Distributionally Robust Optimal Control and MDP Modeling

Alexander Shapiro

School of Industrial and Systems Engineering,
Georgia Institute of Technology,
Atlanta, Georgia 30332-0205,
e-mail: ashapiro@isye.gatech.edu

Abstract. In this paper we discuss Optimal Control and Markov Decision Process (MDP) formulations of multistage optimization problems when the involved probability distributions are not known exactly, but rather are assumed to belong to specified ambiguity families. The aim of this paper is to clarify a connection between such distributionally robust approaches to multistage stochastic optimization.

Key Words: Optimal Control, Markov Decision Process, dynamic programming, Bellman equations, distributional robustness, stochastic games, risk measures, rectangularity, duality

1 Introduction

In this paper we consider Stochastic Optimal Control (SOC) and Markov Decision Process (MDP) formulations of multistage optimization problems when the involved probability distributions are not known exactly, but rather are assumed to belong to specified ambiguity families. In the MDP framework such *distributionally robust* approaches where discussed in a number of recent publications. The aim of this paper is to make clarification of several involved issues; in particular, we discuss a relation, and an essential difference, between the distributionally robust formulations of the SOC and MDP problems. We also discuss their relation to the modern theory of risk measures, a connection with the approach of stochastic games, and the role of the rectangularity assumption. We try to keep the presentation as simple as possible and concentrate on the conceptual issues, while avoiding a discussion of technical details which can be quite delicate and nontrivial.

It appears that first papers on the distributionally robust control of MDPs are [4] and [7]. In [4] the problem is approached in terms of dynamic programming equations and under the assumption of rectangularity. The concept of rectangularity is going back to [2], and was used by various authors in different ways. It is not completely clear from [4] what are the implications of the rectangularity assumption. It seems that similar assumptions were used in [7] in an implicit way. The distributionally robust control of MDPs was discussed further, for example, in [6],[14],[15]. Quite a different approach was suggested in [9], where the construction was made in terms of (coherent) risk measures; we will discuss this in section 3 (see Remark 3.1 in particular). In the MDP framework, stochastic games were introduced in [13], and were discussed extensively (e.g., [5] and references therein). We argue that this is directly related to the distributionally robust analysis of MDPs.

We use the following notation and terminology throughout the paper. For a (random data) process $\xi_1, ...,$ we denote by $\xi_{[t]} = (\xi_1, ..., \xi_t)$ history of the process up to time t. We say that the process ξ_t is *stagewise independent* if random vector ξ_{t+1} is independent of $\xi_{[t]}$ for all $t \geq 1$. When the probability distribution of the random data is discrete with a finite number of possible realizations, it is intuitive to think about the process as a finite scenario tree. A node of such scenario tree at stage t represents the history $\xi_{[t]}$ of the data process.

2 Optimal control model

In this section we consider the classical SOC (discrete time, finite horizon) model (e.g., [1])

$$\min_{u_t \in \mathcal{U}_t(x_t)} \quad \mathbb{E}^Q \left[\sum_{t=1}^{T} c_t(x_t, u_t, \xi_t) + c_{T+1}(x_{T+1}) \right], \tag{2.1}$$

$$\text{s.t.} \quad x_{t+1} = F_t(x_t, u_t, \xi_t), \ t = 1, ..., T. \tag{2.2}$$

Here variables $x_t \in \mathbb{R}^{n_t}$, $t = 1, ..., T+1$, represent state of the system, $u_t \in \mathbb{R}^{m_t}$, $t = 1, ..., T$, are controls, $\xi_t \in \mathbb{R}^{d_t}$, $t = 1, ..., T$, are random vectors, $c_t : \mathbb{R}^{n_t} \times \mathbb{R}^{m_t} \times \mathbb{R}^{d_t} \to \mathbb{R}$, $t = 1, ..., T$, are cost functions, $c_{T+1}(x_{T+1})$ is final cost function, $F_t : \mathbb{R}^{n_t} \times \mathbb{R}^{m_t} \times \mathbb{R}^{d_t} \to \mathbb{R}^{n_{t+1}}$ are (measurable) mappings, and $\mathcal{U}_t : \mathbb{R}^{n_t} \rightrightarrows \mathbb{R}^{m_t}$ are (measurable) multifunctions (point-to-set mappings). Values x_1 and ξ_0 are deterministic (initial conditions); it is also possible to view x_1 as random with a

given distribution, this is not essential for the following discussion. It is well known that the SOC model can be formulated in the MDP terms, we will discuss this later. Unless stated otherwise, we make the following assumption in this section: the distribution of $(\xi_1, ..., \xi_T)$ *does not* depend on the involved decisions. The notation $\mathbb{E}^Q$ emphasizes that the expectation in (2.1) is taken with respect to the probability distribution Q of $(\xi_1, ..., \xi_T)$.

The optimization in (2.1) is performed over policies

$$u_t = \pi_t(x_t, \xi_{[t-1]}),\ t = 1, ..., T, \tag{2.3}$$

satisfying the feasibility constraints $u_t \in \mathcal{U}_t(x_t)$ and state equations (2.2) almost surely (a.s.). For a considered policy, state and control variables are functions of the data process and hence are random, in order to emphasize this we sometimes use the bold face $\boldsymbol{x}_t$ for state variables. It is possible to consider (2.1) - (2.2) as a multistage stochastic programming problem with decision variables $y_t = (x_t, u_t)$ and the data process $\xi_1, ..., \xi_T$.

