Lecture 9 - Discrete MDPs

Guiliang Liu

The Chinese University of Hong Kong, Shenzhen

DDA4230: Reinforcement Learning
Course Page: [Click]

DDA 4230 Resources

Please join our Slack group.

https://join.slack.com/t/slack-us51977/shared_invite/zt-22g8b40v8-0qSs9o0G3~8hXHwWydlCpw

Please check our course page.

https://guiliang.github.io/courses/cuhk-dda-4230/dda_4230.html

Recap: Discrete-time Markov Decision Process (MDP)

Discrete-time Markov decision process (MDP), denoted as the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho_0, \gamma)$.

- $\mathcal{S}$ the state space;
- $\mathcal{A}$ the action space. $\mathcal{A}$ can depend on the state s for $s \in \mathcal{S}$;
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ the environment transition probability function;
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ the reward function;
- $\rho_0 \in \Delta(\mathcal{S})$ the initial state distribution;
- $\gamma \in [0, 1]$ the discount factor.

Note that $\Delta(\mathcal{X})$ denotes the set of all distributions over set $\mathcal{X}$.

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Recap: Discrete-time Markov Decision Process (MDP)

A stationary MDP follows for $t = 0, 1, \ldots$ as below, starting with $s_0 \sim \rho_0$.

- The agent observes the current state s_t;
- The agent chooses an action $a_t \sim \pi(a_t \mid s_t)$;
- The agent receives the reward $r_t \sim \mathcal{R}(s_t, a_t)$;
- The environment transitions to a subsequent state according to the Markovian dynamics $s_{t+1} \sim \mathcal{T}(s_t, a_t)$.

This process generates the sequence $s_0, a_0, r_0, s_1, \ldots$ indefinitely. The sequence up to time t is defined as the trajectory indexed by t, as $\tau_t = (s_0, a_0, r_0, s_1, \ldots, r_t)$.

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Recap: Discrete-time Markov Decision Process (MDP)

The goal is to optimize the expected discounted cumulative return

$$\mathbb{E}_{s_t,a_t,r_t,t\geq 0}[R_0] = \mathbb{E}_{s_t,a_t,r_t,t\geq 0}\Big[\sum_{t=0}^{\infty} \gamma^t r_t\Big]$$

over the agent's policy π.

Discrete Markov chains

A Markov chain, also known as a homogeneous Markov chain, refers to an infinite process $x_1, \ldots, x_T, \ldots$ where

$$\mathbb{P}(x_{t+1} \mid x_t, \ldots, x_1) = \mathbb{P}(x_{t+1} \mid x_t) = \mathbb{P}_{\mathcal{M}}(x' \mid x)$$

holds almost surely for some probability measure $\mathbb{P}_{\mathcal{M}}$.

Discrete Markov chains

Some key definitions:

- The state space $\mathcal{S}$ is countable, e.g., state space $[n] = [1, 2, \ldots, n]$ is finite.
- The transition probability matrix is $P \in \mathbb{R}^{n \times n}$ where the element $P_{ii'}$ on the i-th and i'-th column equals $\mathbb{P}(x' = i' \mid x = i)$. It is P^{π} in MDPs.
- The state value function is $V(s)$, and the value vector is $V \in \mathbb{R}^n$.
- The reward function is $\mathcal{R}(s)$, and the reward vector is $r \in \mathbb{R}^n$.
- The initial state distribution $\rho_0 \in \mathbb{R}^n$.

The occupancy vector ρ_t, where the i-th element of ρ_t denotes $\mathbb{P}(s_t = i \mid s_0 \sim \rho_0)$, is then $\rho_0 P^t$.

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Discrete Markov chains

When $\mathcal{R}(s)$ is deterministic, r is a deterministic vector. The Bellman equation then writes $V = r + \gamma PV$. Since P is a Markov matrix, $I - \gamma P$ is invertible when $\gamma < 1$ and the value function can be solved by

$$V = (I - \gamma P)^{-1} r\,.$$

Discrete Markov chains

Reducible states. For two states $i, i' \in [n]$, if there exists a T such that $\mathbb{P}(i' \in \{s_1, \ldots, s_T\} \mid s_0 = i) > 0$, we say that i' is accessible from i. If i is accessible from i' and i' is accessible from i, we say that i and i' communicate with each other. If for any $i, i' \in [n]$, i and i' communicate with each other, the Markov chain is irreducible.

- For a Markov chain that is reducible, it is intuitive to partition the chain into irreducible components (likewise, to consider each connected component in a graph). It is therefore sensible to assume that the Markov chain is irreducible.

Discrete Markov chains

Periodicity. For $i \in [n]$, the period of state i is the largest integer d satisfying $\mathbb{P}(s_t \neq i \mid s_0 = i,\ t \neq 0 \mod d)$, or infinity if such a largest integer does not exist. When $d = 1$, state i is aperiodic, and otherwise, state i is periodic with period d.

