Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Risk-sensitive partially observable Markov decision processes as fully observable multivariate utility optimization problems

Arsham Afsardeir, [1] Andreas Kapetanis,[2]

Vaios Laschos, [3] Klaus Obermayer [4]

submitted: December 2, 2022

[1] Fakultät Elektrotechnik und Informatik
Technische Universität Berlin
Marchstr. 23
10587 Berlin
Germany

[2] Institute of Mathematics
Technische Universität Berlin
Str. des 17. Juni 136
10623 Berlin
Germany

[3] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: vaios.laschos@wias-berlin.de

[4] Fakultät Elektrotechnik und Informatik
Technische Universität Berlin
Marchstr. 23
10587 Berlin
Germany

No. 2977

Berlin 2022

2020 *Mathematics Subject Classification.* 93E20.

Key words and phrases. Markov decision processes, partial observability, risk sensitivity, utility function, sums of exponentials.

Arsham Afsardeir was partially supported by DFG project OB 102/29-1.
For the biggest part, Vaios Laschos was supported by DFG project OB 102/27-1. For the completion of the work, Vaios Laschos was supported by DFG under Germany´s Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).
Klaus Obermayer was partially supported by DFG project OB 102/27-1.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Risk-sensitive partially observable Markov decision processes as fully observable multivariate utility optimization problems

Arsham Afsardeir, Andreas Kapetanis,

Vaios Laschos, Klaus Obermayer

Abstract

We provide a new algorithm for solving Risk Sensitive Partially Observable Markov Decisions Processes, when the risk is modeled by a utility function, and both the state space and the space of observations are finite. This algorithm is based on an observation that the change of measure and the subsequent introduction of the information space, which is used for exponential utility functions, can be actually extended for sums of exponentials if one introduces an extra vector parameter that tracks the "expected accumulated cost" that corresponds to each exponential. Since every increasing function can be approximated by sums of exponentials in finite intervals, the method can be essentially applied for any utility function, with its complexity depending on the number of exponentials.

1 Introduction

In the classical theory of Markov Decision Processes (MDPs), one deals with controlled Markovian stochastic processes (S_n) taking values on a Borel space $\mathcal{S}$. These processes are controlled via a series of actions (A_n), according to a policy π, that changes the underlying state transition probabilities $\mathrm{P}(S_{n+1}|S_n, A_n)$ of (S_n). The goal is to find a policy π that optimizes the expected value

$$\mathcal{I}_N(s_0, \pi) = \mathbb{E}^{\pi}_{s_0}\left[\sum_{n=0}^{N-1} C(S_n, A_n)\right],$$

where $C : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the cost function, $s_0 \in S_n$ is the initial state , and $N \in \mathbb{N}$. The inclusion of risk-sensitivity and partial observability are natural extensions to this standard model.

In classical MDPs, one makes the assumption that the controlled process (S_n) takes values on a set of states $\mathcal{S}$ which is always accessible to the controller. However, in several real-life applications the real state is not directly observable and only secondary information dependent on the state, can be observed. Partially Observable Markov Decision Processes (POMDPs) are a generalization of MDPs towards incomplete information about the current state. POMDPs extend the notion of MDPs by a set of observations $\mathcal{Y}$ and a set of conditional observation probabilities $Q(\cdot|s)$ given the "hidden" state $s \in \mathcal{S}$. $Q(y|s)$ namely represents the probability of observing y while being in state s. In **risk-neutral POMDPs**, one can introduce a new state space, called **belief (state) space** $\mathcal{X} = \mathcal{P}(\mathcal{S})$, the set of probability measures on $\mathcal{S}$, and a stochastic process (X_n) taking values in $\mathcal{X}$, such that $X_n(s)$ is the probability of S_n being equal to the "hidden" state $s \in \mathcal{S}$ at time n, conditioned on the accumulated observations and actions up to time n. One can treat this new process on the belief space like a Completely Observable Markov Decision Process (COMDP) on $\mathcal{X}$ with classical tools, retrieve optimal or ε-optimal polices (i.e. policies with expected value at most ε-far from the optimal value), and then apply them to the original problem. It is remarkable that, **due to the linearity** of the expectation operator, the belief state is a so-called *sufficient statistic*. Broadly speaking, a sufficient statistic carries adequate information for the controller to make an optimal choice at a specific point in time. It also allows to separate the present cost from the future cost through a Bellman-style equation. For an introduction to sufficient statistics, we refer to (Hinderer, 1970).

DOI 10.20347/WIAS.PREPRINT.2977 Berlin 2022

To introduce risk-sensitivity we will work with the classical theory of expected utility (Jaquette, 1973), where one tries to optimize

$$\mathcal{I}_N(s_0, \pi) = \mathbb{E}^{\pi}_{s_0}\left[U\left[\sum_{n=0}^{N-1} C(S_n, A_n)\right]\right], \tag{1.1}$$

for some increasing and continuous function $U : \mathbb{R} \to \mathbb{R}$. When risk is involved, extra or alternative information is necessary to make an optimal decision. In the case of expected utilities, additional information on the accumulated cost is necessary to make an optimal choice, even in the fully observable case (Bäuerle & Rieder, 2014; Marecki & Varakantham, 2010). An exception to that, is the exponential utility function $U = \exp$ which enjoys nice properties that allow to make optimal choices regardless of the past. Due to these properties, the exponential utility function generates a performance index that belongs to several models of risk at the same time, and it has been extensively studied in many different settings (Borkar & Meyn, 2002; Cavazos-Cadena, 2010; Chung & Sobel, 1987; Di Masi & Stettner, 2007; Dupuis, Laschos & Ramanan, 2019; Fleming & Hernández-Hernández, 1997; Hernández-Hernández & Marcus, 1996; Howard & Matheson, 1972; Levitt & Ben-Israel, 2001).

To our best knowledge, there is only partial progress (Baras & James, 1997; Bäuerle & Rieder, 2017; Bäuerle & Rieder, 2017; Cavazos-Cadena & Hernández-Hernández, 2005; Fan & Ruszczyński, 2018; Fernandez-Gaucherand & Marcus, 1997; Hernández-Hernández, 1999; Marecki & Varakantham, 2010) when it comes to combining risk-sensitivity and partial observability, and most articles study the specific case of the exponential utility function. As it was mentioned in the previous paragraph, this is due to the fact that the extra information of the accumulated cost is needed to make optimal choices. A common workaround to this problem is to assume that the controller is aware of the running cost through some mechanism. For example, the controller observes either the running cost directly (Marecki & Varakantham, 2010), or a part of the whole process that is responsible for the cost (Bäuerle & Rieder, 2017). Fan und Ruszczyński (2018) study cost functions that depend on both observable quantities and beliefs. In Bäuerle und Rieder (2017), a general approach for treating problems with risk measured by a utility function is introduced. It is shown there, that one can use probability measures on the product space $\mathcal{S} \times \mathbb{R}$ as state space. This way, the authors end up with an MDP on the state space $\mathcal{P}(\mathcal{S} \times \mathbb{R})$. As we will explain in the sequel, we take a different route that is computationally less demanding in several cases.

1.1 Our Contribution

It was shown in Bäuerle und Rieder (2017), that by using an exponential function as the utility function, the problem space would shrink dramatically. This proposition is incline to the concept of *information vector* which Cavazos-Cadena and Hernandez-Hernandez discussed (Cavazos-Cadena & Hernández-Hernández, 2005). Both papers show, in two different conceptual systems, that this property of exponential utility function which transforms summation of costs to the product of their utilities, can be exploited to provide sufficient statistics for decision making in a much smaller representation. Consequently the model with exponential utility function has a computational advantage, however, it losses its generality due to the much narrower range of utility functions it can accept.

To extend the ideas related to exponential utility models, in this work we employ the idea of multi-variate optimization and show that by applying exponentials on a finite set of different running costs, we can use the information space approach in a more general way. More specifically, our multi-variate exponential utility model is able to reproduce more complex utility functions and at the same time benefit from simplicity of exponential utilities according to computation burden as well. In our method the exponential running costs are independent from each other, therefore, each term can represent an independent component of a target multivariate utility function. These components can be seen either as building blocks of a target function's formulation (in the case of utility functions equal to sum of exponential terms) or more generally as the elements of function approximator series (in case of approximating the target function).

In comparison to the purely exponential case, the utility functions we treat can be more behaviorally plausible as well. In behavioral economics and finance disciplines it is a common approach to assume people to use a mapping from objective values to subjective utilities and subsequently either apply maximization on the *expected value* of the utilities (Von Neumann & Morgenstern, 1947) or their other distributional properties (Tversky & Kahneman, 1992) (Al-Nowaihi, Bradley & Dhami, 2008). Consequently, the shape of people's utility function has a significant effect on their attitude toward risky choices. The shape of human utility function (Kalyanaram & Winer, 1995; Mosteller & Nogee, 1951) and the effect of contextual parameters on that, like amount of wealth (Markowitz, 1952) and emotions (Bertram, Schulz & Nelson, 2021) has been investigated by different experimental paradigms. For a review see Edwards (1954). In their influential work, *prospect theory*, Kahneman and Tversky proposed an S-shaped utility function which is risk-averse (concave) in gains and risk-seeker(convex) in losses (Kahneman & Tversky, 1979, 2013). They also addressed a set of experiments that confirm different risk tendencies in gain and loss situations. As one can expect, the flexibility of an exponential utility model is not sufficient to produce different risk tendencies between loosing and winning situations by a single function. In contrast, privileging the computational advantage of exponential utility functions, our risk-sensitive model can also address this phenomenon by exploiting either utility functions which are defined by linear combinations of multiple exponential forms (like *sinh(.)*) or utility functions which can be approximated by linear combinations of exponential terms in a specific interval of values (like Sigmoid function, see section 4). Therefore, it becomes possible to shape functions which have both positive and negative second order derivatives on the distinct intervals of their domain simultaneously. Capturing the dissimilarity of risk tendencies among losses and gains is a major advantage of the multi-variate model which gives us more explanatory ability in respect to behavioral modeling in comparison to exponential utility function.

In what follows we argue that it is actually possible to apply similar arguments to treat utility functions that are sums of exponentials, i.e. utility functions of the form

$$\widehat{U}(t) = \sum_{i=1}^{i_{\max}} w^i e^{\lambda^i t}. \tag{1.2}$$

With slight abuse of notation, we observe that for a probability distribution θ_0 on $\mathcal{S}$, we can write

$$\widehat{\mathcal{I}}_N(\theta_0, \widehat{\pi}) = \widehat{\mathbb{E}}^{\widehat{\pi}}_{\theta_0}\left[\sum_{i=1}^{i_{\max}} w^i e^{\lambda^i\left(\sum_{n=0}^{N-1}\widehat{C}(S_n,A_n)\right)}\right] = \sum_{i=1}^{i_{\max}} \widehat{\mathbb{E}}^{\widehat{\pi}}_{\theta_0}\left[w^i e^{\lambda^i\left(\sum_{n=0}^{N-1}\widehat{C}(S_n,A_n)\right)}\right].$$

Then, similar to Cavazos-Cadena und Hernández-Hernández (2005), by applying a change of measure argument, we identify a new state space $\mathcal{X} = \mathcal{P}(\mathcal{S})^{i_{\max}} \times \mathcal{Y}$, a controlled transition matrix $P(x'|x,a)$, and a collection of running costs $C^i : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$, that depend on the next stage as well, such that for the resulting completely observable controlled processes (X_n). With $x_0 = (\theta_0, \dots, \theta_0, y_0)$ and a fixed but arbitrary $y_0 \in \mathcal{Y}$, we have

$$\widehat{\mathcal{I}}_N(\theta_0, \widehat{\pi}) = \mathcal{I}_N(x_0, \pi) = \mathbb{E}^{\pi}_{x_0}\left[\sum_{i=1}^{i_{\max}} w^i e^{\lambda^i\left(\sum_{n=0}^{N-1} C^i(X_n,A_n,X_{n+1})\right)}\right]. \tag{1.3}$$

Note that $(\pi, \mathbb{E})$ and $(\widehat{\pi}, \widehat{\mathbb{E}})$ are connected in a straightforward manner, see Section 2.

To establish our method we exploit the results of Bäuerle und Rieder (2014) and extend that model to introduce multivariate utility functions $\mathcal{U} : \mathbb{R}^d \to \mathbb{R}$, which are component-wise monotone and each variable corresponds to a different running cost. The resulting optimality metric is of the form:

$$\mathcal{I}_N(x, \pi) := \mathbb{E}^{\pi}_x\left[\mathcal{U}\left(\sum_{n=0}^{N-1} \boldsymbol{C}(X_n, A_n, X_{n+1})\right)\right]. \tag{1.4}$$

Note that here C is a vector of different running costs. Criteria like (1.4) can arise when one is trying to solve a multi-objective task, with different running costs, each of the costs contributing in a different manner to a total

cost. As an example, one can think of a policy maker allocating tax money to different public sectors (education, infrastructure, health, etc). For each of them, we get a different cost which can be the position in the global chart or any other comparison metric. However, the total utility, depends on how much each government prioritizes each of these aspects, something that is encoded in the choice of the utility function. In a similar manner to Bäuerle und Rieder (2014), we augment the space to keep track of each accumulated cost term. To provide a more general treatment of the multivariate case, in the lines of Bäuerle und Rieder (2014), we treat the discounted case, i.e.

$$\mathcal{I}_N(x,\pi) := \mathbb{E}_x^\pi \left[\mathcal{U}\left(\sum_{n=0}^{N-1} \boldsymbol{\beta}^n \cdot \boldsymbol{C}(X_n, A_n, X_{n+1}) \right) \right], \tag{1.5}$$

where the dot product denotes the point-wise multiplication. Furthermore we prove that the finite time discounted problem converges to the infinite time one, without the extra assumption demanding that the utility function is either convex or concave, appearing in Bäuerle und Rieder (2014). This also enables the consideration of even more general problems of the form

$$\mathcal{I}_\infty(x,\pi) := \mathbb{E}_x^\pi \left[\mathcal{U}\left(\sum_{n=0}^{\infty} \boldsymbol{\beta}^n \cdot \boldsymbol{C}(X_n, A_n, X_{n+1}) \right) \right]. \tag{1.6}$$

Remark 1.1. *For any two utility functions $\widehat{U}_1, \widehat{U}_2$ that are ε-close on some interval $[N \min_{s,a} \widehat{C}(s,a), N \max_{s,a} \widehat{C}(s,a)]$, an ε-optimal policy for $\widehat{U}_1$ is a 2ε-optimal policy for $\widehat{U}_2$. Therefore, one can apply the method to solve RSPOMDPs with utility functions that can be approximated by functions of the form* (1.2)*. One can easily show that this includes all increasing real functions, by approximating $F(t) = U(log(t))$, by a polynomial $P(t) = \sum_{i=1}^k w_k x^k$ in the interval $[\exp(N \min_{s,a} \widehat{C}(s,a)), \exp(N \max_{s,u} \widehat{C}(s,a))]$, and the composing with the exponential on both sides.*

The rest of the paper is structured as follows: In section 2 our multi-variate utility function and its mathematical modeling and construction are presented. Next, in section 3 we show and prove solution methods for this model and the utility functions from section 2 in both finite and infinite horizon cases. In section 4, we provide an extended version of a famous POMDP, the tiger problem, as a numerical example to explain the model and compare it with the general model of Bäuerle and Rieder. And finally, we discuss about the computational advantage of our model.

