

TrustML Young Scientist Seminars, RIKEN AIP Japan
Aug. 01, 2023

On The Theory of Offline Reinforcement Learning: Data diversity, posterior sampling and beyond

Thanh Nguyen-Tang

Department of Computer Science, Johns Hopkins University

Joint with



Raman Arora¹
JHU CS

Offline RL Intro

Reinforcement Learning Settings



Inventory Management (Madeka et. al., 2022)



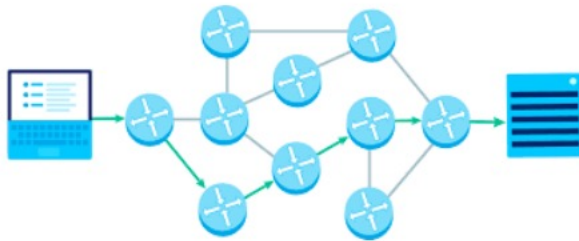
Finite Armed Bandits (Lai and Robbins, 1985)



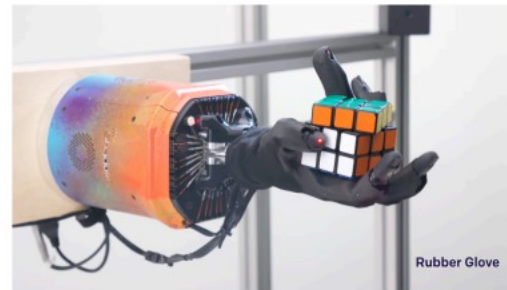
Google X / Waymo [2016]



Deepmind AlphaGo [2016]



Dynamic routing (Awerbuch and Kleinberg, 2004)



OpenAI [2019]



Deepmind AlphaStar [2019]

What if the interaction with the environment is expensive?

Online RL may be risky, unethical or prohibitive in high-stakes applications such as self-driving cars, financial investment and clinical diagnosis

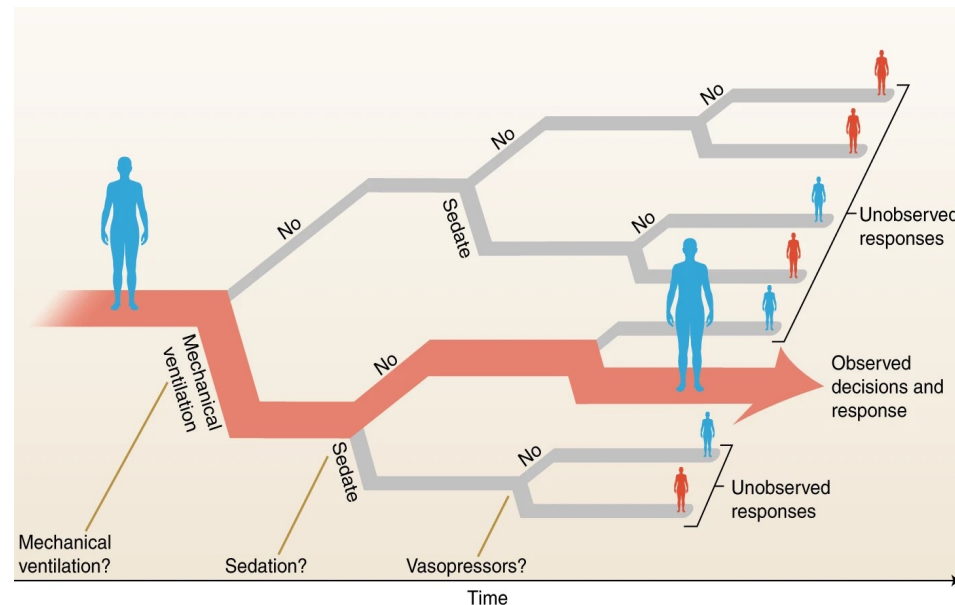


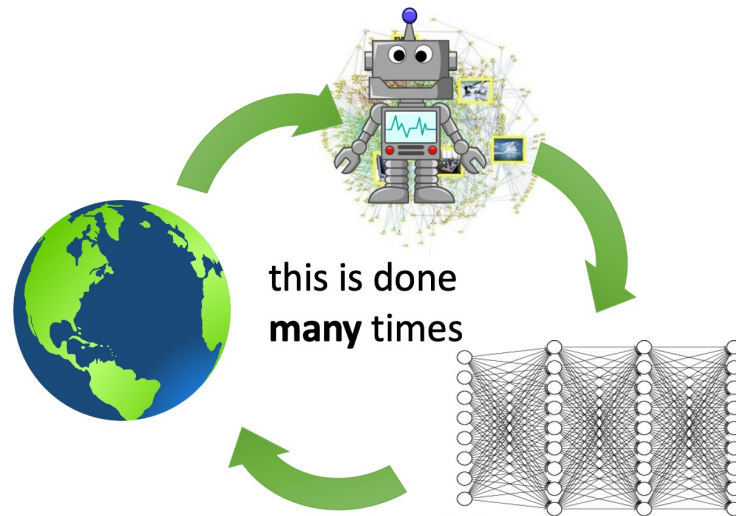
Figure from Gottesman et al. *"Guidelines for reinforcement learning in healthcare"*. Nature Medicine, 2019

