Sequential decisions under uncertainty

KTH/EES PhD course

Lecture 1

- Course website

  http://www.ee.kth.se/~alepro/DecisionCourse

  - Schedule
  - Slides / Lecture Notes
  - Projects
  - Problems and Solutions
  - Grading
  - Readings

- Grading: 70% exams, 30% project

  - Take-home exam 1: after lecture 6 (35%)
  - Take-home exam 2: after lecture 10 (35%)
  - Project (30%)
  - Credits: 8 points

- Contact alepro@kth.se

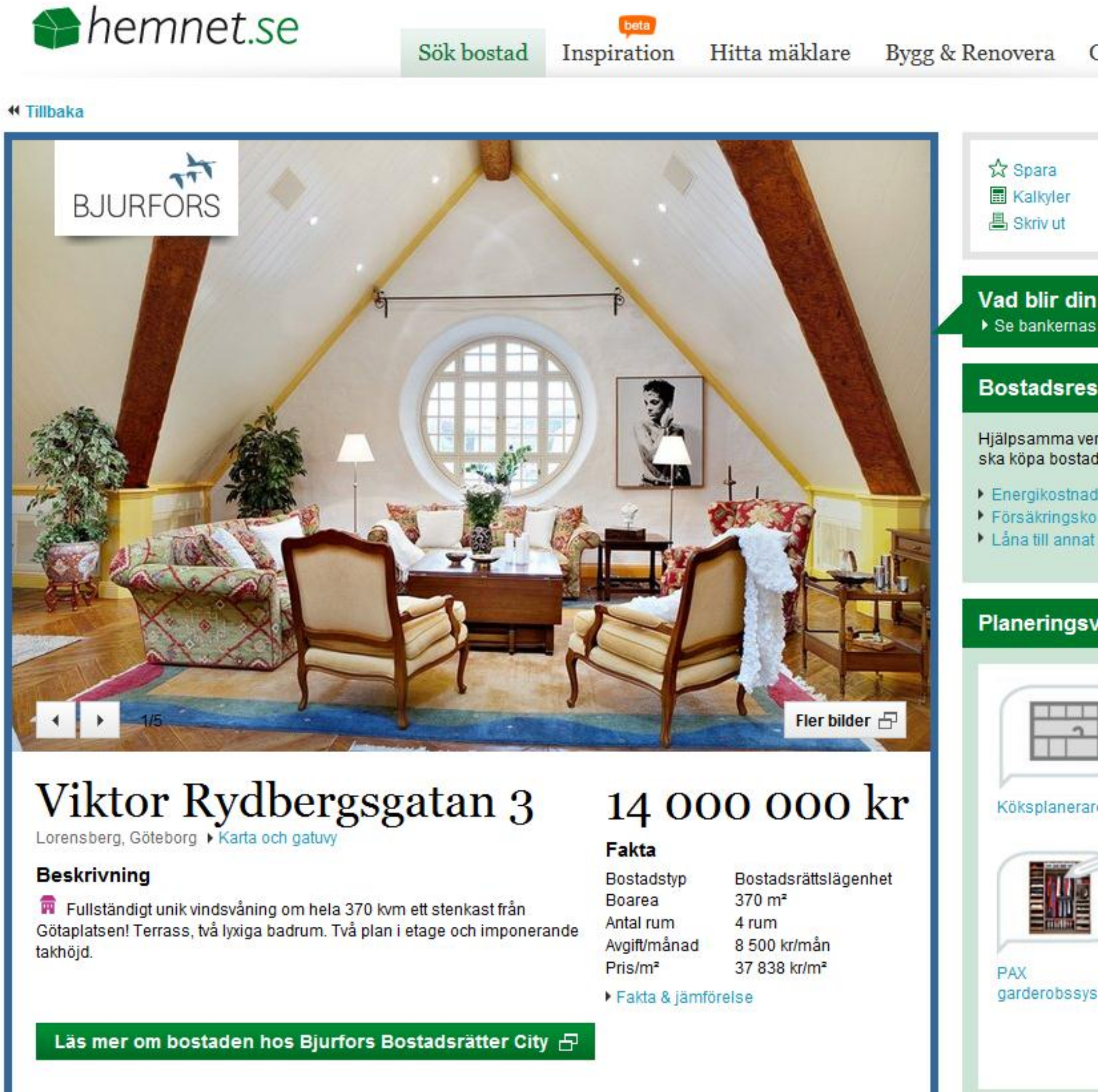
Lecture 1

- Introduction
  - Generic models
  - Examples
  - Schedule

- The secretary problem

Introduction

Buying and selling



Hjälpsamma ver
ska köpa bostad

- Energikostnad
- Försäkringsko
- Låna till annat

Viktor Rydbergsgatan 3

Lorensberg, Göteborg ▸ Karta och gatuvy

Beskrivning

Fullständigt unik vindsvåning om hela 370 kvm ett stenkast från Götaplatsen! Terrass, två lyxiga badrum. Två plan i etage och imponerande takhöjd.

