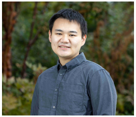
Bilinear Classes: A Structural Framework for Provable Generalization in RL

Talk by: Gaurav Mahajan
(UCSD)

Joint work with:

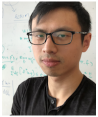


Simon Du

Sham Kakade

Jason Lee

Shachar Lovett



Wen Sun

Ruosong Wang

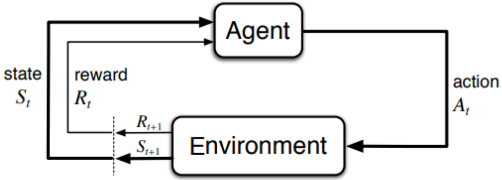
Progress of RL in practice

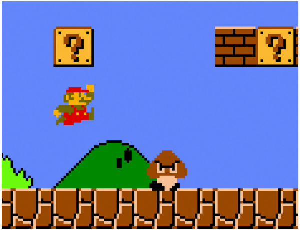
- Lots of recent empirical success.
- Tackling large state spaces is a central challenge in RL.
  - Growing theoretical work on assumptions which allow dealing with large state spaces.
  - Can we unify these assumptions?

This Talk

We aim to understand natural sufficient conditions which capture the learnability in a general class of RL models.

- Part I: Generalization in Reinforcement Learning
  Connections to Supervised Learning
  Various model assumptions for generalization in RL
- Part II: Unifying sufficient conditions
  Is there a common theme to prior assumptions?





- A policy $\pi : \mathcal{S} \to \mathcal{A}$
  - Mario: Always go right!!
- Execute π to obtain a H-step trajectory $s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{H-1}, a_{H-1}, r_{H-1}$
  - Chess: $H \approx 80$, Go: $H = 150$, Dota 2: $H \approx 20000$

Goal

Learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ which maximizes $\mathbb{E}_\pi \left[\sum_{t=0}^{H-1} r_t\right]$.

Part 0: Generalization from Supervised Learning to Reinforcement Learning

Generalization in Supervised Learning

Generalization is possible in the IID supervised learning setting!!

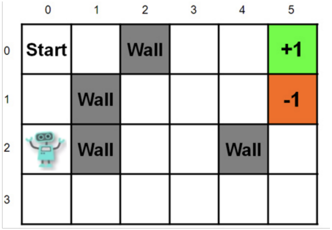
To get ϵ-close to best in hypothesis class $\mathcal{F}$, we need # of samples that is:

- Finite Hypothesis class: $O(\log(|\mathcal{F}|)/\epsilon^2)$.
- Infinite hypothesis classes: $O(\mathrm{VCdim}(\mathcal{F})/\epsilon^2)$.
- Linear Regression in d dimensions: $O(d/\epsilon^2)$

The key idea in SL: uniform convergence / data-reuse.
With a training set, we can simultaneously evaluate the loss of all hypotheses in our class!

Sample Efficient RL in the Tabular Case (no generalization here)

Can we find an ϵ-opt policy with $\mathrm{poly}(\mathcal{S}, \mathcal{A}, H, 1/\epsilon)$ samples?



Theorem (Kearns & Singh '98; ...)

In the episodic setting, $poly(\mathcal{S}, \mathcal{A}, H, 1/\epsilon)$ samples suffice to find an ϵ-opt policy.

- Key Idea: optimism + dynamic programming
- Add bonus for states which are not explored enough.

Q1: Can we find an ϵ-opt policy with no $|\mathcal{S}|$ dependence?

Chess has $|S| \approx 10^{123}$
Dota2 has $S \subset \mathbb{R}^{16000}$!!

- How can we reuse data to estimate the value of all policies in a policy class $\mathcal{F}$?
  Idea: Trajectory tree algorithm acts randomly for length H episodes and then uses importance sampling to evaluate every $f \in \mathcal{F}$.

Theorem (Kearns, Mansour, & Ng '00)

To find an ϵ-best in class policy, the trajectory tree algo uses $O(|\mathcal{A}|^H \log(|\mathcal{F}|)/\epsilon^2)$.

