Constrained Reinforcement Learning

Trust Regions and Primal-Dual Mirror Descent

Yinbin Han
MS&E, Stanford University
email: yinbinha@stanford.edu

MS&E 318: Constrained and Safe AI

Jan 27, 2026

Overview

- Achiam, Joshua, et al. "Constrained policy optimization." International conference on machine learning (2017).
- Chen, Yi, et al. "A primal-dual approach to constrained markov decision processes with applications to queue scheduling and inventory management." Management Science (2025).

Overview

- Achiam, Joshua, et al. "Constrained policy optimization." International conference on machine learning (2017).
- Chen, Yi, et al. "A primal-dual approach to constrained markov decision processes with applications to queue scheduling and inventory management." Management Science (2025).

Roadmap:

- Constrained Markov decision processes
- Trust region policy optimization for CMDP
- Primal-dual mirror descent for CMDP
- Applications to queue scheduling and inventory management

Motivation: Why constraints in RL?

- **Safety-critical systems:** In robotics and industrial control (e.g., datacenter cooling), violating physical limits is catastrophic
  Achiam et al. '17; Dalal et al. '18
- **Operations management:** In hospital scheduling or inventory control, it is often hard to determine relative weights for different objectives (e.g., waiting time vs. staffing cost)
  Girard et al. '20; Singh and Cohn '98
- **Constrained sampling and alignment in GenAI:** Simply maximizing reward leads to reward hacking; hard safety constraints are required in LLMs generation; structural constraints for protein design
  Liu et al. '24; Christopher et al. '25

Constrained MDP

- infinite-horizon discounted CMDP
  - S: state space
  - A: action space
  - $R : S \times A \times S \to \mathbb{R}$: reward functuion
  - $P : S \times A \times S \to \mathbb{R}$: transition probability
  - μ: initial state distribution
  - γ: discounting factor

Constrained MDP

- infinite-horizon discounted CMDP
  - S: state space
  - A: action space
  - $R : S \times A \times S \to \mathbb{R}$: reward functuion
  - $P : S \times A \times S \to \mathbb{R}$: transition probability
  - μ: initial state distribution
  - γ: discounting factor
  - $C : S \times A \times S \to \mathbb{R}^m$: auxiliary cost functions
  - $d \in \mathbb{R}^m$: constraint limits
- **Reward Objective:** $J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})\right]$
- **Safety Constraints:** $J_{C_i}(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t C_i(s_t, a_t, s_{t+1})\right] \leq d_i$ for $i = 1, \ldots, m$
- **Goal:**

$$\pi^* = \arg\max_{\pi \in \Pi} J(\pi) \quad \text{s.t. } \forall i, J_{C_i}(\pi) \leq d_i$$

Taxonomy of CMDP Algorithms

- **Trust Region Methods (CPO):** Guarantees near-constraint satisfaction at each iteration via local approximations
  Schulman et al. '15; Schulman et al. '17; Achiam et al., '17
- **Primal-Dual (PDO):** Utilizes Lagrangian relaxation $\mathcal{L}(\pi, \lambda) = J(\pi) - \sum \lambda_i (J_{C_i} - d_i)$ and updates multipliers via subgradient ascent
  Chow et al. '15; Ding et al. '25; Dong et al. '25
- **Lyapunov Approaches:** Construct scalar potential functions to guarantee global safety through local linear constraints
  Chow et al. '18; Xiao et al. '23
- **Safety Layers:** Directly add an analytical action correction step to the policy to ensure zero violation for all visited states Dalal et al. '18
- **Interior-Point Policy Optimization (IPO):** Augments the objective with logarithmic barrier functions $\phi(x) = -\log(-x)/t$ Liu et al. '19

TRPO for CMDP

- Local search methods work well for MDPs, e.g., TRPO, PPO
- Add constraints to TRPO formulation

$$\pi_{k+1} = \arg\max_{\pi \in \Pi_\theta} \mathbb{E}[A^{\pi_k}(s,a)]$$
$$\text{s.t. } \mathrm{KL}(\pi \| \pi_k) \leq \delta$$
$$J_{C_i}(\pi) \leq \square \leq d_i, \forall i$$

TRPO for CMDP

- Local search methods work well for MDPs, e.g., TRPO, PPO
- Add constraints to TRPO formulation

$$\pi_{k+1} = \arg\max_{\pi \in \Pi_\theta} \mathbb{E}[A^{\pi_k}(s,a)]$$
$$\text{s.t. } \mathrm{KL}(\pi \| \pi_k) \leq \delta$$
$$J_{C_i}(\pi) \leq \square \leq d_i, \forall i$$

Question: A computationally tractable form for $\square$?

