

Introduction to Partially Observable Markov Decision Processes (POMDPs)

CSCI 699 Computational Human-Robot Interaction

Instructor: Stefanos Nikolaidis

Decision Making Problems

	Environment Deterministic	Environment Non- Deterministic
State Known	A* search	Markov Decision Process
State Unknown	Partially Observable Markov Decision Process (special case)	Partially Observable Markov Decision Process

Decision Making Problems

	Environment Deterministic	Environment Non- Deterministic
State Known	A* search	Markov Decision Process
State Unknown	Partially Observable Markov Decision Process (special case)	Partially Observable Markov Decision Process

Markov Decision Process

- In an MDP, an agent interacts with the environment, by taking actions that induce a change in the state of the environment.
- An important assumption in MDPs is that the agent knows the true state of the world, e.g., trust.
- In reality, the true state of the world is not fully observable.

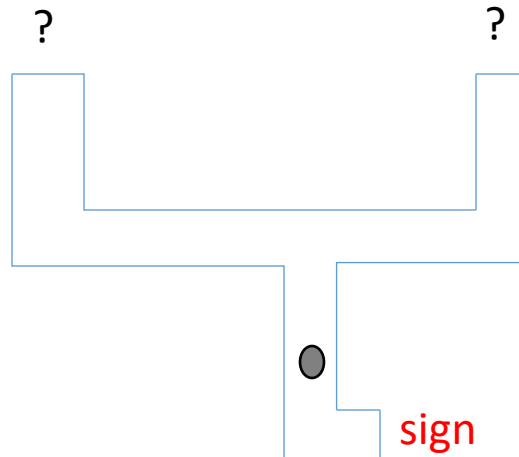
Markov Decision Process

- Example: inferring user intent in robotic wheelchair applications



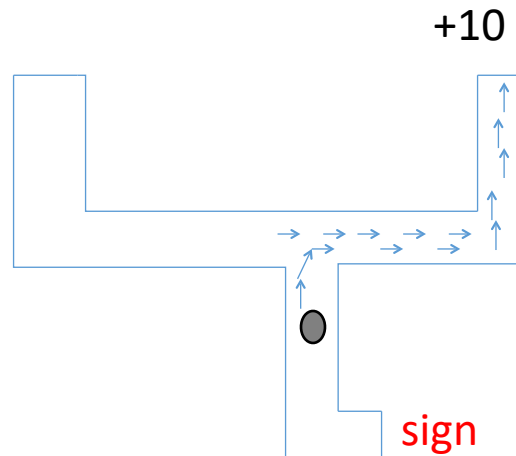
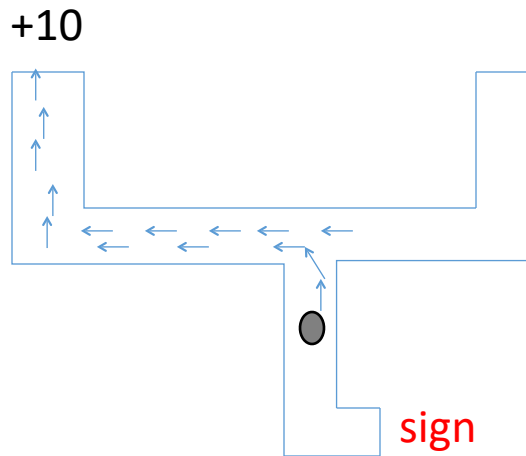
Setting

- Robot navigates in an office building
 - Stochastic environment
 - Partially observable



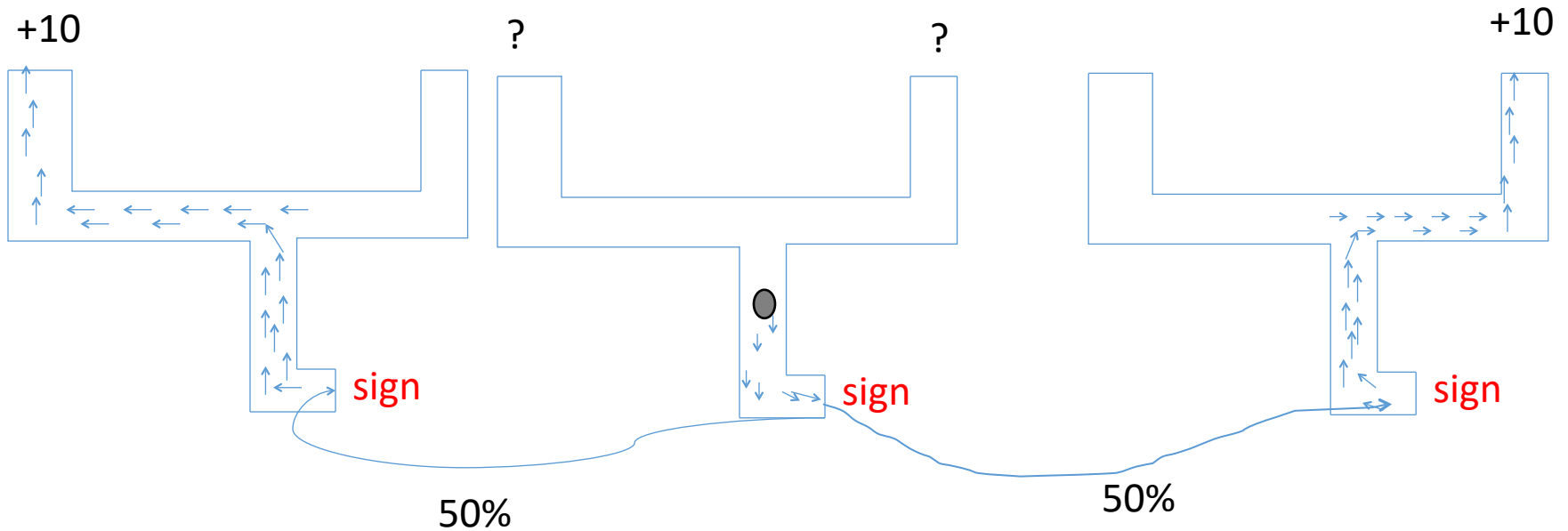
Setting

- Robot navigates in an office building
 - Stochastic environment
 - Partially observable



Setting

- Robot navigates in an office building
 - Stochastic environment
 - Partially observable



Partially Observable Markov Decision Process

- The robot will maintain a *belief state*, which is a probability distribution over states.
- The robot starts with an initial belief, it chooses an action, receives an observation and computes a new belief based on that observation
- For each state and action, the agent receives a reward
- As in the case of MDPs, the goal is to maximize the expected accumulated reward that it will receive over a time horizon

POMDP framework

A partially observable Markov decision process can be described as a tuple $\langle S, A, T, R, \Omega, O \rangle$, where

- S, A, T and R describe an MDP
- Ω is a finite set of observations
- $O: S \times A \rightarrow \Pi(\Omega)$ is the *observation function*

