STOCHASTIC MODELING AND CONTROL
BANACH CENTER PUBLICATIONS, VOLUME 122
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
WARSZAWA 2020

MARKOV DECISION PROCESSES UNDER AMBIGUITY

NICOLE BÄUERLE
Institute for Stochastics, Karlsruhe Institute of Technology (KIT)
D-76128 Karlsruhe, Germany
ORCID: 0000-0003-0077-3444 E-mail: nicole.baeuerle@kit.edu

ULRICH RIEDER
Institute of Optimization and Operations Research, University of Ulm
D-89069 Ulm, Germany
ORCID: 0000-0003-3604-0488 E-mail: ulrich.rieder@uni-ulm.de

Abstract. We consider statistical Markov Decision Processes where the decision maker is risk averse against model ambiguity. The latter is given by an unknown parameter which influences the transition law and the cost functions. Risk aversion is measured either by the entropic risk measure or by the Average Value at Risk. We show how to solve problems of this kind using a general minimax theorem. Under some continuity and compactness assumptions we prove the existence of an optimal (deterministic) policy and discuss its computation. We illustrate our results using an example from statistical decision theory.

1. Introduction. The following experiment has (in a variant) been suggested by Ellsberg (1961) (see e.g. [GS1989]): An agent has to choose between two bets. For this she is shown two urns, each containing 100 balls which are either red or black. Urn A contains 50 red and 50 black balls while there is no further information about urn B. Bet 1 is: ‘the ball drawn from urn A is black’ and bet 2 is: ‘the ball drawn from urn B is black’. In case of winning the bet, the agent receives 100 euros. Empirically it has been observed that most agents prefer bet 1. One explanation for this behaviour is that in case of urn B agents consider a set of possible distributions for the colours of the balls and being ambiguity averse take into account the minimal expected utility.

2010 *Mathematics Subject Classification*: Primary 90C40, 62C05; Secondary 91G99, 90C47.
Key words and phrases: model ambiguity, Bayesian MDP, entropic risk measure, average value at risk, dual representation, minimax theorem.
The paper is in final form and no version of it will be published elsewhere.

DOI: 10.4064/bc122-2

This point of view has become popular in economics and has been formalized later on. In particular one has to specify the possible set of distributions which have to be taken into account. For example [HS2001] consider in the framework of a continuous-time consumption problem the set of distributions $\mathbb{P}$ whose relative entropy with respect to a fixed distribution $\mathbb{P}_0$ is less or equal to a constant. Using a Lagrange approach this is equivalent to penalizing the robust problem with the distance of the distribution to $\mathbb{P}_0$. Optimization criteria like this have been put on an axiomatic basis by [MMR2006].

As far as Markov Decision Processes (MDPs) are concerned, robust approaches have been considered in [I2005] among others. As in [HS2001] model ambiguity is here treated with respect to the whole probability measure of the process. Since the probability measure in MDP theory is a product of transition kernels such robust optimization problems can also be interpreted as games against nature.

In this paper we will take another point of view which is related to the models introduced in [KMM2005]. There, the whole risk is separated into two parts: Model ambiguity and operating risk. One has to operate a system under an unknown probability law which is chosen by nature (from a finite set) in a worst case way. This model ambiguity is incorporated in the optimization criterion in a risk-sensitive way. For further literature on ambiguity see the survey [GR2013].

We will start with a statistical Markov Decision Process where the transition kernel and cost functions depend on an unknown parameter for which we have a prior distribution. Only the states of the process are observable. Since the Ellsberg experiment suggests that the parameter (model) ambiguity should be treated different to uncertainty of the evolution of the process, we will consider the expected cost of a policy as a random variable and use the entropic risk measure for the model ambiguity. This is in this specific setting similar to the approach in [KMM2005], but different to approaches where the entropic risk measure is applied to the product measure of parameter uncertainty and uncertainty of process evolution. The latter approach has been pursued in [BR2017] and extended to robust problems in [RS2018]. Using the dual representation of the entropic risk measure we can show that there is a connection of our risk-sensitive optimization criterion to the robust penalty problem considered in [HS2001]. Relations like this have already been discussed in [O2012]. However, note that in our setting nature chooses only the worst case measure with respect to the parameter uncertainty. Our model includes both the classical Bayesian MDP (with risk neutral attitude towards ambiguity) and the robust MDP as limiting cases. We use the general minimax theorem of Kneser, Fan, Sion (see [S1958]) and results of [S1979] to solve our problem. It is easy to see from our approach that the solution method not only applies to the case where model ambiguity is evaluated by the entropic risk measure, but to any convex risk measure with a suitable dual representation. Thus, we will also consider the case where model ambiguity is evaluated by the Average Value at Risk. This complements studies in which the Average Value at Risk is applied to the whole discounted cost (see e.g. [BO2011, CTMP2015]).

Our paper is organized as follows: In the next section we introduce our statistical MDP with a given prior distribution μ_0 and the optimization criterion which we consider. Alternative representations and interpretations are also discussed. Section 3 is then devoted to the minimax theorem and the existence of optimal polices. It will turn out that under some continuity and compactness assumptions, optimal policies exist and coincide with the optimal policy of a classical Bayesian MDP with different (more pessimistic) prior μ^* instead of μ_0. The model with Average Value at Risk is discussed in Section 4 and yields from a structural point of view the same policy. Section 5 explains how the problem can be solved in an algorithmic way. In the last section we explain our approach using a specific example from statistical decision theory. In this example we are able to derive the optimal policy for the entropic risk measure as well as for the Average Value at Risk.

2. MDP with entropic risk measure for model ambiguity. We suppose that a *statistical Markov Decision Process* is given which we introduce as follows: We assume that the state space E is a Borel space, i.e., a Borel subset of some Polish space endowed with the σ-algebra of Borel sets. Actions can be taken from a set A which is again a Borel space. The set $D_n \subset E \times A$ is a Borel subset for $n \in \mathbb{N}_0$. By $D_n(x) := \{a \in A : (x,a) \in D_n\}$ we denote the feasible actions depending on the state x at time n. We assume that D_n contains the graph of a measurable mapping from E to A. Furthermore, there is a non-empty parameter space Θ endowed with some σ-algebra. The stochastic transition kernel Q_n^ϑ from D_{n-1} to E which determines the distribution of the new state at time n given the current state and action depends on a parameter $\vartheta \in \Theta$. So, $Q_n^\vartheta(B|x,a)$ is the probability that the next state at time n is in $B \in \mathcal{B}(E)$, given the current state is x and action $a \in D_{n-1}(x)$ is taken. Q_0^ϑ is the distribution of the initial state. In what follows we assume that the law of motion is given by

$$Q_0^\vartheta(dx) := q_0^\vartheta(x)\lambda_0(dx),$$
$$Q_n^\vartheta(dx'|x,a) := q_n^\vartheta(x,a,x')\lambda_n(dx').$$

where λ_n are probability measures on E. Moreover, let

$$(\vartheta, x) \mapsto q_0^\vartheta(x),$$
$$(\vartheta, x, a, x') \mapsto q_n^\vartheta(x,a,x')$$

be non-negative measurable functions on $\Theta \times E$ and $\Theta \times D_{n-1} \times E$ for all $n \in \mathbb{N}$, respectively.

