A Unified Framework for Risk-sensitive Markov Control Processes*

Yun Shen† Wilhelm Stannat‡ Klaus Obermayer§

Abstract

We introduce a unified framework for measuring risk in the context of Markov control processes with risk maps on general Borel spaces that generalize known concepts of risk measures in mathematical finance, operations research and behavioral economics. Within the framework, applying weighted norm spaces to incorporate also unbounded costs, we study two types of infinite-horizon risk-sensitive criteria, discounted total risk and average risk, and solve the associated optimization problems by dynamic programming. For the discounted case, we propose a new discount scheme, which is different from the conventional form but consistent with the existing literature, while for the average risk criterion, we state Lyapunov-type stability conditions that generalize known conditions for Markov chains to ensure the existence of solutions to the optimality equation.

1 Introduction

Markov control processes (MCPs, see, e.g., [23, 24] and [31] under the name *Markov decision processes*) are widely applied to model sequential decision making problems of agents. The induced optimal control problem is to find the best policy that minimizes the expected total costs. The core of the MCP-framework consists of two *objective* descriptions of the environment: *transition probabilities* of switching states when performing actions, and immediate *outcomes* (rewards or costs) obtained at states by executing actions. Facing the same environment, however, different agents might have different policies. Therefore, in many applications, it is important to also incorporate the *subjective* perceptions of an agent into the MCP-framework. The subjective outcomes are usually modeled by *utility functions* (see, e.g., [19]),

* *Date:* July 21, 2014. Accepted by the 53^{rd} *IEEE Conference on Decision and Control* as a regular SIAM paper. This work was supported by the BMBF: Bernsteinfokus Lernen TP1, 01GQ0911 for Y. Shen and K. Obermayer, and FKZ01GQ1001B for W. Stannat.

†Fakultät Elektrotechnik und Informatik, Technische Universität Berlin, Marchstr. 23, 10587, Berlin, Germany (yun@ni.tu-berlin.de).

‡Institut für Mathematik, and Bernstein Center for Computational Neuroscience, Technische Universität Berlin, Straße des 17. Juni 136, 10623, Berlin, Germany (stannat@math.tu-berlin.de).

§Fakultät Elektrotechnik und Informatik, and Bernstein Center for Computational Neuroscience, Technische Universität Berlin, Marchstr. 23, 10587, Berlin, Germany (oby@ni.tu-berlin.de).

which can be easily incorporated by simply replacing the immediate outcome with its utility, whereas the subjective transition probabilities require a more sophisticated mathematical framework. They are commonly incorporated in the *risk*.

Coherent/convex risk measures (CRMs, [1, 16]) have been widely employed to model subjective probabilities in mathematical finance since the last decade. Several works (see, e.g, [32, 15, 34, 6] and references therein) extend CRMs to temporal structures in various setups, where mainly finite-horizon problems are considered. On the contrary, in the literature of MCPs, when studying the infinite-horizon risk-sensitive optimal control problems, merely the *entropic map* (see, e.g., [8, 2, 3, 12, 9, 14, 21, 4]) is applied, which is convex and in fact a special type of CRM. All risk measures mentioned in the above literature are coherent/convex based on the assumption that the agent is economically rational and therefore *risk-averse*. This limits applications in the fields of decision-making under risk and behavioral economics, where more general risk measures (see, e.g., [36, 5, 42] and references therein) are applied, since human beings are not always risk-averse. However, the models in these fields can only be applied to one-step decision making problems.

To overcome the limitations mentioned above, 1) we extend the definition of CRMs to include the risk measures considered also in behavioral economics; 2) we apply a *constructive approach* which maintains the Markov property that is necessary for the existence of stationary optimal policies for two infinite-horizon objectives, albeit less general than the risk maps used in [34] and [33].

With the generalized risk measures and constructed risk maps, we provide a unified treatment in the context of MCPs to infinite-horizon risk-sensitive optimal control problems considered in various fields, e.g., optimal control, operations research, finance and behavioral economics. Using weighted norm spaces, we can incorporate unbounded costs in risk-sensitive MCPs also. We prove that two types of objectives, the discounted total risk and the average risk, can be optimized with *dynamic programming* algorithms under proper assumptions. For the case of discounted risk, we apply a new discount scheme which is different from the conventional form but consistent with the one applied in [33], where coherent risk measures are considered. For the average case, we state sufficient conditions, which generalize Lyapunov-type conditions from the literature of Markov chains (see, e.g., [29]), to ensure the existence of solutions to the associated optimality equation.

In the following sections, we introduce the mathematical framework and state our main results. For detailed proofs, we refer the reader to [37] and [38] for verbatim arguments.

2 Risk Maps on Weighted Norm Spaces

Weighted norm spaces. Let $\mathbf{X}$ be a *Borel space*, which is a Borel subset of a complete separable metric space, and its Borel σ-algebra is denoted by $\mathcal{B}(\mathbf{X})$. Let $w : \mathbf{X} \to [1, \infty)$ be a given real-valued $\mathcal{B}(\mathbf{X})$-measurable function. Consider the w-norm $\|u\|_w := \sup_{x\in\mathbf{X}} \frac{|u(x)|}{w(x)}$. Let $\mathscr{B}_w$ be the space of real-valued $\mathcal{B}(\mathbf{X})$ measurable functions with bounded w-norm. It is obvious that $\mathscr{B} \subset \mathscr{B}_w$, where $\mathscr{B}$ denotes

the space of bounded real-valued $\mathcal{B}(\mathbf{X})$-measurable functions. The *w-seminorm* is defined as $\|v\|_{s,w} := \sup_{x\neq y} \dfrac{|v(x)-v(y)|}{d_w(x,y)}$, where $d_w(x,y) := [w(x)+w(y)]\cdot \mathbf{1}_{x\neq y}$. This seminorm is originally used in [20] to study the ergodicity of Markov chains. In particular, when restricting to the space $\mathscr{B}$, i.e., setting $w \equiv 1$, the seminorm is called *span-norm* in [22] or *Hilbert seminorm* in [18].

