CSEP 573

Online Algorithms for Partially Observable Markov Decision Processes

Online Planning

Input of Pomdp Policy: Some belief

- Do we need to know the best action for ***all*** belief states?
- We will reach a very small subset of belief states depending on
  - The initial belief
  - Our actions
  - Observations come from the environment
- Idea: Let's focus only on the belief states that we reach!

Belief Update

Pomdp model: <S, A, T, Z, O, R>

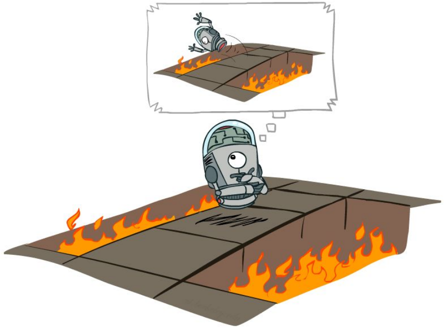
- T(s, a, s') = P(s' | a, s), O(s', a, z) = P(z | a, s')
- At time step *t*, the agent's belief is $b_t(s)$. After action a_{t+1}:

$$b'_t(s) = \sum_{s'} T(s', a_t, s) b_t(s')$$

- After Observation z_{t+1}:

$$b_{t+1}(s) \propto b'_t(s) O(s, a_t, z_{t+1})$$

Offline *vs.* Online (for POMDPs)



Offline Solution

Think hard; compute policy; then act

Online Learning

Compute policy for the current belief; act;

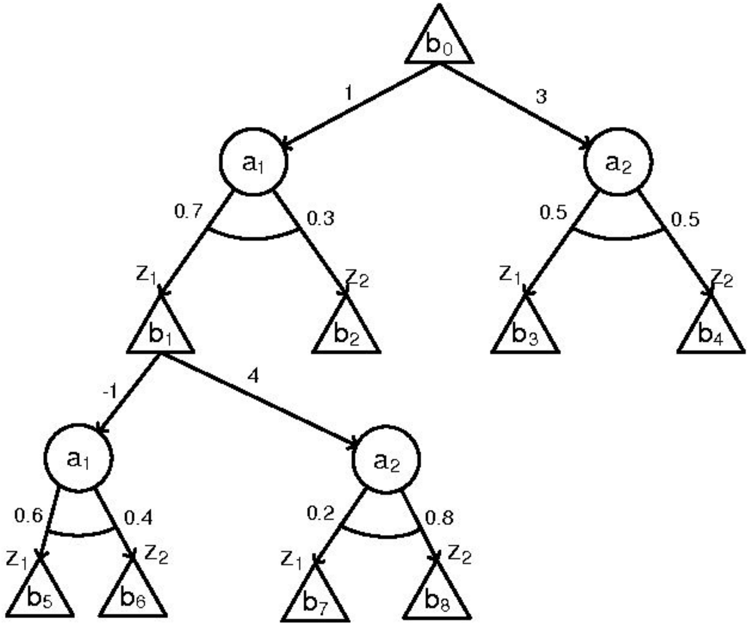
(Figures from Berkeley AI course slides)

Online Planning for POMDP as a search

- Similar to RTDP
  - Root: Current belief state
  - Greedy policy
- Challenges:
  - Different structure: Now we also have observations between states
  - Heuristic: We need a generalizable admissible heuristic

AND-OR Tree

- Actions and observation change the belief
  - Action: we pick!
    - Only the best one (Or)
  - Observation: Not in our hands!
    - Should consider all (AND)



(Ross et al., JAIR 2008)

Admissible Heuristics

- We want to maximize the reward
  - The heuristic should be an upper bound of value function
- Really hard to find one!
- How can we have an admissible heuristic for a POMDP?

  - Based on MDP model of the environment
  - Offline solver can be used iff
    - It is admissible!

