

# Controlled Markov Decision Processes with AVaR Criteria for Unbounded Costs

Kerem Uğurlu

Tuesday 29<sup>th</sup> March, 2016

Department of Mathematics, University of Southern California, Los Angeles, CA 90089  
e-mails:kugurlu@usc.edu

## Abstract

In this paper, we consider the control problem with the Average-Value-at-Risk (AVaR) criteria of the possibly unbounded  $L^1$ -costs in infinite horizon on a Markov Decision Process (MDP). With a suitable state aggregation and by choosing a priori a global variable  $s$  heuristically, we show that there exist optimal policies for the infinite horizon problem for possibly unbounded costs.

*Mathematics Subject Classification:* 90C39, 93E20

*Keywords:* Markov Decision Problem, Average-Value-at-Risk, Optimal Control;

## 1 Introduction

In classical models, the optimization problem has been solved by expected performance criteria. Beginning with Bellman [6], risk neutral performance evaluation has been used via dynamic programming techniques. This methodology has seen huge development both in theory and practice since then (see e.g. [28, 29, 30, 31, 32, 33]). However, in practice expected values are not appropriate to measure the performance criteria. Due to that, risk averse approaches have been begun to forecast the corresponding problem and its outcomes specifically by utility functions (see e.g. [8, 10]). To put risk-averse preferences into an axiomatic framework, with the seminal paper of Artzner et al. [2], the risk assessment gained new aspects for random outcomes. In [2], the concept of *coherent*

*risk measure* has been defined and theoretical framework has been established. Deriving dynamic programming equations for this type of risk-averse operators, risk measures, are not vast. The reason for it is that the Bellman optimality principle is not necessarily true using this type of operators. That is to say the optimization problems are not *time consistent*. We refer the reader to [27] for examples verifying this type of inconsistency. A multistage stochastic decision problem is time-consistent, if resolving the problem at later stages (i.e., after observing some random outcomes), the original solutions remain optimal for the later stages. To overcome this difficulty, in [19], one time step Markovian dynamic risk measures are introduced, hence the operators are only evaluating for one time step and necessarily time-consistent. Another state aggregation method and relevant algorithms are developed in [26] relying on a so-called AVaR decomposition theorem. This approach uses a dual representation of AVaR and hence requires optimization over a space of probability densities when solving an associated Bellman equation. In [4], a different approach is applied and for each path  $\omega$ , the information necessary from the previous time steps is included in the current decision, which is called *state aggregation* in [4]. All these works are studying *bounded* costs in  $L^\infty$ , hence whenever they study infinite time horizon, they verify the existence of optimal policy via a contraction mapping and fixed point argument. To the best of our knowledge, there does not exist a study on the optimal control on MDPs with unbounded costs using coherent risk measures. This paper examines this case.

Our contributions are two fold. First, using the state aggregation idea from [4], we show that in infinite time horizon with possibly unbounded costs that are in  $L^1$ , there exists an optimal stationary policy. Second, we propose a heuristic algorithm to compute the optimal values that is applicable both on continuous and discrete probability spaces that require no technical conditions on the type of distributions as opposed to [4]. We present our results with a numerical example and show that the simulations are consistent with original problem and theoretical expected behaviour of this type of operator.

The rest of the paper is as follows. In Section 2, we give the preliminary theoretical framework. In Section 3, we state our main result and derive the dynamic programming equations for MDP using AVaR criteria for the infinite time horizon. In Section 4 we present an algorithm using our theoretical results and apply it to the classical LQ problem and give the simulation values.

## 1.1 Controlled Markov Decision Processes

We take the control model  $\mathcal{M} = \{\mathcal{M}_n, n \in \mathbb{N}_0\}$ , where for each  $n \in \mathbb{N}_0$ ,

$$\mathcal{M}_n := (X_n, A_n, \mathbb{K}_n, F_n, c_n) \quad (1.1)$$

with the following components:

- $X_n$  and  $A_n$  denote the state and action (or control) spaces, where  $X_n$  take values in a Borel set  $X$  whereas  $A_n$  take values in a Borel set  $A$ .
- For each  $x \in X_n$ , let  $A_n(x) \subset A_n$  be the set of all admissible controls in the state  $x_n = x$ . Then

$$\mathbb{K}_n := \{(x, a) : x \in X_n, a \in A_n(x)\}, \quad (1.2)$$

stands for the set of feasible state-action pairs at time  $n$ , where we assume that  $\mathbb{K}_n$  is a Borel subset of  $X_n \times A_n$ .

- We let  $x_{n+1} = F_n(x_n, a_n, \xi_n)$ , for all  $n = 0, 1, \dots$  with  $x_n \in X_n$  and  $a_n \in A_n$  as described above, with independent random disturbances  $\xi_n \in S_n$  having probability distributions  $\mu_n$ , where the  $S_n$  are Borel spaces.
- $c_n(x, a) : \mathbb{K}_n \rightarrow \mathbb{R}$  stands for the deterministic cost function at stage  $n \in \mathbb{N}_0$  with  $(x, a) \in \mathbb{K}_n$ .

The random variables  $\{\xi_n\}_{n \geq 0}$  are defined on a common probability space  $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n \geq 0}, \mathbb{P})$ , where  $\mathbb{P}$  is the reference probability space with each  $\xi_n$  measurable with respect to sigma algebra  $\mathcal{F}_n$  with  $\mathcal{F} = \sigma(\cup_{n=0}^{\infty} \mathcal{F}_n)$ . Based on the action  $a \in \mathbb{K}_n(x)$  chosen at time  $n$ , we assume that  $A_n$  is  $\mathcal{F}_n = \sigma(X_0, A_0, \dots, X_n)$ -measurable, i.e. our decision might depend entirely on the history  $h_n$ , where  $h_n = (x_0, a_0, x_1, \dots, a_{n-1}, x_n) \in H_n$  is the history up to time  $n$ , where define recursively

$$H_0 := X, \quad H_{n+1} := H_n \times A \times X \quad (1.3)$$

For each  $n \in \mathbb{N}_0$ , let  $\mathbb{F}_n$  be the family of measurable functions  $f_n : H_n \rightarrow A_n$  such that

$$f_n(x) \in A_n(x), \quad (1.4)$$

for all  $x \in X_n$ . A sequence  $\pi = \{f_n\}$  of functions  $f_n \in \mathbb{F}_n$  for all  $n \in \mathbb{N}_0$  is called a policy. We denote by  $\Pi$  the set of all the policies. Then for each policy  $\pi \in \Pi$  and initial state  $x \in \mathbf{X}$ , a stochastic process  $\{(x_n, a_n)\}$  and a probability measure  $\mathbb{P}_x^\pi$  is defined on  $(\Omega, \mathcal{F})$

in a canonical way, where  $x_n$  and  $a_n$  represent the state and the control at time  $n \in \mathbb{N}_0$ . The expectation operator with respect to  $\mathbb{P}_x^\pi$  is denoted by  $\mathbb{E}_x^\pi$ . The distribution of  $X_{n+1}$  is given by the transition kernel  $\mathbb{Q}$  from  $X \times A$  to  $X$  as follows:

$$\begin{aligned} P^\pi(X_{n+1} \in B_x | X_0, \dots, X_n, f_n(X_0, A_0, \dots, X_n)) \\ &= P^\pi(X_{n+1} \in B_x | X_n, f_n(X_0, A_0, \dots, X_n)) \\ &= \mathbb{Q}(B_x | X_n, f_n(X_0, A_0, \dots, X_n)) \end{aligned}$$

for Borel measurable sets  $B_x \subset \mathbf{X}$ . A Markov policy is of the form

$$P^\pi(X_{n+1} \in B_x | X_n, f_n(X_0, A_0, \dots, X_n)) = \mathbb{Q}(B_x | X_n, f_n(X_n)). \quad (1.5)$$