Risk measures. The partial ordering $\leq$ between elements in $\mathscr{B}_w$ is defined as $v \leq u$ if $v(x) \leq u(x)\ \forall x \in \mathbf{X}$. A real number $u \in \mathbb{R}$ can be viewed as a constant-valued function which belongs also to $\mathscr{B}_w$. We now define risk measures on $\mathscr{B}_w$. A mapping $\nu : \mathscr{B}_w \to \mathbb{R}$ is said to be a *risk measure* (see [1, 16]), if (I) (Monotonicity) $\nu(v) \leq \nu(u)$, whenever $v \leq u \in \mathscr{B}_w$; (II) (Translation invariance) $\nu(v+u) = \nu(v)+u$, $\forall u \in \mathbb{R}$; (III) (Centralization) $\nu(0) = 0$.

Within the economic context, u and v can be viewed as costs of two choices. Monotonicity reflects the intuition that if the cost of one choice is *always* (for all events) lower than the cost of another choice, the *risk* of the choice must be also lower. Under the axiom of translation invariance, the sure cost u (equal outcome for every event) after executing decisions, is considered as a sure cost before making decision. This also reflects the intuition that there is no risk if there is no uncertainty. For convenience, the axiom of centralization sets the reference point to 0.

Risk measures can be categorized as follows: a risk measure ν is said to be *convex*, if for all $\alpha \in [0,1]$, $v, u \in \mathscr{B}_w$,

$$\nu(\alpha v + (1-\alpha)u) \leq \alpha\nu(v) + (1-\alpha)\nu(u); \tag{1}$$

concave, if $\tilde{\nu}(\cdot) := -\nu(-\cdot)$ is a convex risk measure; *homogeneous*, if $\nu(\lambda v) = \lambda\nu(v)$ for all $\lambda \in \mathbb{R}_+$ and $v \in \mathscr{B}_w$; *coherent*, if ν is convex and homogeneous.

Risk preference. To judge the *risk-preference* induced by a risk measure, we follow the rule that *diversification* should be preferred if the agent is *risk-averse*. More specifically, suppose an agent has two possible choices with the same probability measure, one of which leads to the future cost v while the other one leads to the future cost u. If the agent *diversifies*, i.e., if one spends only a fraction α of the resources on the first and the remaining amount on the second alternative, the future cost is given by $\alpha u + (1-\alpha)v$. If the applied risk measure is convex, then by (1), the diversification should reduce the risk. Thus, we call the agent's behavior *risk-averse*. Conversely, if the applied risk measure is concave, the induced risk-preference should be *risk-seeking*.

Remark. 1) Note that in our definition, in order to be consistent with the notations of Markov control processes, v and u stand for costs, rather than outcomes (rewards), which are employed in the literature of mathematical finance. Hence, the sign of the inequality in Axiom I is opposite to the sign in the same axiom used in [1] and [16]. Furthermore, in order to incorporate more general types of risk measures applied in behavioral economics, the risk measures in our definition are not necessarily convex/coherent.

2) Comparing with the literature of CRMs, [1, 16], we define risk measures on the weighted space $\mathscr{B}_w$, rather than the space of bounded random variables, L^∞, since the weighted norm space is more suitable for investigating the stability properties of

the underlying Markov chain (see, e.g., [29, 20, 37]) and is also more general than L^∞. We will specify later in Section 4 the choice of w, depending on the form of risk maps and the properties of the underlying Markov control process as well.

3 Risk-sensitive Markov Control Processes

Given two Borel spaces $\mathbf{X}$ and $\mathbf{Y}$, a *stochastic kernel on* $\mathbf{X}$ *given* $\mathbf{Y}$ is a function $\psi(B|y), B \in \mathcal{B}(\mathbf{X}), y \in \mathbf{Y}$ such that i) $\psi(\cdot|y)$ is a probability measure on $\mathcal{B}(\mathbf{X})$ for every fixed $y \in \mathbf{Y}$, and ii) $\psi(B|\cdot)$ is a measurable function on $\mathbf{Y}$ for every fixed $B \in \mathcal{B}(\mathbf{X})$.

A **Markov control process** [24], $(\mathbf{X}, \mathbf{A}, \{\mathbf{A}(x)|x \in \mathbf{X}\}, Q, c)$, consists of the following components: *state space* $\mathbf{X}$ and *action space* $\mathbf{A}$, which are Borel spaces; the feasible action set $\mathbf{A}(x)$, which is a nonempty Borel space of $\mathbf{A}$, $x \in \mathbf{X}$; the *transition model* $Q(B|x,a), B \in \mathcal{B}(\mathbf{X}), (x,a) \in \mathbf{K}$, which is a *stochastic kernel* on $\mathbf{X}$ given $\mathbf{K}$, where $\mathbf{K}$ denotes the set of feasible state-action pairs $\mathbf{K} := \{(x,a)|x \in \mathbf{X}, a \in \mathbf{A}(x)\}$, which is a Borel subset of $\mathbf{X} \times \mathbf{A}$; and the *cost function* c: $\mathbf{K} \to \mathbb{R}$, $\mathcal{B}(\mathbf{K})$-measurable. Random variables are denoted by capital letters, whereas realizations of the random variables are denoted by lowercase letters.

We consider *Markov policies*, $\boldsymbol{\pi} = [\pi_0, \pi_1, \pi_2, \ldots]$, where each *single-step policy* $\pi_t(\cdot|x_t)$, which denotes the probability of choosing action a_t at x_t, $(x_t, a_t) \in \mathbf{K}$, is a stochastic kernel on $\mathbf{A}$ given $\mathbf{X}$. Let Δ denote the set of all stochastic kernels on $\mathbf{A}$ given $\mathbf{X}$, μ, such that $\mu(\mathbf{A}(x)|x) = 1$ and $\Pi_M = \Delta^\infty$ denotes the set of all Markov policies. A policy $f \in \Delta$ is *deterministic* if for each $x \in \mathbf{X}$, there exists some $a \in \mathbf{A}(x)$ such that $f(\{a\}|x) = 1$. A policy $\boldsymbol{\pi}$ is said to be *stationary*, if $\boldsymbol{\pi} = \pi^\infty$ for some $\pi \in \Delta$. For each $x \in \mathbf{X}$ and single-step policy $\pi \in \Delta$, define

$$c^\pi(x) := \int_{\mathbf{A}(x)} c(x,a)\pi(\mathrm{d}a|x), \quad \text{and}$$
$$P^\pi(B|x) := \int_{\mathbf{A}(x)} Q(B|x,a)\pi(\mathrm{d}a|x), B \in \mathcal{B}(\mathbf{X}).$$

