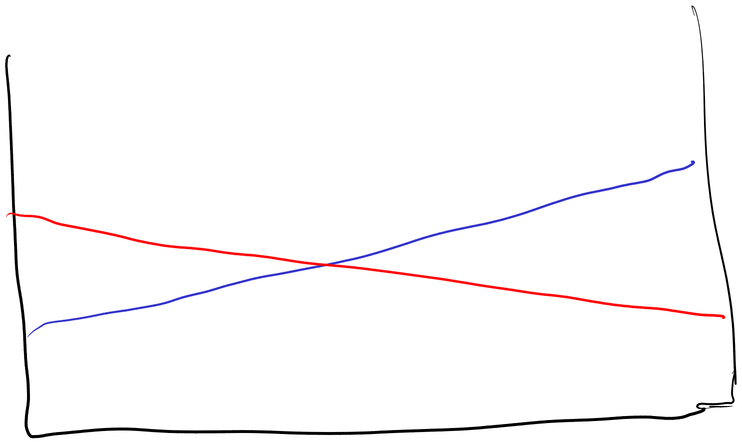
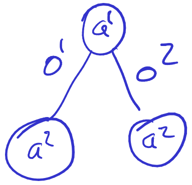
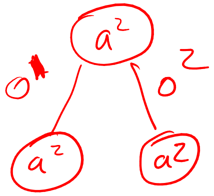
Offline POMDP Algorithms

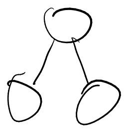


b( )





Last time: POMDP Value Iteration (horizon d)



