CSE 691 习题

德梅萃 · P. 博赛卡斯（Dimitri P. Bertsekas）著
李宇超（Yuchao Li）译

习题 1 [Ber17, 习题 2.1] 考虑由节点（node）$1,\ldots,6$ 以及连接它们的边（edge）构成的图（graph）如图 1所示。请采用动态规划算法计算节点 $1,\ldots,5$ 到节点 6 的最短路径。采用编程或者手算方式均可。*提示*：在此问题中，阶段数目 N 应当设为多少？每阶段中应当包含哪些状态呢？

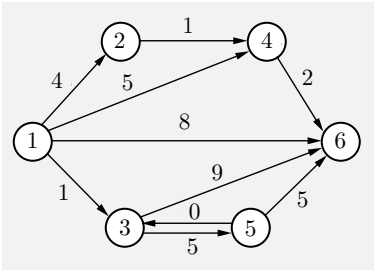

图 1: 习题 1 中涉及的图。标注于边旁的数值表示边长。

习题 2 考虑习题 1 中的最短路径问题。请采用策略前展算法（rollout）给出该问题的近似解。*提示*：可以采用贪心策略作为策略前展中的启发式方法。例如，当处于节点 3 时，可选的下一个节点包括了节点 5 和节点 6。贪心策略比较前往这两个节点的边的长度（即 5 和 9），并选择前往边长较短的后续节点（即对应于边长 5 的节点 5）。

习题 3 [Ber17, 例 3.5.1] 某智力竞赛共有 N 道题目，记作题目 $1,2,\ldots,N$。参赛者可以自由选择其答题次序，当答对题目 i 时，参赛者可以得 R_i 的奖励，并继续回答后续问题。一旦某题目回答错误，参赛者便不可以回答后续问题。小明答对题目 i 的概率为 p_i，那么他应当如何安排答题顺序从而使他期望的收益最大化呢？请采用动态规划给出该问题的解析解。*提示*：在此问题中，什么是状态、控制和系统呢？

习题 4 [Ber22, 习题 1.5] 本习题的目的是通过一维的线性二次型问题

来体现策略迭代与牛顿法的等效性。在此问题中，系统为 $f(x,u)=x+bu$，阶段费用为 $g(x,u)=x^2+ru^2$，其中 $b\neq 0$ 且 $r>0$。

(a) 请验证贝尔曼方程

$$Kx^2=\min_u\left[x^2+ru^2+K(x+bu)^2\right]$$

可以写作等效形式 $H(K)=0$，其中

$$H(K)=K-\frac{rK}{r+b^2K}-1.$$

(b) 现考虑策略迭代算法

$$K_k=\frac{1+rL_k^2}{1-(1+bL_k)^2},$$

其中

$$L_{k+1}=-\frac{bK_k}{r+b^2K_k},$$

且 $\mu(x)=L_0x$ 为起始策略。证明该算法等效于通过牛顿法

$$K_{k+1}=K_k-\left(\frac{\partial H(K_k)}{\partial K}\right)^{-1}H(K_k)$$

求解贝尔曼方程 $H(K)=0$。

(c) 请将上述结论推广到系统为 $f(x,u)=ax+bu$，阶段费用为 $g(x,u)=qx^2+ru^2$，其中 $a,b\neq 0$ 且 $q,r>0$ 的情况。

习题 5 [KG88] 本习题的目的是通过编写代码实现课程中讲解的经典形式的模型预测控制。我们考虑线性系统 $f(x,u)=Ax+Bu$，其中

$$A=\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix},\quad B=\begin{bmatrix}0\\ 1\end{bmatrix}.$$

阶段费用为 $g(x,u)=x'Qx+Ru^2$，其中

$$Q=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},\quad R=1.$$

在此基础上，我们要求状态的每个组分的绝对值小于 5，即状态 x 处于约束集 X 中，且

$$X=\{x\,|\,x=(y,z),\,|y|\leq 5,\,|z|\leq 5\}.$$

控制约束不随状态改变，即 $U(x) = U$，且

$$U = \{u \,|\, |u| \leq 1\}.$$

针对上述问题，请编程实现讲义中式 (1.72)-(1.75) 所表示的经典形式的模型预测控制。其中参数 ℓ 可以设为 5。通过随机生成属于约束集 X 的初始状态 x_0 测试所得模型预测控制的性能。另外，也可以尝试扩大或减小 ℓ 的取值，探索其对于所得策略性能的影响。

习题 6 [Ber17, 习题 3.24，哈代定理] 令 $\{a_1, \ldots, a_n\}$ 和 $\{b_1, \ldots, b_n\}$ 表示实数数列。我们希望给每个 i 分配不同的 j_i，从而最大化 $\sum_{i=1}^{n} a_i b_{j_i}$。

(a) 请采用动态规划给出该问题的解析解。*提示*：在此问题中，什么是状态、控制和系统呢?

