Distributionally robust chance constrained Markov decision process

Jia Liu

Xi'an Jiaotong University

TEL:086-82663166, E-mail: jialiu@xjtu.edu.cn

Joint work with Zhiping Chen, Tian Xia (Xi'an Jiaotong University) and Abdel Lisser (CentraleSupelec, France)

(2023.4.30, Nanning)

Outline

- Introduction to MDP
- Reformulation of K-L divergence based DRCCMDP
- Reformulation of moment-based DRCCMDP
- Dynamical neural network approach for DRCCMDP
- Numerical results on achine replacement problem
- Conclusion

Introduction

- Markov decision processes (MDP) formally describe an environment for reinforcement learning
- Where the environment is fully observable
- i.e. The current state completely characterises the process
- Almost all RL problems can be formalised as MDPs, e.g. Optimal control primarily deals with continuous MDPs; Partially observable problems can be converted into MDPs; Bandits are MDPs with one state
- A state s_t is Markov if and only if

$$P[s_{t+1} \mid s_t] = P[s_{t+1} \mid s_1; \cdots ; s_t]$$

Introduction: MDP

We consider an infinite horizon Markov decision process (MDP) as a tuple $(\mathcal{S}, \mathcal{A}, P, r_0, q, \alpha)$, where:

- $\mathcal{S}$ is a finite state space with $|S|$ states whose generic element is denoted by s .
- $\mathcal{A}$ is a finite action space with $|\mathcal{A}|$ actions and $a \in \mathcal{A}(s)$ denotes an action belonging to the set of actions at state s.
- $P \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{S}|}$ is the distribution of transition probability $p(\overline{s}|s,a)$, which denotes the probability of moving from state s to $\overline{s}$ when the action $a \in \mathcal{A}(s)$ is taken.
- $r_0(s,a)_{s\in\mathcal{S},a\in\mathcal{A}(s)} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ denotes a running reward, which is the reward at the state s when the action a is taken. $r_0 = (r_0(s,a))_{s\in\mathcal{S},a\in\mathcal{A}(s)} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ is the running reward vector.
- $q = (q(s_0))_{s_0\in\mathcal{S}}$ is the probability of the initial state s_0.
- α is the discount factor which satisfies $\alpha \in [0, 1)$.

Introduction: setting of MDP

We consider a discrete time controlled Markov chain $(s_t, a_t)_{t=0}^{\infty}$ defined on the state space $\mathcal{S}$ and the action space $\mathcal{A}$, where s_t and a_t are the state and action at time t, respectively.

- define policy $\pi = (\mu(a|s))_{s\in\mathcal{S}, a\in\mathcal{A}(s)} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ where $\mu(a|s)$ denotes the probability that the action a is taken at state s,
- $\xi_t = \{s_0, a_0, s_1, a_1, ..., s_{t-1}, a_{t-1}, s_t\}$ is the whole historical trajectory by time t.
- history dependent policy, denoted as $\pi_h = (\mu_t(a|s))_{s\in\mathcal{S}, a\in\mathcal{A}(s)}, t = 1, 2, ..., \infty.$
- stationary policy when policy independent of time: there exists a vector $\overline{\pi}$ such that $\pi_h = (\mu_t(a|s))_{s\in\mathcal{S}, a\in\mathcal{A}(s)} = \overline{\pi} = (\overline{\mu}(a|s))_{s\in\mathcal{S}, a\in\mathcal{A}(s)}$ for all t.
- Let Π_h and Π_s be the sets of all possible history dependent policies and stationary policies, respectively.

Introduction: setting of MDP

When the reward $r_0(s,a)$ is random, for a fixed $\pi_h \in \Pi_h$, we consider the discounted expected value function

$$V_\alpha(q,\pi_h) = \sum_{t=0}^{\infty} \alpha^t \mathbb{E}_{q,\pi_h}(r_0(s_t,a_t)), \tag{1}$$

where $\alpha \in [0,1)$ is the given discount factor. The object of the agent is to maximize the discounted expected value function

$$\max_{\pi_h \in \Pi_h} \sum_{t=0}^{\infty} \alpha^t \mathbb{E}_{q,\pi_h}(r_0(s_t,a_t)). \tag{2}$$

Introduction: occupation measures

We denote by $d_\alpha(q, \pi_h, s, a)$ the α-discounted occupation measure such that

$$d_\alpha(q, \pi_h, s, a) = (1-\alpha) \sum_{t=0}^{\infty} \alpha^t p_{q,\pi_h}(s_t = s, a_t = a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s).$$

As the state and action spaces are finite, the occupation measure is a well-defined probability distribution (Theorem 3.1 of Altman, 1999). The discounted expected value function (1) can be written as

$$\begin{aligned} V_\alpha(q, \pi_h) &= \sum_{(s,a)\in\Lambda} \sum_{t=0}^{\infty} \alpha^t p_{q,\pi_h}(s_t = s, a_t = a) r_0(s, a) \\ &= \frac{1}{1-\alpha} \sum_{(s,a)\in\Lambda} d_\alpha(q, \pi_h, s, a) r_0(s, a), \end{aligned}$$

where $\Lambda = \{(s, a) | s \in \mathcal{S}, a \in \mathcal{A}(s)\}$.

