Expert Systems With Applications 129 (2019) 27–36

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Continuous-time reinforcement learning approach for portfolio management with time penalization

Mauricio García-Galicia[a,b], Alin A. Carsteanu[a,b], Julio B. Clempner[a,b,*]

[a] Escuela Superior de Física y Matemáticas, Instituto Politécnico Nacional, Building 9 U.P. Adolfo Lopez Mateos, Col. San Pedro Zacatenco, Mexico City 07730, Mexico

[b] School of Physics and Mathematics, National Polytechnic Institute, Mexico

ARTICLE INFO

*Article history:*
Received 8 September 2018
Revised 21 March 2019
Accepted 30 March 2019
Available online 1 April 2019

*Keywords:*
Portfolio
Reinforcement learning
Transaction costs
Continuous-time
Markov chains

ABSTRACT

This paper considers the problem of policy optimization in the context of continuous-time Reinforcement Learning (RL), a branch of artificial intelligence, for financial portfolio management purposes. The underlying asset portfolio process is assumed to possess a continuous-time discrete-state Markov chain structure involving the simplex and ergodicity constraints. The goal of the portfolio problem is the redistribution of a fund into different financial assets. One general assumption has to be set, namely that the market is arbitrage-free (no price arbitrage is possible) then the problem of how to obtain the optimal policy is solvable. We provide a RL solution based on an actor/critic architecture in which the market is characterized by a restriction called transaction cost, involving time penalization. The portfolio problem in Markov chains is determined by solving a convex quadratic minimization problem with linear constraints. Any Markov chain is generated by a stochastic transition matrices and the mathematical expectations of the rewards. In particular, we estimate the elements of the transition rate matrices and the mathematical expectations of the rewards. This method learns the optimal strategy in order to make a decision on what portfolio weight to take for a single period. With this strategy, the agent is able to choose the state with maximum utility and select its respective action. The optimal policy computation is solved employing a proximal optimization novel approach, which involves time penalization in the transaction costs and the rewards. We employ the Lagrange multipliers approach to include the restrictions of the market and those that are imposed by the continuous time frame. Moreover, a specific numerical example in baking, that fit into the general framework of portfolio, validates the effectiveness and usefulness of the proposed method.



## 1. Introduction

### 1.1. Brief review

Continuous-time Markov decision processes (CTMDP) are consolidated models of systems that are applicable in important fields such as portfolio theory. A CTMDP is a controllable memoryless model that changes states over time according to transition rates that can be affected by actions chosen by an agent. States and actions establish instantaneous rewards, which are considered over a trajectory in order to determine final utilities. Such rewards depend on the current state and action considering the policy that decides the selection of the next actions and states. The expectation of the reward is determined by an average of the MDP. The goal of the agent is to find a policy that maximizes the expectation of the reward over a distribution of the initial state (Sutton & Barto, 1998). Markov jump processes with finite state space are employed in relevant theoretical and application areas of research. The properties of these models are specified by their generator, denoted by $\Lambda$, also known as rate matrix or intensity matrix (Anderson, 1991; Guo & Hernandez-Lerma, 2009). A remarkable application of Markov jump processes in mathematical finance is in portfolio modeling, where the transitions between different states are represented by a Markov jump process.

Since the seminal proposal presented by Markowitz (1952) for portfolio selection models for a single period, there has been a significant effort dedicated to improve the portfolio theory. Several approaches employing different techniques in distinct

* Corresponding author at: School of Physics and Mathematics, National Polytechnic Institute, Mexico.
*E-mail addresses:* alin@esfm.ipn.mx (A.A. Carsteanu), julio@clempner.name (J.B. Clempner).

https://doi.org/10.1016/j.eswa.2019.03.055
0957-4174/

theoretical directions have been proposed (Dominguez & Clempner, 2019; Iorio, Frasso, D'Ambrosio, & Siciliano, 2018; Petropoulos, Chatzis, Siakoulis, & Vlachogiannakis, 2017; Petropoulos, Chatzis, & Xanthopoulos, 2016; Sánchez, Clempner, & Poznyak, 2015a; 2015b; Takano & Gotoh, 2014). Continuous-time mean-variance portfolio theory is also a research line that has received considerable attention. Zhou and Li (2000) improved the work of Li and Ng (2000) restricted to a class of auxiliary stochastic linear-quadratic problems to study the mean-variance problem in continuous time. This line of research interesting mathematical methods are presented by Costa and Araujo (2008),Huang (2008),Sotomayor and Cadenillas (2009), Wu (2013), and Chen (2015). Leccadito, Lozza, and Russo (2007) suggested a Markovian models in portfolio theory and risk management considering a discrete-time optimal allocation models, where they examine the investor's optimal choices either when the returns are uniquely determined by their mean and variance or when they are modeled by a Markov chain theory. Other relevant improvements to mean-variance portfolio are considered in, among many others, (Lim, 2004; Lim & Zhou, 2002; Xia, 2005). Markowitz's problem with transaction cost was tackle by Dai, Xu, and Zhou (2010), Garcia-Galicia, Carsteanu, and Clempner (2018).

