Average Reward Reinforcement Learning

Average Reward Problem

$$\max \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}[r(x_k, u_k)]$$

- For simplicity, assume $r(x_k, u_k)$ is deterministic given x_k, u_k.
- The result can be extended to the case where $r(x_k, u_k)$ is a random variable.

Challenge

Consider given policy π such that the Markov chain has a unique stationary distribution γ

$$V_0^\pi(i) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} E[r(x_k, \pi(x_k))|x_0 = i] = \sum_j \gamma_j r(j, \pi(j)),$$

which is independent of the initial state i since the stationary distribution is the same regardless of the initial state. In other words,

$$V_0^\pi(i) = V_0^\pi(j) \quad \text{for any } i \text{ and } j.$$

Relative Value Function: Intuition

Consider a finite-horizon problem with N steps

$$V_0^{(N)}(i) = \max \sum_{k=0}^{N-1} \mathbb{E}[r(x_k, u_k)|x_0 = i].$$

Define

$$V_m^{(N)}(i) = \max \sum_{k=m}^{N-1} \mathbb{E}[r(x_k, u_k)|x_m = i].$$

Relative Value Function: Intuition

Consider a finite-horizon problem with N steps

$$V_0^{(N)}(i) = \max \sum_{k=0}^{N-1} \mathbb{E}[r(x_k, u_k)|x_0 = i].$$

Define

$$V_m^{(N)}(i) = \max \sum_{k=m}^{N-1} \mathbb{E}[r(x_k, u_k)|x_m = i].$$

The Bellman Equation

$$V_0^{(N)}(i) = \max_u \left(r(i,u) + \sum_j P_{ij}(u) V_1^{(N)}(j) \right)$$

Relative Value Function: Intuition

Suppose the average reward converges to V^*. Define $h_N(i)$ and $h_{N-1}(j)$ such that

$$h_N(i) = V_0^{(N)}(i) - NV^*$$
$$h_{N-1}(j) = V_1^{(N)}(j) - (N-1)V^*$$

Substituting them into the Bellman equation, we have

$$NV^* + h_N(i) = \max_u r(i,u) + \sum_j P_{ij}(u)\left((N-1)V^* + h_{N-1}(j)\right)$$

or

$$V^* + h_N(i) = \max_u r(i,u) + \sum_j P_{ij}(u) h_{N-1}(j).$$

Relative Value Function: Intuition

Suppose as $N \to \infty$, $h_N(i) \to h(i)$ and $h_{N-1}(j) \to h(j)$. Then, we have the following Bellman equation for the average reward problem:

$$V^* + h(i) = \max_u r(i,u) + \sum_j P_{ij}(u) h(j).$$

Relative Value Function: Intuition

Suppose as $N \to \infty$, $h_N(i) \to h(i)$ and $h_{N-1}(j) \to h(j)$. Then, we have the following Bellman equation for the average reward problem:

$$V^* + h(i) = \max_u r(i,u) + \sum_j P_{ij}(u) h(j).$$

$h(i)$: relative value function.
Note that $V_0^{(N)}(i) \approx NV^* + h(i)$ for large N.

Average Reward MDPs

Theorem: Optimality

If there exist V^* and h satisfying the above Bellman equation, then the obtained policy π^* from this equation is the optimal stationary policy and V^* is the optimal average reward.

Proof

Rewrite the Bellman equation as

$$\begin{aligned} V^* &= \max_u r(i,u) + \mathbb{E}[h(x_{k+1})|x_k = i, u_k = u] - h(i) \\ &\geq r(i,\pi_k(i)) + \mathbb{E}[h(x_{k+1})|x_k = i, u_k = \pi_k(i)] - h(i) \quad \forall \pi_k \end{aligned}$$

Taking expectation over x_k and π_k on both sides, we have

$$V^* \geq \mathbb{E}[r(x_k, \pi_k(x_k))] + \mathbb{E}[h(x_{k+1})] - \mathbb{E}[h(x_k)] \quad \forall \pi_k.$$

Average Reward MDPs

With the telescoping sum, we obtain

$$NV^* \geq \sum_{k=0}^{N-1} \mathbb{E}[r(x_k, \pi_k(x_k))] + \mathbb{E}[h(x_N)] - \mathbb{E}[h(x_0)]$$

$$\frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}[r(x_k, \pi_k(x_k))] \leq V^* - \frac{1}{N}\mathbb{E}[h(x_N)] + \frac{1}{N}\mathbb{E}[h(x_0)]$$

If h is bounded or x_k takes values in a finite state space,

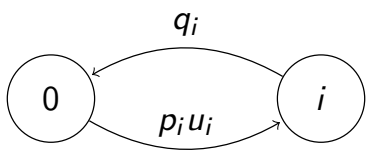
$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{E}[r(x_k, \pi_k(x_k))] \leq V^*$$

which reduces to equality for π^*.

