Robust Offline Reinforcement Learning with Linearly Structured f-Divergence Regularization

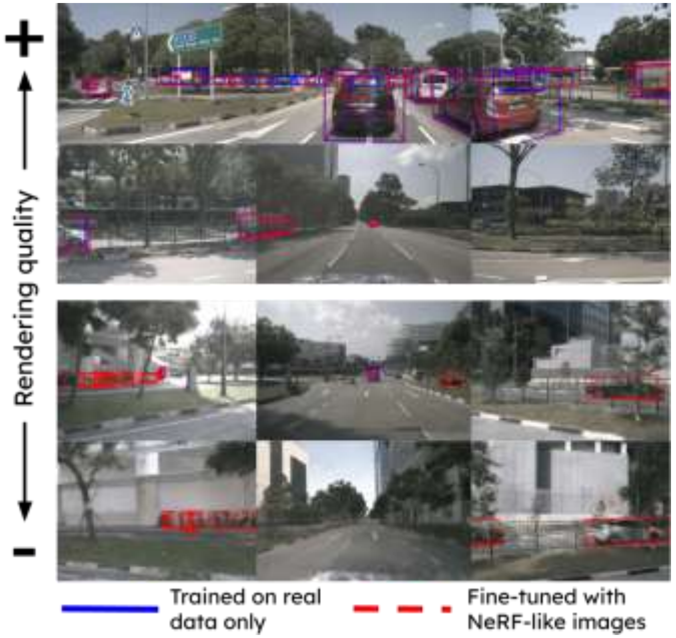
Cheng Tang[1], Zhishuai Liu[2], Pan Xu[2]

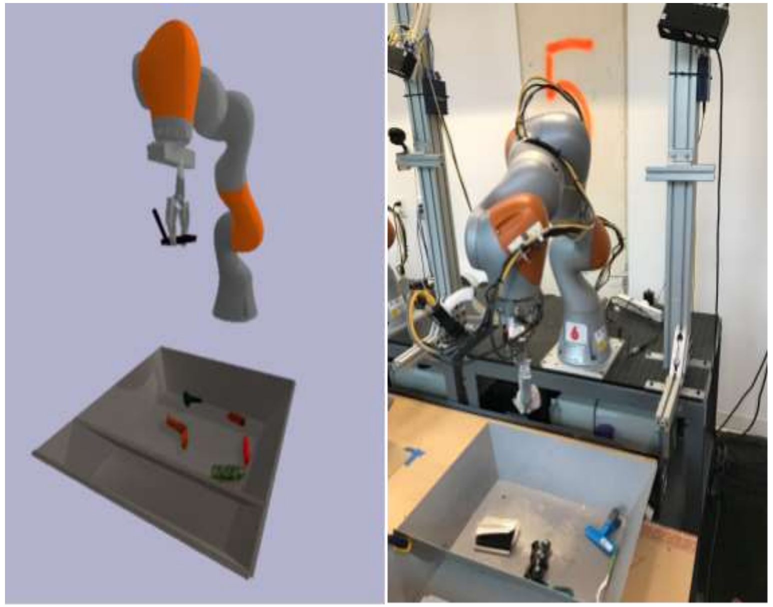
[1] University of Illinois Urbana-Champaign, [2] Duke University

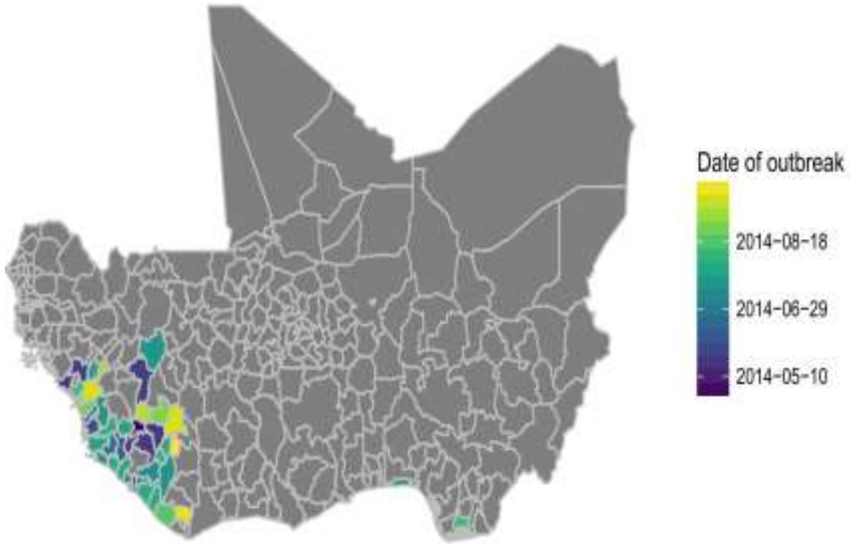
Content

Introduction

■ Offline RL calls for **robust policy**:







- **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:** MDPs + uncertainty set

Value function

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

+

(s, a) - 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

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

(4)

Motivation

- **Drawbacks of d-rectangular DRMDP (hard constraint):**
  - **From theoretical perspective:** need strong assumption on dual variables
  - **From empirical perspective:** solving duality problem in d-DRMDP is time-consuming
  - Existing work considers mainly TV divergence geometry, leaving blanks for cases with KL and χ^2
- **RRMDP: applying regularization penalty term (soft constraint) to measuring the uncertainty**
  - **From Lagrange Duality perspective:** DRMDP $\Leftrightarrow$ RRMDP
  - The forfeit of uncertainty set constraint makes the dual problem easier, leading to potential improvement on computation efficiency and theoretical analysis

Content

Problem Formulation

- **RRMDP (Robust Regularized Markov Decision Process):** $\mathrm{RRMDP}(S, A, H, P^0, r, \lambda, D, F)$

  - Regularized robust parameter λ, probability divergence D, feasible set of all perturbed transition kernels F

  - Regularized robust value function and Q-function:

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

Penalty on divergence with nominal kernel

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

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

- **Linear MDP :**

  - 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

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

Content

Dynamic programming principles

- Robust regularized Bellman Equation:

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

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

- 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 $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

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

Method - Framework

Pessimism based meta 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 R2PVI under general f-divergence

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

init $\hat{V}_{H+1}^{\lambda}(\cdot) = 0$
for episode $h = H, \cdots, 1$ **do**
  Compute $\boldsymbol{\Lambda}_h \leftarrow \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \gamma \mathbf{I}$
  $\hat{w}_{h,i}^{\lambda}(\alpha) \leftarrow \big[\boldsymbol{\Lambda}_h^{-1}\big[\sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau) f^*\big(\frac{\alpha - \hat{V}_{h+1}^{\lambda}(s)}{\lambda}\big)\big]\big]^i$
  ▷ Duality Estimation for general f-divergence
  $\hat{w}_{h,i}^{\lambda} \leftarrow \sup_{\alpha \in \mathbb{R}}\{-\lambda \hat{w}_{h,i}^{\lambda}(\alpha) + \alpha\}$
  Construct the penalty $\Gamma_h(\cdot,\cdot)$
  Estimate $\hat{Q}_h^{\lambda}(\cdot,\cdot) \leftarrow \min\{\langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_h + \hat{\boldsymbol{w}}_h^{\lambda}\rangle - \Gamma_h(\cdot,\cdot), H - h + 1\}^+$.
  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 $\hat{V}_h^{\lambda}(\cdot) \leftarrow \langle \hat{Q}_h^{\lambda}(\cdot,\cdot), \hat{\pi}_h(\cdot|\cdot)\rangle_{\mathcal{A}}$.
end for

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:

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

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

- The upper bound relies on a novel uncertainty function

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

Instance-Independent Upper Bound

■ Comparison of the Suboptimality gap with dataset coverage

Algorithm	Setting	Divergence	Coverage	Suboptimality Gap
DRPVI (Liu & Xu, 2024b)	d-DRMDP	TV	full	$\tilde{O}(dH^2K^{-1/2})$
DROP (Wang et al., 2024a)	d-DRMDP	TV	robust partial	$\tilde{O}(d^{3/2}H^2K^{-1/2})$
P2MPO (TV) (Blanchet et al., 2024)	d-DRMDP	TV	robust partial	$\tilde{O}(d^2H^2K^{-1/2})$
R2PVI-TV **(ours)**	d-RRMDP	TV	regularized partial	$\tilde{O}(d^2H^2K^{-1/2})$
DRVI-L (Ma et al., 2022)	d-DRMDP	KL	robust partial	$\tilde{O}(\sqrt{\underline{\beta}}e^{H/\underline{\beta}}d^2H^{3/2}K^{-1/2})^\star$
P2MPO (KL) (Blanchet et al., 2024)	d-DRMDP	KL	robust partial	$\tilde{O}(e^{H/\underline{\beta}}d^2H^2\rho^{-1}K^{-1/2})^\star$
R2PVI-KL **(ours)**	d-RRMDP	KL	regularized partial	$\tilde{O}(\sqrt{\lambda}e^{H/\lambda}d^2H^{3/2}K^{-1/2})$
R2PVI-χ^2 **(ours)**	d-RRMDP	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, R2PVI achieves nearly same suboptimality gap with SOTA
- For KL divergence, R2PVI needs no extra assumption to guarantee the closeness form solution
- For χ^2 divergence, we are the first to give theoretical guarantee under linear MDP setting

