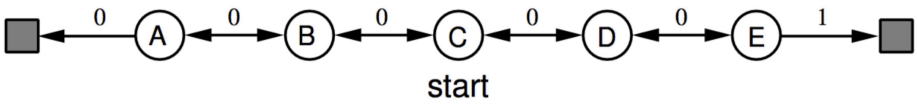
作业 5：无模型方法二——时间差分

陈达贵　深蓝学院

2019-1-4

1. (5 分) **随机游走**。这个作业主要是为了让大家比较 TD(0) 和常量步长 α-MC 的能力。选择的环境是如下的小的 MDP。MDP 的定义如下

   - 状态集 $\mathcal{S}$ 如图所示，每个片段都从中间状态 C 开始。
   - 动作集 $\mathcal{A} = \{\textbf{左}, \textbf{右}\}$
   - 奖励值，当智能体走到最右边的停止状态时，获得 $+1$ 的奖励
   - 状态转移如图所示
   - 衰减系数 $\gamma = 1$



   举个智能体片段的例子：（状态-奖励序列）

$$C, 0, B, 0, C, 0, D, 0, E, 1$$

   给定一个随机策略 π，即往左和往右的概率分别为 0.5。要求评价该策略，即求 $v_\pi(s)$。由于这是一个无衰减的问题，因此某个状态的值函数即为从这个状态开始，最终停在最右边的概率, 即 A, B, C, D, E 的值函数分别为 $\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}$

设定初始值函数 $V(s) = 0.5, s = \{A, B, C, D, E\}$, 且 V(左边停止状态) $= 0, V$(右边停止状态) $= 1$(这是一个 trick，一般终止状态的值函数设置为 0，但是在这个问题中通过设置 V(右边停止状态) $= 1$，可以将所有的奖励设置为 0，这样程序会变简单), 要求实现以下程序，并画出相应的图像

(a) 选择 $\alpha = 0.1$，编程实现 TD 策略评价算法，作出当采样的片段数为 $0, 1, 10, 100$ 时的各个状态的 V 函数的值函数图像 (提示: *值函数图像的画法: 横坐标分别是 A，B，C，D，E，然后纵坐标是相应的值函数，将点连起来得到一条折线。可以先画一条真实的值函数折线作为参考，真实的值函数由于分别是 $\frac{1}{6} \sim \frac{5}{6}$，所以是一条斜率为 $\frac{1}{6}$ 的直线，随着片段数的增加，所画出来的值函数折线会逐渐靠近该直线*)

(b) 实现 TD(0) 算法，并设置 α 分别为 0.15, 0.1, 0.05，画出 *RMS 误差-采样的片段数*的曲线

(c) 实现常量步长的 MC 算法，并设置 α 分别为 0.15，0.1，0.05，画出 *RMS 误差-采样的片段数*的曲线

(d) 对比 TD(0) 算法和常量步长的 MC 算法，比较两者的收敛速度

提示： *RMS 误差是 root mean-squared 的简称，即均方根误差，在这里表示*

$$RMS - error = \frac{1}{N} \sum_{n=1}^{N} \sqrt{\frac{1}{5} \sum_{s \in \{A,B,C,D,E\}} [V(s) - v_\pi(s)]^2}$$

其中 N 表示重复实验的次数，如 $N = 100$, $V(s)$ 表示算法估计值，$v_\pi(s)$ 表示真实值函数。 对于每一个片段 $\in [1, 100]$, 重复实验，求出 RMS-error，以 (片段数，RMS-error) 构成曲线中的一个点，连接不同片段数的点，即得到了对应步长 α 下的曲线

2. (5 分) **悬崖步行**这个作业主要是为了对比 Sarsa，和 Q 学习。下图是是一个悬崖步行的环境，目的是从 S 走到 G，所有的白色格子的移动

都会有-1 的奖励, 而且一旦踏入悬崖 (Cliff) 就会收到-100 的奖励，并回到起点。设定 $\gamma = 1$，即回报值无衰减。要求实现以下算法：

(a) 使用 Sarsa，固定 $\varepsilon = 0.1$, 步长 $\alpha = 0.5$，寻找最优策略。共训练 500 个片段，要求绘制以下图像

- 随着训练片段数的增加，片段的回报值变化的曲线
- 训练结束之后，由 Q 函数导出的贪婪策略 (提示：可手绘一个格子迷宫，然后用箭头表示策略)

(b) 使用 Q 学习，固定 $\varepsilon = 0.1$, 步长 $\alpha = 0.5$，寻找最优策略。总共训练 500 个片段，要求绘制和 Sarsa 类似的图像

(c) 对比 Sarsa 和 Q 学习，哪个算法在训练过程中的回报值更大，哪个找出了最优策略，思考为什么?