Introduction: occupation measures

By [Theorem 3.2 of Altman, 1999], we know that the set of occupation measures corresponding to history dependent policies is equal to that corresponding to stationary ones. We have:

Lemma 1 (Altman, 1999)

The set of occupation measures corresponding to history dependent policies is equal to the set

$$\Delta_{\alpha,q} = \left\{ \tau \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \;\middle|\; \begin{array}{l} \sum_{(s,a)\in\Lambda} \tau(s,a)\left(\delta(s',s) - \alpha p(s'|s,a)\right) = (1-\alpha)q(s'), \\ \tau(s,a) \geq 0, \forall s', s \in \mathcal{S}, a \in \mathcal{A}(s). \end{array} \right\} \quad (3)$$

where $\delta(s',s)$ *is the Kronecker delta, such that the expected discounted value function defined by* (2) *remains invariant to time.*

Introduction: MDP and Constrained MDP

MDP problem with history dependent policies:

$$\max_{\tau} \quad \frac{1}{1-\alpha} \sum_{(s,a)\in\Lambda} \tau(s,a) r_0(s,a) \tag{4a}$$

$$\text{s.t.} \quad \tau \in \Delta_{\alpha,q}. \tag{4b}$$

Constrained MDP can be written as:

$$\max_{\tau} \quad \frac{1}{1-\alpha} \sum_{(s,a)\in\Lambda} \tau(s,a) r_0(s,a) \tag{5a}$$

$$\text{s.t.} \quad \sum_{(s,a)\in\Lambda} \tau(s,a) r_k(s,a) \geq \xi_k, k = 1, 2, ..., K, \tag{5b}$$

$$\tau \in \Delta_{\alpha,q}. \tag{5c}$$

Here $r_k(s,a)_{(s,a)\in\Lambda} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, k = 1, 2, ..., K$ be the running constraint rewards and $r_k = (r_k(s,a))_{(s,a)\in\Lambda} \in \mathbb{R}^{|\Lambda|}$ be the running constraint rewards vector.

Introduction: chance constrained MDP

Joint chance constrained MDP (J-CCMDP) can be defined as:

$$
\begin{aligned}
(\mathrm{J-CCMDP}) \quad \max_{\tau} \quad & \frac{1}{1-\alpha}\mathbb{E}_{F_0}[\tau^\top \cdot r_0] && \text{(6a)}\\
\text{s.t.} \quad & \mathbb{P}_F(\tau^\top \cdot r_k \geq \xi_k, k = 1, 2, ..., K) \geq \hat{\epsilon}, && \text{(6b)}\\
& \tau \in \Delta_{\alpha,q} && \text{(6c)}
\end{aligned}
$$

Individual CCMDP,

$$
\begin{aligned}
(\mathrm{I-CCMDP}) \quad \max_{\tau} \quad & \frac{1}{1-\alpha}\mathbb{E}_{F_0}[\tau^\top \cdot r_0] && \text{(7a)}\\
\text{s.t.} \quad & \mathbb{P}_{F_k}(\tau^\top \cdot r_k \geq \xi_k) \geq \epsilon_k, k = 1, 2, ..., K, && \text{(7b)}\\
& \tau \in \Delta_{\alpha,q}, && \text{(7c)}
\end{aligned}
$$

where F_k is the probability distribution of r_k and $\epsilon_k \in [0, 1]$ is the confidence level of the $k-$th constraint.

Introduction: DRO chance constrained MDP

The joint DRCCMDP (J-DRCCMDP):

$$\max_{\tau} \quad \inf_{F_0 \in \mathcal{F}_0} \frac{1}{1-\alpha} \mathbb{E}_{F_0}[\tau^\top \cdot r_0] \tag{8a}$$

$$\text{s.t.} \quad \inf_{F \in \mathcal{F}} \mathbb{P}_F(\tau^\top \cdot r_k \geq \xi_k, k = 1, 2, ..., K) \geq \hat{\epsilon}, \tag{8b}$$

$$\tau \in \Delta_{\alpha,q}. \tag{8c}$$

Individual DRCCMDP (I-DRCCMDP) :

$$\max_{\tau} \quad \inf_{F_0 \in \mathcal{F}_0} \frac{1}{1-\alpha} \mathbb{E}_{F_0}[\tau^\top \cdot r_0] \tag{9a}$$

$$\text{s.t.} \quad \inf_{F_k \in \mathcal{F}_k} \mathbb{P}_{F_k}(\tau^\top \cdot r_k \geq \xi_k) \geq \epsilon_k, \ \ k = 1, 2, ..., K, \tag{9b}$$

$$\tau \in \Delta_{\alpha,q}, \tag{9c}$$

where $\mathcal{F}_k$ is the ambiguity set for the distribution F_k and $\mathcal{F}$ is the ambiguity set for the unknown joint distribution F of $r_1, r_2, ..., r_k$.

Introduction: chance constrained MDP

Related research:

- Delage and Mannor (2010) studied reformulations of chance constrained MDP (CCMDP) with random rewards or transition probabilities.
- Varagapriya et al. (2022) applied joint chance constraints in constrained MDP and find its reformulations when the rewards follow an elliptical distribution.
- Nguyen et al. (2022) studied individual DRCCMDP with moments-based, ϕ-divergence based and Wasserstein distance based ambiguity sets.

Open questions:

- joint chance constraint in DRCCMDP
- high-kurtosis, fat-tailedness or multimodality of the reference distribution (a-prior information)
- new AI-based solution methods

Outline

- Reformulation of K-L divergence based DRCCMDP
- Reformulation of moment-based DRCCMDP
- Dynamical neural network approach for DRCCMDP

KL divergence

Definition 2

Let D_{KL} denotes the Kullback-Leibler divergence distance

$$D_{\mathrm{KL}}(F_k||\tilde{F}_k) = \int_{\Omega_k} \phi\left(\frac{f_{F_k}(r_k)}{f_{\tilde{F}_k}(r_k)}\right) f_{\tilde{F}_k}(r_k) dr_k,$$

where $\tilde{F}_k$ is the reference distribution of r_k, $f_{F_k}(r_k)$ and $f_{\tilde{F}_k}(r_k)$ are the density functions of the true distribution and the reference distribution of r_k on support Ω_k respectively. $\phi(t)$ is defined as follows

$$\phi(t) = \begin{cases} t\mathrm{log}t - t + 1, & t \geq 0, \\ \infty, & t < 0. \end{cases}$$

KL divergence based ambiguity sets

Assumption 1

The marginal ambiguity sets are

$$\mathcal{F}_k = \left\{ F_k | D_{\mathrm{KL}}(F_k || \tilde{F}_k) \leq \delta_k \right\}, k = 0, 1, ..., K,$$

where $\tilde{F}_k$ is the reference distribution of reward vector r_k, the radius δ_k controls the size of the ambiguity sets.

