Lecture 4:
Rational Decisions

Catholijn M. Jonker

Russell & Norvig
Chapter 16 & 17.1-4

Contents of these slides

I. Making Simple Decisions (Chapter 16)

- Rational preferences
- Utilities
- Money
- Decision networks
- Multi-attribute utility
- Value of information

II. Making Complex Decisions (Chapter 17.1-17.4)

- Sequential decision problems
- Value iteration
- Policy iteration

I. MAKING SIMPLE DECISIONS

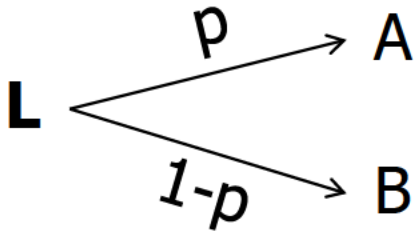
Chapter 16.

Lottery

TU Delft

Preferences

- An agent chooses among prizes (A, B, etc.) and lotteries, i.e., situations with uncertain prizes
- Lottery: **L** = [p, A; (1 – p), B]



- Notation:
  - $A \succ B$ A preferred to B
  - $A \sim B$ indifference between A and B
  - $A \succsim B$ B not preferred to A

Rational preferences

- Idea: preferences of a rational agent must obey constraints.
- Rational preferences ⇒ behavior describable as maximization of expected utility

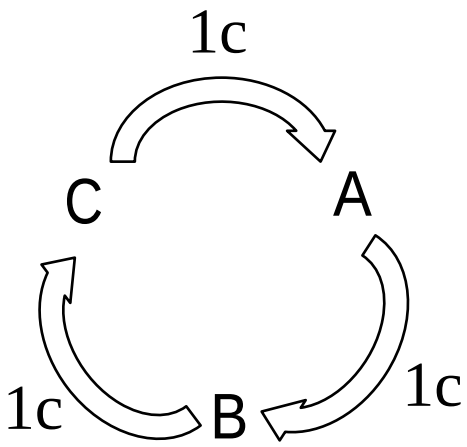
Constraints:

- 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: $A \succ B \succ C \Rightarrow \exists p\ [p, A; 1\text{-}p, C] \sim B$
- Substitutability: $A \sim B \Rightarrow [p, A; 1\text{-}p, C] \sim [p, B; 1\text{-}p, C]$
- Monotonicity: $A \succ B \Rightarrow (p \geq q \Leftrightarrow [p, A; 1\text{-}p, B] \succsim [q, A; 1\text{-}q, B])$

Rational preferences

- Violating the constraints leads to self-evident irrationality
- For example: an agent with intransitive preferences can be induced to give away all its money

- If $B \succ C$, then an agent who has C would pay (say) 1 cent to get B
- If $A \succ B$, then an agent who has B would pay (say) 1 cent to get A
- If $C \succ A$, then an agent who has A would pay (say) 1 cent to get C



Lottery - Utilities

S_n: the state of possessing n€,

Current wealth is k€

EU(accept) = 0.5 * U(S_k) + 0.5 * U($S_{k+3M€}$)

EU(decline) = U($S_{k+1M€}$)

U: first million is worth more than later millions.

Lottery

TU Delft

Lottery - Utilities

Suppose $U(S_k) = 0.5$

$U(S_{k+1M€}) = 0.8$

$U(S_{k+3M€}) = 1$

Then EU(accept) = 0.75

EU(decline) = 0.8

Lottery - Utilities

Suppose $U(S_k) = 0.5$

$U(S_{k+1M€}) = 0.8$

$U(S_{k+3M€}) = 1$

Then EU(accept) = 0.75

EU(decline) = 0.8

Suppose $U(S_k) = 0.8$

$U(S_{k+1M€}) = 0.9$

$U(S_{k+3M€}) = 1$

Then EU(accept) = 0.9

EU(decline) = 0.9

Maximizing expected utility (MEU)

- Theorem (Ramsey, 1931; von Neumann and Morgenstern, 1944): Given preferences satisfying the constraints, there exists a real-valued function U such that:

$$U(A) \geq U(B) \quad \Leftrightarrow \quad A \succsim B$$
$$U([p_1, S_1; \ \ldots \ ; p_n, S_n]) = \Sigma_i \ p_i U(S_i)$$

- MEU principle:

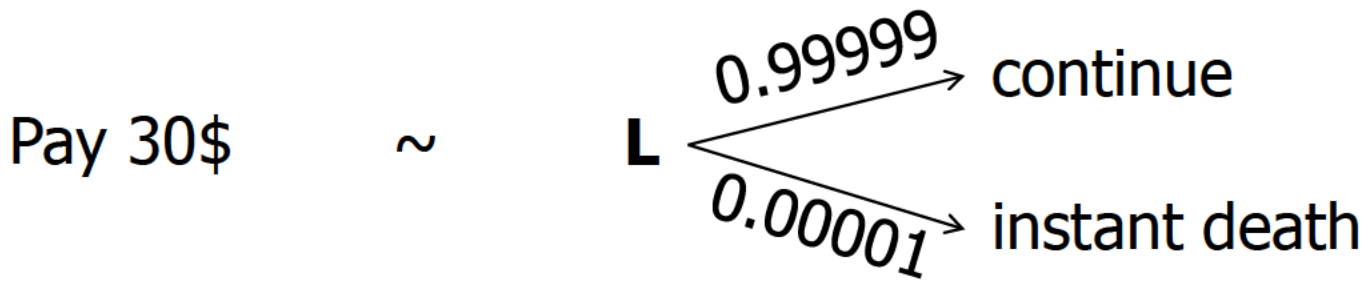
Choose the action that maximizes expected utility!

- Note: an agent can be entirely rational (consistent with MEU) without ever representing or manipulating utilities and probabilities (E.g., a lookup table for perfect tictactoe)

Utilities

- Utilities map states to real numbers. Which numbers?
- Standard approach to assessment of human utilities:

compare a given state A to a standard lottery L_p that has
"best possible prize" $u_\top$ with probability p
"worst possible catastrophe" $u_\bot$ with probability $(1-p)$
adjust lottery probability p until $A \sim L_p$



Utility scales

- Normalized utilities: $u_\top = 1.0, u_\bot = 0.0$
- Note: behavior is invariant w.r.t. +ve linear transformation
- With deterministic prizes only (no lottery choices), only ordinal utility can be determined, i.e., total order on prizes

$$U'(x) = k_1 U(x) + k_2 \quad \text{where } k_1 > 0$$

- Micromorts: one-millionth chance of death

useful for Russian roulette, paying to reduce product risks, etc.

- QALYs: quality-adjusted life years

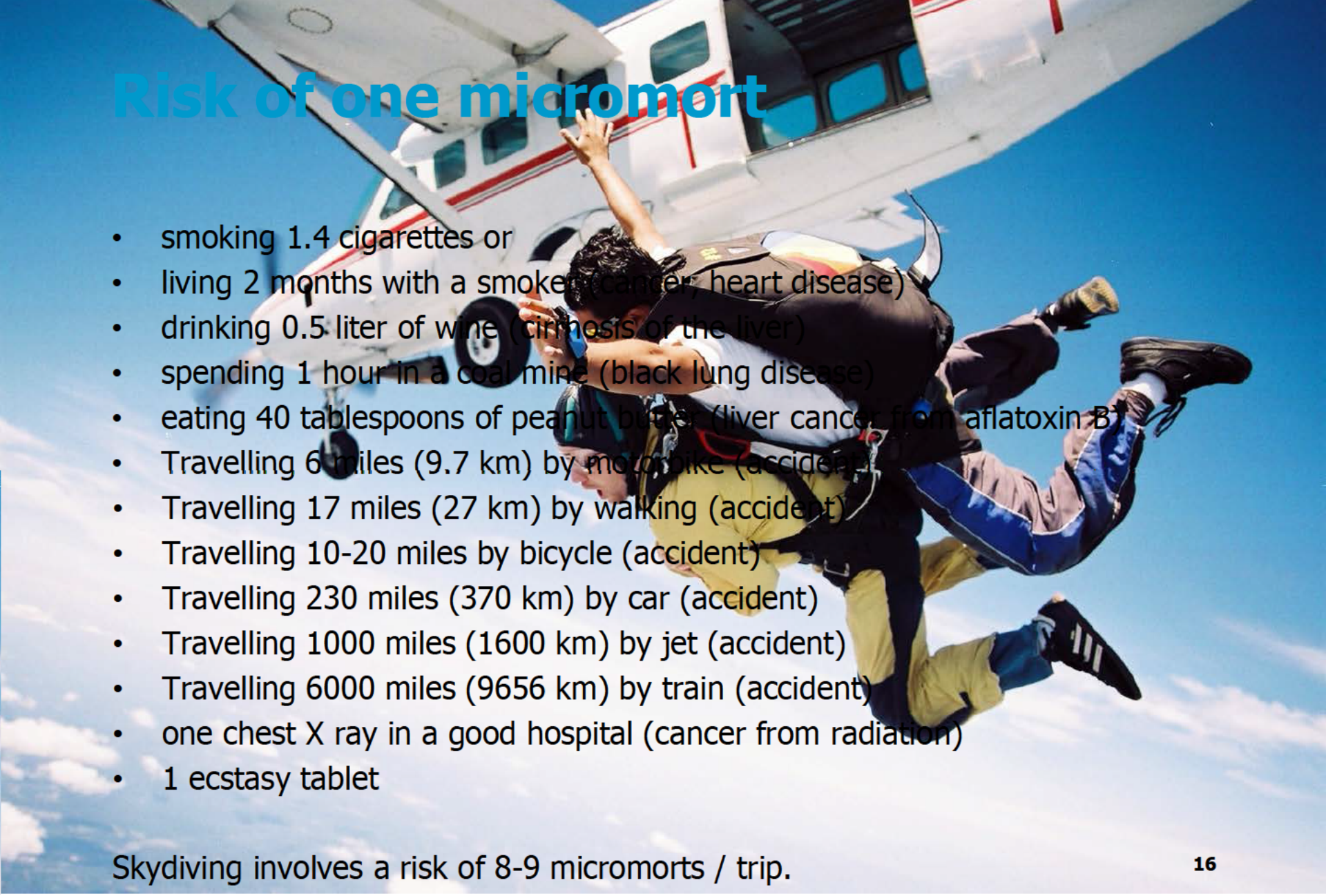
useful for medical decisions involving substantial risk

Micromorts

- A micromort is a unit of risk measuring a one-in-a-million probability of death.
- An application of micromorts is measuring the value that humans place on risk:
  - What is the amount of money one would have to pay a person to get him or her to accept a one-in-a-million chance of death?
  - Or what amount is someone willing to pay to avoid a one-in-a-million chance of death?

