Robust Constrained Offline Reinforcement Learning with Linear Function Approximation

Carnegie Mellon University

上海科技大学 ShanghaiTech University

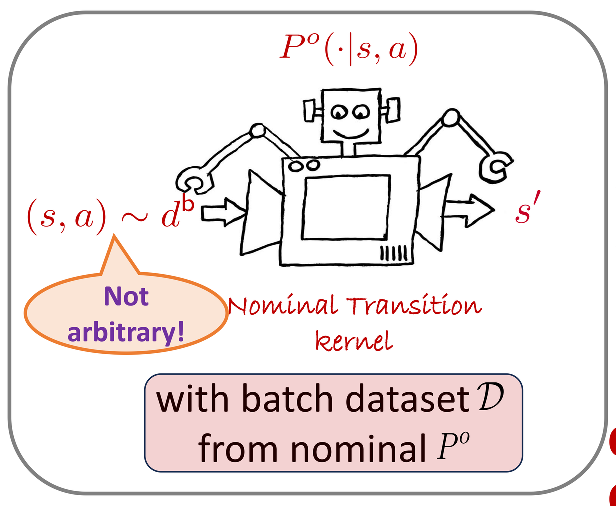
wangwb2023@shanghaitech.edu.cn, hew2@andrew.cmu.edu

Find a PhD Position

Our paper is here!

Motivation

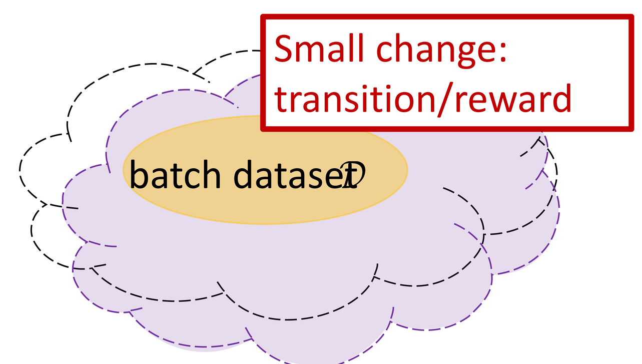
Offline RL: Learning a good policy from batch data



Standard Constrained RL: Learn the optimal policy of the nominal MDP under constraints?

Robust Constrained RL: Learn the robust and safe policy around the nominal MDP?

Challenge I: sim-to-real gaps



Challenge II: safety constraint

Challenge III: sample complexity blows up for large state space

*Can we design a **sample-efficient** algorithm that is **robust to the sim-to-real gap and ensures constraint satisfcation,** even for **large state space?***

Problem Formulation

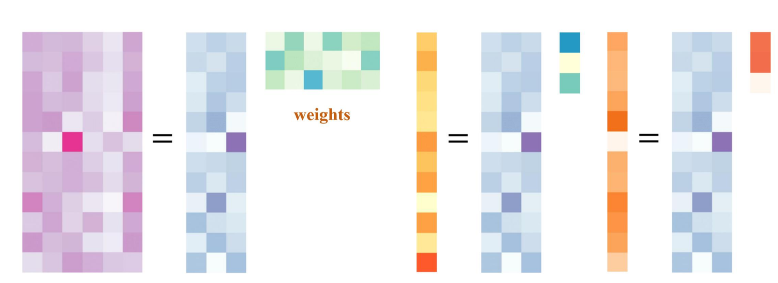
Lin-RCMDPs: distributionally robust linear CMDPs

$$\mathcal{M}_{\text{rob}}=(\mathcal{S},\mathcal{A},H,\mathcal{P}^\rho(P^0),r,g)$$

We use the uncertainty set around the nominal transition kernel to characterize the sim-to-real gap.

- **Linear representations**: The reward function and nominal transition kernel are decomposed as $r_h=\phi(s,a)^\top\theta_{r,h},\ g_h=\phi(s,a)^\top\theta_{g,h},$
  - $\phi(s,a)\in\mathbb{R}^d$: feature mapping $\quad P_h(s'\mid s,a)=\phi(s,a)^\top\mu_h^P(s')$

Note that the number of features d is much smaller than the size of state space.