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

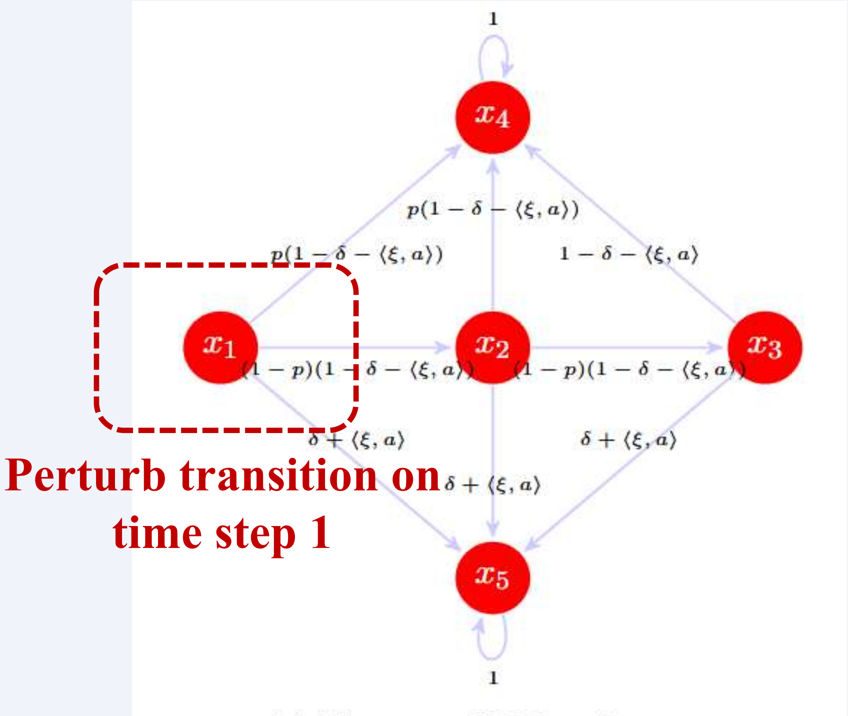
Method	PEVI	DRPVI	DRVI-L	R2PVI (ours)
Framework	MDP	d-DRMDP	d-DRMDP	d-RRMDP
Divergence	/	TV	KL	TV/KL/χ^2

