6)的最大最小化问题可被简化为如式(8)所示的问题 3 和式(9)所示的问题 4：

$$\begin{aligned}\boldsymbol{\mathcal{U}}_i=&\max_{Pi}\ W\log_2\left(1+\frac{p_ih_{ii}}{\sum\limits_{j\neq i}p_jh_{ji}+\varepsilon_{ji}\sqrt{\sum\limits_{j\neq i}p_j^2}+\sigma_0}\right)\\&-u_ip_i-\lambda_0(g_{i0}+\varepsilon_{i0})p_i\end{aligned}\tag{8}$$

$$\left.\begin{aligned}&\boldsymbol{\mathcal{U}}_0=\max_{\lambda_0}\sum_i^N\lambda_0(g_{i0}-\varepsilon_{i0})p_i\\&\text{s.t.}\quad\sum_i^N(g_{i0}+\varepsilon_{i0})p_i\le Z\end{aligned}\right\}\tag{9}$$

3 斯坦伯格博弈模型

斯坦伯格博弈是一种双层博弈的模型，下层效用式(8)和上层效用式(9)一同形成斯坦伯格博弈。博弈的目标是找到斯坦伯格均衡(Stackelberg Equilibrium, SE)，使得上下层用户都不能通过单独改变其策略而得到自身效用的提高。下面分析所提双层斯坦伯格博弈均衡。

首先，将本文的斯坦伯格博弈表示为

$$\breve{\boldsymbol{G}}=\left\{\boldsymbol{N}\cup\{0\},\{\boldsymbol{\lambda}\},\{\boldsymbol{p}_i\}_{i\in\boldsymbol{N}},\{\boldsymbol{\mathcal{U}}_0\},\{\boldsymbol{\mathcal{U}}_i\}_{i\in\boldsymbol{N}}\right\}\tag{10}$$

其中，$\boldsymbol{N}$ 表示下层 SBS 的用户集合，$\{0\}$ 表示上层的 MBS,$\{\boldsymbol{\lambda}\}$ 表示上层的定价策略集，$\{\boldsymbol{p}_i\}$ 表示下层 SBS i 的功率策略集，$\boldsymbol{\mathcal{U}}_0$，$\boldsymbol{\mathcal{U}}_i$ 分别表示上下层的效用函数。

定义 1 斯坦伯格均衡：如果上层 MBS 的策略 λ_0^* 和下层 SBS i 的策略 p_i^* 可以分别最大化上层效用 $\boldsymbol{\mathcal{U}}_0$ 和下层效用 $\boldsymbol{\mathcal{U}}_i$，使得对于每个参与者对任何策略集$\{\lambda_0,p_i\}$，都满足式(11)的关系，则策略集$(\lambda_0^*,p_i^*)\in\boldsymbol{\lambda}\otimes\boldsymbol{P}$ 就是斯坦伯格均衡点，$\otimes$ 代表笛卡尔积。p_{-i}^* 代表除了 SBS i 的其他 SBS 的最佳策略。

$$\left.\begin{aligned}&\boldsymbol{\mathcal{U}}_0\left(\lambda_0^*,p^*\right)\ge\boldsymbol{\mathcal{U}}_0\left(\lambda_0,p^*\right)\\&\boldsymbol{\mathcal{U}}_i\left(p_i^*,\lambda_0^*,p_{-i}^*\right)\ge\boldsymbol{\mathcal{U}}_i\left(p_i,\lambda_0^*,p_{-i}^*\right)\end{aligned}\right\}\tag{11}$$

斯坦伯格均衡是本文所提博弈的稳定解，它意味着没有参与者可以通过单方面的改变策略来提高自己的效用。找出稳定均衡解是非合作博弈建模的基础和首要问题，下面将证明本文提出的博弈具有唯一 SE。

定理 1 本文定义的最优价格策略和最优功率策略$(\lambda_0^*,p_i^*)\in\boldsymbol{\lambda}\times\boldsymbol{P}$ 是所提博弈的唯一 SE。