Three types of objectives are used in the literature of MCPs: finite-stage, discounted and average cost, depicted as $S_T := \sum_{t=0}^{T} c(X_t, A_t), S_\alpha := \sum_{t=0}^{\infty} \alpha^t c(X_t, A_t)$, and $S_A := \limsup_{T\to\infty} \frac{1}{T} S_T$, where $\alpha \in [0,1)$ denotes the discount factor. Suppose we start from one given state $X_0 = x$. The optimization problem is then to minimize the expected objective

$$\inf_{\boldsymbol{\pi} \in \Pi_M} \mathbb{E}^{\boldsymbol{\pi}} \left[\mathcal{S}|X_0 = x\right], \mathcal{S} = S_T, S_\alpha, \text{ or } S_A,$$

by selecting a policy $\boldsymbol{\pi}$. We notice that the finite-stage objective function can be decomposed as follows,

$$\begin{aligned}\mathbb{E}^{\boldsymbol{\pi}}_{X_0}[S_T] =& c^{\pi_0}(X_0) + \mathbb{E}^{\pi_0}_{X_0}[c^{\pi_1}(X_1) + \mathbb{E}^{\pi_1}_{X_1}[c^{\pi_2}(X_2) + \ldots \\ &+ \mathbb{E}^{\pi_{T-1}}_{X_{T-1}}[c^{\pi_T}(X_T)]\ldots]]\end{aligned}$$

where $\mathbb{E}^{\pi_t}_{X_t}[v(X_{t+1})] := \int v(X_{t+1}) P^{\pi_t}(dX_{t+1}|X_t)$ denotes the *conditional expectation* of the function v of the successive state X_{t+1} given current state X_t. Obviously, the conditional expectation plays the key role in the calculation of all three objectives.

Risk maps. In order to incorporate risk, we directly replace the expectation $\mathbb{E}^{\pi_t}_{X_t}$ with a *risk map* $\mathcal{R}^{\pi_t}_{X_t}$ that is defined as follows.

DEFINITION 3.1. *A mapping* $\mathcal{R}(v|x,a) : \mathbf{K} \times \mathscr{B}_w \to \mathbb{R}$ *is said to be a* risk map *on an MCP* $(\mathbf{X}, \mathbf{A}, \{\mathbf{A}(x)|x \in \mathbf{X}\}, Q)$*, if (i) for each* $(x,a) \in \mathbf{K}$*,* $\mathcal{R}(\cdot|x,a) : \mathscr{B}_w \to \mathbb{R}$ *is a risk measure; (ii) for each* $v \in \mathscr{B}_w$*,* $\mathcal{R}(v|\cdot)$ *is* $\mathcal{B}(\mathbf{K})$*-measurable. Furthermore, define* $\mathcal{R}^\pi(v|x) := \int_{\mathbf{A}(x)} \pi(da|x)\mathcal{R}(v|x,a)$.

For convenience, we sometimes write $\mathcal{R}_{x,a}(v) := \mathcal{R}(v|x,a)$ and $\mathcal{R}^\pi_x(v) := \mathcal{R}^\pi(v|x)$. With the replacement, we obtain the *T-stage risk-sensitive* objective

$$\begin{aligned} J^{\boldsymbol{\pi}}_T =& c^{\pi_0}(X_0) + \mathcal{R}^{\pi_0}_{X_0}[c^{\pi_1}(X_1) + \mathcal{R}^{\pi_1}_{X_1}[c^{\pi_2}(X_2) + \dots \\ &+ \mathcal{R}^{\pi_{T-1}}_{X_{T-1}}[c^{\pi_T}(X_T)] \dots]], \end{aligned}$$

and its optimization problem can be solved by *dynamic programming* (see, e.g., [33]), whereas the other two objectives will be defined analogously and discussed in Section 4.

Analogously, a risk map $\mathcal{R}$ is said to be *convex* (resp. *concave, homogenous, coherent*) if $\mathcal{R}_(x,a)$ is convex (resp. *concave, homogenous, coherent*), for all $(x,a) \in \mathbf{K}$.

Remark. 1) Note that since risk maps are subjective representations of objective transition probabilities, as in the above definition, $\mathcal{R}$ depends always on the transition model Q of the underlying MCP. It is obvious that the transition kernel Q is a valid risk map. Thus, the concept of risk maps is a generalization of the conditional expectation. In Definition 3.1, $\mathcal{R}^\pi$ is in fact assumed to be linear to the policy π, which simplifies the optimization problem and is one of the conditions that guarantee the existence of one optimal deterministic policy ("optimal selector").

2) In the mathematical finance literature, there exist various ways to extend the CRM to a temporal structure (see [15, 6, 33] and references therein). The definition is usually selected based on applications. To compare their subtle differences are out of the scope of this paper. The risk maps defined here are similar to the *risk measure generators* in [6] and are implicitly Markovian and time-homogeneous (see also [33]), since $\mathcal{R}$ defined above depends merely on the most recent state and action but not the whole history. The risk maps used in this paper are assumed to be Markovian, since in the MCP-framework the underlying stochastic process is Markovian, while the assumption of time-homogeneity is due to the fact that since we consider the infinite-horizon criteria (see Section 4), as in the literature of MCPs, stationary optimal policies are expected. Hence, to comply with the MCP-framework, it is sufficient to construct an operator which replaces the conditional expectation determined by the transition model Q and policy π.

Examples. There exist several important risk maps in the literature of economics, mathematical finance and control theory. Most of them can be adapted to the framework we introduced above.

