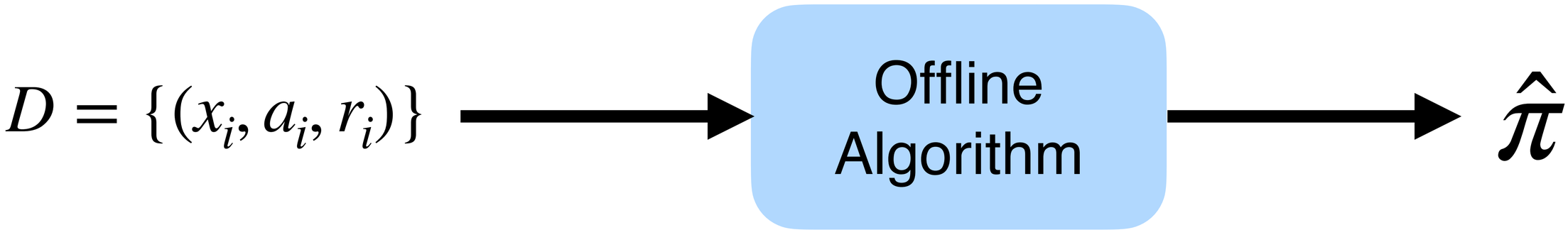
Model Selection in Batch Policy Optimization

Jonathan N. Lee[1,2], George Tucker[2], Ofir Nachum[2], and Bo Dai[2]

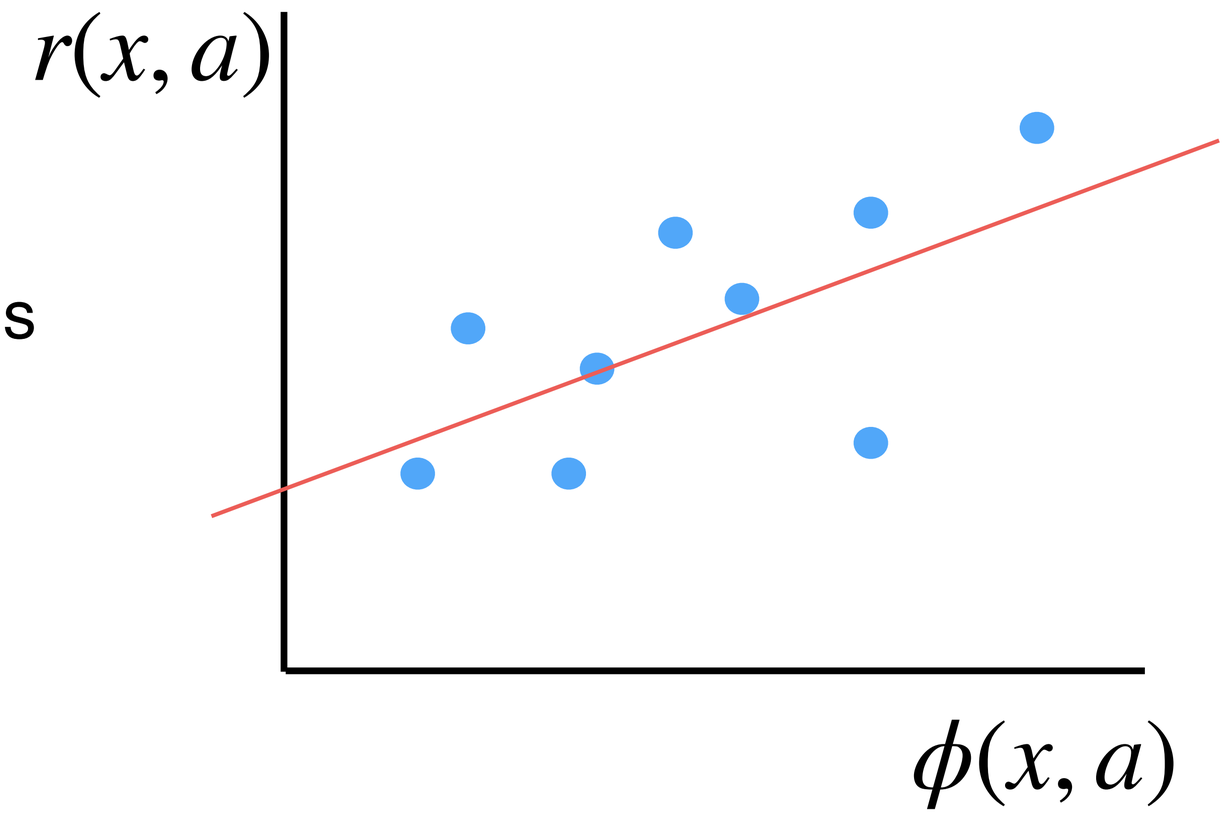
[1]Stanford University
[2]Google Research

