# 基于多智能体深度强化学习的分布式干扰协调

刘婷婷，罗义南，杨晨阳
（北京航空航天大学电子信息工程学院，北京 100191）

**摘 要**：针对干扰网络中的文件下载业务，提出了一种基于多智能体深度强化学习的分布式干扰协调策略。所提策略能够在节点之间只需交互少量信息的条件下，根据干扰环境和业务需求的特点自适应调整传输策略。仿真结果表明，对于任意的用户数和业务需求，所提策略相对于未来信息预测理想时最优策略的用户满意度损失不超过 11%。

**关键词**：多智能体深度强化学习；非实时业务；分布式干扰协调；超密集网络
**中图分类号**：TN929.53
**文献标识码**：A
**doi:** 10.11959/j.issn.1000−436x.2020149

# Distributed interference coordination based on multi-agent deep reinforcement learning

LIU Tingting, LUO Yi’nan, YANG Chenyang
School of Electronic and Information Engineering, Beihang University, Beijing 100191, China

**Abstract:** A distributed interference coordination strategy based on multi-agent deep reinforcement learning was investigated to meet the requirements of file downloading traffic in interference networks. By the proposed strategy transmission scheme could be adjusted adaptively based on the interference environment and traffic requirements with limited amount of information exchanged among nodes. Simulation results show that the user satisfaction loss of the proposed strategy from the optimal strategy with perfect future information does not exceed 11% for arbitrary number of users and traffic requirements.

**Key words:** multi-agent deep reinforcement learning, non-realtime traffic, distributed interference coordination, ultra-dense network

## 1 引言

为了支撑业务的爆炸性增长，第五代及未来移动通信网络的基站越来越密集，小区间干扰严重制约了小区边缘用户的体验以及整个网络的吞吐量[1]。未来无线网络需要支持各种类型的业务，如实时业务（如视频会议、游戏等）和非实时业务（如文件下载、视频点播等）。面向对传输速率、时延以及可靠性有着不同要求的业务，如何有效管理干扰、提升用户体验和网络性能是亟待解决的问题之一。

对于具有时延容忍性的非实时业务，为了保证服务质量并节省网络资源，可以优化未来一段时间内的会话级性能[2]。解决这类问题的直接方法是根据用户的服务质量需求建模预测资源分配问题，在未来信息预测理想的假设下通过求解优化问题来获得最优的资源管理策略（如干扰协调[3]）。因此，网络中的中心节点需要收集并预测所有用户当前和未来的业务需求及信道状态信息，然后把预测的信息当作未来信息的真值进行资源分配，从而得到分配策略。这种集中式预测资源分配或干扰协调求解最优策略的计算复杂度和所需要的信令开销随着网络规模呈指数级增长。此外，当干扰等网络环

**收稿日期**：2020−01−10；**修回日期**：2020−03−07
**基金项目**：国家自然科学基金资助项目（No.61731002, No.61671036）
**Foundation Item:** The National Natural Science Foundation of China (No.61731002, No.61671036)

境发生变化时，还需要重新优化协调策略。

由于集中式干扰协调面临信息预测不准、复杂度高、开销大、可扩展性差等问题，文献[4-13]研究了分布式干扰协调策略。目前，分布式干扰协调主要包括基于分布式优化[4]、博弈论[5-6]以及多智能体强化学习（MARL, multi-agent reinforcement learning）[7-13]的方法。

基于分布式优化的干扰对准策略[4]在最小化加权均方误差的准则下交替优化收发机算法，通过迭代的方式对准干扰，提升网络吞吐量。基于博弈论的干扰协调策略[5-6]把每个基站看作相互博弈的玩家，致力于做出可以最大化自己小区容量的决策。经过反复博弈，所有玩家达到纳什均衡点，从而获得最优传输策略。文献[4-6]中策略的设计目标是提升干扰网络的瞬时吞吐量或和数据率，但没有考虑不同业务的需求。

与分布式优化和博弈论相比，强化学习能解决序贯决策问题。当业务需求和信道状态等信息未知时，智能体通过试错的方式不断与环境进行交互来优化干扰协调策略，使长期累计回报最大。智能体直接根据状态选择动作，能够在动态变化的干扰环境中自适应调整策略[14]。目前，基于强化学习的干扰协调主要包括基于单智能体强化学习的策略[15]和基于MARL 的策略[7-13]。在单智能体强化学习的策略[15]中，智能体需要收集用户的数据率等状态信息来联合优化所有用户的动作，从而获得接近最优的协调策略。因此，该策略与集中式干扰协调类似，存在收集信息信令开销大、计算复杂度高、可扩展性差的问题。在基于 MARL 设计的系统中，每个用户是一个独立的智能体，通过把单智能体的状态与动作进行分解，减小了输入输出的维度，降低了计算复杂度[16]。基于MARL 设计的策略可扩展性强，适用于动态变化的干扰环境。已有文献利用 MARL 设计分布式干扰协调策略来最大化小区长期容量[7-11]、最大化网络覆盖率[12]，以及最小化传输延时[13]。

在基于 MARL 的干扰协调策略中，状态是反映干扰环境和用户需求的重要因素，如何在不交换或少交换信息的条件下选择合适的状态是设计分布式策略的关键。为了表示每个用户受到的干扰程度，文献中考虑的状态变量包括基站与用户间的距离[7-8]、接入基站的用户个数[9]、接收信干噪比（SINR, signal to interference plus noise ratio）的强弱[7,10]、信干比[11]和接收干扰功率[13]。以上状态只考虑了用户受到的干扰，没有考虑用户所产生的干扰。为了更准确地表示用户之间的相互干扰，文献[12]中设计的状态变量包含了对每个用户影响最大的几个干扰源和被干扰源，其中干扰源的影响根据各用户接收的干扰功率大小来衡量，被干扰源的影响通过用户所接入基站对其他小区用户造成的干扰功率占其他小区各用户接收的总干扰功率的百分比来衡量。但是，一个用户对其他用户的干扰百分比无法通过这个用户自身观测得到，需要所有用户交换来自干扰基站的接收功率，从而导致较大的通信开销。文献[7-12]没有考虑业务特点，文献[13]的分布式干扰协调策略则考虑了车联网中的文件下载业务，在设计状态时用需要传输的剩余数据量和剩余传输时间来反映每个用户的业务需求。然而，因为用户之间会竞争资源，所以在设计分布式干扰协调策略时，不仅需要考虑自身的业务需求，还需要考虑其他用户的需求。如何针对业务特点设计有效的分布式干扰协调策略至今尚未解决。

本文面向文件下载业务，研究基于 MARL 的分布式干扰协调策略，主要贡献如下。

1) 所提分布式干扰协调策略在用户数较多、业务需求较高的情况下优于传统的集中式干扰协调策略。

2) 在所设计的 MARL 状态中，综合考虑了干扰环境、业务需求及用户体验的特点，使网络节点之间只需少量的信息交互就能在动态环境下自适应调整传输策略。

3) 所提策略不需要预测任何信息，在任意用户数和业务需求下，相对于未来信息预测理想时最优预测资源分配策略的性能损失小于 11%。

## 2　系统模型

### 2.1　业务模型

考虑一个如图 1 所示的热点区域，$G$ 个基站服务 $K$ 个移动用户。每个用户配置单天线、每个基站配置多个天线，基站 $g$ 有 $M_g$ 个天线，$g=1,\cdots,G$。

