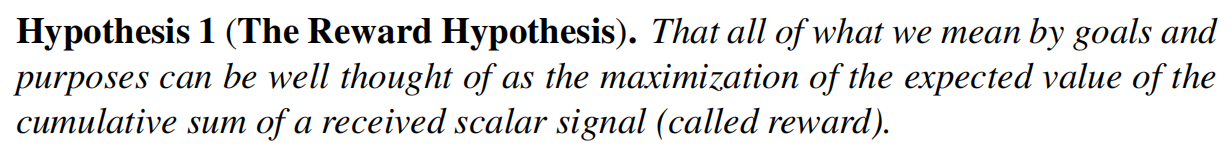
Multi-Agent Reinforcement Learning:

A Report on Challenges and Approaches

**Introduction：**

Reinforcement Learning (RL) refers to both the learning problem and sub-field of

machine learning.假说，所有的目标和假设都是为了最大化期望奖赏。

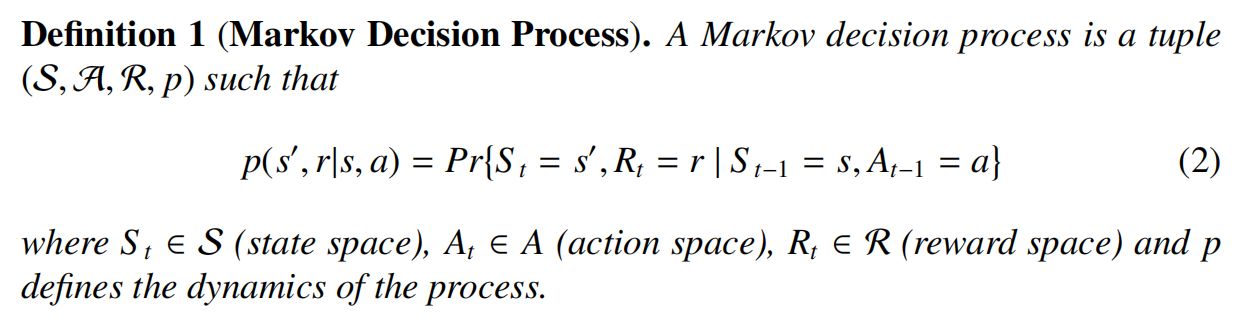


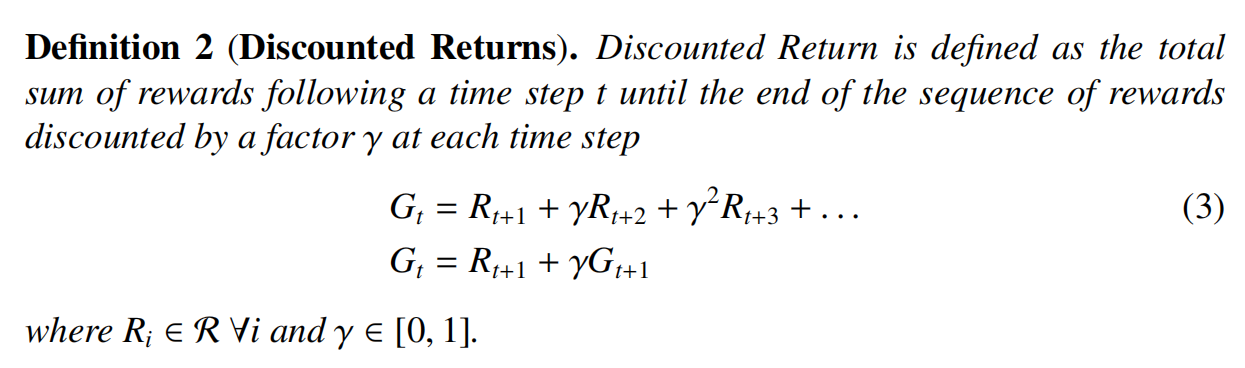
EC is different from RL in that there is no explicit interaction between the environment and the agent. 而许多当今社会很复杂的问题都可以抽象成多智能体学习问题。

