# 基于并行式强化学习的路径规划问题求解

答辩人：王书源

指导教师：李升波

2018年6月15日

# 目录


01 课题背景

02 单智能体强化学习

03 多智能体强化学习

04 并行式强化学习

05 仿真和结果分析

06 总结与展望

# 目录


01 课题背景

02 单智能体强化学习

03 多智能体强化学习

04 并行式强化学习

05 仿真和结果分析

06 总结与展望

# 课题背景

## 强化学习简介

- 强化学习(RL)是机器学习的重要分支
- 目的是使智能体自动、连续的决策
- 基本思想是通过最大化智能体的累积回报，以学习到最优策略

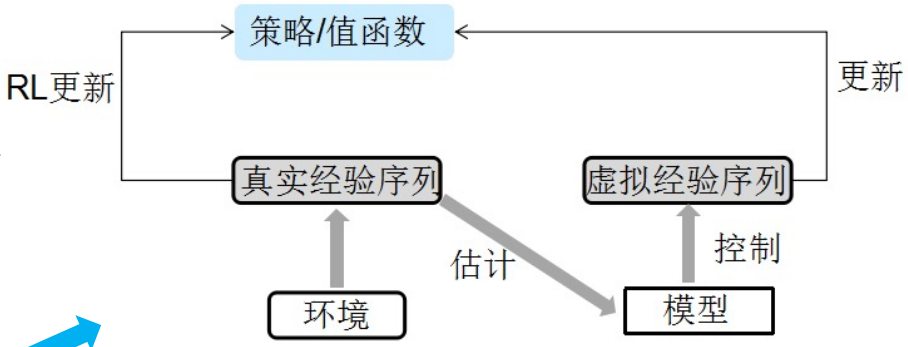
智能体探索环境

## 强化学习的特点

- 在与环境交互中学习，通过试错更新策略
- 无监督学习，不需要环境模型

# 课题背景

- **强化学习的问题**
  - 处理大规模问题时的训练时间长




Dyn-Q框架

- **解决方案**
  - 硬件加速
  - 软件加速
    - Dyn-Q框架
    - 多智能体学习

多智能体系统

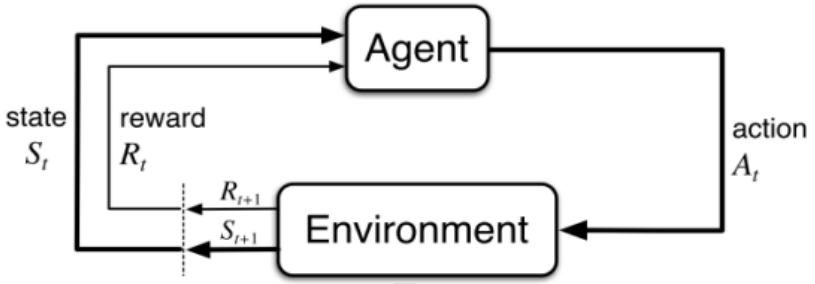
- **课题工作：使用多智能体技术**
  - 加快强化学习(Q-learning)的速度
  - 求解路径规划问题
  - 给出加速性能及影响速度的因素

01 课题背景
02 单智能体强化学习
03 多智能体强化学习
04 并行式强化学习
05 仿真和结果分析
06 总结展望

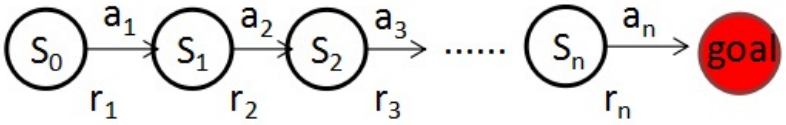
# 单智能体强化学习

## 强化学习框架

- 强化学习基于Markov架构




智能体与环境不断交互

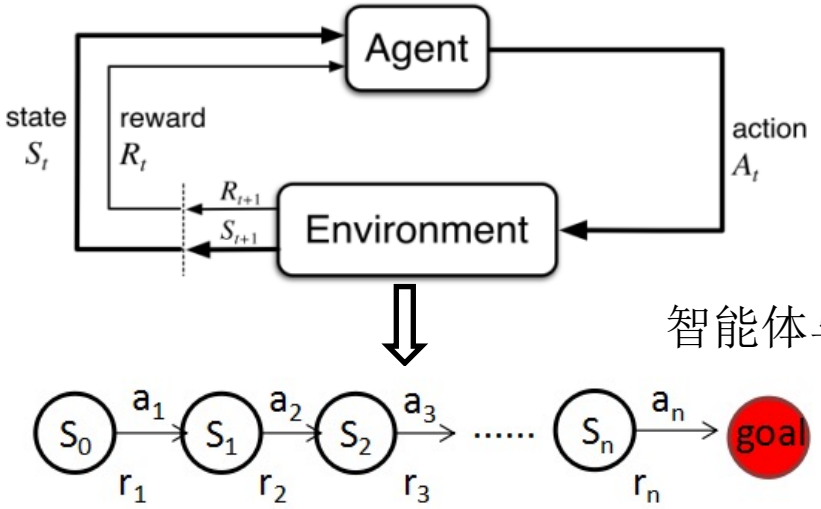



- $S_t$：$t$ 时刻智能体状态
- $\pi$：智能体的行动策略
- $A_t$：$t$ 时刻智能体动作
- $R_t$：$t$ 时刻智能体得到的回报
- $P_{ss'}^{a}$：在状态s执行动作a后下一时刻状态转移至s'的概率

