

Introduction to Dynamic Programming (DP)

Infinite Horizon

Outline

- Dynamic Programming (DP)
 - Review of Finite Horizon Problems
 - Infinite Horizon Problems
 - Notation
 - Value Iteration
 - Policy Iteration
- Introduction to Q-function

Review of Finite Horizon Dynamic Programming

$$\min_{\pi} \quad \mathbb{E} \left[c_N(s_N) + \sum_{t=0}^{N-1} c_t(s_t, a_t, w_t) \right]$$

$$\text{s.t.} \quad s_{t+1} = f_t(s_t, a_t, w_t),$$

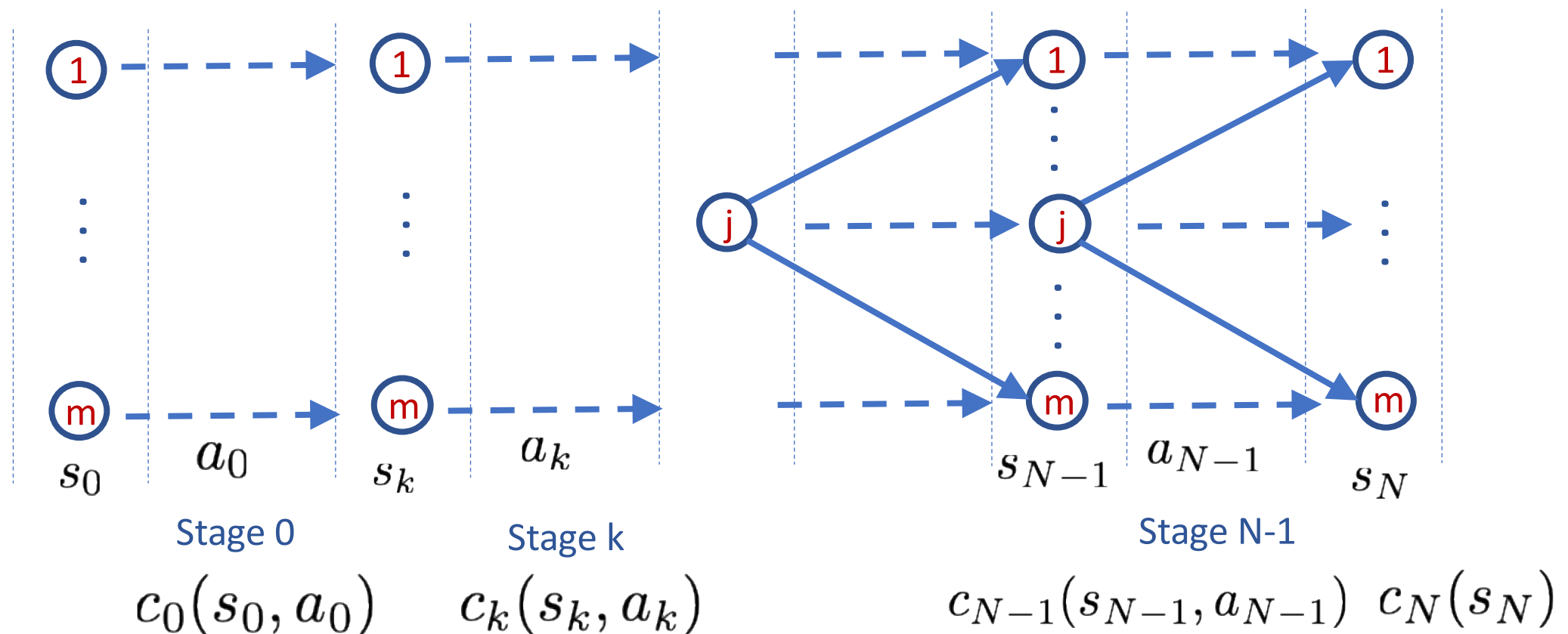
$$a_t = \mu_t(s_t) \in A_t(s_t),$$

$$w_t \sim \mathbb{P}(\cdot \mid s_t, a_t).$$

Solving DPs: Deterministic, Finite-Horizon Case

$$\min_{\pi} \left[c_N(s_N) + \sum_{t=0}^{N-1} c_t(s_t, a_t) \right]$$

$$\text{s.t. } s_{t+1} = f_t(s_t, a_t), \quad a_t = \mu_t(s_t) \in A_t(s_t),$$



Review of Finite Horizon Dynamic Programming

- $V_t(s)$: optimal cost of tail subproblem starting at state s at time t .
- Initialize: $V_N(s) = c_N(s)$
- Backward iteration:

$$V_t(s) = \min_{a \in A_t(s)} \mathbb{E} [c_t(s, a, w) + V_{t+1}(f_t(s, a, w))]$$

- Then $V_0(s_0)$ is the optimal cost and the generated policy is optimal.

The Infinite Horizon Case

Infinite Horizon Discounted Problems

$$\begin{aligned} \min_{\pi} \quad & \lim_{N \rightarrow \infty} \mathbb{E} \left[\sum_{t=0}^{N-1} \gamma^t c(s_t, a_t, w_t) \right] \\ \text{s.t.} \quad & s_{t+1} = f(s_t, a_t, w_t), \quad a_t = \mu(s_t) \in A(s_t), \\ & w_t \sim \mathbb{P}(\cdot \mid s_t, a_t) \end{aligned}$$

Define

$$\begin{aligned} V^{\pi}(s) &= \lim_{N \rightarrow \infty} \mathbb{E} \left[\sum_{t=0}^{N-1} \gamma^t c(s_t, \mu(s_t), w_t) \mid s_0 = s \right] \\ &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, \mu(s_t), w_t) \mid s_0 = s \right] \end{aligned}$$

DP Algorithm: Finite States

Recall: Finite problems: $|S| = \text{finite}$

Transition: $p_{ij}(a) = \mathbb{P}(f(i, a, w) = j)$

Notation

$\mathbb{E}^\mu(\cdot), \mathbb{E}^\pi(\cdot)$ expectation when policy μ or π is employed.

Cost per stage: $c(s, a)$ w.l.o.g.

Define $C_t = C(s_t, a_t)$.

Discounted cost problem: $V_t^\pi(s) = \mathbb{E}^\pi [C_0 + \gamma C_1 + \gamma^2 C_2 + \cdots + \gamma^{t-1} C_{t-1}]$.

Finite Horizon: $V_t^*(s) = \min_{\pi} V_t^\pi(s)$.

Proper definition of Infinite Horizon: $V^\pi(s) = \lim_{t \rightarrow \infty} V_t^\pi(s).$

Interesting question:

$$V^*(s) \triangleq \min_{\pi} V^\pi(s) \stackrel{?}{=} \lim_{t \rightarrow \infty} \min_{\pi} V_t^\pi(s)$$

This entails switching ‘min’ and ‘lim’.

Policy Evaluation (Intuition)

Policy Evaluation

For a fixed policy $\mu(\cdot)$. **We would like to show:**

Discrete States:

$$V(i) = C(i, \mu(i)) + \gamma \sum_{j \in S} p(j \mid i, \mu(i)) V(j)$$

General Systems:

$$V(s) = C(s, \mu(s)) + \gamma \mathbb{E}[V(f(s, \mu(s), w))]$$

Notation: sometimes we write $V^\pi(s)$

Policy Evaluation

$$V_t^\mu(s) = \mathbb{E}[C_0 + \gamma C_1 + \cdots + \gamma^{t-1} C_{t-1}]$$

Define

$V_t^\mu(s)$ is an increasing function.

$$|V_t^\mu(s)| \leq \bar{C} \cdot \sum_{i=0}^{t-1} \gamma^i < \infty$$

This implies convergence to a finite value!

