What Structural Conditions Permit Generalization in Reinforcement Learning?

Joint work with

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett,

Gaurav Mahajan, Ruosong Wang

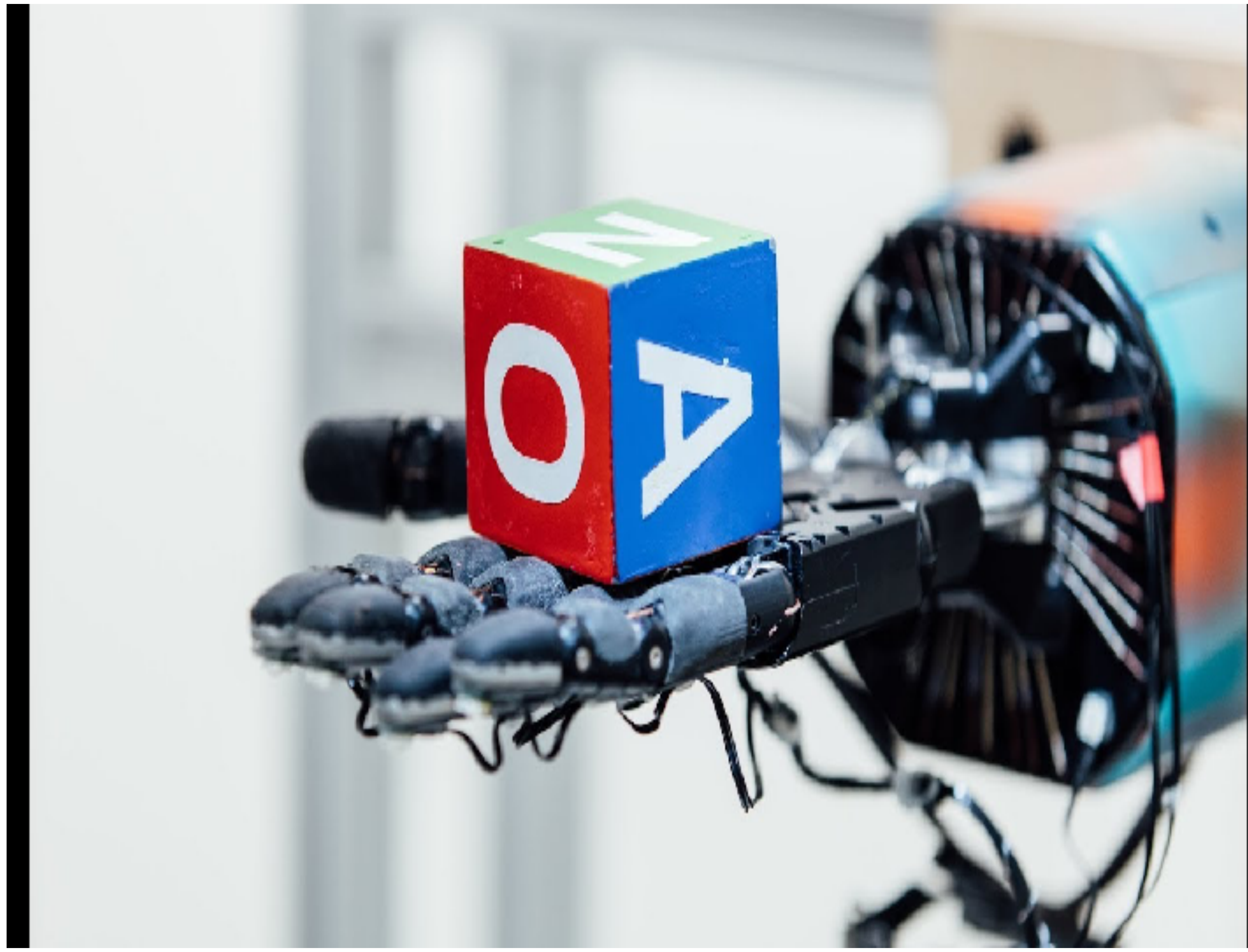





Solving Large-scale RL problems requires generalization

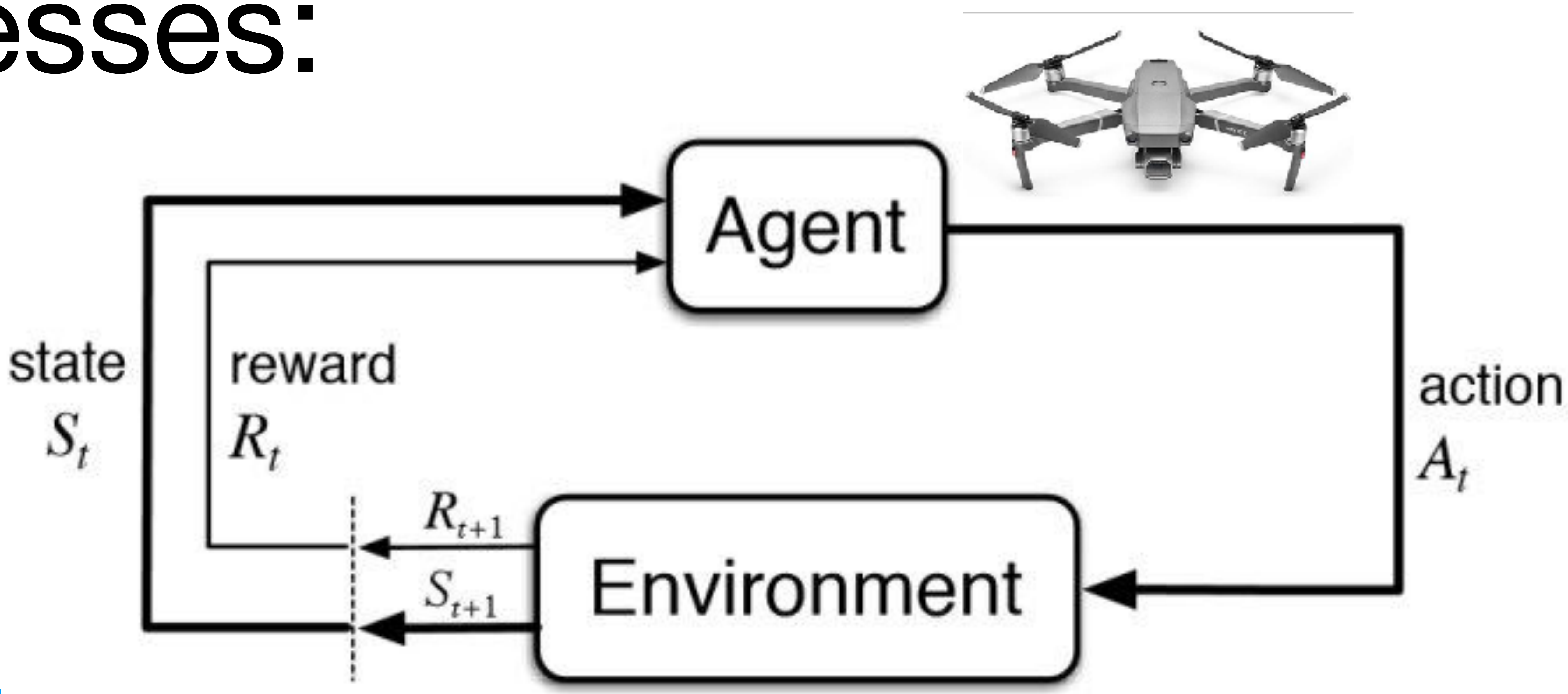
[AlphaZero, Silver et.al, 17]

[OpenAI Five, 18]



[OpenAI,19]

Markov Decision Processes:
a framework for RL



- A policy:

  π : States $\rightarrow$ Actions
- Execute π 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:

$$V_H^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{H-1} r_t \middle| s_0 = s\right], \quad Q_H^\pi(s, a) = \mathbb{E}_\pi\left[\sum_{t=0}^{H-1} r_t \middle| s_0 = s, a_0 = a\right]$$

- Goal: Find a policy π that maximizes our value $V^\pi(s_0)$ from s_0.

  Episodic setting: We start at s_0; act for H steps; repeat…

Provable Generalization in Supervised Learning (SL)

Generalization is possible in the IID supervised learning setting!

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

- "Occam's Razor" Bound (finite hypothesis class): need $O(\log |\mathscr{F}| / \epsilon^2)$

