Regret Bounds for Learning State Representations in Reinforcement Learning

R. Ortner[1] M. Pirotta[2] A. Lazaric[2] R. Fruit[3] O. Maillard[3]

[1]Montanuniversität Leoben [2]Facebook AI Research [3]Sequel Team – INRIA Lille

Abstract

Selecting or designing the state representation is a well-known problem in RL. There are many approaches for feature extraction from high-dimensional observations. Not all these features describe the problem well or show Markovian dynamics.

We consider the *feature RL problem*: given a finite set Φ of models, we have to select online the appropriate model to solve the task. The learning efficiency is measured in terms of speed of convergence toward the optimal solution (i.e., regret). We introduce Ucb-Ms, an optimistic elimination algorithm that performs efficient exploration of the representations.

Example

Planar Navigation: set of state variables

$$\{x_0, y_0, v_0, \dots, x_t, y_t, v_t, goal, \dots\}$$

(position at time t: x_t, y_t; velocity at time t: v_t)

multiple state representations:

- position, orientation, velocity
- current position, previous position, orientation
- position
- last N observations

⚠ Not all these representations:
– are correctly modeling the system
– induce a Markov model

What is the best representation for learning the optimal policy?

I) multiple redundant sensory measures
II) difficult also for experts to identify important variables

Settings

Online Learning

For time $t = 1, 2, \dots$

- execute action $a_t \sim \pi$
- observe reward r_t and observation o_{t+1}

History:
$\mathcal{H}_t = (o_1, a_1, r_1, o_2, \dots, a_t, r_t, o_{t+1})$

State Models

- State-representation model (in short model):

$$\phi : \mathcal{H} \to \mathcal{S}_\phi$$

ϕ? can be an embedding from different neural networks and/or RNN

A state-rep ϕ is Markov if induces an MDP $M(\phi)$

Markov Decision Process (MDP)

$$M = (\mathcal{S}, \mathcal{A}, p, r)$$

Markov:
$P(s_{t+1}, r_t | h_t, a_t) = P(s_{t+1}, r_t | s_t, a_t)$

- Average reward

$$\rho^\pi(M) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T r_t | \pi, M\right]$$

- Optimal policy

$$\pi^\star = \operatorname{argmax}_\pi \rho^\pi(M)$$

- Diameter

$$D(M) = \max_{s\neq s'} \min_\pi \mathbb{E}[\tau_\pi(s, s')]$$

$\tau_\pi(s, s')$ is the first hitting time of s' starting from s

Problem

- Online learning
- The learner has a finite set Φ of state-rep. models
- At least one model $\phi_0 \in \Phi$ is Markov

Goal

Find a policy that performs as well as

$$\pi^\star \in \operatorname{argmax}_\pi \rho^\pi(M(\phi_0))$$

⚠ $\pi^\star, M(\phi_0)$ unknown

▶ Regret

$$R(T) = T\rho^\star(\phi_0) - \sum_{t=1}^T r_t$$

Why is it important?

- Difficult to know a priori if a representation is "reasonable"
- Automatically and quickly discard bad representation
- What is the best representation for learning the optimal policy?
- Learning directly online

solution idea

quickly discard "bad" representations and keep following the optimal policy of models that ***perform*** *well enough*

Ucb-Ms Algorithm

For episodes $k = 1, 2, \dots$

1. For each rep. $\phi \in \Phi_k$, compute optimistic policy $\widetilde{\pi}_{k,\phi}$

$$(\widetilde{M}_{k,\phi}, \widetilde{\pi}_{k,\phi}) = \operatorname*{argmax}_{M\in\mathcal{M}_{k,\phi}, \pi\in\Pi_\phi} \rho^\pi(M)$$

(Maximum average reward accordingly to the uncertainty on the model induced by ϕ)

2. Choose best (optimistic) model:

$$\phi_k = \operatorname*{argmax}_{\phi\in\Phi_k}\{\widetilde{\rho}_{k,\phi}\} = \operatorname*{argmax}_{\phi\in\Phi_k}\{\rho^{\pi_{k,\phi}}(\widetilde{M}_{k,\phi})\}$$

3. Execute best policy $\widetilde{\pi}_{k,\phi_k}$
Repeat until end of episode:
   - Choose action $a_t \sim \widetilde{\pi}_{k,\phi_k}(s_t)$, get reward r_t and observe next state $s_{t+1} = \phi_k(o_{t+1}) \in \mathcal{S}_k$
   - if

$$(t - t_k + 1)\widetilde{\rho}_{k,\phi} - \sum_{t=t_k}^t r_\tau \geq \Gamma_t(\bar{D}) \quad (1)$$

then $\Phi_{k+1} = \Phi_k \setminus \{\phi_k\}$ and terminate episode

$$\Gamma_t(D) \approx DS_{\phi_t}\sum_{s,a}\frac{\nu_{k_t}(s,a)}{\sqrt{N_{k_t}(s,a)}} + D\sqrt{T_{k_t}}$$

Step 1) As UCRL2 [Jaksch et al., 2010], Ucb-Ms builds uncertainty about transitions and rewards for each $\phi \in \Phi$ (i.e., set of plausible MDPs)

Step 2) Optimism at the level of representations

Step 3) Discards models that are not achieving enough reward as "promised"

- Eq. 1 is a **bound on the regret** of a single episode of UCRL2
- An indication that the model is **not Markov**

Guarantees

With probability $1-\delta$, the regret of Ucb-Ms using $\bar{D} \geq D$ is

$$R(T) \leq const \cdot \bar{D}\sqrt{S_{\max}S_\Sigma AT}\log(T\delta)$$

where $S_{\max} = \max_\phi S_\phi$ and $S_\Sigma = \sum_\phi S_\phi$.

- Ucb-Ms adapts to most preferable model
- Needs the knowledge of upper-bound on the true diameter, i.e., $\bar{D} \geq D$
- It reduces to UCRL2 when there is a single model: $\widetilde{O}(DS\sqrt{AT})$
- No need to exactly identify the true model ϕ_0
  - if a non-Markovian model gives as much as reward of a Markovian one $\implies$ no need of discarding it
- Improves regret compared to the state of the art (e.g., [Ortner et al., 2014])

Extensions

▶ Unknown diameter: $\implies$ doubling trick

- if all the models are eliminated
- double the estimate of the diameter

The only difference in the regret is an additional term $\lceil|\Phi|\log_2(D)\rceil$

▶ Effective size S_Φ:
- The entire state space can be covered with only S_Φ states using Φ
- Examples: hierarchical structure

- regret bound scales with S_Φ rather than S_Σ
- $S_\Sigma \gg S_\Phi$

▶ Unknown diameter and state size:

- estimate directly the term DS
- use similar doubling trick
- the regret is upper-bounded by

$$const \cdot DS_{\phi^0}\sqrt{(S_\Phi AT + |\Phi|\log(DS_{\phi^0}))AT}\log(T\delta)$$

References

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Ronald Ortner, Odalric-Ambrym Maillard, and Daniil Ryabko. Selecting near-optimal approximate state representations in reinforcement learning. In *Algorithmic Learning Theory - 25th International Conference, ALT 2014*, pages 140–154, 2014.