Example

- A crowdsourcing worker is presented with type-i job with probability p_i.
- A job of type i can be completed in a time slot with probability q_i (independent of how long it has been with the worker) to complete.
- The agent receives reward R_i for completing a type-i job.
- When the worker is working on a job, she cannot take on a new job.
- Assume that if a job is accepted in time slot k, it cannot be completed in the same time slot.

Find the optimal strategy to accept jobs to maximize average expected reward.



Example

x_k: state of the system at the beginning of time slot k.

$x_k = 0$ means the worker is idle.

$x_k = i$ means the worker is working on a job of type i.

$$u_i = \begin{cases} 1, & \text{accept a type-}i\text{ job} \\ 0, & \text{reject a type-}i\text{ job} \end{cases}$$

Example

x_k: state of the system at the beginning of time slot k.

$x_k = 0$ means the worker is idle.

$x_k = i$ means the worker is working on a job of type i.

$$u_i = \begin{cases} 1, & \text{accept a type-}i\text{ job} \\ 0, & \text{reject a type-}i\text{ job} \end{cases}$$

The Bellman Equation

$$\begin{cases} V^* + h(0) = \sum_i p_i \max(\underbrace{h(i)}_{\text{accept}}, \underbrace{h(0)}_{\text{reject}}), & \text{for state 0} \\ V^* + h(i) = q_i(R_i + h(0)) + (1 - q_i)h(i), & \text{otherwise} \end{cases}$$

Example

Note that adding a constant c to $h(i)\ \forall i$ does not change the above equations, so we set $h(0) = 0$ without loss of generality.

$$V^* = \sum_i p_i \max(0, h(i))$$

and

$$V^* + \cancel{h(i)} = q_i R_i + (\cancel{1} - q_i)h(i) \Rightarrow V^* = q_i(R_i - h(i))$$

Example

Note that adding a constant c to $h(i)\ \forall i$ does not change the above equations, so we set $h(0) = 0$ without loss of generality.

$$V^* = \sum_i p_i \max(0, h(i))$$

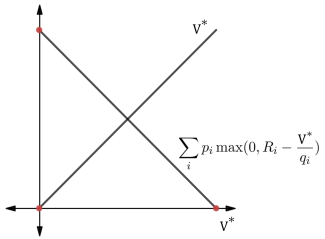
and

$$V^* + \cancel{h(i)} = q_i R_i + (\cancel{1} - q_i)h(i) \Rightarrow V^* = q_i(R_i - h(i))$$

Note that $h(i) = R_i - \frac{V^*}{q_i}$, so we have

$$V^* = \sum_i p_i \max\left\{R_i - \frac{V^*}{q_i}, 0\right\}.$$

Example



An optimal V^* exists since the LHS is $\uparrow$ and RHS is $\downarrow$ in V^*.

$$h(i) = R_i - \frac{V^*}{q_i}$$

Thus the optimal policy is:

$$\begin{cases} \text{accept}, & \text{if } R_i \geq \frac{V^*}{q_i} \\ \text{reject}, & \text{otherwise} \end{cases}.$$

Model-based Algorithm: Relative Value Iteration

Recall the value iteration algorithm:

- Set $V_0(i) = 0 \quad \forall i, \quad k = 0$
- $V_{k+1}(i) = \max_u r(i,u) + \alpha \sum_j P_{ij}(u) V_k(j)$

If we apply the value iteration algorithm for the average reward problem with $\alpha = 1$, then V_k can be thought of as the k-step reward. Thus, V_k will keep increasing and the procedure can be numerically unstable.

Model-based Algorithm: Relative Value Iteration

Let x be an arbitrary state and define

$$\tilde{V}_k(i) = V_k(i) - V_k(x)$$

which is the relative value w.r.t state x. As k increases, if $\frac{V_k(i)}{k} \to V^*$ independent of i, the following procedure will be numerically stable

Relative Value Iteration: Algorithm 1

- Step 1: $\tilde{V}_0(i) = 0 \quad \forall i$
- Step 2: $V_{k+1}(i) = \max_u r(i,u) + \sum_j P_{ij}(u)\tilde{V}_k(j)$
  $\tilde{V}_{k+1}(i) = V_{k+1}(i) - V_{k+1}(x)$
- repeat step 2.