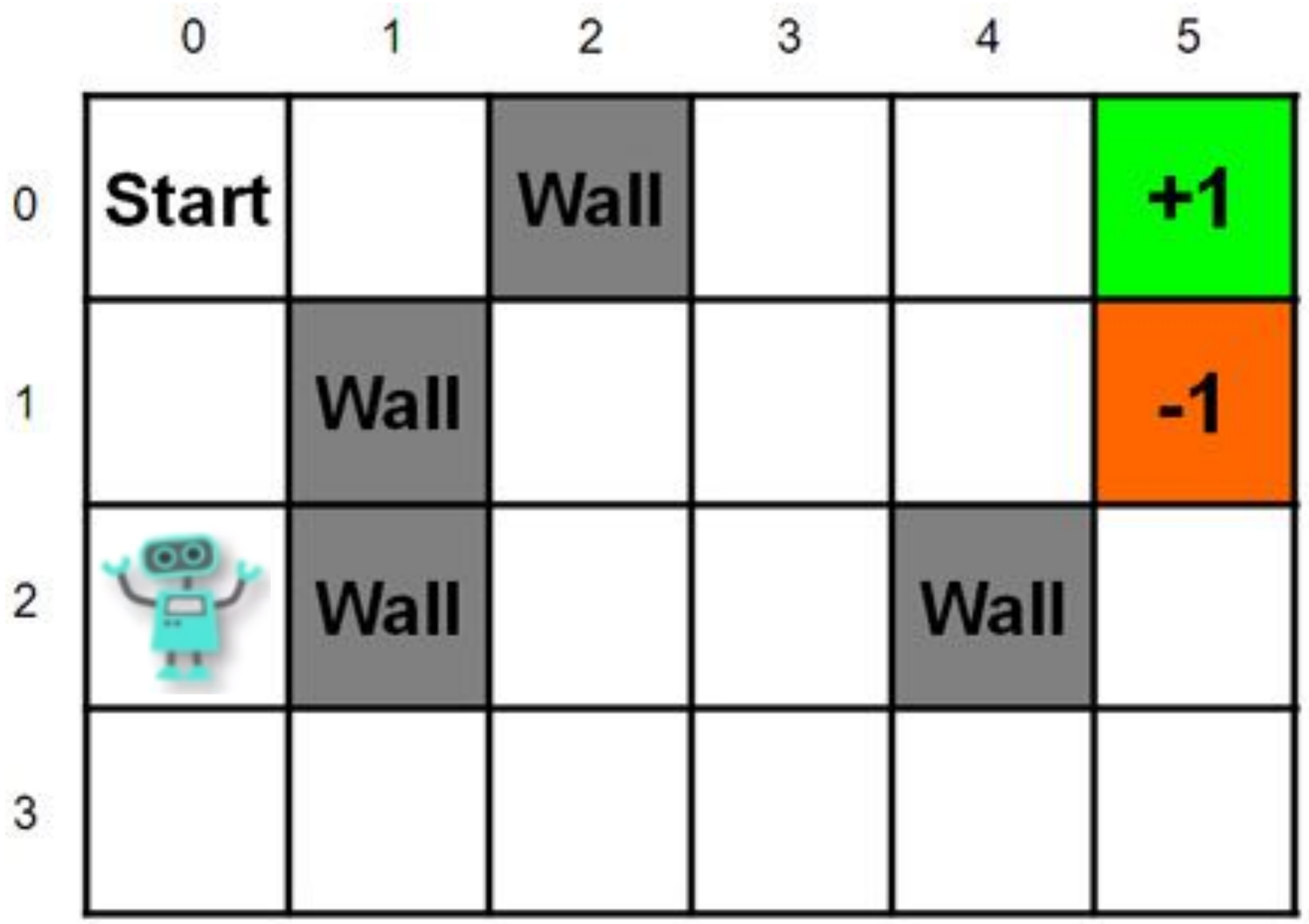
Various Improvements:

- VC dim: need only $O(\text{VC}(\mathscr{F})/\epsilon^2)$
- Classification: linearly separable + margin: $O(\text{margin})/\epsilon^2)$
- Linear Regression in d dimensions: $O(d/\epsilon^2)$
- Deep Learning: the algorithm also determines the complexity control

The key idea in SL: 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)

