Distributionally robust chance constrained Markov decision process with Kullback-Leibler divergence *

Tian Xia[a], Jia Liu[a], Abdel Lisser[b]

a: School of Mathematics and Statistics, Xi'an Jiaotong University, 710049, Xi'an, P. R. China

b: CentraleSupelec, Laboratoire des Signaux et des Systemes, 91190 Gif-sur-Yvette, France

Abstract

In this paper, we consider the distributionally robust chance constrained Markov decision process with random cost and ambiguous cost distribution. We consider both individual and joint chance constraint cases with Kullback-Leibler divergence based ambiguity sets centered at elliptical distribution and elliptical mixture distribution, respectively. We derive tractable reformulation of the distributionally robust individual chance constrained Markov decision process problems and design a sequential convex approximation algorithm for the joint case. We carry out numerical results with a machine replacement problem.

Keyword: Markov decision process, chance constraint, distributionally robust optimization, Kullback-Leibler divergence, elliptical distribution

1 Introduction

Markov decision process (MDP) is an effective mathematical model to find an optimal dynamic policy in a long-term uncertain environment. It has many important applications in healthcare [11], autonomous driving [38], financial markets [3], inventory control [21], game theory [43] and so on. It is worth noting

*This research was supported by National Natural Science Foundation of China under Grant Number 11901449.

that MDP is the mathematical fundamental tool of reinforcement learning, which plays an important role in searching a good policy by interacting with the environment via trail and error.

The randomness of MDP often comes from two perspectives: rewards and transition probabilities. Risk attitude is an important issue when the decision-maker measures the randomness of the reward. Many criteria have been considered in risk-aversion MDP, for instance, mean and variance [41], semi-variance [43], Value-at-Risk [25], Conditional Value-at-Risk [30] and etc. Depending on the randomness of transition probabilities, MDP problems can be classified into two groups: rectangular MDP [32, 31, 39] and nonrectangular MDP [26, 37].

In many real applications of MDP, for instance, autonomous driving or healthcare, the safety requirements play an important role when making a dynamic decision to avoid extreme behaviour out of control [20]. This motivates us to take into account robust constraints in the MDP problem, for instance the constrained MDP (CMDP) [35]. To address the extreme conservation of the robust constraints, we can apply chance constraints. Chance constraints control the extreme loss in a probability, which has been widely applied in shape optimization, game theory, electric market and many other fields [24, 29, 8, 18, 22]. Delage and Mannor [6] studies reformulation of chance constrained MDP (CCMDP) with random rewards or transition probability. Varagapriya et al. [36] apply chance constraints into constrained MDP and find reformulations when the rewards follow an elliptical distribution.

In some applications of CCMDP, the distribution of random parameters is not perfectly known, due to the estimation error or imperfect a-priori knowledge. To address this problem, we can employ the distributionally robust optimization (DRO) approach [13], where the decision maker makes a robust decision with respect to the worst-case distribution in a pre-set ambiguity set. In DRO literature, there are two major types of ambiguity sets: moments-based and distance-based. In moments-based DRO [40, 7], decision maker knows some moments information about of random parameters. In distance-based DRO, the decision maker has a reference distribution and consider a ball centered at it in a probability distance, given that she/he believes that the true distribution of random parameters is close to the reference distribution. Based on the probability distance we choose, there are ϕ-divergence (including Kullback-Leibler (K-L) divergence as an important case) distance based DRO [14, 17] and Wasserstein distance based DRO [9, 42, 5, 16] and etc. Applying the techniques of DRO into CCMDP, we have the distributionally robust

chance constrained MDP (DRCCMDP) problem. Nguyen et al. [28] studied individual DRCCMDP with moments-based, ϕ-divergence based and Wasserstein distance based ambiguity sets. However, the study of DRCCMDP is far from completeness. There are still many important cases worth to research, for instance, the joint chance constraint in DRCCMDP has not been studied, the high-kurtosis, fat-tailedness or multimodality of the reference distribution in distance-based DRCCMDP are not considered.

In this paper, we consider the K-L divergence distance based DRCCMDP (KL-DRCCMDP) when the transition probabilities are known and the reward vector is a random vector whose distribution is partially known. We consider both individual and joint cases of KL-DRCCMDP centered at an elliptical reference distribution. We get reformulations of optimization problems in these two cases. For individual case, the reformulation is convex. While for joint case, the reformulation is not convex. We design a sequential convex approximation algorithm to handle this nonconvex problem. In the last part of the joint case of KL-DRCCMDP, we study the case where the parameter is centered at the elliptical mixture distribution and get its reformulation. Finally we conduct a numerical experiment to test our results, we take the reformulations and algorithm proposed before into the machine replacement problem to work out its optimal policy. The major contributions of this paper are listed below.

- As far as we know, this is the first work studying joint case of DRCCMDP.
- We consider elliptical reference distribution and elliptical mixture reference distribution as the center of the ambiguity sets which can reflect the high-kurtosis, fat-tailedness or multimodality of the a-priori information.
- We propose the sequential convex approximation algorithm to solve the nonconvex reformulation. Numerical results validate the practicability of this algorithm.

In Section 2, we introduce the fundamental model of MDP and bring in five types of MDP in a step-by-step way. In Section 3, we study KL-divergence based MDP, including three main cases: the individual case of KL-DRCCMDP with elliptical reference distributions, the joint case of KL-DRCCMDP with elliptical reference distributions and the joint case of KL-DRCCMDP with elliptical mixture reference distributions. In Section 4, on the basis of the famous Machine replacement problem, we do numerical

experiments on our results of the individual and joint cases with elliptical reference distribution proposed before. In the last section, we give a conclusion of our work in this paper.

2 the model

2.1 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 whose generic element is denoted by s with $|S|$ states.
- $\mathcal{A}$ is a finite action space with $|\mathcal{A}|$ actions and $a \in \mathcal{A}(s)$ denotes the action a 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. And $r_0 = (r_0(s,a))_{s\in\mathcal{S},a\in\mathcal{A}(s)} \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ is the running reward vector.
- $q = (q(s))_{s\in\mathcal{S}}$ represents the probability for the initial state.
- α is the discount factor which satisfies $\alpha \in [0,1)$.

In a generalized MDP, the agent aims at maximizing his value function with respect to the whole trajectory by choosing an optimal policy. By [33], it is worth noting that there are exactly two ways of formulating the agent's objective. One is the average reward formulation, the other is considering the discounting factor $\alpha \in [0,1)$. As we care more about the long term reward obtained from the MDP, we pay more attention on optimizing current rewards over future rewards, so we choose the latter one as the formulation of value function in the following paper.

For a discrete time controlled Markov chain $(s_t, a_t)_{t=0}^{\infty}$ defined on the state space $\mathcal{S}$ and action space $\mathcal{A}$, where s_t and a_t are the state and action at time t respectively, at first when time $t = 0$, the state is $s_0 \in \mathcal{S}$, and the action $a_0 \in \mathcal{A}(s_0)$ is taken according to the initial state's probability q. Then the agent

gains rewards $r_0(s_0, a_0)$ based on the current state and action. When $t = 1$, the state converts to s_1 with the transition probability $p(s_1|s_0, a_0)$. The dynamics of the MDP repeat at state s_1 and continue in the following infinite time horizon. As a result, we are able to get the value function for the whole process.