- When put thus people claim a high number but when inferred from their day-to-day actions (e.g., how much they are willing to pay for safety features on cars) a typical value is around $20.

Risk of one micromort



- smoking 1.4 cigarettes or
- living 2 months with a smoker (cancer, heart disease)
- drinking 0.5 liter of wine (cirrhosis of the liver)
- spending 1 hour in a coal mine (black lung disease)
- eating 40 tablespoons of peanut butter (liver cancer from aflatoxin B)
- Travelling 6 miles (9.7 km) by motorbike (accident)
- Travelling 17 miles (27 km) by walking (accident)
- Travelling 10-20 miles by bicycle (accident)
- Travelling 230 miles (370 km) by car (accident)
- Travelling 1000 miles (1600 km) by jet (accident)
- Travelling 6000 miles (9656 km) by train (accident)
- one chest X ray in a good hospital (cancer from radiation)
- 1 ecstasy tablet

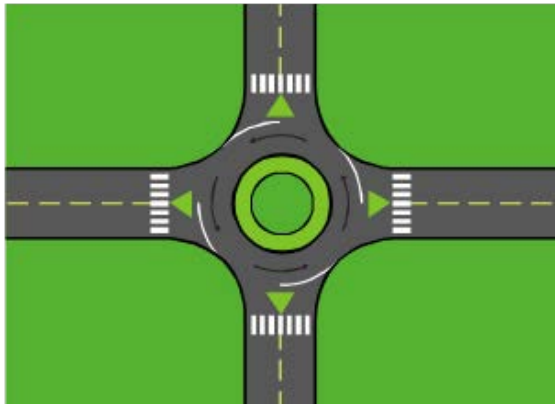
Skydiving involves a risk of 8-9 micromorts / trip.

Running a marathon is 7 micromorts.

Scuba diving involves 5.

Quality Adjusted Life Years (QALYs)

- The quality-adjusted life year (QALY) is a measure of disease burden, including both the quality and the quantity of life lived. It is used in assessing the value for money of a medical intervention.



vs.

Cost of construction

Nr. of deaths on crossing

Quality Adjusted Life Years (QALYs)

- The QALY model requires utility independent, risk neutral, and constant proportional tradeoff behaviour.
- The QALY is based on the number of years of life that would be added by the intervention.
- Each year in perfect health is assigned the value of 1.0 down to a value of 0.0 for death. If the extra years would not be lived in full health, for example if the patient would lose a limb, or be blind or have to use a wheelchair, then the extra life-years are given a value between 0 and 1 to account for this.

QALYs use

- cost-utility analysis : intervention cost / QALYs saved
- Used to allocate healthcare resources
- Controversial: some people will not receive treatment as it is calculated that cost of the intervention is not warranted by the benefit to their quality of life.
- Argument in favor: health care resources are limited, this allocation method is approximately optimal for society, including most patients.

QALY debate about meaning

- Perfect health is hard, if not impossible, to define.
- There are health states worse than death, therefore there should be negative values possible on the health spectrum.
- Measures place disproportionate importance on physical pain or disability over mental health.
- The effects of a patient's health on the quality of life of others (e.g. caregivers or family) do not figure into these calculations.

Twente Airport

Arguments pro:

- Economic impulse to the region
- jobs!

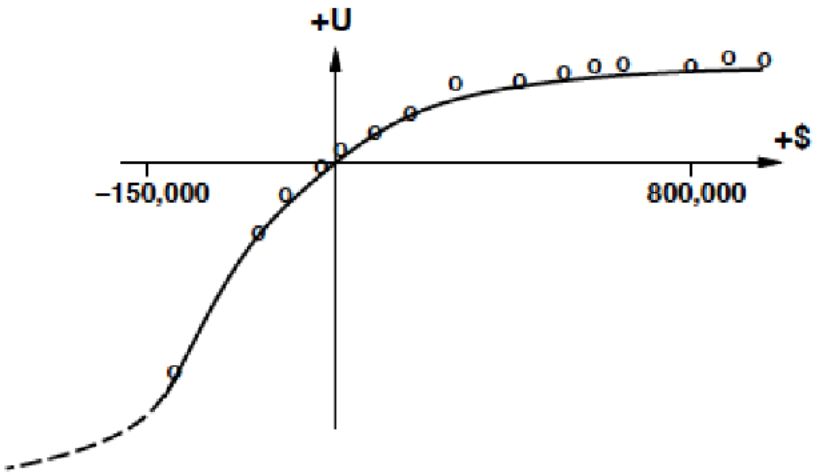
Arguments contra:

- Lot of noise
- Job argument is probably not true
- Will be an empty industry lot
- Tourists can and will fly
  from German airport close by

https://nl.wikipedia.org/wiki/Enschede_Airport_Twente

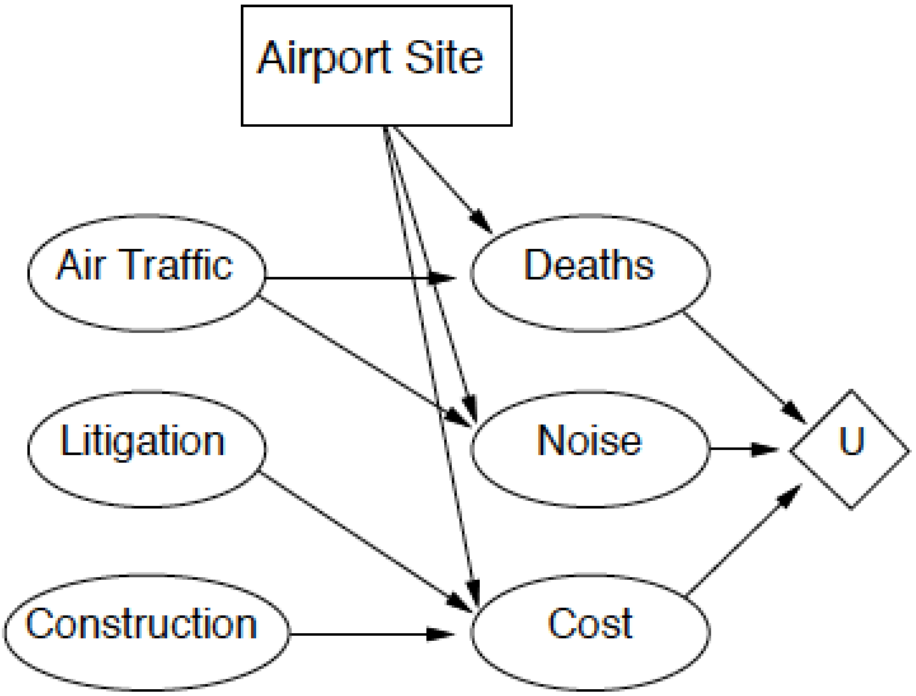
Money

- Money does not behave as a utility function
- Given a lottery L with expected monetary value EMV(L), usually U(L) < U(EMV(L)), i.e., people are risk-averse
- Utility curve: 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, extrapolated with risk-prone behavior:



Decision networks

- Add action nodes and utility nodes to belief networks to enable rational decision making



Algorithm:

- For each value of action node:
  - compute expected value of utility node given action, evidence
- Return MEU action

Multi-attribute utility

- How can we handle utility functions of many variables?
  X1,...,Xn

- E.g., what is U(Deaths, Noise, Cost)?
- How can complex utility functions be assessed from preference behaviour?