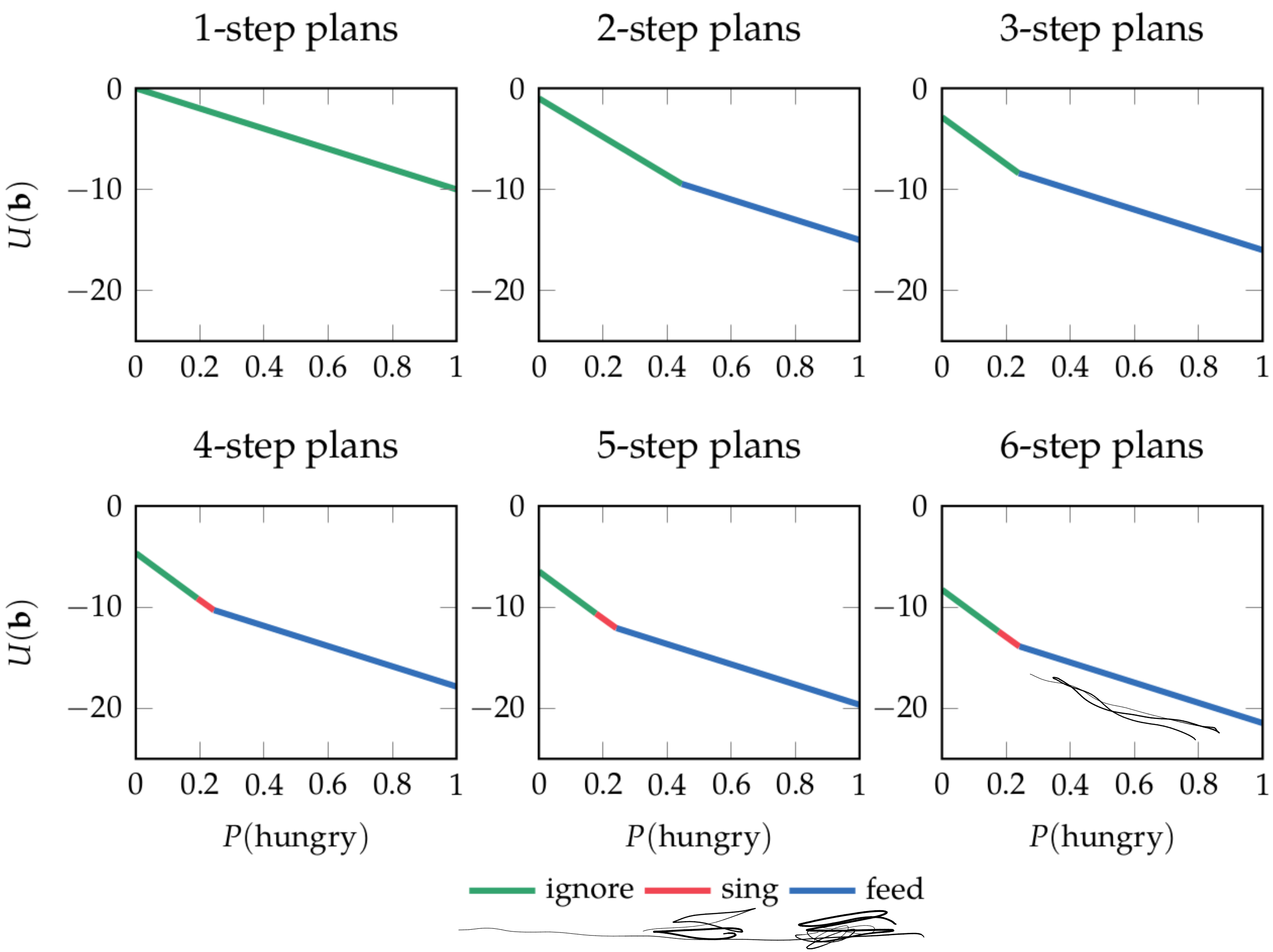
$\Gamma^0 \leftarrow \emptyset$

for $n \in 1 \ldots d$

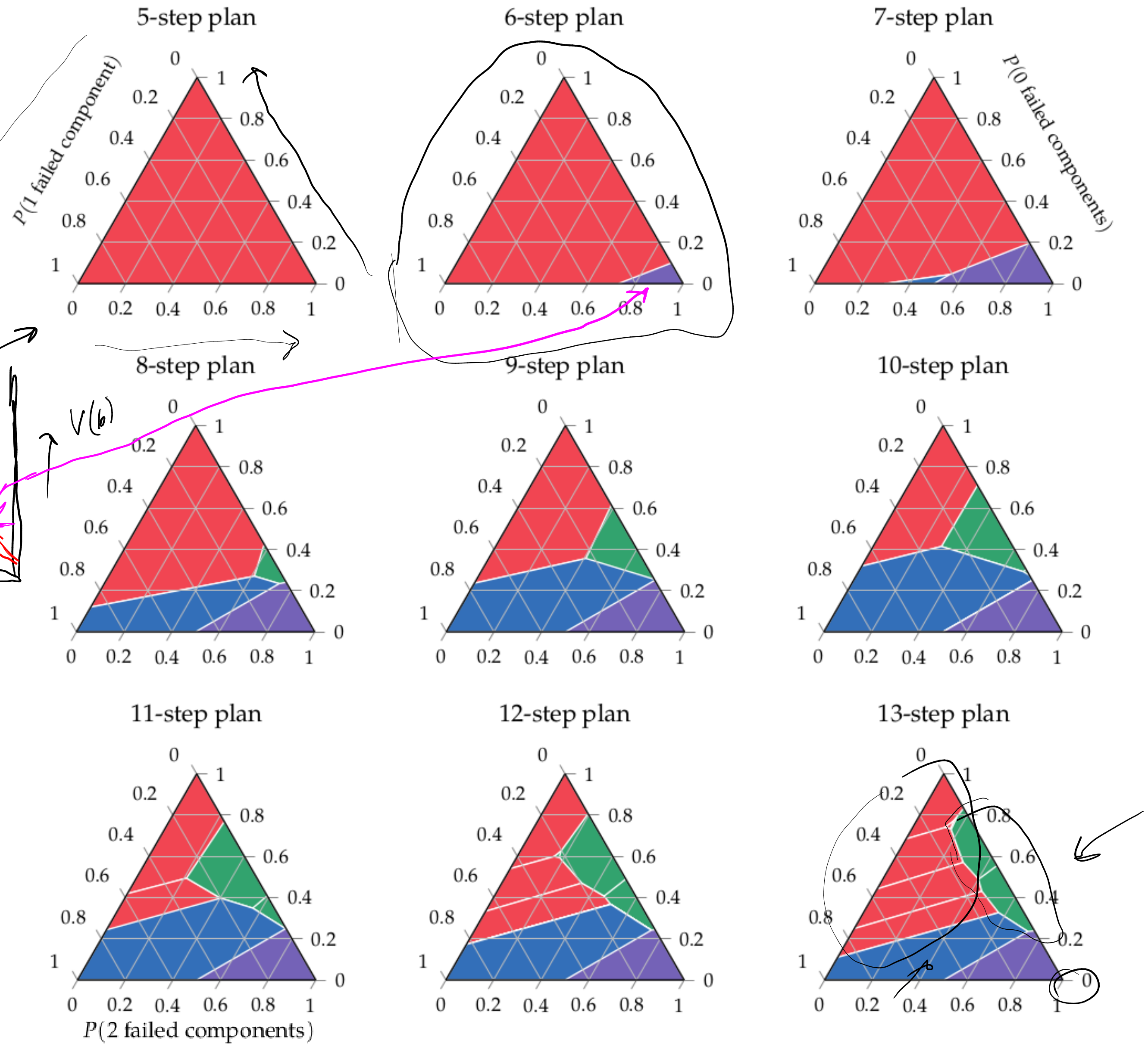
Construct Γ^n by expanding with Γ^{n-1}

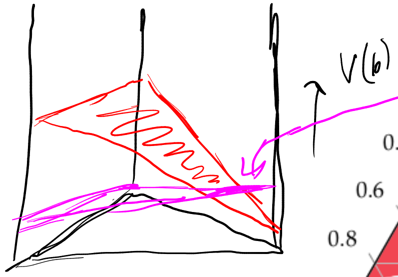
Prune Γ^n

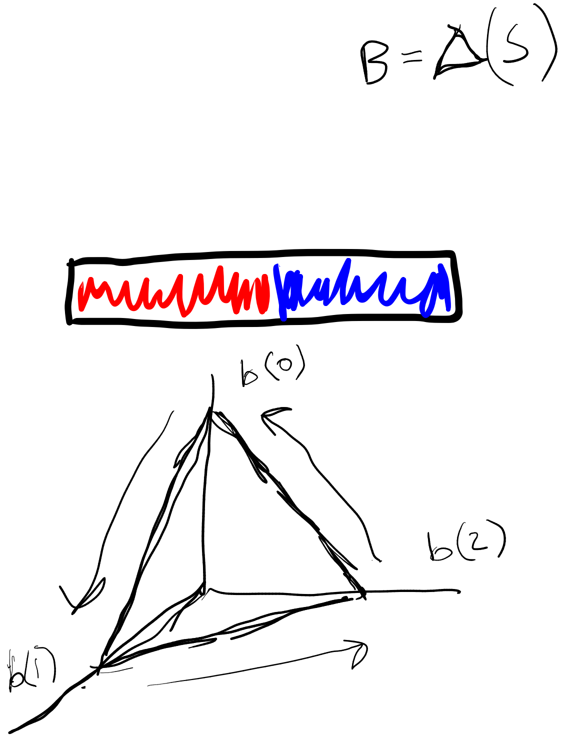
Finite Horizon POMDP Value Iteration



Finite Horizon POMDP Value Iteration







Infinite-Horizon POMDP Lower Bound Improvement

$\frac{R}{1-\gamma}$

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

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

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

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

backup

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

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

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

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

A survey of point based POMDP solvers

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

Point-Based Value Iteration (PBVI)