That is to say, the Markov policy  $\pi = \{f_n\}_{n \geq 0}$  depends only on current state  $X_n$ . We denote the set of all Markovian policies as  $\Pi^M$ . Similarly, the stationary policy is of the form  $\pi = \{f, f, f, \dots\}$  with

$$P^\pi(X_{n+1} \in B_x | X_n, f_n(X_0, A_0, \dots, X_n)) = \mathbb{Q}(B_x | X_n, f(X_n)), \quad (1.6)$$

i.e. we apply the same rule for each time episode  $n$ . Suppose, we are given a policy  $\pi = \{f_n\}_{n=0}^\infty$ , then by Ionescu-Tulcea theorem [7], there exists a unique probability measure  $P^\pi$  on  $(\Omega, \mathcal{F})$ , which ensures the consistency of the infinite horizon problem considered. Hence, for every measurable set  $B \subset \mathcal{F}_n$ ,  $n \in \mathbb{N}_0$ , we denote

$$\begin{aligned} P^\pi(X_1 \in B) &= P(B) \\ P^\pi(X_{n+1} \in B | h_n) &= \mathbb{Q}(B | X_n, f_n(X_0, A_0, \dots, X_n)) \end{aligned}$$

We consider the following cost function  $c_n(x_n, a_n)$  that takes state  $x_n$  and action  $a_n$  at time  $n \in \mathbb{N}_0$  and denote

$$C^\infty := \sum_{n=0}^{\infty} c_n(x_n, a_n), \quad (1.7)$$

for the infinite planning horizon and

$$C^N := \sum_{n=0}^N c_n(x_n, a_n) \quad (1.8)$$

for the finite planning horizon for some terminal time  $N \in \mathbb{N}_0$ . We take that the cost functions  $\{c_n(x_n, a_n)\}_{n \geq 0}$  are non-negative and  $C^N$  and  $C^\infty$  belong to space  $L^1(\Omega, \mathcal{F}, \mathbb{P}_0)$ . We start from the following two well-studied optimization problems for controlled Markov

processes. The first one is called *finite horizon expected value problem*, where we want to find a policy  $\pi = \{f_n\}_{n=0}^N$  with the minimization of the expected cost:

$$\min_{\pi \in \Pi} \mathbb{E}_x^\pi \left[ \sum_{n=0}^N c_n(x_n, a_n) \right]$$

where  $a_n = f_n(x_0, x_1, \dots, x_n)$  and  $c_n(x_n, a_n)$  is measurable for each  $n = 0, \dots, N$ . The second problem is the infinite horizon expected value problem. The objective is to find a policy  $\pi = \{f_n\}_{n=0}^\infty$  with the minimization of the expected cost:

$$\min_{\pi \in \Pi} \mathbb{E}_x^\pi \left[ \sum_{n=0}^{\infty} c_n(x_n, a_n) \right]$$

Under some assumptions the first optimization problem has solution in form of Markov policies, whereas in infinite case the policy is stationary. In both cases, the optimal policies can be found by solving corresponding dynamic programming equations. Our goal is to study the infinite horizon problem, where we use a *risk-averse operator*  $\rho$  instead of the expectation operator and look for stationary optimal policy under some conditions.

We introduce the corresponding risk averse operators that we will be working on throughout the rest of the paper, which is first defined in [2] on essentially bounded random variables in  $L^\infty$  and later extended to random variables on  $L^1$  in [17, 19] with a norm on  $L^1$  introduced in [28].

**Definition 1.1.** *A function  $\rho : L^1 \rightarrow \mathbb{R}$  is said to be a coherent risk measure if it satisfies the following axioms:*

- $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda \rho(X) + (1 - \lambda)\rho(Y) \quad \forall \lambda \in (0, 1), X, Y \in L^1;$
- *If  $X \leq Y$   $\mathbb{P}$ -a.s. then  $\rho(X) \leq \rho(Y)$ ,  $\forall X, Y \in L^1$*
- $\rho(c + X) = c + \rho(X)$ ,  $\forall c \in \mathbb{R}, X \in L^1;$
- $\rho(\beta X) = \beta \rho(X)$ ,  $\forall X \in L^1, \beta \geq 0$ .

The particular risk averse operator that we will be working with is the  $\text{AVaR}_\alpha(X)$

**Definition 1.2.** *Let  $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$  be a real-valued random variable and let  $\alpha \in (0, 1)$ .*

- We define the *Value-at-Risk* of  $X$  at level  $\alpha$ ,  $\text{VaR}_\alpha(X)$ , by

$$\text{VaR}_\alpha(X) = \inf \{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq \alpha\} \quad (1.9)$$

- We define the coherent risk measure, the *Average-Value-at-Risk* of  $X$  at level  $\alpha$ , denoted by  $\text{AVaR}_\alpha(X)$  as

$$\text{AVaR}_\alpha(X) = \frac{1}{1-\alpha} \int_\alpha^1 \text{VaR}_t(X) dt \quad (1.10)$$

We will also need the following alternative representation for  $\text{AVaR}_\alpha(X)$  as shown in [15].

**Lemma 1.1.** *Let  $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$  be a real-valued random variable and let  $\alpha \in (0, 1)$ . Then it holds that*

$$\text{AVaR}_\alpha(X) = \min_{s \in \mathbb{R}} \left\{ s + \frac{1}{1-\alpha} \mathbb{E}[(X - s)^+] \right\}, \quad (1.11)$$

where the minimum is attained at  $s = \text{VaR}_\alpha(X)$ .

**Remark 1.2.** *We note from the representation above that the  $\text{AVaR}_\alpha(X)$  is real-valued for any  $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ .*

**Definition 1.3.** *Let  $L^0(\mathcal{F}_n)$  be the vector space of all real-valued,  $\mathcal{F}_n$ -measurable random variables on the space  $(\Omega, \mathcal{F}, \mathcal{F}_n, \mathbb{P})$  defined above. A one-step coherent dynamic risk measure on  $L^0(\mathcal{F}_{n+1})$  is a sequence of mappings such that*

$$\rho_t : L^0(\mathcal{F}_{n+1}) \rightarrow L^0(\mathcal{F}_n), n = 0, \dots, N - 1. \quad (1.12)$$

that satisfy the followings

- $\rho_n(\lambda X + (1 - \lambda)Y) \leq \lambda \rho_n(X) + (1 - \lambda)\rho_n(Y) \quad \forall \lambda \in (0, 1), Z, W \in L^0(\mathcal{F}_{n+1});$
- *If  $X \leq Y$   $\mathbb{P}$ -a.s. then  $\rho_{n+1}(X) \leq \rho_{n+1}(Y)$ ,  $\forall X, Y \in L^0(\mathcal{F}_{n+1})$*
- $\rho_n(c + X) = c + \rho_n(X)$ ,  $\forall c \in L^0(\mathcal{F}_n)$ ,  $X \in L^0(\mathcal{F}_{n+1});$
- $\rho_n(\beta X) = \beta \rho_n(X)$ ,  $\forall X \in L^0(\mathcal{F}_n)$ ,  $\beta \geq 0$

