Utility function

- Utility defined already in Chapter 2
- Captures an agent’s preference between world states
- Assigns single number to express desirability of a state
- The utility for a state $S$ is denoted by $U(S)$

Maximum Expected Utility (MEU)

- Assume result of actions are non-deterministic
- Result of action $A$ is $Result(A)$
- Given evidence $E$ the prob for each result is

$$P(Result_i(A)|Do(A), E)$$

- A rational agent picks an action that maximizes the expected utility

$$E[U(A|E)] = \sum P(Result_i(A)|Do(A), E)U(Result_i(A))$$

Not as easy as it sounds to apply principle of MEU

- State of the world?
- Computing $P(Result_i(A)|Do(A), E)$?
- Utility of each state? (need to account for consequences)
- Computations can be very involved!
Part A: Single shot decisions
- Utility
- Value of information
- Decision networks

Part B: Sequential decisions
- MDP
- POMDP
- Game Theory

Why MEU
- Why maximize expected utility?
- Is there such a thing as utility
- Look at something more basic
- Agents are assumed at least to have preferences

Preferences
- An agent chooses among \((A, B, \text{ etc.})\)
- Notation:
  - \(A \succ B\) A preferred to B
  - \(A \sim B\) indifference between A and B
  - \(A \succsim B\) B not preferred to A
- What are \(A\) and \(B\)?
- Can be concrete outcome states of actions, but more generally lotteries

Lotteries
- Situations with uncertain prizes
- Lottery \(L = [p, A; (1 – p), B]\)
- Probability \(p\) to get prize \(A\) and \(1 – p\) to get \(B\)
Preference constraints: Axioms of Utility

- **Orderability**
  \((A \succ B) \lor (B \succ A) \lor (A \sim B)\)
- **Transitivity**
  \((A \succ B) \land (B \succ C) \Rightarrow (A \succ C)\)
- **Continuity**
  \(A \succ B \succ C \Rightarrow \exists p \ [p, A; 1 - p, C] \sim B\)
- **Substitutability**
  \(A \sim B \Rightarrow [p, A; 1 - p, C] \sim [p, B; 1 - p, C]\)
- **Monotonicity**
  \(A \succ B \Rightarrow (p \geq q \Rightarrow [p, A; 1 - p, B] \succ [q, A; 1 - q, B])\)
- **Decomposability**
  \([p, A; 1 - p; [q, B; 1 - q, C]] \sim [p, A; (1 - p)q, B; (1 - p)(1 - q), C]\)

Utility principle

- If agent’s preferences obey axioms of utility
  \(\Rightarrow\) exist a real-valued function \(U\) such that
  
  \[U(A) > U(B) \iff A \succ B\]
  
  \[U(A) = U(B) \iff A \sim B\]

- \(U\) is the Utility!

Maximum expected utility principle

- The utility of a lottery is the sum of the utilities of the outcomes weighted with their probability of occurring
  
  \[U([p_1, S_1; \ldots; p_n, S_n]) = \sum_{i} p_i U(S_i)\]

- We want to maximize utility
  \(\Rightarrow\) MEU principle!

More on utility function

- As said before: maps from states to real numbers, \(U(S)\)
- Each agent can have its own opinion on the definition of utility
- Often depends on the situation
  Water typically has high utility in Sahara, but not in the middle of a fresh water lake
- Utility comes from economics
- Is money a good utility measure?
Money as a utility measure

- Given a lottery $L$ with expected monetary value $EMV(L)$, most people agree $U(L) < U(EMV(L))$
- People are risk-averse
- Example: For what probability $p$ am I indifferent between a prize $x$ and a lottery $[p, M; (1-p), 0]$ for large $M$?
- Typical empirical data
  - $U$ proportional to logarithm

Human judgement I

- Two lotteries
  - A: 80% chance of $4000
  - B: 100% chance of $3000
- Which do you prefer A or B?

Human judgement II

Two lotteries
  - C: 20% chance of $4000
  - D: 25% chance of $3000
- Which do you prefer C or D?

Evaluation of human judgement

- $U(A) = 0.8U(4000)$
- $U(B) = U(3000)$
- $U(C) = 0.2U(4000)$
- $U(D) = 0.25U(3000)$
- Looking at it mathematically
  - $U(A) < U(B) \Rightarrow U(C) < U(D)$
  - $U(A) > U(B) \Rightarrow U(C) > U(D)$
- Most people act irrational here
The value of information

- Asking for information one of the most important decisions
- Information is acquired through “sensing actions”
- Information typically cost
- What information to ask for?

