Distributionally robust chance-constrained Markov decision processes with random transition probabilities

Hoang Nam Nguyen, Abdel Lisser, Vikas Vikram Singh

Distributionally robust chance-constrained Markov decision processes with random transition probabilities

Hoang Nam Nguyen[1†], Abdel Lisser[1*†], Vikas Vikram Singh[2†]

[1]Université Paris Saclay, CNRS, CentraleSupelec, L2S, Bât Breguet, 3 rue Joliot Curie, Gif-sur-Yvette, 91190, France.
[2]Department of Mathematics, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, 110016, India.

*Corresponding author(s). E-mail(s): abdel.lisser@centralesupelec.fr;
Contributing authors: hoang-nam.nguyen3@centralesupelec.fr;
vikassingh@maths.iitd.ac.in;
†These authors contributed equally to this work.

Abstract

Markov decision process (MDP) is a decision making framework where a decision maker is interested in maximizing the expected discounted value of a stream of rewards received at future stages at various states which are visited according to a controlled Markov chain. Many algorithms including linear programming methods are available in the literature to compute an optimal policy when the rewards and transition probabilities are deterministic. In this paper, we consider an MDP problem where the reward vector is known and the transition probability vector is a random vector which follow a discrete distribution whose information is not completely known. We formulate the MDP problem using distributionally robust chance-constrained optimization framework under various types of moments based uncertainty sets, and statistical-distance based uncertainty sets defined using ϕ-divergence and Wasserstein distance metric. For each uncertainty set, we propose an equivalent mix-integer bilinear programming problem or a mix-integer semidefinite programming problem with bilinear constraints. As an application, we study a machine replacement problem and perform numerical experiments on randomly generated instances.

Keywords: Markov decision processes, Distributionally robust chance-constrained optimization, Random transition probabilities, Machine replacement problem.

MSC Classification: 90C15 , 90C25 , 90C40 , 90C59.

1 Introduction

An MDP is a decision making framework to model the performance of a stochastic system which evolves over time according to a controlled Markov chain. We consider the case where the system has a finite number of states. At time $t = 0$, the system is at some initial state $s_0 \in S$, where S is a finite state space, according to an initial distribution γ, and a decision maker chooses an action $a_0 \in A(s_0)$, where $A(s_0)$ denotes the set of finite number of actions available to the decision maker at state s_0. As a consequence a reward $R(s_0, a_0)$ is earned and at time $t = 1$, the system moves to a new state s_1 with probability $p(s_0, a_0, s_1)$. The same thing repeats at time $t = 1$ and it continues for the infinite horizon. We assume that the reward and transition probabilities are stationary, i.e., $R(X_t = s, A_t = a) = R(s, a)$ and $P(X_{t+1} = s'|X_t = s, A_t = a) = p(s, a, s')$ for all t; X_t and A_t denote the state and action at time t, respectively. The decision taken at time t, which could be deterministic or randomized, may depend on the history h_t at time t, where $h_t = (s_0, a_0, s_1, \ldots, s_{t-1}, a_{t-1}, s_t)$. Let H_t be the set of all possible histories at time t. A history dependent decision rule f_t at time t is defined as $f_t(h_t) \in \wp(A(s_t))$ for every history h_t with final state s_t, where $\wp(A(s_t))$ denotes the set of probability distributions on the action set $A(s_t)$. A sequence of history dependent decision rules $f^h = (f_t)_{t=0}^{\infty}$ is called a history dependent policy. A history dependent policy $(f_t)_{t=0}^{\infty}$ is called a stationary policy if there exists a decision rule f such that $f_t = f$ for all t and it depends only on the current state. We denote a stationary policy, with some abuse of notations, by f and define $f = (f(s))_{s\in S}$ such that $f(s) \in \wp(A(s))$ for every $s \in S$. As per stationary policy f, whenever the Markov chain visits state s, the decision maker chooses an action a with probability $f(s, a)$. We denote the set of all history dependent and stationary policies by PO_{HD} and PO_S, respectively. A history dependent policy $f^h \in PO_{HD}$ and initial distribution γ defines a probability measure $P_\gamma^{f^h}$ over the state and action trajectories, and $E_\gamma^{f^h}$ denotes the expectation operator corresponding to the probability measure $P_\gamma^{f^h}$. For a given policy f^h and an initial distribution γ, the expected discounted reward at a discount factor $\alpha \in (0, 1)$ is defined as Altman (1999); Puterman et al (1994)

$$\begin{aligned} V_\alpha(f^h, p) &= (1-\alpha)\mathbb{E}^{f^h}\left(\sum_{t=0}^{\infty} \alpha^t R(X_t, A_t)\right) \\ &= \sum_{s\in S}\sum_{a\in A(s)} \hat{m}(f^h, p; s, a)R(s, a). \end{aligned} \tag{1}$$

The set $\{\hat{m}(f^h, p; s, a)\}_{(s,a)}$ is the occupation measure defined by

$$\hat{m}(f^h, p; s, a) = (1-\alpha)\sum_{t=0}^{\infty} \alpha^t P_\gamma^{f^h}(\mathbb{X}_t = s, \mathbb{A}_t = a),\ \forall\ s \in S,\ a \in A(s). \tag{2}$$

It is well known that for a discounted MDP problem there exists an stationary optimal policy. For a given transition probabilities $p(s, a, s')$ for all $s, s' \in S$, $a \in A(s)$, it follows from Theorem 3.2 of Altman (1999) that the set of occupation measures $\hat{m}(f, p)$, for $f \in PO_S$ is equivalent to the following set

$$\mathcal{Q}(p) = \Bigg\{ \rho \in \mathbb{R}^{|\mathcal{K}|} \mid \sum_{(s,a)\in\mathcal{K}} \rho(s,a)\Big(\delta(s', s) - \alpha p(s, a, s')\Big) = (1-\alpha)\gamma(s'),\ \forall\ s' \in S,$$
$$\rho(s, a) \geq 0,\ \forall\ (s, a) \in \mathcal{K} \Bigg\},$$

where $\delta(s', s)$ is the Kronecker delta and $\mathcal{K} = \{(s, a) \mid s \in S, a \in A(s)\}$. For each $\rho \in \mathcal{Q}(p)$, the stationary optimal policy f can be defined as

$$f(s, a) = \frac{\rho(s, a)}{\sum_{a\in A(s)} \rho(s, a)},\ \forall\ (s, a) \in \mathcal{K},$$

whenever the denominator is nonzero $\big($if it is zero, we choose $f(s)$ arbitrarily from $\wp(A(s))\big)$ Altman (1999). We restrict our attention to the stationary policies in the rest of the paper.

The rewards and transition probabilities represented by $|\mathcal{K}|$-dimensional vector $R = (R(s, a))_{(s,a)\in\mathcal{K}}$ and $|\mathcal{K}|.|S|$-dimensional vector $p = (p(s, a, s'))_{(s,a,s')\in\mathcal{K}\times S}$, respectively, are considered as the parameters of an MDP model and are assumed to be exactly known. However, in practice R and p are not known in advance and are estimated from historical data. This leads to errors in the optimal policies Mannor et al (2007). Most efforts to take into account this uncertainty focused on the study of robust MDPs where the rewards or the transition probabilities are known to belong to a prespecified uncertainty set Iyengar (2005); Nilim and El Ghaoui (2005); Varagapriya et al (2022); White III and Eldeib (1994); Wiesemann et al (2012); Ho et al (2022); Goyal and Grand-Clement (2023). However, Delage and Mannor (2010) showed that the robust MDP approach usually leads to conservative policies. For this reason, a chance-constrained Markov decision process (CCMDP) was introduced in Delage and Mannor (2010), where the controller obtains the expected discounted reward with certain confidence. In Delage and Mannor (2010), the case of random rewards and random transition probabilities are considered separately and it is shown that a CCMDP is equivalent to a second-order cone programming (SOCP) problem when the running reward vector follows a multivariate normal distribution and the transition probabilities are exactly known. When the transition probabilities follow Dirichlet distribution and the running rewards are exactly known, the CCMDP problem becomes intractable

and the optimal policies can be computed using approximation methods. Varagapriya et al (2023) considered a constrained MDP problem where the running cost vectors are random vectors and the transition probabilities are known. They formulated it as a joint chance-constrained MDP problem and proposed two SOCP based approximations which give upper and lower bounds to the optimal value of joint chance-constrained MDP problem if the cost vectors follow multivariate elliptical distributions and the dependence among the constraints is driven by a Gumbel-Hougaard copula.

In many practical situations, it is often the case that only a partial information about the underlying distribution is available based on the historical data. In that case, a distributionally robust approach, is used to model the uncertainties, which assumes that the true distribution belongs to an uncertainty set based on its partially available information. Such an approach has been used in modelling the uncertainties of many optimization and game problems Jiang and Guan (2016); Liu et al (2022); Singh et al (2017). There are at least two popular ways to construct an uncertainty set for the distribution of the uncertain parameters. The first one is based on the partial information on moments of the true distribution and the second one is based on the statistical distance between the true distribution and a reference distribution. The moments-based uncertainty sets assume certain conditions on the first two moments Cheng et al (2014); Delage and Ye (2010); Popescu (2007). The statistical distance-based uncertainty sets contain all the distributions which lie inside a ball of small radius and center at a reference distribution which is usually considered to be an empirical distribution or a normal distribution Esfahani and Kuhn (2018); Jiang and Guan (2016). To define a distance between the distributions, either a $\phi-$divergence Ben-Tal et al (2013); Jiang and Guan (2016) or Wasserstein distance metric is used Esfahani and Kuhn (2018); Gao and Kleywegt (2023); Zhao and Guan (2018). A discounted MDP problem with uncertain transition probabilities under distributionally robust optimization framework is considered in the literature Xu and Mannor (2012); Zhi Chen and Haskell (2019). They considered the case where the decision maker aims to find an optimal policy which maximizes worst-case expected discounted reward. To the best of our knowledge, an MDP problem with uncertain transition probabilities under distributionally robust chance constraint based payoff criterion is not considered in the literature.

In this paper, we consider an infinite horizon MDP with discounted payoff criterion where the reward vector is known and transition probabilities are defined by a random vector. The transition probability vector is assumed to follow a discrete distribution which is not completely known and belong to a given uncertainty set. We formulate the random transition probability vector with a distributionally robust chance constraint which guarantees the maximum reward for a given policy with at least a given level of confidence. We call this class of MDP a distributionally robust chance-constrained Markov decision process (DRCCMDP), where we consider both moments and statistical distance based uncertainty sets. The main contributions of the paper are as follows.

