多智体强化学习: 策略前展与策略迭代

译自德梅萃 · P. 博赛卡斯（Dimitri P. Bertsekas）讲座

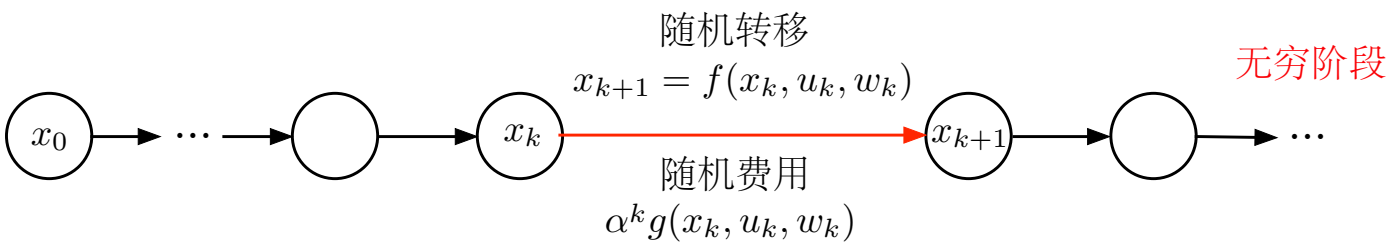
李宇超

yuchao@kth.se

2021 年 6 月 20 日

目录

研究问题：多智体有限状态无穷阶段问题



时不变系统与阶段费用

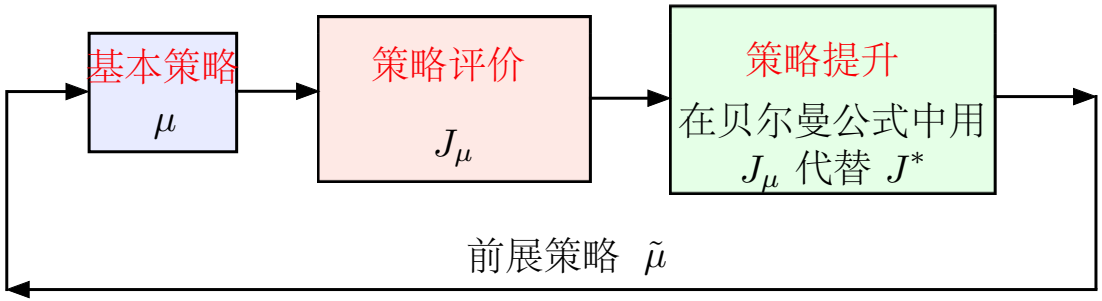
- 系统方程$x_{k+1} = f(x_k, u_k, w_k)$，状态 x_k，控制 $u_k \in U(x_k)$，w_k 随机扰动
- 控制$u = (u_1, \ldots, u_m)$由 m 个元素构成，$U(x) = U_1(x) \times \cdots \times U_m(x)$
- 策略$\mu = (\mu_1, \ldots, \mu_m)$由为从 X 到 U 的映射，且 $\mu_i(x) \in U_i(x)$
- 阶段 k 的费用$\alpha^k g\big(x_k, \mu_1(x_k), \ldots, \mu_m(x_k), w_k\big)$；$\alpha \in (0, 1]$ 为折扣率
- 策略 μ 的费用函数为

$$J_\mu(x_0) = \lim_{N\to\infty} E_{w_k}\left\{ \sum_{k=0}^{N-1} \alpha^k g\big(x_k, \mu_1(x_k), \ldots, \mu_m(x_k), w_k\big) \right\}$$

- 最优费用函数$J^*(x) = \min_\mu J_\mu(x)$
- 最优条件：最小化贝尔曼方程右侧

$$\mu^*(x) \in \arg\min_{(u_1,\ldots,u_m)} E_w\Big\{ g(x, u_1, \ldots, u_m, w) + \alpha J^*\big(f(x, u_1, \ldots, u_m, w)\big) \Big\}$$

策略迭代算法



$$\tilde{\mu}(x) \in \arg\min_{(u_1,\ldots,u_m)} E_w\Big\{g(x,u_1,\ldots,u_m,w) + \alpha J_\mu\big(f(x,u_1,\ldots,u_m,w)\big)\Big\}$$