Proposition (Informal)

For any two policies π and π', it holds

$$J_{C_i}(\pi') \leq J_{C_i}(\pi) + \frac{1}{1-\gamma}\mathbb{E}[A^{\pi}_{C_i}(s,a) + \mathrm{KL}(\pi' \| \pi)]$$

Constrained policy optimization

- CPO formulation:

$$\pi_{k+1} = \arg\max_{\pi \in \Pi_\theta} \mathbb{E}[A^{\pi_k}(s,a)]$$
$$\text{s.t. } \mathrm{KL}(\pi \| \pi_k) \leq \delta$$
$$J_{C_i}(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}[A^{\pi}_{C_i}(s,a)] \leq d_i, \forall i$$

- CPO guarantees approximate feasibility

Constrained policy optimization

- CPO formulation:

$$\pi_{k+1} = \arg\max_{\pi \in \Pi_\theta} \mathbb{E}[A^{\pi_k}(s,a)]$$
$$\text{s.t. } \mathrm{KL}(\pi \| \pi_k) \leq \delta$$
$$J_{C_i}(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}[A^{\pi}_{C_i}(s,a)] \leq d_i, \forall i$$

- CPO guarantees approximate feasibility

Proposition (Informal)

Let $\{\pi_k\}$ be generated by CPO. It holds that

$$J_{C_i}(\pi_{k+1}) \leq d_i + \textit{error}$$

CPO: The Optimization Step

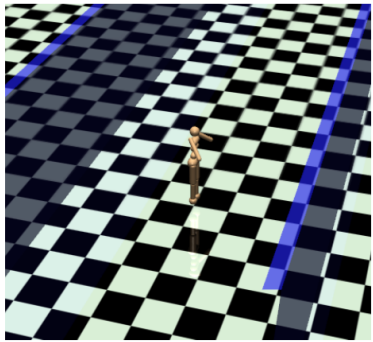
CPO solves a Linearly-Quadratic Constrained Linear Program (LQCLP) at each iteration k:

$$\begin{aligned} \theta_{k+1} = \arg\max_{\theta} \quad & g^T(\theta - \theta_k) \\ \text{s.t.} \quad & J_{C_i}(\pi_k) - d_i + b_i^T(\theta - \theta_k) \leq 0 \quad i = 1, \ldots, m \\ & \frac{1}{2}(\theta - \theta_k)^T H(\theta - \theta_k) \leq \delta \end{aligned}$$

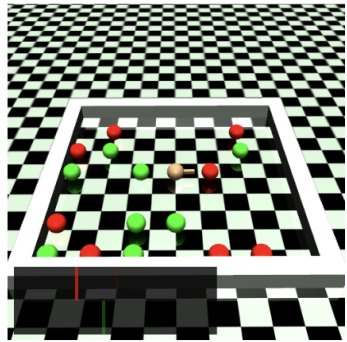
where g is the policy gradient, H is the Fisher Information Matrix, and b_i is the gradient of the i-th constraint.

- **Pros:** efficiently solvable in the dual space when $m \ll \dim(\theta)$
- **Cons:** Heuristic; approximation step; backtracking line search

