Sample-Optimal Parametric Q-Learning Using Linearly Additive Features

Lin F. Yang, Mengdi Wang

PRINCETON UNIVERSITY

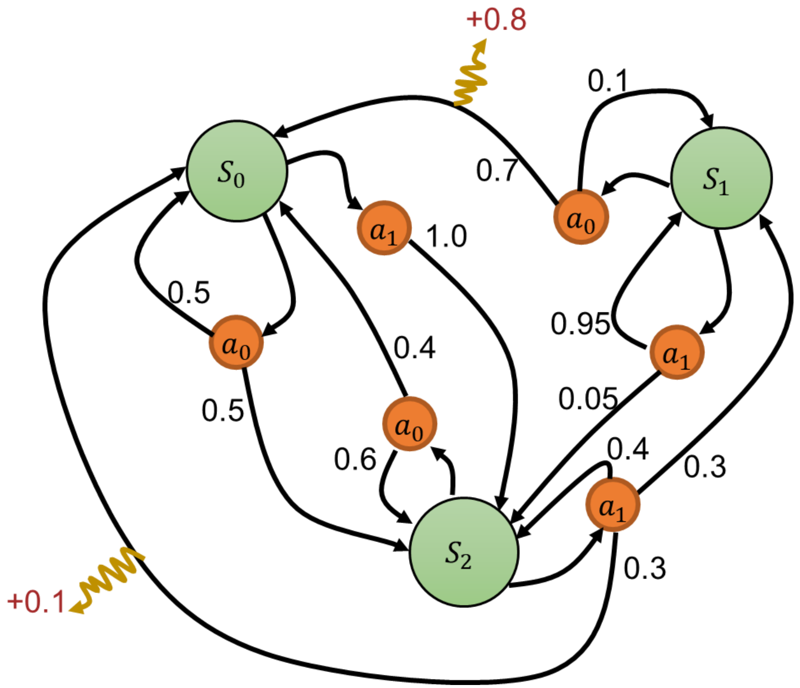
A Basic RL Model: Markov Decision Process

- States: S; Actions: A
- Reward: $r(s, a) \in [0, 1]$
- State transition: $P(s'|s, a)$
- Policy: $\pi : S \to A$

random

$$\max_{\pi} v^{\pi} := \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s^t, a^t)\right]$$

$\gamma \to 1$ Effective Horizon: $(1-\gamma)^{-1}$



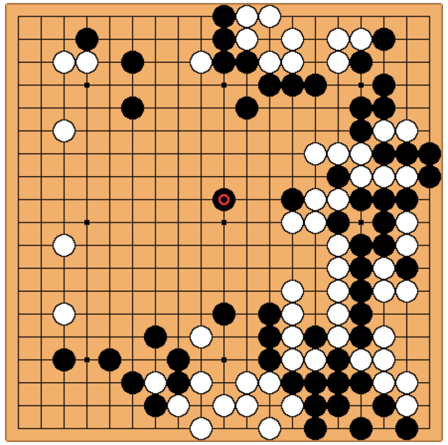
- Optimal policy & value: π^* v^*
- ϵ-optimal policy π : $v^* - v^{\pi} \leq \epsilon$

Curse of Dimensionality

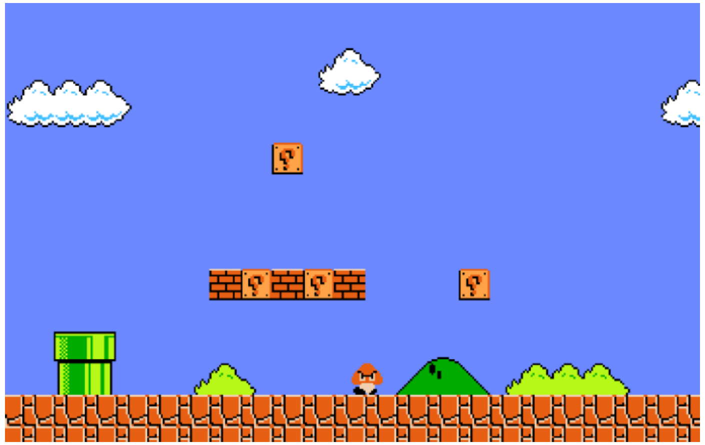
- Optimal sample complexity: $\widetilde{\Theta}[(1-\gamma)^{-3}|S||A|]$

