Homework for "Algorithms for Big-Data Analysis"

Beijing International Center for Mathematical Research
Peking University

February 17, 2024

Note: Please write up your solutions independently. If you get significant help from others, write down the source of references. A formal mathematical proof for all your claims is required.

1. 考虑有限情形的MDP(S, A, P, R, γ)，其中S是有限个离散状态的集合，A是有限个离散动作的集合，R是奖励函数，$\gamma \in (0,1)$是折扣因子。给定时刻t的状态s和动作a，下一时刻转移到状态s'的概率是$P(s' \mid s, a) = P(s_{t+1} = s' \mid s_t = s, a_t = a)$。令$V(s)$为价值函数，定义Bellman 算子$B$：

$$BV(s) = \max_a \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s' \mid s, a) V(s') \right\}, \quad \forall s \in S.$$

   (a) 证明算子B是压缩映射，即：

$$\|BV - BV'\|_\infty \le \gamma \|V - V'\|_\infty,$$

   其中$\|V - V'\|_\infty = \max_s |V(s) - V'(s)|$。

   (b) 从V_0开始执行迭代算法：$V_{k+1} = BV_k$。对于任意$k > 0$，证明：

$$\|V_{n+k} - V_n\|_\infty \le \frac{\gamma^n}{1-\gamma} \|V_1 - V_0\|_\infty.$$

2. 令考虑有限步MDP (S, A, s_1, P, R, H)，其中S为状态集合，A为动作集合，s_1为初始状态，P为转移概率矩阵，R为奖励矩阵，H为终止时间，γ为折扣因子。定义$\pi(s,a)$ 是在状态s根据随机策略π 执行动作a的概率。定义$\tau = (s_1, a_1, s_2, \ldots, a_{H-1}, s_H)$ 是从状态s_1出发，执行策略π产生的轨道(状态-动作对)，即$a_t \sim \pi(s_t, \cdot)$。

   (a) 写出轨道τ的概率表达式$D^\pi(\tau)$。

   (b) 定义$\rho(\pi)$是有限步总奖励的均值，即：

$$\rho(\pi) = \mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_t | \pi, s_1 \right].$$

   令$R(\tau)$是在轨道τ获得的总奖励。写出$\rho(\pi)$关于$R(\tau)$的表达式。

(c) 假设$\pi_\theta(s,a)$是参数化之后的策略，其中θ是参数。证明

$$\nabla_\theta \rho\left(\pi_\theta\right)=\mathbb{E}_\tau\left[R(\tau) \nabla_\theta \log \left(D^{\pi_\theta}(\tau)\right)\right]=\mathbb{E}_\tau\left[R(\tau) \sum_{t=1}^{H-1} \nabla_\theta \log \left(\pi_\theta\left(s_t, a_t\right)\right)\right].$$

(d) 给定状态s_t和参数θ，假设$b_t(s_t)$条件独立于π_θ产生的抽样。证明：

$$\mathbb{E}_\tau\left[\sum_{t=1}^{H-1} b_t\left(s_t\right) \frac{\partial}{\partial \theta_j} \log \left(\pi_\theta\left(s_t, a_t\right)\right) | \theta, s_1\right]=0.$$

(e) 给出满足(d)里条件的一种$b_t(s_t)$。