Policies for POMDPs

Minqing Hu

Background on Solving POMDPs

- MDPs policy: to find a mapping from states to actions
- POMDPs policy: to find a mapping from probability distributions (over states) to actions.
  - belief state: a probability distribution over states
  - belief space: the entire probability space, infinite

Policies in MDP

- *k*-horizon Value function:

$$V_t^{\delta_t}(s_i) = q_i^{\delta_t(s_i)} + \beta \sum_j p_{ij}^{\delta_t(s_i)} V_{t-1}^{\delta_{t-1}}(s_j)$$

- Optimal policy δ^*, is the one where, for all states, s_i and all other policies,

$$V^{\delta^*}(s_i) \geq V^{\delta}(s_i)$$

Finite k-horizon POMDP

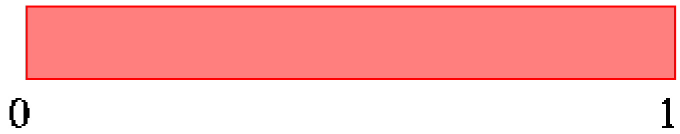
- POMDP: <S, A, P, Z, R, W>
- transition probability: p_{ij}^{a}
- probability of observing *z* after taking action *a* and ending in state s_j: r_{jz}^{a}
- immediate rewards: w_{ijz}^{a}
- Immediate reward of performing action *a* in state s_i:
$$q_i^a = \sum_{j,z} p_{ij}^{z} r_{jz}^{a} w_{ijz}^{a}$$
- Object: to find an optimal policy for finite k-horizon POMDP

$\delta^* = (\delta_1, \delta_2, \ldots, \delta_k)$

A two state POMDP

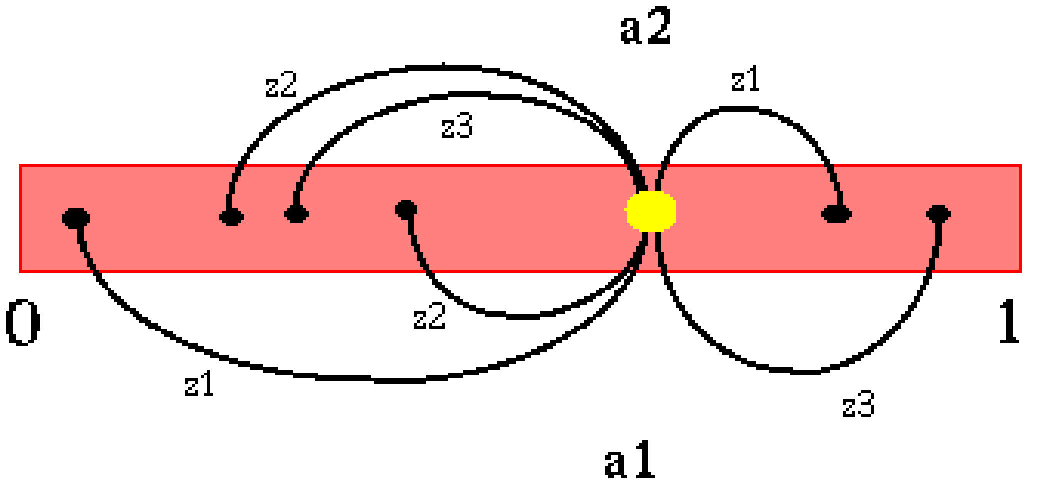
- represent the belief state with a single number *p*.
- the entire space of belief states can be represented as a line segment.

belief space for a 2 state POMDP



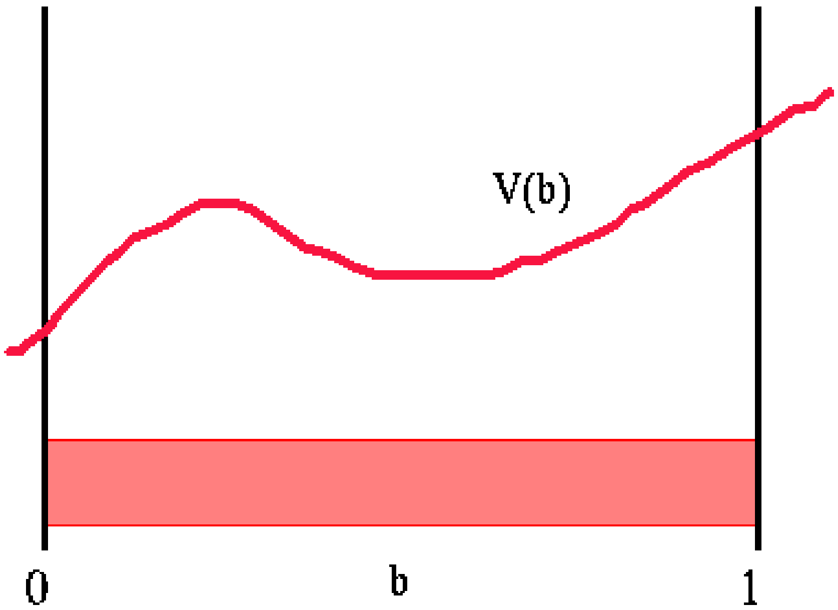
belief state updating

- finite number of possible next belief states, given a belief state
  - a finite number of actions
  - a finite number of observations
- b' = *T*(b| a, z). Given *a* and *z,* b' is fully determined.



