EE365: Markov Decision Processes

Markov decision processes

Markov decision problem

Examples

Markov decision processes

Markov decision processes

- add **input** (or **action** or **control**) to Markov chain with costs
- input selects from a set of possible transition probabilities
- input is function of state (in standard information pattern)

Definition: Dynamical system form

$$x_{t+1} = f_t(x_t, u_t, w_t), \quad t = 0, 1, \ldots, T-1$$

- state $x_t \in \mathcal{X}$
- action or input $u_t \in \mathcal{U}$
- uncertainty or disturbance $w_t \in \mathcal{W}$
- dynamics functions $f_t : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathcal{X}$
- $x_0, w_0, \ldots, w_{T-1}$ are independent RVs

- variation (state dependent input space): $u_t \in \mathcal{U}_t(x_t) \subseteq \mathcal{U}$
  ($\mathcal{U}_t(x_t)$ is set of allowed actions in state x_t at time t)

Policy

- action is function of state:

$$u_t = \mu_t(x_t), \quad t = 0, \dots, T-1$$

- $\mu_t : \mathcal{X} \to \mathcal{U}$ is state feedback function at time t
- $\mu = (\mu_0, \dots, \mu_{T-1})$ is the policy (or control law)

- number of possible policies: $|\mathcal{U}|^{|\mathcal{X}|T}$
  - very large for any case of interest
  - for each $t = 0, \dots, T-1$, for each $x \in \mathcal{X}$, we can choose $\mu_t(x) \in \mathcal{U}$

Closed-loop system

- with policy, ('closed-loop') dynamics is

$$x_{t+1} = F_t(x_t, w_t) = f_t(x_t, \mu_t(x_t), w_t), \quad t = 0, 1, \ldots, T-1$$

- F_t are closed-loop state transition functions
- $x_0, \ldots, x_T$ is Markov

Information patterns

- $u_t = \mu_t(x_t)$ is standard information pattern
  - action is function of current state
  - also called state feedback control
- some nonstandard information patterns:
  - full information (or prescient): $u_t = \mu_t(x_0, w_0, \ldots, w_{T-1})$
  - no information: $u_t = \mu_t()$ (*i.e.*, $u_0, \ldots, u_{T-1}$ are fixed)
  - initial state (also called open-loop): $u_t = \mu_t(x_0)$
  - state and disturbance: $u_t = \mu_t(x_t, w_t)$

Cost function

- total cost is

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

- stage cost functions $g_t : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathbb{R}$
- terminal cost function $g_T : \mathcal{X} \to \mathbb{R}$
- variation: allow g_t to take on value $+\infty$ to encode constraints on state-action pairs ($-\infty$ for rewards, when we maximize)
- we sometimes write J^μ to show dependence of cost on policy

Closed-loop stage cost functions

- closed-loop stage cost functions:

$$G_t(x) = \mathop{\mathbf{E}}_{w_t} g_t(x, \mu_t(x), w_t), \quad t = 0, \ldots, T-1$$

(note that $x_t \perp\!\!\!\perp w_t$)

- closed-loop terminal cost function:

$$G_T(x) = g_T(x)$$

Cost function: Special cases

- deterministic cost: g_t do not depend on w_t
- time-invariant: $g_0, \dots, g_T$ are the same
- terminal cost only: $g_0 = \cdots = g_{T-1} = 0$
- state-control separable (deterministic case):

$$g_t(x_t, u_t, w_t) = q_t(x_t) + r_t(u_t)$$

  - $q_t : \mathcal{X} \to \mathbb{R}$ is state cost function
  - $r_t : \mathcal{U} \to \mathbb{R}$ is action cost function

Value iteration to compute cost

- we can use value iteration to compute J
- (deterministic cost for simplicity)
- take $V_T(x) = g_T(x)$,

$$V_t(x) = g_t(x, \mu_t(x)) + \mathbf{E}\, V_{t+1}(f_t(x, \mu_t(x), w_t)), \quad t = T-1, \ldots, 0$$

  (expectation is over w_t)
- $J = \pi_0 V_0$
- computation cost is $T|\mathcal{X}||\mathcal{W}|$ operations (fewer for sparse transitions)

Concrete form

- $\mathcal{X} = \{1, \ldots, n\}, \mathcal{U} = \{1, \ldots, m\}$
- transition probabilities (time-invariant case) given by

$$P_{ijk} = \mathbf{Prob}(x_{t+1} = j \mid x_t = i,\ u_t = k)$$

- P_{ijk} is probability that next state is j, when current state is i and control action k is taken
- P is 3-D array (often sparse)
- in state i, action chooses next state distribution from choices

$$P_{i,:,k} = [P_{i1k} \ \cdots P_{ink}], \quad k = 1, \ldots, m$$

- for time-varying case, P is 4-D array (!!)

Concrete form

- stage costs (time-invariant case) given by

$$C_{ijk}, \quad i,j = 1,\ldots,n, \quad k = 1,\ldots,m$$

- C_{ijk} is cost when state i transitions to state j with action k
- C is 3-D array (often sparse); can assume that $C_{ijk} = 0$ when $P_{ijk} = 0$
- state-action separable case: $C_{ijk} = q_i + r_k$

Markov decision problem

Markov decision process

- Markov decision process (MDP) defined by
  - (action dependent) state transition functions $f_0, \ldots, f_{T-1}$
  - distributions of $x_0, w_0 \ldots, w_{T-1}$
  - stage cost functions $g_0, \ldots, g_{T-1}$
  - terminal cost function g_T

- policy defined by state feedback functions $\mu_0, \ldots, \mu_{T-1}$
- combining Markov decision problem with policy, we get closed-loop Markov chain with costs

Markov decision problem

- given Markov decision process, cost with policy μ is J^μ
- Markov decision problem: find a policy $\mu^\star$ that minimizes J^μ
- number of possible policies: $|\mathcal{U}|^{|\mathcal{X}|T}$ (very large for any case of interest)
- there can be multiple optimal policies
- we will see how to find an optimal policy next lecture

Examples

Trading

simple trading model for one asset:

- hold (integer) number of shares $q_t \in [Q^{\min}, Q^{\max}]$ in period t
- buy u_t shares at time t, $u_t \in [Q^{\min} - q_t, Q^{\max} - q_t]$, so

$$q_{t+1} = q_t + u_t$$

- price $p_t \in \{P_1, \ldots, P_k\}$ is Markov; p_t known before u_t is chosen
- revenue is $-u_t p_t - T(u_t) - S((q_t)_-)$
  - $T(u_t) \geq 0$ is transaction cost
  - $S((q_t)_-) \geq 0$ is shorting cost
- $q_0 = 0$; we require $q_T = 0$
- maximize total expected revenue over $t = 0, \ldots, T-1$

Trading

MDP model:

- state is $x_t = (q_t, p_t)$
- stage cost is negative revenue
- terminal cost is $g_T(0) = 0$; $g_T(q) = \infty$ for $q \neq 0$
- (trading) policy gives number of assets to buy (sell) as function of time t, current holdings q_t, and price p_t
- presumably, good policy buys when p_t is low and sells when p_t is high

Variations

how do we handle (model) the following, and what assumptions would we need to make?

- price movements that depend on u_t (price impact)
- imperfect fulfillment (*i.e.*, you might not buy or sell the full amount u_t)
- price movements that depend on a 'signal' $s_t \in \{S_1, \ldots, S_r\}$ that you know at time t