EE365: Dynamic Programming

Markov decision problem

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

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

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

Optimal value function

Define the optimal value function

$$V_t^\star(x) = \min_{\mu_t, \mu_{t+1}, \ldots, \mu_{T-1}} \mathbf{E} \left(\sum_{\tau=t}^{T-1} g_\tau(x_\tau, u_\tau) + g_T(x_T) \middle| x_t = x \right)$$

- minimization is over *policies* $\mu_t, \ldots, \mu_{T-1}$
- x_t is known, so we can minimize over *action* u_t and policies $\mu_{t+1}, \ldots, \mu_{T-1}$
- $V_t^\star(x)$ is expected cost-to-go, using an optimal policy, if $x_t = x$
- $J^\star = \sum_x \pi_0(x) V_0^\star(x) = \pi_0 V_0^\star$
- $V_t^\star$ also called Bellman value function, optimal cost-to-go function

Optimal policy

- the policy

$$\mu_t^\star(x) \in \operatorname*{argmin}_u \left(g_t(x,u) + \mathbf{E}\, V_{t+1}^\star(f_t(x,u,w_t))\right)$$

  is optimal
- expectation is over w_t
- can choose any minimizer when minimizer is not unique
- there can be optimal policies not of the form above
- *looks* circular and useless: need to know optimal policy to find $V_t^\star$

Interpretation

$$\mu_t^\star(x) \in \underset{u}{\operatorname{argmin}} \left(g_t(x,u) + \mathbf{E}\, V_{t+1}^\star(f_t(x,u,w_t)) \right)$$

assuming you are in state x at time t, and take action u

- $g_t(x,u)$ (a number) is the current stage cost you pay
- $V_{t+1}^\star(f_t(x,u,w_t))$ (a random variable) is the cost-to-go from where you land, if you follow an optimal policy for $t+1,\ldots,T-1$
- $\mathbf{E}\, V_{t+1}^\star(f_t(x,u,w_t))$ (a number) is the expected cost-to-go from where you land

optimal action is to minimize sum of current stage cost and expected cost-to-go from where you land

Greedy policy

- greedy policy is $\mu_t^{\mathrm{gr}}(x) \in \mathrm{argmin}_u \, g_t(x,u)$
- at any state, minimizes current stage cost without regard for effect of current action on future states
- in optimal policy

$$\mu_t^\star(x) \in \underset{u}{\mathrm{argmin}} \left(g_t(x,u) + \mathbf{E}\, V_{t+1}^\star(f_t(x,u,w_t)) \right)$$

  second term summarizes effect of current action on future states

Dynamic programming recursion

- define $V_T^\star(x) := g_T(x)$
- for $t = T-1, \ldots, 0$,
  - find optimal policy for time t in terms of $V_{t+1}^\star$:

$$\mu_t^\star(x) \in \operatorname*{argmin}_u \left(g_t(x,u) + \mathbf{E}\, V_{t+1}^\star(f_t(x,u,w_t))\right)$$

  - find $V_t^\star$ using $\mu_t^\star$:

$$V_t^\star(x) = \min_u \left(g_t(x,u) + \mathbf{E}\, V_{t+1}^\star(f_t(x,u,w_t))\right)$$

- a recursion that runs backward in time
- complexity is $T|\mathcal{X}||\mathcal{U}||\mathcal{W}|$ operations (fewer when P is sparse)

Variations

- random costs:

$$\mu_t^\star(x) \in \operatorname{argmin}_u \mathbf{E}\left(g_t(x,u,w_t) + V_{t+1}^\star(f_t(x,u,w_t))\right)$$
$$V_t^\star(x) := \mathbf{E}\, g_t(x,\mu_t^\star(x),w_t) + \mathbf{E}\, V_{t+1}^\star(f_t(x,\mu_t^\star(x),w_t))$$

- state-action separable cost $g_t(x,u) = q_t(x) + r_t(u)$:

$$\mu_t^\star(x) \in \operatorname{argmin}_u \left(r_t(u) + \mathbf{E}\, V_{t+1}^\star(f_t(x,u,w_t))\right)$$
$$V_t^\star(x) := q_t(x) + r_t(\mu_t^\star(x)) + \mathbf{E}\, V_{t+1}^\star(f_t(x,\mu_t^\star(x),w_t))$$

- deterministic system:

$$\mu_t^\star(x) \in \operatorname{argmin}_u \left(g_t(x,u) + V_{t+1}^\star(f_t(x,u))\right)$$
$$V_t^\star(x) := g_t(x,\mu_t^\star(x)) + V_{t+1}^\star(f_t(x,\mu_t^\star(x)))$$