Utility-based shortfall [17, Chapter 4]:

$$\mathcal{R}_{x,a}(v) := \sup\left\{m \in \mathbb{R}\middle| \int u(v(y)-m)Q(dy|x,a) \geq 0\right\},$$

where the utility function $u : \mathbb{R} \to \mathbb{R}$ is increasing and nonconstant and satisfies $u(0) = 0$. It can be shown that for each $(x, a) \in \mathbf{K}$, $\mathcal{R}_{x,a}(\cdot)$ satisfies the three axioms of risk measures and if u is convex, then $\mathcal{R}$ is convex as well [17, Chapter 4]. It contains the following special cases:

1) *classical MCPs:* setting $u(x) = x$,

$$\mathcal{R}_{x,a}(v) = \mathbb{E}^Q_{x,a}[v] := \int_{\mathbf{X}} Q(\mathrm{d}y|x,a)v(y).$$

2) *entropic maps* [16]: setting $u(x) = e^{\lambda x} - 1$, $\lambda \neq 0$,

$$\mathcal{R}_{x,a}(v) := \frac{1}{\lambda}\log\left\{\mathbb{E}^Q_{x,a}\left[e^{\lambda v}\right]\right\} \tag{2}$$

where the risk-sensitive parameter $\lambda \in \mathbb{R}$ controls the risk-preference of $\mathcal{R}$: if $\lambda > 0$, $\mathcal{R}$ is convex and therefore risk-averse; if $\lambda < 0$, $\mathcal{R}$ is concave and therefore risk-seeking. This risk map is intensively studied in the field of optimal control ([26, 8, 2, 3, 12, 9, 14, 21, 4]). It has also a connection to the mean-variance trade-off via the Taylor expansion at $\lambda = 0$: $\frac{1}{\lambda}\log \mathbb{E}e^{\lambda Z} = \mathbb{E}Z + \lambda \mathrm{Var}(Z) + O(\lambda^2)$. Suppose that risk is measured by variance. Since the objective is to minimize risk $\mathcal{R}^\pi$, if $\lambda > 0$, the variance is avoided, the agent is risk-averse. On the contrary, if $\lambda < 0$, the variance is preferred, the agent is intuitively risk-seeking. These intuitions coincide with the categorization based on the convexity (concavity) of $\mathcal{R}$.

Robust risk. In [27], to gain the "robustness", the worst cost among a set of probability measures, $\mathscr{P}$, is considered:

$$\mathcal{R}_{x,a}(v) := \sup_{Q_{x,a} \in \mathscr{P}} \mathbb{E}^Q_{x,a}[v] \tag{3}$$

Obviously, $\mathcal{R}$ is coherent and therefore risk-averse, which coincides with the intuition that the worst scenario is considered. One special case of the robust dynamic programming was the *minimax control* (see, e.g., [9]), which considers the worst possible scenario: $\mathcal{R}_{x,a}(v) := Q_{x,a}$-esssup v. In fact, each coherent risk measure can be represented by the form (3) under some regularity conditions for the set $\mathscr{P}$ (see e.g. [10] for essentially bounded spaces and [41] for unbounded ones).

Mean-semideviation trade-off ([30, 35]) considers the trade-off between the one-step conditional mean and semideviation (rather than the deviation of the whole Markov chain [40, 13]),

$$\mathcal{R}_{x,a}(v) := \mathbb{E}^Q_{x,a}[v] + \lambda\left[\mathbb{E}^Q_{x,a}(v - \mathbb{E}^Q_{x,a}[v])^r_+\right]^{1/r} \tag{4}$$

where $r \geq 1$ and $\lambda \in (-1, 1)$ denotes the risk-preference parameter which controls the risk preference of R: if $\lambda > 0$, R is risk-averse; if $\lambda < 0$, R is risk-seeking. Setting

$r = 2$, this map can be viewed as an approximation of the mean-variance trade-off scheme defined in [13].

Choquet integral [7]. To fit the MCP-framework, we first extend *non-additive probability measures* (p.m.) defined in [11]. Φ is said to be a conditional non-additive p.m., if (i) for each $(x, a) \in \mathbf{K}$, $\Phi_{x,a}(\cdot)$ is a non-additive p.m., and (ii) for each $B \in \mathcal{B}(\mathbf{X})$, $\Phi_\cdot(B)$ is $\mathcal{B}(\mathbf{K})$-measurable. Then, given a conditional non-additive p.m. Φ, the Choquet integral equipped with an MCP is defined as

$$\mathcal{R}_{x,a}(v) := \int_{-\infty}^{0} [\Phi_{x,a}(v > t) - 1]dt + \int_{0}^{\infty} \Phi_{x,a}(v > t)dt.$$

It is easy to verify that $\mathcal{R}$ is a homogeneous risk map but not necessarily convex. The well-known *prospect theory* [28] in behavioral economics, which is applied to interpret human behaviors of mixed risk-preferences, can be in fact represented as a Choquet integral [42]. A recent study [39] shows that the key features of the prospect theory can be also captured by the utility-based shortfall with non-convex utility functions.

4 Optimal Risk-sensitive Control

Given a risk map $\mathcal{R}$ and $\boldsymbol{\pi} \in \Pi_M$, we consider the following infinite-horizon risk-sensitive objectives: 1) the *discounted risk*:

$$J_\alpha(x, \boldsymbol{\pi}) := \lim_{T\to\infty} J_{\alpha,T}(x, \boldsymbol{\pi}), \tag{5}$$

$$\begin{aligned} \text{where } J_{\alpha,T}(x, \boldsymbol{\pi}) :=& c^{\pi_0}(x) + \alpha\mathcal{R}_x^{\pi_0}(c^{\pi_1} + \alpha\mathcal{R}^{\pi_1}(c^{\pi_2} + \\ & \ldots + \alpha\mathcal{R}^{\pi_{T-1}}(c^{\pi_T})\ldots)), \end{aligned} \tag{6}$$

and 2) the *average risk*:

$$J(x, \boldsymbol{\pi}) := \limsup_{T\to\infty} \frac{1}{T} J_{1,T}(x, \boldsymbol{\pi}), \boldsymbol{\pi} \in \Pi_M, x \in \mathbf{X}. \tag{7}$$

The optimal control problems for above three risk-sensitive objectives are to minimize the risk among all Markov policies

$$J_\alpha^*(x) := \inf_{\boldsymbol{\pi}\in\Pi_M} J_\alpha(x, \boldsymbol{\pi}), \text{and } J^*(x) := \inf_{\boldsymbol{\pi}\in\Pi_M} J(x, \boldsymbol{\pi}).$$

