Sample-Optimal Parametric Q-Learning Using Linearly Additive Features

Lin F. Yang, Mengdi Wang (Princeton University)

Discounted Markov Decision Process:

- a set of states S
- a set of actions A
- a discount factor $\gamma \in (0,1)$
- a transition probability $P(\cdot\,|s,a)$ at each $s \in S$ and $a \in A$
- a reward function $r(s,a) \in [0,1]$

Goal: find a good *policy* $\pi: S \to A$, such that the following expected reward is at most ϵ-away from the maximum possible

$$\forall s:\ V^{\pi}(s) \coloneqq E\left[\sum_{t=0}^{\infty} \gamma^t r\left(s^t, \pi(s^t)\right) | s^0 = s\right] \geq V^*(s) - \epsilon$$

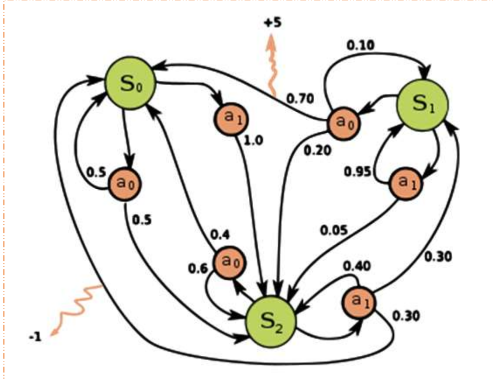
At any state s, an agent plays an action a, the agent will go to the next state s' with some probability $P(s'|s,a)$ and at the same time receive reward $r(s,a)$.

Curse of dimensionality:

- *Go game: #states* $\sim 3^{361}$
- *Autonomous driving: #states* $\sim$ *inifinity*

A basic model: generative model

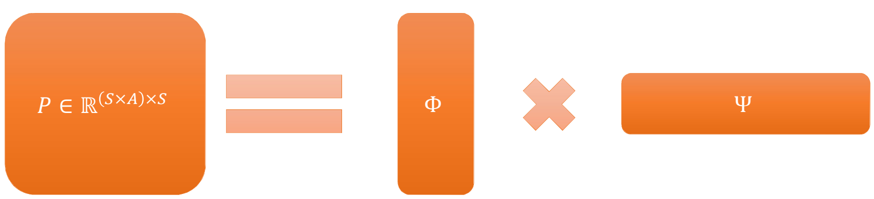
- The agent can query as many samples as possible from any (s,a).
- Each sample costs $O(1)$ time to obtain.



Assumption 1: *features for the transition kernel*

$$P(s'|s,a) = \sum_{k\in[K]} \psi_k(s')\phi_k(s,a)$$

ϕ_k: known features for state-action pairs
ψ_k: unknown linear coefficients



How to use features for provably efficient policy-learning in RL?

- **Q1:** How many observations of state-action-state transitions are necessary for finding an ϵ-optimal policy?
- **Q2:** How many samples are sufficient for finding an ϵ-optimal policy with high probability and how to find it?

Equivalence to linear function approximator and advantages:

- Assumption 1 is equivalent to assuming linear function-approximators of the optimal Q-function with **zero** Bellman-error

Algorithm 1: provable dimension reduction with a parametric Q-learning method

Represent Q-function with parameter w:

$$Q_w \coloneqq r(s,a) + \gamma\phi(s,a)^{\top} w$$

$$V_w(s) \coloneqq \max_{a\in A} Q_w(s,a)$$

$$\pi_w(s) \coloneqq \mathrm{argmax}_{a\in A} Q_w(s,a)$$

Learning w via Q-Learning and linear-regression:

- Find a rep. state-action pairs $\mathcal{L} \subset \mathrm{S} \times \mathrm{A}$, $|\mathcal{L}| = K$ s.t. $\Phi_{\mathcal{L}}$ is regular
- $w^{(0)} \leftarrow 0, i \leftarrow 1$
- For each $(s,a) \in \mathcal{L}$
  - Obtain m samples from $P(\cdot\,|s,a)$: $s_1, s_2, \ldots$
  - Compute empirical average $A^{(i)}(s,a) = m^{-1}\sum_{j=1}^{m} V_{w^{i-1}}(s_j)$
- $w^{(i)} \leftarrow \Phi_{\mathcal{L}}^{-1} A^{(i)},\ i \leftarrow i+1$

Theorem 1: *With*

$$\tilde{O}\left(K(1-\gamma)^{-7}\epsilon^{-2}\right)$$

samples, Algorithm 1 recovers an ϵ-optimal policy with high probability.

Optimality?

Need stronger assumption

Assumption 2: *convex-hull anchors*

there exists $\mathcal{L} \subset S \times A$ such that each $P(\cdot\,|s,a)$ comes is in the convex-hull of $\{P(\cdot\,|s_i,a_i) : (s_i,a_i) \in \mathcal{L}\}$

Theorem 2: *Under Assumption 2, the optimal complexity of an obtaining ϵ-optimal policy is*

$$\widetilde{\Theta}[K(1-\gamma)^{-3}\epsilon^{-2}].$$

https://arxiv.org/abs/1902.04779