Model-based Algorithm: Relative Value Iteration

A variant of relative value iteration: Algorithm 2

$$V_{k+1}(i) = \max_u r(i,u) + \sum_j P_{ij}(u) V_k(j) - V_k(x)$$

Model-based Algorithm: Policy Iteration

- Fix policy π_k
- Compute V_k and h_k from

$$V_k + h_k(i) = r(i, \pi_k(i)) + \sum_j P_{ij}(\pi_k(i)) h_k(j) \quad \forall i$$

$$h_k(0) = 0$$

- Obtain π_{k+1} from

$$\pi_{k+1}(i) \in \arg\max_u r(i, u) + \sum_j P_{ij}(u) h_k(j)$$

- If $V_{k+1} = V_k$ and $h_{k+1} = h_k$, stop.

Model-Free Algorithm: Average Reward Q-learning

Relative Q-Functions

$$h(i) = \max_u \underbrace{r(i,u) + \sum_j P_{ij}(u)h(j) - V^*}_{Q(i,u) \quad \text{(definition)}} = \max_u Q(i,u)$$

Model-Free Algorithm: Average Reward Q-learning

Relative Q-Functions

$$h(i) = \max_u \underbrace{r(i,u) + \sum_j P_{ij}(u)h(j) - V^*}_{Q(i,u) \quad \text{(definition)}} = \max_u Q(i,u)$$

$$\begin{aligned} Q(i,u) &= r(i,u) + \sum_j P_{ij}(u)h(j) - V^* \\ &= r(i,u) + \sum_j P_{ij}(u) \max_v Q(j,v) - V^* \end{aligned}$$

Relative Value Iteration for Q-Learning

Recall value iteration algorithm (Algorithm 2):

$$\begin{aligned} V_{k+1}(i) &= \max_u r(i,u) + \sum_j P_{ij}(u) V_k(j) - V_k(0) \\ &= \max_u r(i,u) + \sum_j P_{ij}(u) \max_v Q_k(j,v) - \max_v Q_k(0,v) \end{aligned}$$

$$Q_{k+1}(i,u) = r(i,u) + \sum_j P_{ij}(u) \max_v Q_k(j,v) - \max_v Q_k(0,v).$$

Relative Value Iteration for Q-Learning

$$Q_{k+1}(i,u) = (1-\beta_k)Q_k(i,u) + \beta_k \left(r(i,u) + \sum_j P_{ij}(u) \max_v Q_k(j,v) - \max_v Q_k(0,v) \right)$$

or

$$Q_{k+1}(i,u) = Q_k(i,u) + \beta_k \left(r(i,u) + \sum_j P_{ij}(u) \max_v Q_k(j,v) - Q_k(i,u) - \max_v Q_k(0,v) \right)$$

Model-based Algorithm: Policy Iteration

- Fix policy π_k
- Compute V_k and h_k from

$$V_k + h_k(i) = r(i, \pi_k(i)) + \sum_j P_{ij}(\pi_k(i)) h_k(j) \quad \forall i$$

$$h_k(0) = 0$$

- Obtain π_{k+1} from

$$\pi_{k+1}(i) \in \arg\max_u r(i, u) + \sum_j P_{ij}(u) h_k(j)$$

- If $V_{k+1} = V_k$ and $h_{k+1} = h_k$, stop.

Convergence of Policy Iteration

Improvement Theorem

Given V and $h(i)$. Suppose that the stationary policy π satisfies

$$r(i, \pi(i)) + \sum_j P_{ij}(\pi(i))h(j) \geq V + h(i) \quad \forall i.$$

Then the long-run average reward of policy π, denoted by V^π satisfies $V^\pi \geq V$. The strict inequality holds if a strict inequality holds for some i, which is a positive recurrent state under π.

Convergence of Policy Iteration

Proof

Multiplying both sides of the inequality by the stationary distribution y_i^π and summing over i, we have

$$\sum_i y_i^\pi r(i, \pi(i)) + \sum_i y_i^\pi \sum_j P_{ij}(\pi(i))h(j) \geq V + \sum_i y_i^\pi h(i).$$

Note that

$$\sum_i y_i^\pi \sum_j P_{ij}(\pi(i))h(j) = \sum_j h(j) \sum_i y_i^\pi P_{ij}(\pi(i)) = \sum_j h(j) y_j^\pi.$$

Therefore, we have

$$V^\pi \geq V.$$

Average Reward MDPs and Discounted MDPs

Blackwell Optimality

A stationary policy π^* is called Blackwell optimal if it is optimal for all $\alpha \in (\bar{\alpha}, 1)$.

