Efficient Exploration in Average-Reward Constrained RL: Achieving Near-Optimal Regret With Posterior Sampling

ICML2024 SLIDES

Danil Provodin, PhD Candidate

Department of Mathematics and Computer Science, Data and AI cluster

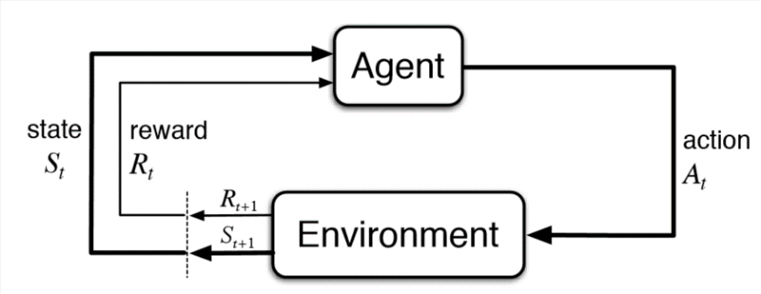
JADS Jheronimus Academy of Data Science

TU/e

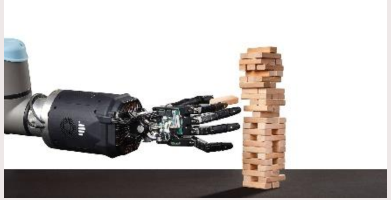
Content

- Motivation and background
- Main results
- Posterior sampling algorithm
- Experiments
- Conclusion and future work

Constrained RL



≠



📌 Multi-dimensional feedback

📌 Restrictions on what policy can do

Constrained MDPs

- CMDP $(\mathcal{S}, \mathcal{A}, p, r, c, \tau)$
  - Finite state space $\mathcal{S}$ and action space $\mathcal{A}$ $(|\mathcal{S}| = S, |\mathcal{A}| = A)$
  - Transition kernel $p(s'|s,a)$
  - Reward function $r(s,a) \in [0,1]$
  - Cost function $c(s,a) \in [0,1]^m$
  - Cost threshold $\tau \in [0,1]^m$
- Policy $\pi: \mathcal{S} \to \Delta(\mathcal{A})$
- Communicating CMDP $\forall s, s'$ there **exists** a stationary policy under which s' is accessible from s in at most D steps (D is diameter)

Objective

- Gain (loss) of policy

$$J^\pi(r,p) = \overline{\lim}_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}_p^\pi \left[r(s_t, a_t) \right];$$

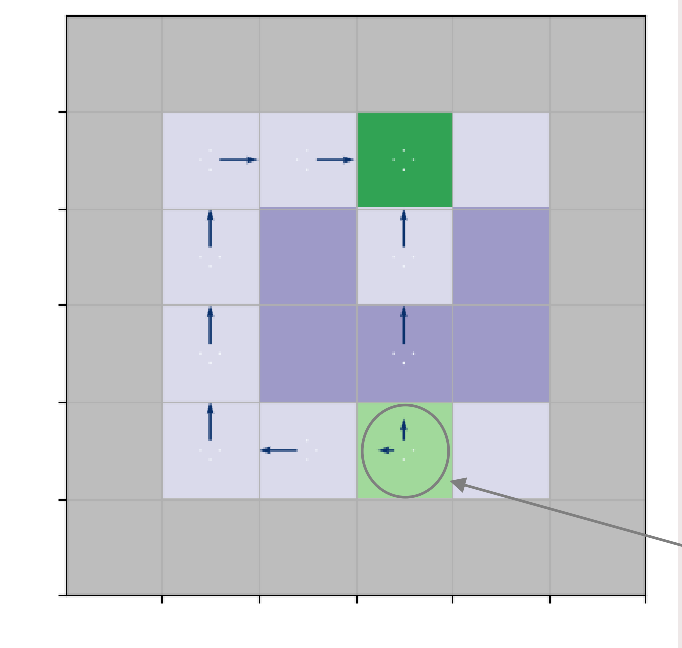
- Optimal policies

constraints

$$\sup_\pi J^\pi(r,p) \quad \text{s.t.} \quad J^\pi(c_i,p) \le \tau_i, \; i = 1, \ldots, m;$$

