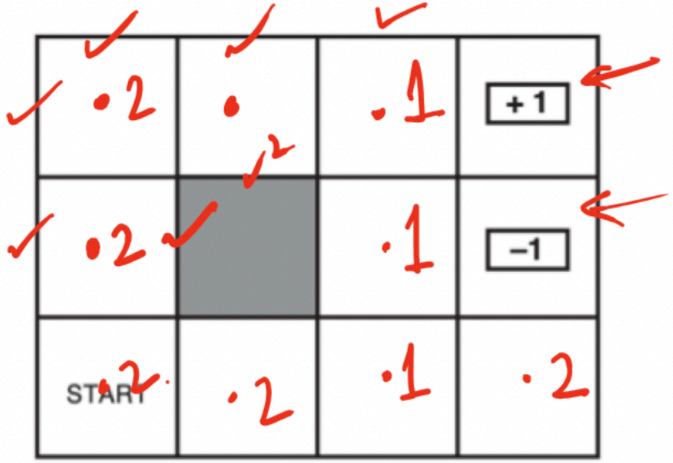
POMDP → Convert → MDP

transition $P(b' \mid b, a, e)$

- The 4 × 3 world can make use of a noisy sensor that can measure the number of adjacent walls, but may give a wrong value with probability 0.1



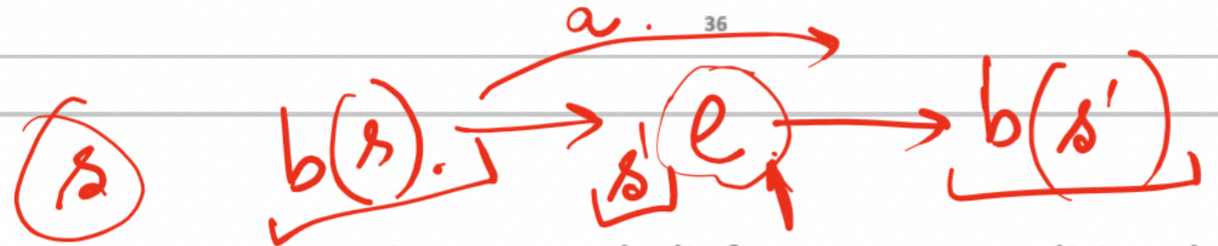
$P(s' \mid s, a)$

4×3.

$b(s)$

$P(b' \mid b, a, e)$

$P(e \mid s')$



- The agent can compute its current belief state as a conditional probability distribution over the given sequence of percepts and actions so far.
- If $b(s)$ was the previous belief state and the agent does action a and then perceives evidence e then the new belief state b' is given by

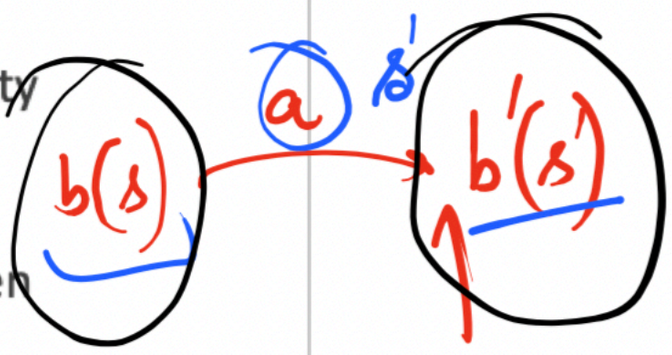
$$b'(s') = \alpha\, P(e \mid s') \sum_s P(s' \mid s, a)\, b(s)$$

$$b' = \text{FORWARD}(b, a, e)$$

deterministic

unnormalized

$$b'_{un}(s') = P(e \mid s') \sum_s P(s' \mid s, a)\, b(s)$$

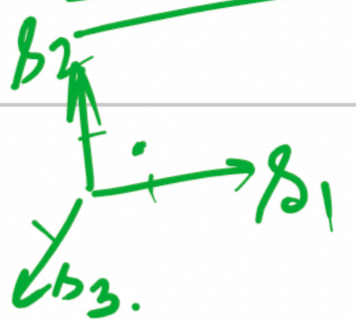


$P(e \mid s')$.

$b(s)$, $b'(s')$ are observable

- Given an evidence, the transition of the belief state is deterministic.

