Sequential decision making under uncertainty

Matthijs Spaan   Francisco S. Melo

Institute for Systems and Robotics

Instituto Superior Técnico

Lisbon, Portugal

Reading group meeting, January 4, 2007

Introduction

This meeting:

- Overview of the field
  - Motivation
  - Assumptions
  - Models
  - Methods
- What topics shall we address?
- Fix a schedule.

Motivation

- Major goal of Artificial Intelligence: build intelligent agents.
- Russell and Norvig (2003): “an agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators”.
- Problem: how to act?
- Example: a robot performing an assigned task.

Applications

Reinforcement learning applications:

- Aibo gait optimization (Kohl and Stone, 2004a,b; Saggar et al., 2006)
- Helicopter control (Bagnell and Schneider, 2001; Ng et al., 2004)
- Airhockey (Bentivegna et al., 2002)
- More on

  `http://neuromancer.eecs.umich.edu/cgi-bin/twiki/view/Main/`

Sequential decision making under uncertainty

Assumptions:

Sequential decisions: problems are formulated as a sequence of "independent" decisions;

Markovian environment: the state at time t depends only on the events at time $t-1$;

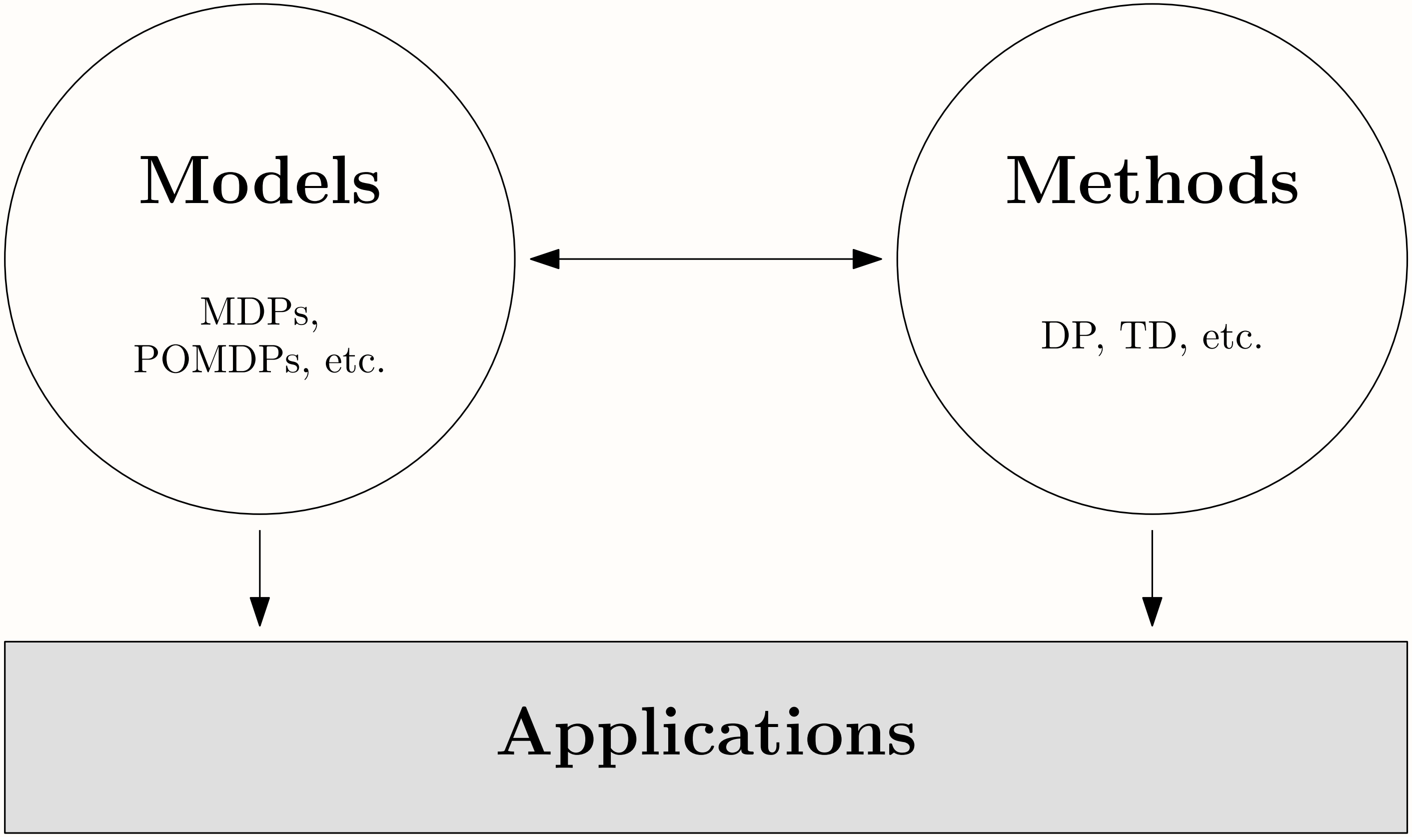
Evaluative feedback: use of a reinforcement signal as performance measure (reinforcement learning);