point-backup(Γ, b)

for $a \in A$

for $o \in O$

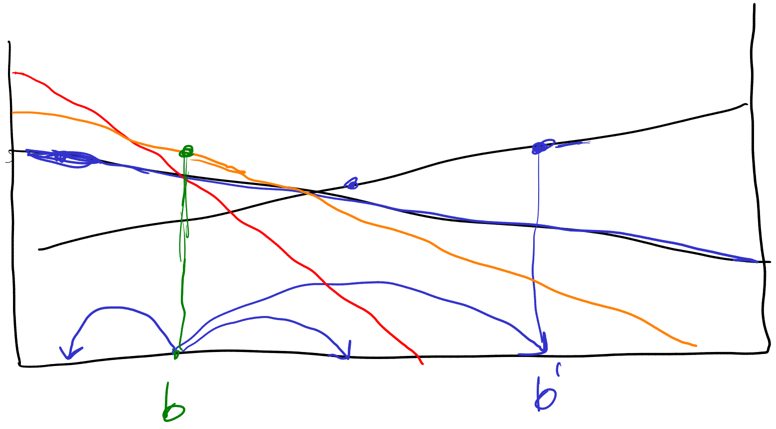
$b' \leftarrow \tau(b, a, o)$

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

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



$\mathcal{B}$

If we perform a backup for each $b \in \mathcal{B}$

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

Original PBVI

how do we choose B

$B \leftarrow b_0$

loop

  for $b \in B$

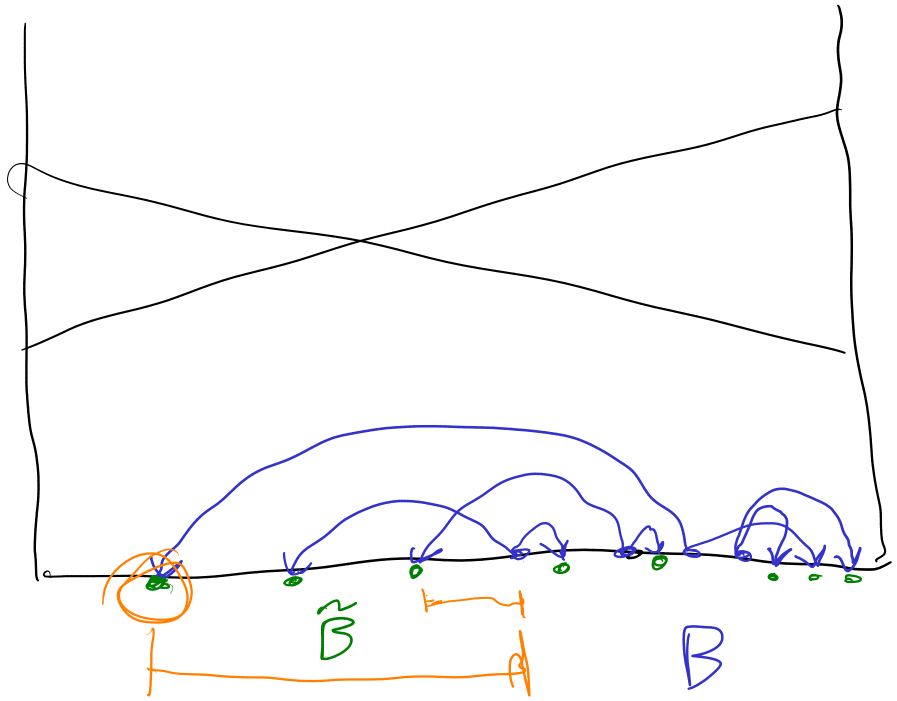
    $\Gamma \leftarrow \Gamma \cup \{\text{point\_backup}(\Gamma, b)\}$

  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:

$B \leftarrow \emptyset$

$b \leftarrow b_0$

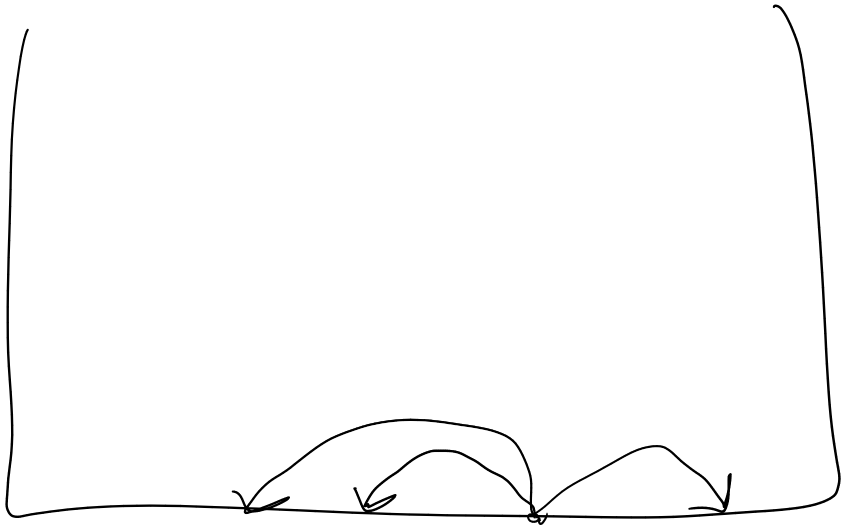
loop until $|B| = n$

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

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

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

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



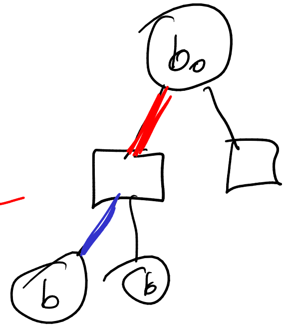
Heuristic Search Value Iteration (HSVI)

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

  explore$(b_0, 0)$

$\overline{V}(b)$ upper bound for $b \in B$

$\underline{V}(b)$ lower bound



function explore(b, t)