Offline reinforcement learning

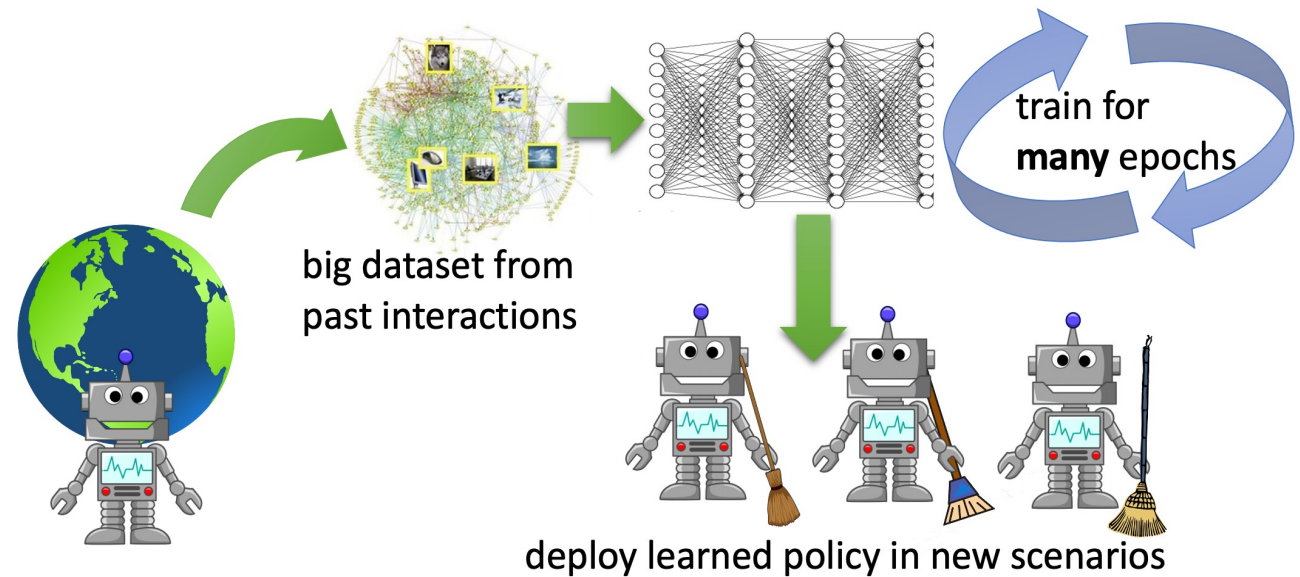
Central question that leads to offline RL

Can we leverage the availability of historical interaction datasets to learn an optimal behavior?

reinforcement learning

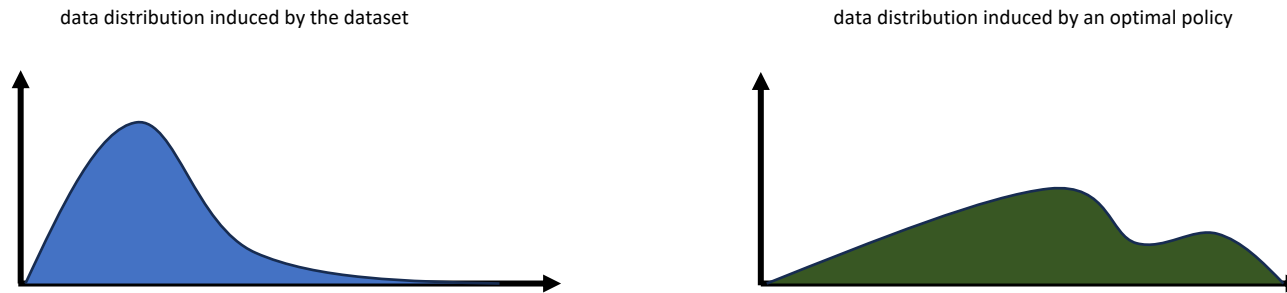


offline reinforcement learning



Challenges in offline RL

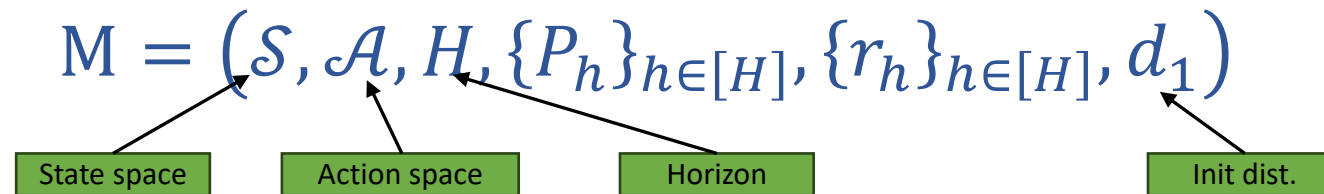
- **Distributional shift:** The distribution of the offline data is different from the data distribution induced by a target policy.



- **Generalization:** How to generalize from scenarios seen in the datasets to new scenarios?
- **Efficiency:** How to design statistical-efficient and oracle-efficient algorithms?