14 000 000 kr

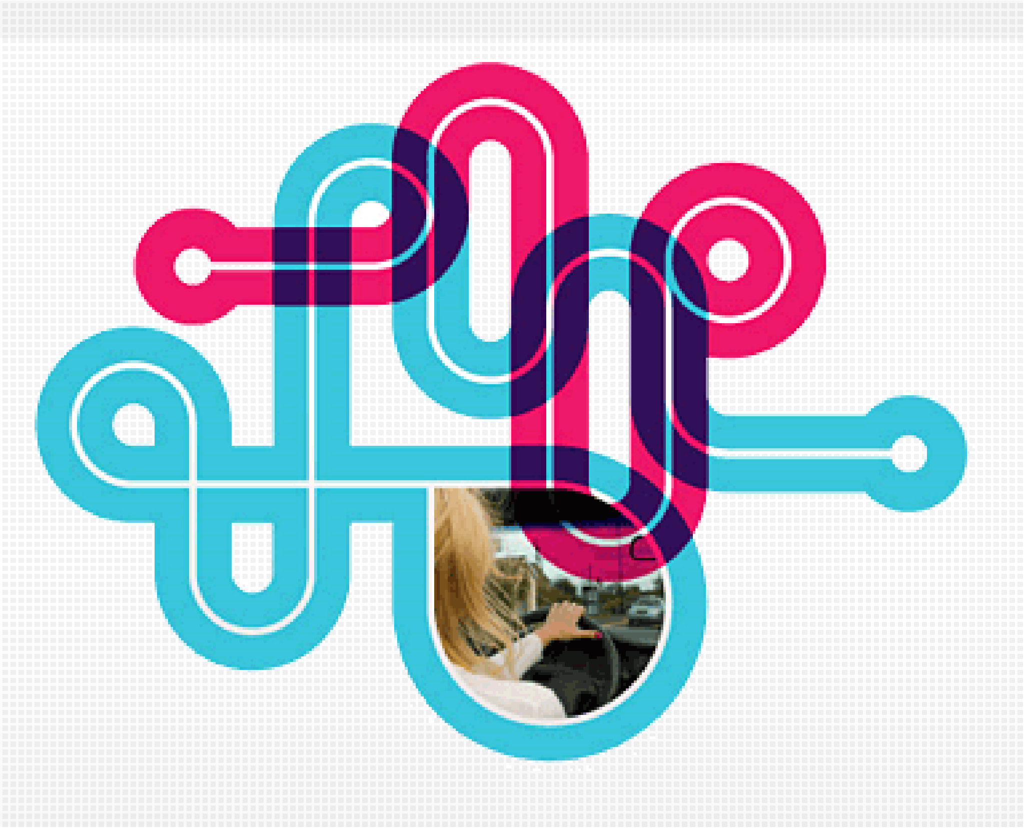
Fakta

Bostadstyp	Bostadsrättslägenhet
Boarea	370 m²
Antal rum	4 rum
Avgift/månad	8 500 kr/mån
Pris/m²	37 838 kr/m²

▸ Fakta & jämförelse

Läs mer om bostaden hos Bjurfors Bostadsrätter City

Routing

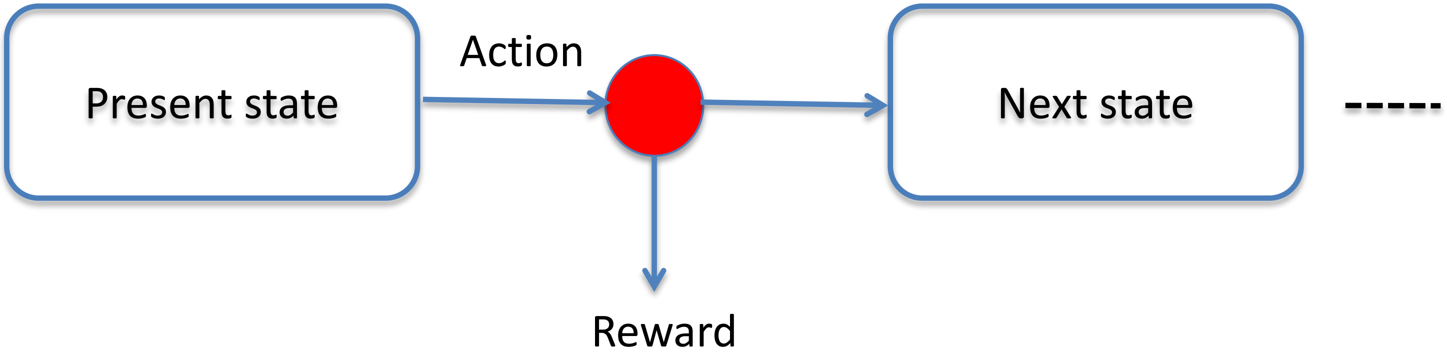


Portfolio optimization

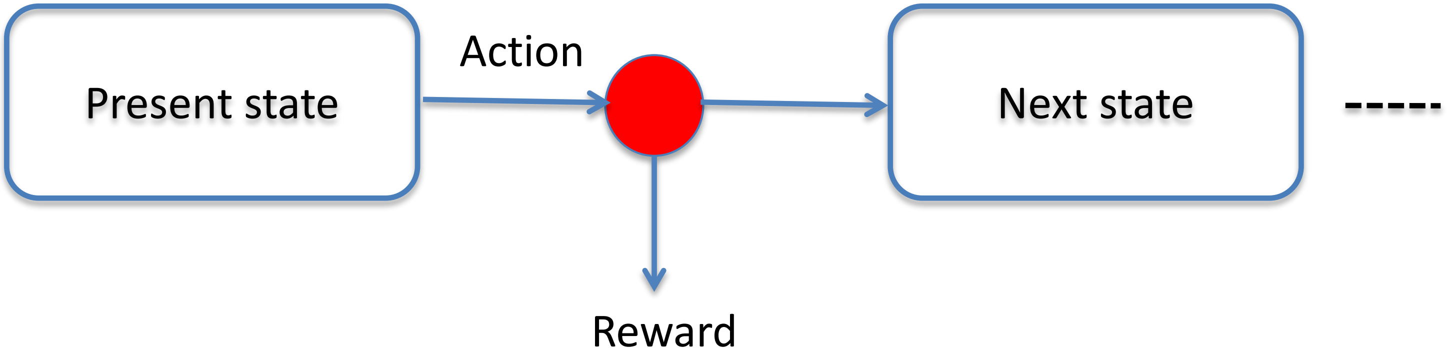
Inventory management

Gambling

Generic model

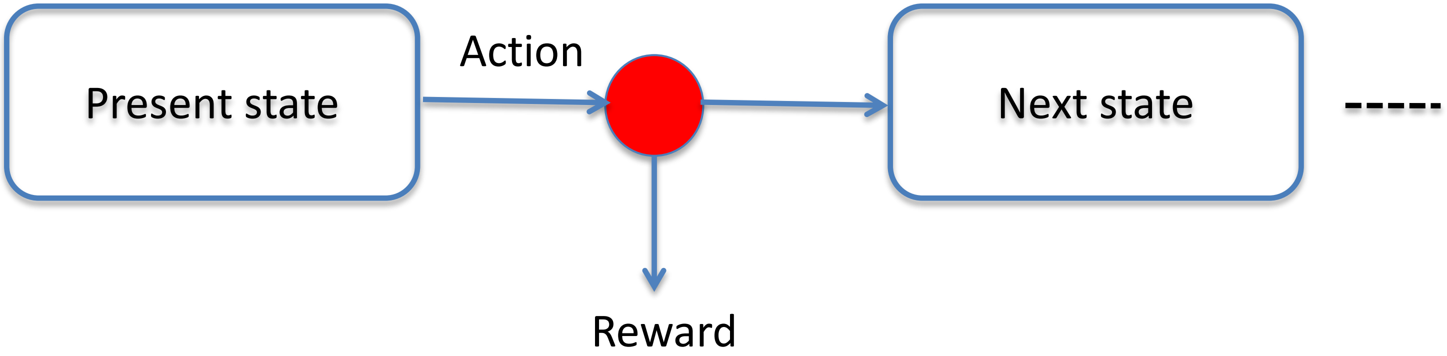


Generic model



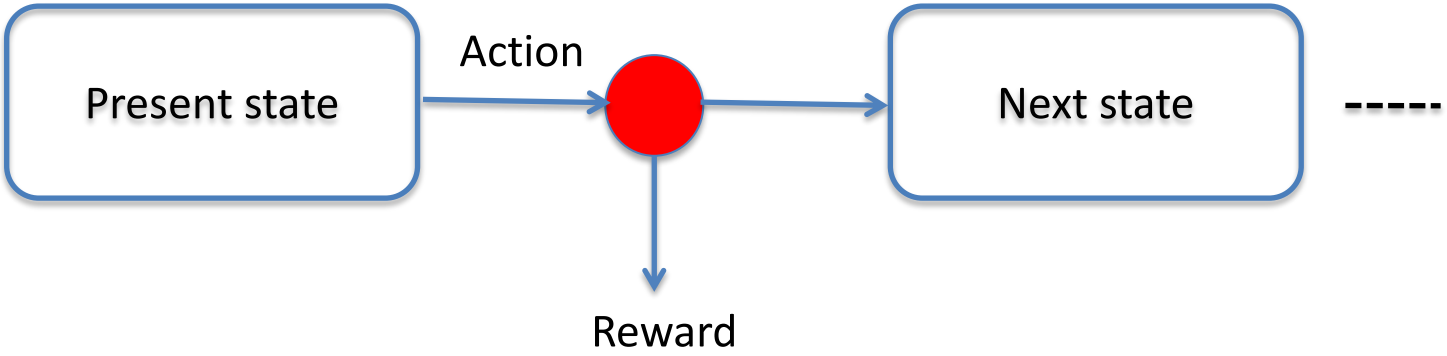
Objective: devise a sequential action policy maximizing your total reward

Generic model



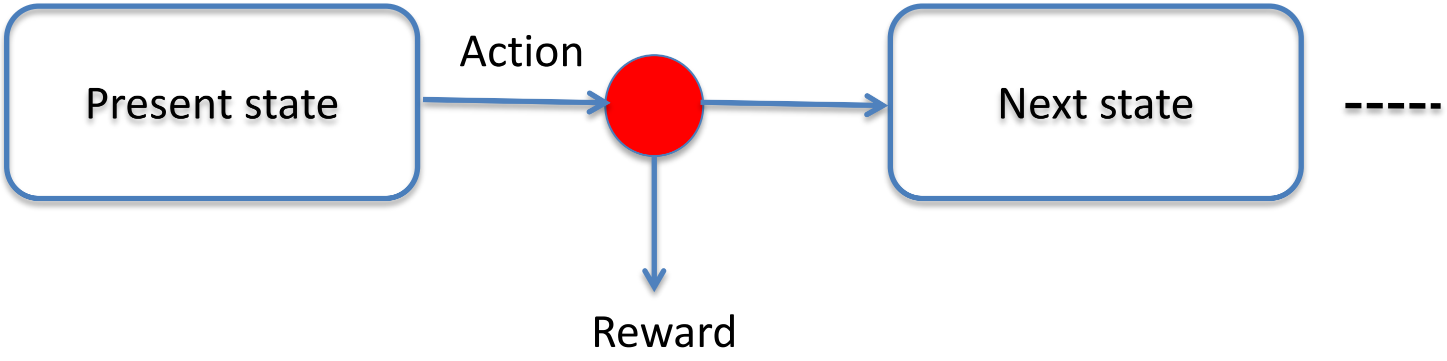
- State s
- Action a
- Reward R(s,a): in general a random variable
- Transitions: P(s,s';a): probability to move from state s to state s' given that the selected action is a

Information



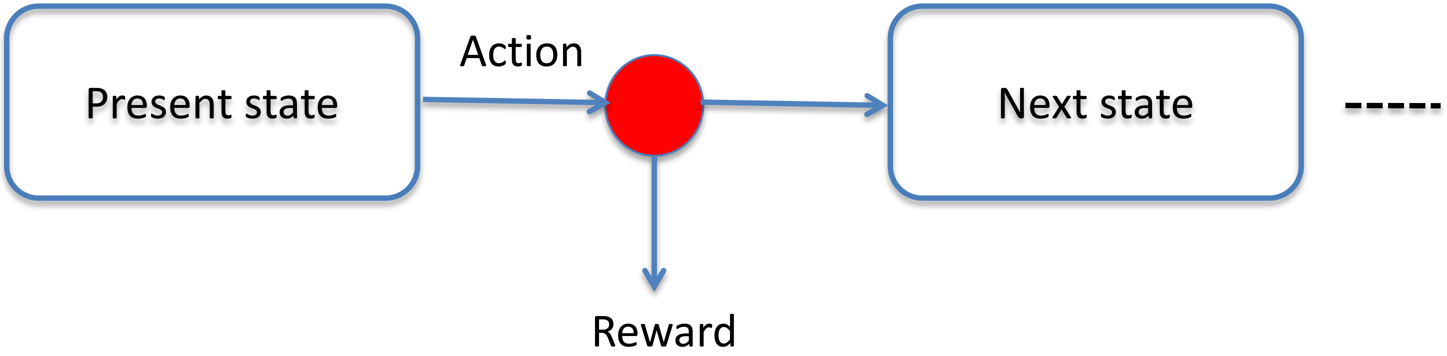
- Markov Decision Process (MDP)
  - Fully observable state and reward
  - Known reward statistics
  - Known transition probabilities
  - At the t-th step, the action is selected depending on the history up to time t: $(s_1,a_1,r_1,\ldots,s_t)$

Information



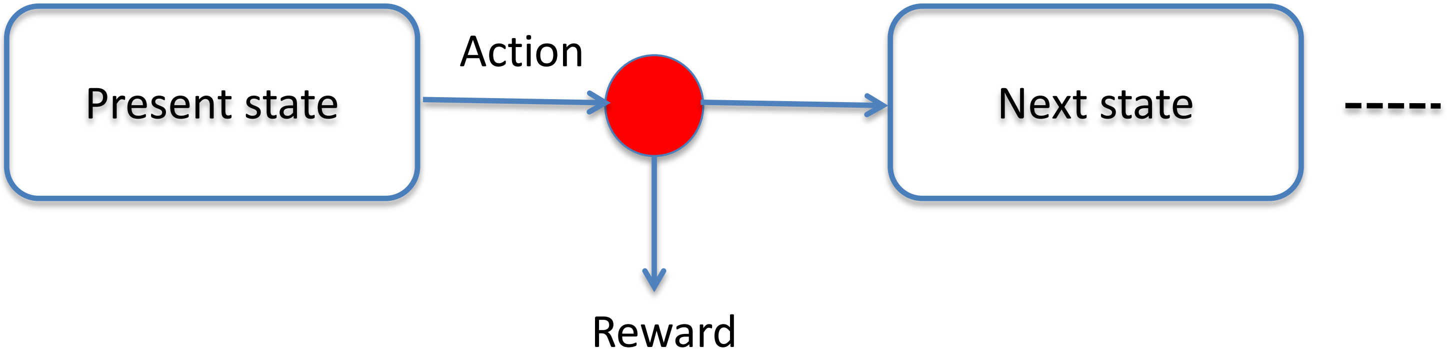
- Partially Observable Markov Decision Process (POMDP)
  - Partially observable state: we observe z, and know P(s|z)
  - Observed rewards
  - Known reward statistics
  - Known transition probabilities
  - At the t-th step, the action is selected depending on the history up to time t: $(z_1,a_1,r_1,\ldots,z_t)$

Information



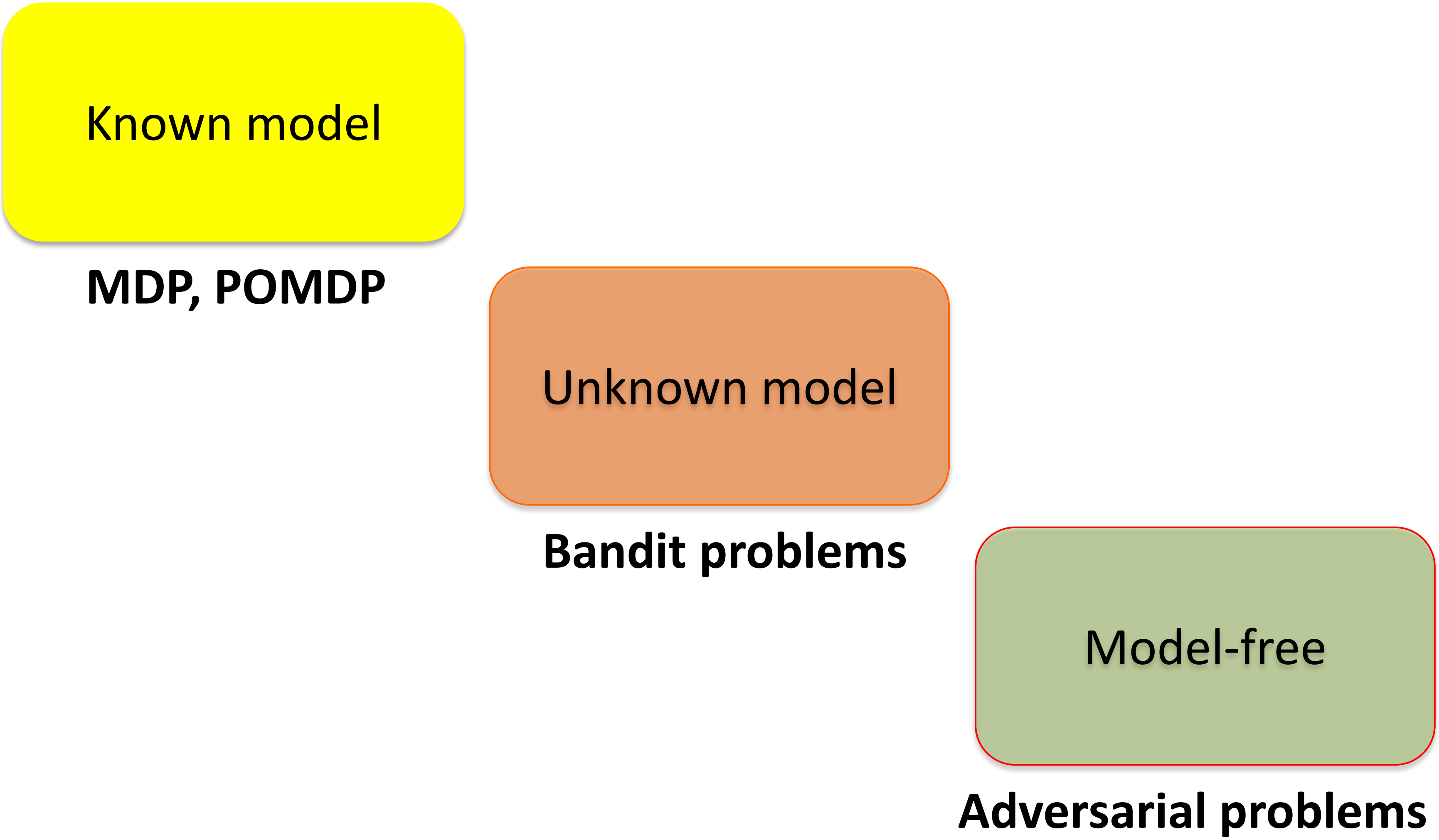
- Bandit problems
  - Observable state and reward
  - Unknown reward statistics
  - Known or unknown transition probabilities
  - At the t-th step, the action is selected depending on the history up to time t: $(s_1,a_1,r_1,\ldots,s_t)$

Information



- Adversarial problems
  - Observable state and reward
  - No model for both reward functions and transitions
  - At the t-th step, the action is selected depending on the history up to time t: $(s_1,a_1,r_1,\ldots,s_t)$

Classification



Known model

MDP, POMDP

Unknown model

Bandit problems

Model-free

Adversarial problems

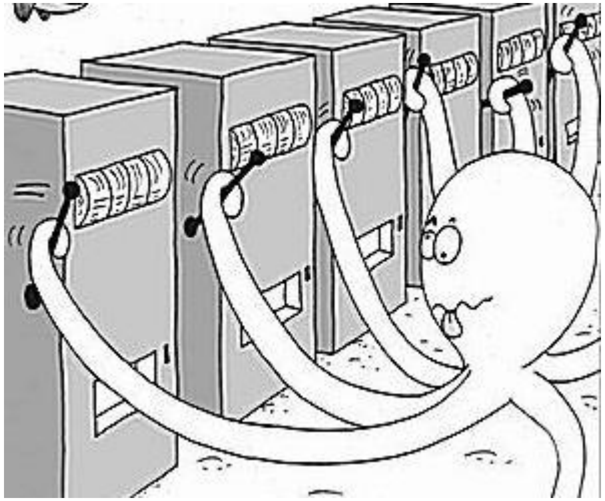
Examples

1. Markov Decision Process (MDP): the secretary problem

2. Classical Multi-armed bandit problem

3. Blind online optimization

Multi-Armed Bandit (MAB)



In a casino, N slot machines can be played. Each time you play machine i, you get a reward X_i = 0 or 1 with initially unknown mean r_i. Objective: sequentially play machines so as to maximize your average reward (over t plays).

MAB: Model

- Rewards
  - When played at step t, machine i brings a random reward $X_i(t)$
  - $X_i(t)$, t=1,2,… are independent and identically distributed with initially unknown mean r_i
  - Rewards are independent across arms

- Model
  - State: no state is needed (i.i.d. rewards), i.e. the state is always 1
  - Action set: {1,…,N}
  - History up to time t: $(a(1),X_{a(1)}(1),\ldots,a(t),X_{a(t)}(t))$
  - A policy maps at step t the history to an action

Regret

- If the average rewards were known, it would be optimal to select the best machine

$$i^\star = \arg \max_{i \in \{1,\ldots,N\}} r_i$$

$$R^\star(t) = t \times r_{i^\star}$$

- Regret of policy π

$$\text{regret}^\pi(t) = R^\star(t) - \sum_{s=1}^{t} r_{\pi(s)}$$

- Goal: find a policy minimizing regret

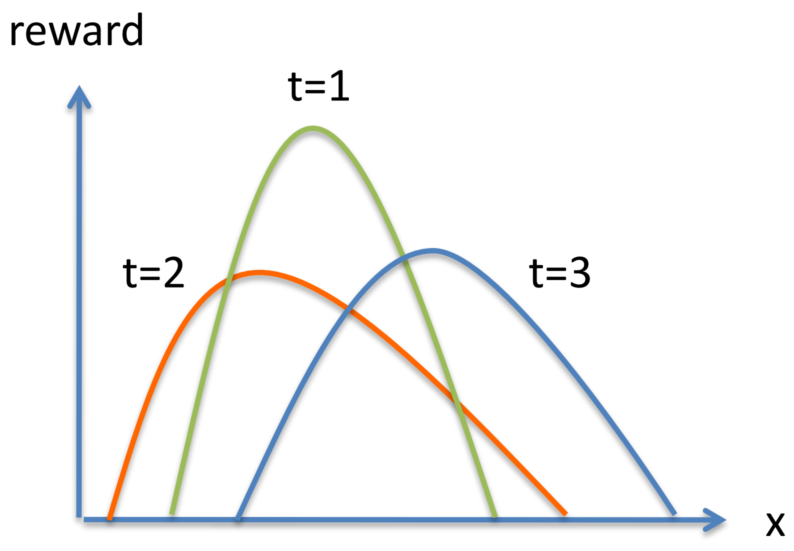
Results

- Exploitation vs. exploitation trade-off: even if a machine has yielded good rewards so far, one needs to explore other machines
- Zero-regret policies: $\frac{1}{t}\text{regret}^{\pi}(t) \to 0, \quad \text{as } t \to \infty$
- Lai-Robbins (1985). An achievable lower bound on regret is:

$$\text{regret}^{\pi}(t) \geq \log(t) \left(\sum_{i : r_i \neq r_{i^\star}} \frac{r_{i^\star} - r_i}{I(r_i, r_{i^\star})} + o(1) \right)$$

$$I(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

Blind online optimization



At the beginning of each year, Volvo has to select a vector x (in a convex set) representing the relative efforts in producing various models (S60, V70, …). The reward is an arbitrarily varying and unknown concave function of x. How to maximize reward over say 50 years?

Model

- Rewards
  - At step t, if the selected vector is x, the reward is $c_t(x)$
  - … and we observe $c_t(x)$ only (the function is not revealed)

- Problem definition
  - State: no state is needed, i.e. the state is always 1
  - Action set: S a convex set
  - History up to time t: $x(1),c_1(x(1)),\ldots,x(t),c_t(x(t))$
  - A policy maps at step t the history to an action

Regret

- Reference: the best offline static policy

$$x^\star = \arg\max_{x \in S} \sum_{s=1}^{t} c_s(x)$$

- Regret of policy π

$$\text{regret}^\pi(t) = \sum_{s=1}^{t} \left(c_s(x^\star) - c_s(x^\pi(s))\right)$$

- Goal: find a policy minimizing regret

Results

- Zero-regret policies exist!

- Flaxman et al. (2005). An algorithm with regret such that:

$$\mathrm{regret}^{\pi}(t) \leq K t^{5/6}$$

PART I - Stochastic models

1. Review of probabilistic tools. Markov chains, Martingales, basic inequalities.
2. Discrete time Markov Decision Processes (MDPs).
   - 2a. Finite time-horizon. Principle of optimality, backward induction.
   - 2b. Infinite time-horizon. Principle of optimality, value / policy iteration, modified policy iteration, linear programming.
3. Solving MDPs - part 1. Exact solutions based on structural properties of the MDP.
4. Solving MDPs - part 2. Some approximation methods.
5. Extensions. Constrained MDPs, Partially Observable MDPs, Decentralized MDPs.
6. Limit theorems. Going from MDPs to deterministic continuous-time control.
7. Optimal stopping time problems.
8. Kalman filter.
9. Prediction with expert advice and Multi-Armed Bandit (MAB) problems.

PART II - Adversarial models and Games.

1. Prediction with expert advice and MAB problems in adversarial scenarios.
2. Sequential decision making in games. Internal regret, Correlated equilibria, Convergence to and selection of Nash Equilibria.
3. Recent advances in online optimization.

Books

- Markov Decision Processes – Discrete Stochastic Dynamic Programming. Martin Puterman. Wiley, 1994

- Prediction, learning, and games. Nicolo Cesa-Bianchi and Gabor Lugosi. Cambridge Univ. Press, 2006

http://www.ii.uni.wroc.pl/~lukstafi/pmwiki/uploads/AGT/Prediction_Learning_and_Games.pdf

The secretary problem

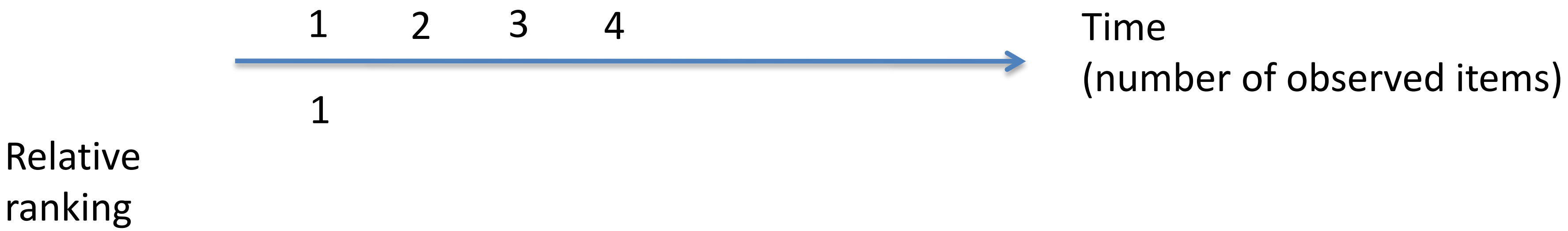
The standard problem

A known number n of items is presented one by one in a random order (all n! ordering are equally likely). The observer is able to rank the observed items. When a new item is presented, he may accept it, in which case the process stops, or reject it and ask for a new item.

The objective of the observer is to maximize the probability of the accepted item to be of highest rank.

The standard problem

- n ranked items (i is the item with i-th rank)
- Observed in a random order σ. σ is a permutation of 1,…,n uniformly distributed on the set of permutation. σ(i): actual rank of the i-th observed item.
- From the observer perspective:



1 2 3 4

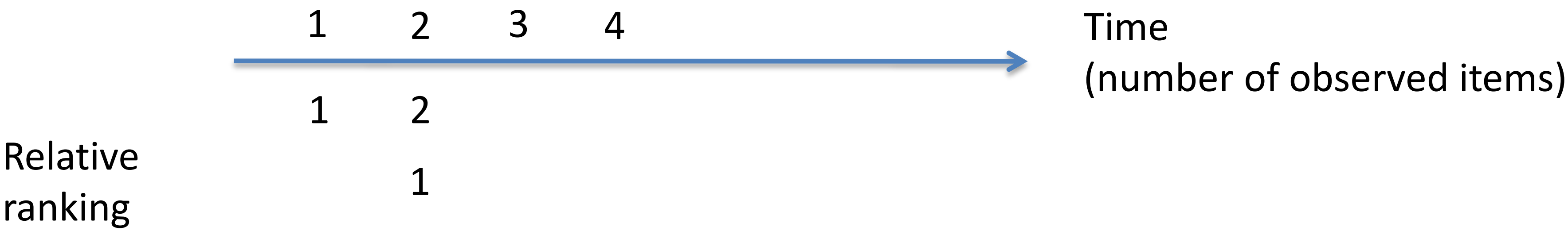
Time
(number of observed items)

1

Relative
ranking

The standard problem

- n ranked items (i is the item with i-th rank)
- Observed in a random order σ. σ is a permutation of 1,…,n uniformly distributed on the set of permutation. σ(i): actual rank of the i-th observed item.
- From the observer perspective:



1 2 3 4

Time
(number of observed items)

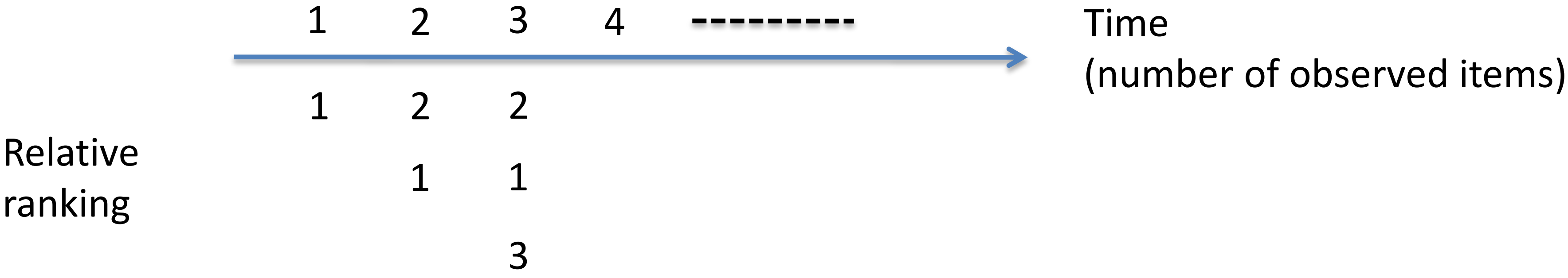
1 2

1

Relative ranking

The standard problem

- n ranked items (i is the item with i-th rank)
- Observed in a random order σ. σ is a permutation of 1,…,n uniformly distributed on the set of permutation. σ(i): actual rank of the i-th observed item.
- From the observer perspective:



1 2 3 4 ---------

Time
(number of observed items)

Relative ranking

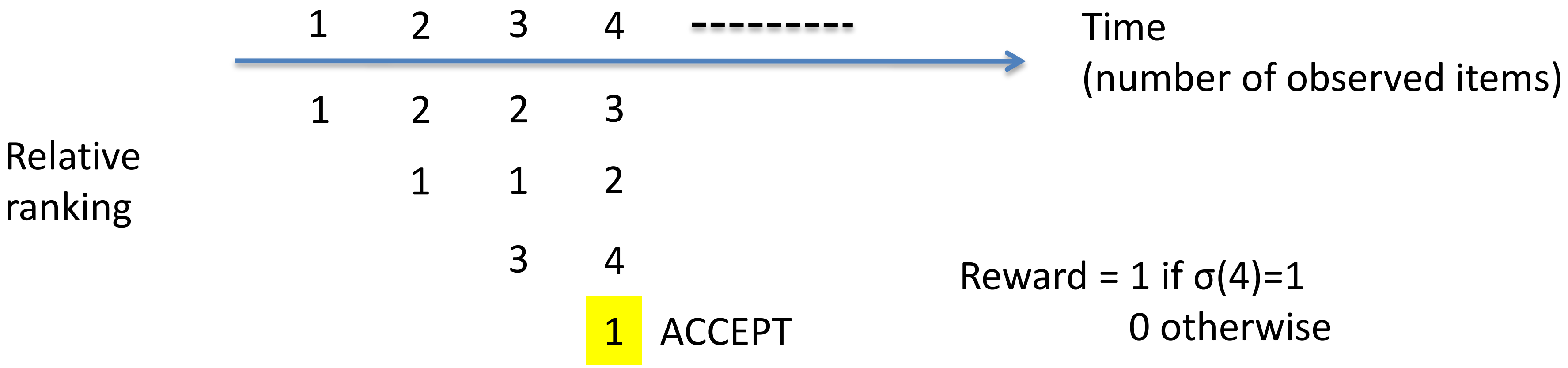
1 2 2

1 1

3

The standard problem

- n ranked items (i is the item with i-th rank)
- Observed in a random order σ. σ is a permutation of 1,…,n uniformly distributed on the set of permutation. σ(i): actual rank of the i-th observed item.
- From the observer perspective:



1 2 3 4 ---------

Time
(number of observed items)

Relative
ranking

1 2 2 3

1 1 2

3 4

1 ACCEPT

Reward = 1 if σ(4)=1
0 otherwise

MDP formulation

- State (r,s): r is the number of items observed so far, s is the relative apparent rank of the last observed item
- Additional state 0: after the observer accepted an item
- Actions: a is either A=accept or R=reject
- Reward R((r,s),A) = 1 if σ(r) = 1, 0 otherwise

  R((r,s),R) = 0
- Transitions:
  - P((r,s),(r+1,s');R) = 1/(r+1) for all s' = 1,…,r+1
  - P((r,s),0;A)=1

Applications

- Choosing the best applicant for a job

- Other common names: marriage problem, beauty contest problem, …

History

- Obscure origin
- Posed in 1955 by A. Gleason
- First published solution by Lindley in 1961

- More in the survey paper. The secretary problem and its extensions: a Review. P.R. Freeman. International Statistical Review, 51, 1983.

Solution

Theorem Let $r^\star = \min\{r \geq 1 : \sum_{k=r+1}^{n} \frac{1}{k-1} \leq 1\}$

The optimal policy consists in first observing $r^\star$ items, and then accepting the first item with apparent rank 1.

Solution

Theorem Let $r^\star = \min\{r \geq 1 : \sum_{k=r+1}^{n} \frac{1}{k-1} \leq 1\}$

The optimal policy consists in first observing $r^\star$ items, and then accepting the first item with apparent rank 1.

When n (and r) is large, $\sum_{k=r+1}^{n} \frac{1}{k-1} \approx \log(n/r)$

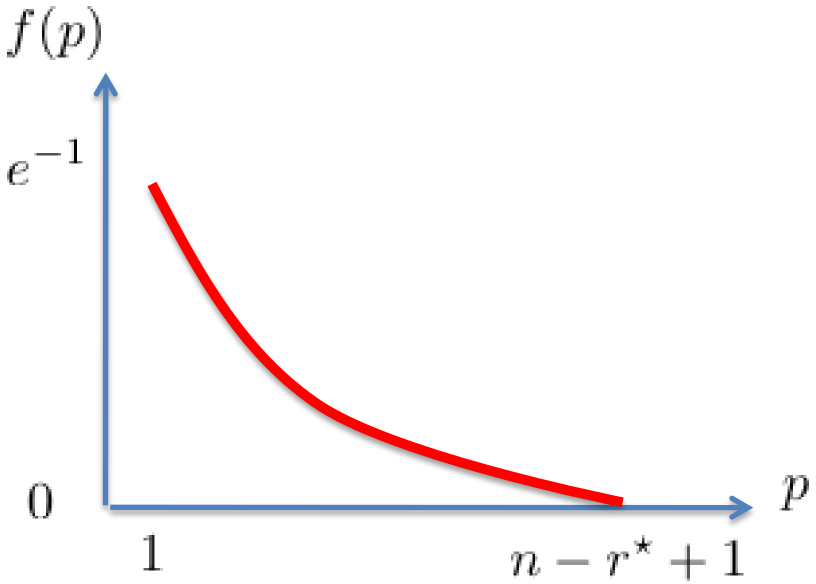
the optimal policy is defined by $r^\star \approx e^{-1}n$

