Offline POMDP Algorithms

Survey of Point Based POMDP Solvers

Name	S	A	O
Exact VI	D (10)	D	D
PBVI			
Perseus			
HSVI			
SARSOP ← Use	D (1000-10,000)		
MCVI	C	D	D/C

Infinite-Horizon POMDP Value Iteration

constant-action policy

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

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

$\Gamma \leftarrow$ blind lower bound

loop

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

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

backup

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

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

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

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

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

Point-Based Value Iteration (PBVI)

backup(Γ, b)

for $a \in A$

for $o \in O$

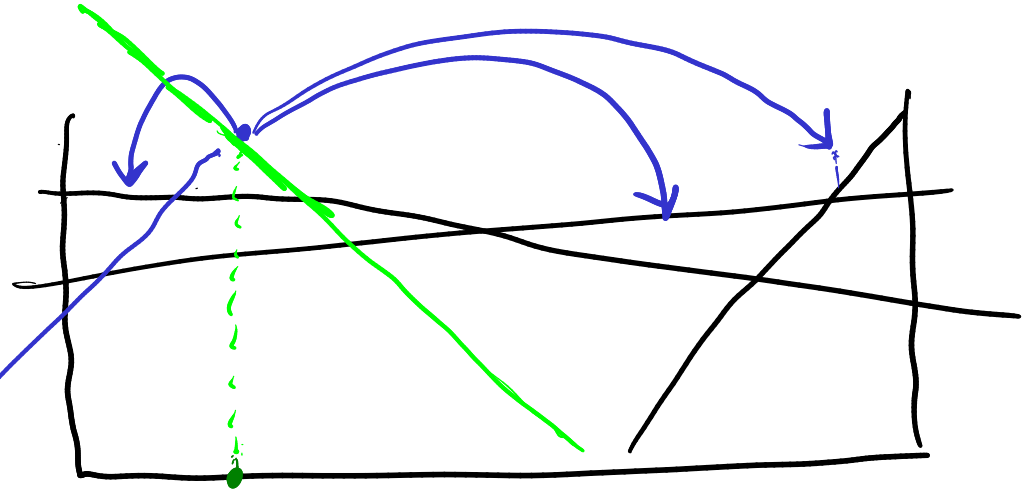
$b' \leftarrow \tau(b, a, o)$

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

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



If we perform a point based backup for every $b \in B$

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

How do we choose B

Original PBVI

$B \leftarrow b_0$

loop

  for $b \in B$

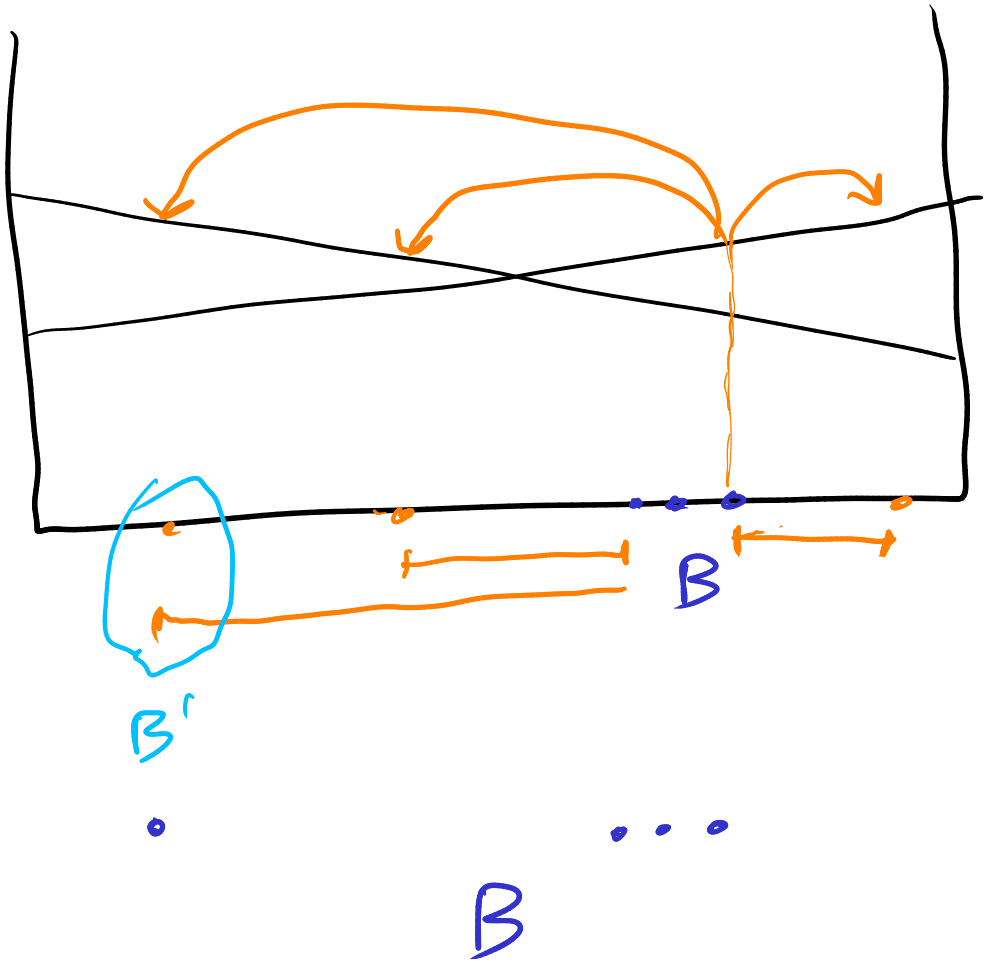
    $\Gamma \leftarrow \Gamma \cup \{\text{point\_backup}(\Gamma, b)\}$

