GEC: A Unified Framework for Interactive Decision Making in MDP, POMDP, and Beyond

Han Zhong

Joint work with
Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang

RL Theory Seminar

Outline

Table of Contents

The agent interacts with the unknown environment and aims to maximize its own reward.

Interactive Decision Making

Can we perform sample-efficient learning for interactive decision making?

Can we perform sample-efficient learning for interactive decision making?

- Exploration-exploitation tradeoff:
  - Naive exploration incurs an exponential sample complexity (Kakade, 2003);
  - Design algorithms with strategic exploration;

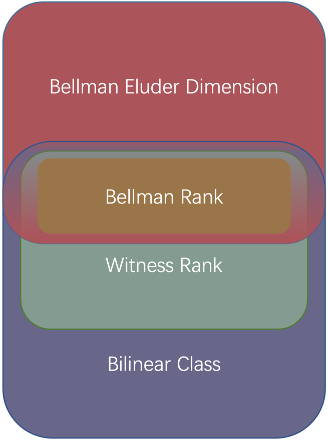
Can we perform sample-efficient learning for interactive decision making?

- Exploration-exploitation tradeoff:
  - Naive exploration incurs an exponential sample complexity (Kakade, 2003);
  - Design algorithms with strategic exploration;
- Large state space:
  - $\Omega(\sqrt{SAH^2T})$ lower bound for tabular RL (Jaksch et al., 2010);
  - Sample-efficient learning for RL with (general) function approximation;

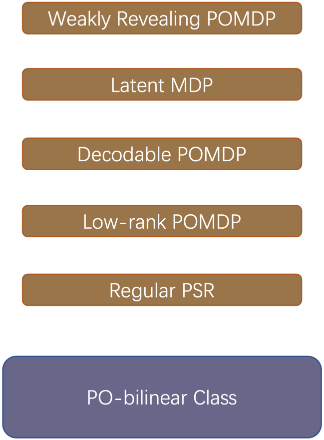
Can we perform sample-efficient learning for interactive decision making?

- Exploration-exploitation tradeoff:
  - Naive exploration incurs an exponential sample complexity (Kakade, 2003);
  - Design algorithms with strategic exploration;
- Large state space:
  - $\Omega(\sqrt{SAH^2T})$ lower bound for tabular RL (Jaksch et al., 2010);
  - Sample-efficient learning for RL with (general) function approximation;
- Partial observations:
  - $\Omega(A^H)$ lower bound for general POMDPs (Krishnamurthy et al., 2016);
  - Identify tractable partially observable RL models and design efficient algorithms.

Previous Works

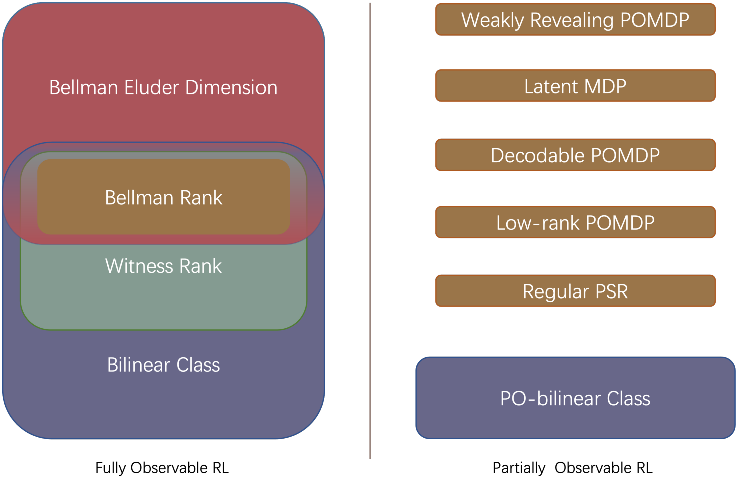


Fully Observable RL



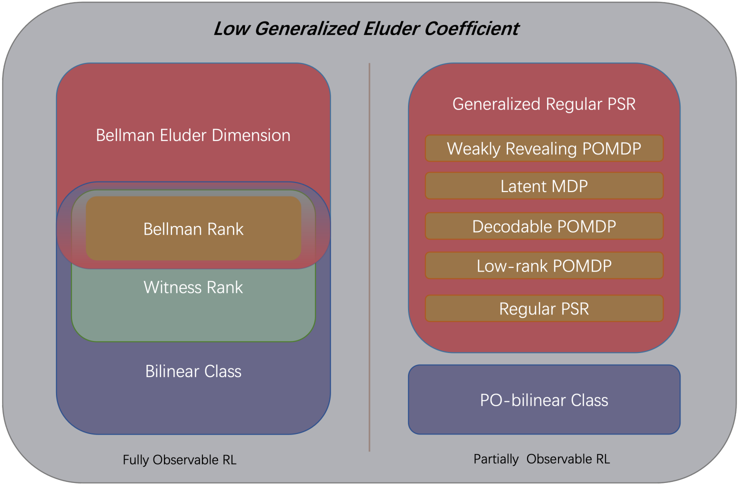
Partially Observable RL

Previous Works



1. Different complexity measures and algorithms;
2. Fully observable RL and partially observable RL are separate.

Our Work



Propose a new complexity measure – Generalized Eluder Coefficient (GEC) – that can capture nearly all known tractable RL problems.

Our Work

Algorithm:

- Generic posterior sampling algorithm;
- Generic UCB-based algorithm;
- Maximize to explore (MEX) algorithm;

Proposed algorithms can be implemented in both model-free and model-based fashion, under both fully observable and partially observable settings.

Our Work

Algorithm:

- Generic posterior sampling algorithm;
- Generic UCB-based algorithm;
- Maximize to explore (MEX) algorithm;

> Proposed algorithms can be implemented in both model-free and model-based fashion, under both fully observable and partially observable settings.

Theory:

- The above three algorithms enjoy the regret of

$$\tilde{\mathcal{O}}(\mathrm{poly}(d_{\mathrm{GEC}}, H) \cdot T^{1/2}) \text{ or } \tilde{\mathcal{O}}(\mathrm{poly}(d_{\mathrm{GEC}}, H) \cdot T^{2/3});$$

- These three algorithms can learn low GEC problems sample-efficiently;
- Match existing regret bounds for Bellman eluder dimension (Jin et al., 2021) and bilinear class (Du et al., 2021).

> A new and unified understanding of both fully observable and partially observable RL.

Table of Contents

Episodic Interactive Decision Making $(\mathcal{O}, \mathcal{A}, H, \mathbb{P}, R)$

- $\mathcal{O}$: observation space;
- $\mathcal{A}$: action space;
- H: length of each episode;
- $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$: $\mathbb{P}_h(o_{h+1} \mid \tau_h)$ denotes the probability of generating the observation o_{h+1} given the history $\tau_h = (o_{1:h}, a_{1:h})$;
- $R = \{R_h : \mathcal{O} \times \mathcal{A} \mapsto \mathbb{R}^+\}_{h\in[H]}$: reward functions;
- Initial observation is sampled from a fixed distribution;
- Assumption: $\sum_{h=1}^{H} R_h \leq 1$.