Example 2.1 (Inventory model) Consider the classical inventory model (cf., [16]):

$$\begin{array}{ll} \min\limits_{u_t \geq 0} & \mathbb{E}^Q \left[\sum\limits_{t=1}^{T} c_t u_t + \phi_t(x_t + u_t, D_t) \right] \\ \text{s.t.} & x_{t+1} = x_t + u_t - D_t,\ t = 1, ..., T, \end{array} \tag{2.4}$$

where $\phi_t(y_t, d_t) := b_t[d_t - y_t]_+ + h_t[y_t - d_t]_+$, and c_t, b_t, h_t are the ordering cost, backorder penalty cost and holding cost per unit, respectively, x_t is the current inventory level, u_t is the order quantity, and D_t is the demand at time t (it is assumed that $b_t > c_t \geq 0$ and $h_t \geq 0$). Here x_t are state variables, u_t are control variables, D_t are random variables with Q being the probability distribution of $(D_1, ..., D_T)$, $x_{t+1} = x_t + u_t - D_t$ are state (balance) equations and $\mathcal{U}_t := \mathbb{R}_+$.

The well known dynamic programming equations for problem (2.1) - (2.2) can be written as follows. At the last stage the value function $V_{T+1}(x_{T+1}) = c_{T+1}(x_{T+1})$ and, going backward in time, the value functions

$$V_t(x_t, \xi_{[t-1]}) = \inf_{u_t \in \mathcal{U}_t(x_t)} \mathbb{E}^{Q|\xi_{[t-1]}} \left[c_t(x_t, u_t, \xi_t) + V_{t+1}\big(F_t(x_t, u_t, \xi_t), \xi_{[t]}\big) \right], \tag{2.5}$$

$t = 1, ..., T$. The minimizers $\bar{u}_t = \pi_t(x_t, \xi_{[t-1]})$, $t = 1, ..., T$, of the right hand side of (2.5), provided that such minimizers exist, define an optimal policy for problem (2.1). The notation $\mathbb{E}^{Q|\xi_{[t-1]}}$ stands for the conditional expectation with respect to the probability distribution Q conditional on $\xi_{[t-1]}$.

In optimal control the random process ξ_t, $t = 1, ..., T$, is often viewed as noise and assumed to be stagewise independent. In that case, the value functions $V_t(x_t)$ depend only on the state variables x_t and the conditional expectation in (2.5) becomes the respective unconditional expectation, that is

$$V_t(x_t) = \inf_{u_t \in \mathcal{U}_t(x_t)} \mathbb{E} \left[c_t(x_t, u_t, \xi_t) + V_{t+1}\big(F_t(x_t, u_t, \xi_t)\big) \right]. \tag{2.6}$$

Also the minimizers $\bar{u}_t = \pi_t(x_t)$, $t = 1, ..., T$, of the right hand side of (2.6) define an optimal policy for problem (2.1). That is, in the stagewise independent case optimization in (2.1) can be performed over policies of the form $u_t = \pi_t(x_t)$, $t = 1, ..., T$.

In the distributionally robust approach the distribution of the involved random process is assumed to belong to a specified family $\mathfrak{M}$ of probability distributions of $(\xi_1, ..., \xi_T)$, rather than to be known exactly. This can be related to a risk-averse setting by duality arguments (cf., [3],[12]). In the distributionally robust formulation the counterpart of equations (2.5) are equations

$$V_t(x_t, \xi_{[t-1]}) = \inf_{u_t \in \mathcal{U}_t(x_t)} \sup_{\mathcal{Q}_t \in \mathfrak{M}} \mathbb{E}^{\mathcal{Q}_t|\xi_{[t-1]}} \left[c_t(x_t, u_t, \xi_t) + V_{t+1}\big(F_t(x_t, u_t, \xi_t), \xi_{[t]}\big) \right]. \tag{2.7}$$

Note that the set $\mathfrak{M}$ in the right hand side of (2.7) can be replaced by the respective set $\mathfrak{M}_{[t]}$ of marginal distributions of $\xi_{[t]}$.

In the *rectangular* setting the set $\mathfrak{M}$ is assumed to be of the form

$$\mathfrak{M} := \{Q_1 \times \cdots \times Q_T : Q_t \in \mathcal{M}_t,\ t = 1, ..., T\}, \tag{2.8}$$

where $\mathcal{M}_t$ is a set of (marginal) probability distributions of ξ_t, $t = 1, ..., T$. The rectangular setting can be considered as a counterpart of the stagewise independent case. In the rectangular case the counterparts of equations (2.6) become

$$V_t(x_t) = \inf_{u_t \in \mathcal{U}_t(x_t)} \sup_{Q_t \in \mathcal{M}_t} \mathbb{E}^{Q_t} \left[c_t(x_t, u_t, \xi_t) + V_{t+1}\big(F_t(x_t, u_t, \xi_t)\big) \right]. \tag{2.9}$$

An optimal policy in the rectangular case is of the form $\bar{u}_t = \pi_t(x_t)$ with

$$\bar{u}_t \in \arg \min_{u_t \in \mathcal{U}_t(x_t)} \sup_{Q_t \in \mathcal{M}_t} \mathbb{E}^{Q_t} \left[c_t(x_t, u_t, \xi_t) + V_{t+1}\big(F_t(x_t, u_t, \xi_t)\big) \right]. \tag{2.10}$$

In the risk-averse approach it is assumed that at every node of the scenario tree at stage $t-1$, represented by the sample path $\xi_{[t-1]}$, is defined a set of probability distributions $\mathcal{M}_t^{\xi_{[t-1]}}$ on the set of child nodes of $\xi_{[t-1]}$ at stage t. The corresponding value functions of the dynamic programming equations are written then as

$$V_t(x_t, \xi_{[t-1]}) = \inf_{u_t \in \mathcal{U}_t(x_t)} \sup_{Q_t \in \mathcal{M}_t^{\xi_{[t-1]}}} \mathbb{E}^{Q_t|\xi_{[t-1]}} \left[c_t(x_t, u_t, \xi_t) + V_{t+1}\big(F_t(x_t, u_t, \xi_t), \xi_{[t]}\big) \right]. \tag{2.11}$$