We assume that running rewards r and transition probabilities p are stationary, that is they both only depend on states and actions rather than on time. We hold the assumption above in the following section of this paper. We define the 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 be taken at state s, and $\xi_t = \{s_0, a_0, s_1, a_1, ..., s_{t-1}, a_{t-1}, s_t\}$ the whole trajectory at time t. Let Θ_t be the set of all possible trajectories of length t. For different discrete time t, sometimes the decisions made by the agent may vary accordingly, thus the chosen policy may vary depending on time. We call this kind of policy the history dependent policy denoted as $\pi_h = (\mu_t(a|s))_{s\in\mathcal{S},a\in\mathcal{A}(s)}, t = 1, 2, ..., \infty$. When the policy is independent of time, we call it stationary policy. That is, 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. Combined with the definition above, when the reward function $r_0(s, a)$ is random, for a fixed $\pi_h \in \Pi_h$, the expected discounted value function is

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

The object of the agent is to solve the following optimization problem

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

With a fixed $\alpha \in [0, 1)$, we denote by $d_\alpha(q, \pi_h)$ the α-discounted occupation measure such that

$$d_\alpha(q, \pi_h, s, a) = (1-\alpha)\sum_{t=0}^{\infty} \alpha^t q(s)p(s|s, a)\mu_t(a|s), \tag{2a}$$

$$= (1-\alpha)\sum_{t=0}^{\infty} \alpha^t p_{q,\pi_h}(s_t = s, a_t = a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s), \tag{2b}$$

which is exactly an α-discounted probability distribution for each state and action pair (s, a). As the state and action spaces are both finite, by Theorem 3.1 in [2], the occupation measure $d_\alpha(q, \pi_h, s, a)$ is a well-defined probability distribution. Thus when taking the occupation measure into account, the

expected discounted value function defined above can be written as

$$V_\alpha(q,\pi_h) = \sum_{t=0}^{\infty} \alpha^t \mathbb{E}_{q,\pi_h}(r_0(s_t,a_t)) \tag{3}$$

$$= \sum_{(s,a)\in\Lambda} \sum_{t=0}^{\infty} \alpha^t p_{q,\pi_h}(s_t=s, a_t=a) r_0(s,a) \tag{4}$$

$$= \frac{1}{1-\alpha} \sum_{(s,a)\in\Lambda} d_\alpha(q,\pi_h,s,a) r_0(s,a), \tag{5}$$

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

By Theorem 3.2 in [2], we know that the set of occupation measures corresponding to history dependent policies is equal to that concerning stationary ones. Furthermore, from [35] we have:

Lemma 1 ([35])**.** *The set of occupation measures corresponding to history dependent policies is equal to the set*

$$\Delta_{\alpha,q} = \left\{ \tau \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \;\middle|\; \begin{array}{c} \sum_{(s,a)\in\Lambda} \tau(s,a)\left(\delta(s',s) - \alpha p(s'|s,a)\right) = (1-\alpha)q(s'), \\ \tau(s,a) \geqslant 0, \forall s', s\in\mathcal{S}, a\in\mathcal{A}(s). \end{array} \right\}, \tag{6}$$

where $\delta(s',s)$ is the Kronecker delta, such that the expected discounted value function defined by (5) *remains the same.*

Remark 1. *We assume that each $\tau\in\Delta_{\alpha,q}$ admits $\tau > 0$ in order to keep optimal policy absolutely continuous w.r.t a uniform sampling policy.*

Therefore the MDP with history dependent policies can be equivalent to the one with stationary ones, that is the optimization problem (1) is equivalent to the following one:

$$\max_{\tau} \quad \frac{1}{1-\alpha} \sum_{(s,a)\in\Lambda} \tau(s,a) r_0(s,a) \tag{7a}$$

$$\text{s.t.} \quad \tau \in \Delta_{\alpha,q}. \tag{7b}$$

2.2 CMDP

In a constrained MDP (CMDP), on the basis of the MDP defined above, we consider the running constraint rewards and the bounds for them additionally. Let $r_k(s,a)_{(s,a)\in\Lambda} : \mathcal{S}\times\mathcal{A}\to\mathbb{R}, k=1,2,...,K$ be the running constraint rewards and k denotes the number of constraints, $r_k = (r_k(s,a))_{(s,a)\in\Lambda} \in \mathbb{R}^{|\Lambda|}$ be

the running constraint rewards vector. Let $\Xi = (\xi_k)_{k=1}^K$ be the bounds for the constraints. A CMDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, R, \Xi, P, q, \alpha)$ where $R = (r_k)_{k=0}^K$.

For more attention on optimizing current rewards than future ones, we apply the discount factor in the expected constrained value function. We have

$$\phi_{k,\alpha}(q, \pi_h) = \frac{1}{1-\alpha} \sum_{(s,a)\in\Lambda} d_\alpha(q, \pi_h, s, a) r_k(s, a) \tag{8}$$

to be the k-th expected constrained value function. By Theorem 1, the object of the agent in a CMDP is to solve the following optimization problem

$$\max_{\tau} \quad \frac{1}{1-\alpha} \sum_{(s,a)\in\Lambda} \tau(s, a) r_0(s, a) \tag{9a}$$

$$\text{s.t.} \quad \sum_{(s,a)\in\Lambda} \tau(s, a) r_k(s, a) \geqslant \xi_k, k = 1, 2, ..., K \tag{9b}$$

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

2.3 RCMDP

Based on the definition of MDP above, we assume that the rewards vectors $r_k, k = 0, 1, ..., K$ are random, and the transition probabilities are known. A most common and useful approach to handle the uncertainty is robust optimization. That is we assume the uncertain parameters are constrained to be in a set, which is uncertain. And we consider the worst-case scenario over the set to solve the original optimization problem.

When applying the approach of robustness, we get the robust constrained MDP (RCMDP), which can be defined by the tuple $(\mathcal{S}, \mathcal{A}, R, \Xi, P, \Omega_R, q, \alpha)$. In this tuple, $\Omega_R = \bigtimes_{k=0}^K \Omega_{r_k}$ and Ω_{r_k} denotes the uncertain set of the running rewards. In a RCMDP with random rewards and deterministic transition probabilities, the agent aims at solving the following optimization problem under the worst-case,

$$\max_{\tau} \quad \inf_{r_0\in\Omega_{r_0}} \quad \frac{1}{1-\alpha} \sum_{(s,a)\in\Lambda} \tau(s, a) r_0(s, a) \tag{10a}$$

$$\text{s.t.} \quad \inf_{r_k\in\Omega_{r_k}} \quad \sum_{(s,a)\in\Lambda} \tau(s, a) r_k(s, a) \geqslant \xi_k, k = 1, 2, ..., K \tag{10b}$$

$$\tau \in \Delta_{\alpha,q}. \tag{10c}$$

2.4 CCMDP

From [6] we know that the generated optimal policies by RCMDP are sometimes overly conservative that we need to turn to better approaches to handle the uncertainty in the optimization problem of a MDP. To make up over conservation caused by the worst-case scenario, we could use chance constraints to ensure that the objective cumulative rewards are beyond certain values with high probability. Based on chance constraints, we get a relatively conservative policy compared with RCMDP, which can be classified into a soft-robust MDP by [4]. We call the MDP using chance constraints to handle the randomness the chance-constrained MDP (CCMDP).

For the k-th random constrained rewards vector $r_k = (r_k(s,a))_{(s,a)\in\Lambda}$, we assume its probability distribution is F_k for $k = 0, 1, ..., K$. Let the confidence vector for the CCMDP be $\epsilon = (\epsilon_k)_{k=1}^K$, where $\epsilon_k \in [0,1]$, then we can define a CCMDP as the tuple $(\mathcal{S}, \mathcal{A}, R, \Xi, P, \mathcal{D}, q, \alpha, \epsilon)$ and $\mathcal{D} = (F_k)_{k=0}^K$. The object of an agent in a CCMDP defined above can be formulated the following two optimization problems:

$$
\begin{aligned}
(\mathrm{I}-\mathrm{CCMDP}) \quad \max_{\tau} \quad & \frac{1}{1-\alpha}\mathbb{E}_{F_0}[\tau^\top \cdot r_0] && \text{(11a)}\\
\text{s.t.} \quad & \mathbb{P}_{F_k}(\tau^\top \cdot r_k \geqslant \xi_k) \geqslant \epsilon_k, k = 1, 2, ..., K && \text{(11b)}\\
& \tau \in \Delta_{\alpha,q}, && \text{(11c)}
\end{aligned}
$$

which is called the individual chance constrained MDP (I-CCMDP),

$$
\begin{aligned}
(\mathrm{J}-\mathrm{CCMDP}) \quad \max_{\tau} \quad & \frac{1}{1-\alpha}\mathbb{E}_{F_0}[\tau^\top \cdot r_0] && \text{(12a)}\\
\text{s.t.} \quad & \mathbb{P}_{\hat{F}}(\tau^\top \cdot r_k \geqslant \xi_k, k = 1, 2, ..., K) \geqslant \hat{\epsilon}, && \text{(12b)}\\
& \tau \in \Delta_{\alpha,q}. && \text{(12c)}
\end{aligned}
$$

which is called the joint chance constrained MDP (J-CCMDP), and $\hat{F}$ denotes the joint probability distribution of $r_1, r_2, ..., r_K$ when $\hat{\epsilon}$ denotes the overall confidence for K constraints.

