Rational agents

Outline

- **Rational agents and decision making**
- Utility theory for decision making
  - Binary relations
  - Preferences
  - Utility
- Making decisions
- Example

Rational Agents

- Let us recall the following agent property:

Rationality

- **Agent's ability** to **act** (i.e., make decision) in a way that **maximizes some utility function**

Rational Agents

- But what is a **utility** function?

- How can we use a **utility** function to make decisions?

Rational Agents

- Before we analyze agents…
- How do we (humans) make decisions?

Making decisions

- **How do we (humans) make decisions?**

- **Example 1:**
  - You won the lottery

  - You have the following decision to make. Either:
    - (Decision1) you receive your prize **today** (EUR 1,000,000.00)
    - (Decision2) you receive **next month**

- What is your decision?

Making decisions

- **Example 2:**
  - You have a plane ticket to Madeira (EUR 100)
  - The flight is overbooked

Making decisions

- The airline must ask for people to volunteer not to fly 'in exchange for benefits’.
- You have the following decision to make. Either, you choose:
  - (Decision1) a rerouting option + EUR 100 in cash
  - (Decision2) to keep your plane ticket (not volunteer)
- What is your decision?

Making decisions

- **Example 3:**
  - You are planning a trip for your next vacation
  - You have the following decisions:
    - (Decision1) Go to Hawaii (EUR 900)
    - (Decision2) Go to Cancun (EUR 500)
  - What is your decision?

Hawaii

Cancun

Making decisions

- While the decision in Example 1 is straightforward
  - I want my money now!

- Decision in Example 2 and 3 depend on preferences. Hence, this can lead to different outcomes.

  - **Many decisions are based on personal preferences!**

Making decisions

Key questions:

- How can I tell an agent what are my “preferences”?
- Can I treat decision making algorithmically?

Outline

- Rational agents and decision making
- Utility theory for decision making
  - **Binary relations**
  - Preferences
  - Utility
- Making decisions
- Example

Bibliography

UTILITY THEORY
FOR
DECISION MAKING

PETER C. FISHBURN

Research Analysis Corporation

JOHN WILEY & SONS, INC

NEW YORK · LONDON · SYDNEY · TORONTO

Binary relations

- A binary relation R on a set of outcomes Y is a set of ordered pairs (x, y) with

$$x, y \in Y$$

- We can also write this binary relation as follows:

$$xRy$$

Binary relations

- Examples of a binary relation
  - Let R_1 mean “is shorter than”
  - John (x) is 1.75m and Harry (y) is 1.85m
  - Then we can say that:

$$(xRy, \text{not } yRx)$$

Some binary relation properties

- Reflexive $\quad$ if xRx for every $x \in Y$
- Irreflexive $\quad$ if not xRx for every $x \in Y$
- Symmetric $\quad$ if $xRy \implies yRx,\;$ for every $x, y \in Y$
- Asymmetric $\quad$ if $xRy \implies$ not $yRx,\;$ for every $x, y \in Y$
- Antisymmetric $\quad$ if $(xRy, yRx) \implies x = y,\;$ for every $x, y \in Y$

Some binary relation properties

- Transitive $$\text{if } (xRy, yRz) \implies xRz, \text{ for every } x, y, z \in Y$$
- Negatively transitive
$$\text{if } (\text{not } xRy, \text{not } yRz) \implies \text{not } xRz, \text{ for every } x, y, z \in Y$$
- Connected or Complete $$\text{if } xRy \text{ or } yRx (\text{possibly both}) \text{ for every } x, y \in Y$$
- Weakly connected $$\text{if } x \neq y \implies (xRy \text{ or } yRx) \text{ throuhout } Y$$

Some binary relation properties

- Relation "is shorter than" is
  - Irreflexive $$\text{if not } xRx \text{ for every } x \in Y$$

    informally: a person cannot be shorter than himself

  - Asymmetric $$\text{if } xRy \implies \text{not } yRx, \text{ for every } x, y \in Y$$

    informally: if person 1 is shorter than person 2 then person 2 is not shorter than person 1

  - Transitive $$\text{if } (xRy, yRz) \implies xRz, \text{ for every } x, y, z \in Y$$

    informally: if person 1 is shorter than person 2 and person 2 is shorter than person 3 then person 1 is shorter than person 3

Outline

- Rational agents and decision making
- Utility theory for decision making
  - Binary relations
  - **Preferences**
  - Utility
- Making decisions
- Example