策略提升的根本特性

$$J_{\tilde{\mu}}(x) \leqslant J_\mu(x), \quad \text{对所有} x$$

策略前展算法即实行一步策略迭代

在 J_μ（近似）已知时，策略前展算法可以通过利用仿真在线执行

策略前展算法另一主要优势：鲁棒性

如果在线执行，则可以通过在线再规划来适应问题参数变化

当 $u = (u_1, \ldots, u_m)$ 时一种新的策略提升方法

标准形式的策略前展算法

$$(\tilde{\mu}_1(x), \ldots, \tilde{\mu}_m(x)) \in \arg \min_{(u_1,\ldots,u_m)} E_w\Big\{ g(x, u_1, \ldots, u_m, w) + \alpha J_\mu\big(f(x, u_1, \ldots, u_m, w)\big) \Big\}$$

其搜索空间随 m 呈指数增长

多智体策略前展算法

连续执行 m 次最小化运算，一次针对一个智体

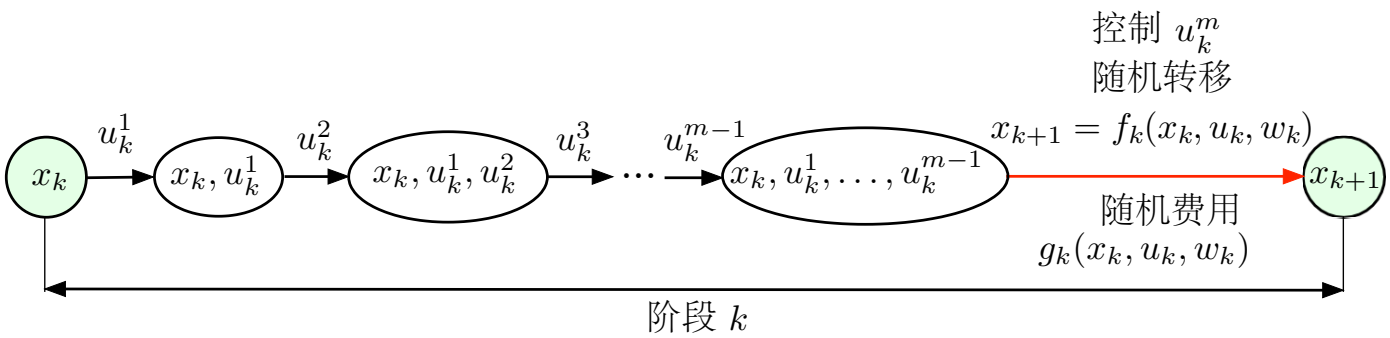
$$\tilde{\mu}_1(x) \in \arg\min_{u_1} E_w\Big\{ g(x, u_1, \mu_2(x) \ldots, \mu_m(x), w) + \alpha J_\mu\big(f(x, u_1, \mu_2(x) \ldots, \mu_m(x), w)\big) \Big\}$$

$$\tilde{\mu}_2(x) \in \arg\min_{u_2} E_w\Big\{ g(x, \tilde{\mu}_1(x), u_2, \mu_3(x) \ldots, \mu_m(x), w) + \alpha J_\mu\big(f(x, \tilde{\mu}_1(x), u_2, \mu_3(x) \ldots, \mu_m(x), w)\big) \Big\}$$

$$\tilde{\mu}_m(x) \in \arg\min_{u_m} E_w\Big\{ g(x, \tilde{\mu}_1(x), \ldots, \tilde{\mu}_{m-1}(x), u_m, w) + \alpha J_\mu\big(f(x, \tilde{\mu}_1(x), \ldots, \tilde{\mu}_{m-1}(x), u_m, w)\big) \Big\}$$

- 搜索空间随 m 呈线性增长
- 巨大的提速！

内在理论：以状态复杂度提升换取控制复杂度降低



原问题的等效问题：将控制元素逐个“展开”

- 以添加 $m-1$ 个额外状态为代价简化控制空间，这些状态对应的费用函数

$$J^1(x_k, u_k^1),\ J^2(x_k, u_k^1, u_k^2), \ldots, J^{m-1}(x_k, u_k^1, \ldots, u_k^{m-1})$$

- 多智体（一次针对一个智体的）策略前展即针对等效问题的标准策略前展
- 状态空间变大不会对策略前展产生负面影响
- 关键的理论事实：费用提升特性保持不变
- 复杂度降低：一步前瞻的分支数量从 n^m 减小到 nm，其中 n 是每个元素 u_k^i 可选的数目

参考文献

- 文献 Bertsekas, Dimitri. "Multiagent reinforcement learning: Rollout and policy iteration." IEEE/CAA Journal of Automatica Sinica 8.2 (2021): 249-272.
- 专著 Bertsekas, Dimitri. P., and John N. Tsitsiklis. "Neuro-dynamic programming. 1996." Athena Scientific (1996).
- 媒体"New algorithm makes it easier for computers to solve decision making problems". EurekAlert! AAAS（美国科学促进会）. 链接