Warm-up as You Walk In

Given

- Set actions (persistent/static)
- Set states (persistent/static)
- Function $T(s, a, s\_prime)$

Write the pseudo code for:

- function $V(s)$ return value

that implements:

$$V(s) = \max_{a \in \text{actions}} \sum_{s' \in \text{states}} T(s, a, s') V(s')$$
Announcements

Assignments:

- P3
  - Due Thu 3/7, 10 pm

Spring Break!

- No recitation this Friday
- HW7 (online): out Wed 3/6, due Tue 3/19
- P4: out after break, due Thu 3/28
Minimax Notation

\[ V(s) = \max_a V(s'), \]
\[ \text{where } s' = \text{result}(s, a) \]

\[ \hat{a} = \arg\max_a V(s'), \]
\[ \text{where } s' = \text{result}(s, a) \]
Expectations

<table>
<thead>
<tr>
<th>Time</th>
<th>Probability</th>
<th>20 min</th>
<th>30 min</th>
<th>60 min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Max node notation

\[ V(s) = \max_a V(s'), \]

where \( s' = \text{result}(s, a) \)

Chance node notation

\[ V(s) = \sum_{s'} P(s') V(s') \]
Piazza Poll 1

Expectimax tree search: Which action do we choose?
A) Left
B) Center
C) Right
D) Eight
Piazza Poll 1

Expectimax tree search:
Which action do we choose?

A) Left
B) Center
C) Right
D) Eight

\[ \hat{a} = \arg\max_a \sum_{s'} P(s'|s, a) V(s') \]
$$V(s) = \max_a \sum_{s'} P(s'|s,a) V(s')$$
Warm-up as You Walk In

Given

- Set actions (persistent/static)
- Set states (persistent/static)
- Function $T(s, a, s')$

Write the pseudo code for:

- function $V(s)$ return value

that implements:

$$V(s) = \max_{a \in \text{actions}} \sum_{s' \in \text{states}} T(s, a, s') V(s')$$
MDP Notation

Standard expectimax:

\[ V(s) = \max_a \sum_{s'} P(s'|s,a)V(s') \]

Bellman equations:

\[ V(s) = \max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V(s')] \]

Value iteration:

\[ V_{k+1}(s) = \max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V_k(s')], \quad \forall \ s \]

Q-iteration:

\[ Q_{k+1}(s,a) = \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma \max_{a'} Q_k(s',a')], \quad \forall \ s, a \]

Policy extraction:

\[ \pi_V(s) = \arg\max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V(s')], \quad \forall \ s \]

Policy evaluation:

\[ V^\pi_{k+1}(s) = \sum_{s'} P(s'|s,\pi(s))[R(s,\pi(s),s') + \gamma V^\pi_k(s')], \quad \forall \ s \]

Policy improvement:

\[ \pi_{\text{new}}(s) = \arg\max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^{\pi_{\text{old}}}(s')], \quad \forall \ s \]
MDP Notation

Standard expectimax:

\[ V(s) = \max_a \sum_{s'} P(s'|s,a) V(s') \]

Bellman equations:

\[ V(s) = \max_a \sum_{s'} P(s'|s,a) [R(s,a,s') + \gamma V(s')] \]

Value iteration:

\[ V_{k+1}(s) = \max_a \sum_{s'} P(s'|s,a) [R(s,a,s') + \gamma V_k(s')], \quad \forall s \]

Q-iteration:

\[ Q_{k+1}(s,a) = \sum_{s'} P(s'|s,a) [R(s,a,s') + \gamma \max_{a'} Q_k(s',a')], \quad \forall s, a \]

Policy extraction:

\[ \pi_V(s) = \arg\max_a \sum_{s'} P(s'|s,a) [R(s,a,s') + \gamma V(s')], \quad \forall s \]

Policy evaluation:

\[ V^\pi_{k+1}(s) = \sum_{s'} P(s'|s,\pi(s)) [R(s,\pi(s),s') + \gamma V^\pi_k(s')], \quad \forall s \]

Policy improvement:

\[ \pi_{new}(s) = \arg\max_a \sum_{s'} P(s'|s,a) [R(s,a,s') + \gamma V^{\pi_{old}}(s')], \quad \forall s \]
Non-Deterministic Search
Example: Grid World