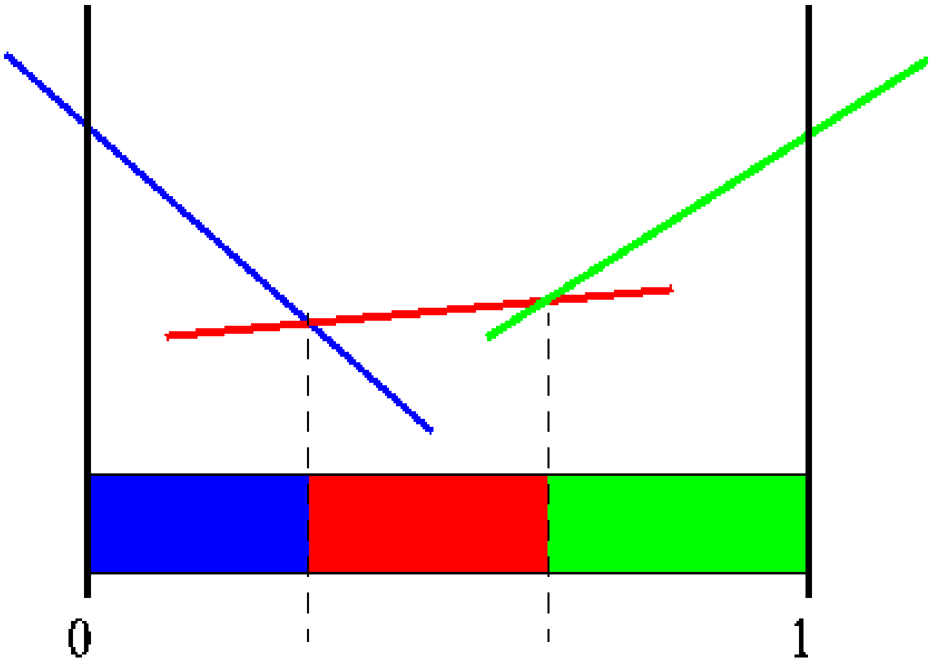
- the process of maintaining the belief state is Markovian: the next belief state depends only on the current belief state (and the current action and observation)
- we are now back to solving a MDP policy problem with some adaptations

- continuous space: value function is some arbitrary function
  - b: belief space
  - V(b): value function
- Problem: how we can easily represent this value function?



Value function over belief space

Fortunately, the finite horizon value function is piecewise linear and convex (PWLC) for every horizon length.



Sample PWLC function

- A Piecewise Linear function consists of linear, or hyper-plane segments
  - Linear function:

$$\sum_i \alpha_i x_i = \alpha_0 x_0 + \alpha_1 x_1 + \ldots + \alpha_N x_N$$

  - K^th linear segment: $\sum_{i=0}^{N} \alpha_i^k x_i$
  - the α-vector

$$\alpha^k = [\alpha_0^k, \alpha_1^k, \ldots, \alpha_N^k]$$

  - each liner or hyper-plane could be represented with $\alpha^k(t)$

- Value function:

$$V_t^*(b) = \max_k \sum_i b_i \alpha_i^k(t)$$

- a convex function

- 1-horizon POMDP problem
  - Single action *a* to execute
  - Starting out belief state *b*
  - Ending belief state *b'*
    - *b'* = *T*(*b* | *a*, *z*)
  - Immediate rewards q_i^a
  - Terminating rewards q_i^0 for state s_i

    Expected terminating reward in *b'*

$$V_0(b') = \sum_i b_i' q_i^0$$

- Value function of t =1

$$V_1^*(b) = \max_{a \in A}[\sum_i b_i q_i^a + \sum_{i,j,z} b_i p_{ij}^a r_{jz}^a q_i^0]$$

- The optimal policy for t =1

$$\delta_1^*(b) = \arg\max_{a \in A}[\sum_i b_i q_i^a + \sum_{i,j,z} b_i p_{ij}^a r_{jz}^a q_i^0]$$

General *k*-horizon value function

- Same strategy for 1-horizon case
- Assume that we have the optimal value function at t – 1, $V_{t-1}^{*}(.)$
- Value function has same basic form as MDP, but
  - Current belief state
  - Possible observations
  - Transformed belief state

- Value function

$$V_t^*(b) = \max_{a \in A}[\sum_i b_i q_i^a + \sum_{i,j,z} b_i p_{ij}^a r_{jz}^a V_{t-1}^*[T(b \mid a, z)]]$$

Piecewise linear and convex?

$$V_t^*(b) = \max_k \sum_i b_i \alpha_i^k(t)$$

Inductive Proof

- Base case: $$V_0\,(b) = \sum_i b_i\, q_i^0$$
- Inductive hypothesis:

$$V_{t-1}^*(b) = \max_k \sum_i b_i \alpha_i^k (t-1)$$

  - transformed to:

$$V_{t-1}^*(T(b \mid a, z)) = \max_k \sum_i b'_i\, \alpha_i^k (t-1)$$

  - Substitute the transformed belief state:

$$b_j^{'} = \frac{\sum_i b_i\, p_{ij}^a r_{jz}^a}{\sum_{i,j} b_i\, p_{ij}^a r_{jz}^a} \qquad V_{t-1}^*(T(b \mid a, z) = \max_k [\frac{\sum_{i,j} b_i\, p_{ij}^a r_{jz}^a \alpha_j^k (t-1)}{\sum_{i,j} b_i\, p_{ij}^a r_{jz}^a}]$$

Inductive Proof (contd)

$$\iota(b,a,z) = \arg\max_{k}[\sum_{i,j} b_i p_{ij}^a r_{jz}^a \alpha_j^k(t-1)]$$

- Value function at step t (using recursive definition)

$$V_t^*(b) = \max_{a \in A} \sum_i b_i [q_i^a + \sum_{j,z} p_{ij}^a r_{jz}^a \alpha_j^{\iota(b,a,z)}(t-1)]$$

