RL 18: A Unified View

Michael Herrmann

University of Edinburgh, School of Informatics

25/03/2014

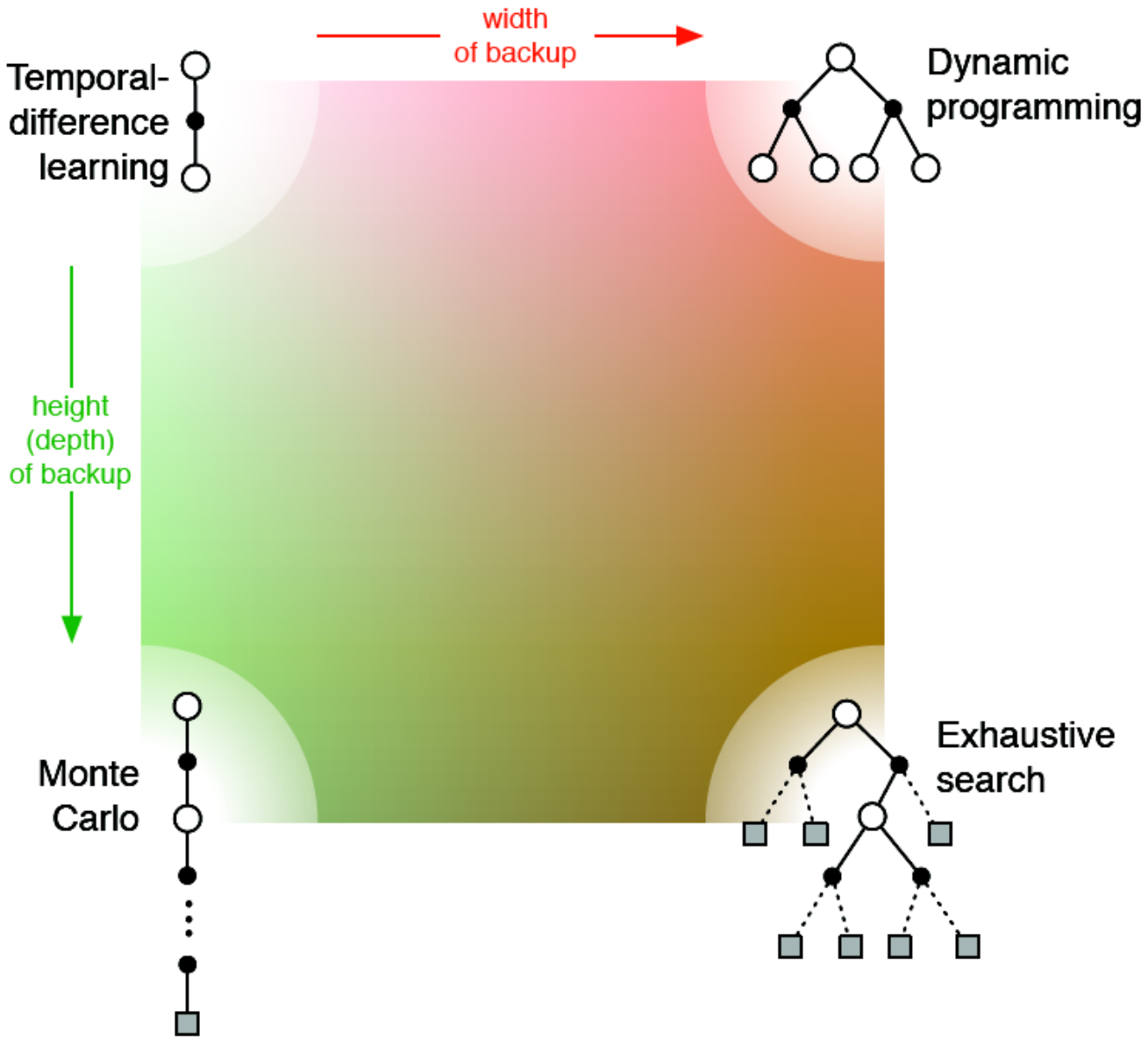
The Unified View

Three main features of RL learning algorithms

1. Estimation of value functions
2. Backing up values along actual or possible state trajectories
3. Generalised policy iteration (GPI): Maintain an approximate value function and an approximate policy, and they continually try to improve each on the basis of the other.

s. S&B, 2nd ed., 16.1

Dimensions of Reinforcement Learning



s. S&B, 2nd ed., 16.1

Exploration and Exploitation

One of the most important issues for the temporal difference learning algorithms is maintaining a balance between exploration and exploitation.

That is, the agent must sometimes choose actions that it believes to be suboptimal in order to find out whether they might actually be good.

This is particularly true in problems which change over time (which is true of most behavioural experiments), since actions that used to be good might become bad, and vice-versa.

The theorems proving that temporal difference methods work usually require much experimentation: all actions must be repeatedly tried in all states.

In practice, it is common to choose policies that always embody exploration (such as choosing a random action some small fraction of the time, but otherwise the action currently believed to be best)

P Dayan, CJCH Watkins (1999) Reinforcement Learning. for *Encyclopedia of Cognitive Science*.

