**强化学习：**智能体、状态、行动、回报

**马尔可夫过程：**

**二元组**

状态空间; 状态转移矩阵，

**马尔可夫回报过程：四元组**

状态转移产生回报。折现因子**;**

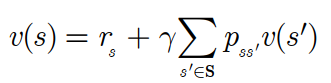
期望状态回报,

累计回报:

状态价值函数:

状态价值就是状态累计回报的期望值。

贝尔曼期望方程:，



**马尔可夫决策过程:**

**五元组**

行动导致状态转移产生回报

行动空间:；

状态转移矩阵:其中；

行动期望回报

策略:,s下选策略a的概率.

,

状态价值:

行动价值:由后续状态价值的加权和计算。



状态价值贝尔曼:

行动价值贝尔曼：

矩阵形式:，

**动态规划**—策略评价(状态价值计算)、策略改进

，

最优状态价值:

最优行动价值:

最优a价值是同时刻最优q价值

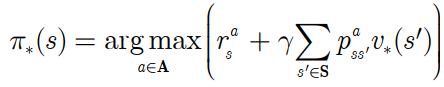
最优策略(**贪心策略**):

策略评价：计算给定策略下的状态价值。

最优策略：最优策略下状态价值取最大，行动价值取最大

策略迭代：交替策略评价和策略改进。

价值迭代：迭代贝尔曼最优方程 

策略提取：

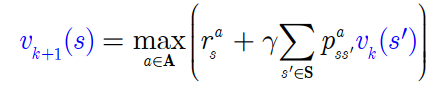
策略迭代：已知策略pi，用动态规划计算出各个行动下的

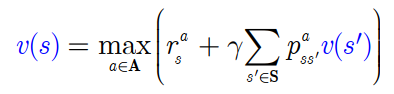
状态价值v（直至收敛，或者算一步改进一次），在某状态

可行的所有策略中选择使得状态价值最大的那个

（贪心策略），然后重复以上过程，得到一列状态、策略。



同步迭代:算完所有s更新一次价值。 

异步迭代:算完一个s就更新一次价值。

**蒙特卡洛—**无P矩阵时状态价值预测、策略改进

已知观测片段，对某个状态，以其出现作为开始

计算出这些幕各自的，对这些取平均值来估计v(s)。



首次访问/每次访问:作为起始的s是否第一次出现。

增量式蒙特卡洛:

定步长蒙特卡洛:

无模型时预测行动价值，

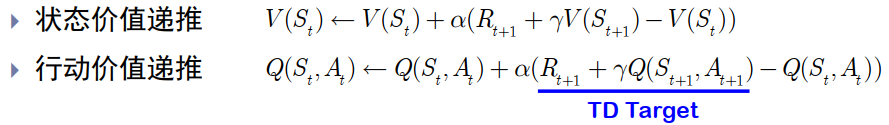
增量式:

定步长:

策略改进:

**epsilon贪心改进**:

**时序差分**—从部分序列学习，不需要运行完整幕

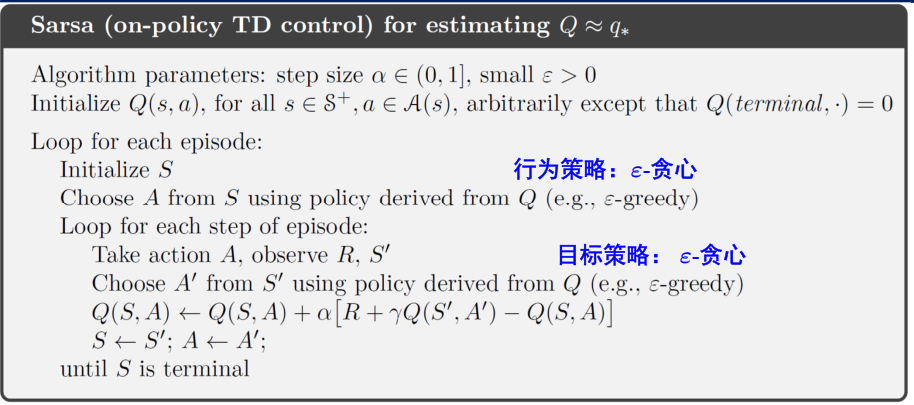


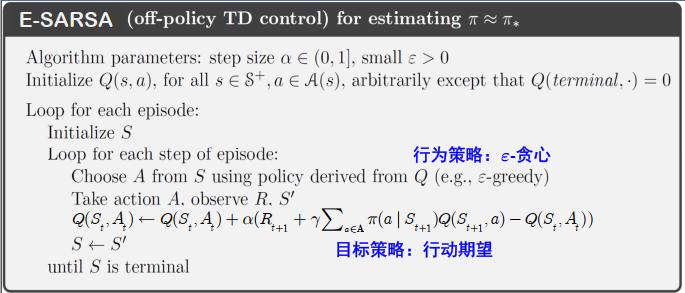
蒙特卡洛从完整序列学习,适用带终止的决策过程

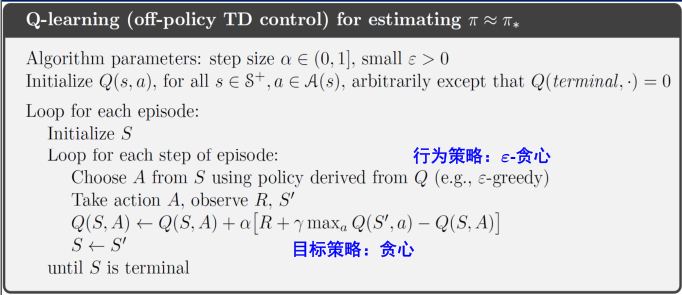
/离线学习(必须获得完整片段来估计累计汇报)/非马更有效；

时序从部分序列学习,可用于无终止状态的决策过程

/在线学习,只需获取下一状态的及时回报/马环境更有效。



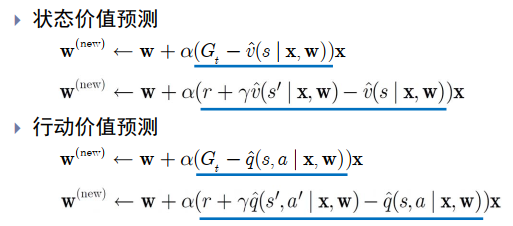




深度强化学习

状态价值函数寻找、优化准则、最优函数

增量式价值函数近似



批量式价值函数近似：

状态价值的蒙特卡洛经验回放:

状态价值的时序差分经验回放:

批量和增量的区别：增量数据只学一次，批量将数据放在库中反复学习