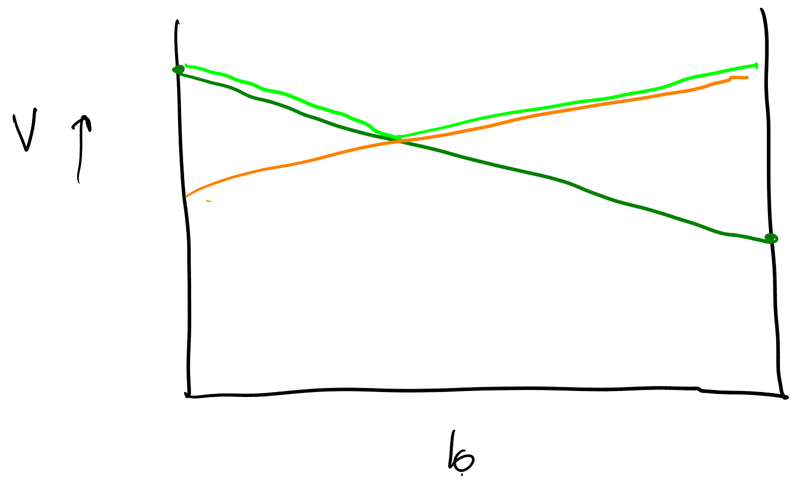
Offline POMDP Algorithms



Last time: POMDP Value Iteration (horizon d)

