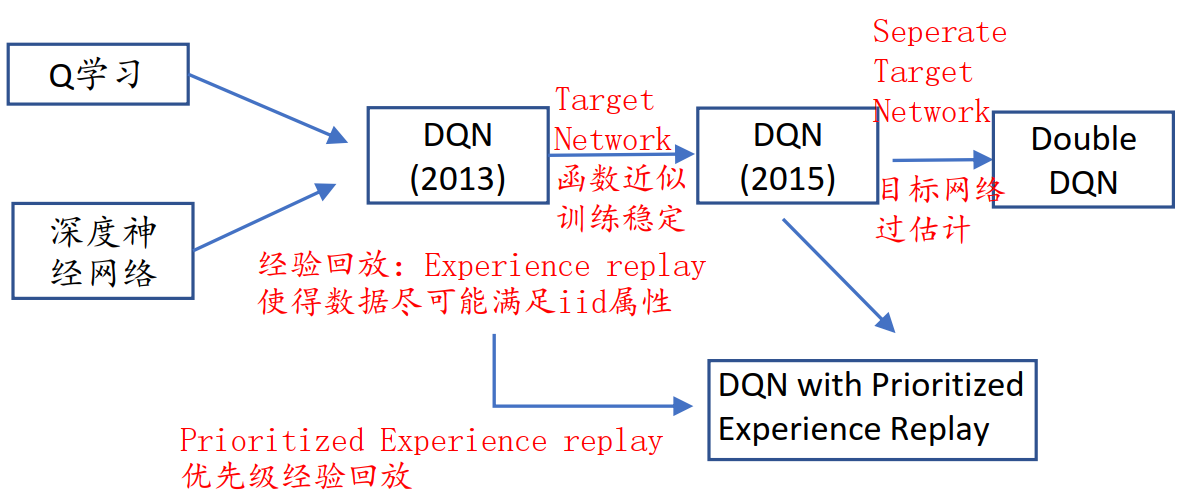
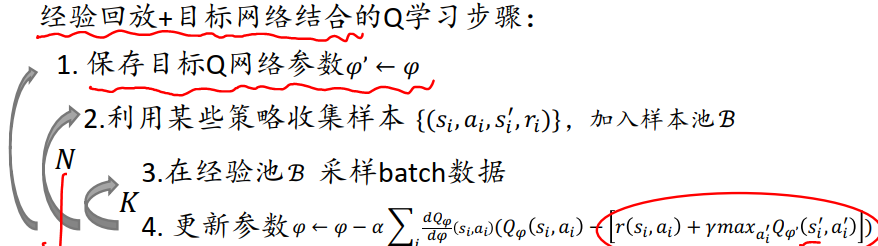
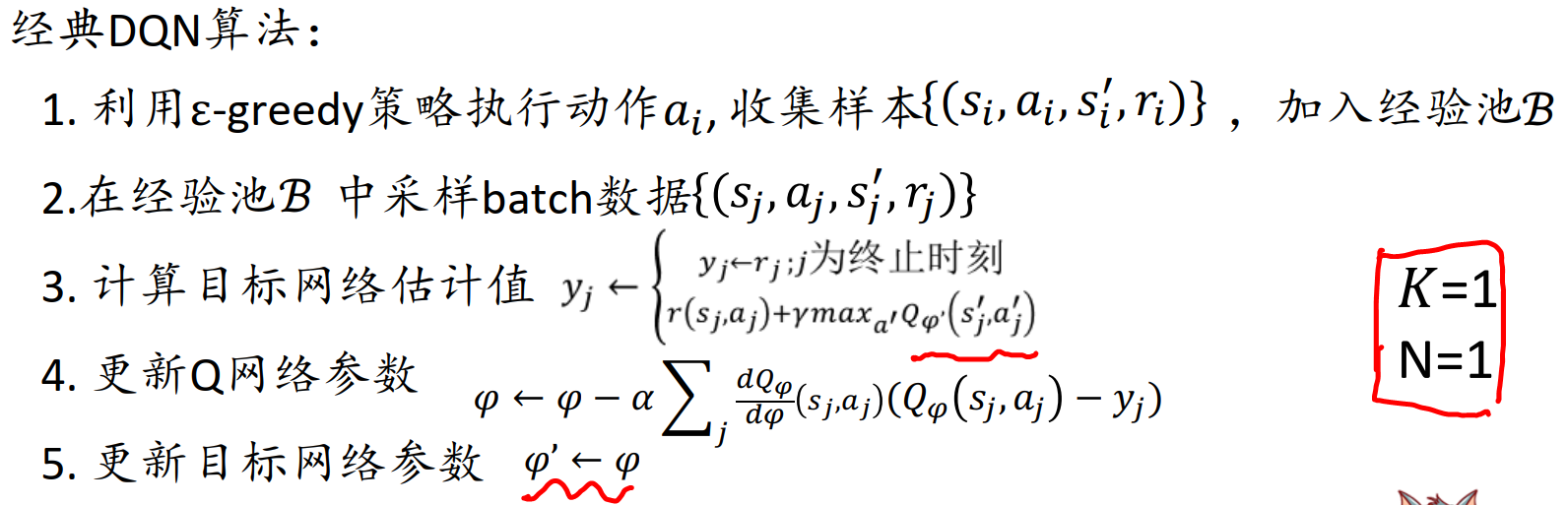
**深度强化学习:**使用深度神经网络作为强化学习的函数近似器。

