Lesson Plan

• Markov decision processes
• Policies and Value functions
• Solving: average reward, discounted reward
• Bellman equations
• Applications
  – Airline meals
  – Assisted living
  – Car driving

Markov decision process (MDP)

- An MDP model contains:
  - A set of possible world states $S$ ($m$ states)
  - A set of possible actions $A$ ($n$ actions)
  - A reward function $R(s,a)$
  - A transition function $P(s,a,s') \in [0,1]$

- The problem is to decide which action to take in each state.
- Can be finite horizon or infinite horizon

Markov property: effect of an action only depend on current state, not prior history

Example: Maintenance (Hillier and Lieberman)

- Four machine states \{0,1,2,3\}
- Actions: do nothing, overhaul, replace; transitions depend on action
- Each action in each state is associated with a cost (= negated reward)
  - E.g., if “overhaul” in state 2 then cost = 4000 due to $2000 maintenance cost and $2000 cost of lost production

<table>
<thead>
<tr>
<th>Decision</th>
<th>State</th>
<th>Expected Cost Due to Producing Defective Items, $</th>
<th>Maintenance Cost, $</th>
<th>Cost (Lost Profit) of Lost Production, $</th>
<th>Total Cost per Week, $</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Do nothing</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1,000</td>
<td>0</td>
<td>0</td>
<td>1,000</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3,000</td>
<td>0</td>
<td>0</td>
<td>3,000</td>
</tr>
<tr>
<td>2. Overhaul</td>
<td>2</td>
<td>0</td>
<td>2,000</td>
<td>0</td>
<td>4,000</td>
</tr>
<tr>
<td>3. Replace</td>
<td>1, 2, 3</td>
<td>0</td>
<td>4,000</td>
<td>2,000</td>
<td>6,000</td>
</tr>
</tbody>
</table>
Policy

- A **policy** \( \mu : S \rightarrow A \) is a mapping from state to action
- We consider **stationary policies** that depend on state but not time

- A **stochastic policy** \( \mu : S \rightarrow \Delta(A) \) maps to a distribution on actions
Following a Policy

- Determine current state $s$
- Execute action $\mu(s)$
- Fixing an MDP model and a policy $\mu$, this defines a Markov chain.
  - The policy induces a distribution on sequences of states
  - An MDP is **ergodic** if the associated Markov chain is ergodic for every deterministic policy

Evaluating a Policy

- How good is a policy $\mu$ in state $s$?
- Expected total reward? This may be infinite!
- How can we compare policies?
Value functions

- A value function $V_\mu(s)$ defines the expected objective value of policy $\mu$ from state $s$.

- Different kinds of objectives
- If there is a finite decision horizon, we can sum the rewards.
- If there is an infinite decision horizon, we can adopt the average reward in the limit or the infinite sum of discounted rewards.

Expected Average Reward Criterion

- Let $V_\mu^{(n)}(s)$ denote the total expected reward of policy $\mu$ for next $n$ transitions from state $s$.
- Expected average reward criterion:
  $$V_\mu(s) = \lim_{n \to \infty} \left[ \frac{1}{n} V_\mu^{(n)}(s) \right]$$

- For an ergodic MDP, then
  $$V_\mu(s) = \sum_{s'} R(s', \mu(s')) \pi_\mu(s'),$$
  where $\pi_\mu(s')$ is the steady-state prob given $\mu$ of being in state $s'$. 
Expected Discounted Reward Criterion

• A reward \( n \) steps away is discounted by \( \gamma^n \), for **discount factor** \( 0 < \gamma < 1 \)
• Discount factor models uncertainty about the model, uncertainty about lifetime or the time value of money.

• Expected discounted reward criterion:
  \[
  V_\mu(s) = \mathbb{E}_{s', s'', \ldots}[ R(s, \mu(s)) + \gamma R(s', \mu(s')) + \gamma^2 R(s'', \mu(s'')) + \ldots ]
  \]

Solving MDPs

• Expected average reward criterion
  – LP
• Expected discounted reward criterion
  – LP
  – Policy iteration
  – Value iteration
Expected Average reward criterion

Given model $M=(S,A,P,R)$, find policy $\mu$ that maximizes expected average reward criterion.

• Assume the Markov chain associated with every policy is ergodic

Representing a Policy

<table>
<thead>
<tr>
<th>state</th>
<th>action</th>
<th>0</th>
<th>1</th>
<th>...</th>
<th>n-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$x(0,0)$</td>
<td>$x(0,1)$</td>
<td>...</td>
<td>$x(0,n-1)$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$x(1,0)$</td>
<td>$x(1,1)$</td>
<td>...</td>
<td>$x(1,n-1)$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>m-1</td>
<td>$x(m-1,0)$</td>
<td>$x(m-1,1)$</td>
<td>...</td>
<td>$x(m-1,n-1)$</td>
<td></td>
</tr>
</tbody>
</table>

$x(s,a)= \begin{cases} 
1 & \text{if action } a \text{ in state } s \\
0 & \text{otherwise} 
\end{cases}$