**RL Methods：**

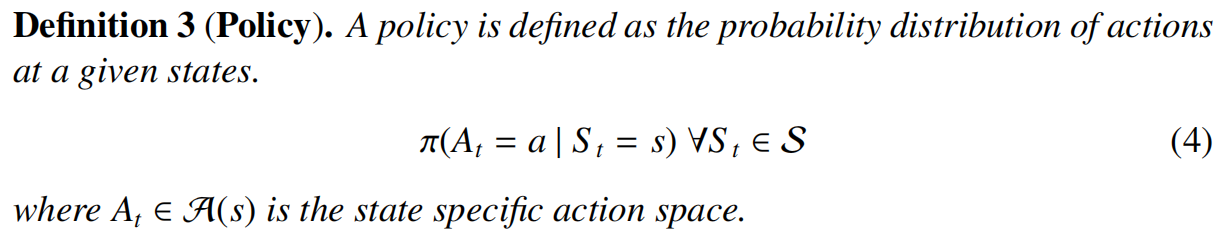
**Basic:**

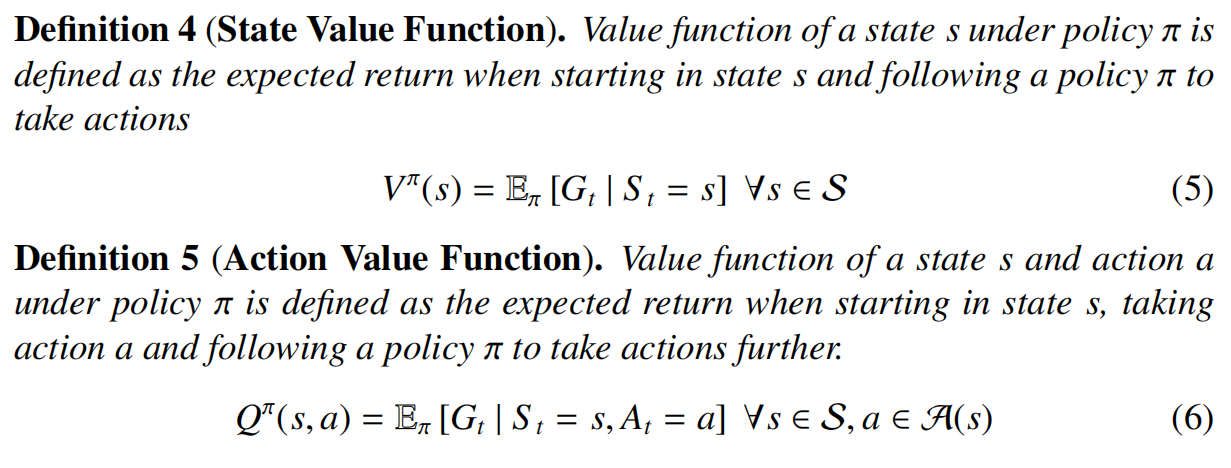
1. **MDP**



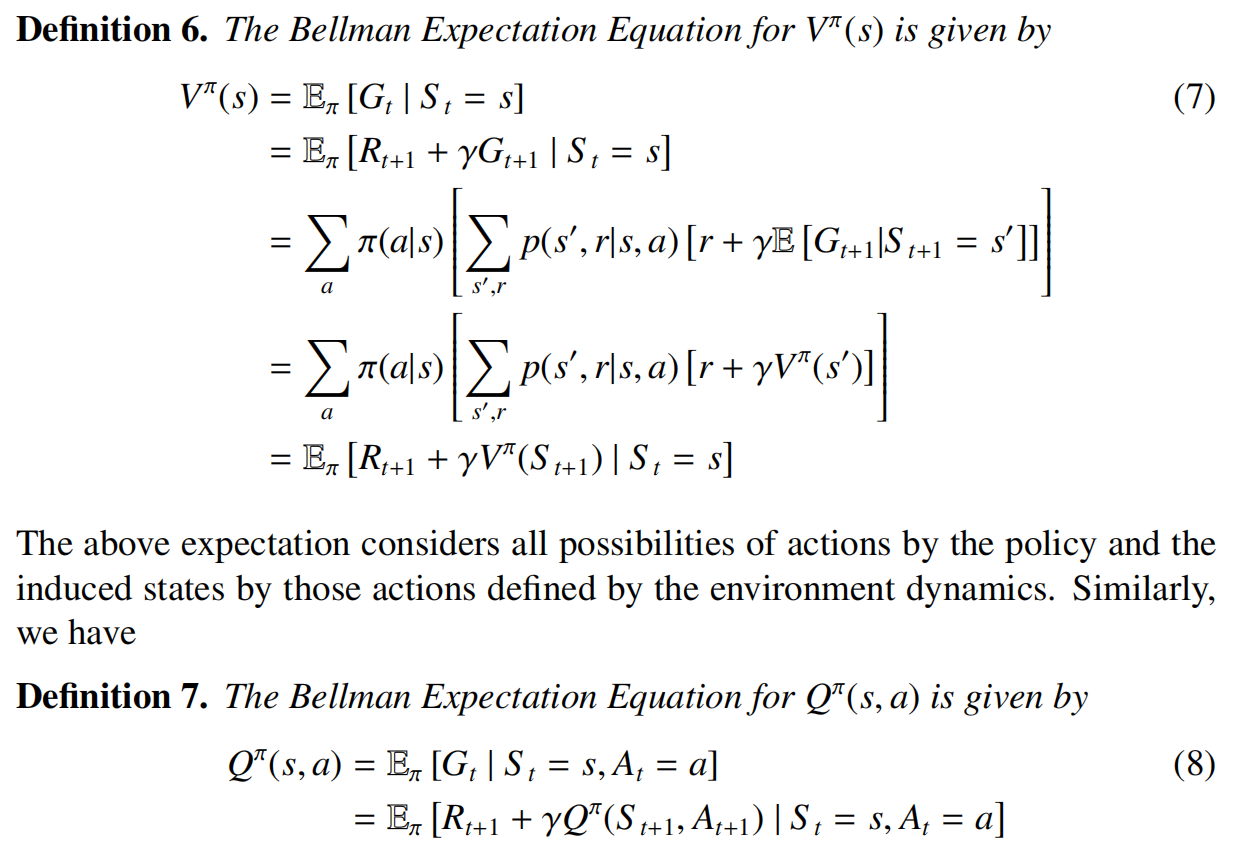


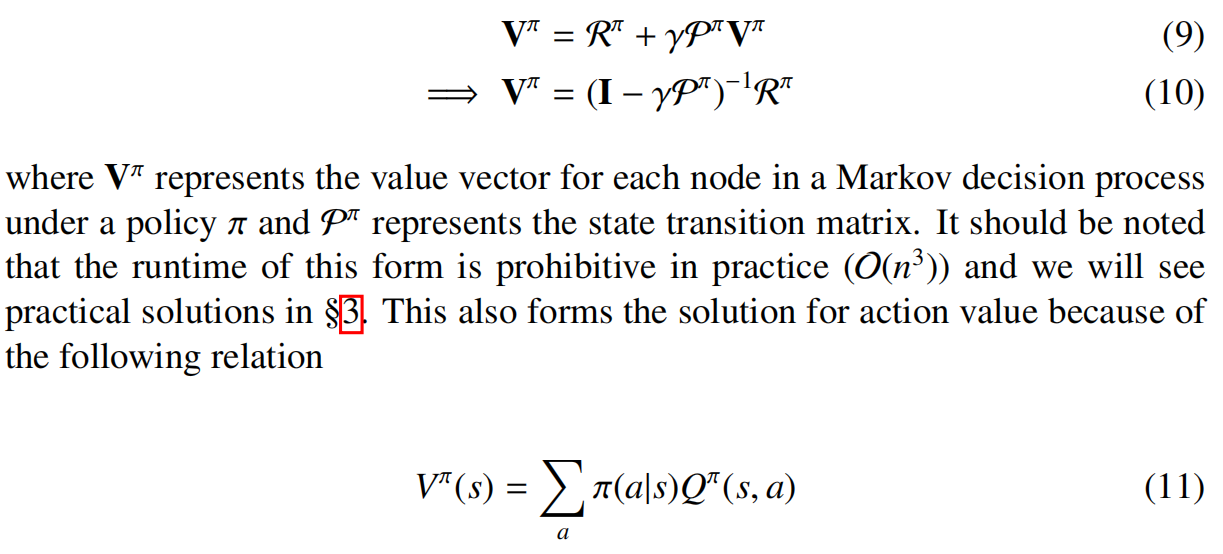
1. **Value function**

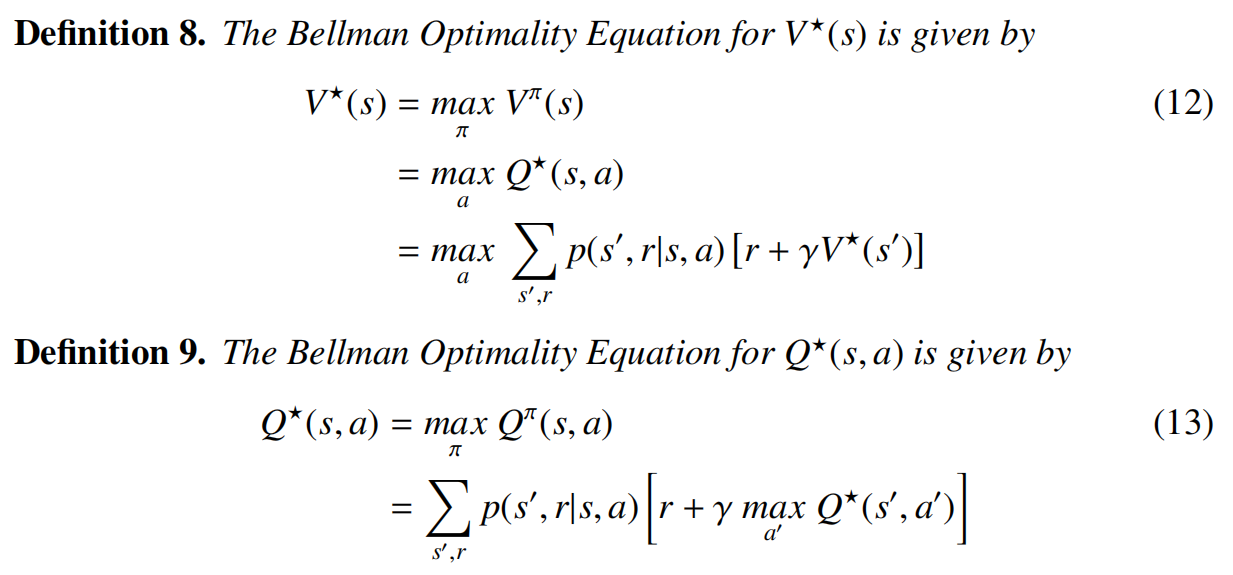




