EE266: Infinite Horizon Markov Decision Problems

Infinite horizon Markov decision problems

Infinite horizon dynamic programming

Example
Infinite horizon Markov decision problems
Infinite horizon Markov decision process

- (time-invariant) Markov decision process: \( x_{t+1} = f(x_t, u_t, w_t) \)
- \( w_t \) IID, independent of \( x_0 \)
- (time-invariant state-feedback) policy: \( u_t = \mu(x_t) \)
- \( x_0, x_1, \ldots \) is Markov
- closed-loop Markov chain: \( x_{t+1} = F(x_t, w_t) = f(x_t, \mu(x_t), w_t) \)
Infinite horizon costs

- total cost:
  \[
  J^{\text{tot}} = \mathbb{E} \sum_{t=0}^{\infty} g(x_t, u_t, w_t) = \lim_{T \to \infty} \mathbb{E} \sum_{t=0}^{T} g(x_t, u_t, w_t)
  \]

- discounted infinite horizon:
  \[
  J^{\text{disc}} = \mathbb{E} \sum_{t=0}^{\infty} \gamma^t g(x_t, u_t, w_t)
  \]

  \(\gamma \in (0, 1)\) is the discount factor

- average stage cost:
  \[
  J^{\text{avg}} = \lim_{T \to \infty} \mathbb{E} \frac{1}{T} \sum_{t=0}^{T} g(x_t, u_t, w_t)
  \]

  (includes cost at absorption as special case)
Infinite horizon costs

- let $P$ be closed-loop transition matrix (which depends on $\mu$)

- total cost (existence can depend on $\pi_0, g$):

$$J_{\text{tot}} = \pi_0 \left( \sum_{t=0}^{\infty} P^t \right) g$$

- discounted cost (always exists):

$$J_{\text{disc}} = \pi_0 \left( \sum_{t=0}^{\infty} \gamma^t P^t \right) g = \pi_0 (I - \gamma P)^{-1} g$$

- average cost (always exists):

$$J_{\text{avg}} = \pi_0 \left( \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} P^t \right) g$$
Infinite horizon Markov decision problems

- choose $\mu$ to minimize $J^{\text{tot}}$, $J^{\text{disc}}$, or $J^{\text{avg}}$
- data are $\pi_0$, $f$, $g$, distribution of $w_t$, and $\gamma$ (for discounted case)
Example: Stopping problem

- \( z_{t+1} = f(z_t, w_t) \) is a Markov chain on \( \mathcal{Z} \), with costs \( g^{\text{hold}}, g^{\text{stop}} : \mathcal{Z} \to \mathbb{R} \)
- augment with a state called D (for DONE): \( \mathcal{X} = \mathcal{Z} \cup \{D\} \)
- actions are \( \mathcal{U} = \{W, S\} \) (WAIT and STOP)
- dynamics: D is absorbing (\( x_t = D \to x_{t+1} = D \)); for \( x_t = z \in \mathcal{Z}, \)
  \[
  x_{t+1} = \begin{cases} 
  f(z, w_t) & u_t = W \\
  D & u_t = S
  \end{cases}
  \]
- stage cost: \( g(D, u) = 0 \); for \( x = z \in \mathcal{Z}, \)
  \[
  g(x, u) = \begin{cases} 
  g^{\text{hold}}(z) & u = W \\
  g^{\text{stop}}(z) & u = S
  \end{cases}
  \]
- minimize total cost \( J^{\text{tot}} \)
- optimal policy tells you whether to wait or stop at each \( z \in \mathcal{Z} \)
Infinite horizon dynamic programming
Total cost: Value function

- define value function

\[ V^*(x) = \min_{\mu} \mathbb{E} \left( \sum_{t=0}^{\infty} g(x_t, u_t, w_t) \mid x_0 = x \right) \]

with \( u_t = \mu(x_t), \ x_{t+1} = f(x_t, u_t, w_t) \)

- gives optimal cost, starting from state \( x \) at \( t = 0 \); can be infinite

- an optimal policy is

\[ \mu^*(x) \in \arg\min_u \mathbb{E} (g(x, u, w_t) + V^*(f(x, u, w_t))) \]

- \( V^* \) is fixed point of Bellman operator:

\[ V^*(x) = \min_u \mathbb{E} (g(x, u, w_t) + V^*(f(x, u, w_t))) \]
Total cost: Value iteration

- value iteration: set $V_0 = 0$; for $k = 0, 1, \ldots$
  
  $$V_{k+1}(x) = \min_u \mathbb{E} \left( g(x, u, w_t) + V_k(f(x, u, w_t)) \right)$$

  ($k$ is an iteration counter, not time)

- define associated policy
  
  $$\mu_k(x) = \arg\min_u \mathbb{E} \left( g(x, u, w_t) + V_k(f(x, u, w_t)) \right)$$

- $V_k \rightarrow V^*$, in the absence of pathologies (ITAP)

- $\mu_k \rightarrow \mu^*$ (more precisely, the total cost with $\mu_k$ converges to optimal) (ITAP)
Total cost: Value iteration