S

Too many states for most cases …



$|S| = 3^{361}$



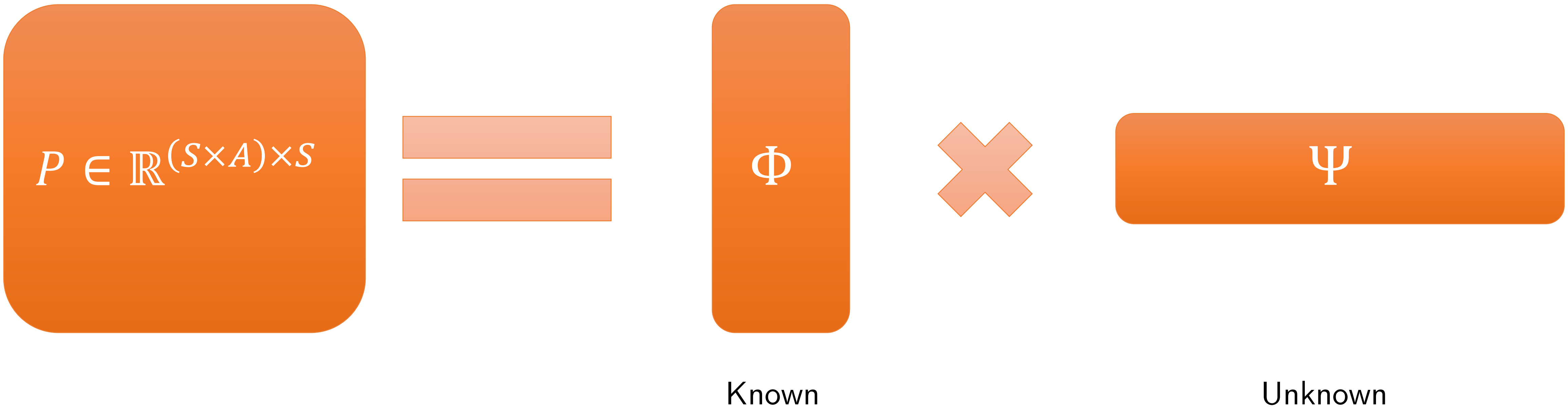
$|S| \geq 256^{256\times 240}$

How to optimally reduce dimensions?

Exploiting structures!

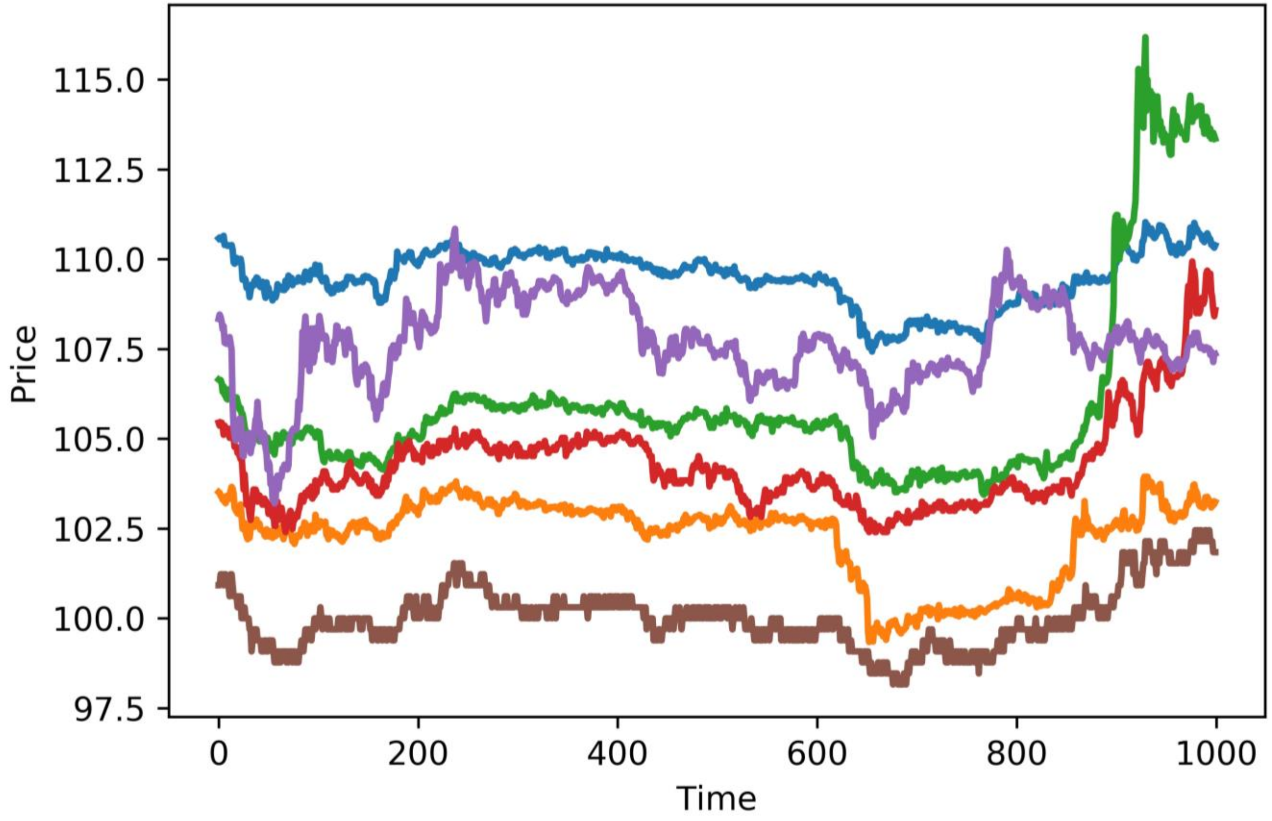
Parametric Q-Learning On Feature-Based MDP

