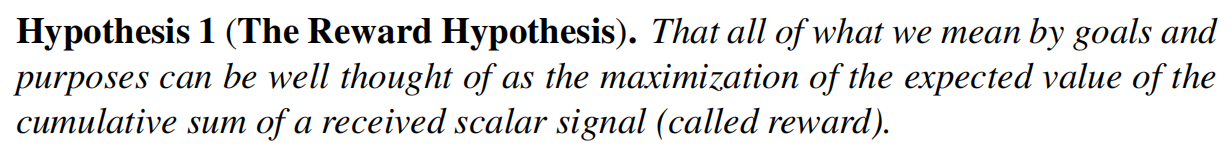
Multi-Agent Reinforcement Learning:

A Report on Challenges and Approaches

**Introduction：**

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

machine learning.



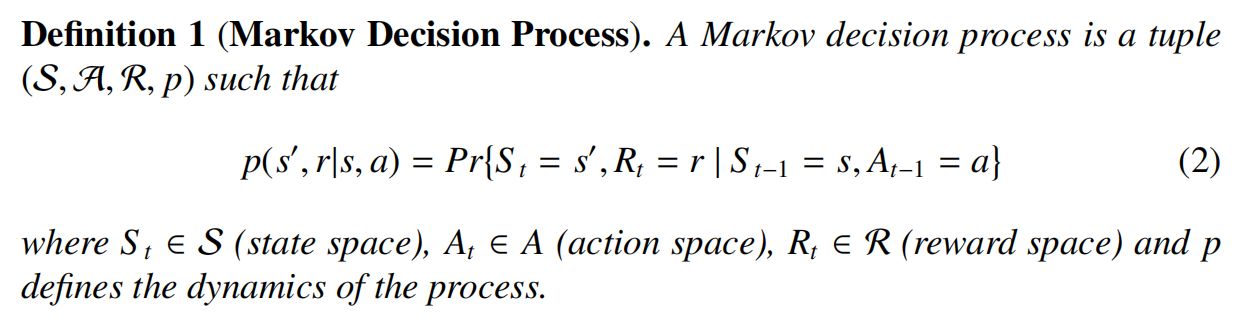
EC is different from RL in that there is no explicit interaction between the environment and the agent.

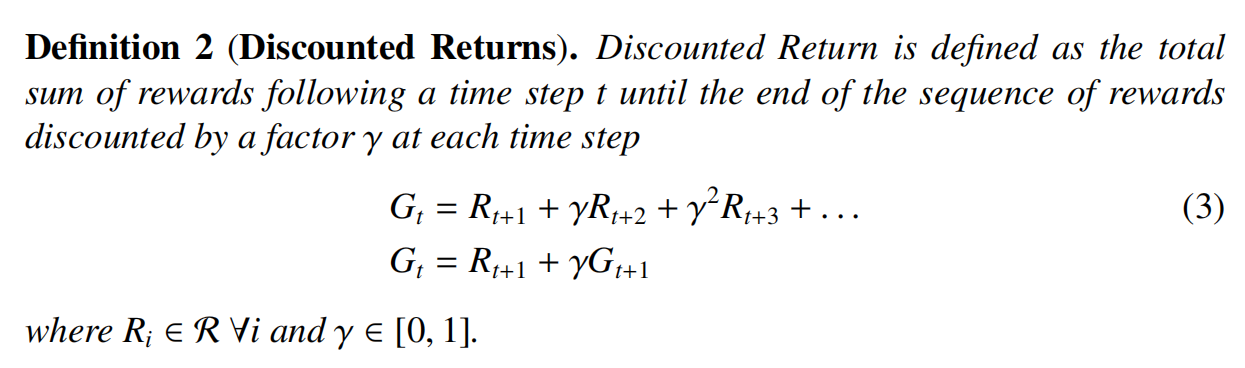
A number of complex problems in today’s society can be modeled as a Multi-Agent Learning problems.