性质 1 当对于所有 SBS 拥有的策略集$\{\boldsymbol{p}_i\}_{i\in\boldsymbol{N}}$ 是整个欧式空间的凸紧集。$\boldsymbol{\mathcal{U}}_i$ 是关于 p_i 的连续拟凸函数，则下层博弈 $\boldsymbol{g}=\{\boldsymbol{N},\{\boldsymbol{p}_i\}_{i\in\boldsymbol{N}},\{\boldsymbol{\mathcal{U}}_i\}_{i\in\boldsymbol{N}}\}$，存在唯一纳什均衡。

证明 因为 SBS 的策略空间被定义为 $P_i=p_{\max}\times p_i,\{p_i:0\le p_i\le1\}$，所以 $\boldsymbol{p}_i$ 是欧式空间的一个凸紧集。下层的效用函数 $\boldsymbol{\mathcal{U}}_i$ 在策略空间是连续的，因此只需要证明其凸性。下层效用函数关于 p_i 求一阶导数可得

$$\begin{aligned}\frac{\partial\boldsymbol{\mathcal{U}}_i}{\partial p_i}=&\frac{W}{\ln2}\frac{h_{ii}}{\sum\limits_{j\neq i}p_jh_{ji}+\varepsilon_{ji}\sqrt{\sum\limits_{j\neq i}p_j^2}+\sigma_0+p_ih_{ii}}\\&-u_i-\lambda_0(g_{i0}+\varepsilon_{i0})\end{aligned}\tag{12}$$

当其等于零时，可求出下层 SBS i 的最优功率，如式(13)所示：

$$\begin{aligned}p_i=&\left[\frac{W}{\ln2}\frac{1}{u_i+\lambda_0(g_{i0}+\varepsilon_{i0})}\right.\\&\left.-\frac{\sum\limits_{j\neq i}p_jh_{ji}+\varepsilon_{ji}\sqrt{\sum\limits_{j\neq i}p_j^2}+\sigma_0}{h_{ii}}\right]^+\end{aligned}\tag{13}$$

从式(13)中可以看出，功率策略 p_i 是关于上层定价的函数，其中 $[\cdot]^+=\max(\cdot,0)$ 表示传输功率非负。然后，再对下层效用函数关于 p_i 求二阶导数得

$$\frac{\partial^2 \boldsymbol{\mathcal{U}}_i}{(\partial p_i)^2}=\frac{-W}{\ln 2}\frac{(h_{ii})^2}{\left(\sum_{j\neq i}p_j h_{ji}+\varepsilon_{ji}\sqrt{\sum_{j\neq i}p_j^2}+\sigma_0+p_i h_{ii}\right)^2}<0 \tag{14}$$

由式(14)可知，下层效用是一个凸函数，一定存在最优极值。　　证毕

性质 2　在式(14)计算出 p_i 的基础上，上层效用式(9)的拉格朗日函数 $\mathcal{L}\boldsymbol{\mathcal{U}}_0$ 关于定价策略 $\{\lambda_0\}$ 是联合凸的。

证明　把式(13) p_i 代入上层效用，列出上层的 KKT 条件，上层带干扰约束问题式(9)的拉格朗日函数可写为

$$\begin{aligned}\mathcal{L}\boldsymbol{\mathcal{U}}_0=&\sum_{i=1}^{R}\left(\frac{W}{\ln 2}\frac{1}{u_i+\lambda_0\left(g_{i0}+\varepsilon_{i0}\right)}\right.\\&\left.-\frac{\sum_{j\neq i}p_j h_{ji}+\varepsilon_{ji}\sqrt{\sum_{j\neq i}p_j^2}+\sigma_0}{h_{ii}}\right)\left(g_{i0}-\varepsilon_{i0}\right)\lambda_0\\&-\alpha\left(\sum_{i=1}^{R}\left(\frac{W}{\ln 2}\frac{1}{u_i+\lambda_0\left(g_{i0}+\varepsilon_{i0}\right)}\right.\right.\\&\left.\left.-\frac{\sum_{j\neq i}p_j h_{ji}+\varepsilon_{ji}\sqrt{\sum_{j\neq i}p_j^2}+\sigma_0}{h_{ii}}\right)\right.\\&\left.\cdot\left(g_{i0}-\varepsilon_{i0}\right)-Z\right)+\beta\lambda_0\end{aligned} \tag{15}$$