- Transition is decomposable $P(s'|s,a) = \sum_{k\in[K]} \phi_k(s,a)^\top \psi_k(s')$



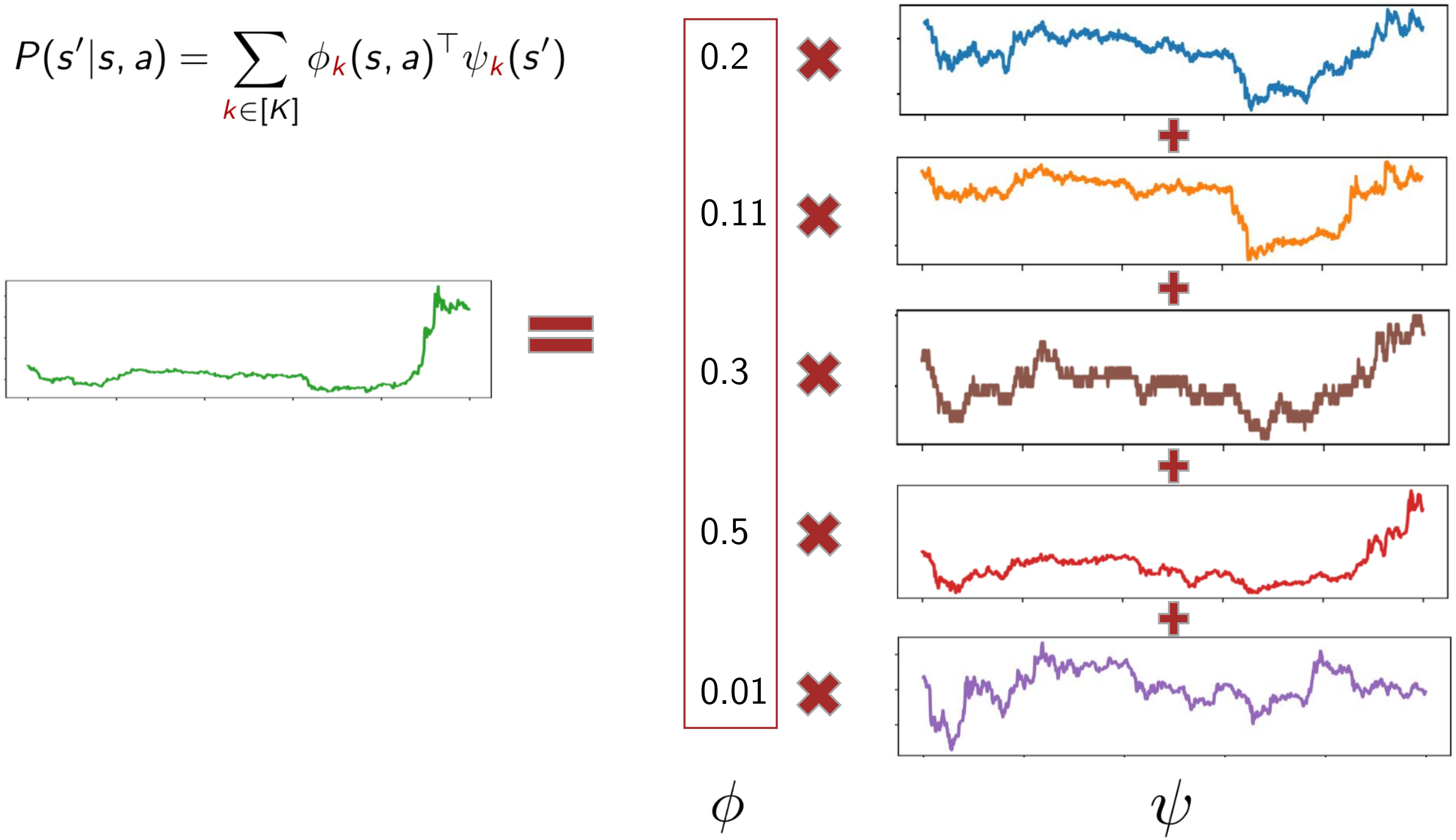
Parametric Q-Learning On Feature-Based MDP