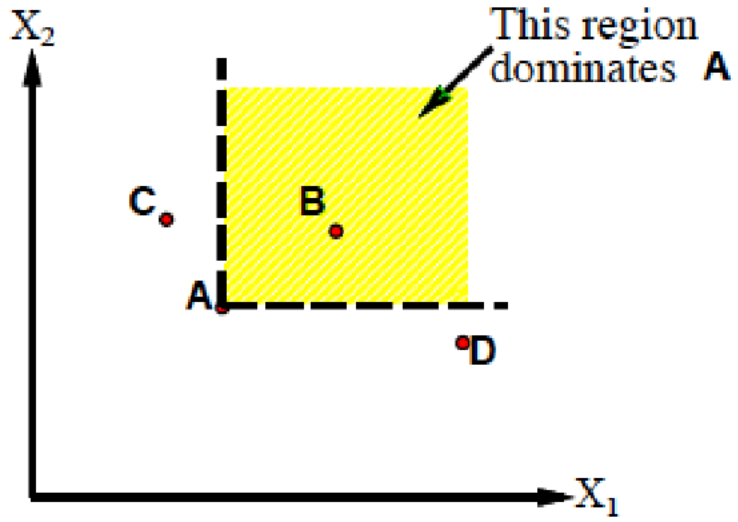
- Idea 1: identify conditions under which decisions can be made without complete identification of U(x1,...,xn)

- Idea 2: identify various types of independence in preferences, and derive consequent canonical forms for U(x1,...,xn)

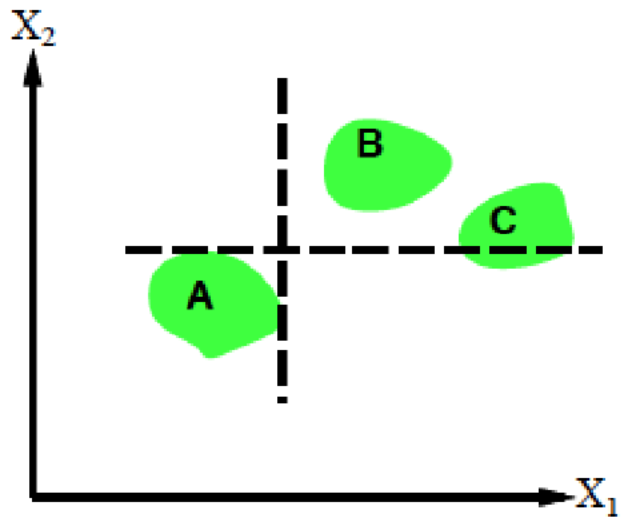
Strict dominance

- Typically define attributes such that U is monotonic in each
- Strict dominance: choice B strictly dominates choice A iff

$\forall i \quad X_i(B) \geq X_i(A)$ (and hence $U(B) \geq U(A)$)



Deterministic attributes



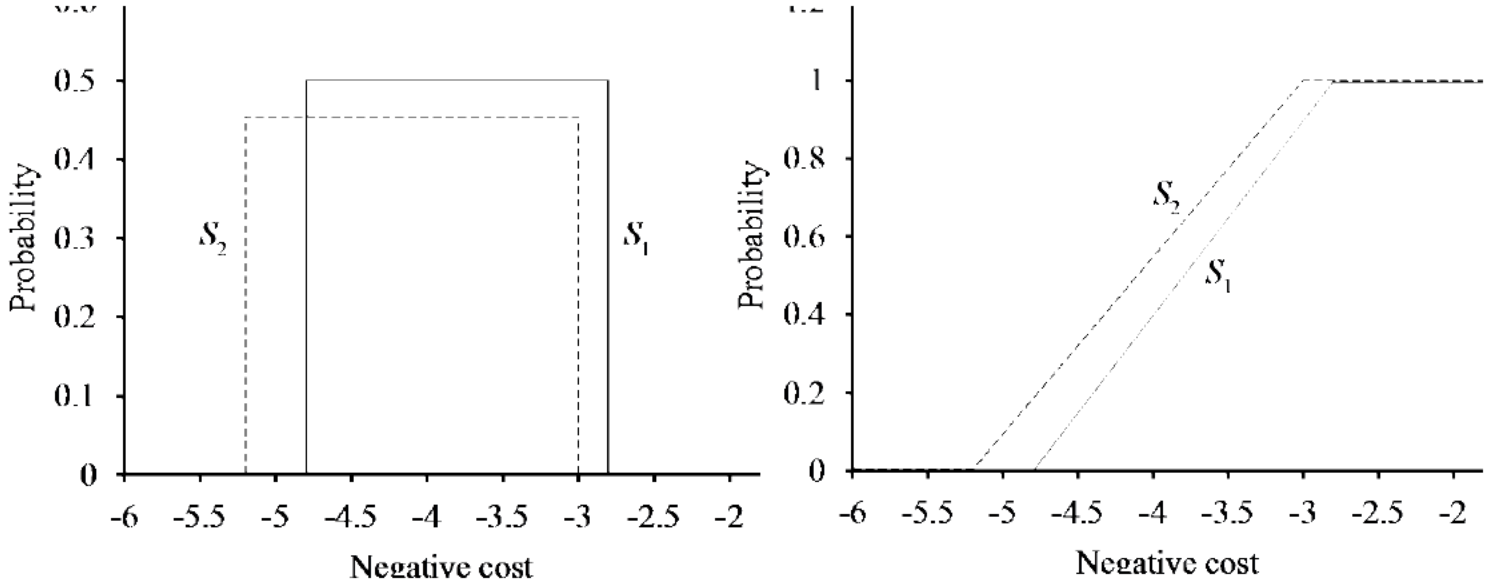
Uncertain attributes

Stochastic dominance

- Distribution p_1 stochastically dominates distribution p_2 iff:

$$\forall x \int_{-\infty}^{x} p_1(t)d(t) \leq \int_{-\infty}^{x} p_2(t)d(t)$$

- x refers to the variable corresponding to the horizontal axis.
- Distributions Cumulative distributions



TU Delft

Stochastic dominance

- Distribution p_1 stochastically dominates distribution p_2 iff:

$$\forall x \int_{-\infty}^{x} p_1(t)d(t) \leq \int_{-\infty}^{x} p_2(t)d(t)$$

- If U is monotonic in x, then A_1 with outcome distribution p_1 stochastically dominates A_2 with outcome distribution p_2:

$$\int_{-\infty}^{\infty} p_1(x)U(x)dx \geq \int_{-\infty}^{\infty} p_2(x)U(x)dx$$

- Multi-attribute case: stochastic dominance on all attributes ⇒ optimal

Stochastic dominance

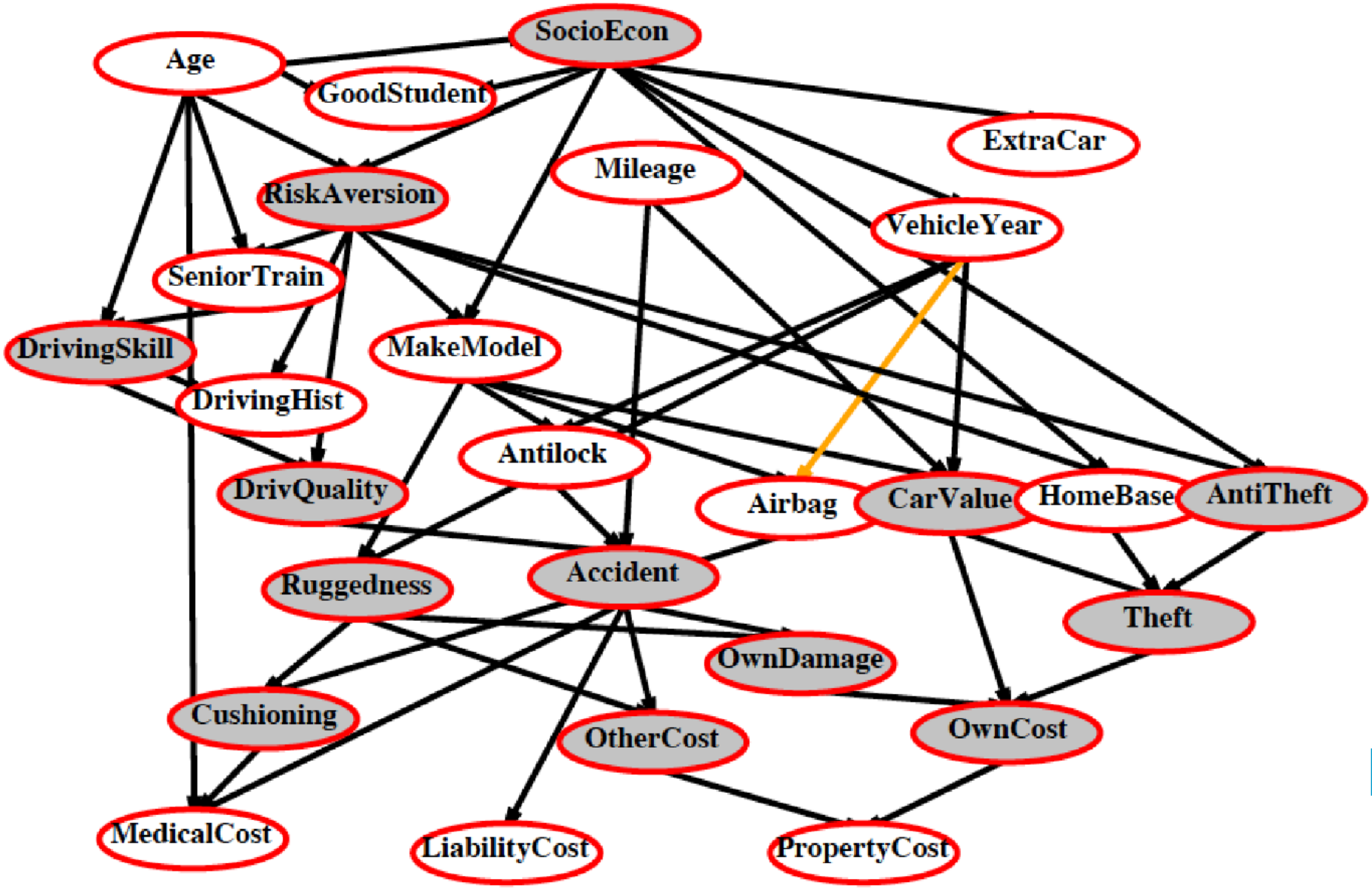
- Stochastic dominance can often be determined without exact distributions using qualitative reasoning
- E.g., construction cost increases with distance from city:

  S1 is closer to the city than S2 ⇒ S1 stochastically dominates S2 on cost
- E.g., injury increases with collision speed

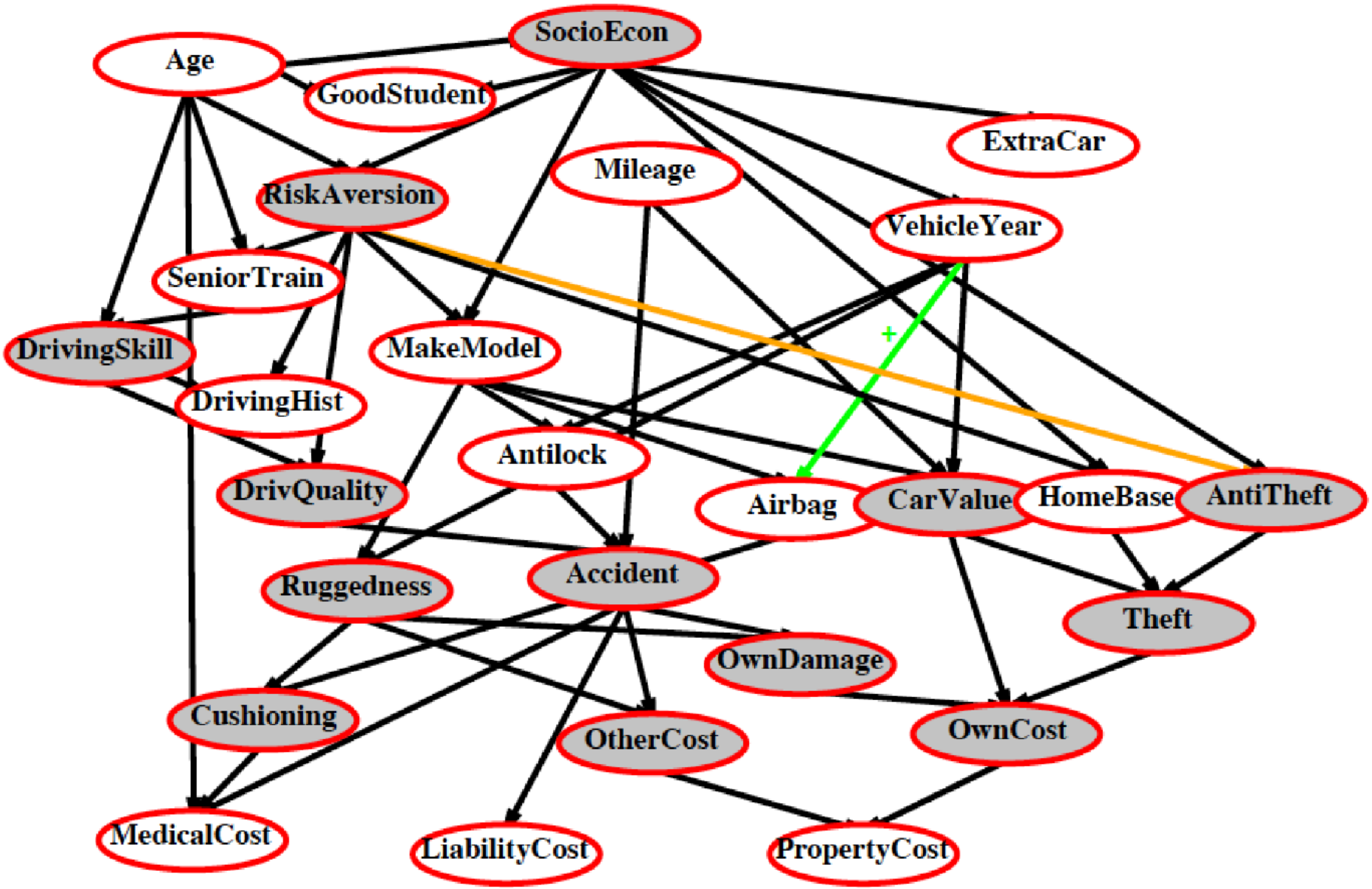
- Can annotate belief networks with stochastic dominance information:
- $X \xrightarrow{+} Y$ (X positively influences Y) means that for every value z of Y's other parents Z: $\forall x1,x2$

