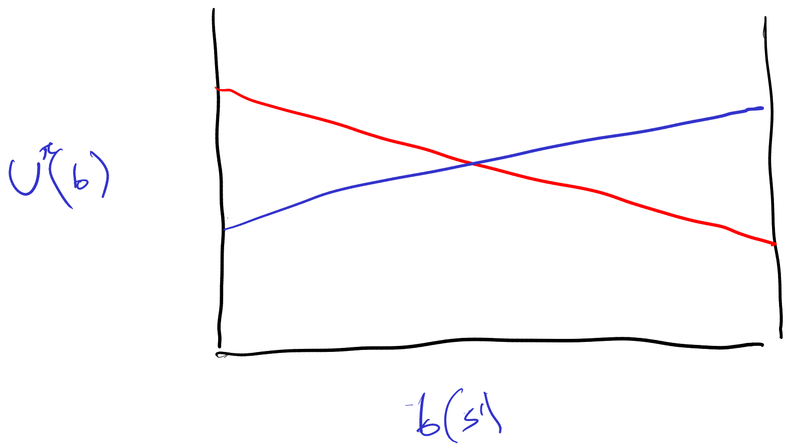
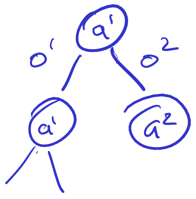
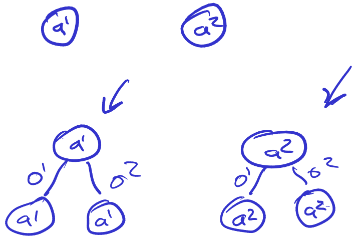
Offline POMDP Algorithms