The agent maintains a *belief state* b

Policy execution

- Given the current belief state b , execute the action $a = \pi^*(b)$
- Receive observation o
- Calculate belief b' and set current belief to b'

Belief Update

$$b'(s') = P(s' | o, a, b)$$

Belief Update

$$\begin{aligned} b'(s') &= P(s' | o, a, b) \\ &= \frac{P(o | s', a, b) P(s' | a, b)}{P(o | a, b)} \end{aligned}$$

(from Bayes' rule)

Belief Update

$$\begin{aligned} b'(s') &= P(s' | o, a, b) \\ &= \frac{P(o | s', a, b) P(s' | a, b)}{P(o | a, b)} \\ &= \frac{P(o | s', a, b) \sum_{s \in \mathcal{S}} P(s' | a, b, s) P(s | a, b)}{P(o | a, b)} \end{aligned}$$

(from Bayes' rule)

(marginalization)

Belief Update

$$\begin{aligned} b'(s') &= P(s' | o, a, b) \\ &= \frac{P(o | s', a, b) P(s' | a, b)}{P(o | a, b)} && \text{(from Bayes' rule)} \\ &= \frac{P(o | s', a, b) \sum_{s \in \mathcal{S}} P(s' | a, b, s) P(s | a, b)}{P(o | a, b)} && \text{(marginalization)} \\ &= \frac{P(o | s', a) \sum_{s \in \mathcal{S}} P(s' | a, s) P(s | b)}{P(o | a, b)} && \text{(from conditional independence)} \end{aligned}$$

Belief Update

$$\begin{aligned} b'(s') &= P(s' | o, a, b) \\ &= \frac{P(o | s', a, b) P(s' | a, b)}{P(o | a, b)} && \text{(from Bayes' rule)} \\ &= \frac{P(o | s', a, b) \sum_{s \in \mathcal{S}} P(s' | a, b, s) P(s | a, b)}{P(o | a, b)} && \text{(marginalization)} \\ &= \frac{P(o | s', a) \sum_{s \in \mathcal{S}} P(s' | a, s) P(s | b)}{P(o | a, b)} && \text{(from conditional independence)} \\ &= \frac{O(s', a, o) \sum_{s \in \mathcal{S}} \mathcal{T}(s, a, s') b(s)}{P(o | a, b)} && \text{(by definition of } \mathcal{T}, O, b) \end{aligned}$$

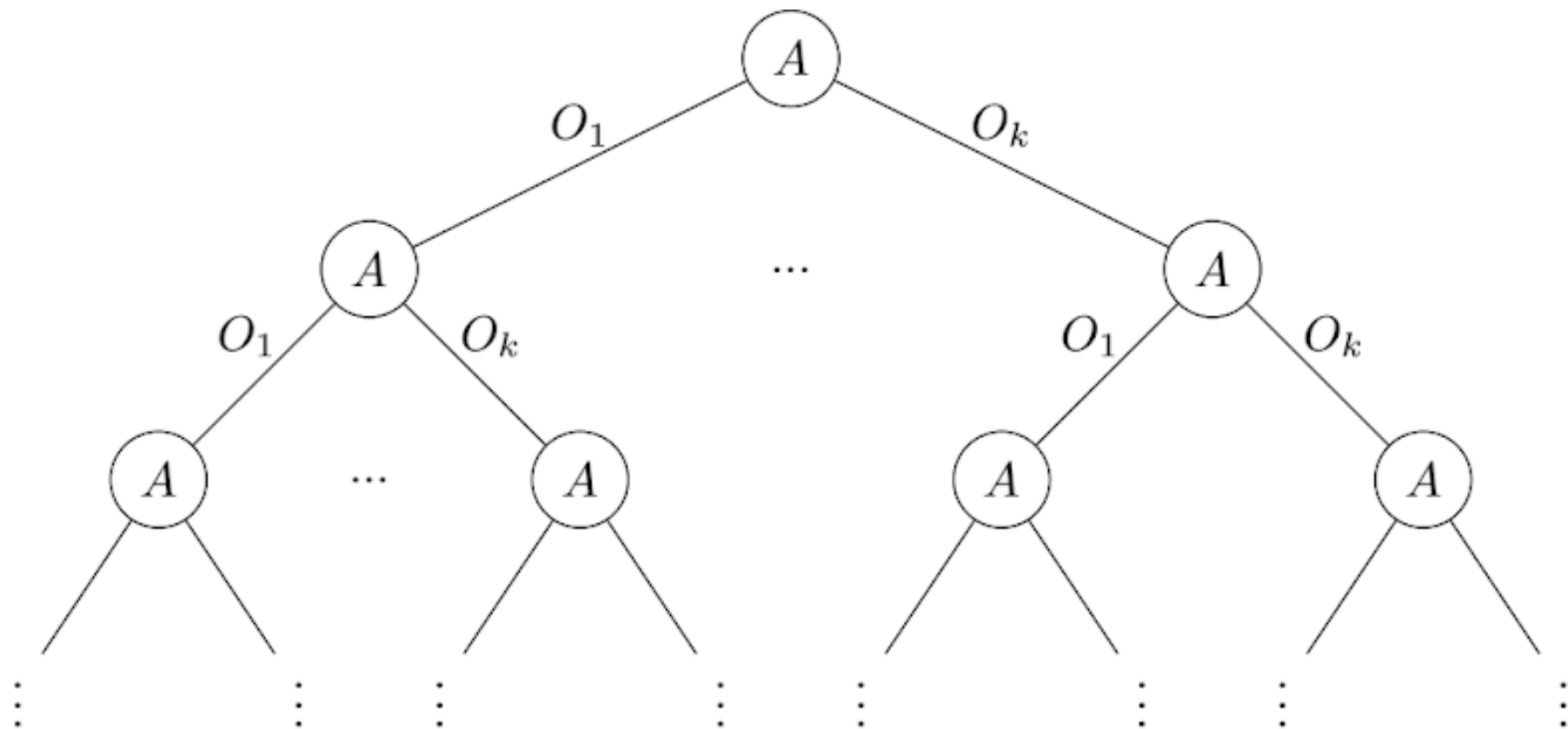
Policy execution

- Given the current belief state b , execute the action $a = \pi^*(b)$
- Receive observation o
- Calculate belief b' and set current belief to b'

Meaning of a policy

- For one timestep, take an action
- For 2 timesteps, take one action, then the next action would depend on the observation we receive
- A policy tree is a tree describing sequences of actions and observations