与性质 1 类似，同样求得上层拉格朗日函数 $\boldsymbol{\mathcal{U}}_0$ 的二阶导数得

$$\frac{\partial^2 \mathcal{L}\boldsymbol{\mathcal{U}}_0}{(\partial\lambda_0)^2}=-\sum_{i=1}^{R}\frac{W}{\ln 2}\frac{2(g_{i0}-\varepsilon_{i0})[u_i+\alpha\lambda_0(g_{i0}+\varepsilon_{i0})]}{(u_i+\lambda_0(g_{i0}+\varepsilon_{i0}))^3}<0 \tag{16}$$

因为上层效用函数关于 λ_0 是连续的，并且关于 λ_0 的二阶导数小于零，所以在下层给定功率的情况下，上层效用关于定价是个凸函数，也一定存在均衡解。　　证毕

由性质 1 和性质 2 可证明上下两层都具有均衡解，所以所提博弈具有 SE，定理 1 成立。

4　分布式双层 Q 学习算法

强化学习是一种动态规划算法，具有处理离散决策问题的优势，主要应用在最优控制理论中。本节将在文献[10,11]的所提强化 Q 学习思路的基础上，针对学习效率低的问题，提出改进型双层 Q 学习算法来找到均衡解。在本文所提的双层博弈框架中，每个参与博弈的用户都具有有限离散策略集。定义用户 i 的可用策略集为 $\boldsymbol{S}_i=(s_{i,1},s_{i,2},\cdots,s_{i,|S_i|})$，$i\in N\cup\{0\}$，$|S_i|$ 表示策略集的个数。定义其在第 t 次迭代时，各策略概率矢量为 $\boldsymbol{\pi}_i^t=(\pi_i^t(s_{i,1}),\pi_i^t(s_{i,2}),\cdots,\pi_i^t(s_{i,|S_i|}))$，同时满足概率和 $\sum_{a=1}^{|S_i|}\pi_i^t(s_{i,a})=1$。这样，用户 i 的期望效用可表示为

$$u_i\left(\boldsymbol{p}_i^t,\boldsymbol{p}_{-i}^t\right)=\mathrm{E}\left[\boldsymbol{\mathcal{U}}_i\left|\boldsymbol{p}_i^t,\boldsymbol{p}_{-i}^t\right.\right]=\sum_{s^t\in S}\boldsymbol{\mathcal{U}}_i(\boldsymbol{s}^t)\prod_{i\in\boldsymbol{N}\cup\{0\}}\pi_{i,a_i}^t \tag{17}$$

其中 $\boldsymbol{s}^t=\{s_{0,a_0}^t,s_{1,a_1}^t,\cdots,s_{N,a_N}^t\}$ 代表各用户 i 基于目前的策略概率集 $\boldsymbol{p}_i^t$ 选出的策略。那么对于上层 MBS、下层 SBS 的最大化效用目标可分别写为 $\max\limits_{\boldsymbol{p}_0}u_0(\boldsymbol{p}_0,\boldsymbol{p}_{-0})$ 和 $\max\limits_{\boldsymbol{p}_i}u_i(\boldsymbol{p}_i,\boldsymbol{p}_{-i})$。

定义 2　当任意策略选择同时满足上下层基站效用 $u_0(\boldsymbol{p}_0^*,\boldsymbol{p}_i^*)\geq u_0(\boldsymbol{p}_0,\boldsymbol{p}_i^*)$ 和 $u_i(\boldsymbol{p}_i^*,\boldsymbol{p}_{-i}^*)\geq u_i(\boldsymbol{p}_i,\boldsymbol{p}_{-i}^*)$ 时，则策略选择 $(\boldsymbol{p}_0^*,\boldsymbol{p}_i^*)$ 是双层学习的稳定策略解，及 SE。

