Robust Offline Reinforcement Learning with Linearly Structured Regularization and f-Divergence

Speaker: Cheng Tang

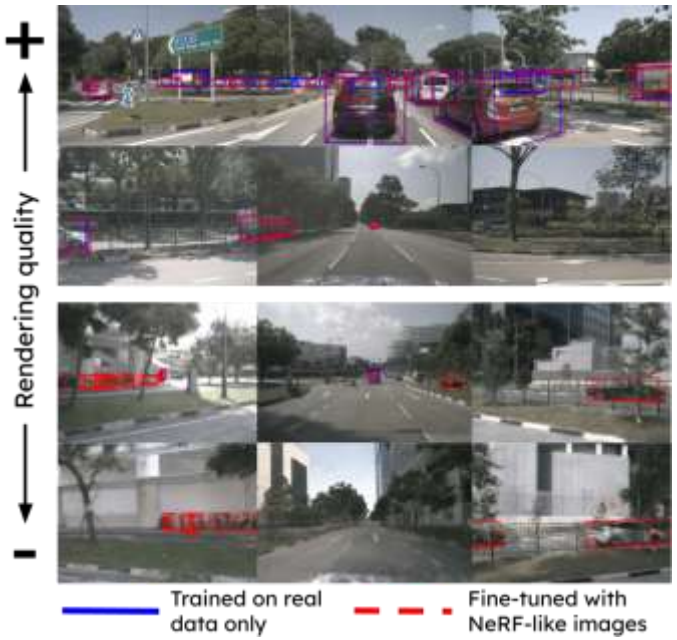
Nov 23th, 2024

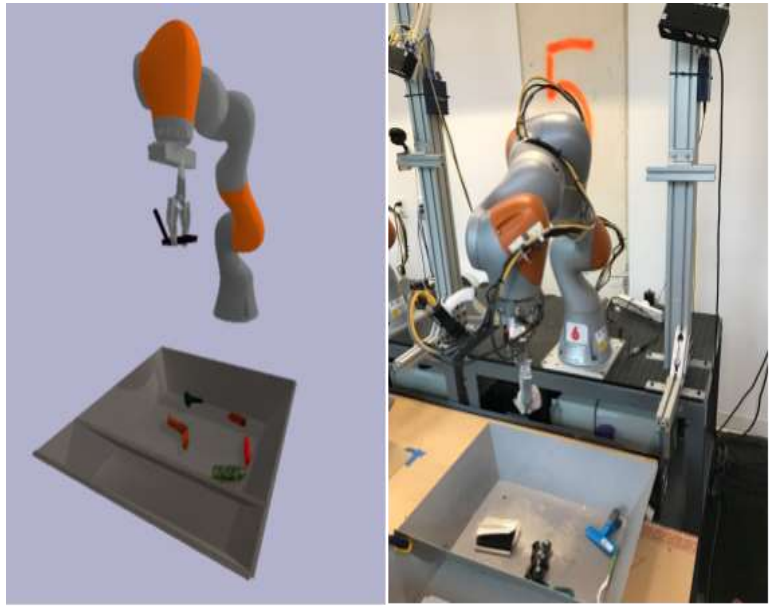
Tsinghua University

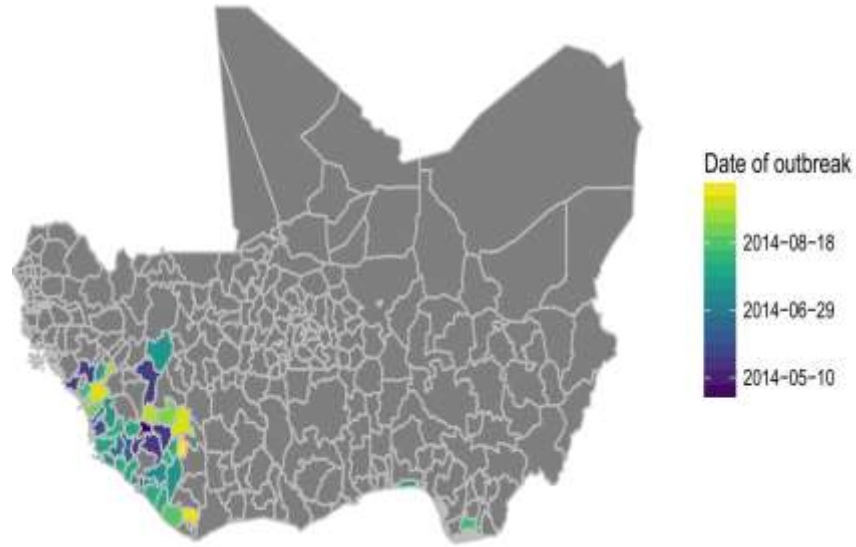
Content

Introduction

■ Sim-to-real gap:







- **Autonomous Driving**
- **Robotic Learning**
- **Disease Control**

[1] Lindström C, Hess G, Lilja A, et al. Are NeRFs ready for autonomous driving? Towards closing the real-to-simulation gap[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 4461-4471.
[2] Bousmalis K, Levine S. Closing the simulation-to-reality gap for deep robotic learning[J]. Google Research Blog, 2017, 1.
[3] Liu Z, Clifton J, Laber E B, et al. Deep spatial q-learning for infectious disease control[J]. Journal of Agricultural, Biological and Environmental Statistics, 2023, 28(4): 749-773.