Now we show that the limit satisfies Bellman's equation

Proof of Bellman Equation for Policy Evaluation

$$\begin{aligned} V^\mu(s) &= \mathbb{E}^\mu[C_0 + C_1 + C_2 + \cdots \mid s] \\ &= \mathbb{E}^\mu[C_0 + \gamma G \mid s], \quad G = C_1 + C_2 + \cdots \\ &= \mathbb{E}_{s'}[\mathbb{E}^\mu[C_0 + \gamma G \mid s, s']] \\ &= \mathbb{E}^\mu[C_0 + \gamma \mathbb{E}_{s'}(G \mid s') \mid s] \\ &= \mathbb{E}^\mu[C_0 + \gamma V^\mu(s') \mid s] \\ &= C_0 + \gamma \mathbb{E}^\mu[V^\mu(s') \mid s] \end{aligned}$$

Infinite Horizon Bellman's equation!

Finite State Discounted DP

Optimal Value Function is the Limit of Finite Problems:

$$V^*(s) = \lim_{t \rightarrow \infty} V_t^*(s)$$

What are the implications of this?

Proof of Bellman's Equation

We use the boundedness of C and the discount factor.

$$\begin{aligned}\gamma^t |\mathbb{E}^\pi[C_t]| &\leq \gamma^t \mathbb{E}^\pi[|C_t|] \\ &\leq \gamma^t b\end{aligned}$$

It follows:

$$\begin{aligned}|V^\pi(i) - V_t^\pi(i)| &\leq \sum_{k=t}^{\infty} \gamma^k b \\ &\leq \frac{\gamma^t}{1-\gamma} b \longrightarrow 0 \quad \text{as } t \rightarrow \infty\end{aligned}$$

Does this mean that

$$\left| \min_{\pi} V^\pi(i) - \min_{\pi} V_t^\pi(i) \right| \rightarrow 0 ?$$

Auxiliary Result

$$\left| \min_x g_1(x) - \min_x g_2(x) \right| \leq \max_x |g_1(x) - g_2(x)|$$

Proof

$$g_1(x) - g_2(x) \geq \min_x g_1(x) - g_2(x),$$

$$|g_1(x) - g_2(x)| \geq g_1(x) - g_2(x) \geq \min_x g_1(x) - g_2(x),$$

$$\implies \max_x |g_1(x) - g_2(x)| \geq \min_x g_1(x) - g_2(x),$$

$$\implies \max_x |g_1(x) - g_2(x)| \geq \min_x g_1(x) - \min_x g_2(x).$$

Now reverse the order of g_1 and g_2 .

Limiting Solution: Bellman's Equation

Given that $V^*(s) = \lim_{t \rightarrow \infty} V_t^*(s)$, it follows

$$V^*(i) = \min_a \left\{ c(i, a) + \gamma \sum_j P(j \mid i, a) V^*(j) \right\}$$

Uniqueness

Suppose V_1^* and V_2^* are both solutions.

$$\begin{aligned} |V_1^*(s) - V_2^*(s)| &= \left| \min_a \left[c(s, a) + \gamma \mathbb{E}V_1^*(s') \right] - \min_a \left[c(s, a) + \gamma \mathbb{E}V_2^*(s') \right] \right| \\ &\leq \gamma \max_a \left| \mathbb{E}V_1^*(s') - \mathbb{E}V_2^*(s') \right| \\ &\leq \gamma \max_s |V_1^*(s') - V_2^*(s')| \\ &\triangleq \gamma r \end{aligned}$$

$$r \leq \gamma r \quad \implies \quad r = 0$$

$$\implies V_1^*(s) = V_2^*(s) \quad \text{for all } s.$$

Greedy Policy and Optimality

$$\mu \text{ is optimal iff } \min_a \left[c(s, a) + \gamma \mathbb{E} V^*(s') \right] = V^*(s).$$

Value Iteration (special case)

$$V_0^*(i) = 0$$

$$V_{t+1}^*(i) = \min_a \left\{ c(i, a) + \gamma \sum_j P(j \mid i, a) V_t^*(j) \right\}$$

We have shown convergence.

What is the interpretation of this result?

Q Function

$$Q(s, a) = c(s, a) + \gamma \mathbb{E}V^*(s')$$

We see from above: $\min_a Q(s, a) = V^*(s)$

Q will have advantages soon.

Question