- A maze-like problem
  - The agent lives in a grid
  - Walls block the agent’s path

- Noisy movement: actions do not always go as planned
  - 80% of the time, the action North takes the agent North (if there is no wall there)
  - 10% of the time, North takes the agent West; 10% East
  - If there is a wall in the direction the agent would have been taken, the agent stays put

- The agent receives rewards each time step
  - Small “living” reward each step (can be negative)
  - Big rewards come at the end (good or bad)

- Goal: maximize sum of rewards
Grid World Actions

Deterministic Grid World

Stochastic Grid World
Markov Decision Processes

An MDP is defined by:
- A set of states \( s \in S \)
- A set of actions \( a \in A \)
- A transition function \( T(s, a, s') \)
  - Probability that \( a \) from \( s \) leads to \( s' \), i.e., \( P(s' | s, a) \)
  - Also called the model or the dynamics
- A reward function \( R(s, a, s') \)
  - Sometimes just \( R(s) \) or \( R(s') \)
- A start state
- Maybe a terminal state

MDPs are non-deterministic search problems
- One way to solve them is with expectimax search
- We’ll have a new tool soon

[Demo – gridworld manual intro (L8D1)]
Demo of Gridworld
What is Markov about MDPs?

“Markov” generally means that given the present state, the future and the past are independent.

For Markov decision processes, “Markov” means action outcomes depend only on the current state:

\[
P(S_{t+1} = s' | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1}, \ldots S_0 = s_0) = P(S_{t+1} = s' | S_t = s_t, A_t = a_t)
\]

This is just like search, where the successor function could only depend on the current state (not the history).
Policies

In deterministic single-agent search problems, we wanted an optimal plan, or sequence of actions, from start to a goal.

For MDPs, we want an optimal policy \( \pi^* : S \to A \)
- A policy \( \pi \) gives an action for each state
- An optimal policy is one that maximizes expected utility if followed
- An explicit policy defines a reflex agent

Expectimax didn’t compute entire policies
- It computed the action for a single state only

Optimal policy when \( R(s, a, s') = -0.03 \) for all non-terminals \( s \)
Which sequence of optimal policies matches the following sequence of living rewards: 
{-0.01, -0.03, -0.04, -2.0}

I. {A, B, C, D}
II. {B, C, A, D}
III. {D, C, B, A}
IV. {D, A, C, B}

Piazza Poll 2
Optimal Policies

\[ R(s) = -2.0 \]

\[ R(s) = -0.4 \]

\[ R(s) = -0.03 \]

\[ R(s) = -0.01 \]
Example: Racing
Example: Racing

A robot car wants to travel far, quickly

Three states: Cool, Warm, Overheated

Two actions: Slow, Fast

Going faster gets double reward
Racing Search Tree
MDP Search Trees

Each MDP state projects an expectimax-like search tree

(s, a, s') called a transition

T(s, a, s') = P(s' | s, a)

R(s, a, s')
Utilities of Sequences
Utilities of Sequences

What preferences should an agent have over reward sequences?

More or less? \[1, 2, 2\] or \[2, 3, 4\]

Now or later? \[0, 0, 1\] or \[1, 0, 0\]
Discounting

It’s reasonable to maximize the sum of rewards
It’s also reasonable to prefer rewards now to rewards later
One solution: values of rewards decay exponentially

Worth Now
Worth Next Step
Worth In Two Steps
Discounting

How to discount?
- Each time we descend a level, we multiply in the discount once

Why discount?
- Sooner rewards probably do have higher utility than later rewards
- Also helps our algorithms converge
Piazza Poll 3

What is the value of $U[2, 4, 8]$ with $\gamma = 0.5$?

A. 3  
B. 6  
C. 7  
D. 14

Bonus: What is the value of $U[8, 4, 2]$ with $\gamma = 0.5$?
Stationary Preferences

Theorem: if we assume stationary preferences:

\[ [a_1, a_2, \ldots] \succ [b_1, b_2, \ldots] \]
\[ \iff \]
\[ [r, a_1, a_2, \ldots] \succ [r, b_1, b_2, \ldots] \]

Then: there are only two ways to define utilities

- Additive utility:
  \[ U([r_0, r_1, r_2, \ldots]) = r_0 + r_1 + r_2 + \cdots \]
- Discounted utility:
  \[ U([r_0, r_1, r_2, \ldots]) = r_0 + \gamma r_1 + \gamma^2 r_2 \cdots \]
Discounting

Given:

- Actions: East, West, and Exit (only available in exit states a, e)
- Transitions: deterministic

For $\gamma = 1$, what is the optimal policy?