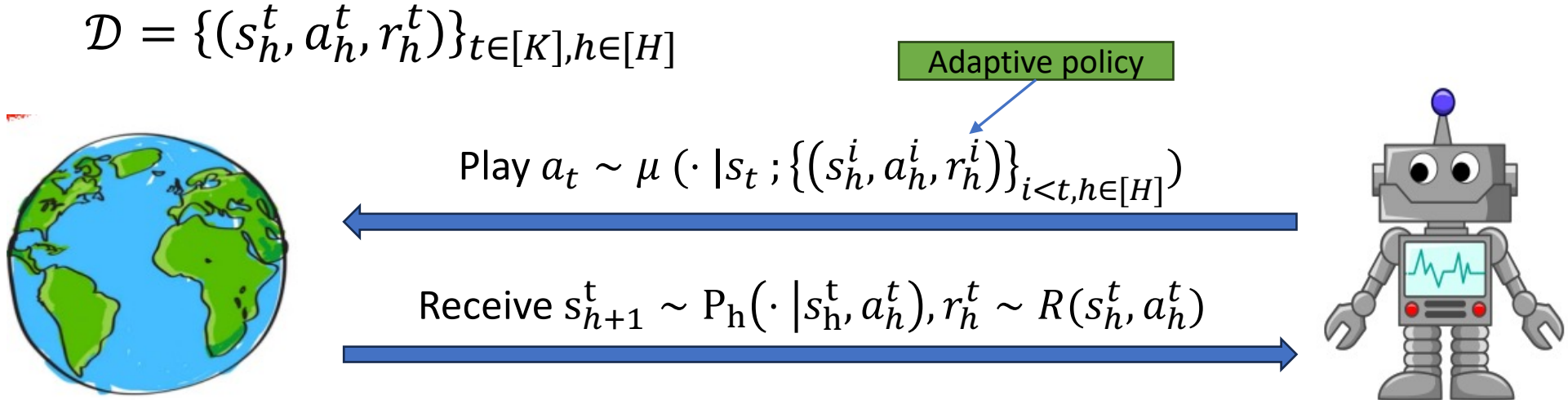
Setup

Episodic time-inhomogeneous Markov decision process



- Transition kernels $P = (P_1, \dots, P_H)$, where $P_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$
- Reward functions $r = (r_1, \dots, r_H)$, where $r_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta([-b, b])$
- On playing $\pi = \{\pi_h\}_{h \in [H]}$ on M , where $\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})$:
 - Observe trajectory: $(s_1, a_1, r_1, \dots, s_H, a_H, r_H)$
where $s_1 \sim d_1, a_h \sim \pi_h(\cdot | s_h), s_{h+1} \sim P_h(\cdot | s_h, a_h), r_h \sim r_h(s_h, a_h)$
 - Assume: the total rewards $|\sum_{h=1}^H r_h| \leq b$ for some constant $b > 0$

Data collection



where μ is an *adaptive* behavior policy

Goal: Find $\hat{\pi} = \text{OfflineRLAlgo}(\mathcal{D})$ that competes with any comparator policy $\pi \in \Pi^{all}$:

$$\text{SubOpt}_{\pi}(\hat{\pi}) = V_1^{\pi}(s_1) - V_1^{\hat{\pi}}(s_1), \text{ where } V_h^{\pi}(s) = \mathbb{E}_{\pi}[\sum_{i=h}^H r_i | s_h = s]$$

Notation: $\mu^t(\cdot | s) = \mu(\cdot | s; \{(s_h^i, a_h^i, r_h^i)\}_{i < t, h \in [H]})$, $\mu = \frac{1}{K} \sum_{t=1}^K \mu^t$

Research questions & our key result summary

Offline RL

Data coverage

- Uniform concentrability: [Munos and Szepesvári, 2008](#), [Chen and Jiang, 2019](#), [Nguyen-Tang et al., 2022b](#)
- Distribution χ^2 mismatch: [Duan et al. 2020](#)
- **Pessimism principle**: "When extrapolate from the offline data, take decision-making on the basis of the worst-case scenarios"
 - Single-policy concentrability coefficients: [Liu et al., 2019](#), [Rashidinejad et al., 2021](#), [Yin and Wang, 2021](#)
 - Relative condition numbers: [Agarwal et al., 2021](#), [Uehara and Sun, 2002](#), [Zanette et al. 2021](#)
 - Bellman residual ratios: [Xie et al., 2021](#)

Algorithms

Most become inefficient in general function approximation

- Lower confidence bounds (LCB) [\[Jin et al., 2021\]](#),
- Version space [\[Xie et al., 2021\]](#) ...

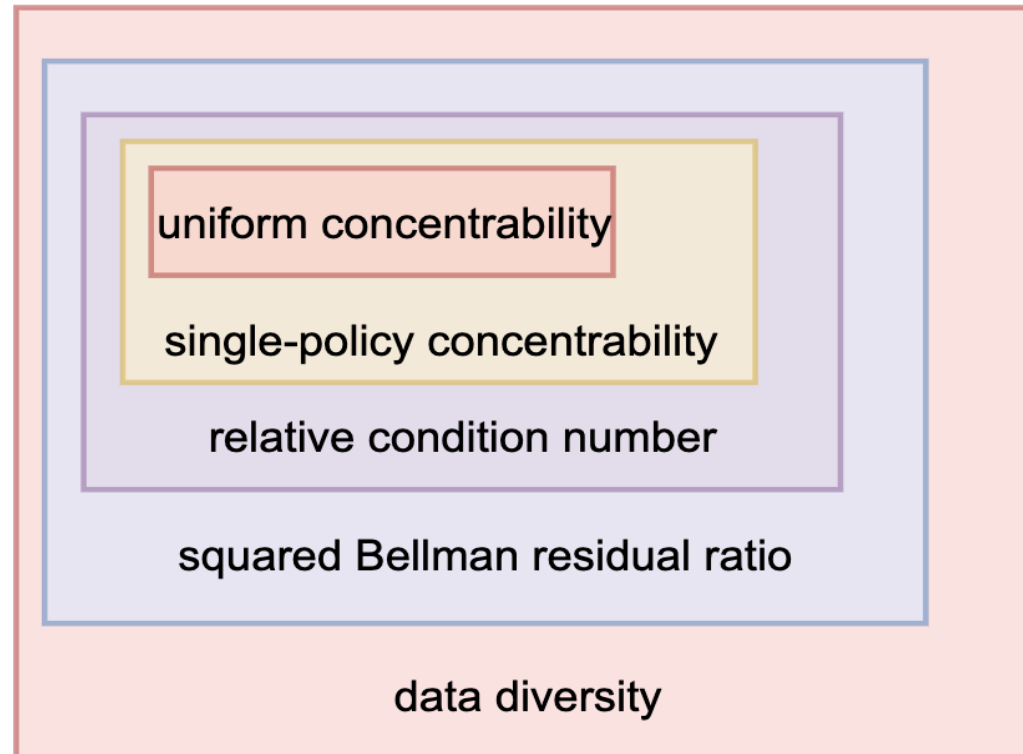
Questions

What **minimal conditions** that enable sample-efficient offline RL?
Can we design **oracle-efficient** algorithms with **competitive guarantees**?