Remarks on the definition of discounted risk. In economics, the time-discount is added to reflect the "time-value" of outcomes: the outcome to be gained in the future is less valuable than the same amount of outcome obtained now. It has similar effects when cost is concerned. Due to its good mathematical properties, exponential discounting scheme, where the cost c_t is multiplied with the time-discount α^t, is widely applied in economics, finance as well as in MCPs. A natural extension of classical discounted MCPs is

$$\begin{aligned} D_\alpha(\boldsymbol{\pi}) =& c^{\pi_0} + \mathcal{R}^{\pi_0}(\alpha c^{\pi_1} + \mathcal{R}^{\pi_1}(\alpha^2 c^{\pi_2} + \\ & \ldots + \mathcal{R}^{\pi_{T-1}}(\alpha^T c^{\pi_T} + \ldots)\ldots)). \end{aligned}$$

However, since the risk map $\mathcal{R}$ is not necessarily homogeneous, a stationary policy that optimizes D_α need not exist. Indeed, it was proved in [8] that for the entropic risk map, which is not homogeneous, the optimal policy might not be stationary if $D_\alpha(\boldsymbol{\pi})$ is optimized w.r.t. $\boldsymbol{\pi}$, though D_α is well-defined for all $\alpha \in [0, 1)$. In our definition, discount factor α is multiplied with $\mathcal{R}$, which has the same "time-discount" effect, where the risk rather than the immediate cost is discounted. Moreover, it is easy to see that, if $\mathcal{R}$ is homogeneous, D_α is equivalent to J_α, the discounted total risk under our definition. Therefore, D_α defined for any homogeneous risk map is merely a special case of our definition. Specifically, the classical discounted MCP is indeed a special case of our defined discounted total risk, since it is homogeneous. D_α was used in [33] and the corresponding optimization problem was solved by a value iteration algorithm, since merely the coherent risk maps were considered. Besides, in the proof of the value iteration algorithm, the representation theorem was used, which is valid merely for coherent risk maps. On the contrary, we will see later that the objective J_α allows a value iteration algorithm for general risk maps. Therefore, we apply J_α rather than D_α.

Define the following operators

$$\mathcal{F}_\alpha^\pi(v) := c^\pi + \alpha \mathcal{R}^\pi(v),\ \mathcal{F}_\alpha(v) := \inf_{\boldsymbol{\pi} \in \Pi_M} \mathcal{F}_\alpha^\pi(v),$$

where $v \in \mathscr{B}_w$ and $\alpha \in [0, 1]$. If $\alpha = 1$, we simply write them as $\mathcal{F}^\pi$ and $\mathcal{F}$ respectively. The operators $(\mathcal{F}_\alpha^\pi)^n$, $n \in \mathbb{N}$, are defined iteratively as $(\mathcal{F}_\alpha^\pi)^0(v) := v$, and $(\mathcal{F}_\alpha^\pi)^n(v) := \mathcal{F}_\alpha^\pi((\mathcal{F}_\alpha^\pi)^{n-1}(v))$, $n = 1, 2, \ldots$, while $\mathcal{F}_\alpha^t$ is defined analogously.

The following assumption is made to ensure the existence of the "selector" in the optimization problem.

Assumption 4.1. *For each $x \in \mathbf{X}$, (i) the cost function $c(x, a)$ is lower semi-continuous on $\mathbf{A}(x)$, (ii) the action space $\mathbf{A}(x)$ is compact, and (iii) the function $u'(x, a) := \mathcal{R}_{x,a}(u)$ is continuous in $a \in \mathbf{A}(x)$ for any $u \in \mathscr{B}_w$.*

In the rest of this section, results are stated without proof. We refer the reader to [38] for detailed proofs, except the part of discounted risk, which is referred to [37].

Upper envelope. Before solving the optimization problems by an iterative approach, we introduce the concept of upper envelope to control the growth of iterations, $\{\mathcal{F}_\alpha^n\}$.

Definition 4.2. *A coherent risk map $\bar{\mathcal{R}}^{(w,C)}$ is said to be an* upper envelope *of a risk map $\mathcal{R}$ given a constant $C > 0$, if for all $v, u \in \mathscr{B}_w^{(C)} := \{v \in \mathscr{B}_w | \|v\|_{s,w} \leq C\}$,*

$$\mathcal{R}_{x,a}(v) - \mathcal{R}_{x,a}(u) \leq \bar{\mathcal{R}}_{x,a}^{(w,C)}(v - u), \forall (x, a) \in \mathbf{X}.$$

Remark. Apparently, if $\mathcal{R}$ is coherent, then $\mathcal{R}$ is an upper envelope of itself for any $C > 0$, due to its sublinearity (for proof see, e.g., [10]).

To ensure the existence of the upper bound C, we assume,

Assumption 4.3. *There exist a $\mathcal{B}(\mathbf{X})$-measurable function $w_0 : \mathbf{X} \to [0, \infty)$, constants $\gamma_0 \in (0, 1)$ and $\tilde{K}_0 > K_0 > 0$ such that (i) $[c(x, a) + \mathcal{R}_{x,a}(w_0)] \vee [-c(x, a) - \mathcal{R}_{x,a}(-w_0)] \leq \gamma_0 w_0(x) + K_0, \forall (x, a) \in \mathbf{K}$; (ii) for all $x, x' \in \mathsf{B}_0 := \{x \in \mathbf{X} | w_0(x) \leq$*

$R_0 := \frac{2K_0}{1-\gamma_0}\}$, $a \in \mathbf{A}(x)$, $a' \in \mathbf{A}(x')$, *the inequality* $\mathcal{R}_{x,a}(v) - \mathcal{R}_{x',a'}(v) \le 2(\tilde{K}_0 - K_0) + \mathcal{R}_{x,a}(w_0) - \mathcal{R}_{x',a'}(-w_0)$ *holds for all* v *satisfying* $|v| \le w_0 + \tilde{K}_0$.

and obtain the following theorem, which gives us a *bounded forward invariant subset*.

THEOREM 4.4. *Suppose Assumption 4.3 holds. Let* $w := 1 + \tilde{K}_0^{-1} w_0$. *Then for all* $\pi \in \Delta$,

$$\|\mathcal{F}^\pi(v)\|_{s,w} \le \tilde{K}_0, \text{ whenever } \|v\|_{s,w} \le \tilde{K}_0.$$

Discounted risk. Under Assumption 4.3, we can restrict to the bounded forward invariant subset $\mathcal{B}_w^{(\tilde{K}_0)}$. We first show that $\mathcal{F}_\alpha$, $\alpha \in [0,1)$, is a contraction map under the w-norm.