- Can we avoid A^H dependence to find an ϵ-best-in-class policy?
  Without further assumptions, NO!!
  Proof: Consider a binary tree with 2^H policies and a sparse reward at a leaf node.

Provable Generalization in RL: Attempt II

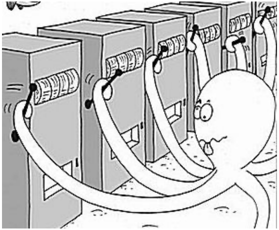
Q2: Can we find an ϵ-opt policy with no $|\mathcal{S}|$, $|\mathcal{A}|$ dependence and $\mathrm{poly}(H, 1/\epsilon,$ "complexity measure"$)$?

- With various stronger assumptions, YES!
  - Linear Bellman Completion: [Munos et al., '05, Zanette et al., '19]
    - Linear MDPs: [Wang & Yang'18]; [Jin et al., '19] (the transition matrix is low rank)
    - Linear Quadratic Regulators (LQR): standard control theory model
  - FLAMBE / Feature Selection: [Agarwal et al., '20]
  - Linear Mixture MDPs: [Modi et al., '20, Ayoub et al., '20]
  - Block MDPs [Du et al., '19]
  - Factored MDPs [Sun et al., '19]
  - Kernelized Nonlinear Regulator [Kakade et al., '20]
  - And more. . .

Part II: What are sufficient conditions for efficient RL?

Is there a common theme to prior settings?

Special case I: Linear bandits ($H = 1$ RL problem) [Abe and Long, 1999]



- [Assumption 1] One step RL ($H = 1$): single state: s_0, large set of actions: $a \in \mathcal{A}$
- [Assumption 2] Linear reward: There exists unknown vector $w^\star \in \mathbb{R}^d$ and known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$

$$\mathbb{E}[r(s_0, a)] = \langle w^\star, \phi(s_0, a) \rangle$$

Polynomial sample complexity is possible here [Auer et al. 2002; Dani et al. 2008]

Special case I: Important structural property

- Linear "value-based" Hypothesis class $\mathcal{F}$:
  set of all (bounded) linear vectors $\mathcal{F} = \{w \in \mathbb{R}^d\}$
  Define for each hypothesis $w \in \mathcal{F}$, $Q_w(s_0, a) = \langle w, \phi(s_0, a) \rangle$,
  (greedy) value $V_w(s_0)$ and (greedy) policy $\pi_w(s_0)$

An important structural property:

- Bilinear Regret: for all $w \in \mathcal{F}$, on policy difference between claimed reward $\mathbb{E}[Q_w]$ and true reward $\mathbb{E}[r]$ satisfies a bilinear form

$$\begin{aligned}\mathbb{E}_{\pi_w}[Q_w(s_0, a) - r] &= \mathbb{E}_{\pi_w}\left[\Big\langle w, \phi(s_0, a) \Big\rangle - \Big\langle w^\star, \phi(s_0, a) \Big\rangle\right] \\ &= \Big\langle w - w^\star,\ \mathbb{E}_{\pi_w}[\phi(s_0, a)] \Big\rangle\end{aligned}$$

- Data reuse: There exists loss function $\ell(s, a, r, w') = Q_{w'}(s, a) - r$ such that the bilinear form for any hypothesis w' is estimable when playing π_w

$$\mathbb{E}_{\pi_w}[\ell(s_0, a, r, w')] = \Big\langle w' - w^\star,\ \mathbb{E}_{\pi_w}[\phi(s_0, a)] \Big\rangle$$

Essentially, we can use data collected under π_w to estimate the bilinear form for w'

Special case II: Linear Bellman complete classes [Munos, 2005]

- [Assumption 1] Linear $Q^\star$: There exists unknown $w^\star \in \mathbb{R}^d$ and known features $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that

$$Q^\star(s,a) = \langle w, \phi(s,a) \rangle$$

- [Assumption 2] Completeness: Let $\mathcal{F}$ be the linear "value-based" hypothesis class. For every $w \in \mathcal{F}$, there exists $\mathcal{T}(w) \in \mathcal{F}$ such that

$$\langle T(w), \phi(s,a) \rangle = r(s,a) + \mathbb{E}_{s' \sim P(s,a)}[\max_{a'} Q_w(s',a')]$$