Remark. In general, λ_n are assumed to be σ-finite measures for all n. But then there exists a probability measure λ_n^* and a finite positive density $f_n(x')$ such that $\lambda_n(dx') = f_n(x')\lambda_n^*(dx')$. Thus, we can replace λ_n by λ_n^* and $q_n^\vartheta(x,a,x')$ by $q_n^\vartheta(x,a,x')f_n(x')$ and without loss of generality we may assume that λ_n are probability measures.

Next, we introduce policies for the decision maker. Here it is important to consider the *set of observable histories* which are defined as follows:

$$H_0 := E$$
$$H_n := \big\{(h_{n-1}, a_{n-1}) : h_{n-1} \in H_{n-1},\ a_{n-1} \in D_{n-1}(x_{n-1})\big\} \times E.$$

An element $h_n = (x_0, a_0, x_1, \ldots, x_n) = (h_{n-1}, a_{n-1}, x_n) \in H_n$ denotes the observable history of the process up to time n which consists of the sequence of states and actions. For a Borel set M we denote by $\mathcal{P}(M)$ the set of all probability measures on M. In what follows we consider MDPs with finite horizon $N \in \mathbb{N}$.

Definition 2.1.

a) A measurable mapping $\pi_n : H_n \to \mathcal{P}(A)$ with the property that $\pi_n(h_n)(D_n(x_n)) = 1$ for $h_n \in H_n$ is called a *randomized decision rule* at stage n.
b) A sequence $\pi = (\pi_0, \pi_1, \dots, \pi_{N-1})$ where π_n is a randomized decision rule at stage n for all $n = 0, \dots, N-1$, is called *policy.* We denote by Π the set of all policies.
c) A decision rule $\pi_n : H_n \to \mathcal{P}(A)$ is called *deterministic* if $\pi_n(h_n) = \delta_{f_n(h_n)}$ for some measurable function $f_n : H_n \to A$ with $f_n(h_n) \in D_n(x_n)$. Here δ_x is the one-point measure on x. A policy is called *deterministic* if all decision rules are deterministic.

A policy $\pi = (\pi_0, \pi_1, \dots, \pi_{N-1})$ induces according to the theorem of Ionescu–Tulcea a probability measure

$$\mathbb{P}^\vartheta_\pi := Q^\vartheta_0 \otimes \pi_0 \otimes Q^\vartheta_1 \otimes \pi_1 \otimes Q^\vartheta_2 \otimes \dots \otimes \pi_{N-1} \otimes Q^\vartheta_N$$

on H_N. Since Q^ϑ_n depends measurably on ϑ, we may infer that for any $\pi \in \Pi$, the mapping $(\vartheta, B) \mapsto \mathbb{P}^\vartheta_\pi(B)$ is a transition probability from Θ into H_N.

The corresponding stochastic decision process is given by $(X_0, A_0, X_1, A_1, \dots, X_N)$ and determines the state-action process.

Next, we define our objective function. For this, consider measurable and *bounded* real-valued cost functions

$$(\vartheta, x, a) \mapsto c^\vartheta_n(x, a)$$

on $\Theta \times D_n$, $n = 0, 1, \dots, N-1$ and a measurable and *bounded* terminal cost

$$(\vartheta, x) \mapsto g^\vartheta_N(x)$$

on $\Theta \times E$. All cost functions may depend on the unknown parameter ϑ. Note that in this case we assume that costs are not observable.

We are now interested in the costs incurred by this decision process over the finite time horizon N. Therefore, we define for a policy π

$$C_{N\pi}(\vartheta) := \mathbb{E}^\vartheta_\pi\Big[\sum_{n=0}^{N-1} c^\vartheta_n(X_n, A_n) + g^\vartheta_N(X_N)\Big]$$

where $\mathbb{E}^\vartheta_\pi$ is the expectation with respect to $\mathbb{P}^\vartheta_\pi$. Note that $\vartheta \mapsto C_{N\pi}(\vartheta)$ is measurable on Θ. Suppose $\mu_0 \in \mathcal{P}(\Theta)$ is a fixed initial belief about the unknown parameter ϑ. In the established theory of Bayesian MDP (see e.g. [BR2011, Section 5]) the aim would be to minimize

$$\int C_{N\pi}(\vartheta)\, \mu_0(d\vartheta) \tag{1}$$

over all policies π. This criterion implies that the decision maker is risk neutral with respect to the operating risk as well as with respect to model ambiguity, given in form of the prior μ_0. In what follows we will now consider the case that the decision maker is risk averse with respect to model ambiguity. More precisely, we consider

$$V_N(\pi) := \frac{1}{\gamma} \ln\Big(\int e^{\gamma C_{N\pi}(\vartheta)}\, \mu_0(d\vartheta)\Big), \tag{2}$$

$$V_N := \inf_\pi V_N(\pi) \tag{3}$$

with $\gamma > 0$. For small γ the criterion is approximately equal to (see [BP2003])

$$V_N(\pi) \approx \int C_{N\pi}(\vartheta)\, \mu_0(d\vartheta) + \frac{1}{2}\gamma\, \mathrm{Var}_{\mu_0}[C_{N\pi}],$$

where $\mathrm{Var}_{\mu_0}[C_{N\pi}]$ is the variance of the random variable $C_{N\pi}(\vartheta)$ when ϑ has distribution μ_0. In particular for $\gamma \downarrow 0$ we obtain in the limit the classical Bayesian MDP (1). For $\gamma > 0$ the variability of the minimal cost in ϑ is penalized. Moreover, we have the following representation for (2), also known as 'dual' representation (see [FS2016, p. 279]) where

$$V_N(\pi) = \sup_{\mu \in \mathcal{P}(\Theta)} \Big\{ \int C_{N\pi}(\vartheta)\, \mu(d\vartheta) - \frac{1}{\gamma} I(\mu \| \mu_0) \Big\}$$

with the usual abbreviation

$$I(\mu\|\nu) := \begin{cases} \int \ln(\frac{d\mu}{d\nu})\, d\mu, & \text{if } \mu \ll \nu, \\ \infty, & \text{otherwise,} \end{cases}$$

