On the Optimal Deterministic Policy Learning in Chance-Constrained Markov Decision Processes

Hongyu Yi [1,2] Chenbei Lu [3,4] Chenye Wu [1]

[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
[2]Department of Electrical and Computer Engineering, University of Washington
[3]Institute for Interdisciplinary Information Sciences, Tsinghua University
[4]Cornell University AI for Science Institute, Cornell University

Introduction and Motivation

- MDPs are widely used for sequential decision-making under uncertainty. Yet in many real systems, good expected performance alone is not enough: one must also ensure that unsafe events occur only with very small probability;
- CCMDPs capture this requirement by imposing probabilistic constraints on cumulative costs. In contrast to standard constrained MDPs, feasibility depends on the *full distribution* of long-term cost, and the feasible policy set is generally nonconvex;
- Most existing methods rely on surrogate risk measures such as CVaR, which may be conservative and do not solve the original CCMDP directly. This motivates the following question:

Can we directly learn the optimal deterministic policy that satisfies the chance constraints?

- This work answers this question positively by developing a model-based learning framework with provable finite-sample guarantees for feasible-set identification and near-optimal deterministic policy learning.

Problem Setup

Consider an infinite-horizon discounted CCMDP with finite state space $\mathcal{S}$, action space $\mathcal{A}$, reward function r, discount factor $\gamma \in [0,1)$, and chance constraints

$$\mathbb{P}^{\pi}\left(\sum_{t=0}^{\infty}\gamma^t c_i(s_t,a_t) \le d_i \;\middle|\; s_0 = s^{[i]}\right) \ge 1-\delta_i, \qquad i \in \mathcal{C}. \tag{1}$$

We seek the optimal deterministic feasible policy satisfying (1):

$$\pi^* \in \arg\max_{\pi\in\Pi_{\text{det}}^*} J(\pi), \tag{2}$$

where

$$J(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]. \tag{3}$$

Key challenge: feasibility depends on the *entire distribution* of cumulative discounted cost, and CCs make the problem nonconvex.

Three-Stage Learning Algorithm

Stage 1: Transition estimation

For each state-action pair (s,a), collect samples and build an empirical transition kernel $\hat{P}$ using $D = N|\mathcal{S}||\mathcal{A}|$ samples.

Stage 2: Offline policy simulation

For each deterministic policy π and each constraint i, define $\hat{x}_{\pi,i,k} = \sum_{t=0}^{T}\gamma^t c_i\left(\tilde{s}_t^{\pi,i,k}, \tilde{a}_t^{\pi,i,k}\right)$, and $(\tilde{s}_t^{\pi,i,k}, \tilde{a}_t^{\pi,i,k})$ the k^{th} trajectory under π, $\hat{P}$, simulate trajectories under $\hat{P}$ and estimate:

$$\hat{p}_{\pi,i}^{N,T,M} = \frac{1}{M}\sum_{k=1}^{M}\mathbf{1}\{\hat{x}_{\pi,i,k} \le d_i\}. \tag{4}$$

Stage 3: Feasibility screening

Keep policy π if $\hat{p}_{\pi,i}^{N,T,M} \ge 1-\delta_i,\ \forall i \in \mathcal{C}$. Then evaluate reward over the estimated feasible deterministic policy set and return the best one.

Essential Definitions

For each deterministic policy $\pi \in \Pi_{\text{det}}$ and constraint i, define the discounted cumulative cost $x_{\pi,i} := \sum_{t=0}^{\infty}\gamma^t c_i(s_t,a_t)$.

Definition 1 (Chance-constraint identification ambiguity).

$$\Delta_{\pi,i} := \left|\mathbb{P}^{\pi}\left(x_{\pi,i} \le d_i \mid s_0 = s^{[i]}\right) - (1-\delta_i)\right|. \tag{5}$$

This measures how close policy π is to the chance-constraint boundary.

Definition 2 (Distributional regularity). Let $\mu_{\pi,i}$ be the distribution of $x_{\pi,i}$.

