作业 6：无模型方法三—时间差分

陈达贵　深蓝学院

2019-1-11

1　文字作业

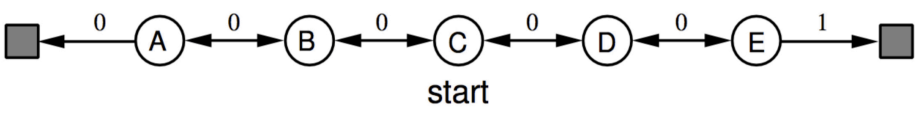
1. (3 分) 上节课我们提到了离策略 TD(0) 优化算法——Q 学习，并解释了为什么 Q 学习不需要重要性采样的原因。本节课我们引入了 n 步 TD 回报值，并给出了 n 步策略评价算法。请回答以下问题：

   (a) 如果使用 n 步 TD 回报值和离策略学习，此时是否需要重要性采样因子？为什么？

   (b) 写出使用 n 步 TD 回报值的对 Q 函数的离策略评价算法，即给定目标策略 π 和行为策略 μ，使用 n 步 TD 回报值去估计 Q 函数

2　编程作业

1. (5 分) 环境定义参考上节课的随机游走环境。



Algorithm 1: n 步 TD 回报值下对 Q 函数的离策略评价算法

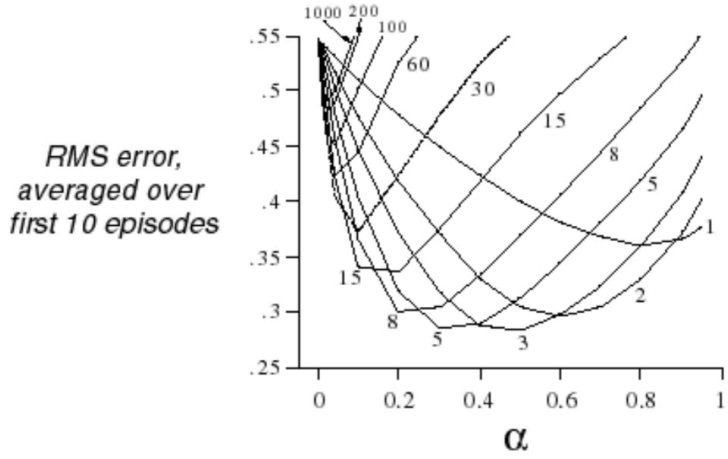
for 每个片段 **do**
  初始化状态 S_0，并存储;
  for $t = 0, 1, 2, \ldots$ **do**
    根据行为策略选择动作 $A_t \sim \mu(\cdot|S_t)$, 并存储;
    观察并存储 R_{t+1}, S_{t+1};
    if S_{t+1} 是终止状态 **then**
      记录片段长度 $T \leftarrow t+1$;
    else
      设置 $T = \infty$;
    end
    $\tau \leftarrow t - n + 1$; /* τ 表示当前需要更新的时间 */
    if $\tau \geq 0$ **then**
      $\rho \leftarrow \prod_{i=\tau+1}^{\min(\tau+n-1, T-1)} \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}$;
      $G \leftarrow \sum_{i=\tau+1}^{\min(\tau+n, T)} \gamma^{i-\tau-1} R_i$;
      if $\tau + n < T$ **then** $G \leftarrow G + \gamma^n Q(S_{\tau+n}, A_{\tau+n})$;
      $Q(S_\tau, A_\tau) \leftarrow Q(S_\tau, A_\tau) + \alpha\rho\left[G - Q(S_\tau, A_\tau)\right]$;
    end
  end
end

(a) 给定随机策略 (左右均 0.5 概率)，采样 10 条轨迹，使用 n 步 TD 算法，计算每个状态的 V 函数。绘制不同 n 下 *RMS 误差*-α 的曲线

(b) 给定随机策略 (左右均 0.5 概率)，采样 10 条轨迹，使用后向视角的 TD(λ) 算法，计算每个状态的 V 函数。绘制不同 λ 下 *RMS 误差*-α 的曲线

(c) 使用 Sarsa(λ) 计算最优 Q 函数和最优策略

(提示: 第一小问中的不同 n 下的 *RMS 误差*-α 的曲线绘制方法可参考下图, 第二小问将不同的 n 替换成不同的 λ 即可)



3 额外任务

1. (2 分) 从下一章开始需要开始逐步进入深度强化学习, 所以要求配置好深度学习框架, 并对其基本使用有所熟悉。下面二选一

(a) Pytorch: 由 Facebook 维护，使用简单，适合学习, 科研与快速实验。(课程推荐)

(b) Tensorflow: 由 Google 维护，使用复杂，学习曲线较陡峭，工业届应用较多。

(提示：安装配置最好参考官网，网络博客中的安装和配置不保证正确且大部分都已经过时，推荐 Anaconda 安装，可以不干扰系统环境。另外有条件的话，可以安装 GPU 版本，需要安装 CUDA(也尽量参考官网))