Different techniques have been presented in the literature for agents to develop automated trading using RL techniques, considered within the artificial research area (for a survey see Cumming (2015)). These include Moody and Saffell (2001), Dempster and Leemans (2006), Heaton, Polson, and Witte (2016), Deng, Bao, Kong, Ren, and Dai (2017), and Pendharkar and Cusatis (2018). These RL algorithms output discrete trading signals on an asset. Important solutions have been suggested, emphasizing the importance of Markov chains solutions for solving Markowitz's portfolio in discrete time and continuous time. Zhou and Yin (2003) presented a continuous-time approach of the Markowitz's mean-variance portfolio selection, solving the efficient portfolio and efficient frontier explicitly. In this sense, Yin and Zhou (2004a) suggested a switching solution presented in a finite-state and discrete-time Markov chain framework for portfolio selection, where the relationship between discrete-time approach and their continuous-time model are explicitly presented. Yiu, Liu, Siu, and Ching (2010) solve the portfolio selection problem considering the maximum value-at-risk constraint and the expected return. The model also considered the volatility of the risky asset switch over time, according to a continuous-time Markov chain. Sánchez, Clempner, and Poznyak (2015b) studied a sparse approach to the mean-variance customer portfolio optimization problem for discrete-time Markov chains proposing Tikhonov's regularization technique to solve employing an iterated method. Sánchez et al. (2015a) suggested a RL method based on an actor-critic architecture for computing the mean-variance customer portfolio. (Petropoulos et al., 2017) employed the multiple FOREX time series forecasting approach for modeling portfolio trading studying the efficacy and the feasibility of developing a stacked generalization system, intelligently combining the predictions of diverse machine learning models. Clempner and Poznyak (2018) presented a solution based on a penalty-regularized approach and studied the applicability of the method for computing the mean-variance Markowitz customer portfolio problem. Garcia-Galicia et al. (2018) suggested a continuous-time optimization solution that enables the computation of the portfolio problem based on a proximal approach, which considers time penalization in the transaction costs and the utility.

This paper is concerned with a RL model in continuous-time discrete-state portfolio management with time penalization for computing scenarios in which the transition kernel and the rewards are known a priori. The market is arbitrage-free with transaction costs, and the underlying asset price process is assumed to possess a Markov chain structure. In this approach all investors have homogeneous expectations, they target the portfolio with the lowest volatility and have the same one-period horizon. Under these assumptions we consider an actor/critic reinforcement learning architecture for improving the effectiveness of the learning process for the portfolio problem. In the case of policy gradient, the dimensionality of the state and action spaces represents a significant challenge, for the reason that the complexity of learning scales exponentially with the number of actions and states (Friedman, Hastie, & Tibshirani, 2001). We employ a policy gradient algorithm, using a proximal approach which is computed deriving the gradients of the portfolio functional and learning the state rate transition probabilities for the Markov jump processes on the way. In this approach, the RL solution considers that the market is characterized by a particular restriction called transaction costs. In particular, we employ a novel algorithm involving time penalization in the transaction costs and the rewards (Trejo, Clempner, & Poznyak, 2018). To the best of our knowledge, is the first time that this algorithm is employed for solving the continuous-time discrete state portfolio management problem. We make use of the Lagrange multipliers to incorporate the restrictions of the market, as well as those that are imposed by the Markov continuous-time frame. This algorithm is the key enabler of a novel RL approach which converges to an optimal point of the continuous-time portfolio problem with time penalization.

A relevant practical issue to be considered in portfolio modeling for Markov jump processes is to determine the generator from time-series data in a RL approach. This issue is one of the topics of the present paper. The classical approach to determine the generator of the chain is to maximize the likelihood function associated with these continuous time data. In this sense, there exists an analytic expression for the maximum likelihood estimator (MLE) of $\Lambda$, which involves quantities easily calculable from the data. MLE is restricted to the case of equidistant observation times lags and several approaches can be found in the literature, e.g., (Bladt & Sørensen, 2005; 2009; Crommelin & Vanden-Eijnden, 2006). For a survey on this subject see Metzner, Dittmer, Jahnke, and Schütte (2007), who summarize, compare, and discuss the generator estimation of Markov jump processes in detail.

### 1.2. Main results

The main goal of this paper is to create a market model that enables a formulation of the utility-based option pricing via a portfolio optimization in the continuous-time case. Within the scope of this paper we present an approach that combines and develops new techniques related to continuous-time discrete-state Markovian setup. The main results of this paper are the following.

- Proposes a continuous-time, discrete-state RL algorithm for solving the problem of policy optimization in the context financial portfolio management.
- Considers an actor/critic reinforcement learning architecture.
- Provides a portfolio management solution in which the market is characterized by transaction costs involving time penalization.
- Estimates the elements of the transition rates matrices and the mathematical expectations of the rewards.
- Computes the optimal policy, employing a novel proximal optimization approach, which involves time penalization in the transaction costs and the rewards.
- Employs the Lagrange-multipliers approach to include the restrictions of the market and those imposed by the Markov chain

All issues outlined above are being studied in particular within the CTMDP framework and the usefulness and effectiveness of the method are validated by a numerical example related to banks.