Rows sum to 1
Representing a Policy

<table>
<thead>
<tr>
<th>state</th>
<th>action</th>
<th>0</th>
<th>1</th>
<th>...</th>
<th>n-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>x(0,0)</td>
<td>x(0,1)</td>
<td>...</td>
<td>x(0,n-1)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>x(1,0)</td>
<td>x(1,1)</td>
<td>...</td>
<td>x(1,n-1)</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>m-1</td>
<td>x(m-1,0)</td>
<td>x(m-1,1)</td>
<td>...</td>
<td>x(m-1,n-1)</td>
<td></td>
</tr>
</tbody>
</table>

Allowing for stochastic policies:
\[ x(s,a) = \text{Prob}\{\text{action}=a \mid \text{state}=s\} \]

Rows sum to 1

Towards an LP formulation

- Define \( \pi(s,a) \) as the steady-state probability of being in state \( s \) and taking action \( a \)

- We have:
  \[
  \pi(s) = \sum_{a'} \pi(s,a') \\
  \pi(s,a) = \pi(s)x(s,a)
  \]

- Given \( \pi(s,a) \), the policy is:
  \[
  x(s,a) = \frac{\pi(s,a)}{\pi(s)} = \frac{\pi(s,a)}{\sum_{a'} \pi(s,a')}
  \]
LP formulation: Average criterion

\[
V = \max \sum_s \sum_a R(s,a)\pi(s,a)
\]
\[
\text{s.t. } \sum_s \sum_a \pi(s,a) = 1
\]
\[
\sum_a \pi(s',a) = \sum_s \sum_a \pi(s,a)P(s,a,s') \quad \forall s' \quad (3)
\]
\[
\pi(s,a) \geq 0 \quad \forall s, \forall a \quad (4)
\]

(1) maximize expected average reward
(2) total unconditional probability sums to one
(3) Balance equations: total prob in state \(s'\) consistent with states \(s\) from which transitions to \(s'\) possible

Optimal Policies are Deterministic

\[
V = \max \sum_s \sum_a R(s,a)\pi(s,a)
\]
\[
\text{s.t. } \sum_s \sum_a \pi(s,a) = 1
\]
\[
\sum_a \pi(s',a) = \sum_s \sum_a \pi(s,a)P(s,a,s') \quad \forall s' \quad (3)
\]
\[
\pi(s,a) \geq 0 \quad \forall s, \forall a \quad (4)
\]

\(nm\) decision variables; \(m+1\) equalities (one of balance equalities is redundant)
simplex terminates with \(m\) basic variables
\(\pi(s)>0\) for each \(s\) since MDP ergodic; and so \(\pi(s,a)>0\)
for at least one \(a\) for each \(s\).

\[\Rightarrow \pi(s,a)>0\] for exactly one \(a\) in each \(s\)
\[\Rightarrow \text{a deterministic policy.}\]
• Note: if the MDP is formulated with costs rather than rewards, we can simply write the object as a minimization

Example: Maintenance

• \(\min 1000\pi(1,1) + 6000\pi(1,3) + 3000\pi(2,1) + 4000\pi(2,2)\)
  \(+ 6000\pi(2,3) + 6000\pi(3,3)\)

  s.t.
  \(\pi(0,1) + \pi(1,1) + \pi(1,3) + \ldots + \pi(3,3) = 1\)
  \(\pi(0,1) - (\pi(1,3) + \pi(2,3) + \pi(3,3)) = 0\)
  \(\pi(1,1) + \pi(1,3) - (7/8\pi(0,1) + 3/4\pi(1,1) + \pi(2,2)) = 0\)
  \(\pi(2,1) + \pi(2,2) + \pi(2,3) - (1/16\pi(0,1) + 1/8\pi(1,1) + \frac{1}{2}\pi(2,1)) = 0\)
  \(\pi(3,3) - (1/16\pi(0,1) + 1/8\pi(1,1) + \frac{1}{2}\pi(2,1)) = 0\)
  \(\pi(0,1), \ldots, \pi(3,3) \geq 0\)

• \(\pi'(0,1)=2/21; \pi'(1,1)=5/7; \pi'(2,2)=2/21; \pi'(3,3)=2/21; \text{rest zero.}\)
• Optimal policy: \(\mu(0)=1, \mu(1)=1, \mu(2)=2, \mu(3)=3; \text{do nothing in 0 and 1, overhaul in 2, and replace in 3.}\)
Solving MDPs

- Expected average reward criterion
  - LP
- Expected discounted reward criterion
  - LP
  - Policy iteration
  - Value iteration

Solving MDPs: Expected discounted reward criterion

Given model \( M=(S,A,P,R) \), find the policy \( \mu \) that maximizes the expected discounted reward, for discount factor \( 0<\gamma<1 \).