Our key contributions

- A new notion of data diversity
 - It generalizes all the prior measures of data coverage
- Unified results that show that: we can go beyond the **version space** algorithms to obtain oracle-efficient algorithms (**regularization opt** and **posterior sampling**) with comparable guarantees
 - First result that shows regularized opt. is comparable to version space algorithm, despite the prior work shows the unfavorable sample complexity of regularized opt for offline RL
 - First algorithm and guarantee for posterior sampling for offline RL, matching the regret bound of the version space algorithm


Data diversity



Contribution #1: Data diversity expands the scenarios for the offline data that enable sample-efficient offline RL

Prior work: Bellman-consistent pessimism

Idea: Instead of calculating point-wise lower bound for the value function, implement pessimism at the initial state over the set of functions consistent with the Bellman equations in the offline data:

$$\hat{\pi} = \arg \max_{\pi \in \Pi} \min_{f \in \{f \in \mathcal{F} : L(f, \pi; \mathcal{D}) \leq \epsilon\}} f^{\pi}(s_0)$$


where

$$\mathcal{E}(f, \pi; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(s, a, s', r) \in \mathcal{D}} (f(s, a) - r - f^{\pi}(s'))^2 - \min_{f' \in \mathcal{F}} \frac{1}{|\mathcal{D}|} \sum_{(s, a, s', r) \in \mathcal{D}} (f'(s, a) - r - f^{\pi}(s'))^2$$

$$f^{\pi}(s) := \mathbb{E}_{a \sim \pi(\cdot | s)}[f(s, a)]$$

Bellman-consistent pessimism

can become vacuous

- Bound: $\text{SubOpt}_\pi(\hat{\pi}) \leq \frac{\sqrt{C_2(\pi) \log(|\mathcal{F}| |\Pi^{all}|)}}{\sqrt{K}}$
- Pro: avoid the overly pessimistic solution in LCB \rightarrow tighter bounds in some cases
- Cons:
 - The algorithm is not efficient (due to search over the version space)
 - The bound scales with $\log |\Pi^{all}|$ where Π^{all} is the class of "comparator" policies \rightarrow exponentially large

Regularized optimization of Xie et al.

- Proposed regularization (policy) optimization to avoid minmax optimization over the version space:

Algorithm 1 PSPI: Pessimistic Soft Policy Iteration

Input: Batch data \mathcal{D} , regularization coefficient λ .

1: Initialize policy π_1 as the uniform policy.

2: **for** $t = 1, 2, \dots, T$ **do**

3: Obtain the pessimistic estimation for π_t as f_t ,

$$f_t \leftarrow \operatorname{argmin}_{f \in \mathcal{F}} (f(s_0, \pi_t) + \lambda \mathcal{E}(f, \pi_t; \mathcal{D})), \quad (4.1)$$

regularization

where $\mathcal{E}(f, \pi_t; \mathcal{D})$ is defined in Eq.(3.1).

4: Calculate π_{t+1} by,

$$\pi_{t+1}(a|s) \propto \pi_t(a|s) \exp(\eta f_t(s, a)), \quad \forall s, a \in \mathcal{S} \times \mathcal{A}.$$

5: **end for**

6: Output $\bar{\pi} := \text{Unif}(\pi_{[1:T]}).$

▷ uniformly mix π_1, \dots, π_T at the trajectory level

- Bound: $\frac{\sqrt{C_2(\pi)} \cdot \sqrt[3]{\log(|\mathcal{F}| |\Pi^{\text{soft}}|)}}{K^{\frac{1}{3}}}$: much slower rate compared to version space algorithms
- very slow rate

Prior work: posterior sampling for offline RL

- Posterior sampling: can be considered as an implicit optimization
 - Amenable to various efficient approximation (e.g., Langevin dynamics)
- Widely used in online RL due to sampling is suitable for exploration
- Never really a great success in offline RL?
 - The only work of posterior sampling for offline RL is by Uehara and Sun '21,
 - but it's model-based + Bayesian regret \rightarrow easier to analyze
 - *model-based* methods: hardly scale in practice
 - *Bayesian regret* has limitations: need realizability in the prior and likelihood assumptions and the Bayesian regret \leq worst-case (over the prior) regret

Algorithm classes for offline RL

Contribution #2:

- Regularized optimization has **comparable** guarantees as version space-based algorithm
- Posterior sampling has **comparable** guarantees as version space-based algorithm

Implications: Practitioner can start to think of using RO and PS (maybe with their approximations) because RO and PS achieve the same guarantees just as VS

Algorithms

Our motivation

Computer Science > Machine Learning

arXiv:2108.08812 (cs)

[Submitted on 19 Aug 2021]

Provable Benefits of Actor–Critic Methods for Offline Reinforcement Learning

Andrea Zanette, Martin J. Wainwright, Emma Brunskill

Download PDF

Actor–critic methods are widely used in offline reinforcement learning practice, but are not so well-understood theoretically. We propose a new offline actor–critic algorithm that naturally incorporates the pessimism principle, leading to several key advantages compared to the state of the art. The algorithm can operate when the Bellman evaluation operator is closed with respect to the action value function of the actor’s policies; this is a more general setting than the low-rank MDP model. Despite the added generality, the procedure is computationally tractable as it involves the solution of a sequence of second-order programs. We prove an upper bound on the suboptimality gap of the policy returned by the procedure that depends on the data coverage of any arbitrary, possibly data dependent comparator policy. The achievable guarantee is complemented with a minimax lower bound that is matching up to logarithmic factors.

Comments: Initial submission; appeared as spotlight talk in ICML 2021 Workshop on Theory of RL

Subjects: **Machine Learning (cs.LG)**

Cite as: [arXiv:2108.08812](https://arxiv.org/abs/2108.08812) [cs.LG]

(or [arXiv:2108.08812v1](https://arxiv.org/abs/2108.08812v1) [cs.LG] for this version)

<https://doi.org/10.48550/arXiv.2108.08812> 

Computer Science > Machine Learning

arXiv:2208.10904 (cs)

[Submitted on 23 Aug 2022]

A Provably Efficient Model–Free Posterior Sampling Method for Episodic Reinforcement Learning

Christoph Dann, Mehryar Mohri, Tong Zhang, Julian Zimmert

Download PDF

Thompson Sampling is one of the most effective methods for contextual bandits and has been generalized to posterior sampling for certain MDP settings. However, existing posterior sampling methods for reinforcement learning are limited by being model–based or lack worst–case theoretical guarantees beyond linear MDPs. This paper proposes a new model–free formulation of posterior sampling that applies to more general episodic reinforcement learning problems with theoretical guarantees. We introduce novel proof techniques to show that under suitable conditions, the worst–case regret of our posterior sampling method matches the best known results of optimization based methods. In the linear MDP setting with dimension, the regret of our algorithm scales linearly with the dimension as compared to a quadratic dependence of the existing posterior sampling–based exploration algorithms.