- New α-vector at step t

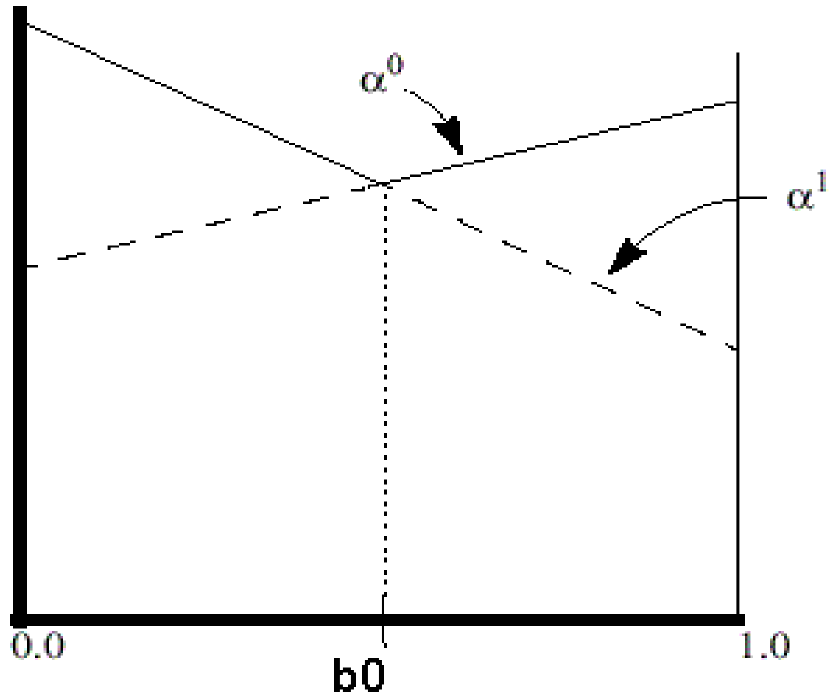
$$\alpha_t^*(b) = q_i^a + \sum_{j,z} p_{ij}^a r_{jz}^a \alpha_j^{\iota(b,a,z)}(t-1)$$

- Value function at step t (PWLC)

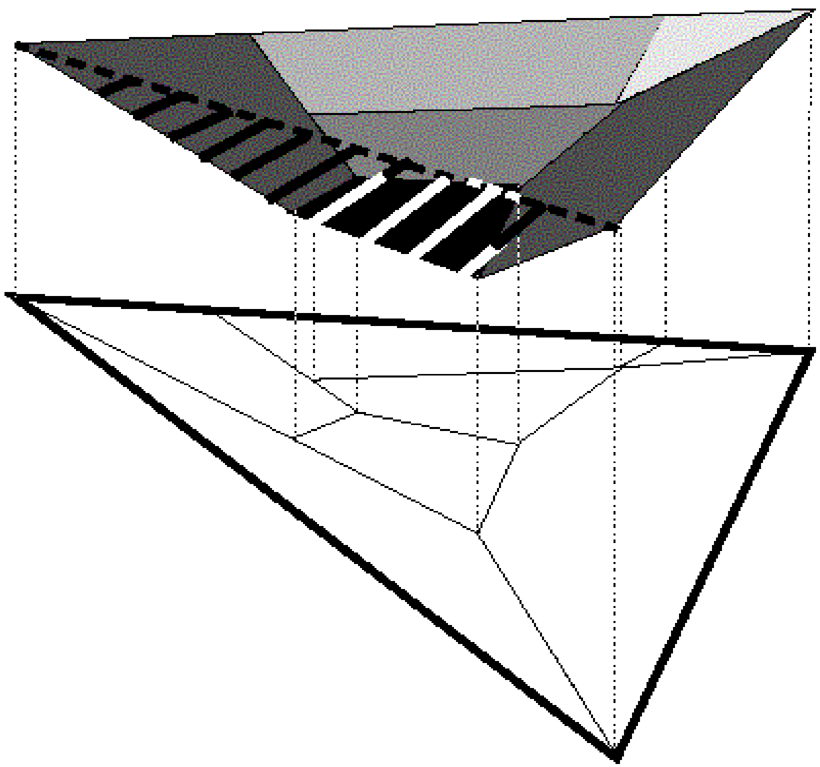
$$V_t^*(b) = \max_k \sum_i b_i \alpha_i^k(t)$$

Geometric interpretation of value function

- |*S*| = 2



Sample value function for |S| = 2



Sample value function for |S| = 3

- $|S| = 3$
- Hyper-planes
- Finite number of regions over the simplex

POMDP Value Iteration Example

- a 2-horizon problem
- assume the POMDP has
  - two states s1 and s2
  - two actions a1 and a2
  - three observations z1, z2 and z3

Horizon 1 value function

Given belief state b = [0.25, 0.75]