for the relative entropy function or Kullback–Leibler distance between two measures $\mu, \nu \in \mathcal{P}(\Theta)$. Note that $\int C_{N\pi}(\vartheta)\, \mu(d\vartheta)$ is finite since c_n^ϑ and g_N^ϑ are assumed to be bounded. From this representation we see that the case $\gamma \uparrow \infty$ corresponds to the case of a robust optimization problem or worst-case optimization problem where we minimize the cost if nature chooses the least favourable measure for the parameter ϑ. For $\gamma > 0$ this means that potentially a whole range of beliefs about ϑ is considered but deviations from the belief μ_0 are penalized. A similar criterion has been used in [HS2001] where preferences of an agent for a bet (r.v.) $X : \Omega \to \mathbb{R}$ are expressed by

$$\sup_{\mathbb{P} \in \mathcal{P}(\Omega)} \Big\{ \mathbb{E}_{\mathbb{P}}[X] - \frac{1}{\gamma} I(\mathbb{P} \| \mathbb{P}_0) \Big\},$$

where Ω is a reference probability space corresponding to the outcomes of the underlying stochastic experiment. In our paper we relate model ambiguity only to the unknown parameter ϑ. This is connected to what in economics is called two-stage approach and where ambiguity typically arises in the first (model) stage. Empirically this has been discovered in the famous Ellsberg experiment. In [KMM2005] it has been suggested to consider

$$\sum_j p_j \xi\big(\mathbb{E}_{\mathbb{P}_j}[U(X)]\big)$$

as a preference function where ξ is an increasing real-valued function which describes the agent's attitude towards ambiguity and U is a utility function. $\mathbb{P}_j$ are here different probability measures which correspond to different models.

In what follows we define

$$C_{N\pi}(\mu) := \int C_{N\pi}(\vartheta)\, \mu(d\vartheta).$$

When we insert the dual representation in (3), then we obtain

$$V_N = \inf_\pi \sup_{\mu \in \mathcal{P}(\Theta)} \Big\{ C_{N\pi}(\mu) - \frac{1}{\gamma} I(\mu \| \mu_0) \Big\}. \tag{4}$$

Though it is well-known how the solution of the inner optimization looks like, namely (see [DMS1999, Section 2])

$$\hat{\mu}(d\vartheta) := \frac{e^{C_{N\pi}(\vartheta)}}{\int e^{C_{N\pi}(\vartheta')}\mu_0(d\vartheta')}\,\mu_0(d\vartheta),$$

it is impossible to solve the outer minimization problem directly, nor get some information about the structure of the optimal policy, since $\hat{\mu}$ depends on the policy, too.

3. Existence of optimal policies. It would be easier to solve the problem if we could interchange the sup and the inf in (4). In order to achieve this we use the general minimax theorem of Kneser, Fan, Sion (see [S1958]). The theorem uses the definition of concave-convexlike functions.

Definition 3.1. A function h on $M \times O$ is called *concave-convexlike*, if

(i) for all $x_1, x_2 \in M$ and $0 \le \alpha \le 1$, there is an $x \in M$ such that
$$\alpha h(x_1, y) + (1-\alpha)h(x_2, y) \le h(x, y), \qquad \text{for all } y \in O.$$
(ii) for all $y_1, y_2 \in O$ and $0 \le \alpha \le 1$, there is an $y \in O$ such that
$$\alpha h(x, y_1) + (1-\alpha)h(x, y_2) \ge h(x, y), \qquad \text{for all } x \in M.$$

Note in particular that any function h on $M \times O$ which is concave in the first component and convex in the second component is concave-convexlike. Then Theorem 4.2 in [S1958] tells us:

Theorem 3.2. *Let M be any space and O be a compact space, h a function on $M \times O$ that is concave-convexlike. If $h(x, y)$ is lower semi-continuous in y for all $x \in M$ then*
$$\sup_x \inf_y h(x, y) = \inf_y \sup_x h(x, y).$$

We would like to apply the theorem to the function

$$L_N(\mu, \pi) := C_{N\pi}(\mu) - \frac{1}{\gamma} I(\mu \| \mu_0)$$

which is defined on $\mathcal{P}(\Theta) \times \Pi$. A topology on Π can be introduced as follows: For any $\pi \in \Pi$ denote by $\mathbb{P}^\lambda_\pi$ the probability measure on H_N defined by

$$\mathbb{P}^\lambda_\pi := \lambda_0 \otimes \pi_0 \otimes \lambda_1 \otimes \ldots \otimes \pi_{N-1} \otimes \lambda_N$$

and let $\Pi^\lambda := \{\mathbb{P}^\lambda_\pi : \pi \in \Pi\} \subset \mathcal{P}(H_N)$ be the set of all probability measures $\mathbb{P}^\lambda_\pi$ which are generated by policies. Recall that λ_n are the dominating measures of the transition law introduced at the beginning of Section 2. On $\mathcal{P}(H_N)$ we consider the ws$^\infty$-topology (see [S1975]), i.e. the coarsest topology such that $\mathbb{P} \mapsto \int g\, d\mathbb{P}$ is continuous for all $g : H_N \to \mathbb{R}$ such that $(a_0, \ldots, a_{N-1}) \mapsto g(x_0, a_0, \ldots, a_{N-1}, x_N)$ is continuous for all $x_0, \ldots, x_N$ and the function g is bounded and measurable. Given the relativization of the ws$^\infty$-topology on Π^λ, we can then endow Π with the inverse image under the mapping $\pi \mapsto \mathbb{P}^\lambda_\pi$ of the topology on Π^λ. This is the coarsest topology on Π for which $\pi \mapsto \mathbb{P}^\lambda_\pi$ is continuous.

For the next statements we need some assumptions for all $n = 0, 1, \ldots, N-1$:

(C1) The set $D_n(x)$ is compact for all x.
(C2) The function $a \mapsto q^\vartheta_n(x, a, x')$ is lower semi-continuous for all $x, x' \in E$ and $\vartheta \in \Theta$.
(C3) The function $a \mapsto c^\vartheta_n(x, a)$ is lower semi-continuous for all $x \in E$ and $\vartheta \in \Theta$.