Subjects: **Machine Learning (cs.LG)**

Cite as: [arXiv:2208.10904](https://arxiv.org/abs/2208.10904) [cs.LG]

(or [arXiv:2208.10904v1](https://arxiv.org/abs/2208.10904v1) [cs.LG] for this version)

<https://doi.org/10.48550/arXiv.2208.10904> 

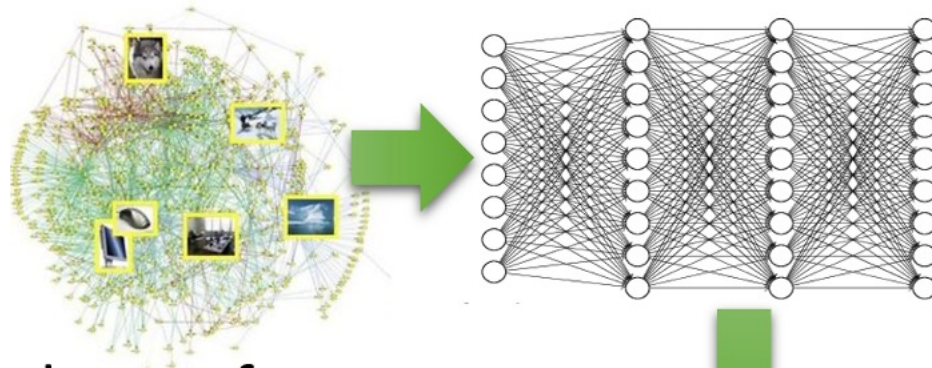
Journal reference: Dann C, Mohri M, Zhang T, Zimmert J. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*. 2021 Dec 6;34:12040–51

- Our algorithm framework generalizes the actor-(pessimistic) critic framework in Zanette et al. to general function approximation;
- Our *Pessimistic* Posterior sampling builds on some components of Dann et al.

Function approximation for value functions

Function class: $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H \subset \{\mathcal{S} \times \mathcal{A} \rightarrow [-b, b]\}^H$

- E.g., a finite function class, a linear function class, neural networks



A generic offline policy optimization (GOPO)

Algorithm 1 $\text{GOPO}(\mathcal{D}, \mathcal{F}, \eta, T, \boxed{\text{CriticCompute}})$:
Generic Offline Policy Optimization Framework

1: **Input:** Offline data \mathcal{D} , function class \mathcal{F} , learning rate $\eta > 0$, and iteration number T
2: Uniform policy $\pi^1 = \{\pi_h^1\}_{h \in [H]}$
3: **for** $t = 1, \dots, T$ **do**
4: $\underline{Q}^t = \boxed{\text{CriticCompute}}(\pi^t, \mathcal{D}, \mathcal{F}, \dots)$
5: $\pi_h^{t+1}(a|s) \propto \pi_h^t(a|s) \exp(\eta \underline{Q}_h^t(s, a)), \forall (s, a, h)$
6: **end for**
7: **Output:** $\hat{\pi} = \text{Uniform}(\{\pi_h^t\}_{t \in [T]})$

Actor (for policy improvement)

Critic: pessimistic estimate of the actor value

- **Actor (line 5):** Use the multiplicative weights algorithm for policy improvement
- **(Pessimistic) Critic (line 4):** Its goal is to give a *pessimistic* value of the current actor (policy)

The (pessimistic) critic function

The critic compute module in GOPO can be:

- Version space-based critic (VSC)
- Regularized optimization-based critic (ROC)
- Posterior sampling-based critic (PSC)

Version space-based critic (VSC)

- Given current critic π^t , construct the version space $\mathcal{F}(\beta; \pi^t)$ that is consistent with the offline data
- Pessimism: Search over the version space for a value function $f \in \mathcal{F}$ that has a minimal initial value

Algorithm 2 VSC($\mathcal{D}, \mathcal{F}, \pi^t, \beta$): Version Space-based Critic

1: $\mathcal{F}(\beta; \pi^t) := \{f \in \mathcal{F} : \hat{L}_{\pi^t}(f_h, f_{h+1}) \leq \inf_{g \in \mathcal{F}} \hat{L}_{\pi^t}(g_h, f_{h+1}) + \beta, \forall h \in [H]\}$

2: $\underline{Q}^t \in \arg \min_{f \in \mathcal{F}(\beta; \pi^t)} f_1^{\pi^t}(s_1)$

3: **Output:** \underline{Q}^t

Pessimistic estimate of value function in the first timestep

$$\hat{L}_{\pi}(f_h, f_{h+1}) = \sum_{t=1}^K \left(f_h(s_h^t, a_h^t) - r_h^t - f_{h+1}^{\pi}(s_{h+1}^t) \right)^2$$

$$f^{\pi}(s) := \mathbb{E}_{a \sim \pi(\cdot|s)}[f(s, a)]$$

Regularized optimization-based critic

- Enforce pessimism via a regularization \rightarrow unconstrained optimization

Algorithm 3 ROC($\mathcal{D}, \mathcal{F}, \pi^t, \lambda$): Regularized Optimization-based Critic

- 1: $\mathcal{L}_{\pi^t}(f) := \sum_{h=1}^H \hat{L}_{\pi^t}(f_h, f_{h+1}) - \inf_{g \in \mathcal{F}} \sum_{h=1}^H \hat{L}_{\pi^t}(g_h, f_{h+1})$
 - 2: $\underline{Q}^t \leftarrow \arg \inf_{f \in \mathcal{F}} \left\{ \lambda f_1^{\pi^t}(s_1) + \mathcal{L}_{\pi^t}(f) \right\}$
 - 3: **Output:** \underline{Q}^t
-

Regularization for first timestep

Practical implementations that works great in practice: Cheng et al. “*Adversarially Trained Actor Critic for Offline Reinforcement Learning*”, 2022