LEMMA 4.5. *Suppose Assumption 4.1 and 4.3 hold. Let* $w := 1 + \tilde{K}_0^{-1} w_0$. *Assume further that there exists a constant* $\bar{w} \in [1, 1/\alpha)$ *such that* $\sup_{a \in \mathbf{A}(x)} \bar{\mathcal{R}}_{x,a}^{(w,\tilde{K}_0)}(w) \le \bar{w} w(x)$. *Then* $\|\mathcal{F}_\alpha(v) - \mathcal{F}_\alpha(u)\|_w \le \bar{w}\alpha \|v - u\|_w$, $\bar{w}\alpha < 1$.

Hence, by Banach's fixed point theorem, starting from some $v \in \mathscr{B}_w$ satisfying $\|v\|_{s,w} \le \tilde{K}_0$, $\mathcal{F}_\alpha^n(v)$ converges to a unique fixed point v^* in $\mathscr{B}_w$ satisfying the *Bellman equation*:

$$v^*(x) = \mathcal{F}_\alpha(v^*|x) = \inf_{a \in \mathbf{A}(x)} \{c(x,a) + \alpha \mathcal{R}(v^*|x,a)\}.$$

Let f be an optimal selector in the right hand side of the above equation. The following theorem indicates the link between the Bellman equation and the optimization problem of discounted risk.

THEOREM 4.6. *Suppose Assumption 4.1 and 4.3 hold. Then* $v^*(x) = J_\alpha^*(x) = J_\alpha(x, f^\infty)$ *for all* $x \in \mathbf{X}$.

Remark. If $\mathcal{R}$ is coherent, then $\bar{\mathcal{R}}^{(w,\tilde{K}_0)} = \mathcal{R}$, and Assumption (i) and (ii) are no longer needed in Lemma 4.5 and Theorem 4.6.

Average risk. We now deal with the average risk based on the following assumption.

Assumption 4.7. *Let* $w_0 : \mathbf{X} \to [0, \infty)$ *and* $w : \mathbf{X} \to [1, \infty)$ *be two real-valued nonnegative* $\mathcal{B}(\mathbf{X})$*-measurable functions satisfying (i)* $\mathscr{B}_{1+w_0} = \mathscr{B}_w$, *(ii) there exist constants* $\gamma \in (0,1)$, $K > 0$ *and an upper envelope* $\bar{\mathcal{R}}^{(w,\tilde{K}_0)}$ *such that* $\bar{\mathcal{R}}_{x,a}^{(w,\tilde{K}_0)}(w_0) \le \gamma w_0(x) + K, \forall (x,a) \in \mathbf{K}$; *and (iii) there exist a constant* $\alpha \in (0,1)$ *and a probability measure* μ *such that for all* $x, x' \in \mathsf{B} := \{x \in \mathbf{X} | w_0(x) \le R, R > \frac{2K}{1-\gamma}\}$, $a \in \mathbf{A}(x)$, $a' \in \mathbf{A}(x')$, *and* $v \ge u \in \mathscr{B}_{1+w_0}$,

$$\bar{\mathcal{R}}_{x,a}^{(w,\tilde{K}_0)}(v) - \bar{\mathcal{R}}_{x,a}^{(w,\tilde{K}_0)}(u) \ge \alpha \int (v(x) - u(x))\, \mu(dx).$$

The following lemma shows that $\mathcal{F}^n \to 0$ under the $(1 + \beta w_0)$-seminorm, as $n \to \infty$.

LEMMA 4.8. *Suppose Assumption 4.1 and 4.3 hold. Assume further that Assumption 4.7 holds with the same w_0 as in Assumption 4.3. Then there exists $\bar{\alpha} \in (0,1)$ and $\beta > 0$ such that for all $v, u \in \mathscr{B}_w^{(\tilde{K}_0)}$, $\|\mathcal{F}(v) - \mathcal{F}(u)\|_{s,1+\beta w_0} \leq \bar{\alpha}\|v-u\|_{s,1+\beta w_0}$.*

Finally, we show the existence of a solution to the *Poisson equation* and its link to the the optimization problem of average risk.

THEOREM 4.9. *Under the same assumption as in Lemma 4.8. Then the following* Poisson equation

$$\rho^* + h(x) = \mathcal{F}_x(h) = \inf_{a \in \mathbf{A}(x)} \{c(x,a) + \mathcal{R}_{x,a}(h)\}$$

has a solution $(\rho^*, h) \in \mathbb{R} \times \mathscr{B}_w$, where ρ^* is unique. Furthermore, $\rho^* = J^*(x) = J(x, f^\infty)$ for all $x \in \mathbf{X}$, where f is an optimal selector in the right hand side of the above equation.

Remark 4.10. If $\mathcal{R}$ is coherent, then $\mathcal{R}$ itself is an upper envelope $\bar{\mathcal{R}}^{(w,C)}$ for any $C > 0$. In this case, Assumption 4.3 is no longer needed in Lemma 4.8 and Theorem 4.9 to determine *a priori* the size of the bounded forward invariant subset, C. For instance, applying the classical MCP, $\bar{\mathcal{R}}_{x,a}^{(w,\tilde{K}_0)}(v) = \mathbb{E}_{x,a}^Q(v)$, and obviously Assumption 4.7(iii) is equivalent to the classical Doeblin's condition. Hence, Assumption 4.7 becomes the classical condition that has been widely used in the MCPs literature (see [25, 24, 43] and references therein) for studying the average cost.

Entropic map. As a special case, applying the entropic map, we state in the next theorem sufficient conditions for Assumption 4.3 and 4.7.