CPO experiments



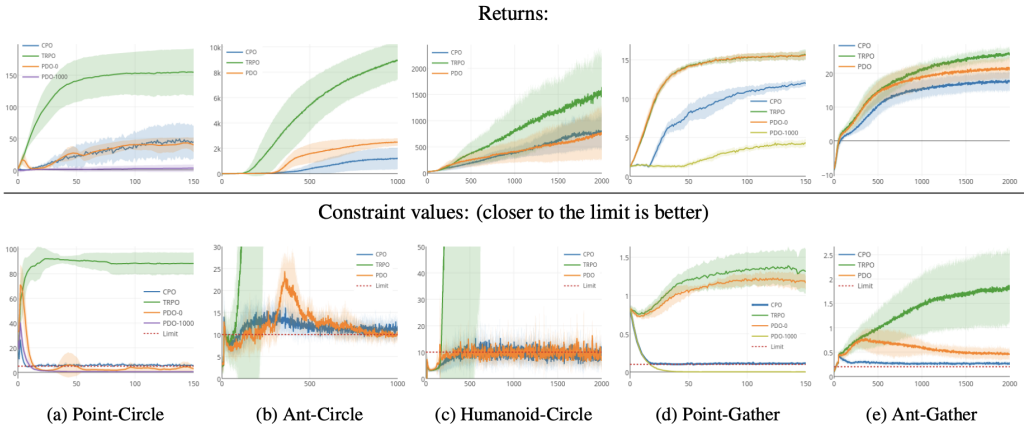
(a) Humanoid-Circle



(b) Point-Gather

- **Circle:** Running in a wide circle, constrained to stay within a safe region
- **Gather:** Collecting green apples, and constrained to avoid red bombs

Numerical results



- CPO pushes constraints to be satisfied faster than PDO
- Return is not sacrified too much by CPO

LP formulation of CMDP

- CPO provides a heuristic for CMDP (Q: When does it fail?)
- A rigorous algorithm with convergence guarantees
- Connection to static constrained optimization

LP formulation of CMDP

- CPO provides a heuristic for CMDP (Q: When does it fail?)
- A rigorous algorithm with convergence guarantees
- Connection to static constrained optimization

$$\min_{\nu} \sum_{s,a} R(s,a) \cdot \nu(s,a)$$

$$s.t. \sum_{s,a} C_i(s,a) \cdot \nu(s,a) \leq d_i, \ \forall i = 1, \ldots, m$$

$$\sum_{s,a} \nu(s,a) = 1, \nu(s,a) \geq 0$$

a linear one-step equation in ν,

where $\nu = \nu^{\pi}$ is state-action occupation measure induced by π

Lagrangian formulation

- Primal or dual forms are challenging for large-scale problems and/or unknown $P(\cdot|s,a)$
- Primal-dual methods have excel in constrained optimization
- Lagrangian:

$$L(\pi,\lambda) = \underbrace{J(\pi)}_{\text{expected cost}} + \sum_{i=1}^{m} \underbrace{\lambda_i}_{\text{multiplier}} \underbrace{(J_{C_i}(\pi) - d_i)}_{\text{constraint}}$$

Lagrangian formulation

- Primal or dual forms are challenging for large-scale problems and/or unknown $P(\cdot|s,a)$
- Primal-dual methods have excel in constrained optimization
- Lagrangian:

$$L(\pi,\lambda) = \underbrace{J(\pi)}_{\text{expected cost}} + \sum_{i=1}^{m} \underbrace{\lambda_i}_{\text{multiplier}} \underbrace{(J_{C_i}(\pi) - d_i)}_{\text{constraint}}$$

- Bilinear in π and ν and strong duality (Slater):

$$\inf_{\pi} \sup_{\lambda \geq 0} L(\pi,\lambda) = L(\pi^*,\lambda^*) = \sup_{\lambda \geq 0} \inf_{\pi} L(\pi,\lambda)$$

Minimax optimization

- Strong duality:

$$\inf_{\pi} \sup_{\lambda \geq 0} L(\pi, \lambda) = L(\pi^*, \lambda^*) = \sup_{\lambda \geq 0} \inf_{\pi} L(\pi, \lambda)$$

- $\inf_\pi L(\pi, \lambda)$ is an **unconstrained** MDP problem for given λ
- Double-loop algorithms work but may be inefficient

Minimax optimization

- Strong duality:

$$\inf_{\pi} \sup_{\lambda \geq 0} L(\pi, \lambda) = L(\pi^*, \lambda^*) = \sup_{\lambda \geq 0} \inf_{\pi} L(\pi, \lambda)$$

- $\inf_\pi L(\pi, \lambda)$ is an **unconstrained** MDP problem for given λ
- Double-loop algorithms work but may be inefficient
- Entropy-regularized policy optimization/policy mirror descent work well for unconstrained RL

  Cen et al. '22; Lan '21; Zhan et al. '23
- Proximal point/gradient methods are widely used in minimax optimization

  Lin, Jin and Jordan '20, '25;

Minimax optimization

- Strong duality:

$$\inf_{\pi} \sup_{\lambda \geq 0} L(\pi, \lambda) = L(\pi^*, \lambda^*) = \sup_{\lambda \geq 0} \inf_{\pi} L(\pi, \lambda)$$

- $\inf_\pi L(\pi, \lambda)$ is an **unconstrained** MDP problem for given λ
- Double-loop algorithms work but may be inefficient
- Entropy-regularized policy optimization/policy mirror descent work well for unconstrained RL