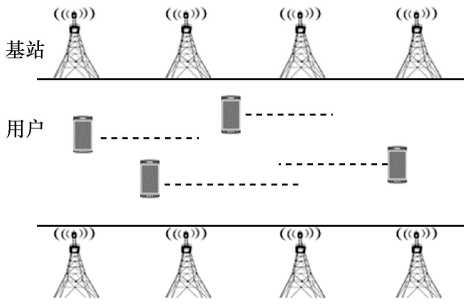


图 1　小区网络示意

$K$ 个用户向基站发出文件下载业务请求。用户 $k$ 在 $t_k^{\text{start}}$ 时刻请求下载数据量为 $B_k$ 的文件，期望在 $t_k^{\text{exp}}$ 时刻之前完成传输，如图 2 所示。

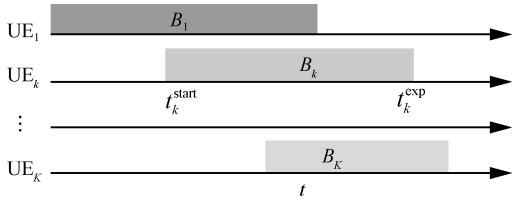


图 2 用户业务需求

对于文件下载业务，如果能够在用户期望的截止时间 $t_k^{\text{exp}}$ 之前完成文件传输，则用户非常满意；如果未能按时完成文件传输，用户不会立即放弃正在下载的业务，而是等待一段时间直到文件传输完成，但用户满意度会随着时延的增加而不断下降。

为了评估用户体验，文献[17]根据实测的超文本传输业务的用户体验提出了一种文件下载业务的满意度性能指标，使用倒 S 型效用函数描述用户的满意度与传输时延的关系。如图 3 所示，若在 $t_k^{\text{exp}}$ 之后完成文件下载，则用户满意度随着传输时延的增加而下降，且满意度的下降速度随着传输时延的增加而变化。当接近用户期望的截止时间时，增加时延会导致满意度有较大的下降；当用户已经等待很久而非常不满意的时候，增加传输时延反而使满意度的下降变得缓慢。

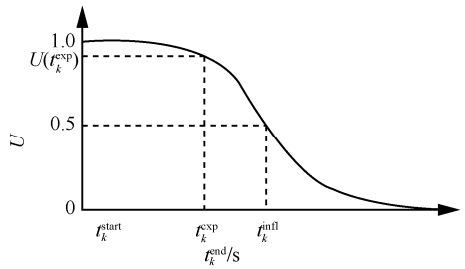


图 3 效用函数

本文采用文献[17]提出的效用函数来描述用户的满意度，则用户 $k$ 的满意度可以表示为

$$U(t_k^{\text{end}})=\frac{1}{1+\exp\left(c(t_k^{\text{end}}-t_k^{\text{infl}})\right)} \tag{1}$$

其中，参数 $c$ 控制曲线的陡峭程度；$t_k^{\text{end}}$ 表示文件实际完成传输的时间；$t_k^{\text{infl}}=\left(t_k^{\text{start}}+q_k(t_k^{\text{exp}}-t_k^{\text{start}})\right)$ 是满意度的转折点，且满足 $U(t_k^{\text{infl}})=0.5$；$q_k>1$，$q_k$ 越大，表明用户 $k$ 可容忍的时延越长。

## 2.2 干扰协调

本节研究基站如何通过选择合适的时间资源服务用户来协调干扰。

首先，确定时间资源分配的粒度。假设平均信道增益在一帧（秒级，记为 $T_f$）内保持不变，在不同帧之间可能发生变化；瞬时信道增益在同一时隙（毫秒级，记为 $T_s$）内保持不变，在不同时隙之间统计独立同分布。因此，一帧内包含 $N_s=\dfrac{T_f}{T_s}$ 个时隙。当基站服务多个用户时，需要给用户分配正交的时间资源来避免干扰。如果以时隙为单位进行资源分配，每个帧内可同时服务多个用户。如果以帧为单位进行资源分配，每个帧内只服务一个用户。文献[2-3]的研究表明，对于非实时业务，平均信道增益是影响用户体验的主要因素，以帧为单位分配资源不仅可以达到与以时隙为单位分配资源相近的性能，还能大幅减少资源分配时需要优化的变量数以及所需的信令开销。因此，本文研究以帧为单位的干扰协调策略。

考虑就近接入，即每个用户接入平均接收功率最大的基站。为了避免频繁切换，每个用户接入的基站在一帧内保持不变。令 $g_k^l$ 表示用户 $k$ 在第 $l$ 帧内接入的基站号，则

$$g_k^l=\arg\max_i\left\{\alpha_{k,i}^l\right\} \tag{2}$$

其中，$\alpha_{k,i}^l$ 表示第 $l$ 帧内基站 $i$ 与用户 $k$ 间的平均信道增益。

用二进制变量 $a_k^l\in\{0,1\}$ 表示资源分配策略，$a_k^l=1$，表示在第 $l$ 帧服务用户 $k$，所接入的基站 $g_k^l$ 需要开启；$a_k^l=0$，表示在第 $l$ 帧不服务用户 $k$。为了不对其他用户产生干扰，基站 $g_k^l$ 保持静默。

因为用户 $k$ 在 $t_k^{\text{start}}$ 时刻发起文件请求，所以它从第 $b_k=\left\lceil\dfrac{t_k^{\text{start}}}{T_f}\right\rceil$ 帧开始被服务，其中 $\lceil\cdot\rceil$ 表示向上取整。当对用户 $k$ 累计传输的数据量超过 $B_k$ 时，文件传输完成。因此，完成时间可以表示为

$$t_k^{\text{end}}=\min\left\{l\middle|\forall\sum_{i=b_k}^{l}D_k^l\geqslant B_k\right\}T_f \tag{3}$$

其中，$D_k^l$ 是第 $l$ 帧内传输给用户 $k$ 的比特数。

当用户 $k$ 请求的文件传输完成后，就不再服务

该用户，即 $a_k^l=0,\forall l>e_k$，$e_k=\dfrac{t_k^{\text{end}}}{T_f}$。因此，协调干扰需要优化的变量为

$$a_k^l\in\{0,1\},\forall b_k\leqslant l\leqslant e_k,k=1,\cdots,K \tag{4}$$

### 2.3　性能指标

当 $a_k^l=1$ 时，用户 $k$ 在第 $t$ 时隙的瞬时接收 SINR 为

$$\gamma_k^{l,t}=\frac{P_{k,g_k^l}^{l,t}}{\sum\limits_{i\in\mathcal{U}^l,i\neq k}a_i^lP_{k,g_i^l}^{l,t}+\sigma_n^2} \tag{5}$$