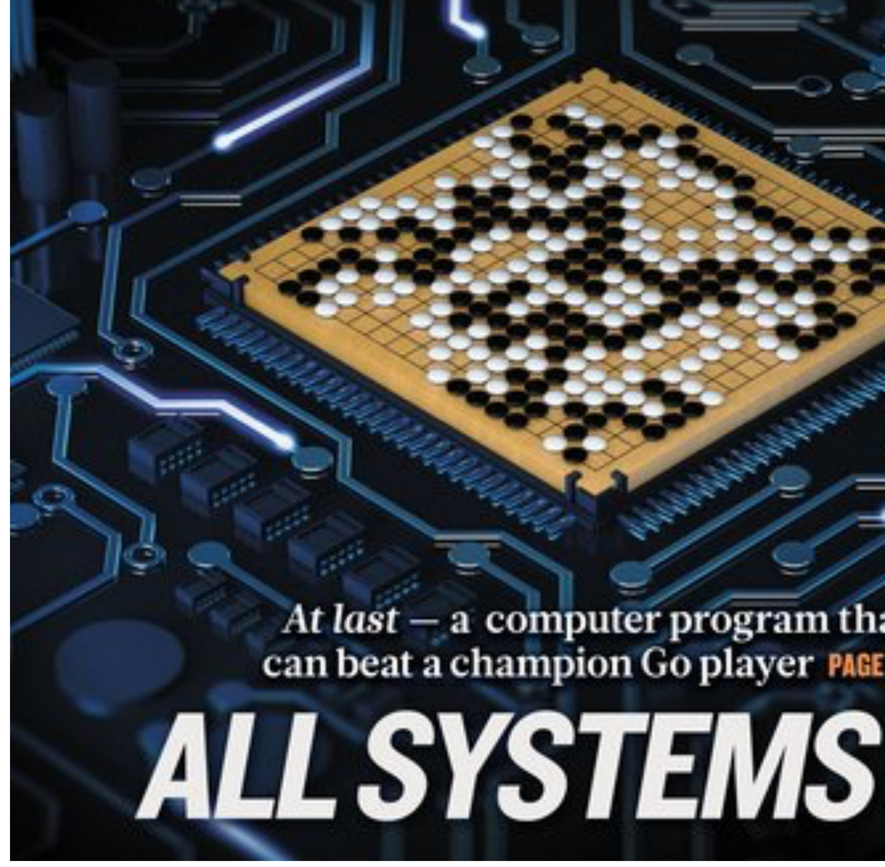


- Thm: In the episodic setting, $poly(S, A, H, 1/\epsilon)$ samples suffice to find an ϵ-opt policy with the E^3 algo. [Kearns & Singh '98]
  also: [Brafman& Tennenholtz '02; K. '03]
  Key idea: optimism + dynamic programming

- Regret guarantees with model based algos:
  [Auer+ '09]
- Provable Q-learning (+bonus):
  [Strehl+ (2006)], [Szita & Szepesvari '10],[Jin+ '18]
- (asymptotically) optimal reget: $Reg(\#episodes) = \sqrt{HSA \cdot \#episodes}$
  [Azar+ '17],[Dann+'17]

I: Provable Generalization in RL

Q1: Can we find an ϵ-opt policy with no S dependence?

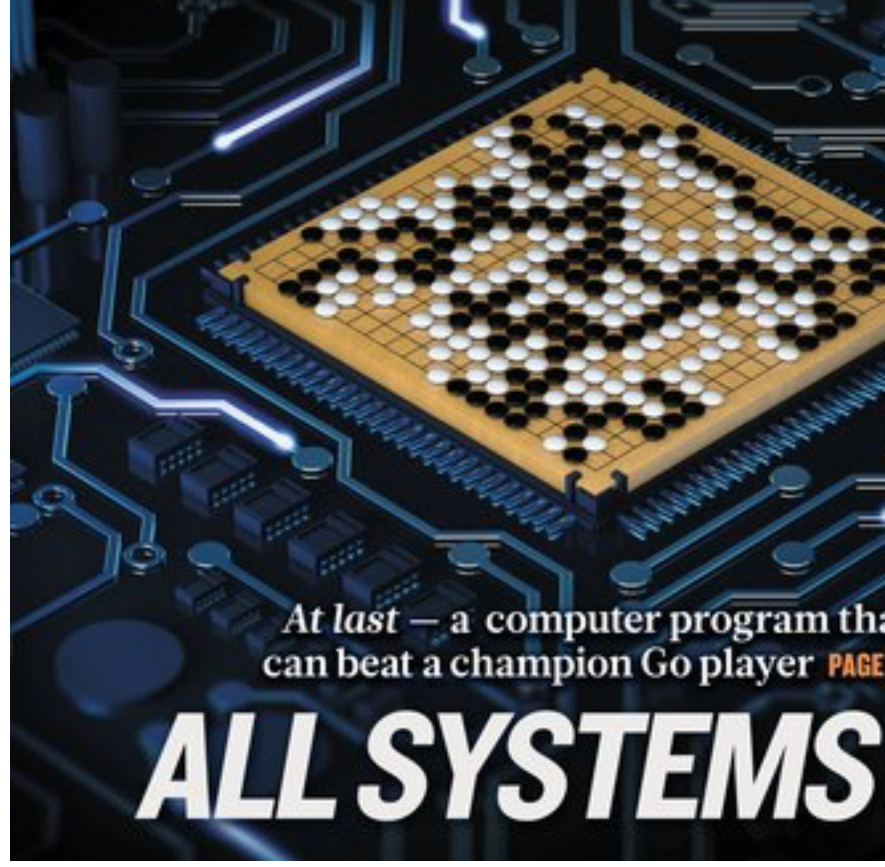


- How can we reuse data to estimate the value of all policies in a policy class $\mathscr{F}$?
  Idea: Trajectory tree algo
  dataset collection: uniformly at random choose actions for all H steps in an episode.
  estimation: uses importance sampling to evaluate every $f \in \mathscr{F}$

- Thm:[Kearns, Mansour, & Ng '00]
  To find an ϵ-best in class policy, the trajectory tree algo uses $O(A^H \log(|\mathscr{F}|)/\epsilon^2)$ samples
  - Only $\log(|\mathscr{F}|)$ dependence on hypothesis class size.
  - There are VC analogues as well.