1. **Bellman equation**







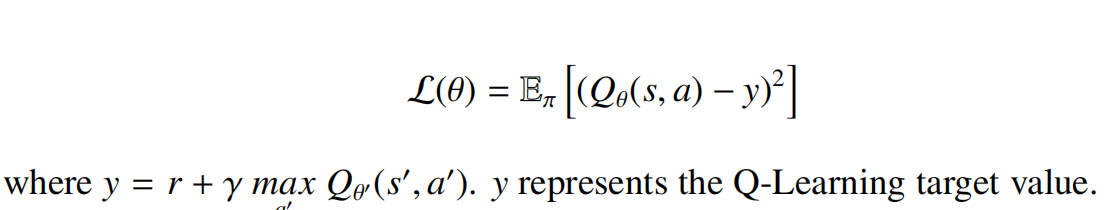
**Methods:**

两种迭代方法解决动态规划问题的有价值迭代和策略迭代，但前提条件是智能体了解环境动态信息。

今天的大多数强化学习技巧可以被归为对于状态、动作对表格的估计和学习状态、动作的选择策略。

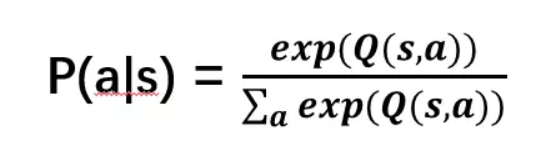
1. **Q-learning:**