### 1.3. Organization of the work

The outline of this paper is as follows. The next section provides a brief overview of mathematical background needed for understanding the rest of the paper related to continuous-time Markov chains and portfolio. Section 3 presents the estimation model considering the estimation of the elements of the transition matrices and the mathematical expectations of the rewards, as well as, the method for computing the optimal policies of the portfolio problem. Section 4 suggests a RL architecture and the corresponding algorithms for computing the estimation model. Section 5 describes the usefulness and effectiveness of the method in a numerical example. Section 6 concludes the paper with some remarks.

## 2. Preliminaries

### 2.1. Basic properties of continuous-time markov decision process

We will consider the continuous-time Markov jump process given by $\{X(t):\ 0\le t\}$ on the finite state space $S=\{s_1,\ldots,s_N\}$ $(N\in\mathbb{N})$ which plays the role of the state space for a Markov chain of interest (Carrillo, Escobar, Clempner, & Poznyak, 2016; Guo & Hernandez-Lerma, 2009). Assuming time-homogeneity, we will denote the transition matrices of this process by $P(X(\tau+t)=s_j \mid X(\tau)=s_i, \{X(u):0\le u<\tau\}) = P(X(\tau+t)=s_j \mid X(\tau)=s_i)$ for all $t\ge 0$ and $\tau\ge 0$.

We will assume that the process is ergodic, implying that there exists a unique vector $\zeta=(\zeta_1,\ldots,\zeta_n)^\intercal$ with positive entries $\zeta_i>0$ such that $\zeta(0)=P(X(0))$ denote the initial probability distribution for the Markov chain under $P$. Likewise, the vector $\zeta(t)=P(X(t))$ stand for the probability distribution of $X$ at time $t$. It can be easily checked that

$$\zeta(\tau+t)=\zeta(0)P(\tau+t)=\zeta(t)P(\tau)=\zeta(\tau)P(t) \tag{1}$$

and it satisfies that $\zeta^\intercal P(t)=\zeta^\intercal,\ \sum_{i=1}^{N}\zeta_i=1$.

This vector is referred to as the stationary distribution of the Markov jump process.

We now impose an important assumption on the family $P(\dot{)}$, specifically, that this family is right-continuous at time $t=0$; that is, $\lim_{t\downarrow 0}=P(0)$. By virtue of the Chapman–Kolmogorov equation, this implies that

$$\lim_{\tau\to 0}P(\tau+t)=P(t),\ \text{ for all } t\ge 0 \tag{2}$$

and thus

$$\lim_{\tau\to 0}P(X_{\tau+t}=s_j|X_t=s_i)=\delta_{ij},\ \text{ for all } t\ge 0 \tag{3}$$

The right-hand side continuity at time $t=0$ of $P(\dot{)}$ implies the right-hand side differentiability at $t=0$ (Rolski, Schmidli, Schmidt, & Teuggels, 1998). Then, the following limits exist

$$\lambda_{ij}=\lim_{t\to 0+}\frac{p_{ij}(t)-p_{ij}(0)}{t}=\lim_{t\to 0+}\frac{p_{ij}(t)-\delta_{ij}}{t} \tag{4}$$

We have for every $i\neq j$ that $\lambda_{ij}\ge 0$ and $-\sum_{i\neq j}^{N}\lambda_{ij}$. The matrix $\Lambda=[\lambda_{ij}]$ is called the infinitesimal generator matrix for a Markov chain.

Since each entry $\lambda_{ij}$ of the matrix $\Lambda$ can be shown to represent the intensity of transition from the state $i$ to the state $j$, the infinitesimal generator matrix $\Lambda$ is also commonly known as the intensity matrix.

This system can be solved by

$$P(t)=P(0)e^{\Lambda t}=e^{\Lambda t}:=\sum_{n=0}^{\infty}\frac{t^n\Lambda^n}{n!} \tag{5}$$

and at the stationary state, the probability transition matrix is defined as $P^*=\lim_{t\to\infty}P(t)$. We conclude that the generator matrix $\Lambda$ uniquely determines all relevant probabilistic properties of a time-homogeneous Markov chain.

The following statements are equivalent:

- $\zeta^\intercal\Lambda=0$
- $\zeta^\intercal P(t)=\zeta^\intercal;\qquad \forall t\ge 0$

The control model associated with the Continuous-Time Markov Decision Processes (CTMDP) that we are concerned with, is a five-tuple (Guo & Hernandez-Lerma, 2009):

$$(S,A,\mathbb{K},\Lambda,U) \tag{6}$$

where $S$ is a finite state space $\{s_1,\ldots,s_N\}$, $N\in\mathbb{N}$; $A$ is a finite set of actions $\{a_1,\ldots,a_M\}$. For each $s\in S$, $A(s)\subset A$ is the non-empty set of admissible actions at state $s\in S$; $\mathbb{K}=\{(s,a)|s\in S, a\in A(s)\}$ is the class of admissible state-action pairs, which is considered as a subspace of $S\times A$; $\Lambda$ is as defined above; $U:S\times\mathbb{K}\to\mathbb{R}$ is a utility function.