  Cen et al. '22; Lan '21; Zhan et al. '23
- Proximal point/gradient methods are widely used in minimax optimization

  Lin, Jin and Jordan '20, '25;

Solution: Primal-dual mirror descent

Primal-dual mirror descent

- **Primal policy update:**

$$\pi_k(\cdot|s) = \arg\min_{\pi} \left\{ \left\langle Q^{\pi_{k-1},\lambda_{k-1}}(,\cdot), \pi(\cdot|s) \right\rangle + \eta_{k-1}^{-1} \mathrm{KL}(\pi \| \pi_{k-1}) \right\}$$

- Closed-form solution:

$$\pi_k(a|s) \propto \pi_{k-1}(a|s) \exp\{-\eta_{k-1} Q^{\pi_{k-1},\lambda_{k-1}}(s,a)\}$$

- **Dual multiplier update:**

$$\lambda_k = \mathrm{Proj}\{\lambda_{k-1} + \eta_{k-1} \partial_\lambda L(\pi_{k-1}, \lambda_{k-1})\}$$

- **Key advantages:** A single-loop algorithm; convergence inherits from policy mirror descent and minimax optimization

Convergence guarantees

Theorem

Under mild conditions, if we choose constant step size $\eta_k \asymp 1/\sqrt{K}$*, it holds*

$$\left\|[J_C(\bar{\pi}_K) - d]^+\right\| = O(1/\sqrt{K}), \quad \|J(\bar{\pi}_K) - J(\pi^*)\| = O(1/\sqrt{K}),$$

where $\bar{\pi}_K$ *is a weighted average of all generated policies.*

Convergence guarantees

Theorem

Under mild conditions, if we choose constant step size $\eta_k \asymp 1/\sqrt{K}$, it holds

$$\left\|[J_C(\bar{\pi}_K) - d]^+\right\| = O(1/\sqrt{K}), \quad \|J(\bar{\pi}_K) - J(\pi^*)\| = O(1/\sqrt{K}),$$

where $\bar{\pi}_K$ is a weighted average of all generated policies.

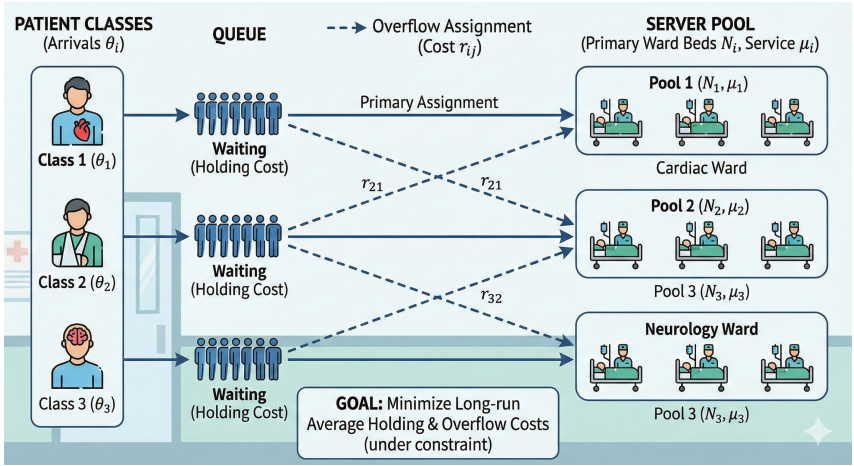
Remarks:

- Same convergence holds for diminishing step size $\eta_k = O(1/\sqrt{k})$ and long-run average CMDPs
- Convergence remains true with value and policy functional approximation, e.g., linear approximation

Queue scheduling setups

- I classes of customers (patients), I pools of primary servers (primary ward beds)
- Pool i has N_i homogeneous servers
- Class i arrives follow Poisson dist with rate θ_i
- Servers in pool i follows Geo(μ_i)
- Waiting in queue incurs holding costs
- Assigning patient i to $j \neq i$ incurs overflow cost r_{ij}
- **Goal:** Minimize long-run average holding cost with long-run average overflow under constraint