Policy Tree

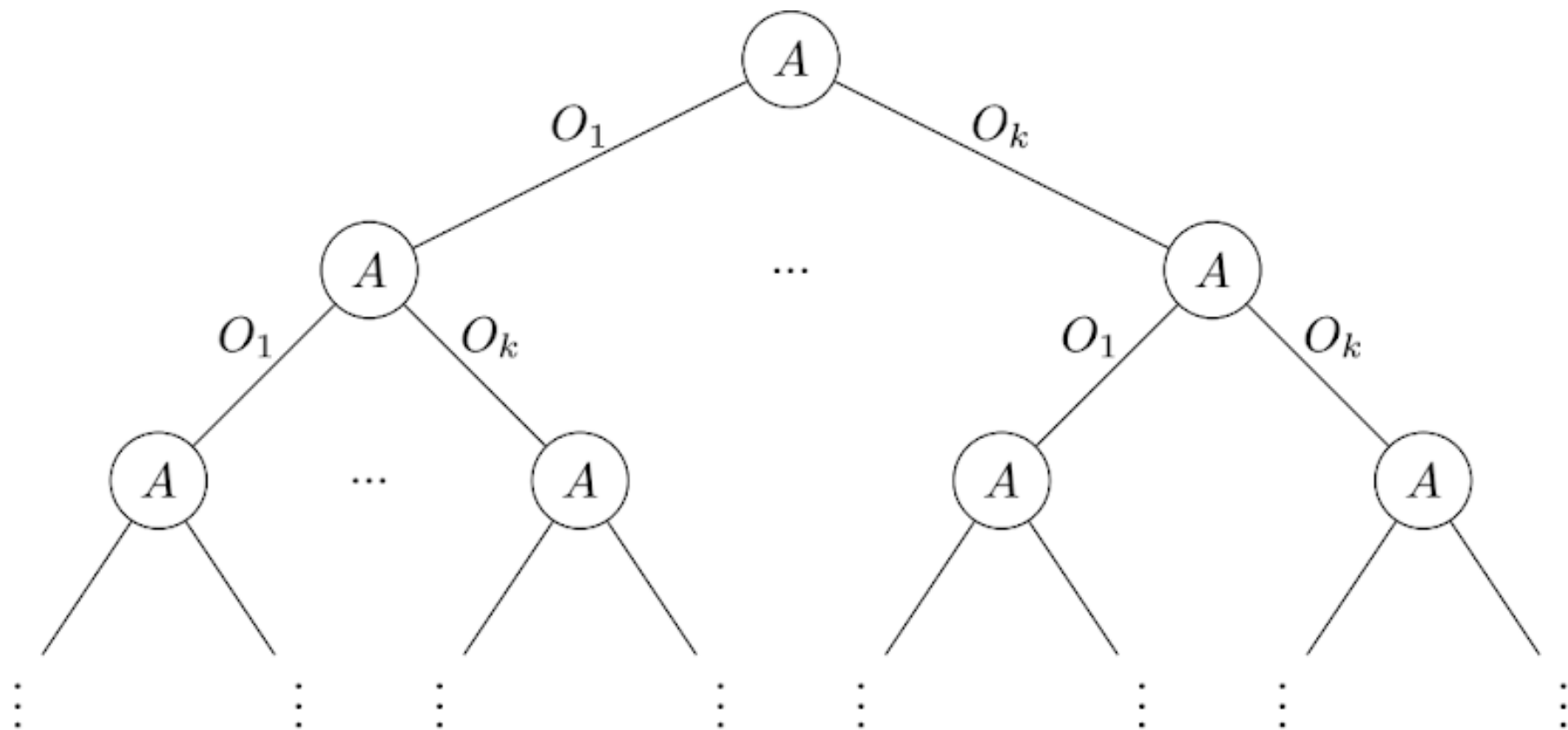


Optimal Policy

- To compute the optimal policy, we need a metric of how good a given policy tree p is.
- The value of executing a policy tree p in state s is the immediate reward by executing the action at the root node of the tree, plus the expected value of the future.

$$V_p(s) = R(s, a(p)) + \gamma \cdot (\textit{Expected value of the future})$$

Optimal Policy



$$V_t^p(s) = R(s, a(p)) + \sum_{s' \in \mathcal{S}} P(s'|s, a(p)) \sum_{o_i \in \Omega} P(o_i|s', a(p)) V_{t-1}^{p, o_i}(s')$$

Value of executing a policy given a belief b

$$V_t^p(b) = \sum_{s \in S} b(s) V_t^p(s)$$

Example

- Everytime there is a crossing, the wheelchair wants to understand whether the user wants to go.
- S: Left / Right
- A: (ask, GL, GR)
- T: $S \times A \rightarrow \Pi(S)$



Example

- Everytime there is a crossing, the wheelchair wants to understand whether the user wants to go.
- S: Left / Right
- A: (ask, GL, GR)
- T: $S \times A \rightarrow \Pi(S)$

Ask	<i>S</i>	<i>L</i>	<i>R</i>
		<i>L</i>	<i>R</i>
<i>S</i>	<i>L</i>	1.0	0.0
	<i>R</i>	0.0	1.0



Example

- Everytime there is a crossing, the wheelchair wants to understand whether the user wants to go.
- S: Left / Right
- A: (ask, GL, GR)
- T: $S \times A \rightarrow \Pi(S)$

			GL		S
				L'	R
	S	L	0.5	0.5	
		R	0.5	0.5	



Example

- Everytime there is a crossing, the wheelchair wants to understand whether the user wants to go.
- S: Left / Right
- A: (ask, GL, GR)
- T: $S \times A \rightarrow \Pi(S)$



Example

- Everytime there is a crossing, the wheelchair wants to understand whether the user wants to go.
- S: Left / Right
- A: (ask, GL, GR)
- T: $S \times A \rightarrow \Pi(S)$
- Ω : {ML, MR}
- O: $S \times A \rightarrow \Omega$



Example

- Everytime there is a crossing, the wheelchair wants to understand whether the user wants to go.
- S: Left / Right
- A: (ask, GL, GR)
- T: $S \times A \rightarrow \Pi(S)$
- Ω : {ML, MR}
- O: $S \times A \rightarrow \Omega$
 - If L and action is “ask”, observe ML with prob 0.9
 - equivalently for R



Example

- Everytime there is a crossing, the wheelchair wants to understand whether the user wants to go.

- S: Left / Right

- A: (ask, GL, GR)

- T: $S \times A \rightarrow \Pi(S)$

- Ω : {ML, MR}

- O: $S \times A \rightarrow \Omega$

- If L and action is “ask”, observe ML with prob 0.9

- equivalently for R

	<i>ML</i>	<i>MR</i>
<i>L</i>	0.9	0.1
<i>R</i>	0.1	0.9



Example

- Everytime there is a crossing, the wheelchair wants to understand whether the user wants to go.
- S: Left / Right
- A: (ask, GL, GR)
- T: $S \times A \rightarrow \Pi(S)$
- Ω : {ML, MR}
- O: $S \times A \rightarrow \Omega$
- R: $S \rightarrow \mathbb{R}$



Example

- Everytime there is a crossing, the wheelchair wants to understand whether the user wants to go.