目标是最小化期望loss以学习最好的状态动作对。



该方法通过大量采样trajectory进行估计，同时为了覆盖所有的策略可能进行ε探索或者玻尔兹曼探索

Boltzmann Exploration：基于Q值概率化，通过概率采样选择动作。

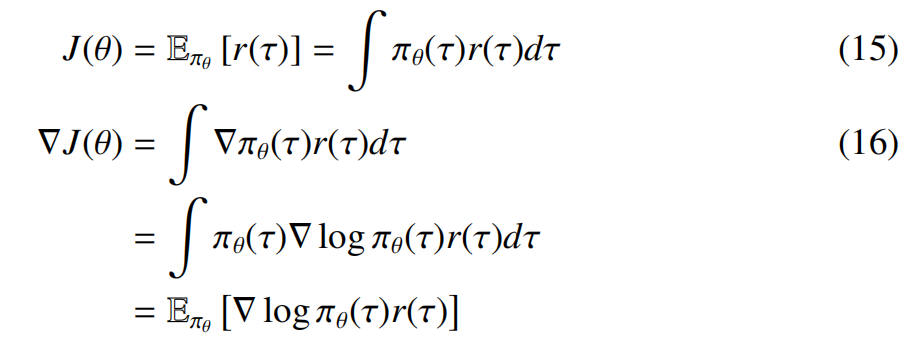


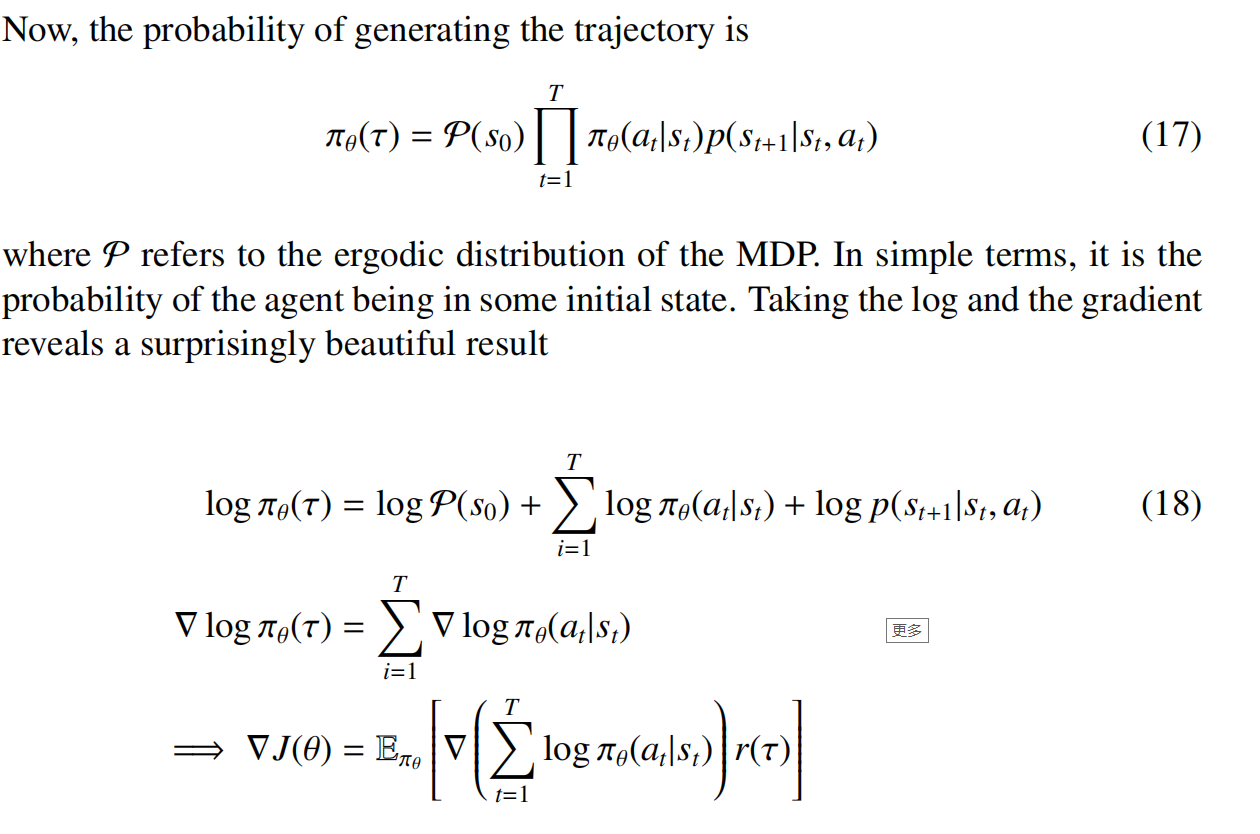
<https://www.jianshu.com/p/dcfd927e7598>