Inpatient overflow example



Queue scheduling numerical results

Table 1 **Performance of different policies for a 2×2 queuing network scheduling problem, scenario 1**

	benchmark type	$\Theta = 1$	$\Theta = 2$	$\Theta = 3$	$\Theta = 4$	optimal mixing	primal-dual
holding cost	$c\mu$	60.9	45.9	31.2	24.2	39.6	**32.6**
	max-pressure	60.8	46.2	32.1	24.5		
overflow cost	$c\mu$	0.87	1.51	1.83	1.96	1.50	1.50
	max-pressure	0.87	1.52	1.84	1.96		

Table 2 **Performance of different policies for a 2×2 queueing network scheduling problem, scenario 2**

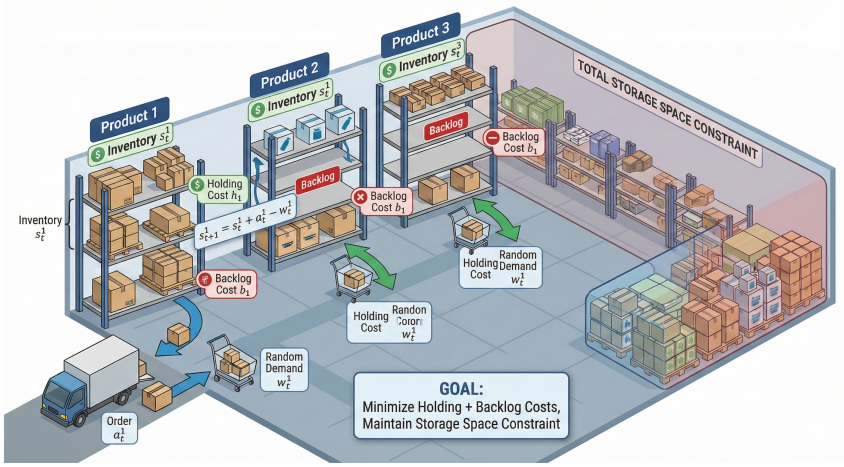
	benchmark type	$\Theta = 0$	$\Theta = 1$	$\Theta = 2$	$\Theta = 3$	optimal mixing	primal-dual
holding cost	$c\mu$	19.2	12.3	10.4	9.4	12.7	**11.4**
	max-pressure	18.7	12.1	10.2	9.5		
overflow cost	$c\mu$	0	0.33	0.50	0.56	0.30	0.30
	max-pressure	0	0.33	0.49	0.56		

- Constraints satisfied, holding cost reduced

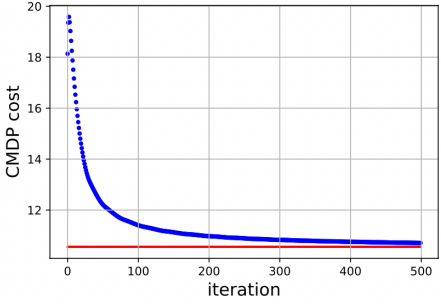
Inventory management setups

- Multi-product newsvendor problem with storage space constraint
- Inventory level s_t^i for $i \in [I]$
- Quantity we order a_t^i
- Random demand w_t^i
- Transition dynamics: $s_{t+1}^i = s_t^i + a_t^i - w_t^i$ (negative is allowed)
- Holding cost h_i; backlog cost b_i
- **Goal:** Minimize holding + backlog costs; maintain storage space constraint

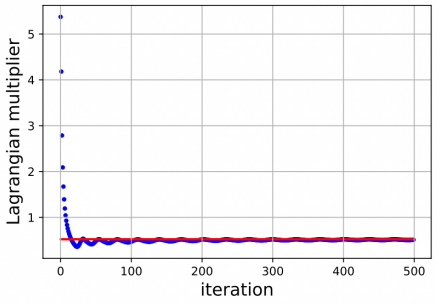
Newvendor example



Newvendor problem numerical results



(a) $\sum_{t=0}^{T-1} \tilde{\eta}_t C(\pi_t)$



(b) $\sum_{t=0}^{T-1} \tilde{\eta}_t \lambda_t$

Figure 4 **Trajectories of the objective and Lagrangian multiplier when $\eta_m = 0.5$**

Dicussion questions

1. Is it possible to improve iteration complexity of primal-dual methods?
2. Can we apply CPO to queue scheduling and inventory management? Any robustness concerns?
3. Which paper/algorithm do you like? An easy-to-implement heuristic algorithm or a mathematically correct but heavy one
4. Other formulations of constrained RL? Any applications?
5. Connection between dynamic optimization and static optimization? Can we apply algorithms from one to the other?