

One-shot Decision-making

Reinforcement Learning and
Planning under Uncertainty

ANU COMP6460/4640, Sem 2, 2008

Scott Sanner

NICTA / RSISE

First.Last@nicta.com.au

Quick Review

Importance of Modeling

- If you remember one thing from this course...

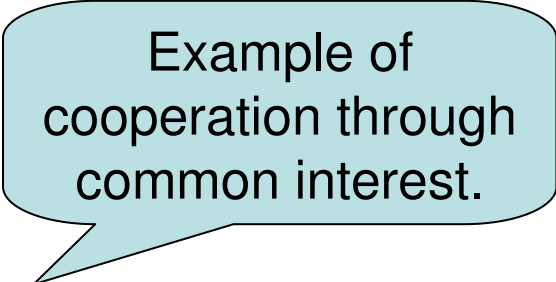
- Formal models
- Formal models
- Formal models
- Formal models

Debdeep will now demonstrate importance of relevant and irrelevant attributes in modelling.

- It just cannot be repeated enough 😊

Models of Decision-making

- To model decision-making, define $\langle \mathbf{S}, \mathbf{O}, \mathbf{A}, \mathbf{Z}, \mathbf{T}, \mathbf{R} \rangle$ and...
 - Observability?
 - Number of agents / actions?
 - Under whose control?
 - Transition distribution
 - Probabilistic (Markovian?, stationary?)
 - Possibilistic (strict uncertainty)
 - Reward model dictates:
 - Cooperative – common / same reward?
 - Adversarial – zero sum?
 - Self-interested – general sum?
 - Objective
 - Horizon (one-shot, sequential)
 - Average vs. discounted
 - Model-based or model-free?
- Once model formalized... can talk about solutions
 - Properties, efficient ways to solve



Example of cooperation through common interest.

Goals, Preferences, and Rewards

Uninterested students
may commence napping.

Goals

- The most easily definable reward
 - Assign any reward \mathbf{R} s.t. $R(\text{win}) > R(\text{lose})$
 - Easily formalized in decision-theoretic context
 - Even in sequential case
(where specify how to trade off rewards over time)

e.g., winning in Go.

- But what if you have multiple goals?
 - How to specify preferences
 - If not all simultaneously achievable
 - (Subset) achieved at different points in time

Reward = Monetary Value?

- Idea: value everything in monetary unit
 - Do we really maximize monetary wealth?
 - 0.1 chance of \$10
vs. 0.000001 chance of \$1,000,000
 - Maybe log transform to account for diminishing returns?
 - Or valuation of winning sensation?
 - Or formalization of risk adversity? (for another day)
 - Caveat: expectations of utility may not hold...
 - But in stochastic setting, we often want to *maximize expected utility!*

Monetary Valuations?

- Imagine previous slide does not exist 😊
 - So assume we maximize expected monetary value
- But still hard to precisely value objects
 - Price of barrel of oil
 - Only a range on valuation relative to buyer / seller
 - price \geq seller's true valuation
 - price \leq buyer's true valuation

Market only recently realized buyer's true valuation much higher.

Preferences and Reward

- Even harder to value non-monetary preferences
 - Apples vs. oranges
 - Box of oranges vs. DVD player
 - Pizza vs. KFC
- But we do have preferences
 - And indifference as equal preference
- So let's forget monetary value temporarily
 - And just specify preferences

Decision / Utility Theory

- Individuals that decide rationally under risk are called “expected utility maximizers”
- Brief History
 - Foundations originated in 17th century with probability theory
 - Bernoulli, de Fermat, Pascal, Huyges
 - Modern theory ~ 1940’s
 - Discrete outcome formalization by von Neumann and Morgenstern
 - Later ~ 1970
 - Fishburn strengthens axioms to support infinite / continuous outcomes
- Use notion of *lottery*
 - *Lottery*: set of outcomes, each with associated probability
 - Then we have preferences over lotteries = distributions!

Useful because all comparisons over prob. w.r.t. *same* outcome set!

Axioms of Preference: Informal

- von Neumann and Morgenstern propose four (widely accepted) axioms of preference:
 - A, B, C are lotteries
 - can always express over same outcomes
 - $A \succeq B$ means “A preferred to B (but not strictly)”
 - $A \sim B$ means “A indifferent to B”
1. **Completeness:** For all A, B, $A \succeq B$ or $B \succeq A$.
 2. **Transitivity:** For all A, B, C, $A \succeq B$ and $B \succeq C \rightarrow A \succeq C$.
 3. **Independence:** For all A, B, $A \succeq B$ and $t \in [0, 1]$
 $\rightarrow tA + (1 - t)C \succeq tB + (1 - t)C$.
 4. **Continuity:** For all A, B, C, $A \succeq B \succeq C$
 $\rightarrow \exists p \in [0, 1]$ such that $B \sim pA + (1 - p)C$.

