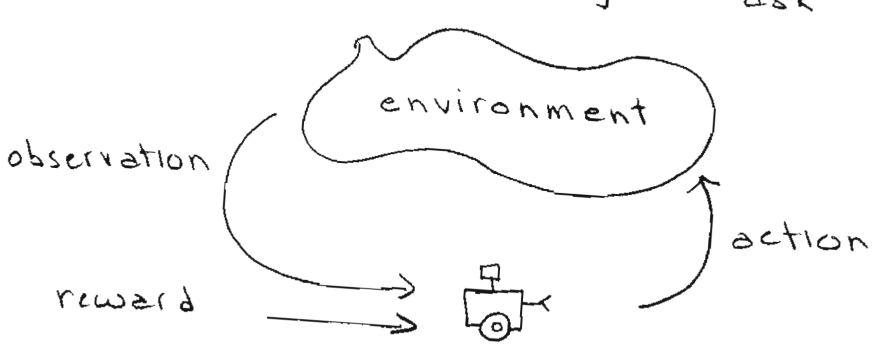
Partially Observable Markov Decision Processes

- Who has already heard of prior to this class?
- If notation is unclear just ask

hand answers to ?s up at start point

?

ideas with domain or idea



POMDP tuple

$< S, A, E, p(s'|s,a), p(e|s',a), r(s,a), \gamma, b_0 >$

assume discrete unless say otherwise:
- S = set of states
- A = set " actions
- E = " " observations (also often called O or Z)

$p(s'|s,a)$ transition model (like MDP)

$p(e|s',a)$ observation model

$r(s,a)$ reward

γ discount factor

in general won't get direct access to states
get observations that provide some information about underlying state
still (like MDP) want to max expected Σ of future rewards
want to compute a policy to achieve this
can think of policy as a function of either of 2 things (& turns out equiv)

π : history $\rightarrow a$

history $= < a^1, z^1, a^2, z^2, \ldots a^t, z^t >$

π : $b \rightarrow a$

b is a belief state
vector of length $|S|$
$b(s)$ represents probability of being in each state s — inf # bs
$\Sigma_s b(s) = 1$ so belief lies in $(|S|-1)$-dim simplex

typically start with an initial b_0
represents prior knowledge
could be uniform (card game, hidden cards from shuffled deck)
could be highly specific (prior prob of topics people call airlines for)
could even be delta function
(robot navig, may know start but perceptual ambiguity
student learning?)

problem w/history is gets longer w/each step
belief state, when updated using Bayes filter, is a
sufficient statistic for history

let $b_{t-1}(s) = p(s \mid h_{t-1})$

$$h_t = [h_{t-1}, a_t, e_t]$$

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

$$= p(s' \mid h_{t-1}, a_t, e_t)$$

$$= \frac{p(s', e_t \mid h_{t-1}, a_t)}{p(e_t \mid h_{t-1}, a_t)} \quad \text{Bayes rule}$$

$$= \frac{\sum_s p(s', e_t, s \mid h_{t-1}, a_t)}{p(e_t \mid h_{t-1}, a_t)} \qquad \sum_a p(a,b) = p(b)$$

$$= \sum_s \underbrace{p(s' \mid s, h_{t-1}, a_t)}_{\substack{\text{Markov} \\ \text{property} \\ = p(s' \mid s, a_t)}} \underbrace{p(e_t \mid s', s, h_{t-1}, a_t)}_{\substack{\text{observ model} \\ \text{depends only on} \\ s', a_t \to p(e_t \mid s', a_t)}} \underbrace{p(s \mid h_{t-1}, a_t)}_{\substack{\text{prior state} \\ \text{can't depend on } a_t \\ = p(s \mid h_{t-1}) = b_{t-1}(s)}} \Big/ p(e_t \mid h_{t-1}, a_t)$$

$$= \frac{p(e_t \mid s', a_t) \sum_s p(s' \mid s, a_t)\, b_{t-1}(s)}{p(e_t \mid h_{t-1}, a_t)}$$

$p(e_t \mid h_{t-1}, a_t) \leftarrow$ = to numerator, summed over s'

so to compute belief state after history h_t sufficient
to calculate using transition & observation models &
prior belief $\longrightarrow$ b suffic stat for history

Anstrom 1965 showed b also sufficient for optimal control

So objective is to compute $\pi : b \to a$ to max expected Σ future rew

- Policies

if only have 1 time step to act, only $|A|$ different policies,
one for each action

What is the value of each of these policies?