# 单智能体强化学习

## 强化学习框架

- 强化学习基于Markov架构




- 智能体在$t$时刻在状态$S_t$按照策略$\pi$执行动作$A_t$，经环境作用，照 $P_{ss'}^a$的概率转移至$S_{t+1}$，得到回报$R_{t+1}$。

# 单智能体强化学习

## 强化学习框架

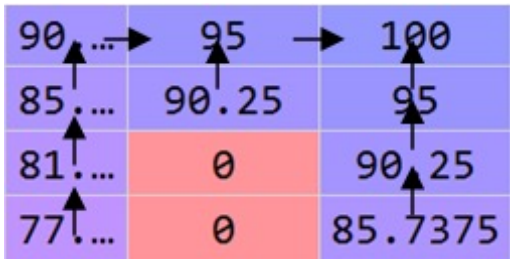
- 累积回报
  - 值函数:在位置s时依照策略π的未来折扣回报的期望值

$$V^{\pi}(s) = E_T[\sum_{i=0}^{\infty} \gamma^i r(s_i, \pi(s_i)) | s_0 = s]$$

  - 状态-动作值函数:在状态s依照策略π采取行为a的未来折扣回报的期望值

$$Q^{\pi}(s,a) = E_T[\sum_{i=0}^{\infty} \gamma^i r(s_i, \pi(s_i)) | s_0 = s, a_0 = a]$$

  – 式中Y为折扣因子，表示未来回报对当前影响




- 目标： $\pi^* = argmax_{\pi} V^{\pi}(s)$

# 单智能体强化学习

## Q-learning

- Q-learning是强化学习中的经典算法
- 单步TD方法，使用自举(boostrapping)的思想
- 离策略，可使用经验回放
- 更新公式：

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma \max_{a} Q(s_{t+1},a) - Q(s,a)\right)$$

$$V(s) = \max_{a \in actions} Q(s,a)$$

- 伪代码：

Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode):
    Initialize $s$
    Repeat (for each step of episode):
        Choose $a$ from $s$ using policy derived from $Q$(e.g.,$\epsilon$-$greedy$)
        Take action $a$ observe $r, s'$
        $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$;
        $s \leftarrow s'$;
    Until $s$ is terminal

01 课题背景
02 单智能体强化学习
03 多智能体强化学习
04 并行式强化学习
05 仿真和结果分析
06 总结与展望

# 多智能体强化学习

## 多智能体系统的优点

- 可并行处理问题
- 鲁棒性强

## 多智能体强化学习分类

- 协同工作式：单个智能体的回报及状态转移会受到其他智能体的动作的影响
  - 典型算法：Distributed-Q,Team-Q
- 非协同工作式：各个智能体独立的完成任务，一般用于加速
  - 按交互信息分类：
    - 交互行动经验 $< s_t, a_t, s_{t+1}, a_{t+1}..... >$
    - 交互值函数表格 $V(s), Q(s,a)$
    - 交互神经网络神经元权重θ

集中式拓扑

分布式拓扑

# 多智能体强化学习

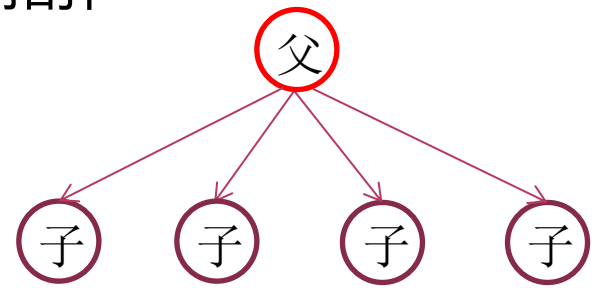
## 非协同工作式多智能体学习

| 文献 | 交互信息种类 | 拓扑 | 特点 |
| --- | --- | --- | --- |
| Tan.M,1993 | 行动经验 | —— | 仅对2个智能体进行了研究，人为指定信息流 |
| Tan.M,1993 | 值函数表 | —— | 不具备筛选功能，仅对值函数取平均，加速性不理想 |
| Alicia,2004 | 值函数表 | 集中式 | 每个智能体预先被分配好待学习的数据域 |
| Z.Abbasi,2008 | 值函数表 | 集中式 | 引入专家机制，让智能体对其他智能体的值函数取加权平均，但加速性结论仅来自单张地图 |
| Mnih,2016 | 深度网络权重向量θ | 集中式 | 多智能体使on-policy的深度强化学习变稳定 |

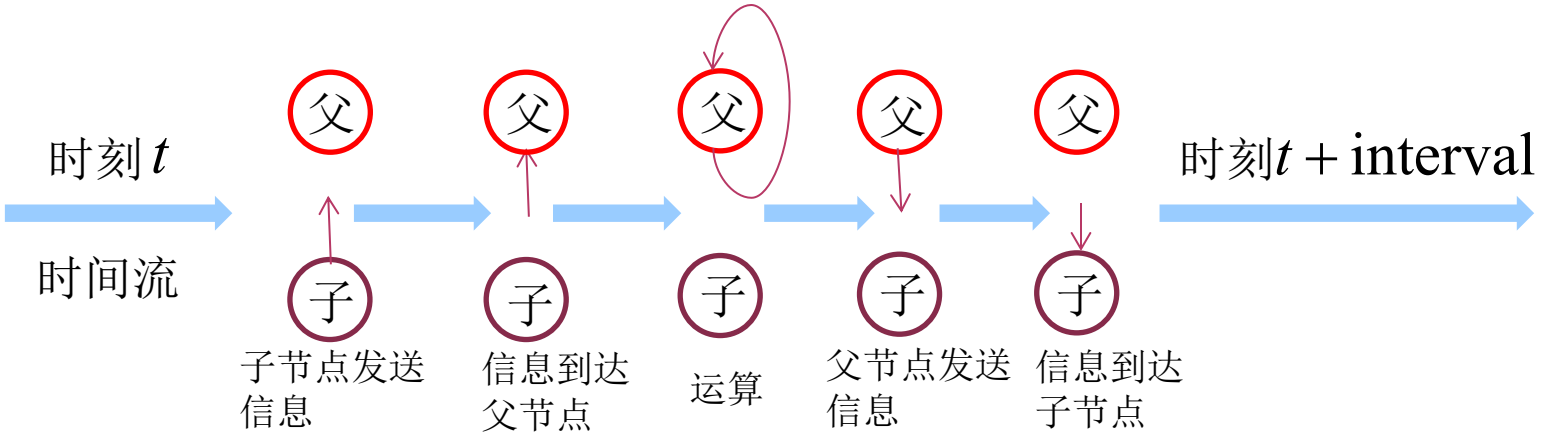
01 课题背景
02 单智能体强化学习
03 多智能体强化学习
04 并行式强化学习
05 仿真和结果分析
06 总结与展望