  $x1 \geq x2 \Rightarrow P(Y|\ x1, z)$ stochastically dominates $P(Y|\ x2, z)$

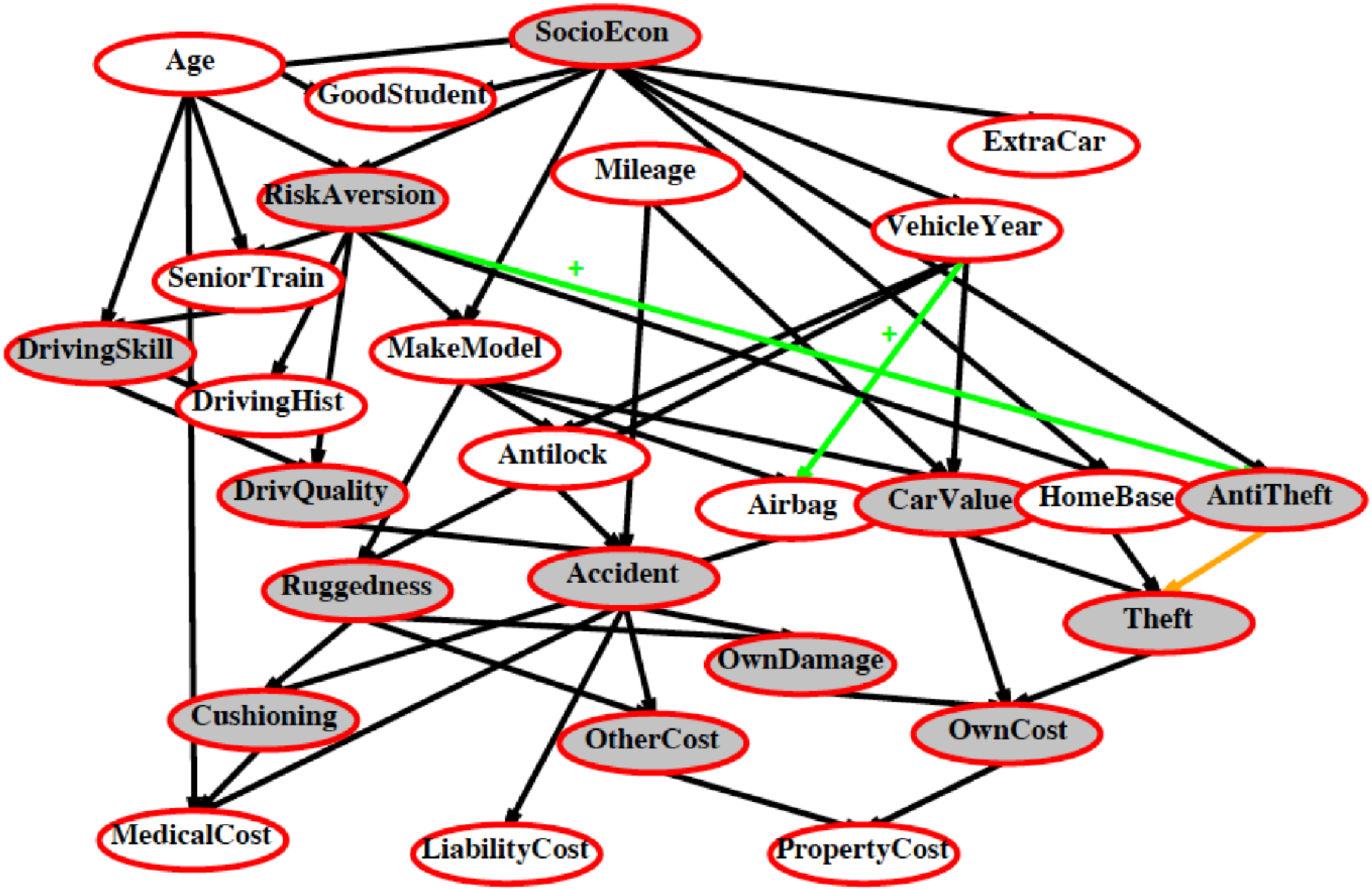
Label the arcs + or -



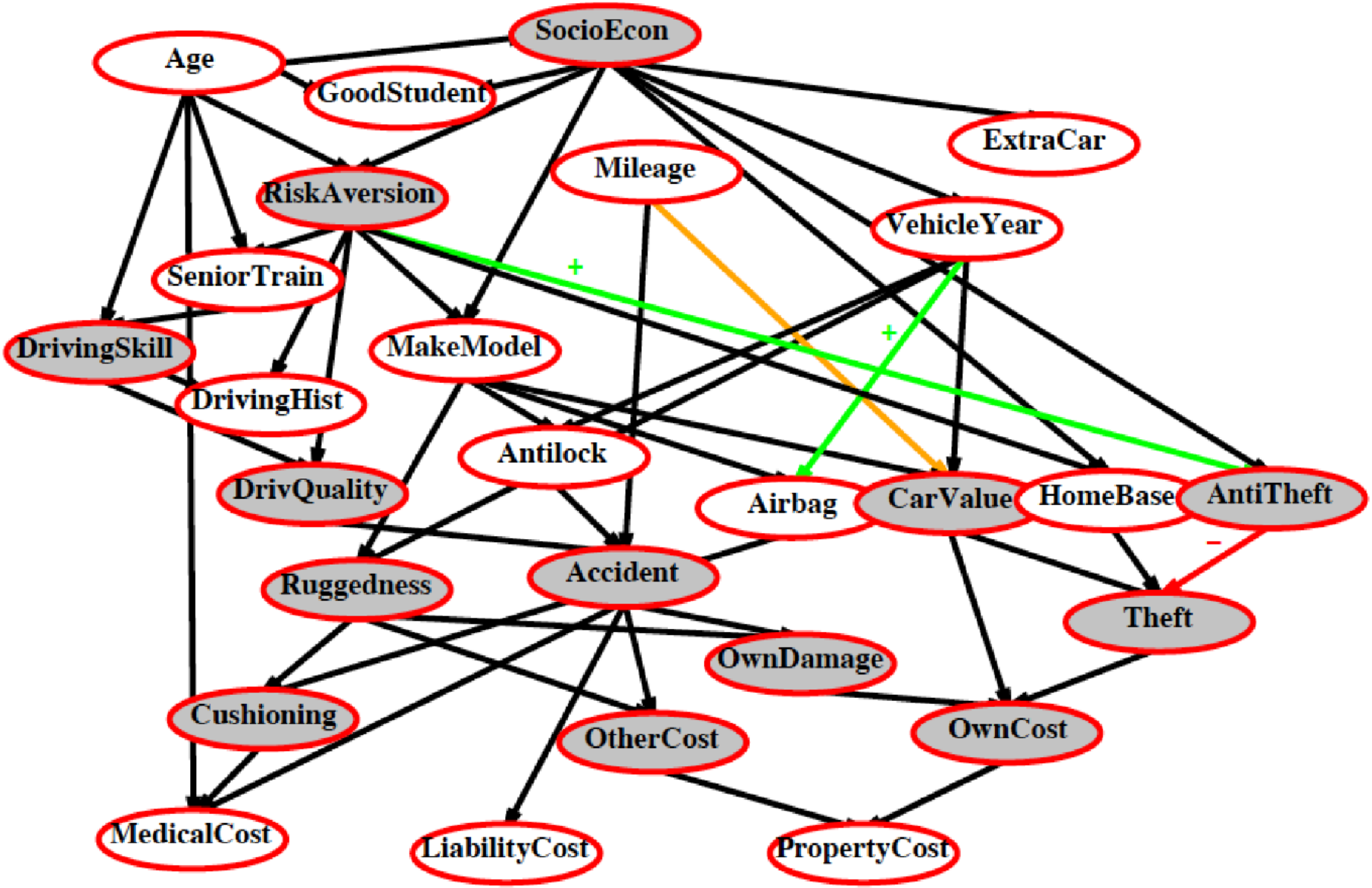
Label the arcs + or -



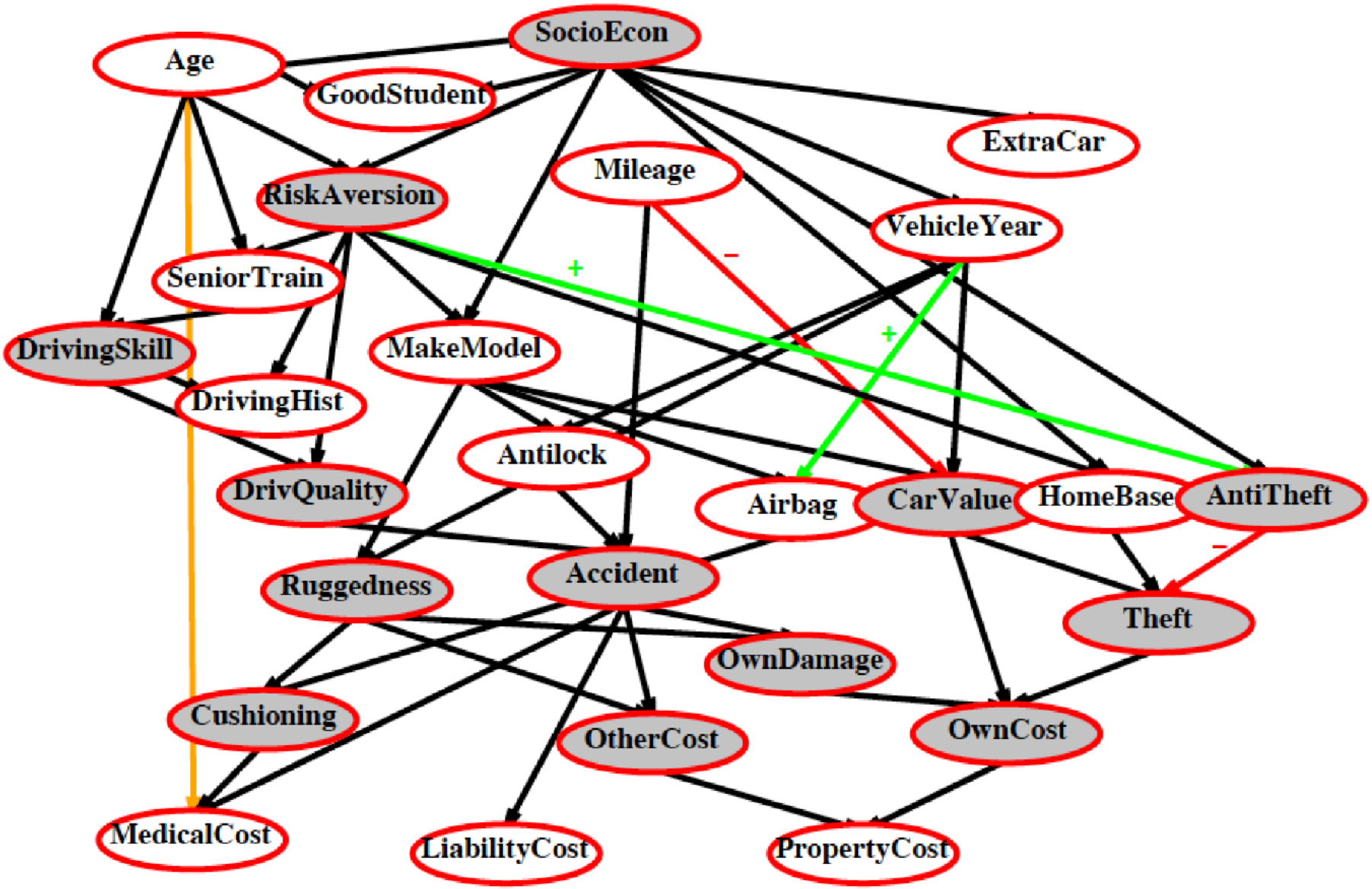
Label the arcs + or -



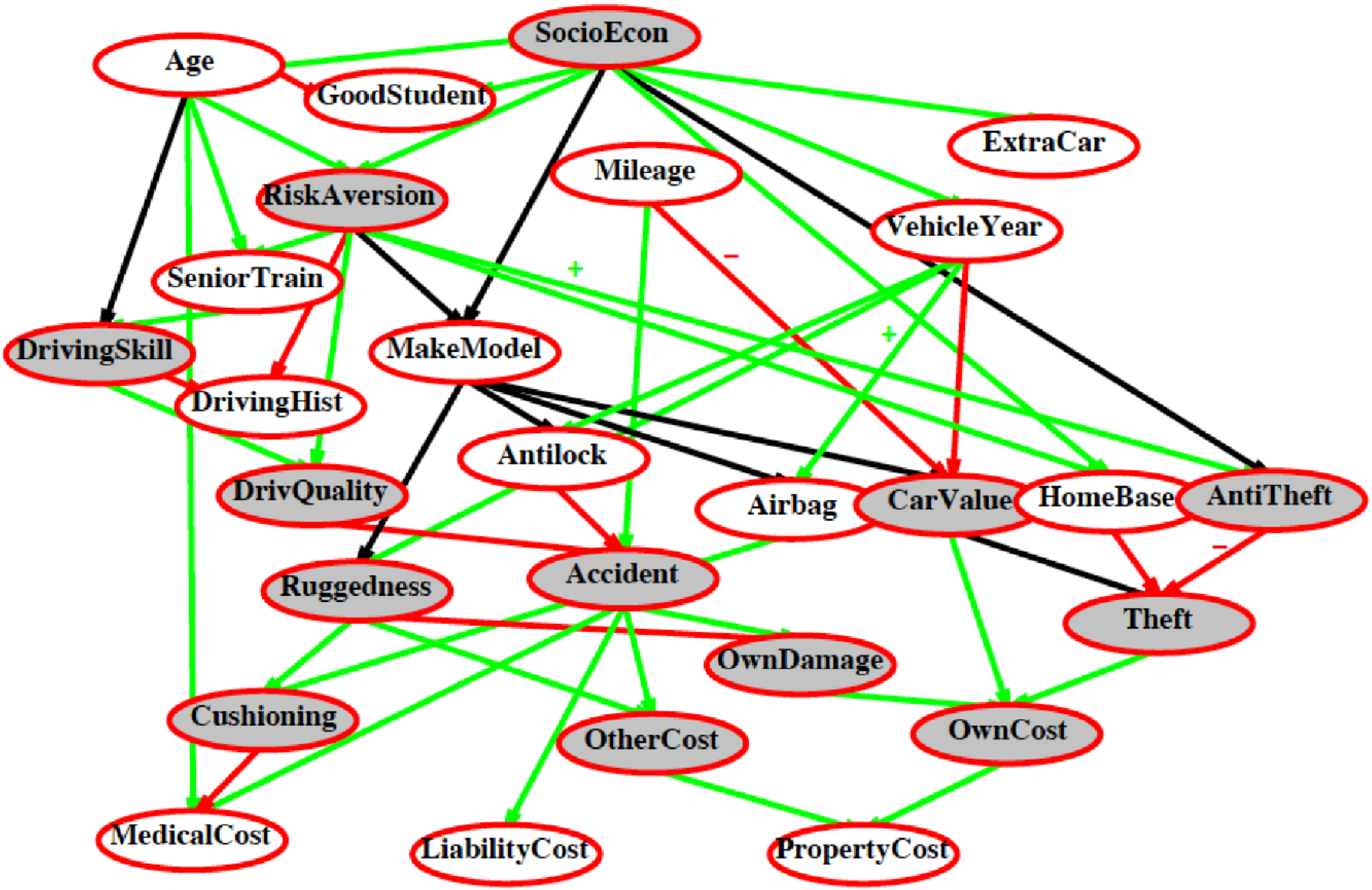
Label the arcs + or -



Label the arcs + or -



Label the arcs + or -



Preference structure: deterministic

- X1 and X2 preferentially independent of X3 iff
  preference between <x1, x2, x3> and <x1', x2', x3>
  does not depend on x3
- E.g., <Noise, Cost, Safety>:
  - <20,000 suffer, $4.6 billion, 0.06 deaths/mpm> vs.
  - <70,000 suffer, $4.2 billion, 0.06 deaths/mpm>
- Theorem (Leontief, 1947): if every pair of attributes is P.I. of its complement, then every subset of attributes is P.I of its complement: mutual P.I..
- Theorem (Debreu, 1960): mutual P.I.

⇒ ∃ additive value function: $V(S) = \sum_i V_i(X_i(S))$

- Hence assess n single-attribute functions; often a good approx.

Preference structure: stochastic

- Need to consider preferences over lotteries:
- X is utility-independent of Y iff
  preferences over lotteries in X do not depend on Y
- That means that if:

[p, (x1, y1); p-1, (x2, y1)] > [p, (x3, y1); p-1, (x4, y1)]
then also [p, (x1, y2); p-1, (x2, y2)] > [p, (x3, y2); p-1, (x4, y2)]

- Mutual U.I.: each subset is U.I of its complement

⇒ ∃multiplicative utility function:

- U = k1U1 + k2U2 + k3U3 + k1k2U1U2 + k2k3U2U3
  + k3k1U3U1 + k1k2k3U1U2U3

Value of information

- Idea: compute value of acquiring each possible piece of evidence
- Can be done directly from decision network
- Example:

There are two different routes through a mountain range: a1, a2.

a1 is a long, straight way, and a2 is a short winding road.