QMDP

- Relax "partial" observability
  - The state of environment is fully observable after one action
    - After one action we solve MDP, i.e use Q-value of MDP
    - We are currently uncertain
      - Use expected Q-value

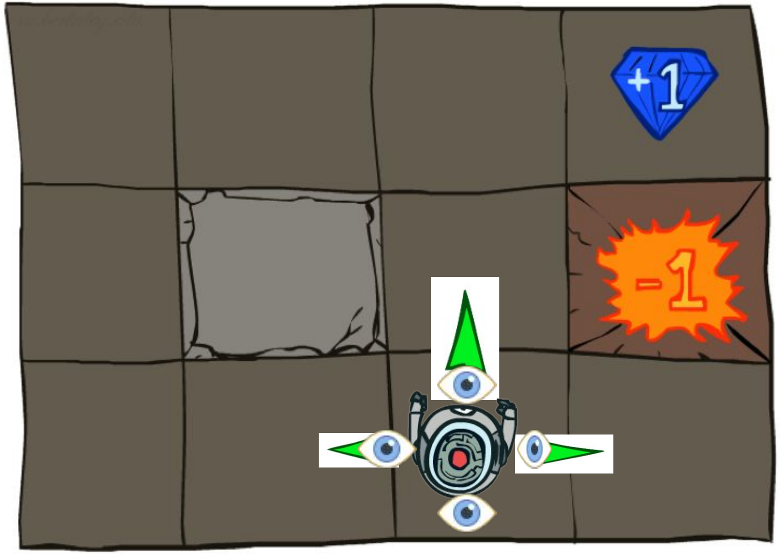
$$Q^{MDP}(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') V^{MDP}(s')$$

$$Q^{MDP}(b_t, a) = \sum_{s} b_t(s) Q^{MDP}(s, a)$$

$$V^{QMDP}(b_t) = argmax_a Q^{MDP}(b_t, a)$$

QMDP Example: Back to Grid World

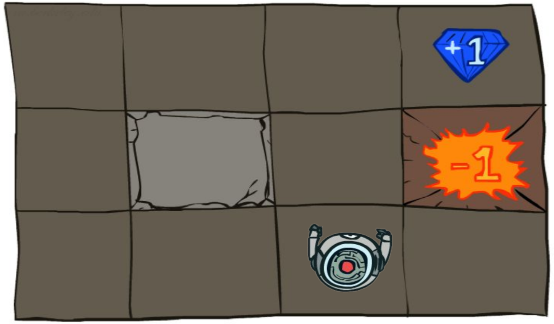
- Our robot does not know its current state
- It has 4 perfect eyes to see the walls of the current state
  – N, E, W: no wall, S: wall
- Actions: North, east, west, south
- Reward of breathing: -0.01 (cost!)
- Observation: only after an action
  – O(s', a, z)



