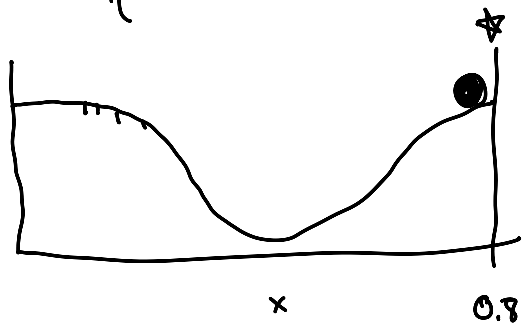
1. (S, A, T, R) MDPs

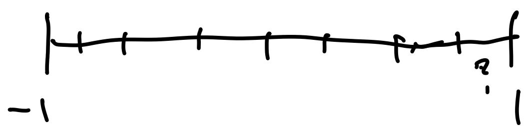
2. reset! RL
   step!
   actions

$s', r, \text{done}, \text{info} = \text{step!}(s, a)$



1. Exploration + Exploitation ←
2. Credit Assignment ←
3. Generalization ←

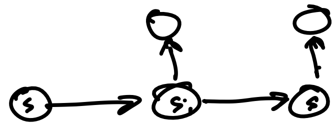
$s' = [0.76, 0.01]$



Review   POMDP – MDP, but only observe o

make decisions based on $h_t = (o_1 \dots o_t)$

HMM



- state dynamics T
- observation model Z

$$b_t(s) = p(s_t = s \mid h_t)$$

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

"Update Belief"

- Discrete Bayesian
- Particle Filters: Discrete, Continuous
  - $O(n)$

Today:

1. More Efficient belief updates on high-dim continuous spaces
2. How is the POMDP Value function related to beliefs?
3. What do POMDP Policies Look Like

Exact if Linear, Gaussian Noise Dynamics + Observations

$$s' \sim \mathcal{N}(As + b, V) \qquad o \sim (Cs' + d, W) \qquad s_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$$

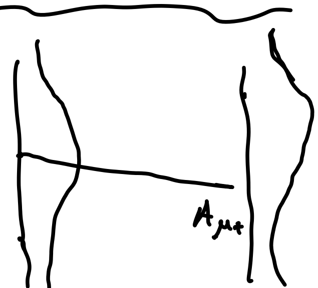
Kalman Filter   5044

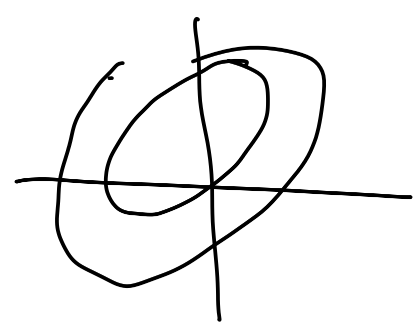
$$b_t = \mathcal{N}(\mu_t, \Sigma_t)$$

$$\Sigma_{t+1} = As\left(\Sigma_t - \Sigma_t C^T (C\Sigma_t C^T + W)^{-1} C\Sigma_t\right) A^T + V$$

$$K = A\Sigma_t C^T (C\Sigma_t C^T + W)^{-1}$$

$$\mu_{t+1} = A\mu_t + K(o - C\mu_t)$$





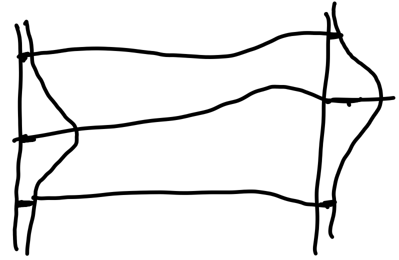
Works for nonliner? Yes, as long as unimodal

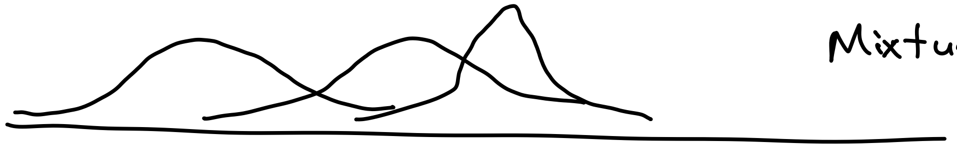
EKF   Linearize Dynamics

$$s' = f(s, w) \quad \leftarrow A$$

$$s' \approx f(\hat{s}) + \frac{\partial f}{\partial s}\Big|_{\hat{s}} (s - \hat{s})$$

UKF





Mixture of Gaussians

2. POMDP

Value Functions? Policies?