**RL Methods：**

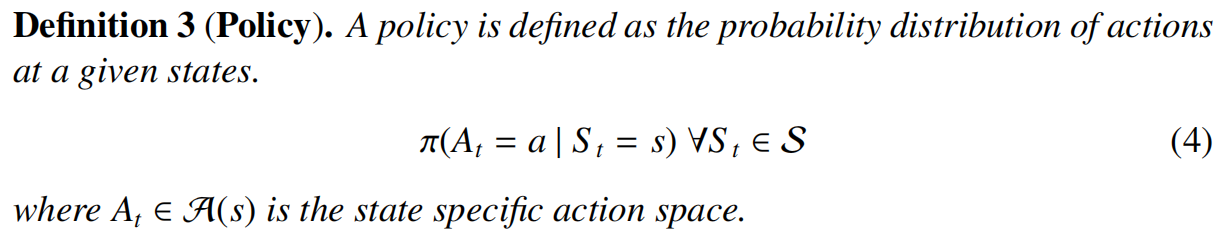
**Basic:**

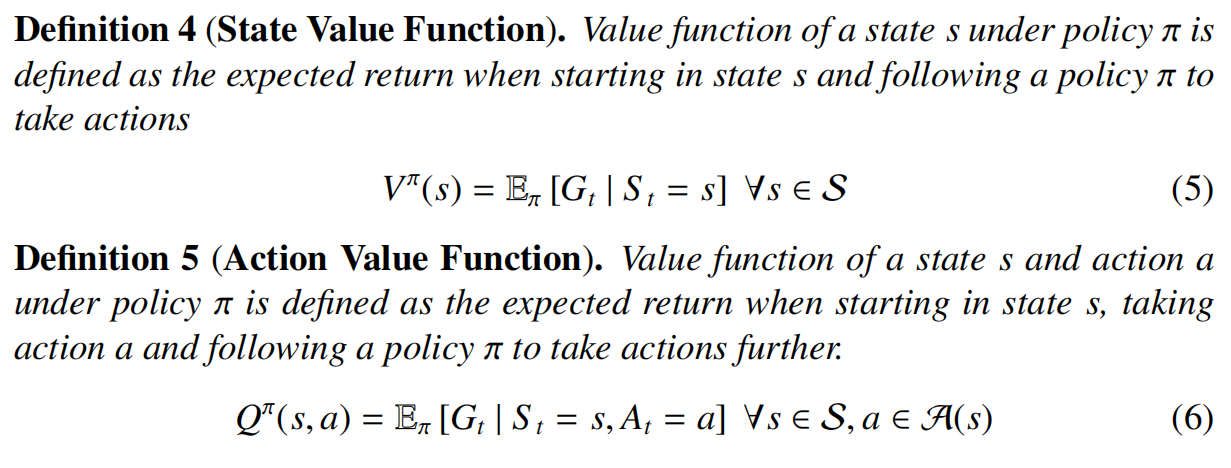
1. **MDP**



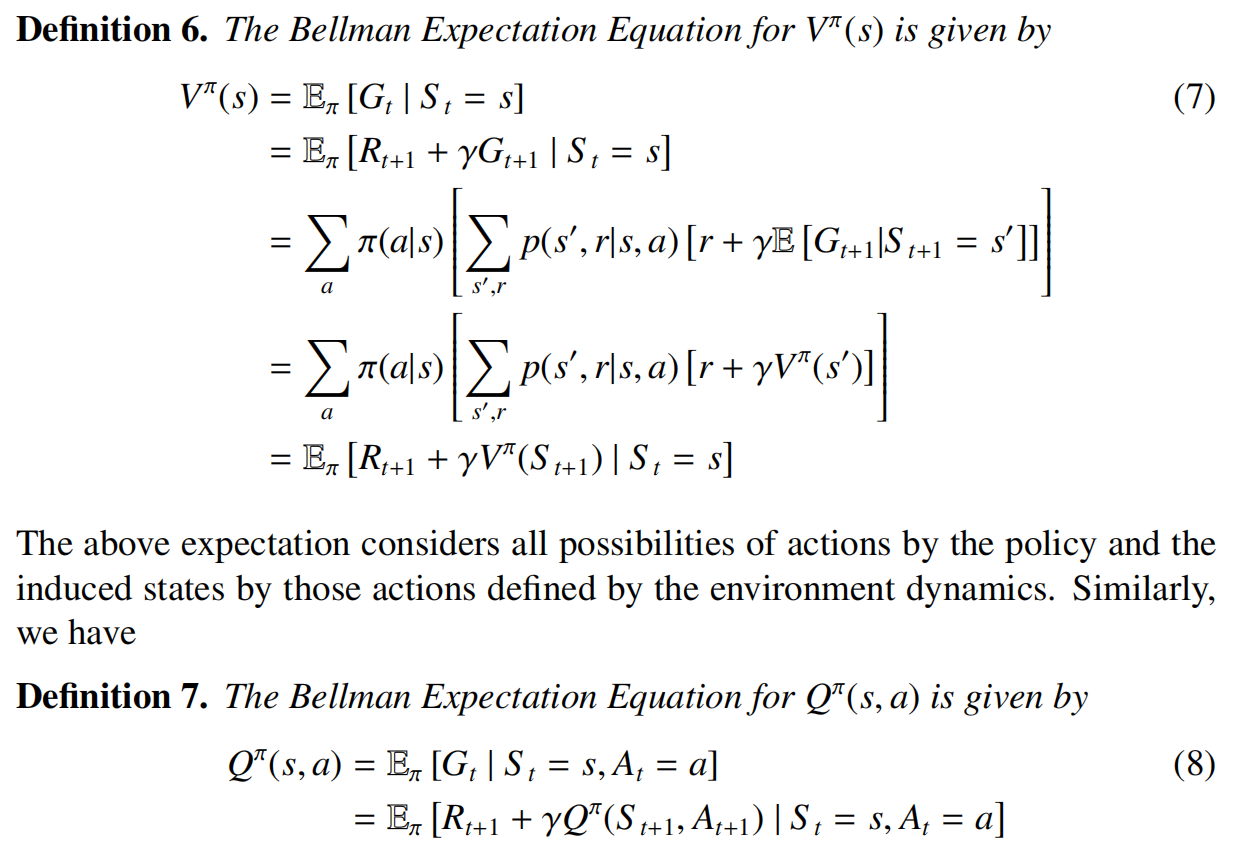


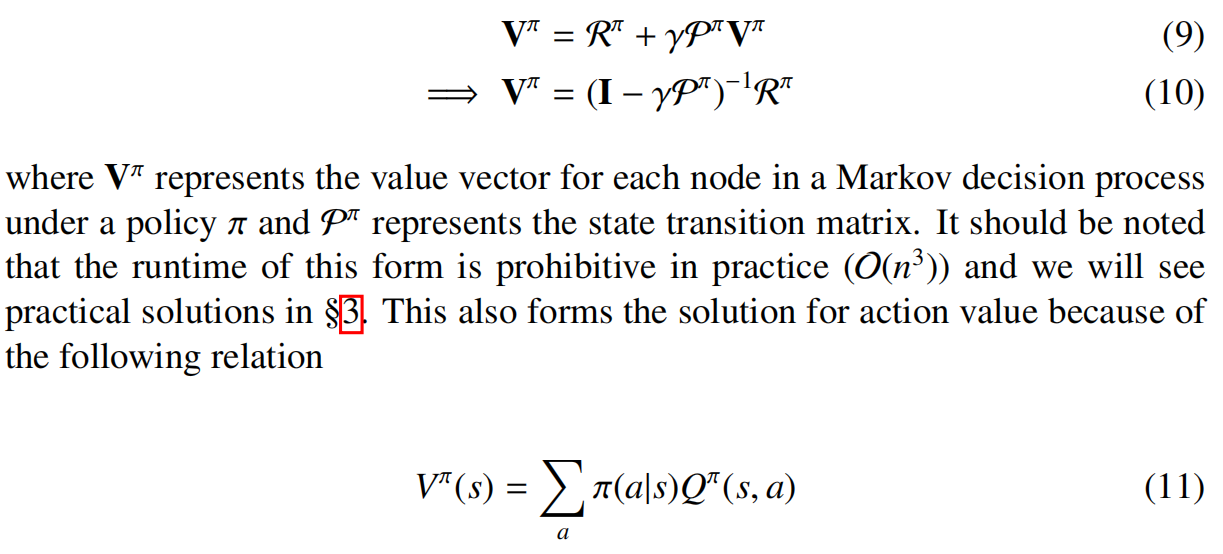
1. **Value function**

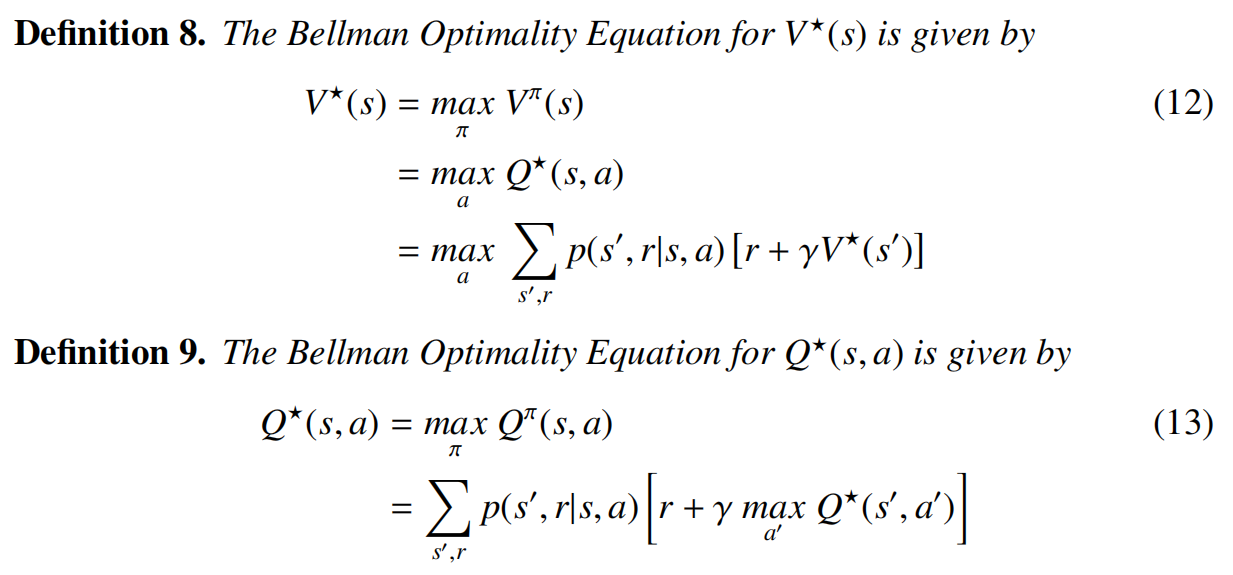




1. **Bellman equation**







**Methods:**

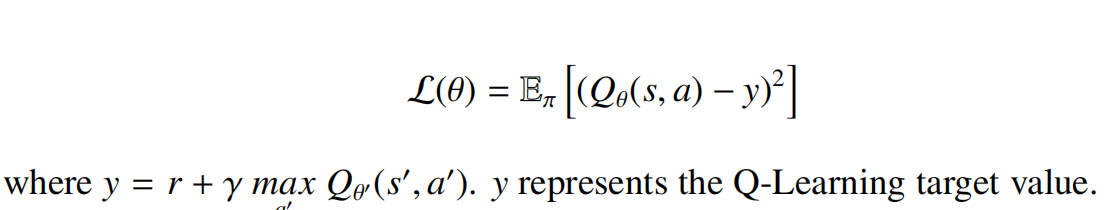