其中，$\mathcal{U}^l=\{k|b_k\leqslant l\leqslant e_k,\forall k\}$ 表示在第 $l$ 帧内可以服务的用户集合；$P_{k,g}^{l,t}=P_g\left|(\boldsymbol{h}_{k,g}^{l,t})^{\mathrm{T}}\boldsymbol{w}_g^{l,t}\right|^2$ 是基站 $g$ 到用户 $k$ 的接收信号功率，$P_g$ 是基站 $g$ 的发射功率；$\boldsymbol{h}_{k,g}^{l,t}\in\mathbb{C}^{M_g\times1}$ 是信道向量，满足 $\mathbb{E}\left\{\left\|\boldsymbol{h}_{k,g}^{l,t}\right\|^2\right\}=M_g\alpha_{k,g}^l$，$\boldsymbol{w}_g^{l,t}=\dfrac{(\boldsymbol{h}_{k,g}^{l,t})^*}{\left\|\boldsymbol{h}_{k,g}^{l,t}\right\|}$ 是基站 $g$ 的预编码向量，为了最大化接收信噪比，考虑最大比传输；$\sigma_n^2$ 是噪声功率。

根据式(5)可以得到用户 $k$ 在第 $l$ 帧内每个时隙的瞬时数据率，进而得到第 $l$ 帧传输的数据量为

$$D_k^l=a_k^lWT_s\sum_{t=1}^{N_s}\log_2\left(1+\frac{P_{k,g_k^l}^{l,t}}{\sum\limits_{i\in\mathcal{U}^l,i\neq k}a_i^lP_{k,g_i^l}^{l,t}+\sigma_n^2}\right) \tag{6}$$

其中，$W$ 是系统的带宽。

把式(6)代入式(3)，可以得到用户 $k$ 的文件传输完成时间 $t_k^{\text{end}}$；再代入式(1)，可以得到用户 $k$ 的满意度，进而得到网络中所有用户的平均满意度为

$$\eta=\frac{1}{K}\sum_{k=1}^{K}U(t_k^{\text{end}}) \tag{7}$$

## 3　基于 MARL 的分布式干扰协调

本节首先把最优干扰协调建模为求解优化问题，然后简要讨论集中式干扰协调存在的问题，最后提出基于 MARL 的分布式干扰协调策略。

### 3.1　问题建模

对于文件下载业务，如式(7)所示，用户满意度在传输完成后才能获得。为了设计最优干扰协调策略，中心节点首先需要预测未来一段时间范围（称为规划窗，长度记为 $T$）内所有用户的信道状态信息和业务需求，然后把预测的信息当作未来信息的真值进行资源分配，这种传输策略称为预测资源分配[3]。假设规划窗内有 $N_f=\dfrac{T}{T_f}$ 个帧，规划窗的长度需要保证所有用户都能传输完所请求的文件。

为了最大化网络中所有用户的满意度，可以把式(7)作为优化目标，即把预测资源分配问题建模为

$$\begin{aligned}&\max_{a_k^l}\ \eta\\&\text{s.t. }\ a_k^l\in\{0,1\},\forall l\in[b_k,e_k],k=1,\cdots,K\end{aligned} \tag{8}$$

式(8)是一个非凸的组合优化问题，需要通过暴力搜索获得最优解，其计算复杂度为 $\mathcal{O}(2^{KN_f})$。

通过求解式(8)，可以得到预测资源分配策略，求解的计算复杂度随着用户数 $K$ 和规划窗长度 $N_f$ 呈指数级增长。另外，中心节点还需要已知在整个规划窗内所有用户与相邻基站间的平均信道增益、以及各个用户的业务需求参数 $\{t_k^{\text{start}},t_k^{\text{exp}},B_k\}$。对已经发出请求的用户，可以获取需求参数，并根据其最近的历史轨迹来预测未来的运动轨迹，再结合信号地图获得未来的平均信道增益；对于没有发出请求的用户，无法获取需求参数和预测移动轨迹。

即使不考虑计算复杂度，且假设需要预测的信息理想，预测资源分配策略只能对规划窗范围内的时间资源进行最优分配，并且若在规划窗内干扰环境和网络规模发生变化则会导致性能损失。

### 3.2　分布式干扰协调

在分布式系统中，基站需要独立决定是否在当前帧服务用户，因此把每个基站视为智能体。本文采用 MARL 来提升用户满意度，首先设计深度 Q 网络（DQN, deep Q network）的动作、状态和奖励函数，然后介绍 DQN 的训练与执行过程。

#### 3.2.1　动作

当采用 MARL 时，基站 $g_k^l$ 根据用户 $k$ 上报的信息决定是否在第 $l$ 帧服务用户 $k$，因此动作变量就是式(4)中的资源分配变量，如式(9)所示。

$$a_k^l=\{0,1\},\forall l\in[b_k,e_k] \tag{9}$$

其中，$b_k$ 和 $e_k$ 分别为用户 $k$ 开始被服务和文件传输结束的帧号。

#### 3.2.2　状态

因为需要根据网络的干扰情况和用户的需求来设计干扰协调策略，所以用户 $k$ 所接入基站 $g_k^l$ 的状态由干扰环境状态和用户需求状态两部分组成，如式(10)所示。

$$s_k^l = \left\{ s_{k-\inf}^l, s_{k-\text{req}}^l \right\} \tag{10}$$

其中，$s_{k-\inf}^l$ 和 $s_{k-\text{req}}^l$ 分别反映干扰环境和用户需求。

**3.2.2.1** 设计与干扰环境有关的状态 $s_{k-\inf}^l$

在 MARL 框架中，每个智能体根据局部观测来进行决策，因此设计状态变量时需要考虑以下因素。1) 可观测。状态中的信息是单个智能体能够独立观测得到的，否则就需要额外的通信开销来交换信息。2) 维度尽可能小。强化学习的训练复杂度随状态空间的规模增加[18]，因此需要忽略影响较小的状态、保留影响较大的状态。3) 表示智能体间的相互影响。

在干扰网络中，由于每个用户不仅受到其他基站的干扰，所接入的基站还对其他用户产生干扰，设计状态变量时需要综合考虑两方面的影响。因此，文献[12]中设计的状态包含了每个干扰源与被干扰源单独产生的影响，导致状态空间的规模太大。事实上，在干扰网络中，直接选择 SINR 和信漏噪比（SLNR, signal to leakage and noise ratio）就可以反映每个用户受到的干扰和所接入基站对其他用户所产生干扰的平均影响。因此，与干扰环境有关的状态可以设计为

$$s_{k-\inf}^l = \{\gamma_k^l, \mu_k^l\} \tag{11}$$

其中，$\gamma_k^l = \frac{1}{N_s}\sum_{t=1}^{N_s}\gamma_k^{l,t}$ 和 $\mu_k^l = \frac{1}{N_s}\sum_{t=1}^{N_s}\mu_k^{l,t}$ 分别为第 $l$ 帧的平均 SINR 与 SLNR，$\gamma_k^{l,t}$ 和 $\mu_k^{l,t}$ 分别为第 $t$ 时隙内的瞬时 SINR 与 SLNR。

根据式(5)可知，接收干扰功率 $P_{k,g}^{l,t}$ 与瞬时信道信息有关，获取这些信息将造成很大的信令开销。因为非实时业务的用户体验主要取决于平均信道增益[2]，所以可用平均信道增益代替瞬时信道增益来计算平均 SINR 和平均 SLNR。第 $l$ 帧的平均接收功率可近似为 $P_{k,g}^l = \frac{1}{N_s}\sum_{t=1}^{N_s}P_{k,g}^{l,t} \approx \mathrm{E}\{P_{k,g}^{l,t}\}$，把其代入式(5)，不难得到近似的平均 SINR 为

$$\gamma_k^l \approx \frac{P_{g_k^l} M_{g_k^l} \alpha_{k,g_k^l}^l}{\sum_{i\in\mathcal{U}^l, i\neq k} a_i^l P_{g_i^l} \alpha_{k,g_i^l}^l + \sigma_n^2} \tag{12}$$