2.5 DRCCMDP

Based on the CCMDP defined above, if the distributions of rewards r_k are unknown, we can apply the approach of robust optimization to handle the uncertainty of $\hat{F}$ and $F_k, k = 0, ..., K$. Thus under the

scenario above, we get the reformulation of a distributionally robust chance-constrained MDP (DRC-CMDP) as the tuple $(\mathcal{S}, \mathcal{A}, R, \Xi, P, \mathcal{D}, \mathcal{F}, \Omega_R, q, \alpha, \epsilon)$, where $\mathcal{F} = (\mathcal{F}_k)_{k=0}^K \bigtimes \hat{\mathcal{F}}$ when $\mathcal{F}_k$ and $\hat{\mathcal{F}}$ denote the ambiguity sets for the random distribution F_k and $\hat{F}$ respectively. Therefore the basic object for an agent in a DRCCMDP can be formulated as the following two optimization problems:

$$
\begin{aligned}
(\mathrm{I-DRCCMDP}) \max_{\tau} \quad & \inf_{F_0 \in \mathcal{F}_0} \quad \frac{1}{1-\alpha} \mathbb{E}_{F_0}[\tau^\top \cdot r_0] && \text{(13a)} \\
\text{s.t.} \quad & \inf_{F_k \in \mathcal{F}_k} \quad \mathbb{P}_{F_k}(\tau^\top \cdot r_k \geqslant \xi_k) \geqslant \epsilon_k, k = 1, 2, ..., K && \text{(13b)} \\
& \qquad\qquad \tau \in \Delta_{\alpha,q}, && \text{(13c)}
\end{aligned}
$$

which is called the individual distributionally robust chance constrained MDP (I-DRCCMDP),

$$
\begin{aligned}
(\mathrm{J-DRCCMDP}) \max_{\tau} \quad & \inf_{F_0 \in \mathcal{F}_0} \quad \frac{1}{1-\alpha} \mathbb{E}_{F_0}[\tau^\top \cdot r_0] && \text{(14a)} \\
\text{s.t.} \quad & \inf_{\hat{F} \in \hat{\mathcal{F}}} \quad \mathbb{P}_{\hat{F}}(\tau^\top \cdot r_k \geqslant \xi_k, k = 1, 2, ..., K) \geqslant \hat{\epsilon}, && \text{(14b)} \\
& \qquad\qquad \tau \in \Delta_{\alpha,q}, && \text{(14c)}
\end{aligned}
$$

which is called the joint distributionally robust chance constrained MDP (J-DRCCMDP).

3 K-L divergence based DRCCMDP

In many real-life problems, it's difficult for us to derive the exact model of any uncertain set $\mathcal{F}_k$ for each k. However through large times of simulation, we could know partial information of the real uncertain sets or derive an approximation from the sample in a data-driven way. In the following sections, we get the tractable reformulation of the main object optimization problem (13) under the following four cases.

Since in moments based distributionally robust optimization (DRO), historical data may not be used efficiently for all information is made use of via moments only, metric based DRO which takes advantage of existing data can make up of this disadvantage greatly. In Kullback-Leibler (K-L) divergence based DRO, the ambiguity set is represented as a ball centered at a reference distribution which is achieved from the historical data. We consider Kullback-Leibler divergence distance [19] as the metric. Next, we give the definition of K-L divergence distance.

Definition 1. *The ambiguity sets are*

$$\mathscr{F}_i = \left\{ F_i | D_{\mathrm{KL}}(F_i || \tilde{F}_i) \leqslant \delta_i \right\}, i = 0, 1, ..., K, \tag{15}$$

where $\delta_i \geqslant 0$ *denotes the radius and* D_{KL} *is the Kullback-Leibler divergence distance defined below,*

$$D_{\mathrm{KL}}(F_i || \tilde{F}_i) = \int_{\Omega_i} \phi(\frac{f_{F_i}(c^i)}{f_{\tilde{F}_i}(c^i)}) f_{\tilde{F}_i}(c^i) dc^i, \tag{16}$$

$\tilde{F}_i$ *is the reference distribution of* c^i*,* $f_{F_i}(c^i)$ *and* $f_{\tilde{F}_i}(c^i)$ *are the density functions of the true distribution and the reference distribution of* c^i *on support* Ω_i *respectively, with the radius* δ_i *controlling the size of the ambiguity sets. And*

$$\phi(t) = \begin{cases} t\log t - t + 1, & t \geqslant 0, \\ \infty, & t < 0. \end{cases} \tag{17}$$

3.1 K-L I-DRCCMDP with Elliptical Reference Distribution

In this subsection, we study the I-DRCCMDP whose ambiguity sets are based on the Kullback-Leibler (K-L)) divergence distance, and we assume that the reference distribution belongs to the elliptical distribution class.

Definition 2 ([27])**.** *A d-dimensional vector* $X \in \mathbb{R}^d$ *follows an elliptical distribution* $E_d(\mu, \Sigma, \psi)$ *if the probability density function (PDF) is* $f(x) = |\Sigma|^{-\frac{1}{2}} g((x-\mu)^\top \Sigma^{-1}(x-\mu))$*, 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 and* $g : \mathbb{R}_+ \to 0$ *is the density generator such that the Fourier transform of* $g(|x|^2)$*, as a generalized function, is equal to* $\psi(|\xi|^2)$*.*

Regarding the properties of the elliptical distribution, we have the following lemma.

Lemma 2 ([27])**.** *Let* $X \sim E_d(\mu, \Sigma, \psi)$ *and take any* $B \in \mathbb{R}^{k \times d}$ *and* $b \in \mathbb{R}^d$*, then*

$$BX + b \sim E_d(B\mu + b, B\Sigma B', \psi). \tag{18}$$

As a special case, if $a \in \mathbb{R}^d$*, then*

$$a'X \sim E_1(a'\mu, a'\Sigma a, \psi). \tag{19}$$

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

In general, the random vector X is said to have a multivariate log-elliptical distribution with parameters μ and Σ if $\log X$ has an elliptical distribution:

$$\log X \sim E_d(\mu, \Sigma, \psi),$$

which can be denoted as $X \sim LE_d(\mu, \Sigma, \psi)$. And we have the following lemma w.r.t. the expectation of log-elliptical distributions.

Lemma 3 ([12]). *Let $\mathbf{X} \sim LE_d(\mu, \Sigma, \psi)$. If the mean of X_k exists, then it is given by*

$$\mathbb{E}(X_k) = e^{\mu_k}\psi(-\frac{1}{2}\sigma_k^2),$$

where μ_k and σ_k^2 denote the mean and variance of X_k respectively.

For some specified elliptical distributions, we have their concrete characteristic generator $\psi : [0, +\infty) \to \mathbb{R}$ correspondingly. Before considering the reformulation of K-L I-DRCCMDP centered at elliptical distributions, we give the following important lemmas.

Lemma 4 ([14]). *Given Definition 1, the objective function in* (13a) *is equivalent to*

$$\inf_{\alpha \in [0,+\infty)} \alpha \log \mathbb{E}_{\tilde{F}_0}\left[\exp(-\frac{\tau^\top r_0}{\alpha})\right] + \alpha\delta_0. \tag{20}$$

Lemma 5 ([17]). *Given Definition 1, the constraint* (13b) *is equivalent to*

$$\mathbb{P}_{\tilde{F}_k}(\tau^\top r_k \geqslant \xi_k) \geqslant \epsilon_k', k = 1, 2, ..., \mathbb{K}, \tag{21}$$

where $\epsilon_k' = \inf_{x \in (0,1)} \{\frac{e^{-\delta_k} x^{\epsilon_k} - 1}{x - 1}\}$.

Based on Lemmas above, we get the reformulation of (13) in this case.

Theorem 1. *Given Definition 1, if the reference distribution $\tilde{F}_k$ for r_k is an elliptical distribution, $\tilde{F}_k \sim \mathbb{E}_{|\Lambda|}(\mu_k, \Sigma_k, \psi_k), k = 0, 1, ..., K$. We assume that ψ_0 is a continuous function, then* (13) *is equivalent*

to

$$\min_{\tau,\alpha} \quad -\tau^\top \mu_0 + \alpha \log\left[\psi_0(-\frac{\tau^\top \Sigma_0 \tau}{2\alpha^2})\right] + \alpha\delta_0, \tag{22a}$$

$$\text{s.t.} \quad \tau^\top \mu_k + \Phi_k^{-1}(1-\epsilon_k')\sqrt{\tau^\top \Sigma_k \tau} \geqslant \xi_k, k = 1, 2, ...K, \tag{22b}$$

$$\alpha \geqslant 0, \tau \in \Delta_{\alpha,q}, \tag{22c}$$

where Φ_k is the cdf of the variable $Z_k \sim E_{|\Lambda|}(0, 1, \psi_k)$ and ϵ_k' is defined the same as that in Lemma 5.