Polynomial sample complexity is possible here [Zanette et al. 2020])

Special case II: Important structural property

Analogous structural property holds here:

- Bilinear Regret: on policy difference between claimed reward $\mathbb{E}[Q_w - V_w]$ and true reward $\mathbb{E}[r]$ satisfies a bilinear form

$$\begin{aligned}
&\mathbb{E}_{\pi_w}[Q_w(s_h,a_h) - r(s_h,a_h) - V_w(s_{h+1})] \\
&= \mathbb{E}_{\pi_w}\left[\left\langle w, \phi(s,a)\right\rangle - \left\langle T(w), \phi(s,a)\right\rangle\right] \\
&= \left\langle w - T(w), \mathbb{E}_{\pi_w}[\phi(s,a)]\right\rangle \\
&= \left\langle w - T(w) - (w^\star - T(w^\star)),\ \mathbb{E}_{\pi_w}[\phi(s,a)]\right\rangle
\end{aligned}$$

- Data reuse: There exists loss function $\ell(\cdot, w')$ such that the bilinear form for any hypothesis w' is estimable when playing π_w

$$\mathbb{E}_{\pi_w}[\ell(s_h,a_h,r_h,s_{h+1},w')] = \left\langle w' - T(w') - (w^\star - T(w^\star)),\ \mathbb{E}_{\pi_w}[\phi(s,a)]\right\rangle$$

Here the loss function is

$$\ell(s_h,a_h,r_h,s_{h+1},w') = Q_{w'}(s_h,a_h) - r_h - V_{w'}(s_{h+1})$$

Special case III: Linear Mixture Model classes [Modi et al., 2020b]

- [Assumption 1] Linear dynamics and rewards: There exists unknown $w^\star \in \mathbb{R}^d$ and known features $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that

$$P(s' \mid s, a) = \langle w^\star,\ \phi(s, a, s') \rangle \quad \text{and} \quad \mathbb{E}[r(s, a)] = \langle w^\star,\ \psi(s, a) \rangle$$

Polynomial sample complexity is possible here [Modi et al., 2020; Ayoub et al., 2020])

Special case III: Important structural property

- Linear "model-based" Hypothesis class $\mathcal{F}$:
  set of all (bounded) linear vectors $\mathcal{F} = \{w \in \mathbb{R}^d\}$
  Define for each hypothesis $w \in \mathcal{F}$, $P_w(s'|s,a) = \langle w, \phi(s,a,s')\rangle$,
  $Q_w(s,a)$, $V_w(s)$ and $\pi_w(s)$ as the optimal functions for model P_w

Analogous structural property holds here:

- Bilinear Regret: on policy difference between claimed reward $\mathbb{E}[Q_w - V_w]$ and true reward $\mathbb{E}[r]$ satisfies a bilinear form

$$\mathbb{E}_{\pi_w}[Q_w(s_h,a_h) - r(s_h,a_h) - V_w(s_{h+1})]$$
$$= \left\langle w - w^\star, \quad \mathbb{E}_{\pi_w}\left[\psi(s_h,a_h) + \sum_{\bar{s}\in\mathcal{S}} \phi(s_h,a_h,\bar{s})V_w(\bar{s})\right]\right\rangle$$

- Data reuse: There exists loss function $\ell_w(\cdot)$ such that the bilinear form for any hypothesis w' is estimable when playing π_w

$$\mathbb{E}_{\pi_w}[\ell(s_h,a_h,r_h,s_{h+1},w')] = \left\langle w' - w^\star, \mathbb{E}_{\pi_w}\left[\sum_{\bar{s}\in\mathcal{S}} \phi(s_h,a_h,\bar{s})V_w(\bar{s})\right]\right\rangle$$

  Here the loss function is

$$\ell_w(s_h,a_h,r_h,s_{h+1},w') = {w'_h}^\top\left(\psi(s_h,a_h) + \sum_{\bar{s}\in\mathcal{S}} \phi(s_h,a_h,\bar{s})V_w(\bar{s})\right) - V_w(s_{h+1}) - r_h$$

BiLinear Classes: structural properties to enable generalization in RL

- Hypothesis class: $\{f \in \mathcal{F}\}$
  with associated state action value $Q_f(s,a)$, (greedy) value $V_f(s)$ and (greedy) policy π_f
  - can be model-based or value-based class.

