Linear Programming for Markov Decision Processes

Zhengyu Jin

Department of Computer Science and Technology
Zhejiang University

2025 年 12 月 10 日

Overview

1. The Construction of LP for MDP

问题背景

折扣马尔可夫决策过程 (Discounted MDP)

考虑由五元组 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \sigma, \gamma)$ 描述的折扣 MDP：

- $\mathcal{S}$: 状态空间
- $\mathcal{A}$: 动作空间
- $\mathcal{P}$: 转移概率
- σ: 奖励函数
- $\gamma \in (0, 1)$: 折扣因子

目标

根据动态规划理论，向量 v^* 是 MDP 的最优价值函数，当且仅当它满足 Bellman 方程

Bellman 方程

Bellman 方程的形式

$$v^*(i) = \max_{a \in \mathcal{A}} \left\{ \gamma \sum_{j \in \mathcal{S}} P_a(i,j) v^*(j) + \sum_{j \in \mathcal{S}} P_a(i,j) r_{ija} \right\}, \quad i \in \mathcal{S} \tag{1}$$

关键性质

- 当 $\gamma \in (0,1)$ 时，Bellman 方程有唯一的不动点解 v^*，它等于 MDP 的最优价值函数
- 策略 π^* 是 MDP 的最优策略，当且仅当它在 Bellman 方程中达到最小化
- 这是一个**非线性**的不动点方程系统

从 Bellman 方程到线性规划

关键观察

Bellman 方程等价于以下 $|\mathcal{S}| \times (|\mathcal{S}||\mathcal{A}|)$ 的线性规划问题：

原问题 (Primal LP)

$$\begin{aligned} \text{minimize} \quad & \xi^T v \\ \text{subject to} \quad & (\mathbf{I} - \gamma P_a) v - r_a \geq 0, \quad a \in \mathcal{A} \end{aligned} \tag{2}$$

其中：

- ξ 是具有正元素的任意向量
- $P_a \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ 是转移矩阵，(i,j) 元素等于 $P_a(i,j)$
- $\mathbf{I}$ 是 $|\mathcal{S}| \times |\mathcal{S}|$ 的单位矩阵
- $r_a(i) = \sum_{j \in \mathcal{S}} P_a(i,j) r_{ija}$, $i \in \mathcal{S}$ 是动作 a 下的期望状态转移奖励

对偶线性规划

对偶问题 (Dual LP)

原问题 (2) 的对偶线性规划为：

$$\begin{aligned} \text{maximize} \quad & \sum_{a \in \mathcal{A}} \lambda_a^T r_a \\ \text{subject to} \quad & \sum_{a \in \mathcal{A}} (\mathbf{I} - \gamma P_a^T) \lambda_a = \xi, \quad \lambda_a \geq 0, a \in \mathcal{A} \end{aligned} \tag{3}$$

价值-策略对偶性

定理 (Value-Policy Duality for Discounted MDP)

假设折扣奖励无限时域 *MDP* 元组 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ 有唯一的最优策略 π^*。则 (v^*, λ^*) 是原问题和对偶问题 *(2), (3)* 的唯一解对，当且仅当

$$v^* = (\mathbf{I} - \gamma P_{\pi^*})^{-1} r_{\pi^*}, \quad \left(\lambda^*_{\pi^*(i),i}\right)_{i\in\mathcal{S}} = (\mathbf{I} - \gamma (P_{\pi^*})^T)^{-1}\xi, \quad \lambda^*_{a,i} = 0 \ \textit{if}\ a \neq \pi^*(i).$$

价值-策略对偶性

证明.

证明基于线性规划的基本性质，即 (v^*, λ^*) 是原问题和对偶问题的最优解对当且仅当：

(a) v^* 是原问题可行的，即 $(\mathbf{I} - \gamma P_a)v^* - r_a \geq 0$ 对所有 $a \in \mathcal{A}$ 成立

(b) λ^* 是对偶问题可行的，即 $\sum_{a \in \mathcal{A}} (\mathbf{I} - \gamma P_a^T)\lambda_a^* = \xi$ 且 $\lambda_a^* \geq 0$ 对所有 $a \in \mathcal{A}$ 成立

(c) (v^*, λ^*) 满足互补松弛条件：

$$\lambda_{a,i}^* \cdot (v_i^* - \gamma P_{a,i} v^* - r_{a,i}) = 0 \quad \forall i \in \mathcal{S}, a \in \mathcal{A}$$

其中 $\lambda_{a,i}^*$ 是 λ_a^* 的第 i 个元素，$P_{a,i}$ 是 P_a 的第 i 行

价值-策略对偶性

证明（续）.

