**Professor Julio B. CLEMPNER, PhD**
**E-mail: julio@clempner.name**
**Instituto Politécnico Nacional, Mexico**

# AN ERGODIC AND TRANSIENT MARKOV MODEL FOR PENALTY REGULARISED PORTFOLIO

***Abstract.*** *A collection of finite-number assets with discontinuous ergodic and transient return models are the subject of this paper, and its objective is to present a novel mean-variance Markowitz portfolio solution for them. We provide a methodology for choosing a portfolio using a penalty regularised Markov algorithm. In order to ensure the convergence to a singular portfolio solution, Tikhonov's regularisation, one of the most well-liked methods for handling discrete ill-posed problems, is crucial. We propose optimality conditions and relations for this system and build equivalent penalty regularised optimisation models. We begin by describing the qualities and conditions that permit the identification of a single solution for the penalty regularised projected utility of a given optimal portfolio. The Markowitz portfolio optimisation problem therefore converges to a unique solution. Together with the limitations, the utility function is also penalised. For both the ergodic and the transient models, we compute a stationary policy. The new strategy is demonstrated using a numerical example.*

***Keywords:*** *Portfolio, Markowitz, Markov chains, ergodic model, transient model, optimisation*

**JEL Classification: G11, C61, C69**

## 1. Introduction

### *1.1. Brief review*

The issue of choosing an ideal portfolio has been the focus of extensive theoretical and computational research in finance. A conceptually sound method of comparing uncertain outcomes is a basic problem in dealing with them. Markov decision processes (MDPs) constitute a powerful paradigm. We address the topic of penalised restrictions in MDPs in the current work. Many different kinds of constraints have been successfully handled using portfolio methods for MDPs. Finding penalised constraints for MDPs that can be studied using analytical formulation is our main objective.

The model of Markov chains plays a fundamental role in the portfolio (Clempner and Poznyak, 2018). The traditional approach, founded by Markowitz (1952), is based on the mean-risk approach, that compares the portfolios with respect to two features: a) the expected return rate (the mean), and b) the variance (the risk), which is given by some scalar measure of the uncertainty of the portfolio return rate. Regarding these two factors, the mean-risk method suggests choosing Pareto-



DOI: 10.24818/18423264/57.4.23.17

efficient portfolios. We incorporate these criteria in a mean-risk portfolio model by giving a parameter as a tradeoff between them. The mean-risk model can be solved numerically extremely effectively as a parametric optimisation problem, which makes this strategy particularly appealing.

### *1.2. Markowitz model*

Let $\{x_1, x_2, \ldots, x_n\} \in X$ the fractions of the initial capital invested in assets $1,2,\ldots,n$, the defining of the preference structure among realistic portfolios is the main challenge in formulating a meaningful portfolio optimisation issue. If we consider the mean return rate $\mathrm{E}[R(x)]$, followed by the subsequent optimisation problem has a useless and easy solution: make every effort asset with the highest predicted rate of return. In these, the process of portfolio optimisation typically turns to two methods.

In the method, a dispersion measure is connected to the portfolio $\mu(R(x))$ is a representation of the return rate's variability $R(x)$. The function in the traditional Markowitz model $\mu(R(x))$ is the return rate's variance

$$\mu(R(x)) = Var[R(x)]$$

but there are a lot of additional options available as well.

Then, the mean-risk portfolio optimisation problem is formulated

$$\max_{x \in X} \mathrm{E}[R(x)] - \gamma Var[R(x)]$$

The nonnegative parameter $\gamma$ represents the preferred mean-to-risk conversion rate. If $\gamma = 0$, there is no value to the risk, and the issue is then reduced to one of maximising the mean. If the value is greater than zero, $\gamma > 0$, we seek a balance between the mean and the risk. Alternately, one can consider a family of problems parametrised by $m$ while minimising the risk function $\mu(R(x))$ and fixing the anticipated return rate $\mathrm{E}[R(x)]$ at some value $m$.

### *1.3. Related Work*

Risk management for MDPs has been studied from a number of perspectives in the literature. In Filar and Kallenberg (1989), penalties for the erratic nature of rewards in MDPs are mentioned. The best policy is obtained by solving a nonlinear programming problem including occupation metrics. More research on the mean-variance trade-off in MDPs is done in Sobel (1994) in the context of Pareto-optimality. The conditional value-at-risk of the total cost is constrained in Asiain et al. (2018) for MDPs with limited horizons. An offline iterative approach is employed to determine the optimum course of action because convex analytical techniques are thought to be inadequate for this sort of problem. Markov risk measures are developed by Ruszczyński (2010) for discounted MDPs with limited and infinite horizons.

By extending the work of Sánchez et al. (2015) by offering a recurrent reinforcement-learning technique modifies policies using preprocessing and an Actor-Critic architecture. Clempner and Poznyak (2018) investigated whether the

mean-variance Markowitz customer portfolio optimisation issue could be solved using the penalty regularised expected utilities technique. A reinforcement-learning technique was created by Asiain et al. (2018) to compute the portfolio for controllable partially observable Markov chains.

Almahdi and Yang (2017) provide an adaptive-based strategy that produces a variable weight allocation system to increase projected drawdown by expanding the work done in recurrent reinforcement learning (Moody and Saffell, 2001). Among reinforcement-based approaches, continuous time and continuous control-based techniques have recently gained popularity. In this way, Aboussalah and Lee's multi-dimensional state space may be addressed by layered deep dynamic recurrent RL (Aboussalah and Lee, 2010). A continuous-time portfolio strategy was examined by García-Galicia et al. (2019). Assuming that the underlying asset portfolio process has a continuous-time discrete-state Markov chain, García-Galicia et al. (2019) investigated continuous-time reinforcement learning in the context of policy optimisation for financial portfolio management.

Domínguez and Clempner (2010) resolve the multi-period mean-variance Markowitz's portfolio optimisation problem. The regularisation of Tikhonov has generated a lot of attention in the application industries. The idea of using Tikhonov's regularisation to generate effective algorithms is currently being researched. In this regard, Vázquez and Clempner (2020) proposed a portfolio strategy based on a Lagrange regularisation technique. There are numerous solutions based on Tikhonov's regularisation that have been developed to solve the Markowitz Mean-Variance Portfolio Mode (Carrasco and Noumon, 2011; Fastrich et al., 2015).