Sequential decision making under uncertainty (1)

Possible variations:

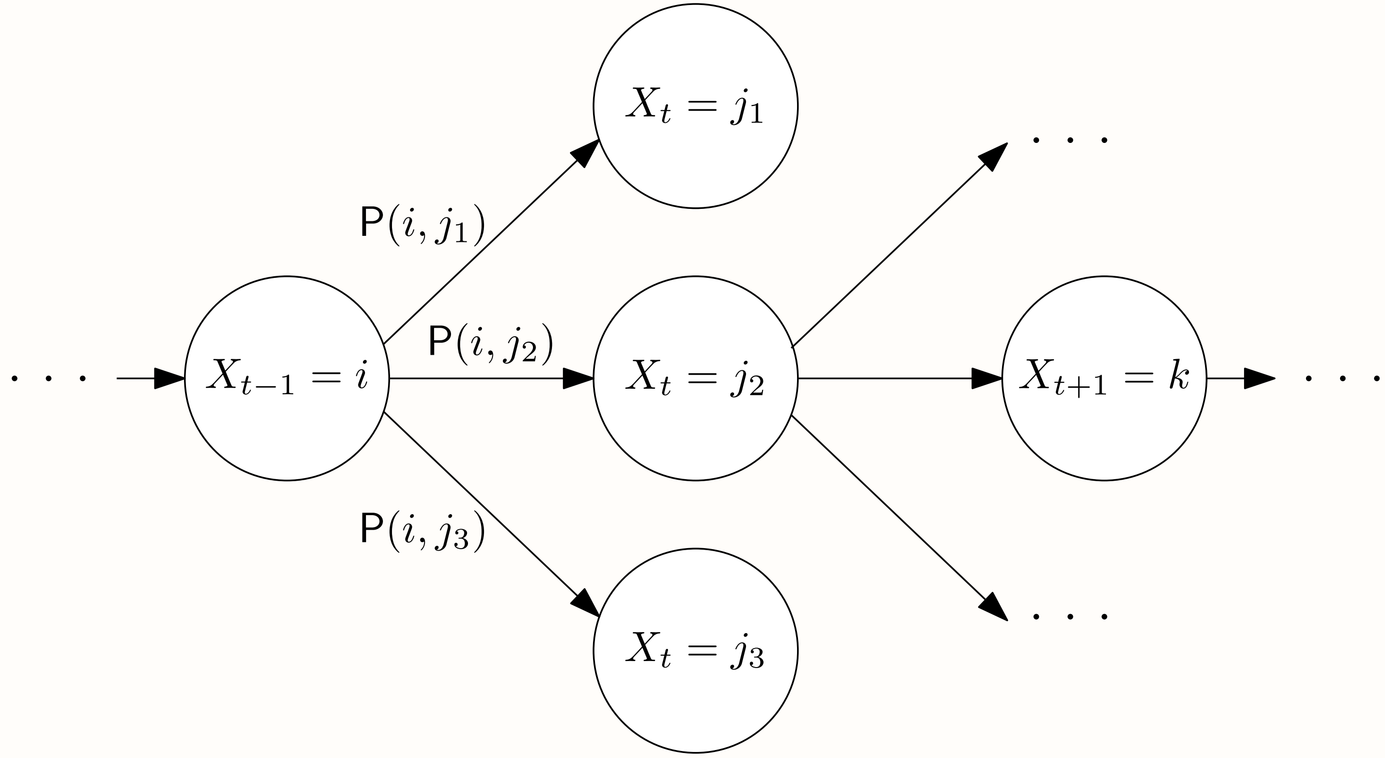
- Type of uncertainty.
- Full vs. partial state observability.
- Single vs. multiple decision-makers.
- Model-based vs. model-free methods.
- Finite vs. infinite state space.
- Discrete vs. continuous time.
- Finite vs. infinite horizon.