In that setting,

$$\mathcal{R}_{t|\xi_{[t-1]}}(\cdot) := \sup_{Q_t \in \mathcal{M}_t^{\xi_{[t-1]}}} \mathbb{E}^{Q_t|\xi_{[t-1]}}[\,\cdot\,],\ t = 1, ..., T, \tag{2.12}$$

is viewed as the (dual representation) of the corresponding (coherent) risk measure conditional on $\xi_{[t-1]}$. Recall that ξ_0 is deterministic, therefore

$$\mathcal{R}_{1|\xi_{[0]}}(\cdot) = \sup_{Q_1 \in \mathcal{M}_1} \mathbb{E}^{Q_1}[\,\cdot\,].$$

We refer to [8] for a detailed discussion of these approaches to distributionally robust and risk-averse multistage optimization. In the rectangular case the sets $\mathcal{M}_t = \mathcal{M}_t^{\xi_{[t-1]}}$ do not depend on the corresponding nodes $\xi_{[t-1]}$. Consequently in the rectangular case the distributionally robust and risk-averse approaches do coincide and the corresponding dynamic programming equations

in both cases can be written in the form (2.9). In both cases we can assume that the sets of considered probability measures are convex (and closed in the weak* topology)

Consider functionals $\mathcal{R}_{t|\xi_{[t-1]}}(\cdot)$ defined in (2.12). With these functionals we can associate the corresponding nested (composite) functional

$$\mathfrak{R}(\cdot) := \mathcal{R}_{1|\xi_{[0]}}\Big(\mathcal{R}_{2|\xi_{[1]}}\big(\cdots \mathcal{R}_{T|\xi_{[T-1]}}(\cdot)\big)\Big). \tag{2.13}$$

(we refer to [8] for a detailed discussion of such construction). Note that the nested functional $\mathfrak{R}$ is defined in terms of the data process ξ_t, and does not depend on the policies (2.3). The assumption that the ambiguity sets of probability distributions do not depend on the policies is essential here. In particular if for a distribution Q of $(\xi_1, ..., \xi_T)$ and every t, the conditional ambiguity set $\mathcal{M}_t^{\xi_{[t-1]}}$ is the singleton consisting of the conditional distribution of ξ_t given $\xi_{[t-1]}$, then $\mathcal{R}_{t|\xi_{[t-1]}} = \mathbb{E}^{Q|\xi_{[t-1]}}$, and hence $\mathfrak{R} = \mathbb{E}^Q$.

Equations (2.11) represent the dynamic programming equations for the following optimization problem

$$\min_{\pi\in\Pi} \mathfrak{R}\left(Z^\pi\right), \tag{2.14}$$

with

$$Z^\pi := \sum_{t=1}^{T} c_t\big(\boldsymbol{x}_t, \pi_t(\boldsymbol{x}_t, \xi_{[t-1]}), \xi_t\big) + c_{T+1}(\boldsymbol{x}_{T+1}) \tag{2.15}$$

being the total cost depending on policy π.

The nested (composite) functional $\mathfrak{R}$ satisfies the axioms of coherent risk measures. Consequently by duality arguments (cf., [10]), it has the following dual representation.

Proposition 2.1 *Suppose that the nested functional $\mathfrak{R}(\cdot)$ is finite valued. Then there exists a set $\widehat{\mathfrak{M}}$ of probability distributions Q of $(\xi_1, ..., \xi_T)$ such that*

$$\mathfrak{R}(\cdot) = \sup_{Q\in\widehat{\mathfrak{M}}} \mathbb{E}^Q[\,\cdot\,]. \tag{2.16}$$

In the above, the set $\widehat{\mathfrak{M}}$ is derived in an abstract way by using the dual representation of the coherent risk measure $\mathfrak{R}$. A constructive description of $\widehat{\mathfrak{M}}$ is quite involved even in the rectangular case (cf., [11]).

It follows that formulation (2.14) can be written in the following minimax form (cf., [8])

$$\min_{\pi\in\Pi} \sup_{Q\in\widehat{\mathfrak{M}}} \mathbb{E}^Q\left[Z^\pi\right]. \tag{2.17}$$

In the formulation (2.17) the expectation $\mathbb{E}^Q$ is taken with respect to distribution $Q \in \widehat{\mathfrak{M}}$, which does not depend on policy $\pi \in \Pi$. This is coming at the cost that the set $\widehat{\mathfrak{M}}$ is different (larger) than the set $\mathfrak{M}$. The set $\widehat{\mathfrak{M}}$ does not have the rectangular structure even if the set $\mathfrak{M}$ is rectangular. Nevertheless, as it was already pointed out, in the rectangular case it suffices to consider policies $\pi \in \Pi$ of the form $u_t = \pi_t(x_t)$, $t = 1, ..., T$, and hence problem (2.17) takes the form

$$\min_{\pi\in\Pi} \sup_{Q\in\widehat{\mathfrak{M}}} \mathbb{E}^Q\left[\sum_{t=1}^{T} c_t\big(\boldsymbol{x}_t, \pi_t(\boldsymbol{x}_t), \xi_t\big) + c_{T+1}(\boldsymbol{x}_{T+1})\right]. \tag{2.18}$$

The dual of problem (2.17) (and hence of problem (2.14)) is obtained by interchanging the minimization and maximization operators. That is, the dual can be written as

$$\max_{Q\in\widehat{\mathfrak{M}}} \inf_{\pi\in\Pi} \mathbb{E}^Q\left[Z^\pi\right]. \tag{2.19}$$

In the rectangular case the policies $\pi_t(x_t)$ in (2.14) - (2.15) do not depend on $\xi_{[t-1]}$. Yet, as it was pointed above, the set $\widehat{\mathfrak{M}}$ could contain probability measures which do not have the rectangular structure (i.e. are not stagewise independent). We have that the optimal value of the dual problem (2.19) cannot be larger than the optimal value of the primal problem (2.17) (problem (2.14)). In case the considered problem is convex, under mild regularity conditions there is no duality gap between the primal and dual problems. There are several ways how this can be rigorously formulated, let us consider the following setting.