A policy is defined as a sequence $\pi=\{\pi(t),t\ge 0\}$ of stochastic kernels $\pi(t)$ such that: for each time $t\ge 0$, $\pi_{k|i}(t)$ is a probability measure on $A$ such that $\pi_{A(s_{(i)})|i}(t)=1$. Let us denote the collection $\{\pi_{k|i}(t)\}$ by $\Pi$. From now on, we will consider only stationary strategies $\pi_{k|i}(t)=\pi_{k|i}$. For each strategy $\pi_{k|i}$ the associated transition rate matrix is defined as:

$$\Lambda(\pi):=[\lambda_{ij}(\pi)]=\sum_{k=1}^{M}\lambda_{j|ik}\pi_{k|i}$$

such that on a stationary state distribution, for all $\pi_{k|i}$ and $t\ge 0$, we have that $P(\pi)=\lim_{t\to\infty}e^{\Lambda(\pi)t}$, see Guo and Hernandez-Lerma (2009), where $P^*(\pi)$ is a stationary, controlled transition matrix.

The utility function depends on the states and actions is given by the values $W_{ik}$, so that the average reward function $R$ in the stationary regime can be expressed as (Clempner & Poznyak, 2014; Poznyak, Najim, & Gomez-Ramirez, 2000)

$$R(z):=\sum_{ik}^{NK}W_{ik}c_{ik} \tag{7}$$

where $c:=[c_{ik}]_{i=\overline{1,N};k=\overline{1,M}}$ is a matrix with elements

$$c_{ik}=\pi_{k|i}\zeta_i \tag{8}$$

and

$$W_{ik}=\sum_{j}^{N}U_{ijk}p_{j|ik} \tag{9}$$

The expected reward function $R$ depends on the states and actions and it is given by the values of $W_{ik}$. The limiting average reward is as follows

$$\begin{aligned}R_t(\pi)&=\lim_{t\to\infty}t^{-1}\mathrm{E}\left\{\sum_{n=0}^{t}\sum_{i}^{N}\sum_{k}^{M}\left(\sum_{j}^{N}U_{ijk}(n)p_{j|ik}(n)\right)\pi_{k|i}(n)\zeta_i(n)\right\}\\&=\lim_{t\to\infty}t^{-1}\mathrm{E}\left\{\sum_{n=0}^{t}\sum_{i}^{N}\sum_{k}^{M}W_{ik}(n)c_{ik}(n)\right\}=\max_{c}R(c)\end{aligned} \tag{10}$$

The variable $c:=[c_{ik}]_{i=\overline{1,N};k=\overline{1,K}}$ is a matrix with elements

$$c_{ik}=\pi_{k|i}\zeta_i \tag{11}$$

which satisfies the following restrictions: a) belongs to the simplex $\Delta^{NM}$, b) satisfies the continuous-time ergodicity constraints:

$$c \in C_{adm} = \begin{cases} \sum_{ik}^{NM} p_{j|ik} c_{ik} - \sum_{k}^{M} c_{jk} = 0 \\ \text{or} \\ \sum_{ik}^{NM} \lambda_{j|ik} c_{ik} = 0 \end{cases}$$

It follows that

$$\zeta_i^{\iota} = \sum_{k}^{M} c_{ik}^{\iota} \qquad \pi_{k|i}^{\iota} = \frac{c_{ik}^{\iota}}{\sum_{k}^{M} c_{ik}} \tag{12}$$

In the ergodic case $\sum_k^M c_{ik}^{\iota} > 0$. The *objective* is

$$R(c) \to \max_{c \in C_{adm}}$$

### 2.2. Portfolio

Let us consider the filtered probability space $\left(\Omega, \mathcal{F}, \{\mathcal{H}_t\}_{t\geq 0}, \mathbb{P}\right)$ where the $\sigma-$algebra $\mathcal{H}_t$ contains all the $\mathbb{P}$-null sets from $\mathcal{F}$, the filtration $\{\mathcal{H}_t\}_{t\geq 0}$ is right-continuous, that is, $\mathcal{H}_{t+1} := \cap_{s>t}\mathcal{H}_s = \mathcal{H}_t$. In case of the financial market, $\mathbb{P}$ is the physical probability that models the randomness of the stock price. Let $\{(W_t, \mathcal{F}_t)\}_{t\geq 0}$ be the standard $\mathbb{R}^m$-valued Brownian motion where $(W_t)_{t\geq 0}$ is given by $W_t = \left(W_t^1, \ldots, W_t^m\right)^{\intercal}$ and $\mathcal{F}_t \subset \mathcal{F}$. We assume that $\mathcal{F}_t$ is generated by $W_t$. $\mathcal{H}_t \subset \mathcal{F}$ stands for the smallest $\sigma-$algebra containing $\mathcal{F}_t := \sigma(\mathcal{F}_s : 0 \leq s < t)$. Let $T > 0$ denote the investment horizon.

$L^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$ represents the set of all $\mathbb{R}^m$-valued, $\mathcal{F}_t$-progressively measurable stochastic processes $f = \{f_t : 0 \leq t \leq T\}$, with the norm $\|f\|_{L^2_{\mathcal{F}}(0,T;\mathbb{R}^m)} := \left(E\int_0^T |f_t|^2 dt\right)^{1/2} < +\infty$. $L^{\infty}_{\mathcal{F}}(0, T; \mathbb{R}^m)$ denotes the set of all uniformly bounded stochastic processes. And $L^2_{\mathcal{F}_T}(\Omega; \mathbb{R}^m)$ is the notation for the set of all $\mathbb{R}^m$-valued, $\mathcal{F}_t$-measurable random variables $\eta$ such $\|\eta\|_{L^2_{\mathcal{F}_T}(\Omega;\mathbb{R}^m)} := \left(E|\eta|^2\right)^{1/2} < +\infty$.