Policy, Value Function, and Learning Objective

- Policy $\pi = \{\pi_h\}_{h \in [H]}$: $\pi_h : (\mathcal{O} \times \mathcal{A})^{h-1} \times \mathcal{O} \to \Delta_{\mathcal{A}}$ is a mapping from an observation-action sequence to a distribution over actions.
- Visitation probability $\mathbb{P}^\pi(\tau_h) = \mathbb{P}(\tau_h) \times \pi(\tau_h)$, where $\mathbb{P}(\tau_h)$ and $\pi(\tau_h)$ are defined by

$$\mathbb{P}(\tau_h) = \prod_{h'=1}^{h} \mathbb{P}_h(o_{h'} \mid \tau_{h'-1}), \ \pi(\tau_h) = \prod_{h'=1}^{h} \pi_{h'}(a_{h'} \mid \tau_{h'-1}, o_{h'}).$$

- Value function:

$$V^\pi := \mathbb{E}_\pi\Big[\sum_{h=1}^{H} r_h\Big].$$

- Optimal policy: $\pi^* = \mathrm{argmax}_\pi V^\pi$, optimal value: $V^* = V^{\pi^*}$.
- Learning objective: An online algorithm predicts $\{\pi^t\}_{t=1}^T$, its *regret* is defined as

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} V^* - V^{\pi^t}.$$

Example 1: MDP

Episodic Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R)$

- $\mathcal{O} = \mathcal{S}$ and $\mathbb{P}_h(x_{h+1} \mid x_{1:h}, a_{1:h}) = \mathbb{P}_h(x_{h+1} \mid x_h, a_h)$;
- Markov policy: $\pi = \{\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}\}$;
- V-function and Q-function

$$V_h^\pi(x) := \mathbb{E}_\pi\Big[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}) \Big| x_h = x\Big],$$

$$Q_h^\pi(x, a) := \mathbb{E}_\pi\Big[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}) \Big| x_h = x, a_h = a\Big].$$

- Optimal policy π^*, optimal Q-function Q^*;
- Bellman optimality equation:

$$Q_h^*(x, a) = (\mathcal{T}_h Q_{h+1}^*)(x, a) := r_h(x, a) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} \max_{a' \in \mathcal{A}} Q_{h+1}^*(x', a');$$

- Bellman residual:

$$\mathcal{E}_h(Q, x, a) = Q_h(x, a) - (\mathcal{T}_h Q_{h+1})(x, a).$$

Example 2: POMDP

Episodic partially observable Markov decision process (POMDP)

$$(\mathcal{S}, \mathcal{O}, \mathcal{A}, H, \mathbb{P}, \mathbb{O} = \{\mathbb{O}_h\}_{h\in[H]}, R),$$

- $\mathbb{P}_h(x_{h+1} \mid x_{1:h}, a_{1:h}) = \mathbb{P}_h(x_{h+1} \mid x_h, a_h)$,
- $\mathbb{O}_h(o \mid x)$ is the probability of observing o at state x and step h;

Learning POMDPs:

- Negative Results:
  - exponential lower bound in the worst-case (Krishnamurthy et al., 2016);
- Positive results:
  - Weakly revealing POMDPs (Jin et al., 2020): $O \geq S$ and $\min_{h\in[H]} \sigma_{\min}(\mathbb{O}_h) \geq \alpha$;
  - Decodable POMDPs (Du et al., 2019; Efroni et al., 2022): $\exists$ unknown encoder $\phi_h^* : \mathcal{O} \mapsto \mathcal{S}$ such that $\phi_h^*(o_h) = x_h$;
  - latent MDP with sufficient test (Kwon et al., 2021), low-rank POMDP (Wang et al., 2022), and regular PSR (Zhan et al., 2022).

Function Approximation

- General function approximation: hypothesis class $\mathcal{H} = \mathcal{H}_1 \times \cdots \mathcal{H}_H$;
- Model-based hypothesis: $f = (\mathbb{P}_f, r_f) \in \mathcal{H}$,
  - $\pi_{h,f}$: optimal policy corresponding to the model f;
  - $V_{h,f}/Q_{h,f}$: optimal value/Q function corresponding to the model f;
  - f^*: true model; $V_{h,f^*} = V_h$, $Q_{h,f^*} = Q_h$;
- Value-based hypothesis (for MDP): $f = \{Q_{h,f}\}_{h \in [H]} \in \mathcal{H}$;
  - $\pi_{h,f}(\cdot) = \mathrm{argmax}_{a \in \mathcal{A}} Q_{h,f}(\cdot, a)$;
  - $V_{h,f}(\cdot) = \max_{a \in \mathcal{A}} Q_{h,f}(\cdot, a)$;
  - $f^* = Q^*$;
- Realizability assumption: $f^* \in \mathcal{H}$.

Table of Contents

Motivation

By the value decomposition lemma (Jiang et al., 2017), we have

$$\underbrace{\sum_{t=1}^{T} V^{*} - V^{\pi_{f^t}}}_{\mathrm{Reg}(T)} = \sum_{t=1}^{T}\sum_{h=1}^{H} \underbrace{\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\text{Bellman residual}} + \underbrace{\sum_{t=1}^{T}\left(V^{*} - V_{f^t}\right)}_{\text{bias}}$$

Motivation

By the value decomposition lemma (Jiang et al., 2017), we have

$$\underbrace{\sum_{t=1}^{T} V^* - V^{\pi_{f^t}}}_{\mathrm{Reg}(T)} = \sum_{t=1}^{T}\sum_{h=1}^{H} \underbrace{\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\text{Bellman residual}} + \underbrace{\sum_{t=1}^{T}\left(V^* - V_{f^t}\right)}_{\text{bias}}$$

$$\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] \quad (\text{if } V^* \leq V_{f^t})$$

Motivation

By the value decomposition lemma (Jiang et al., 2017), we have

$$\underbrace{\sum_{t=1}^{T} V^* - V^{\pi_{f^t}}}_{\mathrm{Reg}(T)} = \sum_{t=1}^{T}\sum_{h=1}^{H} \underbrace{\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\text{Bellman residual}} + \underbrace{\sum_{t=1}^{T}\left(V^* - V_{f^t}\right)}_{\text{bias}}$$

$$\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] \qquad (\text{if } V^* \leq V_{f^t})$$

- UCB-based algorithm: $f^t = \mathrm{argmax}_{f \in \text{confidence set}} V_f$ to ensure optimism;

Motivation

By the value decomposition lemma (Jiang et al., 2017), we have

$$\underbrace{\sum_{t=1}^{T} V^* - V^{\pi_{f^t}}}_{\mathrm{Reg}(T)} = \sum_{t=1}^{T}\sum_{h=1}^{H} \underbrace{\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\text{Bellman residual}} + \underbrace{\sum_{t=1}^{T}\left(V^* - V_{f^t}\right)}_{\text{bias}}$$

$$\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] \qquad (\text{if } V^* \leq V_{f^t})$$

- UCB-based algorithm: $f^t = \mathrm{argmax}_{f \in \text{confidence set}} V_f$ to ensure optimism;
- "Mismatch" between Goal and Guarantee:
  - Goal: f^t performs well on the unseen data τ^t;

$$\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] \text{ is small?}$$

  - Guarantee: f^t is good on the historical dataset $\{\tau^1, \tau^2, \cdots, \tau^{t-1}\}$;

$$\sum_{h=1}^{H}\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}}\left[\mathcal{E}_h(f^t, x_h^s, a_h^s)^2\right] \text{ is small.}$$

Challenge

- Connect the Goal and Guarantee $\approx$ **"generalization"** from the past to the future:
  - Goal: f^t performs well on the unseen data τ^t;

$$\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}} \left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] \text{ is small?}$$

  - Guarantee: f^t is good on the historical dataset $\{\tau^1, \tau^2, \cdots, \tau^{t-1}\}$;

$$\sum_{h=1}^{H} \sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \left[\mathcal{E}_h(f^t, x_h^s, a_h^s)^2\right] \text{ is small.}$$