另一种用于稳定性能和克服灾难性遗忘问题的技术是使用经验回放缓冲区，从该缓冲区中随机采样转换。同时也能破坏从policy产生的trajectory的相关性，能更泛化地学习。

1. **PG methods:**

They explicitly learn a stochastic policy distribution πθ parameterized by θ. If we denote the reward of a trajectory τ generated by policy πθ(τ) as r(τ) :

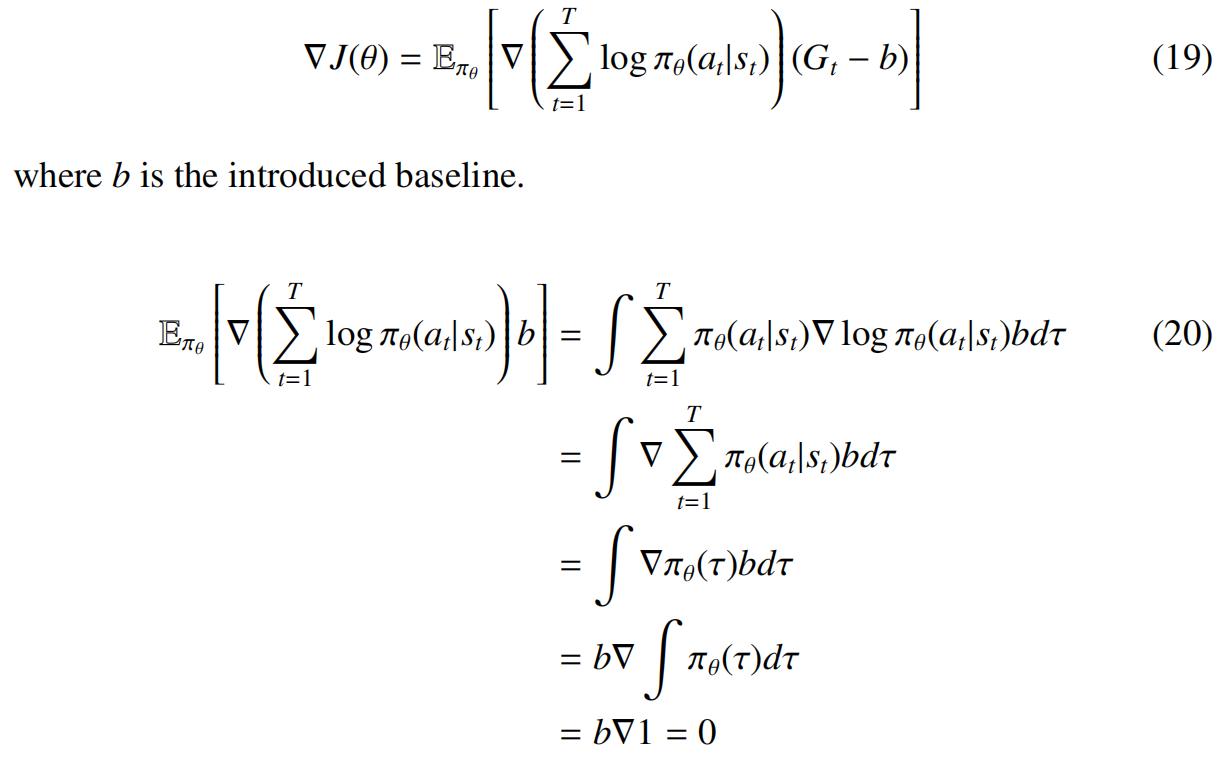




we can now just run Monte-Carlo simulations and approximate the gradient to find the best parameters θ\*.

