Last time: How to perform VI on POMDPs (BellmanBackup)

Today: Most efficient POMDP solutions

<u>Offline</u>

	S	A	O
Exact VI		D	
PBVI			
Perseus			
HSVI			
SARSOP			
MCVI	C	D	D/C
LQG	C	C	C

Survey of PB POMDP solvers (PBVI, Perseus, HSVI)

Bigger

Gaussian/Quad

Exact VI

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

$$\Gamma^a = \bigoplus_{o \in O} \Gamma^{a,o} \qquad \Gamma_1 \oplus \Gamma_2 = \{\alpha_1 + \alpha_2 : \alpha_1 \in \Gamma_1, \alpha_2 \in \Gamma_2\}$$

$r_a[s] = R(s,a)$

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

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

no max

For 1 iteration

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

PBVI   Point Based VI

"BellmanBackup(Γ, b)"

BackupBelief(Γ, b)

for $a \in A$

for $o \in O$

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

$\alpha_{a,o} = \arg\max_{\alpha \in \Gamma} b'^T \alpha$

for $s \in S$

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

return $\arg\max_{\alpha_a} \alpha_a^T b$

Generic

PBVI: use $B.B.(\Gamma, b)$ for b in $\{\boxed{B}\}$

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

O.G. PBVI

$B = \{b_0\}$ $(S, A, T, R, O, Z, \gamma, b_0)$

loop

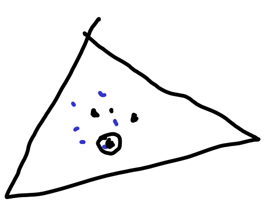
for $b \in B$

$\Gamma = \Gamma \cup BB(\Gamma, b)$

$B' = \emptyset$

for $b \in B$

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



$B' = B' \cup \arg\max_{b' \in \tilde{B}} \|B, b'\|_L$ ← furthest b' away from B

$B = B \cup B'$

Perseus

$B = \emptyset$

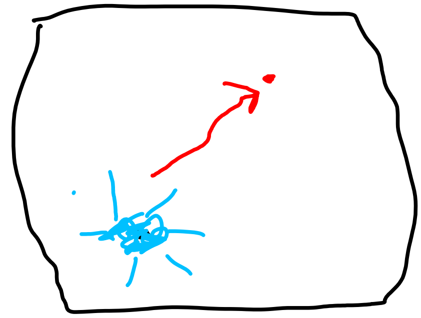
$b = b_0$

loop until $|B| = n$

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

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

$B = B \cup \{\tau(b,a,o)\}$



HSVI Heuristic Search VI

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

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

HSVI

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

explore$(b_0, 0)$

explore(b, t)

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

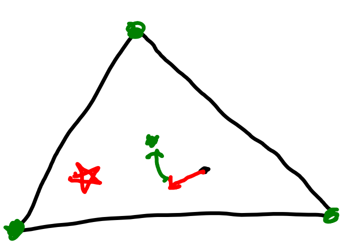
$a^* = \arg\max_{a \in A} Q_{\overline{V}}(b,a)$

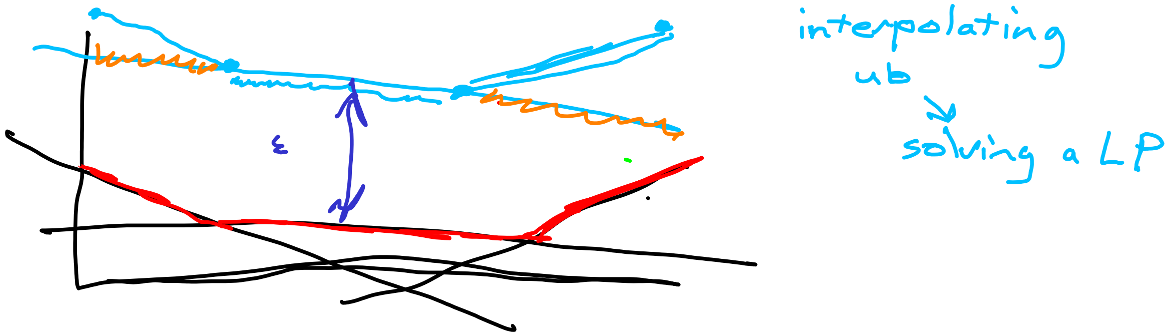
$$o^* = \arg\max_{o \in O} \left(P(o \mid b,a) \left(\underbrace{\overline{V}(\tau(b,a^*,o^*)) - \underline{V}(\tau(b,a^*,o^*)) - \varepsilon\gamma^{-(t+1)}}_{\text{"excess uncertainty"}} \right) \right)$$