Definition

A $(\mathcal{F}, \ell)$ forms an (implicit) Bilinear class if there exists $w_h : \mathcal{F} \to \mathbb{R}^d$ and $\Phi_h : \mathcal{F} \to \mathbb{R}^d$ for all timesteps $h \in [H]$:

- Bilinear regret: on-policy difference between claimed reward and true reward satisfies a bilinear form:

$$\left| E_{\pi_f}\left[Q_f(s_h, a_h) - r(s_h, a_h) - V_f(s_{h+1}) \right] \right| \leq \left| \langle w_h(f) - w_h(f^\star), \Phi_h(f) \rangle \right|$$

- Data reuse: There exists loss function $\ell_f(s_h, a_h, r_h, s_{h+1}, g)$ such that the bilinear form for any hypothesis g is estimable when playing π_f

$$\left| E_{\pi_f}\left[\ell_f(r_h, s_h, a_h, s_{h+1}, g) \right] \right| = \left| \langle w_h(g) - w_h(f^\star), \Phi_h(f) \rangle \right|$$

Theorem 1: Structural Commonalities and Bilinear Classes

Theorem (Du, Kakade, Lee, Lovett, **M.**, Sun, Wang '21)

The following models are bilinear classes for some discrepancy function $\ell(\cdot)$

- *Linear Bellman Completion: [Munos et al. '05, Zanette et al. '19]*
  - *Linear MDPs: [Wang & Yang '18]; [Jin et al.'19] (the transition matrix is low rank)*
  - *Linear Quadratic Regulators (LQR): standard control theory model*
  - *Generalized Linear Bellman Completion: [Wang et al. '2019]*
- *FLAMBE / Feature Selection: [Agarwal et al. '20]*
- *Linear Mixture MDPs: [Modi et al. '20, Ayoub et al. '20]*
- *Block MDPs [Du et al. '19]*
- *Factored MDPs [Sun et al. '19]*
- *Kernelized Nonlinear Regulator [Kakade et al. '20]*
- *And more...*

- (almost) all "named" models (with provable generalization) are bilinear classes
  two exceptions: a) deterministic linear $Q^\star$ [Wen & Van Roy, '13; Du, Lee, **M.**, Wang, '20]
  b) $Q^\star$ **state-action aggregation** [Dong et al. '20]
- Bilinear classes generalize the: Bellman rank [Jiang et al. '17]; Witness rank [Wen et al. '19]
- The framework easily leads to new models (see paper).

The Algorithm: BiLin-UCB

Algorithm 1: BiLin-UCB

Input number of iterations T, estimator function ℓ, batch size m, confidence radius R

Initialize cumulative discrepancy function $\sigma^2(\cdot) = 0$

for *iteration* $t = 0, 1, \ldots, T-1$ **do**

Find the optimistic $f_t \in \mathcal{F}$:

$$f_t := \arg\max_f V_f(s_0) \quad \text{subject to } \sigma^2(f) \leq R$$

Sample m trajectories using π_{f_t} and create a batch dataset of size mH:

$$S = \{(r_h, s_h, a_h, s_{h+1}) \in \text{trajectories}\}$$

Update the cumulative discrepancy function $\sigma^2(\cdot)$

$$\sigma^2(\cdot) \leftarrow \sigma^2(\cdot) + \left(\frac{1}{|S|} \sum_{o \in S} \ell(o, \cdot)\right)^2$$

return: the best policy π_f found

Theorem 2: Generalization in RL

Theorem (Du, Kakade, Lee, Lovett, M., Sun, Wang '21)

Assume $(\mathcal{F}, \ell)$ is a bilinear class with $\Phi_h(f) \in \mathbb{R}^d$, bounded ℓ and the class is realizable, i.e. $Q^\star \in \mathcal{F}$. Using $\frac{d^2}{\epsilon^2} \cdot poly(H) \cdot \log(|\mathcal{F}|) \cdot \log(1/\delta)$ trajectories, the BiLin-UCB algorithm returns an ϵ-opt policy (with prob. $1-\delta$).