Example 2.2 Let the mappings of the state equations (2.2) be of the form

$$F_t(x_t, u_t, \xi_t) := A_t x_t + B_t u_t + b_t,\ t = 1, ..., T, \tag{2.20}$$

where matrices A_t, B_t and vector b_t are functions of ξ_t, $t = 1, ..., T$. Suppose further that the cost functions $c_t(x_t, u_t, \xi_t)$ are convex in (x_t, u_t), and the sets $\mathcal{U}_t(x_t) \equiv \mathcal{U}_t$ do not depend on x_t and are convex for all t. Then the value functions $V_t(x_t, \xi_{[t-1]})$ are convex in x_t.

Indeed, by induction going backward in time suppose that $V_{t+1}(x_{t+1}, \xi_{[t]})$ is convex in x_{t+1}. Then the function

$$\phi_t(x_t, u_t, \xi_{[t]}) := c_t(x_t, u_t, \xi_t) + V_{t+1}\big(A_t x_t + B_t u_t + b_t, \xi_{[t]}\big),$$

is convex in (x_t, u_t). Consequently the corresponding expected value, and hence its maximum, function in the right hand side of (2.11) is convex in (x_t, u_t); and thus its minimum over $u_t \in \mathcal{U}$ is convex.

It follows then that under mild regularity conditions there is no duality gap between the above primal and dual problems (cf., [8]). The inventory model, of Example 2.1, is of that type.

Remark 2.1 The basic assumption that we used so far is that the considered distributions of the data process ξ_t do not depend on our decisions (controls). Consider the stagewise independent risk neutral case (2.1) - (2.2), and suppose now that the distribution of ξ_t can be a function of control u_t, $t = 1, ..., T$. For instance, in the inventory model (of Example 2.1) we can think that distribution of the demand D_t is effected by our decision about order quantity u_t.

We can deal with this in the following way. The random vector ξ_t can be generated by generating a random vector ω_t, whose components are distributed independently of each other and uniformly on the interval [0,1], and then making an appropriate transformation of ω_t. This is how random distributions are constructed in the Monte Carlo method starting with a random number generator. That is, we can represent ξ_t in the form $\xi_t = g_t(u_t, \omega_t)$. Substituting this into (2.1) - (2.2) we obtain the corresponding counterpart, with cost functions $c_t^*(x_t, u_t, \omega_t) = c_t(x_t, u_t, g_t(u_t, \omega_t))$ and mappings $F_t^*(x_t, u_t, \omega_t) = F_t(x_t, u_t, g_t(u_t, \omega_t))$, in terms of the random process ω_t which does not depend on controls u_t. Note, however, that the mappings g_t, and hence the transformed costs c_t^* and state mappings F_t^*, also depend on the distribution

of ξ_t. Therefore in the (rectangular) distributionally robust setting, where ξ_t can have different distributions (from the set $\mathcal{M}_t$), such approach will result in different cost functions and state mappings corresponding to the different distributions of ξ_t. This, of course, is not consistent with the distributionally robust formulation of the SOC model. This observation will be especially important when we will discuss a difference between the distributionally robust approaches to SOC and MDP modeling.

Infinite horizon case Consider stationary infinite horizon case

$$\begin{array}{ll} \min\limits_{u_t \in \mathcal{U}(x_t)} & \mathbb{E}\left[\sum\limits_{t=1}^{\infty} \gamma^{t-1} c(x_t, u_t, \xi_t)\right], \\ \text{s.t.} & x_{t+1} = F(x_t, u_t, \xi_t), \end{array} \tag{2.21}$$

where $\gamma \in (0,1)$ is the discount factor, and ξ_t, $t = 1, ...,$ is an independent identically distributed (iid) sequence of random vectors. The corresponding Bellman equation is

$$V(x) = \inf_{u \in \mathcal{U}(x)} \mathbb{E}\left[c(x, u, \xi) + \gamma V\big(F(x, u, \xi)\big)\right], \tag{2.22}$$

where the expectation is taken with respect to the distribution of random vector ξ.

The distributionally robust counterpart of (2.22) is

$$V(x) = \inf_{u \in \mathcal{U}(x)} \sup_{\mathcal{Q} \in \mathcal{M}} \mathbb{E}^{\mathcal{Q}}\left[c(x, u, \xi) + \gamma V\big(F(x, u, \xi)\big)\right], \tag{2.23}$$

where $\mathcal{M}$ is a set of probability distributions of ξ. The corresponding distributionally robust infinite horizon problem can be formulated as the limit of the respective nested finite horizon problems with horizon $T \to \infty$. The finite horizon problems have the rectangular structure associated with the set $\mathcal{M}$ of marginal distributions. This leads to the following distributionally robust problem associated with equation (2.23):

$$\begin{array}{ll} \min\limits_{u_t \in \mathcal{U}(x_t)} & \mathfrak{R}\big(\sum_{t=1}^{\infty} \gamma^{t-1} c(x_t, u_t, \xi_t)\big), \\ \text{s.t.} & x_{t+1} = F(x_t, u_t, \xi_t), \end{array} \tag{2.24}$$

where $\mathfrak{R}$ is the limit functional of the corresponding nested (composite) distributionally robust (risk-averse) functionals.