**Definition 1.4.** *A dynamic risk measure  $(\rho_n)_{n=0}^{N-1}$  on  $L^0(\mathcal{F}_N)$  is called time-consistent if for all  $X, Y \in L^0(\mathcal{F}_N)$  and  $n = 0, \dots, N - 1$ ,  $\rho_{n+1}(X) \geq \rho_{n+1}(Y)$  implies  $\rho_n(X) \geq \rho_n(Y)$ .*

Another way to define time consistency is from the point of view of optimal policies (see also [20]). Intuitively, the sequence of optimization problems is said to be dynamically consistent, if the optimal strategies obtained when solving the original problem at time  $t$  remain optimal for all subsequent problems. More precisely, if a policy  $\pi$  is optimal on the time interval  $[s, T]$ , then it is also optimal on the sub-interval  $[t, T]$  for every  $t$  with  $s \leq t \leq T$ .

**Remark 1.3.** Given that the probability space is atomless, it is shown in [20] and [14] that the only law invariant coherent risk measure operators  $\rho$  on  $(\Omega, \mathcal{F}, \mathcal{F}_{n=0}^N, \mathbb{P})$ , i.e.

$$X \stackrel{d}{=} Y \Rightarrow \rho(X) = \rho(Y) \quad (1.13)$$

satisfying

$$\rho(Z) = \rho(\rho|\mathcal{F}_1(\dots\rho|\mathcal{F}_{N-1})(Z)), \quad (1.14)$$

for all random variables  $Z$  are  $\text{esssup}(Z)$  and expectation  $\mathbb{E}(Z)$  operators. This suggests that optimization problems with most of the coherent risk measures are not time consistent.

## 2 Main Result

We are interested in solving the following optimization problem in the infinite horizon.

$$\min_{\pi \in \Pi} \text{AVaR}_\alpha^\pi \left( \sum_{n=0}^{\infty} c_n(x_n, a_n) \right), \quad (2.15)$$

**Remark 2.1.** In [4], the infinite horizon with bounded costs are studied and the existence of optimal strategy is obtained via a fixed point argument through contraction mapping. Here, since we deal with cost functions that are in  $L^1$ , this scheme does not work. Hence, we follow a different approach by assuming that there exists at least one policy  $\pi_0$  for which  $\text{AVaR}_\alpha^{\pi_0}(\sum_{n=0}^{\infty} c_n(x_n, a_n)) < \infty$ . The complete assumptions are listed as follows.

First, we put the following assumptions on the problem.

**Assumption 2.2.** For every  $n \in \mathbb{N}_0$ , we impose the following assumptions on the problem:

1. The cost function  $c_n : \mathbb{K}_n \rightarrow \mathbb{R}$  is nonnegative, lower semicontinuous (l.s.c.), that is if  $(x_k, a_k) \rightarrow (x, a)$ , then

$$\liminf_{k \rightarrow \infty} c_n(x^k, a^k) \geq c_n(x, a) \quad (2.16)$$

and inf-compact on  $\mathbb{K}_n$ , i.e., for every  $x \in X_n$  and  $r \in \mathbb{R}$ , the set

$$\{a \in A_n(x) | c_n(x, a) \leq r\} \quad (2.17)$$

is compact.

2. For fixed  $s \in \mathbb{R}$  and for all  $n \in \mathbb{N}_0$ , the function

$$(x_n, a_n) \rightarrow \int v(x_{n+1}, s - c_{n+1}(x_{n+1}, a)) \mathbb{Q}(dx_{n+1} | x_n, a_n) \quad (2.18)$$

is l.s.c. for all l.s.c. functions  $v \geq 0$ .

3. The set  $A_n(x)$  is compact for every  $x \in X_n$  and for every  $n \in \mathbb{N}_0$ .
4. The system function  $x_{n+1} = F_n(x_n, a_n, \xi_n)$  is measurable as a mapping  $F_n : X_n \times A_n \times S_n \rightarrow X_{n+1}$ , and  $(x, a) \rightarrow F_n(x, a, s)$  is continuous on  $\mathbb{K}_n$  for every  $s \in S_n$ .
5. The multifunction also called point-to-set function  $x \rightarrow A_n(x)$ , from  $\mathbf{X}$  to  $A$  is upper semicontinuous (u.s.c.) that is, if  $\{x_n\} \subset \mathbb{X}$  and  $\{a_n\} \subset A$  are sequences such that

$$x_n \rightarrow x^*, \quad a_n \in A(x_n) \quad \text{for all } n, \quad \text{and} \quad a_n \rightarrow a^*, \quad (2.19)$$

then  $a^*$  is in  $A_n(x^*)$ .

6. There exists a policy  $\pi \in \Pi$  such that  $V_0(x, \pi) < M$  for all  $x \in X_0$ .

To solve (2.15), we first rewrite the infinite horizon problem as follows:

$$\begin{aligned} \inf_{\pi} \text{AVaR}_{\alpha}^{\pi}(C^{\infty} | X_0 = x) &= \inf_{\pi \in \Pi} \inf_{s \in \mathbb{R}} \left\{ s + \frac{1}{1 - \alpha} \mathbb{E}_x^{\pi}[(C^{\infty} - s)^+] \right\} \\ &= \inf_{s \in \mathbb{R}} \inf_{\pi \in \Pi} \left\{ s + \frac{1}{1 - \alpha} \mathbb{E}_x^{\pi}[(C^{\infty} - s)^+] \right\} \\ &= \inf_{s \in \mathbb{R}} \left\{ s + \frac{1}{1 - \alpha} \inf_{\pi \in \Pi} \mathbb{E}_x^{\pi}[(C^{\infty} - s)^+] \right\} \end{aligned}$$

Based on this representation, we investigate the inner optimization problem for finite time  $N$  as in [4]. Let  $n = 0, 1, 2, \dots, N$ . We define

$$w_{N\pi}(x, s) := \mathbb{E}_x^{\pi}[(C^N - s)^+], \quad x \in X, \quad s \in \mathbb{R}, \quad \pi \in \Pi, \quad (2.20)$$

$$w_N(x, s) := \inf_{\pi \in \Pi} w_{N\pi}(x, s), \quad x \in X, \quad s \in \mathbb{R}, \quad (2.21)$$