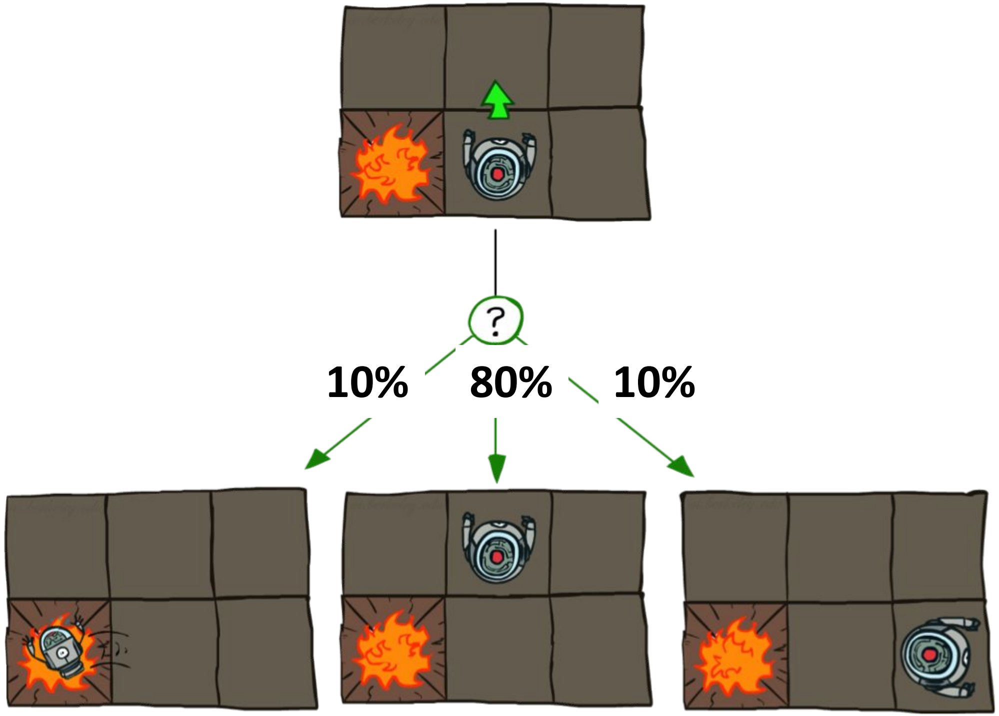
(Figures adapted from Berkeley AI course slides)

QMDP Example: Back to Grid World

- Its initial state:

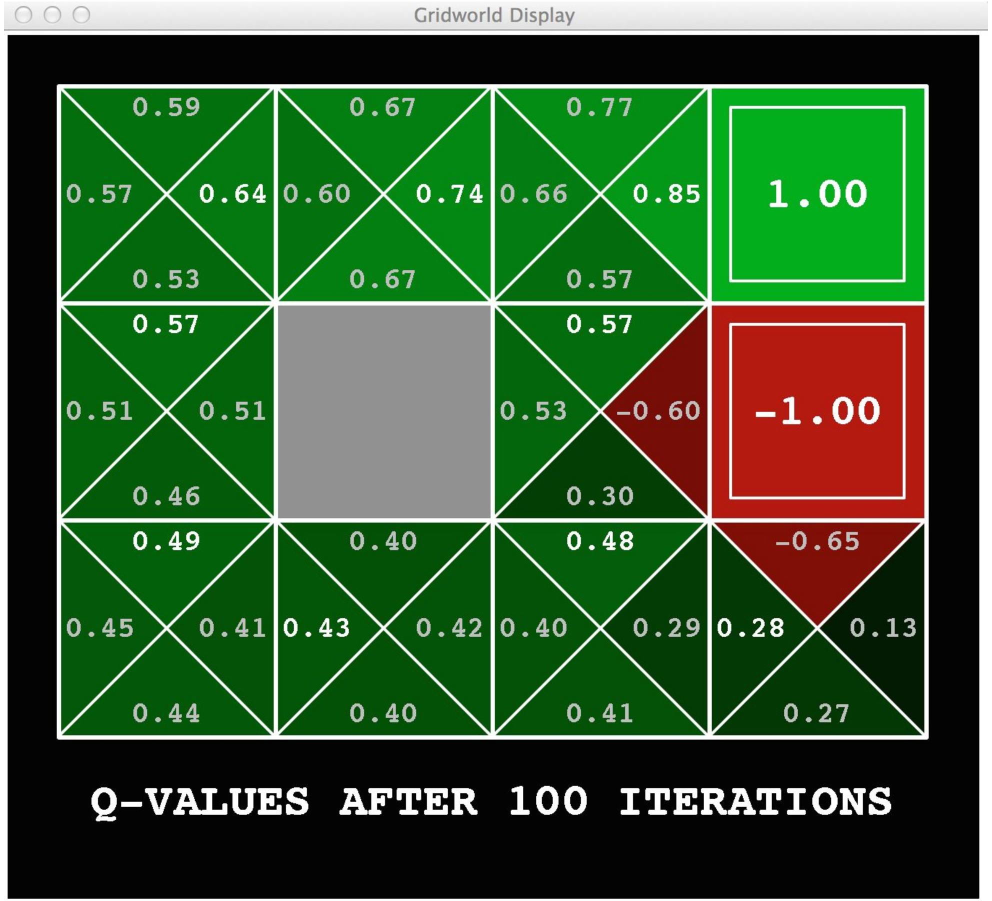


- Its belief state:
  - No clue! (uniform)
  - Each state:1/9



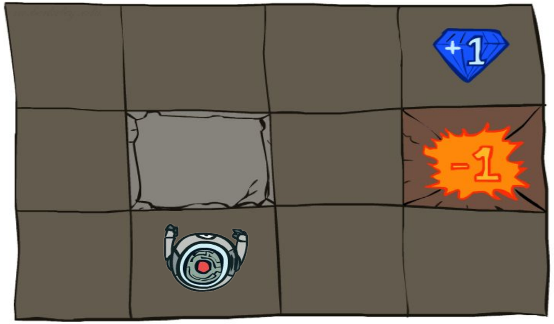
QMDP Example: Back to Grid World

- What is the best action (a_0)?
- Solve the MDP model first!
- $Q(b_0$, east) = .38
- $Q(b_0$, north) = .43
- $Q(b_0$, west) = .49
- $Q(b_0$, south) = .44



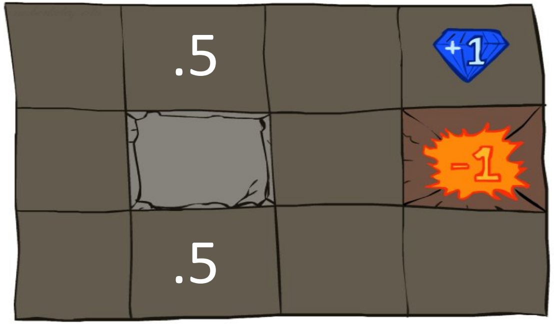
QMDP Example: Back to Grid World

- Its next state:



- Next observation (z_1)?
- Next belief state (b_1)?

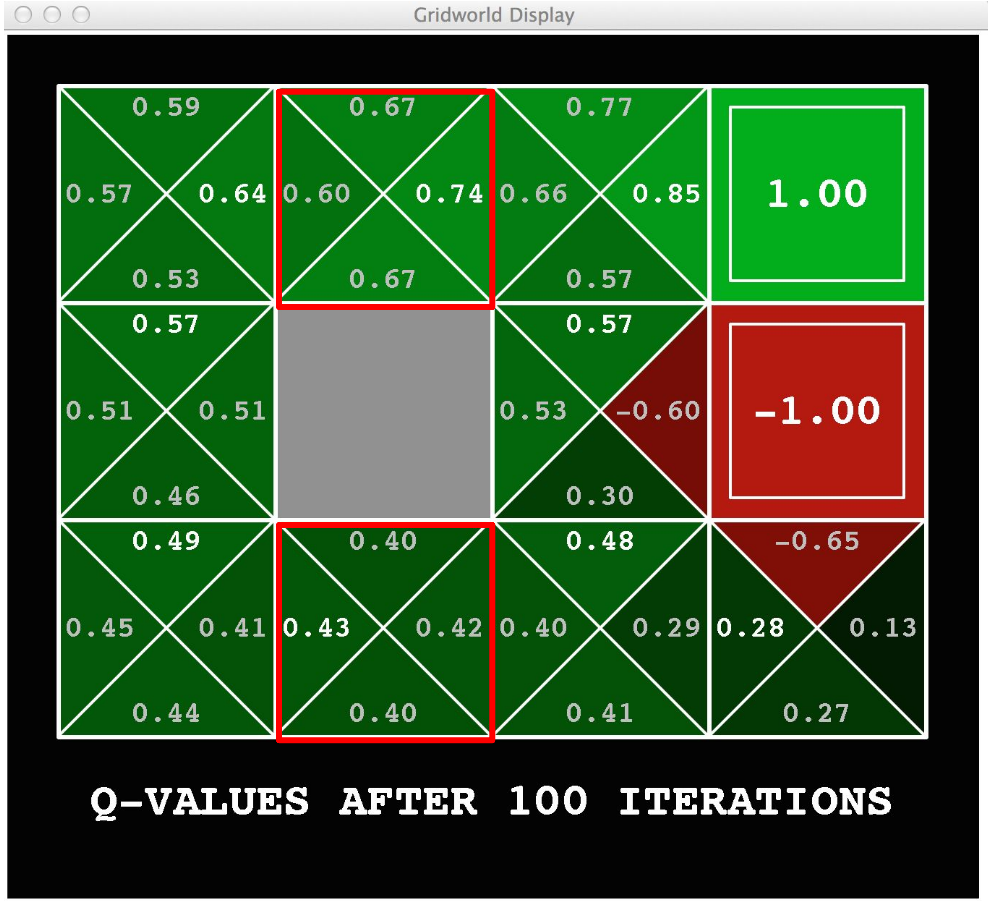
- z_1=N & S: wall, E & W: No wall
- b_1:



QMDP Example: Back to Grid World

What is the best action (a_1)?

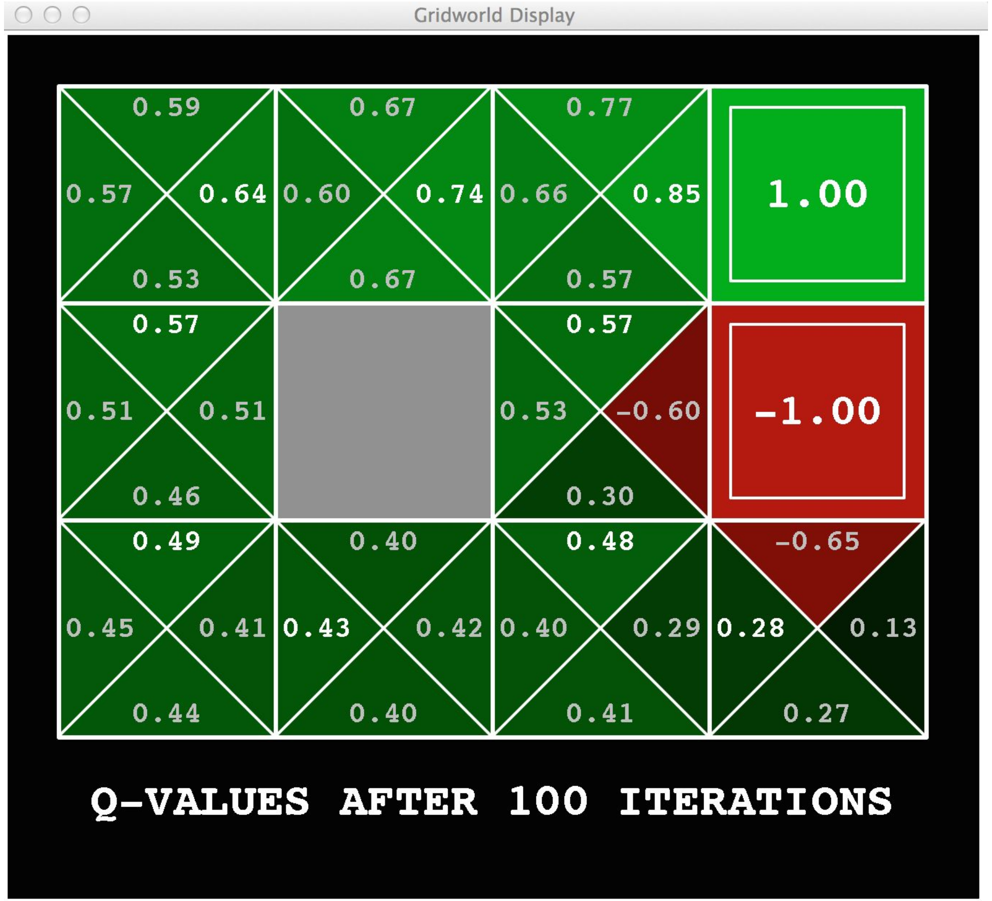
- $Q(b_1, \text{east}) = ½ (0.42) + ½ (0.74)$
  $= .58$
- $Q(b_1, \text{north}) = .54$
- $Q(b_1, \text{west}) = .52$
- $Q(b_1, \text{south}) = .54$