The limit functional can be represented in the form (2.16) for a set $\widehat{\mathfrak{M}}$ of probability measures. Thus problem (2.24) can be written as

$$\begin{array}{ll} \min\limits_{u_t \in \mathcal{U}(x_t)} & \sup\limits_{Q \in \widehat{\mathfrak{M}}} \mathbb{E}^{Q}\big[\sum_{t=1}^{\infty} \gamma^{t-1} c(x_t, u_t, \xi_t)\big], \\ \text{s.t.} & x_{t+1} = F(x_t, u_t, \xi_t). \end{array} \tag{2.25}$$

Similar to (2.19) it is possible to consider the dual of problem (2.25) by interchanging the 'max' and 'min' operators. However, as it was already discussed, the set $\widehat{\mathfrak{M}}$ does not have the rectangular structure. Therefore the dual counterpart of the Bellman equation (2.23), obtained by interchanging the 'min' and 'max' operators in the right hand side of (2.23), does not represent the dual of problem (2.25) (of problem (2.24)).

Stochastic games The dynamic equations can be viewed in terms of stochastic games. In the MDP framework, stochastic games were introduced in Shapley [13] and were studied extensively (see, e.g., the survey [5]). Consider the set Π of policies (2.3) satisfying the feasibility constraints. For a policy $\pi \in \Pi$, we can think about an opponent who chooses at every stage a distribution $Q_t^\pi \in \mathcal{M}_t^{\xi_{[t-1]}}$ for given state x_t and realization $\xi_{[t-1]}$ of the data process. A choice of such probability distributions defines the corresponding policy

$$Q_t^\pi = \gamma_t\big(x_t, \pi_t(x_t, \xi_{[t-1]}), \xi_{[t-1]}\big) \in \mathcal{M}_t^{\xi_{[t-1]}},\ t = 1, ..., T, \tag{2.26}$$

of the opponent. Note again that it is assumed that the ambiguity sets $\mathcal{M}_t^{\xi_{[t-1]}}$ do not depend on our decisions (on policies $\pi \in \Pi$).

For $\pi \in \Pi$ denote by Γ^π the set of feasible policies, of the form (2.26), of the opponent. For $\gamma \in \Gamma^\pi$ consider the corresponding expected value

$$\mathbb{E}^{\pi,\gamma}\left[\,\cdot\,\right] := \mathbb{E}^{Q_1^\pi}\left[\mathbb{E}^{Q_2^\pi|\xi_{[1]}}\left[\cdots \mathbb{E}^{Q_T^\pi|\xi_{[T-1]}}\left[\,\cdot\,\right]\right]\right], \tag{2.27}$$

with Q_t^π of the form (2.26) and x_t satisfying the state equations (2.2). The right hand side of (2.27) is understood in the nested (composite) way associated with the data process and policy $\pi = \{\pi_t(x_t, \xi_{[t-1]})\}_{t=1,...,T} \in \Pi$. Then for the cost function Z^π, given in (2.15), we can write

$$\sup_{\gamma\in\Gamma^\pi} \mathbb{E}^{\pi,\gamma}\left[Z^\pi\right] = \underbrace{\sup_{Q_1\in\mathcal{M}_1} \mathbb{E}^{Q_1}}_{\mathcal{R}_{1|\xi_{[0]}}}\Big[\underbrace{\sup_{Q_2\in\mathcal{M}_2^{\xi_{[1]}}} \mathbb{E}^{Q_2|\xi_{[1]}}}_{\mathcal{R}_{2|\xi_{[1]}}} \cdots \underbrace{\sup_{Q_T\in\mathcal{M}_T^{\xi_{[T-1]}}} \mathbb{E}^{Q_T|\xi_{[T-1]}}}_{\mathcal{R}_{T|\xi_{[T-1]}}}\left[Z^\pi\right]\Big]. \tag{2.28}$$

It follows that

$$\sup_{\gamma\in\Gamma^\pi} \mathbb{E}^{\pi,\gamma}\left[Z^\pi\right] = \mathfrak{R}(Z^\pi), \tag{2.29}$$

and hence equations (2.11) represent the dynamic programming equations for the problem

$$\min_{\pi\in\Pi} \sup_{\gamma\in\Gamma^\pi} \mathbb{E}^{\pi,\gamma}\left[Z^\pi\right], \tag{2.30}$$

and thus problems (2.30) and (2.14) have the same optimal value.

In particular consider the rectangular case with the set $\mathfrak{M}$ of the form (2.8). Then by equations (2.9) - (2.10) it suffices to consider policies of the form $\pi = \{\pi_t(x_t\}_{t=1,...,T}$ and the opponent policies $\gamma_t(x_t, \pi_t(x_t)) \in \mathcal{M}_t$, $t = 1, ..., T$. The corresponding problem (2.30) then takes the form

$$\min_{\pi\in\Pi} \sup_{\gamma\in\Gamma^\pi} \mathbb{E}^{\pi,\gamma}\left[\sum_{t=1}^{T} c_t\big(\boldsymbol{x}_t, \pi_t(\boldsymbol{x}_t), \xi_t\big) + c_{T+1}(\boldsymbol{x}_{T+1})\right]. \tag{2.31}$$

Note that even in the rectangular case the set of policies Γ^π of the opponent depends on the policy $\pi \in \Pi$ and we cannot construct the dual of problem (2.31) simply by interchanging the order of the 'min' and 'max' operators. Nevertheless we can proceed to the dual problem of the form (2.19).