Expected Utility: Informal

- If preference axioms 1-4 hold for \succeq
 - There exists a utility function $U: \text{Lottery} \rightarrow \text{Real}$
 - For all A, B , $A \succeq B \leftrightarrow U(A) \geq U(B)$
 - Holds for all positive affine transforms of U
- “Expected utility hypothesis”:
 - Lottery utility can be represented as expectation over elementary utility function on *outcomes* (not lotteries)
 - We can now assign a real value to each outcome
 - where values are consistent with preferences, axioms
 - Then maximizing expected utility respects preferences!

Summary

Uninterested students
should wake up.

- In this course...we *maximize expected utility*
 - We assume reward **R**: outcome → Real
 - When outcomes are stochastic (i.e., occur with probabilities)
 - We assume rational agents *maximize expected utility*
- We can make these assumptions because...
 - We start with
 - Goals / preferences and 4 axioms
 - Can specify an **R** that respects prefs / axioms
 - Can actually elicit directly through standard gamble queries (see Simon French, *Decision Theory*, 1986)
- Then our math simplifies considerably 😊

One-shot Decision Making

(Let the models begin)

One-shot Decision Making

- $\langle \mathbf{S}, \mathbf{O}, \mathbf{A}, \mathbf{Z}, \mathbf{T}, \mathbf{R} \rangle$

- $s \in \mathbf{S}$: {out/in Pizza Hut, out/in KFC, at Large Billboard}

- $o \in \mathbf{O}$: {see hut-shaped roof, see picture of the Colonel}

- If fully observable, we ignore \mathbf{O}

- $a \in \mathbf{A}$: {get Pizza, get KFC}

- $\mathbf{Z}: \mathbf{S} \times \mathbf{O} \rightarrow [0,1]$

- Use $\mathbf{P}(o|s)$ to represent \mathbf{Z}

- If fully observable, we ignore \mathbf{Z}

- $\mathbf{T}: \mathbf{S} \times \mathbf{A} \times \mathbf{S} \rightarrow [0,1]$

- Use $\mathbf{P}(s'|s,a)$... (note: assuming Markovian)

- $\mathbf{R}: \mathbf{S} \times \mathbf{A} \times \mathbf{S} \rightarrow \text{Real}$

- Use $\mathbf{R}(\bullet, \bullet, \text{inPH})=10$, $\mathbf{R}(\bullet, \bullet, \text{inKFC})=8$, else $\mathbf{R}(\bullet, \bullet, \bullet)=0$



