EE365: Dynamic Programming Proof

Markov decision problem

find policy $\mu = (\mu_0, \ldots, \mu_{T-1})$ that minimizes

$$J^\mu = \mathbf{E}\left(\sum_{t=0}^{T-1} g_t(x_t, u_t) + g_T(x_T)\right)$$

Given

- functions $f_0, \ldots, f_{T-1}$
- stage cost functions $g_0, \ldots, g_{T-1}$ and terminal cost g_T
- distributions of independent random variables $x_0, w_0, \ldots, w_{T-1}$

Here

- system obeys dynamics $x_{t+1} = f_t(x_t, u_t, w_t)$.
- we seek a *state feedback* policy: $u_t = \mu_t(x_t)$
- we consider deterministic costs for simplicity

Bellman operator

define the Bellman (or DP) operator $\mathcal{T}_t$ as

$$\mathcal{T}_t(h)(x) = \min_u \left(g_t(x,u) + \mathbf{E}\, h(f_t(x,u,w_t)) \right)$$

- map operates on any function $h : \mathcal{X} \to \mathbb{R}$
- *define* the optimal value function, for $t = T-1, \ldots, 0$

$$v_T^\star = g_T \qquad v_t^\star = \mathcal{T}_t(v_{t+1}^\star)$$

Performance of the optimal policy

- for the optimal policy $\mu^\star$ we have

$$v_t^\star(x) = g_t(x, \mu_t^\star(x)) + \mathbf{E}\, v_{t+1}^\star(f_t(x, \mu_t^\star(x), w_t)), \quad t = T-1, \ldots, 0$$

- this is value iteration for evaluating $J^\star$, so $J^\star = \pi_0 v_0^\star$

Performance of any policy

- for any policy μ we define the value function for $t = T-1, \ldots, 0$

$$v_T^\mu = g_T \qquad v_t^\mu = g_t(x, \mu_t(x)) + \mathbf{E}\, v_{t+1}^\mu(f_t(x, \mu_t(x), w_t))$$

- the cost achieved is $J^\mu = \pi_0 v_0^\mu$

Optimal policy is better for one step

for any policy μ

$$v_t^\mu \geq \mathcal{T}_t(v_{t+1}^\mu)$$

- *i.e.*, acting optimally for the step at time t is better than using policy μ
- because, for all x

$$\begin{aligned} v_t^\mu(x) &= g_t(x, \mu_t(x)) + \mathbf{E}\, v_{t+1}^\mu(f_t(x, \mu_t(x), w_t)) \\ &\geq \mathcal{T}_t(v_{t+1}^\mu)(x) \end{aligned}$$

- since $\mathcal{T}_t$ minimizes over all choices of $u = \mu_t(x)$

Monotonicity of Bellman operator

The Bellman operator is monotone

$$h \leq \tilde{h} \qquad \Longrightarrow \qquad \mathcal{T}_t(h) \leq \mathcal{T}_t(\tilde{h})$$

- inequalities mean for all x
- to see this, assume $h \leq \tilde{h}$, then for any x and u

$$g_t(x,u) + \mathbf{E}\, h(f_t(x,u,w_t)) \leq g_t(x,u) + \mathbf{E}\, \tilde{h}(f_t(x,u,w_t))$$

- minimizing each side over u gives above

Theorem

suppose

- $v_T^\star = g_T$ and $v_t^\star = \mathcal{T}_t(v_{t+1}^\star)$ for $t = T-1, \ldots, 0$
- μ is any policy
- $v_T^\mu = g_T$ and $v_t^\mu = g_t(x, \mu_t(x)) + \mathbf{E}\, v_{t+1}^\mu(f_t(x, \mu_t(x), w_t))$ for $t = T-1, \ldots, 0$

then for all $t = 0, \ldots, T$

$$v_t^\star \leq v_t^\mu$$

and hence $J^\star \leq J^\mu$

Proof of optimality

- using $v_t^\star = \mathcal{T}_t(v_{t+1}^\star)$, $v_t^\mu \geq \mathcal{T}_t(v_{t+1}^\mu)$, and $v_T^\star = v_T^\mu = g_T$,

$$\begin{aligned} v_t^\mu &\geq \mathcal{T}_t(v_{t+1}^\mu) \\ &\geq \mathcal{T}_t\mathcal{T}_{t+1}(v_{t+2}^\mu) \\ &\vdots \\ &\geq \mathcal{T}_t\mathcal{T}_{t+1}\cdots\mathcal{T}_{T-1}(v_T^\mu) \\ &= \mathcal{T}_t\mathcal{T}_{t+1}\cdots\mathcal{T}_{T-1}(g_T) \\ &= v_t^\star \end{aligned}$$

Summary

- any policy defined by dynamic programming is optimal
- (can replace ‘any’ with ‘the’ when the argmins are unique)
- $v_t^\star$ is minimal for any t, over all policies (*i.e.*, $v_t^\star \le v_t^\mu$)
- there can be other optimal (but pathological) policies; for example we can set $\mu_0(x)$ to be anything you like, provided $\pi_0(x) = 0$