2 RSPOMDPs for sums of exponentials

In this section, we consider risk sensitve POMDPs with a class of utility functions that can be written as weighted sums of exponentials. We will show that it is possible to reformulate the problem in terms of a multi objective risk sensitive MDP with a new performance index that, in turn, can be treated with tools described in the next section.

2.1 The original setting

We start by describing the model for a risk sensitive POMDP, i.e. $(\mathcal{S}, \mathcal{Y}, \mathcal{A}, \widehat{P}, Q, \widehat{C}, \widehat{U})$. $\mathcal{S}, \mathcal{Y}, \mathcal{A}$ will be three **finite** sets equipped with the discrete topology. In the sequel, $\mathcal{S}$ is the *hidden* state space, $\mathcal{Y}$ the set of observations, and $\mathcal{A}$ the set of controls. For every $a \in \mathcal{A}$, we define a transition probability matrix $\widehat{P}(a) = \left[\widehat{P}(s'|s;a)\right]_{s,s' \in S}$. Finally, we denote by $Q = [Q(y|s)]_{y \in \mathcal{Y}, s \in \mathcal{S}}$, with $Q(y|s) > 0, \forall y \in \mathcal{Y}, s \in \mathcal{S}$, the signal matrix and by $\widehat{C} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the cost function.

Now, for each $n \in \mathbb{N}$, let $\widehat{\mathcal{H}}_n$ be the set of histories up to time n, where $\widehat{\mathcal{H}}_0 = \mathcal{P}(S)$, and $\widehat{\mathcal{H}}_n = \widehat{\mathcal{H}}_{n-1} \times \mathcal{A} \times \mathcal{Y}$. We denote by $\mathbf{\Pi}_{\widehat{\mathcal{H}}} := \left\{ \widehat{\pi} = \left(\widehat{f}_0, \widehat{f}_1, \dots \right) \;\middle|\; \widehat{f}_n : \widehat{\mathcal{H}}_n \to A, n \in \mathbb{N} \right\}$ the set of deterministic polices that are functions of the history $\widehat{h}_n = (\theta, a_0, y_1, \dots, a_{n-1}, y_n)$ up to time n. Given $\theta \in \mathcal{P}(S)$, and $\widehat{\pi} \in \mathbf{\Pi}_{\widehat{\mathcal{H}}}$, due to the Ionescu-Tulcea theorem, there exists a unique measure $\widehat{\mathbb{P}}_\theta^{\widehat{\pi}}$ on the Borel sets of $\Omega := \mathcal{S} \times (\mathcal{A} \times \mathcal{S} \times \mathcal{Y})^\infty$ that satisfies:

$$\widehat{\mathbb{P}}_\theta^{\widehat{\pi}}(s_0, a_0, s_1, y_1, a_1, \dots, a_{n-1}, s_n, y_n) := \theta(s_0) \prod_{k=0}^{n-1} \left(\widehat{P}\left(s_{k+1} | s_k; \widehat{f}_k\left(\widehat{h}_k \right) \right) Q\left(y_{k+1} | s_{k+1} \right) \right),$$

The corresponding expectation operator is denoted by $\widehat{\mathbb{E}}_\theta^{\widehat{\pi}}$. Finally, for each $n \in \mathbb{N}$, we define the σ-fields $\widehat{\mathcal{F}}_n$, $\widehat{\mathcal{G}}_n$, by

$$\widehat{\mathcal{F}}_n := \sigma\left((A_k, Y_{k+1}),\ k = 0, 1, \dots, n-1 \right), \quad \widehat{\mathcal{G}}_n := \sigma\left(S_0, (A_k, S_{k+1}, Y_{k+1}),\ k = 0, 1, \dots, n-1 \right).$$

It is straightforward to see that the set of policies $\mathbf{\Pi}_{\widehat{\mathcal{H}}}$, contains exactly the elements $\left(\widehat{f}_n\right)_{n \in \mathbb{N}}$, where $\widehat{f}_n$ are $\widehat{\mathcal{F}}_n$-measurable functions from $\widehat{\mathcal{H}}_n$ to $\mathcal{A}$.

2.2 Utility functions that are sums of exponentials

Let $\{\lambda^i, i = 1, \dots, i_{\max}\} \subseteq \mathbb{R} \setminus \{0\}$ be a finite collection of risk parameters, and $\{w^i, i = 1, \dots, i_{\max}\} \subseteq \mathbb{R}$ be a collection of weights. We define the utility function $\widehat{U} : \mathbb{R} \to \mathbb{R}$ by

$$\widehat{U}(t) := \sum_{i=1}^{i_{\max}} w^i e^{\lambda^i t}, \tag{2.1}$$

and introduce the performance index

$$\widehat{\mathcal{I}}_N(\theta_0, \widehat{\pi}) = \sum_{i=1}^{i_{\max}} w^i \widehat{\mathbb{E}}_{\theta_0}^{\widehat{\pi}} \left[e^{\lambda^i \left[\sum_{n=0}^{N-1} \widehat{C}(S_n, A_n) \right]} \right], \tag{2.2}$$

and the corresponding value fuction

$$\widehat{V}_N(\theta_0) := \inf_{\widehat{\pi} \in \mathbf{\Pi}_{\widehat{\mathcal{H}}}} \widehat{\mathcal{I}}_N(\theta_0, \widehat{\pi}). \tag{$\widehat{P}$}$$

The goal is to minimize $\widehat{\mathcal{I}}_N(\theta_0, \widehat{\pi})$ over all policies $\widehat{\pi} \in \mathbf{\Pi}_{\widehat{\mathcal{H}}}$. We want to show that we can instead work on the completely observable risk sensitive MDP on the space $\mathcal{X}$ with performance index

$$\mathcal{I}_N(x_0, \pi) = \sum_{i=1}^{i_{\max}} w^i \mathbb{E}_{x_0}^{\pi} \left[e^{\lambda^i \left[\sum_{n=0}^{N-1} C^i(X_n, A_n, X_{n+1}) \right]} \right], \tag{2.3}$$

for some reconstructed cost functions C^i ,measure and expectation operator $\mathbb{P}_\theta^\pi$, $\mathbb{E}_\theta^\pi$ and corresponding value function

$$\mathcal{V}_N(x) := \inf_{\pi \in \mathbf{\Pi}_{\mathcal{H}}} \mathcal{I}_N(x, \pi), \tag{P}$$

for set of histories $\mathcal{H}$. The following subsection will establish these constructions and prove the claim. Now problem (P) falls in the framework of Section 3.1 that provides the means to calculate the optimal value and optimal policies as presented in the next chapter.

2.3 Towards a completely observable problem

The aim of this subsection is to prove the following:

Theorem 2.1. *Let a risk sensitive POMDP $(\mathcal{S}, \mathcal{Y}, \mathcal{A}, \widehat{P}, Q, \widehat{C}, \widehat{U})$, with stochastic dynamics as in subsection 2.1 and performance index given in* (2.2)*. Then, there exist operators $G^i : \mathcal{P}(\mathcal{S}) \times \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, $F^i : \mathcal{P}(\mathcal{S}) \times A \times Y \to \mathcal{P}(\mathcal{S})$, $i \in \{1, \dots, i_{\max}\}$, and $\eta : \mathbf{\Pi}_{\widehat{\mathcal{H}}} \to \mathbf{\Pi}_{\mathcal{H}}$, such that the following Completely Observable MDP X_n with performance index $\mathcal{I}_N(x_0, \pi)$ is equivalent the original, i.e.*

$$\mathcal{I}_N(x_0, \eta(\widehat{\pi})) = \widehat{\mathcal{I}_N}(\theta_0, \widehat{\pi}).$$

The completely observable MDP is defined by the following:

1 *The state space is $\mathcal{X} = \mathcal{P}(\mathcal{S})^{i_{\max}} \times \mathcal{Y}$ and the set of actions is $\mathcal{A}$*

2 *The evolution of the information state θ_n^i is given by:*

$$\theta_{n+1}^i = F^i(\theta_n, A_n, Y_{n+1})$$

and under the assumption that Y_n are uniformly distributed on the set $\mathcal{Y}$ we arrive at the following transition rule for $x = (\theta^1, \dots, \theta^{i_{\max}}, y) \in \mathcal{X}, a \in \mathcal{A}$,:

$$P(x'|x; a) := \begin{cases} \frac{1}{|\mathcal{Y}|}, & \text{if } \ x' = (F^1(\theta^1, a, y'), \dots, F^{i_{\max}}(\theta^{i_{\max}}, a, y'), y'), \\ 0, & \text{otherwise.} \end{cases}$$

3 *The cost functions are given by: $C : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$, with:*

$$C^i(x, a, x') = G^i(\theta^i, a, y') + \log(|\mathcal{Y}|^{\frac{1}{\lambda^i}})$$

4 *The set of histories is given by $\mathcal{H}_0 = \mathcal{X}$, and $\mathcal{H}_n = \mathcal{H}_{n-1} \times \mathcal{A} \times \mathcal{X}$. A policy $\pi \in \mathbf{\Pi}_{\mathcal{H}}$ takes the form $\pi = (f_0, \dots, f_n, \dots)$, where $f_n : \mathcal{H}_n \to \mathcal{A}$.*

5 *The optimization problem is governed by the performance index:*

$$\mathcal{I}_N(x_0, \pi) = \sum_{i=1}^{i_{\max}} w^i \mathbb{E}_{x_0}^{\pi} \left[e^{\lambda^i \left[\sum_{n=0}^{N-1} C^i(X_n, A_n, X_{n+1}) \right]} \right],$$

We remark that although η is not a bijection, it essentially behaves like one in the same way as the one explained in Section 3.2.3.

Proof. The transformation to the completely observable problem will be done by introducing a sufficient statistic information space realized through information state measures ψ and θ. The goal is to eliminate the unobservable quantities appearing in

$$\widehat{\mathbb{E}}_{\theta_0}^{\widehat{\pi}} \left[e^{\lambda^i \left[\sum_{n=0}^{N-1} \widehat{C}(S_n, A_n) \right]} \right].$$

Towards this goal a new probability measure that eliminates the dependencies that hold the previous measure to the partially observable case is introduced. Namely, following Cavazos-Cadena und Hernández-Hernández (2005) and Fleming und Hernández-Hernández (1997) there exists a unique probability measure $\mathbb{P}_{\theta}^{\widehat{\pi}}$ on $\widehat{\mathcal{G}}_n$, with expectation operator $\mathbb{E}_{\theta}^{\widehat{\pi}}$, given by:

$$\mathbb{P}_{\theta}^{\widehat{\pi}}(s_0, a_0, s_1, y_1, a_1, \dots, a_{n-1}, s_n, y_n) := \theta(s_0) \prod_{k=0}^{n-1} \left(\frac{1}{|\mathcal{Y}|} \widehat{P}\left(s_{k+1} | s_k; \widehat{f}_k(h_k) \right) \right),$$

for some $\theta \in \mathcal{P}(S)$, $\widehat{\pi} \in \mathbf{\Pi}_{\widehat{\mathcal{H}}}$. Note that $\theta(s_0) = \theta_0$ in our current set-up.

In what follows a relationship between $\widehat{\mathbb{E}}^{\widehat{\pi}}_{\theta_0}$ and $\mathbb{E}^{\widehat{\pi}}_{\theta_0}$ is constructed, in order for the latter to replace the former in the optimization problem.

The first important milestone in that direction makes use of the Radon-Nikodym theorem: Namely it can be noted that on the σ-field $\widehat{\mathcal{G}}_n$, the Radon-Nikodym derivative of $\widehat{P}^{\widehat{\pi}}_{\theta}$ with respect to $P^{\widehat{\pi}}_{\theta}$ is given by

$$\frac{\partial \widehat{\mathbb{P}}^{\widehat{\pi}}_{\theta}}{\partial \mathbb{P}^{\widehat{\pi}}_{\theta}}\Big|_{\widehat{\mathcal{G}}_n} = \prod_{k=0}^{n-1} (|\mathcal{Y}| Q(Y_{k+1}|S_{k+1})) =: Z_n,$$

and therefore

$$\widehat{\mathbb{E}}^{\widehat{\pi}}_{\theta_0}\left[e^{\lambda^i[\sum_{k=0}^n \widehat{C}(S_k,A_k)]}\right] = \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[e^{\lambda^i[\sum_{k=0}^n \widehat{C}(S_k,A_k)]} Z_n\right]. \tag{2.4}$$

After establishing the defining relationship between the two measures for the change of measure in equation (2.4) we now focus on the right part of the equation. This process will also lead to the construction of an information vector, on which the information space of the transformed MDP will be based, following the next steps:

1 First, let us define the positive and $\widehat{\mathcal{F}}_n$-measurable random variable ψ^i_n, by

$$\psi^i_n(s) := \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[\mathbb{1}_{\{S_n=s\}} e^{\lambda^i[\sum_{k=0}^n \widehat{C}(S_k,A_k)]} Z_n \Big| \widehat{\mathcal{F}}_n\right].$$

ψ_n is a a measure on $\mathcal{S}$, i.e. $\psi_n \in \mathcal{M}(\mathcal{S})$. Intuitively it can be understood, as *(random) average accumulated cost* up to time n of all outcomes that share the same observations and choices of controls leading to final state $S_n = s$, given the information observable to or controlled by the agent up to this time step.