The two classic iterative approaches to solve the Dynamic Programming problems

are known as the Value Iteration and Policy Iteration. And the condition is the environment dynamics are known to the agent.

Most Reinforcement Learning techniques today can be broadly classified into either **approximation of the table of state and action values**, **learning a policy distribution for each state or a mixture of the two**.

1. **Q-learning:**

The objective here is to learn the action-value function Qπ(s, a) for policy π by minimizing the expected loss L(θ).

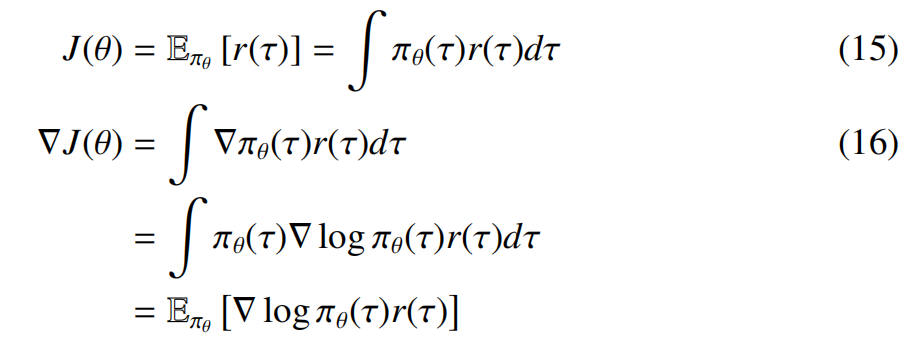


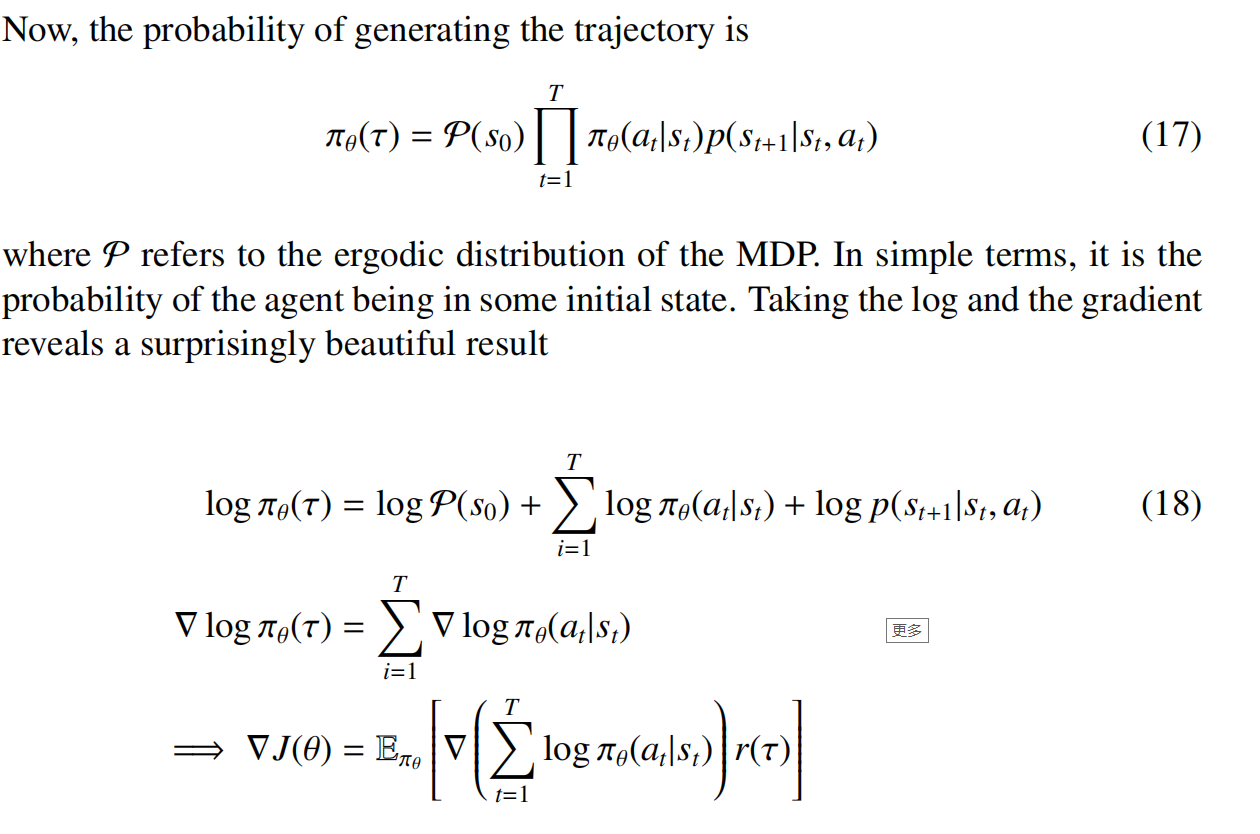
It is approximated by sampling a large number of trajectories following a suitable coverage policy for exploration like ε-greedy and **Boltzmann exploration.??**

