强化学习分类：

通过价值选行为：Q learning，Sarsa，Deep Q Network

直接选行为：Policy Gradients

想象环境并从中学习：Model based RL

Model based比Model free多出了一步，对真实环境建模；同时Model based还可以做出想象，就好像alphaGo下围棋一样

Policy-based是基于概率的，会给出每个动作发生的可能性

Value-based是基于价值的，会给出每个动作的价值，选择价值最高的动作

连续的动作不能使用Value-based，可以使用Policy-based

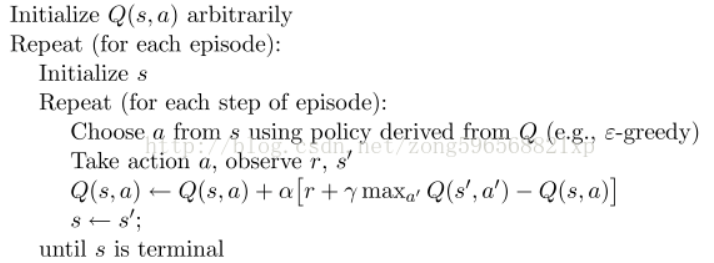
Monte-Carlo回合更新，机器做出一轮动作后进行更新

Temporal-Difference update 单步更新，机器每做出一个动作都进行更新

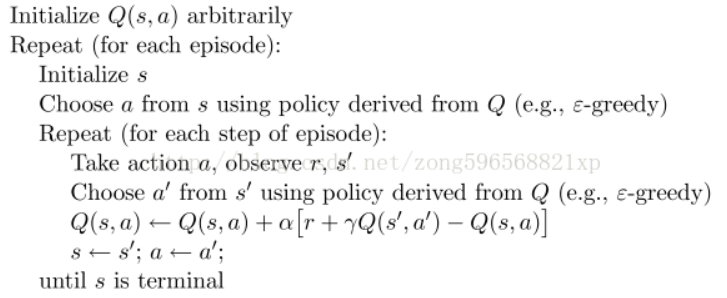
On-Policy 在线学习，边玩边学习

Off-Policy 离线学习，看着别人玩的来学习，或者自己白天玩晚上学习自己白天玩的

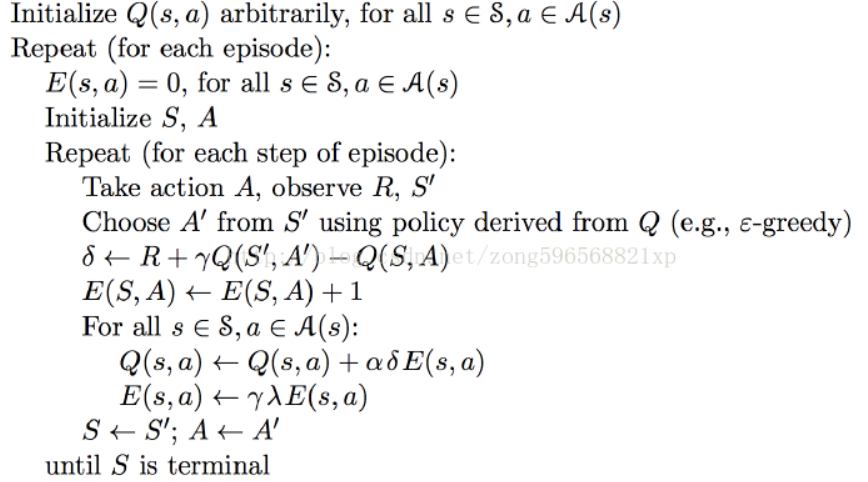
Q-learning算法：



Sarsa算法：



Sarsa-lambda算法：



Sarsa-lambda会在一次任务过程中学到前面所有步骤的回报，因此学习效率更快，但是可能会因为之前有很多赘余的步骤被给予了高回报导致最终的探索过程有效但很冗余。也可能会因为回报太远，在获得回报前agent遇到过很多次挫折导致agent畏首畏尾

Actor-Critic和Policy Gradient的区别：Policy Gradient只能在回合结束进行更新，而Actor-Critic可以通过Critic进行实时的更新

Actor-Critic因为每次更新参数前后都存在相关性，所以网络只能片面的看待问题，导致神经网络容易学不到东西

DDPG:Google DeepMind 提出的一种使用 Actor Critic 结构, 但是输出的不是行为的概率, 而是具体的行为, 用于连续动作 (continuous action) 的预测. DDPG 结合了之前获得成功的 DQN 结构, 提高了 Actor Critic 的稳定性和收敛性.