** 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

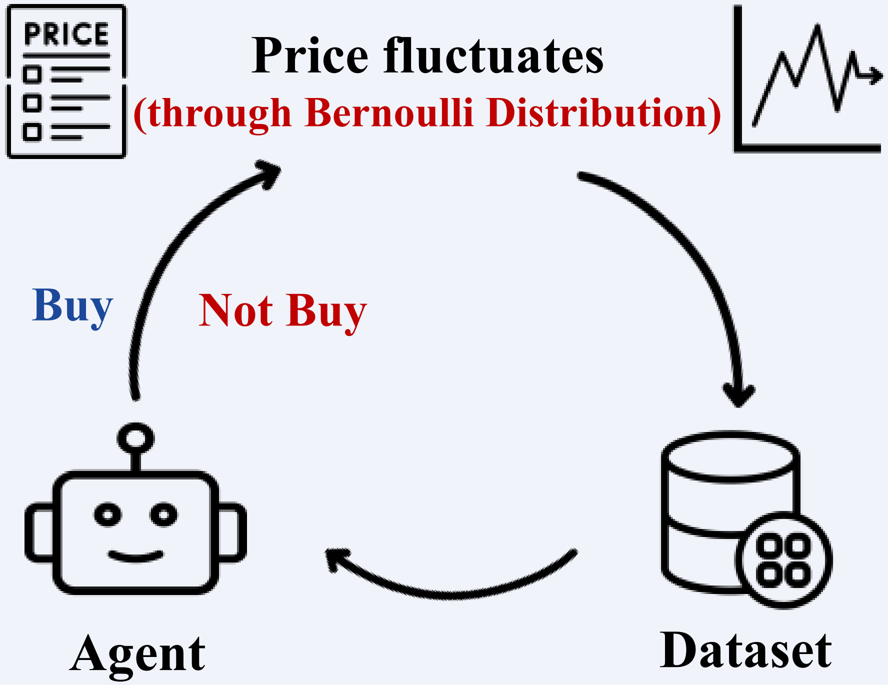
■ Task settings

● Simulated Linear MDP



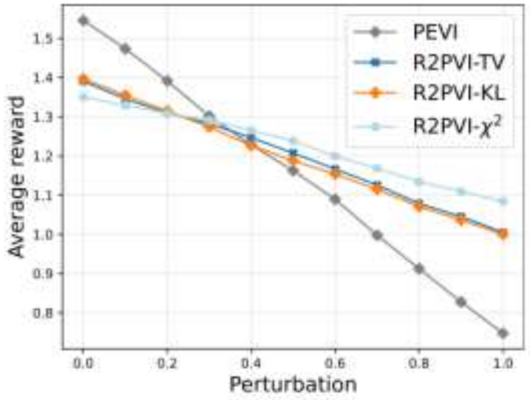
(a) The source MDP environment.

● American Put Option

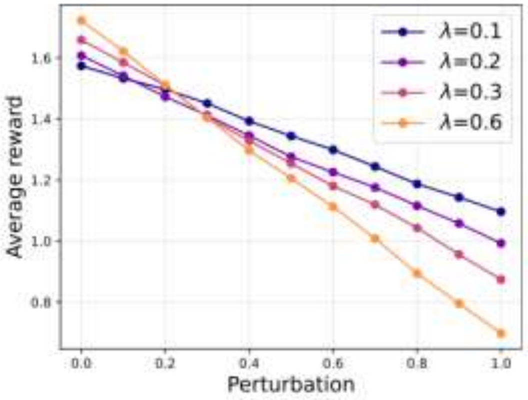


Results

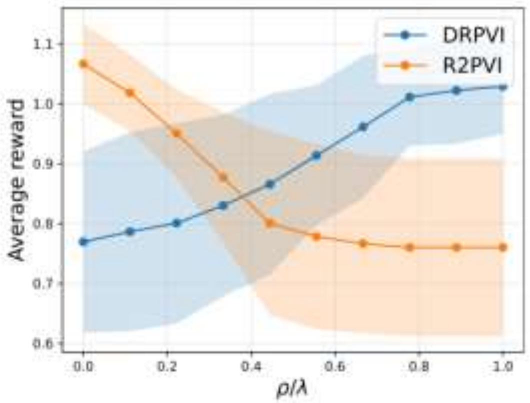
■ Evaluation



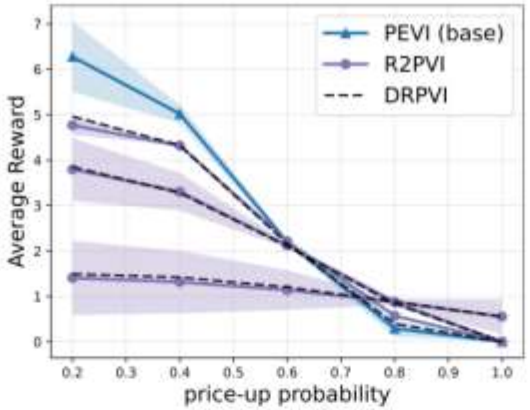
(a) $\lambda = 0.1$



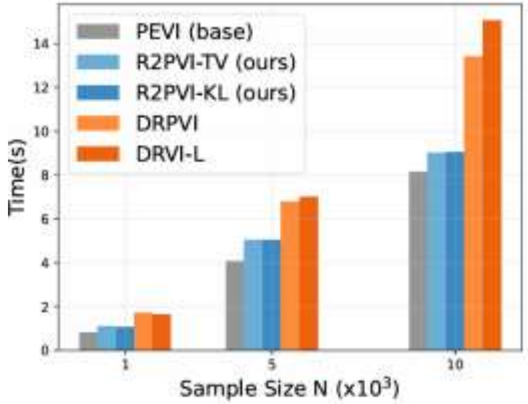
(b) R2PVI



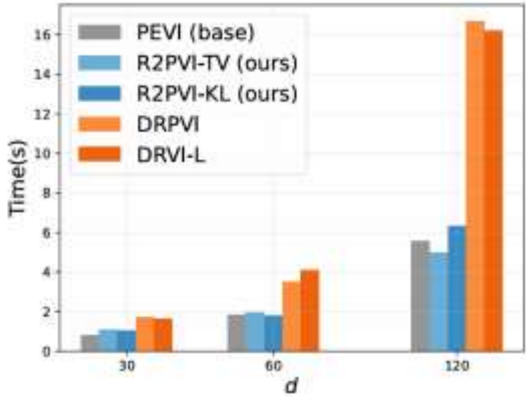
(c) $q = 0.9$



(d) (up to down) $\lambda = 1, 3, 5, \rho = 0.2, 0.1, 0.025.$



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



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