The other technique used to stabilize the performance and overcome the problem of catastrophic forgetting is to use an Experience Replay Buffer from which transitions are sampled at random. This also helps break correlation between the sequential transitions from trajectories generated by the policy π.

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) :

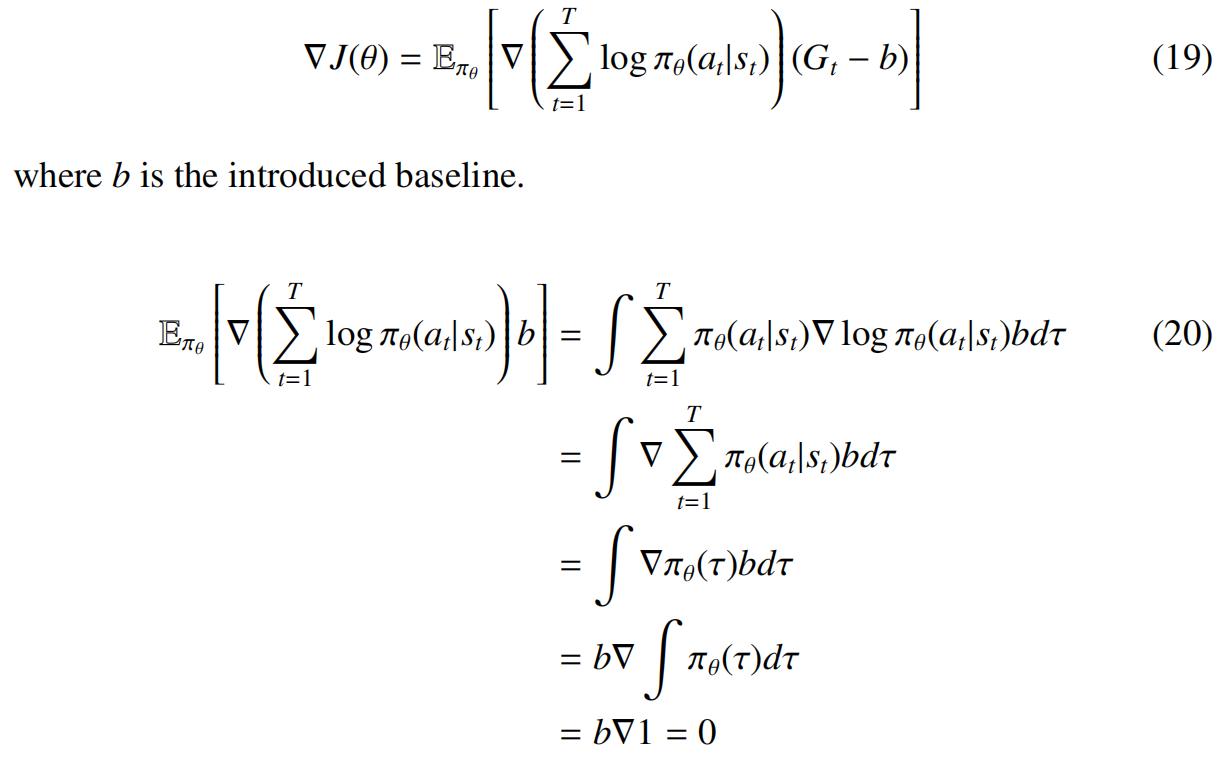




we can now just run Monte-Carlo simulations and approximate the gradient to find the best parameters θ\*.

1. **AC:**

an idea that helps reduce the variance is to instead maximize an objective which keeps track of the relative reward difference.

