MAKING COMPLEX DECISIONS

Outline
- Sequential Decision Problems
- Value Iteration
- Policy Iteration
- Decision-Theoretic Agent Design
- Dynamic Belief/Decision Networks

Based on the textbook by S. Russell & P. Norvig:
Artificial Intelligence, A Modern Approach, Chapter 17

Transition Model
- In a deterministic setting the outcomes of actions are known, and the agent may plan a sequence of actions which moves it to (4, 3).
- This becomes impossible if actions are nondeterministic/unreliable.
- A transition model assigns a probability \( M_{ij} \) to the event that the agent reaches state \( j \) if it performs action \( a \) in state \( i \).

Example. Each one of the four actions North, South, East, and West moves the agent
1. to the intended direction \( d \) with a probability of 0.8, and
2. at right angles to the direction \( d \) with probabilities 0.1 and 0.1.

SEQUENTIAL DECISION PROBLEMS

Example. Consider an agent situated in the following environment:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The agent may perform actions North, South, East, and West in order to move between squares (or states) \((1, 1), \ldots, (4, 3)\).
- Moving towards a wall results in no change in position.
- The operation of the agent stops and it receives a reward/punishment if it reaches a square marked with \(+1/−1\).

Example. If an action sequence \( S = [\text{North}, \text{East}] \) is performed in state \( (3, 2) \) the agent reaches states with following probabilities:

\[
\begin{align*}
P_{(3, 1)} &= 0.1 \times 0.1 = 0.01 \\
P_{(3, 2)} &= 0.8 \times 0.1 = 0.08 \\
P_{(3, 3)} &= 0.8 \times 0.1 + 0.1 \times 0.1 = 0.09 \\
P_{(4, 2)} &= 0.1 + 0.1 \times 0.8 = 0.18 \\
P_{(4, 4)} &= 0.8 \times 0.8 = 0.64 \\
\end{align*}
\]

These are easily inspected from a (partial) reachability graph:
Utilities and Plans

- Utility function $U$ is based on a sequence of states (an environment history) rather than a single state.

Example. In our example, the utility is defined as the value of the terminal state ($+1$ or $-1$) minus $\frac{1}{36}$ of the length of the sequence.

- Considering sequences of actions as long actions implies committing to an entire sequence of actions before executing it.

- In practice, the agent should be able to choose a new action in each state given any additional information provided by sensors.

- In stochastic environments, plans have to be conditional and it may be impossible to set a limit for lengths of conditional plans.

Policies

- We concentrate on accessible environments where the agent's percepts are always sufficient for determining the state it is in.

- A policy is a complete mapping from states to actions.

- Given a policy, it is possible to calculate the expected utility of the possible environment histories generated by that policy.

- It is non-trivial to compute an optimal policy that results in the highest expected utility (recall the MEU principle).

- If the agents knows an optimal policy, then it can choose an action in a deterministic fashion in every state.

Markov Decision Problems

- The problem of calculating an optimal policy in an accessible, stochastic environment with a known transition model is called a Markov decision problem (MDP).

- It is said that the Markov property holds if the transition probabilities depend only on the state (not on previous history).

- In the sequel, we will study two basic techniques for solving MDPs, namely value iteration and policy iteration.

- In an inaccessible environment, the corresponding problem is called a partially observable MDP (or POMDP).

- Solving POMDPs is much more difficult than solving MDPs.
**VALUE ITERATION**

- In value iteration, the idea is to compute the utility $U(s)$ for each state $s$ and to use these utilities for selecting optimal actions.
- It is difficult to determine $U(s)$ because of uncertain actions.
- Let $H(s,p)$ denote the history tree which results when starting from a state $s$ - actions are taken according to a policy $p$.
- Given a transition model $M$, the expected utility of a state $s$ is
  
  $$U(s) = \text{EU}(H(s,policy^*) | M)$$
  
  $$= \sum P(H(s,policy^*) | M)U_h(H(s,policy^*))$$

  where $policy^*$ is an optimal policy defined by $M$ and the utility function $U_h$ on state histories.