Assumption 2

The joint K-L ambiguity set with jointly independent rows is

$$\mathcal{F} := \mathcal{F}_1 \times \cdots \times \mathcal{F}_K = \{F = F_1 \times \cdots \times F_K | F_k \in \mathcal{F}_k, k = 1, ..., K\},$$

where F is the joint distribution of $r_1, r_2, ..., r_K$ with jointly independent marginals $F_1, ..., F_K$, and $\mathcal{F}_k$ is a K-L ambiguity set with reference marginal distribution $\tilde{F}_k$ and radius $\delta_k, k = 1, ..., K$.

KL: elliptical reference distribution

Definition 3 (Fang 2018)

A d-dimensional vector X follows an elliptical distribution $E_d(\mu, \Sigma, \psi)$ if its characteristic function has the form $\mathbb{E}(e^{ib^\top X}) = e^{ib^\top \mu}\psi(b^\top \Sigma b)$, where $\mu \in \mathbb{R}^d$ is the location parameter, $\Sigma \in \mathbb{R}^{d\times d}$ is the dispersion matrix, ψ is the characteristic generator.

Table: The characteristic generator of three different elliptical distributions

Distribution	Gaussian	Laplace	Generalized stable laws
Characteristic generator $\psi(t)$	e^{-t}	$\frac{1}{1+t}$	$e^{-\omega_1 t^{\frac{\omega_2}{2}}}$, $\omega_1, \omega_2 > 0$

KL-individual: reformulation under elliptical

Theorem 4

Consider ambiguity set in Assumption 1. Assume the reference distribution $\tilde{F}_k \sim E_{|\Lambda|}(\mu_k, \Sigma_k, \psi_k), k = 0, 1, ..., K$, Σ_0 *is a positive definite matrix,* ψ_0 *is continuous,* $\inf_{t \le 0} \psi_0(t) \ge e^{-\delta_0}$. *Then (I-DRCCMDP) (9) is equivalent to*

$$\min_{\tau,\alpha} \quad -\tau^\top \mu_0 + \alpha \log\left[\psi_0\left(-\frac{\tau^\top \Sigma_0 \tau}{2\alpha^2}\right)\right] + \alpha\delta_0, \tag{10a}$$

$$\text{s.t.} \quad \tau^\top \mu_k + \Phi_k^{-1}(1 - \tilde{\epsilon}_k)\sqrt{\tau^\top \Sigma_k \tau} \ge \xi_k, k = 1, 2, \ldots, K, \tag{10b}$$

$$\alpha \ge 0, \tag{10c}$$

$$\tau \in \Delta_{\beta,q}, \tag{10d}$$

where Φ_k *is the CDF of the variable* $Z_k \sim E_1(0, 1, \psi_k)$,
$\tilde{\epsilon}_k = \inf_{x \in (0,1)} \{\frac{e^{-\delta_k} x^{\epsilon_k} - 1}{x - 1}\}$.

Ref: [Hu and Hong, 2013, Jiang and Guan, 2016]

KL-joint: reformulation under elliptical

Theorem 5

Consider $\mathcal{F}_0$ *in Assumption 1 and* $\mathcal{F} := \mathcal{F}_1 \times \cdots \times \mathcal{F}_K$ *in Assumption 2. Assume* $\tilde{F}_k \sim E_{|\Lambda|}(\mu_k, \Sigma_k, \psi_k)$, $k = 0, 1, \ldots, K$, Σ_0 *is p.d.,* ψ_0 *is continuous,* $\inf_{t \le 0} \psi_0(t) \ge e^{-\delta_0}$. *(J-DRCCMDP) (8) is equivalent to*

$$
\begin{aligned}
\min_{\tau,\alpha,y} \quad & -\tau^\top \mu_0 + \alpha \log\left[\psi_0\left(-\frac{\tau^\top \Sigma_0 \tau}{2\alpha^2}\right)\right] + \alpha\delta_0, && \text{(11a)}\\
\text{s.t.} \quad & \tau^\top \mu_k + \Phi_k^{-1}(1 - \tilde{y}_k)\sqrt{\tau^\top \Sigma_k \tau} \ge \xi_k, k = 1, 2, \ldots, K, && \text{(11b)}\\
& 0 \le y_k \le 1, k = 1, 2, \ldots, K, && \text{(11c)}\\
& \prod_{k=1}^{K} y_k \ge \hat{\epsilon}, && \text{(11d)}\\
& \alpha \ge 0, && \text{(11e)}\\
& \tau \in \Delta_{\beta,q}, && \text{(11f)}
\end{aligned}
$$

where $\tilde{y}_k = \inf_{x \in (0,1)} \left\{\frac{e^{-\delta_k} x^{y_k} - 1}{x - 1}\right\}$.