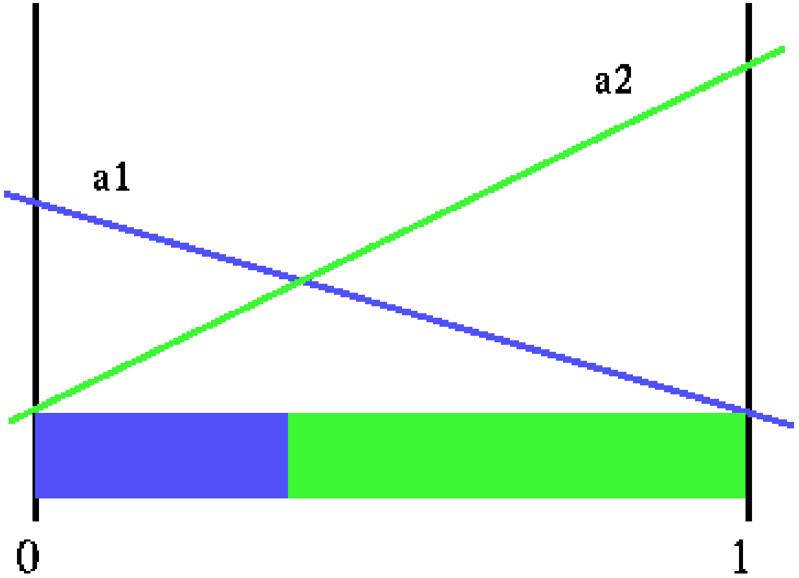
terminating reward = 0

$$q_{s1}^{a1} = 1, q_{s2}^{a1} = 0, q_{s1}^{a2} = 0, q_{s2}^{a2} = 1.5$$

$$V_1^{a1}(b) = 0.25 \text{x}\, 1 + 0.75 \text{ x}\, 0 = 0.25$$

$$V_1^{a2}(b) = 0.25 \text{x}\, 0 + 0.75 \text{ x}\, 1.5 = 1.125$$

- The blue region:
  - the best strategy is a1
- the green region:
  - a2 is the best strategy



Horizon 1 value function

2-Horizon value function

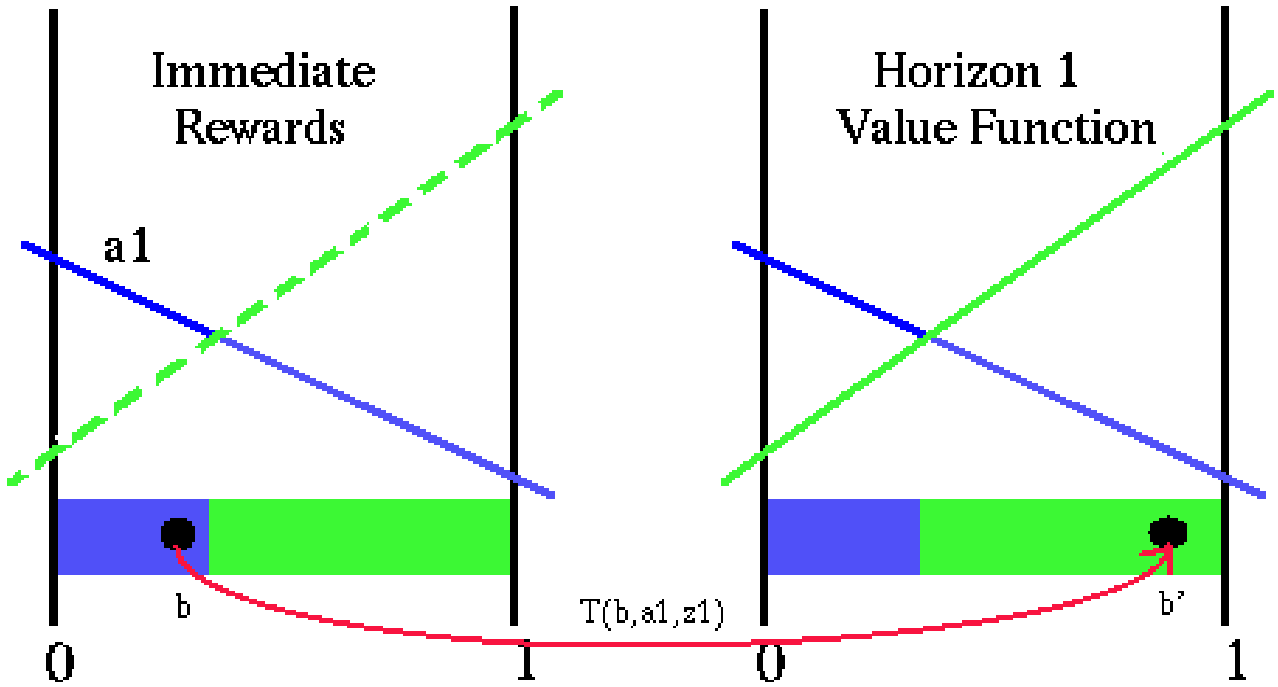
- construct the horizon 2 value function with the horizon 1 value function.
- three steps:
  - how to compute the value of a belief state for a given action and observation
  - how to compute the value of a belief state given only an action
  - how to compute the actual value for a belief state

Step1:

- A restrict problem: given a belief state b, what is the value of doing action a1 first, and receiving observation z1?

- The value of a belief state for horizon 2 is the value of the immediate action plus the value of the next action.

