Last Time

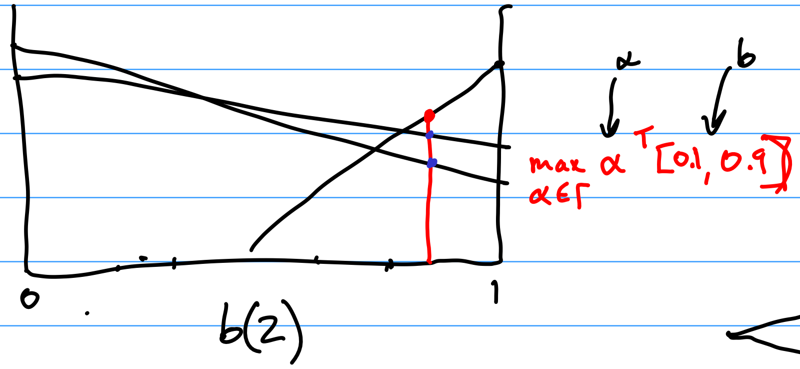
α vectors

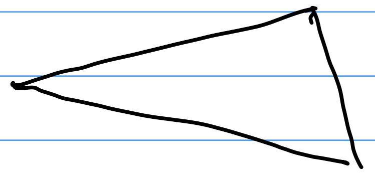
$\underset{x}{\alpha}[s]$ · $[0.4, 0.3, 0.2]$

set Γ

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

$S = \{1, 2\}$





Today

Offline POMDP

Name	S	A	O
Exact VI	D (10)	D	D
PBVI	D	D	D
Perseus	D	D	D
HSVI	D	D	D
SARSOP	D (1000)	D	D
MCVI	C	D	D/C
LQG	C	C	G

"Survey of Point-Based POMDP Solvers"

bigger

(Linear Gaussian Quadratic)

Exact VI

loop

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

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

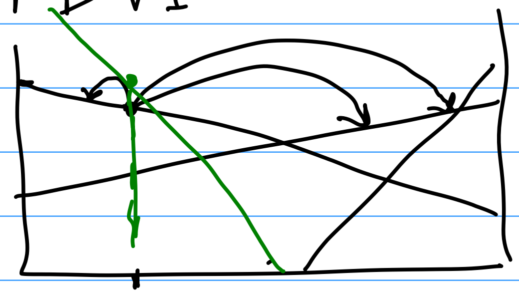
$$\Gamma^{ao} \doteq \left\{ \frac{1}{|O|} r_a + \alpha^{a,o} : \alpha \in \Gamma \right\}$$

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

for 1 iteration

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

PBVI

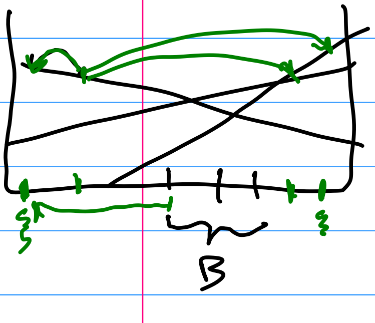


```
BackupBelief($(\Gamma, b)$)
    for $a \in A$
        for $o \in O$
            $b' = \tau(b,a,o)$
            $\alpha_{a,o} = \underset{\alpha \in \Gamma}{\operatorname{argmax}} \alpha^T b'$
        for $s \in S$
            $\alpha_a[s] = R(s,a) + \gamma \sum_{s',o} T(s'|s,a) Z(o'|a,s') \alpha_{a,o}[s']$
    return $\underset{\alpha_a}{\operatorname{argmax}} \alpha_a^T b$
```

$$O(|\Gamma||A||O||S|^2 + |A||S||\Gamma||\underbrace{B}_{\text{set of Beliefs}}|)$$