Consider a market with $m+1$ securities, consisting of a bond and $m$ stocks. The bond price $P_0(t)$ satisfies the (stochastic) ordinary differential equation given by:

$$\left.\begin{aligned} dP_t^0 &= r_t P_t^0 dt, \qquad t \in [0, T] \\ P_t^0 &= p_0 > 0 \end{aligned}\right\} \tag{13}$$

where $r_t$ ($> 0$) is the interest rate (of the bond) is a uniformly bounded, $\mathcal{F}_t$-adapted and scalar-valued process. We assume that the interest rate is constant and equals $r$. The other $m$ assets are stocks whose price processes (stock price dynamics) $P_t^l$, $l = 1, \ldots m$ $(P_t^1, \ldots, P_t^m)$ are modeled by continuous-time processes through the following stochastic differential equation (Yin & Zhang, 1998; Yin & Zhou, 2004b)

$$\left.\begin{aligned} &dP_t^l = P_t^l\left\{\beta^l(S_t)dt + \textstyle\sum_{j=1}^m \sigma^{lj}(S_t)dW_t^j\right\} \\ &t \in [0, T], \ \ l = 1, \ldots m \\ &P_0 = p_l > 0 \end{aligned}\right\} \tag{14}$$

where:

- $S_t$ is a continuous-time and finite state space Markov chain with time step 1, the state space $\left\{s_{(1)}, \ldots, s_{(N)}\right\}$ and the given transition probability matrix $\pi_{j|ik}$ and $\Lambda_{j|ik}$ is a generator for $\pi_{j|ik}$.
- $\beta^l(\cdot) : \mathbb{R} \to \mathbb{R}$ and $\sigma^l(\cdot) : \mathbb{R} \to \mathbb{R}^m$, $l = 1, \ldots m$ are the appreciation rates for the $l$th risky asset and the dispersion (or volatility) rate process of the asset with respect to the $j$th component of the Brownian motion, respectively. Also they are both uniformly bounded, $\mathcal{F}_t$ -adapted and scalar-valued processes.
- $W_t = \left(W_t^1, \ldots, W_t^m\right)^{\intercal}$ is a $m$ -dimensional standard Brownian motion independent of $S_t$.

Thus, the dynamics of $P_t^l$ at time $n$, at which we can observe stock prices, can be written as

$$P_{n+1}^l = P_n^l e^{\beta^l(S_n) - \frac{1}{2}\sum_{j=1}^m \sigma^{lj}(S_n) + \sum_{j=1}^m \sigma^{lj}(S_n)(W_{n+1}^j - W_n^j)}$$

The previous result can be attained from the multi-dimensional Ito's formula.

If we consider an agent (investor) whose total wealth at time $n$ is $u_n$, and the amount of money invested in the $l$th-stock ($l = 1, \ldots m$) is $d_n^l = N_n^l P_n^l$

$$u_n = \sum_{l=0}^{m} d_n^l = \sum_{l=0}^{m} N_n^l P_n^l$$

Now consider an agent with an initial capital endowment $u_0 > 0$ and an investment horizon $[0, T]$. Let $u_n$ denote the total wealth of the agent at time $n \in [0, T]$ ($u_0$ is used indistinctly as a constant value, and as the total wealth at time zero).

**Definition 1.** A portfolio $d$ is said to be an admissible portfolio strategy if $d_n$ is $\mathcal{F}_t-$measurable for $n \in [0, T]$.

We call $(u, d)$ to be an admissible wealth-portfolio pair, and denote the class of admissible portfolio pair, by $\mathcal{A}_{adm}$.

Given a targeted expected return $z$, the agent's objective is to find an admissible portfolio process $d$ such that $E[u_n] = z$ and $Var(u_n)$ is minimized. The problem of finding such a portfolio is called *as the mean-variance portfolio selection* problem, and more rigorously, it is formulated as:

**Definition 2.** The mean-variance portfolio selection problem is a constrained stochastic optimization problem described as

$$Var(u_n) = E[(u_n)^2] - z^2 \underset{mim}{\to}$$

subject to

$$E[u_n] = z, \ \ d \text{ is admissible}, \ \ (u, d) \in \mathcal{A}_{adm}, \ \ u_0$$

Let us denote by $(u^*, d^*)$ as the optimal portfolio for this problem, corresponding to a given $z$, which is called an efficient portfolio.

**Definition 3.** A portfolio $d^*$ is mean-variance efficient if there exists no portfolio parameter $d$ such that

$$E[u(d^*)] \leq E[u(d)] \text{ and } \mathrm{Var}(u^*) \geq \mathrm{Var}(u(d))$$

Furthermore, we call the set of all points $\mathrm{Var}(u^*, z) = E[u^*]$ the efficient frontier.