QMDP Example: Back to Grid World

- It is possible to get reward in 5 moves: .95 > .49
- How is it an upper bound?

- We were lucky!
- It is not going to happen in average
- The agent's prior should align with reality !



Solving POMDP with one heuristic

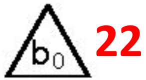
- Use an upper bound such as QMDP as a heuristic
- Use expected reward as the reward from root
- Expand one of the leaves
- Update ancestors
- There is no “goal” state, when to stop?

Solving POMDP with an admissible heuristic

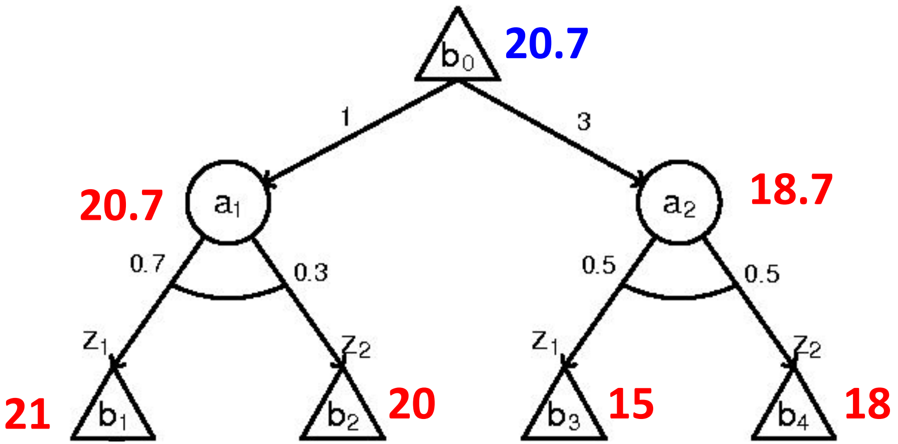
- Expand as long as you can (e.g you are given 1 sec)
  - Your algorithm should handle forced termination!
- You only need to perform one action
- Choose the action with maximum expected value (similar to RTDP)
- Update the new root (next belief state) by the chosen action and given observation (after the action)
- Don’t throw away the tree! Reuse the subtree with the new root!

Example

- Two actions
- Two observations
- Discount factor = .95
- Initial belief state (prior probability of states) = b_0
- $h(b) = V^{QMDP}(b)$
- Let's assume $h(b_0) = 22$
- Red: heuristic as the value function