For $\gamma = 0.1$, what is the optimal policy?

For which $\gamma$ are West and East equally good when in state d?
Infinite Utilities?!

- Problem: What if the game lasts forever? Do we get infinite rewards?

- Solutions:
  - Finite horizon: (similar to depth-limited search)
    - Terminate episodes after a fixed $T$ steps (e.g. life)
    - Gives nonstationary policies ($\pi$ depends on time left)
  - Discounting: use $0 < \gamma < 1$
    \[
    U([r_0, \ldots, r_\infty]) = \sum_{t=0}^{\infty} \gamma^t r_t \leq R_{\text{max}}/(1 - \gamma)
    \]
    - Smaller $\gamma$ means smaller “horizon” – shorter term focus
  - Absorbing state: guarantee that for every policy, a terminal state will eventually be reached (like “overheated” for racing)
Recap: Defining MDPs

Markov decision processes:
- Set of states $S$
- Start state $s_0$
- Set of actions $A$
- Transitions $P(s'|s,a)$ (or $T(s,a,s')$)
- Rewards $R(s,a,s')$ (and discount $\gamma$)

MDP quantities so far:
- Policy = Choice of action for each state
- Utility = sum of (discounted) rewards
Solving MDPs
Optimal Quantities

- The value (utility) of a state $s$:
  \[ V^*(s) = \text{expected utility starting in } s \text{ and acting optimally} \]

- The value (utility) of a q-state $(s,a)$:
  \[ Q^*(s,a) = \text{expected utility starting out having taken action } a \text{ from state } s \text{ and (thereafter) acting optimally} \]

- The optimal policy:
  \[ \pi^*(s) = \text{optimal action from state } s \]
Snapshot of Demo – Gridworld V Values

VALUES AFTER 100 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
Snapshot of Demo – Gridworld Q Values

Q-VALUES AFTER 100 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
Values of States

Fundamental operation: compute the (expectimax) value of a state
- Expected utility under optimal action
- Average sum of (discounted) rewards
- This is just what expectimax computed!

Recursive definition of value:

\[ V^*(s) = \max_a Q^*(s, a) \]

\[ Q^*(s, a) = \sum_{s'} T(s, a, s') \left[ R(s, a, s') + \gamma V^*(s') \right] \]

\[ V^*(s) = \max_a \sum_{s'} T(s, a, s') \left[ R(s, a, s') + \gamma V^*(s') \right] \]
Racing Search Tree
Racing Search Tree
Racing Search Tree

We’re doing way too much work with expectimax!

Problem: States are repeated
- Idea: Only compute needed quantities once

Problem: Tree goes on forever
- Idea: Do a depth-limited computation, but with increasing depths until change is small
- Note: deep parts of the tree eventually don’t matter if $\gamma < 1$
Time-Limited Values

Key idea: time-limited values

Define $V_k(s)$ to be the optimal value of $s$ if the game ends in $k$ more time steps

- Equivalently, it’s what a depth-$k$ expectimax would give from $s$
k=0

Noise = 0.2
Discount = 0.9
Living reward = 0
$k = 1$

VALUES AFTER 1 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
k=2

Noise = 0.2
Discount = 0.9
Living reward = 0
\( k = 3 \)

<table>
<thead>
<tr>
<th>0.00</th>
<th>0.52</th>
<th>0.78</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.43</td>
<td>-1.00</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

VALUES AFTER 3 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
k=4

VALUES AFTER 4 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
k=5

VALUES AFTER 5 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
k=6

VALUES AFTER 6 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
$k=7$

VALUES AFTER 7 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
$k=8$

VALUES AFTER 8 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
$k=10$

Noise = 0.2
Discount = 0.9
Living reward = 0
$k=11$

VALUES AFTER 11 ITERATIONS

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.64</td>
<td>0.74</td>
<td>0.85</td>
<td>1.00</td>
</tr>
<tr>
<td>0.56</td>
<td>0.57</td>
<td>-1.00</td>
<td></td>
</tr>
<tr>
<td>0.48</td>
<td>0.42</td>
<td>0.47</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Noise = 0.2
Discount = 0.9
Living reward = 0
$k=12$