- S: Left / Right
- A: (ask, GL, GR)
- T: $S \times A \rightarrow \Pi(S)$
- Ω : {ML, MR}
- O: $S \times A \rightarrow \Omega$
- R: $S \rightarrow \mathbb{R}$

$$R(s, ask) = -2$$

$$R(L, GL) = 10$$

$$R(L, GR) = -100$$

$$R(R, GR) = 10$$

$$R(R, GL) = -100$$



Example

- Let's start with $T = 1$

$$\begin{aligned} V_1(b) &= \sum_{s \in S} b(s) V_1^p(s) \\ &= \sum_{s \in S} b(s) R(s, a(p)) \end{aligned}$$

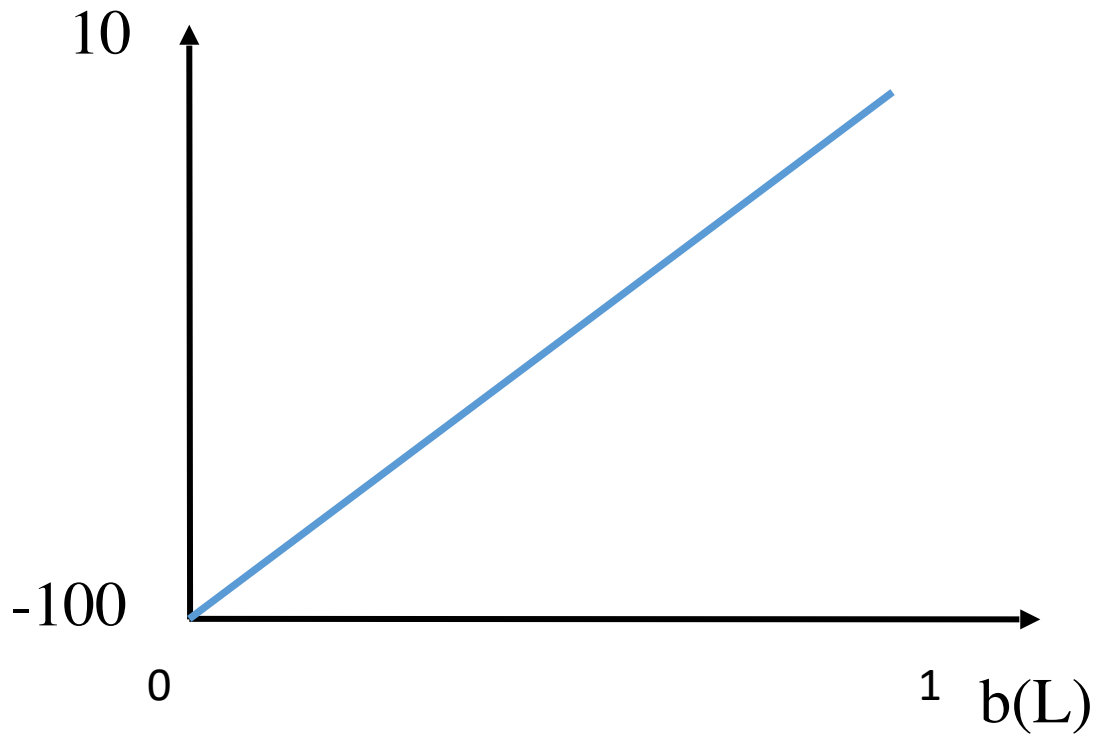
Example

- Let's start with $T = 1$, value of Go-Left (GL)



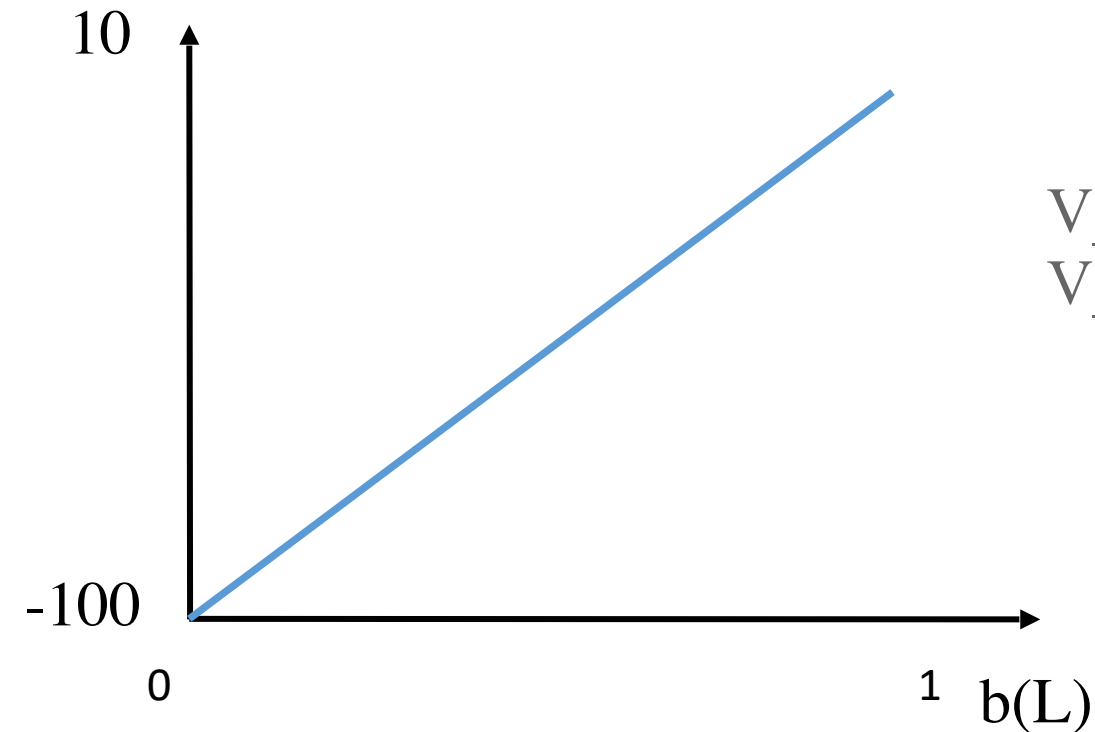
Example

- Let's start with $T = 1$, value of Go-Left (GL)



Example

- Let's start with $T = 1$, value of go-left (GL)