Policy Optimization from Batch Data

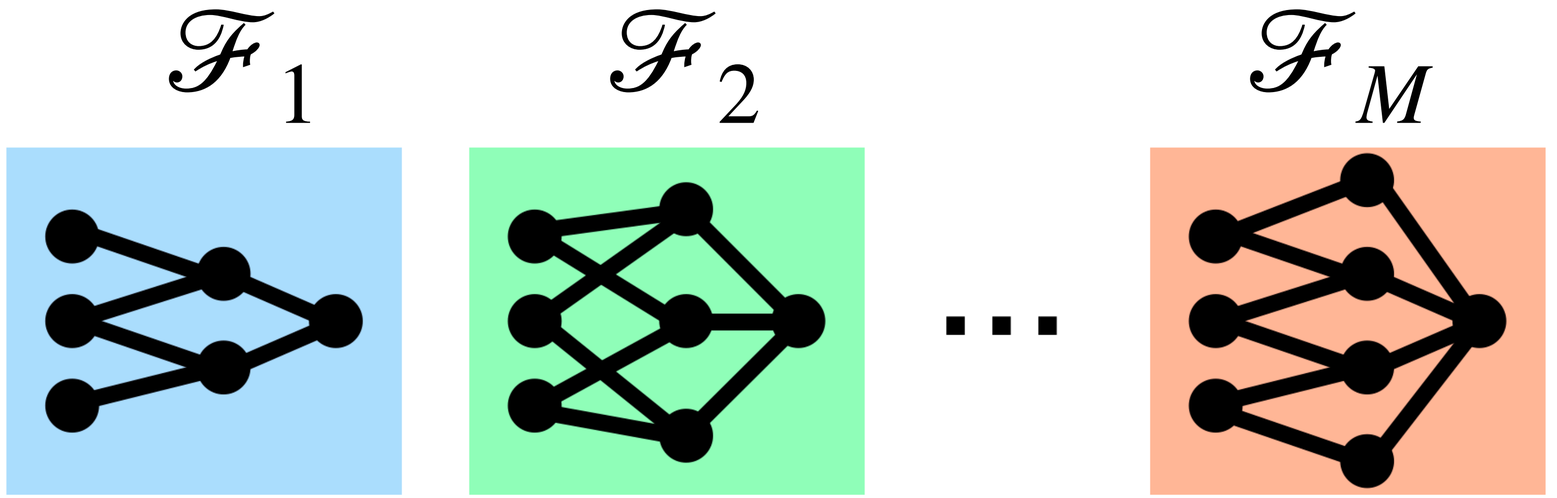


The Role of Function Approximation

- Function approximation is used to generalize and overcome large state-action spaces
- Approximators can have a tremendous number of hyperparameters
- Examples: features in linear regression, layers in neural networks.



Model Selection



How can we identify the best one from data?

Linear Contextual Bandits with Batch Data

Linear Contextual Bandits with Batch Data

Dataset of state-action-reward tuples and linear model class

$$D = \left\{(x_i, a_i, r_i)\right\}_{i \in [n]} \qquad \mathcal{F} = \left\{(x, a) \to \langle \phi(x, a), \theta \rangle \ : \ \theta \in \mathbb{R}^d\right\}$$

Linear Contextual Bandits with Batch Data

Dataset of state-action-reward tuples and linear model class

$$D = \left\{(x_i, a_i, r_i)\right\}_{i \in [n]} \qquad \mathcal{F} = \left\{(x, a) \to \langle \phi(x, a), \theta \rangle \ : \ \theta \in \mathbb{R}^d\right\}$$

A regret bound from least squares + pessimism

$$v(\pi_*) - v(\hat{\pi}) \leq \tilde{\mathcal{O}}\left(\epsilon + \sqrt{\frac{d}{n}} \cdot \mathbb{E}_X \|\phi(X, \pi_*(X))\|_{V^{-1}}\right)$$