假设 (v^*, λ^*) 是原-对偶最优解。因此它满足 (a), (b), (c)，且 v^* 是最优价值向量。

- 由最优价值函数的定义，我们知道 $v_i^* - \gamma P_{\pi^*(i),i} v^* - r_{\pi^*(i),i} = 0$
- 由于 π^* 是唯一的，我们有 $v_i^* > \gamma P_{a,i} v^* + r_{a,i}$ 如果 $a \neq \pi^*(i)$
- 因此，最优对偶变量 λ^* 恰好有 $|\mathcal{S}|$ 个非零元素，对应于原问题的 $|\mathcal{S}|$ 个活跃行约束
- 结合对偶可行性关系 $\sum_{a \in \mathcal{A}} (\mathbf{I} - \gamma P_a^T) \lambda_a^* = \xi$，我们得到

$$(\mathbf{I} - \gamma (P_{\pi^*})^T) \left(\lambda^*_{\pi^*(i),i} \right)_{i \in \mathcal{S}} = \xi$$

价值-策略对偶性

证明（续）.

- $(\mathbf{I}-\gamma(P_{\pi^*})^T)$ 是可逆的
- 我们有 $\left(\lambda^*_{\pi^*(i),i}\right)_{i\in\mathcal{S}} = (\mathbf{I}-\gamma(P_{\pi^*})^T)^{-1}\xi$，结合互补松弛条件可推出 λ^* 是唯一的
- 类似地，我们可以从原问题可行性和松弛条件证明 $v^* = (\mathbf{I}-\gamma P_{\pi^*})^{-1} r_{\pi^*}$

□

定理的重要启示

定理 1 表明了最优对偶解 λ^* 和最优策略 π^* 之间的关键对应关系。特别地，可以从 λ^* 的基中恢复最优策略 π^*：

$$\pi^*(i) = a, \quad \text{如果} \lambda^*_{a,i} > 0$$

换句话说，**寻找最优策略等价于寻找最优对偶解的基。**这表明学习最优策略是随机线性规划的一个特例。

2. Saddle Point Formulation for DMDP

What is Saddle Point?

定义 (鞍点 / Saddle Point)

在数学中，**鞍点**或**极小极大点**是函数图像表面上的一个点，在该点处正交方向上的斜率（导数）都为零（临界点），但该点不是函数的局部极值。

鞍点的一个例子是：在一个临界点上，沿一个轴向（两个峰之间）有相对最小值，而沿交叉轴向有相对最大值。

例如，函数 $f(x, y) = x^2 + y^3$ 在点 $(0, 0)$ 处有一个临界点，该点是鞍点，因为它既不是相对最大值也不是相对最小值，且在 y 方向上没有相对最大值或最小值。

鞍点的几何直观

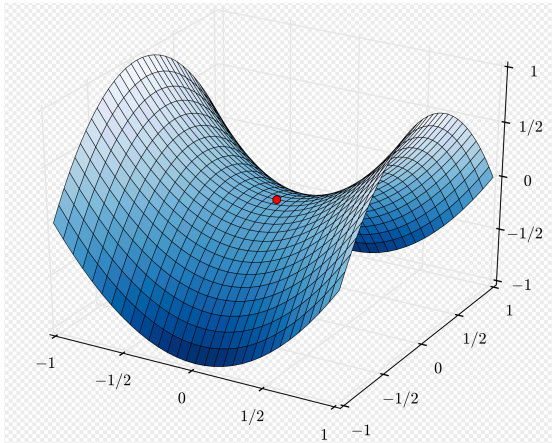

图: 鞍点的几何示意图： $z = x^2 - y^2$

鞍点问题的引入

从 LP 到 Minimax 问题

我们将 LP 问题 (2) 改写为等价的 minimax 问题：

$$\min_{v\in\mathbb{R}^{|\mathcal{S}|}} \max_{\lambda\geq 0} L(v,\lambda) = \xi^T v + \sum_{a\in\mathcal{A}} \lambda_a^T((\gamma P_a - \mathbf{I})v + r_a) \tag{4}$$