Then we obtain:

Lemma 3.3. *Under* (C1)–(C3):

a) Π *is compact.*

b) *The mapping* $\pi \mapsto L_N(\mu, \pi)$ *is lower semi-continuous on* Π *for all* $\mu \in \mathcal{P}(\Theta)$. *Further, for all* $\pi_1, \pi_2 \in \Pi$ *and* $\alpha \in (0,1)$ *there exists a policy* $\pi \in \Pi$ *such that* $L_N(\mu, \pi) = \alpha L_N(\mu, \pi_1) + (1-\alpha) L_N(\mu, \pi_2)$ *for all* $\mu \in \mathcal{P}(\Theta)$.

c) $\mathcal{P}(\Theta)$ *is convex and* $\mu \mapsto L_N(\mu, \pi)$ *is concave on* $\mathcal{P}(\Theta)$.

Note that b) and c) immediately imply that L_N is concave-convexlike.

Proof.

a) This is Corollary 7.3 b) in [S1979].

b) follows from Corollary 7.3 a) in [S1979]. It suffices to show that $C_{N\pi}(\mu)$ is lower semi-continuous on Π. In order to see how our assumptions are needed we give the following sketch of the proof. First note that

$$\begin{aligned} C_{N\pi}(\vartheta) &= \int \Big(\sum_{n=0}^{N-1} c_n^\vartheta(X_n, A_n) + g_N^\vartheta(X_N) \Big)\, d\mathbb{P}_\pi^\vartheta \\ &= \int \Big(\sum_{n=0}^{N-1} \tilde{c}_n^\vartheta(X_0, A_0, \ldots, X_n, A_n) + \tilde{g}_N^\vartheta(X_0, A_0, \ldots, X_N) \Big)\, d\mathbb{P}_\pi^\lambda =: \tilde{C}_N(\mathbb{P}_\pi^\lambda, \vartheta) \end{aligned}$$

where

$$\begin{aligned} \tilde{c}_n^\vartheta(h_n, a_n) &:= c_n^\vartheta(x_n, a_n) q_n^\vartheta(x_{n-1}, a_{n-1}, x_n) \cdots q_0^\vartheta(x_0) \\ \tilde{g}_N^\vartheta(h_N) &:= g_N^\vartheta(x_N) q_{N-1}^\vartheta(x_{N-1}, a_{N-1}, x_N) \cdots q_0^\vartheta(x_0). \end{aligned}$$

We set $\tilde{C}_N(\mathbb{P}_\pi^\lambda, \mu) := \int \tilde{C}_N(\mathbb{P}_\pi^\lambda, \vartheta)\, \mu(d\vartheta) = C_{N\pi}(\mu)$. To show that $\mathbb{P} \mapsto \tilde{C}_N(\mathbb{P}, \mu)$ is lower semi-continuous on Π^λ in the ws$^\infty$-topology we first assume without loss of generality that $c_n^\vartheta \geq 0$. By (C2) and (C3) $\tilde{c}_n^\vartheta$ depends lower semi-continuously on the actions. Hence $\tilde{c}_n^\vartheta$ can be written as an increasing sequence of functions which are bounded and continuous in the actions. Thus $\mathbb{P} \mapsto \tilde{C}_N(\mathbb{P}, \mu)$ is lower semi-continuous as the limit of an increasing sequence of bounded, continuous functions. The lower semi-continuity of $C_{N\pi}(\mu)$ on Π follows since $\pi \mapsto \mathbb{P}_\pi^\lambda$ is continuous.

Note that for the second statement it is important to work with randomized policies.

c) The convexity of the set is obvious. Concavity of the mapping can also be shown: For this purpose let $\mu_i \ll \mu_0$, $i = 1, 2$. According to the Radon–Nikodym theorem μ_i have densities with respect to μ_0, say $\mu_i = \int g_i\, d\mu_0$. Hence for $\alpha \in (0,1)$ the measure $\mu := \alpha\mu_1 + (1-\alpha)\mu_2$ has density $\alpha g_1 + (1-\alpha) g_2$ with respect to μ_0 and we consider

$$I(\mu \| \mu_0) = \int \ln\big(\alpha g_1(\vartheta) + (1-\alpha) g_2(\vartheta)\big)(\alpha g_1(\vartheta) + (1-\alpha) g_2(\vartheta))\, d\vartheta.$$

Since $x \mapsto x\ln(x)$ is convex for $x > 0$, we deduce that $\mu \mapsto -\frac{1}{\gamma} I(\mu\|\mu_0)$ is concave. Since $\mu \mapsto C_{N\pi}(\mu)$ is linear, the statement follows.

This concludes the proof. ■

Consequently, under (C1)–(C3) all assumptions of Theorem 3.2 are now satisfied and we obtain

Theorem 3.4. *Under assumptions* (C1)–(C3) *we have*:

a) min *and* sup *can be interchanged, i.e.*

$$\min_\pi \sup_{\mu\in\mathcal{P}(\Theta)} L_N(\mu,\pi) = V_N = \sup_{\mu\in\mathcal{P}(\Theta)} \min_\pi L_N(\mu,\pi).$$

b) *There exists an optimal policy* π^* *for* (3), *i.e.* $V_N(\pi^*) = V_N$.

Proof. Part a) is a direct consequence of Theorem 3.2 and the fact that $L_N(\mu,\pi)$ is lower semi-continuous in π due to Lemma 3.3 on the compact set Π. Part b) follows from a) since $\pi \mapsto \sup_{\mu\in\mathcal{P}(\Theta)} L_N(\mu,\pi)$ is lower semi-continuous. ∎

For the second main theorem we need further conditions for all $n = 0,1,\dots,N-1$:

(C4) The parameter space Θ is a compact metric space (endowed with the σ-algebra of Borel subsets of Θ).

(C5) For all $x,x' \in E$, the function $(\vartheta,a) \mapsto q_n^\vartheta(x,a,x')$ is lower semi-continuous on $\Theta \times D_{n-1}(x)$.

(C6) For all $x \in E$, the function $(\vartheta,a) \mapsto c_n^\vartheta(x,a)$ is continuous on $\Theta \times D_{n-1}(x)$ and the function $\vartheta \mapsto g_N^\vartheta(x)$ is continuous on Θ.

These assumptions imply that we obtain a worst prior measure (initial belief). Here we endow $\mathcal{P}(\Theta)$ with the weak topology.

Theorem 3.5. *In addition to* (C1) *assume that* (C4)–(C6) *are satisfied. Then*:

a) *There exists a saddle point* (μ^*,π^*) *of the function* $(\mu,\pi) \mapsto L_N(\mu,\pi)$ *and*

$$\min_\pi \max_{\mu\in\mathcal{P}(\Theta)} L_N(\mu,\pi) = L_N(\mu^*,\pi^*) = V_N = \max_{\mu\in\mathcal{P}(\Theta)} \min_\pi L_N(\mu,\pi).$$

b) *The policy* π^* *is an optimal policy for* (3) *and* π^* *is an optimal Bayes policy with respect to* μ^*, *i.e.* $C_{N\pi^*}(\mu^*) = \inf_\pi C_{N\pi}(\mu^*)$.

c) *There exists a deterministic policy* $f^* := (f_0^*,\dots,f_{N-1}^*)$ *with* $C_{Nf^*}(\mu^*) = C_{N\pi^*}(\mu^*)$, *i.e.* f^* *is optimal for* (3).