Introduction

- **Distributionally Robust RL：learn more robust policy through Reinforcement Learning**
  - **d-rectangular DRMDP: model the uncertainty in the dynamics and aim to achieve the best performance under the most adversarial dynamics.**

Value function

$$V_h^{\pi,P}(s) := \mathbb{E}^P\left[\sum_{t=h}^{H} r_t(s_t,a_t)\middle| s_h = s, \pi\right],$$
$$Q_h^{\pi,P}(s,a) := \mathbb{E}^P\left[\sum_{t=h}^{H} r_t(s_t,a_t)\middle| s_h = s, a_h = a, \pi\right]$$

+

Uncertainty set

$$\mathcal{U}_{h,i}^{\rho}(\mu_{h,i}^0) = \{\mu : \mu \in \Delta(\mathcal{S}), D(\mu||\mu_{h,i}^0) \le \rho\}.$$
$$\mathcal{U}_h^{\rho}(P_h^0) = \bigotimes_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{U}_h^{\rho}(s,a;\boldsymbol{\mu}_h^0),$$

=

Robust value function

$$V_h^{\pi,\rho}(s) = \inf_{P\in\mathcal{U}^{\rho}(P^0)} V_h^{\pi,P}(s), \quad \forall (h,s) \in [H]\times\mathcal{S}.$$

- **The robust value function can be seen as the worst value function in an uncertainty set determined by probability divergence *D***

Motivation

- **Drawbacks of d-rectangular DRMDP:**
  - Needs strong assumption on dual variables
  - Existing algorithms rely on approximation to substitute the duality optimization, making it time consuming
  - Existing work consider mainly TV divergence geometry, leaving blanks for cases with KL and χ^2
- **RRMDP: applying regularization penalty term measuring the uncertainty**
  - Related work shows that the robust policy under RRMDP can be equivalent to DRMDP
  - The forfeit of uncertainty set constraint makes the dual problem easier, leading to potential improvement on computation efficiency and theoretical analysis

Content

Problem Formulation

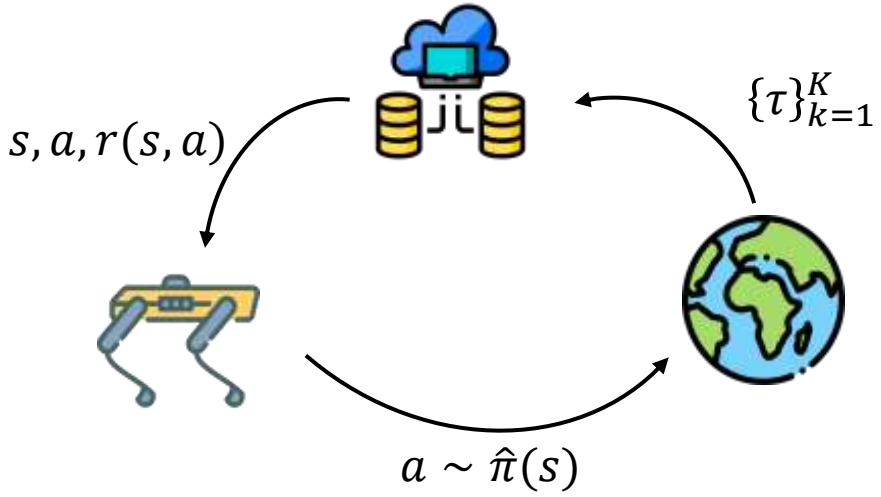
■ Offline MDP (Markov Decision Process): $\mathrm{MDP}(\mathrm{S}, \mathrm{A}, \mathrm{H}, \mathrm{P}^0, \mathrm{r})$

- **State s, action a, reward $r(s,a) \in [0,1]$ (known), nominal kernel $P^0 = \{P_h^0\}_{h=1}^H$,**
- **Value function and Q-function:**

$$V_h^\pi(s) = \mathbb{E}^{P^0}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \middle| s_h = s, \pi\right], \qquad Q_h^\pi(s,a) = \mathbb{E}^{P^0}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \middle| s_h = s, a_h = a, \pi\right].$$

- **Offline dataset and Learning goal: given *K* trajectory $\{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{h=1}^H$ and find policy $\hat{\pi}$ to minimize the**

Suboptimality gap: $\mathrm{SubOpt}(\pi; x) = V_1^{\pi^*}(x) - V_1^\pi(x),$



Problem Formulation

- **RRMDP (Robust Regularized Markov Decision Process):** $\mathrm{RRMDP}(\mathrm{S}, \mathrm{A}, \mathrm{H}, \mathrm{P}^0, \mathrm{r}, \lambda, D, \mathrm{F})$
  - **Regularized robust parameter λ, probability divergence D, feasible set of all perturbed transition kernels** F
  - **Regularized robust value function and Q-function:**

$$V_h^{\pi,\lambda}(s) = \inf_{P\in\mathcal{F}} \mathbb{E}^P\left[\sum_{t=h}^{H}\left[r_t(s_t,a_t) + \lambda D(P_t(|s_t,a_t)\|P_t^0(\cdot|s_t,a_t))\right] \middle| s_h = s, \pi\right],$$

$$Q_h^{\pi,\lambda}(s,a) = \inf_{P\in\mathcal{F}} \mathbb{E}^P\left[\sum_{t=h}^{H}\left[r_t(s_t,a_t) + \lambda D(P_t(\cdot|s_t,a_t)\|P_t^0(\cdot|s_t,a_t))\right] \middle| s_h = s, a_h = a, \pi\right].$$

Penalty on divergence with nominal kernel

  - **Offline dataset and Learning goal: given *K* trajectory** $\{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{h=1}^H$ **and find policy** $\hat{\pi}$ **to minimize the robust Suboptimality gap:**

$$\mathrm{SubOpt}(\hat{\pi}, s_1, \lambda) := V_1^{\star,\lambda}(s_1) - V_1^{\hat{\pi},\lambda}(s_1).$$

Optimal robust regularized value function

RRMDP under linear function class

■ Linear MDP (Markov Decision Process):

- **Known feature mapping** $\phi: s \times a \to R^d$, $\sum_i \phi_i(s,a) = 1$, $\phi_i(s,a) \geq 0$
- **Linear reward function and nominal transition kernel class F**

$$r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \rangle, \; P_h^0(\cdot|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h^0(\cdot) \rangle$$

where $\{\boldsymbol{\theta}_h\}_{h=1}^H$ are known vectors with bounded norm $\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d}$ and $\{\boldsymbol{\mu}_h^0\}_{h=1}^H$ are unknown probability measure vectors over $\mathcal{S}$, i.e., $\boldsymbol{\mu}_h^0 = (\mu_{h,1}^0, \mu_{h,2}^0, \cdots, \mu_{h,d}^0)$, $\mu_{h,i}^0 \in \Delta(\mathcal{S}), \forall i \in [d]$.

