一、问题：

数量众多的智能体（训练组，A），通过某种机制或者学习方案，形成针对于另一群智能体（对抗组，O）的协作攻防的策略。目标，训练组在多次博弈中，获得更多的胜利，且每次胜利步数尽量短。

Key words：

Cooperation：59 paper（in AAMAS）

Competition：16 paper（in AAMAS）

Many-agent or Real Robot

二、问题分析

多人随机博弈，Multi-agent Stochastic games。

（n，S，A1，…，An, T，R1，…, Rn），n为局中人（agent）数量，Ai为局中人i的行为集，T：S x A1 x … An x S ->[0,1]，状态转移函数。Ri：S x A1 x … An x S ->R，局中人i的奖励，Ri（s, a1,…, an, s’）局中人i在状态s，且群体采取（a1,…, an）转移到s’状态所获的奖励。

多人博弈转换为二人博弈。

A为训练组动作集，O为对抗组动作集，a、o均为动作向量，则博弈变为（2，S，A，O，Ra，Ro），对于传统的矩阵博弈|S|=1，不会发生状态转移，传统的强化学习|O|=1，此时对手是环境，环境只有唯一动作，意味着环境是确定性的。

状态的价值估计。

状态s=（la，lo），la为训练组的局中人位置，lo为对抗组局中人位置。

状态势函数V（s），即状态价值函数，表示训练组处于s状态下的优势程度，V（s）越大，优势越大。当准确估计出V（s）后，则每次训练组采取的动作可以进入状态的动作。

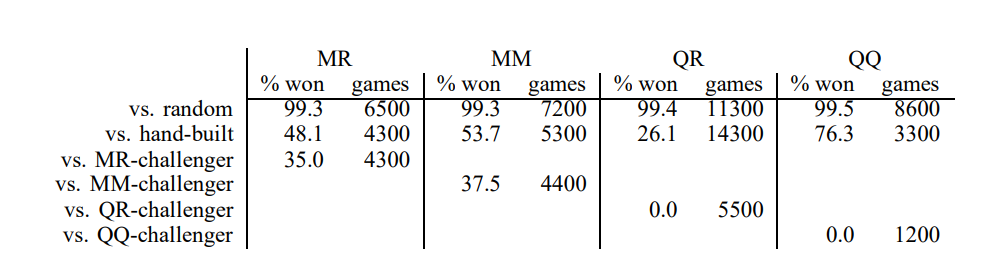
V（s）的估计和O组的动作有关。

纳什均衡在马尔可夫博弈中的估计，极大极小法。假设对手非常聪明，我方应保证在最坏情况下依旧获得最大的收益。

这是非常保守的策略，于是可以引入乐观因子，越乐观，倾向于V（s）估计越大。

此时问题完全套用强化学习的方案来求解，估算Q（s，a, o）。

关于对抗组的选择，已有标准。



未竟事宜，|S|巨大，|A|巨大。

对|S|的合并方案，在状态转移图上合并，或者再得到一个函数f，输入是整个棋盘信息，输出是状态。该问题的状态转移具有局部性。

二、实验场景

动态博弈（Sequential games，dynamic games），In game theory, a sequential game is a game where one player chooses their action before the others choose theirs.

静态博弈（simultaneous games，static）设计。In game theory, a simultaneous game is a game where each player chooses his action without knowledge of the actions chosen by other players.

<http://www-bcf.usc.edu/~brocas/Research/simseq.pdf>

（一）抓捕游戏，A组多agent围捕O组单agent。

（二）对抗游戏，A组多agent同O组多agent，按照如下规则进行博弈：

1）移动，每个agent有上下左右停五个动作；

2）冲突，t回合采取动作后，使得t+1回合同一方格中将会有两个以上的agent；或者若t回合采取动作后，使得t+1回合，两个agent互换了位置。

3）妥协，同组的agent发生第一类冲突，随机选择一个agent进入方格，其他agent采取动作停；发生第二类冲突，双方均采取动作停。

4）战斗，不同组的agent发生冲突，则冲突的agent处于战斗状态，统计周边参加此次 冲突的agent，A组为n，O组为m。A组战斗获胜概率为n/(n+m)，O组获胜概率为m/（n+m）。获胜一方agent留在棋盘，失败一方移走。再根据规则3）处理下一回合agent的位置。