## 通信拓扑介绍

- 集中式通信拓扑




- 其中红色的表示父节点(master)，紫色的表示子节点(slave)。
- 本研究中的集中式拓扑存在通讯间隔(interval)，由人为设定。



# 并行式强化学习

## 算法介绍

- 分布式Q-Learning中每个智能体仅在Q值增大时才更新Q，我们沿用了这个思想：

$$Q_{master}(s,a)=\max_{slavei}\{Q_{slavei}(s,a)\}\quad (i=1,2,3....)$$

假设：回报非负

- 回顾单智能体学习：

$$Q(s,a)\leftarrow Q(s,a)+\alpha(r+\gamma\max_{a}Q(s_{t+1},a)-Q(s,a))$$

$$V(s)=\max_{a\in actions}Q(s,a)$$

- 多智能体学习：

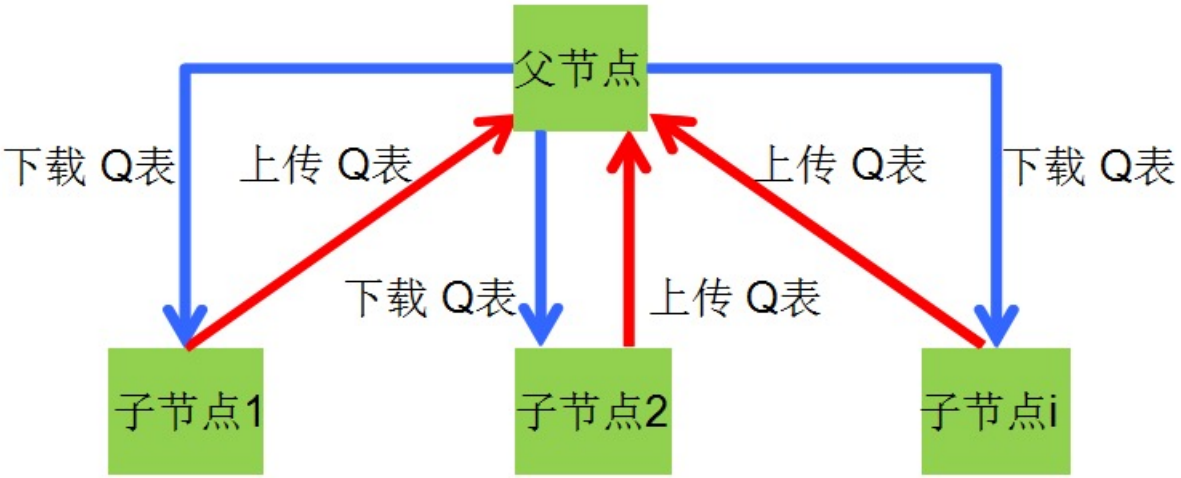
多智能体协同时，仅更新变大的Q值

