WASHINGTON STATE UNIVERSITY

ICML International Conference On Machine Learning

An Optimistic Algorithm for online CMDPS with Anytime Adversarial Constraints

Jiahui Zhu[1], Kihyun Yu [2], Dabeen Lee[2], Xin Liu[3], Honghao Wei[1]

[1]School of Electrical Engineering and Computer Science, Washington State University
[2]Department of Industrial & Systems Engineering, KAIST
[3]School of Information Science & Technology, ShanghaiTech University

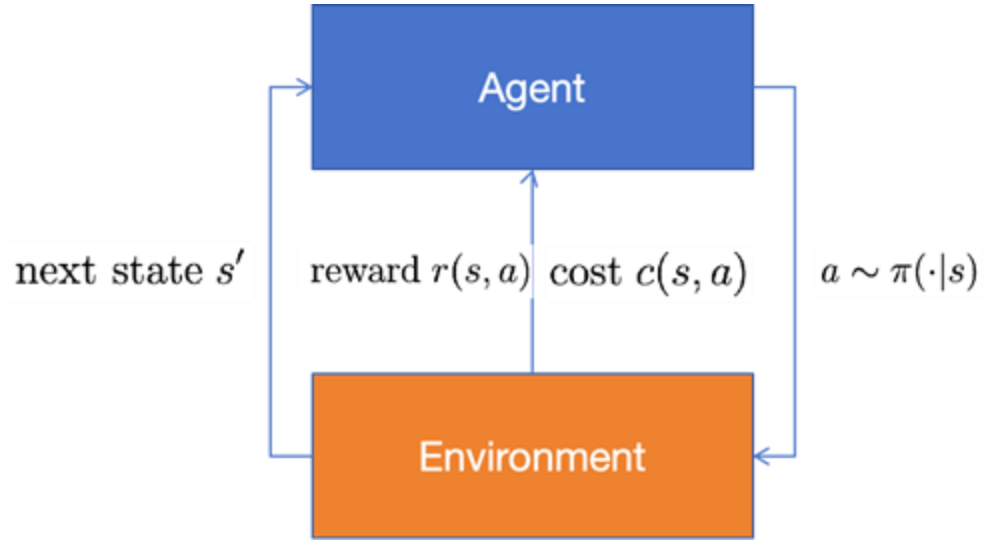
Constrained MDPs Notation[1]:

ICML International Conference On Machine Learning

μ	Initial state distribution
$\mathcal{S}$	State space (finite, $\lvert\mathcal{S}\rvert$ = S)
$\mathcal{A}$	Action space (finite, $\lvert\mathcal{A}\rvert$ = A)
H	Episode horizon length
$\mathcal{P}_h$	Transition kernel $\mathcal{P}_h(s'\lvert s,a)$: $\mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0,1]$
r_k	Reward vector $(r_{\{k,1\}} \dots r_{\{k,H\}})$; $r_{\{k,h\}}$: $\mathcal{S}\times\mathcal{A} \to [0,1]$
d_k	Cost vector $(d_{\{k,1\}} \dots d_{\{k,H\}})$; $d_{\{k,h\}}$: $\mathcal{S}\times\mathcal{A} \to [-1,1]$
Rewards	Stochastic — r_k i.i.d. from fixed distribution R
Costs (Stochastic)	d_k i.i.d. from fixed distribution D
Costs (Adversarial)	d_k chosen online by adversary
Policy π	$\pi = \{\pi_1 \dots \pi_H\}$, $\pi_h(\cdot\lvert s) \in \Delta(\mathcal{A})$

[1] Altman, Eitan. Constrained Markov decision processes. Routledge, 2021

Constrained MDPs Definition[1]:



Given a policy π, we use $V^{\pi}(r_k, p)$ and $V^{\pi}(d_k, p)$ to denote the expected cumulative reward and cost under policy π, starting from state s_1:

$$V^{\pi}(r_k, p) \coloneqq \mathrm{E}\left[\sum_{\mathrm{h}=1}^{H} r_{k,h}(s_h, \mathrm{a}_h) \,\middle|\, s_1, \pi, p\right], \quad V^{\pi}(d_k, p) \coloneqq \mathrm{E}\left[\sum_{\mathrm{h}=1}^{H} \mathrm{d}_{k,h}(s_h, \mathrm{a}_h) \,|s_1, \pi, p\right]$$