- Can we avoid the 2^H dependence to find an ϵ-best-in-class policy?
  Agnostically, **NO**!
  Proof: Consider a binary tree with 2^H-policies and a sparse reward at a leaf node.

II: Provable Generalization in RL



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

- With various stronger assumptions, yes.
  - 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
  - FLAMBE / 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]

This talk:

Structural conditions

Under which Generalization is possible

Warm up

The linear Bellman Complete model:

Def: Given feature map $\phi(s, a) \in \mathbb{R}^d$, Bellman operator has linear closure

Given any $w \in \mathbb{R}^d$, there exists a $\theta \in \mathbb{R}^d$, such that:

$$\forall s, a : \theta^\top \phi(s, a) = r(s, a) + \mathbb{E}_{s' \sim P(s,a)} \left[\max_{a'} w^\top \phi(s', a') \right]$$

$:= T(w)$

Examples:

Linear MDPs, Linear Quadratic Regulator

Warm up

The linear Bellman Complete model:

Generalization is possible here:

$\exists$ an algorithm, finding ϵ-near optimal policy only needs poly$(H, d, 1/\epsilon)$ many samples

Warm up

The linear Bellman Complete model:

what's the structure here that permits generalization?

We can rewrite the average Bellman error (averaged over any roll-in π)

in a bilinear form:

Given a $Q^\star$candidate $w^\top \phi(s,a)$, we have:

$$\mathbb{E}_{s,a\sim\pi}\left[w^\top \phi(s,a) - r(s,a) - \mathbb{E}_{s'\sim P(s,a)} \max_{a'} w^\top \phi(s',a')\right]$$

$$= \mathbb{E}_{s,a\sim\pi}\left[w^\top \phi(s,a) - T(w)^\top \phi(s,a)\right] = \left\langle w - T(w), \mathbb{E}_{s,a\sim\pi}\phi(s,a)\right\rangle$$

Warm up

The linear Bellman Complete model:

what's the unique structure here that permits generalization?

We can also estimate the value of the bilinear form:

Define discrepancy $\ell(s, a, s', w) = w^\top \phi(s, a) - r(s, a) - \max_{a'} w^\top \phi(s', a')$

We have:

$$\mathbb{E}_{s,a\sim\pi}\ell(s, a, s', w) = \left\langle w - T(w), \mathbb{E}_{s,a\sim\pi}\phi(s, a) \right\rangle$$

Note we have data reuse: given data from π, we can evaluate all w

Warm up

The linear Bellman Complete model:

In summary, it has a **bilinear structure**:

For any roll-in policy π, and any w, we have: (1)

$$\mathbb{E}_{s,a\sim\pi}\left[w^\top\phi(s,a) - r(s,a) - \mathbb{E}_{s'\sim P(s,a)}\max_{a'} w^\top\phi(s',a')\right] = \left\langle w - T(w), \mathbb{E}_\pi\phi(s,a)\right\rangle$$

AND (2) there exists a discrepancy function ℓ, s.t.,

$$\mathbb{E}_{s,a\sim\pi}\ell(s,a,s',w) = \left\langle w - T(w), \mathbb{E}_\pi\phi(s,a)\right\rangle$$

Note that the analytical form of bilinear structure is unknown

BiLinear Classes: structural properties to enable generalization in RL

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

Def: A $(\mathscr{F}, \ell)$ forms **an (implicit) Bilinear class** class if there are $W_h \in \mathscr{F} \mapsto \mathscr{H}$, & $X_h \in \mathscr{F} \mapsto \mathscr{H}$ ($\mathscr{H}$ being some Hilbert space):

- Bilinear regret: on-policy difference between claimed reward and true reward
  $$\left| \mathbb{E}_{s_h, a_h \sim \pi_f}\left[Q_f(s_h, a_h) - r(s_h, a_h) - V_f(s_{h+1})\right] \right| \leq \left\langle W_h(f) - W_h(f^\star), X_h(f) \right\rangle$$
- Data reuse: there is discrepancy function $\ell_f(s, a, s', g)$ & policy π_{est} s.t.
  $$\mathbb{E}_{s_h \sim \pi_f, a_h \sim \pi_{est}}\left[\ell_f(s_h, a_h, s_{h+1}, g)\right] = \left\langle W_h(g) - W_h(f^\star), X_h(f) \right\rangle, \forall g \in \mathscr{F}$$

Note: W_h & X_h are implicit—no need to known them

Back to Linear Bellman Complete:

(1) **Bilinear regret:** for any $f(s,a) := w^\top \phi(s,a)$, we have:

$$\mathbb{E}_{s_h,a_h\sim\pi_f}\left[w^\top\phi(s_h,a_h) - r(s_h,a_h) - \mathbb{E}_{s'\sim P(s,a)}\max_{a'} w^\top\phi(s',a')\right] = \left\langle \underbrace{(w - T(w))}_{W_h(f)} - \underbrace{(w^\star - T(w^\star))}_{W_h(f^\star)}, \underbrace{\mathbb{E}_\pi \phi(s,a)}_{X_h(f)} \right\rangle$$