定理 2　在上层 MBS 给定 $\boldsymbol{p}_0$ 的情况下，下层 SBS 一定存在一个混合策略解 $(\boldsymbol{p}_i,\boldsymbol{p}_{-i},\boldsymbol{p}_0)$ 满足 $u_i(\boldsymbol{p}_i^*,\boldsymbol{p}_{-i}^*,\boldsymbol{p}_0)\geq u_i(\boldsymbol{p}_i,\boldsymbol{p}_{-i}^*,\boldsymbol{p}_0)$，从而得到下层的纳什均衡。上述定理 2 的证明请见文献[10]。

在 Q 学习过程中，用户的策略被参数化为 Q 函数，它表示每个特定策略的相对效用。参与博弈的用户每次改变策略都将带来即时回报。通过不断尝试，用户最后会选择最大化长期回报的最优行动策略[12]。定义用户 i 在第 t 次迭代时基于策略概率 $\boldsymbol{p}_i^t$ 所选的策略 s_{i,a_i}^t 的 Q 函数为 $Q(s_{i,a_i}^t)$。通过用户之间的策略和环境交互，得到每个策略的相应回报奖励，更新 Q 函数。在选择策略 s_{i,a_i}^t 后，相应的 Q 值通过式(18)更新：

$$Q_i^{t+1}\left(s_{i,a_i}^{t+1}\right)=\left(1-\kappa_i^t\right)Q_i^t\left(s_{i,a_i}^t\right)+\kappa_i^t u_i\left(s_{i,a_i}^t,\boldsymbol{p}_{-i}^t\right) \tag{18}$$

其中，κ_i^t 代表学习速率，满足 $\sum_{t=0}^{\infty}\kappa_i^t=\infty$，$\sum_{t=0}^{\infty}\left(\kappa_i^t\right)^2<\infty$。

$$u_i\left(s_{i,a_i}^t,\boldsymbol{p}_{-i}^t\right)=\sum_{a_{-i}^t\in S_{-i}}\boldsymbol{\mathcal{U}}_i\left(s_{i,a_i}^t,\boldsymbol{S}_{-i}^t\right)\prod_{j\in\boldsymbol{N}\cup\{0\}/i}\pi_{j,a_j}^t$$

是用户 i 在第 t 次迭代选择策略的期望回报。其中 $\boldsymbol{S}_{-i}^t=\left[s_{0,a_0}^t,s_{1,a_1}^t,\cdots,s_{i-1,a_j}^t,s_{i+1,a_1}^t,\cdots,s_{N,a_N}^t\right]$ 且 $S_{-i}=\otimes_{j\in\boldsymbol{N}\cup\{0\}/i}S_j$。每个基站用户根据式(19)的玻尔兹曼分布来更新其策略。

$$\boldsymbol{p}_i^t\left(s_{i,a_i}\right)=\frac{\exp\left[Q_i^t\left(s_{i,a_i}^t\right)/\psi_i\right]}{\sum_{a_i\in S}\exp\left[Q_i^t\left(s_{i,a_i}^t\right)/\psi_i\right]} \tag{19}$$

其中，$\psi_i > 0$ 是温度系数，用来控制策略选择是倾向探测还是利用。根据式(18)和式(19)，上层 MBS 通过迭代更新对应 Q 函数。假设上层 MBS 每 c 时段更新一次定价策略。在双层学习迭代算法中，作为唯一的公共信息，上层的 MBS 首先向下层所有 SBS 发布定价。下层接收到干扰价格后，通过学习算法找到各自的最优响应功率策略，然后在每个时间段终点反馈回上层 MBS，以便上层 MBS 根据下层上报的功率策略信息更新自己的出价策略。算法是嵌套迭代循环方式，流程如图 2 所示。