■ Offline d-rectangular linear RRMDP (d-RRMDP)

- **Regularized robust value function and Q-function (under linear setting):**

$$V_h^{\pi,\lambda}(s) = \inf_{\boldsymbol{\mu}_t \in \Delta(\mathcal{S})^d, P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t \rangle} \mathbb{E}^{\{P_t\}_{t=h}^H} \left[\sum_{t=h}^H \left[r_t(s_t,a_t) + \lambda \langle \boldsymbol{\phi}(s_t,a_t), \boldsymbol{D}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^0) \rangle \right] \Big| s_h = s, \pi \right],$$

$$Q_h^{\pi,\lambda}(s,a) = \inf_{\boldsymbol{\mu}_t \in \Delta(\mathcal{S})^d, P_t = \langle \boldsymbol{\phi}, \boldsymbol{\mu}_t \rangle} \mathbb{E}^{\{P_t\}_{t=h}^H} \left[\sum_{t=h}^H \left[r_t(s_t,a_t) + \lambda \langle \boldsymbol{\phi}(s_t,a_t), \boldsymbol{D}(\boldsymbol{\mu}_t \| \boldsymbol{\mu}_t^0) \rangle \right] \Big| s_h = s, a_h = a, \pi \right]$$

- **Optimal robust regularized value function and Q-function:**

$$V_h^{\star,\lambda}(s) = \sup_\pi V_h^{\pi,\lambda}(s), Q_h^{\star,\lambda}(s,a) = \sup_\pi Q_h^{\pi,\lambda}(s,a).$$

Dynamic programming principles

■ Robust regularized Bellman Equation:

$$Q_h^{\pi,\lambda}(s,a) = r_h(s,a) + \inf_{\mu_h \in \Delta(\mathcal{S})^d, P_h = \langle \phi, \mu_h \rangle} \left[\mathbb{E}_{s' \sim P_h(\cdot|s,a)} \left[V_{h+1}^{\pi,\lambda}(s') \right] + \lambda \langle \boldsymbol{\phi}(s,a), \boldsymbol{D}(\boldsymbol{\mu}_h \| \boldsymbol{\mu}_h^0) \rangle \right],$$

$$V_h^{\pi,\lambda}(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[Q_h^{\pi,\lambda}(s,a) \right].$$

- This preposition shows the recurrence relationship of regularized robust value function

■ Existence of optimal policy

Proposition 3.3. Under the setting of d-rectangular linear RRMDP, there exists a deterministic and stationary policy $\pi^\star$, such that for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$V_h^{\pi^\star,\lambda}(s) = V_h^{\star,\lambda}(s), Q_h^{\pi^\star,\lambda}(s,a) = Q_h^{\star,\lambda}(s,a). \tag{3.7}$$

- The existence of optimal policy guarantees the solvable of d-RRMDP with greedy policy
- The closeness of linear function class ensures the establishment of the above propositions

Content

Method - Framework

Linearity of Q-function

Proposition 4.1. Under Assumption 3.1, for any $(\pi, s, a, h) \in \Pi \times \mathcal{S} \times \mathcal{A} \times [H]$, we have

$$Q_h^{\pi,\lambda}(s,a) = \langle \phi(s,a), \theta_h + w_h^{\pi,\lambda} \rangle, \tag{4.1}$$

where $w_h^{\pi,\lambda} = \left(w_{h,1}^{\pi,\lambda}, w_{h,2}^{\pi,\lambda}, \cdots, w_{h,d}^{\pi,\lambda}\right)^\top \in \mathbb{R}^d$, and $w_{h,i}^{\pi,\lambda} = \inf_{\mu_{h,i} \in \Delta(s)} \left[\mathbb{E}^{\mu_{h,i}}\left[V_{h+1}^{\pi,\lambda}(s)\right] + \lambda D(\mu_{h,i} \| \mu_{h,i}^0)\right]$.

Pessimism based algorithm

- **Step 1**: estimate w_h^λ by solving dual problem
- **Step 2**: construct pessimism penalty $\Gamma_h(\cdot,\cdot)$
- **Step 3**: compute pessimistic Q-function

Algorithm 1 Robust Regularized Pessimistic Value Iteration (R2PVI)

Require: Dataset $\mathcal{D}$, Regularizer $\lambda > 0$

1: init $\hat{V}_{H+1}^\lambda(\cdot) = 0$
2: **for** episode $h = H, \cdots, 1$ **do**
3: Compute $\Lambda_h \leftarrow \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \gamma \mathbf{I}$
4: Obtain the parameter estimation $\hat{w}_h^\lambda$. ◁ Duality Estimation
TV-divergence: use (4.3)
KL-divergence: use (4.5)
χ^2-divergence: use (4.9)
5: Construct the pessimism penalty $\Gamma_h(\cdot,\cdot)$. ◁ Pessimism
6: Estimate $\hat{Q}_h^\lambda(\cdot,\cdot) \leftarrow \min(\langle \phi(\cdot,\cdot), \theta_h + \hat{w}_h^\lambda \rangle - \Gamma_h(\cdot,\cdot), H - h + 1)^+$.
7: Construct $\hat{\pi}_h(\cdot|\cdot) \leftarrow \operatorname{argmax}_{\pi_h} \langle \hat{Q}_h^\lambda(\cdot,\cdot), \hat{\pi}_h(\cdot|\cdot) \rangle_{\mathcal{A}}$ and $\hat{V}_h^\lambda(\cdot) \leftarrow \langle \hat{Q}_h^\lambda(\cdot,\cdot), \hat{\pi}_h(\cdot|\cdot) \rangle_{\mathcal{A}}$.
8: **end for**

Method – R2PVI-TV

■ Dual form of TV

$$\inf_{\mu\in\Delta(S)} \mathbb{E}_{s\sim\mu} V(s) + \lambda D_{\mathrm{TV}}(\mu\|\mu^0) = \mathbb{E}_{s\sim\mu^0}[V(s)]_{\min_{s'}(V(s'))+\lambda\cdot}$$

Close form solution

■ Specific applicable algorithm

- **Obtain by least square regression:**

$$\hat{\boldsymbol{w}}_h^\lambda = \operatorname*{argmin}_{\boldsymbol{w}\in R^d} \sum_{\tau=1}^{K} \left([\hat{V}_{h+1}^\lambda(s_{h+1}^\tau)]_{\alpha_{h+1}} - \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top \boldsymbol{w}\right)^2 + \gamma\|\boldsymbol{w}\|_2^2,$$

- **Specifically designed penalty with $\boldsymbol{\beta_{TV}}$**

Algorithm 2 Robust Regularized Pessimistic Value Iteration under TV distance (R2PVI-TV)