- Note: no need to assume the Markov chain associated with each policy is ergodic.
LP formulation: Discounted Criterion

Choose any $\beta$ values s.t. $\sum_s \beta(s) = 1$ and $\beta(s) > 0$ for all $s$

(represents the start state distribution, $P(S_0 = s) = \beta(s)$)

\[
V = \max \sum_s \sum_a R(s,a)\pi(s,a) \quad (1)
\]

s.t. $\sum_s \sum_a \pi(s,a) = 1 \quad (2)$

\[
\sum_a \pi(s',a) - \gamma \sum_s \sum_a \pi(s,a)P(s,a,s') = \beta(s') \quad \forall \ s' \quad (2')
\]

$\pi(s,a) \geq 0 \quad \forall s, \forall a \quad (3)$

Policy: $x(s,a) = \pi(s,a) / \sum_{a'} \pi(s,a')$

$p_0(s,a) = \gamma \pi_1(s,a) + \gamma^2 \pi_2(s,a) + \ldots$; $\pi_t(s,a)$ is prob of being in state $s$ and taking action $a$ at time $t$. $\pi_t(s,a)$ is discounted exp time of being in state $s$, taking action $a$.

Fact: value $V$ depends on $\beta$; but the optimal policy is deterministic, and invariant to $\beta$!

Example: Maintenance

- Discount $\gamma = 0.9$, suppose $\beta = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
- $\min 1000\pi(1,1) + 6000\pi(1,3) + 3000\pi(2,1) + 4000\pi(2,2) + 6000\pi(2,3) + 6000\pi(3,3)$

s.t.
\[
\pi(0,1) - 0.9(\pi(1,3) + \pi(2,3) + \pi(3,3)) = \frac{1}{4}
\]

\[
\pi(1,1) + \pi(1,3) - 0.9(7/8\pi(0,1) + 3/4\pi(1,1) + \pi(2,2)) = \frac{1}{4}
\]

\[
\pi(2,1) + \pi(2,2) + \pi(2,3) - 0.9(1/16\pi(0,1) + 1/8\pi(1,1) + 1/8\pi(1,1) + 1/8\pi(2,1)) = \frac{1}{4}
\]

\[
\pi(3,3) - 0.9(1/16\pi(0,1) + 1/8\pi(1,1) + 1/8\pi(2,1)) = \frac{1}{4}
\]

\[
\pi(0,1), \ldots, \pi(3,3) \geq 0
\]

- Optimal policy: $\pi^*(0,1) = 1.21$  $\pi^*(1,1) = 6.66$  $\pi^*(2,2) = 1.07$
  $\pi^*(3,3) = 1.07$, rest all zero. In this example, it is the same as for the average reward criterion.
Fundamental Theorem of MDPs

- **Theorem.** Under discounted reward criterion, policy $\mu^*$ is **uniformly optimal**; i.e., it is optimal for all distributions on start states.

- Note: Not true for expected average reward criterion (e.g., when the MDP is not ergodic.)

Solving MDPs

- **Expected average reward criterion**
  - LP

- **Expected discounted reward criterion**
  - LP
    - Policy iteration
    - Value iteration
Bellman equations

- Basic consistency equations for the case of a discounted reward criterion

- For any policy $\mu$, the value function satisfies:
  $$V_\mu(s) = R(s, \mu(s)) + \gamma \sum_{s' \in S} P(s,\mu(s),s')V_\mu(s') \quad \forall \, s$$

For the optimal policy:
  $$V^*(s) = \max_{a \in A} [R(s,a) + \gamma \sum_{s' \in S} P(s,a,s')V^*(s')] \quad \forall \, s$$

Policy and Value Iteration

- $\mu'(s) := \arg \max_a R(s,a) + \gamma \sum_{s' \in S} P(s,a,s')V_\mu(s')$ (Impr.)
- $V_\mu(s) = R(s, \mu(s)) + \gamma \sum_{s' \in S} P(s,\mu(s),s')V_\mu(s')$ (Eval.)

- **Policy iteration**

  $\mu_0 \xrightarrow{E} V_{\mu_0} \xrightarrow{I} \mu_1 \xrightarrow{E} V_{\mu_1} \xrightarrow{I} \mu_1 \xrightarrow{E} V_{\mu_2} \xrightarrow{I} \ldots$

  Each step provides a strict improvement in policy. Finite # policies. Converges!

- **Value iteration**

  - Just iterate on Bellman equations:
    $$V_{k+1}(s) := \max_a [R(s,a) + \gamma \sum_{s' \in S} P(s,a,s')V_k(s')]$$
    Also converges! Less computationally intensive than policy iteration.
Applications

- Airline meal provisioning
- Assisted Living
- High-speed obstacle avoidance

Airline meal provisioning

(572x25)

15

Applications

- Airline meal provisioning
- Assisted Living
- High-speed obstacle avoidance

Determine quantity of meals to load. Passenger load varies because of stand-by, missed flights... Average costs down 17%, short-catered flights down 33%
Additional applications

Assisted Living (Pollack’ 05)

High speed obstacle avoidance (Ng et al.’ 05)

Summary

• MDPs: Policy maps states to actions, fixing a policy we have a Markov chain
• Average reward and Discounted reward criteria
• Find optimal policies by solving as LPs
  – Deterministic
  – Uniformly optimal (need ergodic if average-reward criterion)
• Can also solve via Policy iteration and value iteration
• Rich applications

\[ R(s) = -|v_{\text{desired}} - v_{\text{actual}}| - K \cdot \text{Crashed} \]