Sequential decision making under uncertainty (2)



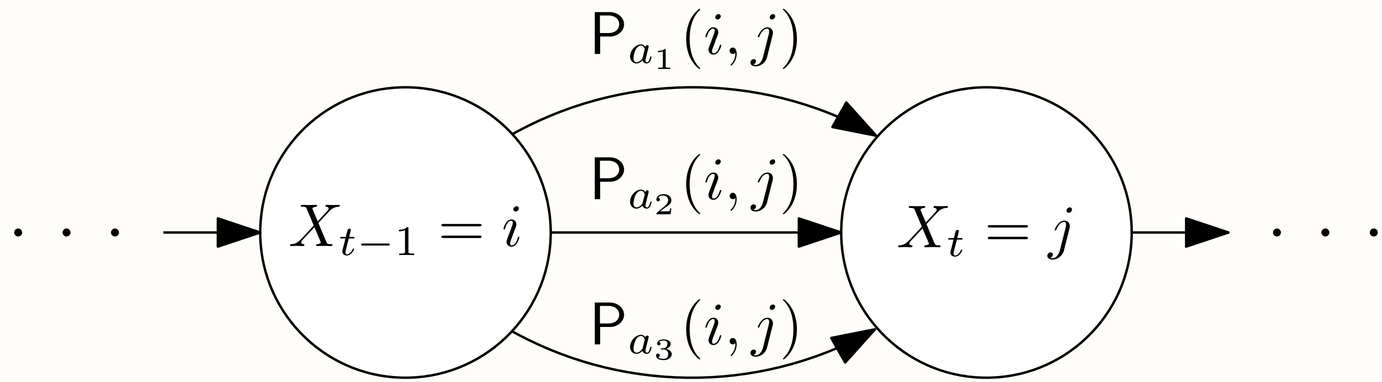
Basic model: Markov chains

The basic model of *Markov chains* describes (first order) discrete-time dynamic systems.



Adding control

In *controlled Markov chains*, the transition probabilities depend on a control parameter a.



Markov decision processes

A *Markov decision process* (MDP) is a controlled Markov chain endowed with a performance criterion (Puterman, 1994; Bertsekas, 2000).

- The decision-maker receives a numerical reward R_t for each time instant t;
- The decision-maker must optimize some long-run optimality criterion, e.g.,