How we choose B

Original PBVI



```
$B = \{b_0\}$
loop $B' = \phi$
    for $b \in B$
        $\Gamma = \Gamma \cup$ BackupBelief$(\Gamma, b)$
    for $b \in B$
        $\tilde{B} = \{\tau(b,a,o) : a \in A, o \in O\}$
        $B' = B' \cup \underset{b \in \tilde{B}}{\operatorname{argmax}} ||B, b'||$
    $B = B \cup B'$
```

add furthest b away from B

$B = \{b^1, b^2, b^3\}$

$B' = \{b^4, b^5\}$

$B \cup B' = \{b^1 \dots b^5\}$

Perseus - randomly choose B

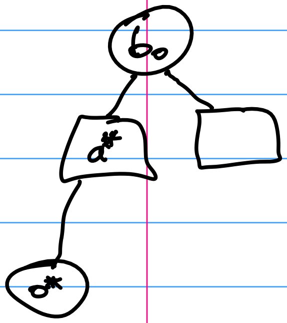
```
$B = \emptyset$
$b = b_0$
loop until $|B| = n$
    $a = rand(A)$
    $o = rand(P(o|b,a))$
    $B = B \cup \{\tau(b,a,o)\}$
```

HSVI Heuristic Search Value Iteration

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

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

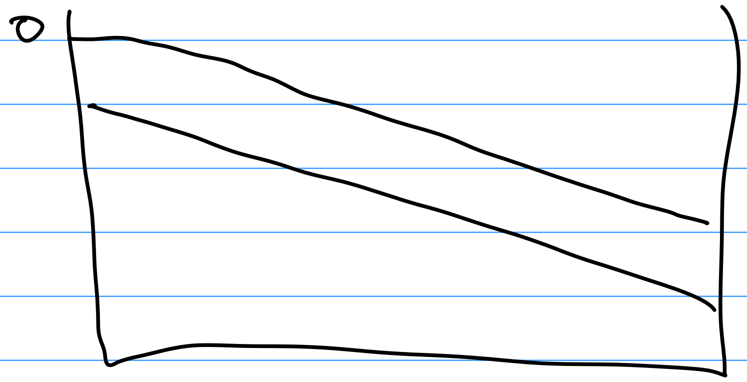
```
while $\bar{V}(b_0) - \underline{V}(b_0) > \varepsilon$
    explore$(b_0, 0)$
```



```
explore$(b,t)$
    if $\bar{V}(b) - \underline{V}(b) > \varepsilon\gamma^t$
        $a^* = \arg\max_a \bar{Q}(b,a)$   ← upper bound
        $o^* = \arg\max_o \left(P(o|b,a)\left(\overbrace{\bar{V}(\tau(b,a,o)) - \underline{V}(\tau(b,a,o)) - \varepsilon\gamma^{(t+1)}}^{\text{excess uncertainty}}\right)\right)$
        explore$(\tau(b,a^*,o^*), t+1)$
        $\underline{\Gamma} = \underline{\Gamma} \cup$ Backup Belief$(\underline{\Gamma}, b)$
        $\bar{V}(b) = B_b[\bar{V}(b)]$
```