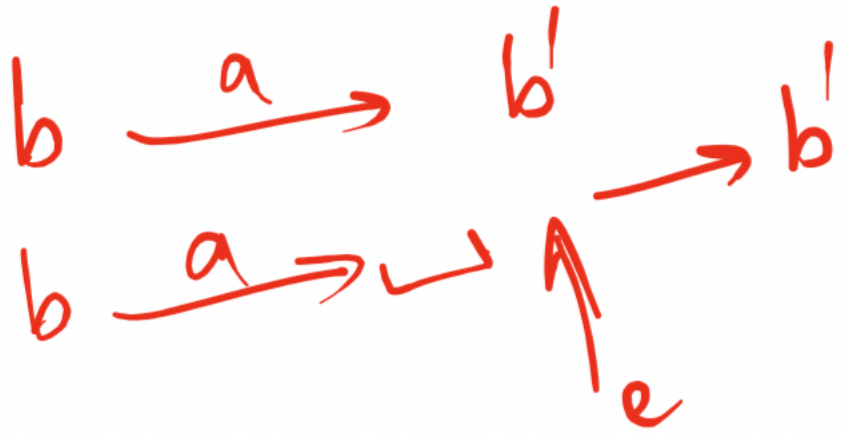
b, b' are continuous.



- The optimal action depends only on the agent's current belief state and is given by $\pi^*(b)$.

- The decision cycle of a POMDP agent is as follows:
  - Given the current belief state, execute the action $a = \pi^*(b)$
  - Receive percept e
  - Set the current belief state to FORWARD(b, a, e)
  - Repeat

- The POMDP belief state space is continuous because the belief state is a probability distribution
- An action changes the belief state and not just the physical state.



Not the physical. $P(s' \mid s, a)$, POMDP

Not the physical. $P(s' \mid s, a)$ POMDP

Transition Model for the belief state space

- Without the evidence e, the transition of the belief state space is non-deterministic

$$P(b' \mid b, a) = P(b' \mid a, b) = \sum_e P(b' \mid e, a, b)\, P(e \mid a, b)$$

$$P(e \mid a, b) = \sum_{s'} P(e \mid a, s', b)\, P(s' \mid a, b)$$

$$= \sum_{s'} P(e \mid s')\, P(s' \mid a, b)$$

$$= \sum_{s'} P(e \mid s') \sum_{s} P(s' \mid s, a)\, b(s) = \sum_{s'} b'_{un}(s')$$

$P(e \mid s)$ noisy sensor model.

b, a → b'?

Transition Model for the belief state space

MDP

- The probability of reaching b' from b, given action a is $P(b' \mid b, a)$

$$P(b' \mid b, a) = \sum_e P(b' \mid e, a, b)\, P(e \mid a, b)$$

$$= \sum_e P(b' \mid e, a, b) \sum_{s'} P(e \mid s') \sum_s P(s' \mid a)\, b(s)$$

- Where $P(b' \mid e, a, b) = 1$ if $b' = \text{Forward}(b, a, e)$ and 0 otherwise.

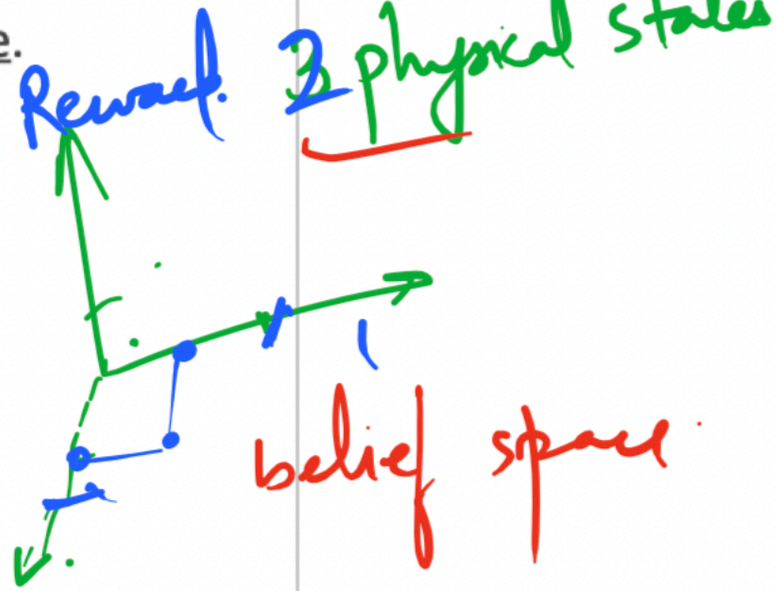
$$= \sum_e P(b' \mid e, a, b) \sum_{s'} b'_{un}(s')$$

Transforming partially observable MDP to observable MDP

- We transform the problem from physical state space to belief state space.
- Transition model in the belief state space $P(b' \mid b, a)$
- The reward function on the belief state space can be defined as

$$\rho(b) = \sum_s b(s) R(s)$$

- The optimal policy for this MDP is $\pi^*(b)$
- The belief state is always observable to the agent.
- BUT, this observable MDP has a high-dimensional continuous state space.