Require: Dataset $\mathcal{D}$, regularizer $\lambda > 0$, $\gamma > 0$ and parameter β

1: init $\hat{V}_{H+1}^\lambda(\cdot) = 0$
2: **for** episode $h = H, \cdots, 1$ **do**
3: $\quad \boldsymbol{\Lambda}_h \leftarrow \sum_{\tau=1}^K \boldsymbol{\phi}(s_h^\tau, a_h^\tau)(\boldsymbol{\phi}(s_h^\tau, a_h^\tau))^\top + \gamma \mathbf{I}$
4: $\quad \alpha_{h+1} \leftarrow \min_{s\in\mathcal{S}}(\hat{V}_{h+1}^\lambda(s)) + \lambda$
5: $\quad \hat{\boldsymbol{w}}_h^\lambda \leftarrow \boldsymbol{\Lambda}_h^{-1}(\sum_{\tau=1}^K \boldsymbol{\phi}(s_h^\tau, a_h^\tau)[\hat{V}_{h+1}^\lambda(s_{h+1}^\tau)]_{\alpha_{h+1}})$
6: $\quad \Gamma_h(\cdot,\cdot) \leftarrow \beta \sum_{i=1}^d \|\phi_i(\cdot,\cdot)\mathbf{1}_i\|_{\boldsymbol{\Lambda}_h^{-1}}$
7: $\quad \hat{Q}_h^\lambda(\cdot,\cdot) \leftarrow \min(\boldsymbol{\phi}(\cdot,\cdot)^\top(\boldsymbol{\theta}_h + \hat{\boldsymbol{w}}_h^\lambda) - \Gamma_h(\cdot,\cdot), H-h+1)^+$
8: $\quad \hat{\pi}_h(\cdot|\cdot) \leftarrow \operatorname{argmax}_{\pi_h}\langle \hat{Q}_h^\lambda(\cdot,\cdot), \pi_h(\cdot|\cdot)\rangle_{\mathcal{A}}$ and $\hat{V}_h^\lambda(\cdot) \leftarrow \langle \hat{Q}_h^\lambda(\cdot,\cdot), \hat{\pi}_h(\cdot|\cdot)\rangle_{\mathcal{A}}$
9: **end for**

Method – R2PVI-KL

■ Dual form of KL

$$\inf_{\mu\in\Delta(S)} \mathbb{E}_{s\sim\mu} V(s) + \lambda D_{\mathrm{KL}}(\mu\|\mu^0) = -\lambda \log \mathbb{E}_{s\sim\mu^0}\left[e^{-\frac{V(s)}{\lambda}}\right].$$

Logarithm may generate error

■ Specific applicable algorithm

- **Obtain by least square regression:**

$$\hat{w}'_h = \operatorname*{argmin}_{w\in\mathbb{R}^d} \sum_{\tau=1}^{K} \left(e^{-\frac{\hat{V}^{\lambda}_{h+1}(s^{\tau}_{h+1})}{\lambda}} - \phi(s^{\tau}_h, a^{\tau}_h)^{\top} w\right)^2 + \gamma\|w\|_2^2.$$

- **Clip the $\hat{w}'_h$ with lower bound of $\mathbb{E}_{s\sim\mu^0} e^{-\hat{V}^{\lambda}_{h+1}(s)/\lambda}$**

- **Specifically designed penalty with β_{KL}**

Algorithm 3 Robust Regularized Pessimistic Value Iteration under KL distance (R2PVI-KL)

Require: Dataset $\mathcal{D}$, regularizer $\lambda > 0, \gamma > 0$ and parameter β

1: init $\hat{V}^{\lambda}_{H+1}(\cdot) = 0$
2: **for** episode $h = H, \cdots, 1$ **do**
3: $\quad \Lambda_h \leftarrow \sum_{\tau=1}^{K} \phi(s^{\tau}_h, a^{\tau}_h)(\phi(s^{\tau}_h, a^{\tau}_h))^{\top} + \gamma \mathbf{I}$
4: $\quad \hat{w}'_h \leftarrow \Lambda_h^{-1}\left(\sum_{\tau=1}^{K} \phi(s^{\tau}_h, a^{\tau}_h) e^{-\frac{\hat{V}^{\lambda}_{h+1}(s^{\tau}_{h+1})}{\lambda}}\right)$
5: $\quad \hat{w}^{\lambda}_h \leftarrow -\lambda \log \max\{\hat{w}'_h, e^{-H/\lambda}\}$
6: $\quad \Gamma_h(\cdot,\cdot) \leftarrow \beta \sum_{i=1}^{d} \|\phi_i(\cdot,\cdot)\mathbf{1}_i\|_{\Lambda_h^{-1}}$
7: $\quad \hat{Q}^{\lambda}_h(\cdot,\cdot) \leftarrow \min(\phi(\cdot,\cdot)^{\top}(\theta_h + \hat{w}'_h) - \Gamma_h(\cdot,\cdot), H - h + 1)^{+}$
8: $\quad \hat{\pi}_h(\cdot|\cdot) \leftarrow \operatorname{argmax}_{\pi_h} \langle \hat{Q}^{\lambda}_h(\cdot,\cdot), \pi_h(\cdot|\cdot)\rangle_{\mathcal{A}}$ and $\hat{V}^{\lambda}_h(\cdot) \leftarrow \langle \hat{Q}^{\lambda}_h(\cdot,\cdot), \hat{\pi}_h(\cdot|\cdot)\rangle_{\mathcal{A}}$
9: **end for**

Method – R2PVI-χ^2

■ Dual form of χ^2

$$\inf_{\mu\in\Delta(S)} \mathbb{E}_{s\sim\mu} V(s) + \lambda D_{\chi^2}(\mu\|\mu^0) = \sup_{\alpha\in[V_{\min},V_{\max}]} \left\{\mathbb{E}_{s\sim\mu^0}[V(s)]_\alpha - \frac{1}{4\lambda}\mathrm{Var}_{s\sim\mu^0}[V(s)]_\alpha\right\}$$