1. We consider three different types of uncertainty sets based on the moments of the random transition probabilities. We show that the DRCCMDP problem can be reformulated as a mixed-integer bilinear programming (MIBP) problem or a

mixed-integer semidefinite programming (MISDP) problem with additional bilinear constraints. The MIBP problems can be solved efficiently using the GUROBI solver, while the MISDP problems with bilinear constraints can be handled by the CUTSDP solver available in the YALMIP toolbox of Matlab, which is time consuming, without any guarantee of the running time.

2. We consider two types of uncertainty sets based on statistical distance between the true distribution of transition probabilities and a reference distribution. The uncertainty sets are constructed using either $\phi-$divergences distance or Wasserstein distance. For $\phi-$divergences distance, we explore four distinct types of ϕ-divergences (Kullbach-Leibler, Variation, Modified χ^2, Hellinger) to construct uncertainty sets. For each uncertainty set, we show that the DRCCMDP problem is equivalent to MIBP problem.
3. We illustrate our theoretical results on a machine replacement problem.

The paper is organized as follows. In Section 2, we define a DRCCMDP under a discounted payoff criterion with random transition probabilities and known reward vector. In Section 2.1 and Section 2.2, we propose equvalent reformulations of DRCCMDP under different moments based and statistical distance based uncertainty sets. We present how to solve reformulations using existing solvers in Section 2.3. The numerical results on a machine replacement problem is given in Section 3. We conclude the paper in Section 4.

2 Distributionally robust chance-constrained Markov decision process

In this section, we consider an MDP model defined in Section 1, where the running reward vector R is exactly known, nonnegative and the transition probability vector p are random variables. For each triplet $(s, a, s') \in \mathcal{K} \times S$, we assume that the $p(s, a, s')$ is an $1-$ dimensional random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Therefore, for each realization $\omega \in \Omega$, the term $p(s, a, s')(w) \in [0, 1]$ and $\sum_{s' \in S} p(s, a, s')(\omega) = 1$. It represents the probability of moving to a new state s', when an action a in taken at state s. Assume that p follows a discrete distribution F_p, whose support is taken by the set of historical data on the transition probabilities. Denote this set by $E_p = \{p_1, p_2, \ldots, p_J\}$. For a given policy $f \in PO_S$ and a realization $\omega \in \Omega$, the expected discounted reward $V_\alpha(f, p)(\omega)$ can be written as

$$V_\alpha(f, p)(\omega) = \sum_{s \in S} \sum_{a \in A(s)} \hat{m}(f, p, s, a)(\omega) R(s, a), \tag{3}$$

where $\hat{m}(f, p)(\omega)$ is an occupation measure corresponding to transition probability vector $p(\omega)$. Since the transition probabilities are random variables, it is clear that $\hat{m}(f, p)$ is a $|\mathcal{K}|-$dimensional random vector and $V_\alpha(f, p)$ is an $1-$dimensional random variable defined on same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We consider the case where decision maker is interested in maximizing the expected discounted reward which can be obtained with at least a given confidence level $(1 - \epsilon)$. This leads to the following

chance-constrained optimization problem

$$\begin{aligned} \text{(CCMDP)} \qquad & \sup_{y\in\mathbb{R},\ f\in PO_S} y \\ \text{s.t.} \quad & \mathbb{P}_p\left(V_\alpha(f,p)\geq y\right)\geq 1-\epsilon, \end{aligned} \tag{4}$$

In most of the practical situations, we only have partial information about the underlying probability distributions of p based on historical data of the transition probabilities. Such situations can be modelled with the distributionally robust optimization approach, where the decision marker believes that the distribution of p belongs to some uncertainty set $\mathcal{D}_p$. To ensure that the chance constraint $\mathbb{P}(V_\alpha(f,p)\geq y)\geq 1-\epsilon$ holds, we assume that it holds for any distribution which belongs to the uncertainty set. This leads to the following distributionally robust chance-constrained optimization problem

$$\begin{aligned} \text{(DRCCMDP)} \qquad & \sup_{y\in\mathbb{R},\ f\in PO_S} y \\ \text{s.t.} \quad & \text{(i)} \inf_{F_p\in\mathcal{D}_p} \mathbb{P}_p\left(V_\alpha(f,p)\geq y\right)\geq 1-\epsilon. \end{aligned} \tag{5}$$

In the following sections, we consider two types of uncertainty sets with moments based uncertainty sets and statistical distance based uncertainty sets.

2.1 Moments based uncertainty sets

We consider the uncertainty sets which are constructed based on available information about the first two moments of the true distribution of p. For each $s'\in S$, consider a $|\mathcal{K}|$-dimensional sub-vector $p(s')=(p(s,a,s'))_{(s,a)\in\mathcal{K}}$ of p. We estimate sample mean vector $\mu\in\mathbb{R}^{|S|\times|\mathcal{K}|}$ of p and positive definite sample covariance matrix $\Sigma(s')$ of $p(s')$ for all $s'\in S$ by observing sufficiently large number of data. We consider three most popular uncertainty sets, based on estimates μ and $\Sigma(s')$, which have been studied in the literature Delage and Ye (2010); Cheng et al (2014); Calafiore and El Ghaoui (2006). They are defined as follows:

1. **Uncertainty set with known mean and known covariance matrix:**

$$\mathcal{D}_1=\left\{F_p\in\mathcal{M}^+_{E_p}\,\middle|\,\begin{array}{l}\text{(i) } \mathbb{E}(\mathbf{1}_{\{p\in E_p\}})=1,\\ \text{(ii) } \mathbb{E}(p)=\mu,\\ \text{(iii) } \mathbb{E}[(p(s')-\mu(s'))(p(s')-\mu(s'))^{\mathrm{T}}]=\Sigma(s'),\ s'\in S.\end{array}\right\}, \tag{6}$$

2. **Uncertainty set with known mean and unknown covariance matrix:**

$$\mathcal{D}_2=\left\{F_p\in\mathcal{M}^+_{E_p}\,\middle|\,\begin{array}{l}\text{(i) } \mathbb{E}(\mathbf{1}_{\{p\in E_p\}})=1,\\ \text{(ii) } \mathbb{E}(p)=\mu,\\ \text{(iii) } \mathbb{E}[(p(s')-\mu(s'))(p(s')-\mu(s'))^{\mathrm{T}}]\preceq\delta_0\Sigma(s'),\ s'\in S.\end{array}\right\}, \tag{7}$$

3. **Uncertainty set with unknown mean and unknown covariance matrix:**

$$\mathcal{D}_3 = \left\{ F_p \in \mathcal{M}_{E_p}^{+} \;\middle|\; \begin{array}{l} \text{(i) } \mathbb{E}(\mathbf{1}_{\{p \in E_p\}}) = 1, \\ \text{(ii) } [\mathbb{E}\,(p(s')) - \mu(s')]^{\mathrm{T}} \Sigma(s')^{(-1)} [\mathbb{E}\,(p(s')) - \mu(s')] \leq \delta_1,\ s' \in S, \\ \text{(iii) } \mathbb{E}[(p(s') - \mu(s'))(p(s') - \mu(s'))^{\mathrm{T}}] \preceq \delta_2 \Sigma(s'),\ s' \in S. \end{array} \right\}, \tag{8}$$

where $\mathcal{M}_{E_p}^{+}$ is the set of all probability distributions on E_p, and $\delta_1 \geq 0, \delta_2, \delta_0 \geq 1$. The notation $A \preceq B$ implies that $B - A$ is a positive semidefinite matrix. We denote the set of $n \times n$ (positive semidefinite) symmetric matrices by ($\mathbf{S}_+^n$) $\mathbf{S}^n$ and $\circ$ denotes the Frobenius product. Using these notations, we present the deterministic reformulation of DRCCMDP problem (5) for each type of moments based uncertainty set defined above.

Theorem 1. *For DRCCMDP problem (5), the following results hold.*

(i) If the true distribution of p belongs to the uncertainty set $\mathcal{D}_1$, then (5) is equivalent to the following deterministic problem

$$\begin{array}{lll} & \sup & y \\ s.t. & (i) & -v - w^T \mu - \sum\limits_{s' \in S} z(s') \circ \Sigma(s') \geq 1 - \epsilon, \\ & (ii) & \mathbf{1}_{\{V_\alpha(f, p_j) \geq y\}} + v + w^T p_j + \sum\limits_{s' \in S} z(s') \circ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^T \geq 0, \\ & & \qquad\qquad \forall\ j = 1, \ldots, J, \\ & (iii) & z(s') \in \mathbf{S}^{|\mathcal{K}|}, \forall\ s' \in S \\ & (iv) & \sum\limits_{a \in A(s)} f(s,a) = 1,\ \forall\ s \in S,\ f(s,a) \geq 0,\ \forall\ (s,a) \in \mathcal{K}. \end{array} \tag{9}$$

(ii) If the true distribution of p belongs to the uncertainty set $\mathcal{D}_2$, then the optimization problem (5) is equivalent to the following deterministic problem

$$\begin{array}{lll} & \sup & y \\ s.t. & (i) & -v - w^T \mu - \sum\limits_{s' \in S} z(s') \circ \delta_0 \Sigma(s') \geq 1 - \epsilon, \\ & (ii) & \mathbf{1}_{\{V_\alpha(f, p_j) \geq y\}} + v + w^T p_j + \sum\limits_{s' \in S} z(s') \circ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^T \geq 0, \\ & & \qquad\qquad \forall\ j = 1, \ldots, J, \\ & (iii) & z(s') \in \mathbf{S}_+^{|\mathcal{K}|}, \forall\ s' \in S, \\ & (iv) & \sum\limits_{a \in A(s)} f(s,a) = 1,\ \forall\ s \in S,\ f(s,a) \geq 0,\ \forall\ (s,a) \in \mathcal{K}. \end{array} \tag{10}$$

(iii) If the true distribution of p belongs to the uncertainty set $\mathcal{D}_3$, then the optimization problem (5) is equivalent to the following deterministic problem

$$
\begin{aligned}
& \sup \quad y \\
s.t. \quad & (i) \ -v - \sum_{s' \in S} z(s') \circ \delta_2 \Sigma(s') \geq 1 - \epsilon, \\
& (ii) \ \mathbf{1}_{\{V_\alpha(f,p_j) \geq y\}} + v - \sum_{s' \in S} Q(s') \circ N_j(s') \\
& + \sum_{s' \in S} z(s') \circ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^T \geq 0, \forall \ j = 1, \ldots, J, \\
& (iii) \ z(s') \in \mathbf{S}_+^{|\mathcal{K}|}, \ Q(s') \in \mathbf{S}_+^{|\mathcal{K}|+1}, \ \forall \ s' \in S, \\
& (iv) \ \sum_{a \in A(s)} f(s,a) = 1, \ \forall \ s \in S, \ f(s,a) \geq 0, \ \forall \ (s,a) \in \mathcal{K}.
\end{aligned} \tag{11}
$$

where

$$
N_j(s') = \left(\begin{array}{c|c} \Sigma(s') & p_j(s') - \mu(s') \\ \hline (p_j(s') - \mu(s'))^T & \delta_1 \end{array} \right). \tag{12}
$$

Proof. For any $f \in PO_S$ and $y \in \mathbb{R}$, we consider the following optimization problem

$$
\inf_{F_p \in \mathcal{D}_p} \mathbb{P}_p \left(V_\alpha(f,p) \geq y \right). \tag{13}
$$

Note that p is a discrete distribution with finite support $E_p = \{p_1, \ldots, p_J\}$, then we can represent its true distribution F_p by its probability mass function which is a $J-$dimensional vector $q = (q_1, \ldots, q_J)$ such that $\sum_{j=1}^J q_j = 1, \ q_j \geq 0$ for all $j = 1, 2, \ldots, J$. By representing optimization problem (13) in terms of variable q, we show the deterministic reformulation of (5) for each uncertainty set.