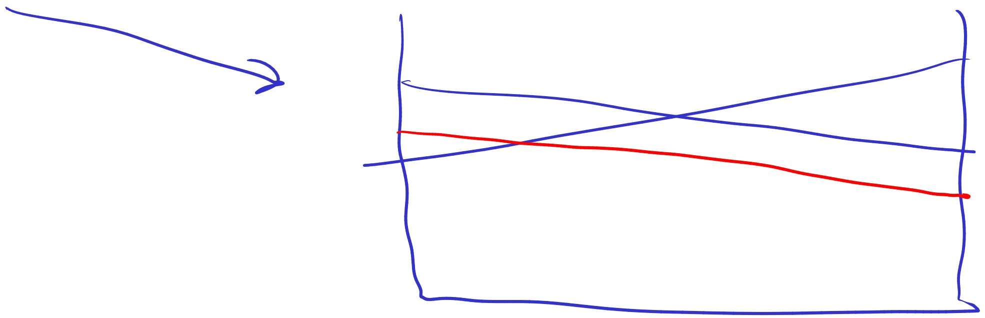




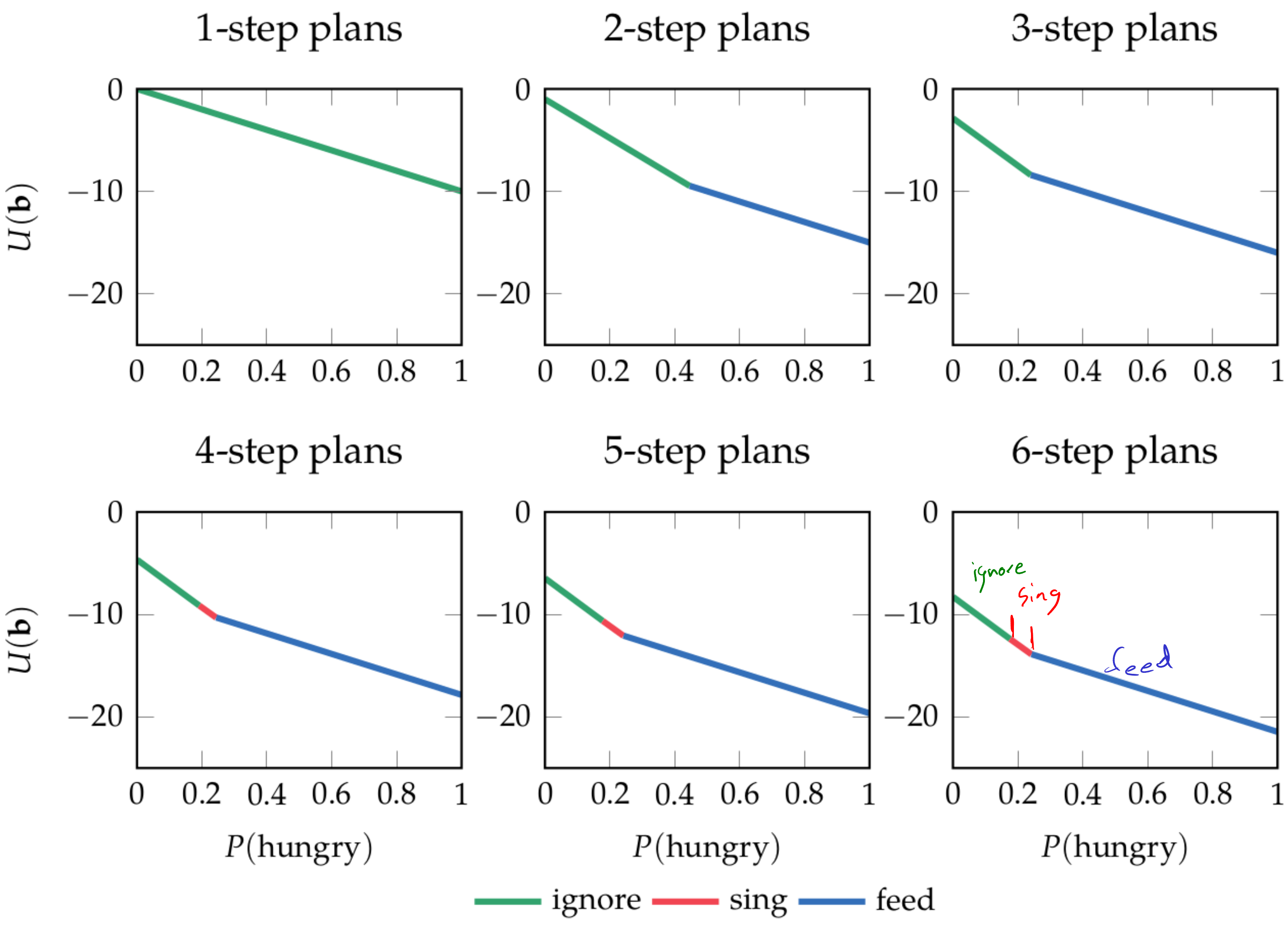
Last time: POMDP Value Iteration (horizon d)

```
$\Gamma^0 \leftarrow \emptyset$
for $n \in 1 \ldots d$
    Construct $\Gamma^n$ by expanding with $\Gamma^{n-1}$
    Prune $\Gamma^n$
```