- Objective specific to each problem, but horizon = 1

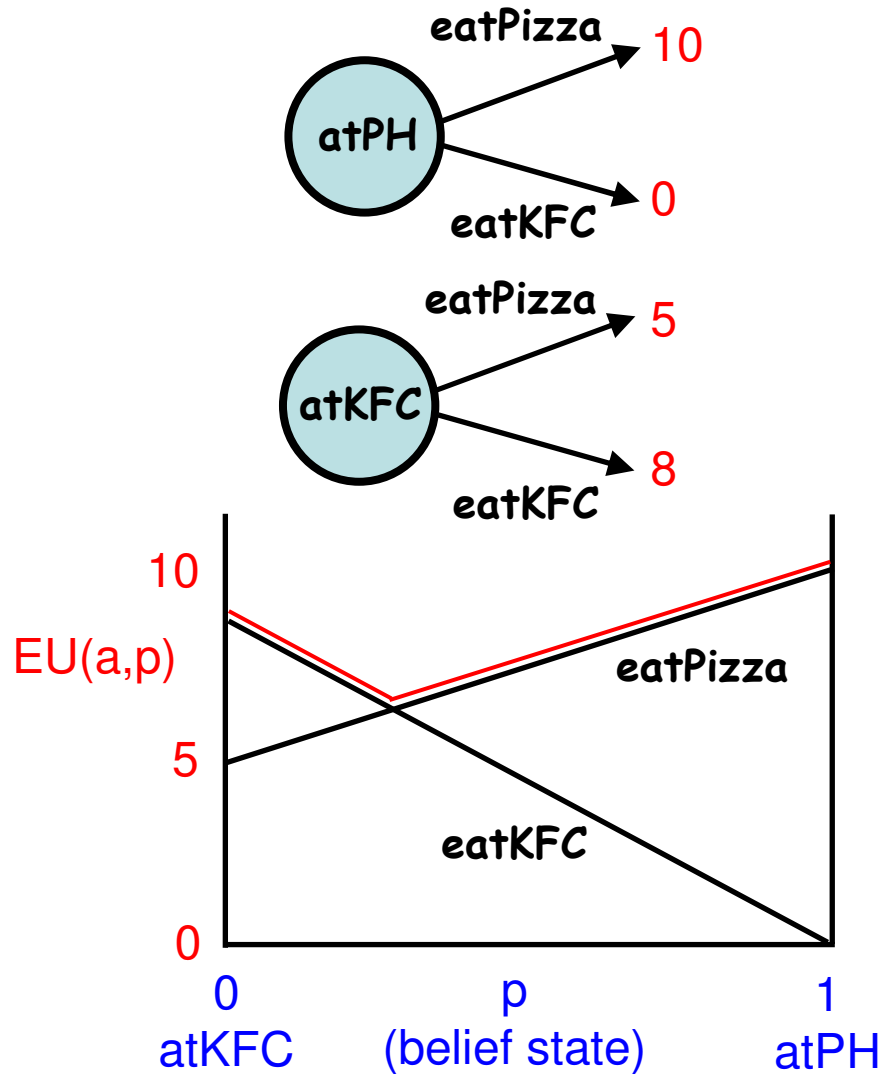
Deterministic, Fully Observable

- Deterministic
 - Means transition function probabilities $\in \{0,1\}$
 - Informally, only get in restaurant if previously in / at restaurant and get appropriate food, e.g.,
 - $P(\text{inPH} \mid \text{outPH}, \text{getPizza}) = 1$
 - $P(\text{inPH} \mid \text{inPH}, \text{getPizza}) = 1$
 - $P(\text{inPH} \mid \text{outPH}, \text{getKFC}) = 0$
 - ...
 - Objective: maximize one-shot expected reward
 - Value: $V(s) = \max_a E_{s' \sim P(s'|a,s)}[R(s,a,s')]$
 - Policy: $\pi(s) = \operatorname{argmax}_a E_{s' \sim P(s'|a,s)}[R(s,a,s')]$
 - Note: trivial expectation because $P(s'|a,s)$ deterministic
- What are the optimal values and policy?

Stochastic, Fully Observable

- Stochastic transition
 - Now have transition function probabilities $\in [0,1]$
 - Informally, only get in restaurant with probability p , e.g.,
 - $P(\text{inPH} \mid \text{outPH}, \text{getPizza}) = p$
 - $P(\text{inPH} \mid \text{inPH}, \text{getPizza}) = 1$
 - $P(\text{inPH} \mid \text{outPH}, \text{getKFC}) = 0$
 - ...
 - Objective: maximize one-shot expected reward
 - Value: $V(s) = \max_a E_{s' \sim P(s'|a,s)}[R(s,a,s')]$
 - Policy: $\pi(s) = \operatorname{argmax}_a E_{s' \sim P(s'|a,s)}[R(s,a,s')]$
- What are the optimal values and policy?
 - Same policy, but some values are different
 - differing transition probabilities lead to differing expectations

Stochastic, Partially Observable



- Redefine $R(s,a)$
 - $R(\text{atPH}, \text{eatPizza})=10$,
 - $R(\text{atKFC}, \text{eatKFC})=8$,
 - $R(\text{atKFC}, \text{eatPizza})=5$
(KFC is trying out a new item)
 - else $R(\bullet, \bullet)=0$
- Observation Function Z
 - Assume a hut-shaped roof seen
 - Define observation probabilities $Z: P(s|o=\text{hut-shaped roof})\dots$
 - $p = P(\text{atPH} \mid \text{hut-shaped roof})$
 - $1-p = P(\text{atKFC} \mid \text{hut-shaped roof})$
- Objective:
 - Expected utility of a given o :
 $EU(a,o) = E_{s \sim P(s|o)} [R(s,a)]$
 - Policy: $\pi(o) = \operatorname{argmax}_a EU(a,o)$
 - policy is function of observation!