Best action

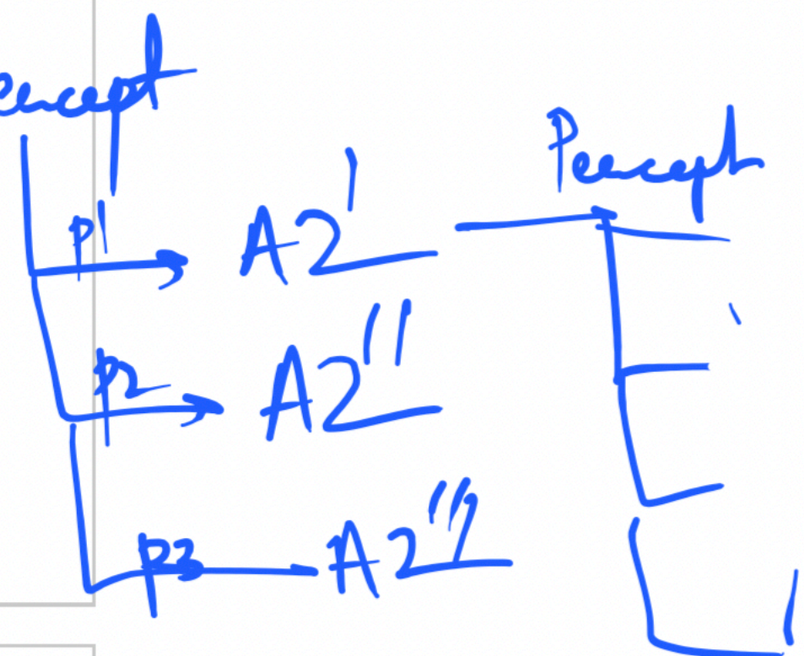
$\pi^*(b)$

plan

Value Iteration for POMDPs

Non trivial.

$\pi(b)$



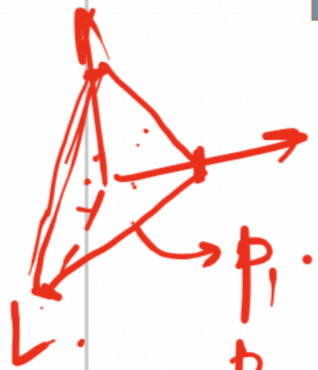
action plan.

What is $\pi^*(b)$?

- The optimal policy for a belief state space is about choosing a conditional plan p.
- The belief state is updated based on the action as well as the percept.
- Let the utility of executing a fixed conditional plan p starting in physical state s is $\alpha_p(s)$
- The expected utility of executing p in belief state b is

$$\sum_s b(s)\,\alpha_p(s) \quad \text{or} \quad b \cdot \alpha_p$$

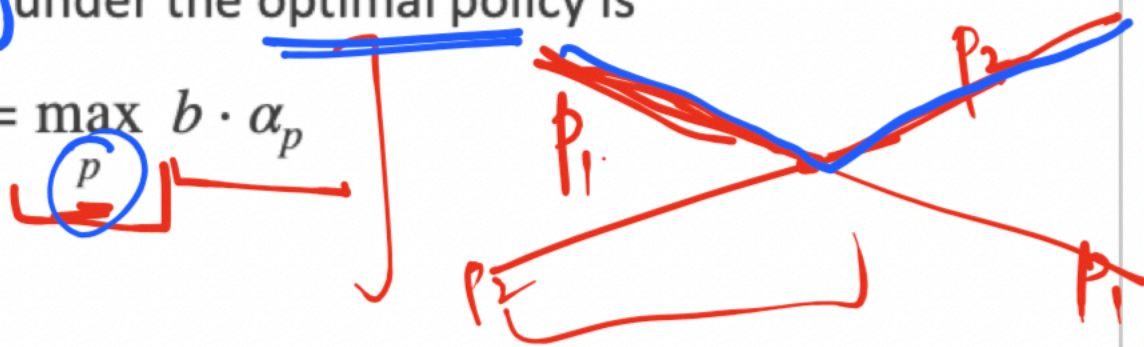
- The expected utility of plan p is a hyperplane in the belief space.