It's winter and snow or ice blockages can be anywhere.

It is possible to obtain satellite reports on the actual state of each road, that would give new expectations.

You carry an injured person. Do you pay for the satellite report?

- Solution: compute expected value of information

  = expected value of best action given the information

  minus expected value of best action without information

General formula for perfect info

Current evidence E, current best action α

Possible action outcomes S_i, potential new evidence E_j

$$EU(\alpha|E) = \max_a \sum_i U(S_i)\ P(S_i|E,a)$$

Suppose we knew $E_j = e_{jk}$, then we would choose $\alpha_{e_{jk}}$ s.t.

$$EU(\alpha_{e_{jk}}|E, E_j = e_{jk}) = \max_a \sum_i U(S_i)\ P(S_i|E,a,E_j = e_{jk})$$

E_j is a random variable whose value is *currently* unknown

$\Rightarrow$ must compute expected gain over all possible values:

$$VPI_E(E_j) = \left(\sum_k P(E_j = e_{jk}|E) EU(\alpha_{e_{jk}}|E, E_j = e_{jk})\right) - EU(\alpha|E)$$

(VPI = value of perfect information)

Properties of VPI

Nonnegative—in expectation, not post hoc

$$\forall j, E \;\; VPI_E(E_j) \geq 0$$

Nonadditive—consider, e.g., obtaining E_j twice

$$VPI_E(E_j, E_k) \neq VPI_E(E_j) + VPI_E(E_k)$$

Order-independent

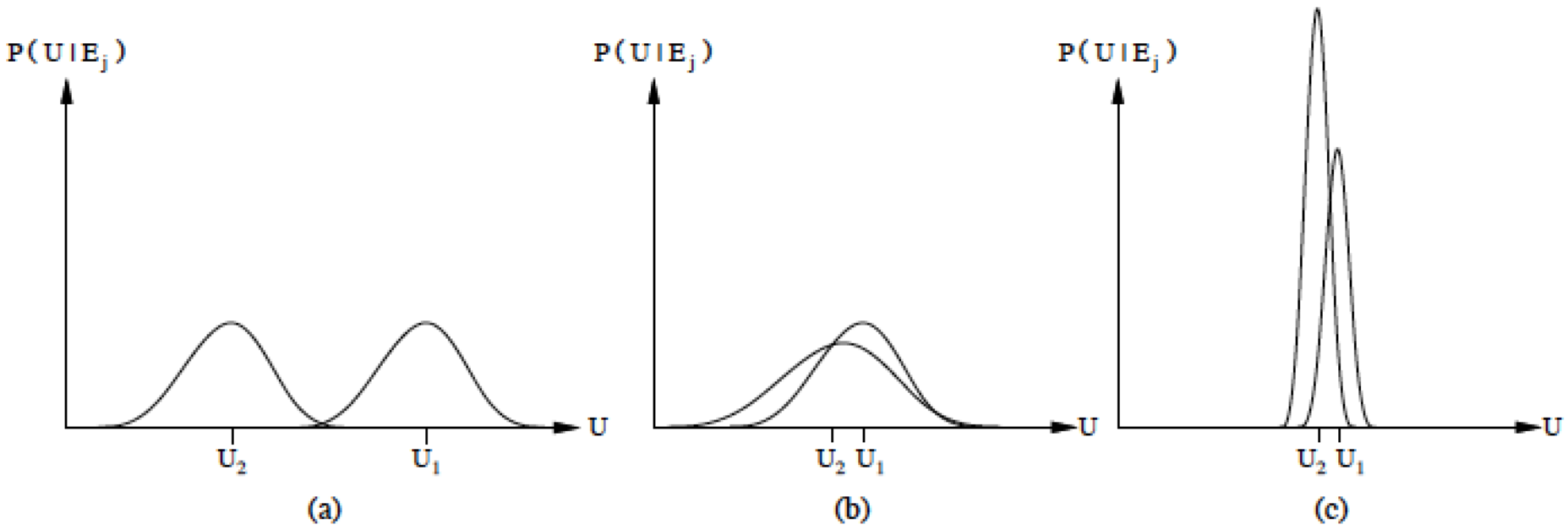
$$VPI_E(E_j, E_k) = VPI_E(E_j) + VPI_{E,E_j}(E_k) = VPI_E(E_k) + VPI_{E,E_k}(E_j)$$

- Note: when more than one piece of evidence can be gathered, maximizing VPI for each to select one is not always optimal

⇒ evidence-gathering becomes a sequential decision problem

Qualitative behavior

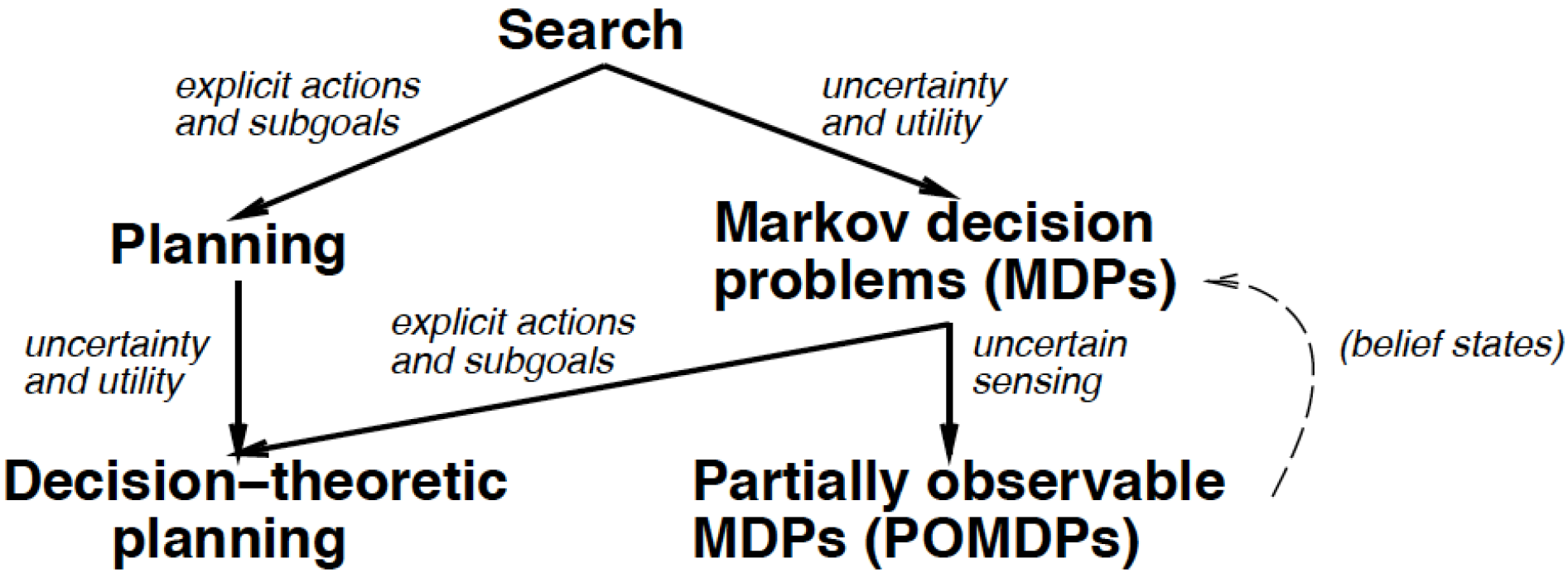
(a) Choice is obvious, information worth little
(b) Choice is nonobvious, information worth a lot
(c) Choice is nonobvious, information worth little