Preferences

- **Strict preference** is a **binary relation** on the set of outcomes, such that

$$x \succ y$$

denotes the proposition that

x* is preferred to *y (or x is better than y)

- We can also use the strict preference to express:

$$x \prec y$$

y* is preferred to *x (or y is better than x)

Preferences

- We can also define **indifference** as the absence of preference

$$x \sim y \iff (\text{not } x \prec y, \text{ not } x \succ y)$$

the two outcome are indifferent

(or x is neither better nor worse than y)

- Indifference might arise in the following situations:
  - One might feel that there is no difference between the outcomes
  - One is uncertain about his preferences

Preferences

- We can also define **preference-indifference** as the union of strict preference and indifference

$$x \preceq y \iff (x \prec y \text{ or } x \sim y)$$

x is not better than y

- Or

$$x \succeq y \iff (x \succ y \text{ or } x \sim y)$$

x is not worse than y

Properties of Preferences

- **Strict preference** $\succ$
  - antisymmetric, transitive, and negatively transitive
- **Indifference** $\sim$
  - reflexive, symmetric, and transitive
- **Preference-indifference** $\succeq$
  - complete and transitive

Rational preference

- A **rational preference** is a binary relation if:
  - complete and transitive

- The **preference-indifference** is complete and transitive
  - Hence a **rational preference**

Outline

- Rational agents and decision making
- Utility theory for decision making
  - Binary relations
  - Preferences
  - **Utility**
- Making decisions
- Example

Utility

- **Why don’t we use (or code) preferences in our agents?**
  - From a computation perspective, they are cumbersome to maintain

- Recall that preferences express an ordering between outcomes
  - Thus, we can express the preferences with an **order-preserving function**

Utility

- **Does this order-preserving function exist?**
  - **Yes, if the preferences are rational**
  - When preferences are rational, **we can sort all outcomes consistently**

Utility

- **Theorem**:

Let X be a set of possible outcomes, and $\succeq$ a rational prefernce on X. Hence, there is a function $u : X \to \mathbb{R}$ such that $u(x) \geq u(y)$ if and only if $x \succeq y$, for all $x, y \in X$

We call u the utility function

Outline

- Rational agents and decision making
- Utility theory for decision making
  - Binary relations
  - Preferences
  - Utility
- **Making decisions**
- Example

Making Decisions

- Agents can use utility to make decisions:
  - Let A be a **set of actions**
  - Given $a \in A$, let $O(a)$ be an **outcome** when an agent selects action a
  - Hence, the **value of action** a is:

$$Q(a) \stackrel{\text{def}}{=} u(O(a))$$

Making Decisions

- So how can an agent make a decision?

$$\underset{a \in A}{\operatorname{argmax}}\, Q(a)$$

$$\underset{a \in A}{\operatorname{argmax}}\, u(O(a))$$

An agent selects an action with the maximum utility


UNCERTAINTY
AHEAD

Making Decisions Under Uncertainty

- So how can an agent make a decision?
  - Let O denote a finite **set of outcomes**
  - Given $o \in O$, let $P(o|a)$ denote the **probability of outcome** o when an agent selects action a
  - Hence, the **expected value of an action** is

$$Q(a) = \mathbb{E}[u(o)|a] = \sum_{o \in O} u(o)P(o|a)$$

Making Decisions Under Uncertainty

- So how can an agent make a decision?

$$\underset{a \in A}{\operatorname{argmax}}\, Q(a)$$

$$\underset{a \in A}{\operatorname{argmax}} \sum_{o \in O} u(o) P(o|a)$$

An agent selects an action with the maximum expected utility

Outline

- Motivation – making decisions
- Utility theory for decision making
  - Binary relations
  - Preferences
  - Utility
- Making decisions
- **Example**