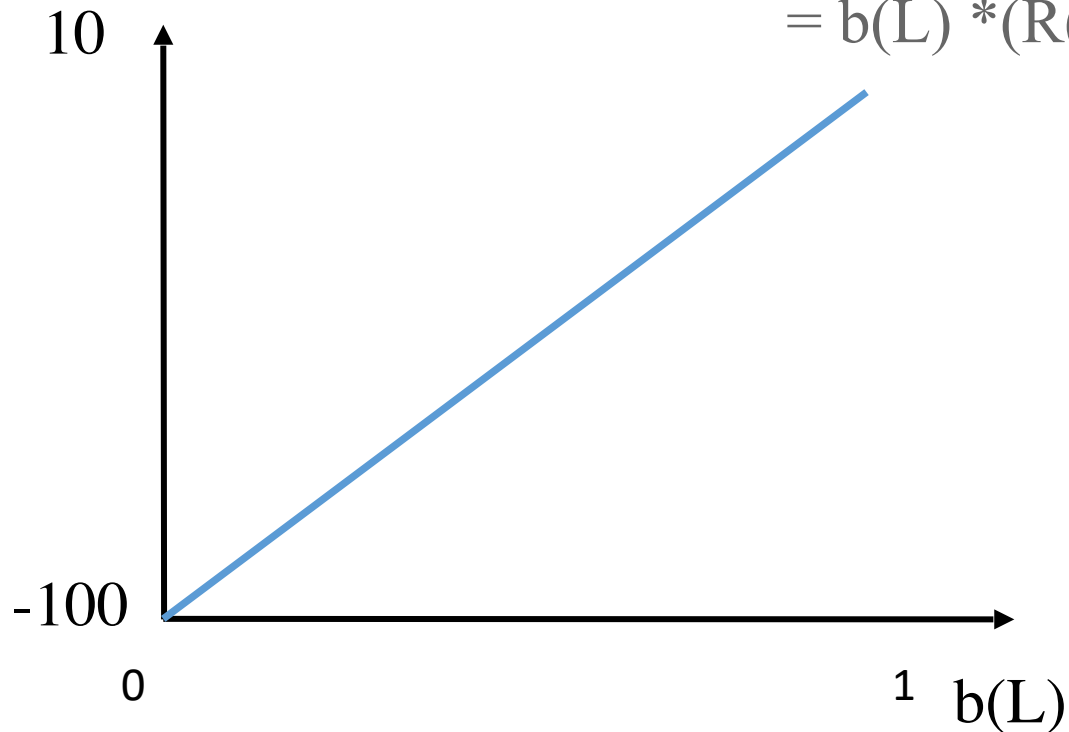
$$V_{-1}^{\{p1\}}(L) = R(L, GL) = 10$$

$$V_{-1}^{\{p1\}}(R) = R(R, GL) = -100$$

Example

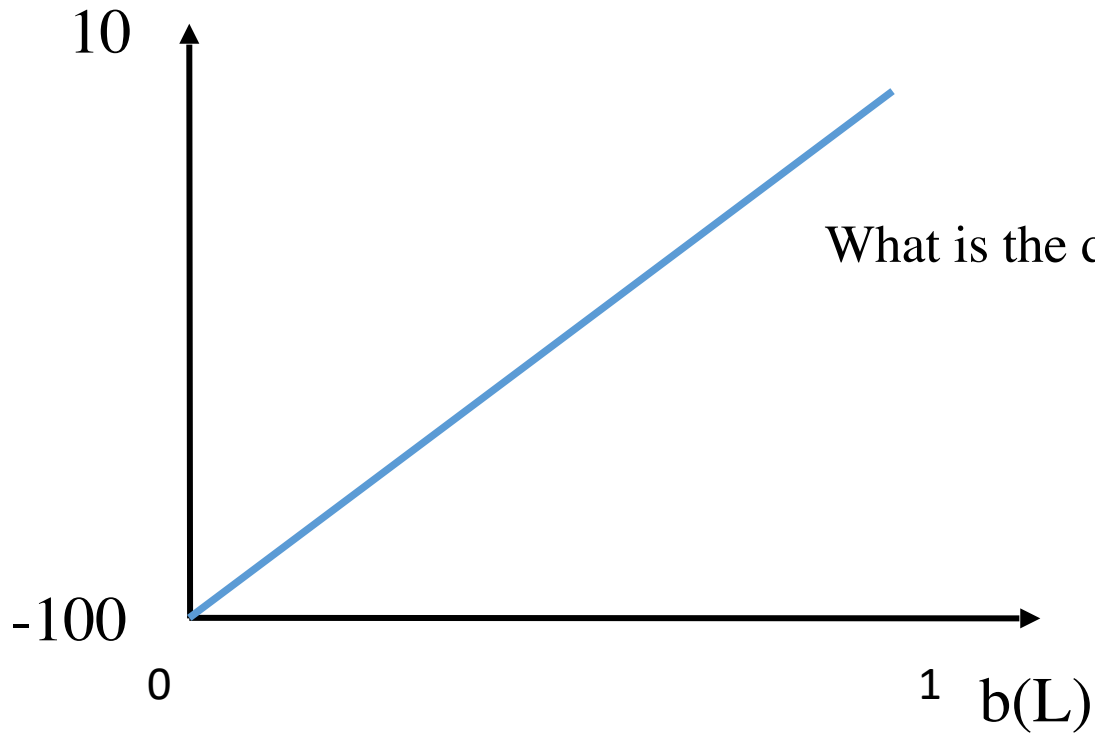
- Let's start with $T = 1$, value of go-left (GL)

$$\begin{aligned} V^{\{p1\}}(b) &= R(R, GL) * b(L) + R(R, GL) * (1-b(L)) \\ &= b(L) * (R(L, GL) - R(R, GL)) + R(R, GL) \end{aligned}$$



Example

- Let's start with $T = 1$, value of go-left (GL)



What is the dimensionality of the belief space?