We work with the Markov Decision Model with a 2-dimensional state space  $\tilde{X} \triangleq X \times \mathbb{R}$ . The second component of the state  $(x_n, s_n) \in \tilde{X}$  gives the relevant information of the history of the process. We take that there is no running cost and we assume that the terminal cost function is given by  $V_{-1\pi}(x, s) := V_{-1}(x, s) := s^-$ . We take the decision rules  $f_n : \tilde{X} \rightarrow A$  such that  $f_n(x, s) \in A_n(x)$  and denote by  $\Pi^M$  the set of Markov policies  $\pi = (f_0, f_1, \dots)$ , where  $f_n$  are decision rules. Here, by Markov policy, we mean that the decision at time  $n$  depends only on the current state  $x$  and as well as on the global variable  $s$  as to be seen in the proof below. We denote for

$$v \in \mathbb{M}(\tilde{X}) := \{v : \tilde{X} \rightarrow \mathbb{R}_+ : \text{measurable}\} \quad (2.22)$$



the operators, for  $n \in \mathbb{N}_0$  and for fixed  $s$

$$Lv(x_n, a) := \int v(x_{n+1}, s - c_n(x_{n+1}, a)) \mathbb{Q}(dx_{n+1} | x + n, a), \quad (x_n, s) \in \tilde{X}, a \in A_n(x) \quad (2.23)$$

and

$$T_f v(x_n) := \int v(x_{n+1}, s - c_n(x_{n+1})) \mathbb{Q}(dx_{n+1} | x_n, f(x_n, s)), \quad (x_n, s) \in \tilde{X}$$

The minimal cost operator of the Markov Decision Model is given by

$$Tv(x) = \inf_{a \in A_n(x)} Lv(x, s, a). \quad (2.24)$$

For a policy  $\pi = (f_0, f_1, f_2, \dots) \in \Pi^M$ . We denote by  $\vec{\pi} = (f_1, f_2, \dots)$  the shifted policy. We define for  $\pi \in \Pi^M$  and  $n = -1, 0, 1, \dots, N$ :

$$\begin{aligned} V_{n+1, \pi} &:= T_{f_0} V_{n\pi}, \\ V_{n+1} &:= \inf_{\pi} V_{n+1\pi} \\ &= TV_n. \end{aligned}$$

A decision rule  $f_n^*$  with the property that  $V_n = T_{f_n^*} V_{n-1}$  is called the minimizer of  $V_n$ . We have *Markovian* policies  $\Pi^M \subset \Pi$  in the following sense: Given the global variable  $s$ , for every  $\sigma = (f_0, f_1, \dots) \in \Pi^M$  we find a policy  $\pi = (g_0, g_1, \dots) \in \Pi$  such that

$$\begin{aligned} g_0(x_0) &:= f_0(x_0, s) \\ g_1(x_0, a_0, x_1) &:= f_1(x_1, s - c_0) \\ &\vdots \\ &\vdots \end{aligned}$$

We remark here that a *Markovian* policy  $\sigma = (f_0, f_1, \dots) \in \Pi^M$  also depends on the history of the process but not on the whole information. The necessary information at time  $n$  of the history  $h_n = (x_0, a_0, x_1, \dots, a_{n-1}, x_n)$  are the state  $x_n$  and the necessary information  $s_n \triangleq s_0 - c_0 - c_1 - \dots - c_{n-1}$ . This dependence of the past and the optimality of the Markovian policy is shown in the following theorem. For convenience, we denote

$$\begin{aligned} V_{0,N}^*(x) &:= \inf_{\pi \in \Pi} \mathbb{E}_x^\pi[(C^N - s)^+] \\ V_{0,\infty}^*(x) &:= \inf_{\pi \in \Pi} \mathbb{E}_x^\pi[(C^\infty - s)^+] \end{aligned}$$

, which corresponds to the optimal value starting at state  $x$  in finite and infinite time horizon accordingly.

**Theorem 2.3.** [4] For a given policy  $\sigma$ , the only necessary information at time  $n$  of the history  $h_n = (x_0, a_0, x_1, \dots, a_{n-1}, x_n)$  are the followings

- the state  $x_n$
- the value  $s_n := s - c_0 - c_1 - \dots - c_{n-1}$  for  $n = 1, 2, \dots, N$ .

Moreover, it holds for  $n = 0, 1, \dots, N$  that

- $w_{n\sigma} = V_{n\sigma}$  for  $\sigma \in \Pi^M$ .
- $w_n = V_n$

If there exist minimizers  $f_n^*$  of  $V_n$  on all stages, then the Markov policy  $\sigma^* = (f_0^*, \dots, f_N^*)$  is optimal for the problem

$$\inf_{\pi \in \Pi} \mathbb{E}_x^\pi[(C^N - s)^+] \quad (2.25)$$

*Proof.* For brevity, suppressing the arguments for cost function  $c_n(x, a)$ , and for  $n = 0$ , we obtain

$$\begin{aligned} V_{0\sigma}(x, s) &= T_{f_0} V_{-1}(x, s) \\ &= \int V_{-1}(x', s - c_0) \mathbb{Q}(dx'|x, f_0(x, s)) \\ &= \int (s - c_0)^- \mathbb{Q}(dx'|x, f_0(x, s)) \\ &= \int (c_0 - s)^+ \mathbb{Q}(dx'|x, f_0(x, s)) \\ &= \mathbb{E}_x^\pi[(C_0 - s)^+] = w_{0\sigma}(x, s) \end{aligned}$$

Next by induction argument and denoting  $f_0(x, s) = a$ , we have

$$\begin{aligned} V_{n+1\sigma}(x, s) &= T_{f_0} V_{n\bar{\sigma}}(x, s) \\ &= \int V_{n\bar{\sigma}}(x', s - c_n(y, a)) \mathbb{Q}(dy|x, a) \\ &= \int \mathbb{E}_{x'}^{\bar{\sigma}}[(C^n - (s - c_{n+1}(y, a)))^+] \mathbb{Q}(dy|x, a) \\ &= \int \mathbb{E}_{x'}^{\bar{\sigma}}[(c_{n+1}(y, a) + C^n - s)^+] \mathbb{Q}(dy|x, a) \\ &= \mathbb{E}_x^\sigma[(C^{n+1} - s)^+] = w_{n+1\sigma}(x, s) \end{aligned}$$

We note that the history of the Markov Decision Process  $\tilde{h}_n = (x_0, s_0, a_0, x_1, s_1, a_1, \dots, x_n, s_n)$  contains history  $h_n = (x_0, a_0, x_1, a_1, \dots, x_n)$ . We denote by  $\tilde{\Pi}$  the history dependent policies of the Markov Decision Process. By ([5], Theorem 2.2.3), we get

$$\inf_{\sigma \in \Pi^M} V_{n\sigma}(x, s) = \inf_{\tilde{\pi} \in \tilde{\Pi}} V_{n\tilde{\pi}}(x, s).$$

Hence, we obtain

$$\inf_{\sigma \in \Pi^M} w_{n\sigma} \geq \inf_{\pi \in \Pi} w_{n\pi} \geq \inf_{\tilde{\pi} \in \tilde{\Pi}} = \inf_{\sigma \in \Pi^M} V_{n\sigma} = \inf_{\sigma \in \Pi^M} w_{n\sigma}$$

We conclude the proof.  $\square$

**Theorem 2.4.** [4] *Under the conditions of the Assumptions 2.1, there exists an optimal Markov policy, in the sense introduced above,  $\sigma^* \in \Pi$  for any finite horizon  $N \in \mathbb{N}_0$  with*

$$\inf_{\pi \in \Pi} \mathbb{E}_x^\pi[(C^N - s)^+] = \mathbb{E}_x^{\sigma^*}[(C^N - s)^+] \quad (2.26)$$

Now we are ready to state our main result.