Main objective

CMDP example



goal state

costly states: swamp

safe states: road

starting position

Agent needs to randomize depending on the budget

Performance measure

- Bayesian regret
  - Define Ω – a set of transitions p such that resulting CMDP is communicating
  - Let f_0 be a prior distribution over Ω
  - Assume that actual transitions $p_* \sim f_0$

main regret

$$BR_+(T, r) = \mathbb{E}_{f_0}\left[\sum_{t=1}^{T}\left(J^*(r, p_*) - r(s_t, a_t)\right)_+\right]$$

constraint violation

$$BR_+(T, c_i) = \mathbb{E}_{f_0}\left[\sum_{t=1}^{T}\left(c_i(s_t, a_t) - \tau_i\right)_+\right], i = 1, \ldots, m.$$

Content

- Motivation and background
- Main results
- Posterior sampling algorithm
- Experiments
- Conclusion and future work

Main result

Theorem:
Suppose CMDP $(\mathcal{S}, \mathcal{A}, p_*, r, c, \tau)$ is communicating with diameter D.
Then there exists an algorithm such that if $T \geq D^4 S^2 A \log(2AT)^2$
the main regret and constraint violation are bounded by:

$$\mathrm{BR}_+(T, r) \leq O\left(DS\sqrt{AT\log(AT)}\right)$$
$$\mathrm{BR}_+(T, c_i) \leq O\left(DS\sqrt{AT\log(AT)}\right), i = 1, \ldots, m$$

- Implies optimal dependency in terms of T and A
- Matches the best-known bound for unconstrained setting - $\tilde{O}(DS\sqrt{TA})$
- First near-optimal bound achieved by computationally tractable algorithm

Content

- Motivation and background
- Main results
- Posterior sampling algorithm
- Experiments
- Conclusion and future work

Feasibility

- Slater's condition

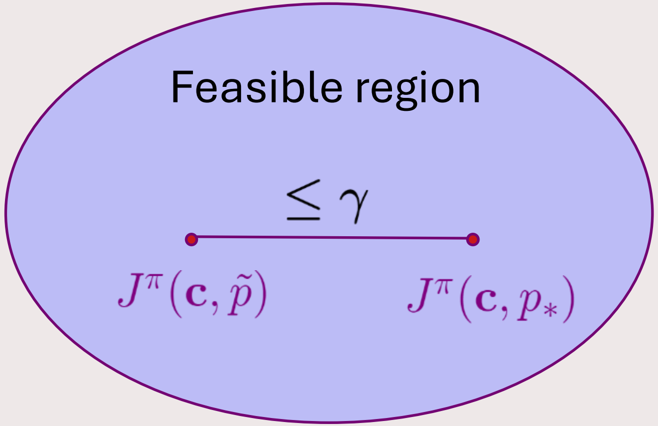
$$\exists \quad \pi : \quad J^{\pi}(\mathbf{c}, p_*) < \tau - \gamma$$

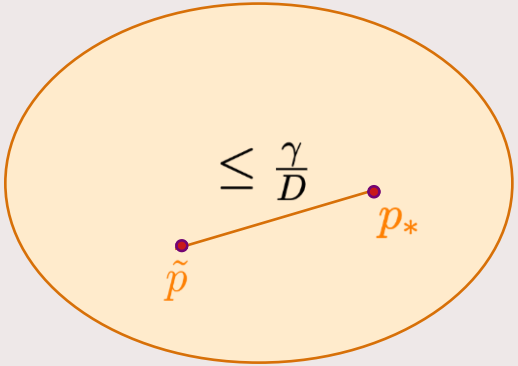
- Relationship between losses and transitions

$$J^{\pi}(\mathbf{c}, \tilde{p}) - J^{\pi}(\mathbf{c}, p_*) \leq D||\tilde{p}(\cdot|s,a) - p_*(\cdot|s,a)||_1$$

difference in losses

deviation between sampled and true transitions





PSConRL

1. Form empirical CMDP $\tilde{p}(s'|s,a) \sim f(\cdot \mid N_{sas'}), \hat{r}(s,a) = \frac{\sum r_{sa}}{N(s,a)},$ $\hat{c}(s,a) = \frac{\sum c_{sa}}{N(s,a)}$
2. If CMDP $\tilde{p}$ is feasible
   a) Solve CMDP: Find $\hat{\pi}$ which is optimal for CMDP $(\mathcal{S}, \mathcal{A}, \tilde{p}, \hat{r}, \hat{c}, \tau)$
3. If CMDP $\tilde{p}$ is not feasible
   b) Explore more: Find $\hat{\pi}$ which explores environment efficiently
4. Execute $\hat{\pi}$ and collect more data*

