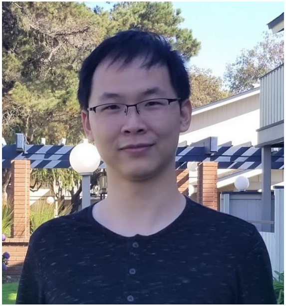
Bellman Eluder Dimension: New Rich Classes of RL Problems, Sample-Efficient Algorithms

based on joint work with Chi Jin and Sobhan Miryoosefi

Qinghua Liu

Princeton University

Collaborators



Chi Jin

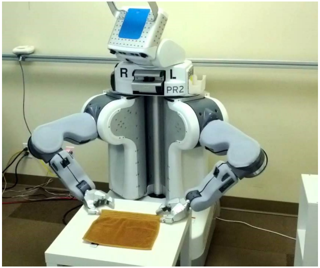
Sobhan Miryoosefi

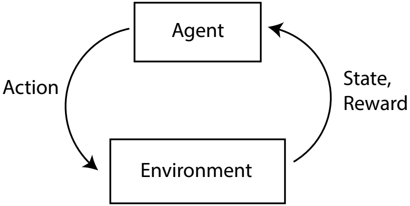
Overview

Sequential Decision Making

Main framework: **Reinforcement Learning (RL)**.

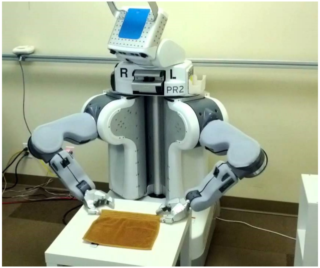
Reinforcement Learning

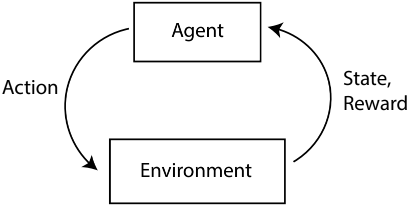




Markov decision process MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$.

Reinforcement Learning

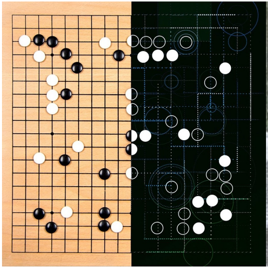




Markov decision process MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$.

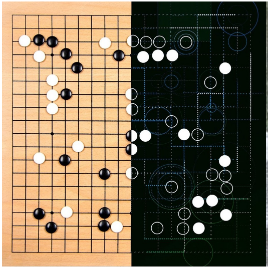
Goal: find the best policy that maximizes the cumulative rewards.

Efficiency



- **Sample efficiency:** collecting samples can be expensive.
- **Computational efficiency:** training deep RL costs weeks.

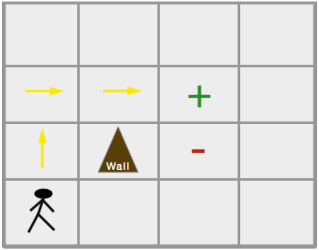
Efficiency

- **Sample efficiency:** collecting samples can be expensive.
- **Computational efficiency:** training deep RL costs weeks.

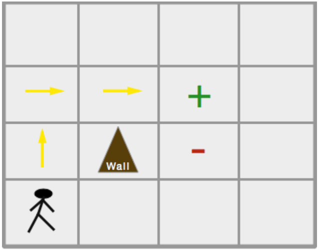
AlphaGo Zero: trained on $\geq 10^7$ games, and took ≥ 1 month.

Classical RL: Tabular Case



The numbers of states & actions are finite and small.

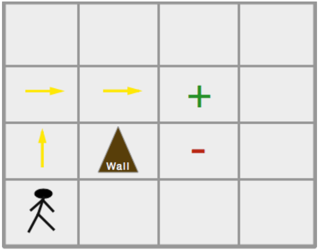
Classical RL: Tabular Case



The numbers of states & actions are finite and small.

Strategy: visit all "reachable" states, and learn directly.

Classical RL: Tabular Case



The numbers of states & actions are finite and small.

Strategy: visit all "reachable" states, and learn directly.