2 Using ψ_n, the notation $\int \psi^i_n = \sum_{s\in\mathcal{S}} \psi^i_n(s)$, the linearity of the expectation operator, and the tower property of conditional expectation we can then rewrite the right side of 2.4:

$$\mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[e^{\lambda^i[\sum_{k=0}^n \widehat{C}(S_k,A_k)]} Z_n\right] =$$

$$\mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[\sum_{s\in\mathcal{S}} \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[\mathbb{1}_{\{S_n=s\}} e^{\lambda^i[\sum_{k=0}^n \widehat{C}(S_k,A_k)]} Z_n \Big| \widehat{\mathcal{F}}_n\right]\right] = \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[\int \psi^i_n\right] = \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[\int \psi^i_0 \prod_{k=1}^n \frac{\int \psi^i_k}{\int \psi^i_{k-1}}\right]. \tag{2.5}$$

3 Then note that setting $\psi^i_0 = \theta_0$, ψ^i_n satisfies the following recursion:

$$\psi^i_n = |\mathcal{Y}| M^i(A_{n-1}, Y_n)\psi^i_{n-1} \tag{2.6}$$

for the matrix $M(a,y) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ given by:

$$M^i(a,y)[s,s'] := \left(e^{\lambda^i \widehat{C}(s,a)} \widehat{P}(s'|s;a) Q(y|s')\right)^{\mathsf{T}}.$$

Inserting the recursion 2.6 in 2.5 yields:

$$\mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[\int \psi^i_n\right] = \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[|\mathcal{Y}|^n \int \psi^i_0 \prod_{k=1}^n \int \left[M^i(A_{k-1}, Y_k) \frac{\psi^i_{k-1}}{\int \psi^i_{k-1}}\right]\right]. \tag{2.7}$$

Note that the integral $\int \left[M^i(A_{k-1}, Y_k) \frac{\psi^i_{k-1}}{\int \psi^i_{k-1}} \right]$ is to be understood in the same way as the previously introduced notation for $\int \psi^i_n$. Also $\int \psi^i_0 = 1$ per definition.

4 Finally we normalize the information measure ψ^i_n by introducing θ^i_n, so that we arrive at an information state that is an element of $\mathcal{P}(\mathcal{S})$:

$$\theta^i_n = \frac{\psi^i_n}{\int \psi^i_n},$$

Replacing ψ^i_n with θ^i_n in 2.7 yields:

$$\mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[\int \psi^i_n\right] = \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[|\mathcal{Y}|^n \prod_{k=1}^{n} \int M^i(A_{k-1}, Y_k)\theta^i_{k-1}\right] = \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[e^{\lambda^i\left(\sum_{k=1}^{n}\left(\frac{1}{\lambda^i}\log\left(\int [M^i(A_{k-1},Y_k)\theta^i_{k-1}]\right)+\log|Y|^{\frac{1}{\lambda^i}}\right)\right)}\right], \tag{2.8}$$

where in the last step we have used the properties of the exponential and logarithmic functions in order to rewrite the operation in a more suitable form.

In steps 1-4 we have thus achieved our goal of rewriting 2.4 only using quantities known to the agent. Namely we have shown:

$$\widehat{\mathbb{E}}^{\widehat{\pi}}_{\theta_0}\left[e^{\lambda^i\left[\sum_{k=0}^{n}\widehat{R}(S_k,A_k)\right]}\right] = \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[e^{\lambda^i\left(\sum_{k=1}^{n}\left(\frac{1}{\lambda^i}\log\left(\int [M^i(A_{k-1},Y_k)\theta^i_{k-1}]\right)+\log|Y|^{\frac{1}{\lambda^i}}\right)\right)}\right]$$

In addition we have constructed an information state $\theta^i_n \in \mathcal{P}(\mathcal{S})$ for the transformed MDP.

As a last step we introduce the notation for the new optimization problem that leads to the direct claim of this theorem. First consider $G : \mathcal{P}(\mathcal{S}) \times \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$, given by:

$$G^i\left(\theta^i, a, y\right) := \frac{1}{\lambda^i}\log\left(\int M^i(u, y)\theta^i\right).$$

This lets us rewrite 2.8:

$$\mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[e^{\lambda^i\left(\sum_{k=1}^{n}\left(\frac{1}{\lambda^i}\log\left(\int [M^i(A_{k-1},Y_k)\theta^i_{k-1}]\right)+\log|Y|^{\frac{1}{\lambda^i}}\right)\right)}\right] = \mathbb{E}^{\widehat{\pi}}_{\theta_0}\left[e^{\lambda^i\left(\sum_{k=1}^{n}\left(G(\theta^i_{k-1},A_{k-1},Y_k)+\log|Y|^{\frac{1}{\lambda^i}}\right)\right)}\right]$$

Furthermore we use use 2.6 on θ^i_n:

$$\theta^i_n = \frac{\psi^i_n}{\int \psi^i_n} = \frac{|\mathcal{Y}|M^i(A_{n-1}, Y_n)\psi^i_{n-1}}{\int |\mathcal{Y}|M^i(A_{n-1}, Y_n)\psi^i_{n-1}} = \frac{M^i(A_{n-1}, Y_n)\theta^i_{n-1}}{\int M^i(A_{n-1}, Y_n)\theta^i_{n-1}}$$

We this recursion we can now write $F : \mathcal{P}(\mathcal{S}) \times A \times Y \to \mathcal{P}(\mathcal{S})$ as a forward operation for the information state:

$$F^i\left(\theta^i, a, y\right) := \frac{M^i(a, y)\theta^i}{\int M^i(a, y)\theta^i}.$$

For the reformulation of the MDP we can now use the cost functions: $C : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$, with:

$$C^i(x, a, x') = G^i(\theta^i, a, y') + \log(|\mathcal{Y}|^{\frac{1}{\lambda^i}})$$

Furthermore, for the policies we have $\eta : \mathbf{\Pi}_{\widehat{\mathcal{H}}} \to \mathbf{\Pi}_{\mathcal{H}}$ such that $(f_n)_{n\in\{0,\dots,N-1\}} = \eta((\widehat{f}_n)_{n\in\{0,\dots,N-1\}})$ satisfies

$$f_n(x_0, a_0, \dots, x_{n-1}, a_{n-1}, x_n) = \widehat{f}_n(\theta_0, a_0, \dots, y_{n-1}, a_{n-1}, y_n).$$

3 MDP with Multivariate Utility Function

In this section, we describe a model for risk sensitive multi-objective sequential decision making on a Borel state and action space with multiple costs and a multivariate utility function. The performance index is the expected multivariate utility, where each variable corresponds to a different running cost. As a generalization to the classical MDP model, we allow for the cost to depend on the subsequent state in addition to the current state-action pair. We thereby follow and extend ideas from Bäuerle und Rieder (2014) and Hernández-Lerma und Lasserre (1996).

3.1 Notation and Assumptions

Throughout this section, we assume that an N-step Markov Decision Process is given by a Borel state space $\mathcal{X}$, a Borel action space $\mathcal{A}$, a Borel set $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{A}$, and a regular conditional distribution P from $\mathcal{X} \times \mathcal{D}$ to $[0,1]$. Given the current state $x \in \mathcal{X}$, we assume that an action $a \in D(x)$ may be chosen, where $D(x) := \{a \in \mathcal{A} \mid (x,a) \in \mathcal{D}\}$ is the set of feasible actions. The transition probability to the next state is then given by the distribution $\mathrm{P}(\cdot|x;a)$, according to the chosen action. The set of histories from up to time n is defined by

$$\mathcal{H}_0 := \mathcal{X}, \qquad \mathcal{H}_n := \mathcal{H}_{n-1} \times \mathcal{A} \times \mathcal{X}, \quad n \in \mathbb{N}$$

and $h_n = (x_0, a_0, \dots, x_n) \in \mathcal{H}_n$ is a historical outcome up to time n.

Definition 3.1. *The set of (history-dependent) policies is defined by*

$$\mathbf{\Pi}_{\mathcal{H}} := \{\pi = (f_0, f_1, \dots) \mid f_n : \mathcal{H}_n \to \mathcal{A}, \quad \forall h_n \in \mathcal{H}_n : f_n(h_n) \in D(x_n), \quad n \in \mathbb{N}\}.$$

Similarly, the set of Markovian policies is defined by

$$\mathbf{\Pi}_{\mathcal{X}} := \{\pi = (g_0, g_1, \dots) \mid g_n : \mathcal{X} \to \mathcal{A}, \quad \forall x \in \mathcal{X} : g_n(x) \in D(x), \quad n \in \mathbb{N}\}.$$

Given an initial state $x \in \mathcal{X}$ and a history-dependent policy $\pi = (f_1, f_2, \dots) \in \mathbf{\Pi}_{\mathcal{H}}$, due to the Ionescu-Tulcea theorem, there exists a probability measure $\mathbb{P}_x^\pi$ on $\mathcal{H}_\infty$ and two stochastic processes $(X_n)_n$, $(A_n)_n$ such that

$$\mathbb{P}_x^\pi(X_0 \in B) = \delta_x(B),\ \mathbb{P}_x^\pi(X_{n+1} \in B \mid H_n, A_n) = \mathrm{P}(X_{n+1} \in B \mid X_n, A_n)$$

and

$$A_n = f_n(H_n)$$

for all Borel sets $B \subseteq \mathcal{X}$. Canonically, H_n, X_n, A_n are the history, state and action at time n. By $\mathbb{E}_x^\pi$ we denote the expectation operator corresponding to $\mathbb{P}_x^\pi$. For more details of this construction, we refer to Bäuerle und Rieder (2014).

Throughout the whole section, we have the following standing assumptions:

Assumption 3.2.

1 *The utility function $\mathcal{U} : \mathbb{R}^{i_{\max}} \to \mathbb{R}$ is continuous, and it exists $0 \le i_\tau \le i_{\max}$, such that $\mathcal{U}$ is component-wise increasing in $\{i < i_\tau\}$ and component-wise decreasing in $\{i > i_\tau\}$.*

2 *$\|\mathcal{U}\|_\infty < \infty$ (Read remark below).*

3 *The sets $D(x), x \in \mathcal{X}$ are compact.*

4 *The map $x \mapsto D(x)$ is upper semi-continuous, i.e. if $x_n \to x \in X$ and $a_n \in D(x_n)$, then (a_n) has an accumulation point in $D(x)$.*

5 *The maps* $(x, a, x') \mapsto C^i(x, a, x'),\ i = 1, \dots, i_{max}$, *are continuous, and it holds* $\underline{c} \leq C^i(\cdot, \cdot, \cdot) \leq \bar{c}$ *for some fixed* $\underline{c}, \bar{c} \in \mathbb{R}$.

6 P *is weakly continuous.*

Remark 3.3. *Due to uniform boundedness of the cost functions* C^i, *and the fact that we work on finite horizon problems with* $N < \infty$ *or infinite horizon problems with discount, we can observe that assumption 2 can be removed without any loss of generality.*

For notational convenience, we define the vector valued function $\boldsymbol{C} : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}^{i_{\max}}$ by

$$\boldsymbol{C}(x, a, x') := \left(C^1(x, a, x'), \dots, C^{i_{\max}}(x, a, x')\right).$$

3.2 Finite Horizon Problems

3.2.1 Performance Index

After we have set the stage for Markov Decision Processes and their policies, we can now define a performance index that is the expected utility of several running costs.

Definition 3.4. *Denote by* N *the number of steps of the MDP. We define the* total cost $\mathcal{I}_N(x, \pi)$ *given an initial state* $x \in \mathcal{X}$, *and a history dependent policy* $\pi \in \mathbf{\Pi}_{\mathcal{H}}$ *by*

$$\mathcal{I}_N(x, \pi) := \mathbb{E}_x^\pi \left[\mathcal{U}\left(\sum_{n=0}^{N-1} \boldsymbol{C}(X_n, A_n, X_{n+1}) \right) \right],$$

and the corresponding value function by

$$\mathcal{V}_N(x) := \inf_{\pi \in \mathbf{\Pi}_{\mathcal{H}}} \mathcal{I}_N(x, \pi). \tag{P}$$

3.2.2 Augmented problem

The aim of what follows is to determine $\mathcal{V}_N$, and optimal policies in (P). To this end, we augment the state space of the MDP to $\mathcal{X} \times \mathbb{R}^{i_{\max}}$. The second component models the so-far accumulated cost of the advancing MDP. In particular, $\widetilde{X}_n := (X_n, \mathcal{R}_n) \in \mathcal{X} \times \mathbb{R}^{i_{\max}}$ taking the value $(x, \boldsymbol{r}) = (x, r^1, \dots, r^{i_{\max}})$ implies that the MDP has advanced to state x and accumulated a cost amounting to r^i in the i-th objective after the first n steps. In order to define transition probabilities of the augmented problem, we introduce the notion of a pushforward measure.