1. **AC:**

一个有助于减少PG方差的想法是最大化一个PG目标，这个目标使用相对回报的差异，即优势函数。

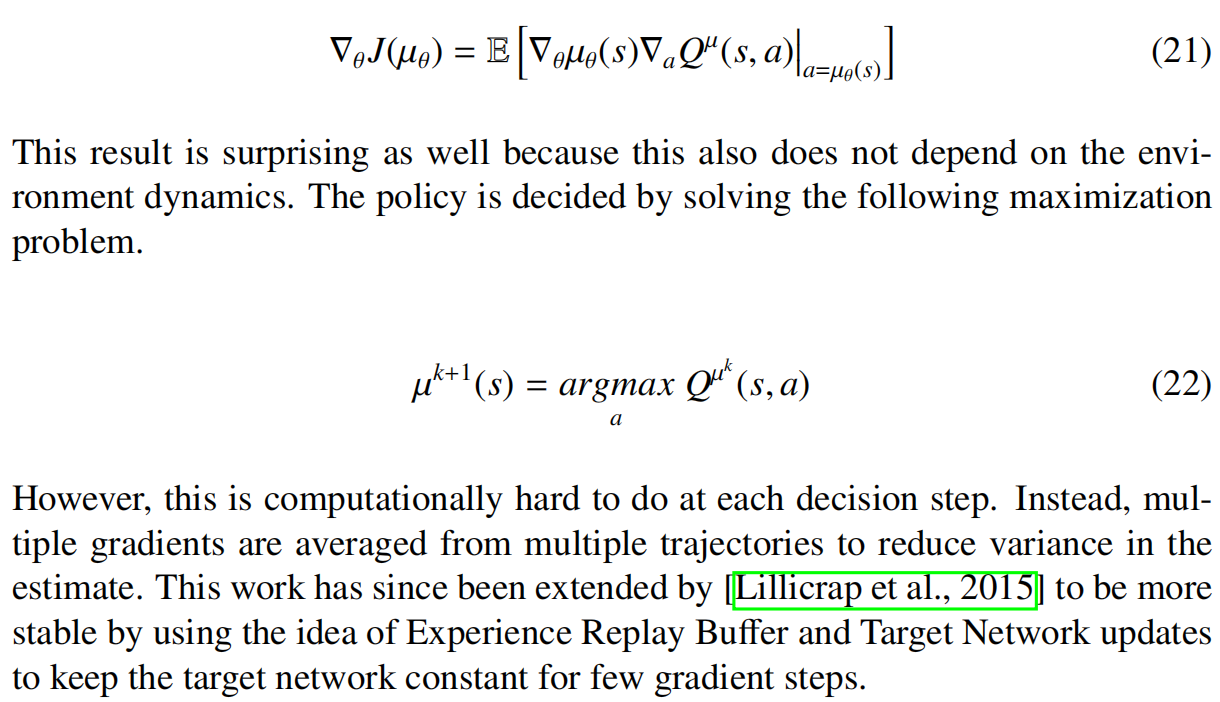


以上计算式表示增加baseline可以保证无偏同时又减小了方差。现代解决这个问题的方法是使用另一个参数化价值函数作为baseline，通常被称为critic。

目标函数和baseline的差被叫做优势函数。

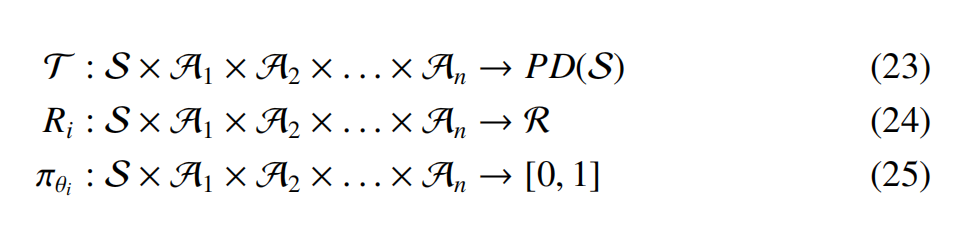
1. **Deterministic PG:**

该方法考虑一个确定性策略形式为a = µθ(s). 因此，DPG就是方差为0的随机策略方法受限的一种情况。



**Deep RL：For Multi-Agent: MADRL**

1. **Joint Action Space**

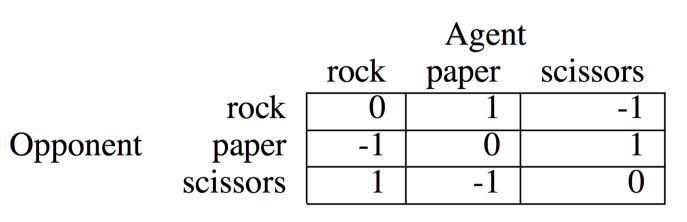


where PD(S) represents the probability distribution over the resultant state space. all the routines above are now exponentially dependent on the action space.