(i) For uncertainty set $\mathcal{D}_1$, the optimization problem (13) can be rewritten as follows

$$
\begin{aligned}
& \inf_{q \geq 0} \quad \sum_{j=1}^J q_j \mathbf{1}_{\{V_\alpha(f,p_j) \geq y\}} \\
\text{s.t.} \quad & \text{(i)} \ \sum_{j=1}^J q_j = 1, \quad \text{(ii)} \ \sum_{j=1}^J q_j p_j = \mu, \\
& \text{(iii)} \sum_{j=1}^J q_j \ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^{\mathrm{T}} = \Sigma(s'), \ \forall \ s' \in S.
\end{aligned} \tag{14}
$$

The dual problem of (14) can be written as follows

$$\begin{aligned}
&\sup_{(v,w,z)} \; -v - w^{\mathrm{T}}\mu - \sum_{s'\in S} z(s') \circ \Sigma(s'), \\
\text{s.t.} \quad &\text{(i) } \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}} + v + w^{\mathrm{T}} p_j + \sum_{s'\in S} z(s') \circ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^{\mathrm{T}} \geq 0, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall\ j = 1, \ldots, J, \\
&\text{(ii) } z(s') \in \mathbf{S}^{|\mathcal{K}|},\ \forall\ s' \in S.
\end{aligned}$$

where $v \in \mathbb{R}$, $w \in \mathbb{R}^{|\mathcal{K}|.|S|}$ and $z(s')$, $s' \in S$, are the dual variables of (i), (ii), and (iii) of (14). Note that (14) is a linear program(LP), and therefore the strong duality holds which in turn implies that the DRCCMDP problem (5) can be reformulated as (9).

(ii) For the case of uncertainty set $\mathcal{D}_2$, the optimization problem (13) can be rewritten as follows

$$\begin{aligned}
&\inf_{q\geq 0} \; \sum_{j=1}^{J} q_j \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}} \\
\text{s.t.} \quad &\text{(i) } \sum_{j=1}^{J} q_j = 1, \quad \text{(ii) } \sum_{j=1}^{J} q_j p_j = \mu, \\
&\text{(iii)} \sum_{j=1}^{J} q_j\ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^{\mathrm{T}} \preceq \delta_0 \Sigma(s'),\ \forall\ s' \in S. \qquad (15)
\end{aligned}$$

The dual problem of (15) is given by

$$\begin{aligned}
&\sup \; -v - w^{\mathrm{T}}\mu - \sum_{s'\in S} z(s') \circ \delta_0\Sigma(s'), \\
\text{s.t.} \quad &\text{(i) } \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}} + v + w^{\mathrm{T}} p_j + \sum_{s'\in S} z(s') \circ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^{\mathrm{T}} \geq 0, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall\ j = 1, \ldots, J, \\
&\text{(ii) } z(s') \in \mathbf{S}_{+}^{|\mathcal{K}|}.
\end{aligned}$$

Since (15) is a semidefinite programming (SDP) problem, the strong duality holds, which ensures that the DRCCMDP problem (5) can be reformulated as (10).

(iii) For the case of uncertainty set $\mathcal{D}_3$, it follows from Schur complement that the constraints (ii) in (8) are equivalent to $N(s') \in \mathbf{S}_{+}^{|\mathcal{K}|+1}$, for any $s' \in S$, where

$$N(s') = \left(\begin{array}{c|c} \Sigma(s') & \mathbb{E}(p(s')) - \mu(s') \\ \hline (\mathbb{E}(p(s')) - \mu(s'))^{\mathrm{T}} & \delta_1 \end{array}\right),$$

Moreover, for any $s' \in S$, we can write $N(s') = \sum_{j=1}^{J} q_j N_j(s')$, where $N_j(s')$ is defined in (12). The rest of the proof follows from the same arguments used in the case of uncertainty set $\mathcal{D}_2$.

□

The deterministic reformulations (9), (10) and (11) have infinite number of constraints due to indicator function $\mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}$. Using the fact that $\mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}$ can take only binary variables, we further reformulate (9), (10) and (11) as mixed integer optimization problems. We present the proof for (9) and for (10) and (11) the proof follows using the same arguments. Let $M = \mathbf{1}_{|\mathcal{K}|}^T R$, where $\mathbf{1}_{|\mathcal{K}|}$ denotes $|\mathcal{K}|-$dimensional vector with all components equal to 1, and $\eta > 0$ be an infinitesimal number. The following theorem gives a mixed-integer reformulation of (9).

Theorem 2. *Suppose initial distribution $\gamma = (\gamma(s))_{s\in S}$ satisfy $\gamma(s) > 0$ for all $s \in S$. Then, the optimization problem (9) can be reformulated as the following mixed-integer problem*

$$
\begin{aligned}
(\textit{Bilinear-K-K}) \quad & \sup_{y,v,w,z,\beta,(\rho_j)_{j=1}^J} \quad y \\
s.t. \quad & (i) \ -v - w^T\mu - \sum_{s'\in S} z(s') \circ \Sigma(s') \geq 1-\epsilon, \\
& (ii) \ \beta_j + v + w^T p_j + \sum_{s'\in S} z(s') \circ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^T \geq 0, \\
& \qquad\qquad\qquad\qquad\qquad\qquad \forall\ j = 1,\ldots,J, \\
& (iii) \ z(s') \in \mathbf{S}^{|\mathcal{K}|}, \ \forall\ s' \in S, \\
& (iv) \ \rho_j^T R - y \leq M\beta_j - \eta(1-\beta_j), \ y - \rho_j^T R \leq M(1-\beta_j), \ j = 1,\ldots,J, \\
& (v) \ \beta_j \in \{0,1\}, \ j = 1,\ldots,J, \\
& (vi) \ \sum_{(s,a)\in\mathcal{K}} \rho_j(s,a)\left(\delta(s',s) - \alpha p_j(s,a,s')\right) = (1-\alpha)\gamma(s'), \ \forall\ s' \in S, \\
& (vii) \ \rho_j(s,a) \geq 0, \ \forall\ (s,a) \in \mathcal{K}, \ j = 1,\ldots,J, \\
& (viii) \ \rho_1(s,a) \sum_{a'\in A(s)} \rho_j(s,a') = \rho_j(s,a) \sum_{a'\in A(s)} \rho_1(s,a'), \\
& \qquad\qquad\qquad\qquad\qquad\qquad \forall\ (s,a) \in \mathcal{K}, \ j = 2,\ldots,J.
\end{aligned}
\tag{16}
$$

Proof. For an $f \in PO_S$, we can write $V_\alpha(f,p_j) = \hat{m}(f,p_j)^{\mathrm{T}} R$, where $\hat{m}(f,p_j)$ is the occupation measure vector defined by (2) corresponding to policy f and transition probabilities vector p_j. Therefore, $V_\alpha(f,p_j) \in [0,M]$ which implies that any optimal solution y of (5) belongs to $[0,M]$. Then, we can add the constraint $y \in [0,M]$ without loss of optimality. We show that $\beta_j = \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}$ is equivalent to following constraints

$$
\text{(i)} \ V_\alpha(f,p_j) - y \leq M\beta_j - \eta(1-\beta_j), \quad y - V_\alpha(f,p_j) \leq M(1-\beta_j), \ j = 1,\ldots,J,
$$

(ii) $\beta_j \in \{0,1\}$. (17)

1. Let $\beta_j = \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}$. If $V_\alpha(f,p_j) < y$, then $\beta_j = 0$, $y - V_\alpha(f,p_j) \leq M$ and $V_\alpha(f,p_j) - y \leq -\eta$ for sufficiently small positive η. If $V_\alpha(f,p_j) \geq y$, then $\beta_j = 1$, $V_\alpha(f,p_j) - y \leq M$ and $y - V_\alpha(f,p_j) \leq 0$. Therefore, the constraints given by (17) are satisfied.
2. Suppose the constraints given by (17) are satisfied. If $\beta_j = 0$, then $V_\alpha(f,p_j) < y$. If $\beta_j = 1$, then $V_\alpha(f,p_j) \geq y$. Therefore, $\beta_j = \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}$.

It follows from the discussion in Section 1 that for all $j = 1,\ldots,J$, the set of occupation measures $\{\hat{m}(f,p_j) \mid f \in PO_S\}$ is equivalent to the set $\mathcal{Q}(p_j)$. Since $\gamma(s) > 0$ for all $s \in S$, $\sum_{a'\in A(s)} \rho_j(s,a') > 0$ for all $j = 1,2,\ldots,J$. This implies that for every $f \in PO_S$, there exists $\rho_j \in \mathcal{Q}(p_j)$ such that

$$V_\alpha(f,p_j) = \rho_j^T R, \ \forall\ j = 1,\ldots,J,$$
$$f(s,a) = \frac{\rho_j(s,a)}{\sum_{a'\in A(s)} \rho_j(s,a')}, \ \forall\ (s,a) \in \mathcal{K}, \ j = 1,\ldots,J.$$

Hence, optimization problem (9) is equivalent to the optimization problem (16). The constraint $(viii)$ ensures that all ρ_j, $j = 1,\ldots,J$ corresponds to the same policy f. □

Remark 1. *If $(y^*, v^*, w^*, z^*, \beta^*, (\rho_j^*)_{j=1}^J$ is an optimal solution of (16), the optimal policy of DRCCMDP problem (5) is defined by*

$$f^*(s,a) = \frac{\rho_1^*(s,a)}{\sum_{a'\in A(s)} \rho_1^*(s,a')}, \ \forall\ (s,a) \in \mathcal{K}.$$