Example: robot coffee machine

robot coffee machine

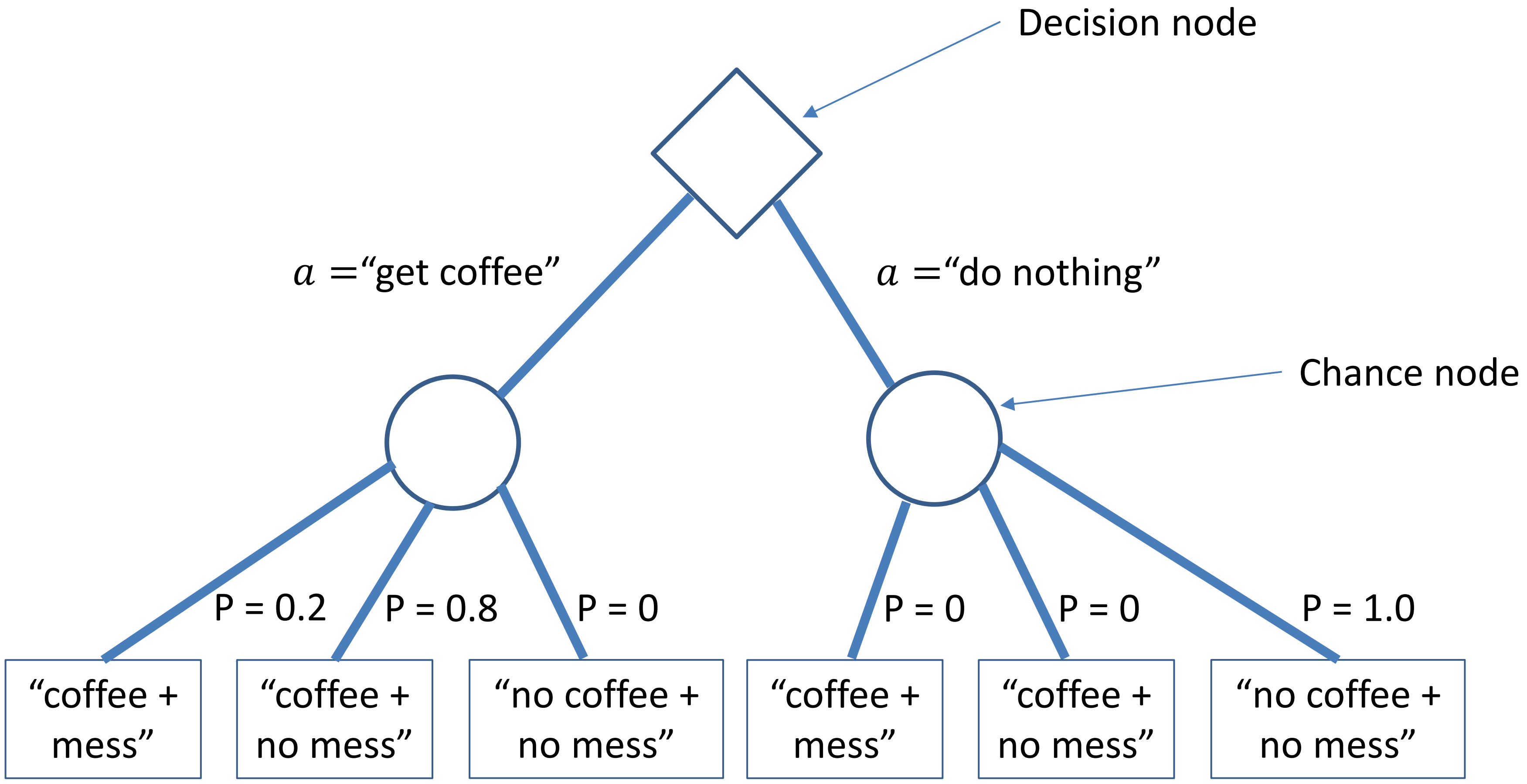
- O = {"coffee + mess", "coffee + no mess", "no coffee + no mess"}
  - set of outcomes
- A = {"get coffee", "do nothing"}
  - set of actions

robot coffee machine

- P is the probability of an outcome
  - $\mathrm{P}(o = \text{“coffee + mess”} | a = \text{“get coffee”})$ = 0.2
  - $\mathrm{P}(o = \text{“coffee + no mess”} | a = \text{“get coffee”})$ = 0.8
  - $\mathrm{P}(o = \text{“no coffee + no mess”} | a = \text{“get coffee”})$ = 0

  - $\mathrm{P}(o = \text{“coffee + mess”} | a = \text{“do nothing”})$ = 0
  - $\mathrm{P}(o = \text{“coffee + no mess”} | a = \text{“do nothing”})$ = 0
  - $\mathrm{P}(o = \text{“no coffee + no mess”} | a = \text{“do nothing”})$ = 1.0