Linear program for CMDPs

* we split interaction into artificial episodes based on doubly-epoch construction technique

Linear program for CMDPs

$$\max_{\mu} \sum_{s,a} \mu(s,a) r(s,a),$$

$$\text{s.t.} \quad \sum_{s,a} \mu(s,a) c_i(s,a) \le \tau_i, \quad i = 1, \dots, m,$$

$$\sum_{a} \mu(s,a) = \sum_{s',a} \mu(s',a) p(s',a,s), \quad \forall s \in \mathcal{S},$$

$$\mu(s,a) \ge 0, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad \textstyle\sum_{s,a} \mu(s,a) = 1;$$

Linear program in occupancy measure μ

$$\pi_*(a|s) = \frac{\mu_*(s,a)}{\sum_{a'} \mu_*(s,a')}$$

Optimal policy

PSConRL

1. Form empirical CMDP $\tilde{p}(s'|s,a) \sim f(\cdot \mid N_{sas'}), \hat{r}(s,a) = \frac{\sum r_{sa}}{N(s,a)}$, $\hat{c}(s,a) = \frac{\sum c_{sa}}{N(s,a)}$
2. If CMDP $\tilde{p}$ is feasible
   a) Solve CMDP: Find $\hat{\pi}$ which is optimal for CMDP $(\mathcal{S}, \mathcal{A}, \tilde{p}, \hat{r}, \hat{c}, \tau)$
3. If CMDP $\tilde{p}$ is not feasible
   b) Explore more: Find $\hat{\pi}$ which explores environment efficiently
4. Execute $\hat{\pi}$ and collect more data*

Reduction to exploration MDPs

* we split interaction into artificial episodes based on doubly-epoch construction technique

Exploration MDP

$(\mathcal{S}, \mathcal{A}, p, c_{\bar{s}})$ for $\bar{s} \in \mathcal{S}$ – set of exploration MDPs

$$c_{\bar{s}}(s,a) = \begin{cases} 1, & \text{if } s \neq \bar{s}; \\ 0, & \text{otherwise.} \end{cases}$$

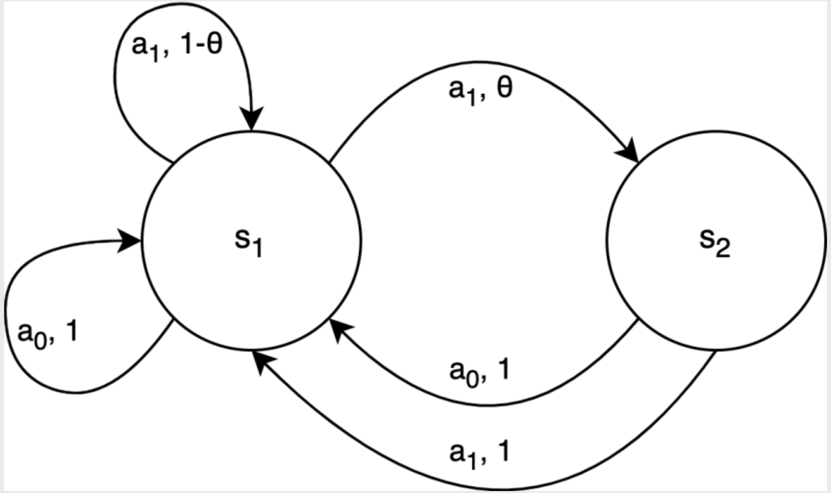
$$J^*(c_{\bar{s}}, p) + v^*(s; c_{\bar{s}}, p) = \min_{a \in \mathcal{A}} \left\{ c_{\bar{s}}(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s'; c_{\bar{s}}, p) \right\}, \forall s \in \mathcal{S}.$$