depends what state in

$V_p(s) = r(s, a(p))$ reward for state s & action of policy p

define the α-vector associated w/policy p as a
$|S|$ length vector where each index i specifies the
value of taking that policy in state s_i

$$\alpha_p = \langle V_p(s_1), V_p(s_2) \ldots V_p(s_{|S|}) \rangle$$

the value of a belief b under a policy is the expected
value over the distribution over states specified by b

$$V_p(b) = \sum_s b(s)\, \alpha_p(s)$$

$$= b \cdot \alpha_p \quad \text{dot product}$$

implies expected value of a particular fixed policy
varies linearly with b (hyperplane in belief space)

the value function over all beliefs b is the supremum
over the set of α-vectors

at each b will choose the policy of the α_p which
maximizes $b \cdot \alpha_p$

implies value function is PWLC

each α-vector is hyperplane → convex

supremum of a set of convex functions is convex

* note that α-vectors / fixed policies effectively partition the belief space into regions that share the same best policy (cool, important ∞ b — finite conditional policies

What about acting for 2 steps?

consider potential conditional policy trees

choose action $\xrightarrow{e_1} \alpha^{11}$, $\rightarrow \alpha^{11}$, $\xrightarrow{e_{|E|}} \alpha^{13}$ — choose a 1-step policy to follow out of set of 1 step policies

for each possible observation

? how many CPTs does this create?

$t=1$ $|A|$

$t=2$ $|A|\,|A|^{|E|}$

computing the value of a 2 step CPT p

$$\alpha_p(s) = r(s, a(p)) + \gamma \sum_{s'} \sum_e \underbrace{p(s'|s, a(p)}_{\text{all states could go to}} \underbrace{p(e|s', a(p))}_{\text{prob seeing observe } e} \underbrace{\alpha_{pe}(s')}_{\text{value of } s' \text{ in 1 step policy tree at } e \text{ branch of tree}}$$

can use this equation for any T-step CPT
(building up from $T=1$ to T)

at each time step value function is the supremum over all
$\alpha \in \Gamma_t$

example (simplified)

doctor diagnosing whether a patient has cancer

state: has cancer or doesn't have cancer

actions: run a test (colonoscopy etc.) diagnose cancer
diagnose no cancer

observations: null, positive test, neg test

transition model: if run a test, assume state says the same
if diagnose, assume patient leaves & a new patient comes in whose probability of cancer is b_0(cancer)

observation model

$p(\text{pos} | \text{test}, s = \text{cancer}) = 0.8$

$p(\text{neg} | \text{test}, s = \text{no cancer}) = 0.9$

$p(\text{null} | \text{either diagnosis}) = 1.0$

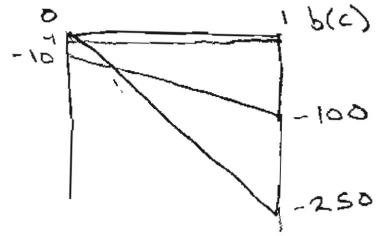
reward

$r(\text{test}, s=\text{cancer/no cancer}) = -1$
$r(\text{diag cancer}, s=\text{cancer}) = -100$
$r(\text{ " \quad " }, s=\text{no cancer}) = -10$
$r(\text{diag no cancer}, s=\text{cancer}) = -250$
$r(\text{ " }, s=\text{no cancer}) = 0$

$b_0 = [.9 \ .1]$

$\gamma = .99$

1 time step to act



$\alpha_{test} = [-1 \ \ -1]$
$\alpha_{diagNC} = [0 \ \ -250]$
$\alpha_{diag\,C} = [-10 \ \ -100]$

note that α_{diagNC} is dominated: no beliefs at which it is best policy

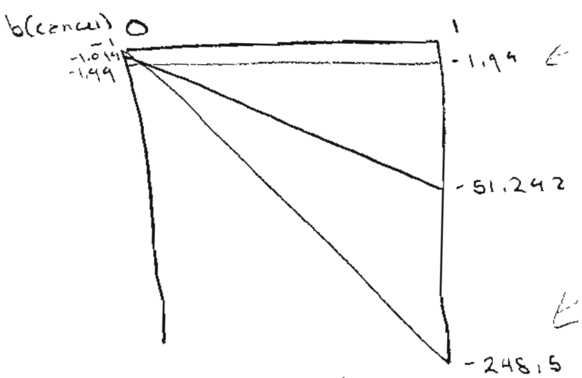
2 time steps
how many CPTs?
potentially 81: $3 \cdot 3^3$
some observations are impossible for some actions
$|A|^2 + 2|A| = 15$

test
neg → no cancer diag
pos → test

$V_p(\text{no cancer}) = -1 + \gamma\,(p(\text{pos}|\text{no cancer})\,\alpha_{test}(\text{no cancer}) + p(\text{neg}|\text{no cancer})\,\alpha_{diag\,no\,C}(\text{no cancer}))$
$= -1 + \gamma\,(.1 \cdot -1 + .9 \cdot 0)$
$= -1 + \gamma(-.1) = -1.099$

$V_p(\text{cancer}) = -1 + \gamma\,(p(\text{pos}|\text{cancer})\,\alpha_{test}(\text{cancer}) + p(\text{neg}|\text{cancer})\,\alpha_{diag\,no}(\text{cancer}))$
$= -1 + \gamma\,(.8 \cdot -1 + .2 \cdot -250)$
$= -51.292$