AND (2) **data-reuse**: there exists a discrepancy function ℓ, s.t.,

$$\mathbb{E}_{\pi_f}\ell(s,a,s',w') = \left\langle (w' - T(w')) - (w^\star - T(w^\star)), \mathbb{E}_{s,a\sim\pi_f}\phi(s,a) \right\rangle$$

Note $W_h(f)$ is unknown as the Bellman backup $T(w)$ is unknown

The Algorithm: BiLin-UCB

For $t = 0 \to T$:

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

$$\arg\max_{f \in \mathcal{F}} V_f(s_0), \text{ s.t., } \sigma_h^2(f) \leq R, \forall h$$

- Sample m trajectories π_{f_t} and create a batch dataset:

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

- Update the cumulative discrepancy function $\sigma_h(\cdot), \forall h$

$$\sigma_h^2(\cdot) \leftarrow \sigma_h^2(\cdot) + \left(\sum_{(s_h, a_h, s_{h+1}) \in D} \ell_{f_t}(s_h, a_h, s_{h+1}, \cdot) / |D| \right)^2$$

Note here we roughly have: $\sigma_h(g) \approx \sum_{i=0}^{t} \left(\mathbb{E}_{s_h, a_h \sim \pi_{f_i}} \ell_{f_i}(s_h, a_h, s_{h+1}, g) \right)^2$

Theorem 2: Generalization in RL

- Theorem: [Du, Kakade., Lee, Lovett, Mahajan, S, Wang '21]
  Assume $\mathscr{F}$ is a bilinear class and the class is realizable, i.e. $f^\star \in \mathscr{F}$.
  Using $\gamma_T^3 \cdot poly(H) \cdot \log(1/\delta)/\epsilon^2$ trajectories, the BiLin-UCB algorithm returns an ϵ-opt policy (with prob. $\geq 1 - \delta$).
  - γ_T is the max. info. gain $\gamma_T := \max_{h, f_0 \ldots f_{T-1} \in \mathscr{F}} \ln \det \left(I + \frac{1}{\lambda} \sum_{t=0}^{T-1} X_h(f_t) X_h(f_t)^\top \right)$
  - $\gamma_T \approx d \log T$ for X_h in d-dimensions

- The proof is "elementary" using the elliptical potential function.
  [Dani, Hayes, K. '08]

Proof Sketch

1. Optimism: $V^\star(s_0) \leq V_{f_t}(s_0)$; this can be verified by showing $f^\star$ is always a feasible solution

2. Using optimism, we can upper bound per-episode regret Bilinear form ("simulation" lemma):

$$V^\star(s_0) - V^{\pi_{f_t}}(s_0) \leq V_{f_t}(s_0) - V^{\pi_{f_t}}(s_0) \leq \sum_{h=0}^{H-1} \left| W_h(f_t) - W_h(f^\star), X_h(f_t) \right|$$

3. If π_{f_t} is really sub-optimal, i.e., $V^\star(s_0) - V^{\pi_{f_t}}(s_0) \geq \epsilon,$ then f_t has large bilinear regret:

$$\exists h, \text{ s.t., } \left| W_h(f_t) - W_h(f^\star), X_h(f_t) \right| \geq \epsilon / H$$

4. Recall in the alg, we have a constraint $\sigma_h(f_t) \leq R$, i.e., $\sum_{i=0}^{t-1} \left(\mathbb{E}_{s_h, a_h \sim \pi_{f_i}} \left[\ell_{f_i}(s_h, a_h, s_{h+1}, f_t) \right] \right)^2 \leq R$

$$\sum_{i=0}^{t-1} \left(W_h(f_t) - W_h(f^\star), X_h(f_i) \right)^2 \leq R$$

$$\Rightarrow (W_h(f_t) - W_h(f^\star))^\top \Sigma_{t;h} (W_h(f_t) - W_h(f^\star)) \leq R + \lambda \qquad \left(\Sigma_{h;t} := \sum_{i=0}^{t-1} X_h(f_i) X_h(f_i)^\top + \lambda I \right)$$

Proof Sketch

3. If π_{f_t} is really **sub-optimal**, i.e., $V^\star(s_0) - V^{\pi_{f_t}}(s_0) \geq \epsilon$, then f_t has large bilinear value:

$$\exists h, \text{ s.t., } \left| W_h(f_t) - W_h(f^\star), X_h(f_t) \right| \geq \epsilon / H$$