同理，平均 SLNR 可近似为

$$\mu_k^l \approx \frac{P_{g_k^l} M_{g_k^l} \alpha_{k,g_k^l}^l}{\sum_{i\in\mathcal{U}^l, i\neq k} P_{g_k^l} \alpha_{i,g_k^l}^l + \sigma_n^2} \tag{13}$$

根据式(12)可知，用户 $k$ 的干扰功率 $I_k^l = \sum_{i\in\mathcal{U}^l, i\neq k} a_i^l P_{g_i^l} \alpha_{k,g_i^l}^l$ 与其他用户所接入基站的动作有关。然而，在分布式系统中，由于所有基站同时做出决策，基站 $g_k^l$ 无法获知其他基站当前的动作。考虑到在密集干扰网络中平均接收干扰功率的变化幅度较小，因此可以把上一帧的干扰功率作为当前帧干扰功率的预测值，即 $I_k^l \approx I_k^{l-1} = \sum_{i\in\mathcal{U}^{l-1}, i\neq k} a_i^{l-1} P_{g_i^{l-1}} \alpha_{k,g_i^{l-1}}^{l-1}$。因为上一帧的动作是已知的，所以上一帧的干扰功率可以准确估计，则当前帧平均 SINR 的预测值为

$$\gamma_k^l \approx \frac{P_{g_k^l} M_{g_k^l} \alpha_{k,g_k^l}^l}{I_k^{l-1} + \sigma_n^2} \tag{14}$$

用户 $k$ 根据式(14)计算平均SINR并上报给接入的基站 $g_k^l$。

对于式(13)中的平均 SLNR，基站 $g_k^l$ 对其他用户产生的总干扰强度主要由其对附近几个用户产生的强干扰决定。因为附近的用户在选择是否接入基站 $g_k^l$ 时会向该基站上报接收功率 $P_{g_k^l}\alpha_{i,g_k^l}^l$，所以基站 $g_k^l$ 可以直接根据式(13)计算平均 SLNR。

**3.2.2.2** 设计与业务需求有关的状态 $s_{k-\text{req}}^l$

对于文件下载业务，文献[13]给出了描述业务需求状态的变量，由一个二元组构成 $\{\overline{B}_k^l, \overline{T}_k^l\}$。

$$\overline{B}_k^l = B_k - \sum_{i=b_k}^{l-1} D_k^i \tag{15}$$

$$\overline{T}_k^l = t_k^{\exp} - (l-1)T_f \tag{16}$$

其中，$\overline{B}_k^l$ 和 $\overline{T}_k^l$ 分别表示用户 $k$ 在第 $l$ 帧开始时刻剩余的数据量和时间。

对于文件下载业务，这样考虑是合理的，因为影响用户性能的不是已经花费了多少时间给用户传输了多少数据量，而是还剩余多少时间、还需要传输多少数据。然而，在分布式干扰网络中，多个基站之间相互竞争资源，基站 $g_k^l$ 不能只根据用户 $k$ 的需求做决策，还需要关注其他用户的需求。因此，若直接用 $K$ 个用户对应的 $\{\overline{B}_k^l, \overline{T}_k^l\}$ 来描述所有用户的需求，不仅会增加描述业务需求的变量数，还会增加共享需求带来的信令开销。而且，这样训练的网络只能工作在给定用户数 $K$ 的场景中，当网络规模（即用户数）动态变化时还需要重新训练。为了使与需求有关的状态能够适应网络规模的动态变化，本文引入一个非负变量 $\xi_k^l$ 来表示为使用户 $k$ 满意在第 $l$ 帧所需的数据率；用

变量 $\xi_{k^-}^l = \sum_{i=1,i\neq k}^{K} \xi_i^l$ 表示其他用户需求的数据率之和。根据 $\xi_k^l$ 和 $\xi_{k^-}^l$，基站 $g_k^l$ 可以准确地了解用户 $k$ 的需求大小及其相对于其他用户需求的紧急程度，从而可以选择一个更激进或更保守的策略。因此，可以把与业务需求有关的状态设计为

$$s_{k-\text{req}}^l = \{\xi_k^l, \xi_{k^-}^l\} \tag{17}$$

由于式(16)中的 $\overline{T}_k^l$ 是相对于期望截止时间 $t_k^{\text{exp}}$（而非真实截止时间 $t_k^{\text{end}}$）的剩余时间，不能直接用 $\frac{\overline{B}_k^l}{\overline{T}_k^l}$ 来描述用户 $k$ 需求的数据率 $\xi_k^l$。首先，当传输时隙接近 $t_k^{\text{exp}}$ 时，$\overline{T}_k^l \to 0$，$\frac{\overline{B}_k^l}{\overline{T}_k^l} \to \infty$，导致神经网络输入的动态范围过大，影响算法的收敛性能。其次，如果用户 $k$ 未能在期望截止时间内完成传输，即 $\overline{T}_k^l < 0$，导致 $\frac{\overline{B}_k^l}{\overline{T}_k^l}$ 是负值，无法反映用户真实的需求。为了解决上述问题，当 $\overline{T}_k^l > 0$ 时，引入一个数据率门限 $R_{\max}$ 来限制每个用户需求的最大数据率。当 $\overline{T}_k^l \leqslant 0$ 时，由于已经超过用户 $k$ 期望的截止时间 $t_k^{\text{exp}}$，用户满意度开始下降。从式(1)中定义的用户满意度可知，如果当前时刻没有超出 $t_k^{\text{exp}}$ 太多，则尽快完成传输可以给用户提供一个较高的满意度，即需要给该用户分配更多资源；如果目前的传输时间已经超出 $t_k^{\text{exp}}$ 太多，用户满意度已经有较大下降，此时给用户分配太多资源并不能有效改善用户满意度，即该用户的业务已经没有那么紧急，可以把资源分配给其他更加紧急的用户。可见，需要根据用户满意度来调整用户对数据率的需求。因为第 $l$ 帧起始时没有完成传输，用户 $k$ 的满意度小于 $U\left((l-1)T_f\right)$，本文采用用户 $k$ 满意度的最大值 $U\left((l-1)T_f\right)$ 对 $R_{\max}$ 进行加权作为用户 $k$ 需求的数据率，即 $\xi_k^l = U\left((l-1)T_f\right)R_{\max}$。根据上面的分析，与用户 $k$ 需求的数据率有关的状态变量可设计为

$$\xi_k^l = \begin{cases} \min\left\{\dfrac{\overline{B}_k^l}{\overline{T}_k^l}, R_{\max}\right\} & \forall \overline{T}_k^l > 0 \\ U((l-1)T_f)R_{\max} & \forall \overline{T}_k^l \leqslant 0 \end{cases} \tag{18}$$

为了理解所设计的表示用户需求的状态，图 4 给出了当给定 $\overline{B}_k^l$ 时，由式(16)计算得到的剩余时间 $\overline{T}_k^l$ 和由式(18)计算得到需求的数据率随时间 $t$ 变化的曲线，此时对应的帧号为 $l = \left\lceil \frac{t}{T_f} \right\rceil$。可以看出，当 $\overline{T}_k^l > 0$ 时，随着 $\overline{T}_k^l$ 逐渐减小至 0，$\xi_k^l$ 先增加至 $R_{\max}$ 后保持不变；当 $\overline{T}_k^l \leqslant 0$ 时，$\xi_k^l$ 随着用户满意度单调下降。

