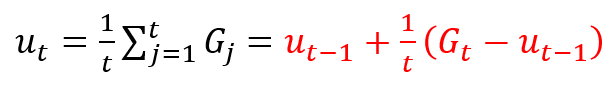
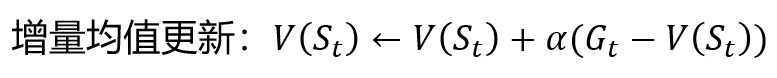
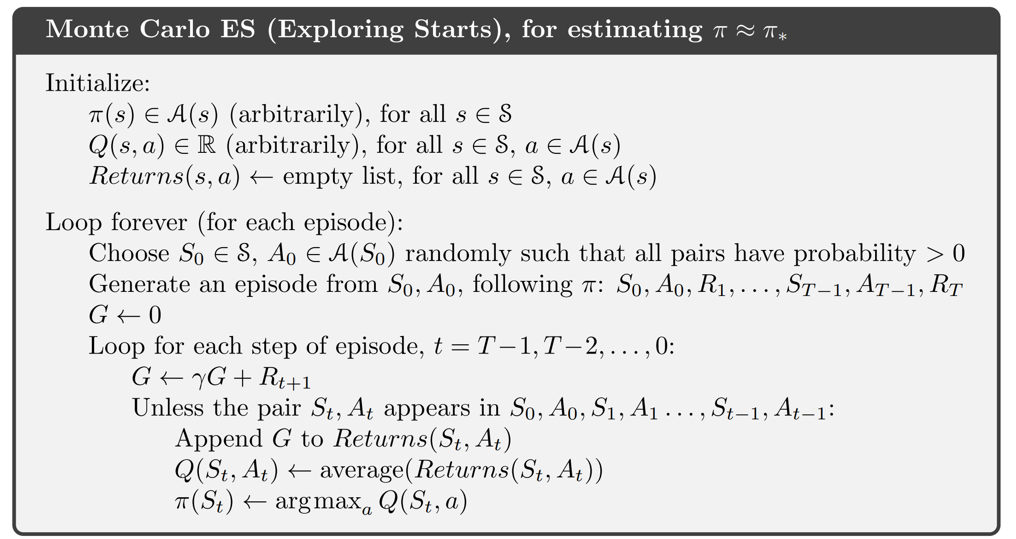
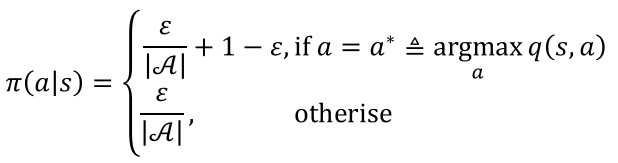
无模型预测与控制：

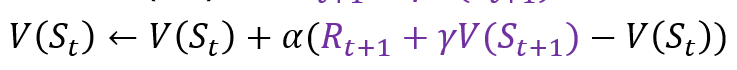
**蒙特卡罗法(Monte-Carlo, MC)**

用于预测：采样多个回合（轨迹），获取实际的回报，再取平均。（高方差，无偏估计，对初始值不敏感）

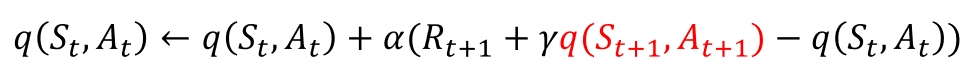
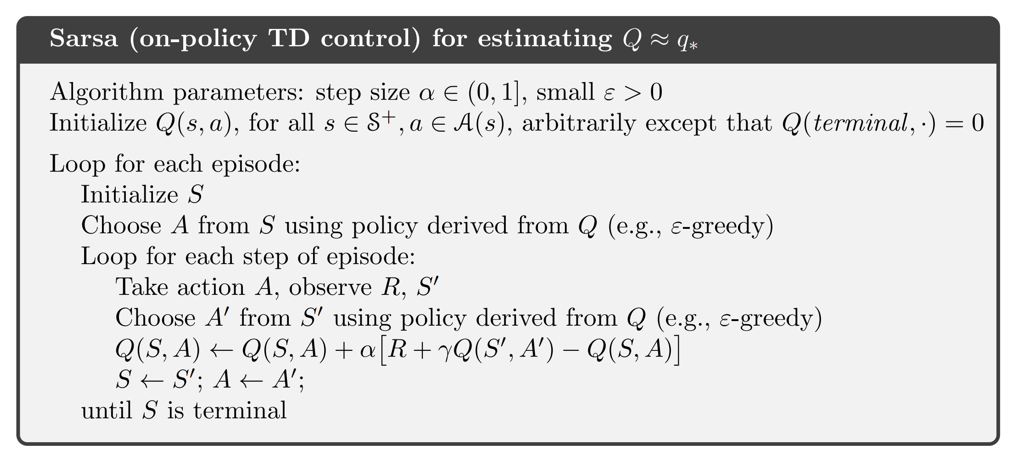
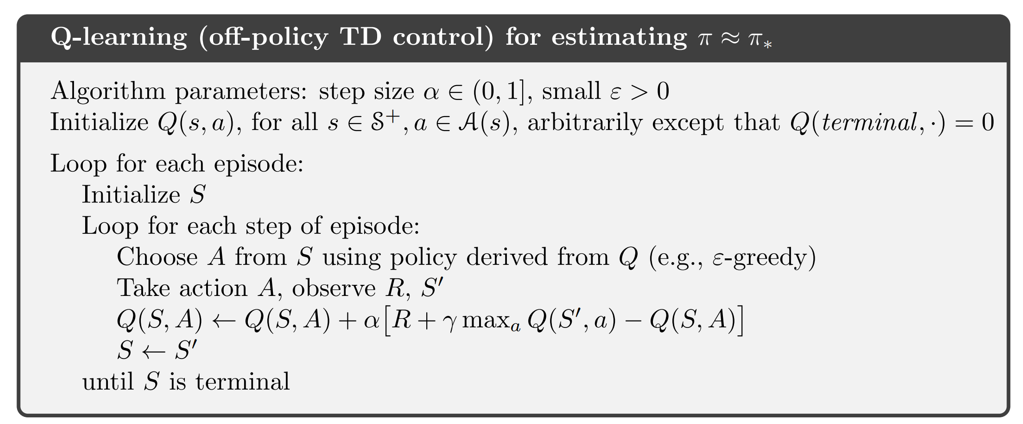
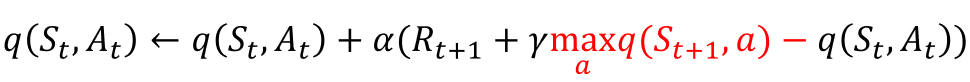
用于策略迭代：

* ES算法：所有状态-动作对都有大于零的概率作为回合的初始。
* 𝜀−贪心探索：以1−𝜀的概率执行贪心策略，即选择最大化𝑞(𝑠,𝑎)对应的动作；以𝜀的概率随机选择一个动作。

**时间差分法(Temporal-Difference, TD)**

 用于预测：在回合过程中进行学习，无需等待回合结束。（低方差，有偏估计，对初始值敏感、自举 (bootstrapping)）

用于控制：

* Sarsa（同策略 (on-policy)：从策略π产生的轨迹来学习策略π）
* Q-learning（异策略 (off-policy): 从另一个探索策略𝜇产生的轨迹中学习目标策略𝜋）