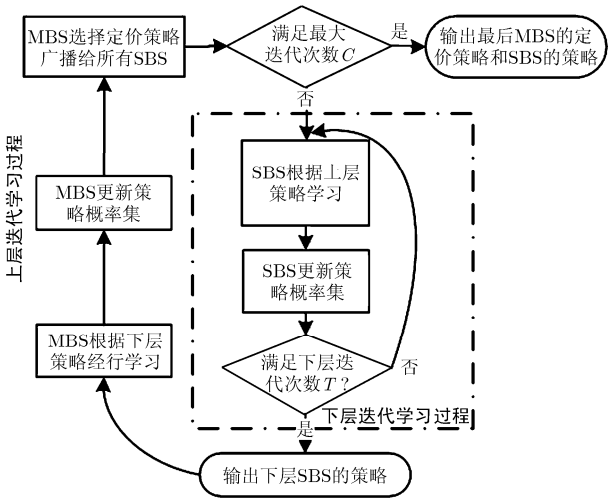


图 2 双层 Q 学习算法流程图

下层 SBS i 的 Q 函数通过式(20)更新：

$$Q_i^{t+1}\left(s_{i,a_i}^{t+1}\right)=\left(1-\kappa_i^t\right)Q_i^t\left(s_{i,a_i}^t\right)+\kappa_i^t\overline{u}_i\left(s_{i,a_i}^t,s_{0,a_0}\right) \quad (20)$$

其中估计的期望效用 $\overline{u}_i\left(s_{i,a_i}^t,s_{0,a_0}\right)$ 可表示为

$$\overline{u}_i^t(s_{i,a_i})=\begin{cases}\dfrac{\boldsymbol{\mathcal{U}}_i\left(s_{i,a_i}^t,S_{-i}^t\right)-\overline{u}_i^t\left(s_{i,a_i}^t,s_{0,a_0}\right)}{n_i^t\left(s_{i,a_i},s_{0,a_0}\right)+1}\\ \quad+\overline{u}_i^{t-1}(s_{i,a_i}^t,s_{0,a_0}), & s_{i,a_i}=s_{i,a_i}^t\\ \overline{u}_i^{t-1}\left(s_{i,a_i}^t\right), & \text{其他}\end{cases} \quad (21)$$

其中 $n_i^t(s_{i,a_i},s_{0,a_0})$ 表示在一个时间段内上下层合并选择为 (s_{i,a_i},s_{0,a_0}) 的次数。可看出上层 MBS 和下层 SBS 的更新是基于不同的时间单位的，上下层用户策略的更新都是基于对方迭代更新后的结果通过 Q 学习得到的。下层在每 T 个时隙结束时执行式(20)，完成其 Q 函数的更新。类似地，上层 MBS 用户在每个时间段 c 结束时执行式(22)，完成其 Q 函数的更新：

$$Q_0^{c+1}\left(s_{0,a_0}\right)=\left(1-\kappa_0\right)Q_0^c\left(s_{0,a_0}\right)+\kappa_0 u_0^c\left(s_{0,a_0},\boldsymbol{p}_{-i}^{cT}\right) \quad (22)$$

在实际算法运行过程中，当用户的策略集相对较大时，收敛的速度将指数增加。而文献[12]的算法，每次只更新一个策略的 Q 值，无法满足双层迭代的速率要求。如果能更高效利用交互信息，在一次迭代中更新所有策略的 Q 值，算法收敛速度会有明显提升。本文提出改进型双层 Q 学习算法，具体步骤如表 1 所示。

表 1 改进型双层 Q 学习算法

改进型双层 Q 学习算法

步骤 1 上层学习过程 c=1:C。初始化所有用户 Q 函数 $Q_i(s_{i,a_i})=0, s_{i,a_i}\in S_i$； $\boldsymbol{p}_i^1$ 为各策略等概率分布。

步骤 2 下层学习过程 t=1:T

(1)在每个时间段开始，MBS 根据其策略概率集 $\boldsymbol{p}_0$，选择一个定价策略 s_{0,a_0}，并广播给所有的下层 SBS。

(2)每个 SBS i 根据自己的策略概率集 $\boldsymbol{p}_i^t$ 选择各自功率策略 $s_{i,ai}$。

(3)每个 SBS i 根据反馈信息计算其效用 $\boldsymbol{\mathcal{U}}_i\left(s_{i,a_i}^t,S_{-i}^t\right)$，并根据式(21)更新其估计期望效用 $\overline{u}_i^t(s_{i,a_i})$。