Definition 3.5. *Given measurable spaces* $(\mathcal{S}, \mathcal{F})$, $(\widetilde{\mathcal{S}}, \widetilde{\mathcal{F}})$, *a measurable mapping* $\mathcal{T} : \mathcal{S} \to \widetilde{\mathcal{S}}$ *and a measure* $\mu : \mathcal{F} \to [0, \infty]$, *the pushforward of* μ *is the measure induced on* $(\widetilde{\mathcal{S}}, \widetilde{\mathcal{F}})$ *by* μ *under* $\mathcal{T}$, *i.e., the measure* $\mathcal{T}_{\#}\mu : \widetilde{\mathcal{F}} \to [0, \infty]$ *is given by*

$$(\mathcal{T}_{\#}\mu)(B) = \mu\left(\mathcal{T}^{-1}(B)\right) \text{ for } B \in \widetilde{\mathcal{F}}.$$

In particular, if a function f is $\widetilde{\mathcal{F}}$-measurable and $\mathcal{T}_{\#}\mu$-integrable, and $f \circ \mathcal{T}$ is μ-integrable, then

$$\int f \, d\mathcal{T}_{\#}\mu = \int f \circ \mathcal{T} \, d\mu.$$

Now, we define the transition kernel $\widetilde{P}$ of the augmented problem by

$$\widetilde{\mathrm{P}}(\cdot|\widetilde{x}; a) = \widetilde{\mathrm{P}}(\cdot|(x, \boldsymbol{r}); a) = (\mathcal{T}_{(x,\boldsymbol{r})})_{\#}\mathrm{P}(\cdot|x, a), \tag{3.1}$$

where

$$\mathcal{T}_{(x,\boldsymbol{r})}(x') = (x', \mathbf{C}(x,a,x') + \boldsymbol{r}). \tag{3.2}$$

If $\mathcal{X}$ is finite, this leads to

$$\widetilde{\mathrm{P}}(\widetilde{x}'|\widetilde{x};a) = \begin{cases} \mathrm{P}(x'|x;a), & \text{if } \widetilde{x} = (x,\boldsymbol{r}), \widetilde{x}' = (x', \boldsymbol{r} + \boldsymbol{C}(x,a,x')), \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

The histories for the augmented MDP are given by

$$\widetilde{\mathcal{H}}_0 := \mathcal{X} \times \mathbb{R}^{i_{\max}}, \qquad \widetilde{\mathcal{H}}_n := \widetilde{\mathcal{H}}_{n-1} \times \mathcal{D} \times \left(\mathcal{X} \times \mathbb{R}^{i_{\max}}\right), \quad n \in \mathbb{N}.$$

The definition of history-dependent policies $\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}$, Markovian policies $\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{X}}}$, and the corresponding decision rules are changed accordingly.

Similar to the previous section, there exist a probability measure $\widetilde{\mathbb{P}}_x^{\pi}$ on $\widetilde{\mathcal{H}}_\infty$ and a coupled stochastic process $(\widetilde{X})_{n\in\mathbb{N}}$ with $\widetilde{X}_n = (X_n, R_n)$, and a stochastic process $(A_n)_{n\in\mathbb{N}}$, such that

$$\widetilde{\mathbb{P}}_{\widetilde{x}}^{\widetilde{\pi}}(\widetilde{X}_0 \in B) = \delta_{\widetilde{x}}(B), \quad \widetilde{\mathbb{P}}_{\widetilde{x}}^{\widetilde{\pi}}(\widetilde{X}_{n+1} \in B \mid \widetilde{H}_n, A_n) = \widetilde{\mathrm{P}}(\widetilde{X}_{n+1} \in B \mid \widetilde{X}_n, A_n)$$

and

$$A_n = \widetilde{f}_n(\widetilde{H}_n)$$

for all Borel sets $B \subseteq \mathcal{X}$, and H_n, X_n, A_n are the history, state and action at time n, given an initial state $\widetilde{x} \in \mathcal{X} \times \mathbb{R}^{i_{\max}}$ and a history-dependent policy $\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}$. By induction, it is easy to prove that $\mathcal{R}_n = \sum_{k=0}^{n-1} \boldsymbol{C}(X_k, A_k, X_{k+1}) + \boldsymbol{r}$, $\widetilde{\mathbb{P}}_{\widetilde{x}}^{\widetilde{\pi}}$-almost surely.

Definition 3.6. *Denote by $N \in \mathbb{N}$ the number of steps of the MDP. For $n = 1, \dots, N$, we define the* total cost $\widetilde{\mathcal{I}}_n((x,\boldsymbol{r}),\widetilde{\pi})$ *for the augmented problem, given the initial state $x \in \mathcal{X}$, initial cost $\boldsymbol{r} \in \mathbb{R}^{i_{\max}}$, and policy $\widetilde{\pi} \in \widetilde{\Pi}$ by*

$$\widetilde{\mathcal{I}}_n(\widetilde{x},\widetilde{\pi}) = \widetilde{\mathcal{I}}_n((x,\boldsymbol{r}),\widetilde{\pi}) := \widetilde{\mathbb{E}}_{\widetilde{x}}^{\widetilde{\pi}}\left[\mathcal{U}\left(\sum_{k=0}^{n-1} \boldsymbol{C}(X_k, A_k, X_{k+1}) + \boldsymbol{r}\right)\right] = \widetilde{\mathbb{E}}_{\widetilde{x}}^{\widetilde{\pi}}\left[\mathcal{U}\left(R_n\right)\right], \tag{3.4}$$

and the corresponding value function by

$$\widetilde{\mathcal{V}}_N(\widetilde{x}) := \inf_{\widetilde{\pi}\in\mathbf{\Pi}_{\widetilde{\mathcal{H}}}} \widetilde{\mathcal{I}}_N(\widetilde{x},\widetilde{\pi}). \tag{$\widetilde{P}$}$$

3.2.3 Policy bijection

For the sequel, let $x \in \mathcal{X}$, and $\widetilde{x} = (x, \mathbf{0})$. Note that policies $\pi = (f_0, f_1, \dots) \in \mathbf{\Pi}_{\mathcal{H}}$ of the original problem consist of functions f_n that are defined on $\mathcal{H}_n$, and policies $\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}$ consist of functions $\widetilde{f}_n$ defined on $\widetilde{\mathcal{H}}_n$. Therefore there is no simple bijectional correspondence between the two sets. However, the set of histories

$$\widetilde{\mathcal{H}}_n^- = \{\widetilde{h}_n \in \widetilde{\mathcal{H}}_n | \, (\exists k \in \{1, \dots, n-1\} : \boldsymbol{r}_k \neq \boldsymbol{r}_{k-1} + \boldsymbol{C}(x_k, a_k, x_{k+1})) \vee (\boldsymbol{r}_0 \neq \mathbf{0})\}, \tag{3.5}$$

is not accessible in the sense that these histories cannot occur. In a more rigorous manner, we have that $\widetilde{\mathbb{P}}_{\widetilde{x}}^{\widetilde{\pi}}(\widetilde{\mathcal{H}}_n^-) = 0$, for all $\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}$. For every policy $\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}$, we may define a new "reduced" policy $\widetilde{\pi}^{\mathrm{red}}$ by

$$\widetilde{f}_n^{\mathrm{red}}(\widetilde{h}_n) = \begin{cases} a^{\mathrm{red}}(x_n), & \text{if } \widetilde{h}_n \in \widetilde{\mathcal{H}}_n^-, \\ \widetilde{f}_n(\widetilde{h}_n), & \text{otherwise,} \end{cases} \tag{3.6}$$

where a^{red} can be any arbitrary but fixed point in $\mathcal{D}(x_n)$. Then for the set $\mathbf{\Pi}^{\mathrm{red}}_{\widetilde{\mathcal{H}}} = \{\widetilde{\pi}^{\mathrm{red}} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}} : \widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}\}$, we can define a bijection to $\mathbf{\Pi}_{\mathcal{H}}$. To do so, for $\pi = (f_0, f_1, \dots) \in \mathbf{\Pi}_{\mathcal{H}}$, we define $\widetilde{\pi}^{\mathrm{red}} = (\widetilde{f}^{\mathrm{red}}_0, \widetilde{f}^{\mathrm{red}}_1, \dots) \in \mathbf{\Pi}^{\mathrm{red}}_{\widetilde{\mathcal{H}}}$, by

$$\widetilde{f}^{\mathrm{red}}_n\left((x_0,0), a_0, \dots, \left(x_{n-1}, \sum_{i=0}^{n-1} \mathbf{C}(x_i,a_i,x_{i+1})\right), a_{n-1}, \left(x_n, \sum_{i=0}^{n} \mathbf{C}(x_i,a_i,x_{i+1})\right)\right) = \begin{cases} a^{\mathrm{red}}, & \text{if } \widetilde{h}_n \in \widetilde{\mathcal{H}}^-_n, \\ f_n(x_0,a_0,\dots,x_{n-1},a_{n-1},x_n), & \text{otherwise.} \end{cases} \tag{3.7}$$

It is easy to see that the value function of (P) coincides with the value function of $(\widetilde{P})$ with $\boldsymbol{r} = \mathbf{0}$, i.e.

$$\mathcal{V}_N(x) = \inf_{\pi \in \mathbf{\Pi}_{\mathcal{H}}} \mathcal{I}_N(x,\pi) = \inf_{\widetilde{\pi} \in \mathbf{\Pi}^{\mathrm{red}}_{\widetilde{\mathcal{H}}}} \widetilde{\mathcal{I}}_N((x,\mathbf{0}),\widetilde{\pi}) = \inf_{\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}} \widetilde{\mathcal{I}}_N((x,\mathbf{0}),\widetilde{\pi}) = \widetilde{\mathcal{V}}_N((x,\mathbf{0})).$$

The next step is to derive a Bellman-style equation for the augmented problem $(\widetilde{P})$. It can be shown that the minimizer of $(\widetilde{P})$ is a Markovian policy.

3.2.4 Bellman operator and first theorem

First, for a fixed $m \in \mathbb{R}$, we define the set

$$\Delta := \Big\{ v : \mathcal{X} \times \mathbb{R}^{i_{\max}} \to \mathbb{R} \;\Big|\; v \text{ is lower semi-continuous, } v(x,\cdot) \text{ is continuous, } \|v\|_\infty < \infty, \; \inf_{x,d}\{v(x,\boldsymbol{r})\} \geq m,$$

and for all $x \in \mathcal{X}$ is component-wise increasing (decreasing) in $\{i < i_\tau\}$ (in $\{i > i_\tau\}$)$\Big\}$.

For $v \in \Delta$ and a Markovian decision rule $\widetilde{g} \in \mathbf{\Pi}_{\widetilde{\mathcal{X}}}$, we define the operators

$$\begin{aligned} T_{\widetilde{g}}[v](\widetilde{x}) = T_{\widetilde{g}}[v](x,\boldsymbol{r}) &= \int v\left(\widetilde{x}'\right) \widetilde{\mathrm{P}}\left(d\widetilde{x}' \mid \widetilde{x}, \widetilde{g}(\widetilde{x})\right) = \int v\left((x',r')\right) \left(\mathcal{T}_{(x,\boldsymbol{r})}\right)_{\#} \mathrm{P}\left(dx' \mid x, \widetilde{g}(x,\boldsymbol{r})\right) \\ &= \int v\left(x', \boldsymbol{C}(x,\widetilde{g}(x,\boldsymbol{r}),x') + \boldsymbol{r}\right) \mathrm{P}(dx' \mid x, \widetilde{g}(x,\boldsymbol{r})), \end{aligned}$$

and

$$T[v](x,\boldsymbol{r}) = \inf_{a \in D(x)} \int v(x', \boldsymbol{C}(x,a,x') + \boldsymbol{r}) P(dx' \mid x, a),$$

whenever the integrals exist. T is called the *minimal cost operator*. We say that a Markovian decision rule $\widetilde{g}$ is a *minimizer* of v if $T_{\widetilde{g}}[v] = T[v]$. In this situation, $\widetilde{g}(x,\boldsymbol{r})$ is a minimizer of

$$\mathcal{D}(x) \ni a \mapsto \int v\left(x', \boldsymbol{C}(x,a,x') + \boldsymbol{r}\right) P(dx' \mid x, a)$$

for every $(x,\boldsymbol{r}) \in \mathcal{X} \times \mathbb{R}^{i_{\max}}$. We may now state the main result of this section:

Theorem 3.7. *Let $\widetilde{\mathcal{V}}_0(x,\boldsymbol{r}) := \mathcal{U}(\boldsymbol{r})$. Then, the following holds:*

a) For any Markovian policy $\widetilde{\pi} = (\widetilde{g}_0, \widetilde{g}_1, \dots) \in \mathbf{\Pi}_{\widetilde{\mathcal{X}}}$, we have the cost iteration

$$\widetilde{\mathcal{I}}_n((x,\boldsymbol{r}),\widetilde{\pi}) = T_{\widetilde{g}_0}[\dots[T_{\widetilde{g}_{n-1}}[\widetilde{\mathcal{V}}_0]]](x,\boldsymbol{r})$$

for all $n = 1, \dots, N$.

b) If an optimal policy exists it is Markovian, i.e.

$$\inf_{\widetilde{\pi}\in\mathbf{\Pi}_{\widetilde{\mathcal{H}}}} \mathcal{I}_N(\widetilde{x},\widetilde{\pi}) = \inf_{\widetilde{\pi}\in\mathbf{\Pi}_{\widetilde{\mathcal{X}}}} \mathcal{I}_N(\widetilde{x},\widetilde{\pi}).$$

c) The operator $T : \Delta \to \Delta$ is well-defined, and for every $v \in \Delta$, there exists a minimizer of $T[v]$.

d) We get the Bellman-style equation

$$\widetilde{\mathcal{V}}_n(x,\boldsymbol{r}) = T[\widetilde{\mathcal{V}}_{n-1}](x,\boldsymbol{r}) = \inf_{a\in D(x)} \int \widetilde{\mathcal{V}}_{n-1}(x', \boldsymbol{C}(x,a,x') + \boldsymbol{r}) P(dx' \mid x, a)$$

for all $n = 1, \dots, N$.

e) If $\widetilde{g}_n^*$ is a minimizer of $\widetilde{\mathcal{V}}_{n-1}$ for $n = 1, \dots, N$, then the history-dependent policy $\pi^* = (f_0^*, \dots, f_{N-1}^*)$, defined by

$$f_n^*(h_n) := \begin{cases} \widetilde{g}_N^*(x_0, 0) & \text{if } n = 0, \\ \widetilde{g}_{N-n}^*\left(x_n, \sum_{k=0}^{n-1} \boldsymbol{C}(x_k, a_k, x_{k+1})\right) & \text{otherwise,} \end{cases}$$

is an optimal policy for problem (P).

For the proof of Theorem 3.7, we need the following lemma:

Lemma 3.8. *Let $v : \mathcal{X} \times \mathbb{R}^{i_{max}} \to \mathbb{R}$ be bounded and lower semi-continuous. Suppose*

1 *$D(x)$ is compact,*

2 *$x \mapsto D(x)$ is upper semi-continuous,*

3 *$(x, \boldsymbol{r}, \widetilde{g}, x') \mapsto v(x', \boldsymbol{C}(x, \widetilde{g}(x,\boldsymbol{r}), x') + \boldsymbol{r})$ is lower semi-continuous.*

Then, Tv is is lower semi-continuous and there exists a minimizer $\widetilde{g}^*$ such that $T_{\widetilde{g}^*} v = Tv$.