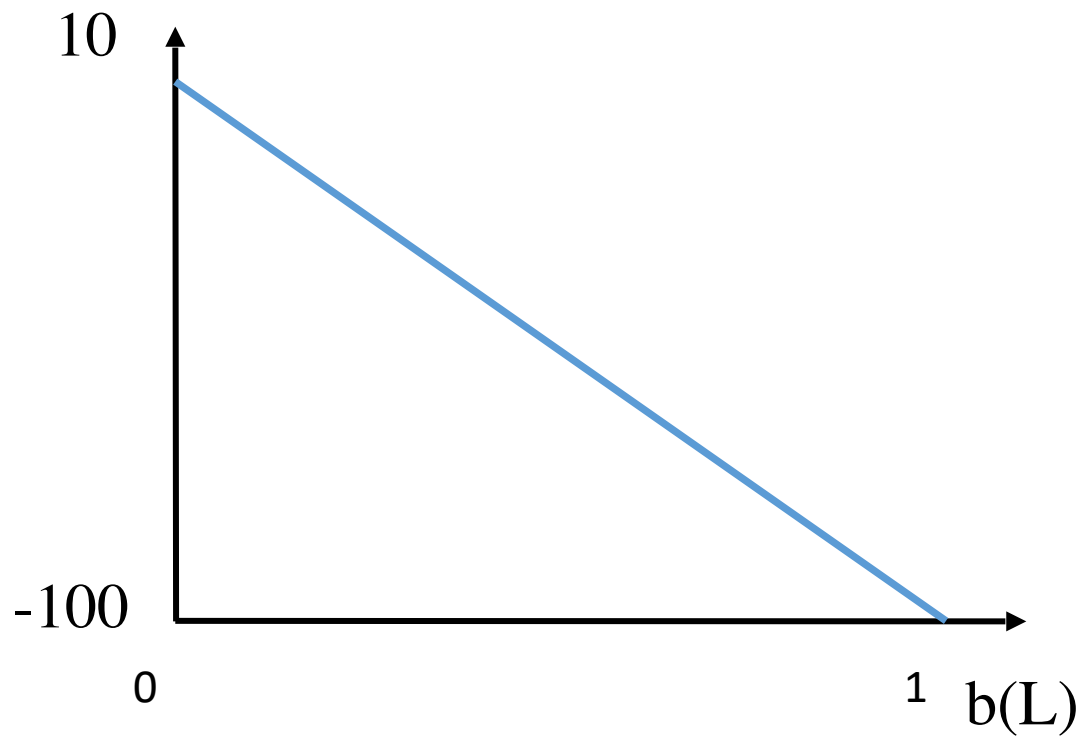
Example

- Let's start with $T = 1$, value of go-right (GR)



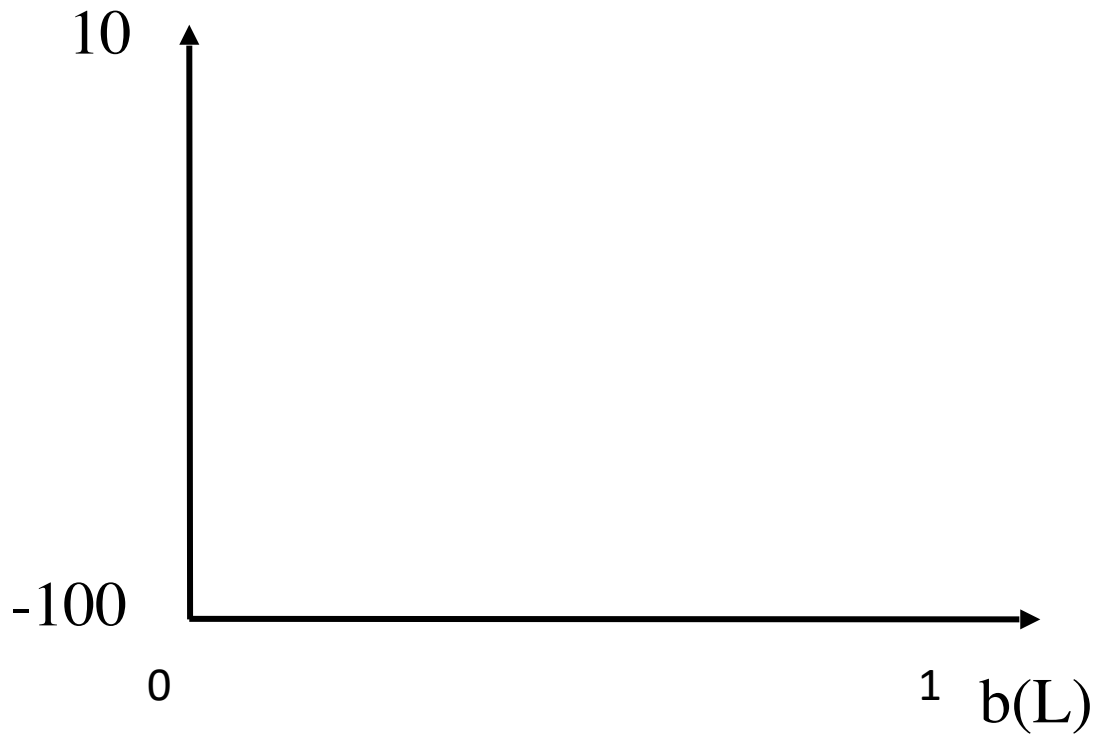
Example

- Let's start with $T = 1$, value of go-right (GR)