Example: value of information

- 5 envelops
- 1 contains 100kr (you do not know which)
- It costs 20kr to buy one (Bad deal for me!)
- How much would you pay someone to tell you what is in one of the envelops?

Example cont’d

- Probability \( \frac{1}{5} \) it is the envelop with 100kr. You can then buy it for 20kr. Profit \( 100 - 20 = 80 \).
- Probability \( \frac{4}{5} \) it is not the right envelops. There are now 4 left only, i.e. \( \frac{1}{4} \) to pick the right one. Expected profit of \( \frac{100}{4} - 20 = 5 \).
- Expected profit given information
  \[ \frac{1}{5} \times 80 + \frac{4}{5} \times 5 = 20 \]
- That is, the information is worth 20kr!!!

For discussion at lecture see
http://www.grand-illusions.com/monty.htm
Value of information

- Information has value if it might change your action
- Value = difference between expected value with and without the information

Value of perfect information (VPI)

Value of current best action $\alpha$

$$E[U(\alpha|E)] = \max_A \sum_i U(\text{Result}_i(A))P(\text{Result}_i(A)|\text{Do}(A), E)$$

Value with information $E_j$

$$E[U(\alpha_{E_j}|E, E_j)] = \max_A \sum_i U(\text{Result}_i(A))P(\text{Result}_i(A)|\text{Do}(A), E, E_j)$$

$E_j$ is a random variable, whose value is unknown. Need to average over all possible values for $E_j$

VPI cont’d

Difference between expected value with the unknown information and the expected value without the information

$$VPI_E(E_j) = \left( \sum_k P(E_j = e_k|E)E[U(\alpha_{e_k}|E; E_j = e_k)] \right) - E[U(\alpha|E)]$$

Properties of VPI

- Nonnegative (in expectation)
  $$\forall j, E \quad VPI_E(E_j) \geq 0$$
- Nonadditive (in general)
  $$VPI_E(E_j, E_k) \neq VPI_E(E_j) + VPI_E(E_k)$$
- Order-independent
  $$VPI_E(E_j, E_k) = VPI_E(E_j) + VPI_{E_j}(E_k) = VPI_E(E_k) + VPI_{E_k}(E_j)$$
### Three cases

- Actions $A_1$ and $A_2$ have utilities $U_1$ and $U_2$ without information.
- Have pdf for $U'_1$ and $U'_2$ with information.
- When is it worth to ask for information?

(a) Information worth little $A_1$ better than $A_2$
(b) Information worth much, not clear which is best
(c) Information worth little, makes little difference

### Information gathering agent

- Should ask questions in reasonable order
- Don’t ask irrelevant questions
- Account for cost of information
- Value of information a good guide for this

### Decision networks (or Influence diagrams)

- Bayesian network with nodes for actions and utilities
  ![Decision Network Diagram](image)
- Chance nodes (ovals) random variables
- Decision nodes (rectangles) choice of action
- Utility nodes (diamonds) agent’s utility function. Child of all nodes directly affecting the utility.

### Semantics of decision networks

- Arcs into chance nodes
  Probabilistic dependence
- Arcs into decision nodes
  available information at the time of the decision
- Arcs into utility nodes
  What parameters the utility depends on
Decision-Theoretic Expert System

- Utility helps distinguish between likelihood and importance
- Ex: medical expert system
  - Most likely: “nothing wrong”
  - Second most likely “dying”
- Must verify and refine a model using a gold-standard
  Ex: medical expert system. Compare with doctors
- Check sensitivity!

Ex: Sequential decision problems

- How do we get to goal +1?
- Assume fully observable environment

Ex: Sequential decision problems cont’d

- What if actions not deterministic?
  - Prob 0.8 to move in desired direction
  - Prob 0.1 90° wrong
- Cannot create plan ahead of time!
Execute predefined plan

- Old plan: Up, Up, Right, Right, Right
- Probability to succeed = 0.32776

\[0.8^5 + (0.14 \cdot 0.8) = 0.32776\]

Model the problem

- Transition model (successor function)?
- Utility?

