Recall
1. deterministic environment
   - complete information
     - suffices to give $\alpha$ sequence of
2. Markov Decision Process 'planning'
   - Stochastic environment
     - Complete information
     - Must react to environment’s random choices using a 'policy'

Partially observed Markov Decision Porcesses (POMDD)
Stochastic environment
Partial information (very general - we assume that the states are factored into attributes and partial information means some of these attributes get hidden)
Still must react to enviroment’s random choices; must use 'reasoning' to cope with missing information

Define A POMDD:
is an interactive process given by sets of states $X$, actions $A$, and observations $O$ along with distributions
- $D(x, \alpha)$ over next states (from X)
- $P(x, \alpha)$ over observation from $\theta$
- $X=0,1^n$, $O$ is a masking process on $X \times A$ that does not hide the action:
  See figure 1
  As before we’ll consider goals of the form, reach $x_T$ s.t $G(x_T)=1$ for some predicate $G$ that is given as input:
  Again we will seek to design policies that cope with ”faults” or random choices of the environment using observations
  We will continue to look at rank-k decision tree policies with the convention that branches are labeled with a literal if it is satisfied we take one branch is failfilled or unknown, we take the other branch
  All we hope to achieve with our policies is that $G$ is $(1-\epsilon)$ - valid when $\pi$ terminates

As we have the problem: Robot find the tube in n rooms:
Informally goal is put the box in room 2 don’t be in room 2 i.e.
only reveals the attributes for the room i s.t: $\partial t_{b,2} \land \neg \partial t_{r,2}$
suppose $O$
$\partial t_{r,i}=1$
In the final steps of our plan:
\[ t=5 \quad 0 \quad 0 \quad 1 \quad \ast \quad \ast \quad 1 \quad 0 \quad 0 \quad \rightarrow G(X_5)=0 \]
t=6 1 0 0 0 0 * * 0 0 \rightarrow G(X_6) = 0

Need to use a frame axion to infer \( \partial t_b(x_b) \)

Notice to use a noisy version of the environment in which, say, the frame axioms might fail up \( \gamma \), we can only conclude \( Pr[G(x_L)] \geq 1 - \gamma \)

Now we wish to learn the environment representation (why? knowledge engineering is tedious error-prove you may not think to include rules that seems "obvious", etc)

**Pitfalls in learning:**

unknown domains

obvious 1st attempt given access to a POMDD capturing a noisy STRIPS domina. find STRIPs rules capturing the domain

Difficulty 1 "Exploration"

Consider the following combination lock" domain we have attributes x1,...xn, C1, ..., Cn and open

There are two actions '0' and '1' Exactly one Ci is 1 at any time.

After taking an action when Ci=1, we enter a state where Ci+1 is 1, Initially Ci=1

The action 'b' when Ci=1 has the effect of setting xi=b

Finally when Cn=1, if (x1 - xn) = (S1-Sn) there is a conditional effect that sets open =0

Let’s suppose effect that we choose \( \{S1,...Sn\} \in \{0,1\}^n \) uniformly at random.

Consider the variant of the environment that does not have the effect that sets open=1

Fix a sequence of actions. We only see open=1 if the sequence of actions was S1, ..., Sn. Since S1, ..., Sn was chosen at random, find \( Pr[S1, ..., Sn= \alpha_1, ..., \alpha_n] = 1/2^n \)

**Over T steps/trials:**

We can take a union bound

\( Pr[open=1 \text{ in any of the T trials}] \leq T/2^n \)

So to get prob \( \geq 1/2 \) of seeing open=1

we need \( 2^n/2\text{trials} \)

Corollary distinguishing \( S \neq S' \) also requires \( \Omega(2^n) \) of trials

Fix, only aim to generate plans w.r.t portions of the environment that have been "explored well" in the training examples / past experience/

**Difficulty 2: Explicit representations**

In general: a Noisy STRIPS environment poses an agnostic learning problem conditional effects. For example, are at least as hard to learn as conjunctions. As we discussed agnostic learning is mostly out of reach for the state-of-the-art.

Fix Don’t try to generate a complete explicit environment representation. We’ll only seek to learn the environment representation. we’ll only seek to learn the environment implicity and we’ll construct a policy using queries against this implicit KB

**Learning from example histories**

We assume that there is some (arbitrary) reference "exploration policy" \( \Pi_0 \) that was used to generate histories for car training data

For example,

- \( \Pi_0 \) might just pick random actions
- \( \Pi_0 \) is a uniformly random past policy
- \( \Pi_0 \) could be be a more intelligent exploration policy(tree to reach new station)
We will assume we're given access to example histories of $\Pi_0$ acting in our environment for $T$ steps. Motivated by HW4 q3, we'll further assume that each step includes the probability with which $\Pi_0$ chose its action.

Def: We say that a sequence of action $\alpha_1,...,\alpha_t$ is $M$-typical for $\mu \in [0,1]$ for the reference policy $\Pi_0$ and the POMDD if the probability that $\Pi_0$ chose each action is $\mu$.

Note in order to guarantee we see a $\mu$-typical sequence of actions we need $1/\mu$ examples.

Our task: given that there is a policy (rank-k decision tree) that only uses typical sequences of actions on any branch and such that under the conditional distribution reached on each branch, some KB that is $(1-\sigma)$-testable allow us to prove $G$ is satisfied on the branch.

Find such a policy using access to histories generated by $\Pi_0$.

Surprising thing we'll get a poly-time algorithm in terms of $1/\mu, 1/\epsilon, 1/\theta, n, T$ and exponential in $k,s$ (space for a resolution proof)
Figure 1: