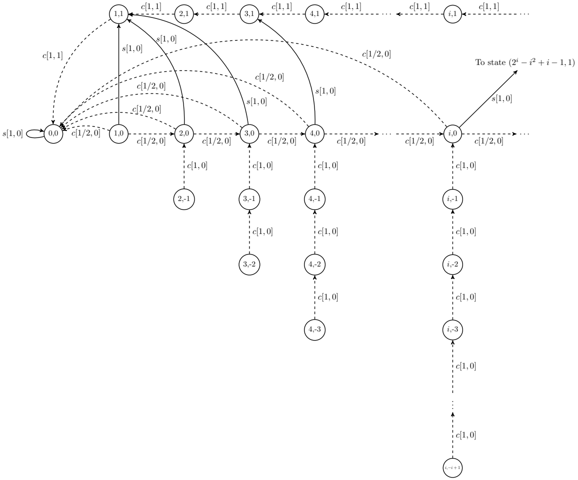
Markov Decision Processes & Complexity Theory: Some Research Directions

Jefferson Huang

Department of Applied Mathematics and Statistics
Stony Brook University

SBU AMS Graduate Reading Group
February 10, 2016

Outline



1. Definition
2. Applications
3. Research Directions

Definition

1. **state** space $\mathbb{X}$
2. sets of available **actions** $A(x)$ at each state x
3. one-step **costs** $c(x, a)$: incurred whenever the state is x and action $a \in A(x)$ is performed
4. transition **probabilities** $p(y|x, a)$: probability that the next state is y, given that the current state is x & action $a \in A(x)$ is performed

Initial Distribution

State $\quad x_0 \xrightarrow[a_0]{p(x_1|x_0,a_0)} x_1 \xrightarrow[a_1]{p(x_2|x_1,a_1)} x_2 \xrightarrow[a_2]{p(x_3|x_2,a_2)} x_3 \longrightarrow \cdots$

Policies & cost criteria

A **policy** φ prescribes an action for every state.

Common **cost criteria**:

- Total (discounted) costs: for $\beta \in [0, 1]$,

$$v_\beta^\varphi(x) := \mathbb{E}_x^\varphi \sum_{n=0}^{\infty} \beta^n c(x_n, a_n)$$

- Average costs:

$$w^\varphi(x) := \limsup_{N\to\infty} \frac{1}{N} \mathbb{E}_x^\varphi \sum_{n=0}^{N} c(x_n, a_n)$$

Optimal policy: minimizes the cost criterion for every initial state.

Example: Gambling

You want to increase your fortune from s dollars to S dollars by repeatedly playing this game:

- You can bet any amount a not exceeding your current fortune.
- If you win, your new fortune is $s + a$; if you lose, your new fortune is $s - a$.
- You win with probability p, and lose with probability $1 - p$.
- Play stops when your fortune is 0 or S.

Goal: Maximize the probability that you get S dollars.

Computing optimal policies

Solving an MDP = computing an (ϵ-)optimal policy

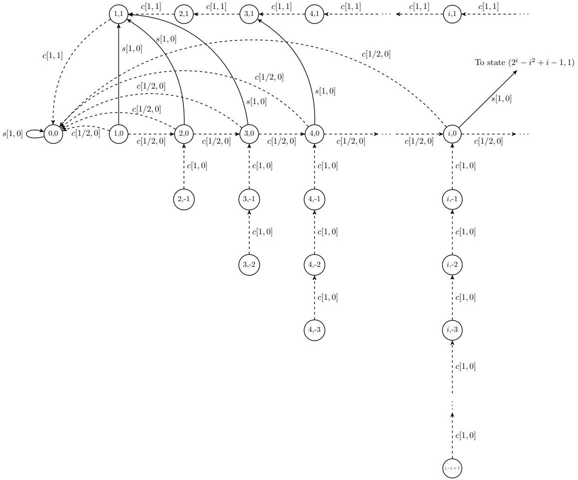
3 main approaches:

1. **Value iteration (VI)** (Shapley 1953)
   - Iteratively approximate the optimal costs from each state.
2. **Policy iteration (PI)** (Howard 1960)
   - Iteratively improve a starting policy.
3. **Linear programming (LP)** (early 1960s)
   - Compute the optimal frequencies with which each state-action pair should be used.

They're closely related:

- PI is a simplex method for the LP (Mine & Osaki 1970)
- VI is a primal-dual method for the LP (Cogill 2016)

Outline



1. Definition
2. Applications
3. Research Directions

Applications

First (?) application of MDPs: Sears mail-order catalogs ($\sim$1958)

Ronald A. Howard (1978):

> *... my one successful application was the original application that sparked my interest in this whole research area.*

Some others:

- **Operations Research:** inventory control, control of queues, vehicle routing, job shop scheduling
- **Power Systems:** voltage & reactive power control, control of storage devices, electric vehicle charging, bidding in electricity markets
- **Healthcare:** medical decision making, epidemic control
- **Finance:** option pricing, portfolio selection, credit granting
- **Computer Science:** wireless sensor networks, cloud computing, reinforcement learning

MDPs & pure mathematics

Ronald A. Howard (1978):

> *The Markov decision process and its extensions have now become principally the province of mathematicians.*

Borel-space MDPs: Blackwell (1965), Strauch (1966)

> $\rightarrow$ connections to **descriptive set theory**: see e.g. Bertsekas & Shreve (1978), Dynkin & Yushkevich (1979)

Motivated **counterexamples** on:

- theory of Borel sets, semicontinuity of minimum functions

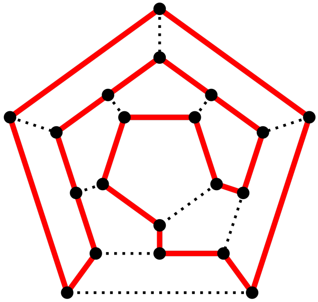
and **new results** on:

- extensions of Berge's Theorems & Fatou's Lemma
- convergence of probability measures, solutions of Kolmogorov's equations

Application: Hamiltonian cycles

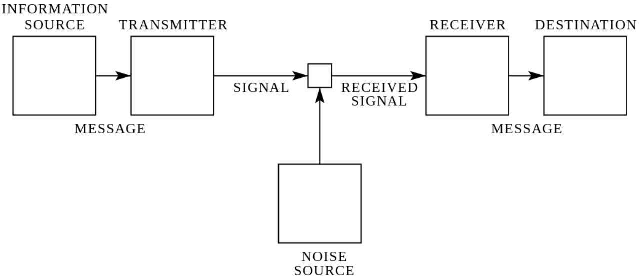
Problem: Is there a cycle that visits every vertex exactly once?

- One of Karp's (1972) 21 NP-complete problems.
- Can be formulated as a constrained MDP (Filar & Krass 1994, Feinberg 2000).



Application: Source coding

Problem: How to compress & de-compress data?



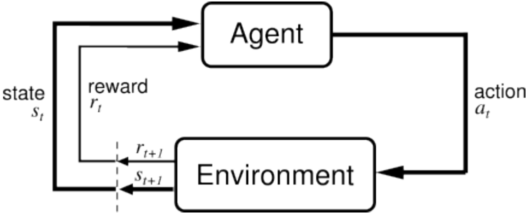
From Claude Shannon's "A Mathematical Theory of Communication".

- Can be formulated as an MDP (e.g. Linder & Yüksel 2014)
    - Minimize average distortion between original and reconstructed data.
- Related to approximating policies for MDPs with infinite state spaces (Saldi Linder Yüksel 2015).

Application: Reinforcement learning

Problem: How can an agent learn to perform well in an unfamiliar environment?

- MDPs provide a modeling framework
- Use simulation to learn about the environment
- Function approximation is used to deal with complex environments

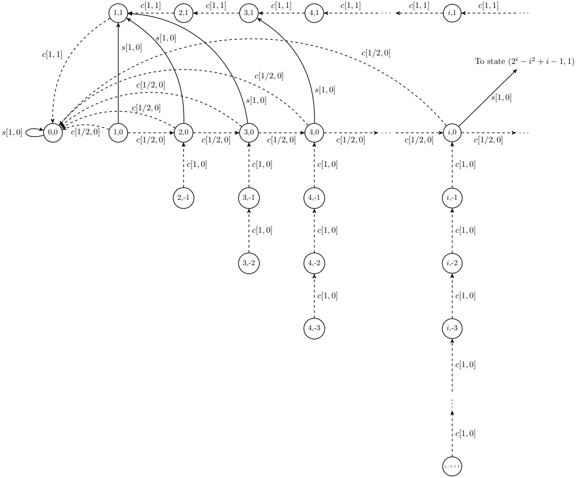


Application: Deep reinforcement learning

See Mnih et al. 2015.

Outline

1. Definition
2. Applications
3. Research Directions

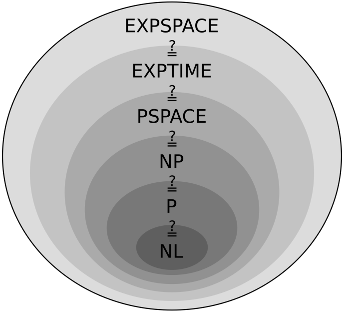


Research direction: NP-mightiness

Definition

An algorithm is **NP-mighty** if it can be used to solve any problem in NP.

- The simplex, network simplex, and successive shortest path algorithms are NP-mighty (Disser & Skutella SODA '15).
- Stronger result using MDPs:
  - Policy iteration (i.e. a simplex method) can be used to solve any problem in PSPACE $\supseteq$ NP (Fearnley & Savani STOC '15).



Research direction: Interior-point methods

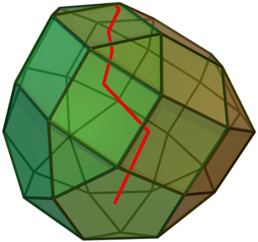
They've played an important role in complexity theory:

- First practical polynomial-time algorithm for linear programming (Karmarkar 1984).
- First strongly polynomial-time algorithm for MDPs (with a fixed discount factor) (Ye 2005).

They seem to work well for MDPs:

- Outperformed policy iteration on test problems and a real-life healthcare problem (Alagoz Ayvaci Linderoth 2015).

They can also be used to solve stochastic games (Hansen & Ibsen-Jensen 2013).



Research direction: Primal-dual methods

An old approach to solving linear programs (Dantzig Ford Fulkerson 1956).

Well-known in combinatorial optimization:

- Hungarian algorithm for the assignment problem
- Dijkstra's algorithm for shortest paths
- Ford-Fulkerson algorithm for maximum flows

Recent work on discounted MDPs (Cogill 2016):

- Includes several forms of value iteration as special cases.
- Leads to an alternative finite algorithm for MDPs.

Outline

1. Definition
2. Applications
3. Research Directions

Contact email:
`jefferson.huang@stonybrook.edu`