VALUES AFTER 12 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
k=100

VALUES AFTER 100 ITERATIONS

Noise = 0.2
Discount = 0.9
Living reward = 0
Computing Time-Limited Values

$V_4(\ ) \ V_4(\ ) \ V_4(\ )$

$V_3(\ ) \ V_3(\ ) \ V_3(\ )$

$V_2(\ ) \ V_2(\ ) \ V_2(\ )$

$V_1(\ ) \ V_1(\ ) \ V_1(\ )$

$V_0(\ ) \ V_0(\ ) \ V_0(\ )$
Value Iteration
Value Iteration

Start with $V_0(s) = 0$: no time steps left means an expected reward sum of zero

Given vector of $V_k(s)$ values, do one ply of expectimax from each state:

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s') \left[ R(s, a, s') + \gamma V_k(s') \right]$$

Repeat until convergence

Complexity of each iteration: $O(S^2A)$

Theorem: will converge to unique optimal values

- Basic idea: approximations get refined towards optimal values
- Policy may converge long before values do
Example: Value Iteration

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_2$</td>
<td>3.5</td>
<td>2.5</td>
<td>0</td>
</tr>
<tr>
<td>$V_1$</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$V_0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Assume no discount!

\[
V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s') \left[ R(s, a, s') + \gamma V_k(s') \right]
\]
Convergence

How do we know the $V_k$ vectors are going to converge?

Case 1: If the tree has maximum depth $M$, then $V_M$ holds the actual untruncated values

Case 2: If the discount is less than 1
- Sketch: For any state $V_k$ and $V_{k+1}$ can be viewed as depth $k+1$ expectimax results in nearly identical search trees
- The difference is that on the bottom layer, $V_{k+1}$ has actual rewards while $V_k$ has zeros
- That last layer is at best all $R_{\text{MAX}}$
- It is at worst $R_{\text{MIN}}$
- But everything is discounted by $\gamma^k$ that far out
- So $V_k$ and $V_{k+1}$ are at most $\gamma^k \max|R|$ different
- So as $k$ increases, the values converge
MDP Notation

Standard expectimax:

\[ V(s) = \max_a \sum_{s'} P(s'|s, a) V(s') \]

Bellman equations:

\[ V(s) = \max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V(s')] \]

Value iteration:

\[ V_{k+1}(s) = \max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V_k(s')], \quad \forall s \]

Q-iteration:

\[ Q_{k+1}(s, a) = \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma \max_{a'} Q_k(s', a')], \quad \forall s, a \]

Policy extraction:

\[ \pi_V(s) = \arg\max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V(s')], \quad \forall s \]

Policy evaluation:

\[ V_{k+1}^\pi(s) = \sum_{s'} P(s'|s, \pi(s))[R(s, \pi(s), s') + \gamma V_k^\pi(s')], \quad \forall s \]

Policy improvement:

\[ \pi_{new}(s) = \arg\max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V_{\pi_{old}}(s')], \quad \forall s \]
MDP Notation

Standard expectimax:
\[ V(s) = \max_a \sum_{s'} P(s'|s, a)V(s') \]

Bellman equations:
\[ V(s) = \max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V(s')] \]

Value iteration:
\[ V_{k+1}(s) = \max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V_k(s')], \quad \forall \ s \]

Q-iteration:
\[ Q_{k+1}(s, a) = \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma \max_{a'} Q_k(s', a')], \quad \forall \ s, a \]

Policy extraction:
\[ \pi_V(s) = \arg\max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V(s')], \quad \forall \ s \]

Policy evaluation:
\[ V_{k+1}^\pi(s) = \sum_{s'} P(s'|s, \pi(s))[R(s, \pi(s), s') + \gamma V_k^\pi(s')], \quad \forall \ s \]

Policy improvement:
\[ \pi_{new}(s) = \arg\max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V_{old}(s')], \quad \forall \ s \]
Next Time: Policy-Based Methods