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

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

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

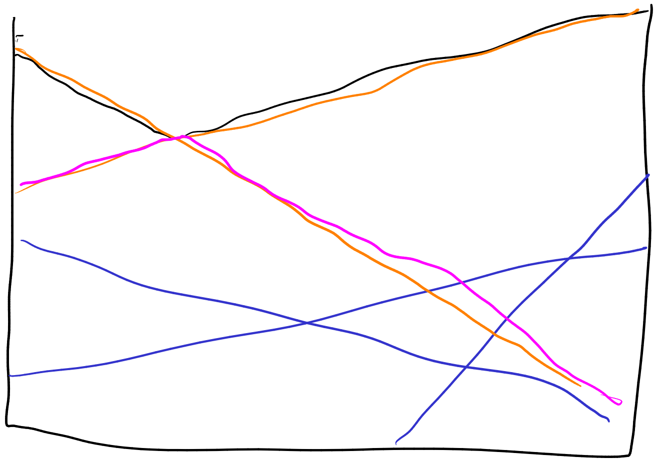
    WEU

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

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

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

Sawtooth Upper Bounds

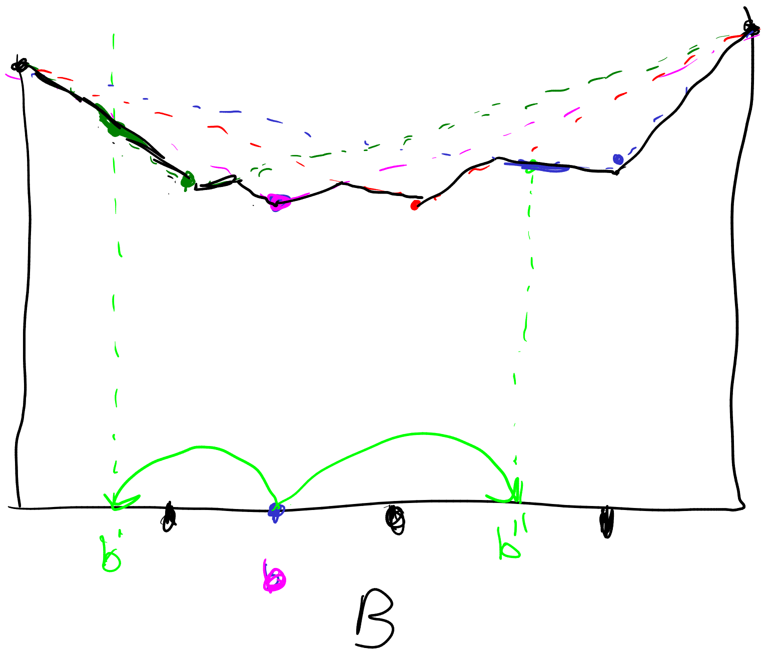


$\cancel{\min_{\alpha} \alpha^T b}$

lower bound

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

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

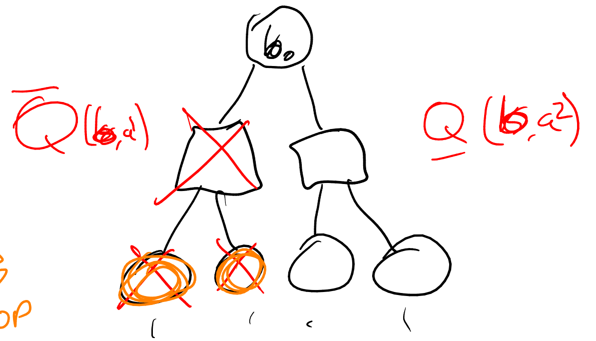


SARSOP

Successive Approximation of Reachable Space under Optimal Policies

Similar to HSVI

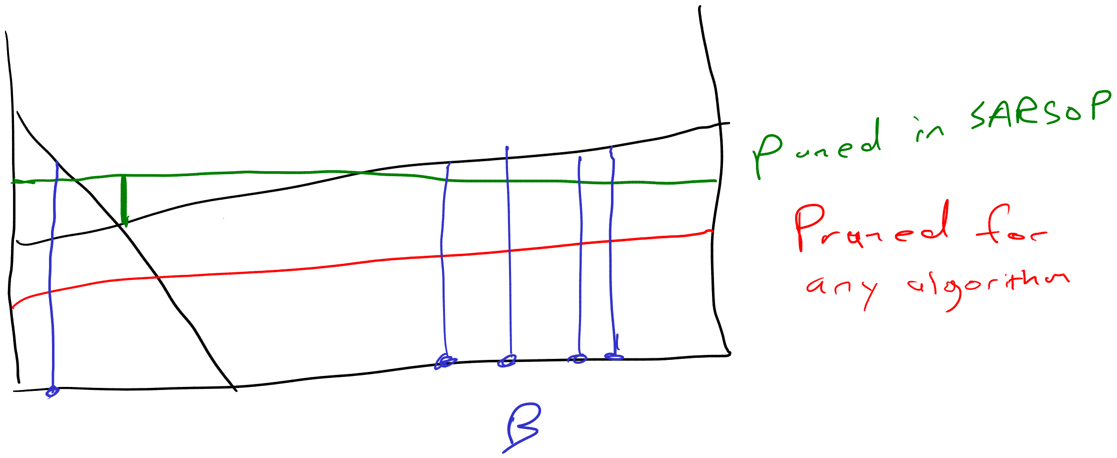
HSVI	SARSOP
$B \subset R$	$B \subset R^*$
reachable	reachable under optimal policy



Not in B for SARSOP

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

then prune all b below (b,a^1)



Witness (α-vector value iteration) : ~20 states

SARSOP: 10,000–100,000 states

Offline POMDP Algorithms

Policy Graphs

Monte Carlo Value Iteration (MCVI)