Efficient Reinforcement Learning in Factored MDPs with Application to Constrained RL

Xiaoyu Chen[1], Jiachen Hu[1], Lihong Li[2] & Liwei Wang[1]

[1]Peking University

[2]Amazon

May 3rd, 2021

Tabular Episodic MDP

- For tabular MDPs, the regret bounds typically depend polynomially on the cardinality of state and action space.
- The matching lower bounds imply that these results cannot be improved without additional assumptions.

Tabular Episodic MDP

- For tabular MDPs, the regret bounds typically depend polynomially on the cardinality of state and action space.
- The matching lower bounds imply that these results cannot be improved without additional assumptions.

Can we take advantage of specific structures to develop more efficient algorithms?

Factored MDPs

A factored MDP is an MDP whose rewards and transitions exhibit certain conditional independence structures.

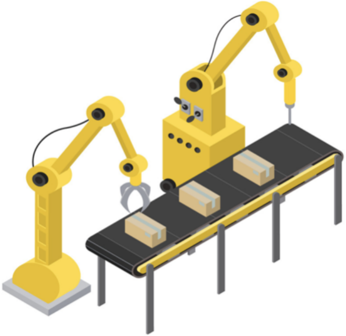
- Reward function: the average of m factored rewards.
  - Each factored reward only depends on a state-action subspace with maximum cardinality J^R.
- Transition function: the multiplication of n factored transitions.
  - Each factored transition only depends on a state-action subspace with maximum cardinality J^P.

We assume the factored structures are known beforehand, and we only need to learn the reward function and transition dynamics.

Motivating Example

A large production line with n machines in sequence:

- $\mathcal{S} \times \mathcal{A} = \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, where $\mathcal{X}_i$ is the state-action subspace of machine i.
- The one-step transition dynamics of each machine can only be influenced by its neighboring machines.



Main Results

FMDP-BF algorithm follows the principle of "optimism in the face of uncertainty".

- We construct the confidence bonus for each factored rewards and transitions separately.
- We maintain both the optimistic and pessimistic value estimation in each episode.

Main Results

FMDP-BF algorithm follows the principle of "optimism in the face of uncertainty".

- We construct the confidence bonus for each factored rewards and transitions separately.
- We maintain both the optimistic and pessimistic value estimation in each episode.

The regret of FMDP-BF scales as $\tilde{O}(\sqrt{J^R T} + \sqrt{nHJ^P T})$

- J^R and J^P is the maximum cardinality of each factored state-action subspace.
- Improving on the previous result of Osband et al. by a factor of $\sqrt{nH|\mathcal{S}_i|}$
- Nearly matching compared with the lower bound we proved.

RLwK

We formulated a natural constrained RL setting called RLwK and applied FMDP-BF to this problem.

- Besides receiving rewards, the agent also suffers a d-dimensional cost c_h sampled from the cost distribution at step h.
- An episode terminates after H steps, or when the cumulative cost of any dimension i exceeds the maximum budget B, whichever occurs first.

- We show that RLwK setting is fundamentally different with previous constrained RL settings by two specific MDP instances.

Summary

- We propose an efficient algorithm called FMDP-BF with near-optimal regret guarantee.
- We formulate a natural constrained RL setting called RLwK, and apply our algorithm to this setting.

Summary

- We propose an efficient algorithm called FMDP-BF with near-optimal regret guarantee.
- We formulate a natural constrained RL setting called RLwK, and apply our algorithm to this setting.

- The regret upper and lower bounds have a gap of approximately $\sqrt{n}$, where n is the number of transition factors.
- For RLwK, our algorithm is computationally inefficient.