Transition model

- Specifies outcome probability
- \(T(s, a, s')\)
  - Probability to reach \(s\) starting from \(s'\) given action \(a\).
- Will assume a first-order Markov process
  - \(T\) depends only on previous state \(s'\) and not the rest of the history

Utility function

- Assume agent gets reward \(R(s)\) for being in state \(s\)
  - \(R([4, 3]) = +1\) (Go to goal)
  - \(R([4, 2]) = -1\) (Avoid trap)
  - \(R(\text{rest}) = -0.04\) (Get to goal quickly)
- Let the utility of the history be the sum of the state utilities for that history
Markov Decision Process

- Problem is called a Markov Decision Process (MDP)
- Assumes fully observable environment
- Defined by:
  - Initial state: $S_0$
  - Transition model: $T(s, a, s')$
  - Reward: $R(s)$

Solution to MDP

- A solution to an MDP cannot be fixed plan
  (nondeterministic world, need to sense state)
- It is a policy $\pi$
- Maps state to action, $a = \pi(s)$
- Can be different path every time

How good is a policy?

- How to measure the quality of a policy
- Measured as expected utility over the history
  (stochastic env -> need to use expectations)
- Optimal policy: highest possible expected utility
- Optimal policy: $\pi^*$

Ex: Optimal policy $\pi^*$

- Optimal policy for previous problem
Optimal policy depends on $T$ and $R$!!

- Different behaviors for different $R$

Utilities of sequences

- Have already considered additive rewards (MEU)
  \[ U_h([s_0, s_1, \ldots, s_n]) = R(s_0) + R(s_1) + \ldots + R(s_n) \]
  - What about infinite sequences?
  - Might get $\infty$ without terminal state?
  - How to compare $\infty$ and $\infty$?

Discounted rewards

- Idea: Give less weights to future rewards
- Use discount factor $\gamma$
  \[ U_h([s_0, s_1, \ldots]) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \ldots \]
  - Gives bounded utility
  - $R(s) \leq R_{\max} \Rightarrow U_h \leq \sum_{i=0}^{\infty} \gamma^i R_{\max} = R_{\max} \frac{1}{1-\gamma}$

Selecting the best policy

- How do we select the best policy?
- Get many state sequences
- As before, maximize expected utility
  \[ \pi^* = \arg \max_{\pi} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) \mid \pi \right] \]
Need method for calculating optimal policy

Key insight: Utility of a state is immediate reward plus discounted expected utility of next state

\[ U(s) = R(s) + \gamma \max_a \sum_{s'} T(s, a, s') U(s') \]

(Assuming that we choose the optimal policy)

Idea: Iterate

- Calculate utility of each state
- Use utilities to select optimal decision in each state

Initialize \( U(s) \) arbitrarily for all \( s \)

Loop until policy is good enough

\[ Q(s, a) := R(s) + \gamma \sum_{s'} T(s, a, s') U(s') \]

end

\[ U(s) := \max_a Q(s, a) \]

end

Bellman equation

- Bellman equation

\[ U_{i+1}(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} T(s, a, s') U_i(s') \]

- Converges to unique optimal solution
- Stop iterations when largest change in utility for a state is small enough
- Can show that

\[ \| U_{i+1} - U_i \| < \epsilon \]

First iteration, assume all \( U(s) = 0 \) and \( \gamma = 0.9 \)

What is the value of cell (3,1)?

Java demo
Part A
Part B
MDP
POMDP
Game Theory

Policy iteration

- See book

Partially observable environment

- What if environment is not observable?
- Cannot execute policy since state is unknown!
- Results in a Partially Observable MDP or POMDP
- Sensors provide observations of environment
- Observation model $O(s, o)$ gives probability of making observation $o$ in state $s$

First attempt at POMDP

- No measurements
- Initial position unknown
- First attempt on a plan:
  - Move left 5 times (likely that you are at left side)
  - Move up 5 time (probably in the upper left corner)
  - Move right 5 time
- 77.5% success rate
- Expected utility only 0.08

Revisit belief state

- Already seen that the belief state can be used in partially observable world (conditional planning, vacuum cleaner)
- Definition: Belief state $b$ is a probability distribution over all possible states.
- $b(s)$ is probability of being in state $s$
- Example: initial belief state 9 equally probable states $\langle \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, 0, 0 \rangle$
Belief state update

- Need to update belief state as we go along
- Assume belief state $b$ and action $a$
- Update:
  \[ b'(s') = \alpha O(s', o) \sum_s T(s, a, s') b(s) \]
- $\alpha$ normalization factor such that $\sum_{s'} b'(s') = 1$

