Learning for exploration/exploitation in reinforcement learning

Castronovo Michael

University of Liège, Belgium

26th of June 2012
Master Thesis Presentation

Introduction

Classical RL problem:

- Single trajectory
- Discounted rewards
- Infinite horizon
- Discrete state/action spaces

This problem is known to be difficult to address, except with a high discount factor or rather small state/action spaces.

How to improve the efficiency of actual techniques ?

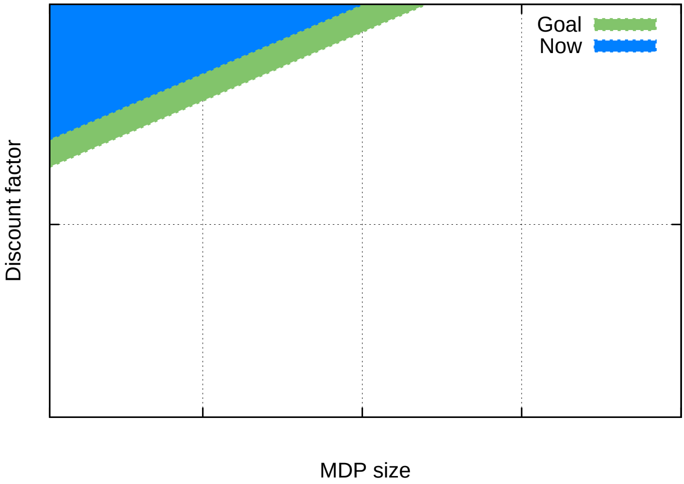
Adding the prior knowledge on the MDP to be played.

- Not actually used
- Available for most applications
- Specific to each type of problem

This can be represented by the knowledge of the distribution from which the MDP to be played will be drawn.

Goal: Discovering new E/E strategies which works better than usual techniques on this distribution.

E/E strategies significantly better than Random:



How?

Defining a rich set of E/E strategies, and searching for the best one in average according to the given MDP distribution.

The chosen approach consists in defining E/E strategies based on short formulas.

Why?

- Simple to define very large spaces of strategies
- Good interpretability
- Easy to use

Background

Let be

- $M = (\mathcal{S}, \mathcal{A}, p_{M,f}(\cdot), \rho_M, p_{M,0}(\cdot), \gamma)$, a MDP
- $\mathcal{S} = \left\{s^{(1)}, \dots, s^{(n_\mathcal{S})}\right\}$, its state space
- $\mathcal{A} = \left\{a^{(1)}, \dots, a^{(n_\mathcal{A})}\right\}$, its action space
- $s_{t+1} \sim p_{M,f}(\cdot | s_t, a_t)$, the transition law (stochastic)
- $r_t = \rho_M(s_t, a_t, s_{t+1})$, the reward distribution (deterministic)

An history $H_t = [s_0, a_0, r_0, \ldots, s_t, a_t, r_t]$ is a vector that gathers the history over the first t steps.

An E/E strategy π:

$$a_t \in \mathcal{A} : a_t \sim \pi(H_{t-1}, s_t)$$

The stochastic discounted return of π:

$$\mathcal{R}^{\pi}_{M}(s_0) = \sum_{t=0}^{\infty} \gamma^t r_t \ ,$$

The average stochastic discounted return of π:

$$J_M^\pi = \mathop{\mathbb{E}}_{p_{M,0}(\cdot), p_{M,f}(\cdot)} [\mathcal{R}_M^\pi(s_0)]$$

The best E/E strategy π, given the prior $p_{\mathcal{M}}(\cdot)$ is the one maximizing:

$$J^\pi = \mathop{\mathbb{E}}_{M' \sim p_{\mathcal{M}}(\cdot)} [J_{M'}^\pi] \ .$$

Formula-based E/E strategies

A formula-based E/E strategy is using a function, ranking each action (like an index-based strategy), in order to choose the next action to perform:

$$\pi^F(H_{t-1}, s_t) \in \arg\max_{a \in \mathcal{A}} F\left(\hat{\rho}(s_t, a), N(s_t, a), \hat{Q}(s_t, a), \hat{V}(s_t), t, \gamma^t\right)$$

The set of all formulas of size K or less is denoted by $\mathbb{F}_{\mathcal{M}}^{K}$ (discrete set).

Finding a high-performance formula-based E/E strategy for a given class of MDPs

Reducing $\mathbb{F}_{\mathcal{M}}^{K}$

Several formulas can lead to the same policy
$\Rightarrow$ Reduction of $\mathbb{F}_{\mathcal{M}}^{K}$ is necessary.

We partition the set $\mathbb{F}_{\mathcal{M}}^{K}$ into equivalence classes, two formulas being equivalent if and only if they lead to the same policy.

For each equivalence class, we then consider one member of minimal length, and we gather all those minimal members into a set $\bar{\mathbb{F}}_{\mathcal{M}}^{K}$.

Since such a set is difficult to compute. Let $\tilde{\mathbb{F}}_{\mathcal{M}}^{K}$ be an approximation of $\bar{\mathbb{F}}_{\mathcal{M}}^{K}$.

Finding a high-performance formula-based E/E strategy for a given class of MDPs

Finding a high-performance formula

Using Monte-Carlo simulations for each formula could reveal itself to be time-inefficient in case of spaces $\tilde{\mathbb{F}}_{\mathcal{M}}^{K}$ of large cardinality.