Baby POMDP

f ¬f

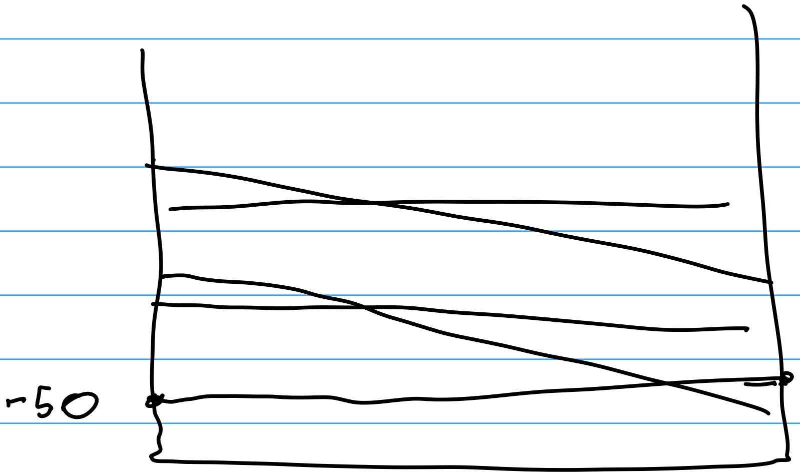


$\frac{R}{1-\gamma}$

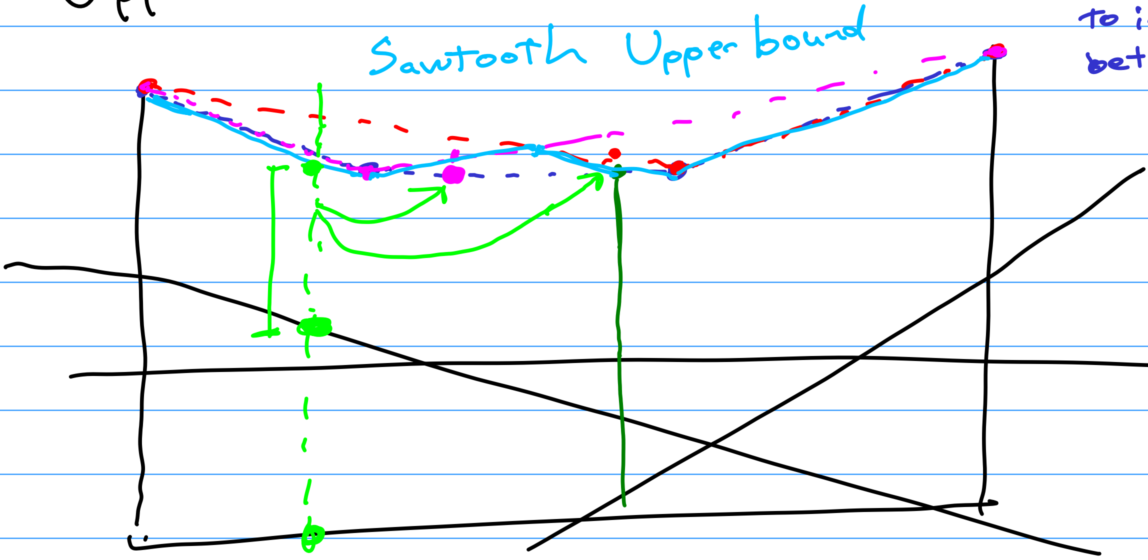
Open-Loop Policy

Never Feed

$\frac{-5}{1-0.9} = -50$



Upper Bounds



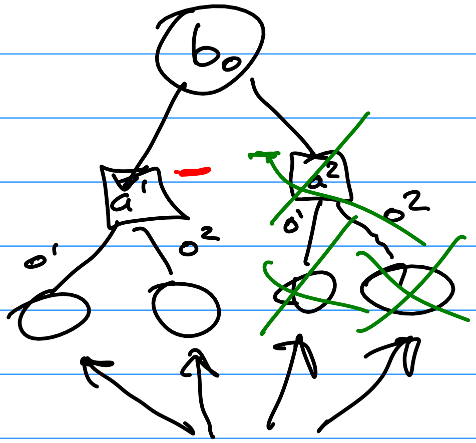
Finding points to interpolate between is an LP