Deterministic, Multiagent

- Multiagent case
 - $a \in \mathbf{A} = \mathbf{A}_{\text{Luke}} \times \mathbf{A}_{\text{Tor}}$
 - $a_L \in \mathbf{A}_{\text{Luke}} = \{\text{get Pizza, get KFC}\}$
 - $a_T \in \mathbf{A}_{\text{Tor}} = \{\text{get Pizza, get KFC}\}$
- For simplicity
 - Assume \mathbf{S} only has one state (so ignore states s, s')
 - Possible to get pizza, KFC from this state
 - We'll ignore transition function \mathbf{T} since only one state
 - Would be strict (possibilistic) uncertainty if did model \mathbf{T}
 - We define $R_{a_L}(a_L, a_T)$ and $R_{a_T}(a_L, a_T)$
 - Different reward for Luke, Tor
 - Luke prefers pizza, Tor prefers KFC, get nothing if disagree
- Objective (general sum)
 - Each agent could deterministically maximize over worst outcome
 - $a_L = \operatorname{argmax}_{a_L} \min_{a_T} R_{a_L}(a_L, a_T)$; $a_T = \operatorname{argmax}_{a_T} \min_{a_L} R_{a_L}(a_L, a_T)$
 - Are purely deterministic strategies always optimal? No.
 - Can do often better by maximin over *randomized* strategies

Deterministic, Multiagent

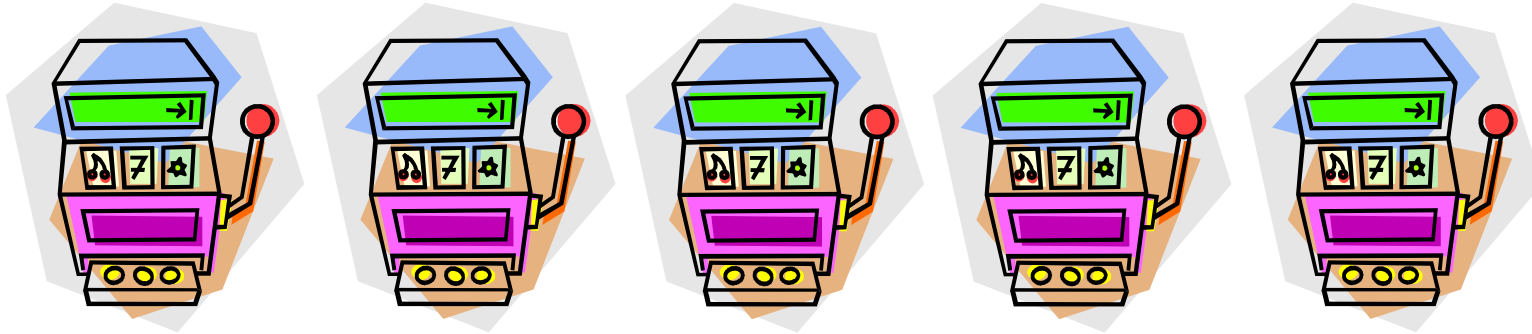
- Normal form game description:

	Tor eatPizza	Tor eatKFC
Luke eatPizza	10 , 8	0 , 0
Luke eatKFC	0 , 0	8 , 10

- 2 Nash equilibrium strategies circled (3rd randomized)
 - No agent would unilaterally defect from these strategies
 - Pure (deterministic) strategies OK here
 - But not for Rock-Paper-Scissors
 - But how do agents coordinate?
 - Introduce concepts like “correlated equilibrium”
 - Agents agree on strategy based on outcome of commonly observed randomizing device; no incentive to defect

Bandit Setting

- k-armed bandit, e.g., $k=5$



- Environment unknown
 - Each bandit gives reward from stationary distribution
 - Assume stateless
 - Or at least that distribution in two large samples is same
- Which arm to pull at every time step?
 - Exploration vs. exploitation tradeoff

Not as much an issue with model-based methods... why?

Bandit Setting

- Upper Confidence Bound (UCB) Algorithm (Auer, Cesa-Bianchi, Fischer, 2002)
 - On n th try, play machine j that maximizes

$$\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$$

Called “exploration bonus”... why?

(sample mean of arm j plus UCB for arm)

- Guarantees optimal logarithmic regret bounds uniformly over time ($\Delta_i = \mu^* - \mu_i$):

$$\text{regret}(n) \leq \left[8 \sum_{i: \mu_i < \mu^*} \left(\frac{\ln n}{\Delta_i} \right) \right] + \left(1 + \frac{\pi^2}{3} \right) \left(\sum_{j=1}^k \Delta_j \right)$$

Conclusion

- Touched on “expected utility hypothesis”
- Discussed simple one-shot decision making problems
 - Just the tip of the iceberg
 - Meant to offer high-level insights into impacts of model decisions
 - Stochasticity
 - Partial observability
 - Multiple agents
 - Unknown environment (bandits)
- But this course is about *sequential decision making*
 - We’ll put multiple agents aside for now
 - We’ll also put partial observability aside until Marcus’s lectures
 - Next up is the fully observable, Markovian decision process...