$$b' = T(b \mid a1, z1)$$

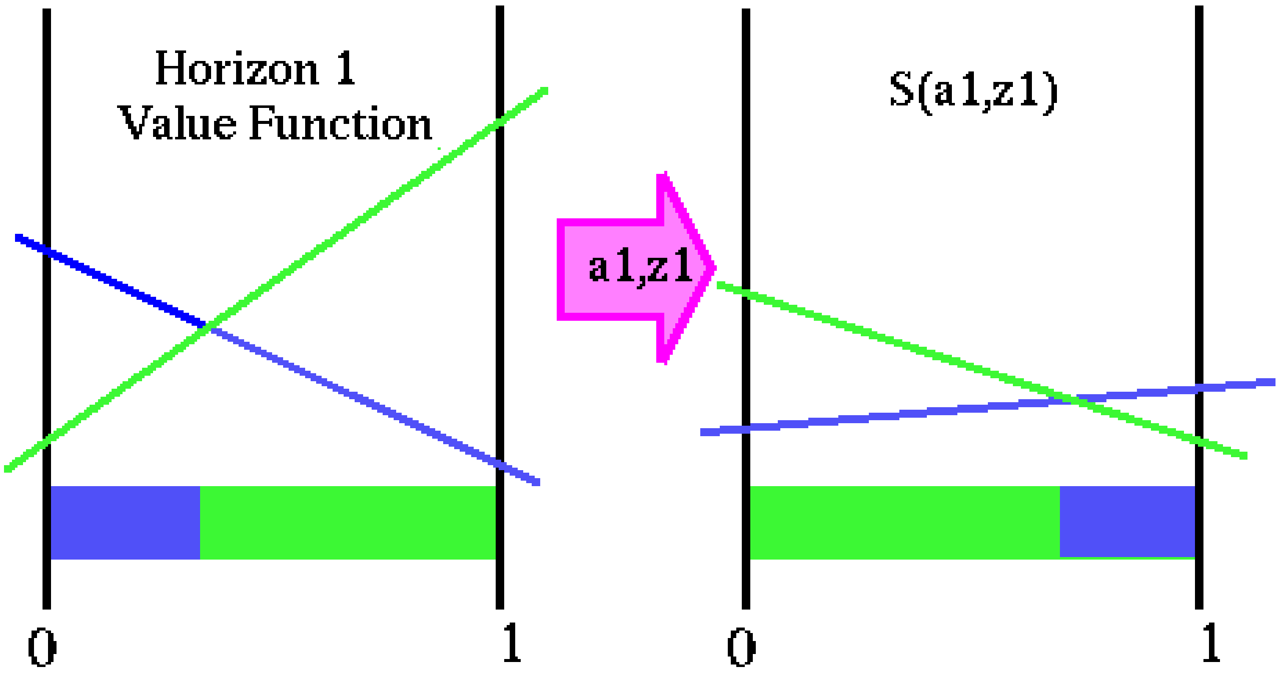


immediate reward function

horizon 1 value function

Value of a fixed action and observation

S(a, z), a function which directly gives the value of each belief state after the action a1 is taken and observation z1 is seen



- Value function of horizon 2 :

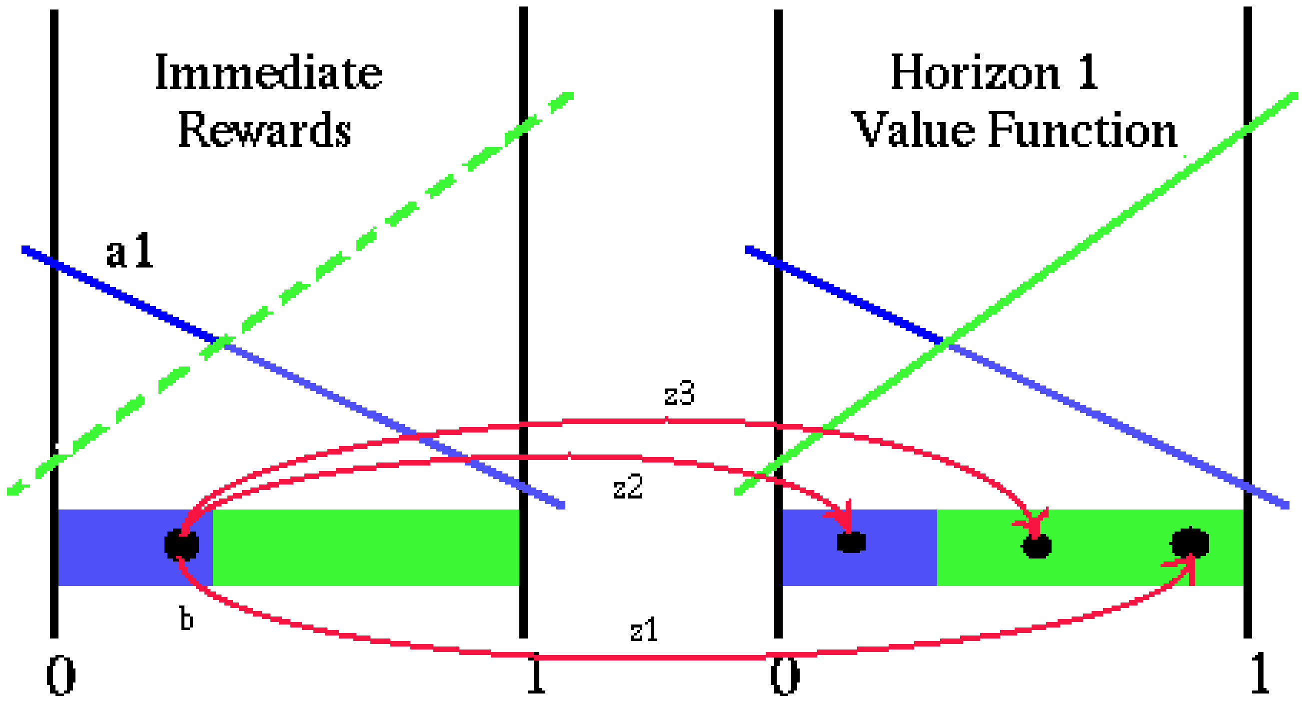
$$V_2^*(b) = \max_{a \in A}[\sum_i b_i q_i^a + \sum_{i,j,z} b_i p_{ij}^a r_{jz}^a V_1^*[T(b \mid a, z)]]$$

Immediate rewards S(a, z)

- Step 1: done

Step 2:

how to compute the value of a belief state given only the action



Transformed value function

So what is the horizon 2 value of a belief state, given a particular action a1?

depends on:

- the value of doing action a1
- what action we do next
  - depend on observation after action a1