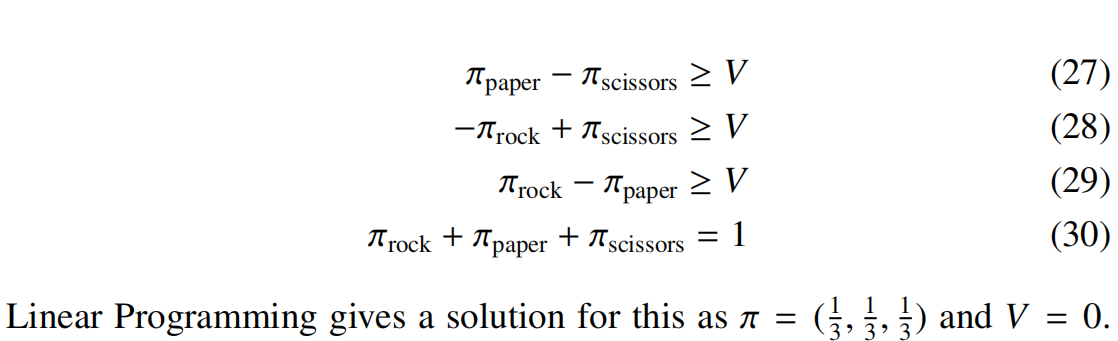
1. **Game-Theoretic Effects (博弈论效应)**

在马尔科夫游戏中，没有一个可支配的确定型决策，因为对手的决策是不确定的。

举例一个随机策略游戏，剪刀石头布，reward如下图所示，o是对手的action，a是agent的action：



线性约束的问题预期回报政策π和总价值回报V如下：



剪刀石头布的最优策略为随机策略。

彻底理解多智能体设置下的学习问题仍然是一个开放性问题。**在多智能体的环境下，评估学习不要对于收敛到纳什均衡过分看重？？？**

然而, Evolutionary Game Theory (进化博弈论)是首选框架。

1. **Credit Assignment (信用分配) and Lazy Agent (懒惰智能体) Problem**

信用分配关心的是系统的成功该如何归因于系统的各部分的贡献。

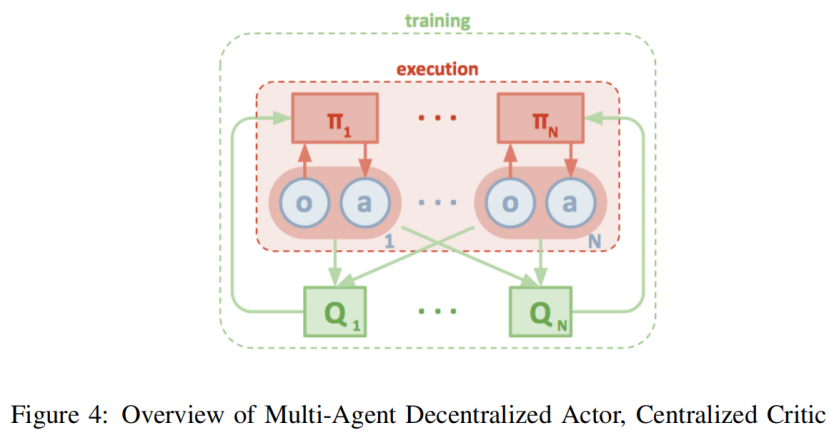
由部分可观测而引起的另一个现象叫做懒惰智能体。当一个智能体学会了一个有用的策略，另一个智能体就会减少探索，因此学习信息被第一个智能体占用并利用了。部分可观测使得一个agent过分学习，一个agent欠学习。

1. **Non-Markov Nature of Environments 非马尔可夫环境性质**

非马尔可夫性质的环境意味着需要利用以前时间的信息，处理这个问题的方法有使用循环网络，允许使用过去的状态历史。

这个问题的现代处理方法采用了相同的方法，即使用门控神经网络或卷积神经网络来构建状态序列的隐藏表示。

Decentralized Actor, Centralized Critic AC with Centralized, 集中训练，分开执行。



要使分开执行成为可能，只需要个体agent可以观测到自己的observation。

有一个集中的critic给每个agent提供完全全局state的observation观测。这有利于解决内部智能体通讯的受限的问题。

它包括MDP的拓展，通常被称为the Decentralized Partially Observable Markov Decision Processes (**Dec-POMDPs**).

这种方法允许使用策略执行中不可用的信息，具有弱收敛性。

最近使用了一种聪明的方法，通过使用政策梯度的反事实基线来估计优势函数。

为了克服这种复杂性，我们往往有必要采取去中心化策略（decentralized policies），即每个智能体仅根据局部的动作观察历史而选择它们自己的动作。此外，即使在联合动作空间不是特别大的情况，学习期间的局部可观察性和通信约束也可能需要使用去中心化策略。

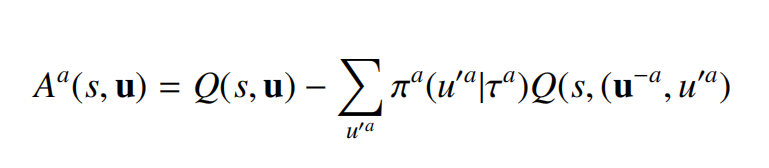
在合作的环境下，联合动作通常只能产生全局奖励，因此每个智能体都很难判断它自己对团队的贡献。然而，这些奖励在合作的情况下通常是不可获取的，且往往不能鼓励单个智能体为更大的集体利益而做出牺牲。这种现象通常会在充满挑战的任务中大幅度地阻碍所智能体的学习，即使相对较少的智能体也会产生这种显现。

首先，COMA 使用一个中心化的 critic（只用于学习期间），同时执行期间只需要 actor。由于学习是中心化的，因此我们可以使用中心化的 critic 限定联合动作和所有可用的状态信息，同时每个智能体的策略只限定在其自己的动作观察历史上。

1. COMA 使用反事实基线（counter-factual baseline）表示不使用事实采用的动作作为基线。**这一想法的灵感来自差异奖励**（difference rewards），**其中每个智能体从目标奖励中学习，目标奖励将比较全局奖励与当那个智能体的动作被默认动作替代时接收的奖励。**尽管差异奖励是执行多智能体信度分配（credit assignment）的有效方法，但是它们需要访问模拟器或者评估的奖励函数，并且总体上不清楚如何选择默认动作。COMA 通过使用中心化的 critic 计算一个特定智能体的优势函数而解决了这一问题，该函数对比了**当前联合动作的评估回报**与**边际化单一智能体动作的反事实基线**，同时保持其他智能体的动作不变。这类似于计算一个贵族效用（aristocrat utility），但避免了策略和原来效用函数之间的递归性相互依赖问题，这主要是因为反事实基线（counter-factual baseline）对策略梯度的期望贡献为零。因此，COMA 并不依赖额外的模拟、近似或关于适当默认行动的假设，它会依赖于集中式的 critic 为每一个智能体计算一个分离的基线，以推理特定智能体动作改变的反事实。