[1] Altman, Eitan. Constrained Markov decision processes. Routledge, 2021

Optimization Problem

The objective for Constrained MDPs is to find a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ ($\Delta(\cdot)$ is a probability simplex) to maximize the expected cumulative rewards while satisfying constraints under both settings:

$$\pi^* \in argmax\, V^{\pi}(\bar{r}, p), \qquad s.t.\ \ V^{\pi}(\bar{d}, p) \leq 0 \ \ (stochastic\ cost)$$
$$s.t.\ \ V^{\pi}(d_k, p) \leq 0, \forall\, k \in [K] \ \ (adversaril\ cost)$$

where $\bar{r} \coloneqq \mathrm{E}_{r\sim R}[r],\ \bar{d} \coloneqq \mathrm{E}_{d\sim D}[d]$. And the goal of the online Constrained MDPs problem is to learn an optimal policy to minimize the cumulative regret and strong cumulative violation of constraints after K episodes, which are defined below:

$$Regret(K) = \sum_{k=1}^{K} [V^{\pi^*}(\bar{r}, p) - V^{\pi^k}(\bar{r}, p)]$$

$$Violation(K) = \sum_{k=1}^{K} [V^{\pi^k}(\bar{d}, p)]^+ \ (stochastic\ cost)$$

$$Violation(K) = \sum_{k=1}^{K} [V^{\pi^k}(d_k, p)]^+ \ (adversarial\ cost)$$

Optimization Problem

Alternatively, the original optimization problem can be naturally reformulated a LP with the occupancy measure $\{q_h^\pi(s,a,s')\}_{h=1}^H$ [1] :

$$\max_{q\in Q} \bar{r}^T q \quad s.t.\ \bar{d}^T q \le 0\ (stochastic\ cost)$$

$$s.t.\ {d_k}^T q \le 0, \forall\, k \in [K] (adversarial\ cost)$$

where $q \in [0,1]^{SAH}$ is the occupancy measure vector and Q is the set of all valid occupancy measures; $\bar{r} \in [0,1]^{SAH}$, $\bar{d} \in [0,1]^{SAH}$ and $d_k \in [0,1]^{SAH}$. And the policy can be reconstructed as:

$$\pi_s^q(a|s) = \frac{q_h(s,a)}{\sum_{a'} q_h(s,a')}$$

To foster exploration in the unknown model, we adopt the principle of optimistic estimation [2], we define the empirical rewards $\hat{r}_h^{k-1}$, costs $\hat{d}_h^{k-1}$ and transition kernels $\hat{p}_h^{k-1}$, then we can construct the optimistic estimates for $\tilde{r}_{k,h}$, $\tilde{d}_{k,h}$ and optimistic occupancy measure set Q.

[1] Altman, Eitan. Constrained Markov decision processes. Routledge, 2021
[2] Auer, Peter, Thomas Jaksch, and Ronald Ortner. "Near-optimal regret bounds for reinforcement learning." Advances in neural information processing systems 21 (2008).

Algorithm: Surrogate Objective Function

We define the following surrogate objective function with exponential potential Lyapunov function:

$$f_k(q) = \alpha\left(-\tilde{r}_k^T q + \Phi'(\lambda_k)\left[\tilde{d}_k^T q\right]^+\right) - \frac{1}{2}\|q - q_k\|^2, \Phi'(x) = \exp(\beta x) - 1, \lambda_k = \lambda_{k-1} + \alpha\left[\tilde{d}_k^T q_k\right]^+$$

Exponential Lyapunov function Benefits:

- Tracks long-term constraint violations
- Encourages adaptive safe exploration
- Penalizes positively violated constraints

Moreover, this objective function design enables us to jointly bound the cumulative regret and constraint violation:

$$\Phi(\lambda_K) + \alpha \sum_{k=1}^{K} (\tilde{r}_k^T q^* - \tilde{r}_k^T q_k) \leq \sum_{k=1}^{K} (f_k(q_k) - f_k(q^*))$$

Thus, this relationship motivates the design of algorithms aimed at minimizing regret $\sum_{k=1}^{K}(f_k(q_k) - f_k(q^*))$.

Algorithm: Optimistic Online Mirror Descent

ICML
International Conference
On Machine Learning

To minimize regret $\sum_{k=1}^{K}(f_k(q_k) - f_k(q^*))$, we adopt the Optimistic Online Mirror Descent algorithm:

- Optimistic Phase: Predicts next occupancy measure using current gradient

$$\hat{q}_{k+1} = arg\min_{q \in Q_k} \eta_k \langle q, \nabla f_k(q_k) \rangle + \mathcal{D}(q, \hat{q}_k)$$

- Refinement Phase: Refines occupancy measure using predicted $\hat{q}_{k+1}$

$$q_{k+1} = arg\min_{q \in Q_k} \eta_{k+1} \langle q, \nabla \hat{f}_{k+1}(\hat{q}_{k+1}) \rangle + \mathcal{D}(q, \hat{q}_{k+1})$$

The **optimistic update mechanism** plays a critical role in tightening performance bounds by integrating historical gradients and occupancy information. Once q_{k+1} is computed, we construct the policy π_{k+1}, execute it, and estimate the reward, cost, and transition kernel.