Proof. By Lemma 17.11 in Hinderer (1970), $(x, \boldsymbol{r}, \widetilde{g}) \mapsto T_{\widetilde{g}} v(x, \boldsymbol{r})$ is lower semi-continuous. The claim then follows from a similar argument to Proposition 2.4.3 in Bäuerle und Rieder (2011).

Proof of Theorem 3.7. The proof is similar to Theorem 2.3.4 and Theorem 2.3.8 in Bäuerle und Rieder (2011) with a different state space, see also Bäuerle und Rieder (2014).

ad a) An easy calculation shows that

$$\begin{aligned}\widetilde{\mathcal{I}}_1((x,\boldsymbol{r}),\widetilde{\pi}) &= \mathbb{E}_x^{\widetilde{\pi}}\Big[\mathcal{U}\left(\boldsymbol{C}(X_0, A_0, X_1) + \boldsymbol{r}\right)\Big] \\ &= \int \mathcal{U}\left(\boldsymbol{C}(x, \widetilde{g}_0(x,\boldsymbol{r}), x') + \boldsymbol{r}\right) P(dx' \mid x, \widetilde{g}_1(x,\boldsymbol{r})) = T_{\widetilde{g}_1}[\widetilde{\mathcal{V}}_0](x,\boldsymbol{r}).\end{aligned}$$

Now, let $\widetilde{\pi}^+ = (\widetilde{g}_2, \dots)$. For $n = 2, \dots, N$, we get

$$\begin{aligned}\widetilde{\mathcal{I}}_n((x,\boldsymbol{r}),\widetilde{\pi}) &= \mathbb{E}_x^{\widetilde{\pi}}\left[\mathcal{U}\left(\sum_{k=0}^{n-1} \boldsymbol{C}(X_k, A_k, X_{k+1}) + \boldsymbol{r}\right)\right] \\ &= \int \mathbb{E}_{x'}^{\widetilde{\pi}^+}\left[\mathcal{U}\left(\sum_{k=0}^{n-2} \boldsymbol{C}(X_k, A_k, X_{k+1}) + \boldsymbol{r} + \boldsymbol{C}(x, \widetilde{g}_1(x,\boldsymbol{r}), x')\right)\right] P(dx' \mid x, \widetilde{g}_0(x,\boldsymbol{r})) \\ &= \int \widetilde{\mathcal{I}}((x',\boldsymbol{r}),\widetilde{\pi}^+) P(dx' \mid x, \widetilde{g}_1(x,\boldsymbol{r})) = T_{\widetilde{g}_1}[\widetilde{\mathcal{I}}_{n-1}(\cdot,\widetilde{\pi}^+)](\widetilde{x}) = T_{\widetilde{g}_1}[\dots[T_{\widetilde{g}_{n-1}}[\widetilde{\mathcal{V}}_0]]](x,\boldsymbol{r}).\end{aligned}$$

The claim follows then by induction.

ad b) This follows from Theorem 2.2.3 in Bäuerle und Rieder (2011).

ad c) Note that every $v \in \Delta$ is bounded from below by m. By our assumptions, we get that $(x, \boldsymbol{r}, \widetilde{g}, x') \mapsto v(x', \boldsymbol{C}(x, \widetilde{g}(x, \boldsymbol{r}), x') + \boldsymbol{r})$ is lower semi-continuous, and bounded from below, i.e. we are in the setting of Lemma 3.8. Thus, $T[v]$ is lower semi-continuous and there exists a minimizer $\widetilde{g}^*$ such that $T_{\widetilde{g}^*}[v] = T[v]$.

For fixed $x \in \mathcal{X}$, and $a \in \mathcal{D}(x)$, the map $\boldsymbol{r} \mapsto \int v(x', \boldsymbol{C}(x, a, x') + \boldsymbol{r}) P(dx' \mid x, a)$ has the same monotonicities with respect to $r_i's$ as v and it is continuous for every $a \in D(x)$. The continuity can be proven with the dominated convergence theorem since $\|v\|_\infty < \infty$. Therefore, the infimum of these maps over all $a \in D(x)$ is upper semi-continuous in $\boldsymbol{r}$. With this, we have shown that $Tv(x, \cdot)$ is upper and lower semi-continuous, and therefore continuous, and respects the same monotonicities as v for all $x \in \mathcal{X}$. Because $v(x, \cdot) \geq m$, we have $T[v](x, \cdot) \geq m$. The boundness assumption follows from the definition of T and the corresponding property of v. We have then shown that $T : \Delta \to \Delta$ is well-defined.

ad d) Let $\widetilde{g}_n^*$ be a minimizer of $\mathcal{V}_{n-1}$ for $n = 1, \ldots, N$ and denote by $\pi^* = (\widetilde{g}_1^*, \ldots, \widetilde{g}_N^*)$ the associated policy.

For $n = 1$, we get that

$$\widetilde{\mathcal{V}}_1(x, \boldsymbol{r}) = \inf_{\widetilde{\pi} \in \boldsymbol{\Pi}_{\widetilde{\mathcal{X}}}} \mathbb{E}_x^{\widetilde{\pi}} \Big[\mathcal{U}\left(\boldsymbol{C}(X_0, A_0, X_1) + \boldsymbol{r}\right) \Big] = \inf_{a \in D(x)} \int \mathcal{U}(\boldsymbol{C}(x, a, x') + \boldsymbol{r}) P(dx' \mid x, a) = T[\widetilde{\mathcal{V}}_0](x, \boldsymbol{r}),$$

and obviously, $\widetilde{\mathcal{V}}_1(x, \boldsymbol{r}) = \widetilde{\mathcal{I}}_1((x, \boldsymbol{r}), \widetilde{\pi}^*)$. Now, assume that $\widetilde{\mathcal{I}}_n((x, \boldsymbol{r}), \widetilde{\pi}^*) = \widetilde{\mathcal{V}}_n(x, \boldsymbol{r})$ for a fixed $n \in \{1, \ldots, N\}$. Then,

$$\begin{aligned}
\widetilde{\mathcal{I}}_{n+1}((x, \boldsymbol{r}), \widetilde{\pi}^*) &= T_{\widetilde{g}_1^*}[\widetilde{\mathcal{I}}_n(\cdot, (\widetilde{\pi}^*)^+)](x, \boldsymbol{r}) && \text{using (a)}, \\
&= T_{\widetilde{g}_1^*}[\widetilde{\mathcal{V}}_n](x, \boldsymbol{r}) && \text{by the induction hypothesis}, \\
&= T[\widetilde{\mathcal{V}}_n](x, \boldsymbol{r}) && \text{by definition of } \widetilde{g}_1^*.
\end{aligned}$$

By taking the infimum, we get

$$\inf_{\widetilde{\pi} \in \boldsymbol{\Pi}_{\widetilde{\mathcal{X}}}} \widetilde{\mathcal{I}}_{n+1}((x, \boldsymbol{r}), \widetilde{\pi}) = \widetilde{\mathcal{V}}_{n+1}(x, \boldsymbol{r}) \leq \widetilde{\mathcal{I}}_{n+1}((x, \boldsymbol{r}), \widetilde{\pi}^*) = T[\widetilde{\mathcal{V}}_n](x, \boldsymbol{r}). \tag{3.8}$$

On the other hand, with an arbitrary policy $\widetilde{\pi} = (\widetilde{g}_1, \ldots, \widetilde{g}_N)$,

$$\begin{aligned}
\widetilde{\mathcal{I}}_{n+1}((x, \boldsymbol{r}), \widetilde{\pi}) &= T_{\widetilde{g}_1}[\widetilde{\mathcal{I}}_n(\cdot, \widetilde{\pi}^+)](x, \boldsymbol{r}) && \text{using (a)}, \\
&\geq T_{\widetilde{g}_1}[\widetilde{\mathcal{V}}_n](x, \boldsymbol{r}) && \text{by the monotonicity of } T, \\
&\geq T[\widetilde{\mathcal{V}}_n](x, \boldsymbol{r}) && \text{by taking the infimum}.
\end{aligned}$$

By taking the infimum, we get

$$\inf_{\widetilde{\pi} \in \boldsymbol{\Pi}_{\widetilde{\mathcal{X}}}} \widetilde{\mathcal{I}}_{n+1}((x, \boldsymbol{r}), \widetilde{\pi}) = \widetilde{\mathcal{V}}_{n+1}(x, \boldsymbol{r}) \geq T[\widetilde{\mathcal{V}}_n](x, \boldsymbol{r}). \tag{3.9}$$

From (3.8) and (3.9), it follows by induction that

$$\widetilde{\mathcal{V}}_n(x, \boldsymbol{r}) = T[\widetilde{\mathcal{V}}_{n-1}](x, \boldsymbol{r}) = \widetilde{\mathcal{I}}_n((x, \boldsymbol{r}), \widetilde{\pi}^*)$$

for all $n = 1, \ldots, N$.

ad e) Consider the Markovian policy $\widetilde{\pi}^* = (\widetilde{g}_1^*, \ldots, \widetilde{g}_N^*)$ as defined in *(d)*. We have just shown that $\widetilde{\mathcal{V}}_N(x, \boldsymbol{r}) = \widetilde{\mathcal{I}}_N((x, \boldsymbol{r}), \widetilde{\pi}^*)$, i.e. $\widetilde{\pi}^*$ is a minimizer of $(\widetilde{P})$, and therefore an optimal policy for the N-step MDP with states in $X \times \mathbb{R}^{i_{\max}}$. The claim follows by (3.7) in Section 3.2.3 where the policy bijection is explored.

3.3 Discounted finite horizon problems

We now consider finite horizon problems with a discount vector $\boldsymbol{\beta} \in (0,1)^{i_{\max}}$, and prove the corresponding analogon to Theorem 3.7. The techniques used are similar to those from Bäuerle und Rieder (2014), where they are applied to *univariate* utility functions. Similar to the previous setting, we define the *total cost* $\mathcal{I}_N(x,\pi)$, given an initial state $x \in \mathcal{X}$, and a history dependent policy $\pi \in \mathbf{\Pi}_{\mathcal{H}}$, by

$$\mathcal{I}_N(x,\pi) := \mathbb{E}_x^{\pi}\left[\mathcal{U}\left(\sum_{n=0}^{N-1} \boldsymbol{\beta}^n \cdot \boldsymbol{C}(X_n, A_n, X_{n+1})\right)\right], \tag{3.10}$$

and the corresponding value function, by

$$\mathcal{V}_N(x) := \inf_{\pi \in \mathbf{\Pi}_{\mathcal{H}}} \mathcal{I}_N(x,\pi). \tag{P}$$

We remark that the dot product appearing in (3.10) is a componentwise product, i.e.

$$\mathbf{a} \cdot \mathbf{b} = (a_1 b_1, \ldots, a_n b_n) \qquad \text{for } \mathbf{a} = (a_1, \ldots, a_n), \mathbf{b} = (b_1, \ldots, b_n).$$

3.3.1 Augmented problem

Again, we consider an augmented state space $\mathcal{X} \times \mathbb{R}^{i_{\max}} \times (0,1)^{i_{\max}}$, where the new components keep track of the decreasing discount factor. Policy augmentation is done similar to the previous section. The new transition kernel $\widetilde{P}$ of the augmented problem is given by

$$\widetilde{\mathrm{P}}(\cdot|\widetilde{x};a) = \widetilde{\mathrm{P}}(\cdot|(x,\boldsymbol{r},\boldsymbol{z});a) = (\mathcal{T}_{(x,\boldsymbol{r},\boldsymbol{z})})_{\#}\mathrm{P}(\cdot|x,a), \tag{3.11}$$

where

$$\mathcal{T}_{(x,\boldsymbol{r},\boldsymbol{z})}(x') = (x', \boldsymbol{z} \cdot \mathbf{C}(x,a,x') + \boldsymbol{r}, \boldsymbol{z} \cdot \boldsymbol{\beta}). \tag{3.12}$$

On the augmented state space, given an initial state $x \in \mathcal{X}$, initial cost $\boldsymbol{r} \in \mathbb{R}^{i_{\max}}$, initial discount rates $\boldsymbol{z} \in (0,1)^{i_{max}}$, and a policy $\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}$ the total cost $\widetilde{\mathcal{I}}_n((x,\boldsymbol{r},z),\widetilde{\pi})$ for $n = 1, ..., N$ is given by

$$\widetilde{\mathcal{I}}_n(\widetilde{x},\widetilde{\pi}) = \widetilde{\mathcal{I}}_n((x,\boldsymbol{r},z),\widetilde{\pi}) := \widetilde{\mathbb{E}}_{\widetilde{x}}^{\widetilde{\pi}}\left[\mathcal{U}\left(\boldsymbol{z} \cdot \sum_{k=0}^{n-1} \boldsymbol{\beta}^k \cdot \boldsymbol{C}(X_k, A_k, X_{k+1}) + \boldsymbol{r}\right)\right]. \tag{3.13}$$

The corresponding value function is

$$\widetilde{\mathcal{V}}_n(\widetilde{x}) := \inf_{\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}} \widetilde{\mathcal{I}}_n(\widetilde{x},\widetilde{\pi}). \tag{$\widetilde{P}$}$$

3.3.2 Bellman operator and second theorem

Let

$$\begin{aligned}\Delta := \big\{ v : \mathcal{X} \times \mathbb{R}^{i_{\max}} \times (0,1)^{i_{\max}} \to \mathbb{R} \;\big|\; & v \text{ is lower semi-continuous,} \\ & v(x,\cdot,\cdot) \text{ is continuous, and componentwise increasing for all } x \in \mathcal{X}, \\ & v(x,\boldsymbol{r},\boldsymbol{z}) \geq \mathcal{U}(\boldsymbol{r}) \text{ for all } (x,\boldsymbol{r},\boldsymbol{z}) \in \mathcal{X} \times \mathbb{R}^{i_{\max}} \times (0,1) \big\}.\end{aligned}$$

For $v \in \Delta$ and a Markovian decision rule $\widetilde{g}$, we define the operators

$$\begin{aligned}T_{\widetilde{g}}[v](\widetilde{x}) = T_{\widetilde{g}}[v](x,\boldsymbol{r},\boldsymbol{z}) &:= \int v\left(\widetilde{x}'\right) \widetilde{\mathrm{P}}\left(d\widetilde{x}' \mid \widetilde{x}, \widetilde{g}(\widetilde{x})\right) = \int v\left((x',\boldsymbol{r}',\boldsymbol{z}')\right) \left(\mathcal{T}_{(x,\boldsymbol{r},\boldsymbol{z})}\right)_{\#}\mathrm{P}\left(dx' \mid x, \widetilde{g}(x,\boldsymbol{r},\boldsymbol{z})\right) \\ &= \int v\left(x', \boldsymbol{z} \cdot \boldsymbol{C}(x, \widetilde{g}(x,\boldsymbol{r},\boldsymbol{z}), x') + \boldsymbol{r}, \boldsymbol{z} \cdot \boldsymbol{\beta}\right) \mathrm{P}(dx' \mid x, \widetilde{g}(x,\boldsymbol{r},z)),\end{aligned}$$

and

$$T[v](x, \boldsymbol{r}, \boldsymbol{z}) = \inf_{a \in D(x)} \int v(x', \boldsymbol{z} \cdot \boldsymbol{C}(x, a, x') + \boldsymbol{r}, \boldsymbol{z} \cdot \boldsymbol{\beta}) \mathrm{P}(dx' \mid x, a),$$

whenever the integrals exist. T is again called the *minimal cost operator*. We may now state the main theorem of this section.