Pessimistic posterior sampling

- Replace optimization by sampling

Algorithm 4 PSC($\mathcal{D}, \mathcal{F}, \pi^t, \lambda, \gamma, p_0$): Posterior Sampling-based Critic

$$1: \underline{Q}^t \sim \hat{p}(f|\mathcal{D}, \pi^t) \propto \underbrace{\exp\left(-\lambda f_1^{\pi^t}(s_1)\right)}_{\text{Pessimistic prior}} p_0(f) \prod_{h \in [H]} \frac{\exp\left(-\gamma \hat{L}_{\pi^t}(f_h, f_{h+1})\right)}{\mathbb{E}_{f'_h \sim p_{0,h}} \exp\left(-\gamma \hat{L}_{\pi^t}(f'_h, f_{h+1})\right)}$$

2: **Output:** \underline{Q}^t

- The likelihood component: encourages the sampled function to have small Bellman errors in the offline data
- The key component: the pessimistic prior $\exp\left(-\lambda f_1^{\pi^t}(s_1)\right)$
 - Encourages the value function sampled from the posterior to have a small value in the initial state, implicitly enforcing pessimism.

Main Results

Extrapolate from one distribution to another under a witness function class

- Offline RL requires *extrapolation* from one data distribution to another
- To measure the extrapolation from distribution p to q under the function class \mathcal{Q} , define “ χ value”:

$$\chi_{\mathcal{Q}}(\epsilon; q, p) = \inf \{ C \geq 0 : (\mathbb{E}_q[g])^2 \leq C \cdot \mathbb{E}_p[g^2] + \epsilon, \forall g \in \mathcal{Q} \}$$



Data diversity in MDP

- Offline learner does not have direct access to the trajectory of a comparator policy $\pi \in \Pi^{all}$
- They can only observe partial information about the goodness of π channeled through the “transferability” with the behavior policy μ

Definition 3. For any comparator policy $\pi \in \Pi^{all}$, we measure the data diversity of the behavior policy μ with respect to a target policy π by χ value

$$\mathcal{C}(\pi; \epsilon) := \max_{h \in [H]} \chi_{(\mathcal{F}_h - \mathcal{F}_h)}(\epsilon; d_h^\pi, d_h^\mu), \forall \epsilon \geq 0, \quad (2)$$

where $\mathcal{F}_h - \mathcal{F}_h$ is the Minkowski difference between the function class \mathcal{F}_h and itself, i.e., $\mathcal{F}_h - \mathcal{F}_h := \{f_h - f'_h : f_h, f'_h \in \mathcal{F}\}$,

witness function class

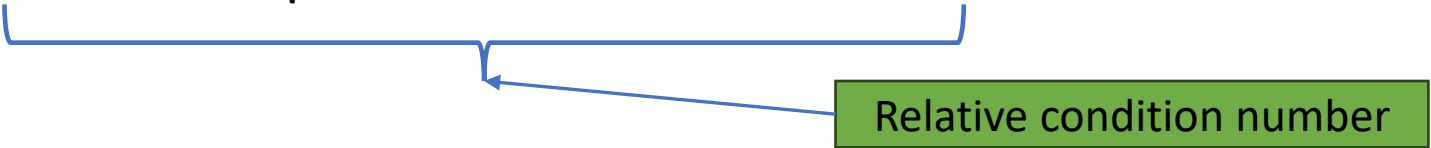
state-action dist. by offline data

state-action dist. by comparator policy π

Here: $d_h^\pi(s, a) = \Pr((s_h, a_h) = (s, a) | \pi)$ is the occupancy state-action density/distribution

Data diversity

- Despite being abstract, data diversity is always upper-bounded by:
 - Single-policy concentrability coefficient
 - Relative condition numbers
 - Bellman residual ratio
 - Generalized Chi-squared divergence
- Property: $\mathcal{C}(\pi; \epsilon)$ is non-increasing with ϵ
- E.g., $\mathcal{F}_h = \{w^T \phi(s, a) | w \in \mathbb{R}^d\}$ and linear MDP with feature ϕ

$$\mathcal{C}(\pi; \epsilon) \leq \mathcal{C}(\pi; 0) \leq \sup_{w \in \mathbb{R}^d} \frac{w^T \mathbb{E}_\pi[\phi(s_h, a_h) \phi(s_h, a_h)^T] w}{w^T \mathbb{E}_\mu[\phi(s_h, a_h) \phi(s_h, a_h)^T] w}$$


Relative condition number

A unified theory for three algorithms

Theorem (Informal): Upper bounds of VS, RO, PS

Under *realizability* and *Bellman completeness*,

- For VS, RO: with high probability,

$$\forall \pi \in \Pi^{all}, V_1^\pi(s_1) - V_1^{\hat{\pi}}(s_1) \leq \frac{H b \sqrt{\text{COMP} \cdot \mathcal{C}(\pi; K^{-1})}}{\sqrt{K}}$$

- For PS:

$$\forall \pi \in \Pi^{all}, \mathbb{E}_{\mathcal{D}} \mathbb{E}_{\hat{\pi} \sim \hat{p}(\cdot|\mathcal{D})} [V_1^\pi(s_1) - V_1^{\hat{\pi}}(s_1)] \leq \frac{H b \sqrt{\text{COMP} \cdot \mathcal{C}(\pi; K^{-1})}}{\sqrt{K}}$$

worst-case bound

where COMP measures the complexity of the function class \mathcal{F} and its induced

policy class $\Pi^{soft} = \Pi_1^{soft} \times \dots \times \Pi_H^{soft}$ where

$$\Pi_h^{soft} = \Pi_h^{soft}(T, \eta) = \left\{ \pi_h(a|s) \propto \exp\left(\eta \sum_{i=1}^t g_i(s, a)\right) : \forall t \in [T], \forall g_i \in \mathcal{F}_h \right\}$$