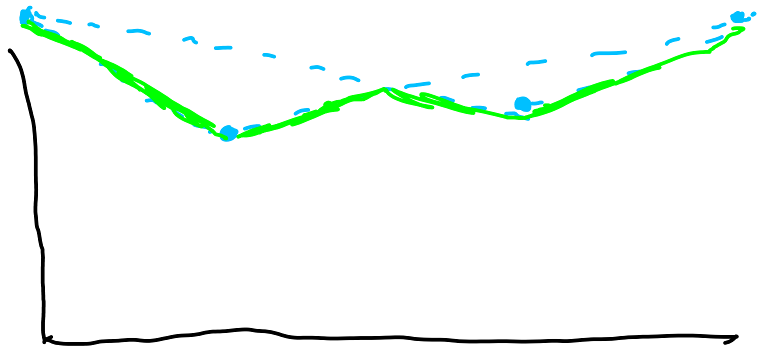
explore$(\tau(b,a^*,o^*), t+1)$

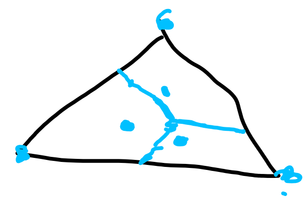
up l.b. $\longrightarrow$ $\underline{\Gamma} = \underline{\Gamma} \cup \underline{B.B.(\Gamma, b)}$

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









$$\tilde{V}(b) = \arg\max_a R(s,a) + \gamma \sum_o P(o|b,a)\tilde{V}(\tau(b,a,o))$$

SARSOP Successive Approx. of Reachable Space under Optimal Policies

---

S is continuous : no α vectors
α-functions

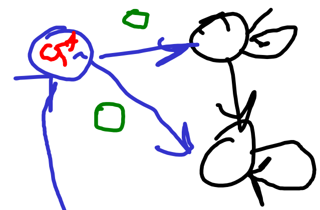
MCVI - Monte Carlo VI

Improve policy graph directly

$$V_G(b) = \max_{v \in G} \int_{s \in S} \alpha_v(s)\, b(s)\, ds$$

$$HV(b) = \max_{a \in A}\left[R(b,a) + \gamma \sum_o P(o|b,a)\, V(\tau(b,a,o))\right]$$

MC-Backup(G, b, N)



$R_a = 0$

$V_{a,o,v} = 0$

for $a \in A$

  for $i \in 1..N$

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

    $s'_i, o_i, r_i = G(s_i, a, w)$

    $R_a \mathrel{+}= r_i$

    for $v \in G$

      $V_{a,o,v} = V_{a,o_i,v} + \text{simulation}(G, v, s'_i)$

  for $o \in O$

    $V_{a,o} = \max_v V_{a,o,v}$

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

  $V_a = \left(R_a + \gamma \sum_o V_{a,o}\right) / N$

$V^* = \max_a V_a$

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

add new vertex u, to G label with a^*

for $o \in O$

  add edge $(u, v_{a^*,o})$, label with o

LQG - Linear Quadratic Gaussian

$$s' \sim N(As + Ba, V)$$

$$o \sim N(Cs + Da, W)$$

$$s_0 \sim N(\mu_0, \Sigma_0)$$

$$R(s,a) = s^T Q s + a^T R a$$

$$b_t = N(\mu_t, \Sigma_t)$$

can prove that optimal policy is simply

$$a_t^* = -K_{LQR}\, \mu_t$$

solution to MDP

"<u>separation</u> principle"