Learning Safe Policies with Zero or Bounded Constraint Violation

Tao Liu*, Ruida Zhou*, Dileep Kalathil, P. R. Kumar, Chao Tian

TEXAS A&M UNIVERSITY®

Example: Car Racing

Objective:

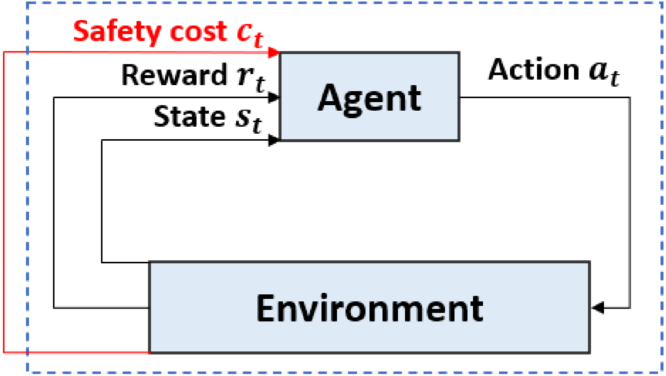
- Maximize the number of laps in a given time

Safety constraints:

- Stay in the lane

Credit to David Merrett

Constrained Markov Decision Process



$$a_t \sim \pi(\cdot | s_t)$$

Constrained MDP Problem

Value Function

For any function g (r or c),

$$V_1^\pi(\mu; g, P) = \mathbb{E}_\pi \left[\sum_{h=1}^{H} g(s_h, a_h) \;\middle|\; s_1 \sim \mu \right].$$

Find an optimal policy π^* for:

$$\begin{aligned} \max_\pi \quad & V_1^\pi(\mu; r, P) \\ s.t. \quad & V_1^\pi(\mu; c, P) \leq \tau. \end{aligned}$$

- Previous works only ensure $\mathcal{O}(\sqrt{K})$ safety cost violation.
- However, it may be important to only allow *zero* or *bounded* constraint violation:

$$\min \quad Regret(K; r) = \sum_{k=1}^{K} \left(V_1^{\pi^*}(\mu; r, P) - V_1^{\pi_k}(\mu; r, P)\right)$$

$$s.t. \quad \mathbb{P}(V_1^{\pi_k}(\mu; c, P) \leq \tau, \forall k \in [K]) \geq 1 - \delta \quad (\textit{zero} \text{ constraint violation})$$

$$\text{OR} \quad Regret(K; c) = \left(\sum_{k=1}^{K} (V_1^{\pi_k}(\mu; c, P) - \tau)\right)_+ = \mathcal{O}(1) \quad (\textit{bounded} \text{ constraint violation}).$$

Main Results

Design safe reinforcement learning algorithms that can

- keep an $\tilde{\mathcal{O}}(\sqrt{K})$ reward regret,
- guarantee *zero* or *bounded* safety constraint violation with high probability.

Comparisons for algorithms on episodic constrained MDPs

Algorithm	Regret	Constraint violation
OPDOP [Ding 2021]	$\tilde{\mathcal{O}}(H^3\sqrt{\mathcal{S}\mathcal{A}K})$	$\tilde{\mathcal{O}}(H^3\sqrt{\mathcal{S}^2\mathcal{A}K})$
OptCMDP [Efroni 2020]	$\tilde{\mathcal{O}}(H^2\sqrt{\mathcal{S}^3\mathcal{A}K})$	$\tilde{\mathcal{O}}(H^2\sqrt{\mathcal{S}^3\mathcal{A}K})$
OptCMDP-bonus [Efroni 2020]	$\tilde{\mathcal{O}}(H^2\sqrt{\mathcal{S}^3\mathcal{A}K})$	$\tilde{\mathcal{O}}(H^2\sqrt{\mathcal{S}^3\mathcal{A}K})$
OptDual-CMDP [Efroni 2020]	$\tilde{\mathcal{O}}(H^2\sqrt{\mathcal{S}^3\mathcal{A}K})$	$\tilde{\mathcal{O}}(H^2\sqrt{\mathcal{S}^3\mathcal{A}K})$
OptPrimalDual-CMDP [Efroni 2020]	$\tilde{\mathcal{O}}(H^2\sqrt{\mathcal{S}^3\mathcal{A}K})$	$\tilde{\mathcal{O}}(H^2\sqrt{\mathcal{S}^3\mathcal{A}K})$
C-UCRL [Zheng 2020]	$\tilde{\mathcal{O}}(T^{\frac{3}{4}})$	0
OptPess-LP [This work]	$\tilde{\mathcal{O}}(\frac{H^3}{\tau - c^0}\sqrt{\mathcal{S}^3\mathcal{A}K})$	0
OptPess-PrimalDual [This work]	$\tilde{\mathcal{O}}(\frac{H^3}{\tau - c^0}\sqrt{\mathcal{S}^3\mathcal{A}K})$	$\mathcal{O}(1)$

Zero constraint violation

Assumption

The agent knows

- a strictly safe policy π^0,
- its safety cost $V_1^{\pi^0}(\mu; c, P) = c^0 < \tau$.