II. MAKING COMPLEX DECISIONS

Chapter 17.1 – 17.4

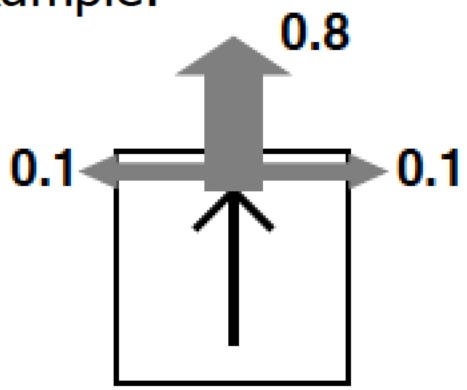
Sequential decision problems

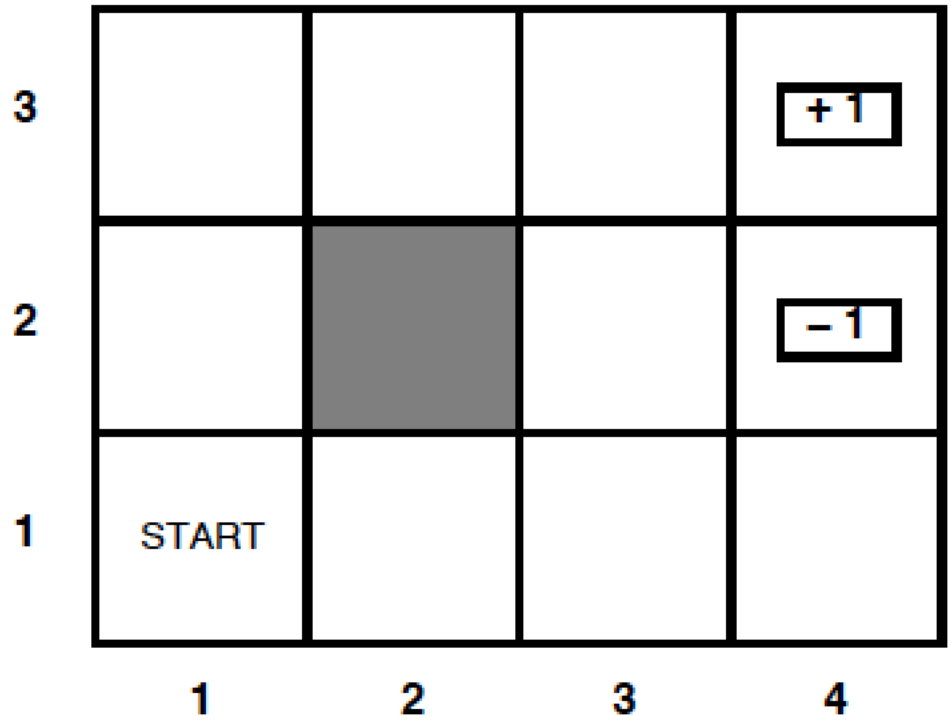


TU Delft

Markov Decision Problem (MDP)

- Example:

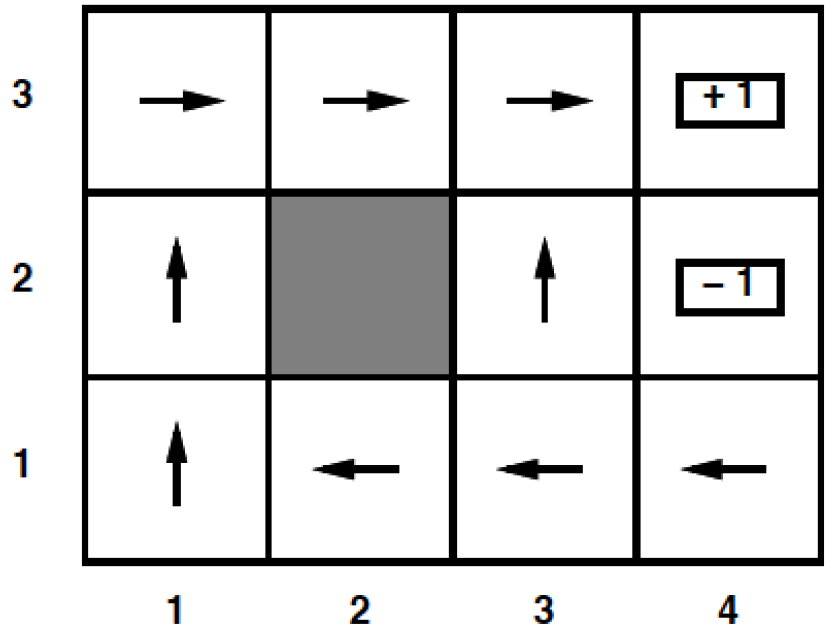




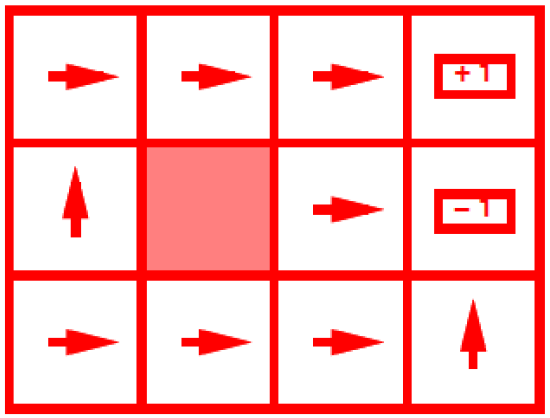
- States: $s \in S$, $a \in A$
- Model: $T(s, a, s') \equiv P(s' \mid s, a)$ = probability that a in s leads to s'
- Reward function: $R(s)$ (or $R(s,a)$, $R(s, a, s')$)
  - -0.04 small penalty for nonterminal states
  - ± 1 for terminal states: +1 instant win, -1 instant death

Solving MDP

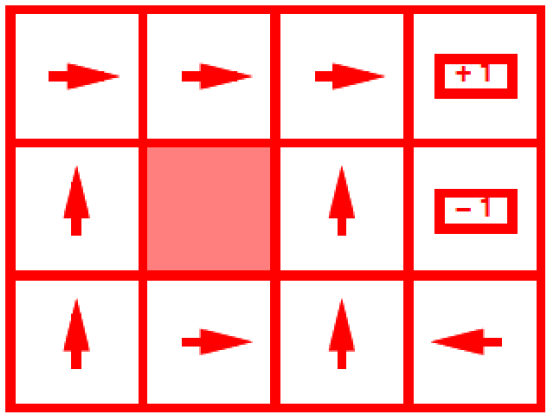
- In search problems, aim is to find an optimal sequence
- In MDPs, aim is to find an optimal policy π(s), i.e., best action for every possible state s (because can't predict where one will end up)
- The optimal policy maximizes the expected sum of rewards
- Optimal policy when state penalty R(s) is -0.04:



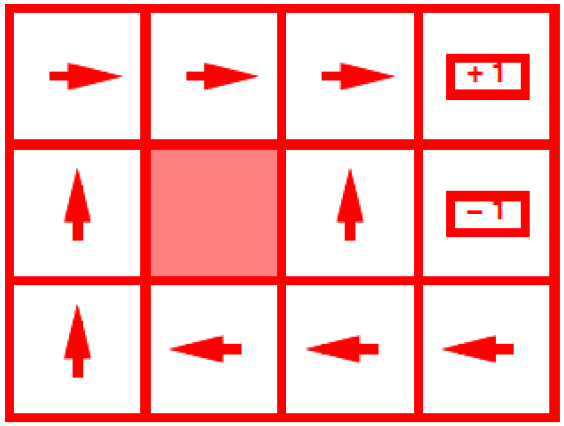
Risk and reward



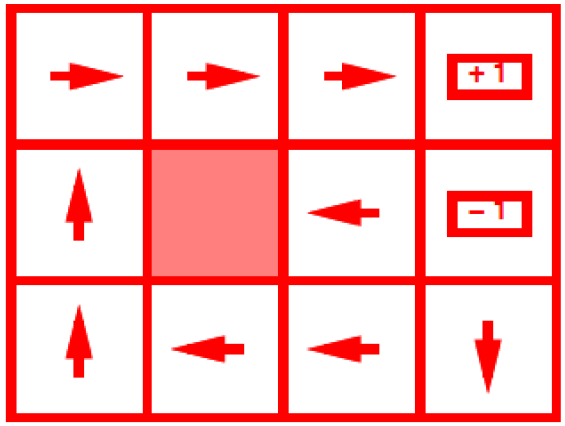
r = [-∞ : -1.6284]



r = [-0.4278 : -0.0850]



r = [-0.0480 : -0.0274]



r=[-0.0218 : 0.000]

TU Delft

Utility of state sequences

- Need to understand preferences between sequences of states
- Typically consider stationary preferences on reward sequences:
- $[r, r0, r1, r2, ...] \succ [r, r0', r1', r2', ...] \Leftrightarrow$

$[r0, r1, r2, ...] \succ [r0', r1', r2', ...]$

Theorem: there are only two ways to combine rewards over time.