$$Q(s,a)\leftarrow Q(s,a)+\alpha(r+\gamma\max_{a}Q(s_{t+1},a)-Q(s,a))$$

$$V(s)=\max_{a\in actions}Q(s,a)$$

## 算法介绍

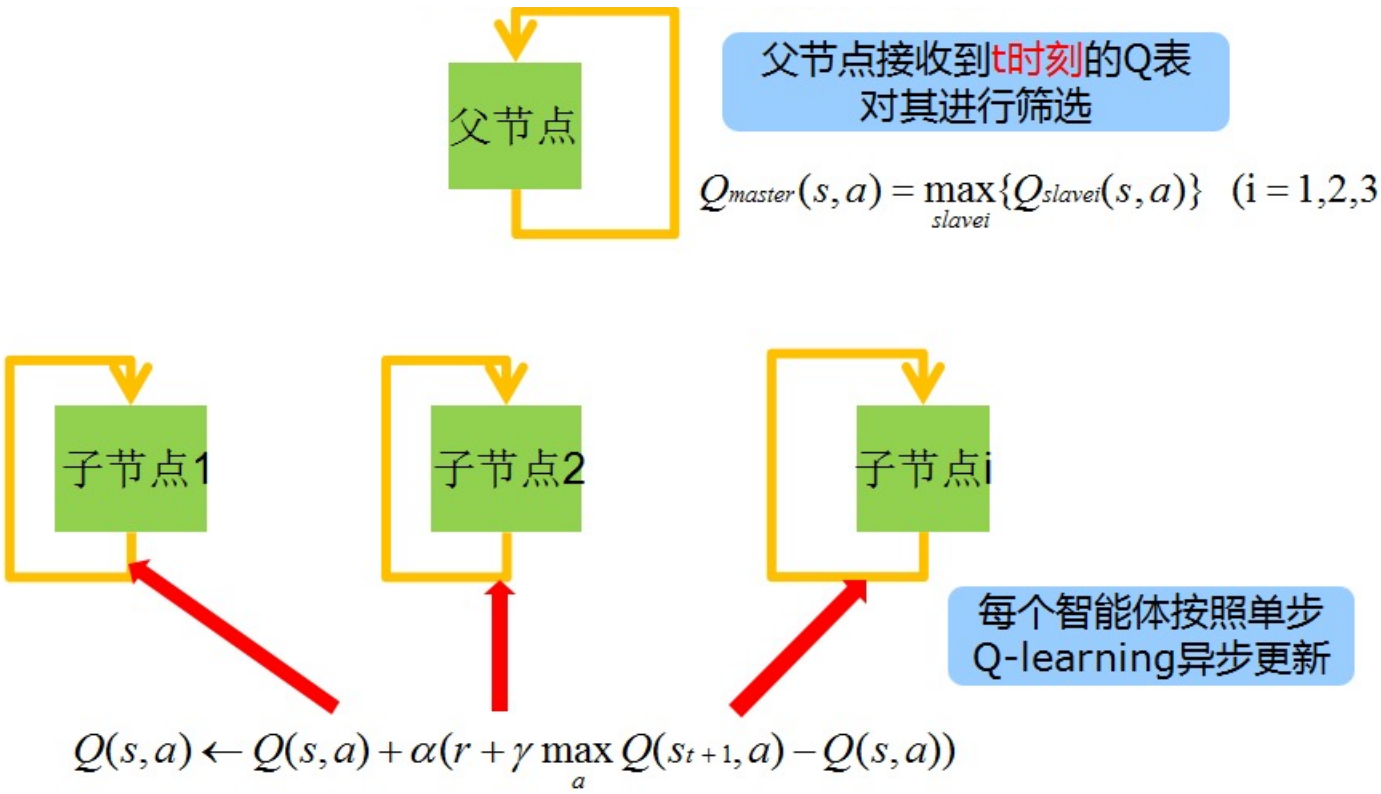
- 时刻$t$：子节点接收到父节点发送的Q表，同时将自己的Q表上传给父节点



# 并行式强化学习

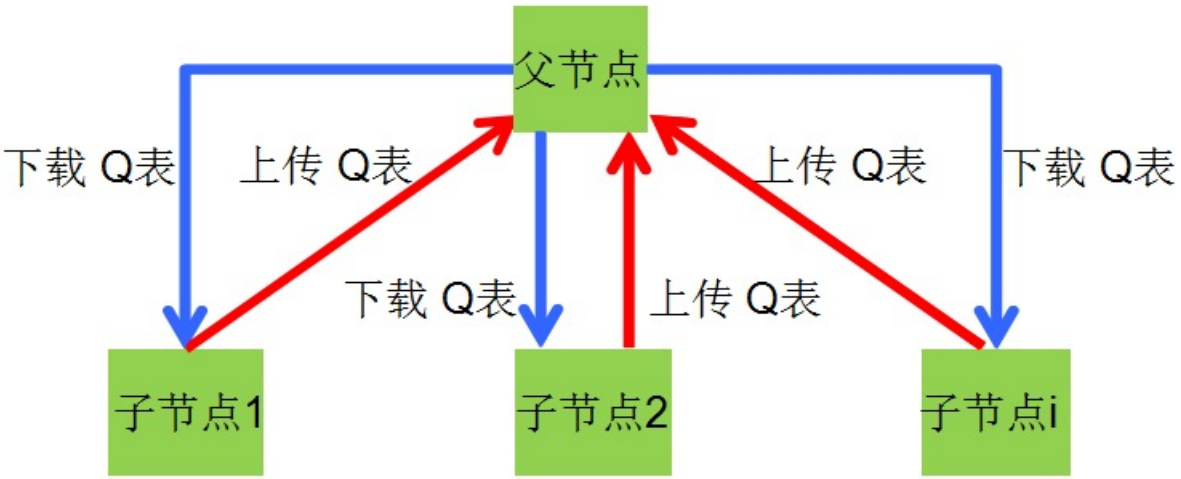
## 算法介绍

- 时刻*t+n(n<interval)*：父节点对t时刻子节点上传的Q表进行合并
  子节点对t时刻下载的表进行更新



## 算法介绍

- 时刻*t+interval*：重复时刻t的操作



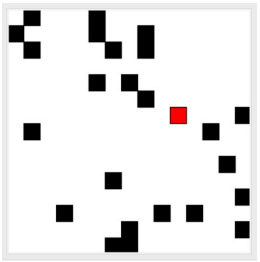
01 课题背景
02 单智能体强化学习
03 多智能体强化学习
04 并行式强化学习
05 仿真和结果分析
06 总结与展望