Challenge

- Connect the Goal and Guarantee $\approx$ **"generalization"** from the past to the future:
  - Goal: f^t performs well on the unseen data τ^t;

$$\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}} \left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] \text{ is small?}$$

  - Guarantee: f^t is good on the historical dataset $\{\tau^1, \tau^2, \cdots, \tau^{t-1}\}$;

$$\sum_{h=1}^{H} \sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \left[\mathcal{E}_h(f^t, x_h^s, a_h^s)^2\right] \text{ is small.}$$

- In supervised learning, $\{z_s\}_{s=1}^{t-1}$ and an unseen z^t are i.i.d. sampled from a fixed distribution $\mathcal{D}_{\text{data}}$;
  - Relizability + low hypothesis complexity (e.g., covering number) ensure generalization;

Challenge

- Connect the Goal and Guarantee $\approx$ **"generalization"** from the past to the future:
  - Goal: f^t performs well on the unseen data τ^t;

$$\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}} \left[\mathcal{E}_h \left(f^t, x_h^t, a_h^t\right)\right] \text{ is small?}$$

  - Guarantee: f^t is good on the historical dataset $\{\tau^1, \tau^2, \cdots, \tau^{t-1}\}$;

$$\sum_{h=1}^{H} \sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \left[\mathcal{E}_h(f^t, x_h^s, a_h^s)^2\right] \text{ is small.}$$

- In supervised learning, $\{z_s\}_{s=1}^{t-1}$ and an unseen z^t are i.i.d. sampled from a fixed distribution $\mathcal{D}_{\text{data}}$;
  - Relizability + low hypothesis complexity (e.g., covering number) ensure generalization;
- In RL, $\tau^1 \sim \pi_{f^1}$, $\tau^2 \sim \pi_{f^2}, \cdots, \tau^t \sim \pi_{f^t}$, distribution shift exists all the time!

Require an additional structure assumption permits "generalization" from the past to the future (in an online manner).

Simplified Generalized Eluder Coefficient

- Generalized Eluder Coefficient (GEC) is the smallest d_{GEC} such that

$$\underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\text{Goal: prediction error}} \lesssim \left[d_{\mathrm{GEC}}\sum_{h=1}^{H}\sum_{t=1}^{T}\underbrace{\sum_{s=1}^{t-1}\mathbb{E}_{\pi_{f^s}}\left[\mathcal{E}_h(f^t, x_h^s, a_h^s)^2\right]}_{\text{Guarantee: training error}}\right]^{1/2}.$$

Simplified Generalized Eluder Coefficient

- Generalized Eluder Coefficient (GEC) is the smallest d_{GEC} such that

$$\underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\text{Goal: prediction error}} \lesssim \left[d_{\mathrm{GEC}}\sum_{h=1}^{H}\sum_{t=1}^{T}\underbrace{\sum_{s=1}^{t-1}\mathbb{E}_{\pi_{f^s}}\left[\mathcal{E}_h(f^t, x_h^s, a_h^s)^2\right]}_{\text{Guarantee: training error}}\right]^{1/2}.$$

- On average, if $f^t \in \mathcal{H}$ is consistent with the historical data, then the prediction error on unseen t-th trajectory would also be small (but is amplified by GEC);

Simplified Generalized Eluder Coefficient

- Generalized Eluder Coefficient (GEC) is the smallest d_{GEC} such that

$$\underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\text{Goal: prediction error}} \lesssim \left[d_{\mathrm{GEC}}\sum_{h=1}^{H}\sum_{t=1}^{T}\underbrace{\sum_{s=1}^{t-1}\mathbb{E}_{\pi_{f^s}}\left[\mathcal{E}_h(f^t, x_h^s, a_h^s)^2\right]}_{\text{Guarantee: training error}}\right]^{1/2}.$$

- On average, if $f^t \in \mathcal{H}$ is consistent with the historical data, then the prediction error on unseen t-th trajectory would also be small (but is amplified by GEC);
- Optimism $(V^* \leq V_{f^t})$ + low GEC + small training error $\approx$ low-regret learning:

$$\begin{aligned}\mathrm{Reg}(T) &\leq \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] \\ &\lesssim \left[d_{\mathrm{GEC}}\sum_{h=1}^{H}\sum_{t=1}^{T}\underbrace{\sum_{s=1}^{t-1}\mathbb{E}_{\pi_{f^s}}\left[\mathcal{E}_h(f^t, x_h^s, a_h^s)^2\right]}_{\text{training error } \leq \beta}\right]^{1/2} \leq \sqrt{d_{\mathrm{GEC}} H T \beta}.\end{aligned}$$

- For LinUCB (Chu et al., 2011), UCRL2 (Jaksch et al., 2010), UCRL-VTR (Ayoub et al., 2020), GOLF (Jin et al., 2021)..., β only has a logarithmic dependency in T.

Generalized Eluder Coefficient

$$\sum_{t=1}^{T} V_{f^t} - V^{\pi_{f^t}} = \sum_{t=1}^{T} \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}} \left[\mathcal{E}_h \left(f^t, x_h, a_h \right) \right]}_{\text{Goal: prediction error}} \lesssim \left[d_{\mathrm{GEC}} \sum_{h=1}^{H} \sum_{t=1}^{T} \underbrace{\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \left[\mathcal{E}_h (f^t, x_h, a_h)^2 \right]}_{\text{Guarantee: training error}} \right]^{1/2}.$$

Generalized Eluder Coefficient

$$\sum_{t=1}^{T} V_{f^t} - V^{\pi_{f^t}} = \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h, a_h\right)\right]}_{\text{Goal: prediction error}} \lesssim \left[d_{\mathrm{GEC}} \sum_{h=1}^{H}\sum_{t=1}^{T} \underbrace{\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}}\left[\mathcal{E}_h(f^t, x_h, a_h)^2\right]}_{\text{Guarantee: training error}}\right]^{1/2}.$$

Definition (Generalized Eluder Coefficient)

Given a hypothesis class $\mathcal{H}$, a discrepancy function $\ell = \{\ell_f\}_{f \in \mathcal{H}}$, an exploration policy class $\Pi_{\exp}$, the generalized eluder coefficient $\mathrm{GEC}(\mathcal{H}, \ell, \Pi_{\exp}, \epsilon)$ is the smallest d $(d \geq 0)$ such that for any sequence of hypotheses and exploration policies $\{f^t, \{\pi_{\exp}(f^t, h)\}_{h \in [H]}\}_{t \in [T]}$:

$$\sum_{t=1}^{T} \underbrace{V_{f^t} - V^{\pi_{f^t}}}_{\text{prediction error}} \leq \left[d \sum_{h=1}^{H}\sum_{t=1}^{T} \underbrace{\left(\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{\exp}(f^s,h)} \ell_{f^s}(f^t, \xi_h)\right)}_{\text{training error}}\right]^{1/2} + \underbrace{2\sqrt{dHT} + \epsilon HT}_{\text{burn-in cost}}.$$

- Flexible choices of discrepancy functions and exploration policies.
- The GEC captures the hardness of exploration-exploitation trade-off by comparing the *out-of-sample* prediction error with the *in-sample* training error;

Generalized Eluder Coefficient: MDP Examples

- Q-type problems :

$$\sum_{t=1}^{T} V_{f^t} - V^{\pi_{f^t}} \leq \left[d_Q \sum_{h=1}^{H} \sum_{t=1}^{T} \left(\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f^t, x_h, a_h)^2 \right) \right]^{1/2}.$$

- V-type problems:

$$\sum_{t=1}^{T} V_{f^t} - V^{\pi_{f^t}} \leq \left[d_V \sum_{h=1}^{H} \sum_{t=1}^{T} \left(\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s} \circ_h \mathrm{Unif}(\mathcal{A})} \mathcal{E}_h(f^t, x_h, a_h)^2 \right) \right]^{1/2},$$

  where $\pi_{f^s} \circ_h \mathrm{Unif}(\mathcal{A})$ means executing π_{f^s} for the first $h-1$ steps and then take a random $a_h \in \mathcal{A}$.