loss

bias function

Bellman optimality eq-n for average reward MDP

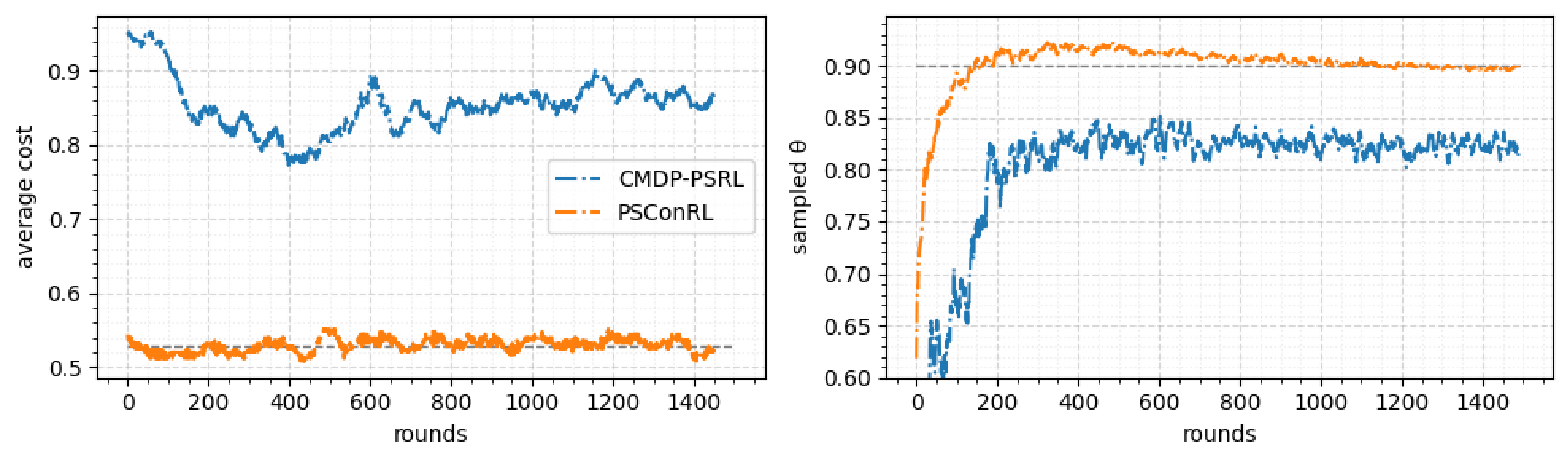
Why extra exploration? PSConRL vs PSRL-CMDP



$$r(s_1,\cdot) = 1, c(s_1,\cdot) = 1$$
$$r(s_2,\cdot) = 0, c(s_2,\cdot) = 0$$

- PSRL-CMDP - posterior sampling algorithm that doesn't reduce to exploration MDPs
- Suitable only for ergodic CMDPs
  - Can’t guarantee feasibility in communicating CMDPs

Why extra exploration?

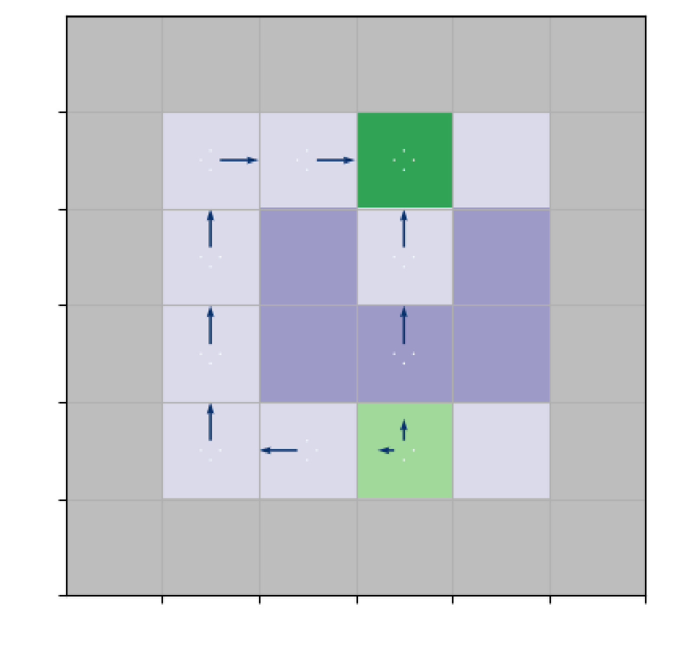


- PSConRL effectively learns the true transition parameter θ
- PSConRL achieves optimal average cost and fluctuates around it
- CMDP-PSRL fails to do so due to its unexplorative nature