KL-joint: reformulation under Gaussian

Proposition 1

Consider $\mathcal{F}_0$ defined in Assumption 1 and $\mathcal{F} := \mathcal{F}_1 \times \cdots \times \mathcal{F}_K$ defined in Assumption 2. If $\tilde{F}_k$ is a Gaussian distribution $N(\mu_k, \Sigma_k)$, $k = 0, 1, \ldots, K$, and Σ_0 is positive definite, then (8) is equivalent to

$$
\begin{aligned}
\min_{\tau, y} \quad & -\tau^\top \mu_0 + \sqrt{2\delta_0 \tau^\top \Sigma_0 \tau}, && \text{(12a)} \\
\text{s.t.} \quad & \tau^\top \mu_k + \Phi_k^{-1}(1 - \tilde{y}_k)\sqrt{\tau^\top \Sigma_k \tau} \geq \xi_k, k = 1, 2, \ldots, K, && \text{(12b)} \\
& 0 \leq y_k \leq 1, k = 1, 2, \ldots, K, && \text{(12c)} \\
& \prod_{k=1}^{K} y_k \geq \hat{\epsilon}, && \text{(12d)} \\
& \tau \in \Delta_{\beta, q}. && \text{(12e)}
\end{aligned}
$$

where $\tilde{y}_k = \inf_{x \in (0,1)} \{\frac{e^{-\delta_k} x^{y_k} - 1}{x - 1}\}$ and Φ_k is the CDF of the standard Gaussian distribution $N(0, 1)$.

KL-joint: sequence approximation

Firstly, we compute $\tilde{y}_k^n = \inf_{x\in(0,1)} \{\frac{e^{-\delta_k} x^{y_k^n} - 1}{x-1}\}$, and update τ by solving

$$\min_{\tau} \quad -\tau^\top \mu_0 + \sqrt{2\delta_0 \tau^\top \Sigma_0 \tau}, \tag{13a}$$

$$\text{s.t.} \quad \tau^\top \mu_k + \Phi_k^{-1}(1-\tilde{y}_k^n)\sqrt{\tau^\top \Sigma_k \tau} \geq \xi_k, k = 1, 2, \ldots, K, \tag{13b}$$

$$\tau \in \Delta_{\beta,q}. \tag{13c}$$

Then we fix $\tau = \tau^n$ and update y by solving

$$\min_{y} \quad \sum_{k=1}^{K} \Gamma_k y_k \tag{14a}$$

$$\text{s.t.} \quad \frac{1}{2} \leq \tilde{y}_k \leq 1 - \Phi\left(\frac{\xi_k - {\tau^n}^\top \mu_k}{\sqrt{\tau^n \Sigma_k {\tau^n}^\top}}\right), k = 1, 2, ..., K, \tag{14b}$$

$$0 \leq y_k \leq 1, k = 1, 2, ..., K, \tag{14c}$$

$$\sum_{k=1}^{K} \log y_k \geq \log \hat{\epsilon}, \tag{14d}$$

$\tilde{y}_k = \inf_{x\in(0,1)} \{\frac{e^{-\delta_k} x^{y_k} - 1}{x-1}\}$.

KL-joint: sequence approximation

We denote $\tilde{y}_k = \chi_k(y_k) := \inf\limits_{x\in(0,1)} \{\frac{e^{-\delta_k} x^{y_k} - 1}{x-1}\}$. By Jiang and Guan 2016, the infimum of $\chi_k(y_k)$ is attained in the interval $(0,1)$. For any $0 \leq y_k \leq 1$, $\chi_k(y_k) > 0$. By the Envelope Theorem (Tercca 2021), $\chi_k(y_k)$ is strictly monotonically decreasing w.r.t. y_k. Thus we can reformulate (14b) as:

$$\chi_k^{-1}\left(1 - \Phi(\frac{\xi_k - {\tau^n}^\top \mu_k}{\sqrt{\tau^n \Sigma_k {\tau^n}^\top}})\right) \leq y_k \leq \chi_k^{-1}(\frac{1}{2}), \tag{15}$$

where $\chi^{-1}(\cdot)$ denotes the inverse function of $\chi(\cdot)$.
We apply the following approximation

$$\Phi^{-1}(x) \approx t - \frac{2.515517 + 0.802853 \times t + 0.010328 \times t^2}{1 + 1.432788 \times t + 0.189269 \times t^2 + 0.001308 \times t^3},$$

$t = \sqrt{-2 \log x}$

KL-joint: algorithm

Algorithm 1: A hybrid algorithm to solve problem (20)

Data: μ_k, Σ_k, δ_k, ξ_k, $\Delta_{\beta,q}$ n_{max}, $\hat{\epsilon}$, $\tilde{\epsilon}$, γ, $k = 0, 1, ..., K$.
Result: τ^n, V^n.

Set $n = 0$;
Choose an initial point $y^0 = [y_1^0, ..., y_K^0]$ feasible for (23c) and (23d);
while $n \leqslant n_{max}$ *and* $\|y^{n-1} - y^n\| \geqslant \tilde{\epsilon}$ **do**

- Compute $\tilde{y}_k^n = \inf\limits_{x \in (0,1)} \{\frac{e^{-\delta_k x^{y_k^n}} - 1}{x - 1}\}$. Solve problem (22) with $\tilde{y}_k^n$. Let τ^n, V^n be an optimal solution and the optimal value of (22) respectively. Let θ^n be the optimal dual multiplier vector to constraints (22b) ;
- Use the line search method to find $y_k^{Up \cdot n} = \chi_k^{-1}(\frac{1}{2})$ and $y_k^{Low \cdot n} = \chi_k^{-1}(1 - \Phi(\frac{\xi_k \tau^{n\top} \mu_k}{\sqrt{\tau^n \Sigma_k \tau^{n\top}}}))$, $k = 1, ..., K$;
- Solve problem (23) where we replace (23b) by $y_k^{Low \cdot n} \leqslant y_k \leqslant y_k^{Up \cdot n}$, $k = 1, ..., K$, and set

$$\Gamma_k = \theta_k^n \cdot (\Phi^{-1})'(1 - \tilde{y}_k^n)\sqrt{\tau^{n\top} \Sigma_k \tau^n};$$

  let $\tilde{y}$ be an optimal solution of problem (23);
- $y^{n+1} \leftarrow y^n + \gamma(\tilde{y} - y^n), n \leftarrow n + 1$. Here, $\gamma \in (0, 1)$ is the step length.