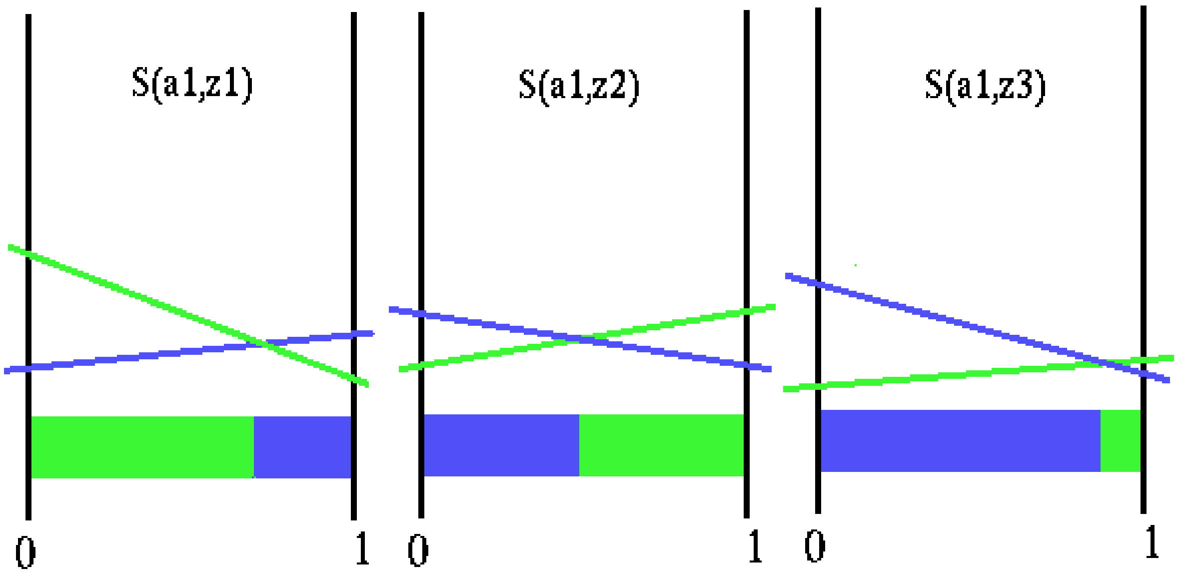
$$z1: V_1^{a1}(b) = 0.8, r_{z1}^{a1} = 0.6$$

$$z2: V_1^{a1}(b) = 0.7, r_{z2}^{a1} = 0.25$$

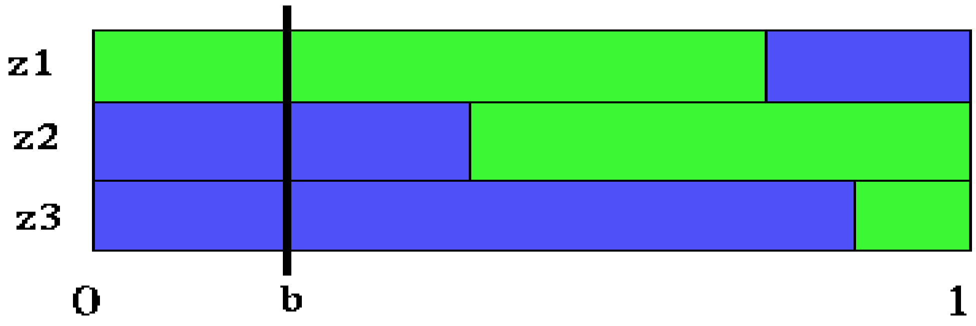
$$z3: V_1^{a1}(b) = 1.2, r_{z3}^{a1} = 0.15$$

$$V_2(b) = (0.6\text{x}0.8 + 0.25\text{x}0.7 + 0.15\text{x}1.2) = 0.835$$

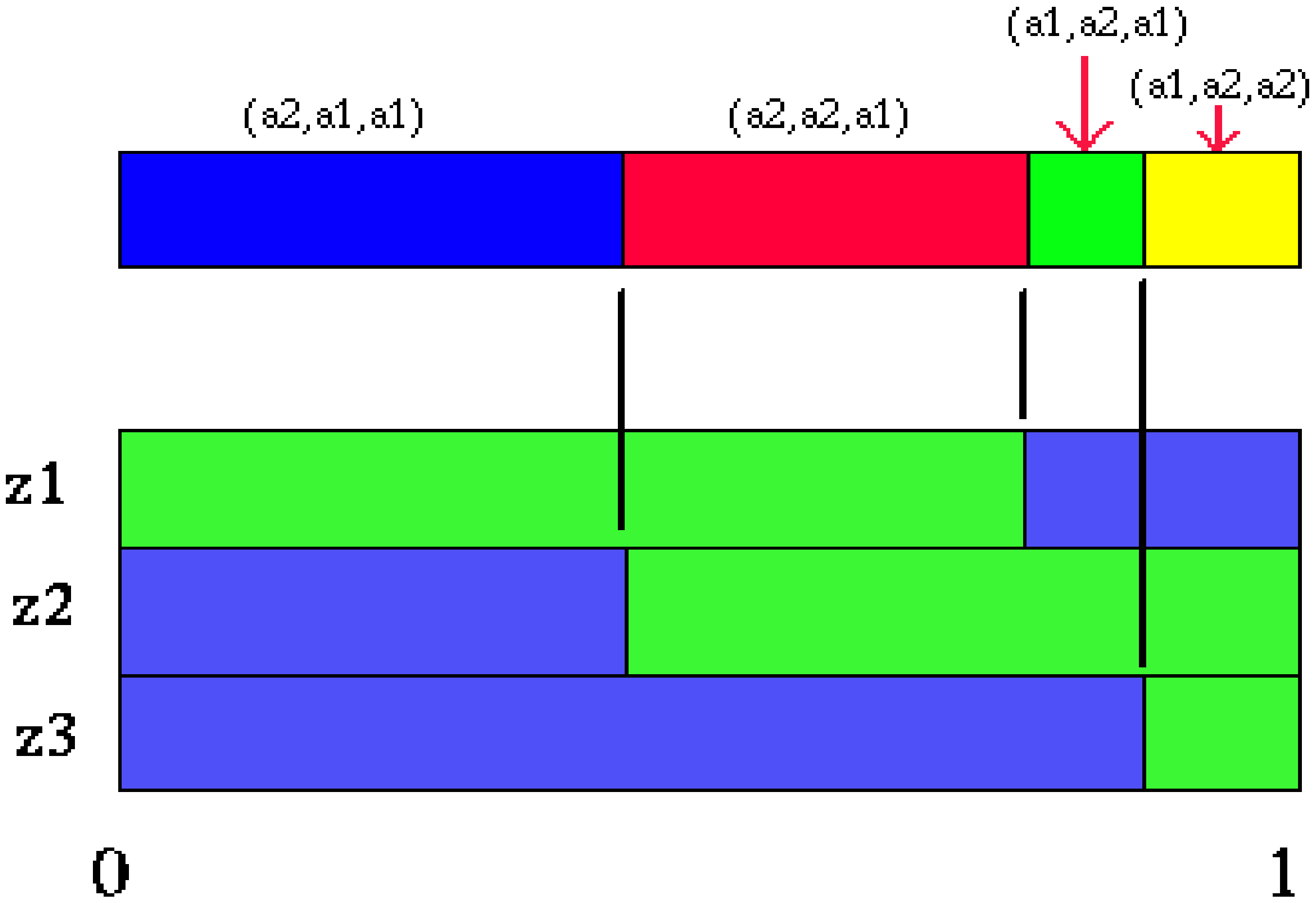
plus the immediate reward of doing action a1 in b