- Objective of the agent:

$$\begin{aligned} min \quad & Regret(K; r) = \sum_{k=1}^{K} (V_1^{\pi^*}(\mu; r, P) - V_1^{\pi_k}(\mu; r, P)) \\ s.t. \quad & \mathbb{P}(V_1^{\pi_k}(\mu; c, P) \leq \tau, \forall k \in [K]) \geq 1 - \delta. \end{aligned}$$

Optimistic *Pessimism* in the Face of Uncertainty

- *Optimistic* reward estimate

$$\bar{r}_h^k(s,a) := \hat{r}_h^k(s,a) + \alpha_r \underbrace{\beta_h^k(s,a)}_{\text{confidence interval}}.$$

Scaling factor

$$\alpha_r := 1 + |\mathcal{S}|H + \frac{4H(1+|\mathcal{S}|H)}{\tau - c^0}.$$

Optimistic Pessimism in the Face of Uncertainty

- *Pessimistic* safety cost estimate

$$\underline{c}_h^k(s,a) := \hat{c}_h^k(s,a) + (1 + H|\mathcal{S}|)\beta_h^k(s,a).$$

- Choose the policy from a "*pessimistically* safe" policy set

$$\Pi^k := \begin{cases} \{\pi^0\} & \text{if } V_1^{\pi^0}(\mu; \underline{c}^k, \hat{P}^k) \geq (\tau + c^0)/2, \\ \{\pi : V_1^\pi(\mu; \underline{c}^k, \hat{P}^k) \leq \tau\} & \text{otherwise,} \end{cases}$$

where $\hat{P}^k$ is the empirical estimate of the transition.

- Use *linear programming* to determine $\pi^k \in \text{argmax}_{\pi \in \Pi^k} V_1^\pi(\mu; \overline{r}^k, \hat{P}^k)$

Lemma (Zero Constraint Violation)

Fix any $\delta \in (0,1)$. With probability at least $(1-\delta)$,

- $V_1^{\pi}(\mu; c, P) \underbrace{\leq}_{\text{pessimism}} V_1^{\pi}(\mu; \underline{c}^k, \hat{P}^k) \underbrace{\leq}_{\text{definition of } \Pi^k} \tau$ *for any k and policy $\pi \in \Pi^k$.*

- Decompose *regret of reward* as:

$$\begin{aligned} Regret(K; r) = & \sum_{k=1}^{K} \mathbb{1}(|\Pi^k| = 1) \left(V_1^{\pi^*}(\mu; r, P) - V_1^{\pi^0}(\mu; r, P) \right) & \text{(burn-in: } \mathcal{O}(1)) \\ & + \sum_{k=1}^{K} \mathbb{1}(|\Pi^k| > 1) \left(V_1^{\pi^*}(\mu; r, P) - V_1^{\pi^k}(\mu; \bar{r}^k, \hat{P}^k) \right) & \text{(optimism: } \leq 0) \\ & + \sum_{k=1}^{K} \mathbb{1}(|\Pi^k| > 1) \left(V_1^{\pi^k}(\mu; \bar{r}^k, \hat{P}^k) - V_1^{\pi^k}(\mu; r, P) \right). & \text{(UCB: } \tilde{\mathcal{O}}(\sqrt{K})) \end{aligned}$$

Theorem

Fix any $\delta \in (0,1)$. With probability at least $(1-\delta)$, OptPess-LP has

- *zero constraint violation,*
-

$$Regret(K;r) = \tilde{\mathcal{O}}\left(\frac{H^3}{\tau - c^0}\sqrt{|\mathcal{S}|^3|\mathcal{A}|K} + \underbrace{\frac{H^5|\mathcal{S}|^3|\mathcal{A}|}{(\tau - c^0)^2 \wedge (\tau - c^0)}}_{burn-in}\right).$$

Bounded Constraint Violation

Assumption

The agent

- knows that there *exists* a strictly safe policy with safety cost c^0,
- but does not know any specific strictly safe policy.

- Objective of the agent:

$$\begin{aligned} min \quad & Regret(K;r) = \sum_{k=1}^{K} \left(V_1^{\pi^*}(\mu; r, P) - V_1^{\pi_k}(\mu; r, P)\right) \\ s.t. \quad & Regret(K;c) = \left(\sum_{k=1}^{K} \left(V_1^{\pi_k}(\mu; c, P) - \tau\right)\right)_+ = \mathcal{O}(1). \end{aligned}$$

Optimistic Pessimism in the Face of Uncertainty

- *Optimistic* estimates of reward and safety cost

$$\tilde{r}_h^k(s,a) := \hat{r}_h^k(s,a) + (1 + H|\mathcal{S}|)\beta_h^k(s,a),$$
$$\tilde{c}_h^k(s,a) := \hat{c}_h^k(s,a) - (1 + H|\mathcal{S}|)\beta_h^k(s,a).$$