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

SARSOP   Successive Approximation of Reachable Space under Optimal Policies

Similar to HSVI

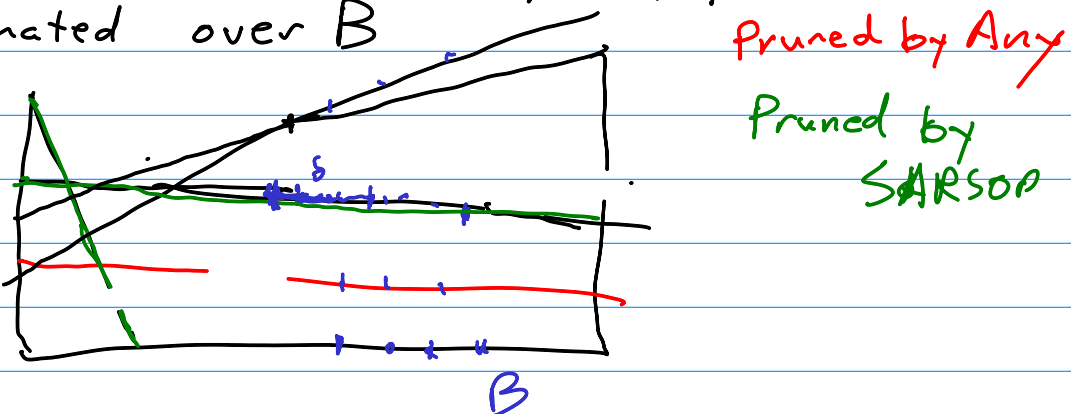
$B \subset R$ $B \subset R^*$



Belief points that could be in B

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

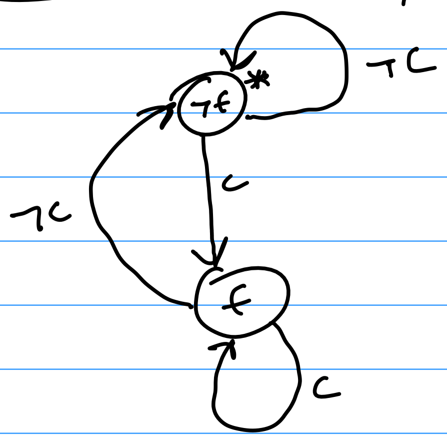
Instead of pruning α that are dominated over the whole belief space, prune α dominated over B



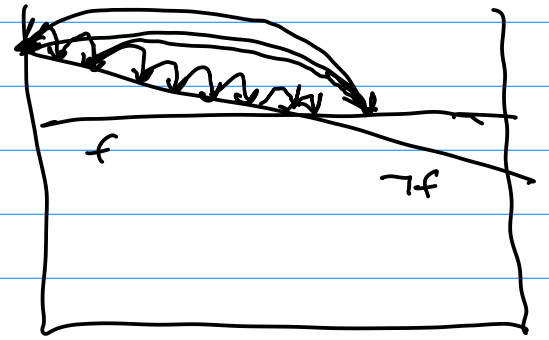
Policy Graph

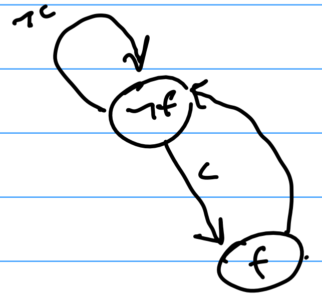
vertex labeled with a
edge labeled with o

Feed when Crying



Policy Graph → α vector
evaluate plan
α vectors → policy graphs

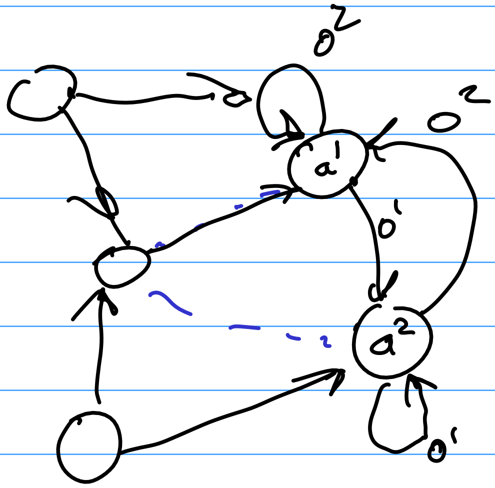




MCVI "Monte Carlo Value Iteration"

$$V_G(b) = \max \int_{s \in S} \alpha_v(s)\, b(s)\, ds$$ ← Approx w/MC

```
MC-Backup$(G, b, N)$
  $R_a = 0$   $V_{a,o,v} = 0$
  for $a \in A$
    for i in 1:N
      $s_i \Leftarrow$ sample$(b)$
      $s_i', o_i, r_i \Leftarrow G(s_i, a)$
      $R_a + r_i$
      for $v \in G$
        $V_{a,o_i,v} = V_{a,o_i,v} + \text{Simulate}(G, v, s', L)$
    for o in O
      $V_{a,o} = \max_{v \in G} V_{a,o,v}$
      $v_{a,o} = \text{argmax}_{v \in G} V_{a,o,v}$
    $V_a = R_a + \gamma \sum_o V_{ao}/N$
  $V^* = \max_a V_a$
  $a^* = \text{argmax } V_a$
  add new node to G labeled with $a^*$
```



V = value

v = node in graph