Transformed value function for all observations



Belief point in transformed value function partitions



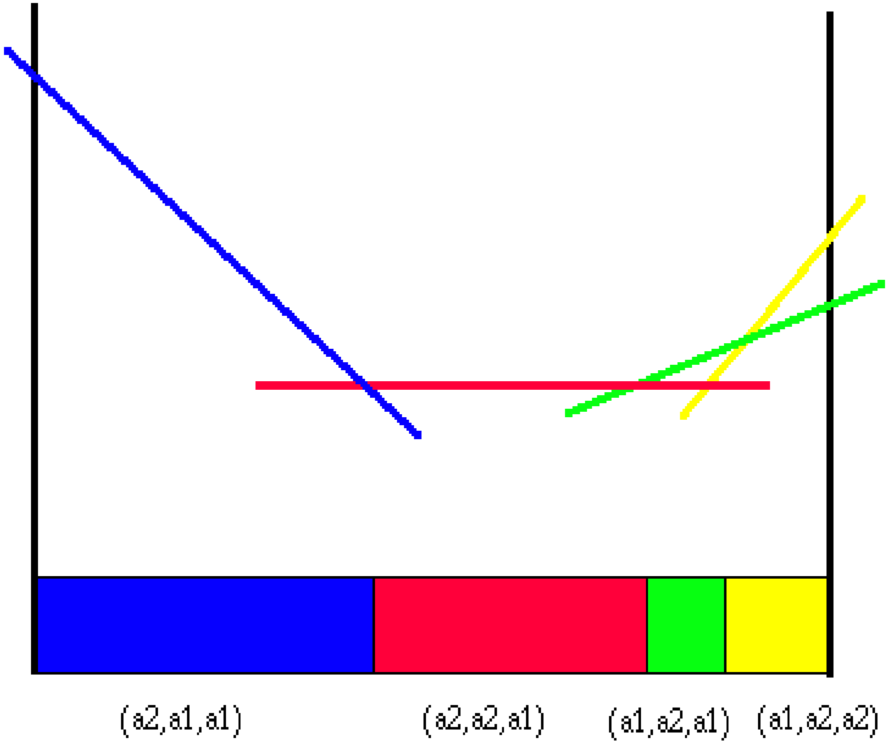
Partition for action a1

- The figure allows us to easily see what the best strategies are after doing action a1.

- the value of the belief point b at horizon 2

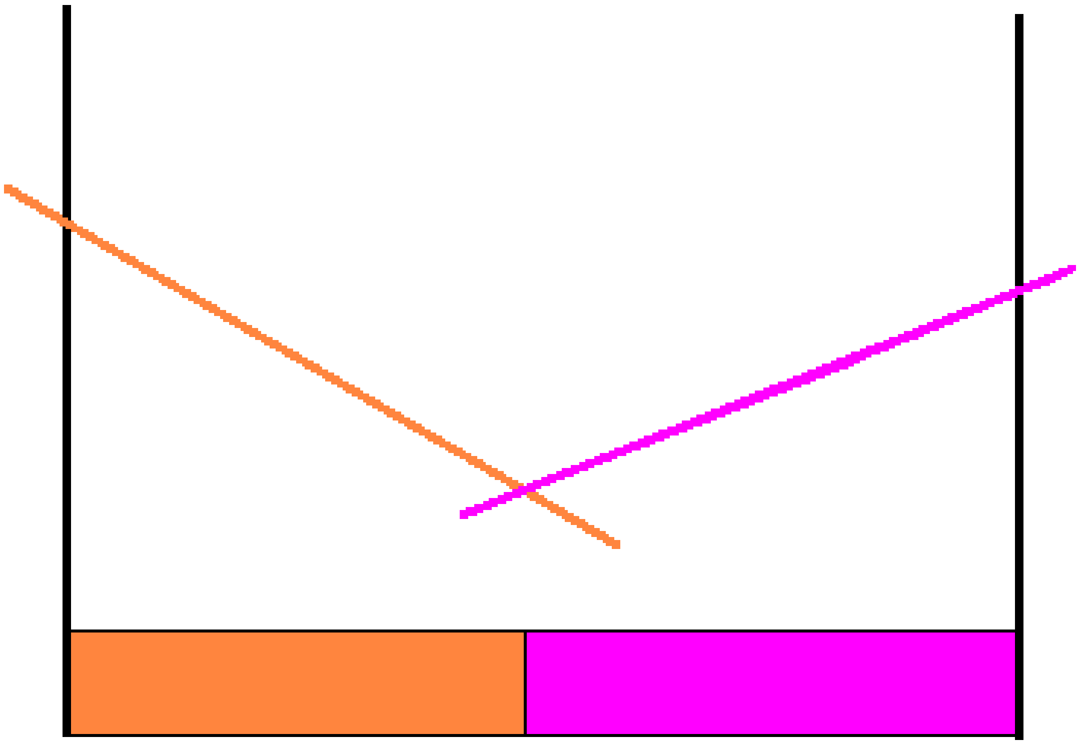
  = the immediate reward from doing action a1 + the value of the functions S(a1,z1), S(a1,z3), S(a1,z3) at belief point b.