(4)每个 SBS i 根据式(8)计算其他 $|S_i|-1$ 个策略的效用 $\boldsymbol{\mathcal{U}}_i(s_{i,k}^t,S_{-i}^t), s_{i,k}\in S_i/s_{i,c_i}$。

(5)每个 SBS i 根据式(20)和式(19)更新其 Q 值和策略概率集。

(6)在 T 时隙结束，所有 SBS 把最后策略传给 MBS。

步骤 3 MBS 计算其第 c 个时间段的效用 $u_0^c(s_{0,a_0},\boldsymbol{p}_{-i}^{cT})$，并根据式(22)和式(19)更新其 Q 值和策略概率集。

步骤 4 MBS 根据其已更新的策略概率集选择上层策略。

步骤 5 $c=c+1$，直到 $c=C$ 最大时间段数。

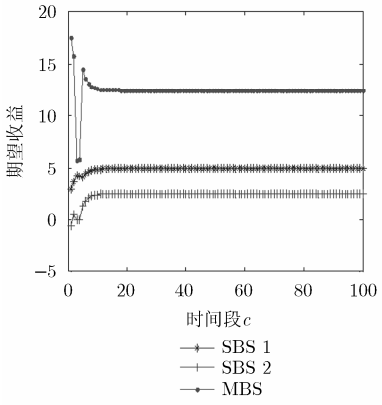
5 仿真结果

本节将通过仿真来分析所提博弈模型和改进型双层Q学习算法的性能。不失一般性，考虑一个MBS和两个SBS组成的异构网络，每个基站服务一个用户。设置SBS1和SBS2对MBS宏用户的标称信道增益分别为 $g_{10}=0.2,\ g_{20}=0.3$，下层SBS间的标称干扰信道增益分别是 $h_{1,2}=h_{2,1}=0.1$，归一化SBS对其自身用户的信道增益为 $h_{1,1}=h_{2,2}=1$。噪声功率 $\sigma_0=0.01$ dBmW。设MBS的干扰价格策略集为[2.5,3.0,3.5,4.0,4.5],SBS的功率分配策略集为 $p_{\max}$[0, 0.10,0.15,0.20,0.25]，其中SBS的最大传输功率 $p_{\max}=100$ dBmW。设置每个时间段由 $T=100$ 个时隙组成，上层迭代时间段数 $C=100$。对于不确定模型，我们假设不确定度是随标称值线性变化 $\Delta g_{i0}=\Theta\times\overline{g}_{i0}$，不确定部分服从均匀分布 $\Theta\sim U(-\theta,\theta)$，$\theta$ 表示不确定值与标称值的比例。因此我们得到不确定界 $\varepsilon=\theta\times\overline{g}_{i0}$。

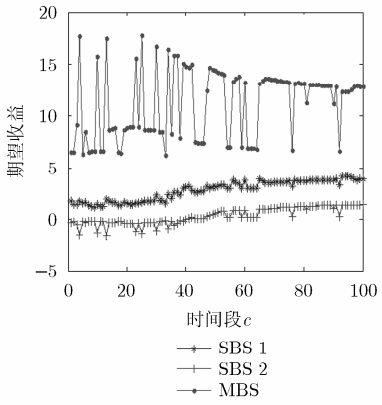
首先研究算法得到SE的收敛性。当不确定度增加时，只是效用函数中的信道数值发生变化，博弈参与者策略的选择有所不同，但收敛形式类似，所以我们以完美信道条件为例说明算法的收敛性。图

3、图4、图5分别给出了Q学习算法改进前后收敛性能比较，可看出改进算法的收敛速度和收敛效果都要好于原算法，且改进算法中各离散策略经过较少迭代便可达收敛到一个纯策略，而原算法只能收敛到一个混合策略。