## 3. Estimation model

### 3.1. Maximum likelihood estimation

Let us consider a controllable CTMDP with a stochastic process given by $\{X(t)\colon 0 \leq t\}$, a finite state space $S = \{s_1, \ldots, s_N\}$, $N \in \mathbb{N}$ and a transition $[p_{(r,i,\tau,j,k)}]_{i,j,k}$, $\tau \geq r$. Let $\lambda_{ijk}$ a transition rate between state $s_i$ and $s_j$ given the action $a_k$ where $\lambda_{ijk} \geq 0$ if $i \neq j$ and $-\sum_{i\neq j}^{N} \lambda_{ijk}$. The sojourn times in the state $s_i$ is given by $-\lambda_{ik}(t)$ and the probability to jump to state $s_j$ is given by $\frac{\lambda_{ijk}(t)}{-\lambda_{iik}(t)}$. We consider that the CTMDP is said to be time homogeneous.

Let us consider the likelihood of observations with the probability to jump from state $s_i$ to $s_j$ at time $\tau_1$, followed by a jump from state $s_j$ to $s_l$ at time $\tau_2$ and so on. The maximum likelihood estimator is defined as follows

$$\mathbb{L}(\Lambda) = e^{-\lambda_{ik}(\tau_2-\tau_1)}\lambda_{ijk}e^{-\lambda_{jk}(\tau_3-\tau_2)}\lambda_{jlk}\ldots. = \prod_{i=1}^{N}\prod_{i\neq j}^{N}\lambda_{ijk}^{\eta_{ijk}(t)}e^{-\lambda_{ik}(Y_{ik}(t))},\ k=1,\ldots,M \tag{15}$$

where $Y_{ik}(t)=\int_0^t 1_{\{X(w)=s_i\}}dw$ is the holding time in the state $s_i$ and action $a_k$ at time $t$ and $\eta_{ijk}(t)$ is the number of times for jump from state $s_i$ to $s_j$ and action $a_k$ in the time interval $[0, t]$. We describe $\eta_{ijk}(t)$ as follows.

We suppose that the CTMDP satisfies the ergodicity condition. We propose a model from experiences that is computed by counting the number $\eta$ of observed experiences defining the following variables recursively as follows:

$$\eta_{ik}(t) = \sum_{n=1}^{t}\chi(X(n)=s_i|A(n)=a_k) = \mathbb{E}\{\chi(X(n))\chi(A(n)\}$$

$$\eta_{ijk}(t) = \sum_{n=1}^{t}\chi(X(n+1)=s_j|X(n)=s_i,A(n)=a_k) = \mathbb{E}\{\chi(X(n+1))\chi(X(n))\chi(A(n)\}$$

where $\chi$ denotes the characteristic function. Now, $\eta_{ik}(t)$ is the number of visits in the observed state $s_j$ for time $n$ applying action $k$ in the estimated process ($n\in\mathbb{N}$), and $\eta_{ijk}(t)$ denotes the total number of times that the process evolves from the observed state $s_i$ to $s_j$ applying action $k$. We have that $\eta_{ik}(t)=\sum_{j=1}^{N}\eta_{ijk}(t)$. The frequency is defined by $f_{ijk}(t)=\frac{\eta_{ijk}(t)}{t}$.

**Remark 1.** When the Markov chain is in discrete time then

$$\hat{p}_{j|ik}(t) = \frac{\eta_{ijk}(t)}{\eta_{ik}(t)} \tag{16}$$

and the frequency $f_{ijk}(t)=\frac{\eta_{ijk}(t)}{t}$ as we expected.

By definition, the maximum likelihood estimator (MLE) maximizes the likelihood function given in Eq. (15). The monotonicity of the log-likelihood of Eq. (15) is given by

$$\log\mathbb{L}(\Lambda) = \sum_{i=1}^{N}\sum_{i\neq j}^{N}\log(\lambda_{ijk})\eta_{ijk}(t) - \sum_{i=1}^{N}\sum_{i\neq j}^{N}\lambda_{ijk}Y_{ik}(t)$$

for $k=1,\ldots,M$. Then, the maximum likelihood estimator for the elements of an infinitesimal generator matrix is explicitly given by the null of the partial derivatives of $\log\mathbb{L}(\Lambda)$ with respect to $\lambda_{ijk}$ and the Hessian matrix of $\log\mathbb{L}(\Lambda)$ evaluated at $\hat{\Lambda}$ is negative definite as

$$\frac{\partial\log\mathbb{L}(\hat{\Lambda})}{\partial\lambda_{ijk}} = 0 \Leftrightarrow \hat{\lambda}_{ijk}(t) = \frac{\eta_{ijk}(t)}{Y_{ik}(t)},\ i\neq j \tag{17}$$

and $\frac{\partial\log\mathbb{L}(\hat{\Lambda})}{\partial\lambda_{ijk}\partial\lambda_{hlk}} = -\frac{\eta_{ijk}(t)}{\hat{\lambda}^2_{ijk}(t)}\chi((s_i,s_j)=(s_h,s_l))$ where

$$\hat{\lambda}_{jlk}(t) = \begin{cases}\frac{\eta_{ijk}(t)}{Y_{ik}(t)}, & if\ i\neq j\\ -\sum_{i\neq j}^{N}\hat{\lambda}_{ijk}(t), & if\ i=j\end{cases} \tag{18}$$

for $Y_{ik}(t)>0$ and $k=1,\ldots,M$.