Proof.

a) The assumption that Θ is compact implies that $\mathcal{P}(\Theta)$ is weakly compact. Moreover, the mapping $\mu \mapsto L_N(\mu,\pi)$ is upper semi-continuous in the weak topology, since $\mu \mapsto \int \mathbb{E}_\pi^\vartheta[C_{N\pi}(\vartheta)]\,\mu(d\vartheta)$ is continuous by our assumptions (see Corollary 8.3 in [S1979] and the addendum for the lengthy proof) and the entropy function $\mu \mapsto I(\mu\|\mu_0)$ is lower semi-continuous with respect to the weak topology (see Theorem 1 in [P1975]). Hence also $\mu \mapsto \inf_\pi L_N(\mu,\pi)$ is upper semi-continuous and attains its supremum on $\mathcal{P}(\Theta)$ at μ^*. The pair (μ^*,π^*) with π^* from Theorem 3.4 b) is then a saddle point of the function $(\mu,\pi) \mapsto L_N(\mu,\pi)$.

Part b) follows directly from a) since $V_N(\pi^*) = V_N$ is equivalent to $C_{N\pi^*}(\mu^*) = \inf_\pi C_{N\pi}(\mu^*)$. Part c) is well-known in Bayesian MDPs and follows with [H1970], Theorem 15.2 together with Lemma 15.1. ∎

Theorem 3.5 has the advantage that it is possible to solve the inner optimization problem $\min_\pi L_N(\mu,\pi)$ explicitly. Since the entropy part does not depend on the policy π, only the part $C_{N\pi}(\mu)$ is interesting and it can be solved with the established theory of

Bayesian MDP (see Section 5). Of course, the resulting optimal policy depends on μ^* which has to be computed in a second step.

Remark. It is possible to consider MDPs with infinite time horizon in the same way, i.e. instead of $C_{N\pi}(\vartheta)$ we take

$$C_{\infty\pi}(\vartheta) := \mathbb{E}_\pi^\vartheta\Big[\sum_{n=0}^\infty c_n^\vartheta(X_n, A_n)\Big]$$

and assume $\sum_{n=0}^\infty \sup_{\vartheta,x,a} |c_n^\vartheta(x,a)| < \infty$ or a weaker convergence assumption. Then we obtain the same results as for finite-stage MDPs with agents who are ambiguity averse.

Remark. The entropic risk measure motivates to penalize the robust MDP formulation by the deviation of the prior from the 'statistically correct' prior μ_0. Instead of taking the relative entropy one could of course take any other distance which is convex. For example if $\Theta \subset \mathbb{R}$, one could take the *Bhattacharyya distance* which for probability measures μ and μ_0 with densities φ and φ_0 is defined by

$$D_B(\mu, \mu_0) := -\log\Big(\int \sqrt{\varphi(x)\varphi_0(x)}\, dx\Big)$$

and is a convex mapping in μ for fixed μ_0. For details see [B1943].

4. MDP with Average Value at Risk for model ambiguity. Instead of the entropic risk measure one may apply any other convex risk measure to penalize model ambiguity. For example convex risk measures have a representation in dual form (see [FS2016, Theorem 4.33]) which can be used to apply the minimax Theorem. In what follows we restrict the discussion to the Average Value at Risk since it is the most popular one in this class. We consider the same Bayesian MDP framework as in Section 2 with a fixed initial belief μ_0 and define the Value at Risk at level $\gamma \in (0,1)$ for the random variable $\vartheta \mapsto C_{N\pi}(\vartheta)$ on Θ as

$$\mathrm{VaR}_\gamma(C_{N\pi}) := \inf\big\{z \in \mathbb{R} : \mu_0(\{\vartheta \in \Theta : C_{N\pi}(\vartheta) \le z\}) \ge \gamma\big\}.$$

Note that we consider the actuarial point of view here where large positive outcomes are bad and γ is usually close to 1. Moreover, note that $\mathrm{VaR}_\gamma(C_{N\pi})$ depends on μ_0.

When model ambiguity is measured by the Average Value at Risk, we obtain as optimization criterion

$$\begin{aligned} V_N(\pi) &:= \frac{1}{1-\gamma}\int_\gamma^1 \mathrm{VaR}_\alpha(C_{N\pi})\, d\alpha, \\ V_N &:= \inf_\pi V_N(\pi). \end{aligned} \tag{5}$$

Note that for a continuous random variable $C_{N\pi}$, the Average Value at Risk is the same as Expected Shortfall and Tail Conditional Expectation. If $\gamma \downarrow 0$ we get in the limit just the expectation and thus the classical risk neutral setting. For $\gamma \uparrow 1$ we obtain in the limit the worst-case risk measure. Using the dual representation of Average Value at Risk (see e.g. [FS2016, Theorem 4.52]) this amounts to

$$V_N = \inf_\pi \sup_{\mu \in \mathcal{Q}_\gamma} C_{N\pi}(\mu)$$

with $C_{N\pi}(\mu) := \int C_{N\pi}(\vartheta)\,\mu(d\vartheta)$ and

$$\mathcal{Q}_\gamma := \left\{\mu \in \mathcal{P}(\Theta) : \mu \ll \mu_0,\ \frac{d\mu}{d\mu_0} \le \frac{1}{1-\gamma}\right\}.$$

The idea here is to proceed in the same way as in Section 3 and use the previously established results. We obtain with some slight changes to the previous section:

Theorem 4.1. *Under* (C1) *and* (C4)–(C6) *we have*:

a) *There exists a saddle point (μ^*, π^*) of the function $(\mu, \pi) \mapsto C_{N\pi}(\mu)$ and*

$$\min_\pi \max_{\mu\in\mathcal{Q}_\gamma} C_{N\pi}(\mu) = C_{N\pi^*}(\mu^*) = V_N = \max_{\mu\in\mathcal{Q}_\gamma} \min_\pi C_{N\pi}(\mu).$$

b) *The policy π^* is an optimal policy for* (5), *i.e. $V_N(\pi^*) = V_N$, and π^* is an optimal Bayes policy with respect to μ^*, i.e. $C_{N\pi^*}(\mu^*) = \inf_\pi C_{N\pi}(\mu^*)$.*

c) *There exists a deterministic policy $f^* := (f_0^*, \dots, f_{N-1}^*)$ with $C_{Nf^*}(\mu^*) = C_{N\pi^*}(\mu^*)$ and f^* is optimal for* (5), *i.e. $V_N(f^*) = V_N$.*