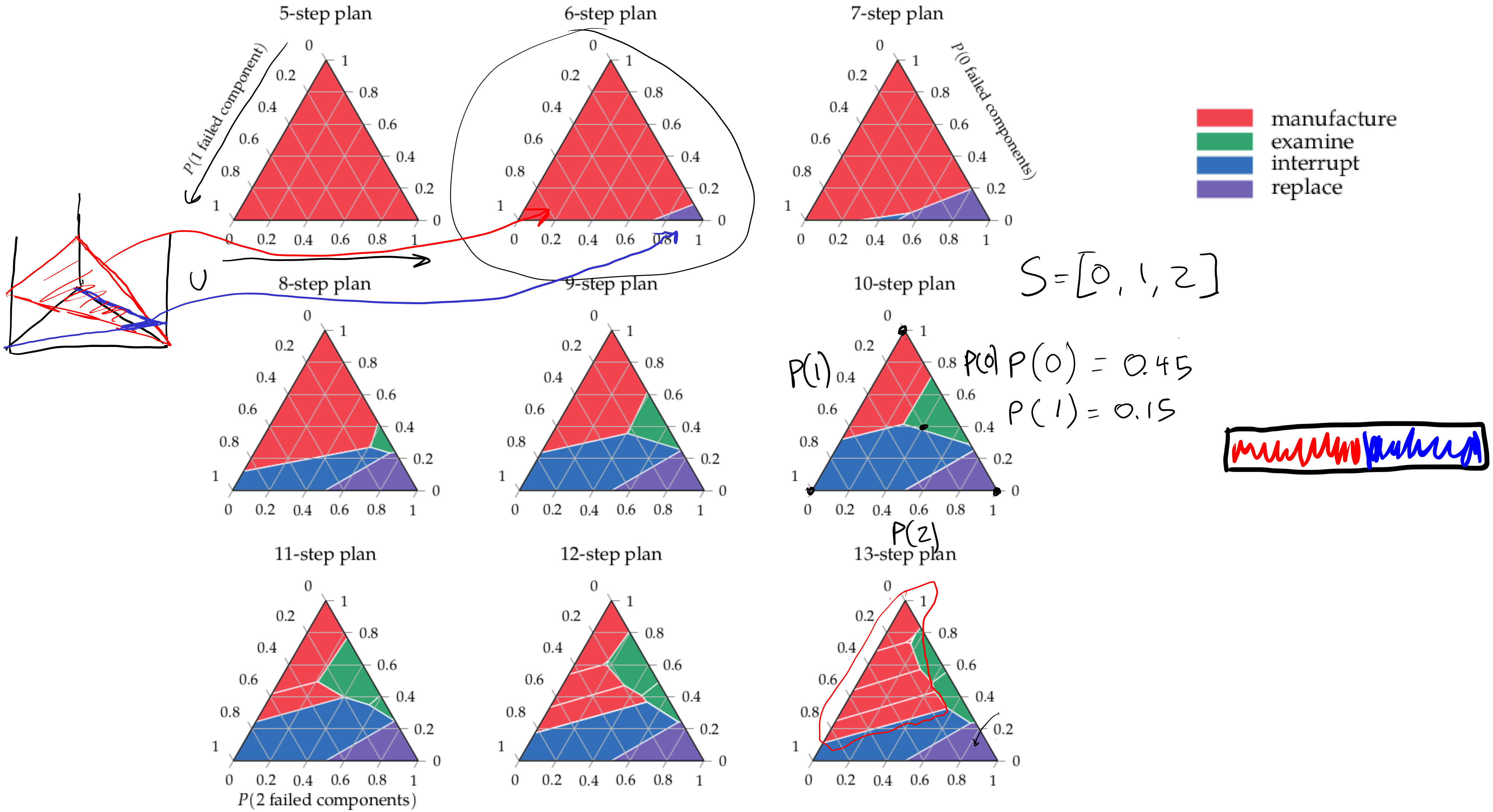




Finite Horizon POMDP Value Iteration



Finite Horizon POMDP Value Iteration



5-step plan

6-step plan

7-step plan

8-step plan

9-step plan

10-step plan

11-step plan

12-step plan

13-step plan

S=[0,1,2]

P(0) P(0) = 0.45

P(1) = 0.15

Infinite-Horizon POMDP Lower Bound Improvement

Infinite-Horizon POMDP Lower Bound Improvement

always execute same action

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

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

```
$\Gamma \leftarrow$ blind lower bound
loop
    $\Gamma \leftarrow \Gamma \cup \text{backup}(\Gamma)$
    $\Gamma \leftarrow \text{prune}(\Gamma)$
```