Optimal Policy $\pi^*(b)$

- The optimal policy in the belief state space chooses a conditional plan with highest expected utility.
- The expected utility of the belief state under the optimal policy is

$$U(b) = U^{\pi^*}(b) = \max_p \; b \cdot \alpha_p$$

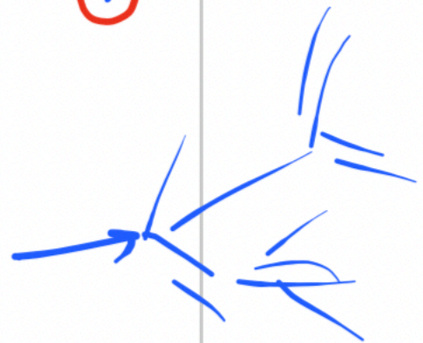


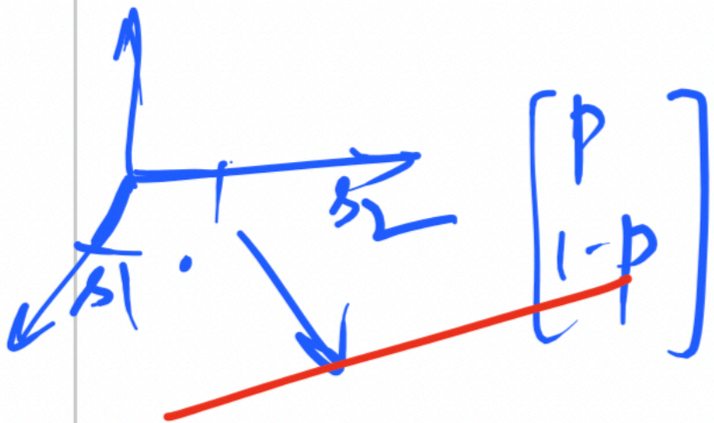
b α_p.

U(b) ← best plan.

Implication of a continuous belief state in POMDPs

- We can expect the agent to execute plan p in belief states that are very close to b.
- If the number of conditional plans being considered is finite, then the belief state space can be divided into regions such that each region has a particular plan that is optimal in that region.
- The utility function is the maximum of a collection of hyperplanes.





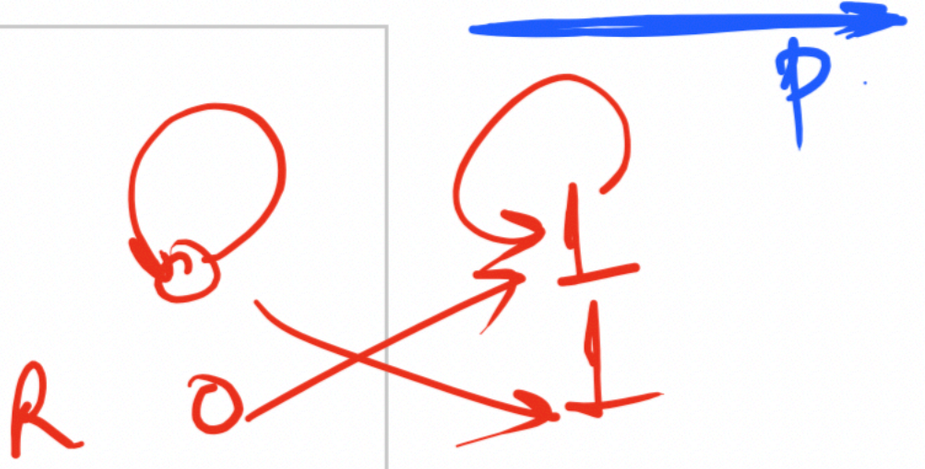
Example: A two-state world

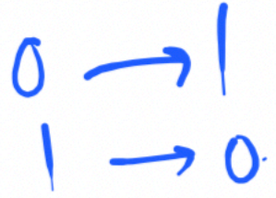
- State with label 0 carries reward $R(0) = 0$
- State with label 1 carries reward $R(1) = 1$
- Action 1: *Stay*