传统在线Q学习问题：1，样本不满足iid；2，无法证明收敛，更新并非梯度下降。

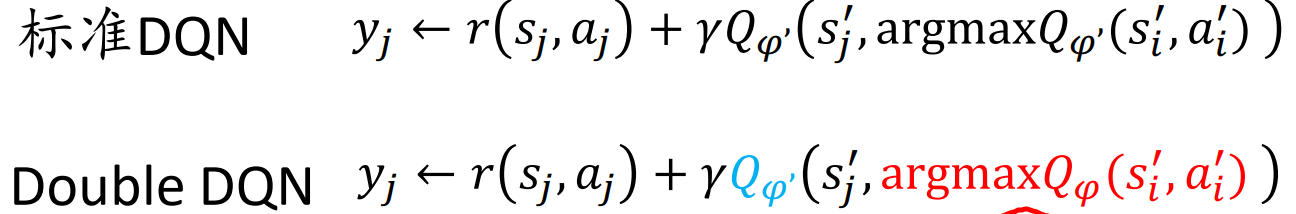




Batch训练降低方差



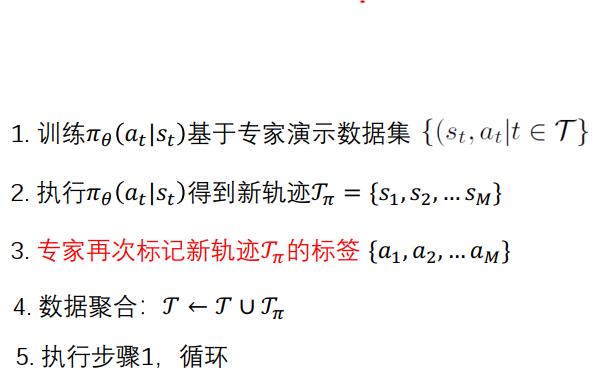
多步回报得更低的方差。Q值估计不准，TD目标值中的max引入正向偏差，导致下一时刻的目标值过估计，解:D**DQN**使用不同的网络来分别计算目标Q网络值和选择动作；当前Q选动作；目标Q值估计目标值。



DQN中经验池均匀采样的效率不高，不能反映样本的价值:带优先级的经验回放，利用TD 误差去衡量优先级，大优先级高。**优先级经验回放带来的问题**：TD 误差对噪声敏感；TD 误差小的得不到更新；过分关注TD 误差大的样本，丧失样本多样性。引入重要性采样权重来平衡“有偏”问题。前期注重优先级高后期注重无偏性。

**Dueling-DQN：**将Q函数分解成不依赖动作的值函数𝑉（s）依赖动作的优势函数𝐴（𝑠, 𝑎）。**分解的原因**：对于很多状态并不需要估计每个动作的值，增加了V函数的学习机会；V函数的泛化性能好，当有新动作加入时，并不需要重新学习；减少了Q 函数由于状态和动作维度差导致的噪声和突变**模仿学习**