Theorem 3.9. *Let $\widetilde{\mathcal{V}}_0(x, \boldsymbol{r}, \boldsymbol{z}) := \mathcal{U}(\boldsymbol{r})$. The following holds:*

a) For any Markovian policy $\widetilde{\pi} = (\widetilde{g}_1, \dots) \in \mathbf{\Pi}_{\widetilde{\mathcal{X}}}$, we have the cost iteration

$$\widetilde{\mathcal{I}}_n((x, \boldsymbol{r}, \boldsymbol{z}), \widetilde{\pi}) = T_{\widetilde{g}_1}[\dots[T_{\widetilde{g}_{n-1}}[\widetilde{\mathcal{V}}_0]]](x, \boldsymbol{r}, \boldsymbol{z})$$

for all $n = 1, \dots, N$.

b) The optimal policy is Markovian, i.e.

$$\inf_{\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{H}}}} \mathcal{I}_N(\widetilde{x}, \widetilde{\pi}) = \inf_{\widetilde{\pi} \in \mathbf{\Pi}_{\widetilde{\mathcal{X}}}} \mathcal{I}_N(\widetilde{x}, \widetilde{\pi}).$$

c) The operator $T : \Delta \to \Delta$ is well-defined, and for every $v \in \Delta$, there exists a minimizer of $T[v]$.

d) We get the Bellman-style equation

$$\widetilde{\mathcal{V}}_n(x, \boldsymbol{r}, \boldsymbol{z}) = T[\widetilde{\mathcal{V}}_{n-1}](x, \boldsymbol{r}, \boldsymbol{z}) = \inf_{a \in D(x)} \int \widetilde{\mathcal{V}}_{n-1}(x', \boldsymbol{z} \cdot \boldsymbol{C}(x, a, x') + \boldsymbol{r}, \boldsymbol{z} \cdot \boldsymbol{\beta}) P(dx' \mid x, a)$$

for all $n = 1, \dots, N$.

e) If $\widetilde{g}_n^*$ is a minimizer of $\widetilde{\mathcal{V}}_{n-1}$ for $n = 1, \dots, N$, then $\widetilde{\pi}^* = (\widetilde{g}_1^*, \dots, \widetilde{g}_N^*)$ is an optimal policy for $(\widetilde{P})$. In this situation, the history-dependent policy $\pi^* = (f_0^*, \dots, f_{N-1}^*)$, defined by

$$f_n^*(h_n) := \begin{cases} \widetilde{g}_N^*(x_0, 0, 1) & \text{if } n = 0, \\ \widetilde{g}_{N-n}^*\left(x_n, \sum_{k=0}^{n-1} \boldsymbol{\beta}^k \cdot \boldsymbol{C}(x_k, a_k, x_{k+1}), \boldsymbol{\beta}^n\right) & \text{otherwise}, \end{cases}$$

is an optimal policy for problem (P).

Proof. The proof is similar to the derivation of Theorem 3.7. We will only prove *(a)* by induction. To that end, note that

$$\begin{aligned} \widetilde{\mathcal{I}}_1((x, \boldsymbol{r}, \boldsymbol{z}), \widetilde{\pi}) &= \widetilde{\mathbb{E}}_x^{\widetilde{\pi}}\Big[\mathcal{U}(\boldsymbol{z} \cdot \boldsymbol{C}(X_0, A_0, X_1) + \boldsymbol{r})\Big] \\ &= \int \mathcal{U}(\boldsymbol{z} \cdot \boldsymbol{C}(x, \widetilde{g}_0(x, \boldsymbol{r}, \boldsymbol{z}), x') + \boldsymbol{r}) P(dx' \mid x, \widetilde{g}_0(x, \boldsymbol{r}, \boldsymbol{z})) \\ &= T_{\widetilde{g}_1}[\widetilde{\mathcal{V}}_0](x, \boldsymbol{r}, \boldsymbol{z}). \end{aligned}$$

Let $\widetilde{\pi}^+ = (\widetilde{g}_2, \widetilde{g}_3, \dots)$. For $n = 2, \dots, N$, we get

$$\widetilde{\mathcal{I}}_n((x, \boldsymbol{r}, \boldsymbol{z}), \widetilde{\pi}) = \mathbb{E}_x^{\widetilde{\pi}}\left[\mathcal{U}\left(\boldsymbol{z} \sum_{k=0}^{n-1} \beta^k \boldsymbol{C}(X_k, A_k, X_{k+1}) + \boldsymbol{r}\right)\right]$$

$$= \int \mathbb{E}_{x'}^{\widetilde{\pi}^+}\left[\mathcal{U}\left(\boldsymbol{z} \cdot \sum_{k=0}^{n-2} \boldsymbol{\beta}^{k+1} \cdot \boldsymbol{C}(X_k, A_k, X_{k+1}) + \boldsymbol{z} \cdot \boldsymbol{C}(x, \widetilde{g}_1(x, \boldsymbol{r}, \boldsymbol{z}), x') + \boldsymbol{r}\right)\right] P(dx' \mid x, \widetilde{g}_1(x, \boldsymbol{r}, \boldsymbol{z}))$$

$$= \int \widetilde{\mathcal{I}}_{n-1}((x', \boldsymbol{z} \cdot \boldsymbol{C}(x, \widetilde{g}_1(x, \boldsymbol{r}, \boldsymbol{z}) + \boldsymbol{r}, \boldsymbol{\beta} \cdot \boldsymbol{z}), \widetilde{\pi}^+) P(dx' \mid x, \widetilde{g}_1(x, \boldsymbol{r}, \boldsymbol{z}))$$

$$= T_{\widetilde{g}_1}[\widetilde{\mathcal{I}}_{n-1}(\cdot, \widetilde{\pi}^+)](\widetilde{x}) = T_{\widetilde{g}_1}[T_{\widetilde{g}_2}[\dots[T_{\widetilde{g}_{n-1}}[\widetilde{\mathcal{V}}_0]]]](x, \boldsymbol{r}, \boldsymbol{z}).$$

The claim follows then inductively.

3.4 Infinite horizon problems

In this section, we study the infinite horizon problem with *discount factor* $\boldsymbol{\beta} \in (0,1)^{i_{\max}}$. For a vector $\boldsymbol{a} \in \mathbb{R}^{i_{\max}}$, we will denote with $\underline{a} = \min\{a_1, \dots, a_{i_{\max}}\}, \overline{a} = \max\{a_1, \dots, a_{i_{\max}}\}$. The notion of T and Δ from the previous section is unchanged. The total cost in this situation reads as

$$\mathcal{I}_\infty(x,\pi) := \mathbb{E}_x^\pi \left[\mathcal{U}\left(\sum_{n=0}^{\infty} \boldsymbol{\beta}^n \cdot \boldsymbol{C}(X_n, A_n, X_{n+1}) \right) \right],$$

and $\widetilde{\mathcal{V}}_\infty$ is defined accordingly. We shall use the following definition:

Definition 3.10. *For a continuous function $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$, we define the modulus of continuity $\omega_{\mathcal{F}}(\delta, R)$ on the ball $B(0,R)$, by*

$$\omega_{\mathcal{F}}(\delta, R) = \sup\left\{ |\mathcal{F}(x) - \mathcal{F}(y)| \;\middle|\; x, y \in B(0,R), \|x-y\|_2 < \delta \right\} \tag{3.14}$$

Note that the modulus of continuity is increasing in both variables, and it holds $\lim_{\delta \to 0} \omega_{\mathcal{F}}(\delta, R) \to 0$.

Theorem 3.11. *Let $\underline{b}(\boldsymbol{r}, \boldsymbol{z}) := \mathcal{U}\left(\boldsymbol{z} \cdot \frac{\underline{c}}{\mathbf{1}-\boldsymbol{\beta}} + \boldsymbol{r} \right)$ and $\bar{b}(\boldsymbol{r}, \boldsymbol{z}) := \mathcal{U}\left(\boldsymbol{z} \cdot \frac{\bar{c}}{\mathbf{1}-\boldsymbol{\beta}} + \boldsymbol{r} \right)$, where $\frac{1}{\boldsymbol{a}} = \left(\frac{1}{a_1}, \dots, \frac{1}{a_{i_{\max}}} \right)$.*

a) Let K be a compact subset of $\mathbb{R}^{i_{\max}}$. Then, $T^n[\underline{b}] \nearrow \widetilde{\mathcal{V}}_\infty$, $T^n[\mathcal{U}] \nearrow \widetilde{\mathcal{V}}_\infty$, and $T^n[\bar{b}] \searrow \widetilde{\mathcal{V}}_\infty$ as $n \to \infty$ uniformly on $\mathcal{X} \times K \times (0,1)^{i_{\max}}$.

b) $\widetilde{\mathcal{V}}_\infty$ is the unique solution of

$$\begin{cases} v = T[v], \\ v \in \Delta, \\ \underline{b}(\cdot,\cdot) \le v(x,\cdot,\cdot) \le \bar{b}(\cdot,\cdot) \text{ for all } x \in \mathcal{X}. \end{cases}$$

c) There exists a decision rule g^* that minimizes $\widetilde{\mathcal{V}}_\infty$.

d) The history-dependent policy $f^* = (f_0^*, f_1^*, \dots)$ given by

$$f_n^*(h_n) = g^*\left(x_n, \sum_{k=0}^{n-1} \boldsymbol{\beta}^k c(x_k, a_k, x_{k+1}), \boldsymbol{\beta}^n \right)$$

is an optimal policy for $\mathcal{V}_\infty$.

Proof. ad a) For $n \in \mathbb{N}$ and $(x, \boldsymbol{r}, \boldsymbol{z}) \in \widetilde{\mathcal{X}}$, it holds

$$\begin{aligned} &\mathcal{U}\left(\boldsymbol{z} \cdot \sum_{k=0}^{\infty} \boldsymbol{\beta}^k \cdot \boldsymbol{C}(x_n, a_n, x_{n+1}) + \boldsymbol{r} \right) - \mathcal{U}\left(\boldsymbol{z} \cdot \sum_{k=0}^{n} \boldsymbol{\beta}^k \cdot \boldsymbol{C}(x_n, a_n, x_{n+1}) + \boldsymbol{r} \right) \\ &\le \omega_{\mathcal{U}}\left(\left\| \boldsymbol{z} \cdot \boldsymbol{\beta}^n \sum_{k=n}^{\infty} \boldsymbol{\beta}^{k-n} \cdot \boldsymbol{C}(x_k, a_k, x_{k+1}) \right\|_2, \left\| \boldsymbol{z} \cdot \sum_{k=0}^{\infty} \boldsymbol{\beta}^k \cdot \boldsymbol{C}(x_n, a_n, x_{n+1}) + \boldsymbol{r} \right\|_2 \right) \\ &\le \omega_{\mathcal{U}}\left(i_{\max} \overline{z} \overline{\beta}^n \frac{\bar{c}}{1-\overline{\beta}}, i_{\max} \overline{z} \frac{\bar{c}}{1-\overline{\beta}} + \overline{r} \right) \end{aligned} \tag{3.15}$$

Now, we get

$$\begin{aligned}\widetilde{\mathcal{V}}_n(x,\boldsymbol{r},\boldsymbol{z}) \le \widetilde{\mathcal{I}}_{n,\widetilde{\pi}}(x,\boldsymbol{r},\boldsymbol{z}) \le \widetilde{\mathcal{I}}_{\infty,\widetilde{\pi}}(x,\boldsymbol{r},\boldsymbol{z}) &= \mathbb{E}_x^{\widetilde{\pi}}\left[\mathcal{U}\left(\boldsymbol{z}\cdot\sum_{k=0}^{\infty}\boldsymbol{\beta}^k\cdot\boldsymbol{C}(X_n,A_n,X_{n+1})+\boldsymbol{r}\right)\right]\\ &\le \mathbb{E}_x^{\widetilde{\pi}}\left[\mathcal{U}\left(\boldsymbol{z}\cdot\sum_{k=0}^{n}\boldsymbol{\beta}^k\cdot\boldsymbol{C}(X_n,A_n,X_{n+1})+\boldsymbol{r}\right)\right]+\omega_{\mathcal{U}}\left(i_{\max}\overline{z}\overline{\beta}^n\frac{\bar{c}}{1-\overline{\beta}},i_{\max}\overline{z}\frac{\bar{c}}{1-\overline{\beta}}+\overline{r}\right)\\ &\le \widetilde{\mathcal{I}}_{n,\widetilde{\pi}}(x,\boldsymbol{r},\boldsymbol{z})+\underbrace{\omega_{\mathcal{U}}\left(i_{\max}\overline{z}\overline{\beta}^n\frac{\bar{c}}{1-\overline{\beta}},i_{\max}\left(\overline{z}\frac{\bar{c}}{1-\overline{\beta}}+\overline{r}\right)\right)}_{=:\varepsilon_n(x,\boldsymbol{r},\boldsymbol{z})}.\end{aligned}$$

Since K is compact, there exists $R>0$ such that $\bar{d}<R$. Then, we have

$$\varepsilon_n(x,\boldsymbol{r},\boldsymbol{z}) \le \omega_{\mathcal{U}}\left(i_{\max}\overline{\beta}^n\frac{\bar{c}}{1-\overline{\beta}},i_{\max}\frac{\bar{c}}{1-\overline{\beta}}+R\right),$$

and since $\overline{\beta}\in(0,1)$, we have $\varepsilon_n \searrow 0$ uniformly. Because $\widetilde{\pi}$ was arbitrary, the previous inequality also holds for the infimum, i.e.

$$\widetilde{\mathcal{V}}_n(x,\boldsymbol{r},\boldsymbol{z}) \le \widetilde{\mathcal{V}}_\infty(x,\boldsymbol{r},\boldsymbol{z}) \le \widetilde{\mathcal{V}}_n(x,\boldsymbol{r},\boldsymbol{z})+\varepsilon_n(x,\boldsymbol{r},\boldsymbol{z}), \tag{3.16}$$

and therefore $\widetilde{\mathcal{V}}_n \nearrow \widetilde{\mathcal{V}}_\infty$.