变量维度

- 原变量 v 的维度为 $|\mathcal{S}|$
- 对偶变量 $\lambda = (\lambda_a)_{a\in\mathcal{A}} = (\lambda_{a,i})_{a\in\mathcal{A},i\in\mathcal{S}}$ 的维度为 $|\mathcal{S}|\cdot|\mathcal{A}|$
- 每个子向量 $\lambda_a \in \mathbb{R}^{|\mathcal{S}|}$ 是对应于约束不等式 $(\mathbf{I}-\gamma P_a)v - r_a \geq 0$ 的向量乘子
- 每个元素 $\lambda_{a,i} > 0$ 是与 $(\mathbf{I}-\gamma P_a)v - r_a \geq 0$ 的第 i 行相关联的标量乘子

修改的鞍点问题

问题修改的动机

为了开发高效的算法，我们对鞍点问题进行如下修改：

修改后的鞍点问题

$$\min_{v\in\mathbb{R}^{|\mathcal{S}|}} \max_{\lambda\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}} \left\{ L(v,\lambda) = \xi^T v + \sum_{a\in\mathcal{A}} \lambda_a^T((\gamma P_a - \mathbf{I})v + r_a) \right\}, \tag{5}$$

$$\text{subject to} \quad v \in \mathcal{V}, \quad \lambda \in \Xi \cap \Delta$$

约束集的定义

约束集

$$\mathcal{V} = \left\{ v \mid v \geq 0, \|v\|_\infty \leq \frac{\sigma}{1-\gamma} \right\},$$
$$\Xi = \left\{ \lambda \mid \sum_{a \in \mathcal{A}} \lambda_{a,i} \geq \xi_i, \forall i \in \mathcal{S} \right\}$$
$$\Delta = \left\{ \lambda \mid \lambda \geq 0, \|\lambda\|_1 = \frac{\|\xi\|_1}{1-\gamma} \right\} \tag{6}$$

关键引理

我们将在后面证明 v^* 和 λ^* 分别属于 $\mathcal{V}$ 和 $\Xi \cap \Delta$（引理 1）。因此，修改后的鞍点问题 (5) 等价于原问题 (4)。

引理 1：最优解的约束

引理 (最优解属于约束集)

假设 (v^*, λ^*) 是线性规划 *(2), (3)* 的原问题和对偶问题解对。则有：

$$\|v^*\|_\infty \le \frac{\sigma}{1-\gamma}, \quad \|v^*\|_2 \le \frac{\sigma\sqrt{n}}{1-\gamma}, \quad \|\lambda^*\|_2 \le \|\lambda^*\|_1 = \frac{\|\xi\|_1}{1-\gamma}, \quad \xi \le \sum_{a\in\mathcal{A}} \lambda_a^*$$

引理 1 的证明（1/3）

证明.

(i) 注意到 $v^* = r_{\pi^*} + \gamma P_{\pi^*} v^*$。我们有

$$\|v^*\|_\infty \le \|r_{\pi^*}\|_\infty + \gamma \|P_{\pi^*}\|_\infty \|v^*\|_\infty \le \|r_{\pi^*}\|_\infty + \gamma \|v^*\|_\infty$$

因此我们有

$$\|v^*\|_\infty \le \frac{\|r_{\pi^*}\|_\infty}{1-\gamma} \le \frac{\sigma}{1-\gamma}$$

(ii) 我们有 $\|v^*\|_2 \le \sqrt{n}\|v^*\|_\infty \le \frac{\sqrt{n}\sigma}{1-\gamma}$

引理 1 的证明（2/3）

证明（续）.