Theoretical Results

Algorithm	Regret	Adversarial Violation	Stochastic Violation	Slater's Condition	Known Safe Policy
[1]	$O(\sqrt{K})$	N/A	$O(\sqrt{K})$	✅	No
[2]	$O(\sqrt{K})$	N/A	$O(\sqrt{K})$	✅	Yes
[3]	$\tilde{O}(\sqrt{K})$	N/A	$\tilde{O}(\sqrt{K})$	✅	No
[4]	$\tilde{O}(K^{0.93})$	N/A	$\tilde{O}(K^{0.93})$	✅	No
[5]	$\tilde{O}(K^{\frac{6}{7}})$	N/A	$\tilde{O}(K^{\frac{6}{7}})$	✅	No
Ours	$\tilde{O}(\sqrt{K})$	$\tilde{O}(\sqrt{K})$	$\tilde{O}(\sqrt{K})$	❌	No

Our Algorithm can obtain the following theoretical results:

$$Regret(K) \leq \tilde{O}\left(\sqrt{NSAH^3K} + S^2AH^3 + \sqrt{C}\sqrt{SAHK} + SAH\right)$$

$$Violation(K) \leq \tilde{O}(\sqrt{NSAH^3K} + S^2AH^3 + \sqrt{C}\sqrt{SAHK} + SAH)(\text{both constraint settings})$$

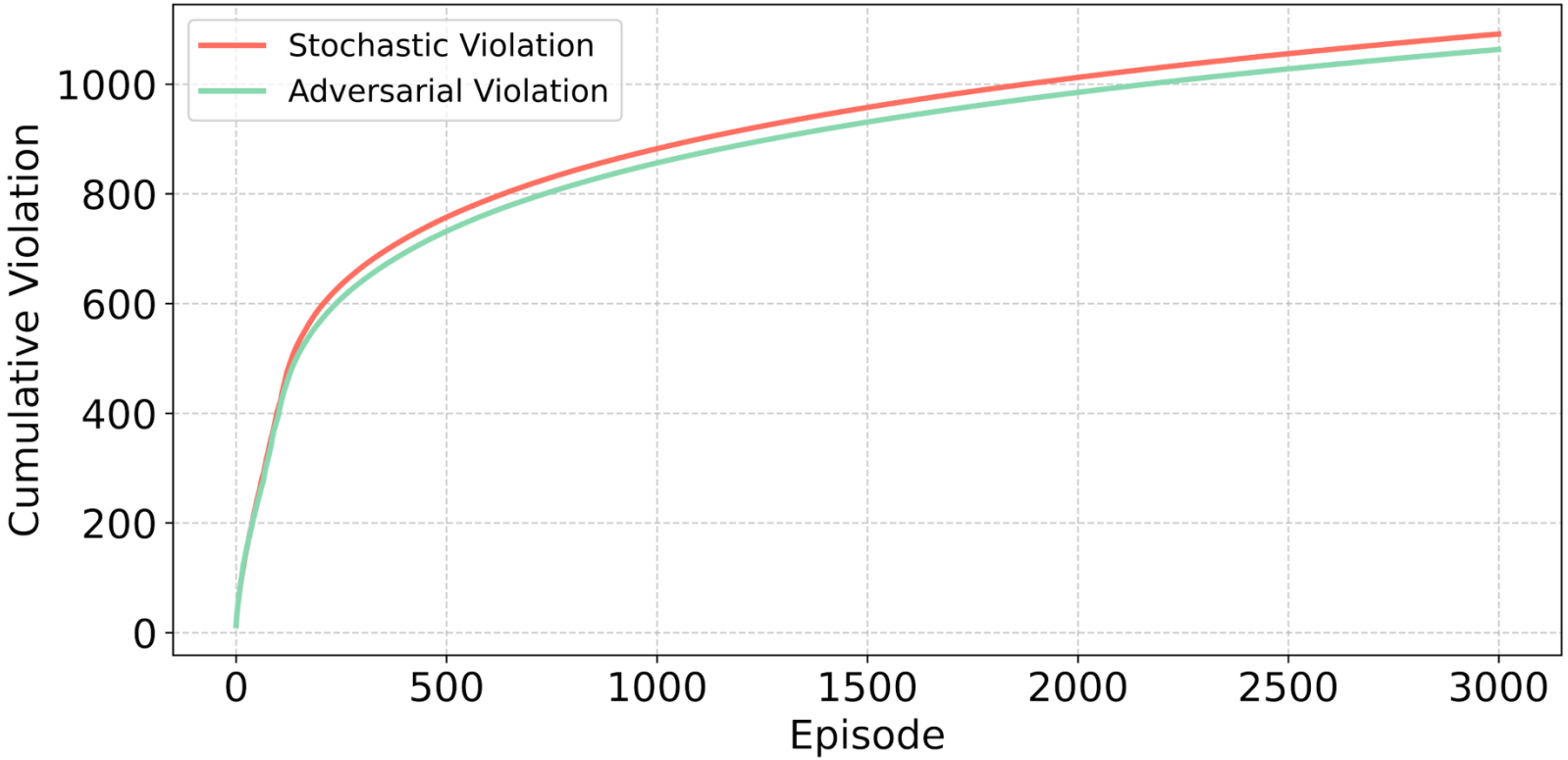
[1] Efroni, Yonathan, Shie Mannor, and Matteo Pirotta. "Exploration-exploitation in constrained mdps." arXiv preprint arXiv:2003.02189 (2020).
[2] Müller, Adrian, Pragnya Alatur, Giorgia Ramponi, and Niao He. "Cancellation-free regret bounds for lagrangian approaches in constrained markov decision processes." arXiv preprint arXiv:2306.07001 (2023).
[3] Stradi, Francesco Emanuele, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. "Optimal Strong Regret and Violation in Constrained MDPs via Policy Optimization." arXiv preprint arXiv:2410.02275 (2024).
[4] Müller, Adrian, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. "Truly no-regret learning in constrained mdps." arXiv preprint arXiv:2402.15776 (2024).
[5] Kitamura, T., Kozuno, T., Kato, M., Ichihara, Y., Nishimori, S., Sannai, A., Sonoda, S., Kumagai, W. and Matsuo, Y., 2024. A policy gradient primal-dual algorithm for constrained mdps with uniform pac guarantees. arXiv preprint arXiv:2401.17780.

