Provably Efficient RL for Linear MDPs under Instantaneous Safe Constraints in Non-Convex Feature Spaces

Joint work with:

Arnob Ghosh, New Jersey Institute of Technology
Ming Shi, University at Buffalo
Fatemeh Nourzad, The Ohio State University
Eylem Ekici, The Ohio State University
Ness B. Shroff, The Ohio State University

Amirhossein Roknilamouki, PhD Student
The Ohio State University, Department of Electrical and Computer Engineering

Motivation & Problem Definition

- **Goal:** Find a policy π that maximizes cumulative rewards while respecting instantaneous safety constraints at every step.

Mathematical Formulation:

$$\max_{\pi} \quad V^{\pi}(s_0) = \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} r_h(s_h, a_h)\right]$$

Subject to instantaneous safety constraints:

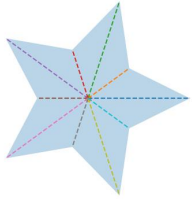
$$c_h(s_h, a_h) \leq \tau, \quad \text{for every step } h \text{ and all episodes.}$$

Examples:

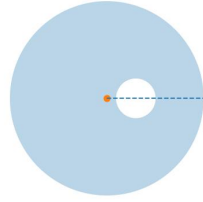
- **Autonomous Driving:** Immediate collision avoidance at every steering or acceleration decision.
- **Robotics:** Instantaneous obstacle avoidance at each robotic arm movement.
- **Healthcare:** Immediate adherence to safe dosage limits for medication.

State of the Art & Proof Gap

- **SLUCB-QVI** [Amani et al. 2021]: claims $\tilde{\mathcal{O}}\big((1+\frac{1}{\tau})\sqrt{d^3H^4K}\big)$ regret under star-convex safety.
- **Proof not correct:** *Uses an invalid covering-number bound from unconstrained RL; see details in Section 5.1.*
- **Star-convexity isn't enough:** Real-world problems such as autonomous driving and robotics often induce **non-star-convex or highly irregular safe decision spaces** (e.g., disjoint regions due to obstacles or kinematic constraints).



Star-Convex Set (all rays from center stay inside)



Non-Star-Convex Set (obstacle creates a hole)

Our Contributions

Objective–Constraint Decomposition (Star-Convex Case)

- We introduce a novel technique, OCD, to bound the covering number in constrained RL problems under star-convexity.
- *Impact:* Resolves Amani et al.'s proof gap; adds complexity factor $\mathcal{O}(\sqrt{\log(\frac{1}{\tau})})$ vs. unconstrained RL.

Lemma 5.3: In non-star-convex problems, the covering number can become arbitrarily large, and OCD (and other star-convex methods) are no longer sufficient.

NCS-LSVI — Two-Phase Algorithm for Non-Star-Convexity

- We develop a new algorithm, **NCS-LSVI**, that enables sublinear regret in non-star-convex environments.
- **Theorem 5.4:** Regret $= \mathcal{O}(\sqrt{K}) + \mathcal{O}\left(\frac{\log(K)}{\varepsilon^2 \iota^2}\right)$

Our Method: OCD for Star-Convexity, NCS-LSVI Beyond

Why ordinary covering arguments fail

- Unconstrained RL: $V_h^k(s) = \max_{a \in \mathcal{A}} Q_h^k(s,a)$ — the action set $\mathcal{A}$ is fixed.
- Constrained RL: $V_h^k(s) = \max_{a \in \widehat{\mathcal{A}}_h^k(s)} Q_h^k(s,a)$, where $\widehat{\mathcal{A}}_h^k(s)$ is data-dependent.

Our Decomposition Strategy (OCD)

- To bound $|V_1(s) - V_2(s)|$, we introduce V_3 using the same Q-function as V_1 and the same feasible set as V_2:

$$|V_1(s) - V_2(s)| \leq \underbrace{|V_1(s) - V_3(s)|}_{\text{objective difference}} + \underbrace{|V_2(s) - V_3(s)|}_{\text{constraint difference}}$$

- **Non-Star-Convex Spaces:** Without star-convexity, this strategy fails; *NCS-LSVI adds an exploration phase to restore provable bounds.*
- Initial exploration step in NCS-LSVI reduces uncertainty about constraints, making OCD valid again.

Conclusion & Future Research

Key Takeaways

- **Geometry matters:** Our work highlights the pivotal role of the decision space's geometry in shaping the complexity of safe RL.
- **Beyond unconstrained RL**: Instantaneous hard constraints demand new tools (e.g., OCD, NCS-LSVI) to keep covering numbers tight.

Future Directions

- Regret bound under Local Point Assumption depends on $\frac{1}{\varepsilon^2 \iota^2}$; whether this is fundamental remains open.
- Future work: extend beyond linear MDPs to handle deep RL with nonlinear feature spaces.
- Another direction: relax the Local Point Assumption to enable safety and low regret in complex environments.