  for $b \in B$

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

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

  $B \leftarrow B \cup B'$



PERSEUS: Randomly Selected Beliefs

Two Phases:

1. Random Exploration
2. Value Backup

Random Exploration:

$B \leftarrow \emptyset$

$b \leftarrow b_0$

loop until $|B| = n$

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

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

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

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

Heuristic Search Value Iteration (HSVI)

$\overline{V}(b)$ upper bound $\quad \underline{V}(b)$ lower bound

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

explore$(b_0, 0)$

function explore(b, t)

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

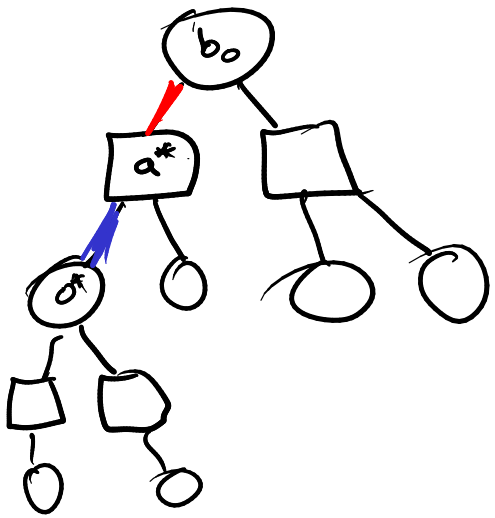
$a^* = \underset{a}{\operatorname{argmax}}\ \overline{Q}(b, a)$ ← maximize upper bound