**Theorem 2.5.** *Under Assumptions 2.1, there exists an optimal Markov policy  $\pi^*$  for the infinite horizon problem 2.15.*

*Proof.* For the policy  $\pi \in \Pi$  stated in the Assumption 2.1, we have

$$\begin{aligned} w_{\infty, \pi} &= \mathbb{E}_x^\pi[(C^\infty - s)^+] \\ &= \mathbb{E}_x^\pi[(C^n + \sum_{k=n+1}^{\infty} C_k - s)^+] \\ &\leq E_x^\pi[(C^n - s)^+] + E_x^\pi[\sum_{k=n+1}^{\infty} C_k], \\ &\leq E_x^\pi[(C^n - s)^+] + M(n), \end{aligned} \quad (2.27)$$

where  $M(n) \rightarrow 0$  as  $n \rightarrow \infty$  due to the Assumption 2.1. Taking the infimum over all  $\pi \in \Pi$  we get

$$w_\infty(x, s) \leq w_n + M(n) \quad (2.28)$$

Hence we get

$$w_n \leq w_\infty(x, s) \leq w_n + M(n) \quad (2.29)$$

Letting  $n \rightarrow \infty$ , we get

$$\lim_{n \rightarrow \infty} w_n = w_\infty \quad (2.30)$$

Moreover, by Theorem 2.3, there exists  $\pi^* = \{f_n\}_{n=0}^N \in \Pi$  such that  $V_\pi^N(x) = V_{0,N}^*(x)$  and we also have by the assumption that  $V_\pi^N(x)$  is l.s.c. By the nonnegativity of the cost functions  $c_n \geq 0$ , we have that  $N \rightarrow V_{0,N}^*(x)$  is nondecreasing and  $V_{0,N}^*(x) \leq V_{0,\infty}^*(x)$  for all  $x \in \mathbf{X}$ . Denote

$$u(x) := \sup_{N > 0} V_{0,N}^*(x). \quad (2.31)$$

Then  $u(x)$  being the supremum of l.s.c. functions is l.s.c. as well. Letting  $N \rightarrow \infty$ , we have  $u(x) \leq V_{0,\infty}^*(x)$ . Hence  $V_{0,\infty}^*(x)$  is l.s.c. as well. We state that the optimal policies are stationary via an induction argument as in Theorem 2.3 and by Theorem 4.2.3 in [13], and hence conclude the proof.  $\square$

We recall that our optimization problem is

$$\inf_{\pi \in \Pi} \text{AVaR}_\alpha^\pi \left( \sum_{n=0}^{\infty} c(x_n, a_n) \right), \quad (2.32)$$

which is equivalent to

$$\inf_{\pi \in \Pi} \text{AVaR}_\alpha^\pi \left( \sum_{n=0}^{\infty} c(x_n, a_n) \right) = \inf_{s \in \mathbb{R}} \left\{ s + \frac{1}{1 - \alpha} \inf_{\pi \in \Pi} \mathbb{E}_x^\pi [(C^\infty - s)^+] \right\} \quad (2.33)$$

Hence, we fix the global variable a priori  $s$  as

$$s = \text{VaR}_\alpha^{\pi_0}(C^\infty), \quad (2.34)$$

where  $\text{VaR}_\alpha^{\pi_0}(C^\infty)$  is decided using the reference probability measure  $\mathbb{P}_0$ . It is claimed in [4] that by fixing global variable  $s$ , the resulting optimization problem would turn out to be over  $\text{AVaR}_\beta(C^\infty)$ , where possibly  $\alpha \neq \beta$ , under some assumptions. But, it is not clear to us, what these conditions would be for that to hold and why it should be necessarily case. Since for each fixed  $s$ , the inner optimization problem in Equation 2.31 has an optimal policy  $\pi(s)$  depending on  $s$ . Hence, as in [4], we focus on the inner optimization problem but by fixing the global variable  $s$  heuristically a priori  $\text{VaR}_\alpha^{\pi_0}(C^N)$  with respect to reference probability measure  $P$  and then solve the optimization problem *for each path*  $\omega$  conditionally with respect to filtration  $\mathcal{F}_n$  at each time  $n \in \mathbb{N}_0$  namely by taking into account whether for that path  $s_n \leq 0$  or  $s_n > 0$ . Hence, by denoting  $s_n = C^n - s$ , the optimization problem reduces to classical risk neutral optimization problem for that path  $\omega$  whenever  $s_n \leq 0$ .

### 3 $s_n(\omega) \leq 0$ case for that particular realization $\omega$

In this section, we are going to solve the case, after time  $n$ , when the risk averse problem reduces to risk neutral problem in that particular realization path  $w$ . Recall that the inner optimization problem is

$$\begin{aligned} V_0^*(x) &= \frac{1}{1-\alpha} \inf_{\pi \in \Pi} \mathbb{E}_x^\pi [(C^\infty - s)^+]. \\ &= \frac{1}{1-\alpha} \inf_{\pi \in \Pi} \mathbb{E}_x^\pi \left[ \left( \sum_{n=N+1}^{\infty} c(x_n, a_n) - (s - C^N) \right)^+ \right] \\ &= \frac{1}{1-\alpha} \inf_{\pi \in \Pi} \mathbb{E}_x^\pi \left[ \left( \sum_{n=N+1}^{\infty} c(x_n, a_n) - s_n \right)^+ \right] \end{aligned} \quad (3.35)$$

$$= \frac{1}{1-\alpha} \inf_{\pi \in \Pi} \mathbb{E}_x^\pi \left[ \mathbb{E}_x^\pi \left[ \left( \sum_{n=N+1}^{\infty} c(x_n, a_n) - s_n \right)^+ \middle| \mathcal{F}_n \right] \right] \quad (3.36)$$

$$= \frac{1}{1-\alpha} \inf_{\pi \in \Pi} \mathbb{E}_x^\pi \left[ \mathbb{E}_x^\pi \left[ \left( \sum_{n=N+1}^{\infty} c(x_n, a_n) - s_n \right)^+ \middle| \{x_n, s_n\} \right] \right] \quad (3.37)$$

Hence, whenever  $s_n(\omega) \leq 0$ , we have obviously a risk neutral optimization problem in that realization path  $\omega$ . Namely,

$$\begin{aligned} &\left( \sum_{i=n+1}^{\infty} \frac{1}{1-\alpha} c_i(x_i, \pi_i)(\omega) - \frac{1}{1-\alpha} s_n(\omega) \right)^+ \\ &= \sum_{i=n+1}^{\infty} \frac{1}{1-\alpha} c_i(x_i, \pi_i)(\omega) - \frac{1}{1-\alpha} s_n(\omega) \end{aligned}$$

where  $n = \min\{m \in \mathbb{N}_0 : s_m(\omega) \leq 0\}$  in that realization path  $\omega$ . Our iterative scheme follows closely [9]. To further proceed, we need the following two technical lemmas.

**Lemma 3.1.** *Fix an arbitrary  $n \in \mathbb{N}_0$ . Let  $\mathbb{K}_n$  be as in assumptions, and let  $u : \mathbb{K}_n \rightarrow \mathbb{R}$  be a given measurable function. Define*

$$u^*(x) := \inf_{a \in A_n(x)} u(x, a), \text{ for all } x \in X_n. \quad (3.38)$$

- *If  $u$  is nonnegative, l.s.c. and inf-compact on  $\mathbb{K}_n$ , then there exists  $\pi_n \in \mathbb{F}_n$  such that*

$$u^*(x) = u(x, \pi_n), \text{ for all } x \in X \quad (3.39)$$

*and  $u^*$  is measurable.*

- If in addition the multifunction  $x \rightarrow A_n(x)$  satisfies the Assumption 2.1, then  $u^*$  is l.s.c.