Abundant theoretical results. Near-optimal algorithms have been designed. [see, e.g., AOM17, JABJ19]

Modern RL: Function Approximation

The number of states in practice is typically $\geq 10^{100}$.

Most states are not visited even once.

Modern RL: Function Approximation

The number of states in practice is typically $\geq 10^{100}$.

Most states are not visited even once.

Strategy: approximate "value" or "policy" by functions in a parametric class $\mathcal{F}$ (such as deep nets).

Challenges in Function Approximation

- *Generalization*: generalize knowledge from the visited states to the unobserved ones.

Challenges in Function Approximation

- *Generalization*: generalize knowledge from the visited states to the unobserved ones.
- *Limited expressiveness*: handle functions outside given class $\mathcal{F}$.

Challenges in Function Approximation

- *Generalization*: generalize knowledge from the visited states to the unobserved ones.
- *Limited expressiveness*: handle functions outside given class $\mathcal{F}$.
- *Exploration*: address exploration vs. exploitation tradeoff.

Challenges in Function Approximation

- *Generalization*: generalize knowledge from the visited states to the unobserved ones.
- *Limited expressiveness*: handle functions outside given class $\mathcal{F}$.
- *Exploration*: address exploration vs. exploitation tradeoff.

Most existing theories focus on special cases under strong assumptions, such as linear approximation [JYWJ20, ZLKB20], LQR [DMM$^+$19].

Main Question

What are the minimal structural assumptions that empower sample-efficient RL?

Main Question

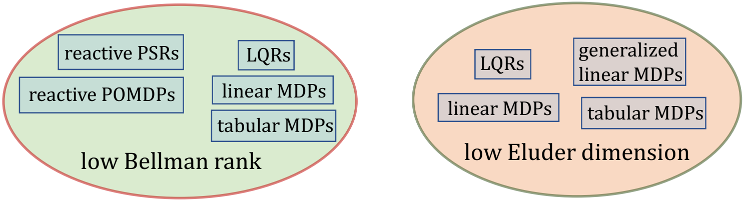
What are the minimal structural assumptions that empower sample-efficient RL?

1. identify a rich class of RL problems.
2. design sample-efficient algorithms for this class.

Previous Attempts

Known classes of RL problems that can be learned sample-efficiently.

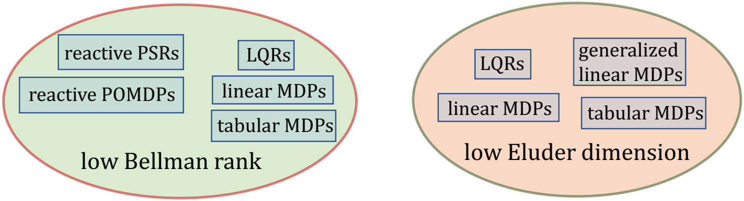
- [JKA$^+$17]: low Bellman rank.
- [WSY20]: low Eluder dimension $+$ completeness.



Previous Attempts

Known classes of RL problems that can be learned sample-efficiently.

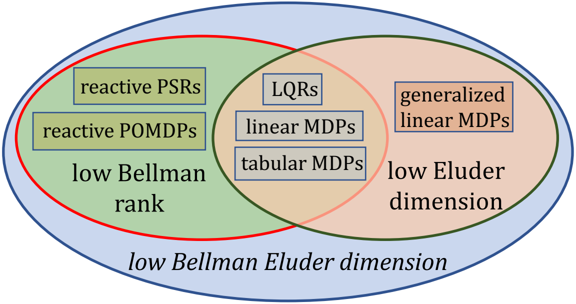
- [JKA$^+$17]: low Bellman rank.
- [WSY20]: low Eluder dimension + completeness.



Definitions and algorithms look very different.