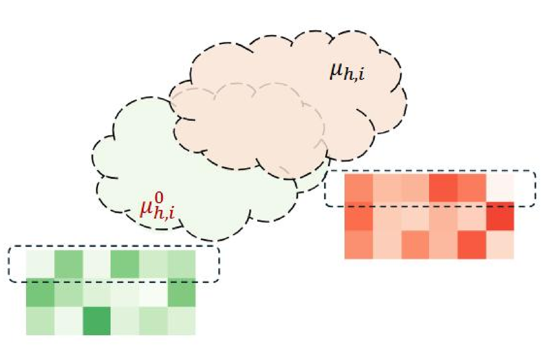


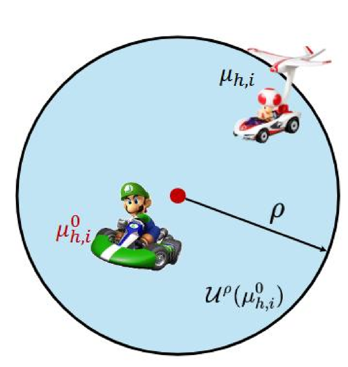
- **The uncertainty set satisfies d-rectangularity assumption:**

$$\mathcal{P}^\rho(P^0)=\left\{\phi(s,a)^\top\mu_h(\cdot):\mu_{h,i}\in\mathcal{U}^\rho(\mu^0_{h,i}),\forall(i,s,a,h)\in[d]\times\mathcal{S}\times\mathcal{A}\times[H]\right\}$$

  - **Decoupling the distribution shift into each feature dimension**.

$$\mathcal{U}^\rho(\mu^0_{h,i}):=\left\{\mu_{h,i}:\frac{1}{2}\|\mu_{h,i}-\mu^0_{h,i}\|\le\rho\ \text{and}\ \mu_{h,i}\in\Delta(\mathcal{S})\right\},\ \forall i\in[d]$$





- **Robust value/Q function:** measure accumulative rewards in the **worst case** of performing in the transition kernel inside the uncertainty set.

$$V^{\pi,\rho}_{r/g,h}(s)=\inf_{P\in\mathcal{P}^\rho(P^0)}V^{\pi,P}_{r/g,h}(s)$$

$$Q^{\pi,\rho}_{r/g,h}(s,a)=\inf_{P\in\mathcal{P}^\rho(P^0)}Q^{\pi,P}_{r/g,h}(s,a)$$

- **Learning goal:** Given the dataset $\mathcal{D}=\{(s_h^\tau,a_h^\tau,r_h^\tau,g_h^\tau,s_{h+1}^\tau)\}_{h\in[H],\tau\in[K]}$ from the nominal environment, find an ϵ —robust policy $\widehat{\pi}$ such that

Sub-optimality gap: $\quad V^{\star,\rho}_{r,1}-V^{\widehat{\pi},\rho}_{r,1}\le\epsilon,\ b-V^{\widehat{\pi},\rho}_{g,1}\le\epsilon$

Comparison with the most relevant work

	State Representation	Blanchet et al. (2024)	Wang et al. (2024)	Ghosh (2024)	This Work
unconstraint	$S\times A$-rectangular (Tabular)	✓	✓	✓	✓
	d-rectangular (Linear)	✓	✓	✗	✓
constraint	$S\times A$-rectangular (Tabular)	✗	✗	✓!	✓
	d-rectangular (Linear)	✗	✗	✗	✓

Table 1: Comparison with the most relevant works in robust RL. ✓indicates that the work is capable of addressing the model with robust partial coverage data, ✓! signifies that the work requires full coverage data to solve the model, and ✗ denotes that the work is not applicable to the model. Light green highlights the models that are either introduced or proven to be tractable in this work.

Performance Guarantees for DROP

- **Arbitrary: without any data coverage assumption**

Theorem 2 (minimal offline data assumption)

Consider any d-rectangular Lin-RCMDP, where the uncertainty is measured by TV distance. With high probability, the policy $\widehat{\pi}$ generated by CROP-VI satisfies

$$V_1^{\star,\rho}-V_1^{\widehat{\pi},\rho}\le\widetilde{O}(dH^2)\max_{P\in\mathcal{P}^\rho(P^0)}\mathbb{E}_{\pi^\star,P}\left[\|\phi_i(s_h,a_h)\mathbb{1}_i\|_{\Lambda_h^{-1}}\right],\quad b-V_{g,1}^{\widehat{\pi},\rho}(\zeta)\le\varepsilon$$

Instance-dependent sub-optimality gap ← depending on the batch data quality

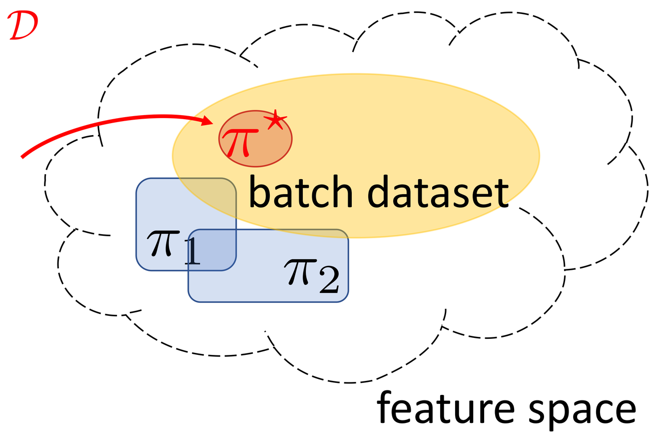
- **Partial feature coverage**
  - **Assumption:** robust single-policy clipped concentrability