Under the optimal policy, the expected reward (the probability of selecting the best item is e^{-1}

Rank of accepted item

- The probability under the optimal policy to accept item with rank p is, for $p = 1, \ldots, n - r^\star + 1,$

$$f(p) = \binom{n}{r^\star - 1}^{-1} \sum_{i=p+1}^{n-r^\star+2} \frac{1}{i-1} \times \binom{n-i}{r^\star - 2}$$



Extensions

- Minimizing the expected rank of the accepted item
- Unknown number of items
- Selecting several items
- Problems with recall
- …

Minimizing the expected rank

- The objective of the observer is to minimize the expected rank of the selected item
- The observer may stop the process after observing an item with relative rank higher than 1
- Optimal policy is threshold-based: stop in state (r,s) if and only if s is less than g(r) – Lindley (1961)
- Limited expected rank – Chow et al. (1964):

Minimizing the expected rank

- The objective of the observer is to minimize the expected rank of the selected item
- The observer may stop the process after observing an item with relative rank higher than 1
- Optimal policy is threshold-based: stop in state (r,s) if and only if s is less than g(r) – Lindley (1961)
- Limited expected rank – Chow et al. (1964):

$$\prod_{j=1}^{\infty}\left(\frac{j+2}{j}\right)^{1/(j+1)} = 3.8695$$

Are we smart enough?

- Sequential Observation and Selection with Rank-Dependent Payoffs: an experimental study. J.N. Bearden et al. Management Science, 2006.

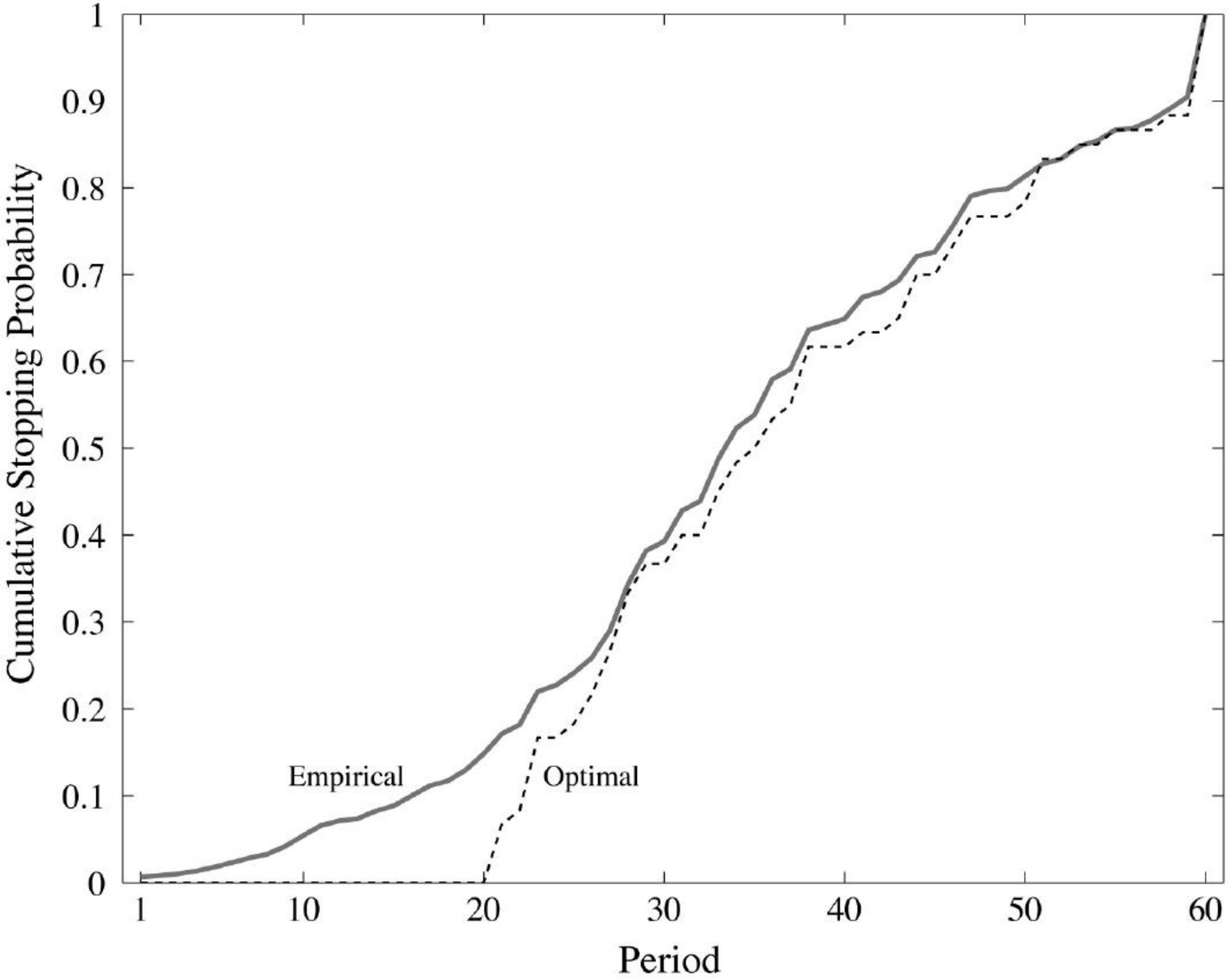
- Are we naturally implementing the optimal decision policy?

- 62 PhD students (Univ. of Arizona)
- Incentive to participate: pay-offs from 5$ to 50$ per experimentation (30 mins)

Set up

- n = 60
- 60 experimentations (out of the 60! possible)
- Each student plays the sequence of 60 item selection game
- Objective: minimize the expected rank of the selected item

We stop too early



… and actually, we do