**适用情况**1.提供多演示轨迹：状态-动作序列2.很容易演示，轨迹容易收集。**目标**：智能体需要找到期望策略，使得该策略下状态-动作轨迹分布尽可能匹配专家演示样本**。输入**：S,A, 𝑃（𝑠′|𝑠, 𝑎）；无𝑅；专家演示样本集 𝑠0, 𝑎0, 𝑠1, 𝑎1, …。**利用监督学习**获得期望策略 (状态到专家动作的映射)输入为演示样本集中的状态，标签为演示样本集中的动作。**行为克隆：**专家演示样本集,在专家演示样本集训练得到期望策略最小化误差**。监督学习成功的基本假设：**训练集和测试集的分布一致, 而行为克隆存在的问题：复合误差，最终导致分布漂移**样本增广：**累计误差的解决办法。**数据聚合：**



**数据聚合的问题**：1.不安全/部分训练的策略，2. 需要大量专家标记,与任务相关。**模仿学习经常面临**：数据分配不匹配；专家非马尔可夫行为。**专家多模态行为一些提升的办法：**1.增广 2. 从一个稳定的轨迹分布中采样3.聚合,4改进模型提升精度5,多的数据。**逆强化学习**假设专家策略𝜋𝐸是最优的，从专家演示样本中恢复出奖赏函数𝑅∗

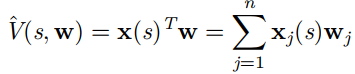
**价值函数逼近**

**好处**:更少的参数;泛化能力: 更少样本学精确价值函数;多样性:多种特征表示和逼近器结构。**价值函数逼近的类型：**1. V(s);2. Q(s, a)3.Q(s)输出所有的 Q(s, a), ∀a

**线性函数逼近:**

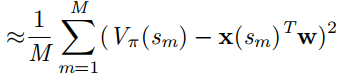
用特征向量 (Feature Vector)代表某一状态:

x(s)=[x1(s), x2(s), x3(s)…xn(s)]T

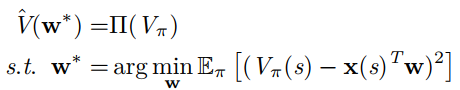


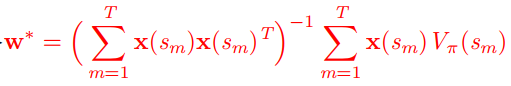
均方误差(MSE):逼近函数和真实价值函数之间的误差(期望损失) d智体在π的状态分布

对线性逼近器, 用样本近似期望损失:



Π 代表从f(x)到 g(x|w)最佳匹配。满足:

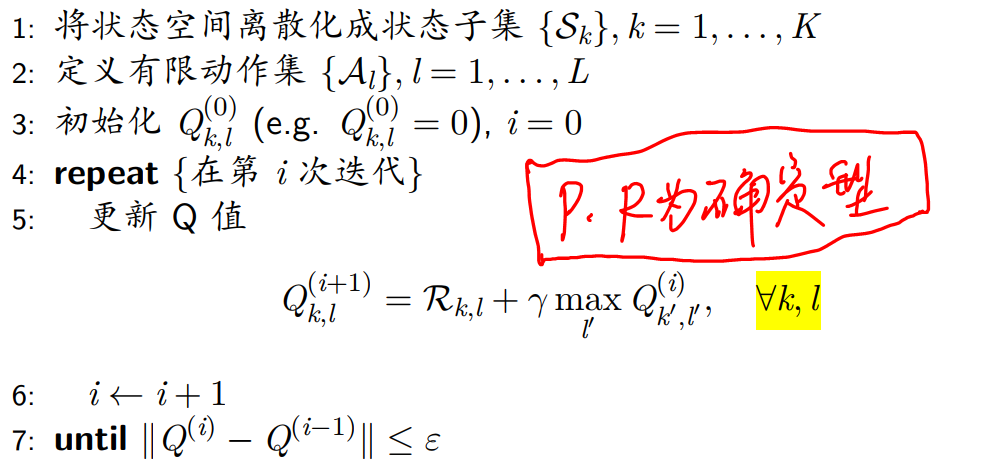


最小二乘法优点: 计算过程简单, 一次性求解最佳权重.求逆,O(n3);当特征空间比较大时, 计算量大, 矩阵可能不满秩, 易有误差

**常见的特征表示方法**

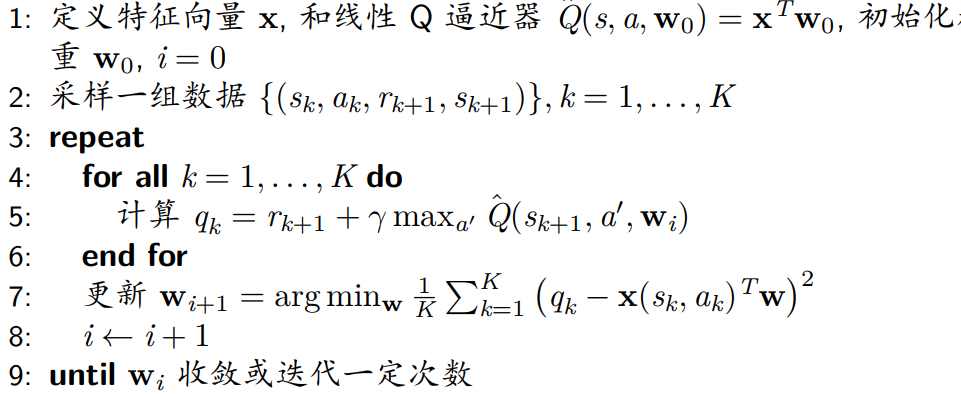
**查表法**;**离散化法**;**粗糙编码**:圆形区域特征可以重叠.**径向基函数**泛化特征值可以是 [0,1], 代表属于该特征的程度,

**价值迭代 + 离散化方法**



收敛性;离散化程度越高, 结果越精确,但是存储空间和计算量增加；维数灾(指数增)

