Last Time

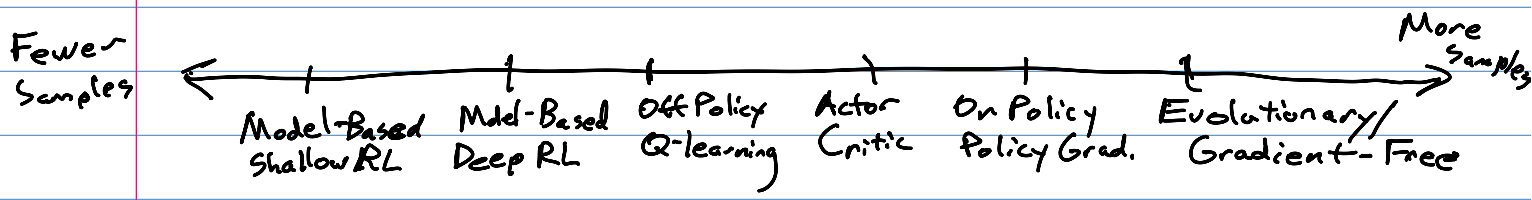
Actor-Critc
Exploration: RND

This Time

POMDP
Bayesian Filters

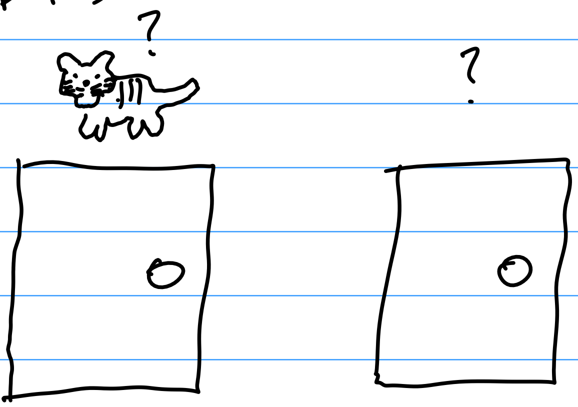
How to choose an RL algorithm (According Sergey Levine)

Sample Efficiency   Ease of Use/Stability



With fast simulator, wall clock time is roughly reversed

POMDPs



S, A, T, R, O, Z

$S = \{TL, TR\}$

$T(s'|s,a) = \begin{cases} 1 & \text{if } s'=s \\ 0 & \text{o.w.} \end{cases}$

$A = \{OL, OR, L\}$

Reward: +10 if open empty door, -100 if open tiger door

$\gamma = 0.99$

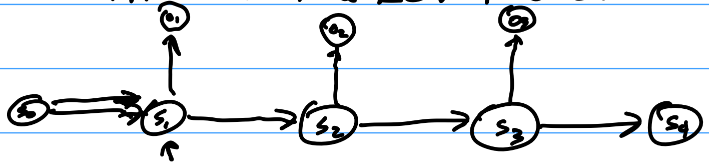
$O = \{TL, TR\}$

$P(o|s) = \begin{cases} 0.85 & \text{if } o=s \\ 0.15 & \text{o.w.} \end{cases} = Z(o|s)$

$$\hat{\pi}(TL) = OR \qquad V^{\hat{\pi}}(b_0) = 0.85 \cdot 10 + 0.15(-100)$$
$$\hat{\pi}(TR) = OL \qquad = -6.5$$

Belief Updating

Hidden Markov Model HMM



$$P(s_1|o_1) = \frac{P(o_1|s_1)\,P(s_1)}{P(o_1)} = \frac{P(o_1|s_1)\sum_{s\in S} P(s_1|s_0=s)P(s_0=s)}{P(o_1)}$$

$$\propto P(o_1|s_1)\sum_{s\in S} P(s_1|s_0=s)P(s_0=s)$$

$$b_t(s) = P(s_t = s \mid h_t)$$
$$h_t = (o_1, \dots, o_t)$$

$b_0(s)$

$$P(s_k|h_k) = \frac{P(o_k|s_k,h_{k-1})\,P(s_k,h_{k-1})}{P(h_k)}$$

$$\propto P(o_k|s_k)\sum_{s_{k-1}} P(s_k|s_{k-1},h_{k-1})\,P(s_{k-1}|h_{k-1})\,\cancel{P(h_{k-1})}$$

$$\propto P(o_k|s_k)\sum_{s_{k-1}} P(s_k|s_{k-1}) \quad \underbrace{P(s_{k-1}|h_{k-1})}_{b_{k-1}}$$

$$b'(s) \propto Z(o|s')\sum T(s'|s)\, b(s)$$

Belief Update

```
b_0
loop
  receive o
  for $s' \in S$
    $b'(s') \leftarrow Z(o|s') \sum_s T(s'|s)\, b(s)$
  $b' = b' / \sum_{s'} b'(s')$
  $b \leftarrow b'$
```

$O(|S|^2)$

$b_0(TL) = 0.5$

$o_1 = TL$

<belief update>

$b_1(TL) = 0.85$

$o_2 = TL$

$b_2(TL) = 0.97$

$o_3 = TR$

$b_3(TL) = 0.85$

$o_1 = TR$

$b_1(TL) = 0.15$