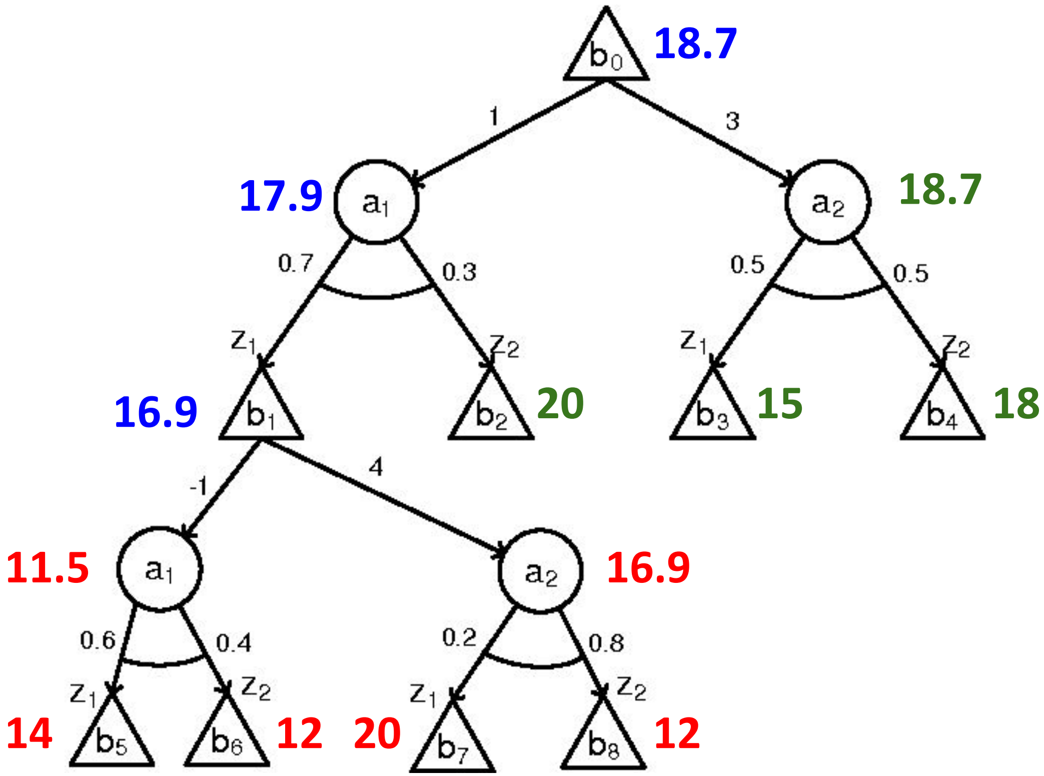
Example



- $h(a_1) = R(b_0, a_1) + .95[P(z_1|b_0, a_1)h(b_1) + P(z_2|b_0, a_1)h(b_2)]$
- $h(b_0) = \max(h(a_1), h(a_2))$
- Blue: Update of value because of children