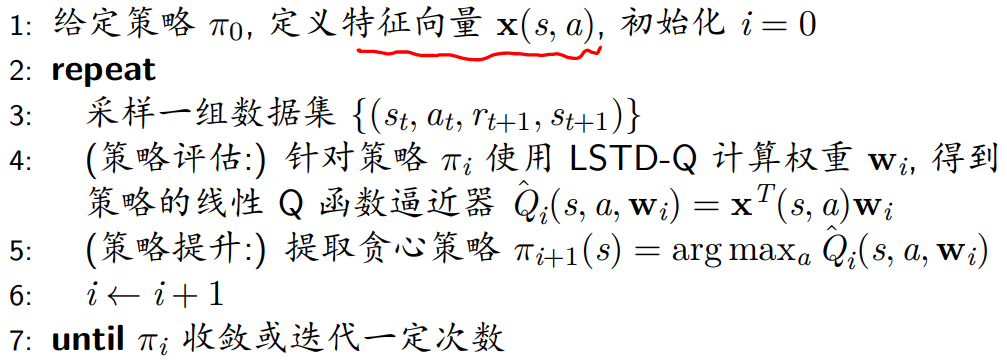
**Fitted Q Iteration（Q\_learning）**

Least-Squared FQI

**策略迭代 + 最小二乘（闭解）**

**根据pi采样?**如果π是确定策略，需要探索的动作,提泛化能力，如果π是随机策略,并且 π(a|s) > 0，可以at ∼ π(st)

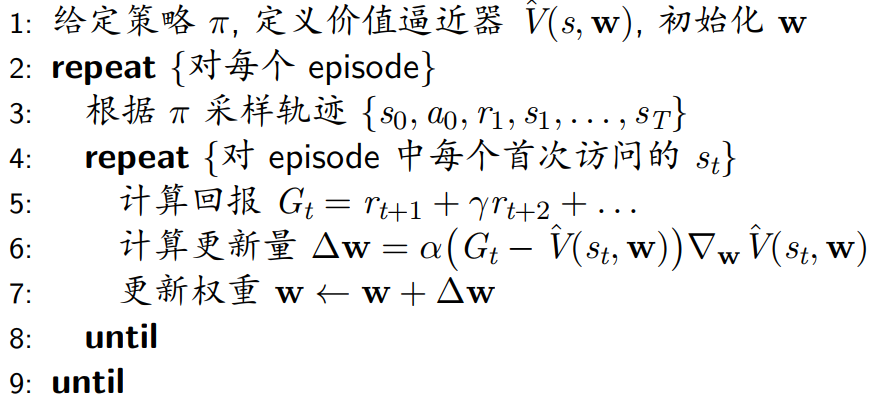
**最小二乘策略迭代 LSPI：**



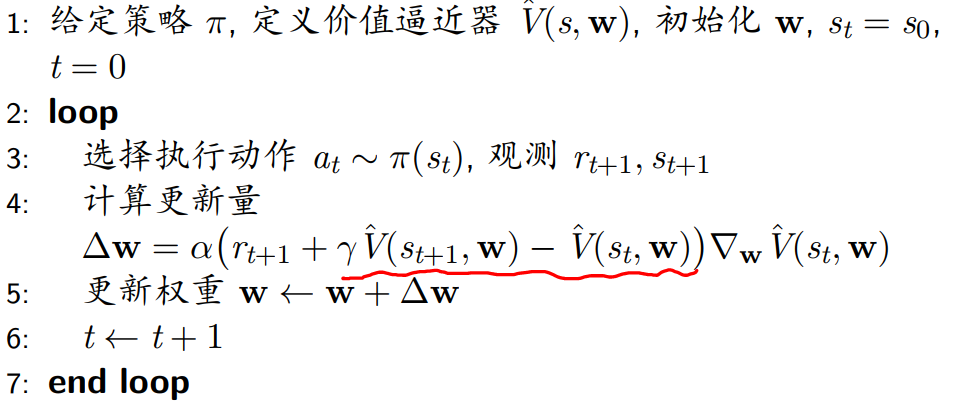
**approximate VI/PI 优缺点**

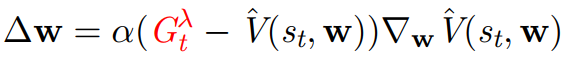
优点:收敛;缺点:每次迭代的计算量大,线性最小二乘法求样本特征矩阵的逆;通常 PI 的迭代次数要比 VI 迭代次数少