- Model-based problems:

$$\sum_{t=1}^{T} V_{f^t} - V^{\pi_{f^t}} \leq \left[d \sum_{h=1}^{H} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \mathbb{E}_{\tilde{\pi}} D_H^2 \left(\mathbb{P}_{h,f^t}(\cdot \mid x_h, a_h), \mathbb{P}_{h,f^*}(\cdot \mid x_h, a_h) \right) \right]^{1/2},$$

  where $\tilde{\pi}$ is either π_{f^s} (Q-type) or $\pi_{f^s} \circ_h \mathrm{Unif}(\mathcal{A})$ (V-type) and $D_H^2(P, Q) = \frac{1}{2} \cdot \mathbb{E}_{x \in P}[(\sqrt{dQ(x)/dP(x)} - 1)^2]$ is the Hellinger divergence.

Relationship with Existing Complexity Measures

- Bellman eluder dimension:

$$\mathrm{GEC} \le \tilde{\mathcal{O}}\left(Hd_Q\right) \quad \text{Q-type}, \qquad \mathrm{GEC} \le \tilde{\mathcal{O}}(AHd_V) \quad \text{V-type};$$

Relationship with Existing Complexity Measures

- Bellman eluder dimension:

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}\left(H d_Q\right) \quad \text{Q-type}, \qquad \mathrm{GEC} \leq \tilde{\mathcal{O}}(A H d_V) \quad \text{V-type};$$

- Bilinear class:

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}(H d_{\mathrm{bil}}) \quad \text{Q-type}, \qquad \mathrm{GEC} \leq \tilde{\mathcal{O}}(A H d_{\mathrm{bil}}) \quad \text{V-type};$$

Relationship with Existing Complexity Measures

- Bellman eluder dimension:

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}\left(H d_Q\right) \quad \text{Q-type}, \qquad \mathrm{GEC} \leq \tilde{\mathcal{O}}(A H d_V) \quad \text{V-type};$$

- Bilinear class:

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}(H d_{\mathrm{bil}}) \quad \text{Q-type}, \qquad \mathrm{GEC} \leq \tilde{\mathcal{O}}(A H d_{\mathrm{bil}}) \quad \text{V-type};$$

- Witness rank:

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}(H d_Q / \kappa_{\mathrm{wit}}^2), \quad \text{Q-type}, \qquad \mathrm{GEC} \leq \tilde{\mathcal{O}}(A H d_V / \kappa_{\mathrm{wit}}^2), \quad \text{V-type}.$$

Relationship with Existing Complexity Measures

- GEC (model-based POMDP version):

$$\sum_{t=1}^{T} V_{f^t} - V^{\pi^t} \leq \left[d_{\mathrm{GEC}} \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s=1}^{t-1} D_H^2 \left(\mathbb{P}_{f^t}^{\pi_{\exp}(f^s,h)}, \mathbb{P}_{f^*}^{\pi_{\exp}(f^s,h)} \right) \right]^{1/2},$$

where $\pi_{\exp}(f^s, h) := \pi_{f^s} \circ_h \mathrm{Unif}(\mathcal{A}) \cdots \circ_H \mathrm{Unif}(\mathcal{A})$.

  - α-revealing POMDPs:

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}\big(\mathrm{poly}(S, A, H, 1/\alpha)\big),$$

  - Decodable POMDPs:

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}\big(\mathrm{poly}(S, A, H)\big),$$

[1] Independent works Liu et al. (2022); Chen et al. (2022) identify similar PSR classes with regular conditions on observable operators (Jaeger, 2000).

Relationship with Existing Complexity Measures

- GEC (model-based POMDP version):

$$\sum_{t=1}^{T} V_{f^t} - V^{\pi^t} \leq \left[d_{\mathrm{GEC}} \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s=1}^{t-1} D_H^2 \Big(\mathbb{P}_{f^t}^{\pi_{\exp}(f^s, h)}, \mathbb{P}_{f^*}^{\pi_{\exp}(f^s, h)} \Big) \right]^{1/2},$$

  where $\pi_{\exp}(f^s, h) := \pi_{f^s} \circ_h \mathrm{Unif}(\mathcal{A}) \cdots \circ_H \mathrm{Unif}(\mathcal{A})$.

  - α-revealing POMDPs:

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}\big(\mathrm{poly}(S, A, H, 1/\alpha)\big),$$

  - Decodable POMDPs:

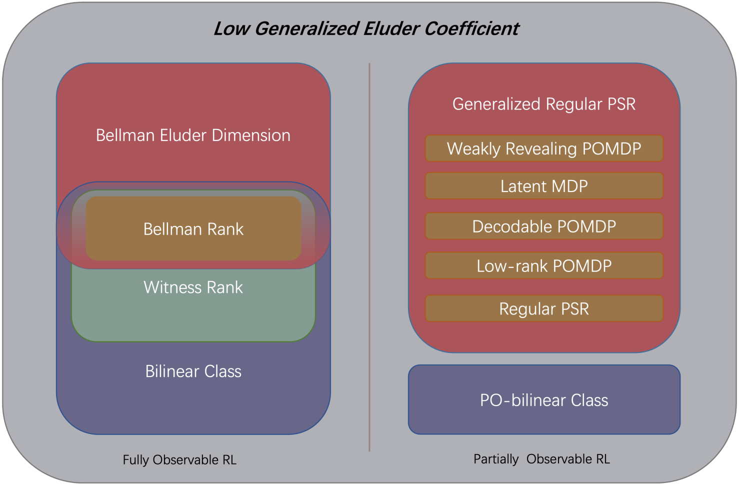
$$\mathrm{GEC} \leq \tilde{\mathcal{O}}\big(\mathrm{poly}(S, A, H)\big),$$

- α-generalized regular PSR (new)[1]:
  - Impose some regular condition on the observable operator representation (Jaeger, 2000) of PSR.
  - Nearly all known tractable POMDPs satisfy this regular condition;
  - With proper exploration policies:

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}\big(\mathrm{poly}(\text{complexity of PSR}, H, A, 1/\alpha)\big)$$

[1] Independent works Liu et al. (2022); Chen et al. (2022) identify similar PSR classes with regular conditions on observable operators (Jaeger, 2000).

Summary of Relationships



GEC captures nearly all known tractable RL problems.