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

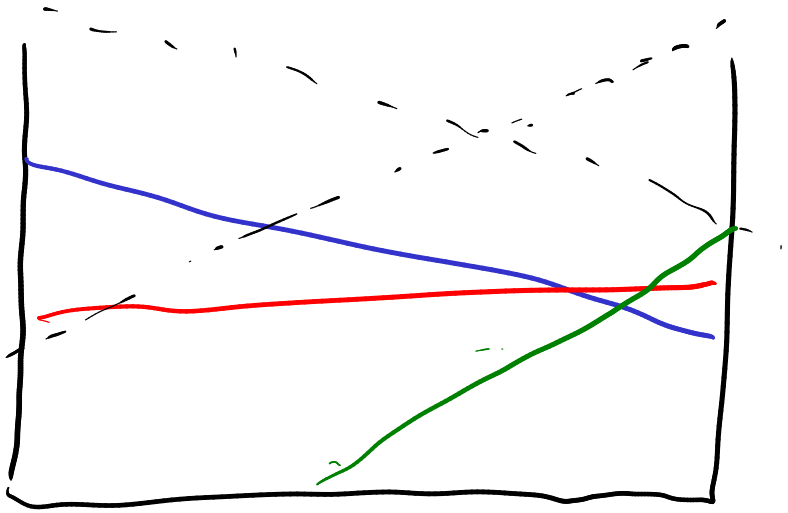
explore$(\tau(b, a^*, o^*), t + 1)$

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

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



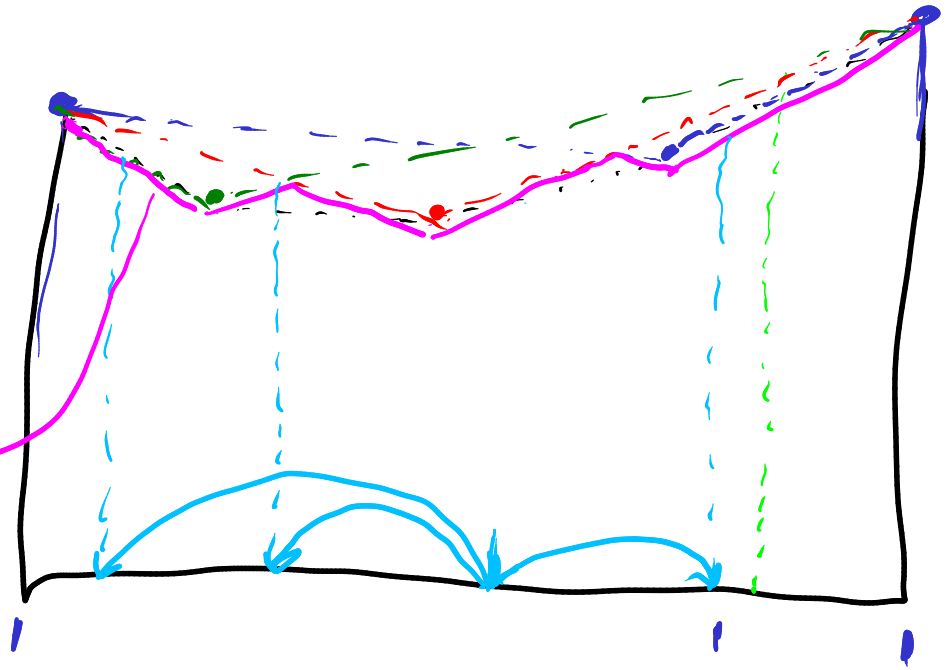
Sawtooth Upper Bounds



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

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

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



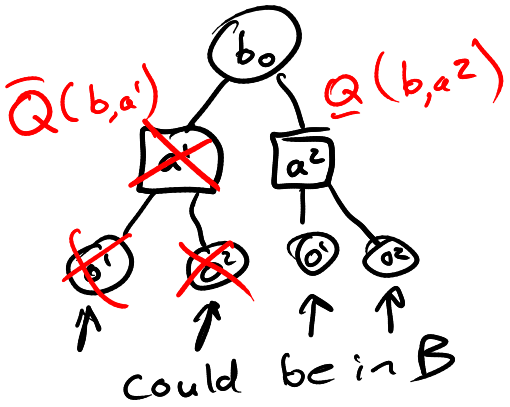
Finding the points to interpolate between requires solving a linear program

SARSOP

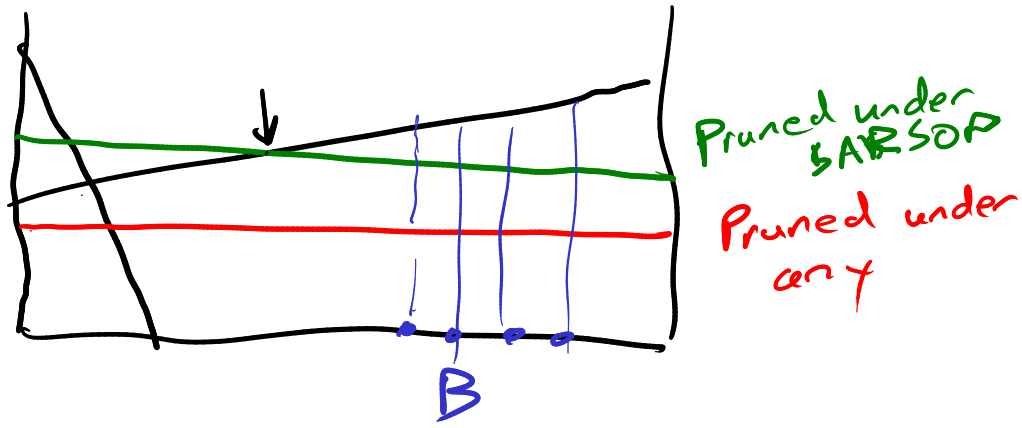
Successive Approximation of Reachable Space under Optimal Policies

Similar to HSVI

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



if $\bar{Q}(b,a^1) < \underline{Q}(b,a^2)$
then prune all b
below (b,a^1) from B



Instead of pruning α that are dominated over whole belief space, prune α dominated over B

Policy Graphs

Monte Carlo Value Iteration (MCVI)