Require to solve max problem

■ Specific applicable algorithm

- estimate w_h^λ by solving dual problem

$$\hat{\mathbb{E}}^{\mu^0_{h,i}}[\hat{V}^\lambda_{h+1}(s)]_\alpha = \left[\operatorname*{argmin}_{w\in R^d} \sum_{\tau=1}^K ([\hat{V}^\lambda_{h+1}(s^\tau_{h+1})]_\alpha - \phi(s^\tau_h,a^\tau_h)^\top w)^2 + \gamma\|w\|_2^2\right]^i_{[0,H]}$$

$$\hat{\mathbb{E}}^{\mu^0_{h,i}}[\hat{V}^\lambda_{h+1}(s)]^2_\alpha = \left[\operatorname*{argmin}_{w\in R^d} \sum_{\tau=1}^K ([\hat{V}^\lambda_{h+1}(s^\tau_{h+1})]^2_\alpha - \phi(s^\tau_h,a^\tau_h)^\top w)^2 + \gamma\|w\|_2^2\right]^i_{[0,H^2]}$$

$$\hat{w}^\lambda_{h,i} = \sup_{\alpha\in[(\hat{V}^\lambda_{h+1})_{\min},(\hat{V}^\lambda_{h+1})_{\max}]} \left\{\hat{\mathbb{E}}^{\mu^0_{h,i}}[\hat{V}^\lambda_{h+1}(s)]_\alpha - \frac{1}{4\lambda}\widehat{\mathrm{Var}}^{\mu^0_{h,i}}[\hat{V}^\lambda_{h+1}(s)]_\alpha\right\}$$

$$= \max_{\alpha\in[(\hat{V}^\lambda_{h+1})_{\min},(\hat{V}^\lambda_{h+1})_{\max}]} \left\{\hat{\mathbb{E}}^{\mu^0_{h,i}}[\hat{V}^\lambda_{h+1}(s)]_\alpha + \frac{1}{4\lambda}(\hat{\mathbb{E}}^{\mu^0_{h,i}}[\hat{V}^\lambda_{h+1}(s)]_\alpha)^2 - \frac{1}{4\lambda}\hat{\mathbb{E}}^{\mu^0_{h,i}}[\hat{V}^\lambda_{h+1}(s)]^2_\alpha\right\}.$$

Algorithm 4 Robust Regularized Pessimistic Value Iteration under χ^2 distance (R2PVI-χ^2)

Require: Dataset $\mathcal{D}$, regularizer $\lambda>0, \gamma>0$ and parameter β

1: init $\hat{V}^\lambda_{H+1}(\cdot)=0$
2: **for** episode $h=H,\cdots,1$ **do**
3: $\quad\Lambda_h \leftarrow \sum_{\tau=1}^K \phi(s^\tau_h,a^\tau_h)(\phi(s^\tau_h,a^\tau_h))^\top + \gamma\mathbf{I}$
4: $\quad\hat{\mathbb{E}}^{\mu^0_{h,i}}[\hat{V}^\lambda_{h+1}(s)]_\alpha \leftarrow [\Lambda_h^{-1}(\sum_{\tau=1}^K \phi(s^\tau_h,a^\tau_h)^\top [\hat{V}^\lambda_{h+1}(s^\tau_{h+1})]_\alpha)]_{[0,H]}$ ◁ Estimated by (4.7)
5: $\quad\hat{\mathbb{E}}^{\mu^0_{h,i}}[\hat{V}^\lambda_{h+1}(s)]^2_\alpha \leftarrow [\Lambda_h^{-1}(\sum_{\tau=1}^K \phi(s^\tau_h,a^\tau_h)^\top [\hat{V}^\lambda_{h+1}(s^\tau_{h+1})]^2_\alpha)]_{[0,H^2]}$ ◁ Estimated by (4.8)
6: $\quad$Estimate $\hat{w}^\lambda_{h,i}$ according to (4.9)
7: $\quad\Gamma_h(\cdot,\cdot) \leftarrow \beta\sum_{i=1}^d \|\phi_i(\cdot,\cdot)\|_{\Lambda_h^{-1}}$
8: $\quad\hat{Q}^\lambda_h(\cdot,\cdot) \leftarrow \min(\phi(\cdot,\cdot)^\top(\theta_h+\hat{w}^\lambda_h) - \Gamma_h(\cdot,\cdot), H-h+1)^+$
9: $\quad\hat{\pi}_h(\cdot|\cdot) \leftarrow \mathrm{argmax}_{\pi_h}\langle\hat{Q}^\lambda_h(\cdot,\cdot),\pi_h(\cdot|\cdot)\rangle_{\mathcal{A}}$ and $\hat{V}^\lambda_h(\cdot) \leftarrow \langle\hat{Q}^\lambda_h(\cdot,\cdot),\hat{\pi}_h(\cdot|\cdot)\rangle_{\mathcal{A}}$
10: **end for**

- Specifically designed penalty with β_{χ^2}

Content

Instance-Dependent Upper Bound

- **We provide Instance-dependent upper bound for our algorithms:**

Theorem 5.2. Under Assumption 3.1, for any $\delta \in (0,1)$, if we set $\gamma = 1$ and $\Gamma_h(s,a) = \beta \sum_{i=1}^{d} \|\phi_i(\cdot,\cdot)\mathbf{1}_i\|_{\mathbf{\Lambda}_h^{-1}}$ in Algorithm 1,

- (TV) $\beta = 16Hd\sqrt{\xi_{\mathrm{TV}}}$, where $\xi_{\mathrm{TV}} = 2\log(1024Hd^{1/2}K^2/\delta)$;
- (KL) $\beta = 16d\lambda e^{H/\lambda}\sqrt{(H/\lambda + \xi_{\mathrm{KL}})}$, where $\xi_{\mathrm{KL}} = \log(1024d\lambda^2K^3H/\delta)$;
- (χ^2) $\beta = 8dH^2(1+1/\lambda)\sqrt{\xi_{\chi^2}}$, where $\xi_{\chi^2} = \log(192K^5H^6d^3(1+H/2\lambda)^3/\delta)$,

then with probability at least $1-\delta$, for all $s \in \mathcal{S}$, the suboptimality of Algorithm 1 satisfies:

$$\mathrm{SubOpt}(\hat{\pi}, s, \lambda) \le 2\beta \sup_{P \in \mathcal{U}^{\lambda}(P^0)} \sum_{h=1}^{H} \mathbb{E}^{\pi^*, P}\Big[\sum_{i=1}^{d} \|\phi_i(s,a)\mathbf{1}_i\|_{\Lambda_h^{-1}} | s_1 = s\Big].$$