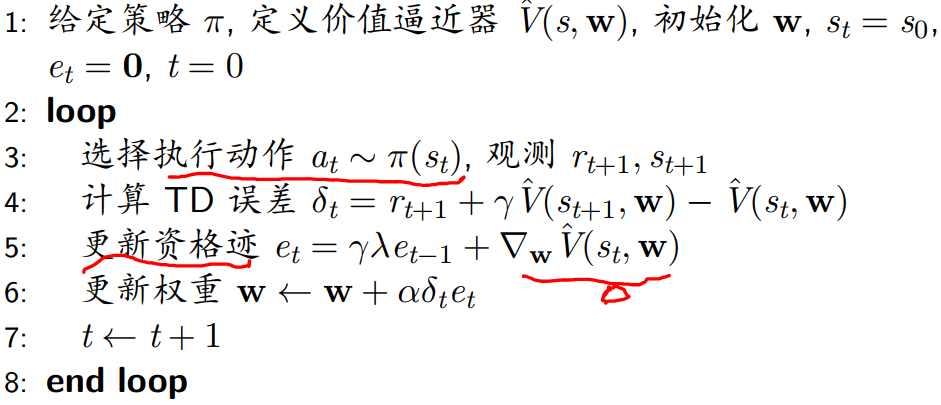
**预测学习 + 随机梯度下降法**

**梯度下降 MC 预测算法**

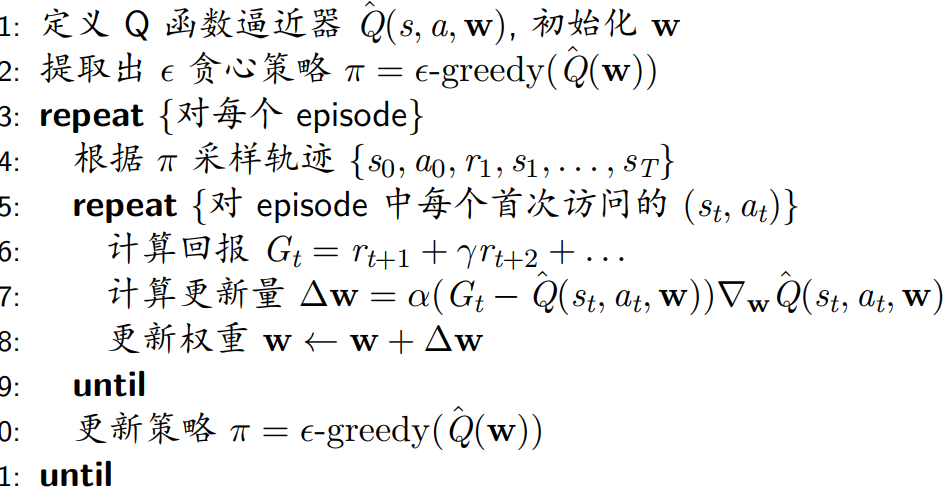
梯度下降 TD 预测学习，



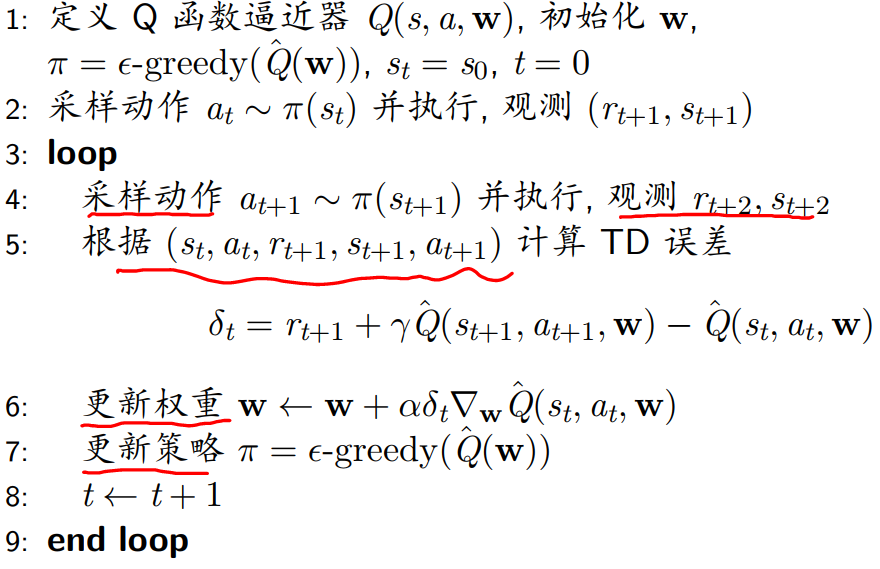
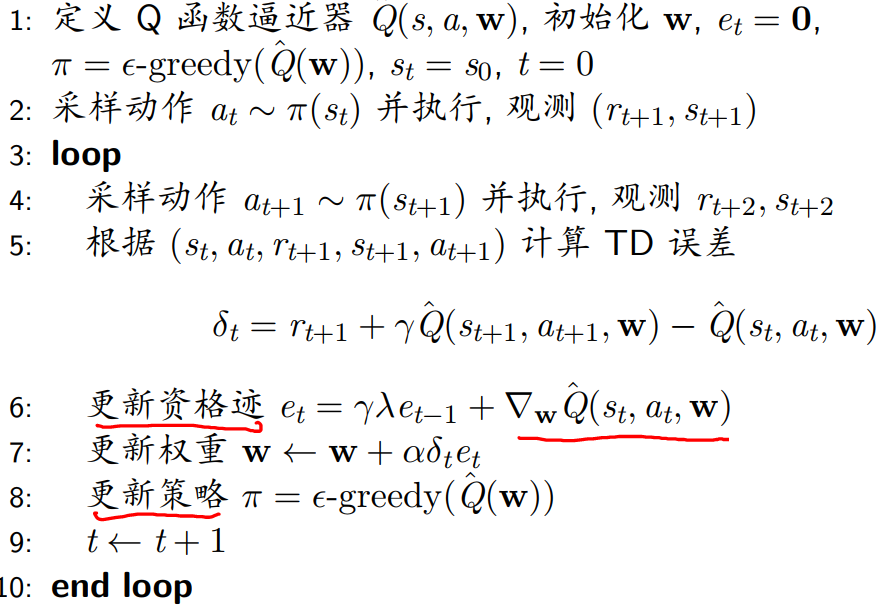
对于前向 TD(λ)：上式为**梯度下降 TD(λ)（后向）算法：**