4. Recall in the alg, we have a constraint $\sigma(f_t) \leq R$, i.e., $\sum_{\tau=0}^{t-1} \left(\mathbb{E}_{s_h, a_h \sim \pi_{f_\tau}} \ell_{f_\tau}(s_h, a_h, s_{h+1}, f_t) \right)^2 \leq R$

$$(W_h(f_t) - W_h(f^\star))^\top \Sigma_{t;h} (W_h(f_t) - W_h(f^\star)) \leq R + \lambda \qquad \left(\Sigma_{h;t} := \sum_{i=0}^{t-1} X_h(f_i) X_h(f_i)^\top + \lambda I \right)$$

5. Finally, combine the two results and using Cauchy-Schwartz, we have:

$$\epsilon / H \leq \left\| W_h(f_t) - W_h(f^\star) \right\|_{\Sigma_{h;t}} \left\| X_h(f_t) \right\|_{\Sigma^{-1}_{h;t}} \leq \sqrt{R + \lambda} \left\| X_h(f_t) \right\|_{\Sigma^{-1}_{h;t}}$$

$$\Rightarrow \left\| X_h(f_t) \right\|_{\Sigma^{-1}_{h;t}} \geq \frac{\epsilon}{H \sqrt{R + \lambda}}$$

In d-dimensional setting, such event cannot happen more than $\widetilde{O}(d/\epsilon^2)$ many times….

- Theorem: [Du, Kakade., Lee, Lovett, Mahajan, S, Wang '21]

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

  - 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
  - FLAMBE / 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]
  - Linear $Q^\star$ & $V^\star$
  - Reactive PSR/POMDP, Generalized linear MDP, and more…..

- (almost) all "named" models (with provable generalization) are bilinear classes

  two exceptions: deterministic linear $Q^\star$; $Q^\star$-state-action aggregation
- Bilinear classes generalize the: **Bellman rank [Jiang+ '17]; Witness rank [Sun+ '19]**

Another Example: Feature Selection for Low-rank MDP

The feature selection problem:

Low-rank MDP: $P(s' \mid s, a) = \mu^\star(s')^\top \phi^\star(s, a),\ \mu^\star$ & $\phi^\star$ unknown

Function approximation for feature: $\phi^\star \in \Psi \subset S \times A \mapsto \mathbb{R}^d$

Function approximation for $Q^\star$: $\mathcal{Q} := \left\{ w^\top \phi(s, a) : w \in \mathrm{Ball}_W, \phi \in \Psi \right\}$

Q: can we find ϵ optimal policy w/ samples poly(d, ln(Ψ), H,$1/\epsilon$)?

Yes & it's in the Bilinear class!

Another Example: Feature Selection for Low-rank MDP

The feature selection problem:

1. Claim: on-policy Bellman error of $f := w^\top \phi$ has Bilinear form:

$$\forall f \in \mathcal{Q} : \mathbb{E}_{s_h, a_h \sim \pi_f} \left[Q_f(s_h, a_h) - r_h - \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q_f(s_{h+1}, a') \right]$$

$$= \mathbb{E}_{s_{h-1}, a_{h-1} \sim \pi_f} \mathbb{E}_{s_h \sim P(\cdot|s_{h-1}, a_{h-1}), a_h \sim \pi_f(\cdot|s_h)} \left[Q_f(s_h, a_h) - r_h - \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q_f(s_{h+1}, a') \right]$$

$$= \mathbb{E}_{s_{h-1}, a_{h-1} \sim \pi_f} \int_{s_h} \phi^\star(s_{h-1}, a_{h-1})^\top \mu^\star(s_h) \, \mathbb{E}_{a_h \sim \pi_f} \left[Q_f(s_h, a_h) - r_h - \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q_f(s_{h+1}, a') \right]$$

$$= \left\langle \underbrace{\mathbb{E}_{s_{h-1}, a_{h-1} \sim \pi_f} \phi^\star(s_{h-1}, a_{h-1}),}_{X_h(f)} \quad \underbrace{\int_{s_h} \mu^\star(s_h) \mathbb{E}_{a_h \sim \pi_f} \left[Q_f(s_h, a_h) - r_h - \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q_f(s_{h+1}, a') \right]}_{W_h(f)} \right\rangle$$

Another Example: Feature Selection for Low-rank MDP

The feature selection problem:

2. Claim: The bilinear regret is estimable using some ℓ

Define $\ell(s, a, s', g) = \dfrac{\mathbf{1}\{a = \pi_g(s)\}}{1/A} \left(Q_g(s, a) - r(s, a) - \max_{a'} Q_g(s', a') \right)$

$$\mathbb{E}_{s_h \sim \pi_f} \mathbb{E}_{a_h \sim U(A)} \ell(s_h, a_h, s'_{h+1}, g) = \left\langle W_h(g) - W_h(f^\star), X_h(f) \right\rangle$$

Another Example: Linear $Q^\star$ & $V^\star$

What's the role of linear function approximation in RL?