# 仿真及结果分析

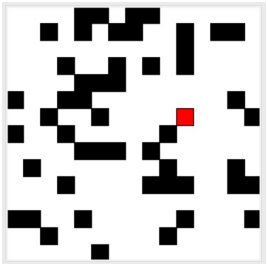
## 问题背景

- 路径规划问题
- 场景初始化

| 障碍物(block)数量 | 有效状态数 | 地图数量 | 生成方式 |
|---|---|---|---|
| 25 | 200 | 25 | 随机 |
| 50 | 175 | 25 | 随机 |
| 75 | 150 | 25 | 随机 |



block=25



block=50

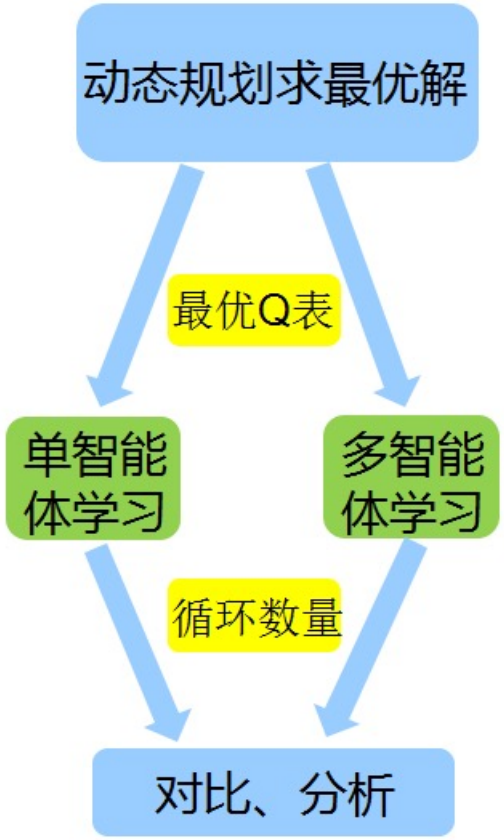
block=75

- 目标：$\pi^* = argmax_\pi V^\pi(s)$ 即在任意初始状态s，都能找到到达终点最短的路径

# 仿真及结果分析

## 问题求解

- 参数设置
  - 折扣因子 $\gamma = 0.95$
  - 学习率 $\alpha = 0.1$
  - 探索率 $\varepsilon = 0.1$
  - 值函数$V(s)$表规格：
    $(15 \times 15 - block) = 225 - block$
  - 动作状态值函数 $Q(s,a)$ 表规格：
    $(15 \times 15 - block) \times 4 = 900 - 4 \times block$
- 终止条件
  - 所有可行状态都学习到了全部的最优策略



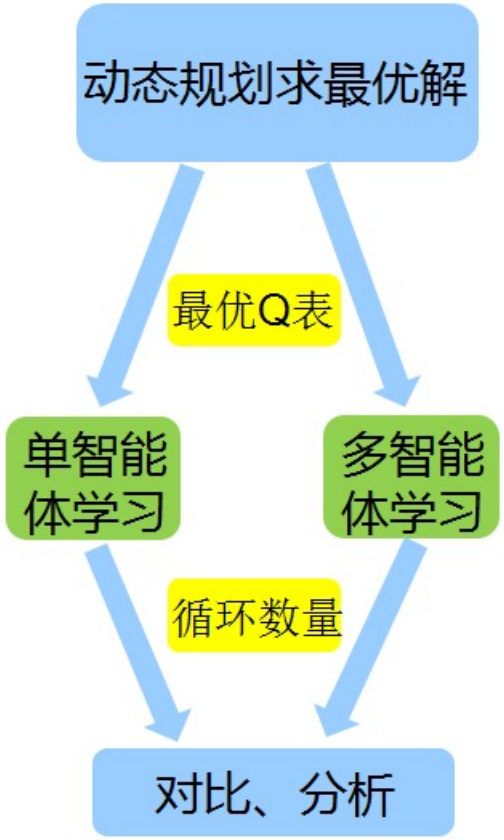
# 仿真及结果分析

## 问题求解

- 多智能体参数设置
  - 通讯间隔：2000，4000，6000，8000，10000，12000（单位：步）

| 10000 | 大间隔 |
| --- | --- |
| 6000 | 中等间隔 |
| 2000 | 小间隔 |

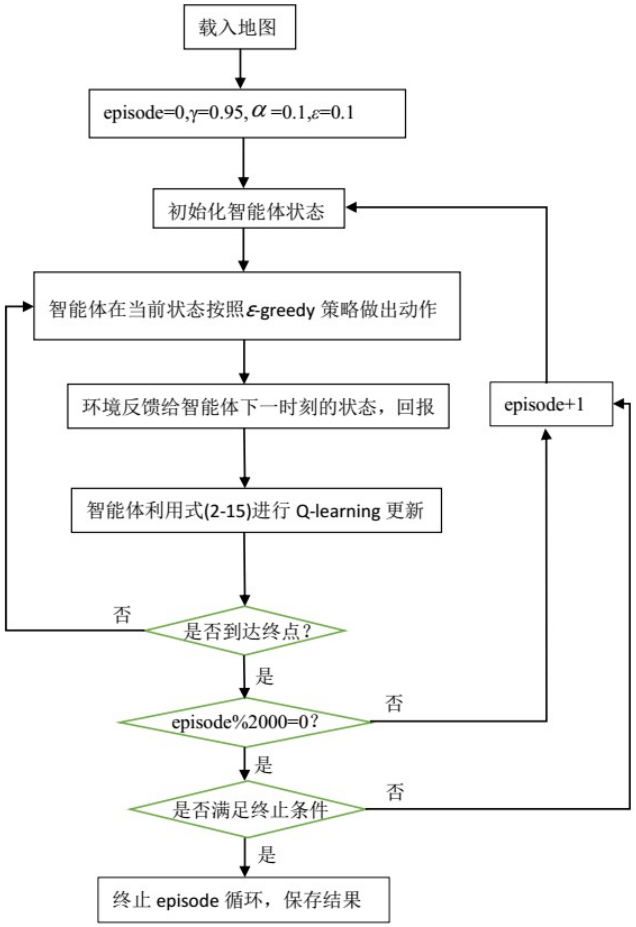
  - 多智能体数目：[2:10]



# 仿真及结果分析

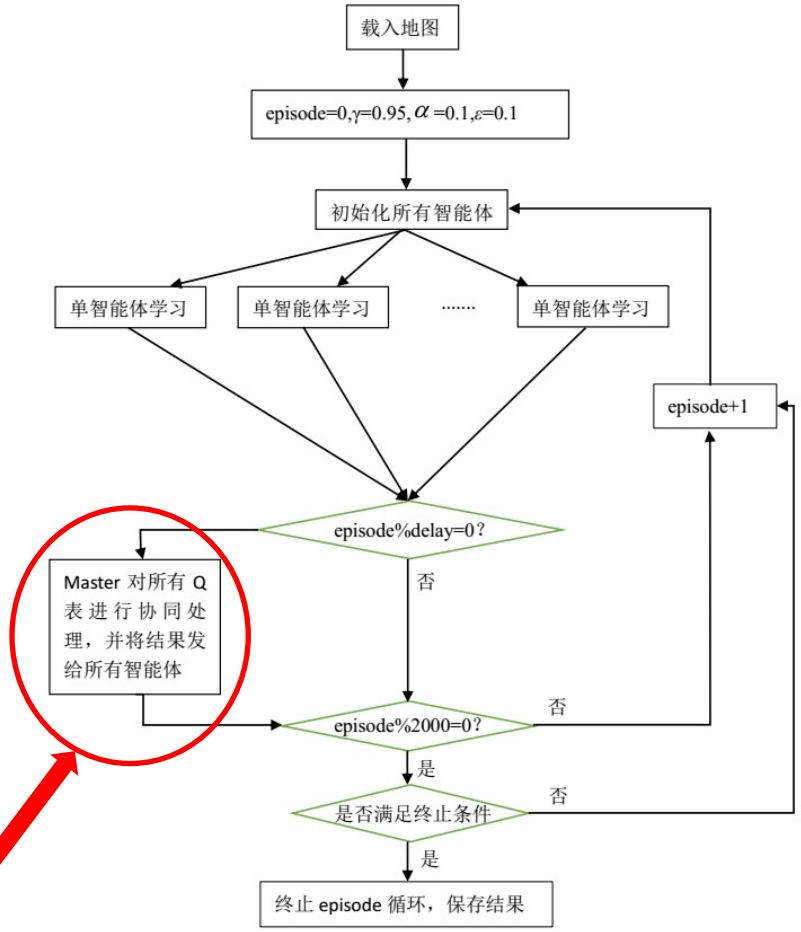
## 问题求解

- 单智能体学习




## 问题求解

- 多智能体学习



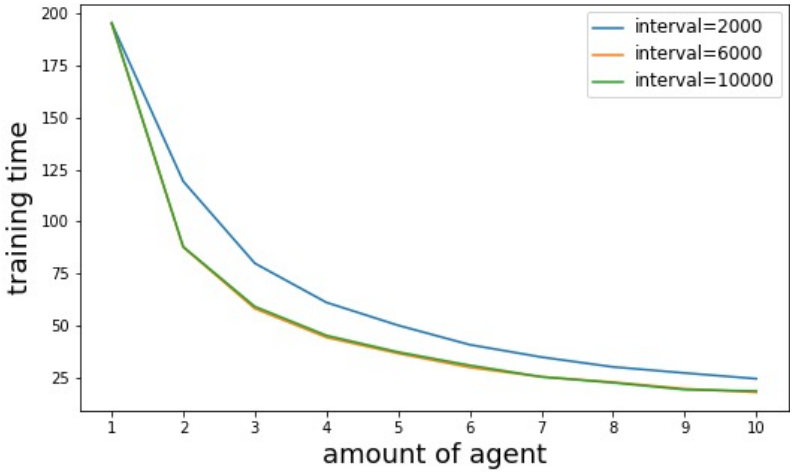

$$Q_{master}(s,a)=\max_{slavei}\{Q_{slavei}(s,a)\}\quad(\mathrm{i}=1,2,3....)$$

# 仿真及结果分析

## 多智能体学习时间与智能体数量关系图

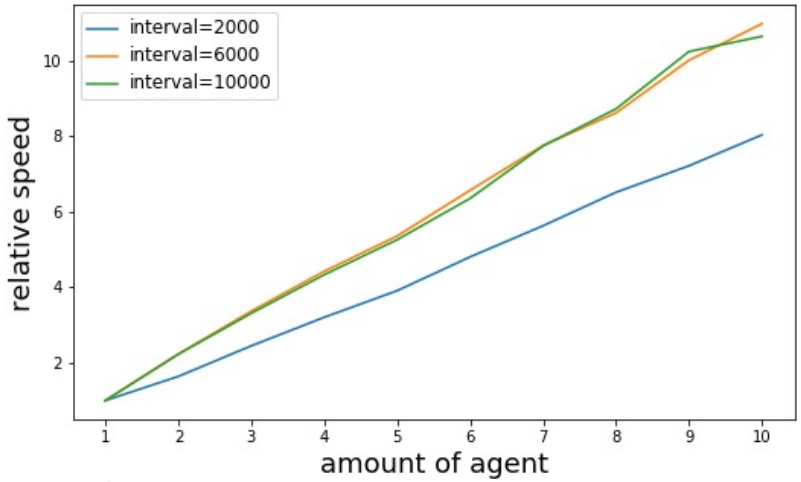
block=50，通讯间隔为小、中、大

- 学习时间与智能体数量近似呈反比规律下降
- 小通讯间隔时学习时间明显长于中、大通讯间隔




## 多智能体学习速度与智能体数量关系图

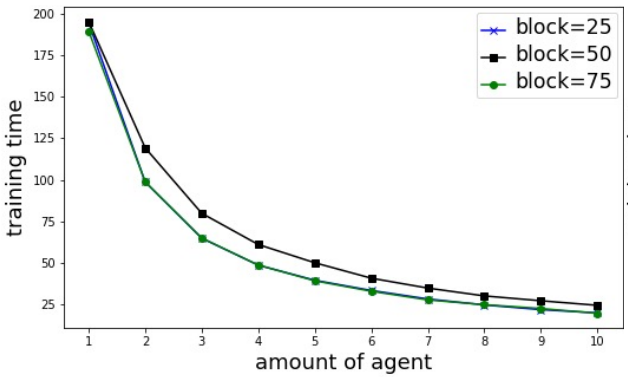
- 多智能体学习速度与智能体数量近似呈线性增长
- 对于中、大通讯间隔，学习速度增长的更快，增长率大于1