图6展示了上层MBS的干扰约束对其收益的影响，在保护上层MBS传输的前提下，上层MBS对于干扰的容忍度越大则收益越多。另外，上层拥有先动优势，下层只是被动接受调整，所以在条件变动的情况下，上层总是获得尽可能多的收益。图7给出了采用鲁棒建模方法前后，随不确定度等级θ变化时，MBS收益情况。随着不确定度θ等级的增加，信道状态相对估计标称值恶化加剧。采用鲁棒建模MBS的收益比原MBS收益有较大改善。对于提出的鲁棒方法，上层MBS考虑了最差信道状态信息，MBS随着信道变化而根据收益情况，自适应改变了自己的相应定价策略，使得总收益总是向着自己有利的方式改变。

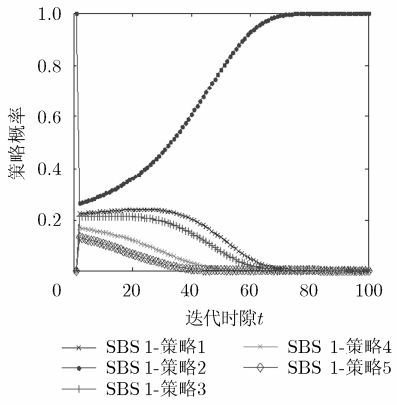


(a)改进型Q学习算法

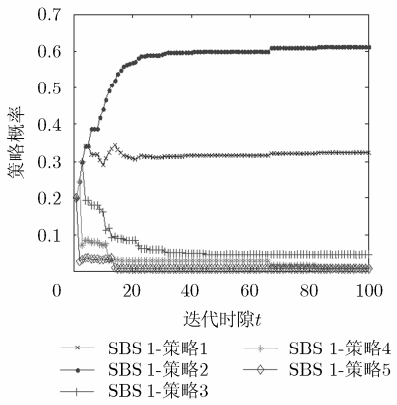


(b)原Q学习算法

图3 两种算法的期望收益比较

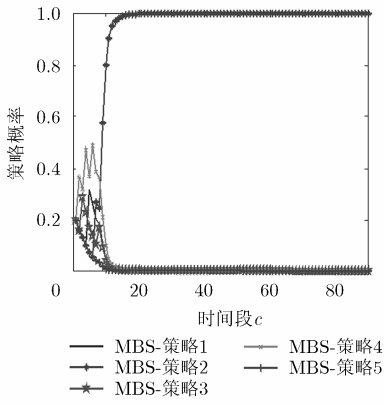


(a)改进型Q学习算法

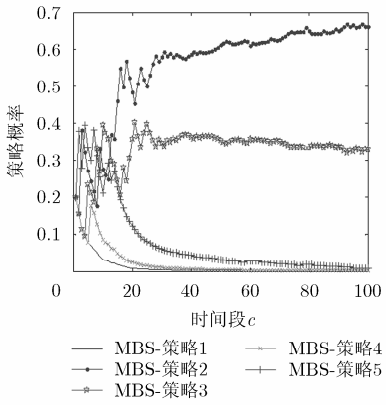


(b)原Q学习算法

图4　SBS1策略的两种算法各策略收敛性比较



(a)改进型Q学习算法



(b)原Q学习算法

图5　MBS策略的两种算法各策略收敛性比较

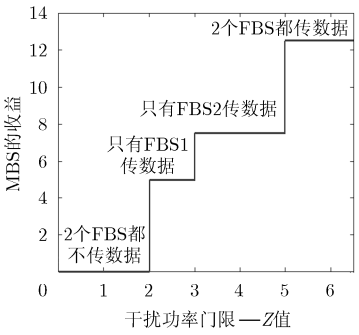


图6 不同干扰门限值下的MBS收益

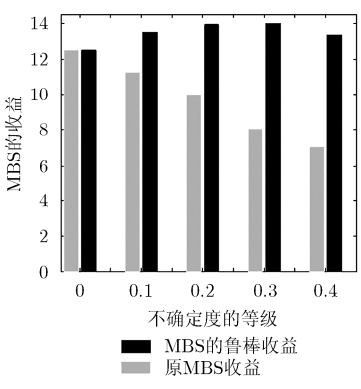


图7 不同不确定度下的MBS收益

6 结论

本文针对无线异构网络中实际信道信息获取不完美，从而导致信道不确定度变化引起的用户收益减少问题，提出了一种基于斯坦伯格博弈模型的分布式鲁棒资源分配方案。证明了所提双层博弈模型均衡的存在性和唯一性。针对用户采用离散策略集方式，提出一种改进型的分布式双层Q学习算法。通过仿真表明，本文设计的鲁棒模型能有效抑制随不确定度变化带来的用户收益减少的问题。与原算法相比，所提算法在收敛速度和性能上都有较大提升。