(b) 尽管该问题有解析解,但接下来我们将通过编程实现策略前展(rollout)算法，从而给出该问题的近似解，并将其与解析解给出的最优做比较，以便观察最优解、基本启发式（base heuristic）给出的解、以及前展策略（rollout policy）给出的解的关系。随机生成实数序列 $\{a_1, \ldots, a_n\}$ 和 $\{b_1, \ldots, b_n\}$，采用贪心策略作为基本启发式，实现策略前展算法。测试 $n = 10, 50, 100$ 时对应的问题，并比较最优解、基本启发式给出的解、以及前展策略给出的解的关系。

(c) 尝试采用其他的基本启发式并采用策略前展算法求解 $n = 10, 50, 100$ 时对应的问题。

习题 7 假设我们采用 20 步前瞻最小化（20-step lookahead minimization）求解某问题，此时的前瞻最小化就对应于在图 2所示的树状图中，找出从最上端的节点（对应于当前状态 x_0）到最下端一层的最短路径，且最下端的节点 x_{20} 还包含终止费用 $\tilde{J}(x_{20})$。除去终止费用外，其余各边的长度均为 0。图中第 20 层共有 21 个节点，其终止费用从左到右依次为 $3, 5, 4, 2, 0, -2, -5, -3, -1, 1, 0, 2, 3, 4, 5, 4, 3, 2, 1, 3, 2$。请通过编程实现增量策略前展（incremental rollout）从而给出该最短路径问题的近似解。基本策略可以是固定选取某一侧的边，也可以是交替选取一侧的边。尝试采用不同的 δ 取值并比较该算法的性能。

习题 8 [Ber17, 第 3.4 节] 考虑涉及 $N = 5$ 阶段的资产出售问题。在该问题中，我们根据买方出价来决定是否出售资产，从而最大化换算到 $N = 5$ 这一的阶段预期收益。该问题中的初始阶段的状态空间为 $X_0 = \{0\}$，后续

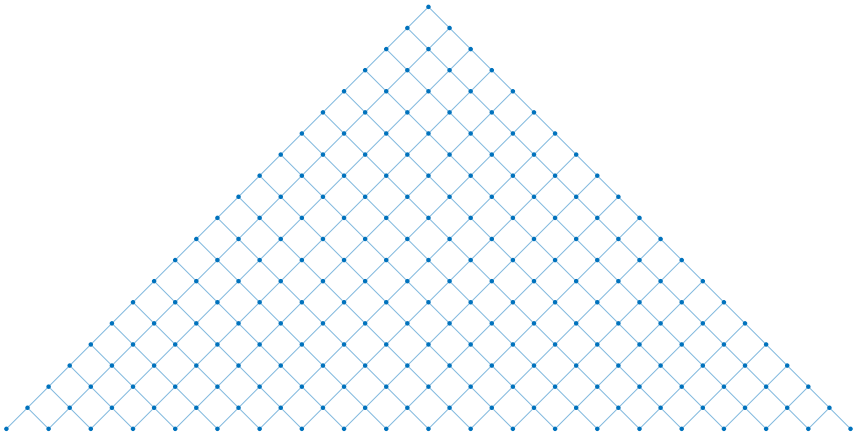

图 2: 习题 7 中涉及的图。该树状图中，$\ell = 20$。最上端的节点为当前状态 x_0，最下一层含有 21 个节点，代表了不同的 x_{20}。

阶段的状态空间为 $X_k = \{8, 9, 10, T\}$，$k = 1, 2, \ldots, N$，其中 T 表示资产已经出售。对于所有阶段，我们都有两个控制选项：$u = 0$ 代表接受上一阶段的报价出售资产，以及 $u = 1$ 代表拒绝已有报价等待本阶段新的随机报价。我们将 k 阶段收到的新报价记为 $w_k \in \{8, 9, 10\}$。那么对于 $x_k \neq T$，系统的状态函数 f_k 为

$$x_{k+1} = \begin{cases} w_k & \text{如果 } u = 1, \\ T & \text{如果 } u = 0. \end{cases}$$

当处于阶段 N 时，对于 $x_N \neq T$，终止收益为 $g_N(x_N) = x_N$。对于 $k \neq N$ 以及 $x_k \neq T$，每阶段收益为

$$g_k(x_k, u_k) = \begin{cases} 0 & \text{如果 } u = 1, \\ (1+r)^{N-k} x_k & \text{如果 } u = 0. \end{cases}$$

其中 $r = 0.1$ 表示利率。显然一旦处于 T，后续阶段也将处于 T，并且不再有收益。我们的目标是找出 $\pi = \{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$ 从而最大化

$$\underset{\substack{w_k \\ k=0,1,\ldots,N-1}}{E} \left\{ g_N(x_N) + \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k), w_k\big) \right\}.$$

假设在 $k = 0, 1, \ldots, N-2$ 阶段时，w_k 为 $8, 9, 10$ 的概率分别为 $0.3, 0.5, 0.2$，而 w_{N-1} 取 $8, 9, 10$ 的概率则为 $0.5, 0.2, 0.3$。请编程求解如下问题。

(a) 请采用动态规划给出该问题的最优解。

(b) 假定基本策略 π 为 $x_k > 8$ 即选择接受报价，请编程计算相应的一步前瞻策略前展策略 $\tilde{\pi}$。请通过解析计算以及采样平均两种方式来计算 $J_{\pi,k}$。样本数可设定为 50。比较其与最优解的关系。

(c) 基于 (b) 中的基本策略 π，请采用讲义中例 2.7.4 的方法，运用蒙特卡罗树搜索（MCTS）计算一步前瞻的策略。样本总量为 20。比较其与最优解的关系。*提示*：为了在探索项 R 中将常数 c 设为 $c = \sqrt{2}$，Q-因子的取值需要缩放到 $[0, 1]$ 的范围内。

参考文献

[Ber17] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, 4 edition, 2017.

[Ber22] Dimitri P. Bertsekas. 策略前展、策略迭代与分布式强化学习 (*Rollout, Policy Iteration, and Distributed Reinforcement Learning*). 清华大学出版社, 2022.

[KG88] S. Sathiya Keerthi and Elmer G. Gilbert. Optimal infinite-horizon feedback laws for a general class of constrained discrete-time systems: Stability and moving-horizon approximations. *Journal of Optimization Theory and Applications*, 57(2):265–293, 1988.