  Stay put with probability 0.9
- Action 2: *Go*

  Go switches to the other state with probability 0.9

- Sensor reports the correct state with probability 0.6

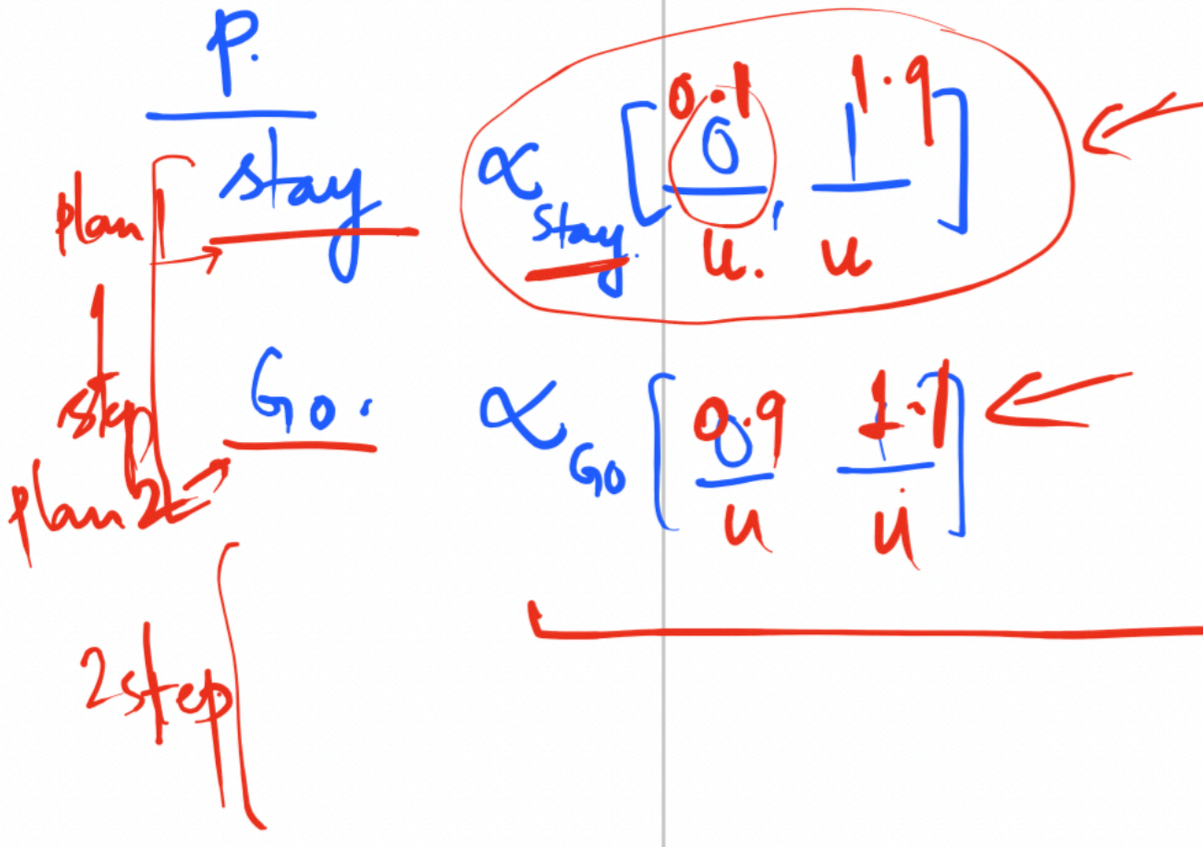




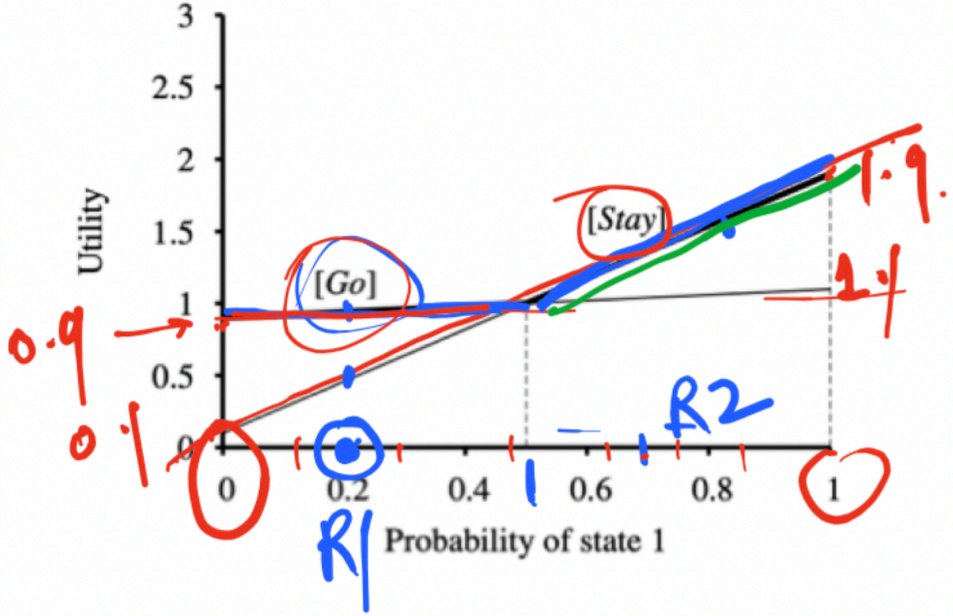
- Next we should assign expected utilities α_p to the states {0,1} based on 1-step plans