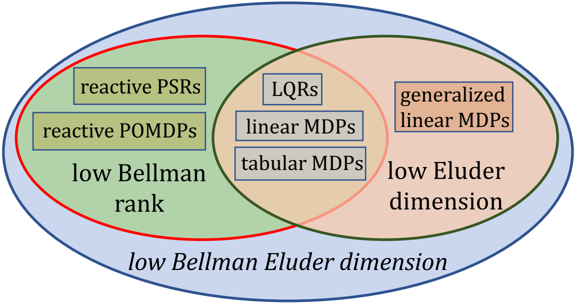
A New Rich Class

This Talk: a new complexity measure—**Bellman Eluder Dimension**.



A New Rich Class

This Talk: a new complexity measure—**Bellman Eluder Dimension**.



contains a majority of known tractable RL problems.

Sample-efficient Algorithms for the New Class

GOLF (new):

- optimization-based, with optimism.
- surprisingly simple and clean.
- regret and sample complexity results match or improve the best existing results for several well-known subclasses.

Sample-efficient Algorithms for the New Class

GOLF (new):

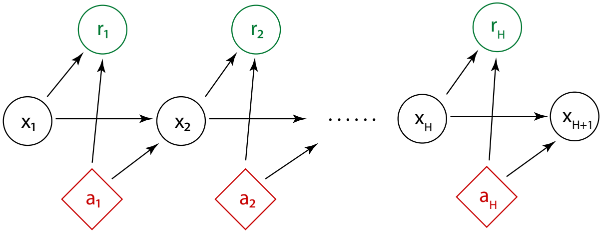
- optimization-based, with optimism.
- surprisingly simple and clean.
- regret and sample complexity results match or improve the best existing results for several well-known subclasses.

OLIVE [JKA$^+$]:

- based on hypothesis elimination.
- new analyses for general classes.

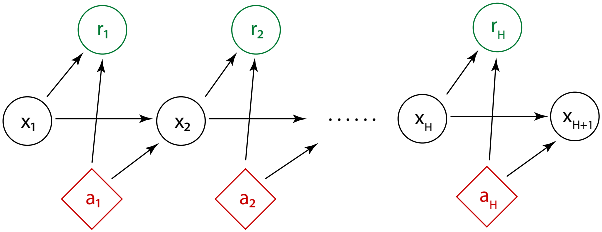
Formal Setups

Episodic MDP



MDP$(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$: Each episode has H steps. Transition probability $\mathbb{P}_h(\cdot|s, a)$, reward $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$.

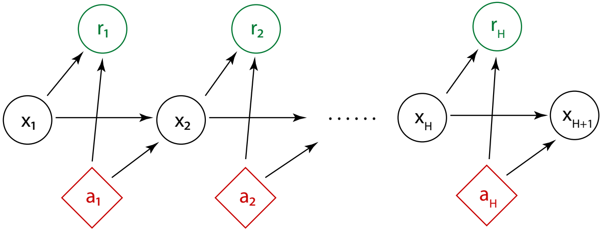
Episodic MDP



MDP$(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$: Each episode has H steps. Transition probability $\mathbb{P}_h(\cdot|s,a)$, reward $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$.

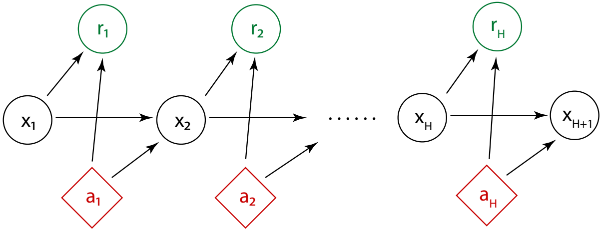
Fixed initial state s_1, the agent only picks action $\{a_h\}_{h=1}^{H}$.

Policy and Value



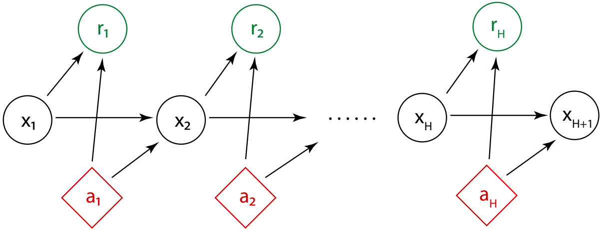
- **Policy:** A map from state to action $\pi : \mathcal{S} \to \mathcal{A}$.