$\Gamma^0 \leftarrow \emptyset$

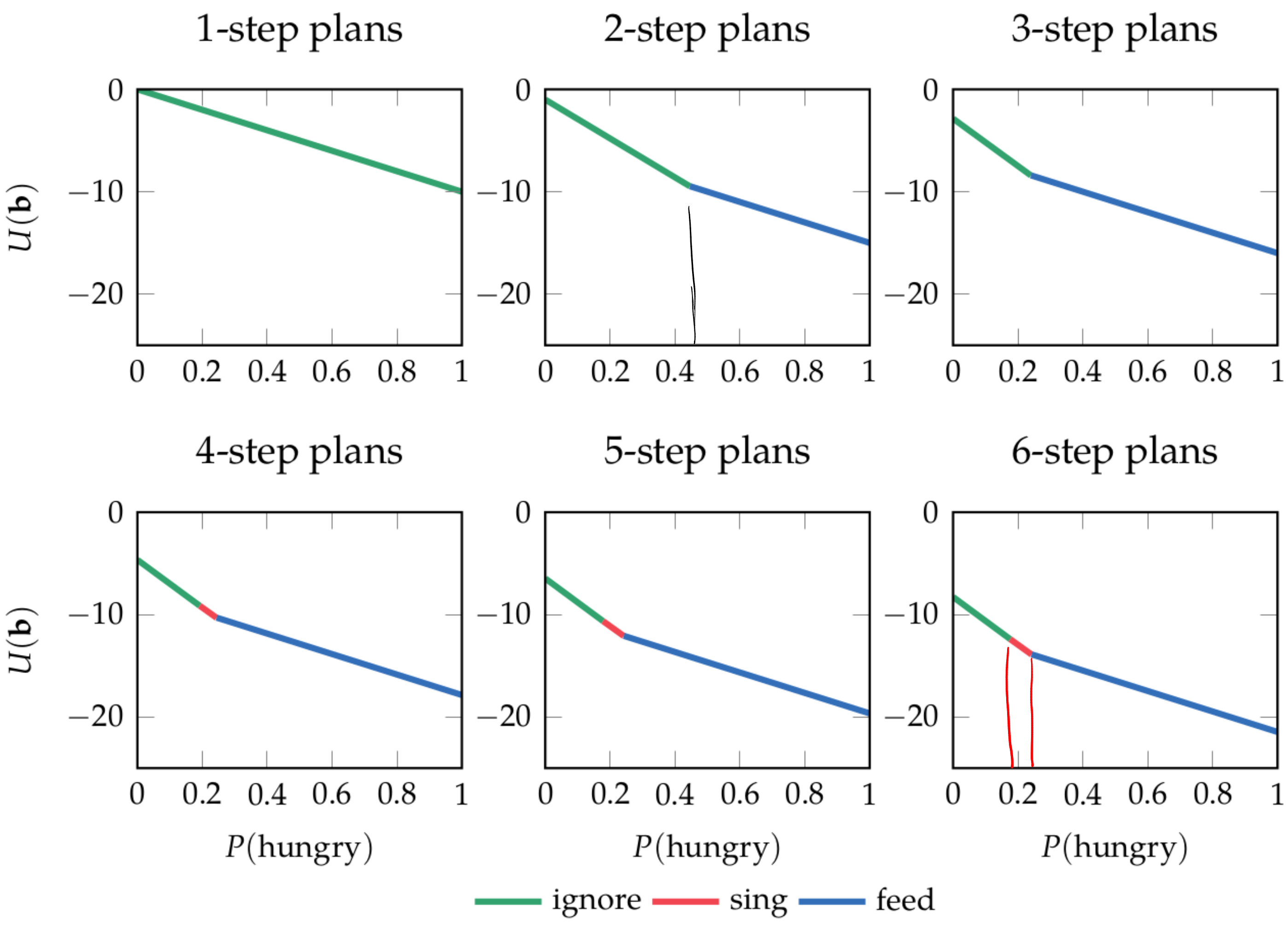
$|S| \approx 10$

for $n \in 1 \dots d$

    Construct Γ^n by expanding with Γ^{n-1}

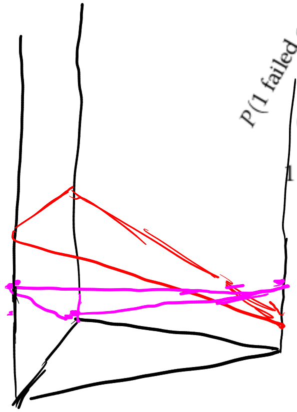
    Prune Γ^n

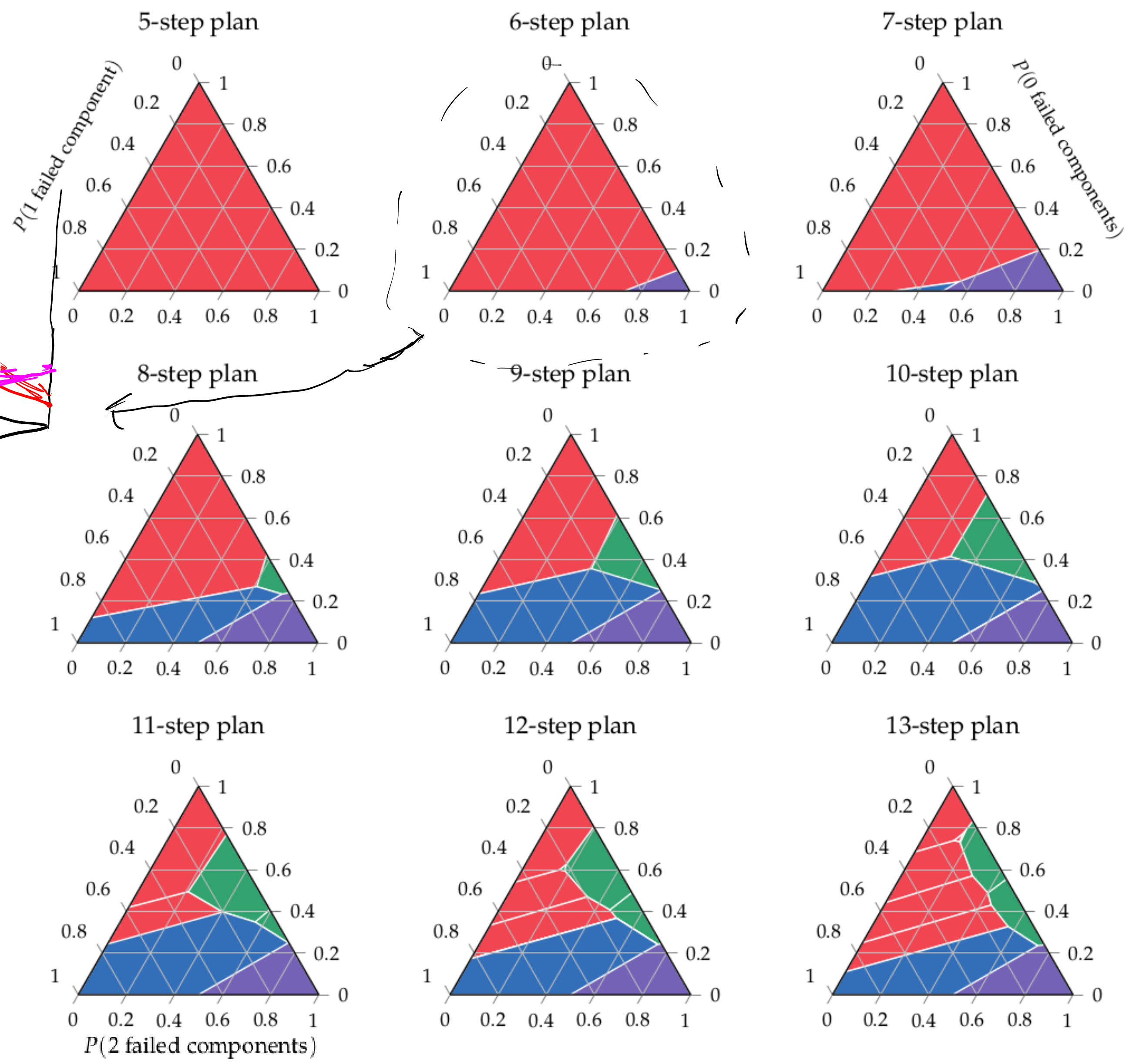
Finite Horizon POMDP Value Iteration

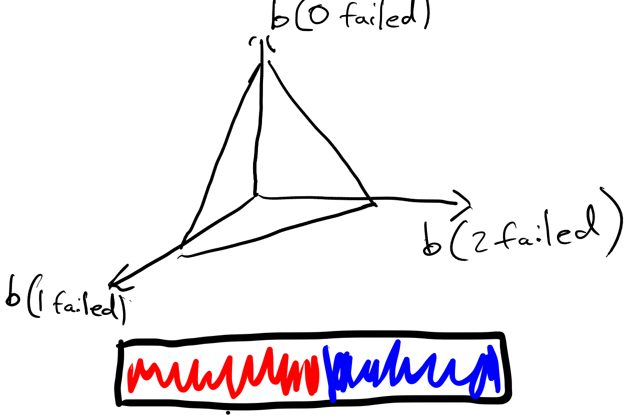


Finite Horizon POMDP Value Iteration

$|S| = 3$







Infinite-Horizon POMDP Lower Bound Improvement

Infinite-Horizon POMDP Lower Bound Improvement

$\alpha_a = (I - \gamma T^a)^{-1} R^a$

always execute same action

$\Gamma \leftarrow$ blind lower bound

loop

$\quad \Gamma \leftarrow \Gamma \cup \text{backup}(\Gamma)$

$\quad \Gamma \leftarrow \text{prune}(\Gamma)$

A survey of point-based POMDP solvers

$O(|\Gamma||A||O||S|^2 + |A||S||\Gamma|^{|O|})$

backup

$\Gamma' = \bigcup_{a \in A} \Gamma^a$

$\Gamma^a = \bigoplus_{o \in O} \Gamma^{a,o}$

$\Gamma^{a,o} = \{ \frac{1}{|O|} R_a + \alpha^{a,o} : \alpha \in \Gamma \}$

$\alpha^{a,o}[s] = \sum_{s'} Z(o \mid a, s') T(s' \mid s, a) \alpha[s']$

$\Gamma^1 \oplus \Gamma^2 = \{ \alpha_1 + \alpha_2 : \alpha_1 \in \Gamma^1, \alpha_2 \in \Gamma^2 \}$