backup

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

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

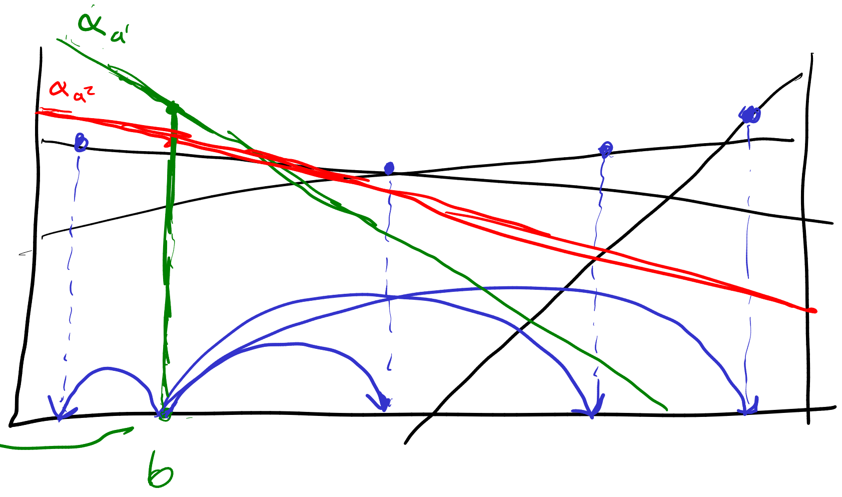
$$\Gamma^{a,o} = \left\{ \frac{1}{|O|} r_a + \alpha^{a,o} : \alpha \in \Gamma \right\}$$

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

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

Point-Based Value Iteration (PBVI)

```
point_backup($\Gamma$, $b$)
    for $a \in A$
        for $o \in O$
            $b' \leftarrow \tau(b, a, o)$
            $\alpha_{a,o} \leftarrow \underset{\alpha \in \Gamma}{\text{argmax}}\ \alpha^\top b'$
        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 $\underset{\alpha_a}{\text{argmax}}\ \alpha_a^\top b$
```



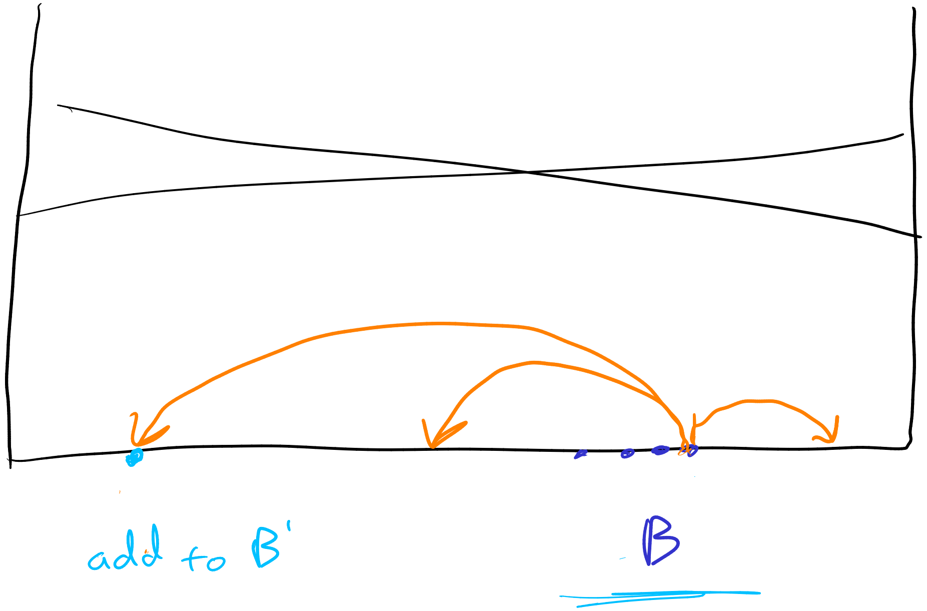
If we have a set of beliefs, B and perform a point-based backup for each $b \in B$