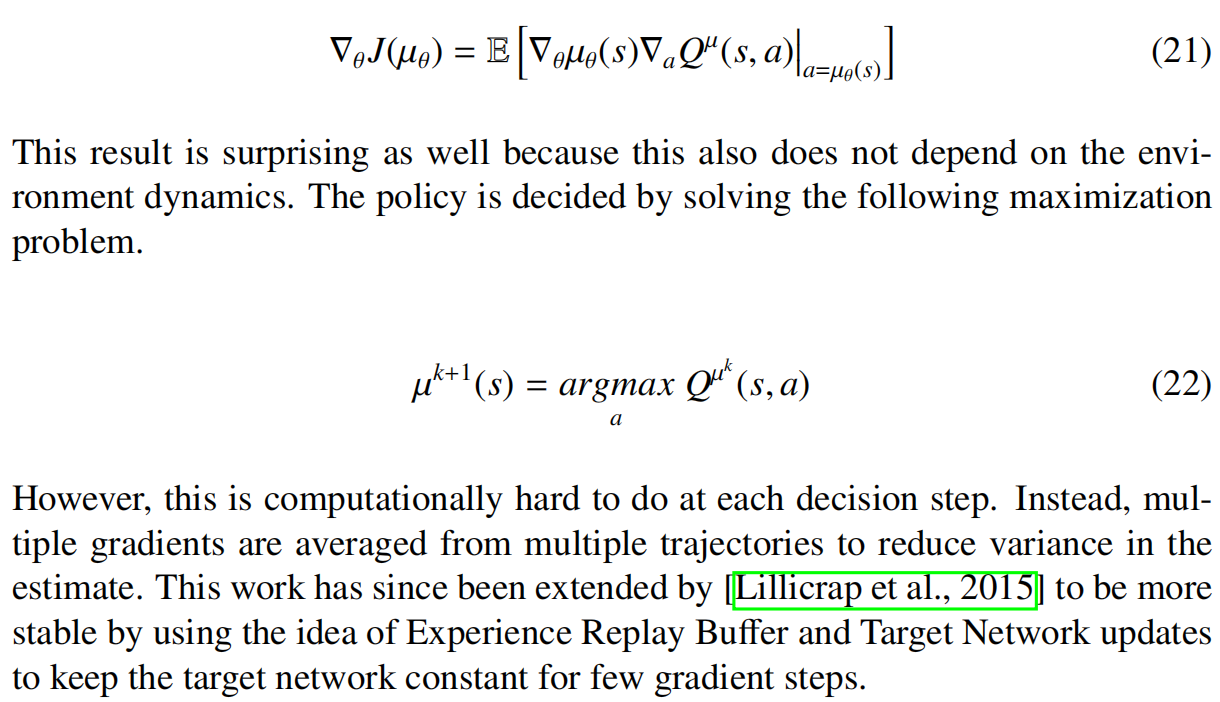


The above calculations show that the addition of a baseline keeps the gradient estimate unbiased while decreasing the variance. modern methods resolve to using another parametric value function Vω(s) as the baseline which is commonly known as the ”critic”.

The difference value of the the objective and the baseline is also known as the advantage estimate. **AC??**

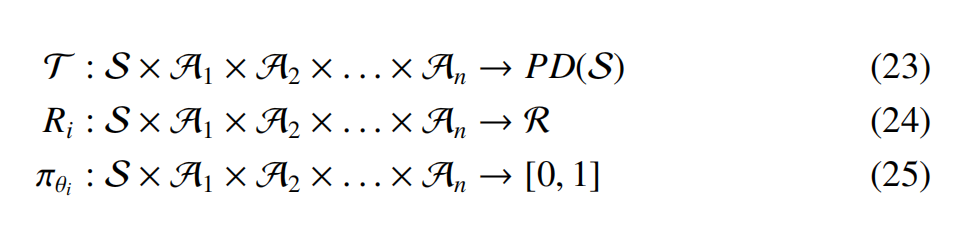
1. **Deterministic PG:**

This approach considers deterministic policies of the form a = µθ(s). It turns out that the Deterministic Policy Gradient is a limiting case of Stochastic Policy Gradients when the variance approaches zero.



**Deep RL：For Multi-Agent: MADRL**

1. **Joint Action Space**

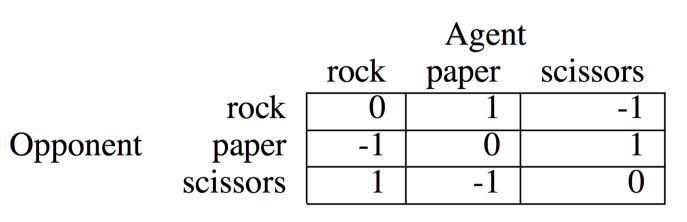


where PD(S) represents the probability distribution over the resultant state space. all the routines above are now exponentially dependent on the action space.