Similarly, the reformulations of DRCCMDP problem (5) for the case of uncertainty sets $\mathcal{D}_2$ and $\mathcal{D}_3$ are given as below

$$\begin{aligned}
\text{(Bilinear-K-U)} \quad & \sup_{y,v,w,z,\beta,(\rho_j)_{j=1}^J} \quad y \\
\text{s.t.} \quad & \text{(i)} \ -v - w^{\mathrm{T}}\mu - \sum_{s'\in S} z(s') \circ \delta_0 \Sigma(s') \geq 1-\epsilon, \\
& \text{(ii)} \ \beta_j + v + w^{\mathrm{T}} p_j + \sum_{s'\in S} z(s') \circ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^{\mathrm{T}} \geq 0, \\
& \qquad\qquad\qquad\qquad\qquad\qquad \forall\ j = 1,\ldots,J, \\
& \text{(iii)} \ z(s') \in \mathbb{S}_+^{|\mathcal{K}|}, \ \forall\ s' \in S, \\
& \text{(iv)} - \text{(viii)} \text{ of (16)}.
\end{aligned} \tag{18}$$

$$\text{(Bilinear-U-U)} \quad \sup_{y,v,Q,z,\beta,(\rho_j)_{j=1}^J} \quad y$$

$$
\begin{aligned}
\text{s.t.} \quad & \text{(i)} \ -v - \sum_{s' \in S} z(s') \circ \delta_2 \Sigma(s') \geq 1 - \epsilon, \\
& \text{(ii)} \ \beta_j + v - \sum_{s' \in S} Q(s') \circ N_j(s') \\
& + \sum_{s' \in S} z(s') \circ [p_j(s') - \mu(s')][p_j(s') - \mu(s')]^{\mathrm{T}} \geq 0, \ \forall \ j = 1, \ldots, J, \\
& \text{(iii)} \ z(s') \in \mathbb{S}_+^{|\mathcal{K}|}, \ Q(s') \in \mathbb{S}_+^{|\mathcal{K}|+1}, \ \forall \ s' \in S, \\
& \text{(iv)} - \text{(viii) of (16)}.
\end{aligned} \tag{19}
$$

2.2 Statistical distance based uncertainty sets

In this section, we consider uncertainty sets defined by ϕ-divergence and Wasserstein distance metrics. In such uncertainty sets, a reference distribution ν is known to the decision maker based on the available estimated data of transition probabilities. The decision maker believes that the true distribution of transition probabilities vector p, denoted by F_p belongs to a ball centered at the reference distribution ν. We assume that both ν and F_p are discrete distributions on same support E_p. Let $(q_j^0)_{j=1}^J$ be the probability mass function of the reference distribution ν, i.e., q_j^0 is the weight of jth atom p_j. We assume that $q_j^0 > 0$ for all $j = 1, \ldots, J$.

2.2.1 Uncertainty set with ϕ -divergence distance

The $\phi-$divergence distance between two discrete probability distributions ν_1 and ν_2 with support E_p is given by

$$
I_\phi(\nu_1, \nu_2) = \sum_{j=1}^{J} \phi\left(\frac{\nu_1(p_j)}{\nu_2(p_j)}\right) \nu_2(p_j),
$$

where $\nu_1(p_j)$ (resp. $\nu_2(p_j)$) is the weight of ν_1 (resp. ν_2) on the jth atom p_j of E_p and ϕ is a convex function on $\mathbb{R}^+$. For general ϕ-divergence with the choices of function ϕ and its properties, we refer to Ben-Tal et al (2013). The uncertainty set of the distribution of p based on ϕ-divergence is defined by

$$
\mathcal{D}_4 = \left\{ F_p \in \mathcal{M}_{E_p}^+ \ \mid \ I_\phi(F_p, \nu) \leq \theta_\phi \right\}, \tag{20}
$$

where $\theta_\phi > 0$ denotes the radius. The following definition of the conjugate of a function is useful for our subsequent analysis.

Definition 1. *The conjugate of ϕ is a function $\phi^* : \mathbb{R} \to \mathbb{R} \cup \infty$ such that*

$$
\phi^*(r) = \sup_{t \geq 0} \left\{ rt - \phi(t) \right\}, \ \forall \ r \in \mathbb{R}.
$$

Table 1 presents four special types of ϕ-divergences with their conjugate.

Table 1 List of selected ϕ-divergences with their conjugate

Divergence	$\phi(t), t \geq 0$	$\phi^*(r)$
Kullback-Leibler	$t\log(t) - t + 1.$	$\mathrm{e}^r - 1$
Variation distance	$\lvert t-1\rvert.$	$\begin{array}{ll} -1, & r \leq -1, \\ r, & -1 \leq r \leq 1, \\ \infty, & r > 1. \end{array}$
Modified χ^2 - distance	$(t-1)^2.$	$\begin{array}{ll} -1, & r \leq -2, \\ r + \frac{r^2}{4}, & r > -2. \end{array}$
Hellinger distance	$(\sqrt{t}-1)^2.$	$\begin{array}{ll} \frac{r}{1-r}, & r < 1, \\ \infty, & r \geq 1. \end{array}$

Lemma 1. *The dual of an optimization problem*

$$\inf_{F_p \in \mathcal{D}_4} \mathbb{P}_{F_p}((V_\alpha(f,p) \geq y), \tag{21}$$

is given by

$$\sup_{\lambda>0, \beta \in \mathbb{R}} \left\{ \beta - \lambda\theta_\phi - \lambda \mathbb{P}_\nu \left(V_\alpha(f,p) \geq y\right) \phi^* \left(\frac{\beta - 1}{\lambda}\right) - \lambda \left[1 - \mathbb{P}_\nu \left(V_\alpha(f,p) \geq y\right)\right] \phi^* \left(\frac{\beta}{\lambda}\right) \right\}, \tag{22}$$

and the strong duality between (21) and (22) holds.

Proof. The primal problem (21) can be written as a following optimization problem

$$\begin{aligned} v_{\mathrm{P}} &= \inf_{q \geq 0} \sum_{j=1}^{J} q_j \mathbf{1}_{\{V_\alpha(f,p_j) \geq y\}} \\ \text{s.t.} \quad &\text{(i)} \ \sum_{j=1}^{J} q_j^0 \phi\left(\frac{q_j}{q_j^0}\right) \leq \theta_\phi, \ \text{(ii)} \ \sum_{j=1}^{J} q_j = 1. \end{aligned} \tag{23}$$

The dual problem of (23) is given by

$$v_{\mathrm{D}} = \sup_{\lambda \geq 0, \beta \in \mathbb{R}} \left\{ \beta - \lambda\theta_\phi + \inf_{q \geq 0} \sum_{j=1}^{J} \left[q_j \mathbf{1}_{\{V_\alpha(f,p_j) \geq y\}} - \beta q_j + \lambda q_j^0 \phi\left(\frac{q_j}{q_j^0}\right) \right] \right\},$$

where λ is the dual variable of the constraint (i) of (23) and β is the dual variable of the constraint (ii) of (23). Note that (23) is a convex optimization problem because ϕ is a convex function. Since $\theta_\phi > 0$, it is clear that for $q = q^0$, (i) of (23) is strictly feasible. Hence, the Slater's condition is satisfied which guarantees the strong duality, i.e., $v_{\mathrm{P}} = v_{\mathrm{D}}$. It remains to show that v_{D} is same as (22). For a given $\beta \in \mathbb{R}$, consider a function $F(\lambda, q) = \sum_{j=1}^{J} \left[q_j \mathbf{1}_{\{V_\alpha(f,p_j) \geq y\}} - \beta q_j + \lambda q_j^0 \phi\left(\frac{q_j}{q_j^0}\right) \right]$. It is clear that $F(\lambda, q)$

is a linear function of λ. Let $G(\lambda) = -\lambda\theta_\phi + \inf_{q\geq 0} F(\lambda, q)$. For any $t \in [0,1]$ and $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, we have

$$\begin{aligned}
&\inf_{q\geq 0} F[t\lambda_1 + (1-t)\lambda_2, q] = \inf_{q\geq 0} \left(tF(\lambda_1, q) + (1-t)F(\lambda_2, q)\right) \\
&\geq t \inf_{q\geq 0} F(\lambda_1, q) + (1-t) \inf_{q\geq 0} F(\lambda_2, q).
\end{aligned}$$

This implies that $G(\lambda)$ is a concave function of λ on $[0, \infty)$. Consider the optimization problem $\sup_{\lambda\geq 0} G(\lambda)$. We prove that $G(0) \leq \sup_{\lambda>0} G(\lambda)$. In fact, if $G(0) > \sup_{\lambda>0} G(\lambda)$. Then, there exists a t sufficiently close to 1 such that $tG(0)+(1-t)G(1) > \sup_{\lambda>0} G(\lambda)$. Since $G(\lambda)$ is a concave function, $G(1-t) \geq tG(0) + (1-t)G(1)$ which in turn implies that $G(1-t) > \sup_{\lambda>0} G(\lambda)$. This gives a contradiction. Therefore, $G(0) \leq \sup_{\lambda>0} G(\lambda)$ and we can restrict on $\lambda > 0$ without loss of optimality in the dual problem v_{D}. Hence, the dual problem of (21) can be rewritten as follows

$$\begin{aligned}
v_{\mathrm{D}} &= \sup_{\lambda>0,\beta\in\mathbb{R}} \left\{ \beta - \lambda\theta_\phi + \inf_{q\geq 0} \sum_{j=1}^{J} \left[q_j \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}} - \beta q_j + \lambda q_j^0 \phi\left(\frac{q_j}{q_j^0}\right)\right]\right\}, \\
&= \sup_{\lambda>0,\beta\in\mathbb{R}} \left\{ \beta - \lambda\theta_\phi + \sum_{j=1}^{J} \inf_{q_j\geq 0} \left[q_j \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}} - \beta q_j + \lambda q_j^0 \phi\left(\frac{q_j}{q_j^0}\right)\right]\right\}, \\
&= \sup_{\lambda>0,\beta\in\mathbb{R}} \left\{ \beta - \lambda\theta_\phi + \sum_{j=1}^{J} \inf_{q_j\geq 0} \lambda q_j^0 \left[\phi\left(\frac{q_j}{q_j^0}\right) - \frac{q_j}{q_j^0}\left(\frac{\beta - \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}}{\lambda}\right)\right]\right\}, \\
&= \sup_{\lambda>0,\beta\in\mathbb{R}} \left\{ \beta - \lambda\theta_\phi - \sum_{j=1}^{J} \sup_{q_j\geq 0} \lambda q_j^0 \left[\frac{q_j}{q_j^0}\left(\frac{\beta - \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}}{\lambda}\right) - \phi\left(\frac{q_j}{q_j^0}\right)\right]\right\}.
\end{aligned}$$