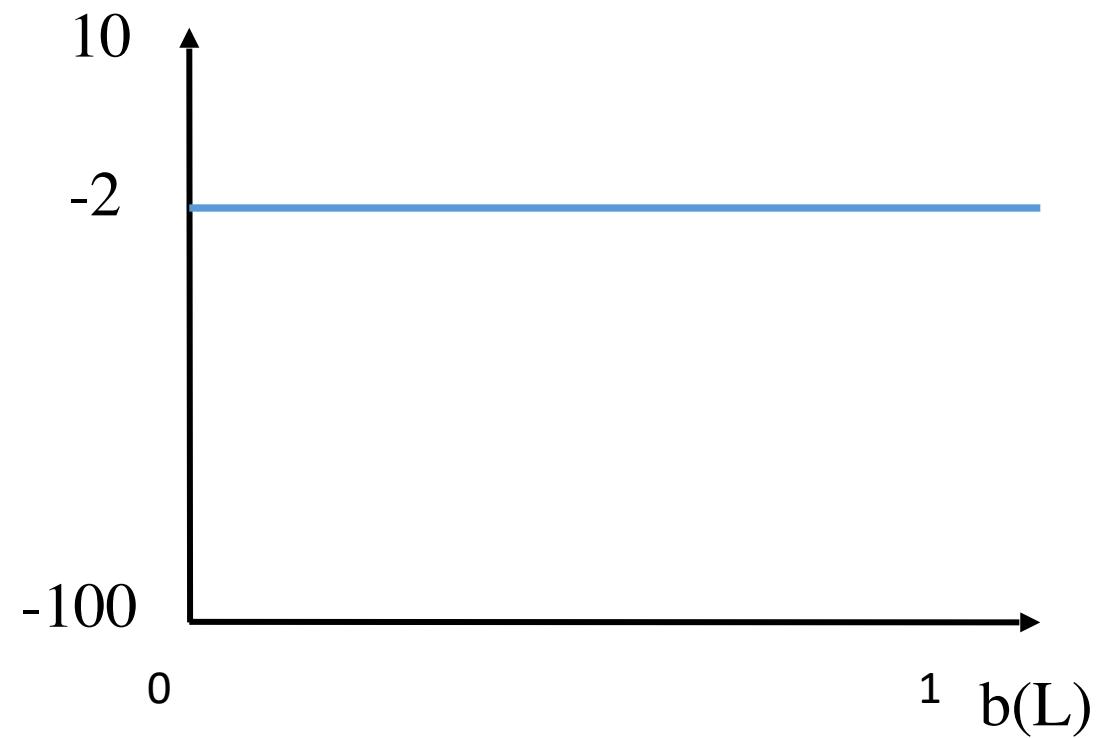
Example

- Let's start with $T = 1$, value of ask



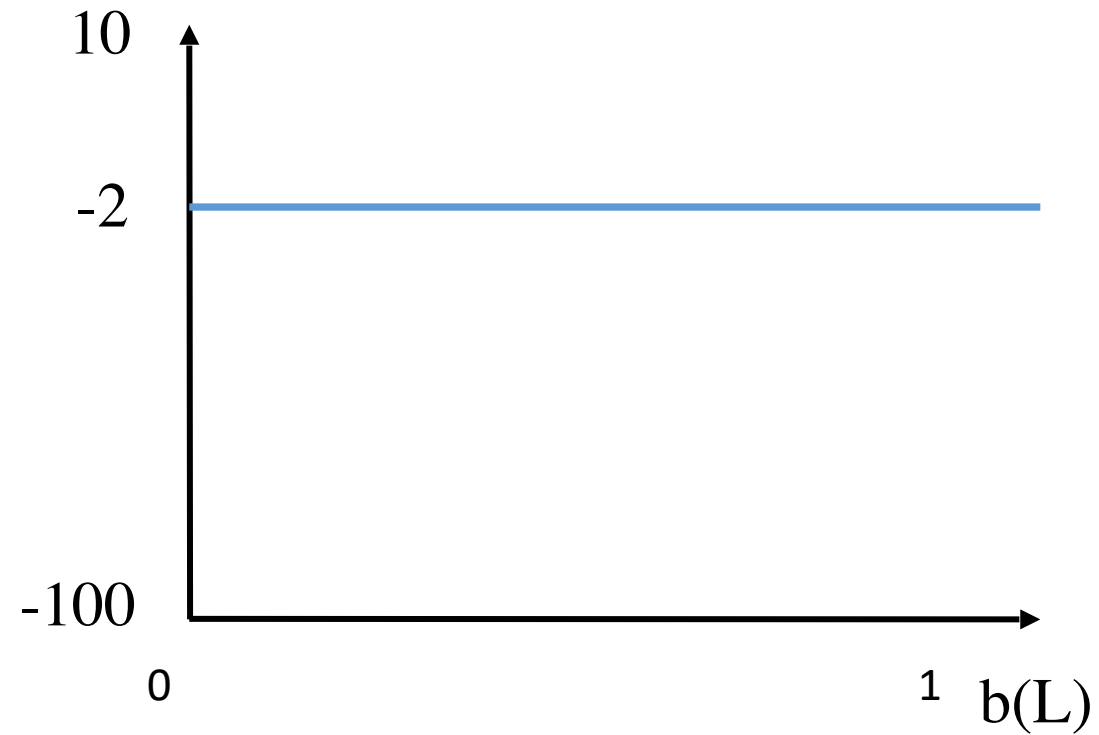
Example

- Let's start with $T = 1$, value of ask



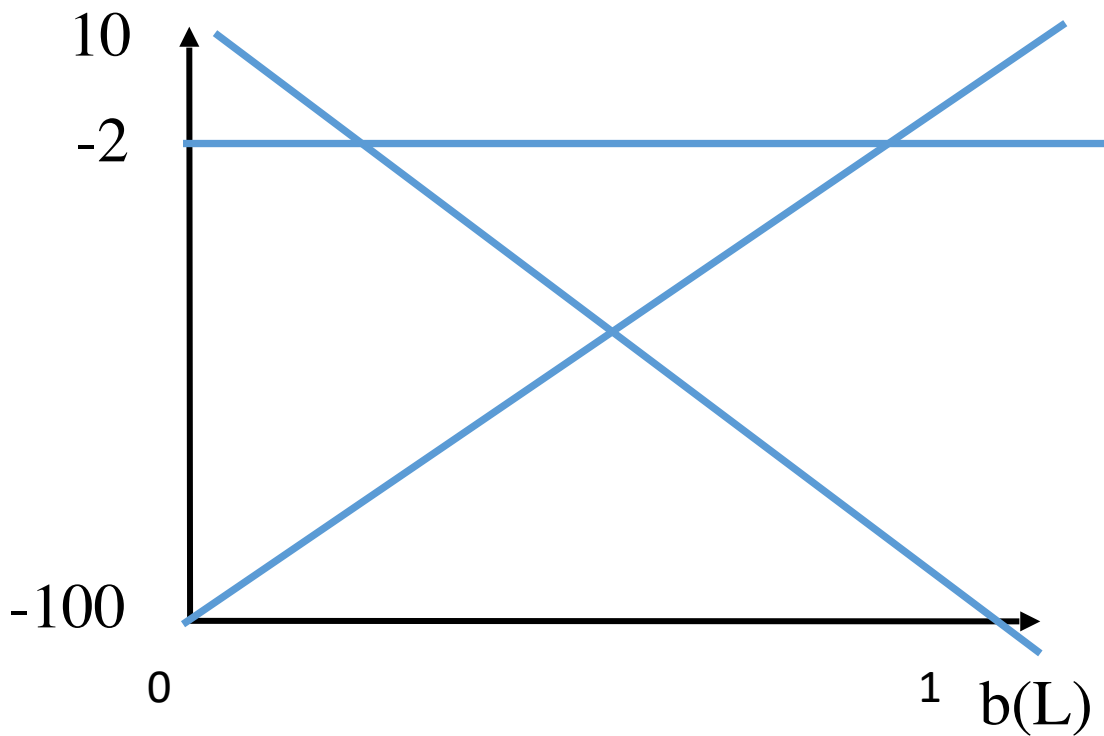
Example

- When is asking better than going left?