$\Phi\left(\mathbf{\Lambda}_h^{-1}, s\right)$: uncertainty function

- **The upper bound relies on a novel uncertainty function**

Information-Theoretic Lower Bound

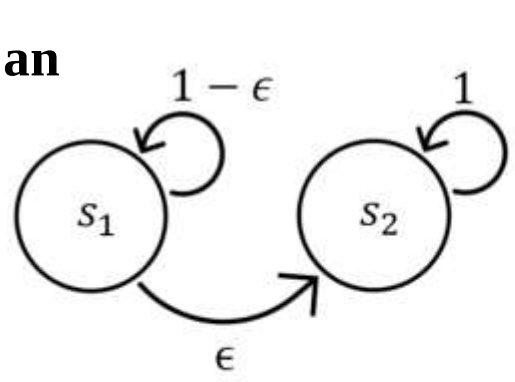
- **We further establish information-theoretic lower bound to illustrate the necessity of $\Phi\left(\boldsymbol{\Lambda}_h^{-1}, s\right)$**

Theorem 6.1. Given a regularizer λ, dimension d, horizon length H and sample size $K > \max\{\tilde{O}(d^6), \tilde{O}(d^3H^2/\lambda^2)\}$, there exists a class of d-rectangular linear RRMDPs $\mathcal{M}$ and an offline dataset $\mathcal{D}$ of size K such that for all $s \in \mathcal{S}$ and any divergence D among $D_{\mathrm{TV}}, D_{\mathrm{KL}}$ and D_{χ^2}, with probability at least $1-\delta$, we have $\inf_{\hat{\pi}} \sup_{M\in\mathcal{M}} \mathrm{SubOpt}(M, \hat{\pi}, s, \lambda, D) \geq c \cdot \Phi(\boldsymbol{\Lambda}_h^{-1}, s)$, where c is a universal constant.

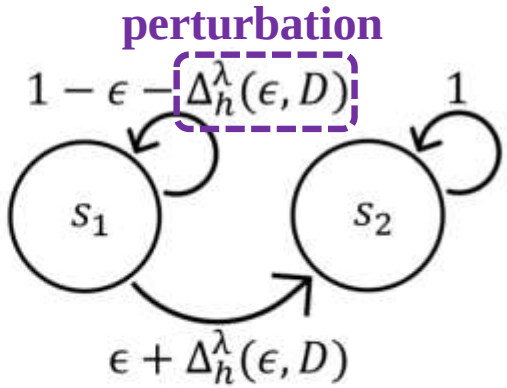
- **The construction of hard instance**

  - **The nominal environment is constructed by inserting an error into the environment with two absorbing states**

  - **The perturbed environment resembles the nominal kernel besides a controllable perturbation**



(a) The nominal environment.



(b) The perturbed environment under time step h.

Instance-Independent Upper Bound

■ Comparison of the Suboptimality gap with dataset coverage

Algorithm	Setting	Robust	Divergence	Coverage	Suboptimality gap
DRPVI	DRMDP	ρ	TV	full	$\tilde{O}(d^{3/2}H^2K^{-1/2})$
DROP	DRMDP	ρ	TV	robust partial	$\tilde{O}(d^{3/2}H^2K^{-1/2})$
P2MPO (TV)	DRMDP	ρ	TV	robust partial	$\tilde{O}(d^2H^2K^{-1/2})$
R2PVI-TV	RRMDP	λ	TV	regularized partial	$\tilde{O}(d^2H^2K^{-1/2})$
DRVI-L	DRMDP	ρ	KL	robust partial	$\tilde{O}(\sqrt{\beta}e^{H/\beta}d^2H^{3/2}K^{-1/2})^\star$
P2MPO (KL)	DRMDP	ρ	KL	robust partial	$\tilde{O}(e^{H/\beta}d^2H^2\rho^{-1}K^{-1/2})^\star$
R2PVI-KL	RRMDP	λ	KL	regularized partial	$\tilde{O}(\sqrt{\lambda}e^{H/\lambda}d^2H^{3/2}K^{-1/2})$
R2PVI-χ^2	RRMDP	λ	χ^2	regularized partial	$\tilde{O}(d^2H^3(1+\lambda^{-1})K^{-1/2})$

** The $\star$ denotes that the result requires an additional assumption on the KL dual variable, which is not required in **R2PVI***

- **For TV divergence, our algorithm achieves nearly same suboptimality gap with SOTA**
- **For KL divergence, our algorithm needs no extra assumption to guarantee the closeness form solution**
- **For χ^2 divergence, we are the first to give algorithms under linear MDP setting with χ^2 divergence**

Content

Experiment

■ We want to explore:

- The **robustness** of R2PVI when facing adversarial dynamics
- The role of **regularizer** λ in determining the robustness of R2PVI
- The **computation cost** of R2PVI compared to other robust algorithms

■ Baselines

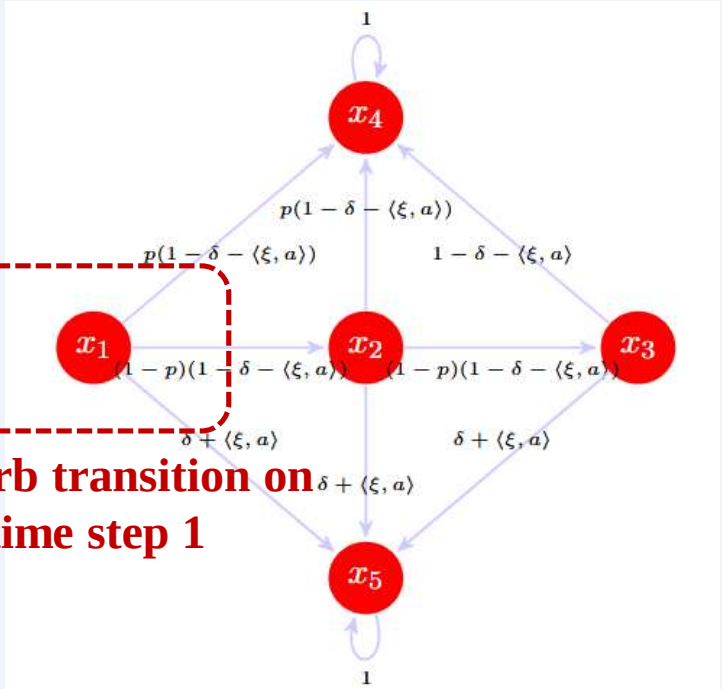
Method	PEVI	DRPVI	DRVI-L	R2PVI (ours)
Framework	MDP	d-DRMDP	d-DRMDP	d-RRMDP
Divergence	/	TV	KL	TV/KL/χ^2

