# An analysis of stochastic game theory for multiagent reinforcement learning

混合策略：用户依照概率来选择行为

Matrix game可以分为竞争和协作，纯竞争为零和博弈，其他的也称为general-sum game 。零和博弈只有一个纳什均衡，用线性规划来解。求general sum的均衡需要解二次规划。

Function value（MG）为博弈的均衡所对应的值，Solve为player的均衡策略。

Stochastic game 定义：最终期望找到稳态的统计策略 

【18,6】证明，零和博弈以及general-sum 博弈均有均衡。

假设stochastic game有唯一的均衡

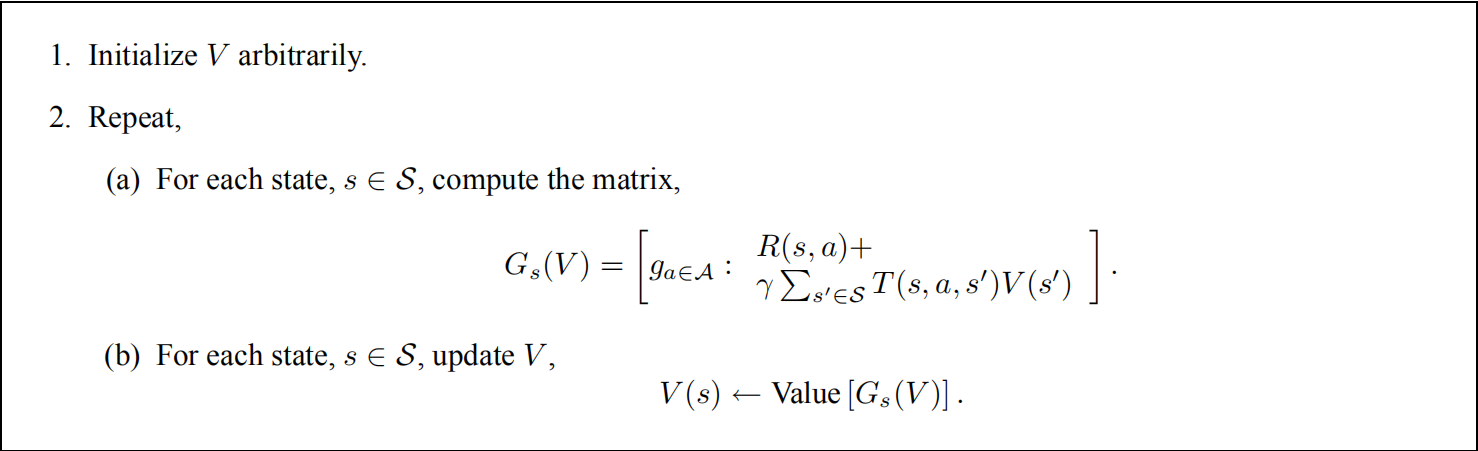
解stochastic game的方法在SG以及学习过程的假设上有所不同，实际上为博弈论与强化学习算法的不同。博弈论要求模型已知，目的在于计算均衡值，而不是找均衡策略。强化学习假设世界未知。