Instantiating our bounds to *finite* function classes

Methods	Sub-optimality Bound	Data
VS [Xie et al., 2021]	$Hb\sqrt{C_2(\pi) \cdot \ln(\mathcal{F} \cdot \Pi^{all})} \cdot K^{-1/2}$	I
RO in [Xie et al., 2021]	$Hb\sqrt{C_2(\pi) \cdot \sqrt[3]{\ln(\mathcal{F} \cdot \Pi^{soft})}} \cdot K^{-1/3}$	I
MBPS in [Uehara and Sun, 2021]	$Hb\sqrt{C^{Bayes} \cdot \ln \mathcal{M} } \cdot K^{-1/2}$ (Bayesian)	I
VS in Algorithm 2	$Hb\sqrt{C(\pi; 1/\sqrt{K}) \cdot \ln(\mathcal{F} \cdot \Pi^{soft})} \cdot K^{-1/2}$	A
RO in Algorithm 3	$Hb\sqrt{C(\pi; 1/\sqrt{K}) \cdot \ln(\mathcal{F} \cdot \Pi^{soft})} \cdot K^{-1/2}$	A
MFPS in Algorithm 4	$Hb\sqrt{C(\pi; 1/\sqrt{K}) \cdot \ln(\mathcal{F} \cdot \Pi^{soft})} \cdot K^{-1/2}$ (frequentist)	A

Much slower rate

standard rate

tighter distribution shift measure

I: episodes are collected **independently**
A: episodes are collected **adaptively**

Instantiating our bounds to *linear* function classes

Method	Sub-optimality bound
PEVI [Jin et al., 2021b]	$Hb \cdot \sqrt{C_{pevi}(\pi)} \cdot d$
PACLE [Zanette et al., 2021]	$Hb\sqrt{C_{pacle}(\pi)} \cdot d$
VC [Xie et al., 2021, Section 3]	$Hb\sqrt{C_{bcp}(\pi)} \cdot d/K$
RO [Xie et al., 2021, Section 4]	$Hb\sqrt{C_{bcp}(\pi)} \sqrt[3]{d/K}$
Ours (VS, RO, PS)	$Hb\sqrt{\mathcal{C}(\pi; 1/\sqrt{K})} \cdot d/K$

Much slower rate

$$\Sigma_h := \lambda I + \sum_{k=1}^K \phi_h(s_h^k, a_h^k) \phi_h(s_h^k, a_h^k)^T,$$

$$\bar{\phi}_h^\pi := \mathbb{E}_\pi[\phi_h(s_h, a_h)],$$

$$\bar{\Sigma}_h := \mathbb{E}_\mu [\phi(s_h, a_h) \phi(s_h, a_h)^T].$$

$$C_{pevi}(\pi) := \max_{h \in [H]} \left(\mathbb{E}_\pi [\|\phi_h(s_h, a_h)\|_{\Sigma_h^{-1}}] \right)^2,$$

$$C_{pacle}(\pi) := \max_{h \in [H]} \|\bar{\phi}_h^\pi\|_{\Sigma_h^{-1}}^2,$$

$$C_{bcp}(\pi) := \max_{h \in [H]} \left(\mathbb{E}_\pi [\|\phi_h(s_h, a_h)\|_{\bar{\Sigma}_h}] \right)^2.$$

$$C_{pevi}(\pi) \geq C_{pacle}(\pi) \approx \mathcal{C}(\pi; 0)/K \leq C_{bcp}(\pi)/K.$$

Table 2: Comparison of sub-optimality bounds when the function class \mathcal{F}_h is linear in $\phi_h : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$.

standard rate

Key Technique Highlight

Key proof techniques

- Data diversity + decoupling argument + concentration inequality
- A new technical argument that handles statistical dependence in the posterior sampling analysis of Dann et al.
 - Idea: carefully incorporate the uniform convergence argument into the in-expectation bounds

Error decomposition

$$\begin{aligned}
 \text{SubOpt}_{\pi}(\hat{\pi}) &= \frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi} \left[(\underline{Q}_h^t - T_h^{\pi^t} \underline{Q}_{h+1}^t)(s_h, a_h) \right] && \text{Bellman error} \\
 &+ \frac{1}{T} \sum_{t=1}^T \underbrace{\underline{Q}_1^t(s_1, \pi^t) - V_1^{\pi^t}(s_1)}_{=:\Delta_t \text{ (where to enforce pessimism)}} && \text{Value gap at initial state} \\
 &+ \underbrace{\frac{1}{T} \sum_{t=1}^T \text{SubOpt}_{\pi}^{M(\underline{Q}^t, \pi^t)}(\pi^t)}_{\leq \frac{H b \sqrt{|\mathcal{A}|}}{\sqrt{T}}} && \text{Regret of online learning}
 \end{aligned}$$

Decoupling lemma

For any $f \in \mathcal{F}, \tilde{\pi} \in \Pi^{soft}, \pi \in \Pi^{all}, \lambda > 0, \epsilon \geq 0$:

$$\sum_{h=1}^H \mathbb{E}_{\pi}[(f_h - T^{\tilde{\pi}} f_{h+1})(s_h, a_h)] \leq \frac{1}{2\lambda} \sum_{h=1}^H \sum_{k=1}^K \mathbb{E}_{\mu^k}[\{(f_h - T^{\tilde{\pi}} f_{h+1})(s_h, a_h)\}^2] + \frac{\lambda H \cdot \mathcal{C}(\pi; \epsilon)}{2K} + H \epsilon$$

Squared Bellman error under offline distribution

Data diversity

Additive error in extrapolation

Proof: Simple rearrangement + AM-GM inequality

For Version space-based algorithm

$$\sum_{k=1}^K \mathbb{E}_{\mu^k} \left[\{(f_h - T_h^{\pi^t} f_{h+1})(s_h, a_h)\}^2 \right] \leq 2 \quad \underbrace{\mathcal{L}_{\pi^t}(f)}_{\text{loss function in VSC}} + b^2 H \cdot \text{COMP} +$$

Freedman's inequality
Empirical squared Bellman error

$$\min_{f \in \mathcal{F}(\beta; \pi^t)} \mathcal{L}_{\pi^t}(f) \leq b^2 H \cdot \text{COMP}$$

$$\underline{Q}_1^t(s_1, \pi^t) - V_1^{\pi^t}(s_1) \leq 0$$

Version space
pessimism

Thus,

$$\text{SubOpt}_{\pi}(\hat{\pi}) \leq \frac{b^2 H \cdot \text{COMP}}{2 \lambda} + \frac{\lambda H \cdot \mathcal{C}(\pi; \epsilon)}{2 K} + H \epsilon + \frac{H b \sqrt{|\mathcal{A}|}}{\sqrt{T}}$$