$$\alpha_{[Stay]}(0) = R(0) + \gamma(0.9R(0) + 0.1R(1)) = 0.1$$
$$\alpha_{[Stay]}(1) = R(1) + \gamma(0.9R(1) + 0.1R(0)) = 1.9$$
$$\alpha_{[Go]}(0) = R(0) + \gamma(0.9R(1) + 0.1R(0)) = 0.9$$
$$\alpha_{[Go]}(1) = R(1) + \gamma(0.9R(0) + 0.1R(1)) = 1.1$$

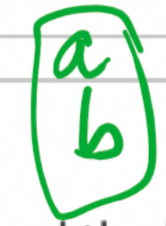
$R(s) + \gamma R(s')$

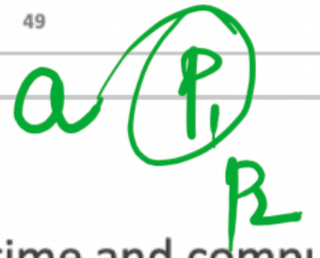


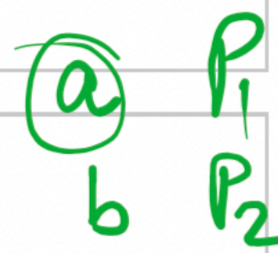
- The hyperplanes for $b \cdot \alpha_{[Stay]}$ and $b \cdot \alpha_{[Go]}$.

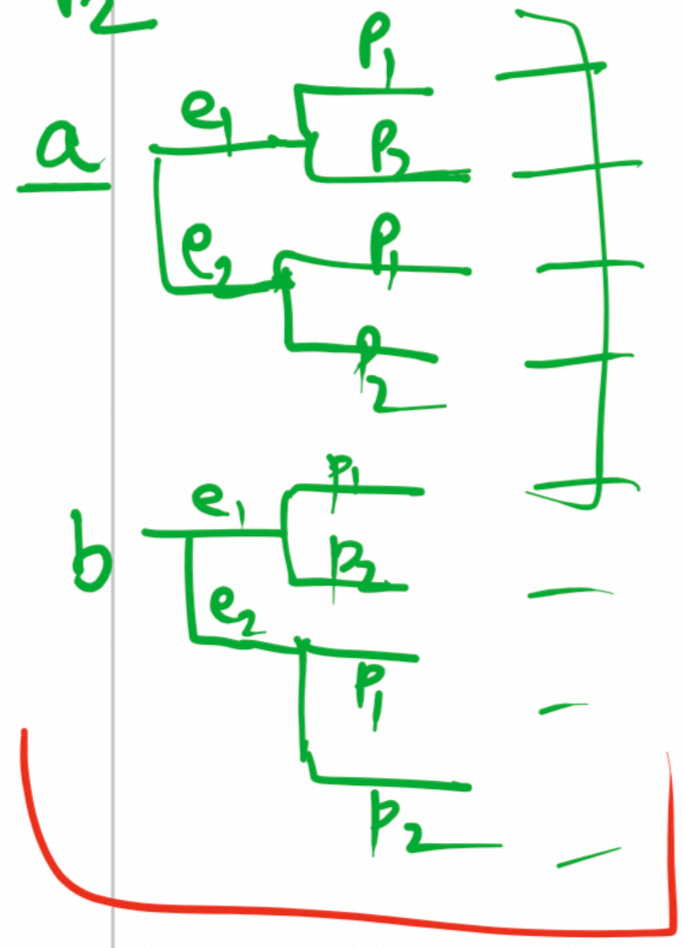


- The bold line represents the utility function for the finite horizon that allows only one action.