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

How to choose B

Original PBVI

```
$B \leftarrow b_0$
loop
    for $b \in B$
        $\Gamma \leftarrow \Gamma \cup \{\text{point\_backup}(\Gamma, b)\}$
    for $b \in B$
        $\tilde{B} \leftarrow \{\tau(b, a, o) : a \in A, o \in O\}$
        $B' \leftarrow B' \cup \left\{ \underset{b' \in \tilde{B}}{\text{argmax}} \, \|B, b'\| \right\}$
    $B \leftarrow B \cup B'$
```



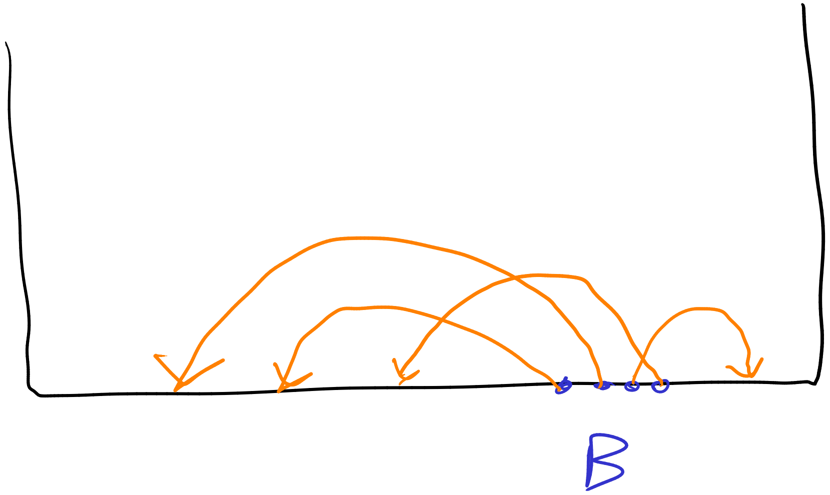
PERSEUS: Randomly Selected Beliefs

Two Phases:

1. Random Exploration
2. Value Backup

Random Exploration:

```
$B \leftarrow \emptyset$
$b \leftarrow b_0$
loop until $\underline{|B| = n}$
    $a \leftarrow \text{rand}(A)$
    $o \leftarrow \text{rand}(P(o \mid b, a))$
    $b \leftarrow \tau(b, a, o)$
    $B = B \cup \{b\}$
```



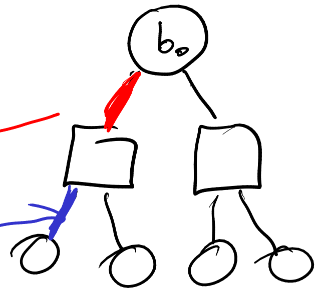
Heuristic Search Value Iteration (HSVI)