- interpretation:
  - solve finite horizon problem over $t = 0, \ldots, k$
  - $\mu_k$ is the policy for $t = 0$ for the finite horizon problem
Discounted cost: Value function

- define value function

\[ V^*(x) = \min_{\mu} \mathbb{E} \left( \sum_{t=0}^{\infty} \gamma^t g(x_t, u_t, w_t) \mid x_0 = x \right) \]

with \( u_t = \mu(x_t) \), \( x_{t+1} = f(x_t, u_t, w_t) \)

- gives optimal cost, starting from state \( x \) at \( t = 0 \); sum always exists

- an optimal policy is

\[ \mu^*(x) \in \arg\min_{\mu} \mathbb{E} (g(x, u, w_t) + \gamma V^*(f(x, u, w_t))) \]

- \( V^* \) is fixed point of Bellman operator:

\[ V^*(x) = \min_{u} \mathbb{E} (g(x, u, w_t) + \gamma V^*(f(x, u, w_t))) \]
Discounted cost: Value iteration

value iteration:

\[ V_{k+1}(x) = \min_u \mathbf{E} \left( g(x, u, w_t) + \gamma V_k(f(x, u, w_t)) \right) \]

converges to \( V^* \) always

reason: Bellman operator

\[ (T h)(x) = \min_u \mathbf{E} \left( g(x, u, w_t) + \gamma h(f(x, u, w_t)) \right) \]

is a \( \gamma \)-contraction:

\[ \| T(h) - T(\tilde{h}) \|_\infty \leq \gamma \| h - \tilde{h} \|_\infty \]
Value iteration for average cost

- we start by defining value iteration: \( V_0 = 0 \); for \( k = 0, 1, \ldots \),
  \[
  V_{k+1}(x) = \min_u \mathbb{E} \left( g(x, u, w_t) + V_k(f(x, u, w_t)) \right)
  \]

- \( V_k \) are value functions for finite horizon total cost problem (indexed in reverse order)

- \( V_k \) does not converge, but associated policy
  \[
  \mu_k(x) = \arg\min_u \mathbb{E} \left( g(x, u, w_t) + V_k(f(x, u, w_t)) \right)
  \]
  does converge, to an optimal policy for average cost problem, ITAP

- we have, as \( k \to \infty \)
  \[
  V_{k+1}(x) - V_k(x) \to J^* \quad \text{independent of } x
  \]
  total cost value function eventually increases by constant \( J^* \) each step
Average cost: Relative value function

- Define relative value function iterate as
  \[ V^\text{rel}_k(x) = V_k(x) - V_k(x') \]

- \( x' \in \mathcal{X} \) is (an arbitrary) reference state: \( V^\text{rel}_k(x') = 0 \)

- Define relative value function as \( V^\text{rel} = \lim_{k \to \infty} V^\text{rel}_k \)

- Optimal policy is
  \[ \mu^*(x) = \arg\min_u \mathbf{E}\left( g(x, u, w_t) + V^\text{rel}(f(x, u, w_t)) \right) \]

- \( V^\text{rel} \) satisfies average cost Bellman equation
  \[ V^\text{rel}(x) + J^* = \min_u \mathbf{E}\left( g(x, u, w_t) + V^\text{rel}(f(x, u, w_t)) \right) \]
Average cost: Relative value iteration

► (relative) value iteration for average cost problem:

\[
\tilde{V}_{k+1}(x) = \min_u \mathbb{E} \left( g(x, u, w_t) + V_{k+1}^{rel}(f(x, u, w_t)) \right)
\]

\[
J_{k+1}(x) = \tilde{V}_{k+1}(x')
\]

\[
V_{k+1}^{rel}(x) = \tilde{V}_{k+1}(x) - J_{k+1}
\]

► \( V_k^{rel} \to V^{rel} \), \( J_k \to J^* \) as \( k \to \infty \)
value iteration: $V_0 = 0$; for $k = 0, 1, \ldots$, 

$$V_{k+1}(x) = \min_u \mathbb{E}(g(x, u, w_t) + V_k(f(x, u, w_t)))$$

(multiply $V_k$ by $\gamma$ for discounted case)

associated policy:

$$\mu_k(x) = \arg\min_u \mathbb{E}(g(x, u, w_t) + V_k(f(x, u, w_t)))$$

for all infinite horizon problems, simple value iteration works

for total cost problem, $V_k$ and $\mu_k$ converge to optimal, ITAP

for discounted cost problem, $V_k$ and $\mu_k$ converge to optimal

for average cost problem, $V_k$ does not converge, but $\mu_k$ does converge to optimal, ITAP
Example
Example: Stopping problem

random walk on a $20 \times 20$ grid, with three target states
Example: Stopping problem

- transitions uniform to neighbors
- holding cost $g^{\text{hold}}(z) = 1$
- stopping at a target state gives a payoff

$$g^{\text{stop}}(z) = \begin{cases} -120 & z = (5, 5) \\ -70 & z = (17, 10) \\ -150 & z = (10, 15) \\ 0 & \text{otherwise} \end{cases}$$
Example: Stopping problem

value function
Example: Stopping problem

optimal policy (red=STOP, white=WAIT)