approximation error

statistical complexity

dataset coverage

An Ideal Model Selection Algorithm

Given M feature maps (model classes), compete with the best one

An Ideal Model Selection Algorithm

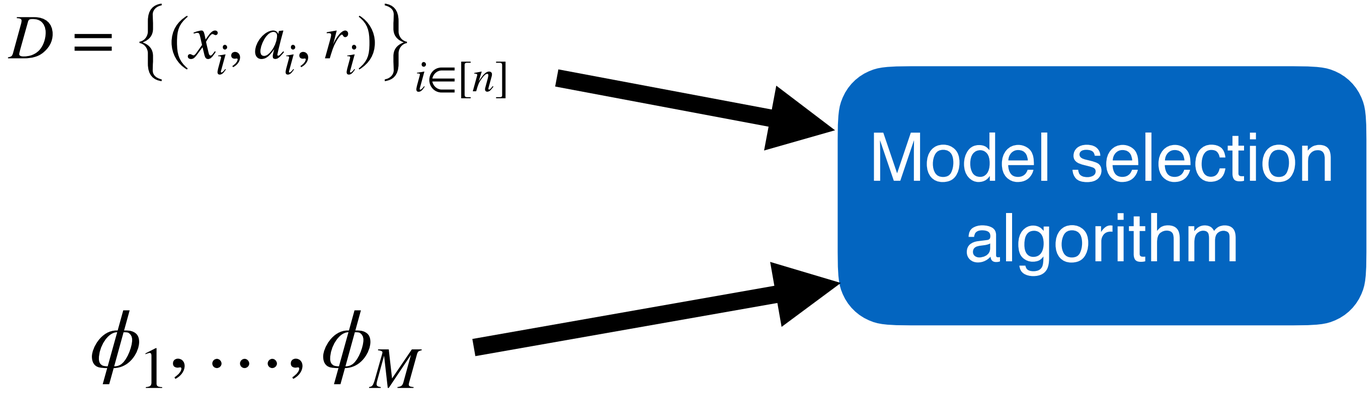
Given M feature maps (model classes), compete with the best one

$$D = \left\{(x_i, a_i, r_i)\right\}_{i \in [n]}$$

$$\phi_1, \ldots, \phi_M$$

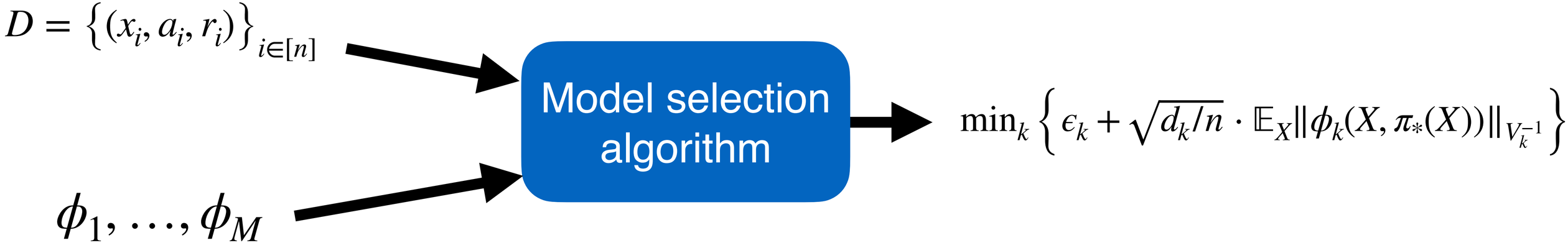
An Ideal Model Selection Algorithm

Given M feature maps (model classes), compete with the best one



An Ideal Model Selection Algorithm

Given M feature maps (model classes), compete with the best one



A Negative Result

For any $\alpha \lesssim \sqrt{n}$ and any model selection algorithm that outputs $\hat{\pi}$, there is an instance such that

learner →

oracle →

$$\frac{\mathbb{E}\left[v(\pi_*) - v(\hat{\pi})\right]}{\min_k \left\{ \epsilon_k + \sqrt{d_k/n} \cdot \mathbb{E}_X \|\phi_k(X, \pi_*(X))\|_{V_k^{-1}} \right\}} \geq \alpha$$

Model selection in batch RL can be much harder than in supervised learning!

Key Idea

- Focus on instances where some model classes are misspecified
  - A model class can be bad in some instances, but very accurate (realizable) in others
- Imbalanced dataset
  - Lots of data about some actions, very little data about others

Learner

Can't distinguish between approximation error and uncertainty due to imbalance

Oracle

Can always select the best realizable model class.

Positive Results for Special Cases

Positive Results for Special Cases

When realizability is always satisfied...

$$v(\pi_*) - v(\hat{\pi}) \lesssim \min_k \left\{ \sqrt{d_k/n} \cdot \mathbb{E}_X \|\phi_k(X, \pi_*(X))\|_{V_k^{-1}} \right\}$$

Positive Results for Special Cases

When realizability is always satisfied...

$$v(\pi_*) - v(\hat{\pi}) \lesssim \min_k \left\{ \sqrt{d_k/n} \cdot \mathbb{E}_X \|\phi_k(X, \pi_*(X))\|_{V_k^{-1}} \right\}$$

When coverage is already good...

$$v(\pi_*) - v(\hat{\pi}) \lesssim \min_k \left\{ \tilde{\epsilon}_k + \sqrt{\mathscr{C} \cdot d_k/n} \right\}$$

concentrability coefficient

Recap

- Approximation, complexity, and coverage contribute to regret in batch/offline learning
- No model selection algorithm to balance all three even in linear contextual bandits
- Relaxing any one of the three contributors leads to strong model selection guarantees

Open questions

- General function approximation?
- Finite horizon reinforcement learning?