Minimize LHS wrt $\lambda > 0$

$$\leq bH \frac{\sqrt{\mathcal{C}(\pi; \epsilon) \cdot \text{COMP}}}{\sqrt{K}} + H \epsilon + \frac{H b \sqrt{|\mathcal{A}|}}{\sqrt{T}}$$

For regularization optimization algorithm

$$\sum_{h=1}^H \mathbb{E}_{\pi} \left[(f_h - T_h^{\pi^t} f_{h+1})(s_h, a_h) \right] + (f_1^{\pi^t}(s_1) - V_1^{\pi^t}(s_1))$$

Decoupling lemma

Squared Bellman error under offline distribution

$$\leq \frac{1}{2\lambda} \sum_{h=1}^H \sum_{k=1}^K \mathbb{E}_{\mu^k} \left[\{(f_h - T_h^{\pi^t} f_{h+1})(s_h, a_h)\}^2 \right] + (f_1^{\pi^t}(s_1) - V_1^{\pi^t}(s_1)) + \frac{\lambda H \cdot \mathcal{C}(\pi; \epsilon)}{2K} + H \epsilon$$

$$= \frac{\sum_{h=1}^H \sum_{k=1}^K \mathbb{E}_{\mu^k} \left[\{(f_h - T_h^{\pi^t} f_{h+1})(s_h, a_h)\}^2 \right] + 2\lambda \cdot (f_1^{\pi^t}(s_1) - V_1^{\pi^t}(s_1))}{2\lambda} + \frac{\lambda H \cdot \mathcal{C}(\pi; \epsilon)}{2K} + H \epsilon$$

Empirical squared Bellman error

Freedman's inequality

$$\leq \frac{2 \left(\mathcal{L}_{\pi^t}(f) + \lambda \cdot (f_1^{\pi^t}(s_1) - V_1^{\pi^t}(s_1)) \right) + b^2 H \cdot \text{COMP}}{2\lambda} + \frac{\lambda H \cdot \mathcal{C}(\pi; \epsilon)}{2K} + H \epsilon$$

Note: $\min_{f \in \mathcal{F}} \{ \mathcal{L}_{\pi^t}(f) + \lambda \cdot (f_1^{\pi^t}(s_1) - V_1^{\pi^t}(s_1)) \} \leq b^2 H \cdot \text{COMP}$

Regularized loss

For posterior sampling –based algorithm

Challenge: data-dependent policy

Likelihood function

- Log partition function

$$Z = \mathbb{E}_{\mathcal{D}} \mathbb{E}_{\pi^t \sim \hat{p}(\cdot|\mathcal{D})} \mathbb{E}_{f \sim \hat{p}(\cdot|\pi^t, \mathcal{D})} \left[\hat{\Phi}(f, \pi^t; \mathcal{D}) + \lambda \left(f_1^{\pi^t}(s_1) - V_1^{\pi^t}(s_1) \right) + \log \hat{p}(f|\mathcal{D}, \pi^t) \right]$$

- Lower bound Z:

Squared Bellman error under offline distribution

$$Z \geq \mathbb{E}_{\mathcal{D}} \mathbb{E}_{\pi^t \sim \hat{p}(\cdot|\mathcal{D})} \mathbb{E}_{f \sim \hat{p}(\cdot|\pi^t, \mathcal{D})} \left[0.125 \gamma \sum_{h=1}^H \sum_{k=1}^K \mathbb{E}_{\mu^k} \left[\{(f_h - T_h^{\pi^t} f_{h+1})(s_h, a_h)\}^2 \right] + \lambda \left(f_1^{\pi^t}(s_1) - V_1^{\pi^t}(s_1) \right) \right] - \gamma H b^2 \cdot \text{COMP}$$

- Upper bound Z:

$$Z \leq \frac{\lambda}{\sqrt{K}} + b^2 \gamma H \cdot \text{COMP}$$

- Decoupling:

$$\begin{aligned} \text{regret} &\leq \frac{\mathbb{E}_{\mathcal{D}} \mathbb{E}_{\pi^t \sim \hat{p}(\cdot|\mathcal{D})} \mathbb{E}_{f \sim \hat{p}(\cdot|\pi^t, \mathcal{D})} \left[0.125 \gamma \sum_{h=1}^H \sum_{k=1}^K \mathbb{E}_{\mu^k} \left[\{(f_h - T_h^{\pi^t} f_{h+1})(s_h, a_h)\}^2 \right] + \lambda \left(f_1^{\pi^t}(s_1) - V_1^{\pi^t}(s_1) \right) \right]}{\frac{\lambda}{\sqrt{K}}} + \frac{0.5 \lambda H \cdot \mathcal{C}(\pi; \epsilon)}{K \gamma} + H \epsilon \\ &\leq \frac{1}{\sqrt{K}} + \frac{b^2 \gamma H \cdot \text{COMP}}{\lambda} + \frac{0.5 \lambda H \cdot \mathcal{C}(\pi; \epsilon)}{K \gamma} + H \epsilon \\ &\leq \frac{1}{\sqrt{K}} + b H \frac{\sqrt{\mathcal{C}(\pi; \epsilon) \cdot \text{COMP}}}{\sqrt{K}} + H \epsilon \end{aligned}$$

Minimize LHS wrt $\lambda > 0$

Summary

- Provide a **data diversity** notion to measure distribution shift in offline RL
- Unified theory that shows **version space**, **regularized optimization** and **posterior sampling** achieve *comparable* guarantees

Other Results in the Paper

- Decoupling argument for offline RL
- New techniques to handle data dependence in the analysis of the worst-case bound of posterior sampling
 - E.g. directly apply to resolve a technical mistake in Xiong et al. 2022. “A self-play posterior sampling algorithm for zero-sum Markov games”

Thank You