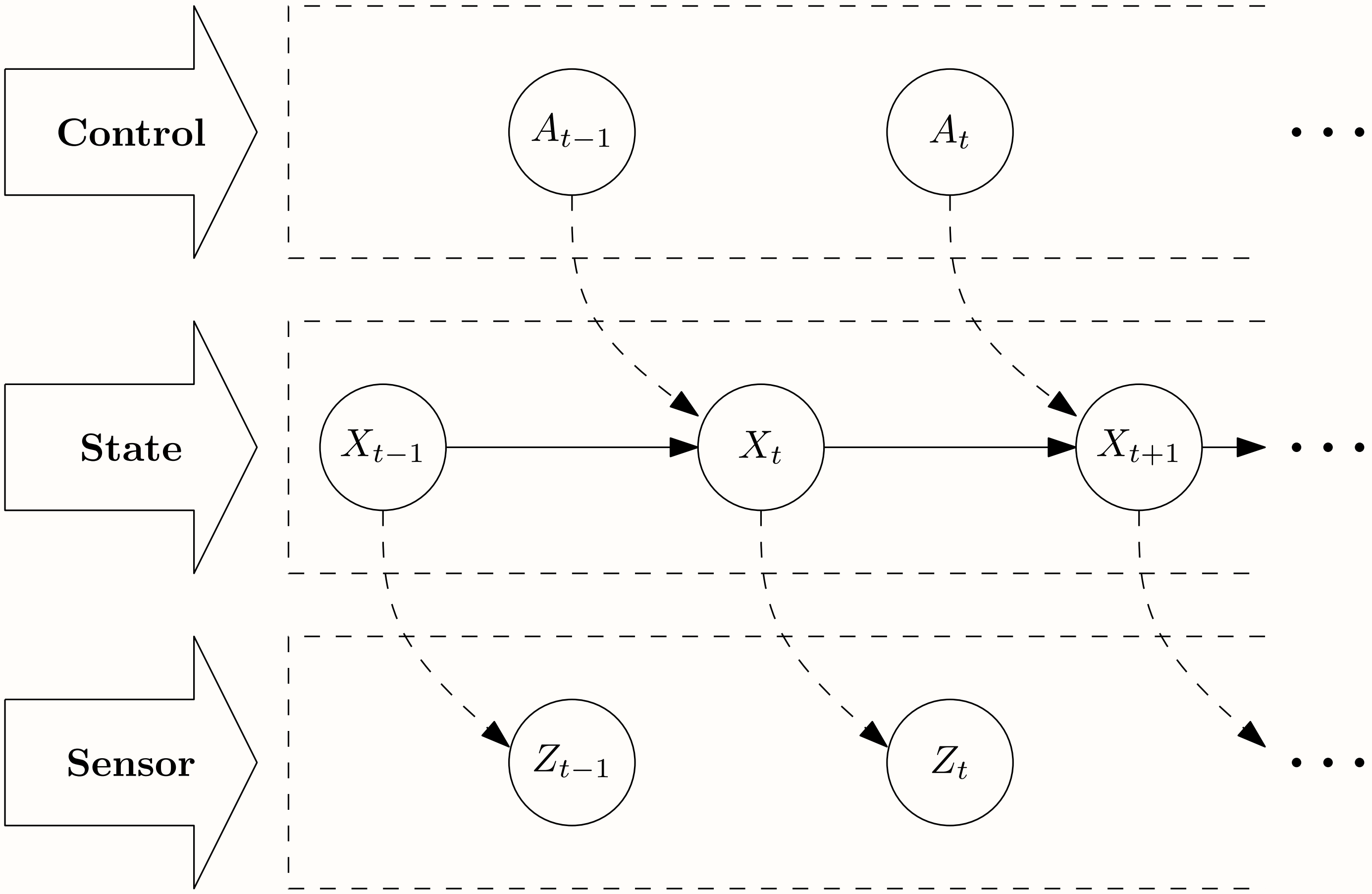
$$J_{\text{av}} = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} R_t\right]; \qquad J_{\text{disc}} = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R_t\right].$$

Considering partial observability

A *partially observable MDP* (POMDP) is an MDP where the decision maker is not able to access all information relevant to the decision-making process (Kaelbling et al., 1998).

- The decision-maker receives an observation Z_t for each time instant t;
- The observation depends on the state of the underlying Markov chain;

Considering partial observability (1)



Multiple decision-makers

- *Stochastic games* (aka Markov games) provide a multi-agent generalization of MDPs (Shapley, 1953);
- In stochastic games, the control parameter depends on the choice of several *independent* decision-makers;
- In stochastic games, each decision-maker (k) can receive a different reward R_t^k at each time instant t.

Multiple decision-makers (1)

In stochastic games, as in MDPs,

- Each decision-maker (k) must optimize its own long-run optimality criterion, e.g.,

$$J_{\text{av}}^k = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} R_t^k\right]; \qquad J_{\text{disc}}^k = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R_t^k\right];$$

- Partial state observability can be considered, leading to the framework of *partially observable stochastic games* (POSGs).

Multiagent models

Fully observable:

- Multiagent MDPs (Boutilier, 1996).

Partially observable:

- Partially observable stochastic games (Hansen et al., 2004).
- Decentralized POMDPs (Bernstein et al., 2002).
- Interactive POMDPs (Gmytrasiewicz and Doshi, 2005).
- Each agent only observes its own observation.