- Such a policy exists for MDPs with finite state and action spaces.
- π^* is optimal for the average reward problem.

Proof

Existence

Assume finite action and state spaces.

- There are only a **finite** number of deterministic policies.
- Given any discounted factor α, there exists a deterministic optimal policy, denoted by π_α.

The two observations above imply there exist a monotonic sequence $\alpha_n \to 1$ and π^* such that $\pi_{\alpha_n}(\cdot) = \pi^*(\cdot)$ for $n \geq \bar{n}$, or

$$Q_{\alpha_n}(i, \pi^*(i)) \geq Q_{\alpha_n}(i, u)$$

for all u and n.

Proof

Existence

Suppose that π^* is not Blackwell optimal. Then for each $_n$ such that $n \geq \bar{n}$, there exists a state, i_n, such that $\pi_{\alpha_n}(i_n) \neq \pi^*(i_n)$, which means

$$Q_{\alpha_n}(i_n, \pi_{\alpha_n}(i_n)) > Q_{\alpha_n}(i_n, \pi^*(i_n)).$$

Since the state and action spaces are both finite, there exists a subsequence of $\{i_n\}$, denoted by $\{i_{n_m}\}$ such that $i_{n_m} = \bar{i}$ for some $\bar{i}$ and $\pi_{\alpha_{n_m}}(i_{n_m}) = \bar{u}$ for some $\bar{i}$ and $\bar{u}$ for all n_m.
In summary, there exists a sequence α_m such that

$$\lim_m \alpha_m = 1 \quad \text{and} \quad Q_{\alpha_m}(\bar{i}, \bar{u}) > Q_{\alpha_m}(\bar{i}, \pi^*(\bar{i})).$$

Proof

Existence

Now define $f(\alpha) = Q_\alpha(\bar{i}, \pi^*(\bar{i}))$ and $g(\alpha) = Q_\alpha(\bar{i}, \bar{u})$. We have two sequences $\{\alpha_n\}$ and $\{\alpha_m\}$ such that

$$f(\alpha_n) \geq g(\alpha_n) \quad \forall n$$
$$f(\alpha_m) < g(\alpha_m) \quad \forall m.$$

In other words, $f(\alpha) - g(\alpha)$ has to cross zero infinitely many times (or has infinitely many roots) over $\alpha \in [0, 1]$. This cannot happen when $f(\alpha) - g(\alpha)$ is a continuous and rational function.

Proof

Average reward optimal

It can be shown that under any stationary policy

$$V_\alpha^\pi(x) = \frac{1}{1-\alpha}\bar{V}^\pi + h^\pi(x) + O(1-\alpha).$$

Let π^* be a Blackwell optimal policy and $\bar{\pi}$ be the average-reward optimal. We have

$$V_\alpha^{\bar{\pi}}(x) \leq V_\alpha^{\pi^*}(x)$$

or

$$\frac{1}{1-\alpha}\bar{V}^{\bar{\pi}} + h^{\bar{\pi}}(x) + O(1-\alpha) \leq \frac{1}{1-\alpha}\bar{V}^{\pi^*} + h^{\pi^*}(x) + O(1-\alpha)$$

or

$$\bar{V}^{\bar{\pi}} + (1-\alpha)h^{\bar{\pi}}(x) + O((1-\alpha)^2) \leq \bar{V}^{\pi^*} + (1-\alpha)h^{\pi^*}(x) + O((1-\alpha)^2)$$

Proof

Average reward optimal

$$\bar{V}^{\bar{\pi}} + (1-\alpha)h^{\bar{\pi}}(x) + o((1-\alpha)^2) \leq \bar{V}^{\pi^*} + (1-\alpha)h^{\pi^*}(x) + o((1-\alpha)^2)$$

By letting $\alpha \to 1$, we have

$$\bar{V}^{\bar{\pi}} \leq \bar{V}^{\pi^*}.$$

Reference

- This lecture is based on R. Srikant's lecture notes on *Average Cost MDPs* available at https://sites.google.com/illinois.edu/mdps-and-rl/lectures?authuser=1
- The proof of Blackwell optimality is based on https://www.dam.brown.edu/people/huiwang/classes/am226/Archive/black.pdf#page=1.00
- The proof of average reward optimal is based on https://ocw.mit.edu/courses/2-997-decision-making-in-large-scale-systems-spring-2004/afcd934c2ae9e684b68a0022c692b9b0_lec_5_v1.pdf