THEOREM 4.11. *Let $\mathcal{R}$ be the entropic map defined in* (2) *with $\lambda = 1$. Suppose the following conditions hold: (i) there exist a function $w_1 : \mathbf{X} \in [1, \infty)$, constants $\gamma_1 \in (0,1)$ and $K_1 > 0$ such that*

$$\mathcal{R}_{x,a}(w_1) \leq \gamma_1 w_1(x) + K_1, \forall (x,a) \in \mathbf{K},$$

(ii) for all $p \in (0,1)$ and all level-sets $\mathsf{C} := \mathscr{B}_{w_1^p}(R)$, $R > 0$, there exist a measure μ_{C} and constants $\lambda_{\mathsf{C}}^+ > \lambda_{\mathsf{C}}^- > 0$ such that $\mu_{\mathsf{C}}(\mathsf{C}) > 0$ and $\forall x \in \mathsf{C}, a \in \mathbf{A}(x)$ and $\mathsf{A} \in \mathcal{B}(\mathbf{X})$,

$$\lambda_{\mathsf{C}}^- \mu_{\mathsf{C}}(\mathsf{A} \cap \mathsf{C}) \leq Q_{x,a}(\mathsf{A} \cap \mathsf{C}) \leq \lambda_{\mathsf{C}}^+ \mu_{\mathsf{C}}(\mathsf{A} \cap \mathsf{C}),$$

and (iii) the cost function c satisfies

$$\bar{c}(x) := \sup_{a \in \mathbf{A}(x)} |c(x,a)| \in \mathscr{B}_{w_1^q} \text{ for some } q \in (0,1).$$

Then Assumption 4.3 holds with $w_0 = w_1^p$ for any $p \in (q,1)$ and some constant $\tilde{K}_0$, and Assumption 4.7 holds with $w = 1 + \tilde{K}_0^{-1} w_0$.

We compare the above conditions with the conditions employed in the most related literature [12], which has the same general settings as this paper, i.e., unbounded costs on general Borel spaces.

a) The assumption (A4) [12, Section 4], which requires the existence of a positive continuous density satisfying $Q(dy|x,a) = \int q(x,a,y)\mu(dy)$ for some reference probability measure μ, in fact, implies the local Doeblin's condition (ii) in the above theorem. Hence, this assumption is more general than the counterpart in [12].

b) The assumption (A3) [12, Section 3] set for the cost function c is implicit and difficult to be verified. On the contrary, the sufficient growth condition for c, is explicit in form of the Lyapunov function w_1 w.r.t. the entropic map. Note that, in the example provided by [12], the assumption (A3) is also verified with the help of a Lyapunov function.

c) As an advantage, in comparison with [12], the convergence rate of iterations towards the solution to the Poisson equation is explicitly specified by $\bar{\alpha}$ in Lemma 4.8.

Examples. We present two examples with the average risk. The first risk map is coherent while the other one is the most widely used convex risk map: the entropic map.

1) Mean-semideviation. Let $\mathcal{R}$ be the mean-semideviation defined in (4) with $r = 2$ and $\lambda \in (0,1)$. Consider a 1-dimensional linear model $x_{n+1} = b(a_t)x_t + w_t$, where w_t is i.i.d. white noise and b is a real-valued $\mathcal{B}(\mathbf{A})$-measurable function satisfying $\sup_{a\in\mathbf{A}}|b(a)| = \epsilon < 1$.

Since $\mathcal{R}$ is coherent, it remains to check Assumption 4.7(ii) and (iii). For (ii), it can be shown [37] that $w_0(x) = x^2$ is the required weight function with $\gamma \in (\epsilon^2, 1)$ and some constant $K > 0$. For (iii), it can be shown [37] that

$$\mathcal{R}_{x,a}(v) - \mathcal{R}_{x,a}(u) \geq (1-\lambda)\int_{\mathbf{X}}(v-u)(y)Q(dy|x,a)$$

holds for all $v \geq u \in \mathscr{B}_{1+w_0}$ and $(x,a) \in \mathbf{K}$. Hence, (iii) holds, since the transition kernel Q has a positive continuous density function w.r.t. the Lebesgue measure.

2) Entropic map. Let $\mathcal{R}$ be the entropic map defined in (2) with $\lambda = 1$. Let $\mathbf{X} = \mathbb{R}^d$. Consider the following discretized ergodic diffusion $\{x_n \in \mathbb{R}^d\}$ (cf. the example in [12]):

$$x_{n+1} = Ax_n + b(x_n, a_n) + D(x_n, a_n)w_n,$$

where $\{w_n \in \mathbb{R}^d\}$ is a sequence of i.i.d. standard white noise, $D : \mathbb{K} \to \mathbb{R}^{d\times d}$ is a continuous bounded matrix-valued function which is uniformly elliptic, i.e., there exist constants $l > 0$ and $L > 0$ such that for all $(x,a) \in \mathbf{K}, \xi \in \mathbb{R}^d$,

$$l\|\xi\|^2 := l\xi^\top\xi \leq \xi^\top D(x,a)D^\top(x,a)\xi \leq L\|\xi\|^2, \tag{8}$$

and $b : \mathbf{K} \to \mathbb{R}^d$ is a continuous bounded vector function, and A is a matrix satisfying that there exists a constant $\tilde{\gamma} \in (0,1)$ such that $\xi^\top A^\top A\xi \leq \tilde{\gamma}\|\xi\|^2$, $\forall \xi \in \mathbb{R}^d$.

We verify now the conditions (i) and (ii) stated in Theorem 4.11. Take one $\gamma \in (\tilde{\gamma}, 1)$ and consider the following function

$$\hat{w}_1(x) = \frac{\epsilon}{2}\|x\|^2, \text{ with some positive } \epsilon \leq \frac{\gamma - \tilde{\gamma}}{\gamma}L^{-1}.$$

It can be shown [38] that there exist $\gamma_1 \in (\gamma, 1)$ and $\hat{K}_1 > 0$ satisfying $R_{x,a}(\hat{w}_1) \leq \gamma_1 \hat{w}_1(x) + \hat{K}_1, \forall (x,a) \in \mathbf{K}$. Hence, the condition (i) in Theorem 4.11 holds with $w_1 := \hat{w}_1 + 1$, γ_1 and $K_1 := \hat{K}_1 + 1 - \gamma_1$. Next, since the transition kernel Q has a positive continuous density function w.r.t. the Lebesgue measure, condition (ii) in Theorem 4.11 is obviously satisfied.

References

[1] P. Artzner, F. Delbaen, J. Eber, and D. Heath, *Coherent measures of risk*, Mathematical Finance, 9 (1999), pp. 203–228.

[2] G. Avila-Godoy and E. Fernández-Gaucherand, *Controlled Markov chains with exponential risk-sensitive criteria: modularity, structured policies and applications*, in Proceedings of the 37th IEEE Conference on Decision and Control, 1998, pp. 778–783.