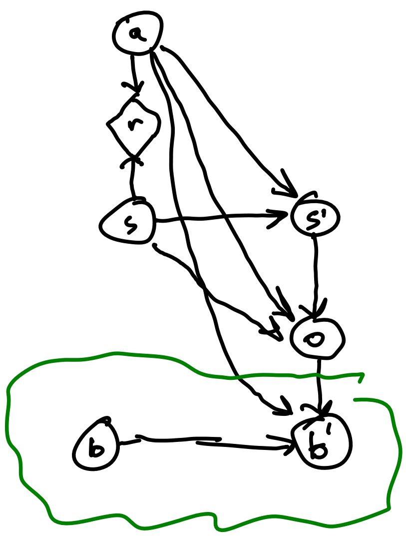
MDP (S, A, T, R, γ) (p_0)

POMDP $(S, A, O, T, R, Z, \gamma)$

O: obs space, Z: obs dist

$Z(o \mid a, s')$ or

$Z(o \mid s, a, s')$



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

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

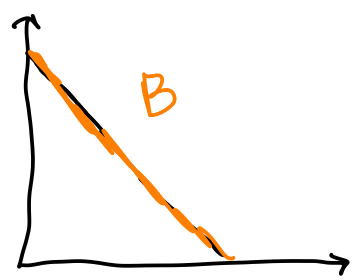
$b : S \rightarrow [0,1]$

$b \in [0,1]^{|S|}$

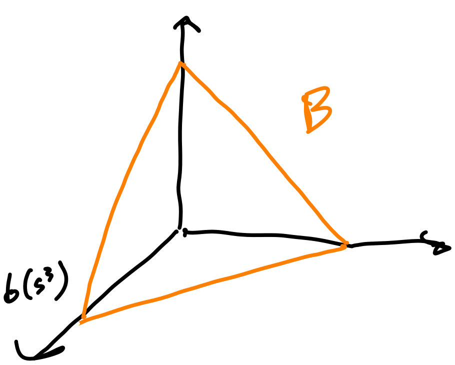
$\pi : B \rightarrow A$

$\pi : H \rightarrow A$

$|S| = 2$



$|S| = 3$



$b(s^3) = 1 - b(s^1) - b(s^2)$

$b(s^1) + b(s^2) + b(s^3) = 1$

Tiger POMDP

$S = \{TL, TR\}$

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

$O = \{TL, TR\}$

T : static until open, reset

Z : 85%

R : +10 good door −100 open tiger

Crying Baby

$S = \{h, \neg h\}$

$T(s' = \neg h \mid s, f) = 1.0$

$A = \{f, \neg f\}$

$O = \{c, \neg c\}$

T: $f \rightarrow \neg h$ $h, \neg f \rightarrow h$

$\neg h, \neg f \rightarrow h$ 10%

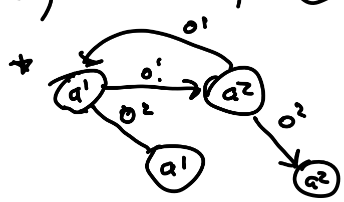
Z: $h \rightarrow c$ 80%

$\neg h \rightarrow c$ 10%

R : f: −5 h: −10

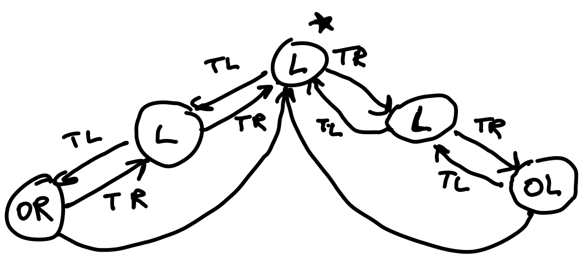
Policies

a) Policy Graph

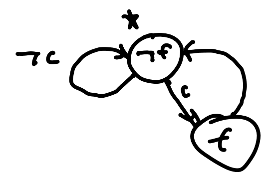