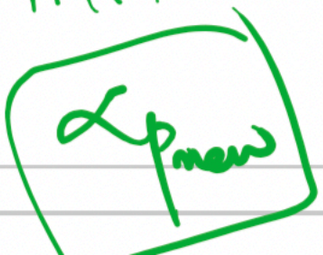




- We extend the horizon one step at a time and compute the utility for the physical state s
- That is, we now compute utilities for conditional plans of depth 2 by considering
  - each possible first action,
  - each possible subsequent percept, and then
  - each way of choosing a depth-1 plan to execute for each percept

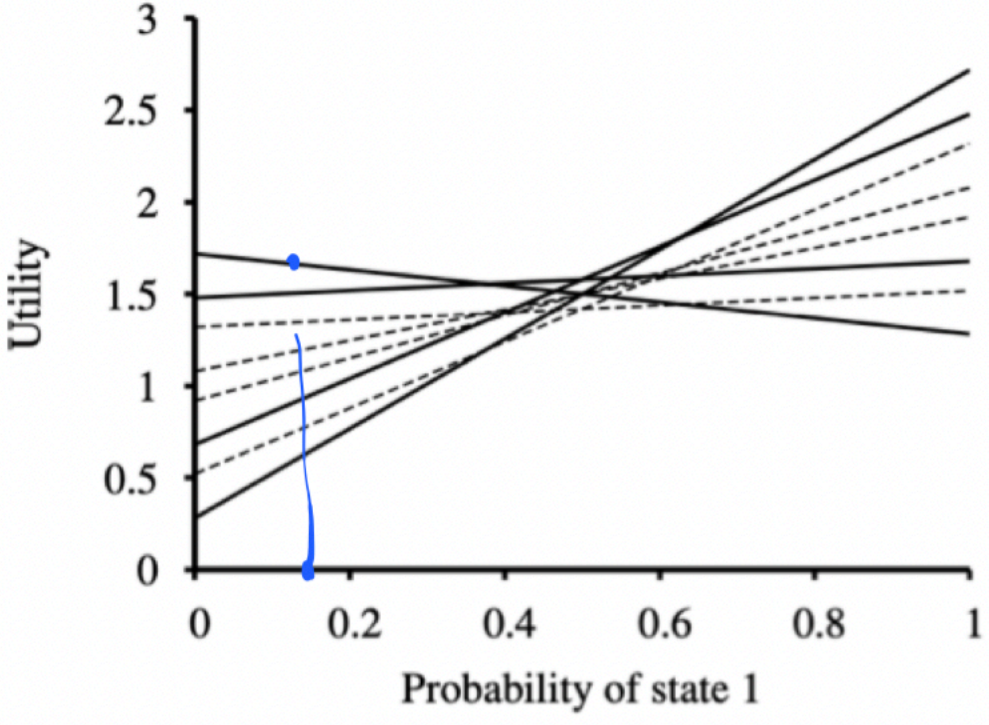
Utilities?

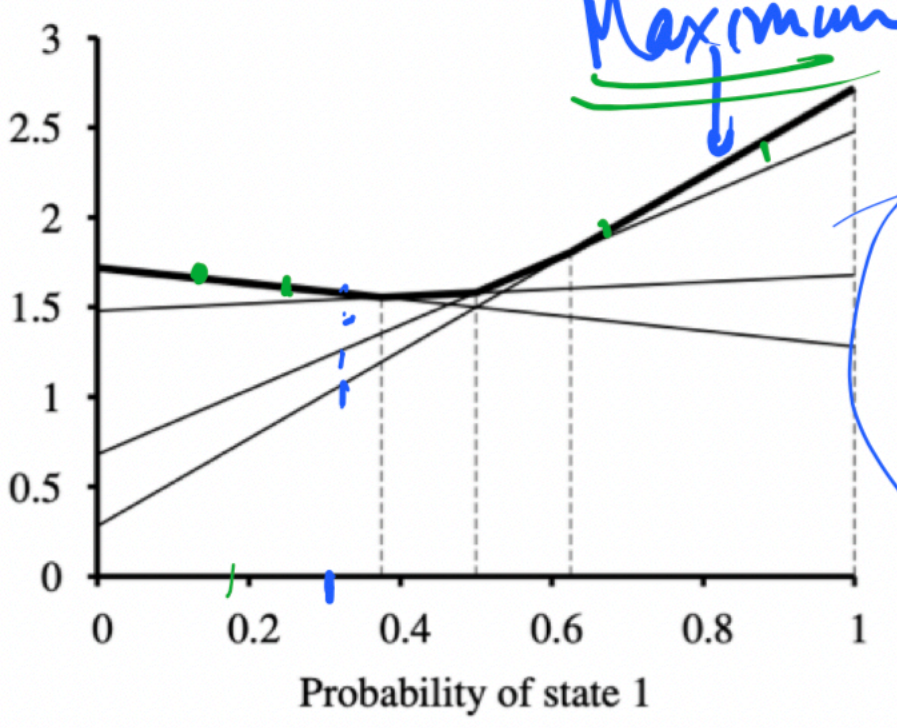
$$[Stay;\ \textbf{if}\ Percept = 0\ \textbf{then}\ Stay\ \textbf{else}\ Stay]$$
$$[Stay;\ \textbf{if}\ Percept = 0\ \textbf{then}\ Stay\ \textbf{else}\ Go] \ldots$$



2 steps.