$\pi^\star$ occupancy distribution; clipping operation

$$\max_{u\in\mathbb{R}^d,h\in[H],i\in d,\ P\in\mathcal{P}^\rho(P^0)}\frac{u^\top\left(\min\{\mathbb{E}_{d_h^{\star,P}}\phi_i^2(s,a),1/d\}\cdot\mathbb{1}_{i,i}\right)u}{u^\top\left(\mathbb{E}_{d_h^b}[\phi(s,a)\phi(s,a)^\top]\right)u}\le\frac{C^\star_{\text{rob}}}{d}$$

distribution of dataset $\mathcal{D}$

$C^\star_{\text{rob}}<\infty$

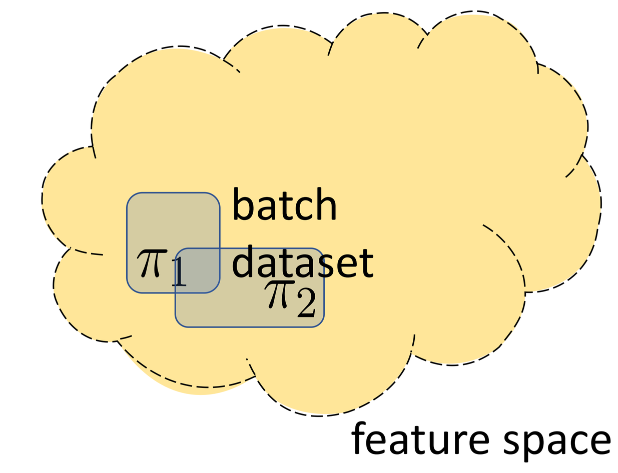


- **Our results:** to achieve the ϵ-sub-optimality gap, , CROP-VI needs at most $\widetilde{O}(C^\star_{\text{rob}}d^2H^4/\epsilon^2)$ samples.

- **Full feature coverage**
  - **Assumption:** $\kappa=\min_{h\in[H]}\lambda_{\min}\left(\mathbb{E}_{d_h^b}[\phi(s,a)\phi(s,a)^\top]\right)>0$

Samples can explore the feature space uniformly well.

  - **Our results:** to achieve the ϵ-sub-optimality gap, CROP-VI needs at most $\widetilde{O}(\frac{d^2H^4}{\kappa\epsilon^2})$ samples.



CROP-VI : Constrained Robust Optimistic-Pessimistic Value Iteration

Input: batch dataset $\mathcal{D}$, feature mappings, parameters $\lambda_0,\ \gamma_0$ → *Two-fold subsampling* → **Preparation:** construct a **temporally independent** dataset → **Initialization:** $\widehat{Q}_{H+1}(\cdot,\cdot)=0\quad\widehat{V}_{H+1}(\cdot)=0$

For $h=H,\ H-1,\ldots,\ 1$
- Step1: construct a rectified dual form for the original constrained problem
- Step1: construct an empirical Bellman operator for Lin-RCMDPs
- Step 2: plan by pessimistic value and optimistic constraint iterations

- **Step 1: construct a rectified dual form for the original constrained problem**

$$\max_\pi V^{\pi,\rho}_{r,1}(s)\quad \text{s.t.}\quad V^{\pi,\rho}_{g,1}(s)\ge b\quad(1)\qquad\Longleftrightarrow\qquad\max_\pi V^{\pi,\rho}_{r,1}(s)-\beta\left(b-V^{\pi,\rho}_{g,1}(s)\right)_+\quad(2)$$

Given $\varepsilon>0$, setting $\beta=H/\epsilon$ ensures that the optimal solution $\widehat{\pi}$ of (2) incurs a constraint violation of at most ϵ, i.e., $(b-V^{\widehat{\pi},\rho}_{g,1}(\zeta))\le\epsilon$. Consequently, with $\beta=H/\epsilon$, if for any infeasible policy π such that $V^{\pi,\rho}_{g,1}(\zeta)-b<\epsilon$, then $\pi^\star$ (i.e., the optimal solution of (1)) is also optimal for (2).