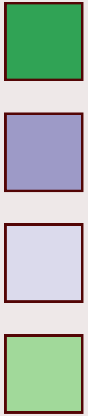
Content

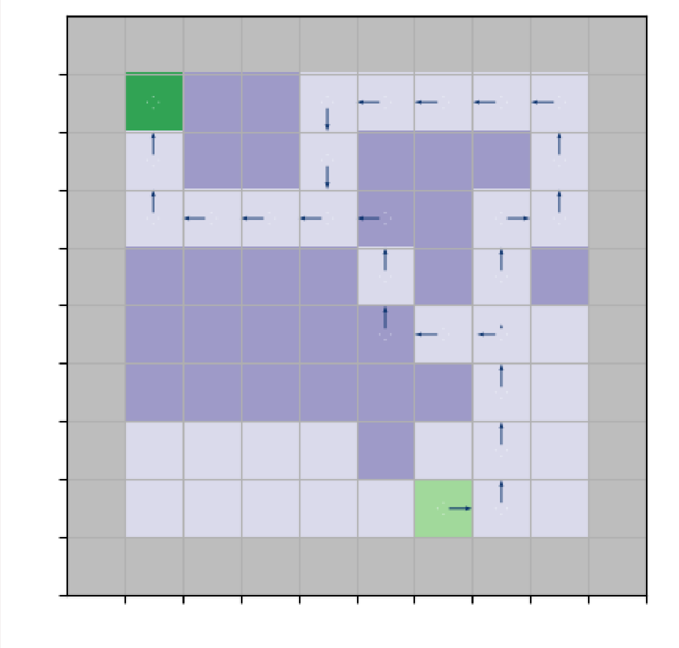
- Motivation and background
- Main results
- Posterior sampling algorithm
- Experiments
- Conclusion and future work