1) Additive utility function:

$U([s0, s1, s2, ...]) = R(s0) + R(s1) + R(s2) + ...$

2) Discounted utility function:

$U([s0, s1, s2, ...]) = R(s0) + \gamma R(s1) + \gamma^2 R(s2) + ...$

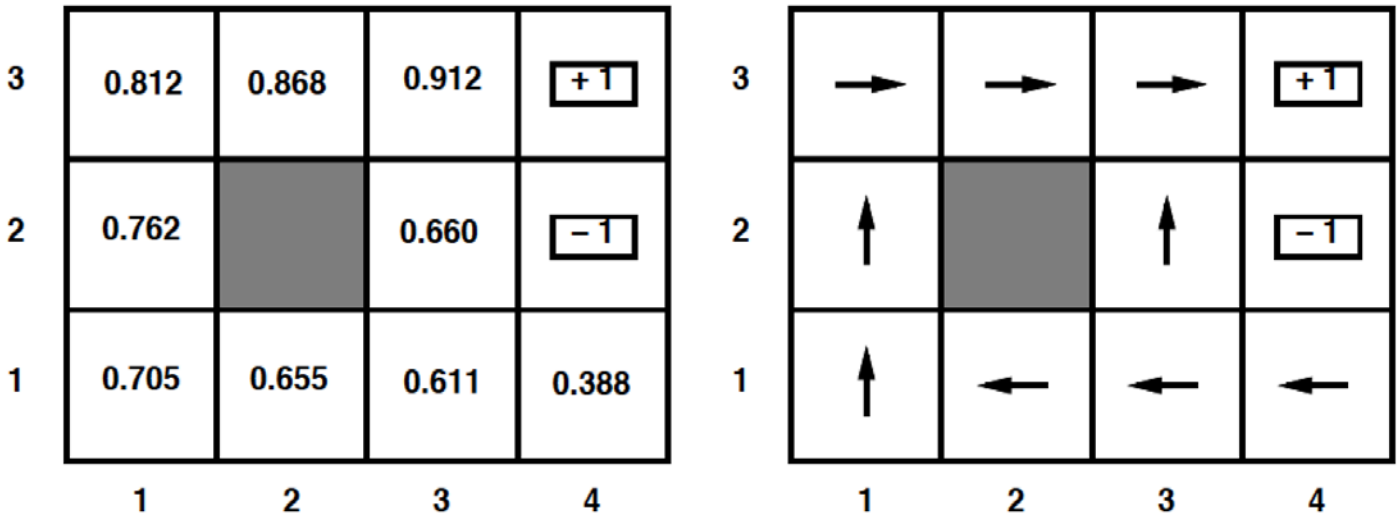
where γ is the discount factor [0-1], preference for current over future rewards

Utility of states

- Utility of a state (a.k.a. its value) is defined to be:

U(s) = expected (discounted) sum of rewards (until termination) assuming optimal actions

- Given the utilities of the states, choosing the best action is just MEU: maximize the expected utility of the immediate successors



Utility of states

Problem: infinite lifetimes ⇒ additive utilities are infinite

1) Finite horizon: termination at a fixed time T
   ⇒ nonstationary policy: π(s) depends on time left
2) Absorbing state(s): w/ prob. 1, agent eventually “dies" for any π
   ⇒ expected utility of every state is finite
3) Discounting: assuming γ < 1, R(s) ≤ Rmax

$$U([s_0, \ldots s_\infty]) = \sum_{t=0}^{\infty} \gamma^t R(s_t) \leq R_{\max}/(1-\gamma)$$

   Smaller γ ⇒ shorter horizon

4) Maximize system gain = average reward per time step

Theorem: optimal policy $\pi^*(s)$ has constant gain after initial transient

- E.g., taxi driver's daily scheme cruising for passengers
- E.g., assembly line has an initial transient period

Dynamic Programming
The Bellman equation

- Definition of utility of states leads to a simple relationship among utilities of neighboring states:
- expected sum of rewards = current reward + expected sum of rewards after taking best action
- Bellman equation (1957):

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

$$\begin{aligned} U(1,1) = & -0.04 \\ & + \gamma \max\{0.8U(1,2) + 0.1U(2,1) + 0.1U(1,1), && \textit{up} \\ & \qquad 0.9U(1,1) + 0.1U(1,2) && \textit{left} \\ & \qquad 0.9U(1,1) + 0.1U(2,1) && \textit{down} \\ & \qquad 0.8U(2,1) + 0.1U(1,2) + 0.1U(1,1)\} && \textit{right} \end{aligned}$$

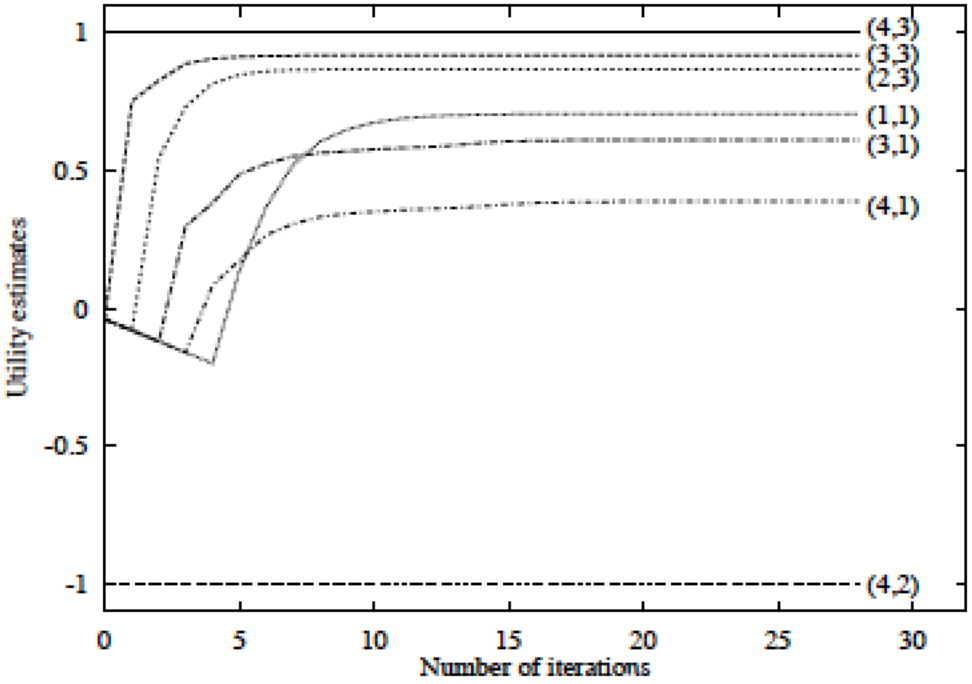
- One equation per state = n nonlinear equations in n unknowns

Value iteration algorithm

Idea: Start with arbitrary utility values

- Update to make them locally consistent with Bellman eqn.
- Everywhere locally consistent ⇒ global optimality
- Repeat for every s simultaneously until "no change"

$$U(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} U(s') T(s, a, s') \qquad \text{for all } s$$



TU Delft

Convergence

- Define the max-norm: ||U|| = $\max_s$|U(s)|,
- So ||U - V|| = maximum difference between U and V
- Let U^t and U^{t+1} be successive approximations to the true utility U

Theorem: For any two approximations U^t and V^t:

$$||U^{t+1} - V^{t+1}|| \leq \gamma\,||U^t - V^t||$$

- I.e., any distinct approximations must get closer to each other
- in particular, any approximation must get closer to the true U
- and value iteration converges to a unique, stable, optimal solution

Theorem: if $||U^{t+1} - U^t|| < \epsilon$, then $||U^{t+1} - U|| < 2\epsilon\gamma/(1-\gamma)$

- I.e., once the change in U^t becomes small, we are almost done.
- MEU policy using U^t may be optimal long before convergence of values

Policy iteration

- Howard, 1960: search for optimal policy and utility values simultaneously
- Algorithm

  π ← an arbitrary initial policy

  repeat until no change in π

    compute utilities given π

    update π as if utilities were correct (i.e., local MEU)

- To compute utilities given a fixed π (value determination):

  $U(s) = R(s) + \gamma \sum_{s'} U(s') T(s, \pi(s), s')$ for all s

- i.e., n simultaneous linear equations in n unknowns, solve in $O(n^3)$

Modified policy iteration

- Policy iteration often converges in few iterations, but each is expensive
- Idea: use a few steps of value iteration (but with π fixed)
  starting from the value function produced the last time
  to produce an approximate value determination step.
- Often converges much faster than pure VI or PI
- Leads to much more general algorithms where Bellman value updates and Howard policy updates can be performed locally in any order
- Reinforcement learning algorithms operate by performing such updates based on the observed transitions made in an initially unknown environment

Partial observability

- POMDP has an observation model O(s, e) defining the probability that the agent obtains evidence e when in state s
- Agent does not know which state it is in

⇒ makes no sense to talk about policy π(s)!!

Theorem (Astrom, 1965):

the optimal policy in a POMDP is a function π(b)

where b is the belief state (probability distribution over states)

- Can convert a POMDP into an MDP in belief-state space, where

T(b, a, b') is the probability that the new belief state is b'

given that the current belief state is b and the agent does a.

- I.e., essentially a filtering update step

Partial observability

- Solutions automatically include information-gathering behavior

- If there are n states, b is an n-dimensional real-valued vector

⇒ solving POMDPs is very (actually, PSPACE-) hard!

- The real world is a POMDP (with initially unknown T and O)