- **Step 2: construct an emperical Bellman operator for Lin-RCMDPS**

Original robust Bellman operator: by strong duality, we have for $j=r,g$

$$(\mathbb{B}^\rho_{j,h}V_{r/g})(s,a)=\phi(s,a)^\top(\theta_{j,h}+\nu_h^{\rho,V_j})$$

the i-th coordinate

$$\nu_{h,i}^{\rho,V_j}:=\max_{\alpha\in[\min_sV_j(s),\max_sV_j(s)]}\left\{\mathbb{E}_{s\sim\mu^0_{h,i}}[V_j]_\alpha(s)-\rho\left(\alpha-\min_{s'}[V_j]_\alpha(s')\right)\right\}\quad\text{with}\ [V_j]_\alpha(s)=\min\{V_j(s),\alpha\}$$

However, we cannot directly have access to the ground-truth $\theta_{j,h}$ and μ_h^0.

Empirical robust Bellman operator: approximate by ridge regression

$$\theta_{j,h}\approx\arg\min_{\theta\in\mathbb{R}^d}\sum_{\tau\in\mathcal{D}_h^0}\left(\phi(s_h^\tau,a_h^\tau)^\top\theta_j-j_h^\tau\right)^2+\lambda_0\|\theta\|_2^2$$

$$\mathbb{E}_{s\sim\mu^0_{h,i}}[V_j]_\alpha(s)\approx\arg\min_{\nu\in\mathbb{R}^d}\sum_{\tau\in\mathcal{D}_h^0}\left(\phi(s_h^\tau,a_h^\tau)^\top\nu-[V_j]_\alpha(s_{h+1}^\tau)\right)^2+\lambda_0\|\nu\|_2^2$$

- **Step 2: plan by pessimistic value iterations**

Following the pessimistic principle, we then estimate the reward Q-function as

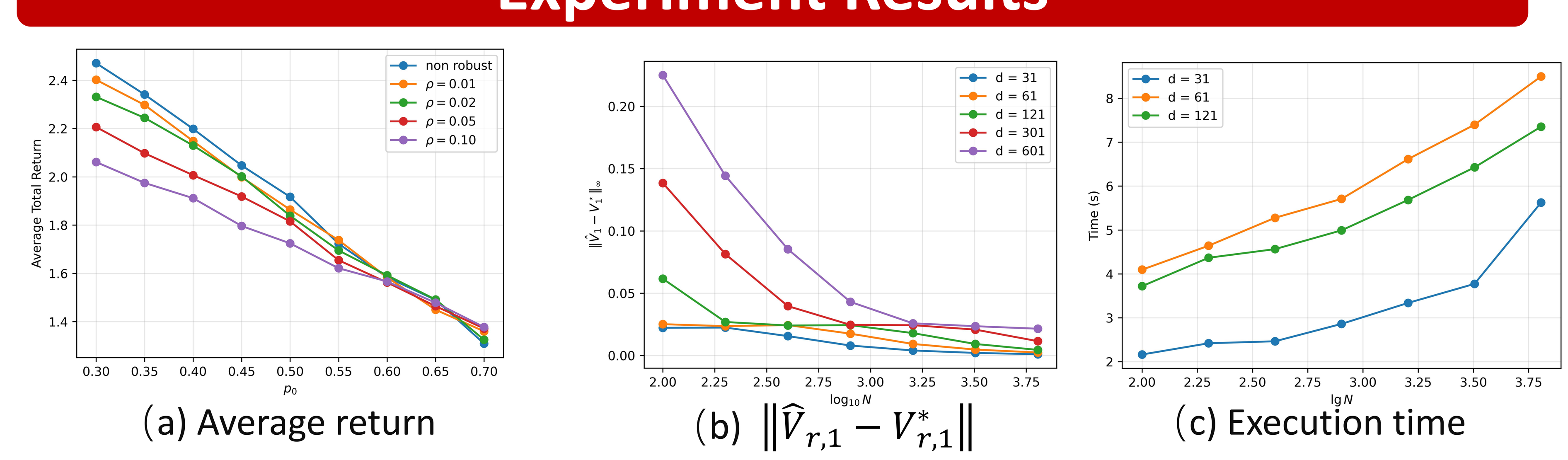
$$\bar{Q}_{r,h}(s,a)=\left(\widehat{\mathbb{B}}^\rho_{r,h}\widehat{V}_{r,h+1}\right)(s,a)-\gamma_0\sum_{i=1}^d\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^i}$$

Penalty function: address uncertainty in each dimension

$$\bar{Q}_{g,h}(s,a)=\left(\widehat{\mathbb{B}}^\rho_{g,h}\widehat{V}_{g,h+1}\right)(s,a)+\gamma_0\sum_{i=1}^d\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^i}$$

Incentive function: balance exploration and exploitation trade-off

Experiment Results



(a) Average return $\quad$ (b) $\|\widehat{V}_{r,1}-V^\star_{r,1}\|$ $\quad$ (c) Execution time