Policy and Value



- **Policy:** A map from state to action $\pi : \mathcal{S} \to \mathcal{A}$.
- **Value:** Expected cumulative reward starting at step h from each state $V_h^\pi(s)$ or each state-action pair $Q_h^\pi(s, a)$.

Policy and Value



- **Policy:** A map from state to action $\pi : \mathcal{S} \to \mathcal{A}$.
- **Value:** Expected cumulative reward starting at step h from each state $V_h^\pi(s)$ or each state-action pair $Q_h^\pi(s, a)$.

- **Objective:** find the optimal policy to maximize the value $V_1^\pi(s_1)$.

Bellman Error

There exists an optimal policy $\pi^\star \to$ optimal value $Q^\star$.

Bellman Error

There exists an optimal policy $\pi^\star \to$ optimal value $Q^\star$.

Bellman optimality equation:

$$Q_h^\star(s,a) = (\mathcal{T}_h Q_{h+1}^\star)(s,a) := r_h(s,a) + \mathbb{E}_{s' \sim \Pr_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q_{h+1}^\star(s',a').$$

Bellman Error

There exists an optimal policy $\pi^\star \to$ optimal value $Q^\star$.

Bellman optimality equation:

$$Q_h^\star(s,a) = (\mathcal{T}_h Q_{h+1}^\star)(s,a) := r_h(s,a) + \mathbb{E}_{s' \sim \Pr_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q_{h+1}^\star(s',a').$$

Bellman error:

$$\mathcal{E}(f,\rho,h) := \mathbb{E}_{(s,a)\sim\rho} \underbrace{(f_h - \mathcal{T}_h f_{h+1})(s,a)}_{\text{Bellman residual function}}$$

Function Approximation

Value function approximation: given function class $\mathcal{F} = \mathcal{F}_1 \times \ldots \times \mathcal{F}_H$, use $f = (f_1, \cdots, f_H) \in \mathcal{F}$ to approximate $Q^\star = (Q_1^\star, \ldots, Q_H^\star)$.

Function Approximation

Value function approximation: given function class $\mathcal{F} = \mathcal{F}_1 \times \ldots \times \mathcal{F}_H$, use $f = (f_1, \cdots, f_H) \in \mathcal{F}$ to approximate $Q^\star = (Q_1^\star, \ldots, Q_H^\star)$.

Common assumptions:

1. realizable: $\forall h \in [H],\ Q_h^\star \in \mathcal{F}_h$.

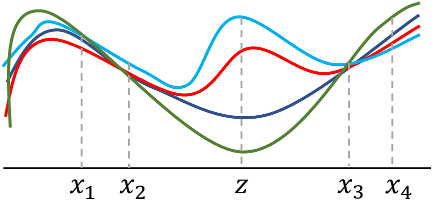
Function Approximation

Value function approximation: given function class $\mathcal{F} = \mathcal{F}_1 \times \ldots \times \mathcal{F}_H$, use $f = (f_1, \cdots, f_H) \in \mathcal{F}$ to approximate $Q^\star = (Q_1^\star, \ldots, Q_H^\star)$.

Common assumptions:

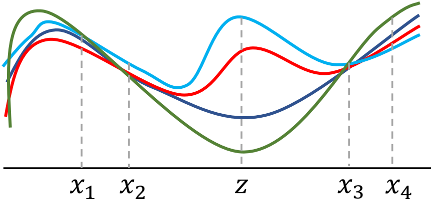
1. realizable: $\forall h \in [H],\ Q_h^\star \in \mathcal{F}_h$.
2. completeness: $\forall h \in [H],\ \mathcal{T}_h \mathcal{F}_{h+1} \subset \mathcal{F}_h$.

Eluder Dimension



Point z is ϵ-independent of $\{x_1, x_2, \ldots, x_n\}$ w.r.t. $\mathcal{F}$ if $\exists f, g \in \mathcal{F}$ such that $\sqrt{\sum_i (f(x_i) - g(x_i))^2} \leq \epsilon$ for all $i \in [n]$, but $f(z) - g(z) > \epsilon$.