$\alpha_p = [-1.099 \ \ -51.292]$



not to scale!

test
neg → test
pos → test

test
neg → diag no cancer
pos → diag no cancer

rest of conditional policy trees are dominated (exists other α-vector with better value for all b)

- Value function backups
  can think of computing value function as enumerating all t+1 step conditional policy trees from prior set of t-step policy trees
  can also think of computing best (t+1)-step policy tree for a particular belief b (& repeating this for all b)

let Γ be the set of t-step α-vectors that make up the t-step value function (PWLC)

back up for belief b is

$$V(b) = \max_a \Big[\textstyle\sum_s r(s,a) b(s) + \gamma \sum_e \underbrace{\max_{\alpha \in \Gamma}}_{} \sum_s \sum_{s'} p(s'|s,a) p(e|s',a) \alpha(s') b(s) \Big]$$

choose the t-step policy to follow that max expected future reward given started in b(s), took a, & saw e

to see in a different way, multiply & divide inner term by $p(e|b,a)$

$$V(b) = \max_a \Big[\textstyle\sum_s r(s,a) b(s) + \gamma \sum_e \max_{\alpha \in \Gamma} p(e|b,a) \sum_{s'} \alpha(s') \underbrace{\dfrac{p(e|s',a) \sum_s p(s'|s,a) b(s)}{p(e|b,a)}}_{= b^{ae}(s')} \Big]$$

just select t-step policy tree at each e branch that maximizes expected value of $b^{ae}(s')$

note that can view this backup as constructing the best (t+1)-step policy for belief b

$$\alpha_b^a = r(s,a) + \gamma \textstyle\sum_e p(e|b,a) \operatorname{argmax}_{\alpha \in \Gamma} (\alpha \cdot b^{ae})$$

$$\alpha' = \operatorname{argmax}_{\alpha_b^a} \alpha_b^a \cdot b$$ (still linear, value func supremum over linear, convex)

if do this for all possible beliefs, not really efficient

only a finite # of m-step conditional policy trees for any m $< \infty$

infinite # of beliefs ((|S|-1)-dim simplex)

So might as well enumerate conditional policy trees & then prune those which are dominated

again, recall αs partition belief space

- can use a linear program

but later see that if not computing a policy over all beliefs, could be helpful

||zm

? is MDP solution an upper or lower bound to POMDP V?

upper bound! aka $V(b) = \sum_s b(s) V^{MDP}(s)$

intuitively: full state knowledge can never hurt you

mathematically

$$V(b) = \max_a \Big[\textstyle\sum r(s,a) b(s) + \gamma \sum_e \max_\alpha \sum_s \sum_{s'} p(s'|s,a) p(e|s',a) \alpha(s') b(s) \Big]$$

$$\le \max_a \Big[\textstyle\sum r(s,a) b(s) + \gamma \sum_e \sum_s \sum_{s'} \max_{\alpha(s')} p(s'|s,a) p(e|s',a) \alpha(s') b(s) \Big]$$

since assume know s (will get to see s')

$$= \max_a \Big[\textstyle\sum r(s,a) b(s) + \gamma \sum_s b(s) \sum_{s'} \max_{\alpha(s')} p(s'|s,a) \alpha(s') \Big]$$

$$= \max_a \textstyle\sum b(s) V^{MDP}(s)$$

? What does PWLC nature of value function imply about uncertainty
often worse to be uncertain
never better to be uncertain vs max of rewards if know s

? Why can't we delude ourselves (just believe we won the lottery)
because belief state represents accurate probability that in each $s \in S$ possible states
using real observation & transition models for state estimation

? In cancer example actions were fairly clearly separated into information gathering (diagnostic test) vs reward max (decide cancer/not). What is an example where actions might more along the continuum? (or a domain)
ex: robot navigation/coastal navigation/Minerva

practicalities
intractable to enumerate all CPTs
\# grows as $|A|^{O(|E|^{d-1})}$

1 step $|A|$
2 steps $|A|\,|A|^{|E|}$
3 steps $|A|(|A|^{|E+1|})^{|E|}$
$= |A|\,|A|^{|E|^2}$

expensive to prune
many of these may be irrelevant
given initial b_0 only a subset of beliefs may be reachable

in work in 1990s (Kaelbling, Littman, Cassandra, Hauskrecht)
most problems very small 4-12 states
this might be sufficient for some problems
breast cancer & colon cancer screening have both been formulated as small POMDPs
most work within AI community has focussed on approximate solutions
exs

? is t-step value function upper or lower bound for (t+1) step value func?
answer: depends!
if r all positive, lower bound
" " negative, upper bound
if interested in ∞ horizon policy can start by initializing starting α-vectors as a lower bound (ex. $r_{min}/1-\gamma$)

maybe do motiv exs before break

55 min