- The proof is "elementary" using the elliptical potential function.[Dani et al., '08]
- Extends to infinite dimensional problems using max info gain γ_T [Auer et al., '02; Srinivas et al., '10; Abbasi-Yadkori et al., '11]

Proof intuition

- The proof follows from this lemma about existence of high quality policy.

Lemma (Existence of high quality policy)

Suppose we run the algorithm for $T \approx d$ iterations. Then, there exists $t \in [T]$ such that the following is true for hypothesis f_t:

$$V^\star - V^{\pi_{f_t}}(s_0) \le 2H\sqrt{d} \cdot \underbrace{H\sqrt{\frac{\log(|\mathcal{F}|)}{m}}}_{\text{SL generalization error of } \ell}$$

Proof of main lemma

- Bilinear regret assumption and Optimism give an upper bound on sub-optimality for all iterations t.

$$V^\star - V^{\pi_{f_t}}(s_0) \leq \sum_{h=0}^{H-1} |\langle w_h(f_t) - w_h(f^\star), \Phi_h(f_t)\rangle| \ .$$

- Our goal then is to show existence of iteration $t \in [T]$ such that

$$\sum_{h=0}^{H-1} |\langle w_h(f_t) - w_h(f^\star), \Phi_h(f_t)\rangle| \quad \text{is small}$$

- To that end, we will show existence of iteration $t \in [T]$ such that for $\Sigma_{0;h} = \lambda I$ and $\Sigma_{t;h} = \Sigma_{0;h} + \sum_{i=0}^{t-1} \Phi_h(f_i)\Phi_h(f_i)^\top$, the following is true

$$\|w_h(f_t) - w_h(f^\star)\|_{\Sigma_{t;h}} \quad \|\Phi_h(f_t)\|_{\Sigma_{t;h}^{-1}} \quad \text{is small for all } h \in [H]$$

Proof of main lemma

- To that end, we will show existence of iteration $t \in [T]$ such that for $\Sigma_{0;h} = \lambda I$ and $\Sigma_{t;h} = \Sigma_{0;h} + \sum_{i=0}^{t-1} \Phi_h(f_i)\Phi_h(f_i)^\top$, the following is true

$$\|w_h(f_t) - w_h(f^\star)\|_{\Sigma_{t;h}} \quad \|\Phi_h(f_t)\|_{\Sigma_{t;h}^{-1}} \quad \text{is small for all } h \in [H]$$

- From our optimization constraint, we get that for all time t (we can set R small because of uniform convergence and Data reuse assumption)

$$\|w_h(f_t) - w_h(f^\star)\|_{\Sigma_{t;h}} \leq R = 2\sqrt{d} \cdot \underbrace{H\sqrt{\frac{\log(|\mathcal{F}|)}{m}}}_{\text{SL generalization error}} \quad \text{for all } h \in [H]$$

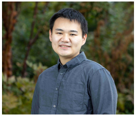
- From Elliptical Potential Lemma, there exists $t \in [T]$ (for $T \approx d$) such that

$$\|\Phi_h(f_t)\|^2_{\Sigma_{t;h}^{-1}} = O(1) \quad \text{for all } h \in [H]$$

Note that for infinite dimensional spaces, we can use max info gain instead.

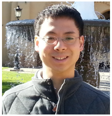
Thanks!

- A generalization theory in RL is possible!
  - linear bandit theory → RL theory (bilinear classes) is rich.
    - covers known cases and new cases
    - leads to simple algorithm and proof



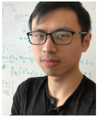
Simon Du

Sham Kakade



Jason Lee

Shachar Lovett



Wen Sun

Ruosong Wang