Point-Based Value Iteration (PBVI)

point_backup(Γ, b)

  for $a \in A$

    for $o \in O$

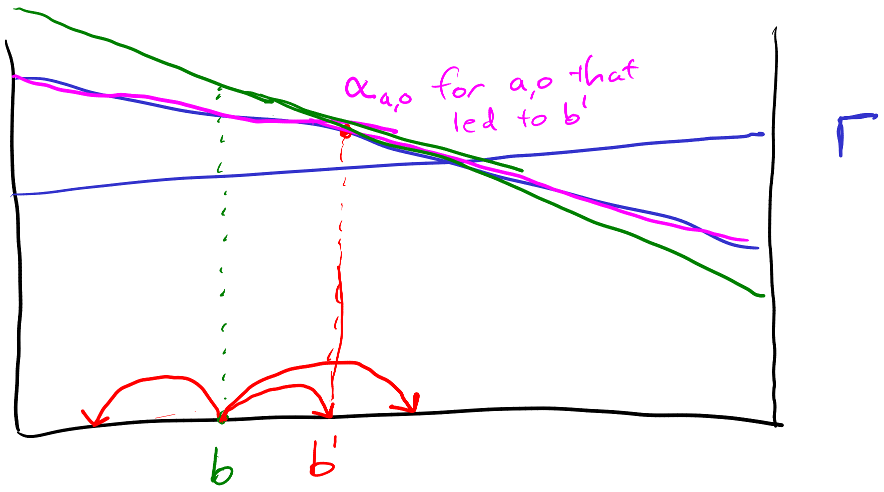
      $b' \leftarrow \tau(b, a, o)$

      $\alpha_{a,o} \leftarrow \underset{\alpha \in \Gamma}{\operatorname{argmax}}\ \alpha^\top b'$

    for $s \in S$

      $\alpha_a[s] = R(s,a) + \gamma \sum_{s',o} T(s' \mid s, a)\, Z(o' \mid a, s')\, \alpha_{a,o}[s']$

  return $\underset{\alpha_a}{\operatorname{argmax}}\ \alpha_a^\top b$



Original PBVI

How do we choose B

$B \leftarrow b_0$

loop

for $b \in B$

$\Gamma \leftarrow \Gamma \cup \{\text{point\_backup}(\Gamma, b)\}$

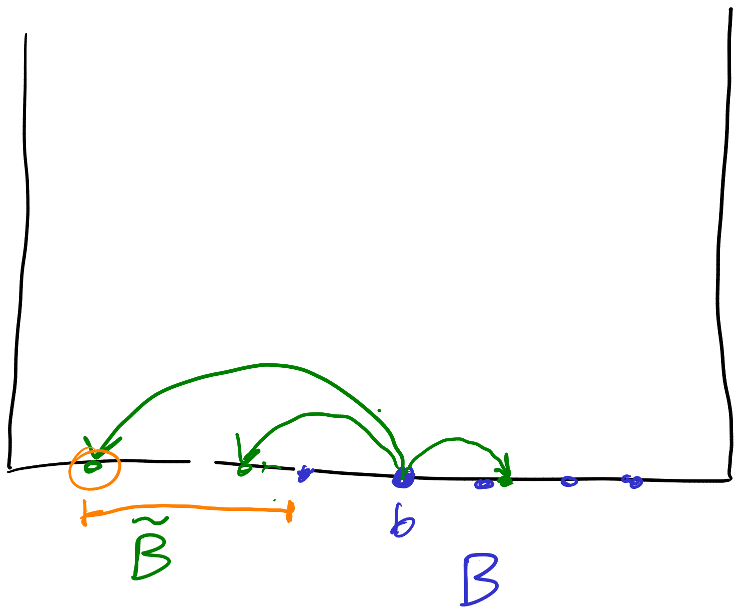
$B' \leftarrow \emptyset$

for $b \in B$

$\tilde{B} \leftarrow \{\tau(b, a, o) : a \in A, o \in O\}$

$$B' \leftarrow B' \cup \left\{ \underset{b' \in \tilde{B}}{\text{argmax}} \; \|B, b'\| \right\}$$