Recall that $\boldsymbol{C}$ is componentwise bounded by $\underline{c},\bar{c}>0$, and therefore, independent of the process $(X_n),(A_n)$, the infinite time cost $\sum_{n=0}^{\infty}\boldsymbol{\beta}^n\cdot\boldsymbol{C}(X_n,A_n,X_{n+1})$ is componentwise bounded by $\frac{\underline{c}}{\mathbf{1}-\boldsymbol{\beta}},\frac{\bar{c}}{\mathbf{1}-\boldsymbol{\beta}}$. We have therefore $\underline{b}\le\widetilde{\mathcal{V}}_\infty\le\bar{b}$. Since T is increasing, we have with the previous result $\widetilde{\mathcal{V}}_{n+1}=T[\widetilde{\mathcal{V}}_n]\le T[\widetilde{\mathcal{V}}_\infty]$, i.e. $\widetilde{\mathcal{V}}_\infty\le T[\widetilde{\mathcal{V}}_\infty]$. Since $\boldsymbol{z}\in(0,\infty)^{i_{\max}}$, we observe that for every triple $(x,\boldsymbol{r},\boldsymbol{z})$ and $a\in\mathcal{D}(x)$, we have

$$\begin{aligned}\varepsilon_n'(x,\boldsymbol{r},\boldsymbol{z},a) &:= \int\varepsilon_n(x',\boldsymbol{z}\cdot\boldsymbol{C}(x,a,x')+\boldsymbol{r},\boldsymbol{z}\cdot\boldsymbol{\beta})\mathrm{P}(dx'\mid x,a)\\ &\le \omega\left(i_{\max}\overline{z}\overline{\beta}^{n+1}\frac{\bar{c}}{1-\overline{\beta}},i_{\max}\left(\overline{z}\overline{\beta}\frac{\bar{c}}{1-\overline{\beta}}+\overline{r}+\overline{z}\,\bar{c}\right)\right)\\ &\le \omega\left(i_{\max}\overline{z}\overline{\beta}^{n+1}\frac{\bar{c}}{1-\overline{\beta}},i_{\max}\overline{z}\frac{\bar{c}}{1-\overline{\beta}}\right)=\varepsilon_{n+1}(x,\boldsymbol{r},\boldsymbol{z}).\end{aligned} \tag{3.17}$$

Now, we get with (3.16)

$$\begin{aligned}T[\widetilde{\mathcal{V}}_\infty](x,\boldsymbol{r},\boldsymbol{z}) &\le T[\widetilde{\mathcal{V}}_n+\varepsilon_n](x,\boldsymbol{r},\boldsymbol{z}) \le T[\widetilde{\mathcal{V}}_n](x,\boldsymbol{r},\boldsymbol{z})+\sup_{a\in\mathcal{D}(x)}\varepsilon_n'(x,\boldsymbol{r},\boldsymbol{z},a)\\ &\le \widetilde{\mathcal{V}}_{n+1}(x,\boldsymbol{r},\boldsymbol{z})+\varepsilon_{n+1}(x,\boldsymbol{r},\boldsymbol{z}),\end{aligned} \tag{3.18}$$

but the last term converges to zero as $n\in\infty$. Therefore we have, $\widetilde{\mathcal{V}}_\infty\ge T[\widetilde{\mathcal{V}}_\infty]$, i.e. $\widetilde{\mathcal{V}}_\infty=T[\widetilde{\mathcal{V}}_\infty]$. We next show that $T^n[\bar{b}]\searrow\widetilde{\mathcal{V}}_\infty$, and $T^n[\underline{b}]\nearrow\widetilde{\mathcal{V}}_\infty$ as $n\to\infty$. First, observe that

$$\begin{aligned}T[\bar{b}](x,\boldsymbol{r},\boldsymbol{z}) &= \inf_{a\in D(x)}\int\mathcal{U}\left(\boldsymbol{z}\cdot\boldsymbol{\beta}\cdot\frac{\bar{c}}{1-\boldsymbol{\beta}}+\boldsymbol{z}\cdot\boldsymbol{C}(x,a,x')+\boldsymbol{r}\right)P(dx'|x,a)\\ &\le \mathcal{U}\left(\boldsymbol{z}\cdot\frac{\bar{c}}{\mathbf{1}-\boldsymbol{\beta}}+\boldsymbol{r}\right)\\ &\le \bar{b}(\boldsymbol{r},z),\end{aligned}$$

and the same holds true for

$$T[\underline{b}](x,\boldsymbol{r},z)\ge\underline{b}(\boldsymbol{r},z).$$

Since T is increasing, the sequences $(T^n[\bar{b}])_n$ and $(T^n[\underline{b}])_n$ are pointwise monotone and bounded, and therefore their pointwise limit exists. By iteration,

$$T^n[\mathcal{U}](x,\boldsymbol{r},\boldsymbol{z}) = \inf_{\pi\in\boldsymbol{\Pi}_{\widetilde{\mathcal{X}}}} \mathbb{E}_x^\pi\left[\mathcal{U}\left(\boldsymbol{z}\cdot\sum_{k=0}^{n-1}\boldsymbol{\beta}^k\cdot\boldsymbol{C}(X_k,A_k,X_{k+1})+\boldsymbol{r}\right)\right]$$

$$T^n\left[\bar{b}\right](x,\boldsymbol{r},\boldsymbol{z}) = \inf_{\pi\in\boldsymbol{\Pi}_{\widetilde{\mathcal{X}}}} \mathbb{E}_x^\pi\left[\mathcal{U}\left(\boldsymbol{z}\cdot\boldsymbol{\beta}^n\frac{\bar{c}}{\mathbf{1}-\boldsymbol{\beta}}+\boldsymbol{z}\cdot\sum_{k=0}^{n-1}\boldsymbol{\beta}^k\cdot\boldsymbol{C}(X_k,A_k,X_{k+1})+\boldsymbol{r}\right)\right].$$

We obtain

$$\begin{aligned}
0 &\le T^n[\bar{b}](x,\boldsymbol{r},\boldsymbol{z}) - T^n[\underline{b}](x,\boldsymbol{r},\boldsymbol{z}) && \text{by monotonicity of } T \text{ and } T(0)=0\\
&\le T^n[\bar{b}](x,\boldsymbol{r},\boldsymbol{z}) - T^n[\mathcal{U}](x,\boldsymbol{r},\boldsymbol{z}) && \text{by monotonicity of } T \text{ and } \mathcal{U}(\boldsymbol{r})\le \underline{b}(x,\boldsymbol{r},\boldsymbol{z})\\
&\le \sup_{\pi\in\Pi_{\widetilde{\mathcal{H}}}} \mathbb{E}_x^\pi\left[\mathcal{U}\left(\boldsymbol{z}\cdot\boldsymbol{\beta}^n\frac{\bar{c}}{\mathbf{1}-\boldsymbol{\beta}}+\boldsymbol{z}\sum_{k=0}^{n-1}\boldsymbol{\beta}^k\cdot\boldsymbol{C}(X_k,A_k,X_{k+1})+\boldsymbol{r}\right)-\mathcal{U}\left(\boldsymbol{z}\cdot\sum_{k=0}^{n-1}\boldsymbol{\beta}^k\cdot\boldsymbol{C}(X_k,A_k,X_{k+1})+\boldsymbol{r}\right.\right.\\
&= \varepsilon_n(x,\boldsymbol{r},\boldsymbol{z}).
\end{aligned}$$

For $n\to\infty$, we obtain

$$\lim_{n\to\infty} T^n[\underline{b}] = \lim_{n\to\infty} T^n[\bar{b}] = \lim_{n\to\infty} T^n[\mathcal{U}] = \widetilde{\mathcal{V}}_\infty,$$

uniformly on compact sets. This proves (a).

ad b) $\widetilde{\mathcal{V}}_\infty$ is lower semi-continuous as a uniform limit on sets of the form $\mathcal{X}\times K\times(0,1)^{i_{\max}}$, where K is a compact subset of $\mathbb{R}^{i_{\max}}$, of lower semi-continuous function. As proved in Theorem 3.7, $T^n[\bar{b}](x,\cdot,\cdot)$ is continuous and componentwise monotonous for all $x\in\mathcal{X}$. Since $T^n[\bar{b}]\searrow\widetilde{\mathcal{V}}_\infty$, $\widetilde{\mathcal{V}}_\infty(x,\cdot,\cdot)$ is upper semi-continuous and therefore continuous, and also preserves the same monotonicities for each $x\in\mathcal{X}$. We have thereby shown $\widetilde{\mathcal{V}}_\infty\in\Delta$.

It remains to show the uniqueness. To that end, suppose that there is $v\in\Delta$, $v\neq\widetilde{\mathcal{V}}_\infty$ such that $v=T[v]$ with $\underline{b}\le v\le\bar{b}$. Then, because T is increasing, $T^n[\underline{b}]\le T^n[v]=v\le T^n[\bar{b}]$, and with $n\in\infty$, we get $\widetilde{\mathcal{V}}_\infty\le v\le\widetilde{\mathcal{V}}_\infty$, i.e. $v=\widetilde{\mathcal{V}}_\infty$, a contradiction. This proves (b).

ad c) The claim follows similar to Theorem 3.7.

ad d) Since $\widetilde{\mathcal{V}}_\infty(x,y,z)\ge\mathcal{U}(y)$, we obtain

$$\widetilde{\mathcal{V}}_\infty = \lim_{n\to\infty} T_{g^*}^n[\widetilde{\mathcal{V}}_\infty] \ge \lim_{n\to\infty} T_{g^*}^n[\mathcal{U}] = \lim_{n\to\infty}\widetilde{\mathcal{I}}_n(\cdot,(g^*,g^*,\dots)) = \widetilde{\mathcal{I}}_\infty(\cdot,(g^*,g^*,\dots))\ge\widetilde{\mathcal{V}}_\infty,$$

Hence f^* is optimal for $\widetilde{\mathcal{I}}_\infty$. This finally proves (d).

4 Numerical example

4.1 Task design

To illustrate our method, we present a generalized version of the repetitive Tiger problem (Kaelbling, Littman & Cassandra, 1998). In the classic Tiger problem, a decision-maker is faced with two close doors, behind one of which is a tiger (punishment) and behind the other is a treasure (reward). In the original version, the agent can either open a door or listen to tiger sound in order to gain more information about the true place of the tiger. The sound signal however, is not fully reliable and with a smaller probability (20%) it can be heard from the wrong door. Each time that agent opens a door, it takes the reward/punishment and the problem resets to the initial

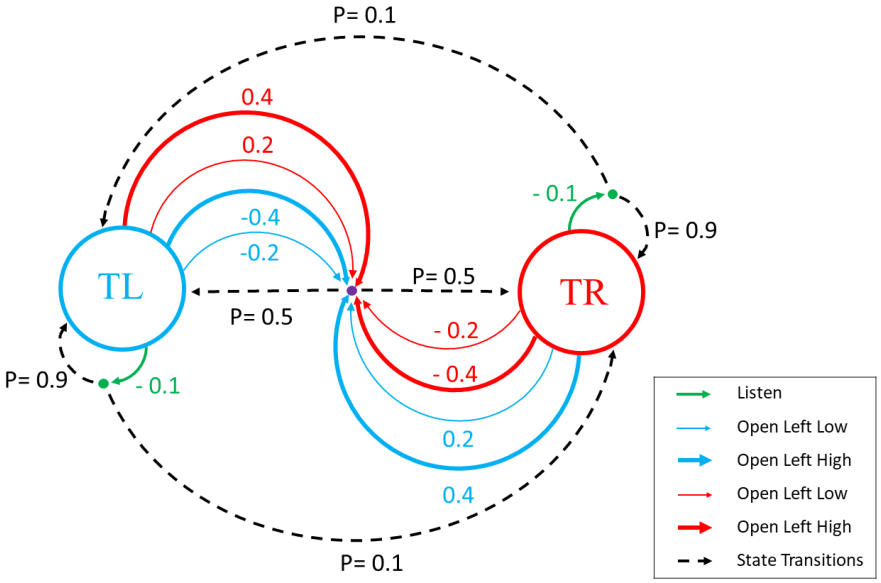

Figure 1: **MDP of the Extended Tiger task.** The states of the MDP are Tiger_right(TR) and Tiger_left(TL). After each *opening* action (red and blue arrows), the environment would set to a state randomly and based on either choosing tiger door or treasure door as well as choosing either high stake or low stake action (thickness of the arrows), the non-observable reward will receive by agent. By doing *listen* action, the agent will pay a small cost and takes an informative signal about the position of the tiger. There is also a small chance that tiger changes its position.

setting. By resetting the problem, the position of the tiger and the treasure would be determined randomly and would remain fixed for the entire next trial.

The version we will discuss here, has been generalized in three aspects: First, the constraint of deterministic state space has been relaxed and the position of the tiger can change during a trial. Here, there is a 10 % chance to change its position at the start of each epoch. In the second extension, the immediate reward (punishment) of opening the doors would not be observable for the agent during the repeats of the problem. Therefore, at any time before the end of the all trials, the agent can only have an estimation about its gains. Only at the end of all trials, the agent would observe the whole accumulated reward and punishments. And third, the action space is expanded in a way that makes the agent able to not only choose the correct option (treasure door) but also bet on its own decision in different levels of investment. Here, we define that the agent can open each door either conservatively (low stake actions), to gain a lower amount of reward and punishment (low stake rewards: $Reward_{corr\_low}$ and $Reward_{incorr\_low}$), or rush to the doors (high stake actions), to gain a bigger reward if it is the treasure, and to take more damage if the tiger is behind the door. (high stake rewards: $Reward_{corr\_High}$ and $Reward_{incorr\_High}$).