Table of Contents

Algorithmic Design to Use GEC

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} V^* - V^{\pi_{f^t}} = \sum_{t=1}^{T} \underbrace{V_{f^t} - V^{\pi_{f^t}}}_{\text{prediction error}} + \sum_{t=1}^{T} \underbrace{V^* - V_{f^t}}_{\text{bias}}$$

$$\lesssim \left[d_{\mathrm{GEC}} \sum_{h=1}^{H} \sum_{t=1}^{T} \underbrace{\left(\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{\exp}(f^s, h)} \ell_{f^s}(f^t, \xi_h) \right)}_{\text{training error}} \right]^{1/2} + \sum_{t=1}^{T} \underbrace{(V^* - V_{f^t})}_{\text{bias}}.$$

Algorithmic Design to Use GEC

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} V^* - V^{\pi_{f^t}} = \sum_{t=1}^{T} \underbrace{V_{f^t} - V^{\pi_{f^t}}}_{\text{prediction error}} + \sum_{t=1}^{T} \underbrace{V^* - V_{f^t}}_{\text{bias}}$$

$$\lesssim \left[d_{\mathrm{GEC}} \sum_{h=1}^{H} \sum_{t=1}^{T} \underbrace{\left(\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{\exp}(f^s, h)} \ell_{f^s}(f^t, \xi_h) \right)}_{\text{training error}} \right]^{1/2} + \sum_{t=1}^{T} \underbrace{(V^* - V_{f^t})}_{\text{bias}}.$$

- How to control the training error?
  - The training error term is not available to the executed algorithm, e.g., the Bellman operator, or the true transition kernel $\mathbb{P}_{f^*}$;
  - We need to approximate the training error by some loss functions and design effective estimation to achieve a low training error.

Algorithmic Design to Use GEC

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} V^* - V^{\pi_{f^t}} = \sum_{t=1}^{T} \underbrace{V_{f^t} - V^{\pi_{f^t}}}_{\text{prediction error}} + \sum_{t=1}^{T} \underbrace{V^* - V_{f^t}}_{\text{bias}}$$

$$\lesssim \Bigg[d_{\mathrm{GEC}} \sum_{h=1}^{H} \sum_{t=1}^{T} \underbrace{\bigg(\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{\exp}(f^s,h)} \ell_{f^s}(f^t, \xi_h) \bigg)}_{\text{training error}} \Bigg]^{1/2} + \sum_{t=1}^{T} \underbrace{(V^* - V_{f^t})}_{\text{bias}}.$$

- How to control the training error?
  - The training error term is not available to the executed algorithm, e.g., the Bellman operator, or the true transition kernel $\mathbb{P}_{f^*}$;
  - We need to approximate the training error by some loss functions and design effective estimation to achieve a low training error.
- How to control the bias term?
  - UCB-based algorithms directly have $V^* - V_{f^t} \le 0$
  - For other algorithms such as posterior sampling, $V^* - V_{f^t} \le 0$ is not directly available.

A Generic Posterior Sampling Framework

Posterior sampling algorithm

- **Optimistic prior (Zhang, 2022):** Choose the prior that favors the hypotheses with higher values

$$p^0(f) \cdot \exp(\gamma V_f), \qquad \gamma > 0.$$

- **Loss function:** Let

$$L_h^{t-1}(f) = \mathcal{L}_h(f, \{f^s\}_{s\in[t-1]}, \{\mathcal{D}_h^s\}_{s\in[t-1]})$$

be a proxy of the unknown training error $\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{\exp}(f^s,h)} \ell_{f^s}(f, \xi_h)$.

- **Posterior:**

$$p^t(f) \propto p^0(f) \cdot \exp\left(\gamma V_f + \sum_{h=1}^{H} L_h^{t-1}(f)\right), \quad f^t \sim p^t(\cdot).$$

- **Data collection:** For any $h \in [H]$, execute $\pi_{\exp}(f^t, h)$ for N_{batch} times and collect samples $\mathcal{D}_h^t$.

Choices of Loss Functions (Model-free case)

Double sampling issue of model-free MDP (Antos et al., 2008):

$$\mathbb{E}_{\pi^s}\underbrace{[Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)]}_{\text{TD error}}^2 = \underbrace{\mathbb{E}_{\pi^s}[\mathcal{E}_h(f, x_h^s, a_h^s)^2]}_{\text{Goal: training error}} + \underbrace{\sigma_{h,f}^2}_{\text{Sampling variance}}$$

1 Minimax formulation (GOLF (Jin et al., 2021), Conditional PS (Dann et al., 2021))[2]

$$L_h^t(f) = -\eta \sum_{s=1}^{t} [Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)]^2$$
$$- \log \mathbb{E}_{\tilde{f}_h \sim p_h^0(\cdot)} \left[\exp\left(-\eta \sum_{s=1}^{t} [Q_{h,\tilde{f}}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)]^2 \right) \right],$$

- The introduced log term cancels the variance;
- The log term requires completeness to deal with;

2 Trajectory average with N_{batch} i.i.d. data (OLIVE (Jiang et al., 2017), BiLin-UCB (Du et al., 2021))

$$L_h^t(f) = -\eta \sum_{s=1}^{t} \left(\frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \left(Q_{h,f}(x_{i,h}^s, a_{i,h}^s) - r_{i,h}^s - V_{h+1,f}(x_{i,h+1}^s)\right) \right)^2;$$

- Sample mean admits a smaller variance: $\mathrm{Var}[\bar{X}_m] = \frac{1}{m}\mathrm{Var}[X]$.

[2] Also used in some works on offline RL (Antos et al., 2008; Chen and Jiang, 2019).

Choices of Loss Function (Model-based case)

- For MDPs, we choose

$$L_h^t(f) = \eta \sum_{s=1}^{t} \log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s),$$

where $\mathcal{D}_h^s = (x_h^s, a_h^s, x_{h+1}^s)$ is the tuple induced by $\pi_{\exp}(f^s, h)$.

- For POMDPs and PSRs, we choose

$$L_h^t(f) = \eta \sum_{s=1}^{t} \log \mathbb{P}_f(\tau_h^s),$$

where $\mathcal{D}_h^s = \tau_h^s$ is the trajectory induced by $\pi_{\exp}(f^s, h)$.

UCB Algorithm

UCB Algorithm

- Given the past $t-1$ iterations, we maintain a confidence set $\mathcal{H}_t \subset \mathcal{H}$ such that $f^* \in \mathcal{H}_t$ with high probability;
- Choose the most optimistic hypothesis f^t:

$$f^t = \operatorname*{argmax}_{f \in \mathcal{H}_t} V_f$$

- Execute exploration policies $\{\pi_{\exp}(f^t, h)\}_{h \in [H]}$ to collect data

UCB Algorithm

UCB Algorithm

- Given the past $t-1$ iterations, we maintain a confidence set $\mathcal{H}_t \subset \mathcal{H}$ such that $f^* \in \mathcal{H}_t$ with high probability;
- Choose the most optimistic hypothesis f^t:

$$f^t = \operatorname*{argmax}_{f \in \mathcal{H}_t} V_f$$

- Execute exploration policies $\{\pi_{\exp}(f^t, h)\}_{h \in [H]}$ to collect data

- Extend previous UCB algorithms (LinUCB, UCRL2, UCRL-VTR, GOLF, BiLinUCB, OMLE, ...) to a more general class (problems with low GEC);
- Theoretical analysis is relatively simple and well-understood;
- Hard to implement: need to solve a constrained optimization problem

Maximize to Explore

Maximize to Explore

- Given the past $t-1$ iterations, we choose a proper loss $L_h^{t-1}(f)$;
- Choose the hypothesis f^t:

$$f^t = \operatorname*{argmax}_f \left\{ V_f - \eta \cdot \sum_{h=1}^{H} L_h^{t-1}(f) \right\}.$$

An optimistic modification of loss minimization problem.

- Execute exploration policies $\{\pi_{\exp}(f^t, h)\}_{h \in [H]}$ to collect data

Easy to implement: only need to optimize an unconstrained objective.