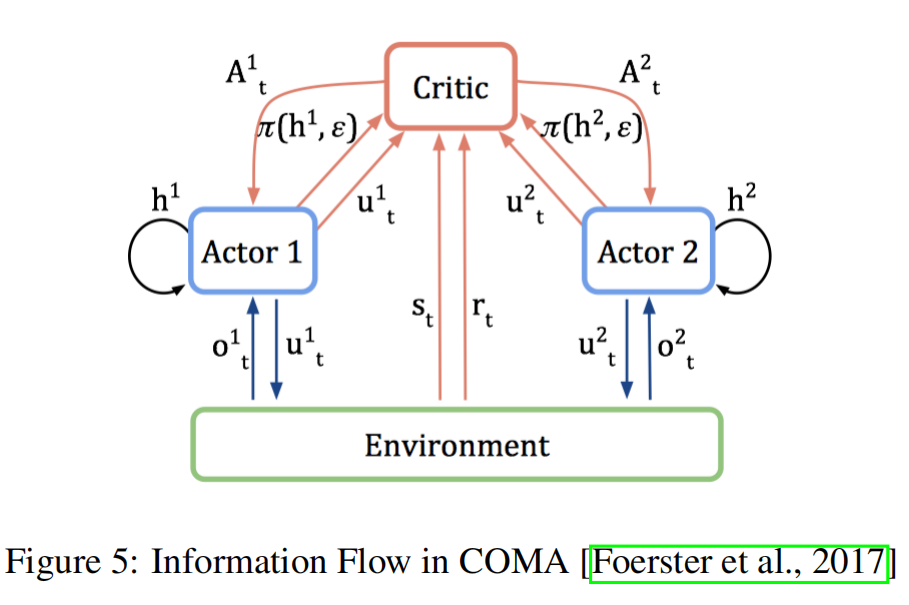
第三，COMA 使用一个 critic 表征以高效地计算反事实基线。在单次前向传播中，以所有其它智能体的动作为条件，它会为给定智能体的所有不同动作计算 Q-values。因为单个集中式的 critic 可用于全部的智能体，所有智能体的所有 Q-values 都能在一个批量的单次前向传播中计算。

<http://www.sohu.com/a/217441065_465975>



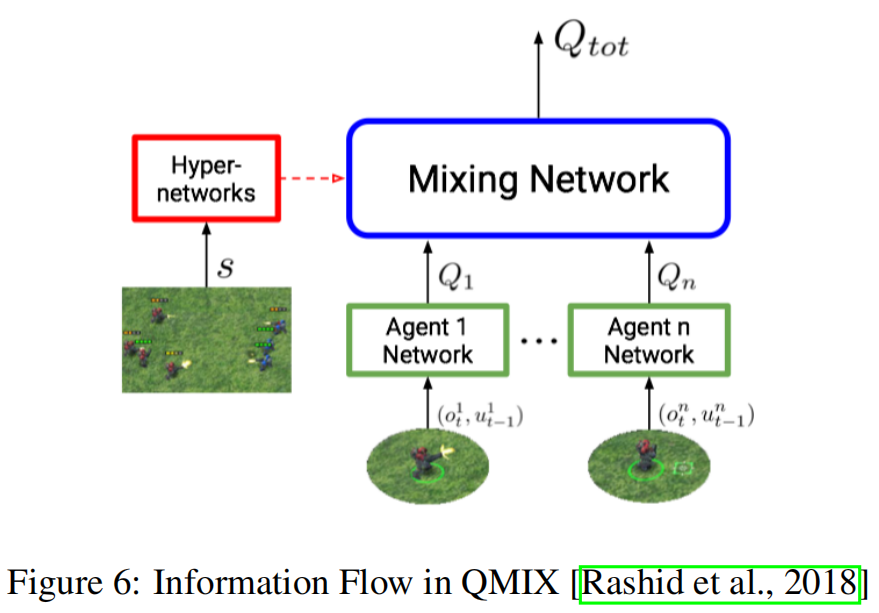
where the advantage estimate is computed for each agent and the baseline marginalizes out the actions (u) of an agent a. This allows the centralized critic to reason about the counter-factual in which only a’s actions change.

训练使用AC的方法。



h表示门控神经网络中存在的actor的隐藏状态，以说明环境的非马尔科夫性质。

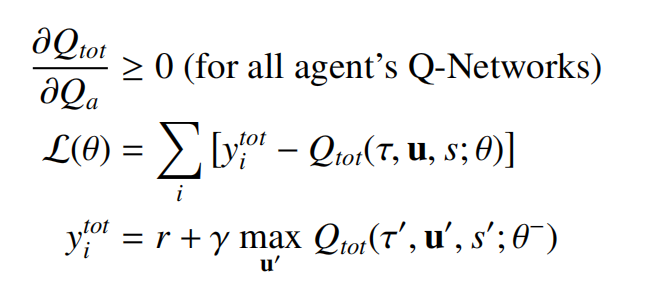
另一种方法使用Q-learning相同的范式只是提出一个新的目标函数以监督loss，主要想法就叫QMIX。使用联合值函数the joint value function的线性分解，线性分解的每个部分作为一个agent的value function，同时保持局部最大和全局最大的值函数value function的单调性。



Hyper Network是一个辅助网络。

Mixing Network通过获取辅助超网络生成的权值的绝对值来加强单调性。

训练使用DQN方法



**Future Work**

A large set of problems still stay open in both the **theoretical** and **applied aspects** of Reinforcement Learning Systems for Multi-Agent Systems.

还是建议以Dec-POMDPs作为理论基础。