Let $t_j = \frac{q_j}{q_j^0}$, using the definition of the conjugate of a function, the dual problem v_{D} can be rewritten as follows

$$\begin{aligned}
v_{\mathrm{D}} &= \sup_{\lambda>0,\beta\in\mathbb{R}} \left\{ \beta - \lambda\theta_\phi - \sum_{j=1}^{J} \lambda q_j^0 \phi^*\left(\frac{\beta - \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}}{\lambda}\right)\right\}, \\
&= \sup_{\lambda>0,\beta\in\mathbb{R}} \left\{ \beta - \lambda\theta_\phi - \lambda\mathbb{P}_\nu\left(V_\alpha(f,p)\geq y\right)\phi^*\left(\frac{\beta-1}{\lambda}\right) - \lambda\left[1 - \mathbb{P}_\nu\left(V_\alpha(f,p)\geq y\right)\right]\phi^*\left(\frac{\beta}{\lambda}\right)\right\},
\end{aligned}$$

which completes the proof. □

Lemma 2. *Consider the DRCCMDP problem (5) under the uncertainty set defined by (20) for the ϕ-divergences listed in Table 2. Then, the DRCCMDP problem (5) can be rewritten as follows*

$$\begin{aligned}
&\sup_{y\in\mathbb{R},\ f\in PO_S} y \\
&s.t. \quad (i)\ \mathbb{P}_\nu\left(V_\alpha(f,p)\geq y\right) \geq g^*(\theta_\phi, \epsilon),
\end{aligned} \tag{24}$$

where the values of θ_ϕ, ϵ *and* $g^*(\theta_\phi, \epsilon)$ *for different* ϕ*-divergences are given in Table 2.*

Table 2 The function f for selected $\phi-$divergences

Divergence	$g^*(\theta_\phi, \epsilon)$	θ_ϕ, ϵ
K-L	$\inf_{x\in(0,1)} \frac{e^{-\theta_\phi x^{1-\epsilon}-1}}{x-1}$	$\theta_\phi > 0$ $0 < \epsilon < 1$
Variation	$1-\epsilon+\frac{\theta_\phi}{2}$	$\theta_\phi > 0$ $0 < \epsilon < 1$
Modified χ^2	$1-\epsilon+\frac{\sqrt{\theta_\phi^2+4\theta_\phi(\epsilon-\epsilon^2)}-(1-2\epsilon)\theta_\phi}{2\theta_\phi+2}$	$\theta_\phi > 0$ $0 < \epsilon < \frac{1}{2}$
Hellinger	$\frac{-B+\sqrt{\Delta}}{2}$, where $B = -(2-(2-\theta_\phi)^2)\epsilon - \frac{(2-\theta_\phi)^2}{2}$, $C = \left(\frac{(2-\theta_\phi)^2}{4}-\epsilon\right)^2$, $\Delta = B^2-4C = (2-\theta_\phi)^2\left[4-(2-\theta_\phi)^2\right]\epsilon(1-\epsilon)$.	$0 < \theta_\phi < 2-\sqrt{2}$ $0 < \epsilon < 1$

Proof. The details of the proof for the Hellinger distance case is given in Appendix A. The proofs for Kullback-Leibler, Variation distance and Modified χ^2 - distance follow from Propositions 2, 3 and 4 of Jiang and Guan (2016), respectively. □

Theorem 3. *Suppose initial distribution* $\gamma = (\gamma(s))_{s\in S}$ *satisfy* $\gamma(s) > 0$ *for all* $s \in S$. *Then, the optimization problem* (24) *can be reformulated as the following mixed-integer optimization problem*

$$\begin{aligned} (\textit{Bilinear-Phi}) \quad & \sup_{y,\beta,(\rho_j)_{j=1}^J} \quad y \\ & s.t. \quad (i)\ \sum_{j=1}^{J} q_j^0\ \beta_j \geq g^*(\theta_\phi,\epsilon), \forall\ j = 1,\ldots,J, \\ & \qquad (iv)-(viii)\ \textit{of}\ (16). \end{aligned} \tag{25}$$

Proof. Using Lemma 2, the optimization problem (24) can be written as

$$\begin{aligned} & \sup \quad y \\ \text{s.t.} \quad & \text{(i)}\ \sum_{j=1}^{J} q_j^0\ \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}} \geq g^*(\theta_\phi,\epsilon), \\ & \text{(ii)}\ \sum_{a\in A(s)} f(s,a) = 1,\ \forall\ s\in S,\ f(s,a)\geq 0,\ \forall\ (s,a)\in\mathcal{K}. \end{aligned} \tag{26}$$

By taking $\beta_j = \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}$ for all $j = 1,\ldots,J$, and using the same arguments used in Theorem 2 the optimization problem (26) is equivalent to the mixed integer optimization problem (25) □

2.2.2 Uncertainty set with Wasserstein distance

The Wasserstein distance $W_d(F_p, \nu)$ between true distribution F_p and reference distribution ν with support on E_p, is given by

$$W_d(F_p,\nu) = \left[\inf\left\{\sum_{i,j=1}^{J}\omega_{ij}||p_i - p_j||_2^d \mid \sum_{j=1}^{J}\omega_{ij} = q_i^0,\ \sum_{i=1}^{J}\omega_{ij} = q_j,\ \omega_{ij}\geq 0, \forall\ i,j\right\}\right]^{\frac{1}{d}},$$

where $d \geq 1$ is some constant. For each i and j, w_{ij} is the weight transported from the ith atom of the reference distribution ν to the jth atom of the true distribution F_p and q_i^0 is the weight of ith atom of reference distribution ν and q_j is the weight of jth atom of F_p. For more details on the Wasserstein distance metric, we refer to Villani et al (2009); Villani (2021). The uncertainty set of p using Wasserstein distance is defined by

$$\mathcal{D}_5 = \left\{F_p \in \mathcal{M}_{E_p}^{+} \mid W_d(F_p,\nu) \leq \theta_W\right\}, \tag{27}$$

where $\theta_W > 0$. We have the following lemma.

Lemma 3. *The DRCCMDP problem (5) under the uncertainty set defined by (27) is equivalent to the following optimization problem*

$$\begin{aligned}
&\sup_{y,v,h,f}\ y\\
s.t.\quad &(i)\ -\sum_{i=1}^{J} q_i^0 v_i - h\theta_W^d \geq 1-\epsilon,\\
&(ii)\ h \geq 0,\\
&(iii)\ \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}} + v_i + hc_{ij}^d \geq 0,\ \forall\ i,j = 1,\ldots,J,\\
&(iv)\ \sum_{a\in A(s)} f(s,a) = 1,\ \forall\ s\in S,\ f(s,a)\geq 0,\ \forall\ (s,a)\in\mathcal{K},
\end{aligned} \tag{28}$$

where $c_{ij}^d = ||p_i - p_j||_2^d$.

Proof. The optimization problem $\inf_{F_p\in\mathcal{D}_5}\mathbb{P}_p\left(V_\alpha(f,p)\geq y\right)$ can be equivalently written as follows

$$\begin{aligned}
&\inf_{\omega\geq 0, q\geq 0}\sum_{j=1}^{J} q_j \mathbf{1}_{\{V_\alpha(f,p_j)\geq y\}}\\
\text{s.t.}\quad &\text{(i)}\ \sum_{j=1}^{J}\omega_{ij} = q_i^0,\ \forall\ i = 1,\ldots,J,\\
&\text{(ii)}\ \sum_{i=1}^{J}\omega_{ij} = q_j,\ \forall\ j = 1,\ldots,J,
\end{aligned}$$

$$\text{(iii)} \quad \sum_{i,j=1}^{J} \omega_{ij} c_{ij}^d \leq \theta_W^d. \tag{29}$$

By replacing $q_j = \sum_{j=1}^{J} \omega_{ij}$ in (29), we have

$$\begin{aligned} &\inf_{\omega \geq 0} \sum_{i,j=1}^{J} \omega_{ij} \mathbf{1}_{\{V_\alpha(f,p_j) \geq y\}} \\ \text{s.t.} \quad &\text{(i)} \sum_{j=1}^{J} \omega_{ij} = q_i^0, \ \forall\ i = 1, \ldots, J \\ &\text{(ii)} \sum_{i,j=1}^{J} \omega_{ij} c_{ij}^d \leq \theta_W^d. \end{aligned} \tag{30}$$

The dual problem of (30) is given by

$$\begin{aligned} &\sup_{v \in \mathbb{R}^J,\ h \geq 0} -\sum_{i=1}^{J} q_i^0 v_i - h\theta_W^d \\ \text{s.t.} \quad &\text{(i)}\ \mathbf{1}_{\{V_\alpha(f,p_j) \geq y\}} + v_i + h c_{ij}^d \geq 0, \ \forall\ i, j = 1, \ldots, J. \end{aligned} \tag{31}$$

Note that (30) is an LP, which ensures that the strong duality holds. Therefore, the DRCCMDP problem (5) is equivalent to (28). □

Theorem 4. *Suppose initial distribution $\gamma = (\gamma(s))_{s \in S}$ satisfy $\gamma(s) > 0$ for all $s \in S$. Then, the optimization problem (28) can be reformulated as the following mixed-integer optimization problem*

$$\begin{aligned} (\textit{Bilinear-Wasserstein}) \quad &\sup_{y,v,h,\beta,(\rho_j)_{j=1}^J} \quad y \\ \textit{s.t.} \quad &(i),\ (ii)\ \textit{of}\ (28) \\ &(ii)\ \beta_j + v_i + h c_{ij}^d \geq 0, \ \forall\ i, j = 1, \ldots, J, \\ &(iv) - (viii)\ \textit{of}\ (16). \end{aligned} \tag{32}$$

Proof. By taking $\beta_j = \mathbf{1}_{\{V_\alpha(f,p_j) \geq y\}}$ for all $j = 1, \ldots, J$, and using the same arguments used in Theorem 2 the optimization problem (28) is equivalent to the mixed integer optimization problem (32) □

2.3 Solving mixed integer programming problems with bilinear and positive semidefinite cone constraints

The equivalent mixed-integer programming problems proposed in Sections 2.1 and 2.2 to solve DRCCMDP problem (5) for different types of moments based and statistical distance based uncertainty sets are quite similar. For example, they have the

same objective function, and the same set of integer and bilinear equality constraints. Additionally, these problems have (i) linear constraints corresponding to moments based uncertainty set defined by (6) and statistical distance based uncertainty sets defined by (20) and (27), (ii) linear constraints as well as positive semidefinite cone constraints corresponding to moments based uncertainty sets defined by (7) and (8). The equivalent optimization problems corresponding to the uncertainty sets defined by (6), (20) and (27) are called MIBP problems. The equivalent optimization problems corresponding to the uncertainty sets defined by (7) and (8) are called MISDP problems with bilinear constraints.