## Solutions from game theory

研究不同状态的值函数V(s)，

### Shaply

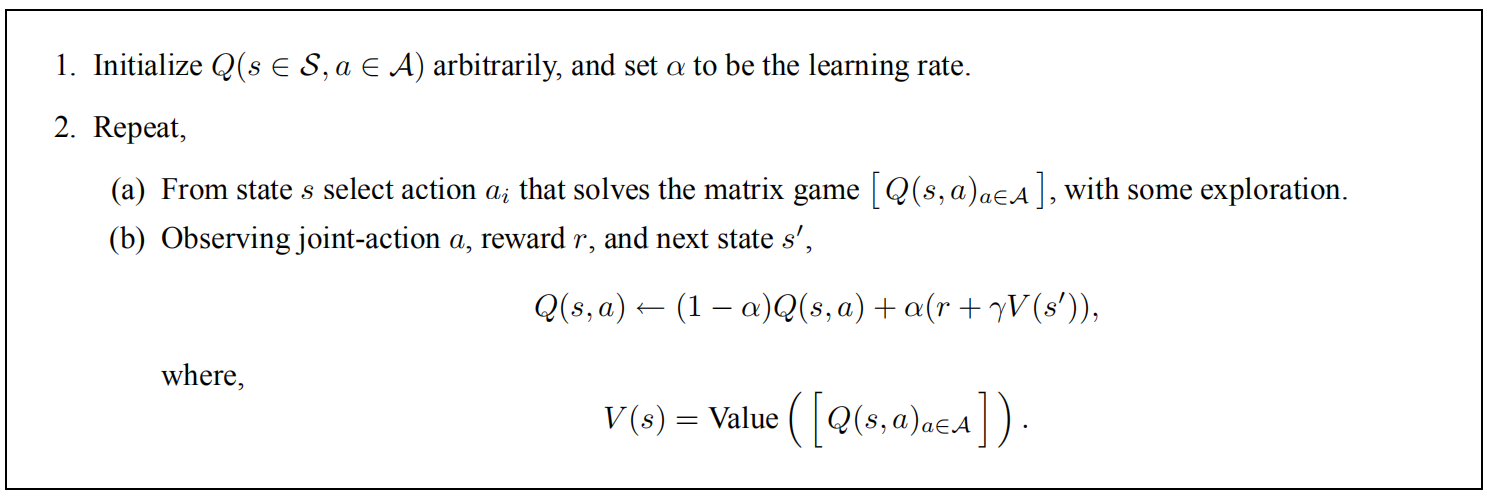
Shaply在【18】中证明零和Stochastic Game中有均衡



## Solutions from RL

### Minimax-Q

Litmma【13】将传统的MDP中的Q-learning 扩展到零和stochastic game中，实际上是Shaply的“value iteration”的off-policy 强化学习等效