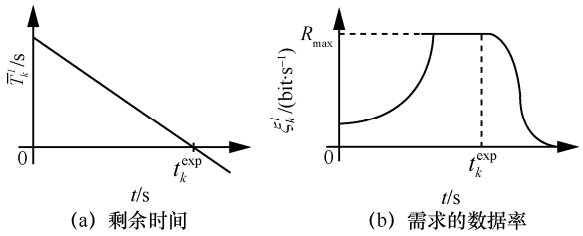


图 4　给定剩余数据量时的剩余时间与需求的数据率

这样设计的基站 $g_k^l$ 对用户 $k$ 的状态变量 $s_{k-\text{req}}^l = \{\xi_k^l, \xi_{k^-}^l\}$ 不仅包含了用户 $k$ 需求的数据率 $\xi_k^l$，还包含了其他用户需求的数据率之和 $\xi_{k^-}^l$。当网络规模扩大时，$\xi_{k^-}^l$ 随着用户个数 $K$ 而增加，因此，$\xi_{k^-}^l$ 的大小可以反映网络规模的动态变化，从而使所设计的干扰协调策略对网络变化具有稳健性。

在实现的过程中，每个用户在第 $l$–1 帧结束之后先根据式(15)和式(16)计算自身的剩余数据量和剩余时间，再根据式(18)计算 $\xi_k^l$。为了获得 $\xi_{k^-}^l$，所有用户要向接入基站发送自己需求的数据率。为了减小基站间共享信息的信令开销，基站 $g_k^l$ 把用户 $k$ 需求的数据率上报给中心节点。中心节点计算所有用户需求的数据率之和 $\xi_{\text{sum}}^l = \sum_{k=1}^{K} \xi_k^l$，并广播给网络中的所有基站，状态获取流程如图 5 所示。基站 $g_k^l$ 可以获得其他用户的需求之和为 $\xi_{k^-}^l = \xi_{\text{sum}}^l - \xi_k^l$。

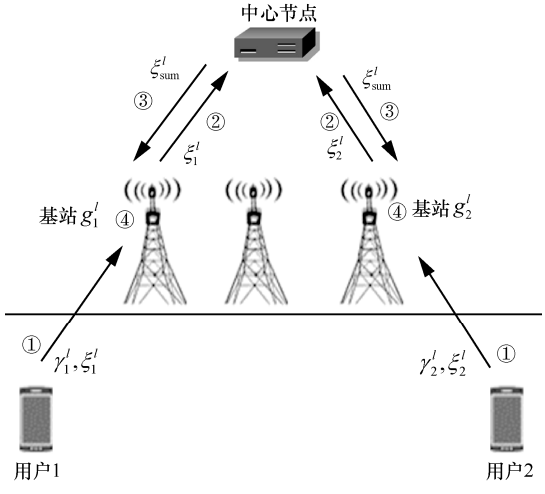


图 5　状态获取流程

这里的中心节点可以是覆盖多个基站的宏基站，或是管理多个基站的核心网网关，主要负责协调交换信息和训练 DQN。

根据式(11)、式(13)、式(14)和式(17)、式(18)，可以得到基站 $g_k^l$ 对用户 $k$ 的全部状态为

$$s_k^l = \{\gamma_k^l, \mu_k^l, \xi_k^l, \xi_{k^-}^l\} \tag{19}$$

图 5 说明了基站获取式(19)中状态信息的流程，具体包含以下步骤。

**步骤 1** 用户 $k$ 向所接入基站 $g_k^l$ 上报信干噪比 $\gamma_k^l$ 和需求的数据率 $\xi_k^l$。

**步骤 2** 基站 $g_k^l$ 向中心节点上传 $\xi_k^l$。

**步骤 3** 中心节点计算所有用户需求的数据率之和 $\xi_{\text{sum}}^l$，并广播给所有基站。

**步骤 4** 基站 $g_k^l$ 计算 $\mu_k^l$、$\xi_{k^-}^l$，并得到关于用户 $k$ 的状态 $\{\gamma_k^l, \mu_k^l, \xi_k^l, \xi_{k^-}^l\}$。

在整个流程中基站与中心节点间只需要交换少量的信息即可获得所需的状态信息，从而做出决策。

#### 3.2.3 奖励函数

在 MARL 框架中，设计奖励函数的关键因素包括：1) 把网络整体的优化目标拆分成每个基站的优化目标；2) 奖励函数需要准确评估每个基站选择的动作对本小区用户和对其他小区用户的影响。从式(8)可知，要优化的用户满意度只有在用户完成传输之后才能获得。如果把用户 $k$ 的满意度直接作为基站 $g_k^l$ 的奖励函数，则存在如下的问题：1) 基站 $g_k^l$ 只有在用户 $k$ 的文件下载完成之后才能得到奖励，这种延迟奖励会降低学习效率；2) 不能准确评估当前决策对其他用户满意度的影响。

由于式(18)中的用户状态已经把满意度的影响考虑到用户需求的数据率之中，设计奖励函数只需要关注有效传输了多少数据。与用户满意度相比，传输数据量更易于评估基站 $g_k^l$ 的当前策略对用户 $k$ 和其他小区用户的影响，因此本文把用户有效传输的数据量作为奖励。如果在第 $l$ 帧没有服务用户 $k$，即 $a_k^l = 0$，则基站 $g_k^l$ 获得的奖励为 0，即 $r_k^l = 0$；如果 $a_k^l = 1$，则奖励 $r_k^l$ 由两部分组成，一部分是第 $l$ 帧已经传输的数据量，即式(6)中的 $D_k^l$，另一部分是基站 $g_k^l$ 服务用户 $k$ 产生的干扰对其他传输用户数据量所带来的损失。

根据式(6)，不难得到基站 $g_k^l$ 服务用户 $k$ 对用户 $j$ 造成的数据量损失为

$$\Delta D_{j\backslash k}^l = WT_s \sum_{t=1}^{N_s} \log_2\left(1 + \frac{P_{j,g_i^l}^{l,t}}{I_{j\backslash k}^{l,t} + \sigma_n^2}\right) - D_j^l \tag{20}$$

其中，$I_{j\backslash k}^{l,t} = I_j^{l,t} - P_{j,g_k^l}^{l,t}$ 表示基站 $g_k^l$ 静默时，用户 $j$ 的接收干扰功率。

则奖励函数可表示为

$$r_k^l = a_k^l\left(D_k^l - \sum_{j=1, j\neq k}^{K} a_j^l \Delta D_{j\backslash k}^l\right) \tag{21}$$

对应的累计回报为

$$G_k^l = \sum_{\tau=l}^{e_k} \gamma^{\tau-l} r_k^\tau \tag{22}$$

其中，$\gamma \in [0,1]$ 是折扣因子。

由式(21)可知，当 $a_k^l = 0$ 时，基站 $g_k^l$ 没有任何奖励。而当 $a_k^l = 1$（在第 $l$ 帧服务用户 $k$）时，如果 $r_k^l > 0$，则表明用户 $k$ 增加的数据量超过了基站 $g_k^l$ 所产生的干扰给其他用户带来的损失，即在第 $l$ 帧服务用户 $k$ 会带来增益；否则用户 $k$ 增加的数据量小于其传输对其他用户造成的损失，即在第 $l$ 帧服务用户 $k$ 严重影响了其他用户的性能。