The underlying MDP of the experiment is depicted in fig 1. Here, the probabilities of getting observations (Y_{tiger_right} and Y_{tiger_left}) are depend on the agent's actions and the successive new state. As mentioned before, by doing the *listen* action there is a smaller probability that agent takes a false signal. Also, by doing each of *open* actions (regardless of their correctness or their stake type) the MDP will be reset and a random signal, with probability of 0.5 for pointing to each state, would be received. In other words, the signal which received after the re-initializing the problem, is uninformative in purpose of detecting the tiger's position. The observation function of the experiment's POMDP is shown in table 1.

Action	Next_state	y_1: Tiger_right	y_2: Tiger_left
$listen$	$tiger\_right$	0.8	0.2
$listen$	$tiger\_left$	0.2	0.8
$open\_right(low/high)$ or $open\_left(low/high)$	$tiger\_right$	0.5	0.5
$open\_right(low/high)$ or $open\_left(low/high)$	$tiger\_left$	0.5	0.5

Table 1: **Tabular representation of the Observation Function.**

4.2 Simulation results

In order to illustrate the competency of our method in replicating different risk sensitive behaviors, in this section we have tested our method on the extended tiger task by using four different utility functions which are composed by linear combinations of of exponential functions. Three out of them are adjusted to address risk-neutral, risk-seeking and risk-aversion decision-making. It means the utility functions are (near-)linear, convex and concave in the interval of possible rewards respectively. It should be mentioned that as in our method e^0 is not a valid term, linear combinations cannot replicate an exact linear function ($\frac{d^2U}{dx^2} \neq 0$). However, for the sake of being more intuitive we used a utility function with an infinitesimal second derivative in the interval of rewards to show the ability of the model to mimic different patterns of risk sensitivity (fig 2.a). Last but not least, we have tested our model on the task with a sigmoid utility function. As mentioned before, it is assumed that human uses S-shaped utility functions, like sigmoid function, in face with losses and gains. Therefore, regarding computational modeling of behavior, it is a crucial ability for a risk sensitive model to mimic such functions. Like the linear case, sigmoid function cannot be expressed by linear combinations of exponential functions. However, we can approximate the sigmoid in a specific interval by using a combination which contains enough numbers of exponential functions. Here, We fitted weighted sum of five exponential terms: $0.381 * e^{0.2906}$, $0.404 * e^{0.2876}$, $-0.427 * e^{-0.0091}$, $-0.182 * e^{0.6537}$, $0.322 * e^{0.2982}$(fig 2.b).

The simulation results have been presented in table 2. The results clearly show the effect of utility functions' shape on the risk attitude of the simulated agent. In table 2 2 we only present selected actions in trials with depth of planning equal to either one or two steps. As we have used deterministic greedy policy in our simulations, choosing between different action types (listening, low stake door openings and high stake door openings) in plannings with maximum depth of one or two only depends on environment dynamics, utility function, discount factor and initial wealth. In other words, in planning with depth one or two, choosing the type of actions(and not their directions) is independent from the observations from the environment side. Therefore, we can easily fix the other dynamics and examine only the effect of utility functions. In this simulations, we have fixed the environments dynamics to the above-mentioned values, with no discounting and the initial wealth equals to zero.

In the extended tiger problem, one can assume that high/low stake opening actions represent riskiness of the decisions. While the expected reward of them are equal, the deviation of outcomes in high stake cases are higher. Table 2 shows that in the maximum planning-depth of two, risk-neutral agent prefers to gather more information (and pay its cost) in the first step, and then in the second step open the door with higher probability of being treasure in high stake mode. However, the risk averse agent(u_3) prefers to do the second action more conservatively and open the door in the low stake mode while in contrast with them, the risk-seeker agent (u_2) prefers to perform risky actions in each step. The sigmoid-agent also behaves like the risk-neutral case, however it should be considered that S-shaped utilities make agents risk-averse toward positive accumulated outcomes and risk-averse in face with negative valuations of total expected rewards. In one step planning conditions, paying the certain cost of listening rather than doing a risky action with higher expected return seems irrational for all of used utility functions. However, risk-averse and risk-neutral cases prefer to choose low stake actions in a fifty-fifty situation while the risk-seeker and sigmoid agents prefer to risk more.

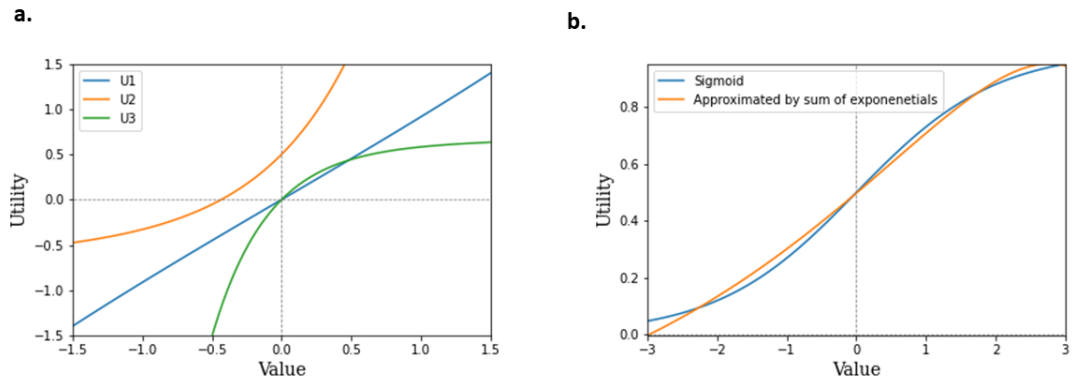

Figure 2: **Utility functions. a.** Utility functions produce risk_neutral, risk_seeking and risk_aversion decisions in the showed interval respectively. risk_neutral: $U_1 = 1.5e^{0.3} - 1.5e^{-0.3}$, risk_seeker: $U_2 = e^{1.5} - 0.5e^{-0.1}$, risk_avers: $U_3 = 0.6e^{0.05} - 0.6e^{-2.5}$. **b.** Approximation of Sigmoid function by weighted sum of five exponential functions.

5 Discussion

The method we introduced works for problems which have finite set of states while using any increasing utility function. Our method calculates the exact utility values in case of weighted sum of exponential utility functions and approximate them for any other monotone function, contrasting Bäuerle und Rieder (2017) which is more general and doesn't need to approximate values. However the resulting augmented state space in Mutlivariate utility method is $\mathcal{P}(\mathcal{X})^{i_{\max}} \times \mathcal{Y} \times \mathbb{R}^{i_{\max}} \subset \mathbb{R}^{|\mathcal{X}| \times 2 \times i_{\max}} \times \mathcal{Y}$, where $i_{\max}$ is the number of exponential functions that make up the utility function, see (1.2). In Bäuerle und Rieder (2017), the resulting state space is $\mathcal{P}(\mathcal{X} \times \mathbb{R})$ which is an infinite dimensional space, and even in cases where the wealth space is discretized appropriately, one ends up with a dimension of $|\mathcal{X}|\cdot$(partition size). Our method therefore has a clear computational advantage when $i_{\max}$ is small. In the general case of approximating utility function, the lower dimensionality of Multivariate method brings the trade-off between accuracy of approximation and computational tractability to attention. One can expect that by increasing the number of exponential terms in the approximated utility function, the accuracy of approximated utility values would improve (become more similar to their exact non-approximated values) but in the cost of an increase in state space dimensionality. Moreover by increasing the depth of planning, the method introduced by Bäuerle und Rieder (2017) would also face with the trade-off between lack of accuracy and increase of state space complexity in case of using partitioned wealth-axis. Because, when the maximum depth of planning grows the possible amounts of wealth would also increase. Both mentioned accuracy/complexity trade-offs are heavily dependent on the dynamics of the problem as well as the utility function and can be subject of further studies. Last but not least, our proposed model is eligible to apply on problems which would be defined in a multi-variate manner. In this work we only discussed the ability of the Multivariate model to address monotone utility functions in a class of problems which only have one objective (wealth), however the problems with different separate running costs like: resource allocation in different governmental sectors or maximizing the overall utility of an economic actor while she uses different diminishing marginal utility functions in different goods or aspects of life are another area that our method can address and could be investigated more in terms of computational efficiency.

	Depth: 1	Depth: 2	
	Step 1	Step 1	step2
U_1: risk-neutral	Low	Listen	**High**
U_2: risk-seeking	**High**	**High**	**High**
U_3: risk-averse	Low	Listen	Low
U_4: Sigmoid	**High**	Listen	**High**

Table 2: **Actions with best value for each step under different utility functions**

Statements and Declarations

The funding of each author of this research has been declared on the footer of the first page. The Authors have no competing interests to declare.

References

Al-Nowaihi, A., Bradley, I. & Dhami, S. (2008). A note on the utility function under prospect theory. *Economics letters*, *99* (2), 337–339.

Baras, J. S. & James, M. R. (1997). Robust and Risk-Sensitive Output Feedback Control for Finite State Machines and Hidden Markov Models. *Journal of Mathematics, Systems, Estimation and Control*, *7* (3), 371–374.

Bäuerle, N. & Rieder, U. (2017). Partially observable risk-sensitive stopping problems in discrete time. *arXiv preprint arXiv:1703.09509*.

Bertram, L., Schulz, E. & Nelson, J. D. (2021). Subjective probability is modulated by emotions.

Borkar, V. S. & Meyn, S. P. (2002). Risk-Sensitive Optimal Control for Markov Decision Processes with Monotone Cost. *Mathematics of Operations Research*, *27* (1), 192–209.

Bäuerle, N. & Rieder, U. (2011). *Markov Decision Processes with Applications to Finance*. Berlin, Heidelberg: Springer.

Bäuerle, N. & Rieder, U. (2014). More Risk-Sensitive Markov Decision Processes. *Mathematics of Operations Research*, *39* (1), 105–120.

Bäuerle, N. & Rieder, U. (2017). Partially Observable Risk-Sensitive Markov Decision Processes. *Mathematics of Operations Research*, *42* (4), 1180–1196.

Cavazos-Cadena, R. (2010). Optimality equations and inequalities in a class of risk-sensitive average cost Markov decision chains. *Mathematical Methods of Operations Research*, *71* (1), 47–84.

Cavazos-Cadena, R. & Hernández-Hernández, D. (2005). Successive approximations in partially observable controlled Markov chains with risk-sensitive average criterion. *Stochastics*, *77* (6), 537–568.

Chung, K.-J. & Sobel, M. J. (1987). Discounted MDP's: Distribution Functions and Exponential Utility Maximization. *SIAM Journal on Control and Optimization*, *25* (1), 49–62.

Di Masi, G. B. & Stettner, L. (2007). Infinite Horizon Risk Sensitive Control of Discrete Time Markov Processes under Minorization Property. *SIAM Journal on Control and Optimization*, *46* (1), 231–252.

Dupuis, P., Laschos, V. & Ramanan, K. (2019). Exit time risk-sensitive control for systems of cooperative agents. *Mathematics of Control, Signals, and Systems*, *31* (3), 279–332.

Edwards, W. (1954). The theory of decision making. *Psychological bulletin*, *51* (4), 380.

Fan, J. & Ruszczyński, A. (2018). Risk measurement and risk-averse control of partially observable discrete-time Markov systems. *Mathematical Methods of Operations Research*, *88* (2), 161–184.

Fernandez-Gaucherand, E. & Marcus, S. I. (1997). Risk-sensitive optimal control of hidden Markov models: Structural results. *IEEE Transactions on Automatic Control*, *42* (10), 1418–1422.

Fleming, W. H. & Hernández-Hernández, D. (1997). Risk-Sensitive Control of Finite State Machines on an Infinite Horizon I. *SIAM Journal on Control and Optimization*, *35* (5), 1790–1810.

Hernández-Hernández, D. (1999). Partially Observed Control Problems with Multiplicative Cost. In W. M. McEneaney, G. Yin & Q. Zhang (Hrsg.), *Stochastic Analysis, Control, Optimization and Applications* (S. 41–55). Basel: Birkhäuser.

Hernández-Hernández, D. & Marcus, S. I. (1996). Risk sensitive control of Markov processes in countable state space. *Systems & Control Letters*, *29* (3), 147–155.

Hernández-Lerma, O. & Lasserre, J.-B. (1996). *Discrete-Time Markov Control Processes: Basic Optimality Criteria* (Bd. 30). New York: Springer.

Hinderer, K. (1970). *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter* (Bd. 33). Berlin, Heidelberg: Springer.

Howard, R. A. & Matheson, J. E. (1972). Risk-Sensitive Markov Decision Processes. *Management Science*, *18* (7), 356–369.

Jaquette, S. C. (1973). Markov Decision Processes with a New Optimality Criterion: Discrete Time. *The Annals of Statistics*, *1* (3), 496–505.

Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101* (1-2), 99–134.

Kahneman, D. & Tversky, A. (1979). On the interpretation of intuitive probability: A reply to jonathan cohen.

Kahneman, D. & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part i* (S. 99–127). World Scientific.

Kalyanaram, G. & Winer, R. S. (1995). Empirical generalizations from reference price research. *Marketing science*, *14* (3_supplement), G161–G169.

Levitt, S. & Ben-Israel, A. (2001). On Modeling Risk in Markov Decision Processes. In A. M. Rubinov & B. M. Glover (Hrsg.), *Optimization and Related Topics* (Bd. 47, S. 27–40). Springer US.

Marecki, J. & Varakantham, P. (2010). Risk Sensitive Planning in Partially Observable Environments. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (S. 1357–1368). Toronto, Canada.

Markowitz, H. (1952). The utility of wealth. *Journal of political Economy*, *60* (2), 151–158.

Mosteller, F. & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, *59* (5), 371–404.

Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5* (4), 297–323.

Von Neumann, J. & Morgenstern, O. (1947). Theory of games and economic behavior, 2nd rev.