Reinforcement Learning in Rich-Observation MDPs using Spectral Methods

Kamyar Azizzadenesheli, Alessandro Lazaric, Animashree Anandkumar

*University of California, Irvine (UCI) †Institut National de Recherche en Informatique et en Automatique, (INRIA)

Reinforcement Learning

Agent-Environment interactions under uncertainty:
- Policy \( \pi(a|y) \) : \( X \rightarrow A \)
- Goal: \( \max \theta = \max \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} r_i \)
- No prior knowledge
- Learning (Exploring)
- Planning (Exploiting)
- Undiscounted average reward

SL-UC

Start

Initial policy \( \pi \)

Observe current \( y \)

Apply the action on the environment \( \pi(a|y) \)

Observe reward \( r \)

Is the number of samples at least for one pair of \( (s, a) \) doubled?

Yes

No

Find the optimal policy w.r.t. optimistic aux-model

Construct the plausible set of Aux-MDP

Observation \( y \)

Action \( a \)

Reward \( r \)

State \( s \)

Large MDPs

Structured MDPs
- Rich-Observation MDP (ROMDP)
- Injective mapping from \( s \) to \( y \)
- Known mapping \( \Rightarrow \text{Regret}(T) = \tilde{O}\left(D_T X \sqrt{AT}\right) \)
- No Prior knowledge \( \Rightarrow \) Learn the mapping

Tensor Decomposition:
- Multiview model condition on middle action and middle state

Tensor Moments
- \( V_i = \mathbb{P}(y|s_i, a_i = 1) \)
- \( V_i(0) = \mathbb{E}(y|s_i, a_i = 1) \in \mathbb{R}^{Y \times X} \)

\( E[\omega_1 \otimes \omega_2 \otimes \nu_1 \otimes \nu_2 | s_i] = \sum_{a_i \in A} \omega_1(a_i) \otimes V_i(0) \)

Parameter Learning

Second and Third order moments given middle action

\( M_i^{(2)} = \sum_{a_i \in A} \omega_1(a_i) \otimes V_i(0) \otimes V_i(0) \)

\( M_i^{(3)} = \sum_{a_i \in A} \omega_1(a_i) \otimes V_i(0) \otimes V_i(0) \otimes V_i(0) \)

Confidence intervals

\( \| \hat{O}_{i} - \hat{O}_{i} \| = \tilde{O}\left(\frac{\sqrt{Y \cdot T}}{\sqrt{N}}\right) \)

Multiview Model

Random ROMDPs: \( X = 5, A = 4, Y = 20 \)

Random MDPs: \( X = 10, 20, 30 \), and \( A = 4 \)

Results

Theorem: SL-UC achieves a regret of

\( \text{Regret}(T) = \tilde{O}\left(D_T X \sqrt{AT}\right) \)

- Observation independent regret,
- Optimal regret (UCRL)

\( \text{Regret}(T) = \tilde{O}\left(D_T Y \sqrt{AT}\right) \)

- Per epoch computation reduction \( O(Y^3) \rightarrow O(X) \)
- Linearly reducing the number of epochs.

Clustering rate: A random ROMDP \( (X = 4, A = 4, Y = 20) \)

Average Reward

Random MDPs: with \( X = [10, 20, 30] \), and \( A = 4 \)

UCRL: The Optimal algorithm
- DQN: 3 hidden layers, 30 hyperbolic tangent units at each layer with RMSprop update
- Gridworld: [Johnson et al., 2016]