Proof. By Lemma 4 and 5, problem (13) is equivalent to

$$\min_{\tau} \quad \inf_{\alpha\in[0,+\infty)} \quad \alpha \log \mathbb{E}_{\tilde{F}_0}\left[\exp(-\frac{\tau^\top r_0}{\alpha})\right] + \alpha\delta_0, \tag{23a}$$

$$\text{s.t.} \quad \mathbb{P}_{\tilde{F}_k}(\tau^\top r_k \geqslant \xi_k) \geqslant \epsilon_k', k = 1, 2, ..., K, \tag{23b}$$

$$\tau \in \Delta_{\alpha,q}, \tag{23c}$$

where ϵ_k' is defined in Lemma 5.

By Lemma 2, as $\tilde{F}_0$ is an elliptical distribution for r_0, $-\frac{\tau^\top r_0}{\alpha}$ still follow an elliptical distribution with mean $-\frac{\tau^\top \mu_0}{\alpha}$ and variance $\frac{\tau^\top \Sigma_0 \tau}{\alpha^2}$. By Lemma 3, $\exp(-\frac{\tau^\top r_0}{\alpha})$ follows a log-elliptical distribution with mean $e^{-\frac{\tau^\top \mu_0}{\alpha}}\psi_0(-\frac{\tau^\top \Sigma_0 \tau}{2\alpha^2})$. Therefore (23a) is equivalent to

$$\min_{\tau} \inf_{\alpha\in[0,+\infty)} -\tau^\top \mu_0 + \alpha \log\left[\psi_0(-\frac{\tau^\top \Sigma_0 \tau}{2\alpha^2})\right] + \alpha\delta_0. \tag{24}$$

If ψ_0 is continuous w.r.t. α when $\alpha \geqslant 0$, then the inner function of (24) is continuous w.r.t. α. So there exists $\alpha^* \in [0, +\infty)$ such that when $\alpha = \alpha^*$, the inner infimum term of (24) reach its optimal value. (24) is equivalent to

$$\min_{\tau,\alpha} \quad -\tau^\top \mu_0 + \alpha \log\left[\psi_0(-\frac{\tau^\top \Sigma_0 \tau}{2\alpha^2})\right] + \alpha\delta_0, \tag{25a}$$

$$\text{s.t.} \quad \alpha \geqslant 0. \tag{25b}$$

(23b) is equivalent to $\mathbb{P}_{\tilde{F}_k}(\frac{\tau^\top r_k - \tau^\top \mu_k}{\sqrt{\tau^\top \Sigma_k \tau}} \geqslant \frac{\xi_k - \tau^\top \mu_k}{\sqrt{\tau^\top \Sigma_k \tau}}) \geqslant \epsilon_k'$, $k = 1, 2, ..., K$. Let $Z_k = \frac{\tau^\top r_k - \tau^\top \mu_k}{\sqrt{\tau^\top \Sigma_k \tau}}$, and we know that $Z_k \sim E_{|\Lambda|}(0, 1, \psi_k)$. We denote $\Phi_k(z) = \mathbb{P}(Z_k \leqslant z)$ be the cdf of Z_k, and (23b) is equivalent to $\frac{\xi_k - \tau^\top \mu_k}{\sqrt{\tau^\top \Sigma_k \tau}} \leqslant \Phi_k^{-1}(1-\epsilon_k')$, which is just

$$\tau^\top \mu_k + \Phi_k^{-1}(1-\epsilon_k')\sqrt{\tau^\top \Sigma_k \tau} \geqslant \xi_k, k = 1, 2, ...K. \tag{26}$$

We finish the proof. □

3.2 K-L J-DRCCMDP with Elliptical Reference Distribution

In this section, we assume that different rows of the true distribution are jointly dependent and the reference distribution is jointly independent.

Definition 3. *The joint K-L uncertainty set with jointly dependent rows is*

$$\mathscr{F} = \left\{ F | D_{KL}(F||\tilde{F}) \leqslant \delta \right\}, \tag{27}$$

where $\tilde{F}$ is the reference joint distribution for $r_1, r_2, ..., r_K$ with marginals $\tilde{F}_1, ..., \tilde{F}_K$, and $\tilde{F}_1, ..., \tilde{F}_K$ are jointly independent.

By Definition 3 and Lemma 5, we know that constraint (14b) is equivalent to

$$\mathbb{P}_{\tilde{F}}(\tau^\top r_k \geqslant \xi_k, k = 1, 2, ..., K) \geqslant \epsilon', \tag{28}$$

where $\epsilon' = \inf_{x \in (0,1)} \{ \frac{e^{-\delta} x^\epsilon - 1}{x - 1} \}$.

Theorem 2. *Given $\mathscr{F}_0$ defined in Definition 1 and $\mathscr{F}$ defined in Definition 3, if $\tilde{F}_k$ follows an elliptical distribution as $\tilde{F}_k \sim E_{|\Lambda|}(\tilde{\mu}_k, \tilde{\Sigma}_k, \tilde{\psi}_k)$ for $k = 0, 1, ...K$ and $\tilde{\psi}_0$ is a continuous function.* (14) *is equivalent to*

$$\min_{\tau, \alpha, y} \quad -\tau^\top \tilde{\mu}_0 + \alpha \log [\tilde{\psi}_0(-\frac{\tau^\top \tilde{\Sigma}_0 \tau}{2\alpha^2})] + \alpha \delta_0, \tag{29a}$$

$$\text{s.t.} \quad \tau^\top \tilde{\mu}_k + \tilde{\Phi}_k^{-1}(1 - y_k') \sqrt{\tau^\top \tilde{\Sigma}_k \tau} \geqslant \xi_k, k = 1, 2, ...K, \tag{29b}$$

$$0 \leqslant y_k \leqslant 1, k = 1, 2, ...K, \tag{29c}$$

$$\prod_{k=1}^{K} y_k \geqslant \hat{\epsilon}, \tag{29d}$$

$$\alpha \geqslant 0, \tau \in \Delta_{\alpha, q}. \tag{29e}$$

Proof. As $\mathscr{F}_0$ stem from the Definition 1 and (14a) is the same as (13a), we have (14a) is equivalent to

$$\min_{\tau, \alpha} \quad -\tau^\top \tilde{\mu}_0 + \alpha \log [\tilde{\psi}_0(-\frac{\tau^\top \tilde{\Sigma}_0 \tau}{2\alpha^2})] + \alpha \delta_0, \tag{30a}$$

$$\text{s.t.} \quad \alpha \geqslant 0. \tag{30b}$$

As the variable r_k is independent of each other, the inner constraint (14b) is equivalent to

$$\prod_{k=1}^{K} \inf_{\tilde{F}_k \in \tilde{\mathcal{F}}_k} \mathbb{P}_{\tilde{F}_k}(\tau^\top \cdot r_k \geqslant \xi_k) \geqslant \hat{\epsilon}. \tag{31}$$

By introducing auxiliary variables $y_k \in \mathbb{R}_+$, (31) is equivalent to

$$\mathbb{P}_{\tilde{F}_k}(\tau^\top \cdot r_k \geqslant \xi_k) \geqslant y'_k, k = 1, 2, ...K, \tag{32}$$

$$\prod_{k=1}^{K} y_k \geqslant \hat{\epsilon}, 0 \leqslant y_k \leqslant 1, k = 1, 2, ...K, \tag{33}$$

where $y'_k = \inf_{x\in(0,1)} \{\frac{e^{-\delta_k x^{y_k}}-1}{x-1}\}$. By Lemma 3 and the discussion in Theorem 1, (32) is equivalent to

$$\tau^\top \tilde{\mu}_k + \tilde{\Phi}_k^{-1}(1 - y'_k)\sqrt{\tau^\top \tilde{\Sigma}_k \tau} \geqslant \xi_k, k = 1, 2, ...K, \tag{34}$$

where $\tilde{\Phi}_k$ is the cdf of the variable $Z_k \sim E_{|\Lambda|}(0, 1, \tilde{\psi}_k)$. Combining (30),(33) and (34), we finish the proof.