**How to Derive an Optimal Policy?**

- It is required that the utility function $U_h$ on histories is separable:
  
  $$U_h([s_0, \ldots, s_n]) = f(s_0, U_h([s_1, \ldots, s_n]))$$
  
  for some $f$.
- The simplest form of a separable utility function is additive:
  
  $$U_h([s_0, s_1, \ldots, s_n]) = R(s_0) + U_h([s_1, \ldots, s_n])$$

  where $R$ is a reward function on individual states $s$.
- Given an additive utility function $U_h$, an optimal policy $policy^*$ in state $i$ can be defined by the standard MEU principle:
  
  $$policy^*(i) = \arg \max_{a} \sum_{j} M_{ij}^a U(j)$$

  $$U(i) = R(i) + \max_{a} \sum_{j} M_{ij}^a U(j)$$

**Dynamic Programming**

- In value iteration, the idea is to compute the utility $U(s)$ for each state $s$ and to use these utilities for selecting optimal actions.

**Value Iteration Algorithm**

- There is an iterative procedure that approximates the utilities of states to any degree of accuracy.
- The next estimate $U_{i+1}(i)$ is based on the old utility estimates of the neighboring states: $U_{i+1}(i) = R(i) + \max_{a} \sum_{j} M_{ij}^a U_{i}(j)$.
- There is no bound on the length of actions sequences.
As \( t \) grows, the utility values will converge to stable values given certain conditions on the environment.

Given a stabilized utility function, the corresponding optimal policy [shown by Bellman and Dreyfus, 1962] is easy to compute.

![Convergence](image)

Un fortunately, it is difficult to estimate how long the value iteration algorithm should be run to get an optimal policy.

The progress of value iteration can be measured using root mean square error (RMS error) if the correct values are known.

Alternatively, policies can be evaluated using policy loss, i.e., the difference of expected utility with respect to the optimal policy.

![Root mean square error](image)

A modified value iteration algorithm can be used.

Unfortunately, value iteration may converge very slowly.

Another approach is to solve utilities directly using equations

\[
U(i) = R(i) + \sum_j M_{ij}^p U(j)
\]

that characterize stabilized utility values (\( \forall i: U_{t+1}(i) = U_t(i) \)).

Example. The utilities of states (3, 2) and (3, 3) are solved as follows:

\[
\begin{align*}
  u_{(3,2)} &= -0.04 + 0.8u_{(3,3)} + 0.1u_{(3,2)} - 0.1 \\
  u_{(3,3)} &= -0.04 + 0.8 + 0.1u_{(3,3)} + 0.1u_{(3,2)} \\
  &\Rightarrow -0.8u_{(3,3)} = -0.9u_{(3,2)} - 0.14 \\
  &\Rightarrow 8.1u_{(3,3)} = 0.9u_{(3,2)} + 6.84 \\
  &\Rightarrow u_{(3,3)} = \frac{0.7}{8.1} \approx 0.0876, \\
  u_{(3,2)} &= \frac{0.8u_{(3,3)} - 0.14}{0.9} \approx 0.660.
\end{align*}
\]

![Policy iteration](image)
Once the utilities of all states are known, it is straightforward to update the current policy using the MEU principle.

function POLICY-ITERATION(M, R) returns a policy
inputs: M, a transition model
        R, a reward function on states
local variables: U, a utility function, initially identical to R
                P, a policy, initially optimal with respect to U
repeat
    U ← VALUE-DETERMINATION(P, U, M, R)
    unchanged? ← true
    for each state i do
        if max \( \sum M_{ij} U(j) - \sum M_{ij} U(i) \) then
            P(i) ← arg max \( \sum M_{ij} U(j) \)
            unchanged? ← false
    end
until unchanged?
return P

DECISION-THEORETIC AGENT DESIGN

Recall the schematic design of decision theoretic agents performing decision cycles repeatedly:

function DECISION-THEORETIC-AGENT(percept) returns action
    calculate updated probabilities for current state based on
    available evidence including current percept and previous action
    calculate outcome probabilities for actions
    given action descriptions and probabilities of current states
    select action with highest expected utility
    given probabilities of outcomes and utility information
    return action

The components of the cycle are refined gradually in the sequel.
We begin with the problem of determining the current state.

How Immortal Agents Decide What to Do?

The total reward obtained by a policy can easily be unbounded if the lifetime of an agent is not limited.

In discounting, rewards received in the future are considered less valuable than rewards received in the current time step.

Given a discount factor \( 0 \leq \gamma < 1 \), the sum \( U(H) = \sum_{i=1}^{\infty} \gamma^i R_i \) of rewards \( R_1, R_2, \ldots \) (bounded by \( R \)) in a history \( H \) converges.