** We don't compare DROP and P2MPO mentioned in the upper bound due to the lack of experiment and code base in such works.*

Experiment

■ Task settings

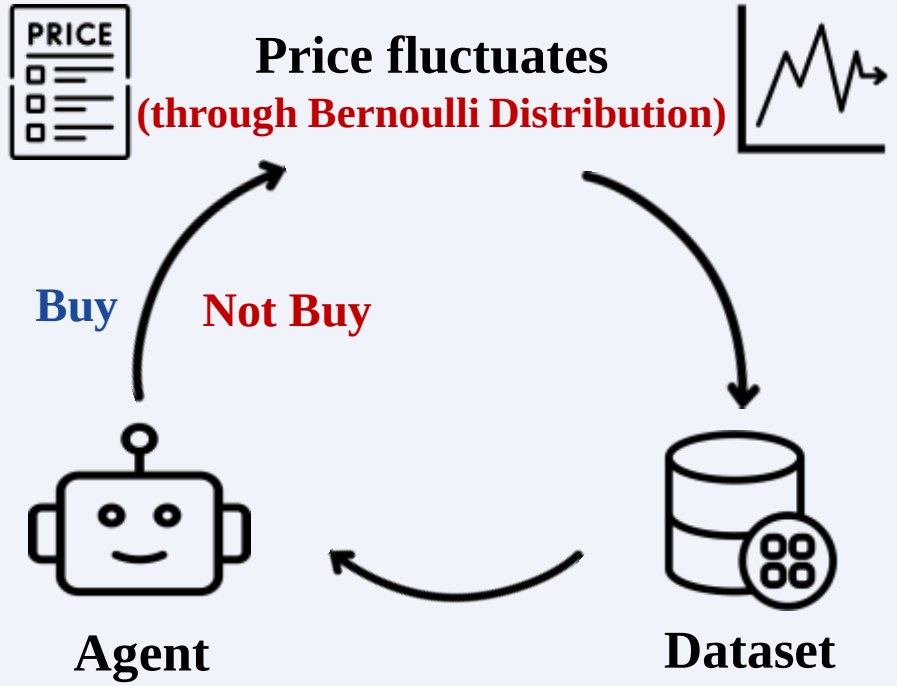
● Simulated Linear MDP



(a) The source MDP environment.