The application industries are quite interested in the regularisation of Tikhonov. It is still being investigated if Tikhonov's regularisation can be used to create efficient algorithms. In this context, a portfolio approach based on a Lagrangian regularisation technique was put out by Vázquez and Clempner (2020). Carrasco and Noumon (2010) and Fastrich et al. (2015) have created a number of techniques based on Tikhonov's regularisation to deal with the Markowitz portfolio mode. Using a finite or infinite horizon and continuous or discrete time might result in different results in the literature (Akian et al., 2001; Cvitanic and Karatzas, 1996; Davis and Norman, 1990; Liu, 2020; Ziemba, 2012; Nowak and Romaniuk, 2018; Mwanakatwe et al., 2019).

### *1.4. Main results*

The Markowitz portfolio optimisation problem is addressed in this paper using Markov chains.

- Optimises the portfolio performance considering ergodic and transient constraints. This approach has a fundamental advantage over the classical models.
- Proposes a novel method for choosing a portfolio using a penalty regularised Markov technique.

- Guarantees the convergence to a unique portfolio solution dealing with discrete ill-posed problems using Tikhonov's regularisation.
- Provides comparable penalty regularised optimisation models and optimal requirements for this system.
- Starts by outlining the characteristics and circumstances under which a single solution can be found for the penalty regularised projected utility of a certain optimum portfolio.
- Employs the regularised poly-linear optimisation problem's parameters to solve the problem.
- Shows that the portfolio problem converges to a unique solution. Along with the limitations, the utility function is also penalised.
- Computes a stationary policy for both, the ergodic and the transient model
- Demonstrates the new approach using a numerical example.

### *1.5. Organisation of the paper*

This paper is organised as follows. In section 2, we consider preliminaries for MDPs. In section 3 we develop the background for ergodic and transient systems. Section 4 formulates the portfolio problem. In section 5, we discuss the regularised penalty function for Markowitz portfolio. We illustrate our method in section 6 with a portfolio optimisation example, and then conclude the paper in section 7.

## 2. Preliminaries

A *Markov chain* (MC) is a tuple $(\mathcal{S},\mathcal{A},P)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s'|s,a)_{s',s\in\mathcal{S},a\in\mathcal{A}}$ is the transition probability distribution where $P(s'|s,a)\in 0,1]$ and $\sum_{s'\in\mathcal{S}}P(s'|s,a)=1$. A *Markov decision process* is a tuple $MDP=(MC,c)$ where $c(s',s,a)$ is the utility function. During the process, the agent observes the current state of the environment $s_n$, applies control $a_n$, and incurs in a utility $c_n$ that is a known function of time $(s_{n+1},s_n,a_n)$ at each discrete time $n=1,2,\ldots\in N\subseteq\mathbb{N}$.

The transition matrix $P(s_{n+1}|s_n,a_n)$ and the distribution $P(s_0)$ describe the dynamics of a Markov chain, where $P(s_n)\in\Delta(\mathcal{S})$ and $\Delta(\mathcal{S})$ is employed to indicate the collection of all probability distributions over $\mathcal{S}$. The Markov chains are self-contained. The absolute values of the utility function $u(s',s,a)$ are bounded by some constant. We assume that each Markov chain $(P(s_n),P(s_{n+1}|s_n,a_n))$ is irreducible, recurrent and aperiodic (ergodic), and that $P$ is its unique invariant distribution. Then, we have $P(s_{n+1})=\sum_{s_n\in\mathcal{S}}\sum_{a_n\in\mathcal{A}}P(s_{n+1}|s_n,a_n)P(s_n)$. There exists, as well, a state $s^*$ which is recurrent for every distribution $P$.

Let $\pi:\mathcal{S}\to\Delta(\mathcal{A})$ be a *stationary policy*, where $\Delta(\mathcal{A})$ is the $\mathcal{A}$-simplex, which maps the state-space $\mathcal{S}$ to a probability distribution on the action-space $\mathcal{A}$ and determines randomised actions based on the current state $s_n$. Let $\Pi_{adm}$ be the admisible set of Markov policies, i.e.

$$\Pi_{adm} = \left\{ \pi(a_n|s_n) \left| \sum_{a_n \in \mathcal{A}} \pi(a_n|s_n) = 1, s_n \in \mathcal{S}, a_n \in \mathcal{A} \right. \right\}.$$

The utility at state vector $s_n$, policy $\pi(a_n|s_n)$ and probability $P(s_n)$ is given by:

$$\mathcal{C}(\pi) = \sum_{s_n \in \mathcal{S}} \sum_{a_n \in \mathcal{A}} \left( \sum_{s_{n+1} \in \mathcal{S}} c(s_{n+1}, s_n, a_n) P(s_{n+1}|s_n, a_n) \right) \pi(a_n|s_n) P(s_n) =$$

$$\sum_{s_n \in \mathcal{S}} \sum_{a_n \in \mathcal{A}} W(s_n, a_n) \pi(a_n|s_n) P(s_n),$$

where

$$W(s_n, a_n) = \sum_{s_{n+1} \in \mathcal{S}} c(s_{n+1}, s_n, a_n) P(s_{n+1}|s_n, a_n).$$

A *policy* $\{\pi_n\}_{n \geq 0}$ is called a *optimal* if for each $n \geq 0$ minimises the conditional mathematical expectation of the utility-function $\mathcal{C}(\pi)$ by solving

$$\pi^* := \arg \max_{\pi \in \Pi_{adm}} \mathbb{E}\{\mathcal{C}(\pi)|\mathcal{H}_n\}$$

cuch that $\mathcal{C}(\pi)$ is the average utility function.

For ergodic Markov chains, the strong law of large numbers governs not only the objective requirement of average utility but also the stronger equality

$$\lim_{N \to \infty} N^{-1} \sum_{n=1}^{N} \mathcal{C}(\pi_n) = \max_{\pi \in \Pi_{adm}} \mathcal{C}(\pi)$$

Note that the limit value on the left-hand side of the last equality is independent of the chain's initial state. A policy $\pi^* \in \Pi_{adm}$ is average optimal if $\mathcal{C}(\pi) = \max_{\pi \in \Pi_{adm}} \mathcal{C}(\pi)$. It is well known that there exists an average optimal deterministic policy $\pi^*$.

For any policy $\pi \in \Pi_{adm}$ and any positive integer $N$, we denote *state-action frequencies* in the first $N$ periods by $x(s_n, a_n)$

$$x(s, a) = \lim_{N \to \infty} N^{-1} \sum_{n=1}^{N} \sum_{s_{n+1} \in \mathcal{S}} P(s_n) P(s_{n+1}|s_n, a_n) \pi(a_n|s_n)$$

The $MDP$ is reduced to a stochastic process after the initial state and strategy are determined. A discrete-time *stochastic process* is a collection of random variables that are included into a random set $\{X_n | n \in N\}$. The value $X_n \in S$ is the state of the process at time $n$.

The distribution vector $P(s_n)$ is defined as

$$P(s_{n+1}) = \sum_{s_{n+1} \in \mathcal{S}} \sum_{s_n \in \mathcal{S}} \left( \sum_{a_n \in \mathcal{A}} P(s_{n+1}|s_n, a_n) \pi(a_n|s_n) \right) P(s_n)$$

In the ergodic case, we ha that , $P(s_n) \underset{n \to \infty}{\to} P(s)$. We shall now focus on stationary distributions.

## 3. Erogodicity and Non-ergodic

### *3.1. Ergodic Markov chains*

Linear programming provides the tools needed to compute the optimal policies in problem (1). We only consider the stationary policy situation. Let us present this problem in terms of ergodic Markov chains with only communicating states, that is, there is a probability that for a finite number of transitions each state $s'$ can be reached from any state $s$ applying action $a$.

Let us first introduce the following definitions

1. the optimal action $a^* = a(s^*)$
2. the limiting probabilities of the states under the optimal policy is given by $x(s,a) = \pi(a|s)P(s)$
3. the maximum of the objective function is given

$$C^*(x(s,a)) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} W(s,a)x^*(s,a),$$

such that

$$\left.\begin{array}{l} 1) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} x(s,a) = 1 \\ 2) \sum_{a\in\mathcal{A}} x(s,a) > 0 \\ 3) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left[\kappa_{s',s} - P(s'|s,a)\right] x(s,a) = 0, s' \in \mathcal{S} \end{array}\right\} (*)$$

where $\kappa_{s',s}$ is Kroneker's variable and we denote the above set of constraints $(*)$ by $X_{adm}$.

Then, we have that the linear programming problem can be expressed

$$4) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} W(s,a)x(s,a) \to \max_{x\in X_{adm}}$$

**Theorem 1.** *The policy $\pi(s,a)$ is optimal if and only if $x(s,a)$ is a solution to the linear programming problem (3).*

*Proof.* Let $x(s,a)$ be a feasible solution to problem (3) and $x(s,a) = \pi(a|s)P(s)$:

1) Each $x(s,a)$ represents the "long-run average frequency" that belong to the simplex

$$\Delta = \left\{ x(s,a) \left| \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} x(s,a) = 1, x(s,a) > 0 \right. \right\}.$$

2) We have that

$$\sum_{a\in\mathcal{A}} x(s,a) = \sum_{a\in\mathcal{A}} \pi(a|s)P(s) = P(s) \sum_{a\in\mathcal{A}} \pi(a|s) = P(s) > 0.$$

3) The ergodicity constraint satisfies that

$$P(s') = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} P(s'|s,a)P(s)$$

such that

$$\sum_{a\in\mathcal{A}} x(s',a) - \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} P(s'|s,a)x(s,a) = \sum_{s\in\mathcal{S}} \left[\sum_{a\in\mathcal{A}} (x(s',a) - P(s'|s,a)x(s,a))\right] =$$

$$\sum_{s_n\in\mathcal{S}} \left[\sum_{a_n\in\mathcal{A}} \left[\kappa_{s',s} - P(s'|s,a)\right]x(s,a)\right] = 0$$

4) The objetive function satisfies that taking $x(s,a) = \pi(a|s)P(s)$

$$\mathcal{C}(\pi) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left(\sum_{s'\in\mathcal{S}} c(s',s,a)P(s'|s,a)\right)\pi(a|s)P(s) =$$

$$\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} W(s,a)\pi(a|s)P(s) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} W(s,a)x(s,a) = C(x(s,a)).$$

As a result, we have

$$\min_{x\in X_{adm}} C(x(s,a)) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} W(s,a)x^*(s,a)$$

For recovering the relationship of interest $\pi(a|s)$ and $P(s)$ we have that $x(s,a) = \pi(a|s)P(s)$, then

$$\sum_{a\in\mathcal{A}} x(s,a) = \sum_{a\in\mathcal{A}} \pi(a|s)P(s) = P(s)\sum_{a\in\mathcal{A}} \pi(a|s) = P(s)$$

as a result

$$P(s) = \sum_{a\in\mathcal{A}} x(s,a)$$

Now

$$x(s,a) = \pi(a|s)P(s) = \pi(a|s)\sum_{a\in\mathcal{A}} x(s,a)$$

which implies

$$\pi(a|s) = \frac{x(s,a)}{\sum_{a'\in\mathcal{A}} x(s,a')}$$

The linear programming problem is more complicated for non-ergodic controlled Markov chains.

### *3.2. Non-ergodic controlled Markov chains*

Let $\tau_{ss}$ represent the return time to state $s$ for $X_0 = s$:

$$\tau_{ss} = \min\{n \geq 1 | X_n = s : X_0 = s\}$$

denotes that the Markov chain started in state $s$ it determines the amount of time (number of steps) until it returns to state $s$. The non return is defined by $\tau_{ss} = \infty$ if $X_n \neq s$ for $n \geq 1$. Denote by $\mu_s := P(\tau_{ss} < \infty)$ the probability of ever returning to state $s$ given that the chain started in state $s$. A state $s$ is called *recurrent* if $\mu_s = 1$ and *transient* if $\mu_s < 1$. If $\mu_s = 1$ then the chain will return to state $s$ infinitily often and if $\mu_s < 1$ it will only be visited a finite (random) number of times. Counting the total number of visits to state $s$ for $X_0 = s$ is

$$\eta_s = \sum\nolimits_{n=0}^{\infty} \mathbb{I}\{X_n = s : X_0 = s\}$$

where $\mathbb{I}\{X_n = s\}$ is the indicator function denoting if the $n - th$ state is $s$ having a geometric distribution $P(\eta_s = n) = \mu_s^{n-1}(1 - \mu_s), n \geq 1$. The expected number of visits is thus given by $\mathbb{E}\{\eta_s\} = (1 - \mu_s)^{-1}$ $(\sum_{n=0}^{\infty} P^n(s,s))$ for which a state $s$ is called recurrent if $\mathbb{E}\{\eta_s\} = \infty$ $(\sum_{n=0}^{\infty} P^n(s,s) = \infty)$ and transient if $\mathbb{E}\{\eta_s\} < \infty$ $(\sum_{n=0}^{\infty} P^n(s,s) < \infty)$. When the limits exist, let $s'$ denote the long-run proportion of time that the chain spends in state $s'$:

$$P(s') = \lim_{N \to \infty} \frac{1}{N} \sum\nolimits_{n=1}^{N} \mathbb{I}\{X_n = s'\}$$

denote the *long-run average frequency of state* $s'$.

Let us introduce for the stationary Markov strategies $\pi(a|s)$ the function $f : \mathcal{S} \to 2^{\mathcal{A}} \backslash \{\emptyset\}$ of the current state. The values $x(s,a)$ denote the action $a$'s choice probabilities in the state $x(s,a)$ or, in other words, the policy in the class of randomised strategies is being sought. The ergodicity case reveals that there exists a unique $f(s)$ for any $s$ such that $x(s,a) > 0$ and $x(s,a) = 0$ when $a \notin f(s)$ (the state will never be visited). $x(s,a)$ is identified as the l*ong-run average frequency of the state-action pair* $(s,a)$.

Let $E \subseteq \mathcal{S}$ be the reachable subset of states and and let $\mathcal{R} = \{R_1, R_2, \ldots, R_l\}$ be the set of closed connected recurrent set of states. The set $E$ and $\mathcal{R}$ can be computed in linear-time. Let $R = \cup_{i=1}^{l} R_i$, and the set $E \backslash R$ consists of transient states.

We will try to find an auxiliary variable $x(s,a)$ which solve the following linear programming problem

$$C(x(s,a)) \to \max_{x \in X_{adm}}$$

$$C(x) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} W(s,a)x(s,a)$$

where

$$\left.\begin{aligned}
&1)\sum_{s\in\mathcal{R}} \sum_{a\in\mathcal{A}} x(s,a) = 1\\
&2)\sum_{a\in\mathcal{A}} x(s,a) \geq 0, a \in \mathcal{A}, s \in R\\
&3)x(s,a) > 0, \qquad a \in f(s), s \in R\\
&4)x(s,a) = 0, \qquad a \notin f(s), s \in R\\
&5)\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left[\kappa_{s',s} - P(s'|s,a)\right]x(s,a) = 0, s' \in R
\end{aligned}\right\}(*)$$

$$\left.\begin{aligned}
&6)\sum_{a\in\mathcal{A}} y(s',a) - \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} P(s'|s,a)y(s,a) = \xi(s'), s' \in E\backslash R\\
&7)y(s') - \sum_{s\in E\backslash R} \sum_{a\in\mathcal{A}} P(s'|s,a)y(s,a) = 0, s' \in R\\
&8)\sum_{a\in\mathcal{A}} y(s,a) > 0, s \in R\\
&9)y(s,a) = 0, \qquad a \notin f(s), s \in E\backslash R\\
&10)\sum_{s\in R} \sum_{a\in\mathcal{A}} x(s,a) = \sum_{s\in R} y(s)
\end{aligned}\right\}(+)$$

The numbers $\xi(s')$ are an initial distribution of the chain (it is convenient to let $\xi(s')$ a uniform distribution). The solution defines the desired optimal variables $x(s,a)$ and $y(s,a)$, which refer to the states of the recurrent class and to the transient class, respectively.

For determining the relations between policies and feasible solutions $(x,y)$, we define a stationary policy as follows

$$\pi(s,a) = \begin{cases} \dfrac{x(s,a)}{\sum_{a\in\mathcal{A}} x(s,a)} & a \in \mathcal{A}, s \in R \\ \dfrac{y(s,a)}{\sum_{a\in\mathcal{A}} y(s,a)} & a \in \mathcal{A}, s \in E\backslash R \end{cases}$$

## 4. Portfolio

### *4.1. Mean-Variance Portfolio*

We are therefore proceeding with the notion of portfolio. The variance is given by

$$V(\pi)=\sum_{s_n\in\mathcal{R}}\sum_{a_n\in\mathcal{A}}[W^2(s,a)\pi(a|s)P(s)-C^2(x(s,a))]$$

where $W(s,a)$ are the rewards of the original MDP-process. The notion of variance based on the long-run frequency of occurrence of state-action which establish the rewards is defined

$$V(x)=\sum_{s_n\in\mathcal{R}}\sum_{a_n\in\mathcal{A}}[W(s,a)-C(x(s,a))]^2x(s,a),.$$

and

$$V(y)=\sum_{s_n\in E\setminus R}\sum_{a_n\in\mathcal{A}}[W(s,a)-C(y(s,a))]^2y(s,a).$$

This definition is equivalent to

$$V(x)=\lim_{N\to\infty}N^{-1}\sum\nolimits_{n=1}^{N}\mathbb{E}[W_n(s,a)-C(x(s,a))]^2$$

and

$$V(y)=\lim_{N\to\infty}N^{-1}\sum\nolimits_{n=1}^{N}\mathbb{E}[W_n(s,a)-C(y(s,a))]^2$$

where $\mathbb{E}(\cdot)$ denotes the expectation operator.

**Formulation of the problem**. We consider the following optimisation portfolio problem

$$\max_{\pi\in\Pi_{adm}}C(\pi)-\gamma V(\pi) \tag{1}$$

or in term of the long-run frequency we have

$$\left.\begin{array}{l}\max\limits_{x\in X_{adm}}C(x)-\gamma V(x)\\ \max\limits_{y\in Y_{adm}}C(y)-\gamma V(y)\end{array}\right\} \tag{2}$$

where $\gamma\in(0,\infty]$.

The Eq.(1) represents the average reward ($C(\pi)$) penalised by variance ($V(\pi)$) induced by the policy ($\pi\in\Pi_{adm}$). The Eq.(2) can be expressed as the following convex quadratic form

$$\max_{x\in X_{adm}}\left[\sum_{s\in R}\sum_{a\in\mathcal{A}}W(s,a)x(s,a)-\gamma\sum_{s\in R}\sum_{a\in\mathcal{A}}W^2(s,a)x(s,a)+\right.$$

$$+\gamma\left(\sum_{s\in R}\sum_{a\in\mathcal{A}}W(s,a)x(s,a)\right)^2\Bigg]+\max_{y\in Y_{adm}}\Bigg[\sum_{s\in E\setminus R}\sum_{a\in\mathcal{A}}W(s,a)y(s,a)-$$

$$-\gamma\sum_{s\in E\setminus R}\sum_{a\in\mathcal{A}}W^2(s,a)y(s,a)+\gamma\left(\sum_{s\in E\setminus R}\sum_{a\in\mathcal{A}}W(s,a)y(s,a)\right)^2\Bigg] \text{ where } \gamma\in(0,\infty]$$

The following theorems arise.

**Theorem 2.** *The policy $\pi^*$ is Pareto optimal*

*Proof.* Let us assume that $\pi^*$ is not optimal then there exist $\pi\in\Pi_{adm}$ $(\pi^*\neq\pi)$ such that $C(\pi)-V(\pi)\geq C(\pi^*)-V(\pi^*)$. Then, from Eq.(1) we have that $C(\pi)-\gamma V(\pi)>C(\pi^*)-\gamma V(\pi^*)$ for $\gamma\in(0,\infty]$. This is a contradiction to the fact that $\pi^*$ is the maximises Eq.(1).

**Theorem 3.** *Let the policy $\pi^*$ be optimal in Eq.(1), then $x^*$ is optimal in Eq.(2) and the maximum of Eq.(1) is the same as in Eq.(2)*

*Proof.* We have the following:

$$C(\pi^*)-\gamma V(\pi^*)=\max_{\pi\in\Pi_{adm}}C(\pi)-\gamma V(\pi)=$$

$$\max_{x\in X_{adm}}\Bigg[\sum_{s\in R}\sum_{a\in\mathcal{A}}W(s,a)x(s,a)$$

$$-\gamma\left(\sum_{s\in R}\sum_{a\in\mathcal{A}}W^2(s,a)x(s,a)-\left(\sum_{s\in R}\sum_{a\in\mathcal{A}}W(s,a)x(s,a)\right)^2\right)+$$

$$\max_{y\in Y_{adm}}\Bigg[\sum_{s\in E\setminus R}\sum_{a\in\mathcal{A}}W(s,a)y(s,a)-\gamma\sum_{s\in E\setminus R}\sum_{a\in\mathcal{A}}W^2(s,a)y(s,a)+$$

$$\gamma\left(\sum_{s\in E\setminus R}\sum_{a\in\mathcal{A}}W(s,a)y(s,a)\right)^2\Bigg]=$$

$$\Bigg[\sum_{s\in R}\sum_{a\in\mathcal{A}}W(s,a)x^*(s,a)-\gamma\sum_{s\in R}\sum_{a\in\mathcal{A}}W^2(s,a)x^*(s,a)$$

$$+\gamma\left(\sum_{s\in R}\sum_{a\in\mathcal{A}}W(s,a)x^*(s,a)\right)^2\Bigg]+$$

$$\Bigg[\sum_{s\in E\setminus R}\sum_{a\in\mathcal{A}}W(s,a)y^*(s,a)-\gamma\sum_{s\in E\setminus R}\sum_{a\in\mathcal{A}}W^2(s,a)y^*(s,a)$$

$$\gamma\left(\sum_{s\in E\setminus R}\sum_{a\in\mathcal{A}}W(s,a)y^*(s,a)\right)^2\Bigg]$$

Then, given that $x^*(s,a)=\pi^*(a|s)P^*(s)$ ( $s\in R$ ) and $y^*(s,a)=\pi^*(a|s)P^*(s)$ ($s\in E\backslash R$) the maximum of Eq.(1) is the same as in Eq.(2).

**Remark 1.** *The oposite statement holds. Let $x^*$ optimal in Eq.(2), then the policy $\pi^*$ is optimal in Eq.(1) and the maximum of Eq.(1) is the same as in Eq.(2).*

### *4.2. Nonlinear programming format reformulation*

Introduce the following vectors

$$x=\mathrm{col}[x(s,a)],\quad y=\mathrm{col}[y(s,a)].$$

Then the problem (8) can be represented as

$$f(x,y):=\widetilde{W}x+\frac{\gamma}{2}\left[\left(\widetilde{W}x\right)^2-\bar{W}x\right]+\widetilde{W}y+\frac{\gamma}{2}\left[\left(\widetilde{W}y\right)^2-\bar{W}y\right]\to\max_{\substack{y\in Y_{adm}\\ x\in X_{adm}}}$$

where

$$\widetilde{W}=\mathrm{diag}\left(W_{1|1},\dots,W_{1||A|};W_{2|1},\dots,W_{2||A|};\dots;W_{|R||1},\dots,W_{|R||A|}\right)$$
$$\bar{W}=\mathrm{diag}\left(W^2_{1|1},\dots,W^2_{1||A|};W^2_{2|1},\dots,W^2_{2||A|};\dots;W^2_{|R||1},\dots,W^2_{|R||A|}\right)$$
$$\widetilde{W}^2=\bar{W}$$
$$\widehat{W}=\mathrm{diag}\left(W_{1|1},\dots,W_{1||A|};W_{2|1},\dots,W_{2||A|};\dots;W_{|E\backslash R||1},\dots,W_{|E\backslash R||A|}\right)$$
$$W=diag\left(W^2_{1|1},\dots,W^2_{1||A|};W^2_{2|1},\dots,W^2_{2||A|};\dots;W^2_{|E\backslash R||1},\dots,W^2_{|E\backslash R||A|}\right)$$
$$\widehat{W}^2=W$$

The set of constraints $X_{adm}$ and $Y_{adm}$ looks as

$$X_{adm}:=\left\{x\in\mathbb{R}^{|R||A|}\middle|\bar{e}_j^{\mathsf{T}}x-\bar{p}_j^{\mathsf{T}}x=0(j=1,\dots,|R|),\right.$$

$$\left.x_s>0,e^{\mathsf{T}}x=1\right\},s\in R$$

where $e^{\mathsf{T}}:=(1,\dots,1)\in\mathbb{R}^{|R||A|}$ and

$$\bar{e}_j^{\mathsf{T}}:=\left(\underbrace{0,\dots,0}_{1};0,\dots,0;\dots;0,\dots,0;\underbrace{1,\dots,1}_{j};0,\dots,0;\dots;\underbrace{0,\dots,0}_{|R|}\right)\in\mathbb{R}^{|R||A|}$$

$$\bar{p}_j^{\mathsf{T}}:=\left(\underbrace{p_{j|11},\dots,p_{j|1|A|}}_{1};\underbrace{p_{j|21},\dots,p_{j|2|A|}}_{2};\dots;\underbrace{p_{j||R|1},\dots,p_{j||R||A|}}_{|R|}\right)\in\mathbb{R}^{|R||A|},$$

$$j=\overline{1,|R|}$$

and

$$Y_{adm}=\left\{y\in\mathbb{R}^{|E\backslash R||A|}\middle|\bar{e}_j^{\mathsf{T}}y-\bar{p}_j^{\mathsf{T}}y=\xi(s'),\bar{e}_j^{\mathsf{T}}y-\bar{p}_j^{\mathsf{T}}y=0,\right.$$
$$\left.y_s>0,e^{\mathsf{T}}y=1,e^{\mathsf{T}}x-e^{\mathsf{T}}y=0\right\},s\in E\backslash R$$

### 5. Regularised penalty function Markowitz portfolio

Let us consider the Markowitz portfolio model based on a *regularised penalty function* as following

$$\widetilde{\Phi}_{k,\delta}(x,y) := f(x,y) - \frac{k}{2}\left[\sum_{j=1}^{|R|}\left[\left(\bar{e}_j - \bar{p}_j\right)^{\mathsf{T}}x\right]^2 + (e^{\mathsf{T}}x-1)^2 + \sum_{j=1}^{|E\setminus R|}\left[\left(\bar{e}_j - \bar{p}_j\right)^{\mathsf{T}}y - \xi(s')\right]^2 + \sum_{j=1}^{|E\setminus R|}\left(\bar{e}_j - \bar{p}_j\right)^{\mathsf{T}}y + (e^{\mathsf{T}}x - e^{\mathsf{T}}y)^2\right] - \frac{1}{2}\delta(\|x\|^2 + \|y\|^2)$$

where the parameters $k$ and $\delta$ are positive. Here the function $[z]_+$ is defined as $[z]_+ := \max(z,0)$.

Notice also that

$$\arg\max_{\substack{y\in Y_{adm}\\ x\in X_{adm}}} \widetilde{\Phi}_{k,\delta}(x,y) = \arg\max_{\substack{y\in Y_{adm}\\ x\in X_{adm}}} \Phi_{\mu,\delta}(x,y)$$

where $\mu := k^{-1} > 0$ and

$$\Phi_{\mu,\delta}(x,y) := \mu f(x,y) - \frac{1}{2}\left[\sum_{j=1}^{|R|}\left[\left(\bar{e}_j - \bar{p}_j\right)^{\mathsf{T}}x\right]^2 + (e^{\mathsf{T}}x-1)^2 + \sum_{j=1}^{|E\setminus R|}\left[\left(\bar{e}_j - \bar{p}_j\right)^{\mathsf{T}}y - \xi(s')\right]^2 + \sum_{j=1}^{|E\setminus R|}\left(\bar{e}_j - \bar{p}_j\right)^{\mathsf{T}}y + (e^{\mathsf{T}}x - e^{\mathsf{T}}y)^2\right] \quad (3)$$

Let $\Phi_{\mu,\delta}(x,y)$ be the associated Markowitz portfolio function that satisfies Eq. (3). If the penalty parameter $\mu$ as well as the regularising parameter $\delta$ tend to zero in a special way, then we may expect that $x^*(\mu,\delta)$, $y^*(\mu,\delta)$, which are the solutions of the Markowitz portfolio optimisation problem under fixed parameters $\mu,\delta > 0$

$$\Phi_{\mu,\delta}(x,y) \to \max_{\substack{y\in Y_{adm}\\ x\in X_{adm}}}, \quad (4)$$

tend to the set $X^* \times Y^*$ of all solutions of the original Markowitz portfolio optimisation problem (4), that is,

$$\rho\{x^*(\mu,\delta), y^*(\mu,\delta); X^* \otimes Y^*\} \underset{0<\delta,\mu\downarrow 0,\frac{\mu}{\delta}\downarrow 0}{\to} 0$$

where $\rho\{a; X^* \otimes Y^*\}$ is the Hausdorff distance between the point $a = (a_x, a_y)$ and the set $X^* \otimes Y^*$ defined as

$$\rho\{a; X^* \otimes \Upsilon^*\} = \min_{x^*\in X^*, y^*\in Y^*}\left(\|a_x - x^*\|^2 + \left\|a_y - y^*\right\|^2\right)$$

## 6. Portfolio

### *6.1. Mean-Variance Portfolio*

We make the assumption that investors will target the portfolio with the lowest risk under the one-period horizon and anticipate the returns to follow the same probability distribution. We think that there is no inflation or change in interest rates, and that the markets are stable. We assume transaction costs in trading and infinite trading by investors on an arbitrage-free market in order to be more realistic. We suppose that the system consists of fourteen states, the first seven correspond to the ergodic system, and the rest to the transient system. Fixing the initial values for $\gamma_0 = 2.0 \times 10^{-3}$ and $\delta_0 = 31.99$. We have the resulting portfolio is given by

$$\pi(s,a) = \begin{cases} \begin{bmatrix} 0.8151 & 0.1849 \\ 0.3162 & 0.6838 \\ 0.6926 & 0.3074 \\ 0.4363 & 0.5637 \\ 0.2941 & 0.7059 \\ 0.9930 & 0.0070 \\ 0.6592 & 0.3408 \end{bmatrix} & a \in \mathcal{A}, s \in R, \\ \begin{bmatrix} 0.7509 & 0.2491 \\ 0.6884 & 0.3116 \\ 0.1347 & 0.8653 \\ 0.6030 & 0.3970 \\ 0.4587 & 0.5413 \\ 0.3021 & 0.6979 \\ 0.7729 & 0.2271 \end{bmatrix} & a \in \mathcal{A}, s \in E \backslash R \end{cases}$$

To turn a profit is the investor's main goal. An intelligent investor will try to select the portfolio with the lowest risk that meets this objective. To do this, we construct a mean-variance diagram for each potential portfolio of a hazardous asset, with the points denoting the returns $\mathcal{C}$ and risk $\mathcal{V}ar$ (variance). The convergence of the utility is shown in Fig. 1, the variance is shown in Fig. 2, and the convergence of the functional is plotted in Fig. 3. The convergence of the norm is seen in Fig. 4. The convergence of the portfolio policies is seen in Figures 5 and 6.

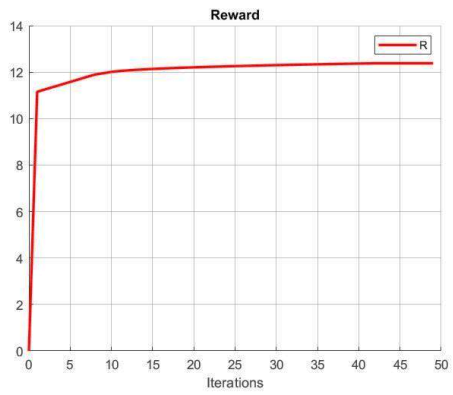



**Figure 1. Utility value of the portfolio**

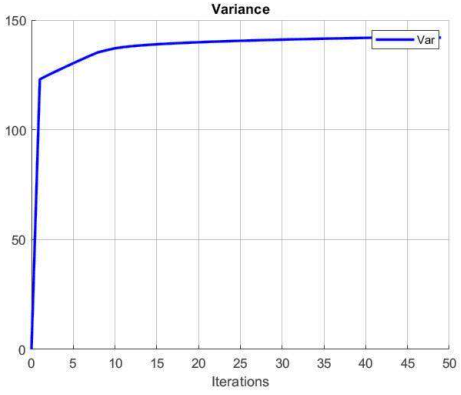
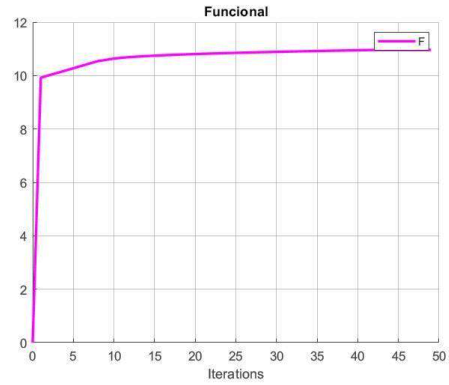



**Figure 2. Variance value of the portfolio**



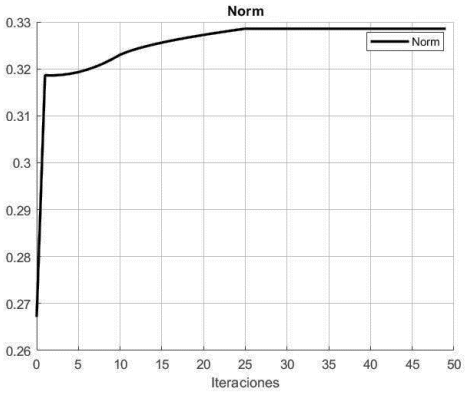
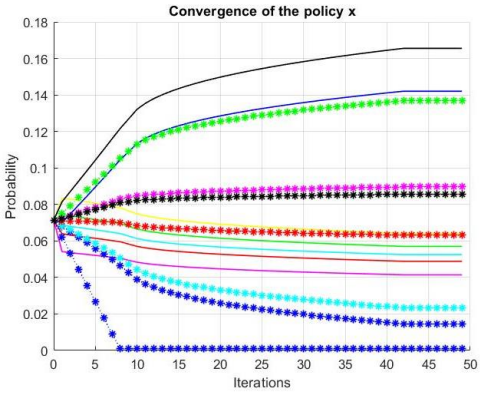




**Figure 3. Functional value of the Portfolio** **Figure 4. Convergence of the norm**



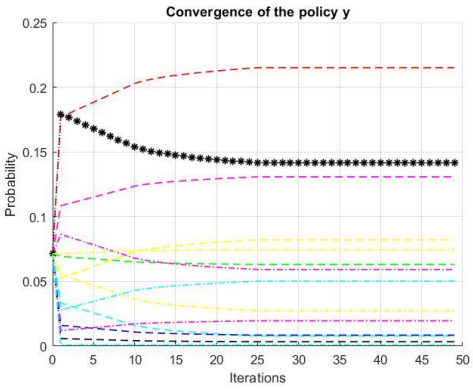




**Figure 5. Convergence of the policy x** **Figure 6. Convergence of the policy y**

## 7. Conclusions

We introduced a novel method for choosing a portfolio. For a collection of finite-number assets with discontinuous ergodic and transient return models, the aim of this work is to provide a novel mean-variance Markowitz portfolio solution. We provide a technique for portfolio selection using a penalty regularised Markov approach. We have demonstrated that one of the most widely used methods for addressing discrete ill-posed problems, Tikhonov's regularisation, is crucial in this situation for ensuring the convergence to a singular portfolio solution. For this system, we offered equivalent penalty regularised optimisation models and optimality specifications. We began by describing the qualities and conditions that permit the identification of a single solution for the penalty regularised portfolio. Finally, we demonstrated that the initial Markowitz portfolio optimisation problem

converges to a singular solution. The utility function also suffers penalties in addition to the restrictions. We calculated a stationary policy for the ergodic and the transient models, respectively. A numerical example exemplifies the novel strategy and proves the effectiveness of the technique. Future difficulties will be the expansion of the method to multivariate and multistage benchmarks and outcomes.