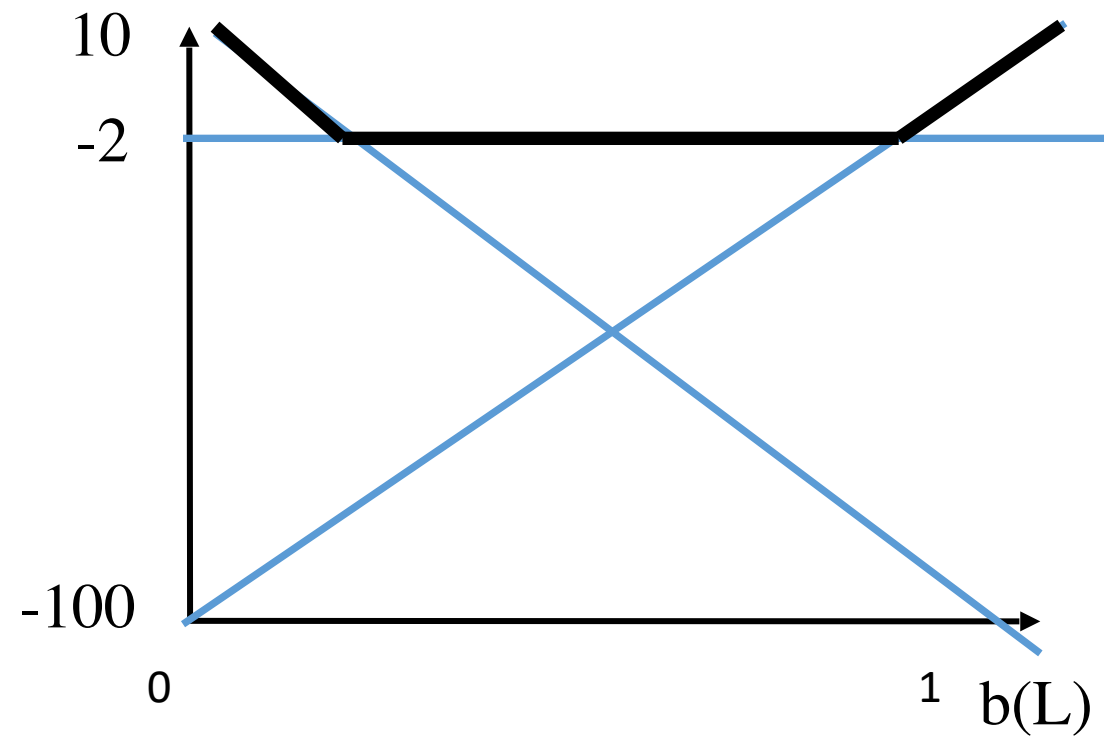
Example

- When is asking better than going left?



Example

- When is asking better than going left?



Example

- What about $T = 2$?

$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$

Example

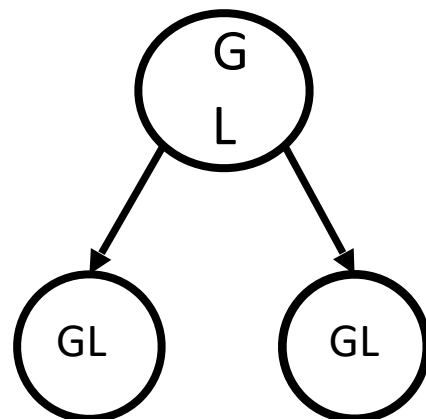
- What about $T = 2$?

$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$

- That's a two-step policy. It depends on the action at the first timestep, and the subtree on the second timestep. Now, how many possible subtrees we have?

Example

$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$



Example

$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$

$$\begin{aligned} V_2^{GL}(L) &= R(L, GL) + T(L|L, GL) * O(ML|L, GL) * V_1^{GL, ML}(L) \\ &\quad + T(L|L, GL) * O(MR|L, GL) * V_1^{GL, MR}(L) \\ &\quad + T(R|L, GL) * O(ML|R, GL) * V_1^{GL, ML}(R) \\ &\quad + T(R|L, GL) * O(MR|R, GL) * V_1^{GL, MR}(R) \end{aligned}$$

Example

$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$

$$\begin{aligned} V_2^{GL}(L) &= R(L, GL) + T(L|L, GL) * O(ML|L, GL) * V_1^{GL, ML}(L) \\ &\quad + T(L|L, GL) * O(MR|L, GL) * V_1^{GL, MR}(L) \\ &\quad + T(R|L, GL) * O(ML|R, GL) * V_1^{GL, ML}(R) \\ &\quad + T(R|L, GL) * O(MR|R, GL) * V_1^{GL, MR}(R) \end{aligned}$$

$$\begin{aligned} V_2^{GL}(L) &= 10 + 0.5 * 0.5 * 10 \\ &\quad + 0.5 * 0.5 * 10 \\ &\quad + 0.5 * 0.5 * (-100) \\ &\quad + 0.5 * 0.5 * (-100) \\ &= -35 \end{aligned}$$

Example

$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$

$$\begin{aligned} V_2^{GL}(R) &= R(R, GL) + T(L|R, GL) * O(ML|L, GL) * V_1^{GL, ML}(L) \\ &\quad + T(L|R, GL) * O(MR|L, GL) * V_1^{GL, MR}(L) \\ &\quad + T(R|R, GL) * O(ML|R, GL) * V_1^{GL, ML}(R) \\ &\quad + T(R|R, GL) * O(MR|R, GL) * V_1^{GL, MR}(R) \end{aligned}$$

Example

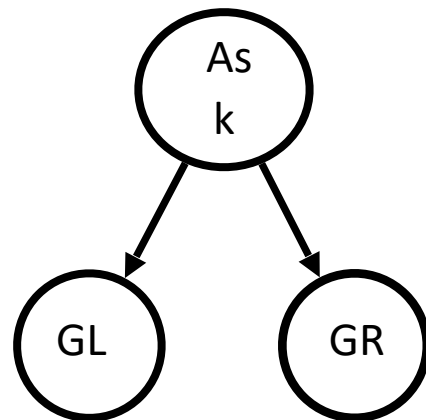
$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$

$$\begin{aligned} V_2^{GL}(R) &= R(R, GL) + T(L|R, GL) * O(ML|L, GL) * V_1^{GL, ML}(L) \\ &\quad + T(L|R, GL) * O(MR|L, GL) * V_1^{GL, MR}(L) \\ &\quad + T(R|R, GL) * O(ML|R, GL) * V_1^{GL, ML}(R) \\ &\quad + T(R|R, GL) * O(MR|R, GL) * V_1^{GL, MR}(R) \end{aligned}$$

$$\begin{aligned} V_2^{GL}(R) &= -100 + 0.5 * 0.5 * 10 \\ &\quad + 0.5 * 0.5 * 10 \\ &\quad + 0.5 * 0.5 * (-100) \\ &\quad + 0.5 * 0.5 * (-100) \\ &= -145 \end{aligned}$$

Example

$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$



Example

$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$

$$\begin{aligned} V_2^{ask}(L) &= R(L, ask) + T(L|L, ask) * O(ML|L, ask) * V_1^{GL, ML}(L) \\ &\quad + T(L|L, ask) * O(MR|L, ask) * V_1^{GR, MR}(L) \\ &\quad + T(R|L, ask) * O(ML|R, ask) * V_1^{GL, ML}(R) \\ &\quad + T(R|L, ask) * O(MR|R, ask) * V_1^{GR, MR}(R) \end{aligned}$$

Example

$$V_2^p(s) = R(s, a(p)) + \sum_{s' \in S} \mathcal{T}(s, a(p), s') \sum_{o_i \in \Omega} O(s', a(p), o_i) V_1^{p, o_i}(s')$$

$$\begin{aligned} V_2^{ask}(L) &= R(L, ask) + T(L|L, ask) * O(ML|L, ask) * V_1^{GL, ML}(L) \\ &\quad + T(L|L, ask) * O(MR|L, ask) * V_1^{GR, MR}(L) \\ &\quad + T(R|L, ask) * O(ML|R, ask) * V_1^{GL, ML}(R) \\ &\quad + T(R|L, ask) * O(MR|R, ask) * V_1^{GR, MR}(R) \end{aligned}$$

$$\begin{aligned} V_2^{ask}(L) &= -2 + 1 * 0.9 * 10 \\ &\quad + 1 * 0.1 * (-100) \\ &= -3 \end{aligned}$$

Example