Discounting conforms to a preference-independence assumption called stationarity: if \( R_1 = S_1 \) holds for two reward sequences \( R_1, R_2, \ldots \) and \( S_1, S_2, \ldots \), then these sequences should be preference ordered in the same way as \( R_2, R_3, \ldots \) and \( S_2, S_3, \ldots \).

An optimal policy yields a constant system gain in the long run.

Determining the Current State of the World

In general, it is assumed that a set of state variables \( X_t \) (indexed by time \( t \)) refers to the current state of the world.

Given the percept history \( E_1, \ldots, E_t \) and the previous actions \( A_1, \ldots, A_{t-1} \), we are interested in the probability distribution

\[
Bel(X_t) = P(X_t | E_1, \ldots, E_t, A_1, \ldots, A_{t-1}).
\]

The direct evaluation of \( Bel(X_t) \) is out of the question, as it requires conditioning on many variables.

Conditional independence statements can be introduced in order to simplify the expression for \( Bel(X_t) \).
Simplifying Assumptions

- Assuming the Markov property, we obtain
  \[ P(X_t \mid X_1, \ldots, X_{t-1}, A_1, \ldots, A_{t-1}) = P(X_t \mid X_{t-1}, A_{t-1}) \].
- The Markov property can be established by introducing state variables that record relevant information from percepts.
  **Example.** If the robot is battery-powered, then the state variable `BatteryLevel` is needed to restore the Markov property.
- Percepts are causally determined by the state of the world:
  \[ P(E_t \mid X_1, \ldots, X_t, A_1, \ldots, A_{t-1}, E_1, \ldots, E_{t-1}) = P(E_t \mid X_t) \].
- The action taken depends only on the percepts received to date:
  \[ P(A_{t-1} \mid A_1, \ldots, A_{t-2}, E_1, \ldots, E_{t-1}) = P(A_{t-1} \mid E_1, \ldots, E_{t-1}) \].

Calculating the State Estimate \( \widehat{Bel}(X_t) \)

- The calculation takes place in two phases:
  1. **Prediction phase:** the prior probability distribution \( \widehat{Bel}(X_t) \) based on the previous state and how actions affect states:
     \[ \sum_{X_{t-1}} P(X_t \mid X_{t-1}, A_{t-1}) Bel(X_{t-1} = x_{t-1}) \].
  2. **Estimation phase:** the effect of the most recent percept \( E_t \) is incorporated to the distribution \( \widehat{Bel}(X_t) \) by Bayesian updating:
     \[ Bel(X_t) = \alpha P(E_t \mid X_t) \widehat{Bel}(X_t) \]
     where \( \alpha \) is a normalization constant.
- The equations for \( Bel \) and \( \widehat{Bel} \) form a generalization of Kalman filtering – a technique of classical control theory.

The Complete Decision-Theoretic Design

- The remaining steps of the decision cycle are straightforward.
- The **sensor model** \( P(E_t \mid X_t) \) describes how the environment generates the sensor data.
- The **action model** \( P(X_t \mid X_{t-1}, A_{t-1}) \) gives the effects of actions.

Sensing in Uncertain Worlds

- A sensor model is **stationary** if it holds for all \( t \) that
  \[ P(E_t \mid X_t) = P(E_t \mid X) \].
- A fixed model \( P(E_t \mid X) \) can be used at each time step.
- All variables affecting the sensors have to be included in \( X \).
- A sensor model is easily implemented as a conditional probability table in a belief network (the figure on the left hand side):
The values of sensors are causally related to the state of the world.

A perfect sensor corresponds to a purely deterministic CPT.

Possible noise and errors in the sensor are taken into account in the probabilities of incorrect readings.

**Example.** In the burglar-alarm example, both JohnCalls and MaryCalls can be viewed as sensors for the Alarm state variable.

- Typically, each sensor only measures some small aspects of the total state (as illustrated in the figure on the right hand side).
- Decomposing the overall sensor model into several components may reduce the size of the CPTs required.

**Example.** Measuring Pressure and Temperature with sensors that measure Pressure/Temperature and Pressure \times Temperature leads to complicated sensor models that depend on both state variables.

---

**Sensor Failures**

- It may be difficult to detect a sensor failure.
- To handle sensor failures in the first place, the possibility of failure has to be taken into account in the sensor model.
- Sensor fusion may discount the readings of a failed sensor.
- One possibility is to add a detailed sensor failure model:

---

**Sensor Fusion**

- There are often several sensors measuring the same state variable.