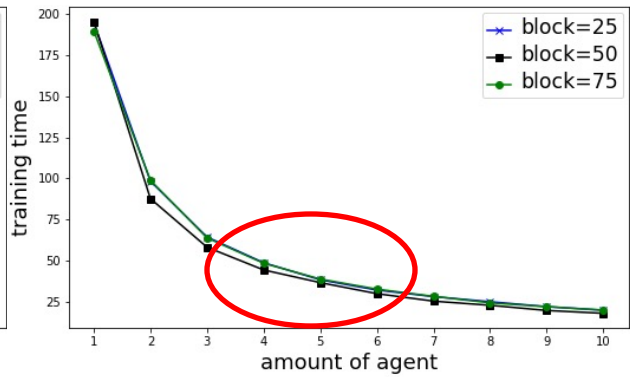


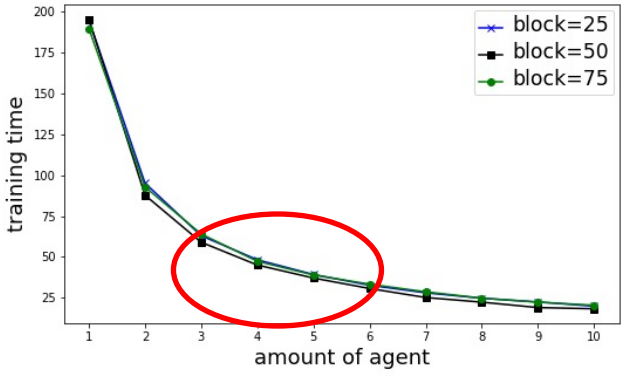
# 仿真及结果分析

## 不同障碍物数目下学习时间与智能体数量关系图

从左至右通讯间隔分别为小间隔、中间隔、大间隔

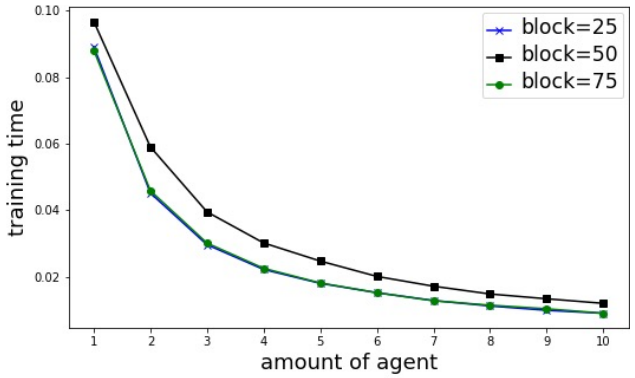


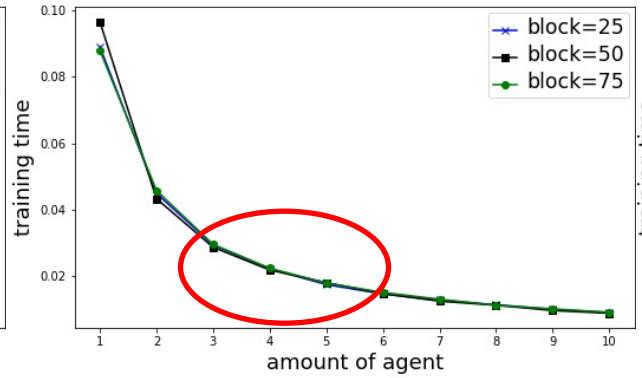



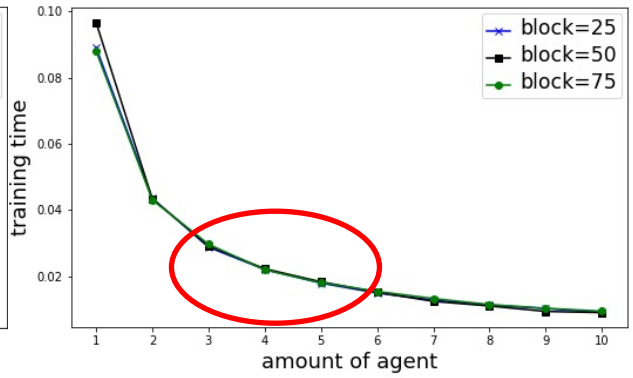




## 平均每个状态每个最优步学习时间与智能体数量关系图

从左至右通讯间隔分别为小间隔、中间隔、大间隔



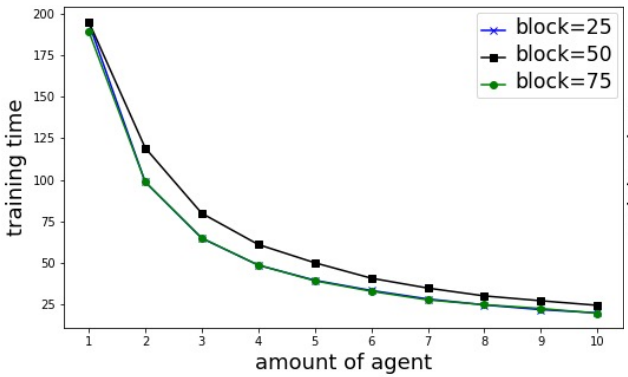



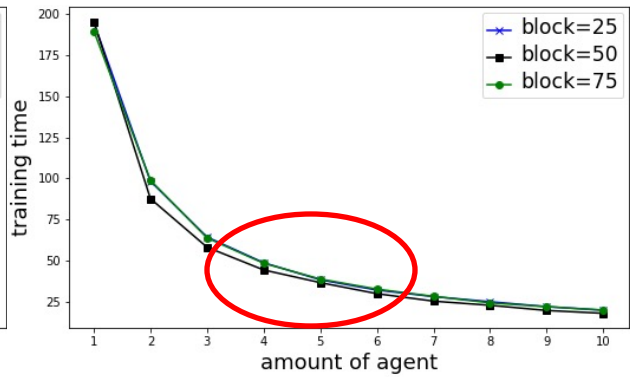



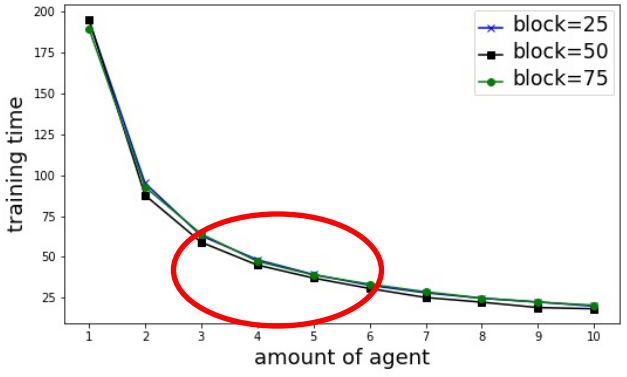
# 仿真及结果分析

## 不同障碍物数目下学习时间与智能体数量关系图

地图不同，可行状体数不同，平均每个状态到达终点的距离不同

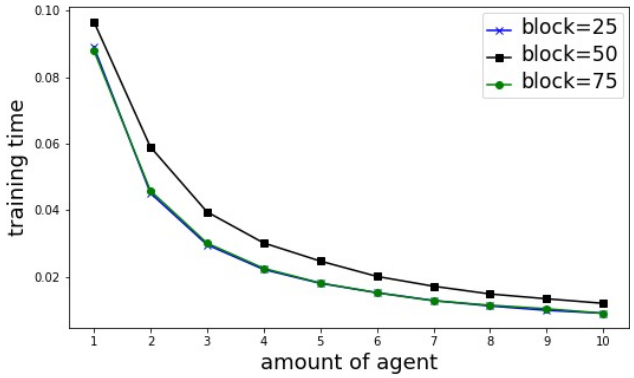