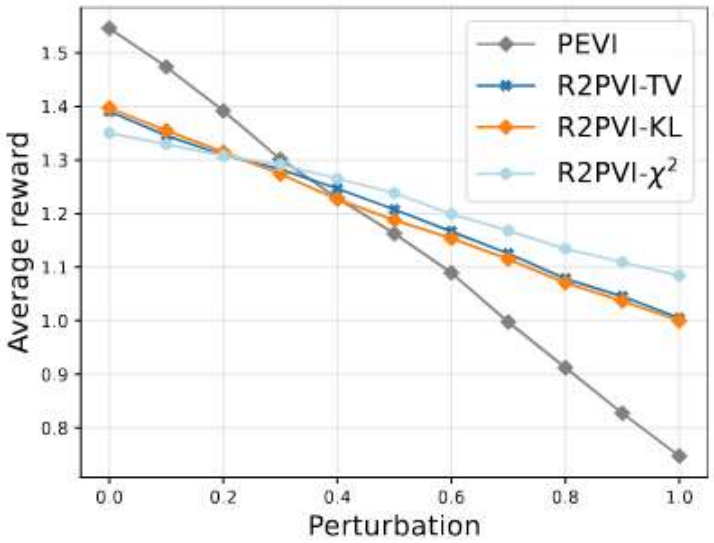
Perturb transition on time step 1

● American Put Option

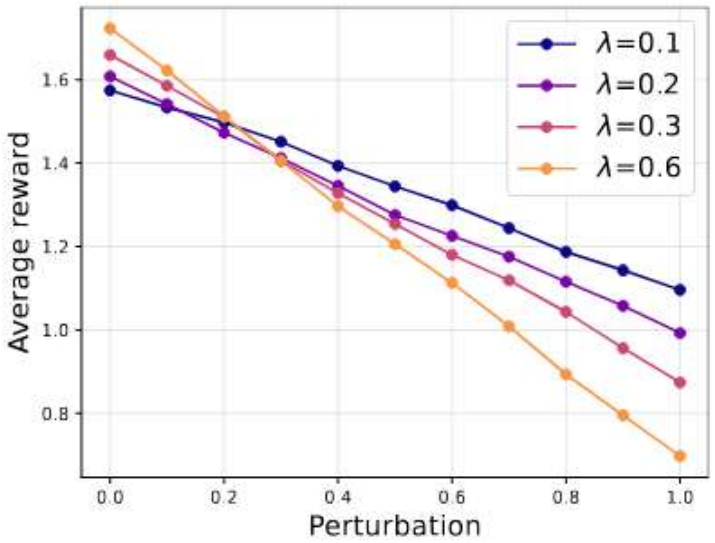


Simulated Linear MDP

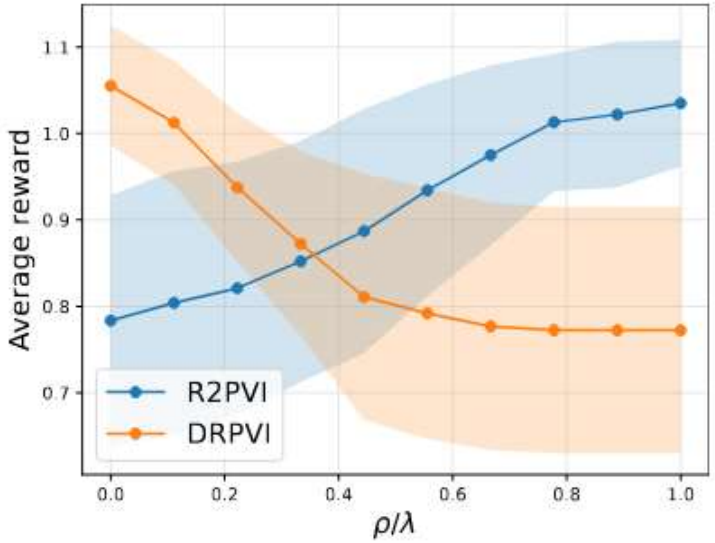
■ Evaluation



(a) $\lambda = 0.1$



(b) R2PVI

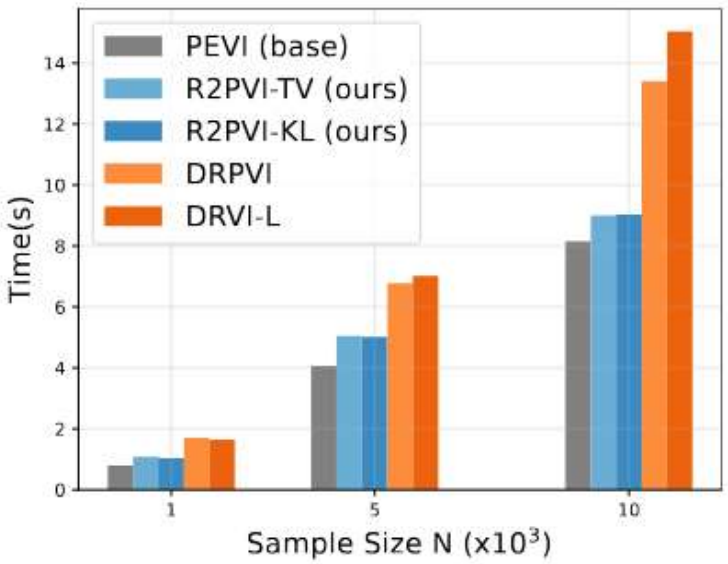


(c) q=0.9

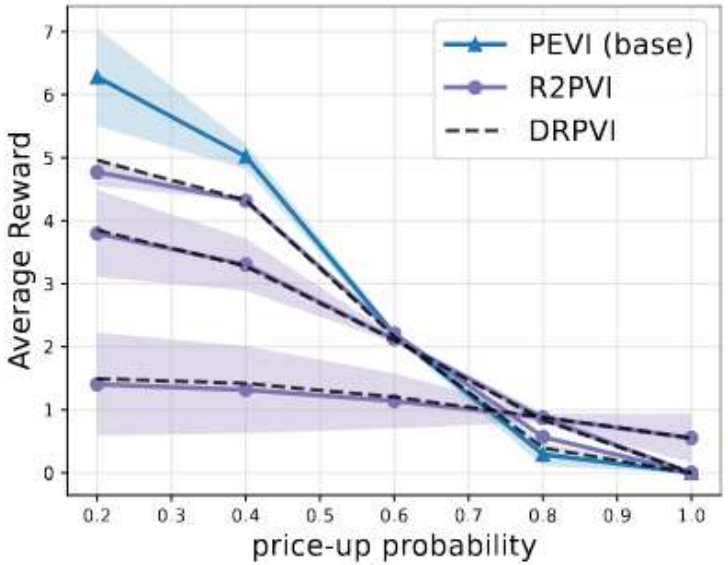
- Compared to non robust algorithm (PEVI), R2PVI learns **robust** policy under all divergence measure
- Robust parameter λ serves as **regularization** to adjust the robustness of the policy
- λ plays a similar role in d-RRMDP as **inverse robust parameter** $1/\rho$ in d-DRMDP

American Put Option

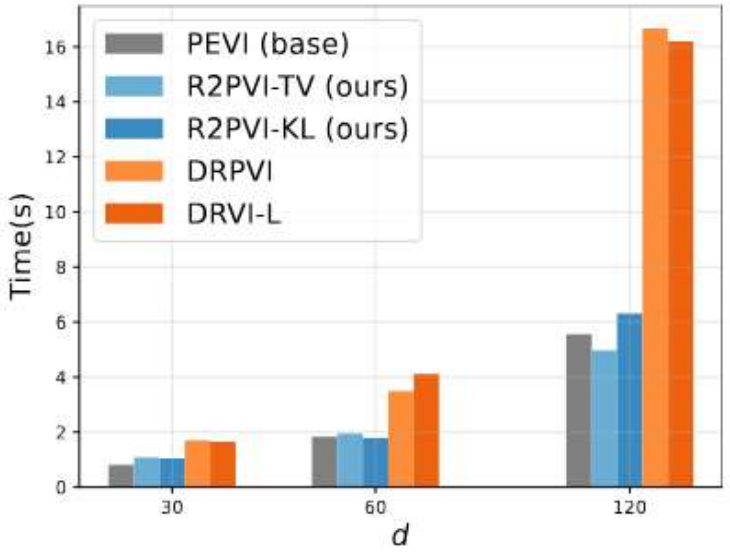
■ Evaluation



(a) execution time w.r.t N.



(b) $\lambda = 1, 3, 5$, $\rho = 0.025, 0.1, 0.2$.



(c) execution time w.r.t d.

- When N and d are large, the computation cost of DRPVI and DRVI-L increase rapidly
- The computation cost of R2PVI is as low as PEVI.
- R2PVI can achieve the same robust performance with DRPVI and DRVI-L

Content

Conclusion

■ Contribution

- We propose a novel d-RRMDP framework and establish dynamic planning principles
- We derive dual formulations of Q-functions under TV, KL, χ^2 divergence, and admit their linear representations
- We design meta-algorithms, R2PVI, in our setting and provide specific applicable algorithms under TV, KL, χ^2 divergence
- We provide instance-dependent upper bounds of our algorithms with a general form

$$\beta \sup_{P\in\mathcal{U}^\lambda(P^0)} \sum_{h=1}^{H} \mathbb{E}^{\pi^\star,P}\left[\sum_{i=1}^{d} \|\phi_i(s,a)\mathbf{1}_i\|_{\Lambda_h^{-1}} | s_1 = s\right],$$

and then construct theoretical-lower bound to illustrate that the general form is intrinsic

- We conduct extensive experiment to illustrate robustness and time efficiency of our algorithms

Thank you!