Solution methods: MDPs

Model based

- Basic: dynamic programming (Bellman, 1957), value iteration, policy iteration.
- Advanced: prioritized sweeping, function approximators.

Model free, reinforcement learning (Sutton and Barto, 1998)

- Basic: Q-learning, TD(λ), SARSA, actor-critic.
- Advanced: generalization in infinite state spaces, exploration/exploitation issues.

Techniques for partially observable environments

Model based (POMDP)

- Exact methods (Monahan, 1982; Cheng, 1988; Cassandra et al., 1994; Zhang and Liu, 1996)
- Heuristic methods: based on MDP solution.
- Approximate methods: gradient descent, policy search, point-based techniques.

Other topics

- Predictive State Representations (Littman et al., 2002).
- Reinforcement learning in POMDPs, PSRs.

Multiagent methods

Model based:

- Hansen et al. (2004)'s dynamic programming.
- JESP (Nair et al., 2003).
- Bayesian game approximation (Emery-Montemerlo et al., 2004).

Model free:

- Minimax-Q (Littman, 1994)
- FriendFoe-Q (Littman, 2001)
- Nash-Q, multi-agent DYNA-Q, correlated-Q.
- Learning coordination.

Reading group

Questions to be answered:

- What topics shall we cover?
- When shall we meet? How often?
- Schedule, volunteers?

References

J. A. Bagnell and J. G. Schneider. Autonomous helicopter control using reinforcement learning policy search methods. In *Proceedings of the 2001 IEEE International Conference on Robotics and Automation*, pages 1615–1620, 2001.

R. Bellman. *Dynamic programming*. Princeton University Press, 1957.

D. C. Bentivegna, A. Ude, C. G. Atkeson, and G. Cheng. Humanoid robot learning and game playing using PC-based vision. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'02)*, pages 2449–2454, October 2002.

D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA, 2nd edition, 2000.

C. Boutilier. Planning, learning and coordination in multiagent decision processes. In *Theoretical Aspects of Rationality and Knowledge*, 1996.

A. R. Cassandra, L. P. Kaelbling, and M. L. Littman. Acting optimally in partially observable stochastic domains. In *Proc. of the National Conference on Artificial Intelligence*, 1994.

H. T. Cheng. *Algorithms for partially observable Markov decision processes*. PhD thesis, University of British Columbia, 1988.

R. Emery-Montemerlo, G. Gordon, J. Schneider, and S. Thrun. Approximate solutions for partially observable stochastic games with common payoffs. In *Proc. of Int. Joint Conference on Autonomous Agents and Multi Agent Systems*, 2004.

P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

E. A. Hansen, D. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *Proc. of the National Conference on Artificial Intelligence*, 2004.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

N. Kohl and P. Stone. Machine learning for fast quadrupedal locomotion. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI'04)*, pages 611–616, July 2004a.

N. Kohl and P. Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation (ICRA'04)*, pages 2619–2624, May 2004b.

M. L. Littman. Friend-or-foe q-learning in general-sum games. In *International Conference on Machine Learning*, 2001.

M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 1994.

M. L. Littman, R. S. Sutton, and S. Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

G. E. Monahan. A survey of partially observable Markov decision processes: theory, models and algorithms. *Management Science*, 28(1), Jan. 1982.

R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *Proc. Int. Joint Conf. on Artificial Intelligence*, 2003.

A. Y. Ng, A. Coates, M. Diel, V. Ganapathi, J. Schulte, B. Tse, E. Berger, and E. Liang. Inverted autonomous helicopter flight via reinforcement learning. In *Proceedings of the 2004 International Symposium on Experimental Robotics (ISER'04)*, 2004.

M. L. Puterman. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

S. J. Russell and P. Norvig. *Artificial Intelligence: a modern approach*. Prentice Hall, 2nd edition, 2003.

M. Saggar, T. D'Silva, N. Kohl, and P. Stone. Autonomous learning of stable quadruped locomotion. In *Proceedings of the 2006 International RoboCup Symposium (to appear)*, 2006.

L. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39:1095–1100, 1953.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

N. L. Zhang and W. Liu. Planning in stochastic domains: problem characteristics and approximations. Technical Report HKUST-CS96-31, Department of Computer Science, The Hong Kong University of Science and Technology, 1996.