Eluder Dimension



Point z is ϵ-independent of $\{x_1, x_2, \ldots, x_n\}$ w.r.t. $\mathcal{F}$ if $\exists f, g \in \mathcal{F}$ such that $\sqrt{\sum_i (f(x_i) - g(x_i))^2} \leq \epsilon$ for all $i \in [n]$, but $f(z) - g(z) > \epsilon$.

Eluder dimension [RV13] $\dim_{\mathrm{E}}(\mathcal{F}, \epsilon)$:

The length of the longest sequence $\{x_j\}_{j=1}^n$ such that $\exists \epsilon' \geq \epsilon$ where x_i is ϵ'-independent of $\{x_j\}_{j=1}^{i-1}$ for all $i \in [n]$.

Bellman Eluder dimension

Distributional Eluder Dimension

Points $\rightarrow$ distributions.

Distribution μ is ϵ-independent of $\{\nu_1, \nu_2, \dots, \nu_n\}$ w.r.t. $\mathcal{F}$ if $\exists f \in \mathcal{F}$ such that $\sqrt{\sum_i (\mathbb{E}_{\nu_i} f)^2} \leq \epsilon$ for all $i \in [n]$, but $|\mathbb{E}_\mu f| > \epsilon$.

Distributional Eluder Dimension

Points $\rightarrow$ distributions.

Distribution μ is ϵ-independent of $\{\nu_1, \nu_2, \ldots, \nu_n\}$ w.r.t. $\mathcal{F}$ if $\exists f \in \mathcal{F}$ such that $\sqrt{\sum_i (\mathbb{E}_{\nu_i} f)^2} \leq \epsilon$ for all $i \in [n]$, but $|\mathbb{E}_\mu f| > \epsilon$.

Distributional Eluder dimension $\dim_{\mathrm{DE}}(\mathcal{F}, \Pi, \epsilon)$:

The length of the longest sequence $\{\nu_j\}_{j=1}^n \subset \Pi$ such that $\exists \epsilon' \geq \epsilon$ where ν_i is ϵ'-independent of $\{\nu_j\}_{j=1}^{i-1}$ for all $i \in [n]$.

Bellman Eluder Dimension

Bellman Eluder (BE) dimension $\dim_{\mathrm{BE}}(\mathcal{F}, \Pi, \epsilon)$

$$\dim_{\mathrm{BE}}(\mathcal{F}, \Pi, \epsilon) := \max_{h\in[H]} \dim_{\mathrm{DE}} \left((I - \mathcal{T}_h)\mathcal{F}, \Pi_h, \epsilon \right)$$

- $(I - \mathcal{T}_h)\mathcal{F} := \{f_h - \mathcal{T}_h f_{h+1} : \ f \in \mathcal{F}\}$: Bellman residuals at step h.
- $\Pi = \{\Pi_h\}_{h=1}^{H}$: a collection of distributions over $\mathcal{S} \times \mathcal{A}$.

Bellman Eluder Dimension

Bellman Eluder (BE) dimension $\dim_{\mathrm{BE}}(\mathcal{F}, \Pi, \epsilon)$

$$\dim_{\mathrm{BE}}(\mathcal{F}, \Pi, \epsilon) := \max_{h\in[H]} \dim_{\mathrm{DE}}\left((I - \mathcal{T}_h)\mathcal{F}, \Pi_h, \epsilon\right)$$

- $(I - \mathcal{T}_h)\mathcal{F} := \{f_h - \mathcal{T}_h f_{h+1} : f \in \mathcal{F}\}$: Bellman residuals at step h.
- $\Pi = \{\Pi_h\}_{h=1}^{H}$: a collection of distributions over $\mathcal{S} \times \mathcal{A}$.

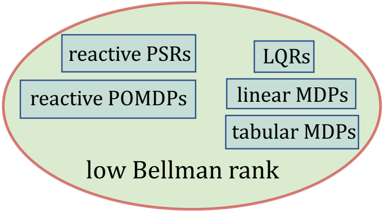
Typical choices of Π:

- $\mathcal{D}_{\mathcal{F}}$: distributions generated by executing π_f greedy w.r.t $f \in \mathcal{F}$.
- $\mathcal{D}_{\Delta}$: all Dirac distributions over $\mathcal{S} \times \mathcal{A}$.