虽然从形式上看，式(21)中的奖励函数只反映了干扰状态的影响，而没有直接反映用户需求，但是在后面的仿真中可以看到，式(22)中的累计回报能够随着用户需求 $\xi_k^l$ 而增加，因此所设计的策略能根据用户需求的大小调整资源分配策略，这是因为所设计的状态已经反映了用户需求。

#### 3.2.4 训练与执行

在 MARL 框架中，最直接的训练方法是把其他智能体视为环境的一部分，每一个智能体独立训练自己的神经网络。然而，由于每个智能体只能观测部分的环境状态且环境受其他智能体决策的影响，导致这种方法不容易收敛。本文考虑一种为了解决上述问题常用的集中式训练方法[16]。在集中式训练过程中，所有基站把经验 $(s,a,r,s')$ 上传给中心节点，中心节点把经验存储到回放池 $\mathcal{D}$ 中，用经验回放池中的数据训练神经网络，然后把模型参数分发给网络中的基站。因为所有用户的任务相似，所以这种共享经验的方式可以加快收敛速度。

在 DQN 训练过程中造成发散的原因包括[19]：1) 状态随时间演变，相邻时间步的状态具有较高的相关性；2) 神经网络参数的微小更新导致策略发生很大变化，从而使样本分布变化；3) 神经网络参数

的更新导致策略更新，使优化目标随着训练过程一直在改变。采用经验回放可以有效解决前 2 个问题。为了解决第三个问题，一般采用 2 个神经网络来同时训练：用一个在线网络拟合动作值函数，记为 $Q(s,a;\theta)$；同时考虑一个目标神经网络，记为 $\hat{Q}(s,a;\theta^-)$。在线网络参数 $\theta$ 更新一段时间之后再更新目标网络的参数 $\theta^-$，从而降低目标网络与在线网络之间的相关性，避免优化目标一直变化。采用文献[20]提出的双重深度 Q 网络（DDQN, double DQN）训练方法。从经验回放池 $\mathcal{D}$ 中随机抽取一小批经验 $(s,a,r,s')$ 作为样本集合 $\mathcal{B}$ 进行训练，则这批样本上的损失函数为

$$L(\theta)=\sum_{(s,a,r,s')\in\mathcal{B}}\left(y^{\mathrm{DDQN}}-Q(s,a;\theta)\right)^2 \tag{23}$$

其中，目标值为

$$y^{\mathrm{DDQN}}=r+\gamma\hat{Q}\left(s',\arg\max_{a'}Q(s',a';\theta);\theta^-\right) \tag{24}$$

根据损失函数，利用梯度下降法对参数 $\theta$ 进行更新。

## 4　仿真结果

本节通过仿真评估所提出的分布式 DQN 策略的性能。无线网络的仿真参数如表 1 所示[21]。精调后的 DQN 算法超参数如表 2 所示。

**表 1　　无线网络仿真参数**

| 参数 | 数值 |
|---|---|
| 网络区域 | 20 m × 500 m 矩形区域 |
| 系统带宽/MHz | 20 |
| 基站数 | 40 |
| 基站发射功率/dBm | 30 |
| 基站发射天线数 | 2 |
| 用户个数 | 5～30 |
| 噪声功率/dBm | −95 |
| 路径损耗 | $38.4+36.8\log_{10}d$ |
| 规划窗长度/s | 300 |
| 帧长/s | 1 |
| 时隙长/ms | 100 |
| 用户速度/(m·s$^{-1}$) | 1.5～6.0 |
| 用户请求的文件大小 $B_k$ /MB | 100～500 |
| 用户期望的传输时间 $(T_k^{\exp}-T_k^{\mathrm{start}})$/s | 30～60 |

仿真环境如图 1 所示，基站均匀分布在道路两侧，用户从随机的起始位置出发，沿着道路做匀速直线运动，速度在 1.5～6.0 m/s 均匀分布，在用户运动经过的道路上的随机位置发出业务请求。为了便于比较，假设所有用户请求的文件大小相同，但发起请求的起始时刻 $t_k^{\mathrm{start}}$ 和期望截止时刻 $t_k^{\exp}$ 不同。用户的满意度还受延时容忍 $q$ 和曲线陡峭程度 $c$ 的影响，在仿真中设置 $q=1.5$，即当实际传输时间是期望传输时间的 1.5 倍，即 $\dfrac{t_k^{\mathrm{end}}-t_k^{\mathrm{start}}}{t_k^{\exp}-t_k^{\mathrm{start}}}=1.5$ 时，用户满意度下降到 0.5；曲线陡峭程度设为 $c=9.190\,2$，从而满足 $U(t_k^{\exp})=0.99$（表示用户若在 $t_k^{\exp}$ 时刻完成传输，则其满意度为 0.99）。为了评估最优干扰协调策略的性能，考虑规划窗长度为 300 s，保证所有用户能在规划窗内完成传输。在 DQN 训练阶段，使用 Adam 优化算法更新在线网络参数 $\theta$。

**表 2　　DQN 超参数**

| 参数 | 数值 |
|---|---|
| 输入层神经元个数 | 4 |
| 隐藏层神经元个数 | 8 层，每层 256 个 |
| 输出层神经元个数 | 2 |
| 激活函数 | ReLU |
| 学习率 | 0.001 |
| 折扣因子 | 0.1 |
| 探索率$\varepsilon$ | 0.01 |
| 经验回放池容量 | 20 000 |
| 批规模 | 256 |

为了分析所设计的分布式 DQN 当需求的数据率不同时的累计回报，图 6 给出了在固定 SINR、SLNR 和其他用户需求的数据率之和 $\xi_{k^-}^l$ 的条件下累计回报的期望值 $Q(s,a=1)$ 随着需求的数据率变化的曲线。

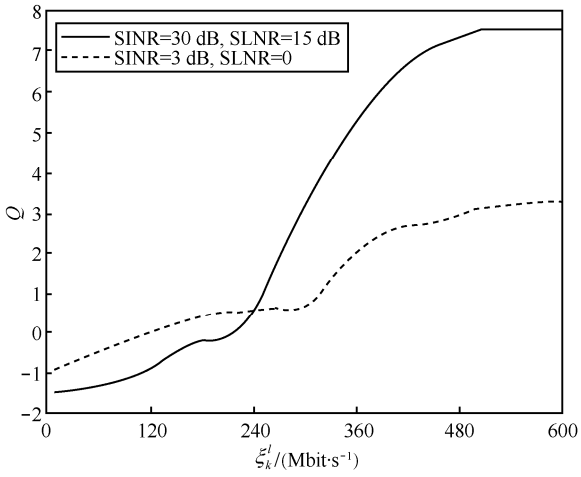


图 6　$Q(s,a=1)$随着需求的数据率变化的曲线

当 $Q(s,a=1)>0$ 时，服务该用户；当 $Q(s,a=1)<0$ 时，不服务该用户。图 6 考虑了 2 种干扰环境，一种环境的网络中干扰相对较小（SINR 和 SLNR 都比较高，如图 6 中实线所示），另一种环境的网络中干扰比较严重（SINR 和 SLNR 都比较低，如图 6 中虚线所示）。由图 6 可见，随着用户需求的数据率增加，$Q(s,a=1)$ 的取值逐渐增大，这时用户被服务的机会也越大。另外，用户被服务的机会与干扰环境密切相关，当网络中干扰比较严重时分布式策略会减少用户被服务的机会；而当网络中干扰较轻时，则会增加用户被服务的机会。可见，尽管式(21)中的奖励函数没有直接反映用户的需求，所设计的分布式策略确实可以根据用户需求和干扰环境自适应调整传输策略。