Marsrover environments



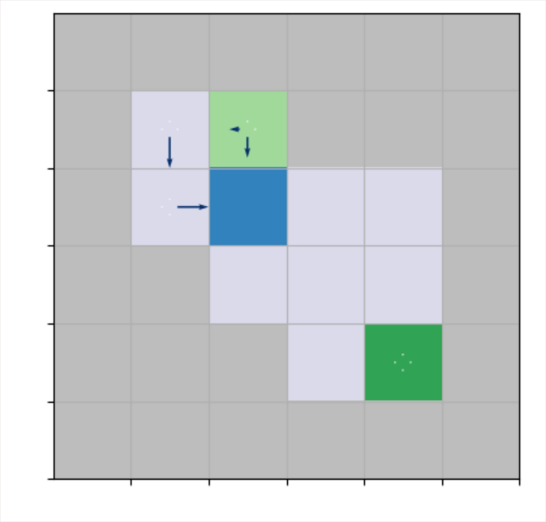
4x4



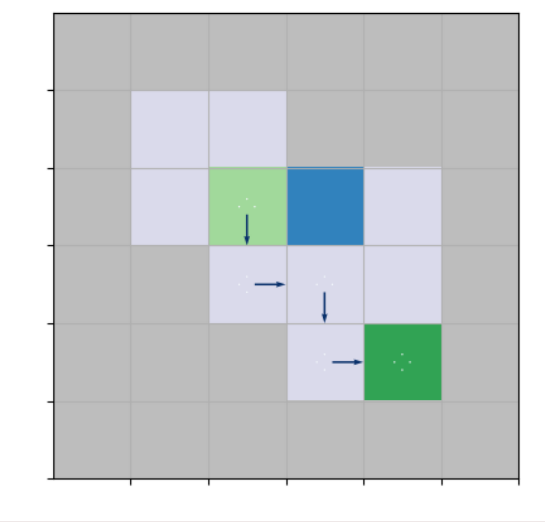


8x8

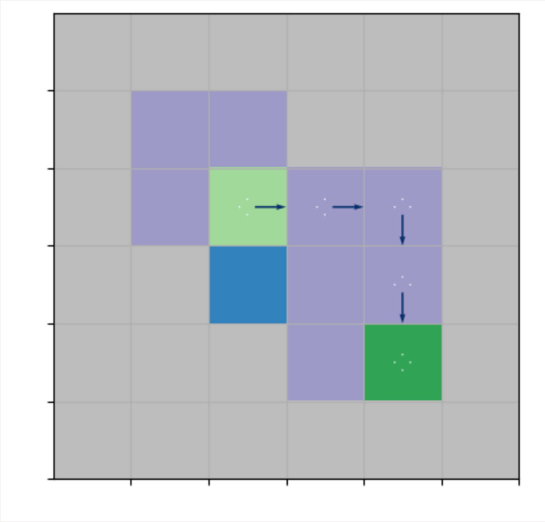
Box environment



Starting configuration

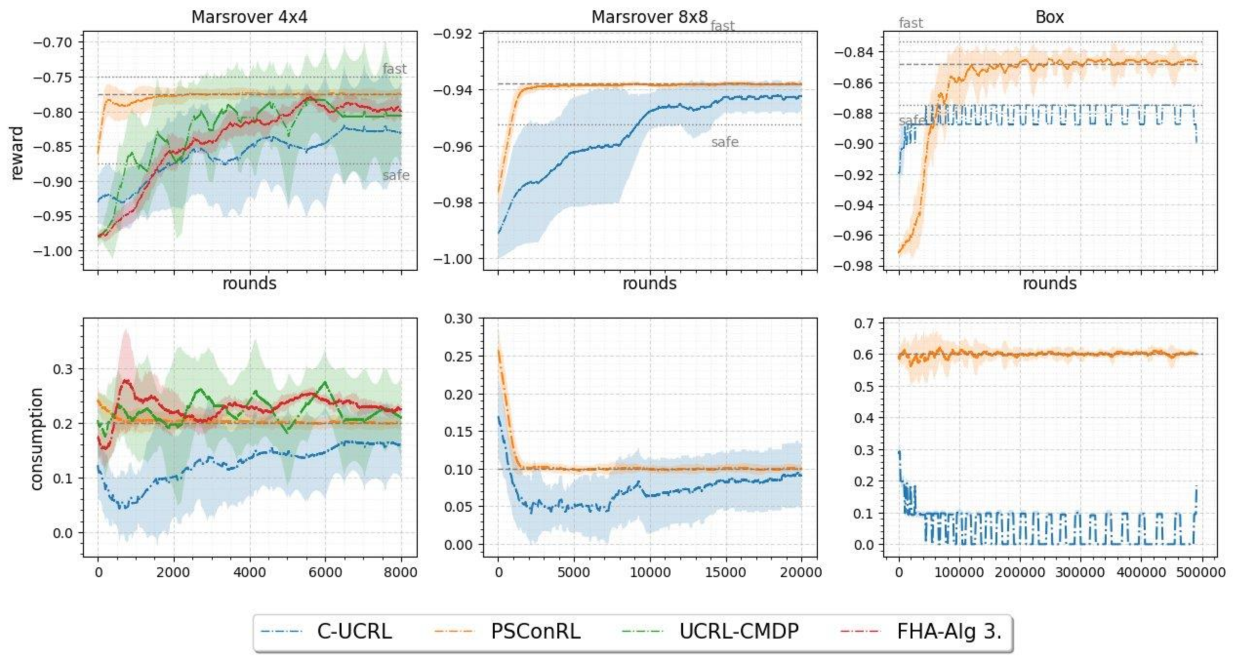


Move box right



Move box left

Empirical reward and cost



- PSConRL converges to optimal performance significantly ahead of baselines
- Optimistic algorithms UCRL-CMDP, FHA-Alg 3 fail to scale beyond the smallest environment
- C-UCRL is too conservative for constrained RL

Content

- Motivation and background
- Posterior sampling algorithm
- Main theoretical results
- Experiments
- Conclusion and future work

Takeaways

- PSConRL is practical and computationally efficient
  - (compared to optimistic algorithms)
  - It doesn't require any additional knowledge from the environment
  - It has polynomial time complexity in problem parameters
- PSConRL introduces a novel efficient exploration mechanism
  - PSConRL enjoys near-optimal Bayesian regret bound
  - PSConRL vs. CMDP-PSLR comparison highlights that the exploration step is essential for effective learning in communicating CMDPs
- A novel analysis of feasibility in constrained RL
  - First feasibility guarantees that don't rely on brute force optimization
  - Holds for frequentist setting

Future work

- Limitations of the current work
  - Bayesian regret to frequentist regret
  - Asymptotic regret bound
  - Finite S and A

Thank you for listening!

Questions?