We emphasize again that it is possible to write the left hand side of (2.28) in the form of the respective right hand side, since it is assumed that the ambiguity sets do not depend on policies $\pi \in \Pi$. The equation (2.29) then follows with the respective functional (coherent risk measure) $\mathfrak{R}$ which does not depend on policies $\pi \in \Pi$. In turn by (2.16) this implies existence of the set $\widehat{\mathfrak{M}}$ of probability measures and allows construction of the dual problem (2.19).

3 MDP formulation

In this section we discuss the setting of MDP modeling. It is well known that in the risk neutral case the SOC model can be formulated in the MDP form, and in fact both models in a certain sense are equivalent. We argue that in the distributionally robust setting there is an essential difference between the SOC and MDP formulations. Suppose for the sake of simplicity that the state space $\mathcal{S}_t$, of possible values of x_t, and the action (control) set $\mathcal{U}_t(x_t)$ are finite, $t = 1, ..., T$. Consider history

$$\mathfrak{h}_t = (x_1, u_1, ..., x_t, u_t),\ t = 1, ..., T,$$

of the decision process. Transition probability, of moving from state $x_t \in \mathcal{S}_t$, at stage t, to state $x_{t+1} \in \mathcal{S}_{t+1}$, conditional on the history $\mathfrak{h}_{t-1}$ and action u_t, is $p_t^{\mathfrak{h}_{t-1}}(x_{t+1}|x_t, u_t)$. For the sake of simplicity we restrict our discussion below to *deterministic* policies.

In terms of the history of the decision process, the value functions of the dynamic programming equations can be written as

$$V_t(x_t, \mathfrak{h}_{t-1}) = \inf_{u_t \in \mathcal{U}_t(x_t)} \sum_{x_{t+1} \in \mathcal{S}_{t+1}} p_t^{\mathfrak{h}_{t-1}}(x_{t+1}|x_t, u_t) \Big[r_t(x_t, u_t, x_{t+1}) + V_{t+1}\big(x_{t+1}, \underbrace{\mathfrak{h}_{t-1}, x_t, u_t}_{\mathfrak{h}_t}\big) \Big], \quad (3.1)$$

where $r_t(x_t, u_t, x_{t+1})$ are cost functions. The corresponding policy is

$$u_t = \pi_t(\mathfrak{h}_{t-1}, x_t) \in \mathcal{U}_t(x_t),\ t = 1, ..., T. \quad (3.2)$$

In the Markovian case the respective transition probability does not depend on $\mathfrak{h}_{t-1}$. Consequently the value function $V_t(x_t)$ does not depend on $\mathfrak{h}_{t-1}$, with

$$V_t(x_t) = \inf_{u_t \in \mathcal{U}_t(x_t)} \sum_{x_{t+1} \in \mathcal{S}_{t+1}} p_t(x_{t+1}|x_t, u_t) \left[r_t(x_t, u_t, x_{t+1}) + V_{t+1}\big(x_{t+1}\big) \right]. \quad (3.3)$$

The optimization problem, corresponding to the dynamic programming equations (3.1), can be written as

$$\min_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{t=1}^{T} r_t(x_t, u_t, x_{t+1}) + r_{T+1}(x_{T+1}) \right], \quad (3.4)$$

where Π is the set of feasible policies of the form (3.2). In the Markovian case optimization in (3.4) is over policies $u_t = \pi_t(x_t)$ satisfying the feasibility constraints $u_t \in \mathcal{U}_t(x_t)$, and with transition probabilities $p_t(x_{t+1}|x_t, u_t)$ and initial condition given by x_1. In general when transition probabilities $p_t^{\mathfrak{h}_{t-1}}(x_{t+1}|x_t, u_t)$ also depend on history $\mathfrak{h}_{t-1}$, optimization in (3.4) is over feasible policies of the form (3.2).

As it was discussed in section 2 there are two somewhat natural ways to formulate the risk-averse and distributionally robust counterparts of the risk neutral model (cf., [8]). We adopt below the risk-averse approach. We can formulate the following counterpart in the MDP setting. With history $\mathfrak{h}_t = (\mathfrak{h}_{t-1}, x_t, u_t)$ we associate a set $\mathcal{P}_t^{\mathfrak{h}_{t-1}}(x_t, u_t)$ of transition probabilities. The respective dynamic programming equations are

$$V_t(x_t, \mathfrak{h}_{t-1}) = \inf_{u_t \in \mathcal{U}_t(x_t)} \sup_{p_t^{\mathfrak{h}_{t-1}}(\cdot\,|x_t,u_t) \in \mathcal{P}_t^{\mathfrak{h}_{t-1}}(x_t,u_t)} \sum_{x_{t+1} \in \mathcal{S}_{t+1}} p_t^{\mathfrak{h}_{t-1}}(x_{t+1}|x_t,u_t)\Big[r_t(x_t,u_t,x_{t+1}) + V_{t+1}\big(x_{t+1}, \underbrace{\mathfrak{h}_{t-1}, x_t, u_t}_{\mathfrak{h}_t}\big)\Big], \quad (3.5)$$

and the corresponding optimal policies of the form (3.2).