$\Rightarrow$ Formalizing this research as a $N-$armed bandit problem.

To each formula $F_n \in \tilde{\mathbb{F}}^K_{\mathcal{M}}$ ($n \in \{1, \ldots, N\}$), we associate an arm.

Pulling the arm n consists first in randomly drawing a MDP M according to $p_{\mathcal{M}}(\cdot)$ and an initial state s_0 for this MDP according to $p_{M,0}(\cdot)$.

The reward associated to arm n is the empirical discounted return $\mathcal{R}^\pi_M(s_0)$.

The bandit problem is used to identify high-quality formula(s).

Experimental results

Random MDPs:

- $|\mathcal{S}| = 20$, $|\mathcal{A}| = 5$, $\gamma = 0.995$
- For each state-action pair, there is $0.1\,|\mathcal{S}|$ reachable states (2 for $|\mathcal{S}| = 20$).
- Each transition provides a constant reward, randomly chosen in $]0;1]$ at the MDP generation.

Formula space ($K = 5$):

- Variables:
  $\hat{\rho}(s_t, a),\ N(s_t, a),\ \hat{Q}(s_t, a),\ \hat{V}(s_t),\ t,\ \gamma^t$
- Constants:
  $1,\ 2,\ 3,\ 5,\ 7$
- Operators:
  $+,\ -,\ \times,\ /,\ |\,.\,|,\ \log(.),\ \sqrt{.},\ \min(.,\,.),\ \max(.,\,.)$

Baselines		Learned strategies	
Name	J^π	Formula	J^π
Optimal	65.3	$(N(s,a) \times \hat{Q}(s,a)) - N(s,a)$	30.3
Random	10.1	$\max(1, (N(s,a) \times \hat{Q}(s,a)))$	22.6
Greedy	20.0	$\hat{Q}(s,a)$ (= Greedy)	20.0
ϵ-Greedy($\epsilon = 0$)	20.0	$\min(\gamma^t, (\hat{Q}(s,a) - \hat{V}(s)))$	19.4
R-max ($m = 1$)	27.7	$\min(\hat{\rho}(s,a), (\hat{Q}(s,a) - \hat{V}(s)))$	19.4

Table: Performance of the top-5 learned strategies with respect to baseline strategies.

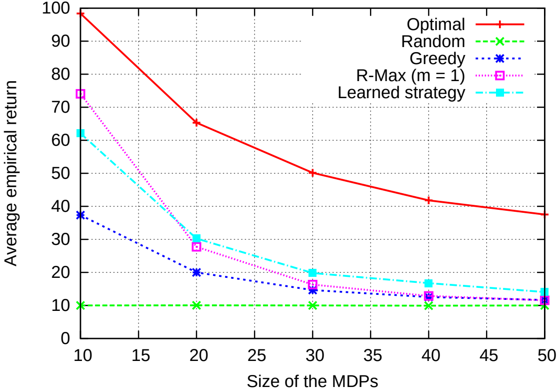


Figure: Performances of the learned and the baseline strategies for different distributions of MDPs that differ by the size of the MDPs belonging to their support.

Conclusions

- We outperformed usual approaches
- ... even on larger MDPs (good robustness)

Further improvements:

- Approximating $\bar{\mathbb{F}}_{\mathcal{M}}^{K}$ more precisely
- Considering larger and/or continuous formula spaces
- Generalizing the approach to continuous state/action spaces

References I

J. Asmuth, L. Li, M.L. Littman, A. Nouri, and D. Wingate.
A Bayesian sampling approach to exploration in reinforcement learning.
In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 19–26. AUAI Press, 2009.

J.Y. Audibert, R. Munos, and C. Szepesvári.
Tuning bandit algorithms in stochastic environments.
In *Algorithmic Learning Theory*, pages 150–165. Springer, 2007.

P. Auer and R. Ortner.
Logarithmic online regret bounds for undiscounted reinforcement learning.
In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, page 49. The MIT Press, 2007.

P. Auer, N. Cesa-Bianchi, and P. Fischer.
Finite-time analysis of the multiarmed bandit problem.
Machine learning, 47(2):235–256, 2002.

R.I. Brafman and M. Tennenholtz.
R-max – a general polynomial time algorithm for near-optimal reinforcement learning.
The Journal of Machine Learning Research, 3:213–231, 2002.

T. Jaksch, R. Ortner, and P. Auer.
Near-optimal regret bounds for reinforcement learning.
The Journal of Machine Learning Research, 11:1563–1600, 2010.

References II

M. Kearn and S. Singh.
Near-optimal reinforcement learning in polynomial time.
Machine Learning, 49:209–232, 2002.

F. Maes, L. Wehenkel, and D. Ernst.
Automatic discovery of ranking formulas for playing with multi-armed bandits.
In *9th European workshop on reinforcement learning*, Athens, Greece, September 2011.

F. Maes, L. Wehenkel, and D. Ernst.
Learning to play K-armed bandit problems.
In *International Conference on Agents and Artificial Intelligence*, Vilamoura, Algarve, Portugal, February 2012.

P. Poupart, N. Vlassis, J. Hoey, and K. Regan.
An analytic solution to discrete Bayesian reinforcement learning.
In *Proceedings of the 23rd international conference on Machine learning*, pages 697–704. ACM, 2006.

C.J. Watkins and P. Dayan.
Q-learning.
Machine Learning, 8(3-4):179–192, 1992.