The MIBP problems can be solved efficiently by GUROBI 9.0 or higher version. It deals bilinear constraints by relaxing them using linear constraints based on McCormick lower and upper envelopes McCormick (1976), which depend on the local bounds of the variables present in the bilinear terms. In our case, all the mixed integer programming problems have following bilinear constraints in common

$$\rho_1(s,a) \sum_{a' \in A(s)} \rho_j(s,a') = \rho_j(s,a) \sum_{a' \in A(s)} \rho_1(s,a'),\ \forall\ (s,a) \in \mathcal{K},\ j = 2,\ldots,J. \tag{33}$$

Let

$$\begin{aligned}
&\sum_{a' \in A(s)} \rho_j(s,a') = M_j(s),\ \forall\ j = 1,\ldots,J,\ s \in S,\\
&\rho_1(s,a) M_j(s) = C_j(s,a),\ \forall\ j = 2,\ldots,J,\ s \in S,\ a \in A(s),\\
&\rho_j(s,a) M_1(s) = D_j(s,a),\ \forall\ j = 2,\ldots,J,\ s \in S,\ a \in A(s).
\end{aligned}$$

By introducing auxiliary variables $M_j(s),\ C_j(s,a),\ D_j(s,a)$, the constraints (33) are equivalent to the following set of constraints

$$\begin{aligned}
&\text{(i)} \sum_{a' \in A(s)} \rho_j(s,a') = M_j(s),\ \forall\ j = 1,\ldots,J,\ s \in S,\\
&\text{(ii)}\ C_j(s,a) = D_j(s,a),\ \forall\ j = 2,\ldots,J,\ s \in S,\ a \in A(s),\\
&\text{(iii)}\ \rho_1(s,a) M_j(s) = C_j(s,a),\ \forall\ j = 2,\ldots,J,\ s \in S,\ a \in A(s),\\
&\text{(iv)}\ \rho_j(s,a) M_1(s) = D_j(s,a),\ \forall\ j = 2,\ldots,J,\ s \in S,\ a \in A(s).
\end{aligned} \tag{34}$$

We propose tight bounds for $\rho_j(s,a)$ and $M_j(s)$ which are used in generating the McCormick envelopes of bilinear terms present in (iii) and (iv) of (34). It is well known that for the occupation measure ρ_j there exists a policy f such that (see Altman (1999))

$$\rho_j(s,a) = f(s,a) M_j(s),\ \forall\ s \in S,\ a \in A(s),\ j = 1,\ldots,J,$$

and

$$M_j = (M_j(s))_{s \in S} = (1-\alpha)\gamma^{\mathrm{T}} \left[I - \alpha P_j(f)\right]^{(-1)},$$

$$= (1-\alpha)\gamma^{\mathrm{T}} \left[I + \sum_{i=1}^{\infty} (\alpha P_j(f))^i \right], \ \forall\ j = 1, \ldots, J,$$

where $P_j(f)$ is a transition probability matrix induced by transition probabilities p_j and stationary policy f, which is given by

$$P_j(f)(s, s') = \sum_{a \in A(s)} f(s, a) p_j(s, a, s'), \ \forall\ s, s' \in S.$$

Since $\sum_{a \in A(s)} f(s, a) = 1$, we have $P_j(f)(s, s') \geq \min_{a \in A(s)} p_j(s, a, s')$, for any $s, s' \in S$. Let $P_{j,\min}(s, s') = \min_{a \in A(s)} p_j(s, a, s')$ and $P_{j,\min} = (P_{j,\min}(s, s'))_{s,s' \in S}$ be a $|S| \times |S|$ matrix. We derive lower bound for $M_j(s)$ as follows

$$\begin{aligned} M_j(s) &= (1-\alpha)\gamma^{\mathrm{T}} \left[I + \sum_{i=1}^{\infty} (\alpha P_j(f))^i \right]_s \\ &\geq (1-\alpha)\gamma^{\mathrm{T}} \left[I + \sum_{i=1}^{\infty} (\alpha P_{j,\min})^i \right]_s = (1-\alpha)\gamma^{\mathrm{T}} \left[I - \alpha P_{j,\min} \right]_s^{(-1)}. \end{aligned}$$

Denote $M_j^l(s) = (1-\alpha)\gamma^{\mathrm{T}} \left[I - \alpha P_{j,\min} \right]_s^{(-1)}$ and $M_j^u(s) = 1 - \sum_{s' \neq s} M_j^l(s')$. Since $\sum_{s \in S} M_j(s) = 1$, a lower bound and an upper bound of $M_j(s)$ are given by

$$M_j^l(s) \leq M_j(s) \leq M_j^u(s), \ \forall\ j = 1, \ldots, J, \ s \in S. \tag{35}$$

Since $0 \leq \rho_j(s, a) \leq M_j(s)$, for all $s \in S, \ a \in A(s), \ j = 1, \ldots, J$, the lower and upper bounds of $\rho_j(s, a)$ are given by

$$0 \leq \rho_j(s, a) \leq M_j^u(s), \ \forall\ j = 1, \ldots, J, \ s \in S, \ a \in A(s). \tag{36}$$

The MISDP problems with bilinear constraints are difficult to solve in general. To the best of our knowledge, there is no commercial solver which can solve efficiently this type of optimization problem. We propose using CUTSDP solver, which is an internal solver in YALMIP toolbox of MATLAB. One of the core ideas in YALMIP is to rely on external solvers for time-consuming tasks. A positive semidefinite constraint $A \succeq 0$, where A is an $n \times n$ matrix, is equivalent to an infinite number of linear constraints $x^{\mathrm{T}} A x \geq 0$, for all $x \in \mathbb{R}^n$. CUTSDP solver uses a cutting plane (outer approximation) approach, which relaxes positive semidefinite constraint by finite number of linear constraints, solves the relaxation problem and add violated linear cuts to the model and repeat the same procedure. The main idea of CUTSDP is to appropriately generate several vectors x to construct a finite number of linear constraints, that leads to an outer approximation of the original problem. The approximation problem is a MIBP problem which is solved by an external solver (such as GUROBI). If the solution of approximation problem does not satisfy the original positive semidefinite constraint,

the solver will add a cutting plane based on a negative eigenvalue, and repeats the process, with a hope that after some iterations, it will eventually satisfy the semidefinite constraint.

3 Machine replacement problem

In this section, we present a series of numerical experiments performed on a machine replacement problem to compare the approaches discussed earlier. These comparisons aim to evaluate the performance and efficiency of the different reformulations and solvers in solving the respective problems. By conducting these comparisons, we can gain insights into the strengths and limitations of each approach and make informed decisions based on the specific problem characteristics. All the numerical results below are performed using Matlab R2023a on an Intel Core i5-1135G7, Processor 2.4 GHz (8M Cache, up to 4.2 GHz), RAM 16G, 512G SSD.

We consider a machine replacement problem where a machine in a factory has a life-time of N years. At every stage a maintenance of the machine is scheduled but a factory owner can decide whether to repair the machine or not. There is a high probability that the machine behaves like a new one if it is being repaired and its life gets reduced by a year if it is not being repaired. The factory owner incurs maintenance cost if he decides to repair the machine. It can be modelled as an MDP problem where the life of a machine represents the state of underlying Markov chain, i.e., there are $N+1$ states. The first state represents a brand new machine. At each state there are two actions: i) "repair", ii) "do not repair". Figure 1 presents a case of fixed transition probabilities of the Markov chain with respect to each action. The maintenance cost corresponding to every state-action pair is not exactly known and is realised after the decision is made. Therefore, it is modelled with a random variable. We assume that for every state action pair (s,a), the maintenance cost is defined as $\hat{c}(s,a) = K + \hat{Z}(s,a)$, where K represents the fixed cost and $\hat{Z}(s,a)$ represents a variable cost which is a random variable. The machine generates a revenue $L(s,a)$ at state-action pair (s,a) and the profit for each $(s,a) \in \mathcal{K}$ is given by

$$\hat{R}(s,a) = L(s,a) - K - \hat{Z}(s,a). \tag{37}$$

We define the reward vector R by taking the expected value of $\hat{R}$, i.e., $R(s,a) = \mathbb{E}(\hat{R}(s,a)),\ \forall\ s \in S,\ a \in A(s)$. The reward vector is given by

$$R(s,a) = L(s,a) - K - \mu_{\hat{Z}}(s,a), \tag{38}$$

where $K = 10$ and $L, \mu_{\hat{Z}}$ are given in Table 3, $\mu_{\hat{Z}}$ is the mean vector of $\hat{Z}$. We randomly simulate 100 transition probabilities $p_j,\ j = 1, \ldots, 100$ as follows. For any $s, s' = 1, \ldots, N+1$ and $a = 1, 2$, we consider the following cases.

- **Case 1:** If $a = 1$, i.e., the decision is to repair the machine. Assume that we are actually at state s. If $2 \leq s \leq N+1$, we come back to state 1 with high probability $q_{1,s}$ and stay at state s with probability $1 - q_{1,s}$. Then, $p(s,a,s') = 1 - q_{1,s}$ if $s' = s$,

$p(s,a,s') = q_{1,s}$ if $s' = 1$ and $p(s,a,s') = 0$, otherwise. We randomly generate $q_{1,s}$ on $[0.7, 1]$. If $s = 1$, then $p(s,a,s') = 1$ if $s' = 1$ and $p(s,a,s') = 0$, otherwise.

- **Case 2:** If $a = 2$, i.e., the decision is to do not repair the machine. If $1 \leq s \leq N$, we move to the next state $s+1$ with high probability $q_{2,s}$ and stay at actual state s with probability $1 - q_{2,s}$. Then, $p(s,a,s') = 1 - q_{2,s}$ if $s' = s$, $p(s,a,s') = q_{2,s}$ if $s' = s+1$ and $p(s,a,s') = 0$, otherwise. We randomly generate $q_{2,s}$ on $[0.7, 1]$. If $s = N+1$, then $p(s,a,s') = 1$ if $s' = N+1$ and $p(s,a,s') = 0$, otherwise.

The transition probability vector p is $|S|.|\mathcal{K}|$-dimensional random vector with support on $E_p = \{p_1, \ldots, p_J\}$. Let μ be the sample mean of p and $\Sigma(s')$ be the sample covariance matrix of $p(s')$, for any $s' \in S$, where μ and $\Sigma(s')$ are given as follows

$$\mu = \frac{1}{J}\sum_{j=1}^{J} p_j,$$

$$\Sigma(s') = \frac{1}{J-1}\sum_{j=1}^{J}(p_j(s') - \mu(s'))(p_j - \mu(s'))^{\mathrm{T}}.$$

The factory owner is interested in maximizing the expected discounted profit. We assume that the factory owner has a finite number of the same machines which are modelled using the same Markov chain. Therefore, we compute the optimal repair policy with respect to a single machine and the same repair policy can be applied for all other machines.