The interpretation is the following. $\eta_{ijk}(t)$ counts the number of jumps (transitions) from state $s_i$ to $s_j$ applying action $a_k$ during the time interval of the observation. As well as, $Y_{ik}(t)$ is the time spent before leaving state $s_i$ applying action $a_k$. Then, the maximum likelihood estimator of all transition matrices can be obtained applying the exponential matrix generator over $\hat{\Lambda}$

$$\hat{p}_{j|ik}(t) = e^{\hat{\Lambda}t}\ \text{for}\ t>0 \tag{19}$$

**Remark 2.** If the process has never been in state $s_i$, then the information of $\lambda_{ijk}$ is empty and the maximum likelihood estimator of $\lambda_{ijk}$ does not exist.

### 3.2. Utility model

Same as for the utility model, we keep a running average of the utility observed upon taking each action in each state as follows

$$\hat{U}_{ijk}(t) = \frac{U_{ijk}(t)}{\eta_{ijk}(t)} \tag{20}$$

given

$$\hat{U}_{ijk}(t) = \sum_{n=1}^{t}\nu_U(n)\chi\left(X(n+1)=s_j|X(n)=s_i,A(n)=a_k\right)$$

such that $\nu_U := U_{ijk}+(\bar{u})\varrho_U$, for $\bar{u}\leq U_{ijk}$ and $\varrho_U=\text{rand}([-1,1])$ where $U_{ijk}(t)$ is incremented/decremented by $\bar{u}$ multiplied by a $\varrho_U$, $-1\leq\varrho_U\leq 1$.

### 3.3. Policy model

The *portfolio optimization* problem involves a model-user's tolerance for risk and it is described as follows:

$$\Phi(c) := U(c) - \frac{\alpha}{2}\text{Var}(c) \to \max_{c\in C_{adm}} \tag{21}$$

and $\alpha$ is *the risk-aversion* parameter. Employing the Lagrange principle, the mean-variance Markowitz portfolio can be written in the following manner:

$$\begin{aligned}&\mathbb{L}_{\xi_n,\delta_n}(c,\mu_0,\mu_{N+1},\mu_{N+2}|c_n) :=\\ &-\frac{\delta}{2}\tau(c_{ik})\|c_{ik}-c_{ik}(n)\|^2 + \frac{1}{\tau(c_{ik})}\left\{\left[\sum_{i=1}^{N}\sum_{k=1}^{M}W_{ik}c_{ik}\right.\right.\\ &\left.+\frac{\kappa}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}W_{ik}c_{ik}\sum_{\hat{i}=1}^{N}\sum_{\hat{k}=1}^{M}W_{\hat{i}\hat{k}}c_{\hat{i}\hat{k}} - \frac{\kappa}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}W_{ik}^2c_{i|k}\right]\\ &-\sum_{j=1}^{N}\mu_{0,j}\left[\left(\sum_{i=1}^{N}\sum_{k=1}^{M}\pi_{j|ik}c_{ik}-\sum_{k=1}^{M}c_{jk}\right)-b_{eq,j}\right]\\ &-\mu_{N+1}\left(\sum_{i=1}^{N}\sum_{k=1}^{M}c_{ik}-b_{eq,N+1}\right)\\ &\left.-\sum_{j=1}^{N}\mu_{N+2,j}\left[\left(\sum_{i=1}^{N}\sum_{k=1}^{M}\lambda_{j|ik}c_{ik}\right)-b_{eq,N+1+j}\right]\right\}\end{aligned} \tag{22}$$

The Eq. (22) involves the time penalization in the trajectory and the time penalization of the utility considering the restriction related to the simplex and ergodicity. Following Trejo et al. (2018), let us take

$$\begin{aligned}\|(c-c^*)\|^2_{\Lambda=diag(\Lambda_1,\ldots,\Lambda_M)} &= \sum_{k=1}^{M}\left\|\left(c^{(k)}-c^{(k)*}\right)\right\|^2\\ &= \sum_{k=1}^{M}\left(c^{(k)}-c^{(k)*}\right)^{\intercal}\Lambda_k\left(c^{(k)}-c^{(k)*}\right)\end{aligned}$$

where

$c=\left(c^1,\ldots,c^M\right)^{\intercal}\in\mathbb{R}^{NM}$, $c^{(k)}=(c_{1k},\ldots,c_{Nk})^{\intercal}\in\mathbb{R}^N$,

$k=1,M$ and

$\Lambda_k := \frac{1}{2}\left[\tilde{\Lambda}_k+\tilde{\Lambda}_k^{\intercal}\right]$, $\tilde{\Lambda}_k := \left[\tau_{j|ik}\right]$, $\tilde{\Lambda}_k\in\mathbb{R}^{NN}$

$$\tau_{j|ik} := \begin{cases}\frac{1}{\left|\sum_{i\neq j}\lambda_{j|ik}\right|}, & if\ i=j\\ \frac{1}{\lambda_{j|ik}}, & if\ i\neq j\end{cases}$$