- There are 8 distinct depth 2 plans and their utility for physical state s



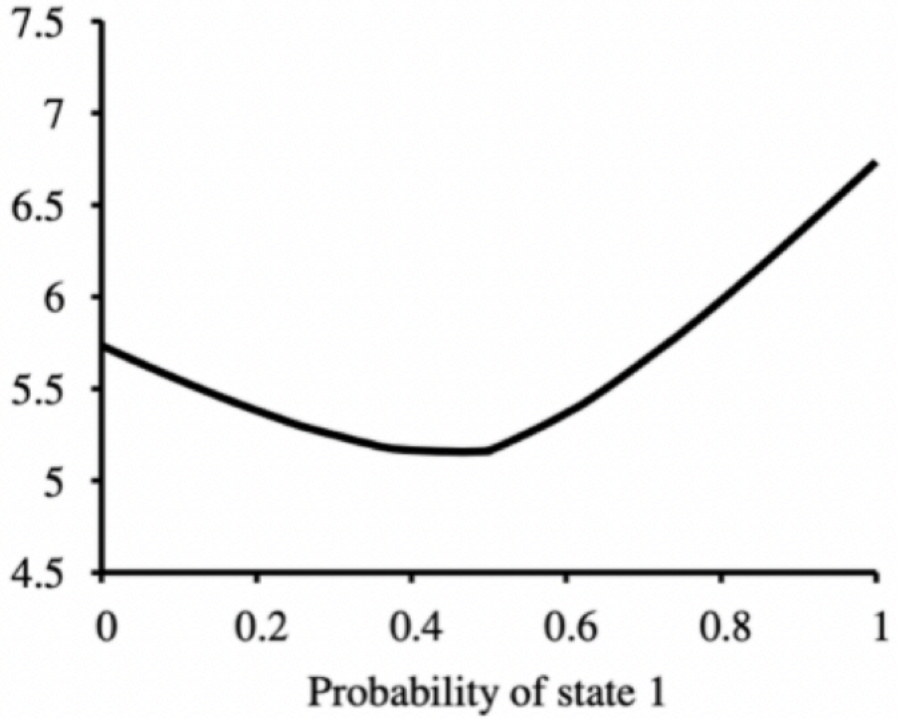


$\pi^*(s)$

$\pi^*(b)$

- The dashed plans are suboptimal across the entire belief space

- If we extend the horizon to to 8-step plan the utility function is



Piece-wise constant,
Convex

Backing up the utility values for a depth d plan

- Let p be a depth d conditional plan whose initial action is a and whose depth $d-1$ sub plan for percept e is $p.e$

$$\alpha_p(s) = R(s) + \gamma \left(\sum_{s'} P(s'|s,a) \sum_{e} P(e|s') \alpha_{p.e}(s') \right)$$

$$\alpha_p(s) = R(s) + \gamma \left(\sum_{s'} P(s'|s,a) U_{p.e}(s') \right)$$

Value Iteration Algorithm

function POMDP-VALUE-ITERATION($pomdp, \epsilon$) **returns** a utility function
  inputs: $pomdp$, a POMDP with states S, actions $A(s)$, transition model $P(s' \mid s, a)$,
    sensor model $P(e \mid s)$, rewards $R(s)$, discount γ
    ϵ, the maximum error allowed in the utility of any state
  local variables: U, U', sets of plans p with associated utility vectors α_p

  $U' \leftarrow$ a set containing just the empty plan $[\,]$, with $\alpha_{[]}(s) = R(s)$
  repeat
    $U \leftarrow U'$
    $U' \leftarrow$ the set of all plans consisting of an action and, for each possible next percept,
      a plan in U with utility vectors computed according to $\alpha_p(s) = R(s) + \gamma \left(\sum_{s'} P(s' \mid s, a) \sum_{e} P(e \mid s') \alpha_{p.e}(s') \right)$
    $U' \leftarrow$ REMOVE-DOMINATED-PLANS(U')
  until MAX-DIFFERENCE(U, U') $< \epsilon(1-\gamma)/\gamma$
  return U

1-step augmentation