Convergent Policy Optimization for Safe Reinforcement Learning

Ming Yu, Zhuoran Yang, Mladen Kolar, Zhaoran Wang

Abstract

- **Fact:** In real-world applications of RL, we need to take into consideration the safety of the agent (constrained MDP)
- **Challenge:** Both the objective and constraint function are nonconvex and involve expectation without closed form expression
- **Algorithms:** Optimize a sequence of convex relaxation problems, motivated by [1]
- **Theoretical result:** Convergence of subsequence to a stationary point almost surely
- **Extension:** Actor-Critic method and parallel / multi-agent RL problem with safety constraint

Optimization

- Constrained MDP (CMDP) setup (D_0 is a constant):

$$\begin{aligned} \underset{\theta\in\Theta}{\text{minimize}} \quad & J(\theta) = \mathbb{E}_{\pi_\theta}\Big[-\sum_{t\geq 0}\gamma^t \cdot r(s_t,a_t)\Big], \\ \text{subject to} \quad & D(\theta) = \mathbb{E}_{\pi_\theta}\Big[\sum_{t\geq 0}\gamma^t \cdot d(s_t,a_t)\Big] \leq D_0 \end{aligned} \tag{1}$$

- In each iteration k at θ_k, we sample a trajectory and obtain sample reward and constraint value:

$$J^*(\theta_k) = -\sum_t \gamma^t \cdot r(s_t,a_t) \qquad \text{and} \qquad D^*(\theta_k) = \sum_t \gamma^t \cdot d(s_t,a_t)$$

and their gradients $\nabla_\theta J^*(\theta_k)$ and $\nabla_\theta D^*(\theta_k)$

- We approximate $J(\theta)$ and $D(\theta)$ at θ_k by the quadratic surrogates:

$$\widetilde{J}(\theta,\theta_k,\tau) = J^*(\theta_k) + \langle \nabla_\theta J^*(\theta_k), \theta-\theta_k\rangle + \tau\|\theta-\theta_k\|_2^2,$$
$$\widetilde{D}(\theta,\theta_k,\tau) = D^*(\theta_k) + \langle \nabla_\theta D^*(\theta_k), \theta-\theta_k\rangle + \tau\|\theta-\theta_k\|_2^2.$$

where $\tau > 0$ is any fixed constant and

$$\overline{J}^{(k)}(\theta) = (1-\rho_k)\cdot \overline{J}^{(k-1)}(\theta) + \rho_k \cdot \widetilde{J}(\theta,\theta_k,\tau),$$
$$\overline{D}^{(k)}(\theta) = (1-\rho_k)\cdot \overline{D}^{(k-1)}(\theta) + \rho_k \cdot \widetilde{D}(\theta,\theta_k,\tau).$$

- In each iteration, we solve

$$\overline{\theta}_k = \arg\min_\theta \overline{J}^{(k)}(\theta) \qquad \text{subject to} \qquad \overline{D}^{(k)}(\theta) \leq D_0, \tag{2}$$

if it is feasible; otherwise we solve the feasibility problem

$$\overline{\theta}_k = \underset{\theta,\alpha}{\arg\min} \ \alpha \qquad \text{subject to} \qquad \overline{D}^{(k)}(\theta) \leq D_0 + \alpha. \tag{3}$$

- Update θ_k by

$$\theta_{k+1} = (1-\eta_k)\cdot\theta_k + \eta_k \cdot \overline{\theta}_k, \tag{4}$$

Algorithm

Algorithm 1 Successive convex relaxation algorithm for CMDP

```
Input: Initial value θ_0, τ, {ρ_k}, {η_k}.
for k = 1, 2, 3, ... do
    Obtain sample J*(θ_k), D*(θ_k) by Monte-Carlo sampling.
    Obtain sample ∇_θ J*(θ_k), ∇_θ D*(θ_k) by policy gradient theorem.
    if problem (2) is feasible then
        Obtain θ̄_k by solving (2).
    else
        Obtain θ̄_k by solving (3).
    end if
    Update θ_{k+1} by (4).
end for
```

Assumptions

- **Assumption 1. [Step size]** We have $\lim_{k\to\infty}\sum_k \eta_k = \infty$, $\lim_{k\to\infty}\sum_k \rho_k = \infty$ and $\lim_{k\to\infty}\sum_k \eta_k^2 + \rho_k^2 < \infty$. Furthermore, we have $\lim_{k\to\infty}\eta_k/\rho_k = 0$ and η_k is decreasing.
- **Assumption 2. [Smooth objective and constraint]** For any realization, $J^*(\theta)$ and $D^*(\theta)$ are continuously differentiable as functions of θ. Moreover, $J^*(\theta)$, $D^*(\theta)$, and their derivatives are uniformly Lipschitz continuous.

