Utilities and rationality

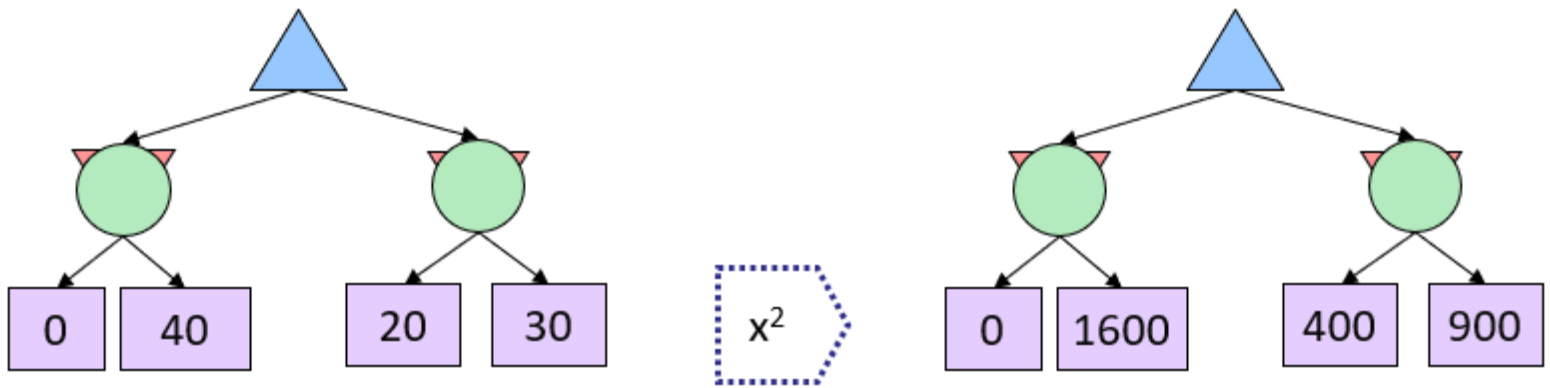
- We kept talking about **utility** but did not yet give a good definition what it is
  - Let us say that the values associated with the terminal outcomes are the utility
  - Agents pursue higher utility, so these values determine the behavior of the agent
- We say that a **rational agent** is one that chooses actions that **maximize its expected utility**, given its knowledge

Rationality

- So we defined rationality as utility maximization
  - You can also consider it as a definition of the kind of agent intelligence we are considering in this class
- The choice of utility function is up to the agent
  - power, art, scientific discovery
  - pleasure, procreation, public respect
  - number of followers, winning an election, winning a sport competition
  - a weighted sum of these or a preference ordered list of these
- There are a lot of choices, but not every possible choice allows a rational agent.

"Expected"?

- For minmax reasoning, utility is simple. The greater the utility outcome, the better.
  - So only **ordering** matters.
  - "insensitivity to monotonic transformations"
- But for expectimax, we take expectations, so suddenly the absolute values matter as well.
  - We should define utilities such that we can add and average them!


0
40
20
30
x2
0
1600
400
900

Utilities and preferences

- We define utilities as functions on states $U(s) \in \mathbb{R}$
- We will say that the utilities describe the **preferences of the agents**
  - It is a way to summarize the goals of the agent
- Two strategies to build an intelligent agent
  - **Behavior specification:** Describe its behavior for each state - i.e. write the $\pi(s) \rightarrow a$ function directly.
  - **Utility specification:** Provide a utility function $U(s)$. A rational agent will choose its own actions in the pursuit of the goal of maximizing expected utility.

Problems

- Problems with behavior specification:
  - **Framing problem:** need to handle a very large number of cases
    - Buy milk, unless you already have it, it is 2am, it is a hurricane, it is a zombie attack, ...
  - The relationship between the behavior and good outcomes difficult to prove.
- Problems with utility specification:
  - Where does the utility coming from?
  - Can every rational behavior be expressed as utilities?
  - **Theorem:** any rational preferences over states can be summarized as a utility function.

Preferences

- Prizes: A, B
- Lotteries: situations with uncertain prizes: $L = [p, A; (1 - p)B]$
- The agent prefers A denoted by $A \succ B$
  - A good way to think about it is that the agent would pay at least \$0.01 to get A instead of B
- The agent is **indiferent** denoted by $A \sim B$

Rational preferences

- What kind of preferences can be considered rational?
- Let us imagine an agent with $A \succ B, B \succ C, C \succ A$.
- Such an agent can be induced to give away all his money!
  - This happens because the preferences are not transitive
  - "Dutch book" auctions in horse races.