Proof. First note that Lemma 3.3 holds in the same way since $\mathcal{Q}_\gamma$ is convex. The statements follow as in the proof of Theorem 3.5 since $\mathcal{Q}_\gamma$ is weakly compact (see [FS2016, Corollary 4.38]). ■

5. Solving the Bayesian Dynamic Decision Problem. Theorems 3.5 and 4.1 provide algorithms to solve MDPs with ambiguity. More precisely, the possibility to interchange min and max implies that we can first solve $\min_\pi L_N(\mu, \pi)$ and $\min_\pi C_{N\pi}(\mu)$ and then maximize over μ. This inner optimization problem however is a standard Bayesian Dynamic Decision Problem which can be solved with well-known tools. We will give a sketch here. For a detailed explanation how these problems can be solved, see [BR2011, Section 5]. First, let $\mu_0 \in \mathcal{P}(\Theta)$ be fixed. We consider the problem $\inf_\pi C_{N\pi}(\mu_0) = C_N(\mu_0)$. Note that in order to obtain the optimal policy, we finally have to replace μ_0 by the optimal μ^* in the solution procedure. The problem can be solved by a state space augmentation. The state which has to be considered is (x, μ) where $x \in E$ and $\mu \in \mathcal{P}(\Theta)$ is the current belief (conditional distribution) about ϑ. This belief has to be updated as follows:

$$\mu_0(x)(C) := \frac{\int_C q_0^\vartheta(x)\,\mu_0(d\vartheta)}{\int_\Theta q_0^\vartheta(x)\,\mu_0(d\vartheta)}, \qquad C \in \mathcal{B}(\Theta),$$

$$\Phi_n(x, \mu, a, x')(C) := \frac{\int_C q_n^\vartheta(x, a, x')\,\mu(d\vartheta)}{\int_\Theta q_n^\vartheta(x, a, x')\,\mu(d\vartheta)}, \qquad C \in \mathcal{B}(\Theta),\ n = 1, \dots, N.$$

$\Phi_n(x, \mu, a, x')$ gives the new belief, if the previous belief was μ, the previous state was x, the new state is x' and action a is chosen. $\mu_0(x)$ is the new belief directly after the observation of the first state. Thus, starting with the prior μ_0 we obtain a sequence of beliefs $\mu_n(h_{n-1}, a_{n-1}, x_n) := \Phi_n(x_{n-1}, \mu_{n-1}(h_{n-1}), a_{n-1}, x_n)$ depending on the observations and the history of the process: $\mu_0(x_0), \mu_1(x_0, a_0, x_1), \mu_2(h_2), \dots$. The state transition kernel is given by

$$Q_n^X(B|x, \mu, a) := \int Q_n^\vartheta(B|x, a)\,\mu(d\vartheta), \quad B \in \mathcal{B}(E),\ n = 1, \dots, N.$$

Under well-known continuity and compactness assumptions (see e.g. [BR2011, Theorem 2.4.10]) it is then possible to show that the value

$$C_N(\mu_0) = \int\int J_0(x, \mu_0(x))\, Q_0^\vartheta(dx)\, \mu_0(d\vartheta)$$

can be computed recursively by

$$J_N(x,\mu) := \int g_N^\vartheta(x)\, \mu(d\vartheta),$$

$$J_n(x,\mu) := \inf_{a\in D_n(x)}\Big\{\int_\Theta c_n^\vartheta(x,a)\,\mu(d\vartheta) + \int J_{n+1}(x', \Phi_{n+1}(x,\mu,a,x'))\, Q_{n+1}^X(dx'|x,\mu,a)\Big\}$$

for $n = N-1,\dots,0$. If we denote by g_n^* the (deterministic) minimizer of J_{n+1} on the right-hand side of the equation for $n = 0,1,\dots,N-1$, then the deterministic policy $\pi^* := (f_0^*,\dots,f_{N-1}^*)$ is optimal for the given problem with

$$f_n^*(h_n) := g_n^*(x_n, \mu_n(h_n)), \quad h_n \in H_n, \quad n = 0,1,\dots,N-1,$$

i.e. $C_{N\pi^*}(\mu_0) = C_N(\mu_0)$.

6. An example.

We consider the following example which is taken from [dG1970, Example 2, Section 12.6]. A statistician observes (sequentially) a sequence of at most $N \geq 2$ Bernoulli random variables with unknown success probability ϑ and has to determine from this the true value of ϑ. Suppose that ϑ is either $\frac{1}{3}$ or $\frac{2}{3}$. She has an initial belief μ_0 about the two probabilities. After each observation two actions are available: Either stop the observation process and choose a terminal decision (whether the true probability is either $\frac{1}{3}$ or $\frac{2}{3}$) or make a further observation of the Bernoulli trial. After N observations the statistician has to take a decision. The cost of one observation is 1 and if the decision is correct, there is no cost. For a wrong terminal decision one has to pay the amount of 10. What is the optimal Bayesian strategy? We assume that the statistician is risk averse and uses the criteria presented in this paper.

We use Theorem 3.5 and Theorem 4.1, to solve these problems. We have to take as the state space the current belief about the two hypothesis. Since the parameter set is $\Theta = \{\frac{1}{3}, \frac{2}{3}\}$ these beliefs are only two-point distributions. In what follows we assume that the interval $[0,1]$ is the state space where $\mu \in [0,1]$ is the current belief that $\vartheta = \frac{1}{3}$ is the true parameter. The action space is $A = \{1,2\}$ where $a = 1$ means to take another observation and $a = 2$ means to stop the observation process and choose a terminal decision (which is then the hypothesis with higher belief). Note that since Θ and A are finite, (C1)–(C6) are satisfied. In this case the cost is given by

$$c(\mu) := \min\{10\mu, 10(1-\mu)\}.$$

In case we decide to take another observation and the observation is a ‘success’ (indicated by ‘1’) we obtain the following new belief:

$$\Phi(\mu, 1) = \frac{\mu/3}{\mu/3 + 2(1-\mu)/3} = \frac{\mu}{2-\mu}\,.$$

In case we observe a 'failure' (indicated by '0') we obtain for the new belief

$$\Phi(\mu,0)=\frac{2\mu/3}{2\mu/3+(1-\mu)/3}=\frac{2\mu}{1+\mu}\,.$$

Then we get from the Bayesian MDP theory the following recursion for $n=1,\dots,N$:

$$\begin{aligned}C_0(\mu)&=c(\mu),\\ C_n(\mu)&=\min\Big\{c(\mu);1+\Big(\frac{1}{3}\mu+\frac{2}{3}(1-\mu)\Big)C_{n-1}\Big(\frac{\mu}{2-\mu}\Big)\\ &\qquad+\Big(1-\frac{1}{3}\mu-\frac{2}{3}(1-\mu)\Big)C_{n-1}\Big(\frac{2\mu}{1+\mu}\Big)\Big\}.\end{aligned}$$