- The sensor values are conditionally independent of each other, given the actual value of the state variable.
- **Sensor fusion** or **data fusion** is about interpreting and putting together perceptual information from multiple sensors.

**Example.** If the readings from gauges are $13.6^\circ K (\pm 0.5^\circ K)$ and $14.4^\circ K (\pm 0.5^\circ K)$, the temperature is between $13.9^\circ K$ and $14.1^\circ K$.

---

**Dynamic Belief Networks**

- A dynamic belief network (DBN) represents how the state of the environment evolves over time.
- In analogy to sensor models, a stationarity assumption is made: the distribution $P(X_t|X_{t-1},A_{t-1})$ is the same for all $t$.
- Moreover, the agent is assumed to be passively monitoring and predicting a changing environment (i.e., it performs no actions).
- A state evolution model where a sequence of $X_t$ values is based on a fixed distribution $P(X_t|X_{t-1})$ is called a Markov chain.
- DBNs will be generalized for the decision of actions later on.
Structure of Dynamic Belief Networks

> For each time step $t$, there is one node for each state variable $X_t$ and sensor variable $Y_t$, including appropriate interconnections.

The task is to calculate the probability distribution for $State_t$, given the evidence for ..., $Percept_{t-1}$, $Percept_t$.

Probabilistic projection means estimating how the state of the environment (i.e., $State_{t+n}$ with $n > 0$) evolves in the future.

Prediction and Estimation with Belief Networks

> The prediction and estimation phases of the refined decision cycle can be implemented as operations on belief networks.

It is sufficient to consider two time steps $t$ and $t-1$ (also called the slices of the network):

(a) **Prediction**: using $Bel(X_{t-1})$ calculate the prior distribution

   
   $Bel(X_t) = \sum_{X_{t-1}} P(X_t \mid X_{t-1} = x_{t-1}) Bel(X_{t-1} = x_{t-1}).$

(b) **Rollup**: remove the slice for $t-1$ from the network and add the prior probability tables (based on $Bel(X_t)$) for $X_t$.

(c) **Estimation**: add the new percept $E_t$, calculate $Bel(X_t)$ by updating the network, and add the slice for $t+1$.

After these three steps, the network is ready for the next cycle.

Example. Let us extend the sensor failure model presented earlier by adding state evolution models for the state variables *Weather*, *Terrain*, *SensorFailure* and *LanePosition*:

The model for the variable *SensorFailure* determines that the sensor usually stays broken once it gets broken.
**DYNAMIC DECISION NETWORKS**

- Any dynamic belief network can be converted into a **dynamic decision network** by adding utility nodes and decision nodes.

- The goal is to calculate the value of $D_t$ by the MEU principle.

- The utility of a decision sequence $d$ is a weighted sum of utilities associated with each possible percept sequence given $d$.

- The probabilities of percept sequences given $d$ are used as weights.

- Dynamic decision networks solve POMDPs only approximately.

---

**SUMMARY**

- A **optimal policy** associates an optimal decision with every state that the agent might reach.

- **Value iteration** and **policy iteration** are two methods for calculating optimal policies.

- Unbounded action sequences can be dealt with **discounting**.

- **Dynamic belief networks** can handle sensing and updating over time, and provide a direct implementation of the update cycle.

- **Dynamic decision networks** can solve sequential decision problems arising for agents in complex, uncertain domains.

---

**Discussion**

DDNs provide solutions to many problems arising in AI systems:

- Uncertainty is handled correctly, and sometimes efficiently.

- Continuous streams of sensor input can be dealt with.

- Unexpected events are supported, since fixed plans are not used.

- Noisy and failing sensors can be modeled.

- The relevance of information can be estimated before acquisition.

- Relatively large state spaces can be handled if states can be represented by state variables with sparse connections.

- There are techniques for approximative reasoning.

---

**QUESTIONS**

1. Recall the belief network that you designed for representing the ball tracking mechanism of a soccer playing agent.

   - Is it possible to identify a state evolution model and a sensor model from your network?

   - If not, reconstruct the network by keeping these in mind.

2. Continue the analysis of soccer playing agents.

   - Can you identify other problems in this domain that can be considered as real sequential decision problems?

   - Try to formalize such a problem as a dynamic decision network.