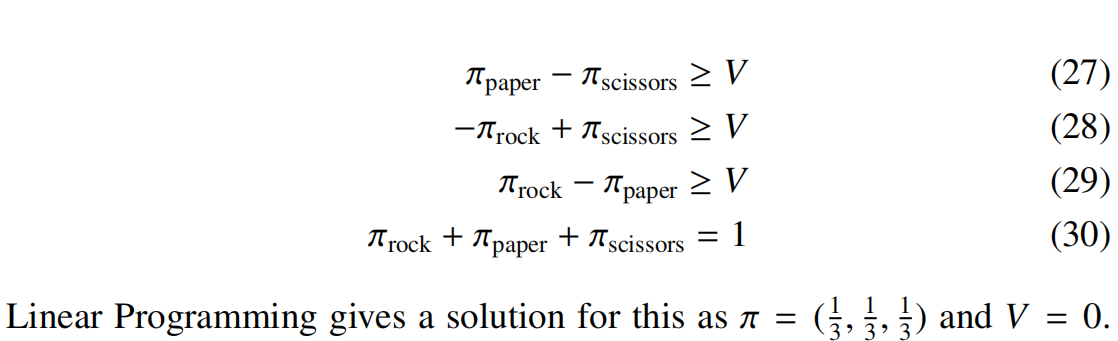
1. **Game-Theoretic Effects (博弈论效应)**

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：



The linear constraints on the problem for expected rewards for a policy π and total value pay-off V:



剪刀石头布的最优策略为随即策略。

A thorough understanding of the learning problem in multi-agent settings is still an open problem.

**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？？？**

Instead, Evolutionary Game Theory (进化博弈论) is emerging as the preferred framework.

1. **Credit Assignment (信用分配) and Lazy Agent (延迟代理) Problem**

Credit Assignment is how the success of an overall system can be attributed to the various contributions of a systems components. 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。

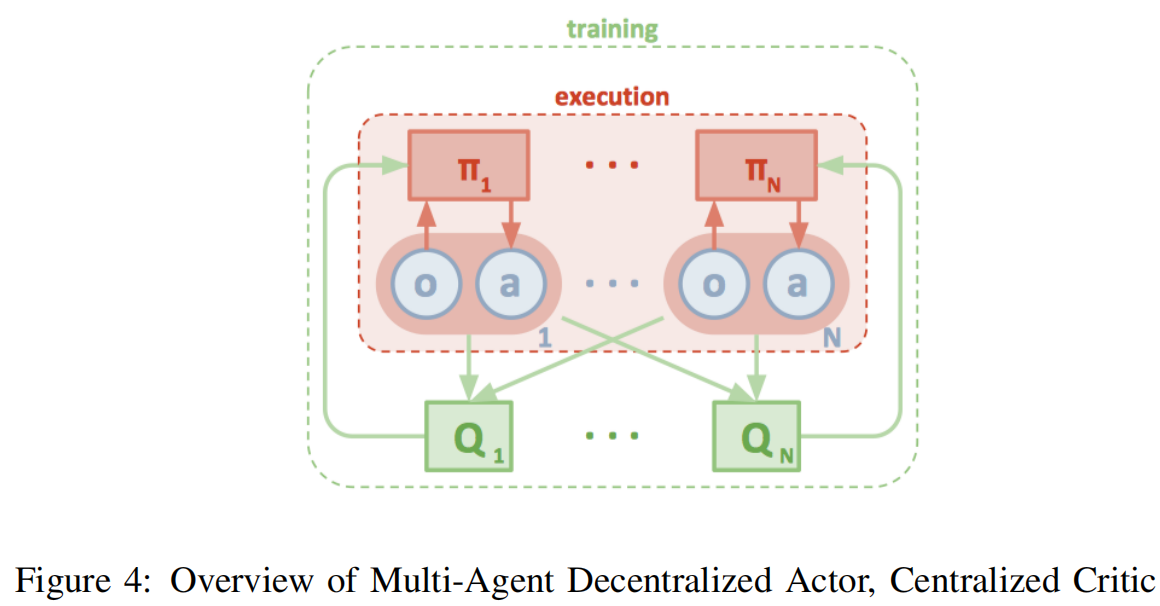
Another phenomena which arises due to partial observation is called the “Lazy Agent Problem”. When one agent learns a useful policy, the second agent can be discouraged from exploration so as to not affect the first agent’s performance. 部分可观测使得一个agent过分学习，一个agent欠学习。

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

**This problem was addressed using Recurrent Networks, which is allowed the state representation over sequences of state history.**

**Modern treatment of this problem has followed the same approach to build a hidden representation of state sequence with Gated Neural Networks or Convolutional Neural Networks.???**

**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。

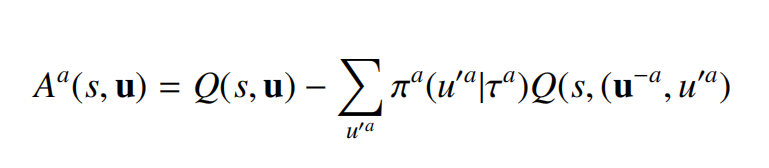
有一个集中的critic给每个agent提供完全全局state的observation观测。

This helps work around the constraint of inter-agent communications.

It includes an extension of MDPs known as the Decentralized Partially Observable Markov Decision Processes (**Dec-POMDPs**).