Working through this recursion finally yields:

$$C_0(\mu)=\begin{cases}10\mu, & 0\le\mu\le\frac{1}{2},\\ 10(1-\mu), & \frac{1}{2}<\mu\le 1.\end{cases}$$

and $C_n=C_1$ for all $n\in\mathbb{N}$, where

$$C_1(\mu)=\begin{cases}10\mu, & 0\le\mu\le\frac{13}{30},\\ \frac{13}{3}, & \frac{13}{30}<\mu\le\frac{17}{30},\\ 10(1-\mu), & \frac{17}{30}<\mu\le 1.\end{cases}$$

The optimal decision at the beginning is to take another observation $(a=1)$ if $\mu\in(\frac{13}{30},\frac{17}{30})$, otherwise take a terminal decision $(a=2)$ immediately. After one observation the statistician will always take a terminal decision.

6.1. Problem with entropic risk measure. When we want to solve the problem with risk aversion against ambiguity measured by the entropic risk measure, we have to consider the following problem for $N\ge 1$ where $\mu_0\in(0,1)$ is the initial belief

$$V_N=\sup_{0<\mu<1}\Big\{C_1(\mu)-\frac{1}{\gamma}\,I(\mu\|\mu_0)\Big\}.$$

Using the fact that the function C_1 is symmetric, i.e. $C_1(\mu)=C_1(1-\mu)$, this boils down to

$$\begin{aligned}V_N=\max\Big\{&\sup_{0<\mu<\frac{13}{30}}\Big\{10\mu-\frac{1}{\gamma}\Big(\mu\ln\Big(\frac{\mu}{\mu_0}\Big)+(1-\mu)\ln\Big(\frac{1-\mu}{1-\mu_0}\Big)\Big)\Big\};\\ &\sup_{\frac{13}{30}<\mu\le\frac{1}{2}}\Big\{\frac{13}{3}-\frac{1}{\gamma}\Big(\mu\ln\Big(\frac{\mu}{\mu_0}\Big)+(1-\mu)\ln\Big(\frac{1-\mu}{1-\mu_0}\Big)\Big)\Big\}\Big\}.\end{aligned}$$

When the statistician is risk averse with parameter $\gamma=0.1$ and has initial belief $\mu_0=0.1$ about the hypothesis $\vartheta=\frac{1}{3}$ she will rather solve the Bayesian MDP with $\mu^*=0.232$. Observe that $\mu=\frac{1}{2}$ is the most risk averse choice of the prior. In Figure 1 the optimal μ^* is plotted as a function of the risk aversion γ for different μ_0. Note that

$$\begin{aligned}V_N&=C_1(\mu^*)-\frac{1}{\gamma}I(\mu^*\|\mu_0) && \text{if }\gamma>0\\ V_N&=C_1(\mu_0) && \text{if }\gamma=0.\end{aligned}$$

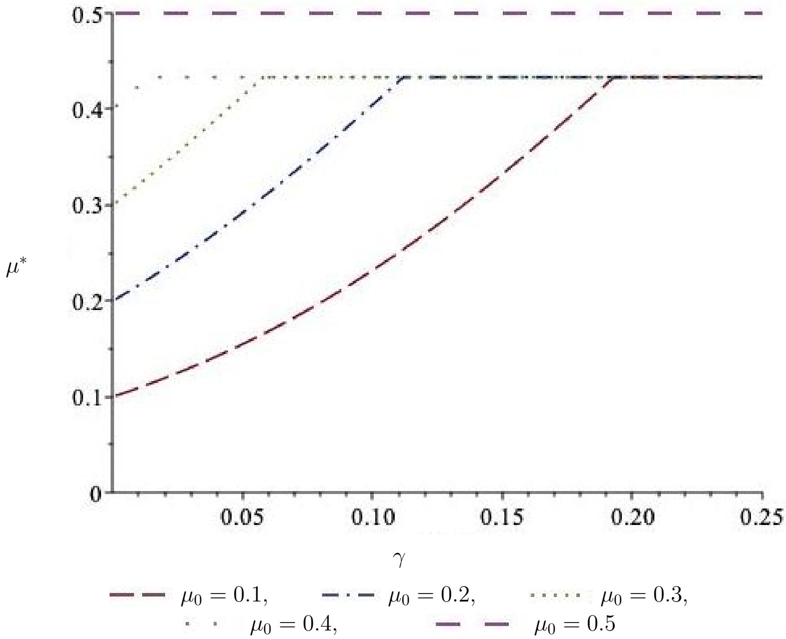

Fig. 1. μ^* as a function of γ for different μ_0

What we observe in the example is that

(i) For $\mu_0 \in [0, \frac{1}{2}]$ we have $\mu_0 \le \mu^* \le \frac{1}{2}$.
(ii) $\lim_{\gamma \downarrow 0} \mu^*(\gamma) = \mu_0$.
(iii) $\lim_{\gamma \uparrow \infty} \mu^*(\gamma) \in [\frac{13}{30}, \frac{1}{2}]$.

The interpretation of (i) is that a risk averse statistician will always shift the statistically correct prior in direction of the uniform distribution. The case $\lim_{\gamma \downarrow 0}$ is in the limit the classical Bayesian MDP with original prior μ_0. The case $\lim_{\gamma \uparrow \infty}$ corresponds to the robust optimization where the most unfavourable prior is chosen. In this example the most unfavourable prior is any prior in the interval $[\frac{13}{30}, \frac{1}{2}]$ since this requires another observation.