end

Moment-based ambiguity set

Moment-based ambiguity set

$$\mathcal{F}_k = \left\{ F_k \middle| \begin{array}{l} (\mathbb{E}_{F_k}[r_k] - \mu_k)^\top (\Sigma_k)^{-1} (\mathbb{E}_{F_k}[r_k] - \mu_k) \le \rho_{1,k}, \\ \mathrm{Cov}_{F_k}[r_k] \preceq_S \rho_{2,k} \Sigma_k. \end{array} \right\}, \tag{16}$$

We then assume that different rows in the joint chance constraint are independent of each other and consider the following ambiguity set for the joint distribution

$$\mathcal{F} := \mathcal{F}_1 \times \cdots \times \mathcal{F}_K = \{F = F_1 \times \cdots \times F_K | F_k \in \mathcal{F}_k, k = 1, ..., K\}, \tag{17}$$

where F is the joint distribution for independent random vectors $r_1, ..., r_K$ with marginals $F_1, ..., F_K$.

Moment-based: reformulation

Proposition 2

Given the ambiguity set $\mathcal{F}$ defined in (17)*, the J-DRCCMDP problem* (8) *can be reformulated as:*

$$\min_{\tau \in \mathbb{R}_+^{|\Lambda|}, h \in \mathbb{R}_+^K} \quad \frac{1}{1-\alpha}\left[-\tau^\top \mu_0 + \sqrt{\rho_{1,0}}\|(\Sigma_0)^{\frac{1}{2}}\tau\|\right] \tag{18a}$$

$$\text{s.t.} \quad \tau^\top \mu_k - \left(\sqrt{\frac{h_k}{1-h_k}}\sqrt{\rho_{2,k}} + \sqrt{\rho_{1,k}}\right)\|(\Sigma_k)^{\frac{1}{2}}\tau\| \geq \xi_k, \quad k = 1, 2, ..., K, \tag{18b}$$

$$0 \leq h_k \leq 1, k = 1, 2, ..., K, \tag{18c}$$

$$\prod_{k=1}^{K} h_k \geq \hat{\epsilon}, \tag{18d}$$

$$\tau \in \Delta_{\alpha,q}. \tag{18e}$$

Moment-based: reformulation

By logarithmic transformation:

$$\begin{aligned}
\min_{\tilde{\tau},h} \quad & -\mu_0^\top e^{\tilde{\tau}-\log(1-\alpha)\cdot 1_{|\Lambda|}} + \left\|(\Sigma_0)^{\frac{1}{2}} e^{\tilde{\tau}+\left(\frac{1}{2}\log(\rho_{1,0})-\log(1-\alpha)\right)1_{|\Lambda|}}\right\| \\
\text{s.t.} \quad & \mu_k^\top e^{\tilde{\tau}} - \left\|(\Sigma_k)^{\frac{1}{2}} e^{\tilde{\tau}+\log\left(\sqrt{\frac{e^{\tilde{h}_k}}{1-e^{\tilde{h}_k}}}\sqrt{\rho_{2,k}}+\sqrt{\rho_{1,k}}\right)\cdot 1_{|\Lambda|}}\right\| \geq \xi_k, k = 1,2,...,K \\
& \tilde{h}_k \leq 0, \;\; k = 1,2,...,K, \\
& \sum_{k=1}^K \tilde{h}_k \geq \log(\hat{\epsilon}), \\
& \tilde{\tau} \in \tilde{\Delta}_{\alpha,q},
\end{aligned}$$

where

$$\tilde{\Delta}_{\alpha,q} = \left\{\tilde{\tau} \in \mathbb{R}^{|\Lambda|} \;\middle|\; \sum_{(s,a)\in\Lambda} e^{\tilde{\tau}(s,a)} \left(\delta(s',s) - \alpha p(s'|s,a)\right) = (1-\alpha)q(s'), \forall s', s, a \right. \tag{20}$$

Moment-based: algorithm

Algorithm 1: Sequential convex approximation algorithm(Problem (13))

Data: μ_k, Σ_k, $\rho_{1,k}$, $\rho_{2,k}$, ξ_k, $\Delta_{\alpha,q}$, n_{max}, γ, $\hat{\epsilon}$, L, $k = 0, 1, ..., K$.

Result: τ^n, V^n.