为了评估所设计的分布式干扰协调策略（简称为“分布式 DQN”）性能，与如下几种已有的策略进行比较：1) 为了评估考虑用户满意度时干扰协调的最优性能，假设存在一个超级智能体，能在资源规划窗开始时准确预测所有用户的业务请求信息 $\{t_k^{\text{start}},t_k^{\text{exp}},B_k\}$ 和平均信道，借鉴文献[3]中的方法求解式(8)（简称为“集中式（有需求）”）；2) 为了评估相对于传统非预测集中式干扰协调的增益，考虑只根据当前时刻的平均信道来协调干扰的策略，并采取随机资源分配来服务用户，这种策略以尽力而为的方式服务、无法在用户期望的时间内下载完所请求的文件（简称为“集中式（无需求）”）；3) 为了评估分布式干扰协调的增益，考虑只根据当前信道进行传输的无干扰协调策略（简称为“无协调（无需求）”）。

首先，比较不同策略所需的信息量。对于文件下载业务，所需交互的信息量与用户完成传输所需的时间有关。为了便于比较，这里假设所有用户都使用整个规划窗的资源进行传输。对于“无协调（无需求）”策略，由于不协调干扰，不需要交互任何信息。对于“集中式（无需求）”策略，基站需要向中心节点上报所有开启基站与用户间在所有帧的平均信道增益，因此需要交互的信息量为 $N_f K^2 C$，其中 $C$ 表示量化一个实数标量所需要的比特数。对于“集中式（有需求）”策略，除了上述的平均信道信息，还需要每个用户的业务请求信息 $\{t_k^{\text{start}},t_k^{\text{exp}},B_k\}$，因此中心节点需要在规划窗开始时预测的信息量为 $(3K+N_f K^2)C$。对于分布式 DQN 策略，根据图 5 所示的信息交换流程，在每一帧中用户向接入基站上报信干噪比 $\gamma_k^l$ 和需求的数据率 $\xi_k^l$，交互的信息量为 $2KC$；基站向中心节点上报需求的数据率 $\xi_k^l$，中心节点广播所有用户需求的数据率之和 $\xi_{\text{sum}}^l$，因此交互的信息量为 $N_f(3K+1)C$。表 3 比较了不同策略所需的信息量，可见当用户数 $K$ 较大时，分布式 DQN 策略所需的信息远远小于集中式干扰管理方案。

**表 3　不同策略需要交互/预测的信息量**

| 传输策略 | 获取信息方式 | 信息量 |
|---|---|---|
| 无协调（无需求） | — | 0 |
| 集中式（无需求） | 交互 | $N_f K^2 C$ |
| 集中式（有需求） | 预测 | $(3K+N_f K^2)C$ |
| 分布式 DQN | 交互 | $N_f(3K+1)C$ |

图 7 和图 8 分别比较了用户请求的文件大小 $B$ 和用户数 $K$ 不同时的平均满意度 $\eta$。表 4 和表 5 分别为根据图 7 和图 8 的结果得到的分布式 DQN 策略相对于其他 3 种策略的性能增益。

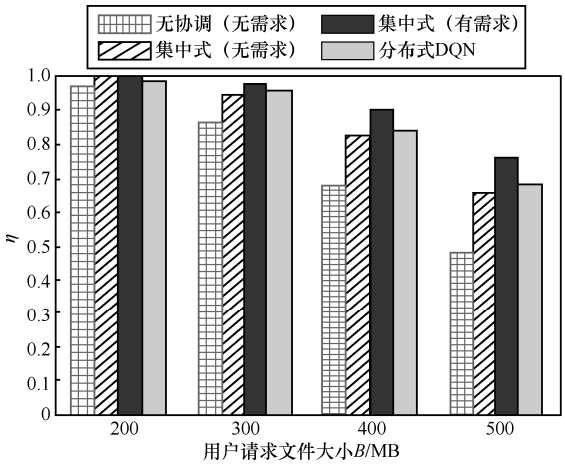


图 7　用户请求文件大小不同时的满意度（$K$=20）

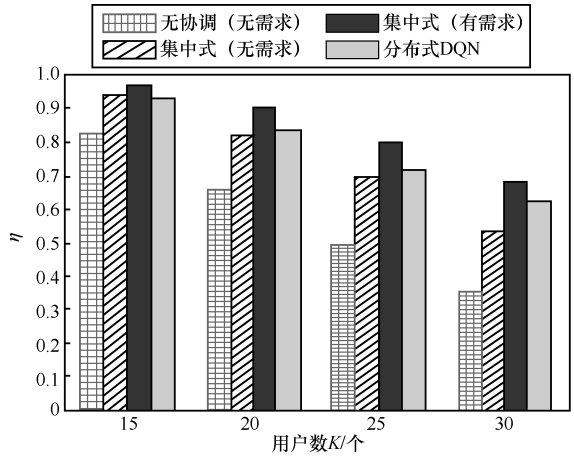


图 8　用户数不同时的满意度（$B$=400 MB）

**表 4　文件大小不同时分布式策略的性能增益($K$=20)**

| $B$/MB | 无协调（无需求） | 集中式（无需求） | 集中式（有需求） |
|---|---|---|---|
| 200 | 2.28% | −0.28% | −0.67% |
| 300 | 11.08% | 1.42% | −1.93% |
| 400 | 26.27% | 2.29% | −6.81% |
| 500 | 42.39% | 3.91% | −9.94% |

**表 5　用户数不同时分布式策略的性能增益($B$=400 MB)**

| $K$/个 | 无协调（无需求） | 集中式（无需求） | 集中式（有需求） |
|---|---|---|---|
| 15 | 13.13% | −0.74% | −3.88% |
| 20 | 26.27% | 2.29% | −6.81% |
| 25 | 45.85% | 2.89% | −10.39% |
| 30 | 66.89% | 15.64% | −8.54% |

从上述结果可见，分布式 DQN 策略相对于“无协调（无需求）”和“集中式（无需求）”策略的性能增益随着用户需求 $B$ 和用户数 $K$ 而增加。这是由于无论 $B$ 还是 $K$ 增加，网络中的干扰都变得更严重。“无协调（无需求）”没有协调干扰，而“集中式（无需求）”没有利用不同用户需求的差异性来调整干扰协调策略。分布式 DQN 策略根据用户需求和网络干扰环境动态地调整传输策略，如图 6 所示，因此干扰越严重能提供的性能增益越大。