6.2. Problem with Average Value at Risk. We can also consider this example with the ambiguity measured by the Average Value at Risk. Here we have to solve

$$\max_{\mu \in \mathcal{M}_\gamma} C_1(\mu), \qquad \mathcal{M}_\gamma := \left\{ \mu \in (0,1) : \frac{\mu}{\mu_0} \le \frac{1}{1-\gamma}, \ \frac{1-\mu}{1-\mu_0} \le \frac{1}{1-\gamma} \right\}.$$

The set $\mathcal{M}_\gamma$ corresponds to $\mathcal{Q}_\gamma$. The maximum point μ^* as a function of γ and $\mu_0 \le \frac{1}{2}$ is given by (in case on non-uniqueness we give the whole range of optimal values)

$$\mu^* = \begin{cases} \dfrac{\mu_0}{1-\gamma}, & \gamma \le 1 - \dfrac{30}{13}\mu_0 \\ \left(\dfrac{13}{30}, \dfrac{\mu_0}{1-\gamma} \right), & \gamma \in \left(1 - \dfrac{30}{13}\mu_0, \ 1 - \dfrac{30}{17}\mu_0 \right) \\ \left(\dfrac{13}{30}, \dfrac{17}{30} \right), & \gamma \ge 1 - \dfrac{30}{17}\mu_0. \end{cases}$$

Due to symmetry reasons we restrict again to the case $\mu_0 \leq \frac{1}{2}$. In Figure 2 the optimal μ^* is plotted as a function of γ for different μ_0. Note that

$$\begin{aligned} V_N &= C_1(\mu^*) \quad \text{if } \gamma > 0 \\ V_N &= C_1(\mu_0) \quad \text{if } \gamma = 0. \end{aligned}$$

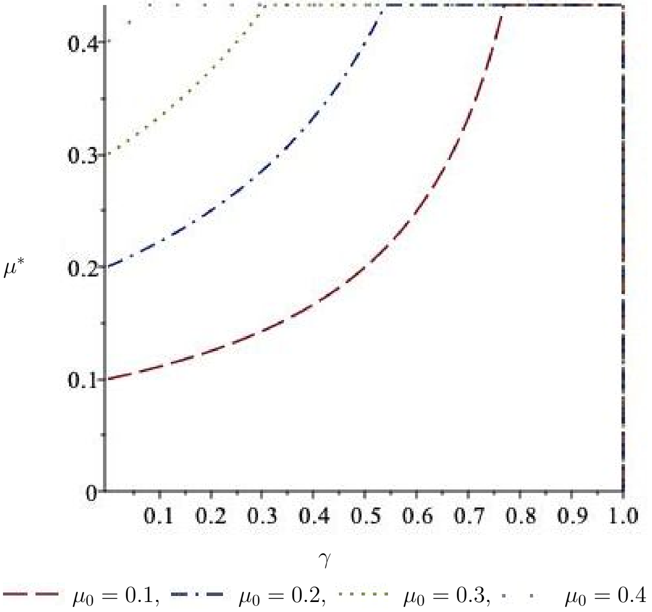

Fig. 2. μ^* as a function of γ for different μ_0

We again observe in this case that

(i) For $\mu_0 \in [0, \frac{1}{2}]$ we have $\mu_0 \leq \mu^* \leq \frac{1}{2}$.
(ii) $\lim_{\gamma \downarrow 0} \mu^*(\gamma) = \mu_0$.
(iii) $\lim_{\gamma \uparrow 1} \mu^*(\gamma) \in [\frac{13}{30}, \frac{1}{2}]$.

Though in this case the interpretation of γ is different, the general behaviour of the optimal μ^* is the same.

7. Conclusion. In this paper we present a proposal to deal with model ambiguity for MDPs. Using a dual representation and a general minimax theorem we are able to solve the ambiguity problem. The solution procedure is illustrated by an example taken from statistical decision theory. The approach is closely related to robust MDPs.

References

[BO2011] N. Bäuerle, J. Ott, *Markov decision processes with average-value-at-risk criteria*, Math. Methods Oper. Res. 74 (2011), 361–379.

[BR2011] N. Bäuerle, U. Rieder, *Markov Decision Processes with Applications to Finance*, Springer, Heidelberg, 2011.

[BR2017] N. Bäuerle, U. Rieder, *Partially observable risk-sensitive Markov decision processes*, Math. Oper. Res. 42 (2017), 1180–1196.

[B1943] A. Bhattacharyya, *On a measure of divergence between two statistical populations defined by the probability distributions*, Bull. Calcutta Math. Soc. 35 (1943), 99–109.

[BP2003] T. Bielecki, S. R. Pliska, *Economic properties of the risk sensitive criterion for portfolio management*, Rev. Account. Fin. 2 (2003), no. 2, 3–17.

[CTMP2015] Y. Chow, A. Tamar, S. Mannor, M. Pavone, *Risk-sensitive and robust decision-making: a CVaR optimization approach*, in: Advances in Neural Information Processing Systems 28, 2015, 1522–1530.

[dG1970] M. H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, New York, 1970.

[DMS1999] G. B. Di Masi, L. Stettner, *Risk-sensitive control of discrete-time Markov processes with infinite horizon*, SIAM J. Control Optim. 38 (1999), 61–78.

[FS2016] H. Föllmer, A. Schied, *Stochastic Finance: an Introduction in Discrete Time*, 4th edition, Walter de Gruyter, Berlin, 2016.

[GS1989] I. Gilboa, D. Schmeidler, *Maxmin expected utility with non-unique prior*, J. Math. Econom. 18 (1989), 141–153.

[GR2013] M. Guidolin, F. Rinaldi, *Ambiguity in asset pricing and portfolio choice: A review of the literature*, Theory and Decision 74 (2013), 183–217.

[HS2001] L. Hansen, T. J. Sargent, *Robust control and model uncertainty*, Amer. Economic Review 91 (2001), no. 2, 60–66.

[H1970] K. Hinderer, *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Oper. Res. and Math. Systems 33, Springer, Berlin, 1970.

[I2005] G. N. Iyengar, *Robust dynamic programming*, Math. Oper. Res. 30 (2005), 257–280.

[KMM2005] P. Klibanoff, M. Marinacci, S. Mukerji, *A smooth model of decision making under ambiguity*, Econometrica 73 (2005), 1849–1892.

[MMR2006] F. Maccheroni, M. Marinacci, A. Rustichini, *Ambiguity aversion, robustness, and the variational representation of preferences*, Econometrica 74 (2006), 1447–1498.

[O2012] T. Osogami, *Robustness and risk-sensitivity in Markov decision processes*, in: Advances in Neural Information Processing Systems 25, 2012, 233–241.

[P1975] E. Posner, *Random coding strategies for minimum entropy*, IEEE Trans. Inform. Theory 21 (1975), 388–391.

[RS2018] M. Rasouli, S. Saghafian, *Robust partially observable Markov decision processes*, preprint, 2018.

[S1975] M. Schäl, *On dynamic programming: compactness of the space of policies*, Stochastic Process. Appl. 3, no. 4, 1975, 345–364.

[S1979] M. Schäl, *On dynamic programming and statistical decision theory*, Ann. Statist. 7 (1979), 432–445.

[S1958] M. Sion, *On general minimax theorems*, Pacific J. Math. 8 (1958), 171–176.