□

Proposition 1. *Given $\mathscr{F}_0$ defined in Definition 1 and $\mathscr{F}$ defined in Definition 3, if $\tilde{F}_k$ follows the Gaussian distribution as $\tilde{F}_k \sim E_{|\Lambda|}(\tilde{\mu}_k, \tilde{\Sigma}_k, \tilde{\psi}_k)$ for $k = 0, 1, ...K$, (14) is equivalent to*

$$\min_{\tau,y} \quad -\tau^\top \tilde{\mu}_0 + \sqrt{2\delta_0 \tau^\top \tilde{\Sigma}_0 \tau}, \tag{35a}$$

$$\text{s.t.} \quad \tau^\top \tilde{\mu}_k + \tilde{\Phi}_k^{-1}(1 - y'_k)\sqrt{\tau^\top \tilde{\Sigma}_k \tau} \geqslant \xi_k, k = 1, 2, ...K, \tag{35b}$$

$$0 \leqslant y_k \leqslant 1, k = 1, 2, ...K, \tag{35c}$$

$$\prod_{k=1}^{K} y_k \geqslant \hat{\epsilon}, \tau \in \Delta_{\alpha,q}. \tag{35d}$$

where $y'_k = \inf_{x\in(0,1)} \{\frac{e^{-\delta_k x^{y_k}}-1}{x-1}\}$ and $\tilde{\Phi}_k$ is the cdf of the standard Gaussian distribution. If $\tilde{F}_k$ follows the Laplace distribution for $k = 0, 1, ...K$, (14) is infeasible.

Proof. When $\tilde{F}_0$ follows the Gaussian distribution, $\tilde{\psi}_0(t) = e^{-t}$ for $t \geqslant 0$, and the inner function of (30) can be written as $-\tau^\top \tilde{\mu}_0 + \frac{\tau^\top \tilde{\Sigma}_0 \tau}{2\alpha} + \alpha\delta_0$, which reaches its minimum value when $\alpha = \sqrt{\frac{\tau^\top \tilde{\Sigma}_0 \tau}{2\delta_0}}$, the optimal value is $-\tau^\top \tilde{\mu}_0 + \sqrt{2\delta_0 \tau^\top \tilde{\Sigma}_0 \tau}$. Therefore (14) is equivalent to

$$\min_{\tau,y} \quad -\tau^\top \tilde{\mu}_0 + \sqrt{2\delta_0 \tau^\top \tilde{\Sigma}_0 \tau}, \tag{36a}$$

$$\text{s.t.} \quad \tau^\top \tilde{\mu}_k + \tilde{\Phi}_k^{-1}(1 - y'_k)\sqrt{\tau^\top \tilde{\Sigma}_k \tau} \geqslant \xi_k, k = 1, 2, ...K, \tag{36b}$$

$$0 \leqslant y_k \leqslant 1, k = 1, 2, ...K, \tag{36c}$$

$$\prod_{k=1}^{K} y_k \geqslant \hat{\epsilon}, \tau \in \Delta_{\alpha,q}. \tag{36d}$$

where $\tilde{\Phi}_k$ is the cdf of the standard Gaussian distribution.

When $\tilde{F}_0$ follows the Laplace distribution, $\tilde{\psi}_0(t) = \frac{1}{1+t}$ for $t > -1$, the inner function of (30) can be written as $-\tau^\top\tilde{\mu}_0 + \alpha\log(\frac{2\alpha^2}{2\alpha^2-\tau^\top\tilde{\Sigma}_k\tau}) + \alpha\delta_0$. Let $\Upsilon_1(\alpha) = -\tau^\top\tilde{\mu}_0 + \alpha\log(\frac{2\alpha^2}{2\alpha^2-\tau^\top\tilde{\Sigma}_k\tau}) + \alpha\delta_0$, $\Upsilon_1'(\alpha) = \log(\frac{2\alpha^2}{2\alpha^2-\tau^\top\tilde{\Sigma}_k\tau}) + \delta_0 > 0$. So $\Upsilon_1(\alpha)$ is monotonically increasing. However when $\alpha \to \sqrt{\frac{\tau^\top\tilde{\Sigma}_k\tau}{2}}$, $\alpha\log(\frac{2\alpha^2}{2\alpha^2-\tau^\top\tilde{\Sigma}_k\tau}) \to +\infty$. Thus (14) is infeasible.

□

Next we talk about the solution of the optimization problem (35). As y_k and τ are both random variables, so (35b) is a non-convex formulation and (35) is not convex. Like Algorithm 1, we still refer to the sequential convex approximation method to handle this non-convex problem. We decompose the problem (35) into the following two subproblems where a subset of variables is fixed alternatively. Firstly, we fix $y = y^n$ and update τ by

$$\min_{\tau} \quad -\tau^\top\tilde{\mu}_0 + \sqrt{2\delta_0\tau^\top\tilde{\Sigma}_0\tau}, \tag{37a}$$

$$\text{s.t.} \quad \tau^\top\tilde{\mu}_k + \tilde{\Phi}_k^{-1}(1-y_k^{n\prime})\sqrt{\tau^\top\tilde{\Sigma}_k\tau} \geqslant \xi_k, k = 1, 2, ...K, \tag{37b}$$

$$\tau \in \Delta_{\alpha,q}, \tag{37c}$$

where $y_k^{n\prime} = \inf_{x\in(0,1)}\{\frac{e^{-\delta_k}x^{y_k^n}-1}{x-1}\}$, and then we fix $\tau = \tau^n$ and update y by

$$\min_{y} \quad \sum_{k=1}^{K}\psi_k y_k \tag{38a}$$

$$\text{s.t.} \quad \frac{1}{2} \leqslant y_k' \leqslant 1 - \Phi(\frac{\xi_k - \tau^{n\top}\mu_k}{\sqrt{\tau^n\Sigma_k\tau^{n\top}}}), k = 1, 2, ..., K, \tag{38b}$$

$$0 \leqslant y_k \leqslant 1, k = 1, 2, ..., K, \tag{38c}$$

$$\sum_{k=1}^{K}\log y_k \geqslant \log\hat{\epsilon}, \tag{38d}$$

where ψ_k is a given searching direction and $y_k{}' = \inf_{x\in(0,1)}\{\frac{e^{-\delta_k}x^{y_k}-1}{x-1}\}$. We say y_k' is a function of y_k as its formulation is regardless of x, say $y_k' = \chi(y_k)$. By the proof of proposition 4 in [17], the infimum of $\chi(y_k)$ is attained in the interval $(0, 1)$. For any $0 \leqslant y_k \leqslant 1$, $\chi(y_k) > 0$. By the envelope theorem [34], $\chi(y_k)$ monotonically decreases with the increasing of y_k. Thus we can transfer (38b) into the following terms:

$$\chi^{-1}\left(1 - \Phi(\frac{\xi_k - \tau^{n\top}\mu_k}{\sqrt{\tau^n\Sigma_k\tau^{n\top}}})\right) \leqslant y_k \leqslant \chi^{-1}(\frac{1}{2}), \tag{39}$$

where $\chi^{-1}(y)$ denotes the value $x \in (0,1)$ such that $\chi(x) = y$ for any $y > 0$. And from the monotonicity of function χ, both sides of (39) are unique at the interval $(0,1)$. Using the sequential convex approximation method, we have the following algorithm to solve the problem (35).

Algorithm 1: Sequential convex approximation (Problem (35))

Data: y^0 feasible for (35c) and (35d), n_{max}, δ_k, $k = 0, 1, ..., K$.

Result: τ^n, V^n.

