第四课 价值函数近似 下

1. 线性vs非线性值函数$approximator$
    - 线性估计模型在正确的（好的）特征下具有较好的效果
    - 线性模型需要对特征进行人为设计
    - 需要一个更加丰富的估计模型来在无需对特征进行设计的情况下就能学习
    - 非线性估计器：$deep\ neural\ network(DNN)$
2. $deep\ reinforcement\ learning$
    - DNN用于估计值函数，策略和$world\ model$
    - 代价用随机梯度下降进行优化
    - 挑战
        - 效率低：太多要优化的参数
        - 训练中不稳定和发散的死亡三角：非线性FA，自举法（$bootstrapping$），$off\ policy\ training$
3. $deep\ Q\ network(DQN)$
    - 用神经网络逼近行为值函数
    - 大规模MDP的$Q-learning\ with\ VFA$可能面临两个问题
        - 样本间有相关性（如雅达利游戏中不同样本间可能仅有像素级别的差距）
        - 非静态$target$（就是在$target$中包括了参数$\vec{w}$）
    - 克服以上两个问题的方法
        - $experience\ replay$
        - $fixed\ Q\ target$
4. $experience\ replay$

    每次更新参数$\vec{w}$时，用到的状态序对(s,a,r,s')不是按顺序，而是在全部的序对中随机取样（可能在不同轨迹中）
5. $fixed\ Q\ target$

    更新$\vec{w}$时需要计算$Q\ target$，此时其中带有参数会影响梯度下降的更新。在此，我们延后$target$中参数的变化，在对参数进行多次更新后再更新$target$中的$\vec{w}$