5）胜利，规定回合内，棋盘上不存在O组agent则，A组获胜。棋盘上不存在A组agent则，O组获胜。超出规定回合，平局。

简化的对抗游戏，1）移动，每个agent有上下左右停五个动作；2）每回合，同组的agent依次采取动作，不允许重叠；3）战斗，当不同组的agent处于相邻的方格，则自动发生战斗，统计周边参加此次冲突的agent，A组为n，O组为m。A组战斗获胜概率为n/(n+m)，O组获胜概率为m/（n+m）。获胜一方agent留在棋盘，失败一方移走。再根据规则3）处理下一回合agent的位置。（仍旧保证了完全信息和静态博弈的特征）

重新形式化定义场景。先形式化定义这个问题，多人完全信息静态博弈

（三）多人球赛

四、baseline的问题

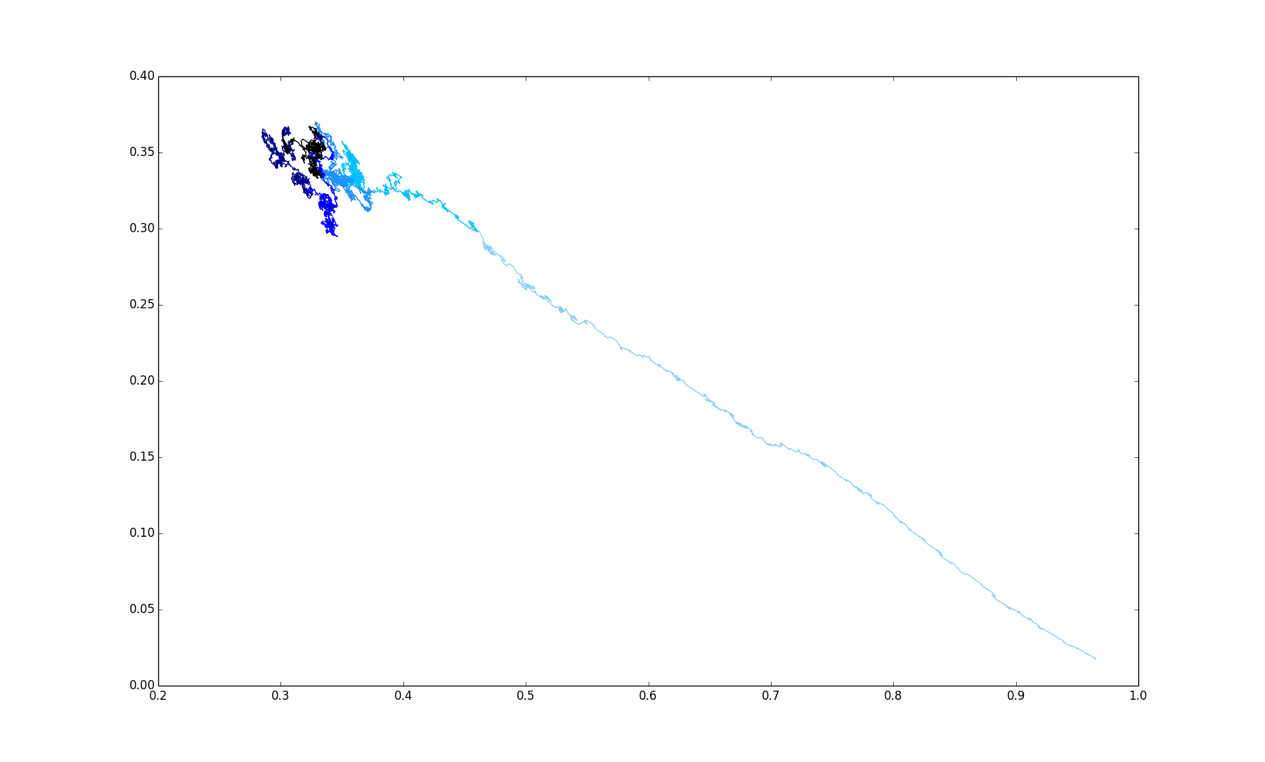
有三处可供参考的baseline，openai的MADDPG论文中给出了合作竞争实验，openai的roboschool，双人博弈实验，Magent平台的实验。