Set $n = 0$;

Choose an initial point y^0 feasible for (35c) and (35d);

while $n \leqslant n_{max}$ *and* $\|y^{n-1} - y^n\| \geqslant \tilde{\epsilon}$ **do**

Solve problem (37); let τ^n, θ^n, V^n denote an optimal solution, an optimal solution of the Lagrangian dual variable θ and the optimal value of (37), respectively;

Divide the interval $(0,1)$ into 50 equal parts, and use the line search method to find the one that is closest to $\chi^{-1}(\frac{1}{2}), \chi^{-1}(1 - \Phi(\frac{\xi_k - \tau^{n\top}\mu_k}{\sqrt{\tau^n \Sigma_k \tau^{n\top}}})), k = 0, 1, ..., K$ among 50 equal-part points respectively, say $\tilde{\chi}^{-1}(\frac{1}{2}), \tilde{\chi}^{-1}(1 - \Phi(\frac{\xi_k - \tau^{n\top}\mu_k}{\sqrt{\tau^n \Sigma_k \tau^{n\top}}})), k = 0, 1, ..., K$;

Solve problem (38) substituting (38b) into $\tilde{\chi}^{-1}(1 - \Phi(\frac{\xi_k - \tau^{n\top}\mu_k}{\sqrt{\tau^n \Sigma_k \tau^{n\top}}})) \leqslant y_k \leqslant \tilde{\chi}^{-1}(\frac{1}{2})$ for each k, and

$$\psi_k = \theta_k^n \cdot (\Phi^{-1})'(1 - y_k^{n\prime}) \sqrt{\tau^{n\top} \tilde{\Sigma}_k \tau^n}; \tag{40}$$

let $\tilde{h}$ denote an optimal solution of substituted (38);

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

end

Note that in practical numerical experiments, the function $(\Phi^{-1})'$ does not have a closed form, we apply the approximation results of the standard Gaussian quantile function which holds a error bound of 4.5×10^{-14} in ([1], Page.933) to approximate Φ^{-1} here. That is, we take

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

Similar to the discussion on Algorithm 1, Algorithm 2 still converge to a stationary point in a finite number of iterations.

From Theorem 2 in [23], Algorithm 1 converges in a finite number of iterations and the returned value

V^n is an upper bound of problem (**??**). Algorithm 1 can be seen as a particular case of the alternate convex search or block-relaxation methods [10]. When these sub-problems are all convex, the objective function is continuous, the feasible set is closed, the alternate convex search algorithm converges monotonically to a partial optimal point (Theorem 4.7 [10]). When the objective function is a differentiable and biconvex function, (h, τ) is a partial optimal point if and only if (h, τ) is a stationary point (Corollary 4.3 [10]). Thus for the reason that $\|(\Sigma_k)^{\frac{1}{2}}\tau^n\|$ is a convex function for any $k = 1, 2, ..., K$, Algorithm 1 converges to a stationary point.

3.3 K-L J-DRCCMDP with Elliptical mixture Distribution

In this section, we assume the center reference distribution of the K-L J-DRCCMDP is a Gaussian mixture distribution. We study the reformulation of problem (14) under this case. As for the variable vector r_k, the pdf f_k of r_k is defined by $f_k(r_k) = \sum_{j=1}^{J_k} \omega_j^k f_j^k(r_k)$, where $f_j^k(r_k)$ is the density function which follows $E_{|\Lambda|}(\mu_j^k, \Sigma_j^k, \psi_j^k)$ and $\sum_{j=1}^{J_k} \omega_j^k = 1$.

Theorem 3. *Given $\mathscr{F}_0$ defined in Definition 1 and $\mathscr{F}$ defined in Definition 3, if r_k follows the elliptical mixture distribution for each k, that is the pdf of r_k is $f_k(r_k) = \sum_{j=1}^{J_k} \omega_j^k f_j^k(r_k)$, where $f_j^k(r_k)$ is the density function which follows $E_{|\Lambda|}(\tilde{\mu}_j^k, \tilde{\Sigma}_j^k, \tilde{\psi}_j^k)$ and $\sum_{j=1}^{J_k} \omega_j^k = 1$. We assume that $\tilde{\psi}_j^0$ is a continuous function.* (14) *is equivalent to*

$$\min_{\tau,\alpha,y,l} \quad \alpha \log\left[\sum_{j=1}^{J_0} \omega_j^0 \exp\left(-\frac{\tau^\top \tilde{\mu}_j^0}{\alpha}\right)\psi_j^0\left(-\frac{\tau^\top \tilde{\Sigma}_j^0 \tau}{2\alpha^2}\right)\right] + \alpha\delta_0, \tag{41a}$$

$$\text{s.t.} \quad \tau^\top \tilde{\mu}_j^k + (\tilde{\Phi}_j^k)^{-1}\left(1 - \frac{l_j^k}{\omega_j^k}\right)\sqrt{\tau^\top \tilde{\Sigma}_j^k \tau} \geqslant \xi_k, j = 1, 2, ...J_k; k = 1, 2, ...K, \tag{41b}$$

$$\sum_{j=1}^{J_k} l_j^k \geqslant y_k, k = 1, 2, ..., K, \tag{41c}$$

$$0 \leqslant y_k \leqslant 1, 0 \leqslant l_j^k \leqslant 1, j = 1, 2, ...J_k; k = 1, 2, ...K, \tag{41d}$$

$$\sum_{k=1}^{K} y_k \geqslant \hat{\epsilon}, \alpha \geqslant 0, \tag{41e}$$

$$\tau \in \Delta_{\alpha,q}, \tag{41f}$$

where $\tilde{\Phi}_j^k$ is the cdf of the variable $Z_j^k \sim E_{|\Lambda|}(0, 1, \tilde{\psi}_j^k)$.

Proof. By Lemma 5, (14a) is equivalent to

$$\min_{\tau} \inf_{\alpha\in[0,+\infty)} \alpha \log \mathbb{E}_{\tilde{F}_0}\left[\exp(-\frac{\tau^\top r_0}{\alpha})\right] + \alpha\delta_0. \tag{42}$$

We assume that the pdf of r_0 is defined by $f_0(r_0) = \sum_{j=1}^{J_0} \omega_j^0 f_j^0(r_0)$, where $f_j^0(r_0)$ is the density function which follows $E_{|\Lambda|}(\tilde{\mu}_j^0, \tilde{\Sigma}_j^0, \tilde{\psi}_j^0)$ and $\sum_{j=1}^{J_0} \omega_j^0 = 1$.

Let $b_0 = -\frac{\tau^\top r_0}{\alpha}$, we study the mean value of $\exp(b_0)$:

$$\begin{aligned}
\mathbb{E}[\exp(b_0)] &= \int_{c=0}^{+\infty} \exp(b_0)\mathbb{P}(\exp(b_0) = c)dc, && (43a)\\
&= \int_{c=0}^{+\infty} \exp(b_0)\mathbb{P}(b_0 = \ln c)dc, && (43b)\\
&= \int_{c=0}^{+\infty} \exp(b_0)\sum_{j_0} \omega_j^0 \mathbb{P}_j^0(b_0 = \ln c)dc, && (43c)\\
&= \sum_j \omega_j^0 \int_{c=0}^{+\infty} \exp(b_0)\mathbb{P}_j^0(b_0 = \ln c)dc, && (43d)\\
&= \sum_j \omega_j^0 \int_{c=0}^{+\infty} \exp(b_0)\mathbb{P}_j^0(\exp(b_0) = c)dc, && (43e)\\
&= \sum_j \omega_j^0 \exp(-\frac{\tau^\top \tilde{\mu}_j^0}{\alpha})\psi_j^0(-\frac{\tau^\top \tilde{\Sigma}_j^0 \tau}{2\alpha^2}), && (43f)
\end{aligned}$$

where the third equation is because b_0 follows the elliptical mixture distribution, the last equation is by Lemma 3. Following the same process in Theorem 2, by introducing auxiliary variables $y_k \in \mathbb{R}_+$, (14b) is equivalent to (32) and (33). By Proposition 2 in [15], we have

$$(32) \iff \sum_{j=1}^{J_k} \omega_j^k \mathbb{P}_{\tilde{F}_j^k}(\tau^\top \cdot r_k \geqslant \xi_k) \geqslant y_k, k = 1, 2, ...K. \tag{44}$$

With Theorem 2, through adding auxiliary variables $l_j^k \in \mathbb{R}_+$, (32) is equivalent to

$$\omega_j^k \mathbb{P}_{\tilde{F}_j^k}(\tau^\top \cdot r_k \geqslant \xi_k) \geqslant l_j^k, j = 1, 2, ..., J_k; k = 1, 2, ...K, \tag{45}$$

$$\sum_{j=1}^{J_k} l_j^k \geqslant y_k, \sum_{k=1}^{K} y_k \geqslant \hat{\epsilon}, 0 \leqslant y_k \leqslant 1, 0 \leqslant l_j^k \leqslant 1, j = 1, 2, ...J_k; k = 1, 2, ...K. \tag{46}$$

Also $\omega_j^k \mathbb{P}_{\tilde{F}_j^k}(\tau^\top \cdot r_k \geqslant \xi_k) \geqslant l_j^k$ is equivalent to

$$\tau^\top \tilde{\mu}_j^k + (\tilde{\Phi}_j^k)^{-1}(1 - \frac{l_j^k}{\omega_j^k})\sqrt{\tau^\top \tilde{\Sigma}_j^k \tau} \geqslant \xi_k, j = 1, 2, ...J_k; k = 1, 2, ...K, \tag{47}$$

where $\tilde{\Phi}_j^k$ is the cdf of the variable $Z_j^k \sim E_{|\Lambda|}(0, 1, \tilde{\psi}_j^k)$. We finish the proof.