- Each line segment is constructed by adding the immediate reward line segment to the line segments for each future strategy.



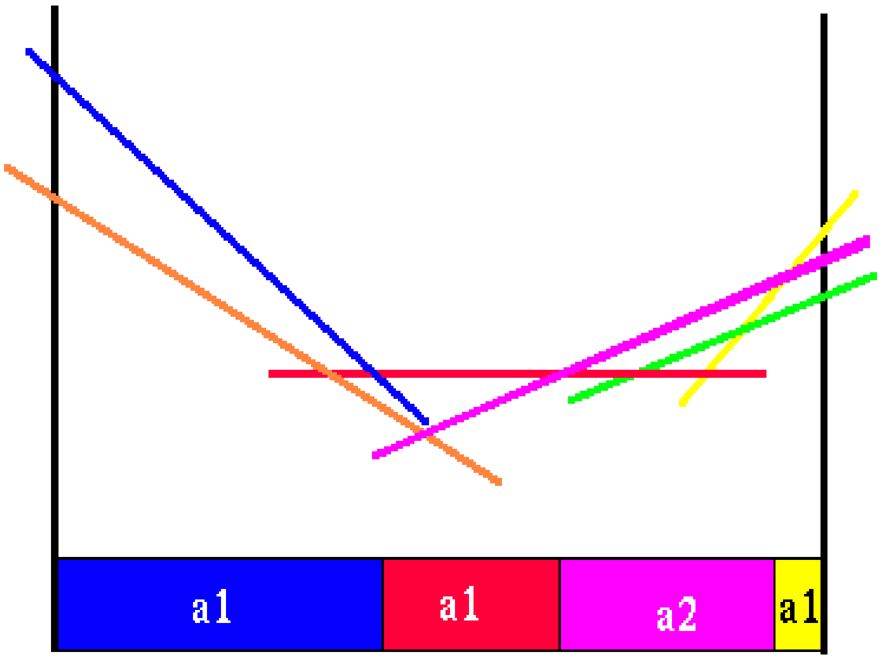
Horizon 2 value function and partition for action a1

Repeat the process for action a2

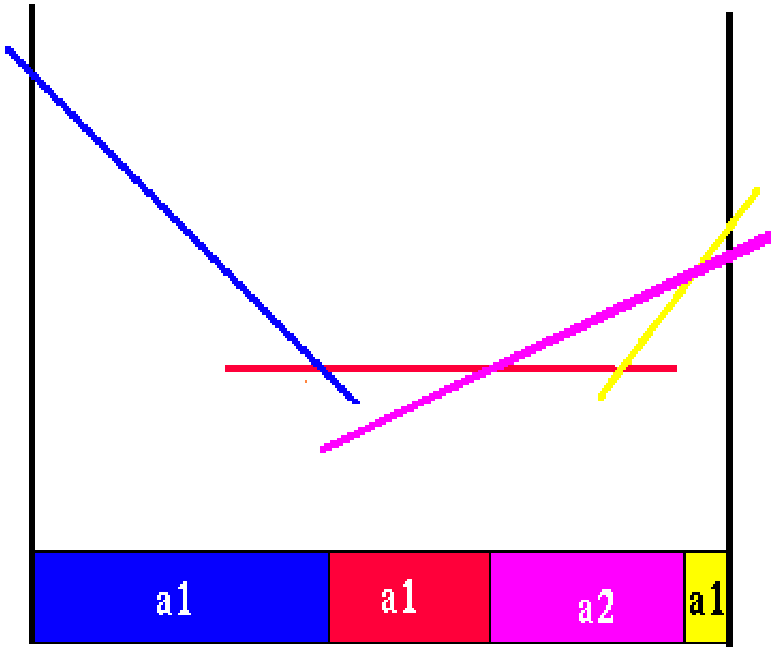


Value function and partition for action a2

Step 3: best horizon 2 policy



Combined a1 and a2 value functions



Value function for horizon 2

- Repeat the process for value functions of 3-horizon,…, and k-horizon POMDP

$$V_t^*(b) = \max_{a \in A}[\sum_i b_i q_i^a + \sum_{i,j,z} b_i p_{ij}^a r_{jz}^a V_{t-1}^*[T(b \mid a, z)]]$$

Alternate Value function interpretation

- A decision tree
  - Nodes represent an action decision
  - Branches represent observation made
- Too many trees to be generated!