Poster

Efficient Exploration in Average-Reward Constrained Reinforcement Learning: Achieving Near-Optimal Regret With Posterior Sampling

Danil Provodin · Maurits Kaptein · Mykola Pechenizkiy

Hall C 4-9

[Abstract]

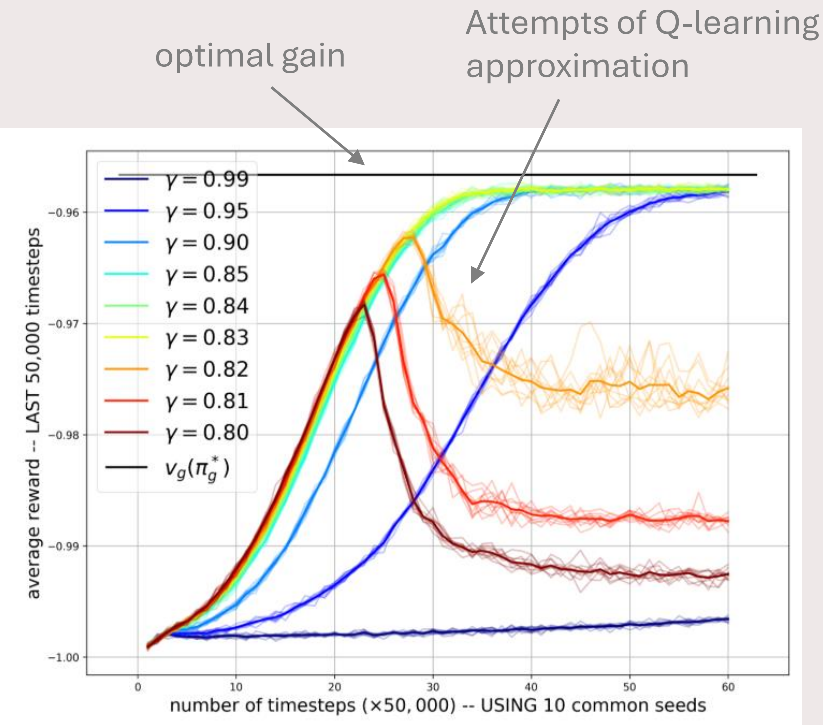
Wed 24 Jul 1:30 p.m. CEST — 3 p.m. CEST (Bookmark)

Open for Collaboration

Excited to explore new collaboration opportunities. If you're interested in working together, please feel free to reach out.

Contact: d.provodin@tue.nl
LinkedIn: linkedin.com/in/danil-provodin/

Why average-reward criterion?



- Discounted MDPs are ubiquitous in RL
  - Sometimes discount factor γ is inherent part of the problem
  - Or a problem has a small horizon
- Often we care about long-term performance (infinite horizon)
  - γ becomes part of the solution method, artificial discounting

📕 Examining average and discounted reward optimality criteria in reinforcement learning

Comparison to the existing literature

	Algorithm	Main Regret	Constraint violation	CMDP class	Required knowledge	Computation
frequentist	C-UCRL (Zheng & Ratliff, 2020)	$\tilde{O}(mSAT^{3/4})$	0	ergodic	safe policy π and p	efficient
	UCRL-CMDP (Singh et al., 2023)	$\tilde{O}(T_M\sqrt{SA}T^{2/3})$	$\tilde{O}(T_M\sqrt{SA}T^{2/3})$	ergodic	T	inefficient
	Alg. 3 (Chen et al., 2022)	$\tilde{O}(sp(p)(S^2AT^2)^{1/3})$	$\tilde{O}(sp(p)(S^2AT^2)^{1/3})$	weakly communicating	$sp(p)$, T	inefficient
	Alg. 4 (Chen et al., 2022)	$\tilde{O}(sp(p)S\sqrt{AT})$	$\tilde{O}(sp(p)S\sqrt{AT})$	weakly communicating	$sp(p)$, T	intractable
Bayesian	CMDP-PSRL (Agarwal et al., 2022)	$\tilde{O}(T_MS\sqrt{AT})$	$\tilde{O}(T_MS\sqrt{AT})$	ergodic	-	efficient
	PSConRL (this paper)	$\tilde{O}(DS\sqrt{AT})$	$\tilde{O}(DS\sqrt{AT})$	communicating	-	efficient
	lower bound (Singh et al., 2023)	$\Omega(\sqrt{DSAT})$	$\Omega(\sqrt{DSAT})$	-	-	-

Empirical regret