Set $n = 0$;

Choose an initial point h^0 feasible for (21c)-(21d);

while $n \leqslant n_{max}$ *and* $\|h^{n-1} - h^n\| \geqslant L$ **do**

  Solve problem (20); let τ^n, θ^n, V^n be an optimal solution, the optimal Lagrangian dual variable and the optimal value of (20), respectively;

  Solve problem (21) with

$$\mathscr{A}_k = \frac{{\tau^n}^\top \mu_k - \xi_k}{\|(\Sigma_k)^{\frac{1}{2}}\tau^n\|\sqrt{\rho_{2,k}}} - \sqrt{\frac{\rho_{1,k}}{\rho_{2,k}}}, \quad \psi_k = \theta_k^n \frac{\|(\Sigma_k)^{\frac{1}{2}}\tau^n\|}{2(1-h_k^n)}\sqrt{\frac{\rho_{2,k}}{h_k^n(1-h_k^n)}};$$

  let $\tilde{h}$ be an optimal solution of (21);

  $h^{n+1} \leftarrow h^n + \gamma(\tilde{h} - h^n), n \leftarrow n+1$. Here, $\gamma \in (0,1)$ is the step length.

end

DNN approach: introduction

DNN approach is a machine learning technique to solve optimization problems, initiated by Hopfield and Tank (1985). DNN solve

- linear programming (Jun Wang 1993, Youshen Xia 1996)
- second-order cone programming (Chun-Hsu Ko et al. 2011, Nazemi 2020),
- quadratic programming (Xia 1996, Nazemi 2014, 2021)
- nonlinear programming (Forti et al. 2004, Xin-Yu Wu et al. 2004)
- minimax problems (Nazemi 2011, Xing-Bao Gao, Li-Zhi Liao 2004)
- stochastic game problems (Wu, Lisser 2021),
- geometric programming problems (Tassouli, Lisser 2023).

DNN approach: introduction

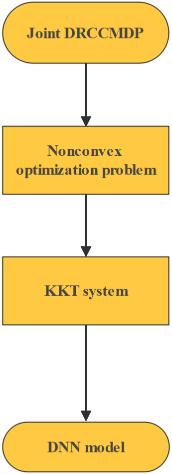

Figure: Flowchart of the DNN approach for solving J-DRCCMDP

DNN approach: introduction

These ODE systems have been shown to have global convergence properties, meaning that the state solutions converge to the optimal solution as the independent variable approaches infinity.

- Dissanayake et al. (1994) were the first to use a neural network to approximate the solution of differential equations, where the loss function contains two terms that satisfy the initial/boundary condition and the differential equation.
- Lagaris et al. (1998) developed a trial solution method that ensures initial conditions are always satisfied.
- Flamant et al. (2020) take the parameter of ODE system models as an input variable to the neural network, allowing a neural network to be the solution of multiple differential equations, namely solution bundles.

Bi-convex reformulation

we first get the following bi-convex reformulation with respect to τ and x.

$$\min_{\tau \in \mathbb{R}_+^{|\Lambda|}, x \in \mathbb{R}_-^K} \quad \frac{1}{1-\alpha}\left[-\tau^\top \mu_0 + \sqrt{\rho_{1,0}}\|(\Sigma_0)^{\frac{1}{2}}\tau\|\right] \tag{21a}$$

$$\text{s.t.} \quad \tau^\top \mu_k - \left(\sqrt{\frac{e^{x_k}}{1-e^{x_k}}}\sqrt{\rho_{2,k}} + \sqrt{\rho_{1,k}}\right)\|(\Sigma_k)^{\frac{1}{2}}\tau\| \geq \xi_k, \quad k = 1, 2, ..., K \tag{21b}$$

$$x_k \leq 0, k = 1, 2, ..., K, \tag{21c}$$

$$\sum_{k=1}^{K} x_k \geq \log \hat{\epsilon}, \tag{21d}$$

$$\tau \in \Delta_{\alpha,q} \tag{21e}$$

Bi-convex reformulation

We can then write for short

$$
\begin{aligned}
\min_{\tau \in \mathbb{R}_+^{|\Lambda|}, x \in \mathbb{R}_-^K} \quad & f(\tau) && \text{(22a)} \\
\text{s.t.} \quad & \phi_k(\tau, x) \leq 0, k = 1, ..., K, && \text{(22b)} \\
& g_k(x) \leq 0, k = 1, ..., K, && \text{(22c)} \\
& h(x) \leq 0, && \text{(22d)} \\
& \omega_s(\tau) \leq 0, s \in S, && \text{(22e)} \\
& -\omega_s(\tau) \leq 0, s \in S, && \text{(22f)} \\
& \nu(\tau) \leq 0. && \text{(22g)}
\end{aligned}
$$

where $f(\tau) = \frac{1}{1-\alpha}\left[-\tau^\top \mu_0 + \sqrt{\rho_{1,0}}\|(\Sigma_0)^{\frac{1}{2}}\tau\|\right]$,

$\phi_k(\tau, x) = \left(\sqrt{\frac{e^{x_k}}{1-e^{x_k}}}\sqrt{\rho_{2,k}} + \sqrt{\rho_{1,k}}\right)\|(\Sigma_k)^{\frac{1}{2}}\tau\| - \tau^\top \mu_k + \xi_k$,

$g_k(x) = x_k,\ k = 1, ..., K,\ h(x) = \log\hat{\epsilon} - \sum_{k=1}^K x_k$,

$\omega_s(\tau) = \sum_{(s',a')\in\Lambda} \tau(s', a')\left(\delta(s, s') - \alpha p(s|s', a')\right) - (1-\alpha)q(s),\ s \in S$

and $\nu(\tau) = -\tau$.