- Add a *pessimistic term* ϵ_k:

$$\max_{\pi} \quad V_1^{\pi}(\mu; r, P)$$
$$\text{s.t.} \quad V_1^{\pi}(\mu; c, P) + \epsilon_k \le \tau.$$

Primal-Dual Method

- *Lagrangian:*

$$L^k(\pi, \lambda) := V_1^{\pi}(\mu; r, P) + \lambda(\tau - \epsilon_k - V_1^{\pi}(\mu; c, P)).$$

- *Policy Update (Dynamic Programming):*
$\pi^k \in \text{argmax}_{\pi \in \Pi} \hat{V}_1^{\pi}(\mu; \tilde{r}^k, \hat{P}^k) - \frac{\lambda^k}{\eta^k} \left(\hat{V}_1^{\pi}(\mu; \tilde{c}^k, \hat{P}^k) - \tau \right).$

- *Dual Update:*
$\lambda^{k+1} = \left(\lambda^k + \hat{V}_1^{\pi^k}(\mu; \tilde{c}^k, \hat{P}^k) + \epsilon_k - \tau \right)_+.$

- Decompose *constraint violation* as:

$$Regret(K; c) = \left(\sum_{k=1}^{K} \left(V_1^{\pi^k}(\mu; c, P) - \hat{V}_1^{\pi^k}(\mu; \tilde{c}^k, \hat{P}^k) \right) + \sum_{k=1}^{K} \left(\hat{V}_1^{\pi^k}(\mu; \tilde{c}^k, \hat{P}^k) - \tau \right) \right)_+$$

$$\leq \left(\sum_{k=1}^{K} \left(V_1^{\pi^k}(\mu; c, P) - \hat{V}_1^{\pi^k}(\mu; \tilde{c}^k, \hat{P}^k) \right) + \lambda^{K+1} - \underbrace{\sum_{k=1}^{K} \epsilon_k}_{\text{compensate previous terms}} \right)_+ .$$

- Decompose *regret of reward* as:

$$
\begin{aligned}
Regret(K; r) = & \underbrace{\sum_{k=1}^{C''} \left(V_1^{\pi^*}(\mu; r, P) - V_1^{\pi^k}(\mu; r, P) \right)}_{\text{burn-in: } \mathcal{O}(1)} \\
& + \underbrace{\sum_{k=C''}^{K} \left(V_1^{\pi^*}(\mu; r, P) - V_1^{\pi^{\epsilon_k,*}}(\mu; r, P) \right)}_{\tilde{\mathcal{O}}(\sqrt{K})} + \underbrace{\sum_{k=C''}^{K} \left(V_1^{\pi^{\epsilon_k,*}}(\mu; r, P) - \hat{V}_1^{\pi^{\epsilon_k,*}}(\mu; \tilde{r}^k, \hat{P}^k) \right)}_{\text{optimism: } \leq 0} \\
& + \underbrace{\sum_{k=C''}^{K} \left(\hat{V}_1^{\pi^{\epsilon_k,*}}(\mu; \tilde{r}^k, \hat{P}^k) - \hat{V}_1^{\pi^k}(\mu; \tilde{r}^k, \hat{P}^k) \right)}_{\tilde{\mathcal{O}}(\sqrt{K})} + \underbrace{\sum_{k=C''}^{K} \left(\hat{V}_1^{\pi^k}(\mu; \tilde{r}^k, \hat{P}^k) - V_1^{\pi^k}(\mu; r, P) \right)}_{\text{UCB: } \tilde{\mathcal{O}}(\sqrt{K})}.
\end{aligned}
$$

Theorem

Fix any $\delta \in (0,1)$. Then, OptPess-PrimalDual has

$$Regret(K;r) = \tilde{\mathcal{O}}\left(\frac{H^3}{\tau - c^0}\sqrt{|\mathcal{S}|^3|\mathcal{A}|K} + \underbrace{\frac{H^5|\mathcal{S}|^3|\mathcal{A}|}{(\tau - c^0)^2}}_{burn-in}\right),$$

$$Regret(K;c) = \mathcal{O}\left(C''(H-\tau) + H^2\sqrt{|\mathcal{S}|^3|\mathcal{A}|C''}\right) = \mathcal{O}(1),$$

where $C'' = \mathcal{O}(\frac{H^4|\mathcal{S}|^3|\mathcal{A}|}{(\tau-c^0)^2}\log\frac{H^4|\mathcal{S}|^3|\mathcal{A}|}{(\tau-c^0)^2\delta'})$ *does not depend on* K.

Concluding remarks:

- It is possible to
  - keep an $\tilde{\mathcal{O}}(\sqrt{K})$ reward regret,
  - guarantee *zero* or *bounded* safety constraint violation under some mild assumptions.

- The general idea of *"Optimistic Pessimism in the Face of Uncertainty"* is useful for safe exploration.

Thank You