$B \leftarrow B \cup B'$



PERSEUS: Randomly Selected Beliefs

Two Phases:

1. Random Exploration
2. Value Backup

Random Exploration:

$B \leftarrow \emptyset$

$b \leftarrow b_0$

loop until $|B| = n$

$\quad a \leftarrow \text{rand}(A)$

$\quad o \leftarrow \text{rand}(P(o \mid b, a))$

$\quad b \leftarrow \tau(b, a, o)$

$\quad B = B \cup \{b\}$

Heuristic Search Value Iteration (HSVI)

$\underline{\Gamma}$

while $\overline{V}(b_0) - \underline{V}(b_0) > \epsilon$

explore$(b_0, 0)$

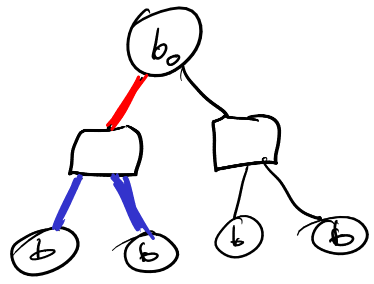
$\underline{V}(b_0)$ $\overline{V}(b_0)$

$\underline{\Gamma} = \{\alpha_1, \alpha_2 \ldots \alpha_n\}$

$B = \{b_1, b_2 \ldots b_m\}$

$\overline{V}$ —

$\underline{V}$ —

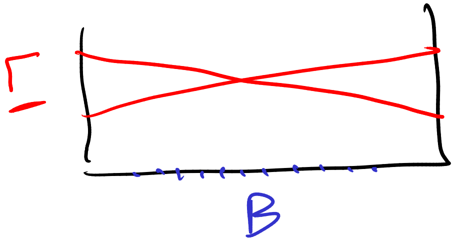


explore(b, t)

if $\overline{V}(b) - \underline{V}(b) > \epsilon\gamma^t$

promising

$a^* = \operatorname*{argmax}_a \overline{Q}(b, a)$



observations
1. likely
2. lots of uncertainty

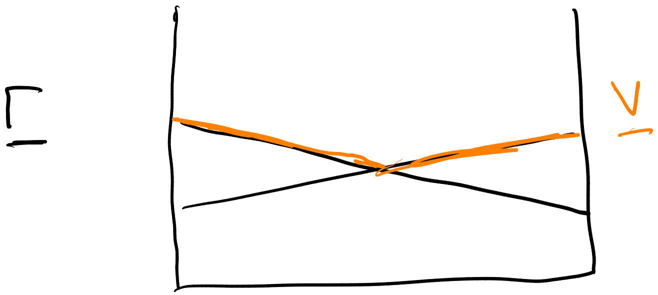
$o^* = \operatorname*{argmax}_o P(o \mid b, a^*) \left(\overline{V}(\tau(b, a^*, o)) - \underline{V}(\tau(b, a^*, o)) - \epsilon\gamma^t\right)$

Weighted Excess Uncertainty

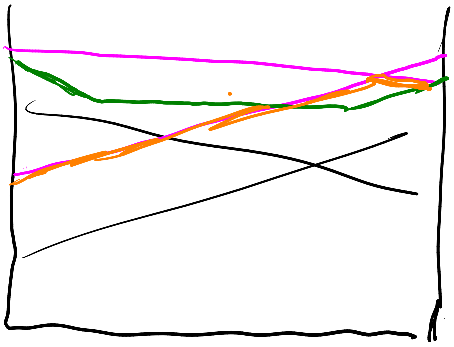
explore$(\tau(b, a^*, o^*), t + 1)$

$\underline{\Gamma} \leftarrow \underline{\Gamma} \cup \text{point\_backup}(\underline{\Gamma}, b)$