□

4 Numerical experiments

4.1 Machine replacement problem

We take the "Machine replacement problem" as our example, which was introduced by Delage and Mannor [6], and was studied in [35, 39, 11, 31]. In a machine replacement problem, we consider the maintenance cost along with the opportunity cost. The maintenance cost comes from the production losses when the machine is under repair. The opportunity cost comes in two forms: one is due to the operation consumption for machines, such as the required electricity fees and fuel costs when the machine is working; the other is incurred due to the production of inferior quality products. These three costs are unknown in advance and we just know the mean values and corresponding covariance matrix for each cost. Suppose the owner possesses a fixed number of the same machines, and we assume that each machine is modeled with the same model. Therefore we only consider one machine and the same repair policy for it can be applied uniformly for all the machines. The states represent the age of a machine. At each state there are two possible actions, i.e. repair or do not repair. Three considered costs above are incurred at every state. The known transition probabilities for the whole MDP are the same as those in [35] and are available in Figure 1 of [35].

In all numerical experiments, we take the discount factor $\alpha = 0.9$ and assume that the initial distribution q is uniformly distributed. We consider the case of 10 states. The mean values of three considered costs are summarized in Table 4.1. For example, if at state 1 the 'repair' action is used, the mean values of r_0, r_1, r_2 are $-10, -15, 0$ respectively. If the action 'do not repair' is used, the mean values of three costs are $0, -10, -40$ respectively. The last two states are risky states such that the mean values of costs are lower at these states [6]. The covariance matrices of costs are all assumed to be diagonal. Concretely, the covariance matrix of r_0 is $\Sigma_0 = diag([0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 3, 5, 2, 8, 9])$, the covariance matrix of r_1 is $\Sigma_1 = diag([0.5, 5, 0.5, 0.5, 0.5, 5, 0.5, 5, 0.5, 0.5, 0.5, 5, 0.5, 0.5, 0.5, 0.5, 8, 9, 8, 9])$ and the the covariance matrix of r_2 is $\Sigma_2 = diag([0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04,$

$0.04, 0.04, 0.04, 4, 9, 8, 8.5, 10]$).

In our numerical experiments, we still study the DRCCMDP model. In particular, for I-DRCCMDP model, we set $\xi_1 = \xi_2 = -40$ and $\epsilon_1 = \epsilon_2 = 0.8$ in (13b). For J-DRCCMDP, we set $\xi_1 = \xi_2 = -40$ and $\hat{\epsilon} = 0.8$ in (14b). Next we list our results and conclusions from numerical experiments on Moments based DRCCMDP and K-L divergence based DRCCMDP respectively.

Table 1: The mean value of considered costs

States	Maintenance cost		Operation consumption cost		Inferior quality cost	
	$r_0(s, a_1)$	$r_0(s, a_2)$	$r_1(s, a_1)$	$r_1(s, a_2)$	$r_2(s, a_1)$	$r_2(s, a_2)$
1	-10	0	-15	-10	0	-40
2	-10	0	-15	-30	0	-40
3	-10	0	-15	-40	0	-50
4	-10	0	-15	-50	0	-50
5	-10	0	-15	-70	-15	-50
6	-10	0	-15	-80	-15	-55
7	-10	0	-15	-80	-15	-55
8	-10	0	-15	-80	-15	-55
9	-40	-85	-50	-200	-30	-80
10	-40	-95	-50	-200	-30	-100

4.2 Numerical results on Moments based DRCCMDP

For Moments based DRCCMDP, let $\rho_{1,0} = 20, \rho_{1,1} = \rho_{2,1} = 20, \rho_{1,2} = \rho_{2,2} = 15$. Then when focusing on the I-DRCCMDP, we solve the convex optimization problem (**??**) using the MOSEK solver in YALMIP toolbox of MATLAB given the data needed above. When focusing on the J-DRCCMDP, we turn to (**??**), (**??**). Based on the Algorithm 1, we set the initial points $h_1^0 = 0.93, h_2^0 = 0.95$ and $n_{max} = 100, \tilde{\epsilon} = 10^{-4}, \gamma = 0.9$, we use the MOSEK solver to work out this problem. The results are listed in Table 4.2, from which we conclude that the repair probability of the machine increases with its age and we must repair it at the last four states for both I-DRCCMDP and J-DRCCMDP in this model.

Table 2: Optimal policies of Moments based DRCCMDP

States		1	2	3	4	5	6	7	8	9	10
I-DRCCMDP	repair	0	0	0	0.2077	0.5068	0.9534	1	1	1	1
	do not repair	1	1	1	0.7923	0.4932	0.0466	0	0	0	0
J-DRCCMDP	repair	0	0	0.0679	0.3287	0.3624	0.8176	1	1	1	1
	do not repair	1	1	0.9321	0.6713	0.6376	0.1824	0	0	0	0

4.3 Numerical results on K-L divergence based DRCCMDP

In this section, we particularly consider the case where the reference distribution of the K-L divergence based ambiguity set is a standard Gaussian distribution. Same as Moments based DRCCMDP, we let $\rho_{1,0} = \rho_{1,1} = \rho_{2,1} = 20, \rho_{1,2} = \rho_{2,2} = 15$. When focusing on I-DRCCMDP in this case, we aim to solve the convex optimization problem (22). Following the same proof process in Proposition 1 when the reference distribution is a standard Gaussian distribution, we can simplify (22) into the following one:

$$\min_{\tau} \quad -\tau^\top \mu_0 + \sqrt{2\delta_0 \tau^\top \Sigma_0 \tau}, \tag{48a}$$

$$\text{s.t.} \quad \tau^\top \mu_k + \Phi_k^{-1}(1-\epsilon_k')\sqrt{\tau^\top \Sigma_k \tau} \geqslant \xi_k, k = 1, 2, ...K, \tag{48b}$$

$$\tau \in \Delta_{\alpha,q}. \tag{48c}$$

We consider six different cases when $\delta_0 = \delta_1 = \delta_2 = 0.5, 0.4, 0.3, 0.2, 0.1, 0.01$ respectively, and we solve the convex optimization problem (48) using the GUROBI solver in MATLAB. We list the results on the probability of "repair" action for each state under K-L I-DRCCMDP in Figure 4.3. As there are in total two actions to choose for each state, the probability of "do not repair" action can be computed by subtracting that of "repair" action with 1. Thus the trend of the probability that "do not repair" be taken for each state is just the same as that of "repair", and we omit it taken for granted. From Figure 4.3, we see the asymptotic convergence of the probability for each state when $\delta_0, \delta_1, \delta_2$ are taken decreasingly from 0.5 to 0.01. Also we see that for all six radii we choose, the probability of "repair" in last three states are all 1, which is in correspondence with the fact that the machine gets aging with the state forward.