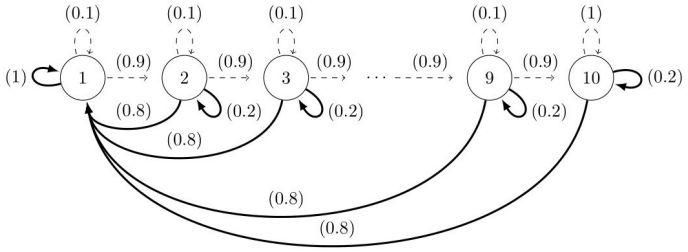

Fig. 1 Machine replacement *MDP* with two actions: "repair" (with solid lines) and "do not repair" (with dashed lines)

In our numerical experiments, we set the number of states to 10, the threshold value $\epsilon = 0.1$, the discount parameter $\alpha = 0.85$ and the initial distribution of states γ to be uniformly distributed. For the above instance, $|\mathcal{K}| = 20$. We use the above μ and $\Sigma(s')$, $s' \in S$ for all the moments based uncertainty sets. The reference distribution in all statistical distance based uncertainty sets is assumed to be uniformly distributed on E_p. We summarize the other parameters related to all the uncertainty sets in Table 4.

Table 3 Random cost $\hat{Z}$ and Revenue L

State(s) \ Action(a)	"Repair" $\mu_{\hat{Z}}(s,1)$	"Do not repair" $\mu_{\hat{Z}}(s,2)$	"Repair" $L(s,1)$	"Do not repair" $L(s,2)$
1	10	0	30	30
2	10.1	0	30	29.9
3	10.2	0	30	29.8
4	10.3	0	30	29.7
5	10.4	0	30	29.6
6	10.5	0	30	29.5
7	10.6	0	30	29.4
8	10.7	0	30	29.3
9	10.8	0	30	29.2
10	10.9	5	30	29.1

Table 4 Other parameters

Known mean unknown covariance	$\delta_0 = 0.9$
Unknown mean unknown covariance	$\delta_1 = \delta_2 = 1$
$\phi-$divergence	$\theta_\phi = 0.01$
Wasserstein distance	$\theta_W = 0.01,\ d = 1$

We compute the optimal policies of the DRCCMDP problem (5) for each uncertainty

Table 5 Optimal policies of *DRCCMDP* with different uncertainty sets

State(s) \ Optimal policies	Known known (p,1-p)	Known unknown (p,1-p)	Unknown unknown (p,1-p)	$\phi-$divergence (Kullback-Leibler) (p,1-p)	Wasserstein (p,1-p)
1	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
2	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
3	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
4	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
5	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
6	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
7	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
8	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
9	(0.77, 0.23)	(0.75, 0.25)	(0.71, 0.29)	(0.14, 0.86)	(0.05, 0.95)
10	(0.88, 0.12)	(0.89, 0.11)	(0.9, 0.1)	(0.91, 0.09)	(0.91, 0.09)
Optimal value	53.34	53.78	54.01	61.62	62.19

set by solving corresponding equivalent mixed-integer optimization problem using the existing solvers mentioned in Section 2.3. The optimal policies corresponding to all the uncertainty sets are summarized in Table 5, where p is the probability of "repair" action and $1-p$ is the probability of "do not repair" action. It is clear that the optimal repair policy corresponding to all the uncertainty sets for first eight states is same. At

state 9, the factory owner decides to repair for moments based uncertainty sets with high probability whereas for statistical distance based uncertainty sets repair action is taken with small probability. At last state repair action is taken with high probability for all types of uncertainty sets. The optimal value of the DRCCDMP problem is more for statistical distance based uncertainty sets as compared to moments based uncertainty sets. We present the time analysis by considering the number of states for all the uncertainty sets between 10 and 5000. All the parameters are taken similar to the case of 10 states. The results are presented in Figure 2 which shows that DRCCMDP problem with statistical distance based uncertainty sets can be solved efficiently up to 5000 states while it takes significantly longer computation time to solve the model with moments based uncertainty sets. For uncertainty set $\mathcal{D}_1$, the Gurobi solver perform efficiently up to 1000 states whereas for uncertainty sets $\mathcal{D}_2$ and $\mathcal{D}_3$ CUTSDP solver manages only up to 100 states.

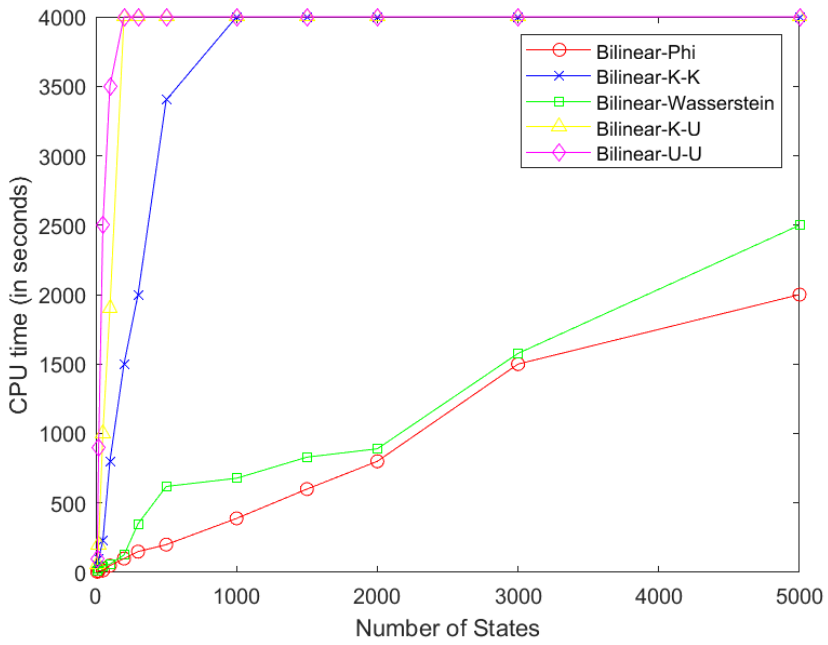

Fig. 2 CPU time (in seconds) vs number of states.

4 Conclusions

We study a DRCCMDP problem with random transition probabilities under various moments and statistical distance based uncertainty sets defined using ϕ-divergence and Wasserstein distance metric. We propose equivalent MIBP problems and MISDP problem with bilinear constraints for the DRCCMDP problem depending on the choice

of the uncertainty set. All these optimization problems can be solved efficiently using commercial solver GUROBI, except ones with both bilinear and positive semidefinite constraints. Using randomly generated data, the numerical experiments are performed on a machine replacement problem up to 5000 states which shows that a DRCCMDP problem with statistical distance based uncertainty sets can be solved very efficiently, while it takes more time in the case of moments based uncertainty sets.

Acknowledgement

This research was supported by the French government under the France 2030 program, reference ANR-11-IDEX-0003 within the OI H-Code.

A Proof of Lemma 2 - Case Hellinger distance

Given a policy $f \in PO_S$ and $y \in \mathbb{R}$, let $O = \{p \in E_p \mid V_\alpha(f,p) \geq y\}$. From Table 1, the conjugate of ϕ has the following form

$$\phi^*(r) = \begin{cases} \frac{r}{1-r}, & \text{if } r < 1, \\ \infty, & \text{if } r \geq 1. \end{cases} \tag{39}$$

Let

$$L = \sup_{\lambda>0,\beta\in\mathbb{R}} \left\{ \beta - \lambda\theta_\phi - \lambda\phi^*\left(\frac{-1+\beta}{\lambda}\right)\mathbb{P}_\nu(O) - \lambda\phi^*\left(\frac{\beta}{\lambda}\right)(1-\mathbb{P}_\nu(O)) \right\}. \tag{40}$$

The constraint (i) of (5) is equivalent to

$$L \geq 1 - \epsilon. \tag{41}$$

We consider two cases as follows:
Case 1: Let $\frac{\beta}{\lambda} < 1$. Since $\lambda > 0$, the following inequality holds

$$\frac{\beta-1}{\lambda} < \frac{\beta}{\lambda} < 1.$$

From (39), we have: $\phi^*\left(\frac{\beta}{\lambda}\right) = \frac{\beta}{\lambda-\beta}$, $\phi^*\left(\frac{\beta-1}{\lambda}\right) = \frac{\beta-1}{\lambda+1-\beta}$. Consequently, it follows from (40) that: $L = \sup_{\lambda>0,\beta<\lambda} \left\{ \mathbb{P}_\nu(O)\frac{\lambda^2}{(\lambda-\beta)(\lambda-\beta+1)} - \frac{\beta^2}{\lambda-\beta} - \lambda\theta_\phi \right\}$. Let $\eta = \lambda - \beta$. Then, we can write

$$L = \sup_{\lambda>0,\eta>0} \left\{ \lambda^2\left(\frac{\mathbb{P}_\nu(O)}{\eta(\eta+1)} - \frac{1}{\eta}\right) + \lambda(2-\theta_\phi) - \eta \right\}.$$

Let $g(\lambda,\eta) = \lambda^2\left(\frac{\mathbb{P}_\nu(O)}{\eta(\eta+1)} - \frac{1}{\eta}\right) + \lambda(2-\theta_\phi) - \eta$. It is a second-order polynomial of λ and the coefficient of λ^2 is negative because $0 \leq \mathbb{P}_\nu(O) \leq 1$ and $\eta > 0$. It is well

known that the maximum value of a second order polynomial $f(x) = ax^2 + bx + c$ with $a < 0$ is $c - \frac{b^2}{4a}$ and it holds at $x = \frac{-b}{2a}$. Hence, the maximum value of $g(\lambda, \eta)$ holds at $\lambda^* = \frac{\eta(\eta+1)(2-\theta_\phi)}{2(1+\eta-\mathbb{P}_\nu(O))}$. Since $\theta_\phi < 2$, $\lambda^* > 0$. Therefore, for a given $\eta > 0$, the optimal value L holds at λ^* and $L = c - \frac{b^2}{4a}$, where $c = -\eta$, $b = 2 - \theta_\phi$, $a = \frac{\mathbb{P}_\nu(O)}{\eta(\eta+1)} - \frac{1}{\eta}$, which implies that

$$L = \sup_{\eta > 0} \left\{ -\eta + \frac{(2-\theta_\phi)^2 \eta(\eta+1)}{4(\eta + 1 - \mathbb{P}_\nu(O))} \right\}. \tag{42}$$