Further Dimensions of RL

- Exploration
  - Relation to back-up
    - On-policy
    - Off-policy
  - Choice of exploratory actions
    - ε-greedy: Random (sometimes)
    - Boltzmann (soft-max): Biased by earlier experience
    - Optimistic initialisation
    - Self-motivated
- Function approximation
  - Tabular
  - State aggregation
  - Linear methods
  - Nonlinear methods

Further Dimensions of RL (I)

- **Definition of return**
  - episodic or continuing
  - discounted or undiscounted
- **Action values vs. state values vs. afterstate values**
  - What kind of values should be estimated?
  - If only state values are estimated, then either a model or a separate policy is required for action selection.
- **Synchronous vs. asynchronous**
  - Are the backups for all states performed simultaneously or one by one in some order?
- **Replacing vs. accumulating traces**
  - If eligibility traces are used, which kind is most appropriate?

Further Dimensions of RL (II)

- **Real vs. simulated**
  - Should one backup real experience or simulated experience?
  - If both, how much of each?
- **Location of backups**
  - What states or state-action pairs should be backed up? Model-free methods can choose only among the states and state-action pairs actually encountered, but model-based methods can choose arbitrarily.
- **Timing of backups**
  - Should backups be done as part of selecting actions, or only afterward?
- **Memory for backups**
  - How long should backed-up values be retained? Should they be retained permanently, or only while computing an action selection, as in heuristic search?

Non-stationary RL

- In on-line algorithms, the agent is allowed to gather experience in the real-world system. Information about the system becomes available gradually with time.
- This is in contrast to simulation-based applications in which the distributions of the underlying random variables are assumed to be known.
- If the distribution functions of the underlying random variables cannot be estimated accurately and if trial runs of the real-world system are not too expensive, on-line algorithms are more suitable.
- Attempts at the problem for if transition probabilities
  - are noisy (Givan et al., 2000; Satia and Lave, 1973; White and Eldeib, 1994)
  - change with time (Szita et al., 2002)

A. Gosavi

Current research (according to wikipedia)

- Learning and acting under partial information, e.g., using predictive state representation
  - Littman, Michael L.; Richard S. Sutton; Satinder Singh (2002). "Predictive representations of state. NIPS 14, 1555-1561.
  - S. Singh, M. R. James, M. R. Rudary (2004) Predictive state representations: A new theory for modeling dynamical systems. Proc. 20th Conf. Uncertainty in AI, 512-519.
- Adaptive methods which work with fewer parameters, parameter optimisation
- Scaling: Improving existing value-function and policy search methods for large or continuous action spaces
- Modular and hierarchical reinforcement learning
- Transfer learning
- Lifelong learning
- Multiagent or distributed reinforcement learning

A few more interesting aspects (JFYI)

- Evolutionary RL: Grefenstette, J. J., Moriarty, D. E., & Schultz, A. C. (2011). Evolutionary algorithms for reinforcement learning. arXiv preprint arXiv:1106.0221.
- ML-based improvements
  - Linear-programming-based approaches
  - Bayesian RL
  - Kernel-based RL
- Multi-objective RL
- Hybrid algorithms

A few more interesting aspects (JFYI)

- S. Thiebaux, C. Gretton, J. Slaney, D. Price and F. Kabanza (2006) Decision-theoretic planning with non-Markovian rewards. J. Artif. Intell. Res. **25**, 17-74.
- Trung Thanh Nguyen, Zhuoru Li, Tomi Silander and Tze-Yun Leong (2013) Online feature selection for model-based reinforcement learning. *ICML*.
- J. Asmuth, L. Li, M. L. Littman, A. Nouri, D. Wingate (2009) A Bayesian sampling approach to exploration in reinforcement learning. *25th Conf. Uncertainty in AI*, 19-26.
- B. C. Silva and A. G. Barto (2012) TD-DeltaPi: A model-free algorithm for efficient exploration. *26th AAAI Conf. on AI*.
- E. Hazan and C. Seshadhri (2009) Efficient learning algorithms for changing environments. *ICML*.
- Reinforcement Learning based on human-generated rewards http://www.cs.utexas.edu/~pstone/Papers/bib2html-links/iui13-knox.pdf

Successes of reinforcement learning
in real-life applications

- http://umichrl.pbworks.com/w/page/7597597/Successes of Reinforcement Learning
- http://rl-community.org/wiki/Successes_Of_RL

Acknowledgements & References

Most of the material of today's lecture was adapted from Sutton and Barto's Reinforcement Learning book (draft of 2nd edition).

See also: umichrl.pbworks.com/w/page/7597585/Myths of Reinforcement Learning

More literature (review articles, JFYI):

Panait, L., & Luke, S. (2005) Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems* **11**:3, 387-434.

Busoniu, Lucian, Robert Babuska, and Bart De Schutter (2008) A comprehensive survey of multiagent reinforcement learning." *IEEE Transactions on Systems, Man, and Cybernetics*, **38**:2, 156-172.

Singh, S., Barto, A. G., & Chentanez, N. (2005) Intrinsically motivated reinforcement learning. Defence Technical Information Center.

F. L. Lewis, K. G. Vamvoudakis (2011) Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data. *IEEE TA on Systems, Man, and Cybernetics* **41**:1, 14-25.