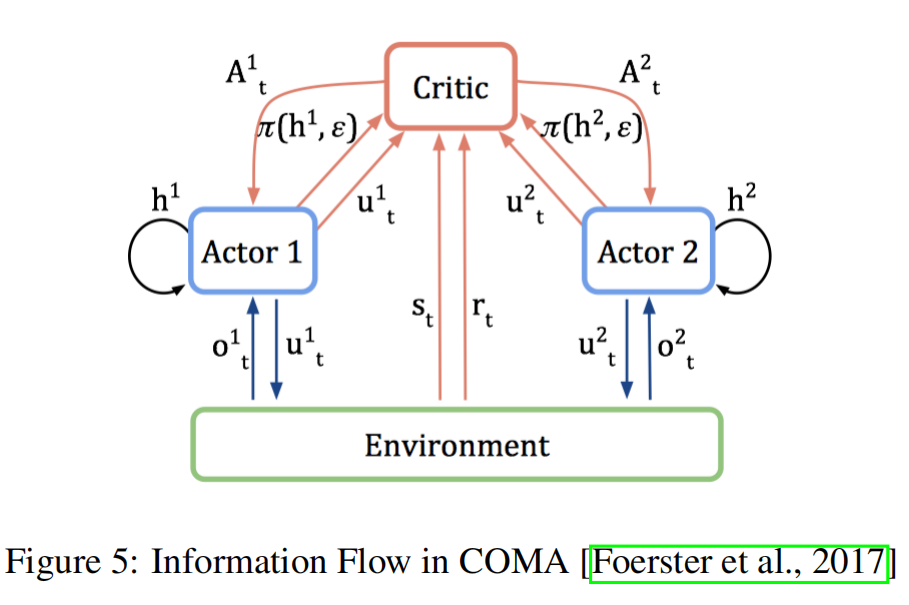
An approach where the agents are allowed to rehearse with the unavailable information during policy execution and also present weak convergence guarantees. 这种方法允许使用策略执行中不可用的信息，具有弱收敛性。

**One recent approach introduces a clever approach to estimate the advantage estimate by using a counter-factual baseline for policy gradients??**



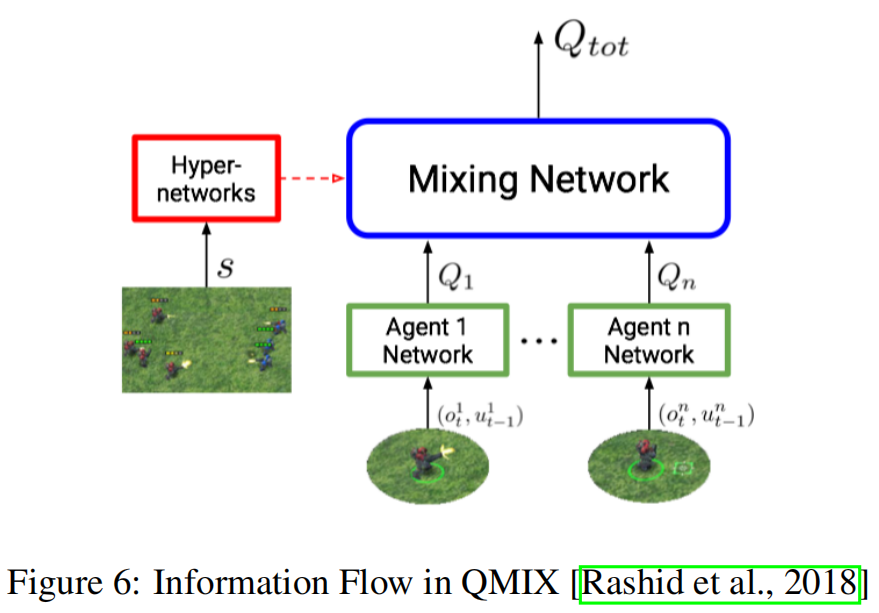
where the advantage estimate is computed for each agent and the baseline marginalizes out the actions (u) of an agent a. This allows the centralized critic to reason about the counter-factual in which only a’s actions change.

训练使用AC的方法。



h represents the hidden state of the actors which are existent for the Gated Neural Networks to account for the Non-Markovian nature of the environments.

Another approach applies the same paradigm to Q-Learning by proposing a new objective for the supervised loss. The key idea is named 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。

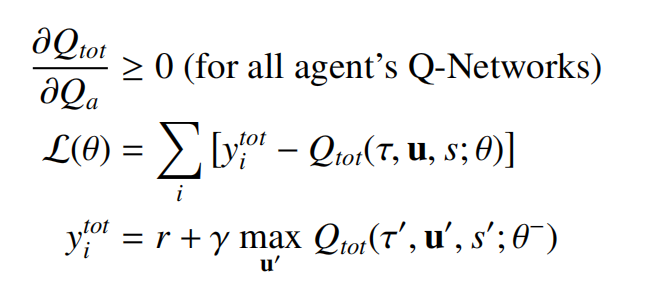


Hyper Network is an auxiliary network. 是一个辅助网络。

Mixing Network enforces the monotonicity 单调性 by taking absolute values of the weights generate by an auxiliary hyper-network.

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.

I am also optimistic about Dec-POMDPs to be the fundamental theoretical framework.

**competitive self-play??**