In the rectangular case the set $\mathcal{P}_t(x_t, u_t)$ does not depend on the history $\mathfrak{h}_{t-1}$ and the dynamic equations become

$$V_t(x_t) = \inf_{u_t \in \mathcal{U}_t(x_t)} \sup_{p_t(\cdot\,|x_t,u_t) \in \mathcal{P}_t(x_t,u_t)} \sum_{x_{t+1} \in \mathcal{S}_{t+1}} p_t(x_{t+1}|x_t,u_t)\Big[r_t(x_t,u_t,x_{t+1}) + V_{t+1}\big(x_{t+1}\big)\Big], \quad (3.6)$$

with optimal policy of the form $\bar{u}_t = \pi_t(x_t)$,

$$\bar{u}_t \in \arg\min_{u_t \in \mathcal{U}_t(x_t)} \sup_{p_t(\cdot\,|x_t,u_t) \in \mathcal{P}_t(x_t,u_t)} \sum_{x_{t+1} \in \mathcal{S}_{t+1}} p_t(x_{t+1}|x_t,u_t)\left[r_t(x_t,u_t,x_{t+1}) + V_{t+1}\big(x_{t+1}\big)\right]. \quad (3.7)$$

Stochastic games Similar to (2.30), the optimization problem corresponding to the dynamic equations (3.5) can be formulated in terms of stochastic games. For a policy $\pi \in \Pi$ the opponent chooses, at stage t, transition probability

$$p_t^{\mathfrak{h}_{t-1}}(\cdot\,|x_t, \pi_t(x_t, \mathfrak{h}_{t-1})) \in \mathcal{P}_t^{\mathfrak{h}_{t-1}}(x_t, \pi_t(x_t, \mathfrak{h}_{t-1})). \quad (3.8)$$

Such choice defines the corresponding policy (depending on $\pi \in \Pi$) for the opponent. In the rectangular case the policy $\pi_t(x_t)$ and $p_t(\cdot\,|x_t, \pi_t(x_t))$ in (3.8) do not depend on $\mathfrak{h}_{t-1}$. For $\pi \in \Pi$ denote by Υ^π the set of feasible policies of the opponent. Then the distributionally robust counterpart of problem (3.4) can be written in the following minimax form

$$\min_{\pi \in \Pi} \sup_{\upsilon \in \Upsilon^\pi} \mathbb{E}^{\pi,\upsilon}\left[\sum_{t=1}^{T} r_t(x_t,u_t,x_{t+1}) + r_{T+1}(x_{T+1})\right], \quad (3.9)$$

with u_t of the form (3.2) determined by policy π. Problem (3.9) can be considered as the MDP analogue of the respective optimal control problem (2.31).

Remark 3.1 In general the set Υ^π depends on $\pi \in \Pi$. For a policy $\pi \in \Pi$ and the respective actions $u_t = \pi_t(x_t, \mathfrak{h}_{t-1})$, let $\mathcal{H}_t^\pi$ be the set of the corresponding histories $\mathfrak{h}_t = (x_1, u_1, ..., x_t, u_t)$, $x_\tau \in \mathcal{S}_\tau$, $\tau = 1, ..., t$. Since the states and action spaces are assumed to be finite, the set $\mathcal{H}^\pi = \mathcal{H}_T^\pi$ can be viewed as the set of scenarios on the scenario tree. To each history $\mathfrak{h}_t \in \mathcal{H}_t^\pi$ corresponds the node, at stage t, of the scenario tree, and the set $\mathcal{P}_t^{\mathfrak{h}_{t-1}}(x_t, u_t)$, $u_t = \pi_t(x_t, \mathfrak{h}_{t-1})$, of transition

probabilities. For transition probability $p_t^{\mathfrak{h}_{t-1}}(\cdot\,|x_t,u_t) \in \mathcal{P}_t^{\mathfrak{h}_{t-1}}(x_t,u_t)$, we view $p_t^{\mathfrak{h}_{t-1}}(x_{t+1}|x_t,u_t)$ as the (conditional) probability of moving from node $\mathfrak{h}_t$ to its child node $(\mathfrak{h}_t, x_{t+1}, u_{t+1})$, $x_{t+1} \in \mathcal{S}_{t+1}$, $u_{t+1} = \pi_{t+1}(x_{t+1}, \mathfrak{h}_t)$, at stage $t+1$. In that way every choice of the transition (conditional) probabilities $p_t^{\mathfrak{h}_{t-1}}(\cdot\,|x_t,u_t)$ defines a probability distribution supported on the scenario tree $\mathcal{H}^\pi$. The considered set of these probability distributions can be described in terms of the corresponding nested risk measure similar to the derivations of section 2 (in an abstract form of coherent risk measures, this is basically the approach suggested in [9]).

Consequently in a way similar to the derivation of Proposition 2.1, we conclude that there exists a set $\widehat{\mathfrak{P}}^\pi$ of probability distributions supported on the scenario tree $\mathcal{H}^\pi$ such that problem (3.9) can be written as

$$\min_{\pi\in\Pi} \sup_{p\in\widehat{\mathfrak{P}}^\pi} \mathbb{E}^{\pi,p}\left[\sum_{t=1}^{T} r_t(x_t,u_t,x_{t+1}) + r_{T+1}(x_{T+1})\right], \tag{3.10}$$

with $u_t = \pi_t(x_t, \mathfrak{h}_{t-1})$. As it was discussed in section 2, the set $\widehat{\mathfrak{P}}^\pi$ could have a complicated structure even in the rectangular case.

Remark 3.2 Formulations (3.9) and (3.10) may look similar, however there is an important difference which might be hidden in the notation. The expectation $\mathbb{E}^{\pi,\upsilon}$ in (3.9) is understood to be taken with respect to policy $\pi \in \Pi$ and policy $\upsilon \in \Upsilon^\pi$ of the opponent. On the other hand, for a given $\pi \in \Pi$, the set $\widehat{\mathfrak{P}}^\pi$ in formulation (3.10) is the set of probability distributions defined on the scenario tree (depending on π) of the histories of the decision process.

Remark 3.3 Problem (3.10) can be viewed as the MDP analogue of the SOC formulation (2.17). However, there is an essential difference - the set $\widehat{\mathfrak{P}}^\pi$ of probability distributions in (3.10) depends on policy $\pi \in \Pi$. This is because in the distributionally robust MDP framework it does not seem to be possible to separate construction of the transition probabilities from our actions (decisions). In the framework of the SOC model, this is highlighted in the discussion of Remark 2.1. Formulation (3.10) can be also compared, for example, with the respective formulation in [6, p.106], where it is written in terms of dynamic equations associated with the considered histories of the decision process.