Solving POMDP

- Key insight: Optimal action depends on belief state and not actual state!
- Optimal policy $\pi^*(b)$
- Decision cycle:
  1. Execute action $a = \pi^*(b)$
  2. Receive observation
  3. Update belief state

Transfer MDP to POMDP

- Introduce
  - $\tau(b, a, b')$ - probability of reaching belief state $s'$ from $b$ given action $a$
  - $\rho(b) = \sum_s b(s) R(s)$ - expected reward
- $\tau(b, a, b')$ and $\rho(b)$ define an observable MDP
- Optimal MDP strategy $\pi^*(b)$ is also optimal for original POMDP

Second attempt on POMDP

- No observations $\Rightarrow$
  Problem is deterministic in belief space
- The policy is a fixed sequence
- Optimal sequence is:
  Left, Up, Up, Right, Up, Right, Up, Up, Right, Up, Right, Up, Right, Up
- Expected utility 0.38 (0.08 before)
So it is simple?

- Sounds “simple” at first, BUT
- Belief state is probability distribution!
- Compare MDP and POMDP in 4x3 world
- MDP: The position of the agent
  1 discrete with 11 possible values
- POMDP: 11 dimensional vector!!!
  11 continues variables

Example 2: Tiger Problem

- “Classic” POMDP example
- State: Tiger behind left door or right door
- Actions: Open left, open right or listen
- Listening costs and is uncertain
- What to do?
- Check out Tony’s POMDP page

POMDP

- POMDPs in general are very hard to solve
- With observations and a few dozen states it is often infeasible

Game Theory

- Other agents also introduce uncertainty (what decision do they make?)
- Important area not only to make money at gambling
- Two main areas:
  - Agent design:
    - What is the best strategy for a player?
  - Mechanism design:
    - How do we construct rules such that the best policy for the individual agents are also for the good for all
Single Move Games

- Main components:
  - Players or agents
  - Actions
  - Payoff-matrix: Gives utility for each player for each combination of actions
- Strategy = policy

Example: Prisoner's Dilemma

Burglars Alice and Bob are caught
Interrogate separately by the police
Rules of the game:
- Both confess: 5 years each
- Both refuse to confess: 1 year each
- Testify: One 10 year, other go free

What to do?

Prisoner's dilemma cont’d

Each want to maximize expected utility
Payoff matrix

<table>
<thead>
<tr>
<th></th>
<th>Alice:testify</th>
<th>Alice:refuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob:testify</td>
<td>A=-5,B=-5</td>
<td>A=-10,B=0</td>
</tr>
<tr>
<td>Bob:refuse</td>
<td>A=0,B=-10</td>
<td>A=-1,B=-1</td>
</tr>
</tbody>
</table>

Alice analysis:
- Bob testifies: Best to testify (-5 > -10)
- Bob refuses: Best to testify (0 > -1)

Best to testify!
Bob feels the same way.
Nash equilibrium

- Nash equilibrium:
  Set of strategies such that no player can benefit by changing her strategy while the other players keep their strategies unchanged
- Alice and Bob are stuck in a Nash equilibrium

Repeated game

- What if Alice and Bob faces the same situation many times
- Will the strategy be the same?

Repeated game cont'd

- Two cases:
  - Know how many times
  - Don't know how many times

Repeated game I

- Alice and Bob will face each other 100 times
- Know that 100th is the last and the best is to use same as before than
- Ends up with always testifying
Repeated game II

- Alice and Bob has a 99% chance of meeting again
- 100 expected encounters
- Encourages more cooperative strategies, e.g.
  - Perpetual punishment
  - tit-for-tat

Example: Auction (Mechanism design)

- Auction
- Single good
- Each player has utility value $v_i$ for the good
- Utility value known only to the bidder
- Bidder make bids $b_i$, highest bid wins

Example: English auction

- Bids are incremented
- Continues until only one bidder
- Bidder with highest $v_i$ wins
- Pays $b_m + d$ ($b_m$ highest among other, d increment)
- Strategy: bid as long as price below your value
- High communication cost!

Example: Sealed bid auction

- Each bidder make single bid
- Highest bid wins
- Strategy: bid $\min(v_i, b_m + \epsilon)$ ($b_m$ believed max of other bids)
- Player with highest $v_i$ might not get the good
Example: Vickrey auction

- Same as sealed-bid auction but winner pays second highest bid
- Strategy: bid your own value $v_i$
- Simple and minimal communication

THE END