Controlled Markov Decision Processes with AVaR Criteria for Unbounded Costs

Kerem Uğurlu

Monday 28th November, 2016

Department of Applied Mathematics, University of Washington, Seattle, WA 98195
e-mails:keremu@uw.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 aversive 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 necessariliy 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 method, called state aggregation, 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 to state aggregation is applied and for each path ω, the information necessary from the previous time steps is included in the current decision. All these works are studying *bounded* costs in L^∞, hence whenever they study infinite time horizon, they verify the existence of optimal policy via a contraction mapping and fixed point argument. In [34], several weaker conditions and the notion of weak time consistency are introduced. These are characterized by the existence of dual representations making it easier to solve dynamic programming equations, but these approaches only hold in L^∞. To the best of our knowledge, there are few papers related to minimizing AVaR or other risk measures in L^p spaces with $1 \leq p < \infty$ ([35, 36, 37]). This paper is in that direction. We study the optimal control on MDPs with possibly unbounded costs on L^1 using coherent risk measures.

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. We also present examples in real life scenarios related to insurance and finance and give the complete recipe to apply our scheme.

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.

Notation. Given a *Borel space*, namely a Borel subset of a complete separable metric space Y, its Borel sigma-algebra is denoted by $\mathcal{B}(Y)$ and "measurable" means "Borel-measurable". Moreover, $L(Y)$ stands for the family of lower semicontinuous (l.s.c.) functions on Y, bounded from below, and $L(Y)_+$ denotes the subclass of nonnegative functions in $L(Y)$.

2 The Control Model

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

$$\mathcal{M}_n := (X, A, \mathbb{K}_n, Q, F_n, c_n) \tag{2.1}$$

with the following components:

- X and A denote the state and action (or control) spaces. X and A are assumed to be Borel spaces.

- For each $x_n \in X$, let $A(x_n) \subset A$ be the set of all admissible controls in the state x_n. Then

  $$\mathbb{K}_n := \{(x_n, a_n) : x_n \in X, a_n \in A(x)\}, \tag{2.2}$$

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

- We assume that $\mathbb{K}_n$ is a Borel subset of $X \times A$, and that it contains the graph of a measurable function $\pi : X \to A$ (the latter condition ensures that the set $\mathbb{F}_n$ defined below is nonempty)

- We let

  $$x_{n+1} = F_n(x_n, a_n, \xi_n), \tag{2.3}$$

  for all $n = 0, 1, ...$ with $x_n \in X$ and $a_n \in A$ as described above, with independent random disturbances $\xi_n \in S_n$ having probability distributions μ_n, where the S_n are Borel spaces and F_n is a given measurable function, system equation, from $\mathbb{K}_n \times S_n$ to X.

- $c_n(x_n, a_n, \xi_n) : \mathbb{K}_n \times S_n \to \mathbb{R}$ stands for the deterministic cost-per-stage function at stage $n \in \mathbb{N}_0$ with $(x_n, a_n) \in \mathbb{K}_n$ and for fixed ξ_n, $c_n(\cdot, \cdot, \xi_n)$ is assumed to be l.s.c. and nonnegative.

- The transition law $Q(B|x,a)$, where $B \in \mathcal{B}(X)$ and $(x,a) \in \mathbb{K}_n$ is a stochastic kernel on X given $\mathbb{K}_n$ (see [38, 39] for further details). That is, for each pair $(x,a) \in \mathbb{K}_n$, $Q(\cdot|x,a)$ is a probability measure on X, and for each $B \in \mathcal{B}(X)$, $Q(B|\cdot)$ is a measurable function on $\mathbb{K}_n$.

To state one of our main assumptions, we give first the following definition.

Definition 2.1. *A real valued function v on $\mathbb{K}_n$ is said to be inf-compact on $\mathbb{K}_n$, if the set*

$$\{a \in A_n(x) | v(x,a) \leq c\} \tag{2.4}$$

is compact for every $x \in X$ and $c \in \mathbb{R}$. As an example, if the sets $A(x)$ are compact and $v(x,a)$ is l.s.c. in $a \in A(x)$ for every $x \in X$, then v is inf-compact on $\mathbb{K}_n$. Conversely, if v is inf-compact on $\mathbb{K}_n$, then v is l.s.c. in $a \in A(x)$ for every $x \in X$.

Assumption 2.2. *(a) $c_n(x,a)$ is non-negative, l.s.c. and inf-compact on $\mathbb{K}_n$ for fixed ξ_n.*

(b) The transition law Q is weakly continuous; i.e. for any continuous and bounded function u on X, the map

$$(x,a) \to \int_X u(y) Q(dy|x,a) \tag{2.5}$$

is continuous on $\mathbb{K}_n$.

(c) The multifunction (or set-valued map) $x \to A(x)$ is l.s.c.; i.e. if $x_m \to x$ in X as $m \to \infty$ and $a \in A(x)$, then there are $a_m \in A(x_m)$ such that $a_m \to a$ as $m \to \infty$.

(d) The system function $x_{n+1} = F_n(x_n, a_n, \xi_n)$ is continuous on $\mathbb{K}_n$ for every $\xi_n \in S_n$.

Remark 2.3. *A function v belongs to $L(X)$ if and only if there is a sequence of continuous and bounded functions u_m on X such that $u_m \uparrow v$. By using this fact, we can restate Assumption 2.1 (b) as: For any $v \in L(X)$, the map $(x,a) \to \int v(y) Q(dy|x,a)$ is l.s.c. and bounded from below on $\mathbb{K}_n$. We note also that if $(x,a) \to F(x,a,s)$ in Equation 2.3 is continuous on $\mathbb{K}_n$ fir every $s \in S_n$, then Assumption 2.2 (b) holds. It is also known that Assumption 2.2. (c) holds, if $\mathbb{K}_n$ is convex. ([40], cf. Lemma 3.2). Moreover, the latter convexity condition holds in many real life scenarios related to control problems like in inventory/production systems, water resources management, etc. ([41, 42, 43, 39]).*

Definition 2.4. *We let $\mathbb{F}$ denote the family of measurable functions f from X to A such that $f(x) \in A(x)$ for all $x \in X$. We let x_n and a_n denote, respectively, the state of the system and the control action applied at time $n = 0, 1, ...$ A rule to choose the control action a_n at time n is called a control policy. More formally, a control policy π is a sequence $\{f_n\}$ such that for each $n = 0, 1, ..., \pi_n(\cdot|h_n)$ is a conditional probability on $\mathcal{B}(A)$, given the history $h_n := (x_0, a_0, ..., x_{n-1}, a_{n-1}, x_n)$, that satisfies the constraint $f_n(A(x_n)|h_n) = 1$. The class of all policies is denoted by Π. A sequence $\{f_n\}$ of functions $f_n \in \mathbb{F}$ is called a Markov policy if*

$$f_n : X \to A. \tag{2.6}$$

A Markov policy $\{f_n\}$ is said to be a stationary policy, if it is of the form $f_n \equiv f$ for all $n = 0, 1, ...$ for some $f \in \mathbb{F}$. Furthermore, $\pi = \{\pi_n\}$ is said to be

- *a deterministic policy, if there is a sequence $\{f_n\}$ of measurable functions $f_n : H_n \to A$ such that for all $h_n \in H_n$ and $n = 0, 1, 2, ...,$ we have $f_n \in A(x_n)$ and $\pi_n(\cdot|h_n)$ is concentrated at $f_n(h_n)$, i.e.*

  $$\pi_n(C|h_n) = I_C(f_n(h_n)), \tag{2.7}$$

  for all $C \in \mathcal{B}(A)$.

- *a deterministic Markov policy, if there is a sequence $\{f_n\}$ of functions $f_n \in \mathbb{F}$ such that $\pi_n(\cdot|h_n)$ is concentrated at $f_n(x_n) \in A(x_n)$ for all $h_n \in H_n$ and $n = 0, 1, 2, ..,$.*

- *a deterministic stationary policy, if there is a function $f \in \mathbb{F}$ such that $\pi_n(\cdot|h_n)$ is concentrated at $f(x_n) \in A(x_n)$ for all $n \in \mathbb{N}_0$.*

Remark 2.5. *In this paper, our admissible policies $\pi = \{\pi_n\}$ are restricted to deterministic policies.*

Let $(\Omega, \mathcal{F})$ be the measurable space consisting of the sample space $\Omega := \prod_{n=1}^{\infty}(X \times A)$ and the corresponding Borel σ-algebra on Ω is denoted by $\mathcal{F}$. Then, for an arbitrary policy $\pi \in \Pi$ and initial state $x \in X$, by Ionescu-Tulcea Theorem [7], there exists a unique probability measure P_x^π on $(\Omega, \mathcal{F})$, which is concentrated on the set of all sequences $(x_0, a_0, x_1, a_1, ...)$ with $(x_n, a_n) \in \mathbb{K}_n$ for all $n = 0, 1, ...$ Moreover, P_x^π satisfies that $P_x^\pi(x_0 = x) = 1$, and for every $n = 0, 1, ...$

$$P_x^\pi(a_n \in C|h_n) = \pi_n(C|h_n) \tag{2.8}$$

$$P_x^\pi(x_{n+1} \in B|h_n, a_n) = Q(B|x_n, a_n), \tag{2.9}$$

for every $C \in \mathcal{B}(A)$ and $B \in \mathcal{B}(X)$. $(\Omega, \mathcal{F}, P_x^\pi, \{x_n\})$ is called a discrete time Markov control process. The expectation operator with respect to P_x^π is denoted by $\mathbb{E}_x^\pi$.

Remark 2.6. *If $\pi = \{f_n\}$ is a Markov policy, then the state process $\{x_n\}$ is a Markov process with transition kernel $Q(\cdot|x, f_n(x))$; that is*

$$P_x^\pi(x_{n+1} \in B|x_0, x_1, ..., x_n) = P_x^\pi(x_{n+1} \in B|x_n) = Q(B|x_n, f_n(x_n)), \tag{2.10}$$

for all $B \in \mathcal{B}(X_n)$ and $n = 0, 1, ...$ In particular if $f \in \mathbb{F}$ is a stationary policy, then $\{x_n\}$ has a time-homogeneous transition kernel $Q(B|x_n, f_n(x_n))$.

3 Coherent Risk Measures

Evaluation Criteria. We consider the cost functions denoted by

$$C^\infty := \sum_{n=0}^{\infty} c_n(x_n, a_n, \xi_n), \tag{3.11}$$

for the infinite planning horizon and

$$C^N := \sum_{n=0}^{N} c_n(x_n, a_n, \xi_n) \tag{3.12}$$

for the finite planning horizon for some terminal time $N \in \mathbb{N}_0$. 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:

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

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:

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

Under some assumptions, the first optimization problem has solution in form of Markov policies, whereas in infinite case the optimal 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* ρ instead of the expectation operator and look for an 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^∞ and later extended to random variables on L^1 in [17, 19] with a norm on L^1 introduced in [28]. Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a measurable space and let $X \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ be a real-valued random variable. A function $\rho : L^1 \to \mathbb{R}$ is said to be a *coherent risk measure* if it satisfies the following axioms:

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

Remark 3.1. *We note that under the fourth property (homogeneity), the first property (convexity) is equivalent to sub-additivity.*

The particular risk averse operator that we will be working with is the $\text{AVaR}_\alpha(X)$. Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a measurable space and let $X \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ be a real-valued random variable and $\alpha \in (0,1)$.

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

$$\text{VaR}_\alpha(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X \leq x) \geq \alpha\} \tag{3.13}$$

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

$$\text{AVaR}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \text{VaR}_t(X)dt \tag{3.14}$$

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

Lemma 3.2. *Let $X \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ be a real-valued random variable and let $\alpha \in (0,1)$. Then it holds that*

- $$\mathrm{AVaR}_\alpha(X) = \min_{s\in\mathbb{R}} \left\{ s + \frac{1}{1-\alpha}\mathbb{E}[(X-s)^+] \right\}, \tag{3.15}$$
  where the minimum is attained at $s = \mathrm{VaR}_\alpha(X)$.

- $\mathrm{AVaR}_\alpha(X) = \sup_{\mu\in\mathcal{M}} E^\mu[X]$, *where* $\mathcal{M}$ *is the set of absolutely continuous probability measures with densities satisfying* $0 \le \frac{d\mu}{d\mathbb{P}} \le 1/\alpha$.

Remark 3.3. *We note from the representations above that the* $\mathrm{AVaR}_\alpha(X)$ *is real-valued for any* $X \in L^1(\Omega, \mathcal{G}, \mathbb{P})$. *We further note that*

$$\lim_{\alpha\to 0} \mathrm{AVaR}_\alpha(X) = \mathbb{E}[X] \tag{3.16}$$

$$\lim_{\alpha\to 1} \mathrm{AVaR}_\alpha(X) = \operatorname{ess\,sup} X \le \infty. \tag{3.17}$$

Since we are dealing with dynamic decision process, we should introduce a concept of so called *time consistency*. One approach is to define time consistency from the point of view of optimal policies strategies. In that regard, we cite [44]: "The sequence of optimization problems is said to be dynamically consistent if the optimal strategies obtained when solving the original problem at time t_1 remain optimal for all subsequent problems". A similar definition is given in [45]: "Optimality of the decision at a state of the process at time $t \in 1, ..., T1$ should not involve states which do not follow that state, i.e., cannot happen in the future." [20] describes the concept of *time consistency* as "if the decision process is represented by the corresponding scenario tree, this means that if at a time t, we are at a certain node of the tree, then optimality of our future decisions should not depend on scenarios which do not pass through this node."

Remark 3.4. *Given* $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n=0}^N, \mathbb{P})$ *be the measurable space with* $\{\mathcal{F}_n\}_{n=0}^N$ *being the filtration,* $\mathcal{F} = \sigma(\cup_{n=0}^N \mathcal{F}_n)$ *being the* σ*-algebra, and* Ω *and* $\mathbb{P}$ *being the probability space and the probability measure respectively, if the probability space is atomless, it is shown in [20] and [14] that the only law invariant coherent risk measure operators* ρ *on i.e.*

$$X \overset{d}{=} Y \Rightarrow \rho(X) = \rho(Y) \tag{3.18}$$

satisfying the "telescoping property"

$$\rho(Z) = \rho(\rho_{|\mathcal{F}_1}(...\rho_{|\mathcal{F}_{N-1}})(Z)), \tag{3.19}$$

for all random variables Z *measurable on* $((\Omega, \mathcal{F}, \mathbb{P}))$ *are* $\operatorname{esssup}(Z)$ *and expectation* $\mathbb{E}(Z)$ *operators. We refer the reader to [20] to further investigate the expression in Equation 3.19. This suggests that optimization problems with most of the coherent risk measures are not time consistent.*

4 Main Result

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

$$\min_{\pi\in\Pi} \mathrm{AVaR}_\alpha^\pi(\sum_{n=0}^{\infty} c_n(x_n, a_n, \xi_n)), \tag{4.20}$$

Remark 4.1. *In [4], the infinite horizon with bounded costs with a discount factor of $0 < r < 1$ 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.*

Assumption 4.2. *There exists a policy $\pi_0 \in \Pi$ such that for the risk neutral case the optimization problem is finite for any $x \in X$. Namely,*

$$\mathbb{E}_x^{\pi_0}\big(\sum_{n=0}^{\infty} c_n(x_n, a_n, \xi_n)\big) < \infty. \tag{4.21}$$

Remark 4.3. *By Lemma 3.2 above, this immediately necessitates that for that policy π_0*

$$\mathrm{AVaR}_\alpha^{\pi_0}\bigg(\sum_{n=0}^{\infty} c_n(x_n, a_n)\bigg) < \infty, \tag{4.22}$$

since $\mathrm{AVaR}_\alpha(X) \le \frac{1}{\alpha}\mathbb{E}(X)$ for any random variable $X \in L^1(\Omega, \mathcal{G}, \mathbb{P})$.

To solve 4.20, we first rewrite the infinite horizon problem as follows:

$$\inf_\pi \mathrm{AVaR}_\alpha^\pi(C^\infty | X_0 = x) = \inf_{\pi\in\Pi}\inf_{s\in\mathbb{R}}\left\{s + \frac{1}{1-\alpha}\mathbb{E}_x^\pi[(C^\infty - s)^+]\right\} \tag{4.23}$$

$$= \inf_{s\in\mathbb{R}}\inf_{\pi\in\Pi}\left\{s + \frac{1}{1-\alpha}\mathbb{E}_x^\pi[(C^\infty - s)^+]\right\} \tag{4.24}$$

$$= \inf_{s\in\mathbb{R}}\left\{s + \frac{1}{1-\alpha}\inf_{\pi\in\Pi}\mathbb{E}_x^\pi[(C^\infty - s)^+]\right\} \tag{4.25}$$

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

$$w_{N\pi}(x, s) := \mathbb{E}_x^\pi[(C^N - s)^+], \quad x \in X, \quad s \in \mathbb{R}, \pi \in \Pi, \tag{4.26}$$

$$w_N(x, s) := \inf_{\pi\in\Pi} w_{N\pi}(x, s), \quad x \in X, \quad s \in \mathbb{R}, \tag{4.27}$$

We work with the Markov Decision Model with a 2-dimensional state space $\widetilde{X} \triangleq X \times \mathbb{R}$. The second component of the state $(x_n, s_n) \in \widetilde{X}$, s_n gives the relevant information of the

history of the process, hence we *aggregate the state*. We take that there is no running cost and we assume that the terminal cost function is given by $V_{-1\pi}(x,s) := V_{-1}(x,s) := s^-$. Further, we take the decision rules $f_n : \widetilde{X} \to A$ such that $f_n(x,s) \in A(x)$ and denote by Π^{pM} the set of *pseudo-Markovian* policies $\pi = (f_0, f_1, ...,)$, where f_n are decision rules. Here, by *pseudo-Markovian*, we mean that the decision at time n depends only on the current state x_n and as well as on the variable $s_n \in \mathbb{R}$, where s_n is also updated at each time episode n as to be seen in the proof of Theorem 4.4 below. We denote for

$$v \in \mathbb{M}(\widetilde{X}) := \{v : \widetilde{X} \to \mathbb{R}_+ : \text{measurable}\} \tag{4.28}$$

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

$$T_a(x_n, s, a) := \int v(x_{n+1}, s - c_n(x_{n+1}, a))\mathbb{Q}(dx_{n+1}|x_n, s, a), \;\; (x_n, s) \in \widetilde{X}, a \in A_n(x) \tag{4.29}$$

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

$$Tv(x) = \inf_{a \in A(x)} T_a(x, s, a). \tag{4.30}$$

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

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

A decision rule f_n^* with the property that $V_n = T_{f_n^*} V_{n-1}$ is called the minimizer of V_n. The necessary information at time n of the history $h_n = (x_0, a_0, x_1, ..., a_{n-1}, x_n)$ are the state x_n and the necessary information $s_n \triangleq s_0 - c_0 - c_1 - ... - c_{n-1}$. This dependence of the past and the optimality of the pseudo Markovian policy is shown in Theorem 4.4. 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, respectively.

Theorem 4.4. *[4] For a given policy π, the only necessary information at time n of the history $h_n = (x_0, a_0, x_1, ..., a_{n-1}, x_n)$ are the followings*

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

Moreover, it holds for $n = 0, 1, ..., N$ *that*

- $w_{n\pi} = V_{n\pi}$ *for* $\pi \in \Pi^{pM}$.
- $w_n = V_n$

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

$$\inf_{\pi \in \Pi} \mathbb{E}_x^{\pi}[(C^N - s)^+] \tag{4.31}$$

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

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

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

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

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

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

Hence, we obtain

$$\inf_{\pi\in\Pi^{pM}} w_{n\pi} \geq \inf_{\pi\in\Pi} w_{n\pi} \geq \inf_{\widetilde{\pi}\in\widetilde{\Pi}} w_{n\widetilde{\pi}} = \inf_{\pi\in\Pi^{pM}} V_{n\pi} = \inf_{\pi\in\Pi^{pM}} w_{n\pi}$$

We conclude the proof. □

Theorem 4.5. *[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*

$$\inf_{\pi\in\Pi} \mathbb{E}_x^{\pi}[(C^N - s)^+] = \mathbb{E}_x^{\sigma^*}[(C^N - s)^+] \tag{4.32}$$

Now, we are ready to state our main result.

Theorem 4.6. *Under Assumptions 2.1, there exists an optimal Markov policy π^* for the infinite horizon problem 2.15.*

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

$$\begin{aligned} w_{\infty,\pi} &= \mathbb{E}_x^{\pi}[(C^{\infty} - s)^+] \\ &= \mathbb{E}_x^{\pi}[(C^n + \sum_{k=n+1}^{\infty} C_k - s)^+] \\ &\leq E_x^{\pi}[(C^n - s)^+] + E_x^{\pi}[\sum_{k=n+1}^{\infty} C_k], \\ &\leq E_x^{\pi}[(C^n - s)^+] + M(n), \end{aligned} \tag{4.33}$$

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

$$w_{\infty}(x, s) \leq w_n + M(n) \tag{4.34}$$

Hence we get

$$w_n \leq w_{\infty}(x, s) \leq w_n + M(n) \tag{4.35}$$

Letting $n \to \infty$, we get

$$\lim_{n\to\infty} w_n = w_{\infty} \tag{4.36}$$

Moreover, by Theorem 4.4, there exists $\pi^* = \{f_n\}_{n=0}^N \in \Pi$ such that $V_{\pi}^N(x) = V_{0,N}^*(x)$. By the nonnegativity of the cost functions $c_n \geq 0$, we have that $N \to V_{0,N}^*(x)$ is nondecreasing and $V_{0,N}^*(x) \leq V_{0,\infty}^*(x)$ for all $x \in X$. Denote

$$u(x) := \sup_{N>0} V_{0,N}^*(x). \tag{4.37}$$

Letting $N \to \infty$, we have $u(x) \leq V^*_{0,\infty}(x)$. □

We recall that our optimization problem is

$$\inf_{\pi \in \Pi} \text{AVaR}^{\pi}_{\alpha}(\sum_{n=0}^{\infty} c(x_n, a_n, \xi_n)), \tag{4.38}$$

which is equivalent to

$$\inf_{\pi \in \Pi} \text{AVaR}^{\pi}_{\alpha}(\sum_{n=0}^{\infty} c(x_n, a_n)) = \inf_{s \in \mathbb{R}} \left\{ s + \frac{1}{1-\alpha} \inf_{\pi \in \Pi} \mathbb{E}^{\pi}_{x}[(C^{\infty} - s)^{+}] \right\} \tag{4.39}$$

Hence, we fix the global variable a priori s as

$$s = \text{VaR}^{\pi_0}_{\alpha}(C^{\infty}), \tag{4.40}$$

where $\text{VaR}^{\pi_0}_{\alpha}(C^{\infty})$ is decided using the reference probability measure $\mathbb{P}_0$.

Remark 4.7. *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 4.23 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 $\text{VaR}^{\pi_0}_{\alpha}(C^N)$ with respect to reference probability measure P and then solve the optimization problem for each path ω 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 ω whenever $s_n \leq 0$.*

5 $s_n(\omega) \leq 0$ case for that particular realization ω

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 ω. Recall that the

inner optimization problem is

$$\begin{aligned}
V_0^*(x) &= \frac{1}{1-\alpha}\inf_{\pi\in\Pi}\mathbb{E}_x^\pi[(C^\infty - s)^+]. \\
&= \frac{1}{1-\alpha}\inf_{\pi\in\Pi}\mathbb{E}_x^\pi\Big[\big(\sum_{n=N+1}^{\infty} c(x_n,a_n) - (s - C^N)\big)^+\Big] \\
&= \frac{1}{1-\alpha}\inf_{\pi\in\Pi}\mathbb{E}_x^\pi\Big[\big(\sum_{n=N+1}^{\infty} c(x_n,a_n) - s_n\big)^+\Big] \qquad (5.41)\\
&= \frac{1}{1-\alpha}\inf_{\pi\in\Pi}\mathbb{E}_x^\pi\Big[\mathbb{E}_x^\pi\Big[\big(\sum_{n=N+1}^{\infty} c(x_n,a_n) - s_n\big)^+|\mathcal{F}_n\Big]\Big] \qquad (5.42)\\
&= \frac{1}{1-\alpha}\inf_{\pi\in\Pi}\mathbb{E}_x^\pi\Big[\mathbb{E}_x^\pi\Big[\big(\sum_{n=N+1}^{\infty} c(x_n,a_n) - s_n\big)^+|\{x_n,s_n\}\Big]\Big] \qquad (5.43)
\end{aligned}$$

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

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

where $n = \min\{m \in \mathbb{N}_0 : s_m(\omega) \le 0\}$ in that realization path ω. To further proceed, we need the following two technical lemmas.

Lemma 5.1. *Fix an arbitrary $n \in \mathbb{N}_0$. Let $\mathbb{K}_n$ be as in Assumption 2.2, and let $u : \mathbb{K}_n \to \mathbb{R}$ be a given measurable function. Define*

$$u^*(x) := \inf_{a\in A_n(x)} u(x,a), \ \textit{for all } x \in X_n. \qquad (5.44)$$

- *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*

$$u^*(x) = u(x,\pi_n), \ \textit{for all } x \in X \qquad (5.45)$$

  and u^* is measurable.

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

Proof. See [25]. □

Lemma 5.2. *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 $N \to \infty$, then*

$$\lim_{N\to\infty} \min_{a\in A_n(x)} w_{n,N}(x,a) = \min_{a\in A_n(x)} w_n(x,a) \tag{5.46}$$

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 $c(x_i, a_i, \xi_i)$ for all $i \in \mathbb{N}_0$, we have obviously

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

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

and similarly for the infinite horizon problem, we have

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

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

Definition 5.3. *A sequence of functions $u_n : X \to \mathbb{R}$ on a realization path ω at time n is called a solution to the optimality equations if*

$$u_n(x)(\omega) = \inf_{a\in A(x)} \{c_n(x, a, \xi_n)(\omega) + \mathbb{E}[u_{n+1}[F_n(x, a, \xi_n)]]\}, \tag{5.47}$$

where

$$\mathbb{E}[u_{n+1}[F_n(x, a, \xi_n)]] = \int_{S_n} u_{n+1}[F_n(x, a, s)]\mu_n(ds). \tag{5.48}$$

We introduce the following notation for simplicity.

$$P_n u(x)(\omega) := \min_{a\in A_n(x)} \{c_n(x, a)(\omega) + \mathbb{E}[u_{n+1}[F_n(x, a, \xi_n)]\}, \tag{5.49}$$

for all $x \in X$, and for every $n \in \mathbb{N}_0$. Let $L_n(X_n)$ be the family of l.s.c. non-negative functions on X_n.

Lemma 5.4. *Under the Assumption 2.2, the followings hold.*

- *P_n maps $L_{n+1}(X)$ into $L_n(X)$.*
- *For every $u_{n+1} \in L_{n+1}(X)$, there exists an optimal action $a_n^* \in A(x)$ attaining the minimum in 5.47, i.e.*

$$P_n u(x)(\omega) := \{c_n(x, a_n, \xi_n)(\omega) + \mathbb{E}[u_{n+1}[F_n(x, a_n, \xi_n)]\}, \tag{5.50}$$

Proof. Let $u_{n+1} \in L_{n+1}(X)$. Then by Assumption 2.2, for fixed ω, we have that the function

$$(x, a, \omega) \to c_n(x, a, \omega) + \mathbb{E}[u_{n+1}[F_n(x, a_n, \xi_n)] \tag{5.51}$$

is non-negative and l.s.c. and by Lemma 5.1, there exists $\pi_n \in \mathbb{F}_n$ that satisfies Equation 5.49 and $P_n u$ is l.s.c. So we conclude the proof. □

By dynamic programming principle, we express the optimality equations in 5.47 as

$$V_m^* = P_m V_{m+1}^*, \tag{5.52}$$

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

Lemma 5.5. *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 Lemma 5.4, there exists a policy $\pi = \{f_m\}_{m \geq n}$ such that for all $m \geq n$

$$u_m(x) \geq c_m(x_m, a_m, \xi_i) + u_{m+1}(x_{m+1}). \tag{5.53}$$

By iterating, we have

$$u_m(x) \geq \sum_{i=m}^{N-1} c_i(x_i, a_i, \xi_i) + u_{m+N}(x_{m+N}), \tag{5.54}$$

Hence we have

$$u_m(x) \geq V_{m,N}(x, \pi), \tag{5.55}$$

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

Theorem 5.6. ***(Value Iteration)*** *Suppose that assumptions hold, then for every $m \geq n$ and $x \in X$,*

$$V_{n,N}^*(x) \uparrow V_n^*(x), \tag{5.56}$$

as $N \to \infty$.

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 $t = N-1, N-2, ..., n$, and let

$$J_t(x) := \inf_{a \in A_t(x)} \{c_t(x, a) + J_{t+1}[F_t(x, a, \xi)]\}. \tag{5.57}$$

By backward iteration, for $t = N-1, ..., n$, there exists $\pi_t \in \mathbb{F}_m$ such that $\pi_m(x) \in A_m(x)$ attains the minimum in the Equation 5.57, and $\{\pi_{N-1}, \pi_{N-2}, ..., \pi_n\}$ is an optimal policy. Moreover, J_n is the optimal cost for

$$J_n(x) := V^*_{n,N}(x_n), \tag{5.58}$$

Hence, we have

$$V^*_{n,N}(x) = \min_{a_n \in A(x)} \{c_n(x_n, a_n, \xi_n) + V^*_{n+1,N}[F_n(x_n, a_n, \xi_n)]\}. \tag{5.59}$$

By Lemma 5.2, we have

$$V^*_n(x) = \min_{a \in A_n(x)} \{c_n(x, a) + V^*_{n+1}[F_n(x, a, \xi)]\}. \tag{5.60}$$

Moreover, cost functions $c_n(x_n, a_n, \xi_n)$ 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. □

6 Examples and Applications

In the examples below, we emphasize that we *do not* find the optimal solution verified theoretically above. Using that the variable s_0 is the indicator to apply dynamic programming or not, we divide the problem into two sub-problems. Until dynamic programming can be applied, we confine ourselves to *greedy* algorithm and solve the optimization problem at that time step n. After, we are allowed to apply dynamic programming we switch to that scheme and accumulate the total cost for the problem.

6.1 LQR Problem

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 equation

$$F(x_n, a_n, \xi_n) = x_n + a_n + Z_n \tag{6.61}$$

$$x_{n+1} = x_n + a_n + Z_n, \tag{6.62}$$

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 $c(x_n, a_n, \xi_n) = x_n^2 + a_n^2$ for $n = 0, 1, ..., N-1$, hence it is continuous in both a_n and x_n, and nonnegative satisfying the Assumption 2.2. We also assume that the control constraint sets $A(x)$ with $x \in X$ are all equal to $A = [0, 1]$, where $X = \mathbb{R}$. Thus, under the above assumptions , we wish to find a policy that minimizes the performance criterion

$$J(\pi, x) := \text{AVaR}_\alpha^\pi\bigg(\sum_{n=0}^{N-1}(x_n^2 + a_n^2)\bigg), \tag{6.63}$$

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

$$\begin{aligned} &K_N = 0 \\ &K_n = \bigg[1 - (1 + K_{n+1})^{-1} K_{n+1}\bigg] K_{n+1} + 1, \text{ for } n = 0, ..., N-1 \\ &f_n(x) = -(1 + K_{n+1})^{-1} K_{n+1} \\ &J_n(x) = K_n x^2 + \sum_{i=n+1}^{N-1} K_i, \text{ for } n = 0, ..., N-1 \end{aligned} \tag{6.64}$$

for every $x \in X$. (see e.g. [13]). When we use AVaR operator, we proceed as follows. First, we choose the global variable s_0 a-priori and fix it. Our scheme suggests that when $s_0 \leq 0$, then the problem reduces to risk neutral model. Hence, the variable s_0 determines our *risk avereness* level. Ideally,

$$s_0 := \text{VaR}_\alpha^{\pi^*}(\sum_{n=0}^{N-1} c(x_n, a_n)),$$

for an optimal policy π^*. Instead, heuristically, we take that

$$s_0 = \inf\bigg\{x \in \mathbb{R} : \mathbb{P}\bigg(\sum_{n=0}^{N-1} Z_n^2\bigg)\bigg\}, \tag{6.65}$$

where $Z_n \backsim \mathcal{N}(0,1)$ as above. We note that our initial s_0 is positive and $\sum_{n=0}^{N-1} Z_n^2$ has χ^2 distribution with $n-1$ degrees of freedom. We start at time $n = 0$. If $s_0 > 0$,

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. Then, we update global variable s with $s - c_n(x_n, a_n)$, namely, $s - x_n^2$. 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 $M = 10000$ and find that our scheme preserves the monotonicity property of $\text{AVAR}_\alpha(X)$ operator, namely we have $\text{AVaR}_\alpha(X) \leq \text{AVaR}_\alpha(Y)$, whenever $X \leq Y$. 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 $\alpha_1 \leq \alpha_2$. That is to say, increasing our initial risk aversion level s_0 a priori, we see that the value function is increasing correspondingly, as expected. Our algorithm also soatisfies that for $\alpha = 0$, we have the risk neutral value functions which is consistent with $\lim_{\alpha \to 0} \text{AVaR}_\alpha(X) = \mathbb{E}[X]$. We give the pseudocode of this algorithm below and present our simulation results afterwards.

procedure LQ-AVaR Algorithm
  $s = \text{VaR}_\alpha^{\pi_0}(\sum_{n=0}^{N-1} Z_n^2)$
  $x = 0$
  $V^{dyn} = 0$
  $V(x) = 0$
  for each $n \in N - 1$ **do**
    if $s \leq 0$ **then**
      apply Dynamic Programming from state x_n onwards as in Equation 6.64
      Update V^{dyn}
    else
      Choose $a_n = 0$
      Update $s = s - x_n^2$
      Update $c_n = x_n^2 + a_n^2$
      Update $x_{n+1} = x_n + a_n + \xi_n(\omega)$
      Update $V(x) = V(x) + c_n$
    end if
  end for
return $V(x) + V^{dyn}$
end procedure

6.2 Simulation Results

α	N	Value	α	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

α	N	Value	α	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

α	N	Value	α	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

α	N	Value	α	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

α	N	Value	α	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

α	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

6.3 Inventory-Production System

Consider an inventory-production system, in which x_n is the stock level at time n, a_n the quantity ordered (or produced) at time n and ξ_n stands for the demand at time n. The "disturbance" or "exogenous" variable ξ_n is the demand during that period. We assume ξ_n to be i.i.d. random variables. We take that $A = X = \mathbb{R}$. Hence, we allow "negative" stock levels by assuming that excess demand is backlogged and filled when additional inventory becomes available. Thus, the system equation is of the form

$$x_{n+1} = x_n + a_n - \xi_n, \tag{6.66}$$

for $n = 0, 1, , ...$ We note that $F(x_n, a_n, \xi_n)$ is continuous on $\mathbb{K}_n$ as required in the Assumption 2.2 for our framework. We wish to minimize the operation cost and use our scheme for that. Suppose one-stage cost function is of the form

$$c(x_n, a_n, \xi_n) = b \cdot a_n + h \cdot \max(0, x_{n+1}) + p \cdot \max(0, -x_{n+1}), \tag{6.67}$$

where b stands for the unit production cost, h is the unit handling cost for excess inventory, and p stands for the penalty for unfilled demand with $p > b$, where these unit costs are all positive, where we note that for fixed ξ_n the cost functions $c(x_n, a_n, \xi_n)$ is continuous and inf-compact, hence necessarily satisfy the Assumption 2.2. Furthermore, we take that the demand variables ξ_n are non-negative, i.i.d. random variables, independent of the initial stock X_0; their probability distribution function is denoted by ν, that is, $\nu(s) := P(\xi_0 \leq s)$, for every $s \in \mathbb{R}$ with $\nu(s) = 0$, if $s < 0$. We also assume that the mean demand $\mathbb{E}(\xi_0)$ is finite. Moreover, $c(x, a, \xi)$ is continuous in (x, a) for fixed ξ and non-negative, hence satisfy the requirements in Assumption 2.2. It is well known that in risk neutral case the minimization problem

$$\min_{\pi \in \Pi} \mathbb{E}_x^{\pi} \left[\sum_{n=0}^{N} c(x_n, a_n, \xi_n) \right], \tag{6.68}$$

has an optimal Markovian policy $\pi = \{f_n\}_{n=0}^{\infty}$ which satisfies the following optimality equations

$$f_n(x) = \begin{cases} 0, & \text{if } x \geq K_n \\ K_n - x, & \text{if } x < K_n \end{cases} \tag{6.69}$$

for some threshold constant K_n updated at each time n retrieved from the corresponding dynamic programming equations and value functions. We refer the reader to [13] for further details. In the risk averse case, we are interested in solving the following optimization problem.

$$\min_{\pi \in \Pi} \text{AVaR}_{\alpha}^{\pi} \left[\sum_{n=0}^{N} c(x_n, a_n, \xi_n) \right], \tag{6.70}$$

we use our scheme. Namely, as in our previous example of LQR problem, we choose the positive variable s_0 a priori, first. Depending on our *risk avereness* level, as we increase s_0, the risk avereness increases.

Next, we determine a_0 as

$$a_0 = \arg\min_{a_n \in \mathbb{R}} \left(b \cdot a_n + h \cdot \max(0, x_{n+1}) + p \cdot (0, -x_{n+1} - s_0)^+ \right). \tag{6.71}$$

Then, we calculate c_0 and update $s_1 = s_0 - c_0$. If $s_1 \leq 0$ apply dynamic programming onwards. Otherwise, simulate ξ_1. Update x_1 and solve the one step optimization problem as in Equation 6.71. Update c_1 and let $s_2 = s_1 - c_1$ and check whether s_2 is negative or not and repeat the procedure. We give the algorithm of this scheme below.

procedure INVENTORY-ALGORITHM
    Choose $s_0 > 0$ heuristically based on the risk avereness level.
    $x_0 > 0$
    $V^{dyn} = 0$
    $V(x) = 0$
    for each $n \in N - 1$ **do**
        if $s \leq 0$ **then**
            apply Dynamic Programming from x_n onwards with the as in Equation 6.69
            Update V^{dyn}
        else
            Determine a_n by Equation 6.71.
            Update c_n as in Equation 6.67.
            Update $s_{n+1} = s_n - c_n$.
            Simulate ξ_n.
            Update $x_{n+1} = x_n + a_n - \xi_n(\omega)$
            Update $V(x) = V(x) + c_n$
        end if
    end for
return $V(x) + V^{dyn}$
end procedure

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 L1* , 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 Lp-spaces*, Mathematical Methods in Operations Research, 475-495.

[18] Ruszcynski, 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, 16311658.

[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, 265293.

[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, 667678.

[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, 907909.

[34] Roorda, B. and Schumacher, J.(2016), *Weakly time consistent concave valuations and their dual representations*. Finance and Stochastics, 20, 123-151.

[35] Goovaerts, M.J. and Laeven, R.(2008), *Actuarial risk measures for financial derivative procing*. Insurance: Mathematics and Economics, 42, 540-547.

[36] Godin, F.(2016), *Minimizing CVaR in global dynamic hedging with transaction costs*(2016), Quantitative Finance, 6, 461-475.

[37] Balbas, A., Balbas, R. and Garrido, J.(2010), *Extending pricing rules with general risk functions*, European Journal of Operational Research, 201, 23-33.

[38] Bertsekas, D.P., Shreve, S.(1978), *Stochastic Optimal Control:The Discrete Time Case*. Academic Press, New York, 1978.

[39] Hernandez-Lerma, O.(1989), *Adaptive Markov Control Processes*, Springer-Verlag. New York.

[40] Hernandez-Lerma, O. and Rungladier, W.(1994), *Monotone approximations for convex stochastic control problems*, Journal of Mathematical System, Estimation and Control, 99-144.

[41] Bensoussan, A.(1982), *Stochastic control in discrete time and applications to the theory of production*, Math. Programminsg Study, 18, 43-60.

[42] Bertsekas, D.P.(1978), *Dynamic programming: deterministic and stochastic models*. Prentice-Hall, Englewood Cliffs, New Jersey.

[43] Dynkin, E.B. and Yushkevich, A.A.(1979), *Controlled markov processes*. Springer-Verlag, New York.

[44] P. Carpentier, J.P. Chancelier, G. Cohen, M. De Lara and P. Girardeau, Dynamic consistency for stochastic optimal control problems, *Annals of Operations Research*, 200 (2012), 247-263.

[45] A. Shapiro, On a time consistency concept in risk averse multi-stage stochastic programming, Operations Research Letters 37 (2009) 143-147.