Relation to Bellman Rank

Bellman rank (type-I) is the minimum integer d, so that $\forall h \in [H]$, $\exists \phi_h, \psi_h : \mathcal{F} \to \mathbb{R}^d$, $\forall f, g \in \mathcal{F}$:

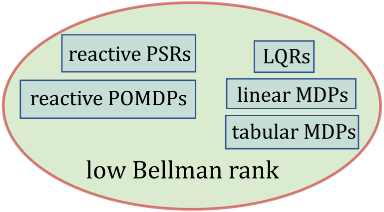
$$\mathcal{E}(f, \pi_g, h) := \mathbb{E}_{\pi_g}[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)] = \langle \phi_h(f), \psi_h(g) \rangle.$$



Relation to Bellman Rank

Bellman rank (type-I) is the minimum integer d, so that $\forall h \in [H]$, $\exists \phi_h, \psi_h : \mathcal{F} \to \mathbb{R}^d$, $\forall f, g \in \mathcal{F}$:

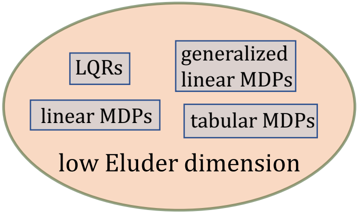
$$\mathcal{E}(f, \pi_g, h) := \mathbb{E}_{\pi_g}[(f_h - \mathcal{T}_h f_{h+1})(s_h, a_h)] = \langle \phi_h(f), \psi_h(g) \rangle.$$



low Bellman rank $\subset$ low BE dimension

$$\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon) \;\leq\; \mathcal{O}(\text{Bellman rank} \cdot \log(1/\epsilon)).$$

Relation to Eluder Dimension

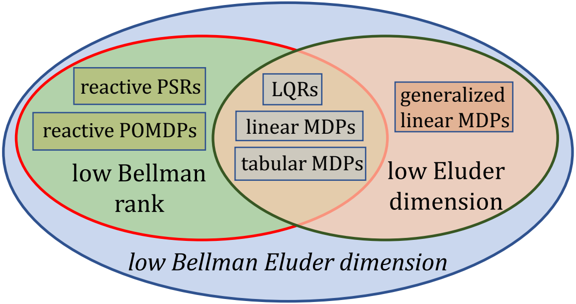


low Eluder dimension $\subset$ low BE dimension

Assume completeness,

$$\dim_{\mathrm{BE}}(\mathcal{F}, \mathcal{D}_\Delta, \epsilon) \;\leq\; \dim_{\mathrm{E}}(\mathcal{F}, \epsilon).$$

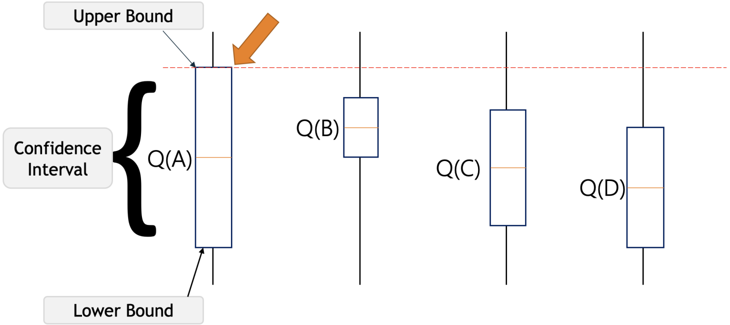
Summary of Relations



The class of low BE dimension problems contains a majority of known RL problems learnable in polynomial samples.

Sample-Efficient Algorithm

Upper Confidence Bounds Algorithm



1. pull arms optimistically, **2. collect rewards**

3. update confidence intervals.

GOLF Algorithm

Global Optimism based on Local Fitting (GOLF)

for $k = 1, ..., K$

1. **optimistic planning**

   $\pi^k = \pi_{f^k}$, where $f^k = \mathrm{argmax}_{f \in \mathcal{B}} f_1(s_1, \pi_f(s_1))$.