KKT system

The partial optimum is a KKT point and KKT system is

$$\nabla f(\tau^*) + \sum_{k=1}^{K} \beta_k \nabla_\tau \phi_k(\tau^*, x^*) + \sum_{s\in S} (\theta_{1,s} - \theta_{2,s}) \nabla \omega_s(\tau^*) + \varrho \nabla \nu(\tau^*) = 0, \tag{23a}$$

$$\sum_{k=1}^{K} \beta_k \nabla_x \phi_k(\tau^*, x^*) + \sum_{k=1}^{K} \chi_k \nabla g_k(x^*) + \zeta \nabla h(x^*) = 0, \tag{23b}$$

$$\beta_k \geq 0, \quad \beta_k \phi_k(\tau^*, x^*) = 0, \quad \beta_k \phi_k(\tau^*, x^*) = 0, \quad k = 1, ..., K, \tag{23c}$$

$$\chi_k \geq 0, \quad \chi_k g_k(x^*) = 0, \quad k = 1, ..., K, \tag{23d}$$

$$\zeta \geq 0, \quad \zeta h(x^*) = 0, \tag{23e}$$

$$\theta_{1,s} \geq 0, \quad \theta_{1,s}\omega_s(\tau^*) = 0, \quad \theta_{2,s} \geq 0, \quad \theta_{2,s}\omega_s(\tau^*) = 0, \quad s \in S, \tag{23f}$$

$$\varrho \geq 0, \quad \varrho\nu(\tau^*) = 0, \tag{23g}$$

dynamical equations for KKT

construct the dynamical equations for KKT system as

$$
\begin{aligned}
\frac{d\tau}{dt} =& -\Big(\nabla f(\tau) + \nabla_\tau \phi(\tau,x)^\top (\beta + \phi(\tau,x))^+ + \nabla\omega(\tau)^\top(\theta_1 + \omega(\tau))^+ \\
& - \nabla\omega(\tau)^\top(\theta_2 - \omega(\tau))^+ + \nabla\nu(\tau)^\top(\varrho + \nu(\tau))^+\Big), \\
\frac{dx}{dt} =& -\Big(\nabla_x \phi(\tau,x)^\top(\beta + \phi(\tau,x))^+ + \nabla g(x)^\top(\chi + g(x))^+ + \nabla h(x)^\top(\zeta + h(x))^+\Big), \\
\frac{d\beta}{dt} =& (\beta + \phi(\tau,x))^+ - \beta, \\
\frac{d\chi}{dt} =& (\chi + g(x))^+ - \chi, \\
\frac{d\zeta}{dt} =& (\zeta + h(x))^+ - \zeta, \\
\frac{d\theta_1}{dt} =& (\theta_1 + \omega(\tau))^+ - \theta_1, \\
\frac{d\theta_2}{dt} =& (\theta_2 - \omega(\tau))^+ - \theta_2, \\
\frac{d\varrho}{dt} =& (\varrho + \nu(\tau))^+ - \varrho.
\end{aligned} \tag{24}
$$

dynamical system of DNN

Let $z = (\tau, x, \beta, \chi, \zeta, \theta_1, \theta_2, \varrho)$, then the dynamical system can be written as

$$\begin{cases} \frac{dz}{dt} = \kappa\varphi(z), \\ z(t_0) = z_0, \end{cases} \tag{25}$$

where

$$\varphi(z) = \begin{bmatrix} \varphi_1(z) \\ \varphi_2(z) \\ \varphi_3(z) \\ \varphi_4(z) \\ \varphi_5(z) \\ \varphi_6(z) \\ \varphi_7(z) \\ \varphi_8(z) \end{bmatrix} = \begin{bmatrix} -(\nabla f(\tau) + \nabla_\tau \phi(\tau, x)^\top (\beta + \phi(\tau, x))^+ + \nabla\omega(\tau)^\top (\theta_1 + \omega(\tau))^+ \\ \quad -\nabla\omega(\tau)^\top (\theta_2 - \omega(\tau))^+ + \nabla\nu(\tau)^\top (\varrho + \nu(\tau))^+) \\ -(\nabla_x \phi(\tau, x)^\top (\beta + \phi(\tau, x))^+ + \nabla g(x)^\top (\chi + g(x))^+ \\ \quad +\nabla h(x)^\top (\zeta + h(x))^+) \\ (\beta + \phi(\tau, x))^+ - \beta \\ (\chi + g(x))^+ - \chi \\ (\zeta + h(x))^+ - \zeta \\ (\theta_1 + \omega(\tau))^+ - \theta_1 \\ (\theta_2 - \omega(\tau))^+ - \theta_2 \\ (\varrho + \nu(\tau))^+ - \varrho \end{bmatrix},$$

DNN approach

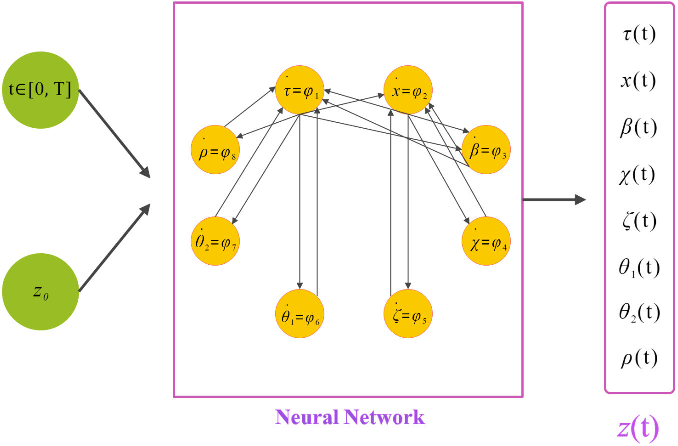

Figure: Flowchart of the DNN approach for solving J-DRCCMDP

existence of solution

Theorem 6

Let $(\tau^*, x^*, \beta^*, \chi^*, \zeta^*, \theta_1^*, \theta_2^*, \varrho^*)$ *be an equilibrium point of the neural network, then* (τ^*, x^*) *is a KKT point. On the other hand, if* (τ^*, x^*) *is a KKT point, then there exists* $\tau^* \geq 0, x^* \geq 0, \beta^* \geq 0, \chi^* \geq 0, \zeta^* \geq 0, \theta_1^* \geq 0, \theta_2^* \geq 0, \varrho^* \geq 0$ *such that* $(\tau^*, x^*, \beta^*, \chi^*, \zeta^*, \theta_1^*, \theta_2^*, \varrho^*)$ *is an equilibrium point of the DNN model.*

Theorem 7

For any initial point $z_0 = (\tau_0, x_0, \chi_0, \zeta_0, \theta_1^0, \theta_2^0, \varrho_0)$*, there exists a unique continuous solution* $z(t) = (\tau(t), x(t), \chi(t), \zeta(t), \theta_1(t), \theta_2(t), \varrho(t))$ *for the DNN model.*

Stability analysis

Lemma 8

The Jacobian matrix $\nabla\varphi(z)$ is a negative semidefinite matrix.

