Reinforcement Learning in High Dimensional Systems
(and why “reward” is not enough...)

Sham M. Kakade
University of Washington & Microsoft Research
Progress of RL in Practice

[AlphaZero, Silver et.al, 17]

[OpenAI Five, 18]
Let’s start with Supervised Learning (SL)
Provable Generalization in Supervised Learning (SL)

Generalization is possible in the IID supervised learning setting!

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

- Finite hypothesis class: need $O(\log( |\mathcal{F}|)/\epsilon^2)$
- Linear hypothesis classes $\mathcal{F}$:
  - Linear regression: $O(\text{dimension}/\epsilon^2)$; Classification (margin bounds): $O(\text{margin}/\epsilon^2)$;
- Neural Hypothesis Classes: $O(\text{size of weights} \times \text{# layers} / \epsilon^2)$
- VC dim: $O(\text{VC}(\mathcal{F})/\epsilon^2)$

The key idea in SL: data reuse

With a training set, we can simultaneously evaluate the loss of all hypotheses in our class.
What about RL?
Markov Decision Processes: a framework for RL (standard notation)

- A policy: \( \pi : \text{States} \rightarrow \text{Actions} \)
- Execute \( \pi \) to obtain a trajectory:
  \[ s_0, a_0, r_0, s_1, a_1, r_1 \ldots s_{H-1}, a_{H-1}, r_{H-1} \]
- Cumulative \( H \)-step reward:
  \[
  V^\pi_H(s) = \mathbb{E}_\pi \left[ \sum_{t=0}^{H-1} r_t \mid s_0 = s \right], \quad Q^\pi_H(s, a) = \mathbb{E}_\pi \left[ \sum_{t=0}^{H-1} r_t \mid s_0 = s, a_0 = a \right]
  \]
- Goal: Find a policy \( \pi \) that maximizes our value \( V^\pi(s_0) \) from \( s_0 \).
- Episodic setting: We start at \( s_0 \); act for \( H \) steps; repeat… (so we must balance exploration/exploitation)
Sample Efficient RL in small, unknown MDPs

- $S = \#\text{states}, A = \#\text{actions}, H = \#\text{horizon}$
- Thm [Kearns & Singh '98]: In the episodic setting, the $E^3$ algo finds an $\epsilon$-opt policy with $\text{poly}(S, A, H, 1/\epsilon)$ samples.
  - No generalization here due to $\text{poly}(S)$ dependence.

- Many improvements on the rate:
  - [Brafman & Tennenholtz '02][K. '03][Auer + '09][Agrawal, Jia '17]
  - minimax rates: [Azar + '13],[Dann & Brunskill '15]
  - provable Q-learning: [Strehl + (2006)], [Szita & Szepesvari '10],[Jin + '18]
Provable Generalization in RL?

• Suppose our hypothesis class $\mathcal{F}$ is a set of policies.
• Can we find an $\epsilon$-opt policy with no $S$ dependence, poly $H$, and $\log(|\mathcal{F}|)$ dependence?
  • No: We need $\min(2^H, \log(|\mathcal{F}|))$ samples (for no $S$ dependence) [Kearns, Mansour, & Ng ’00][K’ 03]
  • Proof:
    • Consider a binary tree with a single rewarding leaf
    • We have $2^H$ policies
    • We have to try them all
  • Unlike SL, data reuse not possible!
Outline

What are necessary representational and distributional conditions that permit provably sample-efficient offline reinforcement learning?

• Part I: Lower bounds (necessity)
  Is RL possible under linear realizability?

• Part II: Upper bounds (sufficiency)
  Are there unifying conditions that are sufficient?
Lower bounds:
What is necessary?
Approx. Dynamic Programming
with **Linear Function Approximation**

- Idea: Approximate the $Q(s, a)$ values with linear basis functions,
  $Q(s, a) = w \cdot \phi(s, a)$, where $\phi(s, a) \in \mathbb{R}^d$ and $d \ll S, A$.