## REFERENCES


[1] **Aboussalah, A.M., Lee, C.G. (2020),** ***Continuous control with stacked deep dynamic recurrent reinforcementlearning for portfolio optimization***. *Expert Systems with Applications*, 140,112891;

[2] **Akian, M., Sulem, A., Taksar, M. (2001)**, ***Dynamic optimization of long-term growth rate for a portfolio with transaction costs and logarithmic utility***. *Mathematical Finance*, 11(2),152–188;

[3] **Almahdi, S., Yang, S.Y.** (**2017**), ***An adaptive portfolio trading system: A risk-return portfolio optimizationusing recurrent reinforcement learning with expected maximum drawdown***. *Expert Systems with Applications*, 87, 267–279;

[4] **Asiain, E., Clempner, J.B., Poznyak, A.S. (2018),** ***A Reinforcement Learning Approach for Solving the Mean Variance Customer Portfolio in Partially Observable Models***. *International Journal on Artificial Intelligence Tools*, 27(08), 1850034;

[5] **Borkar, V., Jain, R. (2010), Risk-constrained markov decision processes**. In *49th IEEE Conference on Decision and Control (CDC)*, 2664–2669, Atlanta, GA, USA;

[6] **Carrasco, M., Noumon, N. (2011),** ***Optimal portfolio selection using regularization***. *Citeseer, Tech. Rep.*

[7] **Clempner, J.B., Poznyak, A.S. (2018),** ***Sparse Mean-Variance Customer Markowitz Portfolio Selection for Markov Chains: A Tikhonov's Regularization Penalty Approach***. *Optimization and Engineering*, 19, 383–417;

[8] **Cvitanic, J., Karatzas, I. (1996),** ***Hedging and portfolio optimization under transaction costs: A martingale approach***. *Mathematical Finance*, 6, 133–166;

[9] **Davis, M., Norman, A. (1990),** ***Portfolio selection with transaction costs*****.** *Mathematics of Operations Research*, 15(4), 676–713;

[10] **Domínguez, F., Clempner, J.B. (2019),** ***Multiperiod mean-variance customer constrained portfolio optimization for finite discrete-time Markov chains***. *Economic Computation and Economic Cybernetics Studies*, 1, 39–56;

[11] **Fastrich, B., Paterlini, S., Winker, P. (2015),** ***Constructing optimal sparse portfolios using regularization methods***. *Computational Management Science*, 12(3), 417–434;

[12] **Filar, J.A., Kallenberg, L.C.M., Lee, H-M. (1989),** ***Variance-penalized markov decision processes***. *Mathematics of Operations Research*, 14(1), 147–161;

[13] **Garcia-Galicia, M., Carsteanu, A.A., Clempner, J.B. (2019),** ***Continuous-time learning method for customer portfolio with time penalization***. *Expert Systems with Applications*, 129, 27–36;

[14] **Garcia-Galicia, M., Carsteanu, A.A., Clempner, J.B. (2019).** ***Continuous-time mean variance portfolio with transaction costs: A proximal approach involving time penalization***. *International Journal of General Systems*, 48, 91–111;

[15] **Liu, H. (2005),** ***Optimal consumption and investment with transaction costs and multiple risky assets***. *Journal of Finance*, 59(1), 289–338;

[16] **Markowitz, H. (1952), Portfolio Selection.** ***Journal of Finance***, 7, 77–91;

[17] **Moody, J., Saffell, M. (2001),** ***Learning to trade via direct reinforcement***. *IEEE transactions onneural Networks*, 12(4), 875–889;

[18] **Mwanakatwe, P.K., Song, L., Hagenimana, E., Wang, X. (2019),** ***Management strategies for a defined contribution pension fund under the hybrid stochastic volatility model***. *Computational and Applied Mathematics*, 38(45);

[19] **Nowak, P., Romaniuk, M. (2018),** ***Valuing catastrophe bonds involving correlation and cir interest rate model***. *Computational and Applied Mathematics*, 37, 365–394;

[20] **Ruszczyński, A. (2010),** ***Risk-averse dynamic programming for markov decision processes***. *Mathematical Programming*, 125, 235–261;

[21] **Sánchez, E.M., Clempner, J.B., Poznyak, A.S. (2015),** ***A Priori-Knowledge/Actor-Critic Reinforcement Learning Architecture for Computing the Mean-variance Customer Portfolio: The Case of Bank Marketing Campaigns***. *Engineering Applications of Artificial Intelligence*, 46, Part A, 82–92;

[22] **Sánchez, E.M., Clempner, J.B., Poznyak, A.S. (2015),** ***Solving the Mean-Variance Customer Portfolio in Markov Chains Using Iterated Quadratic/Lagrange Programming: A Credit-Card Customer-Credit Limits Approach.*** *Expert Systems with Applications*, 42(12), 5315–5327;

[23] **Sobel, M.J. (1994),** ***Mean-variance tradeoffs in an undiscounted mdp***. *Operations Research*, 42(1), 175–183;

[24] **Vazquez, E., Clempner, J.B. (2020),** ***Customer portfolio model driven by continuous-time markov chains: An L2 lagrangian regularization method***. *Economic Computation And Economic Cybernetics Studies And Research*, 54(2), 23–40;

[25] **Ziemba, W., (2012),** ***Calendar anomalies and arbitrage.*** *World Scientific Publishing.*