$\overline{V}(b) = B_b[\overline{V}(b)]$



How do we represent an upper bound with alpha vectors

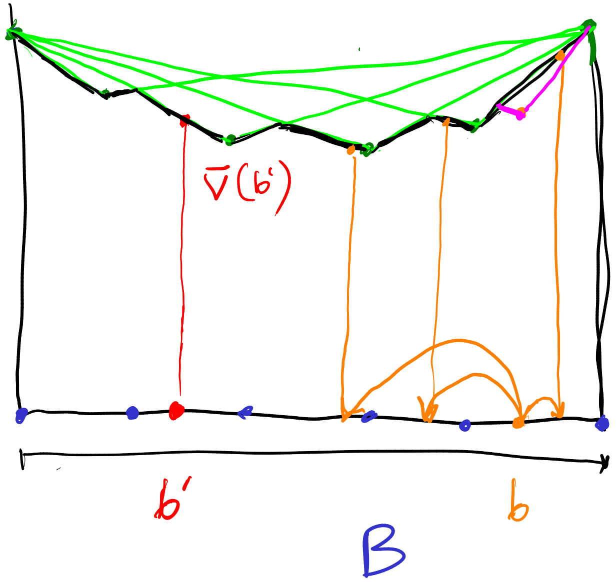


$$\bar{V}(b) = \min_{\alpha \in \bar{\Gamma}} \alpha^T b$$

$$\underline{V}(b) = \max_{\alpha \in \underline{\Gamma}} \alpha^T b$$

Sawtooth Upper Bounds

for each $b \in \underline{B}$, store $\bar{V}(b)$
and the vertices of belief simplex

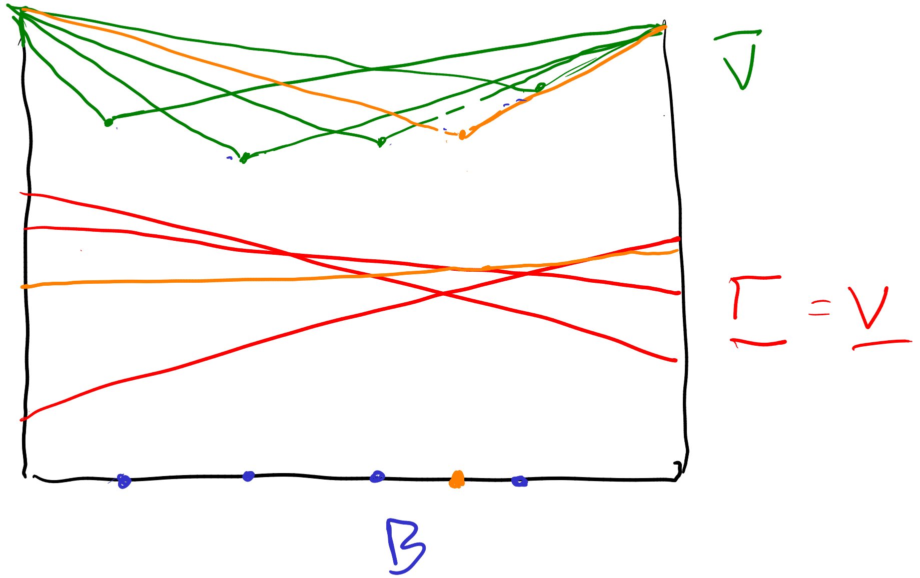


$\bar{V}(b)$ for $b \in B \cup$ vertices

$$B_b[\bar{V}](b) = \max_a \left\{ R(b,a) + \gamma \sum_o P(o|b,a)\bar{V}(\tau(b,a,o)) \right\}$$