A set of axioms that ensure rationality

- Orderability

$$(A \succ B) \vee (B \succ A) \vee (A \sim B)$$

- Transitivity

$$(A \succ B) \wedge (B \succ C) \Rightarrow (A \succ C)$$

- 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

$$A \succ B \Rightarrow (p \geq q \Leftrightarrow [p, A; 1-p, B] \succeq [q, A; 1-q, B])$$

Maximum expected utility principle

- Given preferences satisfying the axioms, there exists a utility function $U(s)$ such that

$$U(A) \geq U(B) \Leftrightarrow A \succeq B$$

$$U([p_1, S_1; \ldots; p_n, S_n]) = \sum_i p_i U(S_i)$$

(Ramsey 1931, von Neumann & Morgenstern 1944)

Utilities and humans

- We can try to **elicit** human preferences by presenting humans with lotteries
  - Most studies concluded that *humans are not rational*
  - Psychologists had a field day with this!
- Some of it might be that we have limited computing power $\rightarrow$
  - "Bounded rationality", "Satisficing" actions
  - Herbert Simon 1978 Nobel Prize in Economics
- But some of it would be also be that we do not restrict our thinking at the specific setting of the problem
  - i.e. only two choices with no other implications, no repeated games and no temporal setting

Examples of utility calculations in human affairs

- **Micromorts**: 10^{-6} chance of death
  - How much are you willing to pay for a 7th airbag in your car?
- **QALY**: quality adjusted life-years, useful for medical decisions
  - Who gets the heart transplant etc.

Money

- You can calculate an expected monetary value (EVM) of a transaction by calculating the expectation of probabilities
- But money does **not** behave as a utility function
- Most people are **risk averse**
  - A decrease in money by X triggers a greater utility change than the same increase
- When deep in dept, people are **risk prone**

Insurance

- How much are you willing to buy this lottery: $[0.5, \$10000; 0.5, \$0]$
  - i.e. the **certainty equivalent**
- The difference between the certainty equivalent and the EVM is the **insurance premium**
- Why does this work out for the insurance company?
  - They have a different utility curve (more rational)
  - They average over a different lotteries.

Rewards

- Utilities are associated with end states.
- Our definition of rationality based on utilities is simply based on what kind of end state we prefer.
- But often this is not the real problem the agent faces
  - Usually, the preferred end state is clear, the problem is how to get there.
  - Getting to the preferred end-state is a multi-step affair, and in each step we need to take the right action.

Rewards (cont'd)

- Idea: associate a **reward** with every step, for instance in the form $R(s, a, s') \in \mathbb{R}$
- In the simplest form: the reward is zero everywhere, except when reaching the terminal state, when it is the utility.
    - This is not helpful.
- Ideally, we have positive / negative rewards on the states / transitions on the way to goals.
    - Your homework and midterm grades are rewards, the final grade is utility.

Rewards and utilities

- Rewards and utilities can be tied together in several ways
- Eg. the utility is the sum of rewards

$$U = \sum_t R(t)$$

- Utility is discounted rewards

$$U = \sum_t \gamma^t R(t)$$

- Other combinations are possible, but these two have formal, mathematical and computational benefits.

Who decides on the rewards?

- Some rewards are natural: for instance, actual money the agent gets by visiting certain states.
    - Many scenarios have a large reward at the end, basically defining the utility.
- Some rewards can be virtual: we add create them artificially in order to facilitate learning.
- There is a specific technique called **reward shaping** that adds rewards to intermediate states, to make it easier to learn the solution
    - For instance, in a maze, instead of just having a reward at the end, we add rewards every time we get closer to it

Problem: reward hacking

- If we made mistakes in deciding the reward, the learning agent might find ways to exploit it:
  - Collect rewards without actually reaching a prefered state
- Examples:
  - In a boat racing game, the agent might learn to repeatedly circle reward targets without finishing the race.
  - The robot trained to stack blocks might learn to knock blocks on the floor if the reward counts "blocks not touching the table"
- The more complex a reward function, the more likely that a learning agent learns to hack it.

Utilities: building agents and robots

- Note that we can build a perfectly rational agent by behavior specification, without ever representing utilities.
- Historically, it had been difficult to build agents by utility specification
  - But this is changing, as we are moving towards more ML and less hardcoded behaviors
- How do you specify the utilities for a self-driving car?
  - Traffic rules?
  - Optimize time to goal, energy consumption?
  - Safety?

Utilities and AGI

- What should be the utilities of an artificial general intelligence?
- **Alignment problem**: the AGI should share preferences with humanity / smart humans / important humans / **me!**
- Couple of issues:
  - **Can** we specify the utility of humans?
  - Who gets to specify it? Likely differs from person to person.
  - Are we happy with the human utility function? Eg. pleasure seeking behavior?
  - Wouldn't we better of just placing limits on actions? Eg. Asimov's three laws of robotics.
  - Specification gaming
  - Many others...