2. **data collection**

   execute π^k to collect a trajectory $(s_1, a_1, \ldots, s_H, a_H)$.

3. **update confidence set $\mathcal{B}$.**

output π^{out} sampled uniformly from $\{\pi^k\}_{k=1}^K$.

GOLF Algorithm

Global Optimism based on Local Fitting (GOLF)

for $k = 1, ..., K$

1. **optimistic planning**

   $\pi^k = \pi_{f^k}$, where $f^k = \mathrm{argmax}_{f \in \mathcal{B}} f_1(s_1, \pi_f(s_1))$.

2. **data collection**

   execute π^k to collect a trajectory $(s_1, a_1, \ldots, s_H, a_H)$.

3. **update confidence set $\mathcal{B}$.**

output π^{out} sampled uniformly from $\{\pi^k\}_{k=1}^K$.

Key idea: global optimism + local confidence set

GOLF Algorithm II

Confidence set $\mathcal{B} = \bigcap_h \mathcal{B}_h$:

$$\underbrace{\mathcal{B}_h}_{\text{local confidence set}} = \left\{ f \in \mathcal{F} : \quad \underbrace{\mathcal{L}_{\mathcal{D}_h}(f_h, f_{h+1})}_{\text{proxy to Bellman error}} \leq \underbrace{\inf_{g \in \mathcal{F}_h} \mathcal{L}_{\mathcal{D}_h}(g, f_{h+1})}_{\text{“ERM”}} + \underbrace{\beta}_{\text{relaxation}} \right\}$$

$$\mathcal{L}_{\mathcal{D}_h}(\phi, \psi) = \sum_{(s,a,r,s') \in \mathcal{D}_h} [\phi(s,a) - r - \max_{a' \in \mathcal{A}} \psi(s', a')]^2.$$

Theoretical Guarantees

Theorem [JLM21]

Assume realizability and completeness. **GOLF** outputs an $\mathcal{O}(\epsilon)$-optimal policy in $\widetilde{\mathcal{O}}(H^2 d \log(\mathcal{N}_{\mathcal{F}})/\epsilon^2)$ episodes.

- $d = \min_{\Pi \in \{\mathcal{D}_\Delta, \mathcal{D}_{\mathcal{F}}\}} \dim_{\mathrm{BE}}\left(\mathcal{F}, \Pi, \epsilon/H\right)$, is the BE dimension.
- $\mathcal{N}_{\mathcal{F}}$: $\mathcal{O}(\epsilon)$-covering number of $\mathcal{F}$ in $\|\cdot\|_\infty$.

Theoretical Guarantees

Theorem [JLM21]

Assume realizability and completeness. **GOLF** outputs an $\mathcal{O}(\epsilon)$-optimal policy in $\widetilde{\mathcal{O}}(H^2 d \log(\mathcal{N}_{\mathcal{F}})/\epsilon^2)$ episodes.

- $d = \min_{\Pi \in \{\mathcal{D}_\Delta, \mathcal{D}_{\mathcal{F}}\}} \dim_{\mathrm{BE}}\left(\mathcal{F}, \Pi, \epsilon/H\right)$, is the BE dimension.
- $\mathcal{N}_{\mathcal{F}}$: $\mathcal{O}(\epsilon)$-covering number of $\mathcal{F}$ in $\|\cdot\|_\infty$.

- regret guarantee: $\mathrm{Regret}(K) \le \widetilde{\mathcal{O}}(H\sqrt{dK \log \mathcal{N}_{\mathcal{F}}})$.

Theoretical Guarantees

Theorem [JLM21]

Assume realizability and completeness. **GOLF** outputs an $\mathcal{O}(\epsilon)$-optimal policy in $\widetilde{\mathcal{O}}(H^2 d \log(\mathcal{N}_{\mathcal{F}})/\epsilon^2)$ episodes.