Some context:
- [Tesauro, ’95], [de Farias & Van Roy ’03], [Wen & Van Roy ’13]

What conditions must our basis functions (our representations) satisfy in order for his approach to work?
- Let’s look at the most basic question with “linearly realizable $Q^*$”
  - Analogous to (bandit) linear regression (when $H = 1$)
Linearly Realizable Values is Not Sufficient for RL

Linearly realizable values: suppose \( Q_h^*(s, a) = w_h^* \cdot \phi(s, a) \)

Sub-optimality gap (a “margin”): For all \( a \neq \pi^*(s) \), \( V^*(s) - Q^*(s, a) \geq \Delta_{\text{min}} \)

Theorem: [Wang, Wang, K. ‘21] There exists a class of MDPs with linearly realizable values + constant sub-optimality gap s.t. any online RL algorithm requires \( \min(\Omega(2^d), \Omega(2^H)) \) samples to obtain a 0.1-near optimal policy (with prob. \( \geq 0.9 \)).

- Theorem [Weisz, Amortila, Szepesvári ’21]: With only linearly realizable values, the lower bound still holds (even in a generative model).
- Theorem [Du, K., Wang, Yang ‘20]: With linearly realizable values + constant gap + generative model, there is a sample efficient algorithm.
Linearly Realizable Policies are also Not Sufficient for RL

Linearly realizable policies: $\pi^*(s) = \arg\max_a w^* \cdot \phi(s, a)$

Large “margin”: Suppose $\|w^*\| \leq \text{const}$ (and $\|\phi\| \leq 1$)

Theorem [Du, K., Wang, Yang ‘20]: There exists a class of MDPs with linearly realizable policies + large margin s.t. any online RL algorithm requires $\min(\Omega(2^d), \Omega(2^H))$ samples to obtain a 0.1-near optimal policy (with prob. $\geq 0.9$).

- For (bandit) classification and regression ($H = 1$), learning is $\text{poly}(d)$ for $H = 1$
The Construction: a Hard MDP Family
(A “leaking complete graph”)

- $m$ is an integer (we will set $m \approx 2^d$)
- the state space: $\{ \bar{1}, \ldots, \bar{m}, f \}$
- call the special state $f$ a “terminal state”.
- at state $\bar{i}$, the feasible actions set is $[m] \setminus \{i\}$
- at $f$, the feasible action set is $[m-1]$.
  i.e. there are $m-1$ feasible actions at each state.
- each MDP in this family is specified by an index
  $a^* \in [m]$ and denoted by $M_{a^*}$.
  i.e. there are $m$ MDPs in this family.

**Lemma:** For any $\gamma > 0$, there exist $m = \lfloor \exp(\frac{1}{8} \gamma^2 d) \rfloor$ unit vectors $\{v_1, \ldots, v_m\}$ in $\mathbb{R}^d$ s.t. $\forall i, j \in [m]$ and $i \neq j$, $|\langle v_i, v_j \rangle| \leq \gamma$.

**We will set $\gamma = 1/4$.**
(proof: Johnson-Lindenstrauss)
Upper bounds: What are sufficient conditions?
Special case: linear Bellman complete classes
(let's make stronger assumptions)

- Linear hypothesis class: \( \mathcal{F} = \{ Q_w : Q_w(s, a) = w \cdot \phi(s, a) \} \)
- Bellman "backup" operator: \( \mathcal{T}(Q)(s, a) = r(s, a) + E_{s' \sim P(s'|s,a)}[\max_{a'} Q(s', a')] \)
- Linear Completeness [Munos, '05]: \( Q \in \mathcal{F} \implies \mathcal{T}(Q) \in \mathcal{F} \)
- Linear completeness is much stronger than linearly realizability!
  - Adding a feature to \( \phi \) can break the completeness property.
  - It is fundamentally related to the underlying dynamics model \( P(s'|s, a) \)
- Theorem [Zanette+ '19]: Sample efficient RL, \( \text{poly}(d, H, 1/\epsilon) \), is possible with Bellman complete, linear \( \mathcal{F} \).
- Are there other conditions when sample efficient RL is possible?
Sufficiency: under what conditions is generalization in RL possible?