- **Maximal local density:** if $f_{\pi,i}$ denotes the density of the absolutely continuous part of $\mu_{\pi,i}$, define:

$$f_{\max}^* := \max_{i\in[|\mathcal{C}|],\,\pi\in\Pi_{\text{det}}} \sup_x f_{\pi,i}(x). \tag{6}$$

- **Quantile ambiguity gap:** let:

$$q_{1-\delta_i}^{\pi,i} := \inf\{x\in\mathbb{R} : \mu_{\pi,i}((-\infty,x]) \ge 1-\delta_i\}, \tag{7}$$

and let $D_{\pi,i} := \{x\in\mathbb{R} : \mu_{\pi,i}(\{x\}) > 0\}$ be the set of atoms of $\mu_{\pi,i}$. Define:

$$d_{\text{gap}}^{\pi,i} := \inf_{x\in D_{\pi,i}} |x - q_{1-\delta_i}^{\pi,i}|, \qquad d_{\text{gap}}^{*} := \min_{\pi\in\Pi_{\text{det}},\, i\in[|\mathcal{C}|]} d_{\text{gap}}^{\pi,i}.$$

These quantities characterize the statistical hardness of feasibility identification in CCMDPs.

Main Results

The probability estimation error is decomposed into three terms:

$$\hat{p}_{\pi,i}^{N,T,M} - p_{\pi,i}^* = \kappa_{\pi,i}^{(1)} + \kappa_{\pi,i}^{(2)} + \kappa_{\pi,i}^{(3)}, \tag{8}$$

where:

- $\kappa_{\pi,i}^{(1)}$ is model estimation error,
- $\kappa_{\pi,i}^{(2)}$ is Monte Carlo sampling error,
- $\kappa_{\pi,i}^{(3)}$ is truncation error from finite horizon T.

The bounds scale as

$$|\kappa_{\pi,i}^{(1)}| \lesssim \frac{f_{\max}^*}{1-\gamma}\sqrt{\frac{\log(\cdot)}{N}}, \qquad |\kappa_{\pi,i}^{(2)}| \lesssim \sqrt{\frac{\log(\cdot)}{M}}, \tag{9}$$

$$|\kappa_{\pi,i}^{(3)}| \lesssim \frac{f_{\max}^*\gamma^{T+1}}{1-\gamma}. \tag{10}$$

Feasible-set identification with high probability $1-\delta$:

$$D = O\left(\frac{(f_{\max}^*)^2|\mathcal{S}|^2|\mathcal{A}|\log\left(|\mathcal{S}||\mathcal{A}||\mathcal{C}|/(\delta(1-\gamma))\right)}{\Delta_{\min}^2(1-\gamma)^4}\right). \tag{11}$$

Thus, larger N, M, and T respectively improve model estimation, probability estimation, and reduce truncation bias.

Simulation

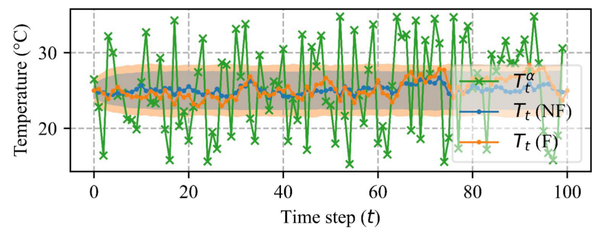
The discrete-time dynamics of the TCLs:

$$T_{t+1} = \alpha T_t + (1-\alpha)\left(T_t^{\alpha} - Ru_t\right) + \epsilon,\ \forall t, \tag{12}$$

subject to:

$$P\left(\sum_{t=1}^{\infty}\gamma_0^t c_0(T_t,u_t) \le Y_0\right) \ge 1-\delta_T, \tag{13}$$

$$-u_{\text{set}} \le u_t \le u_{\text{set}}. \tag{14}$$



References

[1] H. Yi, C. Lu, and C. Wu, "On the optimal deterministic policy learning in chance-constrained markov decision processes," *IEEE Control Systems Letters*, vol. 9, pp. 2217–2222, 2025.