- Transition is decomposable $P(s'|s,a) = \sum_{k \in [K]} \phi_k(s,a)^\top \psi_k(s')$



Parametric Q-Learning On Feature-Based MDP

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



A Simple Regression Based Algorithm

- Generative Model: we are able to samples from any (s,a)

  Represent Q-function with parameter $w \in \mathbb{R}^K$:

$$Q_w := r(s,a) + \gamma \phi(s,a)^\top w$$

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

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

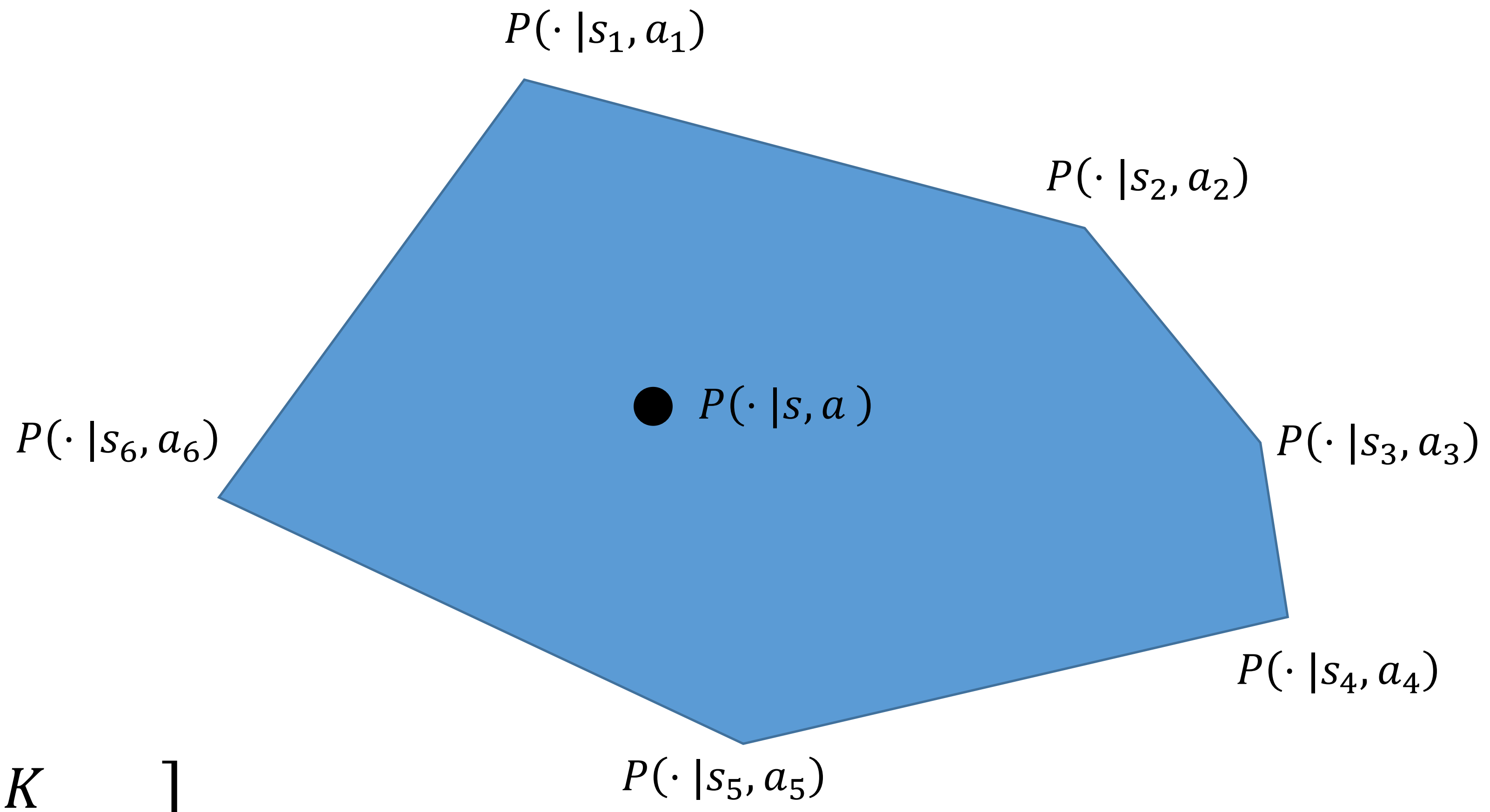
- Learn w with modified Q-learning

  Sample complexity (K: feature dimension):

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

Sample Optimality?

- Anchor condition:



Sample complexity:

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

ArXiv: 1902.04779. Poster: 117