(iii) 类似地，我们注意到 $\lambda^* = \xi + \gamma(P_{\pi^*})^T\lambda^*$。注意到 $\lambda^* \geq 0$ 且 $\xi \geq 0$，我们有

$$\|\lambda^*\|_1 = e^T\lambda^* = e^T\xi + \gamma e^T(P_{\pi^*})^T\lambda^* = e^T\xi + \gamma e^T\lambda^* = \|\xi\|_1 + \gamma\|\lambda^*\|_1$$

因此我们有 $(1-\gamma)\|\lambda^*\|_1 = \|\xi\|_1$ 从而 $\|\lambda^*\|_2 \leq \|\lambda^*\|_1 = \frac{\|\xi\|_1}{1-\gamma}$

引理 1 的证明（3/3）

证明（续）.

(iv) 我们使用对偶可行性约束得到

$$\left(\sum_{a\in\mathcal{A}}\lambda_a^*\right)(\mathbf{I}-\gamma P_{\pi^*}^T)=\sum_{a\in\mathcal{A}}\lambda_a^*(\mathbf{I}-\gamma P_a^T)=\xi$$

这意味着 $\sum_{a\in\mathcal{A}}\lambda_a^*\geq\xi$ 在元素意义上成立。 □

3. Stochastic Primal-Dual Methods for Reinforcement Learning

3.1 Algorithm Description

MDP 的无模型学习设定

目标

我们希望开发的算法不仅适用于显式给定的 MDP 模型，而且也适用于强化学习。特别地，我们关注**无模型学习设定**（model-free learning setting）。

无模型学习设定

- **已知信息:** 状态空间 $\mathcal{S}$、动作空间 $\mathcal{A}$、奖励上界 σ 和折扣因子 γ（或时域 H）
- **未知信息:** 转移概率 $\mathcal{P}$ 和奖励函数 r
- **采样预言机 (Sampling Oracle, SO):** 接受输入 (i, a) 并生成新状态 j（概率为 $P_a(i, j)$）和随机奖励 $\hat{r}_{ija} \in [0, \sigma]$（期望值为 r_{ija}）

算法说明

基本思想

受价值-策略对偶性（定理 1, 2）的启发，我们开发了一类用于 Bellman 方程鞍点表达式的随机原-对偶方法。

SPD-dMDP 算法特点

- 维护最优价值函数的运行估计（即原解）和最优策略（即对偶解）
- 当从采样预言机中抽取新的状态和奖励观测时，对价值和策略估计进行简单更新
- 原变量更新：基于梯度下降
- 对偶变量更新：基于梯度上升
- 投影步骤：确保对偶变量满足约束 $\Xi \cap \Delta$

SPD-dMDP 算法（1/3）

算法 1: Stochastic Primal-Dual Algorithm for Discounted MDP

输入: Sampling Oracle SO, $n = |\mathcal{S}|$, $m = |\mathcal{A}|$, $\gamma \in (0,1)$, $\sigma \in (0,\infty)$

初始化:

- $v^{(0)} : \mathcal{S} \mapsto \left[0, \frac{\sigma}{1-\gamma}\right]$ 任意初始化
- $\lambda^{(0)} : \mathcal{S} \times \mathcal{A} \mapsto \left[0, \frac{\|\xi\|_1 \sigma}{1-\gamma}\right]$ 任意初始化
- 设 $\xi = \frac{\sigma}{\sqrt{n}} e$

对于 $k = 1, 2, \ldots, T$ **执行:**

- 从 $\mathcal{S}$ 中均匀采样 i
- 从 $\mathcal{A}$ 中均匀采样 a
- 从 SO 中根据 (i, a) 条件采样 j 和 $\hat{r}_{ija}$
- 设 $\beta = \sqrt{n/k}$