Summary of Algorithm Design

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} V^* - V^{\pi_{f^t}} = \sum_{t=1}^{T} V_{f^t} - V^{\pi_{f^t}} + \sum_{t=1}^{T} V^* - V_{f^t}$$

$$\lesssim \Bigg[d_{\mathrm{GEC}} \sum_{h=1}^{H} \sum_{t=1}^{T} \underbrace{\Bigg(\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{\exp}(f^s,h)} \ell_{f^s}(f^t, \xi_h) \Bigg) \Bigg]^{1/2}}_{\text{training error}} + \sum_{t=1}^{T} \underbrace{(V^* - V_{f^t})}_{\text{bias}}.$$

- How to control the training error?
  - Choose proper loss functions to approximate the training error.
  - Choose proper exploration policies to collect data.
- How to control the bias term?
  - Optimistic posterior sampling
  - UCB-based algorithm
  - Maximize to explore (MEX)

Theory

Theorem ((Zhong et al., 2022; Liu et al., 2023))

The above three algorithms enjoy the following regret bounds:

1. *Value-based approach for MDPs*
   - *Minimax formulation with Realizability + Completeness:* $\tilde{\mathcal{O}}\big(\sqrt{d_{\mathrm{GEC}} \cdot HT \cdot \log|\mathcal{H}|}\big)$;
   - *Trajectory average with Realizability:* $\tilde{\mathcal{O}}\big((d_{\mathrm{GEC}}^2 H \log|\mathcal{H}|)^{1/3} \cdot T^{2/3}\big)$[a];
2. *Model-based approach for MDP, POMDP, and PSR:*
   - *Realizability:* $\tilde{\mathcal{O}}\big(\sqrt{d_{\mathrm{GEC}} \cdot HT \cdot \log|\mathcal{H}|}\big)$.

[a] Also holds for PO-bilinear class.

- Interactive decision making with low GEC is learnable.
- Matches existing bound for Bellman eluder dimension (Jin et al., 2021) and Bilinear class (Du et al., 2021).

Optimistic modification + Low GEC + Effective training error estimation
$\approx$ Sample-efficient learning.

Table of Contents

Comparison with Decision Estimation Coefficient (Foster et al., 2021)

Similarities:

- Universality: subsume most of the known tractable RL problems;
- Reduction-based idea: convert regret minimization to new target;

Differences:

- Different reduction ideas: in-sample estimation v.s. online learning;
- Different policy selection strategies: simple strategy v.s. minimax subroutine;
- Algorithm design:
  - GEC: flexible in algorithmic design: i) Posterior sampling, ii) UCB-based algorithm, and iii) Maximize to explore;
  - DEC: restrictive algorithm design: Estimation to decision-making (E2D);
- Regret upper bound:
  - GEC: match the best-known results;
  - DEC: suboptimal $T^{3/4}$ regret bound (Foster et al., 2022) for bilinear class;
- Lower bound: DEC also characterizes the lower bound of the RL problems.

Conclusion

- New complexity measure – GEC – that can capture nearly all known tractable interactive decision making problems.

  reduce the out-of-sample prediction error to the in-sample training error.

- Three efficient algorithms for interactive decision making with low GEC.

  optimistic modification + an effective sequential estimation of training error.

A new and unified understanding for both fully observable and partially observable RL.

Thank you!

Agarwal, A. and Zhang, T. (2022a). Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659*.

Agarwal, A. and Zhang, T. (2022b). Non-linear reinforcement learning in large action spaces: Structural conditions and sample-efficiency of posterior sampling. *arXiv preprint arXiv:2203.08248*.

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR.

Chen, F., Bai, Y., and Mei, S. (2022). Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms. *arXiv preprint arXiv:2209.14990*.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings.

Dann, C., Mohri, M., Zhang, T., and Zimmert, J. (2021). A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12040–12051.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. (2019). Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR.

Efroni, Y., Jin, C., Krishnamurthy, A., and Miryoosefi, S. (2022). Provable reinforcement learning with a short-term memory. In *International Conference on Machine Learning*, pages 5832–5850. PMLR.

Foster, D. J., Golowich, N., Qian, J., Rakhlin, A., and Sekhari, A. (2022). A note on model-free reinforcement learning with the decision-estimation coefficient. *arXiv preprint arXiv:2211.14250*.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*.

Jaeger, H. (2000). Observable operator models for discrete stochastic time series. *Neural computation*, 12(6):1371–1398.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR.

Jin, C., Kakade, S., Krishnamurthy, A., and Liu, Q. (2020). Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539.

Jin, C., Liu, Q., and Miryoosefi, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34.

Kakade, S. M. (2003). *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom).

Krishnamurthy, A., Agarwal, A., and Langford, J. (2016). Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. (2021). Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34:24523–24534.

Liu, Q., Netrapalli, P., Szepesvari, C., and Jin, C. (2022). Optimistic mle–a generic model-based algorithm for partially observable sequential decision making. *arXiv preprint arXiv:2209.14997*.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z., and Wang, Z. (2023). One objective to rule them all: A maximization objective fusing estimation and planning for exploration. *arXiv preprint arXiv:2305.18258*.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. (2019). Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2022). Embed to control partially observed systems: Representation learning with provable sample efficiency. *arXiv preprint arXiv:2205.13476*.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. (2022). Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*.

Zhang, T. (2022). Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. (2022). Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*.

Backup Slides

Relationship with Eluder Dimension (Jin et al., 2021)

Definition (ϵ-independence between distributions)

Let $\mathcal{G}$ be a function class defined on $\mathcal{X}$, and $\nu, \mu_1, \cdots, \mu_n$ be probability measures over $\mathcal{X}$. We say ν is ϵ-independent of $\{\mu_1, \mu_2, \cdots, \mu_n\}$ with respect to $\mathcal{G}$ if there exists $g \in \mathcal{G}$ such that $\sqrt{\sum_{i=1}^n (\mathbb{E}_{\mu_i}[g])^2} \leq \epsilon$ but $|\mathbb{E}_\nu[g]| > \epsilon$.
The distributional eluder dimension $\dim_{\mathrm{DE}}(\mathcal{G}, \Pi, \epsilon)$ is the length of the longest sequence $\{\rho_1, \cdots, \rho_n\} \subset \Pi$ such that there exists $\epsilon' \geq \epsilon$ with ρ_i being ϵ'-independent of $\{\rho_1, \cdots, \rho_{i-1}\}$ for all $i \in [n]$.

- Let $(I - \mathcal{T}_h)\mathcal{H} := \{(x,a) \to (f_h - \mathcal{T}_h f_{h+1})(x,a) : f \in \mathcal{H}\}$, $(I - \mathcal{T}_h)V_{\mathcal{H}} := \{x \to (f_h - \mathcal{T}_h f_{h+1})(x, \pi_{f_h}(x)) : f \in \mathcal{H}\}$ be the set of Q/V type Bellman residuals induced by $\mathcal{H}$ at step h;
- The Q/V-type ϵ-BE dimension of $\mathcal{H}$ with respect to Π is defined as

$$d_Q/d_V := \max_{h \in [H]} \left\{\dim_{\mathrm{DE}}\big((I - \mathcal{T}_h)\mathcal{H}/\dim_{\mathrm{DE}}(I - \mathcal{T}_h)\mathcal{H}_V, \Pi_h, \epsilon\big)\right\}.$$

- We have $\mathrm{GEC} \leq \tilde{O}\left(H d_Q\right)$ and $\mathrm{GEC} \leq \tilde{O}(A H d_V)$.