$\overline{V}(b)$ upper bound

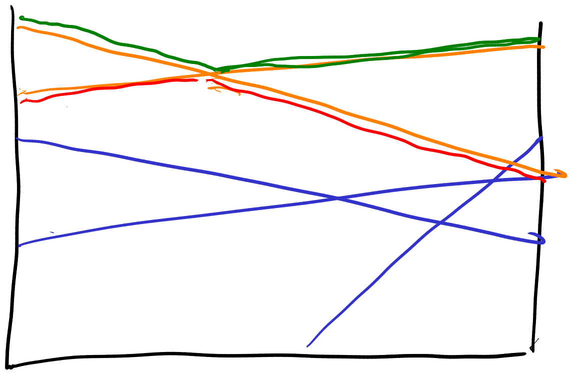
$\underline{V}(b)$ lower bound

```
while $\overline{V}(b_0) - \underline{V}(b_0) > \epsilon$
    explore($b_0, 0$)

function explore(b, t)
    if $\overline{V}(b) - \underline{V}(b) > \epsilon\gamma^t$
        $a^* = \operatorname*{argmax}_a \overline{Q}(b, a)$
        $o^* = \operatorname*{argmax}_o P(o \mid b, a) \overbrace{\left(\overline{V}(\tau(b, a^*, o)) - \underline{V}(\tau(b, a^*, o)) - \epsilon\gamma^t\right)}^{\text{Excess Uncertainty}}$
        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

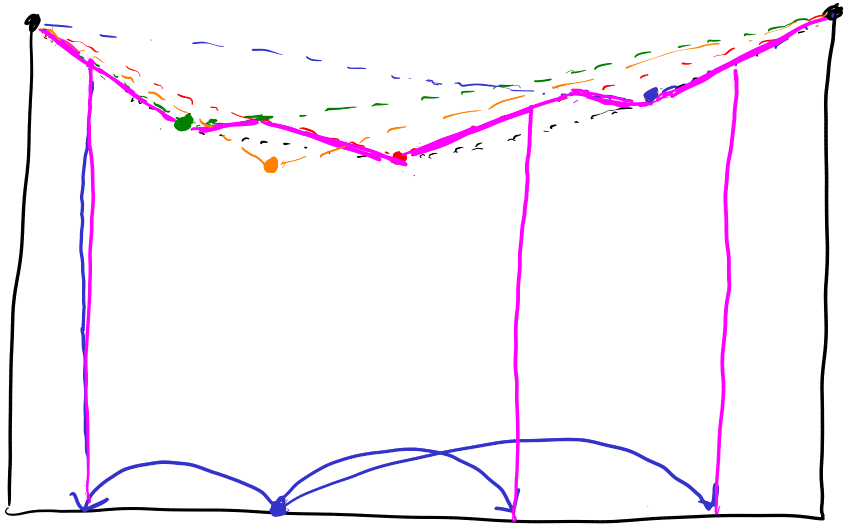


~~$\min_{\alpha} \alpha^T b$~~ X

lower bound

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

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



SARSOP

Successive Approximation of Reachable Space under Optimal Policies

Similar to HSVI

HSVI

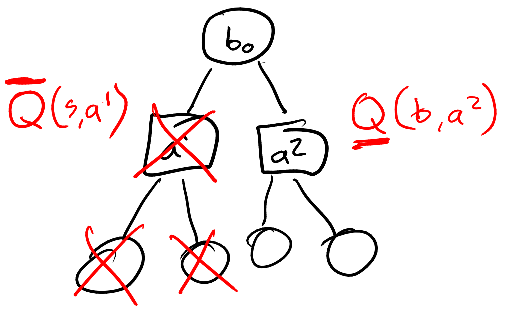
$B \subset R$

↑ reachable

SARSOP

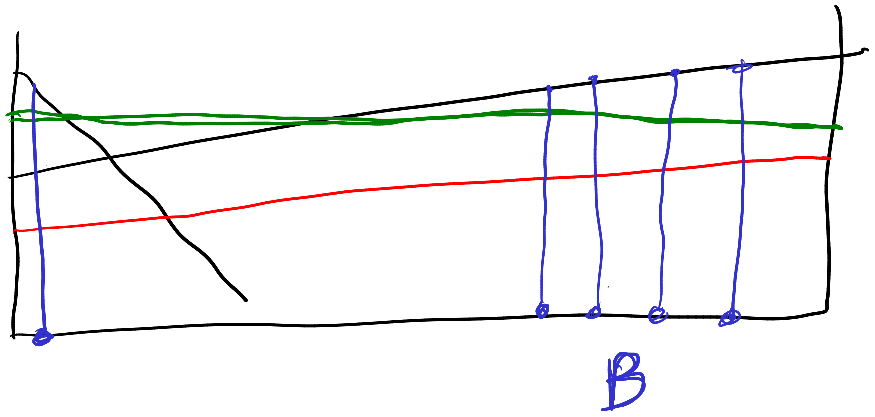
$B \subset R^*$

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



Pruned under SARSOP b/c not optimal for any $b \in B$

Prune under any algorithm

Offline POMDP Algorithms

Policy Graphs

Monte Carlo Value Iteration (MCVI)