Lemma 9 (Rockafellar, Wets 2009)

A differentiable mapping $F:\mathbb{R}^n\to\mathbb{R}^n$ is monotonic if and only if the Jacobian matrix $\nabla F(x),\forall x\in\mathbb{R}^n$ is positive semidefinite.

monotonic: $(x-y)^\top(F(x)-F(y))\geq 0,\forall x,y\in\mathbb{R}^n$.

Theorem 10

Define $V(z)=\|\varphi(z)\|^2+\frac{1}{2}\|z-z^*\|^2$, we have $\frac{dV(z(t))}{dt}\leq 0$, i.e., DNN model is stable in the Lyapunov sense and converges to $(\tau^*,x^*,\beta^*,\chi^*,\zeta^*,\theta_1^*,\theta_2^*,\varrho^*)$, where (τ^*,x^*) is a KKT point.

Numerical experiments: machine replacement problem

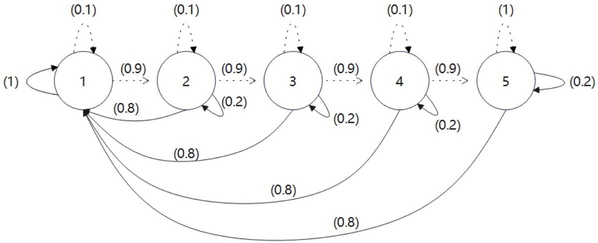

Figure: The transition probabilities for the MDP

- c_0: opportunity cost comes from the potential production losses when the machine is under repair.
- c_1: operational consumption of machines, such as the required electricity fees and fuel costs when the machine is working;
- c_2: the production of inferior quality products.

Numerical experiments: setting

Table: The mean values of three costs

States	Maintenance cost		Operation consumption cost		Inferior quality cost	
	$c_0(s, a_1)$	$c_0(s, a_2)$	$c_1(s, a_1)$	$c_1(s, a_2)$	$c_2(s, a_1)$	$c_2(s, a_2)$
1	1	0	1.5	8	0	5
2	1	0	1.5	8	0	5
3	1	0	1.5	8	0	8
4	4	30	5	100	1.5	30
5	4	70	5	200	3	50

Numerical results: optimal policy

Table: Optimal policies of Moments based J-DRCCMDP

States		1	2	3	4	5
DNN	repair	1.7576e-08	2.8942e-08	≈ 1	≈ 1	≈ 1
	do not repair	≈ 1	≈ 1	3.2052e-08	4.4931e-07	2.7283e-07
SCA	repair	3.5698e-10	4.8275e-10	≈ 1	≈ 1	≈ 1
	do not repair	≈ 1	≈ 1	2.1067e-11	2.2559e-10	5.0490e-10

Numerical results: convergence quality

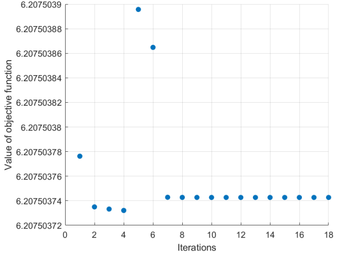

Figure: Objective value for SCA algorithm

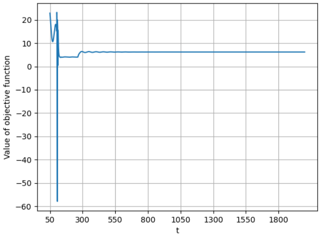

Figure: Objective value for DNN approach

Numerical results: generalization performance

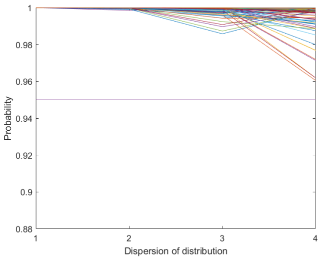

Figure: SCA

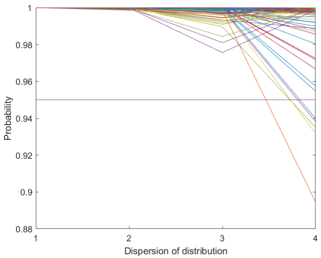

Figure: DNN

Figure: Out-of-sample values $\mathbb{P}_{\mathcal{K}^j}(\tau^\top r_k \geq \xi_k, k = 1, 2, ..., K), j = 1, 2, ..., 100$ with randomly generated distributions $\mathcal{K}^j$, $j = 1, 2, ..., 100$, where the optimal solutions τ are obtained by DNN approach and SCA algorithm, respectively.

Conclusions

Summary:

- Apply DRO-CC in MDP
- Joint CC with two kinds of ambiguity sets
- DNN approach for DRO-CC-MDP

Limitation:

- ambiguity reward; deterministic transaction probability
- environment is fully observable

Ongoing:

- joint ambiguity in transaction probability and reward
- environment is NOT fully observable (reinforcement learning)
- quantitative convergence, error estimation of the dynamic system

Thank you!

Contract: jialiu@xjtu.edu.cn