Relation to the control model It is well known in the risk neutral setting, that the SOC model can be formulated in the MDP framework, and in fact both models are equivalent. In the distributionally robust setting we can proceed as follows. Consider the ambiguity sets $\mathcal{M}_t^{\xi_{[t-1]}}$ of probability distributions, discussed in section 2 in the risk-averse setting. Consider histories $\mathfrak{h}_t$, $t = 1,...,T$, associated with a policy π and the corresponding controls (actions) u_t of the form (3.2). (Note that these histories depend on policy π through the equations (3.2); we omit the superscript π in $\mathfrak{h}_t$ for the sake of simplicity of the notation.) For $t = 1,...,T$ define transition probabilities

$$p_t^{\mathfrak{h}_{t-1}}(x_{t+1}|x_t,u_t) := Q_{t|\xi_{[t-1]}}\left\{x_{t+1} = F_t(x_t,u_t,\xi_t)\right\},\ Q_{t|\xi_{[t-1]}} \in \mathcal{M}_t^{\xi_{[t-1]}}. \tag{3.11}$$

This defines the respective set $\mathcal{P}_t^{\mathfrak{h}_{t-1}}(x_t, u_t)$ of transition probabilities associated with the set $\mathcal{M}_t^{\xi_{[t-1]}}$ (these sets of transition probabilities depend on $\pi \in \Pi$, we omit this in the notation for the sake of simplicity). That is, at stage $t = 1$ the set $\mathcal{P}_1(x_1, u_1)$ consists of transition probabilities

$$p_1(x_2|x_1, u_1) = Q_1\{x_2 = F_1(x_1, u_1, \xi_1)\},\ Q_1 \in \mathcal{M}_1.$$

At stage $t = 2$ the set $\mathcal{P}_2^{\mathfrak{h}_1}(x_2, u_2)$ consists of transition probabilities

$$p_2^{\mathfrak{h}_1}(x_3|x_2, u_2) = Q_{2|\xi_{[1]}}\{x_3 = F_2(x_2, u_2, \xi_2)\},\ Q_{2|\xi_{[1]}} \in \mathcal{M}_2^{\xi_{[1]}},$$

with $\mathfrak{h}_1 = (x_1, \pi_1(x_1))$, and so on for the later stages.

In this construction there are no additional implications involved in considering settings where probability distributions of the ambiguity sets depend on our actions. This is in contrast with the SOC model, discussed in section 2, where the assumption of independence of the distributions of the data process from our actions, was essential for the construction of the set $\widehat{\mathfrak{M}}$ for the minimax problem (2.17) (compare with the discussion of Remark 2.1).

Note again that in the rectangular case the transition probabilities (3.11) do not depend on $\mathfrak{h}_{t-1}$ (on $\xi_{[t-1]}$) and are given by (Markovian case)

$$p_t(x_{t+1}|x_t, u_t) = Q_t\left\{x_{t+1} = F_t(x_t, u_t, \xi_t)\right\}, \tag{3.12}$$

where $Q_t \in \mathcal{M}_t$ with $\mathcal{M}_t$ being the set of (marginal) probability distributions of ξ_t.

By solving equation $x_{t+1} = F_t(x_t, u_t, \xi_t)$ for ξ_t, we can write the corresponding cost at stage t as $r_t(x_t, u_t, x_{t+1})$. For instance in the inventory model (Example 2.1) we can write $D_t = x_t + u_t - x_{t+1}$. In the (risk neutral) Markovian case we simply can take $r_t(x_t, u_t) := \mathbb{E}[c(x_t, u_t, \xi_t)]$ with the expectation is taken with respect to the (marginal) distribution of ξ_t.

Acknowledgement The author is indebted to Eugene Feinberg for helpful discussions about the contents of this paper.

References

[1] D.P. Bertsekas and S.E. Shreve. *Stochastic Optimal Control, The Discrete Time Case*. Academic Press, New York, 1978.

[2] L. G. Epstein and M. Schneider. Recursive multiple-priors. *Journal of Economic Theory*, 113(1):1–31, 2003.

[3] H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter, Berlin, 2nd edition, 2004.

[4] G.N. Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research*, 30:257–280, 2005.

[5] A. Jaśkiewicz and A. S. Nowak. Zero-sum stochastic games. In T. Basar and G. Zaccour, editors, *Handbook of dynamic game theory*. Springer, 2016.

[6] S. Mannor and H. Xu. Data-Driven Methods for Markov Decision Problems with Parameter Uncertainty. *INFORMS Tutorials in Operations Research*, pages 101 – 129, 2019.

[7] A. Nilim and L. El Ghaoui. Robust control of Markov Decision Processes with uncertain transition matrices. *Operations Research*, 53:780–798, 2005.

[8] A. Pichler and A. Shapiro. Mathematical foundations of distributionally robust multistage optimization. *https://arxiv.org/abs/2101.02498*, 2020.

[9] A. Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical Programming*, 125:235 – 261, 2010.

[10] A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. *Mathematics of Operations Research*, 31:433–452, 2006.

[11] A. Shapiro. Rectangular sets of probability measures. *Operations Research*, 64:528–541, 2016.

[12] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, 2nd edition, 2014.

[13] L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39:1095–1100, 1953.

[14] W. Wiesemann, D. Kuhn, and B. Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153 – 183, 2013.

[15] H. Xu and S. Mannor. Distributionally robust markov decision processes. *Mathematics of Operations Research*, 37:288 – 300, 2012.

[16] P.H. Zipkin. *Foundations of Inventory Management*. McGraw-Hil, 2000.