RL: Find Optimal Policy

1. Markov Decision Process

return = discounted sum of all rewards

2. The Bellman equation

- return
  - bootstrapping
    - Method 2:

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} + \begin{bmatrix} \gamma v_2 \\ \gamma v_3 \\ \gamma v_4 \\ \gamma v_4 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}$$

$$\mathbf{v} = \mathbf{r} + \gamma \mathbf{P} \mathbf{v}$$

which can be rewritten as

Deterministic situation

- General form

$$v_\pi(s) = \mathbb{E}[R_{t+1}|S_t = s] + \gamma \mathbb{E}[G_{t+1}|S_t = s],$$
$$= \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r + \gamma v_\pi(s')] = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) r + \gamma \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) v_\pi(s'),$$
$$= \sum_a \pi(a|s) \left[\sum_{s'} p(s'|s,a) r + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') \right], \quad \forall s \in \mathcal{S}.$$

mean of immediate rewards / mean of future rewards

Highlights: symbols in this equation

- $v_\pi(s)$ and $v_\pi(s')$ are state values to be calculated. Bootstrapping!
- $\pi(a|s)$ is a given policy. Solving the equation is called policy evaluation.
- $p(s'|s,a)$ and $p(s'|s,a)$ represent the dynamic model. What if the model is known or unknown?

- state value

$$v_\pi(s) = \mathbb{E}[G_t | S_t = s]$$

- policy evaluation
  - closed form实际不用，逆计算量大

$$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

  - iterative solution 从V0开始，迭代收敛

$$v_{k+1} = r_\pi + \gamma P_\pi v_k$$

- action value

$$q_\pi(s,a) = \mathbb{E}[G_t | S_t = s, A_t = a]$$

value iteration & policy iteration

- value iteration
  - 找到
    - step1: policy update
    - step2: value update
- policy iteration
- truncated policy iteration

3. BOE: Bellman Optimal Equation

- Optimal state value
- Optimal policy
  - definition
    - 改变r: reward
      - 所有reward进行线性变化ar+b，policy不改变
    - 改变γ: discount rate
      - γ小，近视
  - analyzing
- BOE

$$v = \max_\pi (r_\pi + \gamma P_\pi v)$$

  - algorithm to solve the problem
    - value iteration algorithm
  - existence
  - uniqueness
  - optimality

$$v^* = r_{\pi^*} + \gamma P_{\pi^*} v^*$$

Contraction Mapping Theorem

Theorem (Contraction Mapping Theorem)

For any equation that has the form of $x = f(x)$, if f is a contraction mapping, then

- *Existence: there exists a fixed point x^* satisfying $f(x^*) = x^*$.*
- *Uniqueness: The fixed point x^* is unique.*
- *Algorithm: Consider a sequence $\{x_k\}$ where $x_{k+1} = f(x_k)$, then $x_k \to x^*$ as $k \to \infty$. Moreover, the convergence rate is exponentially fast.*