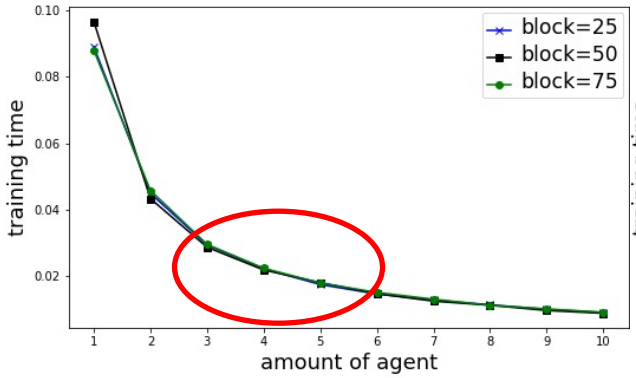



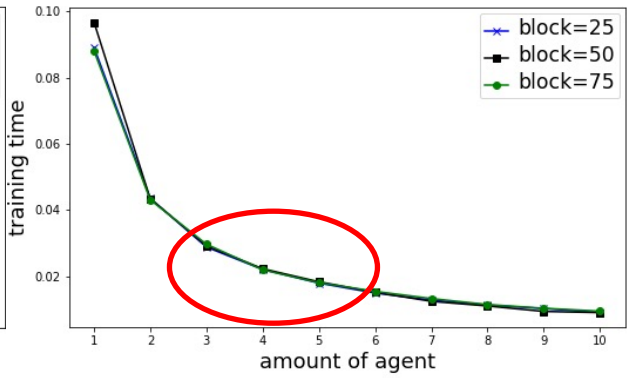




## 平均每个状态每个最优步学习时间与智能体数量关系图

对于中、大间隔，处理后的结果与地图种类无关



01 课题背景
02 单智能体强化学习
03 多智能体强化学习
04 并行式强化学习
05 仿真和结果分析
06 总结与展望

# 总结与展望

## 工作总结

- 调研了多智能体强化学习并改进了单智能体Q-learning算法
- 搭建了多智能体仿真平台
- 分析仿真结果，得到结论：
  - 算法的学习速度随智能体数量呈斜率为1的速度增长
  - 较大通讯间隔时多智能体学习速度会更快
  - 不同障碍物数目下，平均每个状态每个最优步的学习时间是相等的
  - 基于上一点结论，可知当地图变化时，影响速度的2大核心因素为地图中可行状态数及到达终点的平均步数。因此实际应用中若已知地图模式则可以估算学习时间

## 展望

- 本课题采用集中式通信拓扑，未来可考虑研究分布式拓扑的多智能体学习

# 谢谢各位，
# 敬请指正！

www.idlab-tsinghua.com