Theoretical Results

Algorithm	Regret	Adversarial Violation	Stochastic Violation	Slater's Condition	Known Safe Policy
[1]	$O(\sqrt{K})$	N/A	$O(\sqrt{K})$	✅	No
[2]	$O(\sqrt{K})$	N/A	$O(\sqrt{K})$	✅	Yes
[3]	$\widetilde{O}(\sqrt{K})$	N/A	$\widetilde{O}(\sqrt{K})$	✅	No
[4]	$\widetilde{O}(K^{0.93})$	N/A	$\widetilde{O}(K^{0.93})$	✅	No
[5]	$\widetilde{O}(K^{\frac{6}{7}})$	N/A	$\widetilde{O}(K^{\frac{6}{7}})$	✅	No
Ours	$\widetilde{O}(\sqrt{K})$	$\widetilde{O}(\sqrt{K})$	$\widetilde{O}(\sqrt{K})$	❌	No

- We establish **optimal** $\widetilde{O}(\sqrt{K})$ **regret** and **constraint violation** bounds under **minimal assumptions**: **No Slater's Condition** and **No Known Safe Policy** required.
- This is the **first result** achieving optimal order in total episodes K for **online CMDPs with anytime adversarial constraints**.
- Furthermore, when a **generative model** (i.e., a perfect simulator) is available, the regret can be tightly bounded by $O(1)$.

Experiment Results



We test the algorithm in a synthetic, finite-horizon CMDP with 5 states, 3 actions, horizon H = 5, Dirichlet(α = 0.5) transitions, uniform rewards, and either stochastic costs (uniform in [-1,1]) or adversarial costs drawn each episode from {-1.0, -0.6, -0.2, 0, 0.2, 0.6, 1.0}. The initial state is chosen uniformly and the cumulative-cost constraint is fixed at 0; we measure cumulative constraint violation over K = 3000 episodes.

Conclusion

- **OMDPD algorithm**: First method to handle online safe RL with anytime adversarial constraints—no Slater condition or pre-known safe policy required.
- **Optimal guarantees**: Provably achieves $\widetilde{O}\left(\sqrt{K}\right)$ bounds on both regret and strong constraint violation.
- **Broader impact**: Advances CMDP theory and offers a robust blueprint for safe decision-making in dynamic, adversarial environments.

Paper Link: https://arxiv.org/pdf/2505.21841

Thanks for your listening !