参 考 文 献

[1] ZAHIR T, ARSHAD K, NAKATA A, *et al*. Interference management in femtocells[J]. *IEEE Communication Survey & Tutorials*, 2013, 15(1): 293–311. doi: 10.1109/SURV.2012.020212.00101.

[2] HAN Zhu, NIYATO D, SAAD W, *et al*. Game Theory in Wireless and Communication Networks[M]. Cambridge: UK, Cambridge University Press, 2012: 88–91.

[3] 扶奉超, 张志才, 路兆铭, 等. Femtocell 双层网络中基于 Stackelberg 博弈的节能功率控制算法[J]. 电子科技大学学报, 2015, 44(3): 363–368.
FU Fengchao, ZHANG Zhicai, LU Zhaoming, *et al*. Energy-efficient power control algorithm based on Stackelberg game in two-tier femtocell Networks[J]. *Journal of University of Electronic Science and Technology of China*, 2015, 44(3): 363–368.

[4] LASHGARI M, MAHAM B, KEBRIAEI H, *et al*. Distributed power allocation and interference mitigation in two-tier femtocell networks: A game-theoretic approach[C]. Wireless Communications and Mobile Computing Conference, Dubrovnik, Croatia, 2015: 55–60.

[5] DUONG N D, MADHUKUMAR A S, and NIYATO D. Stackelberg Bayesian game for power allocation in two-tier networks[J]. *IEEE Transactions on Vehicular Technology*, 2016, 65(4): 2341–2354. doi: 10.1109/TVT.2015.2418297.

[6] ZHU Kun, HOSSAIN E, and ANPALAGAN A. Downlink power control in two-tier cellular OFDMA networks under uncertainties: A robust Stackelberg game[J]. *IEEE Transactions on Communications*, 2015, 63(2): 520–535. doi: 10.1109/TCOMM.2014.2382095.

[7] 吴敏, 何勇. 鲁棒控制理论[M]. 北京: 高等教育出版社, 2010.

[8] ZHANG H, VENTURINO L, PRASAD N, *et al*. Weighted sum-rate maximization in multi-cell networks via coordinated scheduling and discrete power control[J]. *IEEE Journal on Selected Areas in Communications*, 2011, 29(6): 1214–1224. doi: 10.1109/JSAC.2011.110609.

[9] YANG K, WU Y, and HUANG J. Distributed robust optimization for communication networks[C]. IEEE Infocom Conference, Phoenix, AZ, USA, 2008: 1157–1165. doi: 10.1109/ INFOCOM.2008.171.

[10] FUDENBURG D and TIROLE J. Game Theory[M]. Cambridge, MA, USA, The MIT Press, 1991: 29–34.

[11] CHEN X, ZHANG H, CHEN T *et al*. Improving energy efficiency in femtocell networks: A hierarchical reinforcement learning framework[C]. IEEE International Conference on Communications (ICC), Budapest, Hungary, 2013: 2241–2245. doi: 10. 1109/ICC.2013.6654861.

[12] WATKINS C and DAYAN P. Q-learning[J]. *Journal of Machine Learning Research*, 1992, 8(1): 279–292.

邵鸿翔： 男，1983 年生，博士生，讲师，研究方向为异构无线网络资源分配、博弈论、电磁频谱管理.

赵杭生： 男，1962 年生，博士，博士生导师，研究方向为异构无线网络资源分配、电磁频谱管理.

孙有铭： 男，1988 年生，博士生，研究方向为异构无线网络、超密集组网、资源分配、强化学习.

孙丰刚： 男，1982 年生，博士生，讲师，研究方向为无线通信传输技术、阵列信号处理.