[3] V. Borkar and S. Meyn, *Risk-sensitive optimal control for Markov decision processes with monotone cost*, Mathematics of Operations Research, (2002), pp. 192–209.

[4] R. Cavazos-Cadena, *Optimality equations and inequalities in a class of risk-sensitive average cost Markov decision chains*, Mathematical Methods of Operations Research, 71 (2010), pp. 47–84.

[5] A. Chateauneuf and M. Cohen, *Cardinal extensions of the EU model based on the Choquet integral*, Decision-making Process, (2008), pp. 401–433.

[6] P. Cheridito and M. Kupper, *Composition of time-consistent dynamic monetary risk measures in discrete time*, International Journal of Theoretical and Applied Finance, 14 (2011), pp. 137–162.

[7] G. Choquet, *Theory of capacities*, in Annales de l'institut Fourier, vol. 5, 1953.

[8] K. Chung and M. Sobel, *Discounted MDPs: distribution functions and exponential utility maximization*, SIAM Journal on Control and Optimization, 25 (1987), p. 49.

[9] S. Coraluppi and S. Marcus, *Mixed risk-neutral/minimax control of discrete-time, finite-state Markov decision processes*, IEEE Transactions on Automatic Control, 45 (2000), pp. 528–532.

[10] F. Delbaen, *Coherent risk measures on general probability spaces*, Advances in Finance and Stochastics Essays in Honour of Dieter Sondermann, (2000), pp. 1–37.

[11] D. Denneberg, *Non-additive Measure and Integral*, Kluwer Academic Publishers, 1994.

[12] G. Di Masi and L. Stettner, *Infinite horizon risk sensitive control of discrete time Markov processes under minorization property*, SIAM Journal on Control and Optimization, 46 (2008), p. 231.

[13] J. Filar, L. Kallenberg, and H. Lee, *Variance-penalized Markov decision processes*, Mathematics of Operations Research, (1989), pp. 147–161.

[14] W. Fleming and D. Hernández-Hernández, *Risk sensitive control of finite state machines on an infinite horizon. I*, in Proceedings of the 36th IEEE Conference on Decision and Control, IEEE, 1997, pp. 3407–3412.

[15] H. Föllmer and I. Penner, *Convex risk measures and the dynamics of their penalty functions*, Statistics & Decisions, 24 (2006), pp. 61–96.

[16] H. Föllmer and A. Schied, *Convex measures of risk and trading constraints*, Finance and Stochastics, 6 (2002), pp. 429–447.

[17] ———, *Stochastic Finance*, Walter de Gruyter & Co., Berlin, 2004. Extended edition.

[18] S. Gaubert and J. Gunawardena, *The Perron-Frobenius theorem for homogeneous, monotone functions*, Transactions American Mathematical Society, 356 (2004), pp. 4931–4950.

[19] C. Gollier, *The Economics of Risk and Time*, The MIT Press, 2004.

[20] M. Hairer and J. Mattingly, *Yet another look at Harris' ergodic theorem for Markov chains*, in Seminar on Stochastic Analysis, Random Fields and Applications VI, Springer, 2011, pp. 109–117.

[21] D. Hernández-Hernández and S. Marcus, *Risk sensitive control of Markov processes in countable state space*, Systems & Control Letters, 29 (1996), pp. 147–155.

[22] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer, 1989.

[23] O. Hernández-Lerma and J. Lasserre, *Discrete-time Markov Control Processes: Basic Optimality Criteria*, Springer, 1996.

[24] ———, *Further Topics on Discrete-Time Markov Control Processes*, Springer Verlag, 1999.

[25] O. Hernández-Lerma, R. Montes-De-Oca, and R. Cavazos-Cadena, *Recurrence conditions for Markov decision processes with Borel state space: a survey*, Annals of Operations Research, 28 (1991), pp. 29–46.

[26] R. Howard and J. Matheson, *Risk-sensitive Markov decision processes*, Management Science, 18 (1972), pp. 356–369.

[27] G. Iyengar, *Robust dynamic programming*, Mathematics of Operations Research, (2005), pp. 257–280.

[28] D. Kahneman and A. Tversky, *Prospect theory: an analysis of decision under risk*, Econometrica, 47 (1979), pp. 263–292.

[29] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag London Ltd., London, 1993.

[30] W. Ogryczak and A. Ruszczyński, *From stochastic dominance to mean-risk models: Semideviations as risk measures*, European Journal of Operational Research, 116 (1999), pp. 33–50.

[31] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Inc., 1994.

[32] B. Roorda, J. Schumacher, and J. Engwerda, *Coherent acceptability measures in multiperiod models*, Mathematical Finance, 15 (2005), pp. 589–612.

[33] A. Ruszczyński, *Risk-averse dynamic programming for Markov decision processes*, Mathematical Programming, (2010), pp. 1–27.

[34] A. Ruszczyński and A. Shapiro, *Conditional risk mappings*, Mathematics of Operations Research, 31 (2006), pp. 544–561.

[35] ———, *Optimization of risk measures*, Probabilistic and Randomized Methods for Design under Uncertainty, (2006), pp. 119–157.

[36] L. Savage, *The Foundations of Statistics*, Dover Publications, 1972.

[37] Y. Shen, W. Stannat, and K. Obermayer, *Risk-sensitive Markov Control Processes*, SIAM Journal on Control and Optimization, (2013), pp. 3652–3672.

[38] ———, *Risk-sensitive Markov control processes with strictly convex risk maps*, Arxiv preprint arXiv:1403.3321, (2014).

[39] Y. Shen, M. Tobia, T. Sommer, and K. Obermayer, *Risk-sensitive reinforcement learning*, Neural Computation, 26 (2014), pp. 1298–1328.

[40] M. Sobel, *The variance of discounted Markov decision processes*, Journal of Applied Probability, (1982), pp. 794–802.

[41] G. Svindland, *Convex risk measures beyond bounded risks*, PhD thesis, Ludwig-Maximilians-Universität München, 2009.

[42] A. Tversky and D. Kahneman, *Advances in prospect theory: cumulative representation of uncertainty*, Journal of Risk and Uncertainty, 5 (1992), pp. 297–323.

[43] O. Vega-Amaya, *The average cost optimality equation: a fixed point approach*, Bol. Soc. Mat. Mexicana, 9 (2003), pp. 185–195.