*Proof.* See [25]. □

**Lemma 3.2.** For every  $N > n \geq 0$ , let  $w_n$  and  $w_{n,N}$  be functions on  $\mathbb{K}_n$ , which are nonnegative, l.s.c. and inf-compact on  $\mathbb{K}_n$ . If  $w_{n,N} \uparrow w_n$  as  $N \rightarrow \infty$ , then

$$\lim_{N \rightarrow \infty} \min_{a \in A_n(x)} w_{n,N}(x, a) = \min_{a \in A_n(x)} w_n(x, a) \quad (3.40)$$

for all  $x \in X$ .

*Proof.* See [13] page 47. □

For  $n = \min\{m \in \mathbb{N}_0 : s_m(\omega) \leq 0\}$ , taking the beginning state as  $x_n(\omega)$  and calculating the minimal cost from that state  $x_n(\omega)$  onwards, by nonnegativity of cost functions  $c(x_i, a_i)$  for all  $i \in \mathbb{N}_0$ , we have obviously

$$V_{n,N}^*(x_n(\omega)) := \inf_{\pi \in \Pi} \int \left( \sum_{i=n}^N c(x_i, a_i) - s_n(\omega) \right)^+ \mathbb{Q}(dx' | x, f_0(x, s))$$

$$V_{n,N}^*(x_n(\omega)) := \inf_{\pi \in \Pi} \int \left( \sum_{i=n}^N c(x_i, a_i) - s_n(\omega) \right) \mathbb{Q}(dx' | x, f_0(x, s))$$

and similarly for the infinite horizon problem, we have

$$V_n^*(x_n(\omega)) := \inf_{\pi \in \Pi} \int \left( \sum_{i=n}^{\infty} c(x_i, a_i) - s_n(\omega) \right)^+ \mathbb{Q}(dx' | x, f_0(x, s))$$

$$V_n^*(x_n(\omega)) := \inf_{\pi \in \Pi} \int \left( \sum_{i=n}^{\infty} c(x_i, a_i) - s_n(\omega) \right) \mathbb{Q}(dx' | x, f_0(x, s))$$

**Definition 3.1.** A sequence of functions  $u_n : X_n \rightarrow \mathbb{R}$  on a realization path  $\omega$  at time  $n$  is called a solution to the optimality equations if

$$u_n(x)(\omega) = \inf_{a \in A_n(x)} \{c_n(x, a)(\omega) + \mathbb{E}[u_{n+1}[F_n(x, a, \xi_n)]]\}, \quad (3.41)$$

where

$$\mathbb{E}[u_{n+1}[F_n(x, a, \xi_n)]] = \int_{S_n} u_{n+1}[F_n(x, a, s)] \mu_n(ds). \quad (3.42)$$

First, we introduce the following notations for simplicity. Let  $L_n(X)$  be the family of l.s.c. non-negative functions on  $X$ . Moreover, denote

$$P_n u(x)(\omega) := \min_{a \in A_n(x)} \{c_n(x, a)(\omega) + \mathbb{E}[u_{n+1}[F_n(x, a, \xi_n)]]\}, \quad (3.43)$$

for all  $x \in X$ .

**Lemma 3.3.** *Using the Assumption 2.1, then*

- $P_n$  maps  $L_{n+1}(X)$  into  $L_n(X)$ .
- For every  $u \in L_{n+1}(X)$ , there exists  $a_n^* \in \mathbb{F}_n$  such that  $a_n \in A_n(x)$  attains the minimum in 3.41, i.e.

$$P_n u(x)(\omega) := \{c_n(x, a_n)(\omega) + \mathbb{E}[u_{n+1}[F_n(x, a_n, \xi_n)]]\}, \quad (3.44)$$

*Proof.* Let  $u \in L_{n+1}(X)$ . Then by assumptions we have that the function

$$(x, a, \omega) \rightarrow c_n(x, a, \omega) + \mathbb{E}[u_{n+1}[F_n(x, a, \xi_n)]] \quad (3.45)$$

is non-negative and l.s.c. and by Lemma 3.1, there exists  $\pi_n \in \mathbb{F}_n$  that satisfies Equation 3.52 and  $P_n u$  is l.s.c. So we conclude the proof.  $\square$

By dynamic programming principle, we express the optimality equations in 3.41 as

$$V_m^* = P_m V_{m+1}^*, \quad (3.46)$$

for all  $m \geq n$ . We continue with the following lemma.

**Lemma 3.4.** *Using the Assumption 2.1, consider a sequence  $\{u_m\}$  of functions  $u_m \in L_m(X)$  for  $m \in \mathbb{N}_0$ , then the following is true. If  $u_n \geq P_n u_{n+1}$  for all  $m \geq n$ , then  $u_m \geq V_m^*$  for all  $m \geq n$ .*

*Proof.* By previous lemma, there exists a policy  $\pi = \{\pi_m\}_{m \geq n}$  such that for all  $m \geq n$

$$u_m(x) \geq c_m(x, \pi_m) + u_{m+1}(x_{m+1}^\pi). \quad (3.47)$$

By iterating, we have

$$u_m(x) \geq \sum_{i=m}^{N-1} c_i(x_i^\pi, \pi_i) + u_{m+N}(x_{m+N}^\pi), \quad (3.48)$$

Hence we have

$$u_m(x) \geq V_{m,N}(x, \pi), \quad (3.49)$$

for all  $N > 0$ . By letting  $N \rightarrow \infty$ , we have  $u_m(x) \geq V_m(x, \pi)$  and so  $u_m \geq V_m^*$ . Hence, we conclude the proof.  $\square$

**Theorem 3.5.** *Suppose that assumptions hold, then for every  $m \geq n$  and  $x \in X$ ,*

$$V_{n,N}^*(x) \uparrow V_n^*(x), \quad (3.50)$$

as  $N \rightarrow \infty$  and  $V_n^*$  is l.s.c.

*Proof.* We justify the statement by appealing to dynamic programming algorithm, we have  $J_N(x) := 0$  for all  $x \in X_N$ , and by going backwards for  $t = N - 1, N - 2, \dots, n$ , and let

$$J_t(x) := \inf_{a \in A_t(x)} \{c_t(x, a) + J_{t+1}[F_t(x, a, \xi)]\}. \quad (3.51)$$

By backward iteration, for  $t = N - 1, \dots, n$ , there exists  $\pi_t \in \mathbb{F}_m$  such that  $\pi_m(x) \in A_m(x)$  attains the minimum in the Equation 3.59, and  $\{\pi_{N-1}, \pi_{N-2}, \dots, \pi_n\}$  is an optimal policy. Moreover,  $J_n$  is the optimal cost for

$$J_n(x) := V_{n,N}^*(x_n), \quad (3.52)$$

Hence, we have

$$V_{n,N}^*(x) = \min_{a \in A_n(x)} \{c_n(x, a) + V_{n+1,N}^*[F_n(x, a, \xi)]\}. \quad (3.53)$$