Theoretical result

- **Theorem.** Suppose Assumptions 1 and 2 are satisfied, and θ_0 is a feasible point. For subsequence $\{\theta_{k_j}\}$ of $\{\theta_k\}$ that converges to some $\widetilde{\theta}$, there exist $\widehat{J}(\theta)$ and $\widehat{D}(\theta)$ satisfying

$$\lim_{j\to\infty}\overline{J}^{(k_j)}(\theta) = \widehat{J}(\theta) \qquad \text{and} \qquad \lim_{j\to\infty}\overline{D}^{(k_j)}(\theta) = \widehat{D}(\theta).$$

Suppose there exists θ such that $\widehat{D}(\theta) < D_0$ (i.e. the Slater's condition holds), then $\widetilde{\theta}$ is a **stationary point** of the original problem (1) almost surely.

- If θ_0 is not feasible, then the following Assumption 3 is needed to exclude convergence to undesired stationary point
- **Assumption 3.** Suppose (θ_S, α_S) is a stationary point of the optimization problem

$$\underset{\theta,\alpha}{\text{minimize}} \quad \alpha \qquad \text{subject to} \qquad D(\theta) \leq D_0 + \alpha.$$

We have that θ_S is a feasible point of the original problem.

Application to linear quadratic regulator (LQR)

- We consider the infinite-horizon, discrete-time LQR problem.
- Denote x_t as the state and u_t as the control. We have the state transition and the control sequence

$$x_{t+1} = Ax_t + Bu_t + v_t,$$
$$u_t = -Fx_t + w_t$$

- Random initial state $x_0 \sim \mathcal{D}$
- Optimization problem with parameter F:

$$\begin{aligned} \text{minimize} \quad & J(F) = \mathbb{E}\Big[\sum_{t\geq 0} x_t^\top Q_1 x_t + u_t^\top R_1 u_t\Big], \\ \text{subject to} \quad & D(F) = \mathbb{E}\Big[\sum_{t\geq 0} x_t^\top Q_2 x_t + u_t^\top R_2 u_t\Big] \leq D_0. \end{aligned}$$

Experiment

- Constraint values and objective values for one realization

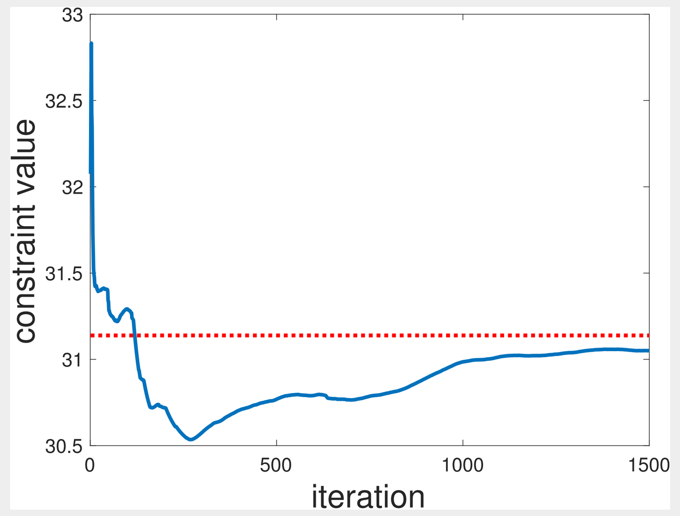

Figure: Constaint value $D(\theta_k)$

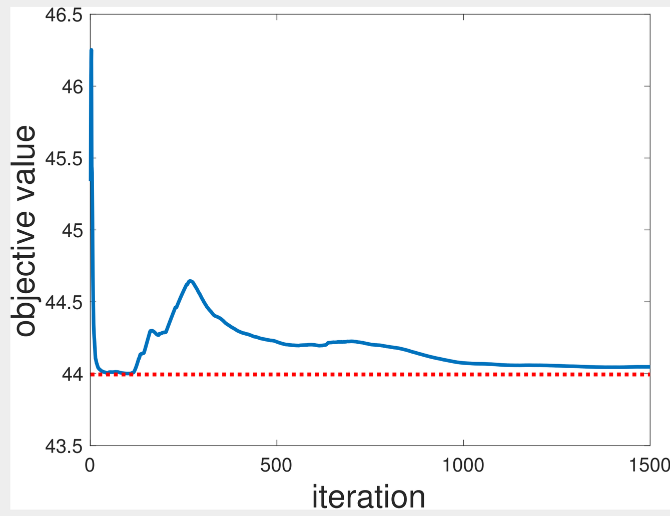

Figure: Objective value $J(\theta_k)$

- Compare with Lagrangian method (50 replicates)

	min value	# iterations	approx # iterations
Our method	30.689 ± 0.114	2001 ± 1172	604.3 ± 722.4
Lagrangian	30.693 ± 0.113	7492 ± 1780	5464 ± 2116

Table: Comparison of our method with Lagrangian method

- Our method requires less number of policy updates
- Code available at `https://github.com/ming93/Safe_reinforcement_learning`

Reference

[1] An Liu, Vincent Lau, and Borna Kananian. Stochastic successive convex approximation for non-convex constrained stochastic optimization. *IEEE Transactions on Signal Processing, 2019*

ming93@uchicago.edu