We have linear MDP, Linear Bellman complete...

The most natural one should just be $Q^\star$ being linear-realizable:

$$Q^\star(s,a) = (w^\star)^\top \phi(s,a), \text{ under known feature } \phi$$

However generalization is **impossible** here:

- Theorem [Weisz, Amortila, Szepesvári '21]: There exists an MDP w/ linear $Q^\star$, s.t any online RL algorithm requires $\Omega(\min(2^d, 2^H))$ samples to output a near optimal policy

Additional Example: Linear $Q^\star$ & $V^\star$

What's the role of linear function approximation in RL?

However, if we further assume $V^\star(s) = (\theta^\star)^\top \psi(s)$, then we will be ok:

- Theorem [Du, Kakade., Lee, Lovett, Mahajan, S, Wang '21]
  For any MDP with both $Q^\star$ and $V^\star$ being realizable (under some known features, e.g., RKHS), there exists an algorithm that learns with # of samples:
  $$\frac{d^3 \text{poly}(H)}{\epsilon^2}$$

Additional Example: Linear $Q^\star$ & $V^\star$

Linear $Q^\star$ & $V^\star$ has Bilinear Structure

Function classes for $Q^\star \in \mathcal{Q} := \left\{ w^\top \phi(s,a) : w \in \mathrm{Ball}_w \right\}, V^\star \in \mathcal{V} := \left\{ \theta^\top \psi(s) : \theta \in \mathrm{Ball}_B \right\}$

Key step: pre-process function class

$$\left\{ Q, V \right\} := \left\{ (w, \theta) : \forall s, \max_a w^\top \phi(s,a) = \theta^\top \psi(s) \right\}$$

(1) on-policy Bellman error of any $f := (w, \theta)$ has bilinear form:

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[w^\top \phi(s_h, a_h) - r_h - \mathbb{E}_{s' \sim P(s_h, a_h)} \theta^\top \psi(s') \right] = \left\langle W_h([w, \theta]) - W_h([w^\star, \theta^\star]), X_h([w, \theta]) \right\rangle$$

(2) there exists $\ell(s, a, s', [w, \theta]) = w^\top \phi(s,a) - r(s,a) - \theta^\top \psi(s')$, s.t.,

$$\mathbb{E}_{s_h, a_h \sim \pi_f} \left[\ell(s_h, a_h, s'_{h+1}, [w', \theta']) \right] = \left\langle W_h([w', \theta']) - W_h([w^\star, \theta^\star]), X_h(f) \right\rangle, \forall [w', \theta']$$

Final Example: Linear Mixture Model (model-based)

The linear Mixture Model:

Function class: $\mathscr{F} = \left\{ P : P(s' \mid s, a; \theta) = \theta^\top \phi(s, a, s'), \theta \in \mathrm{Ball}_W \right\}$

Why this model is ever interesting?

Imagine we have d simulators, $P^i(s' \mid s, a), i \in [d]$,

we assume the ground truth is the linear mixture of d simulators:

$$P^\star(s' \mid s, a) = \sum_{i=1}^{d} \theta^\star[i] P^i(s' \mid s, a)$$

Final Example: Linear Mixture Model (model-based)

The linear Mixture Model:

Function class: $\mathcal{F} = \left\{ P : P(s' \mid s, a; \theta) = \theta^\top \phi(s, a, s'), \theta \in \mathrm{Ball}_W \right\}$

(Notation: $f \in \mathcal{F}$ is a potential transition, and $f^\star := P^\star$)

Claim 1: For any $f \in \mathcal{F}$, the on-policy Bellman error of Q_f under π_f has bilinear form:

$$\mathbb{E}_{\pi_f}\left[Q_f(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s' \sim f^\star(s_h, a_h)} V_f(s') \right] = \left\langle W_h(f) - W_h(f^\star), X_h(f) \right\rangle$$

Final Example: Linear Mixture Model

The linear Mixture Model:

Function class: $\mathscr{F} = \left\{ P : P(s' \mid s, a; \theta) = \theta^\top \phi(s, a, s'), \theta \in \mathrm{Ball}_W \right\}$

(Notation: $f \in \mathscr{F}$ is a potential transition, and $f^\star := P^\star$)

Claim 2: For any $f \in \mathscr{F}$, there exists discrepancy $\ell_f(s, a, s', g)$ to measure bilinear form:

$$\ell_f(s, a, s', g) = \mathbb{E}_{s' \sim g(s,a)} \left[V_f(s') \right] - V_f(s')$$

$$\text{s.t., } \mathbb{E}_{\pi_f} \ell_f(s_h, a_h, s'_{h+1}, g) = \left\langle W_h(g) - W_h(f^\star), X_h(f) \right\rangle$$