Decision Tree

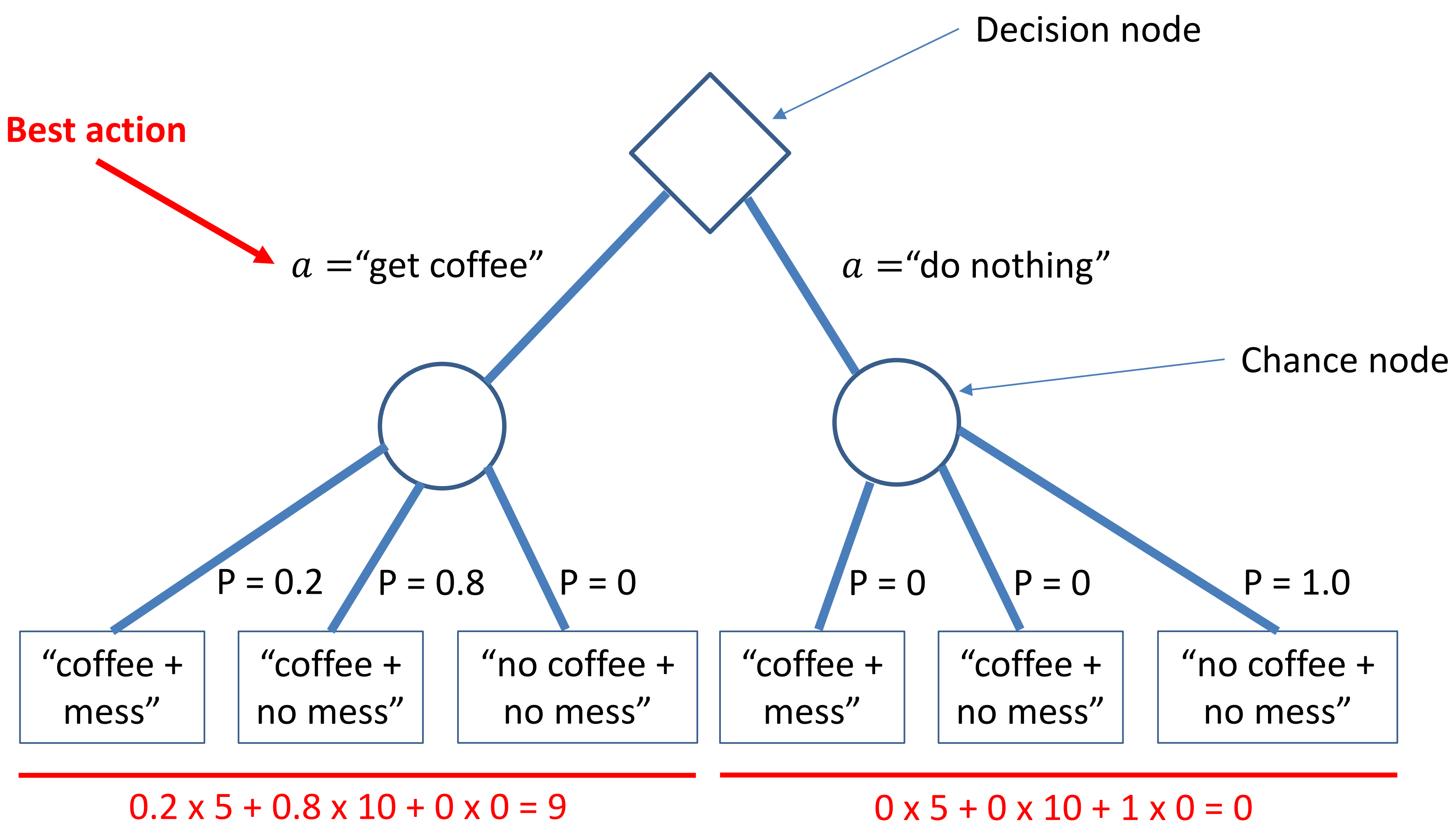


robot coffee machine

- Now let us assume a different utility function:
  - $u(s =$ "coffee + mess") = 5
  - $u(s =$ "coffee + no mess") = 10
  - $u(s =$ "no coffee + no mess") = 0

I love coffee!

Decision Tree



Final remarks

- We have only considered **decision-making problems** that has **ONE agent**
- What if our environment has **two or more agents**?
  - **Two or more utility-maximizing agents** whose actions can **affect each other's utility**
  - We need a decision-making framework: **GAME THEORY!**

Thank You

rui.prada@tecnico.ulisboa.pt