Denoting  $u(x) = \sup_{N > n} V_{n,N}^*(x)$ , we have  $u(x)$  is l.s.c. By Lemma 3.2, we have

$$V_n^*(x) = \min_{a \in A_n(x)} \{c_n(x, a) + V_{n+1}^*[F_n(x, a, \xi)]\}. \quad (3.54)$$

Moreover, cost functions  $c_n(x, a)$  being nonnegative, we have  $u(x) \leq V_n^*(x)$ . But by definition, we have  $V_n^*(x) \leq u(x)$ . Hence, we conclude the proof.  $\square$

## 4 Numerical Example

We are going to illustrate our theoretical framework through a numerical example. We treat the classical LQ-problem using risk sensitive AVaR operator to illustrate our results below and give a heuristic algorithm that specifies the decision rule at each time episode  $n$  based on our results above. We solve the classical linear system with a quadratic one-stage cost problem with AVaR Criteria. Suppose we take  $X = \mathbb{R}$  with a linear system

$$x_{n+1} = x_n + a_n + Z_n, \quad (4.55)$$

with  $x_0 = 0$ ,  $Z_n$  is i.i.d. standard normal i.e.  $Z_n \sim \mathcal{N}(0, 1)$ . We take one stage cost functions as  $c(x_n, a_n) = x_n^2 + a_n^2$  for  $n = 0, 1, \dots, N - 1$ . We also assume that the control



constraint sets  $A_n(x)$  with  $x \in X$  are all equal to  $A_n = \mathbb{R}$ . Thus, under the above assumptions, we wish to find a policy that minimizes the performance criterion

$$J(\pi, x) := \text{AVaR}_\alpha^\pi \left( \sum_{n=0}^{N-1} (x_n^2 + a_n^2) \right), \quad (4.56)$$

It is well known that in risk neutral case using dynamic programming, the optimal policy  $\pi^* = \{f_0, \dots, f_{n-1}\}$  and the value function  $J_n$  satisfy the following dynamics.

$$\begin{aligned} K_N &= 0 \\ K_n &= \left[ 1 - (1 + K_{n+1})^{-1} K_{n+1} \right] K_{n+1} + 1, \text{ for } n = 0, \dots, N-1 \\ J_n(x) &= K_n x^2 + \sum_{i=n+1}^{N-1} K_i, \text{ for } n = 0, \dots, N-1 \end{aligned} \quad (4.57)$$

(see e.g. [13]). When we use, AVaR operator, we proceed as follows. First, we choose the global variable  $s$  a-priori and fix it. We take  $a_n = 0$  for  $n = 0, \dots, N-1$ , i.e.  $\pi_0 = \{0, 0, \dots, 0\}$  and let

$$\begin{aligned} s &:= \text{VaR}_\alpha \left( \sum_{n=0}^{N-1} c(x_n, a_n) \right) \\ &:= \inf \left\{ x \in \mathbb{R} : \mathbb{P} \left( \sum_{n=0}^{N-1} X_n^2 \leq x \right) \geq \alpha \right\}. \end{aligned}$$

If  $s > 0$ , then we choose  $a_n = 0$  at time  $n$ . This means  $c(x_n, a_n) = x_n^2 + a_n^2$  is minimal for that time  $n$  in a *greedy* way. We remark here that by setting  $a_n = 0$ , we have  $\sum_{n=0}^{N-1} X_n^2$  has  $\chi^2$  distribution with  $n-1$  degrees of freedom. Then, we update global variable  $s$  with  $s - c_n(x_n, a_n) = s - x_n^2$ . Next, we simulate the random variable  $\xi_n(\omega)$  and get  $x_{n+1} = x_n + \xi_n(\omega)$ . If  $s \leq 0$ , then our problem reduces to risk neutral case. We repeat the procedure until end horizon  $N$ . We simulated our algorithm for  $N = 10000$  and find that our scheme preserves the monotonicity property of  $\text{AVaR}_\alpha(X)$  operator, namely  $\text{AVaR}_\alpha(X) \leq \text{AVaR}_\alpha(Y)$ . Moreover, we also see that with respect to risk aversion the corresponding value functions increase as well, namely  $\text{AVaR}_{\alpha_1}(X) \leq \text{AVaR}_{\alpha_2}(Y)$  whenever  $\alpha_1 \leq \alpha_2$ . We also have by our algorithm that for  $\alpha = 0$  we have the risk neutral value functions which is consistent with  $\text{AVaR}_\alpha(X)_{\lim \alpha \rightarrow 0} = \mathbb{E}[X]$ . We give the pseudocode of this algorithm below and present our numerical results afterwards.

```

1: procedure LQ-AVAR ALGORITHM
2:    $s = \text{VaR}_\alpha^{\pi_0}(\sum_{n=0}^{N-1} X_n^2)$ 
3:    $x = 0$ 
4:    $V^{dyn} = 0$ 
5:    $V(x) = 0$ 
6:   for each  $n \in N - 1$  do
7:     if  $s \leq 0$  then
8:       apply Dynamic Programming from state  $x_n$  onwards as in Equation 4.57
9:       Update  $V^{dyn}$ 
10:    else
11:      Choose  $a_n = 0$ 
12:      Update  $s = s - x_n^2$ 
13:      Update  $c_n = x_n^2 + a_n^2$ 
14:      Update  $x_{n+1} = x_n + a_n + \xi_n(\omega)$ 
15:      Update  $V(x) = V(x) + c_n$ 
16:    end if
17:  end for
18: return  $V(x) + V^{dyn}$ 
19: end procedure

```

## 4.1 Simulation Results

$\alpha$	$N$	Value	$\alpha$	$N$	Value
0	5	7.33303167	0.1	5	14.0959365
0	10	15.4231355	0.1	5	30.1207678
0	15	23.5133055	0.1	5	44.2908071
0	20	31.6034754	0.1	5	61.0531863
0	25	39.6936453	0.1	5	72.8025974
0	30	47.7838153	0.1	5	87.8589529
0	35	55.8739852	0.1	5	104.57686
0	40	63.9641552	0.1	5	118.657453
0	45	72.0543251	0.1	5	131.609203
0	50	80.1444951	0.1	5	147.581156

$\alpha$	$N$	Value	$\alpha$	$N$	Value
0.2	5	12.7832559	0.3	5	15.0884687
0.2	10	30.2933167	0.3	10	34.418539
0.2	15	45.8463747	0.3	15	52.9767429
0.2	20	63.2351183	0.3	20	63.435157
0.2	25	77.005527	0.3	25	81.0304336
0.2	30	95.862442	0.3	30	100.093363
0.2	35	105.469191	0.3	35	110.798357
0.2	40	124.158071	0.3	40	128.664487
0.2	45	137.95591	0.3	45	142.315159
0.2	50	148.692589	0.3	50	154.596272

$\alpha$	$N$	Value	$\alpha$	$N$	Value
0.4	5	15.7236254	0.5	5	13.4512086
0.4	10	32.7566005	0.5	10	35.954209
0.4	15	47.4891141	0.5	15	49.7899661
0.4	20	68.6376918	0.5	20	67.0747353
0.4	25	83.0223834	0.5	25	83.582299
0.4	30	97.9839512	0.5	30	101.501654
0.4	35	116.642985	0.5	35	110.798357
0.4	40	129.137922	0.5	40	129.393742
0.4	45	142.574767	0.5	45	146.171546
0.4	50	157.641565	0.5	50	162.326472