Let $u = \eta + 1 - \mathbb{P}_\nu(O)$, then $\eta > 0$ is equivalent to $u > 1 - \mathbb{P}_\nu(O)$ and we can write

$$\begin{aligned} L = \sup_{u > 1-\mathbb{P}_\nu(O)} \Bigg\{ & \left(\frac{(2-\theta_\phi)^2}{4} - 1 \right) u + \frac{(2-\theta_\phi)^2 \mathbb{P}_\nu(O)(\mathbb{P}_\nu(O) - 1)}{4} \frac{1}{u} \\ & + 1 - \mathbb{P}_\nu(O) + \frac{(2-\theta_\phi)^2 (2\mathbb{P}_\nu(O) - 1)}{4} \Bigg\} = \sup_{u > 1-\mathbb{P}_\nu(O)} G(u), \end{aligned}$$

where $G(u) = a_1 u + \frac{b_1}{u} + c_1$ such that

$$\begin{aligned} a_1 &= \frac{(2-\theta_\phi)^2}{4} - 1, \; b_1 = \frac{(2-\theta_\phi)^2 \mathbb{P}_\nu(O)(\mathbb{P}_\nu(O) - 1)}{4}, \\ c_1 &= 1 - \mathbb{P}_\nu(O) + \frac{(2-\theta_\phi)^2 (2\mathbb{P}_\nu(O) - 1)}{4}. \end{aligned}$$

Since $0 < \theta_\phi < 2$ and $0 \leq \mathbb{P}_\nu(O) \leq 1$, $a_1 < 0$ and $b_1 \leq 0$. It is clear that G is decreasing on (u^*, ∞), increasing on $(-u^*, u^*)$ and decreasing on $(-\infty, -u^*)$, where

$$\begin{aligned} u^* &= \sqrt{\frac{b_1}{a_1}} = \sqrt{\frac{(2-\theta_\phi)^2}{4 - (2-\theta_\phi)^2} \mathbb{P}_\nu(O)(1 - \mathbb{P}_\nu(O))}, \\ G(u^*) &= a_1 u^* + \frac{b_1}{u^*} + c_1 = -2\sqrt{a_1 b_1} + c_1. \end{aligned} \tag{43}$$

If $u^* \leq 1 - \mathbb{P}_\nu(O)$, we deduce that $(1 - \mathbb{P}_\nu(O), \infty) \subset (u^*, \infty)$. Since G is decreasing on (u^*, ∞), it implies that G is decreasing on $(1 - \mathbb{P}_\nu(O), \infty)$. Hence, the optimal value of G is attained when $u = 1 - \mathbb{P}_\nu(O)$, i.e, $\eta = 0$. From (42), $L = 0$ which violates the constraint (41). Therefore, $u^* > 1 - \mathbb{P}_\nu(O) > 0$. Since, G is decreasing on (u^*, ∞) and increasing on $(1 - \mathbb{P}_\nu(O), u^*)$, then $u = u^*$ is the optimal solution of $G(u)$ and $L = -2\sqrt{a_1 b_1} + c_1$. Therefore,

$$\begin{aligned} L = & -2\sqrt{\frac{(2-\theta_\phi)^2}{4} \left(1 - \frac{(2-\theta_\phi)^2}{4} \right) \mathbb{P}_\nu(O)(1 - \mathbb{P}_\nu(O))} \\ & + 1 - \mathbb{P}_\nu(O) + \frac{(2-\theta_\phi)^2 (2\mathbb{P}_\nu(O) - 1)}{4}. \end{aligned}$$

Then, (41) is rewritten equivalently as follows

$$-2\sqrt{\frac{(2-\theta_\phi)^2}{4}\left(1-\frac{(2-\theta_\phi)^2}{4}\right)\mathbb{P}_\nu(O)(1-\mathbb{P}_\nu(O))} \geq \left(1-\frac{(2-\theta_\phi)^2}{2}\right)\mathbb{P}_\nu(O)+\frac{(2-\theta_\phi)^2}{4}-\epsilon. \tag{44}$$

By taking the square on both side of (44), we get

$$(2-\theta_\phi)^2\left(1-\frac{(2-\theta_\phi)^2}{4}\right)\mathbb{P}_\nu(O)(1-\mathbb{P}_\nu(O)) \leq \left[\left(1-\frac{(2-\theta_\phi)^2}{2}\right)\mathbb{P}_\nu(O)+\frac{(2-\theta_\phi)^2}{4}-\epsilon\right]^2. \tag{45}$$

By rewriting (45), we get the following second-order inequality in $\mathbb{P}_\nu(O)$

$$\big(\mathbb{P}_\nu(O)\big)^2 + B\ \mathbb{P}_\nu(O) + C \geq 0,$$

which is equivalent to: $\big(\mathbb{P}_\nu(O)-x_{\max}\big)\big(\mathbb{P}_\nu(O)-x_{\min}\big)\geq 0$, where $x_{\max}=\frac{-B+\sqrt{\Delta}}{2}$, $x_{\min}=\frac{-B-\sqrt{\Delta}}{2}$ and B, C, Δ are given in Table 2. It is clear that (44) is equivalent to either $\mathbb{P}_\nu(O)\geq x_{\max}$ or $\mathbb{P}_\nu(O)\leq x_{\min}$. Moreover, $x_{\max}$ and $x_{\min}$ are solutions of the following two equalities

$$-2\sqrt{\frac{(2-\theta_\phi)^2}{4}\left(1-\frac{(2-\theta_\phi)^2}{4}\right)x(1-x)} = \left(1-\frac{(2-\theta_\phi)^2}{2}\right)x+\frac{(2-\theta_\phi)^2}{4}-\epsilon, \tag{46}$$

$$2\sqrt{\frac{(2-\theta_\phi)^2}{4}\left(1-\frac{(2-\theta_\phi)^2}{4}\right)x(1-x)} = \left(1-\frac{(2-\theta_\phi)^2}{2}\right)x+\frac{(2-\theta_\phi)^2}{4}-\epsilon. \tag{47}$$

Since $\theta_\phi < 2-\sqrt{2}$, we deduce that $1-\frac{(2-\theta_\phi)^2}{2}<0$. Therefore, we have

$$\left(1-\frac{(2-\theta_\phi)^2}{2}\right)x_{\min}+\frac{(2-\theta_\phi)^2}{4}-\epsilon > \left(1-\frac{(2-\theta_\phi)^2}{2}\right)x_{\max}+\frac{(2-\theta_\phi)^2}{4}-\epsilon,$$

which implies that $x_{\max}$ is a solution of (46) and $x_{\min}$ is a solution of (47). Hence, the condition $\mathbb{P}_\nu(O)\leq x_{\min}$ implies that

$$\left(1-\frac{(2-\theta_\phi)^2}{2}\right)\mathbb{P}_\nu(O)+\frac{(2-\theta_\phi)^2}{4}-\epsilon \geq \left(1-\frac{(2-\theta_\phi)^2}{2}\right)x_{\min}+\frac{(2-\theta_\phi)^2}{4}-\epsilon > 0,$$

which violates the constraint (44). Then, (44) is equivalent to $\mathbb{P}_\nu(O)\geq x_{\max}$, i.e., the constraint (i) of (5) is equivalent to: $\mathbb{P}_\nu(\rho^T\hat{R}\geq y)\geq\frac{-B+\sqrt{\Delta}}{2}$.

Case 2: Let $1 \leq \frac{\beta}{\lambda}$. From (39), $\phi^*\left(\frac{\beta}{\lambda}\right) = \infty$, which in turn implies that $L = -\infty$ and it violates the constraint (41).

References

Altman E (1999) Constrained Markov Decision Processes. Routledge, New York

Ben-Tal A, Den Hertog D, De Waegenaere A, et al (2013) Robust solutions of optimization problems affected by uncertain probabilities. Management Science 59(2):341–357

Calafiore GC, El Ghaoui L (2006) On distributionally robust chance-constrained linear programs. Journal of Optimization Theory and Applications 130(1):1–22

Cheng J, Delage E, Lisser A (2014) Distributionally robust stochastic knapsack problem. SIAM Journal on Optimization 24(3):1485–1506

Delage E, Mannor S (2010) Percentile optimization for Markov decision processes with parameter uncertainty. Operations Research 58(1):203–213

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research 58(3):595–612

Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming 171(1-2):115–166

Gao R, Kleywegt A (2023) Distributionally robust stochastic optimization with Wasserstein distance. Mathematics of Operations Research 48(2):603–655

Goyal V, Grand-Clement J (2023) Robust Markov decision processes: Beyond rectangularity. Mathematics of Operations Research 48(1):203–226

Ho CP, Petrik M, Wiesemann W (2022) Robust ϕ-divergence MDPs. In: Koyejo S, Mohamed S, Agarwal A, et al (eds) Advances in Neural Information Processing Systems, vol 35. Curran Associates, Inc., pp 32680–32693

Iyengar GN (2005) Robust dynamic programming. Mathematics of Operations Research 30(2):257–280

Jiang R, Guan Y (2016) Data-driven chance constrained stochastic program. Mathematical Programming 158(1-2):291–327

Liu J, Lisser A, Chen Z (2022) Distributionally robust chance constrained geometric optimization. Mathematics of Operations Research 47(4):2950–2988

Mannor S, Simester D, Sun P, et al (2007) Bias and variance approximation in value function estimates. Management Science 53(2):308–322

McCormick GP (1976) Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems. Mathematical programming 10(1):147–175

Nilim A, El Ghaoui L (2005) Robust control of Markov decision processes with uncertain transition matrices. Operations Research 53(5):780–798

Popescu I (2007) Robust mean-covariance solutions for stochastic optimization. Operations Research 55(1):98–112

Puterman M, et al (1994) Markov Decision Processes. John Wiley & Sons, Inc., New York

Singh VV, Jouini O, Lisser A (2017) Distributionally robust chance-constrained games: Existence and characterization of Nash equilibrium. Optimization Letters 11(7):1385–1405

Varagapriya V, Singh VV, Lisser A (2022) Constrained Markov decision processes with uncertain costs. Operations Research Letters 50(2):218–223

Varagapriya V, Singh VV, Lisser A (2023) Joint chance-constrained Markov decision processes. Annals of Operations Research 322(2):1013–1035

Villani C (2021) Topics in optimal transportation, vol 58. American Mathematical Soc.

Villani C, et al (2009) Optimal transport: old and new, vol 338. Springer

White III CC, Eldeib HK (1994) Markov decision processes with imprecise transition probabilities. Operations Research 42(4):739–749

Wiesemann W, Kuhn D, Rustem B (2012) Robust Markov decision processes. Mathematics of Operations Research 38(1):153–183

Xu H, Mannor S (2012) Distributionally robust Markov decision processes. Mathematics of Operations Research 37(2):288–300

Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. Operations Research Letters 46(2):262–267

Zhi Chen PY, Haskell WB (2019) Distributionally robust optimization for sequential decision-making. Optimization 68(12):2397–2426