〈2. Dynamic Programming Algorithm〉

20232863 배가흥.

- Contents
  - DP 알고리즘 명제
  - DP 알고리즘 명제 증명

* DP 알고리즘의 명제

· 초기 state인 S_1에 대해서, N-1 번째부터 1 까지 거꾸로 S_t, a_t를 길어간다?

↳ $V_1^*(S_1) = V_1(S_1)$ (미래를 알고 있으므로)

<매 t 마다 S_t, a_t 기반해서 비용인 g_t가 반환됨>

$$\left[\begin{array}{l} V_N(S_N) = g_N(S_N) \quad \text{← } a_N\text{이 존재할 수 없으므로 } S_N\text{만 고려.} \\ V_t(S_t) = \min_{a_t \in A_S}\left[g_t(S_t, a_t) + V_{t+1}(f(S_t, a_t)) \right] \quad \forall t = \{1, \dots, N-1\} \end{array}\right]$$

현재 비용도 ↓, 다음 단계 비용도 ↓ ~최적의 V

현재 비용 / a_t 함으로써 t+1의 비용.

· 만약 $a_t^* = d_t^*(S_t)$가 $V_t(S_t)$식을 최소화하면,

$\pi^* = (d_1^*, d_2^*, \dots, d_{N-1}^*)$ 은 optimal

* DP 알고리즘 명제 증명

· $\pi = (d_1, d_2, \dots, d_{N-1})$과 truncated policy $\pi^t = (d_t, d_{t+1}, \dots, d_{N-1})$에 대해,

· $V_t^*(S_t)$가 최적의 cost-to-go 함수라 가정. → $\pi^* = (d_1^*, \dots, d_{t-1}^*)$이 Optimal

· $V_t^*(S_t) = \min_{\pi^t}\left\{ g_N(S_N) + \sum_{i=t}^{N-1} g_t(S_t, a_t) \right\}$

· $t = N$일 때, $V_N^*(S_N) = g_N(S_N)$, $V_t^*(S_t) = V_t(S_t)$

벨만 방정식에 의해 생성.

· t와 모든 S_{t+1}에 대해, $V_{t+1}^*(S_{t+1}) = V_{t+1}(S_{t+1})$이라 가정한다면,

↳ $\pi^t = (d_t, \pi^{t+1})$, $V_t^*(S_t) = \min_{(d_t, \pi^{t+1})}\left\{ g_N(S_N) + g_t(S_t, a_t) + \sum_{i=t+1}^{N-1} g_i(S_i, a_i) \right\}$

미래 / 현재 / 그 사이

$V_t^*(S_t)$: t시점 기준

$$= \min_{d_t}\left\{ g_t(S_t, a_t) + \min_{\pi^{t+1}}\left\{ g_N(S_N) + \sum_{i=t+1}^{N-1} g_i(S_i, a_i) \right\} \right\}$$

$$\Rightarrow V_{t+1}^*(S_{t+1}) = V_{t+1}(S_{t+1})$$

$$= \min_{d_t}\left\{ g_t(S_t, a_t) + V_{t+1}(S_{t+1}) \right\}$$

$V_{t+1}(S_{t+1}) = f_t(S_t, a_t)$

$$= \min_{a_t \in A_S}\left\{ g_t(S_t, a_t) + V_{t+1}(f_t(S_t, a_t)) \right\}$$

$$= V_t(S_t) \qquad \therefore V_1^*(S_1) = V_1(S_1)$$ ∎

* State value function, Action value function

· 지금부터 (t) 기대되는 Return, 지금 행동으로부터 기대되는 Return

* optimal policy

· 지금부터 기대되는 return을 maximize

* 정의 (가치함수, optimal policy)

Return $G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots$

① $V(S_t) \triangleq \int_{a_t:a_\infty} G_t \, p(a_t, S_{t+1}, a_{t+1}, \cdots | S_t) \, da_t:a_\infty$ (지금부터 시작되서 기대되는 리턴)

↳ 조건부 확률

② $Q(S_t, a_t) \triangleq \int_{S_{t+1}:a_\infty} G_t \, p(S_{t+1}, a_{t+1}, S_{t+2}, a_{t+2}, \cdots | S_t, a_t) \, dS_{t+1}:a_\infty$

③ $V(S_t)$를 maximize하는 $\begin{pmatrix} p(a_t|S_t) \\ p(a_{t+1}|S_{t+1}) \\ \vdots \\ p(a_\infty|S_\infty) \end{pmatrix}$ 가 optimal policy.

* Bellman Equation

$$V(S_t) \triangleq \int_{a_t:a_\infty} G_t \underbrace{p(a_t, S_{t+1}, a_{t+1}, \cdots | S_t)}_{②} \, da_t:a_\infty$$

① 베이지안 룰 적용 $p(x,y) = p(x|y)\,p(y)$

$p(x,y|z) = p(x|y,z)\,p(y|z)$

$$p(S_{t+1}, a_{t+1}, \cdots | S_t, a_t)\, p(a_t|S_t)$$

$$①-① = \int_{a_t} \underbrace{\int_{S_{t+1}:a_\infty} G_t \, p(S_{t+1}, a_{t+1}, \cdots | S_t, a_t) \, dS_{t+1}:a_\infty}_{Q(S_t, a_t)} \, p(a_t|S_t) \, da_t$$

$$= \int_{a_t} Q(S_t, a_t)\, p(a_t|S_t) \, da_t$$

② $p(a_{t+1}, \ldots \mid s_t, a_t, s_{t+1})\, p(a_t, s_{t+1} \mid s_t)$

① - ②