$\alpha$	$N$	Value	$\alpha$	$N$	Value
0.6	5	16.1387319	0.7	5	16.1602546
0.6	10	34.3455866	0.7	10	35.5536365
0.6	15	52.6591864	0.7	15	54.9161278
0.6	20	71.5636394	0.7	20	76.0232881
0.6	25	85.7706824	0.7	25	93.5905457
0.6	30	105.396058	0.7	30	106.406181
0.6	35	123.231726	0.7	35	121.878061
0.6	40	134.8542	0.7	40	139.969018
0.6	45	149.150083	0.7	45	150.071055
0.6	50	160.836396	0.7	50	165.494609

$\alpha$	$N$	Value	$\alpha$	$N$	Value
0.8	5	14.4071662	0.9	5	14.5457655
0.8	10	37.4974225	0.9	10	40.7872526
0.8	15	57.3475455	0.9	15	61.4048039
0.8	20	75.7915348	0.9	20	82.3070707
0.8	25	92.5621339	0.9	25	97.6919741
0.8	30	109.284529	0.9	30	115.164378
0.8	35	123.941457	0.9	35	128.772272
0.8	40	147.458903	0.9	40	146.842669
0.8	45	160.066665	0.9	45	163.259278
0.8	50	174.394	0.9	50	178.658011

$\alpha$	$N$	Value
1	5	19.5096313
1	10	54.9595218
1	15	119.689535
1	20	210.271252
1	25	287.815856
1	30	417.29921
1	35	603.178852
1	40	912.998288
1	45	839.972665
1	50	1108.49149

## References

- [1] ACCIAIO, B., PENNER, I. (2011). *Dynamic convex risk measures.*, In G. Di Nunno and B. ksendal (Eds.), *Advanced Mathematical Methods for Finance*, Springer, 1-34.
- [2] ARTZNER, P., DELBAEN, F., EBER, J.M., HEATH, D. (1999). *Coherent measures of risk*, *Math. Finance* 9, 203-228.
- [3] AUBIN, J.-P., FRANKOWSKA, H. (1978). *Set-Valued Analysis* Birkhauser, Boston, 1990.
- [4] BAUERLE, N., OTT J. (2011). *Markov Decision Processes with Average-Value-at-Risk Criteria*, *Mathematical Methods of Operations Research*, 74, 361-379.

- [5] BAUERLE, N. , RIEDER, U. (2011). *Markov Decision Processes with applications to finance*, Springer.
- [6] BELLMAN, R. (1952). *On the theory of dynamic programming* Proc. Natl. Acad. Sci 38, 716.
- [7] BERTSEKAS, D., SHREVE, S.E. (1978). *Stochastic Optimal Control. The Discrete Time Case*, Math. Program. Ser. B 125:235-261.
- [8] CHUNG,K.J.,SOBEL,M.J. (1987). *Discounted MDPs: distribution functions and exponential utility maximization* SIAM J. Control Optimization., 25, 49-62.
- [9] EKELAND, I., TEMAM, R. (1974). *R. Convex Analysis and Variational Problems*, Dunnod.
- [10] FLEMING,W., SHEU,S. (1999). *Optimal long term growth rate of expected utility of wealth* Ann. Appl. Prob.,9. 871-903.
- [11] FILIPOVIC, D. AND SVINDLAND, G. (2012). *The canonical model space for law-invariant convex risk measures is  $L^1$*  , Mathematical Finance 22(3), 585-589.
- [12] GUO, X., HERNANDEZ-LERMA, O. (2012). *Nonstationary discrete-time deterministic and stochastic control systems with infinite horizon*, International Journal of Control, vol. 83, pp 1751-1757.
- [13] HERNANDEZ-LERMA,O., LASSERRE, J.B. (1996). *Discrete-time Markov Control Processes. Basic Optimality Criteria.*, Springer,New York.
- [14] KUPPER, M., SCHACHERMAYER, W. (2009). *Representation results for law invariant time consistent functions*,Mathematics and Financial Economics 189-210.
- [15] ROCKAFELLAR, R.T , URYASEV, S. (2002). *Conditional-Value-at-Risk for general loss distributions*, Journal of Banking and Finance 26, 1443-1471.
- [16] ROCKAFELLAR, R.T., WETS, R.J.-B. (1998). *Variational Analysis.*, Springer, Berlin.
- [17] RUSCHENDORF, L., KAINA, M. (2009). *On convex risk measures on  $L^p$ -spaces*, Mathematical Methods in Operations Research, 475-495.
- [18] RUSZCZYNSKI, A. (1999). *Risk-averse dynamic programming for Markov decision processes*, Math. Program. Ser. B 125:235-261.
- [19] RUSZCZYNSKI, A. AND SHAPIRO, A. (2006). *Optimization of convex risk functions*, Mathematics of Operations Research, vol. 31, pp. 433-452.
- [20] SHAPIRO, A. (2012). *Time consistency of dynamic risk measures*, Operations Research Letters, vol. 40, pp. 436-439.
- [21] XIN, L., SHAPIRO, A. (2009). *Bounds for nested law invariant coherent risk measures*, Operations Research Letters, vol. 40, pp. 431-435.
- [22] SHAPIRO, A. (2015). *Rectangular sets of probability measures*, preprint.
- [23] EPSTEIN, L. G. AND SCHNEIDER, M. (2003). *Recursive multiple-priors*, Journal of Economic Theory, 113, 1-31.

- [24] IYENGAR, G.N. (2005). *Robust Dynamic Programming*, Mathematics of Operations Research, 30, 257-280.
- [25] RIEDER, U. (1978). *Measurable Selection Theorems for Optimisation Problems*, Manuscripta Mathematica, 24, 115-131.
- [26] G. C. PFLUG AND A. PICHLER, *Time-inconsistent multistage stochastic programs: Martingale bounds*, European J. Oper. Res., 249 (2016), pp. 155-163.
- [27] M.STADJE AND P. CHERIDITO, *Time-inconsistencies of Value at Risk and Time-Consistent Alternatives*, Finance Research Letters. (2009) 6, 1, 40-46.
- [28] A. PICHLER, *The Natural Banach Space for Version Independent Risk Measures*, Insurance: Mathematics and Economics. (2013) 53, 405-415.
- [29] ENGWERDA, J.C., *Control Aspects of Linear Discrete Time-varying Systems*, International Journal of Control, (1988) 48, 1631-1658.
- [30] KEERTHI, S.S., AND GILBERT, E.G. *Optimal Infinite-horizon Feedback Laws for a General Class of Constrained Discrete-time Systems*. Journal of Optimization and Theory Applications, (1988), 57, 265-293.
- [31] GUO, X.P., LIU, J.Y., AND LIU, K. (2000), *The Average Model of Nonhomogeneous Markov Decision Processes with Non-uniformly Bounded Rewards*. Mathematics of Operation Research, (2000) 25, 667-678.
- [32] BERTSEKAS, D.P., AND SHREVE, S.E. *Stochastic Optimal Control: The Discrete Time Case*, (1978), New York: Academic Press.
- [33] KEERTHI, S.S., AND GILBERT, E.G. *An Existence Theorem for Discrete-time Infinite-horizon Optimal Control Problems*. IEEE Transactions on Automatic Control, (1985), 30, 907-909.