可以开发一套棋盘博弈的游戏库，作为完全信息、不完全信息、静态博弈、动态博弈问题算法的baseline。

2018/10/15

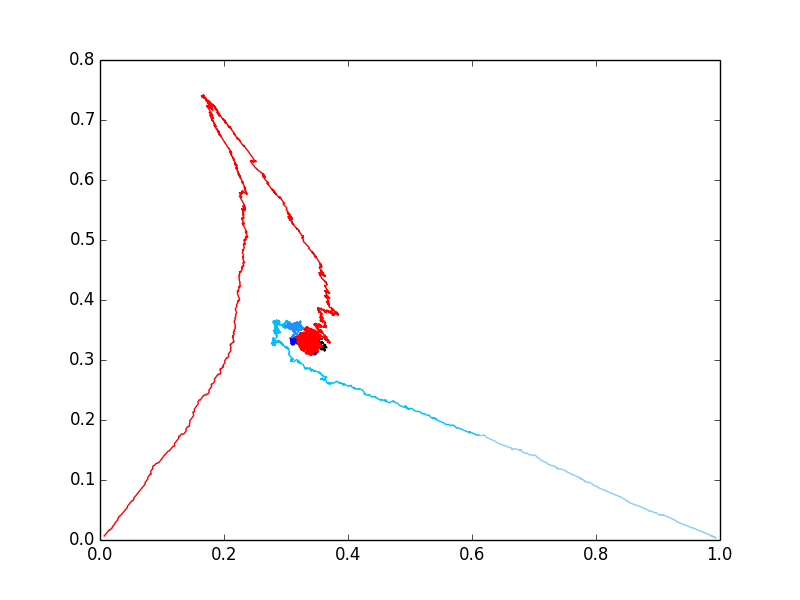
nash均衡实验

实验图片1，figure1。剪刀石头布实验，对手测量，剪刀石头布等概率。

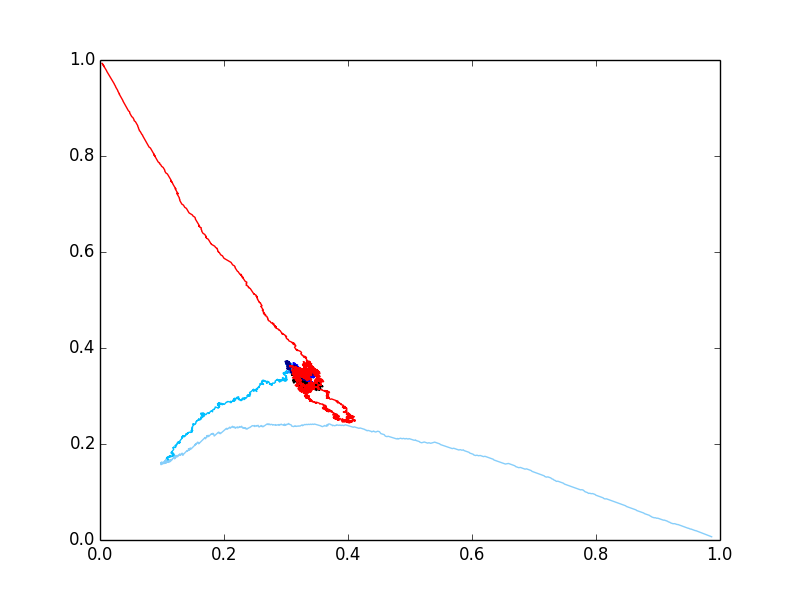


只有q-learning，数据输出直接对Q值做softmax，初始状态q(1,j)=1，横坐标剪刀的概率，纵坐标石头的概率。

实验图片2，两个人均用q-learning，蓝色初始，一直出剪刀，红色初始，一直出布。



实验图片3，两人均用q-learning，蓝色初始，一直剪刀，红色初始，一直石头。



实验图片4，两人均用q-learning，蓝色初始，一直剪刀，红色初始，一直剪刀。