Next we focus on J-DRCCMDP in this case. Based on Algorithm 2, we set the initial points $y_1^0 =$

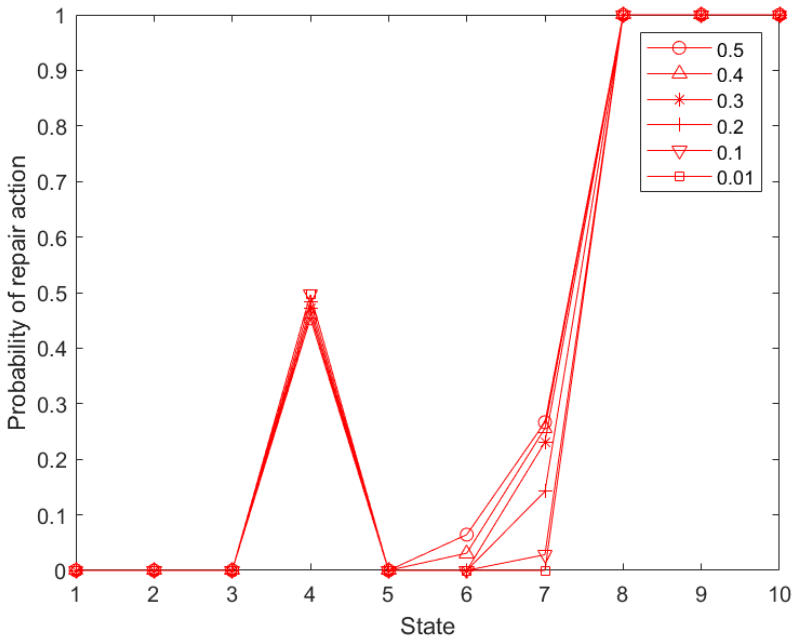

Figure 1: The probability of the action "repair" for each state under K-L I-DRCCMDP

$0.95, y_2^0 = 0.91$ and $n_{max} = 50$, $\tilde{\epsilon} = 10^{-4}$, $\gamma = 0.9$. For the line search method which we use to get $\tilde{\chi}^{-1}(\cdot)$, let the number of intervals be 50 and the approximation accuracy be 10^{-3}. We consider six cases with different radii when $\delta_0 = \delta_1 = \delta_2 = 10^{-4}, 5 \times 10^{-5}, 10^{-5}, 5 \times 10^{-6}, 10^{-6}, 0$. We use the MOSEK solver to work out this problem and we list our numerical results on probability of "repair" for each state in this case in Figure 4.3, from which we observe the asymptotic convergence of the probability for each state as the radius decreases to 0.

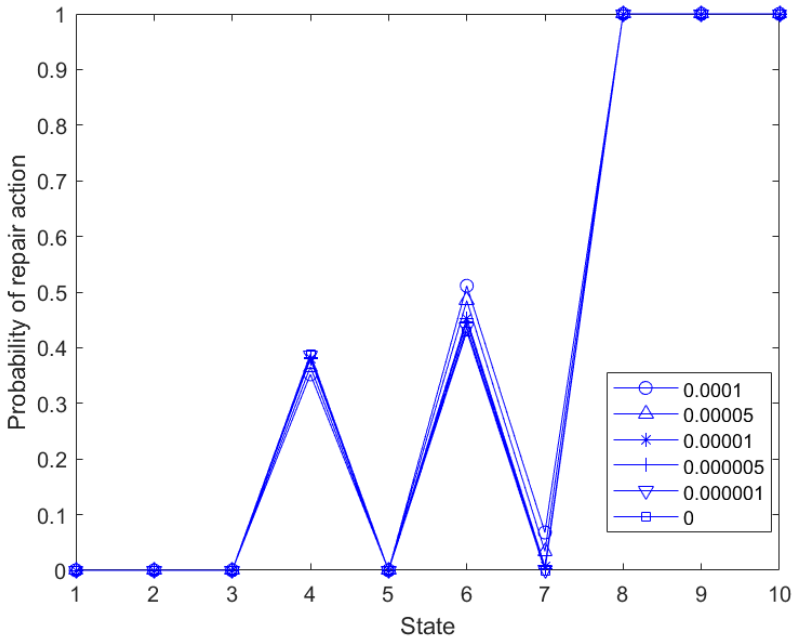

Figure 2: The probability of the action "repair" for each state under K-L J-DRCCMDP

References

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1964.

[2] Eitan Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

[3] Souradeep Chakraborty. Capturing financial markets to apply deep reinforcement learning. *arXiv preprint arXiv:1907.04373*, 2019.

[4] Shiyu Chen and Yanjie Li. An overview of robust reinforcement learning. In *2020 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, pages 1–6. IEEE, 2020.

[5] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over wasserstein balls. *Operations Research*, 2022.

[6] Erick Delage and Shie Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, (1):203–213, 2010.

[7] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[8] Yury Dvorkin. A chance-constrained stochastic electricity market. *IEEE Transactions on Power Systems*, 35(4):2993–3003, 2019.

[9] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 2022.

[10] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical methods of operations research*, 66(3):373–407, 2007.

[11] Vineet Goyal and Julien Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

[12] Mahmoud Hamada and Emiliano A Valdez. Capm and option pricing with elliptically contoured distributions. *Journal of Risk and Insurance*, 75(2):387–409, 2008.

[13] Grani A Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming*, 151(1):35–62, 2015.

[14] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pages 1695–1724, 2013.

[15] Zhaolin Hu, Wenjie Sun, and Shushang Zhu. Chance constrained programs with gaussian mixture models. *IISE Transactions*, 54(12):1117–1130, 2022.

[16] Ran Ji and Miguel A Lejeune. Data-driven distributionally robust chance-constrained optimization with wasserstein metric. *Journal of Global Optimization*, 79(4):779–811, 2021.

[17] Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1):291–327, 2016.

[18] Shiyi Jiang, Jianqiang Cheng, Kai Pan, Feng Qiu, and Boshi Yang. Data-driven chance-constrained planning for distributed generation: A partial sampling approach. *IEEE Transactions on Power Systems*, 2022.

[19] James M Joyce. Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer, 2011.

[20] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[21] Diego Klabjan, David Simchi-Levi, and Miao Song. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management*, 22(3):691–710, 2013.

[22] Simge Küçükyavuz and Ruiwei Jiang. Chance-constrained optimization under limited distributional information: a review of reformulations based on sampling and distributional robustness. *EURO Journal on Computational Optimization*, page 100030, 2022.

[23] Jia Liu, Abdel Lisser, and Zhiping Chen. Stochastic geometric optimization with joint probabilistic constraints. *Operations Research Letters*, 44(5):687–691, 2016.

[24] Jia Liu, Abdel Lisser, and Zhiping Chen. Distributionally robust chance constrained geometric optimization. *Mathematics of Operations Research*, 2022.

[25] Shuai Ma and Jia Yuan Yu. State-augmentation transformations for risk-sensitive reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4512–4519, 2019.

[26] Shie Mannor, Ofir Mebel, and Huan Xu. Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

[27] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.

[28] Hoang Nam Nguyen, Abdel Lisser, and Vikas Vikram Singh. Distributionally robust chance-constrained markov decision processes. *arXiv preprint arXiv:2212.08126*, 2022.

[29] Shen Peng, Abdel Lisser, Vikas Vikram Singh, Nalin Gupta, and Eshan Balachandar. Games with distributionally robust joint chance constraints. *Optimization Letters*, 15(6):1931–1953, 2021.

[30] LA Prashanth. Policy gradients for cvar-constrained mdps. In *International Conference on Algorithmic Learning Theory*, pages 155–169. Springer, 2014.

[31] Sivaramakrishnan Ramani and Archis Ghate. Robust markov decision processes with data-driven, distance-based ambiguity sets. *SIAM Journal on Optimization*, 32(2):989–1017, 2022.

[32] Jay K Satia and Roy E Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

[33] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[34] Gonçalo Terça and David Wozabal. Envelope theorems for multistage linear stochastic optimization. *Operations Research*, 69(5):1608–1629, 2021.

[35] V Varagapriya, Vikas Vikram Singh, and Abdel Lisser. Constrained markov decision processes with uncertain costs. *Operations Research Letters*, 50(2):218–223, 2022.

[36] V Varagapriya, Vikas Vikram Singh, and Abdel Lisser. Joint chance-constrained markov decision processes. *Annals of Operations Research*, pages 1–23, 2022.

[37] Jie Wang, Rui Gao, and Hongyuan Zha. Reliable off-policy evaluation for reinforcement learning. *Operations Research*, 2022.

[38] Junqing Wei, John M Dolan, Jarrod M Snider, and Bakhtiar Litkouhi. A point-based mdp for robust single-lane autonomous driving behavior under uncertainties. In *2011 IEEE International Conference on Robotics and Automation*, pages 2586–2592. IEEE, 2011.

[39] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[40] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[41] Li Xia. Risk-sensitive markov decision processes with combined metrics of mean and variance. *Production and Operations Management*, 29(12):2808–2827, 2020.

[42] Weijun Xie. On distributionally robust chance constrained programs with wasserstein distance. *Mathematical Programming*, 186(1):115–155, 2021.

[43] Zhihui Yu, Xianping Guo, and Li Xia. Zero-sum semi-markov games with state-action-dependent discount factors. *Discrete Event Dynamic Systems*, 32(4):545–571, 2022.