Chapter 9: Rationality and utility

DIT410/TIN172 Artificial Intelligence

Peter Ljunglöf
modifed from slides by Poole & Mackworth

8 May, 2015

Outline

1. *Preferences and Utility (9.1)*
   - Rationality
   - Are humans rational?

Outline

1. *Preferences and Utility (9.1)*
   - Rationality
   - Are humans rational?

Outcomes and preference

- An outcome is the result of an action.
- To be able to choose between actions, an agent has to have a preference of some actions over others.
  - if the agent does not have preferences over anything, it does not matter what the agent does

Example

- what are the outcomes of the travel-in-Romania problem?
- what are the preferences of these outcomes?

Formalising outcomes

$\succeq$ is the basic relation over outcomes: $o_1 \succeq o_2$ means that outcome o_1 is (at least) as desirable as outcome o_2.

From this we can define the following relations:

- $o_1 \sim o_2$ is defined as $o_1 \succeq o_2$ and $o_2 \succeq o_1$
- $o_1 \succ o_2$ is defined as $o_1 \succeq o_2$ and $o_2 \not\succeq o_1$

Note that outcomes do not have to be numbers – they can be arbitrarily complex objects!

Lotteries

A lottery is a probability distribution over outcomes:

$$[p_1 : o_1, p_2 : o_2, \ldots, p_k : o_k]$$

(where o_i are outcomes, and $p_i \geq 0$, and $\sum p_i = 1$)

- This lottery specifies that outcome o_i occurs with probability p_i.
- We will assume that outcomes include lotteries,
  i.e., that a lottery can be an outcome itself.

Outline

1. *Preferences and Utility (9.1)*
   - Rationality
   - Are humans rational?

Rational agents

An agent is rational if it obeys the axioms of rationality:

Completeness $o_1 \succeq o_2$ or $o_2 \succeq o_1$ (for all o_1, o_2)

Transitivity if $o_1 \succeq o_2$ and $o_2 \succeq o_3$, then $o_1 \succeq o_3$

Monotonicity if $o_1 \succ o_2$ and $p > q$, then
$[p : o_1, 1 - p : o_2] \succeq [q : o_1, 1 - q : o_2]$

Decomposability an agent is indifferent between lotteries that have the same probabilities over the same outcomes, even if one or both is a lottery over lotteries

Continuity if $o_1 \succeq o_2$ and $o_2 \succeq o_3$, then there is a p such that
$o_2 \sim [p : o_1, 1 - p : o_3]$

Substitutability if $o_1 \sim o_2$ then the agent is indifferent between lotteries that only differ by o_1 and o_2

Rational agents and utility

If an agent is rational, its preferences can be measured by a real-valued utility function over outcomes.

Theorem

For every rational agent there is a utility function u, such that:

- $o_i \succeq o_j$ *if and only if* $u(o_i) \geq u(o_j)$, *and*
- *utilities are linear with probabilities:*

$$\begin{aligned} & u([p_1 : o_1, p_2 : o_2, \ldots, p_k : o_k]) \\ = \quad & p_1 \cdot u(o_1) + p_2 \cdot u(o_2) + \cdots + p_k \cdot u(o_k) \end{aligned}$$

Risks

Most people are *risk averse* when it comes to money:

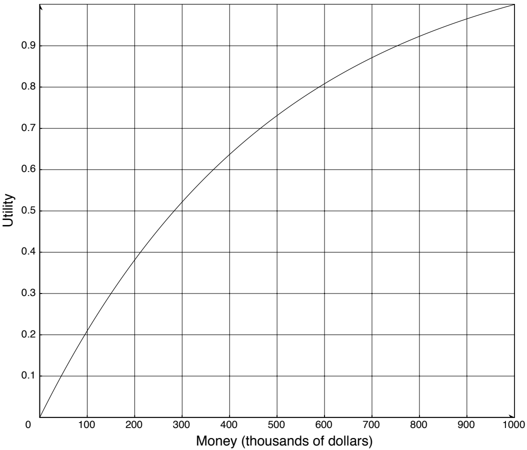
- would you prefer 100 000 kr in your hand, or
- 50% chance of winning 150 000 kr?

But if the reward is big enough, you might prefer to gamble:

- would you prefer 100 000 kr in your hand, or
- 50% chance of winning 1 000 000 kr?