- $d = \min_{\Pi \in \{\mathcal{D}_\Delta, \mathcal{D}_{\mathcal{F}}\}} \dim_{\mathrm{BE}} \left(\mathcal{F}, \Pi, \epsilon/H\right)$, is the BE dimension.
- $\mathcal{N}_{\mathcal{F}}$: $\mathcal{O}(\epsilon)$-covering number of $\mathcal{F}$ in $\|\cdot\|_\infty$.
- regret guarantee: $\mathrm{Regret}(K) \le \widetilde{\mathcal{O}}(H\sqrt{dK \log \mathcal{N}_{\mathcal{F}}})$.

GOLF learns low BE dimension problem sample-efficiently!

Relation to Prior Works

Guarantees for GOLF when restricted to following subclasses:

- **Linear function approximation**: regret $\widetilde{\mathcal{O}}(Hd_{\mathrm{lin}}\sqrt{K})$

  matches [ZLKB20].

Relation to Prior Works

Guarantees for GOLF when restricted to following subclasses:

- **Linear function approximation**: regret $\widetilde{\mathcal{O}}(Hd_{\mathrm{lin}}\sqrt{K})$

  matches [ZLKB20].

- **Low Eluder dimension**: regret $\widetilde{\mathcal{O}}(H\sqrt{d_{\mathrm{E}}K\log\mathcal{N}_{\mathcal{F}}})$

  improves over [WSY20] by $\sqrt{d_{\mathrm{E}}}$.

Relation to Prior Works

Guarantees for GOLF when restricted to following subclasses:

- **Linear function approximation**: regret $\widetilde{\mathcal{O}}(Hd_{\mathrm{lin}}\sqrt{K})$
  matches [ZLKB20].
- **Low Eluder dimension**: regret $\widetilde{\mathcal{O}}(H\sqrt{d_{\mathrm{E}}K\log\mathcal{N}_{\mathcal{F}}})$
  improves over [WSY20] by $\sqrt{d_{\mathrm{E}}}$.
- **Low Bellman rank**: sample complexity $\widetilde{\mathcal{O}}(H^2 d_{\mathrm{br}}\log(\mathcal{N}_{\mathcal{F}})/\epsilon^2)$
  improves over [JKA+17] by d_{br}.
  but requires completeness.

OLIVE Algorithm

A hypothesis elimination-based algorithm proposed in [JKA$^+$17].

OLIVE Algorithm

A hypothesis elimination-based algorithm proposed in [JKA+17].

Theorem [JLM21]

Assume realizability. **OLIVE** finds an ϵ-optimal policy within $\widetilde{\mathcal{O}}(H^3 d^2 \log \mathcal{N}_{\mathcal{F}}/\epsilon^2)$ episodes.

- $d = \dim_{\mathrm{BE}}\left(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon/H\right)$.
- $\mathcal{N}_{\mathcal{F}}$: $\mathcal{O}(\epsilon)$-covering number of $\mathcal{F}$.

OLIVE Algorithm

A hypothesis elimination-based algorithm proposed in [JKA$^+$17].

Theorem [JLM21]

Assume realizability. **OLIVE** finds an ϵ-optimal policy within $\widetilde{\mathcal{O}}(H^3 d^2 \log \mathcal{N}_{\mathcal{F}}/\epsilon^2)$ episodes.

- $d = \dim_{\mathrm{BE}}\left(\mathcal{F}, \mathcal{D}_{\mathcal{F}}, \epsilon/H\right)$.
- $\mathcal{N}_{\mathcal{F}}$: $\mathcal{O}(\epsilon)$-covering number of $\mathcal{F}$.

Comparing to GOLF: worse sample complexity, no $\mathcal{D}_{\Delta}$, no regret guarantees, but does not require completeness.

Conclusion

Summary

New rich class of tractable RL problems—low BE dimension.

it contains a majority of known tractable RL problems.

Summary

New rich class of tractable RL problems—low BE dimension.

it contains a majority of known tractable RL problems.

New sample-efficient alg for low BE dimension problems—GOLF.

simple, clean, and with sharp rate.

Summary

New rich class of tractable RL problems—low BE dimension.

it contains a majority of known tractable RL problems.

New sample-efficient alg for low BE dimension problems—GOLF.

simple, clean, and with sharp rate.

New simpler analysis for OLIVE for general low BE dimension problems.

Thank You!