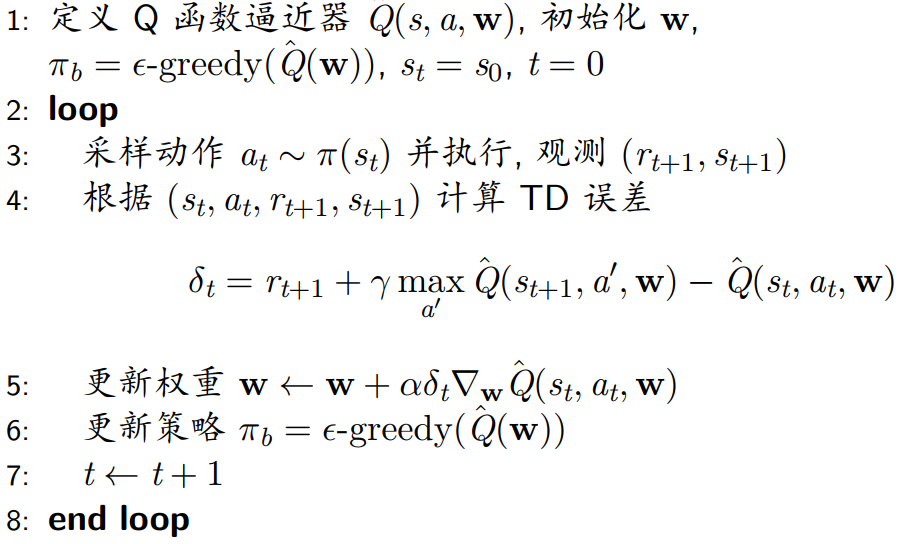


**梯度下降 MC 控制算法**



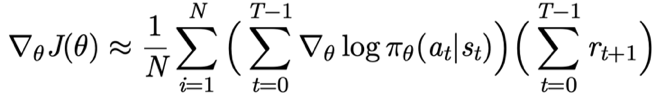
**梯度下降 Sarsa 算法**

**梯度下降 Sarsa(λ) 算法****梯度下降的 Q 学习算法**

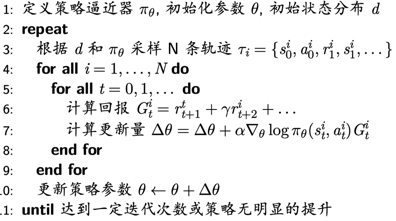


**策略梯度**

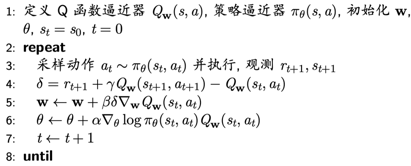
**策略RL好处:**更好的收敛性;有效解决大规模动作集或连续动作空间问题;能够学习随机策略。缺点:收敛到局部最优;策略评估费力且方差大.**策略逼近器不同形式:** p(a|s)=pi(a|s,)，a=pi(s,),.**不同优化目标。优化问题**:无梯度优化 (爬山法);梯度优化.**有限差分法(FD):**对策略参数每个维度增加微小扰动计算.

**解析法求策略梯度:**采样N条轨迹,用样本近似策略梯度: 

**REINFORCE算法（蒙特卡洛策略梯度算法）:**利用时间连贯性改变解析法的:t时刻之前的奖励与t时刻的策略无关，所以不影响t时刻的梯度更新;同时将上述扩展到连续运行场景。实际算法不可能产生无穷长的轨迹,截断计算。

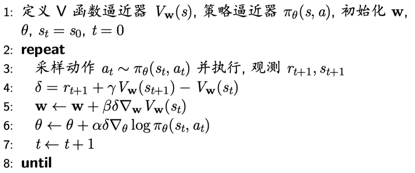
****

**Actor-Critic:** Q函数逼近器为Critic,策略逼近器为Actor.Critic更新价值函数的权重;Actor利用Q(降低方差)更新策略梯度。**Actor**的策略梯度基于Critic定，故让大一些,Critic学的快一些.



**AC算法Q和pi满足条件则收**: Q和pi兼容；Q的权重w等于真实的最小均方误差解

**策略梯度引入基准:**在回报上减去一个baseline不改变梯度. b值可以方差降到最低。对于episodic:b=.连续b()=。**优势函数:** .



**自然梯度:**最速上升方向不再根据欧式距离决定，而根据J的策略概率分布决定.KL散度在概率分布空间上描述两个分布的距离.

**确定型AC**:以上算法为有限动作集MDPs即a~.本算法的策略则是确定性的a=.动作的好坏又Q(s,)反应.