- In an irreducible Markov chain, all states have the same period. An irreducible Markov chain is aperiodic if all states are aperiodic.
- Mathematically, a chain is aperiodic if and only if P^t contains only positive elements for some positive integer t.

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Discrete Markov chains

Ergodicity. A Markov chain that is irreducible and aperiodic must be ergodic. We commonly assume a chain to be ergodic without loss of generality.

- For the rest of the course, unless otherwise specified, we assume the Markov chains to be ergodic. In MDPs however, in general, there exist policies such that the chain induced by the policies are not ergodic.

Policy Evaluation With a Known Model

Model-based Policy Evaluation:

- When at least one of P and r is known, the problem is policy evaluation with a known model.
- When both P and r are unknown we can make an effort to estimate a P' such that P and P' are close in some measure of discrepancy (or r', respectively).

If otherwise and we only utilize the access to the environment transition, the method is categorized as model-free policy evaluation.

Policy Evaluation With a Known Model

Policy Evaluation (PE): compute the value function given a fixed policy. When at least one of P and r is known, the problem is policy evaluation with a known model.

- Under discrete state and action spaces,
- When both P and r are known.
- When $\gamma < 1$

The solution is: $V = (I - \gamma P)^{-1} r$.

Model-based Policy Evaluation

When both P and r are unknown and we estimate a P' such that P and P' are close (or r', respectively), the problem is model-based policy evaluation.

Lemma

Assume that $0 \le r \le 1$. Let $\varepsilon \in (0, \frac{1}{1-\gamma})$. There is an absolute constant c such that once one has collected at least:

$$N \geq \frac{\gamma}{(1-\gamma)^4} \frac{n^2 m \log(cnm/\delta)}{\varepsilon^2}$$

samples for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ pair, then we could estimate $\hat{P}$ and $\hat{Q}^\pi$ such that with probability at least $1-\delta$, $\|P(\cdot \mid s,a) - \hat{P}(\cdot \mid s,a)\|_1 \le (1-\gamma)^2 \varepsilon$. For every (s,a) pair, and $\|Q^\pi - \hat{Q}^\pi\|_\infty \le \varepsilon$ for every policy π.

The Bellman Optimality Equation

Some assumptions:

- Let the reward function be deterministic.
- Let the reward is bounded by $[0,1]$ in discrete MDPs.
- Let the reward function $\mathcal{R}(s,a)$ defined by a matrix $r \in \mathbb{R}^{n\times m}$, where the element at the i-th row and the j-th column denotes $\mathcal{R}(i,j)$.
- Let P_j be the transition matrix for the policy that chooses action j at every state.

Recall that in discrete MDPs a value function V is optimal if and only if the Bellman optimality equation:

$$V = r + \max_j \gamma P_j V$$

is satisfied with $P \in \{P_1, \ldots, P_m\}$.

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

The Bellman Optimality Equation

By exhausting the action set under the max operator and numbering the actions from 1 to m, the Bellman optimality equation is formulated into the below linear program:

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & \mathrm{e}^T V \\ \text{subject to} \quad & (I - \gamma P_j)V - r_j \geq 0\,, \quad j = 1, \ldots, m\,, \end{aligned}$$

where e is the all-one vector and $\mathrm{e}^T V$ is a dummy objective. Linear programming is in P and can be solved in $\mathrm{poly}(n, m)$. We consider a problem solved if we can cast it to a linear program. Though, this requires P_i to be known.

The Bellman Optimality Equation

The dual of the above linear program is

$$\begin{aligned} \underset{\lambda_1,\dots,\lambda_m}{\text{maximize}} \quad & \sum_j \lambda_j^T r_j \\ \text{subject to} \quad & \sum_j (I - \gamma P_j^T)\lambda_j = \mathrm{e}\,, \\ & \lambda_j \geq 0, \quad j = 1, \dots, m\,. \end{aligned}$$

The dual formulation optimizes $\lambda_1, \dots, \lambda_m$, which could be regarded as the policy.

The Bellman Optimality Equation

Lemma

There exists an optimal dual solution λ_j^*, $j = 1, \ldots, m$, an optimal deterministic policy $\pi^*(\cdot)$, and the corresponding transition matrix P^*, such that

$$\sum_j \lambda_j^* = (I - \gamma P^{*T})^{-1} \mathrm{e},$$

and the i-th entry of λ_j^* equals to the i-th entry of $\sum_j \lambda_j^*$ if $\pi^*(i) = j$, and zero otherwise.

The Bellman Optimality Equation

Lemma

The ℓ^1-norm $\|\sum_j \lambda_j^*\|_1$ of the dual optimum is exactly $n/(1-\gamma)$.

The Bellman Optimality Equation

Lemma

The stochastic policy $\pi(j \mid i) = \lambda_j'^{(i)} / \sum_{j'} \lambda_{j'}'^{(i)}$ *achieves a value* V' *such that* $e^T V' = \sum_j \lambda_j'^T r_j$.

Question and Answering (Q&A)