Relationship with Bilinear Class (Du et al., 2021)

Definition (Bilinear Class)

We say the RL problem is in a Bilinear class if there exist functions $W_h : \mathcal{H} \to \mathcal{V}$ and $X_h : \mathcal{H} \to \mathcal{V}$ for a Hilbert space $\mathcal{V}$, such that $\forall f \in \mathcal{H}$ and $h \in [H]$, we have

$$\left|\mathbb{E}_{\pi_f} \mathcal{E}_h(f, x_h, a_h)\right| \leq \left|\langle W_h(f) - W_h(f^*), X_h(f)\rangle\right|,$$
$$\left|\mathbb{E}_{x_h \sim \pi_f, a_h \sim \tilde{\pi}}\left[l_f(g, \zeta_h)\right]\right| = \left|\langle W_h(g) - W_h(f^*), X_h(f)\rangle\right|,$$

where l is a loss function with $\zeta_h = (x_h, a_h, r_h, x_{h+1})$ and $\tilde{\pi}$ is either π_f (Q-type) or π_g (V-type). The complexity of a bilinear class is characterized by the information gain: $\gamma_T(\epsilon, \mathcal{X}) = \sum_{h=1}^{H} \gamma_T(\epsilon, \mathcal{X}_h)$ with $\mathcal{X}_h = \{X_h(f) : f \in \mathcal{H}\}$.

- With $\ell_{f'}(f, x_h, a_h) = \left|\mathbb{E}_{x_{h+1}|x_h, a_h} l_{f'}(f, \zeta_h)\right|^2$, we have

$$\mathrm{GEC} \leq 2\gamma_T(\epsilon, \mathcal{X}) \quad \text{Q-type}, \qquad \mathrm{GEC} \leq 2A\gamma_T(\epsilon, \mathcal{X}), \quad \text{V-type}.$$

Relationship with Witness Rank (Sun et al., 2019)

Definition (Q-type/V-type Witness Rank)

Given a discriminator class $\mathcal{V} = \{\mathcal{V}_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]\}_{h\in[H]}$. We say an MDP has witness rank d if given two models $f, g \in \mathcal{H}$, there exists $X_h : \mathcal{H} \to \mathbb{R}^d$ and $W_h : \mathcal{H} \to \mathbb{R}^d$ such that

$$\max_{v\in\mathcal{V}_h} \mathbb{E}_{x_h\sim\pi_f, a_h\sim\tilde{\pi}}[\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)} v(x_h,a_h,x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)} v(x_h,a_h,x')] \geq \langle W_h(g), X_h(f)\rangle,$$

$$\kappa_{\mathrm{wit}} \cdot \mathbb{E}_{x_h\sim\pi_f, a_h\sim\tilde{\pi}}[\mathbb{E}_{x'\sim\mathbb{P}_{h,g}(\cdot|x_h,a_h)} V_{h+1,g}(x') - \mathbb{E}_{x'\sim\mathbb{P}_{h,f^*}(\cdot|x_h,a_h)} V_{h+1,g}(x')] \leq \langle W_h(g), X_h(f)\rangle,$$

where $\tilde{\pi}$ is either π_f (Q-type) or π_g (V-type), and $\kappa_{\mathrm{wit}} \in (0,1]$.

- With details as in the model-based examples, we have

$$\mathrm{GEC} \leq 4 d_Q H \cdot \log(\frac{\epsilon + T}{\epsilon})/\kappa_{\mathrm{wit}}^2, \quad \text{Q-type},$$
$$\mathrm{GEC} \leq 4 d_V A H \cdot \log(\frac{\epsilon + T}{\epsilon})/\kappa_{\mathrm{wit}}^2, \quad \text{V-type}.$$

Example 3: Predictive State Representations (PSR)

Predictive State Representation (PSR)

- History $\tau_h = (o_{1:h}, a_{1:h}) = (o_1, a_1, \dots, o_h, a_h)$;
- Test (future) $t_{h+1} = (o_{h+1:h+W}, a_{h+1:h+W-1})$, where length $W \in \mathbb{N}^+$;
- System dynamics matrix $\mathbb{D}_h$: i) tests as rows and histories as columns; and ii) the (t_{h+1}, τ_h)-th entry of $\mathbb{D}_h$ is equal to $\mathbb{P}(t_{h+1}|\tau_h)$;
- PSR rank d_{PSR}: $d_{\mathrm{PSR}} = \max_{h\in[H]} d_{\mathrm{PSR},h}$, where $\mathrm{Rank}(\mathbb{D}_h) = d_{\mathrm{PSR},h}$;
- Observable Operator Representation (Jaeger, 2000): given a PSR with a core test set $\{\mathcal{U}_h\}_{h\in[H]}$, there exists a set of matrices $\{\mathbf{M}_h(o,a) \in \mathbb{R}^{|\mathcal{U}_{h+1}|\times|\mathcal{U}_h|}\}_{o\in\mathcal{O},a\in\mathcal{A},h\in[H]}, \mathbf{q}_0 \in \mathbb{R}^{|\mathcal{U}_1|}$ that can characterize its dynamics:

$$\mathbb{P}(\tau_H) = \mathbf{M}_H(o_H, a_H)\mathbf{M}_{H-1}(o_{H-1}, a_{H-1}) \cdots \mathbf{M}_1(o_1, a_1)\mathbf{q}_0.$$

Connection with POMDP

- $d_{\mathrm{PSR}} \le S$: $\mathbb{D}_h = [\mathbb{P}(t_{h+1}|\tau_h)] = [\mathbb{P}(t_{h+1}|s_{h+1})] \times [\mathbb{P}(s_{h+1}|\tau_h)]$
- For one step revealing/decodable POMDPs, we can choose $\mathcal{U}_h = \mathcal{O}$

$$\mathbf{M}_h(o_h, a_h) = \underbrace{\mathbb{O}_{h+1}}_{\mathbb{R}^{O\times S}} \underbrace{\mathbb{T}_{h,a_h}}_{\mathbb{R}^{S\times S}} \underbrace{\mathrm{diag}\left(\mathbb{O}_h(o_h \mid \cdot)\right)}_{\mathbb{R}^{S\times S}} \underbrace{\mathbb{O}_h^{\dagger}}_{\mathbb{R}^{S\times O}} \in \mathbb{R}^{O\times O}, \quad \mathbf{q}_0 = \mathbb{O}_1\mu_1 \in \mathbb{R}^O.$$

Relationship with Existing Complexity Measures

Definition (α-Generalized Regular PSR)

1. For any $h \in [H]$ and $\mathbf{x} \in \mathbb{R}^{|\mathcal{U}_h|}$, it holds that

$$\max_{\pi} \sum_{o_{h:H}, a_{h:H}} |\mathbf{M}_H(o_H, a_H) \cdots \mathbf{M}_h(o_h, a_h)\mathbf{x}| \cdot \pi(o_{h:H}, a_{h:H}) \le \frac{\|\mathbf{x}\|_1}{\alpha},$$

   where $\tau_{h:H} = (o_{h:H}, a_{h:H}) \in (\mathcal{O} \times \mathcal{A})^{H-h+1}$.

2. For any $h \in [H-1]$ and $\mathbf{x} \in \mathbb{R}^{|\mathcal{U}_h|}$, it holds that

$$\max_{\pi} \sum_{(o_h, a_h) \in \mathcal{O} \times \mathcal{A}} \|\mathbf{M}_h(o_h, a_h)\mathbf{x}\|_1 \cdot \pi(o_h, a_h) \le \frac{|\mathcal{U}_{A,h+1}|}{\alpha} \|\mathbf{x}\|_1,$$

   where $\mathcal{U}_{A,h+1}$ is the the action sequences in the core test set $\mathcal{U}_{h+1}$.[a]

[a] Independent works Liu et al. (2022); Chen et al. (2022) identify similar PSR classes with regular conditions on observable operators.