当用户数为 30，请求的文件大小为 400 MB 时，相对于“无协调（无需求）”和“集中式（无需求）”策略，分布式 DQN 策略分别可以提供 66%和 15%以上的增益，说明在用户数多、业务需求大的情况下，所提出的分布式策略增益明显；对所仿真的任意用户数和文件大小，所提出的不需要预测信息的分布式策略相对于未来平均信道预测理想时“集中式（有需求）”策略的性能损失不超过 11%。

因为网络中不断有用户离开,或者有新用户的请求到达，所以网络中的用户数是动态变化的。为了进一步评估分布式 DQN 策略在动态变化环境中的性能，图 9 给出了当训练时的用户数与测试时用户数不一致时的性能。考虑训练时的用户数为 20，而测试时用户数从 5 增加到 30。从仿真结果可见，即使用户数不一致，所达到的性能与用户数一致时的性能几乎相同，说明分布式 DQN 策略对用户数的变化具有较强的稳健性。这是由于所设计的 DQN 状态中包含了 SINR、SLNR 以及其他用户需求的数据率之和，这些变量都随着网络中的用户数变化而变化，基站能够根据这些状态的变化自适应地调整传输策略。此外，在训练 DQN 时也考虑了网络中不断有用户离开及有新用户的请求到达的场景，值函数是在不同帧内用户数不同的条件下学习到的。

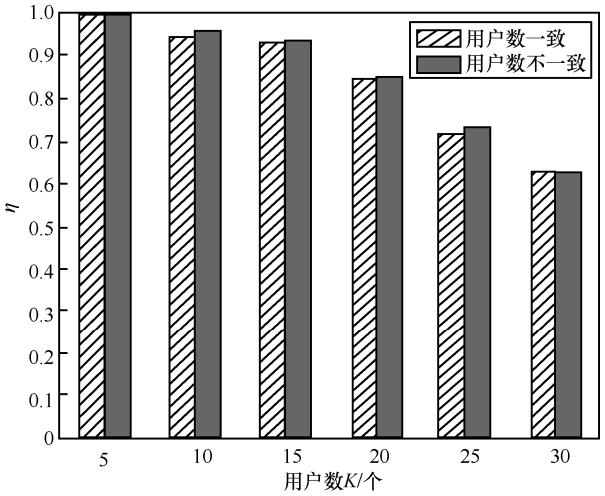


图 9　训练与测试时用户数不一致时的满意度（$B$=400 MB）

## 5　结束语

本文针对文件下载业务提出了基于多智能体强化学习的分布式干扰协调策略，设计了强化学习算法的状态和奖励函数。所提出的策略能够在不同的干扰环境和用户业务需求下自适应调整传输策略、提升用户满意度。该策略不需要预测任何信息，且网络节点之间只需要交互少量的信息。从仿真结果可见，所设计的分布式干扰协调策略能在用户数较多、业务需求较大的情况下相对于传统的集中式干扰协调明显提高用户满意度，对于任意的用户数和文件大小，相对于未来信息预测理想时最优策略的性能损失不超过 11%，并且对于用户数变化具有稳健性。

**参考文献：**

[1] TENG Y, LIU M, YU F R, et al. Resource allocation for ultra-dense networks: a survey, some research issues and challenges[J]. IEEE Communications Surveys & Tutorials, 2019, 21(3):2134-2168.

[2] YAO C, YANG C, XIONG Z. Energy-saving predictive resource planning and allocation[J]. IEEE Transactions on Communications, 2016, 64(12): 5078-5095.

[3] GUO K, LIU T, YANG C, et al. Interference coordination and resource allocation planning with predicted average channel gains for Het-Nets[J]. IEEE Access, 2018, 6(1): 60137-60151.

[4] GOMADAM K, CADAMBE V R, JAFAR S A. Approaching the capacity of wireless networks through distributed interference alignment[J]. IEEE Transactions on Information Theory, 2011, 57(6): 3309-3322.

[5] XU C, SHENG M, WANG X, et al. Distributed subchannel allocation for interference mitigation in OFDMA femtocells: a utility-based learning approach[J]. IEEE Transactions on Vehicular Technology, 2014, 64(6): 2463-2475.
[6] WANG X, ZHANG H, TIAN Y, et al. Optimal distributed interference mitigation for small cell networks with non-orthogonal multiple access: a locally cooperative game[J]. IEEE Access, 2018, 6(1): 63107-63119.
[7] GALINDO-SERRANO A, GIUPPONI L. Distributed Q-learning for aggregated interference control in cognitive radio networks[J]. IEEE Transactions on Vehicular Technology, 2010, 59(4): 1823-1834.
[8] AMIRI R, MEHRPOUYAN H, FRIDMAN L, et al. A machine learning approach for power allocation in HetNets considering QoS[C]// IEEE International Conference on Communications. Piscataway: IEEE Press, 2018:1-7.
[9] ZHANG Y, KANG C, MA T, et al. Power allocation in multi-cell networks using deep reinforcement learning[C]// IEEE Vehicular Technology Conference. Piscataway: IEEE Press, 2018:1-6.
[10] ZHAO N, LIANG Y C, NIYATO D, et al. Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks [J]. IEEE Transactions on Wireless Communications, 2019, 18(11): 5141-5152.
[11] XU Y, YU J, HEADLEY W C, et al. Deep reinforcement learning for dynamic spectrum access in wireless networks[C]// IEEE Military Communications Conference. Piscataway: IEEE Press, 2018:1-6.
[12] NASIR Y S, GUO D. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(10): 2239-2250.
[13] YE H, LI G Y. Deep reinforcement learning for resource allocation in V2V communications[J]. IEEE Transactions on Vehicular Technology, 2019, 68(4): 3163-3173.
[14] XIONG Z, ZHANG Y, NIYATO D, et al. Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges[J]. IEEE Vehicular Technology Magazine, 2019, 14(2): 44-52.
[15] LI H, GAO H, LYU T, et al. Deep Q-learning based dynamic resource allocation for self-powered ultra-dense networks[C]// IEEE International Conference on Communications Workshops. Piscataway: IEEE Press, 2018:1-6.
[16] NGUYEN T T, NGUYEN N D, NAHAVANDI S. Deep reinforcement learning for multi-agent systems: a review of challenges, solutions and applications[J]. arXiv preprint arXiv:1812.11794, 2018.
[17] PROEBSTER M, KASCHUB M, WERTHMANN T, et al. Context-aware resource allocation for cellular wireless networks[J]. EURASIP Journal on Wireless Communications and Networking, 2012, 2012(1): 216-235.
[18] SI J, BARTO A G, POWELL W B, et al. Reinforcement learning in large, high-dimensional state spaces[M]. Piscataway: IEEE Press, 2004.
[19] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
[20] VAN H H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]// Proceedings of AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 2094-2100.
[21] 3GPP TSG RAN. Further advancements for E-UTRA physical layer aspects[S]. (2010-03)[ 2020-01-10].

[作者简介]

**刘婷婷**（1982− ），女，陕西西安人，博士，北京航空航天大学副教授，主要研究方向为基于机器学习和无线大数据的干扰管理、资源规划和信息预测。

**罗义南**（1995− ），男，辽宁丹东人，北京航空航天大学硕士生，主要研究方向为超密集网络中的分布式干扰协调。

**杨晨阳**（1965− ），女，浙江杭州人，博士，北京航空航天大学教授、博士生导师，主要研究方向为基于机器学习、无线大数据的缓存、传输资源管理、超可靠低延时通信等。