start at *
loop
  take a at current node
  observe o
  traverse graph on edge o

Tiger

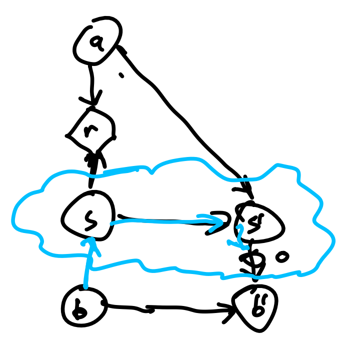


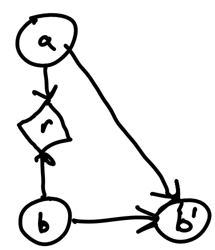
Feed when crying



b) Alpha Vectors

Important : A POMDP is an MDP on the belief space





Even if S, A, O are discrete

B is continuous

b^1 close b^2 $\longrightarrow$ $Q(b^1, a)$ close $Q(b^2, a)$

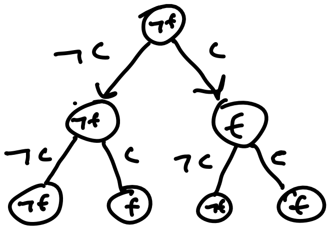
α vector |S| - dimensional vector

$\alpha \in \mathbb{R}^{|S|}$ each entry $Q^P(\delta_s, P_0)$

$\alpha^P[s^2] = Q^P(\delta_{s^2}, P_0)$

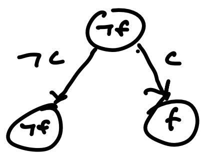
Conditional plan: history based policy with fixed steps

2-step c.p. for crying baby



$$V(b) = \max_{\alpha \in \Gamma} b^T \alpha \quad \leftarrow \sum_{s \in S} b(s)\alpha[s]$$

1 step c.p. $\tilde{p}$



if $s = h$

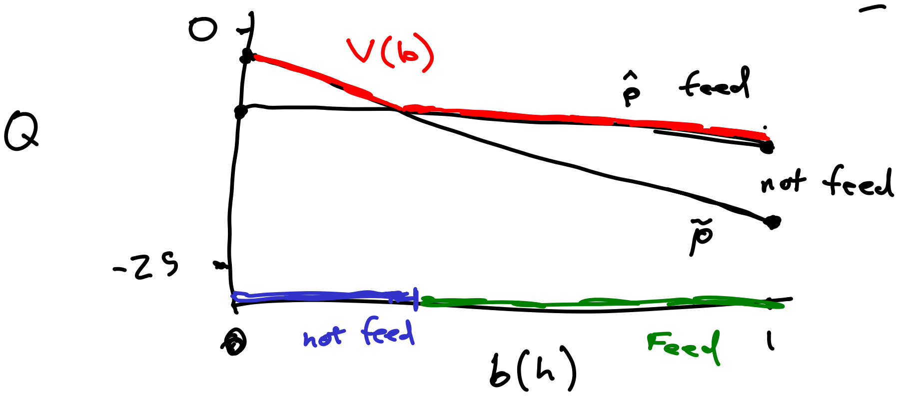
$$Q^{\tilde{p}}(h, \neg f) = -10 + \gamma\left(0.8 \times -15 + 0.2 \cdot -10\right)$$

$$= -22.6$$

if $s = \neg h$

$$Q^{\tilde{p}}(\neg h, \neg f) = 0 + \gamma \quad \begin{array}{cc} h & c \\ \neg h & \neg c \end{array}$$

$$= -1$$



1step

Thursday:

Value Iteration