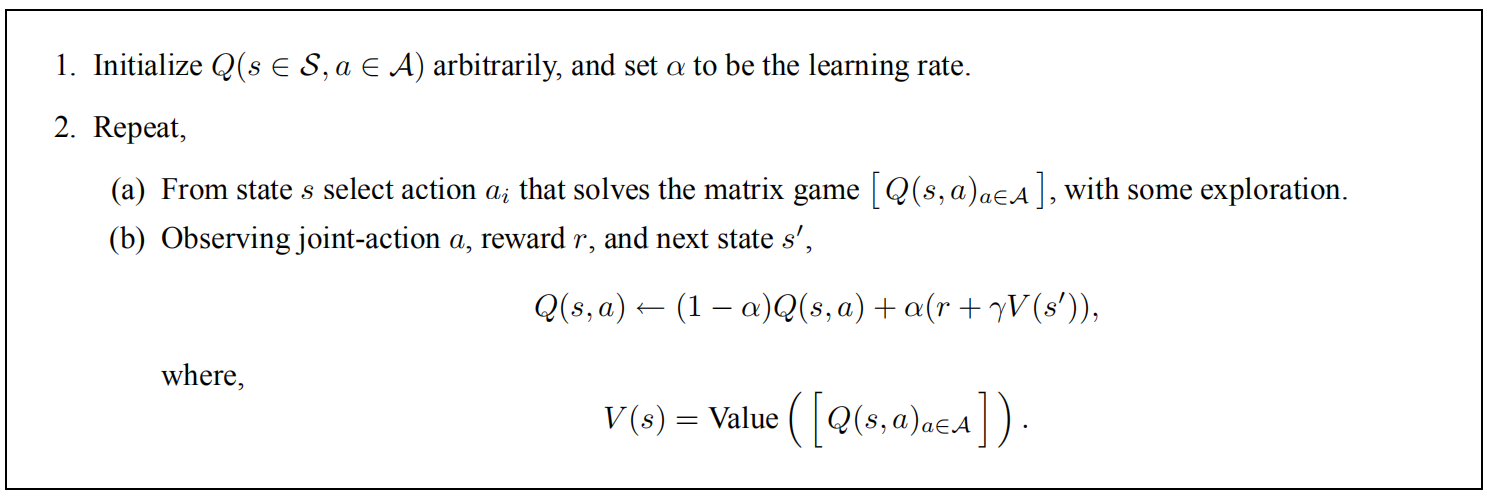


Stochastic game的解决方法主要分为两部分：Matrix Game Solver以及Temporal Differencing

化为两类：博弈论的（假设模型已知）、强化学习的（模型未知的）

### Nash Q

Hu在【10】中将minimax-Q 扩展到general-sum 博弈，同样由下面表格给出



特点：

1. 要求每个agent获得其他agent的Q值，零和博弈中的线性规划替代为二次规划
2. 算法需要一些限制性假设，其中最强的假设在于对中间过程中遇到的matrix game的结构做了假设；其他的假设比如该博弈只有唯一的均衡，对于general-sum stochastic game 不一定成立，该假设的目的在于获取算法的收敛鲁棒性，也就是不论其他人何种决策，自己一定收敛到均衡。

## Observation

1. On-policy 需要一个agent 执行它所学习的策略
2. 收敛性证明比较难：收敛性证明中，大部分都是假设零和博弈，仅有的其他算法Nash-Q尽管不假设零和，仍旧需要假设博弈有唯一的均衡以及对payoff的假设。
3. 最终的发现在于所有的算法对于矩阵博弈都采用闭式的解，前提假设是排除了多均衡的情况，如果一个博弈有多个均衡，最佳的策略必须依赖于其他agent的策略，此时静态的对手独立的算法如（线性规划、二次规划）都不能正常工作。

## Other solution

对于matrix game solving部分，采用学习的方法来解而不用闭式解。原因在于第一，如果一旦有多个均衡，那么需要观测其他用户的行为。实际上，stochastic game当中，均衡解的数目可能以指数增长。第二，当其他agent有限制约束，比如physical或者rational,因此采用opponent-dependent matrix game solver

下面举例opponent-dependent matrix game solver

### Fictitious play

【16,25】

### Opponent Modeling

【22,5】

## 未来方向

1. general-sum game with many equilibria尚无算法解
2. 当其他agent有明显的物理性以及合理性限制时

# Multiagent Reinforcement Learning: Theoretical framework and an algorithm

将littman零和推广到general sum stochastic game ，仍旧为2 player，经过证明，若只有一个nash 均衡则会收敛到该均衡。

# Asymmetric multiagent reinforcement learning

之前的Markov game都是假设为对称学习结构，这篇文章介绍非对称学习结构。Leader 能帮助followers去选择最佳的行为，这种架构在半中心化的Multiagent system中出现。

**当前的Multiagent RL分为三类【6,24,14,13,5,26】**

1. 对值函数的直接梯度
2. 估计值函数计算均衡‘
3. 直接策略梯度

第一个多状态Markov博弈学习的方法由littman提出【19】，两个用户零和博弈

【11】中，Hu Wellman将零和博弈推广到general sum Markov game，但是必须在非常限制性的条件下，仍旧是两个用户

【20】中，littman提出新方法，增加了一些agents角色信息并放松了限制

**【30】中，wang 提出一个方法使得Team Markov收敛到最佳纳什均衡**

【7】中，Sandholm 提出一个算法收敛到纳什均衡，并且学会与稳态的对手博弈

对于计算finite games的纳什均衡仍旧是难点，为了克服该问题

Greenwald Hall提出multiagent RL用相关均衡代替纳什均衡【10】

**Multiagent reinforcement learning中一个完全不同的方法是Collective Intelligence结构【34】，该方法可扩展并且对于大规模问题也很鲁棒，应用于【35,33】**