- There are many others cases where sample efficient RL possible:
  - **Linear Bellman Completion**: [Munos, '05, Zanette+ '19]
    - Linear MDPs: [Wang & Yang’18]; [Jin+ ’19] (the transition matrix is low rank)
  - Linear Quadratic Regulators (LQR): standard control theory model
  - FLAMBRE / Feature Selection: [Agarwal, K., Krishnamurthy, Sun ’20]
  - Linear Mixture MDPs: [Modi+’20, Ayoub+ ’20]
  - Block MDPs [Du+ ’19]
  - Factored MDPs [Sun+ ’19]
  - Kernelized Nonlinear Regulator [K.+ ’20]
- And more…..
- Are there commonalities?

**Theorem [Du, K., Lee, Lovett, Mahajan, Sun, Wang ’19]:**
All the “named” models above are special cases of bilinear classes (see paper for formal def).
Also, provable generalization is possible for bilinear classes.

- Bilinear classes generalize the **Bellman rank [Jiang+ ‘17]**
- Proof techniques come from linear bandits framework [Dani, Hayes, K. ’08]
- Bilinear classes work for model based and model free settings
Bilinear Classes: A Structural Framework for Sample Efficient RL

- Two exceptions: linear $Q^*$ with deterministic dynamics; $Q^*$-state aggregation
- The framework leads to new models (see paper).
Def: BiLinear Classes

- For each hypothesis $f \in \mathcal{F}$, suppose there are associated $Q_f(s, a), V_f(s), \pi_f$
- The hypothesis class $\mathcal{F}$ can be model based or model-free class.

Def: A $(\mathcal{F}, \ell')$ forms an (implicit) Bilinear class class if:
- Bilinear regret: on-policy difference between claimed reward and true reward
  \[
  \left| E_{\pi_f}[Q_f(s_h, a_h) - r(s_h, a_h) - V_f(s_{h+1})] \right| \leq \langle w_h(f) - w_h^*, \Phi_h(f) \rangle
  \]
- estimation (the on-policy case): there is a discrepancy function $\ell'_f(s, a, s', g)$ s.t.
  \[
  \forall g, \quad E_{\pi_f}[\ell'_f(s_h, a_h, s_{h+1}, g)] = \langle w_h(g) - w_h^*, \Phi_h(f) \rangle
  \]

Data reuse: the key is that $\ell(\cdot, g)$ can be estimated simultaneously $\forall g \in \mathcal{F}$
Special case: Linear $Q^*$, $V^*$ is sufficient for RL

Linearly $Q^*, V^*$: suppose $Q^*(s, a) = w^*_Q \cdot \phi_Q(s, a)$ and $V^*(s) = w^*_V \cdot \phi_V(s)$

Theorem [Du, K., Lee, Lovett, Mahajan, Sun, Wang '19]:
Suppose the linear $Q^*, V^*$ assumption is satisfied (with known features) then sample efficient RL is possible.

- This assumption is subtle. It does impose much stronger constraints than just linear $Q^*$. 
Thanks!

• A generalization theory in RL is possible and different from SL!
  • necessary: linear realizability insufficient. need much stronger assumptions.
  • sufficient: bilinear classes is a more general framework.
    • covers known cases/new cases
    • FLAMBE: [Agarwal+ ’20] feature learning possible in this framework.
  • related: offline RL has similar challenges
    [Wang, Foster, K. ’20], [Zanette ’21], [Wang, Wu, Salakhutdinov, K., 2021]

See https://rltheorybook.github.io/ for forthcoming book!