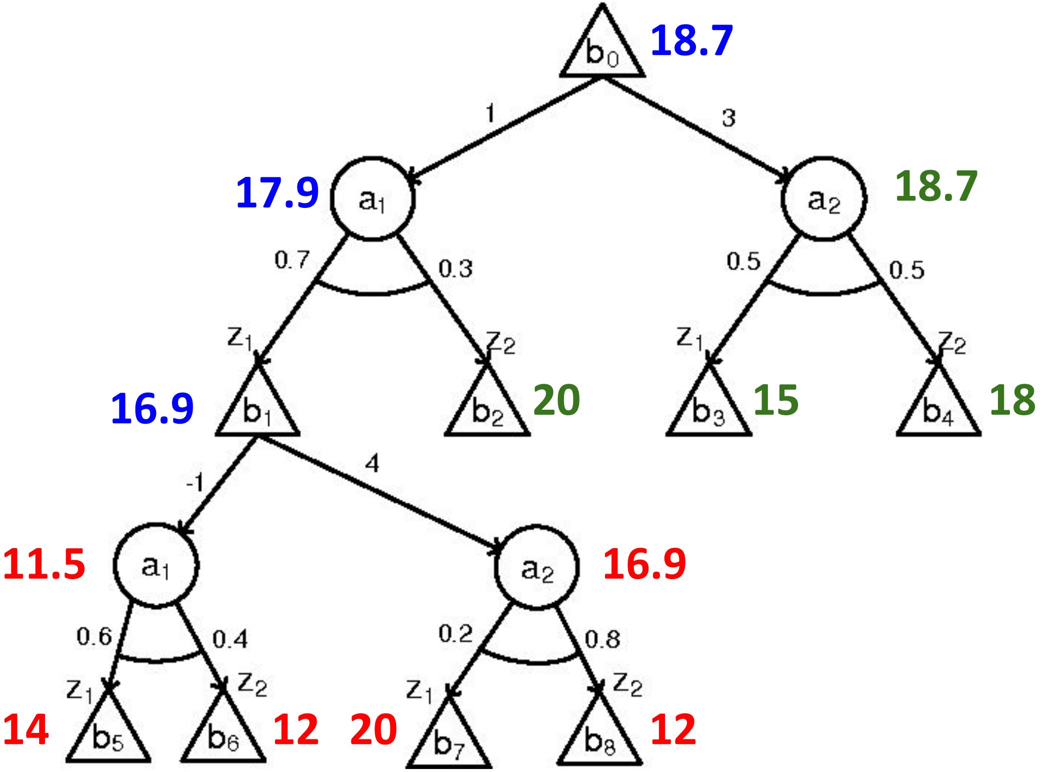
Example

- Green: No change



Example

- Best action?
- Next root?



Expand

- Each update has one action and one observation
  - Only “Or” nodes represent belief
  - “And” nodes are intermediate nodes, like b’
  - Expand One “Or” node at a time
    - Expand all of its children (“And” nodes)
- Expands smarter not longer!
  - Ideas?

Expand smarter!

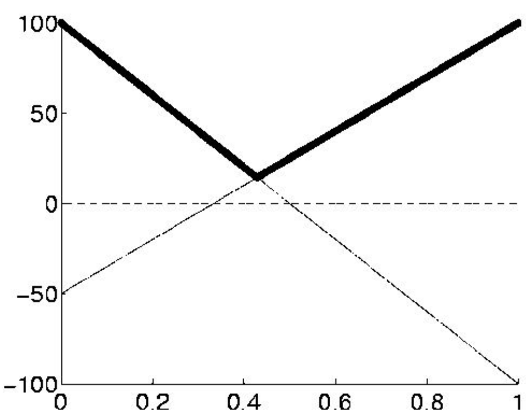
- Breadth first search
  - When is it effective?
- Expand a node with maximum h(b)
- Using two heuristics
  - One upper bound, one lower bound
  - Prune!

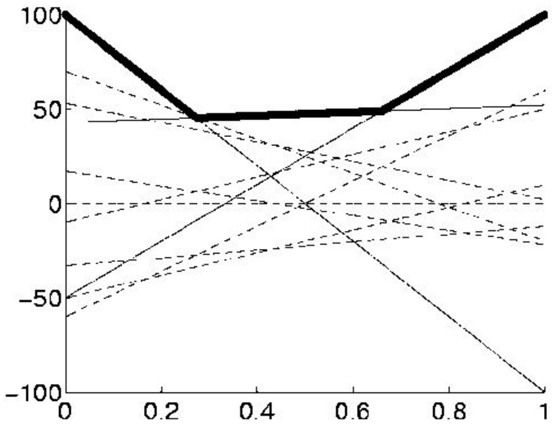
Pruning

- Do not expand a node iff:
  - There exists another node with lower bound greater than upper bound of this node
- Update the lower bound as well
  - Based on lower bounds of children
- One way to find bugs in your algorithm:
  - Upper bound should not increase with update
  - Lower bound should not decrease with update

PBVI as a heuristic

- PBVI is still relatively computationally expensive
- We can run it for a very short time and use it as a heuristic
- PBVI is lower bound
  - Value function is piecewise linear and convex





Error Minimization Search

- Upper bound is larger than the true value U >= V*
- Lower bound is smaller than the true value L =< V*
  - L =< V* =< U
- What is U-L?
  - What does small U-L mean?

AEMS

- Expand a node based on these criteria:
  - Error (U-L)
  - Depth
  - Edges (actions and observations) in the path from root