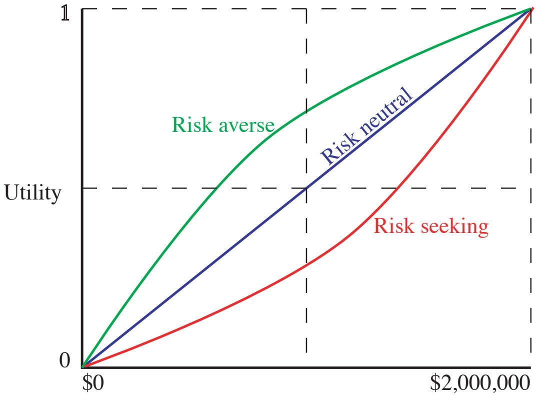
Risks and utility

A risk averse agent can be simulated using a *concave* utility function:



Risk averse vs. risk seeking

Risk averse vs. risk neutral vs. risk seeking utility functions:



Outline

1. *Preferences and Utility (9.1)*
   - Rationality
   - Are humans rational?

Are you rational?

Which would you prefer, A or B?

A: 11% chance of winning 1 000 000 kr
89% risk of getting nothing at all

B: 10% chance of winning 1 500 000 kr
90% risk of getting nothing at all

Are you rational? (part 2)

Which would you prefer, A or B?

A: to receive 1 000 000 kr

B: 10% chance of winning 1 500 000 kr
89% chance of winning 1 000 000 kr
1% risk of getting nothing at all

Humans are not rational

Maurice Allais tried this experiment in 1953 and showed that people preferred B over A in the first case, but A over B in the second case.

This is impossible to catch in a utility function – both cases are instances of this one:

A is a lottery $[0.89 : x,\ 0.11 : 1\,000\,000]$

B is a lottery $[0.89 : x,\ 0.10 : 1\,500\,000,\ 0.01 : 0]$

In the first case, $x = 0$, and in the second case, $x = 1\,000\,000$.

- People seem to have a preference for certainty
- It is inconsistent with the rationality axioms to have $B_1 \succ A_1$ and $A_2 \succ B_2$
- This is called the *Allais paradox*

Humans are not rational (part 2)

Tversky and Kahneman tried the following experiment in 1974:

Example

A disease is expected to kill 600 people. Which of the following two alternative programs for handling this would you favour?

Program A: 200 people will be saved

Program B: with probability $1/3$, 600 people will be saved, and with probability $2/3$, no one will be saved

Humans are not rational (part 2)

Tversky and Kahneman tried the following experiment in 1974:

Example

A disease is expected to kill 600 people. Which of the following two alternative programs for handling this would you favour?

Program C: 400 people will die

Program D: with probability $1/3$, no one will die, and
with probability $2/3$, 600 will die

Humans are not rational (part 2)

Tversky and Kahneman tried the following experiment in 1974:

Example

A disease is expected to kill 600 people. Which of the following two alternative programs for handling this would you favour?

Program A: 200 people will be saved

Program B: with probability $^{1}/_{3}$, 600 people will be saved, and with probability $^{2}/_{3}$, no one will be saved

In Tversky's and Kahneman's experiment,
72% chose A over B .

Humans are not rational (part 2)

Tversky and Kahneman tried the following experiment in 1974:

Example

A disease is expected to kill 600 people. Which of the following two alternative programs for handling this would you favour?

Program C: 400 people will die

Program D: with probability $^1/_3$, no one will die, and
with probability $^2/_3$, 600 will die

In Tversky's and Kahneman's experiment,
72% chose A over B, but only 22% chose C over D.

Humans are not rational (part 2)

Tversky and Kahneman tried the following experiment in 1974:

Example

A disease is expected to kill 600 people. Which of the following two alternative programs for handling this would you favour?

Program C: 400 people will die

Program D: with probability $1/3$, no one will die, and
with probability $2/3$, 600 will die

In Tversky's and Kahneman's experiment,
72% chose A over B, but only 22% chose C over D.

However, these are exactly the same choice,
only described in a different way!

Why rationality?

So, humans behave irrationally...

- ...but we already knew that
- the axioms of rationality are still very useful in many scenarios
- utility theory is a nice and simple mathematical theory

However, there are alternatives, such as *prospect theory*