1. 除了可以用线性值函数来近似值函数

还可以用神经网络来近似值函数

线性：θ = θ - α[Q(s,a;θ)-r-γQ(s’,a’;θ)] Q(s,a;θ)对θ进行求导

神经网络则是 θ = θ - α\*损失函数对θ进行求导

通过两者对比发觉，重点不是如何进行具体的参数更新公式

关键是在策略评估时，需要获取单个或多个 s, a ,r，s’, (a’) 对

根据该对，输入为 s, a 输出则定义为 r + Q(s’,a’;θ)

输出根据当前值函数可得出

1. 个人理解：强化学习，关于神经网络模拟

重点在于采样对的获取，获取之后，可以视为标签数据，然后进行模型训练

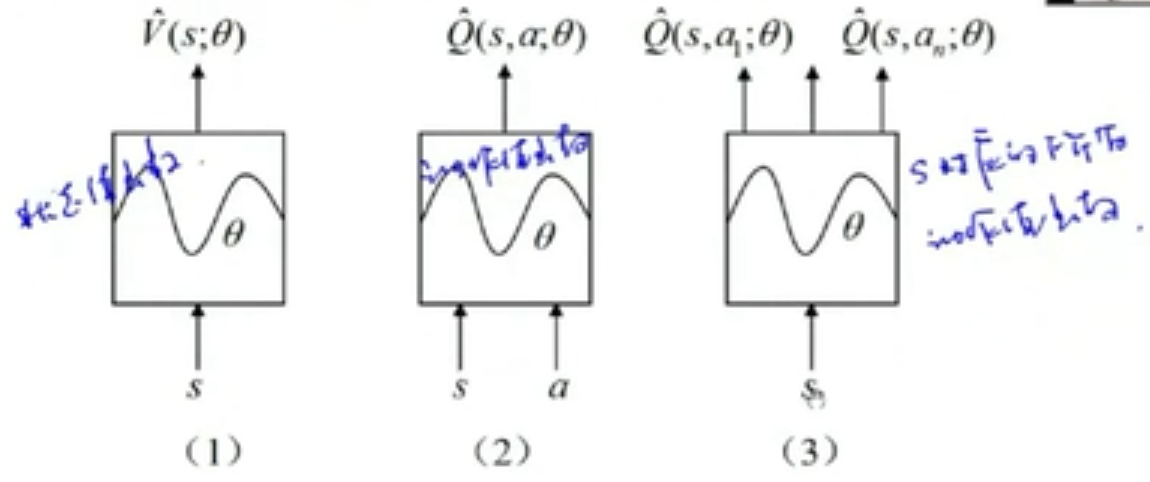
个人疑惑：关于策略如何表示，如果使用动作值函数间接表示

即将状态+所有动作组合输入到动作值函数，然后查看获取的结果

依据当前状态所有动作的评分，进行动作选取，或者说策略优化

并且如果选择策略优化，该如何优化策略

1. 对于以上问题及教程给出理解，有大概三种神经网络近似值函数



1. 输入为s，输出为该状态值
2. 输入为s，a，输出为动作值
3. 输入为状态，输出为所有动作值函数，即每个动作都给出评分

个人疑惑，对于策略，该如何进行调整

策略评估后给出的结果是单个状态，所有动作的评分

个人理解，给出的值函数即策略，需要对序列训练后的值函数进行调整

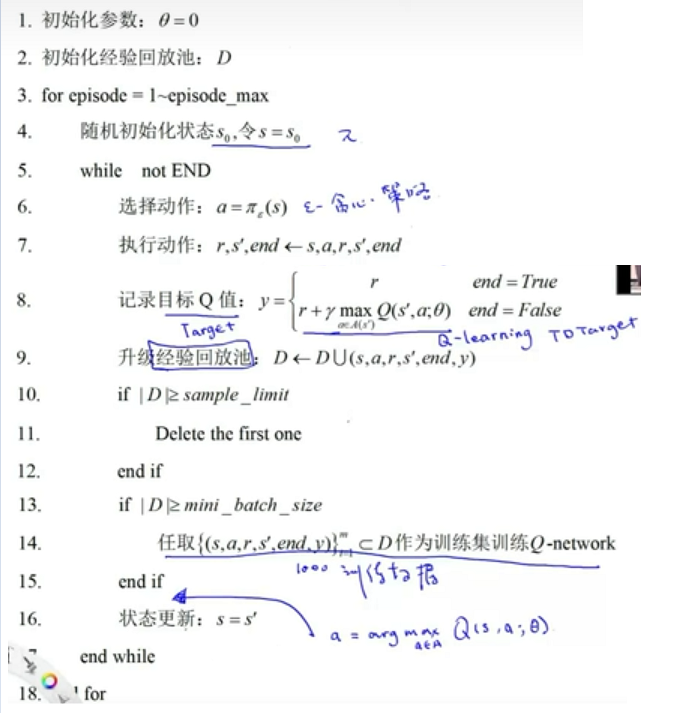
使得对应动作评分最大化

4. 训练思想，如果为终止状态

那么当前状态值评估就直接为r，而非r+γ下一个状态，下一个状态没有

理解仍然以动作值函数进行，而非通过第三种，需要多种同一状态数据

优化算法，有随机梯度下降，Adam等等



谋求按对算法理解

1. 有经验回放池，该池用于存储数据
2. 需要满足两个条件
   1. 经验回放池中的数据样本量不能超过1000条，每次都删除最开始的经验

个人觉得应该是数据量较小时，选取前面数据概率较大，已经被选取过了，一般就要删除

* 1. 数据量要大于最小batch\_size数量

1. 每一次动作符合条件的话，都会随机取多条条数据作为Q-network的训练数据集

这整个过程 都是采样，策略评估，并非训练过程

如果需要训练，需要放在采取数据集一端

在经过训练后，需要对策略进行更新，使得对于某动作的值函数最大化