SPD-dMDP 算法（2/3）

算法 1（续）：更新步骤

更新原变量:

$$v^{(k)}(i) \leftarrow \max\left\{\min\left\{v^{(k-1)}(i) - \beta\left(\frac{1}{m}\xi(i) - \lambda_a^{(k-1)}(i)\right), \frac{\sigma}{1-\gamma}\right\}, 0\right\}$$

$$v^{(k)}(j) \leftarrow \max\left\{\min\left\{v^{(k-1)}(j) - \gamma\beta\lambda_a^{(k-1)}(i), \frac{\sigma}{1-\gamma}\right\}, 0\right\}$$

$$v^{(k)}(s) \leftarrow v^{(k-1)}(s) \quad \forall s \neq i, j$$

更新对偶变量:

$$\lambda_a^{(k-\frac{1}{2})}(i) \leftarrow \lambda^{(k-1)}(a, i) + \beta\left(\gamma v^{(k-1)}(j) - v^{(k-1)}(i) + \hat{r}_{ija}\right)$$

$$\lambda^{(k-\frac{1}{2})}(a', i') \leftarrow \lambda^{(k-1)}(a', i'), \quad \forall (a', i') \text{ s.t. } a' \neq a \text{ or } i' \neq i$$

SPD-dMDP 算法（3/3）

算法 1（续）：投影和输出

投影对偶变量:

$$\lambda^{(k)} \leftarrow \Pi_{\Xi \cap \Delta} \lambda^{(k-\frac{1}{2})}$$

其中 Ξ 和 Δ 由方程 (6) 给出

输出:

- 平均对偶迭代: $\bar{\lambda} = \frac{1}{T} \sum_{k=1}^{T} \lambda^{(k)}$
- 随机化策略 $\hat{\pi}$，其中:

$$\mathbf{P}(\hat{\pi}(i) = a) = \frac{\bar{\lambda}_a(i)}{\sum_{a \in \mathcal{A}} \bar{\lambda}_a(i)}$$

3.2 Main Results

主要结果

定理 (PAC Duality Gap)

对于任何 $\epsilon > 0$，$\delta \in (0,1)$，令 $\hat{\lambda} = (\hat{\lambda}_a)_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ 为 *SPD-dMDP* 算法 *1* 生成的平均对偶迭代，使用以下每次迭代的样本数量：

$$\Omega\left(\frac{|\mathcal{S}|^3|\mathcal{A}|^2\sigma^4}{(1-\gamma)^4\epsilon^2}\ln\left(\frac{1}{\delta}\right)\right).$$

那么对偶迭代 $\hat{\lambda}$ 满足：

$$\sum_{a\in\mathcal{A}}(\hat{\lambda}_a)^T(v^* - \gamma P_a v^* - r_a) \le \epsilon$$

以至少 $1-\delta$ 的概率成立。

样本复杂度

定理 (PAC Sample Complexity)

对于任何 $\epsilon > 0$，$\delta \in (0, 1)$，令 *SPD-dMDP* 算法 *1* 使用以下每次迭代的样本数量进行迭代：

$$\Omega\left(\frac{|\mathcal{S}|^4|\mathcal{A}|^2\sigma^2}{(1-\gamma)^6\epsilon^2}\ln\left(\frac{1}{\delta}\right)\right),$$

那么输出策略 $\hat{\pi}$ 以至少 $1-\delta$ 的概率是绝对 ϵ-最优的。

精确恢复最优策略

定理 (Exact Recovery of The Optimal Policy)

对于任何 $\epsilon > 0$，$\delta \in (0,1)$，令 *SPD-dMDP* 算法 *1* 使用以下样本数量进行迭代：

$$\Omega\left(\frac{|\mathcal{S}|^4|\mathcal{A}|^4\sigma^2}{d^2(1-\gamma)^4}\ln\left(\frac{1}{\delta}\right)\right).$$

令 $\hat{\pi}^{Tr}$ 通过将随机化策略 $\hat{\pi}$ 舍入到最近的确定性策略获得，即：

$$\hat{\pi}^{Tr}(i) = \mathit{argmax}_{a\in\mathcal{A}}\hat{\lambda}_{a,i}, \quad i \in \mathcal{S}.$$

那么 $\mathbf{P}(\hat{\pi}^{Tr} = \pi^*) \geq 1-\delta$。

References


Chen, Y., & Wang, M. (2016). Stochastic Primal-Dual Methods and Sample Complexity of Reinforcement Learning. arXiv preprint arXiv:1612.02516.

The End