- Any revealing POMDP is an $\alpha/\sqrt{S}$-generalized regular PSR.
- Any decodable POMDP is a 1-generalized regular PSR.
- Latent MDPs with the full-rank test, low-rank POMDPs, regular PSR, ...

Generalized Regular PSR Examples

GEC (model-based POMDP/PSR version):

$$\sum_{t=1}^{T} V_{f^t} - V^{\pi^t} \leq \left[d_{\mathrm{GEC}} \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s=1}^{t-1} D_H^2 \left(\mathbb{P}_{f^t}^{\pi_{\exp}(f^s,h)}, \mathbb{P}_{f^*}^{\pi_{\exp}(f^s,h)} \right) \right]^{1/2},$$

where $\pi_{\exp}(f^s, h) := \pi_{f^s} \circ_h \mathrm{Unif}(\mathcal{A}) \circ_{h+1} \mathrm{Unif}\left(\mathcal{U}_{A,h+1}\right)$ and $\mathcal{U}_{A,h+1} = \mathcal{A}^{m-1}$ for m-step revealing/decodable POMDPs.

Theorem (GEC of Generalized Regular PSR)

For α-generalized regular PSR

$$\mathrm{GEC} \leq \tilde{\mathcal{O}}\left(\frac{d_{\mathrm{PSR}} \cdot A^3 U_A^4 H}{\alpha^4} \right),$$

where d_{PSR} is the PSR rank and $U_A = \max_{h \in [H]} |\mathcal{U}_{A,h}|$.

Decision-Estimation Coefficient

DEC (Foster et al., 2021) is another complexity measure that is very general to cover most of the known tractable problems. We consider a set of models $\mathcal{M}$ and Hellinger distance D_H^2:

$$\mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}_t) = \inf_{p_t \in \Delta(\Pi)} \underbrace{\sup_{M \in \mathcal{M}}}_{\text{worst-case}} \mathbb{E}_{\pi_t \sim p_t}[\underbrace{\mathrm{Reg}_t^M}_{\text{regret when M is true model}} - \gamma \cdot \underbrace{D_{\mathrm{H}}^2(M(\pi_t), \widehat{M}_t(\pi_t))}_{\text{Easy to control}}],$$

- Convert our target (not easy to control) within one iteration to something we know how to control (assumption 4.1 of (Foster et al., 2021)):

$$\mathbb{E}_{\pi_t \sim p_t} \mathrm{Reg}_t \leq \mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}_t) + \gamma \mathbb{E}_{\pi_t \sim p_t} D_H^2\left(M^*(\pi_t), \widehat{M}_t(\pi_t)\right),$$

  where $\widehat{M}_t$ is a sequence of estimation and p_t is the solution in the definition of DEC.
- DEC is the worst-case cost for such a transformation from a game viewpoint and I think that is why DEC is also very close to the lower bound;
- We have

$$\mathbb{E}\mathrm{Reg}(T) \leq \underbrace{\sum_{t=1}^{T} \mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}_t)}_{\text{Cost of transformation}} + \underbrace{\gamma \cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p^t}\left[D_H^2\left(M^*(\pi_t), \widehat{M}_t(\pi_t)\right)\right]}_{\text{New target: online learning}}. \quad (1)$$

Decoupling Coefficient

Decoupling coefficient (Zhang, 2022; Agarwal and Zhang, 2022b,a) is a complexity measure that has applied to model-free/model-based RL and contextual bandit. We illustrate the main idea by the contextual bandit version. We consider a value class $\mathcal{F} = \{f : \mathcal{S} \times \mathcal{A} \to [-1, 1]\}$:

$$\mathbb{E}_{f^t \sim q^t, a^t = a^{f^t}(x^t))} \underbrace{V_{1,f^t}(x^t) - V_1^*(x^t, a^{f^t}(x^t))}_{\text{Feel-good regret}}$$

$$\leq \frac{d_{\mathrm{DC}}}{4\mu} + \mu \underbrace{\mathbb{E}_{a^t \sim q^t(a^t | x^t, S^{t-1})} \mathbb{E}_{f^t \sim q^t} (Q_{1,f^t}(x^t, a^t) - Q_1^*(x^t, a^t))^2}_{\text{Easy to control}}.$$

where we use $a^f(x) := \operatorname{argmax}_{a' \in \mathcal{A}} Q_{1,f}(x, a')$. DC shares similar spirits with DEC but is different in:

1. Feel-good term: $V_{1,f^t}(x^t, a^{f^t}(x^t)) - V_1^*(x^t, a^{f_*}(x^t))$: we favor f with large value;
2. Flexible choice of policy distribution: suppose that $f^t \sim q^t$:
   - DC directly picks $\pi_t = \pi_{f^t}$: $p^t(\pi) := \sum_{f \in \mathcal{H}: \pi_f = \pi} q^t(f)$;
   - DEC solves the minimax problem of definition to get:

$$p^t(\pi) = \operatorname*{argmin}_{p \in \Delta(\Pi)} \sup_{f \in \mathcal{H}} \mathbb{E}_{\pi_t \sim p^t} [\underbrace{\mathrm{Reg}_t^M}_{\text{regret when f is true model}} - \gamma \cdot \mathbb{E}_{f^t \sim q^t} \underbrace{D_{\mathrm{H}}^2(f(\pi_t), f^t(\pi_t))}_{\text{Easy to control}}];$$

3. Flexible choice of notion of new target.

Reduction-based Framework

- GEC reduces out-of-sample V_{1,f^t} to in-sample error estimation:
  1. A low GEC: model-based + model-free;
  2. An effective in-sample error estimator;
  3. Handle the difference between $V_{1,f}$ and V_1^*;

$$\mathrm{Reg}(T) \lesssim \Big[d_{\mathrm{GEC}} \cdot \sum_{t=1}^{T}\sum_{s=1}^{t-1} \ell^s(f^t)\Big]^{1/2} \le \underbrace{\gamma \sum_{t=1}^{T}\sum_{s=1}^{t-1} \ell^s(f^t)}_{\text{New target: in-sample estimation}} + \frac{1}{\gamma}\cdot d_{\mathrm{GEC}}.$$

- DEC reduces out-of-sample V_1^* to another out-of-sample target:
  1. A low DEC: model-based;
  2. An effective online learning oracle;

$$\mathbb{E}\mathrm{Reg}(T) \le \underbrace{\sum_{t=1}^{T} \mathrm{dec}_\gamma^H(\mathcal{M}, \mu^t)}_{\text{Cost of transformation}} + \underbrace{\gamma \cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p^t}\mathbb{E}_{\widehat{M}_t \sim \mu^t}\Big[D^{\pi_t}(\widehat{M}_t \| M^*)\Big]}_{\text{New target: online learning}}.$$

- DC/O-DEC reduces out-of-sample V_{1,f^t} to another optimistic out-of-sample target:
  1. A low complexity measure: model-based + model-free;
  2. An effective online learning oracle;
  3. Handle the difference between $V_{1,f}$ and V_1^*.

$$\mathbb{E}\mathrm{Reg}(T) \le \underbrace{\sum_{t=1}^{T} \mathrm{odec}_\gamma^D(\mathcal{M}, \mu^t)}_{\text{Cost of transformation}} + \underbrace{\gamma \cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p^t}\mathbb{E}_{\widehat{M}_t \sim \mu^t}\Big[D^{\pi_t}(\widehat{M}_t \| M^*) - \gamma^{-1}\Delta V_{1,\widehat{M}_t}(x_1)\Big]}_{\text{New target: online learning with feel-good term}}.$$