$$\frac{\bar{R}}{1-\gamma} \quad = \max_{a,s} R(s,a)$$

What HSVI is working with



SARSOP

Successive Approximation of Reachable Space under Optimal Policies

Similar to HSVI

HSVI

$B \subset R$

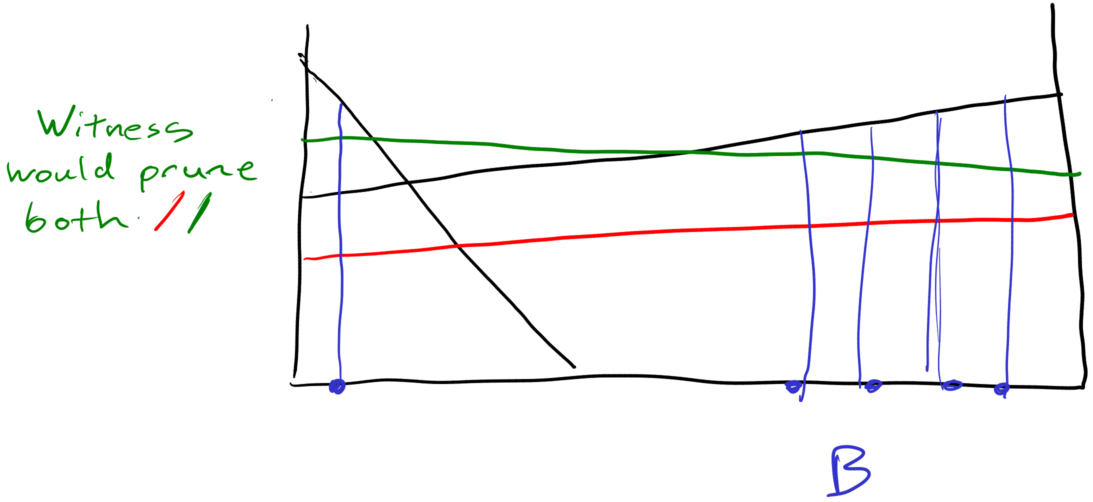
reachable

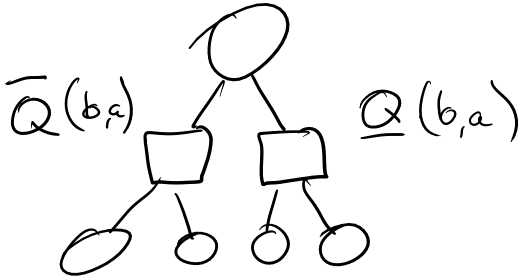
SARSOP

$B \subset R^*$

beliefs reachable under optimal policy

+ a few other tricks





if $\bar{Q}(b,a^1) < \underline{Q}(b,a^2)$

then remove all b. below (b,a^1) from B

Witness: ~20 states

SARSP: 10,000 – 100,000 states

Offline POMDP Algorithms

Policy Graphs

Monte Carlo Value Iteration (MCVI)

MC- Backup (G, b, N)

$R_a = 0 \quad V_{a,o,v} = 0$

for $a \in A$

for i in $1:N$

$s_i \leftarrow \text{sample}(b)$

$s_i', o_i, r_i \leftarrow G(s_i, a)$

$R_a + r_i$

for $v \in G$

$V_{a,o_i,v} = V_{a,o_i,v} + \text{Simulate}(G, v, s_i', L)$

for o in O

$V_{a,o} = \max\limits_{v \in G} V_{a,o,v}$

$v_{a,o} = \operatorname{argmax}\limits_{v \in G} V_{a,o,v}$

$V_a = R_a + \gamma \sum\limits_o V_{ao}/N$

$V^* = \max\limits_a V_a$

$a^* = \operatorname{argmax} V_a$

add new node to G labeled with a^*