Then, we have that

$$c^* = \arg\min_{c\in C_{adm}} \left\{\tfrac{\delta_n}{2}\|(c-c^*)\|^2_{\Lambda=diag(\Lambda_1,\ldots,\Lambda_M)} + \gamma_n(\Phi(c)-\Phi(c_n))\right\} \tag{23}$$

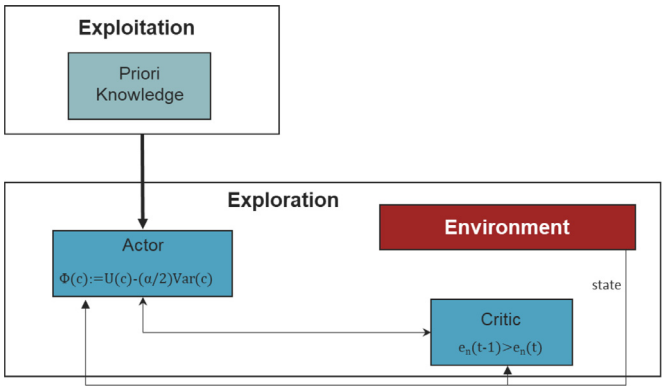


**Fig. 1.** RL architecture.

$$\mu^* = \arg\min_{\mu}\left\{\tfrac{\delta_n}{2}\|(\mu-\mu^*)\|^2 + \gamma_n(\Phi(c)-\Phi(c_n))\right\} \tag{24}$$

## 4. RL architecture

We consider a actor/critic RL architecture (Asiain, Clempner, & Poznyak, 2018) for improving the effectiveness of learning process for the portfolio problem (see Fig. 1). An advantage presented in this approach is that the actor/critic architectures split up the data structures represented by the "actor" and employed for the control policy of the value function governed by the "critic". In this framework, the selection of an action requires minimal computation. The exploitation uses a fast-tracked learning algorithm, which employs a fix strategy and a priori knowledge.

A straightforward method to implement the RL architecture would proceed as follows:

1. Choose an arbitrary generator $\Lambda$ and a time $t$.
2. Compute the corresponding transition matrix $p_{j|ik} = e^{\Lambda t}$.
3. Produce a time series $Y_{ik}(t)$ by sampling from $\Lambda$.
4. Compute $\hat{p}_{j|ik}(t)$ given $\hat{\Lambda}_{j|ik}(t)$
5. Compute $\hat{U}_{j|ik}(t)$
6. Compute and estimate errors.

In this process, we establish $t=0$ to be the initial time and $s(0)=s_i$ be the initial state. We also consider the sequence of discrete states that a CTMDP visits by letting $\lambda_{j|ik}$ be the matrix of the transition rates. By taking $p_{j|ik} = e^{\Lambda t}$ for $t>0$ we obtain a discrete-time transition matrix. In terms of computing $\hat{\Lambda}$, we fix an initial estimation error $\epsilon_0 > 0$. In every iteration of the process for computing $\hat{\Lambda}$, we first establish the optimal portfolio policy $\pi_{k|i}$ according to the estimation rule given in Eq. (23). Then, randomly we select an action $a(t)=a_k$ from $\pi_{k|i}$ for fixed state $s_i$. We think about CTMDPs in terms of jump chains and holding times, the jump to the next state is according to the transition matrix $p_{j|ik}$. So, we get the next state $s_j$ randomly from $\hat{p}_{j|ik}(t)$ for fixed state $s_i$ and action $a_k$. Next, the values of $\eta_{ijk}(t)$ are being updated. In CTMDP the jumps between states take place at random times. Then, we compute $Y_{ik}(t) = -\frac{\ln(\varrho)}{\lambda_{i|ik}}$, considering $\varrho$ a random value. Subsequently, we compute $\hat{\lambda}_{(j|ik)}(t)$ and update $\hat{\Lambda}_{j|ik}(t)$ according to the update rule given in Eq. (18). Then, we compute $\hat{p}_{j|ik}(t)$ according to $e^{\hat{\Lambda}(t)}$ and $\hat{U}_{j|ik}(t)$ using the update rule given in Eq. (20). Finally, we calculate the mean square error $\epsilon(t)$ by using $\epsilon(t) = \sum_{k=1}^{M} tr\big((\hat{p}_k(t-1)-\hat{p}_k(t))^{\intercal}(\hat{p}_k(t-1)-\hat{p}_k(t))\big)$. The process continues while $\epsilon > \epsilon_0$. At the end, we obtain $\hat{\Lambda}_{j|ik}$ and $\hat{U}_{j|ik}$ for the portfolio management problem.

Following the RL architecture given in Fig. 1 the computational procedure for the estimation of the transition rate matrix $\hat{\Lambda}_{j|ik}$ and the utility $\hat{U}_{j|ik}$ is described in the following algorithms.

**Algorithm 1** Priori knowledge.

Let $t=0$ and so $s(0)=s_i$ be the initial state.
Let $\lambda_{j|ik}$ be the matrix of the transition rates
Choose a generator $\Lambda$ and a time lag $t$ and compute the corresponding transition matrix $p_{j|ik} = e^{\Lambda t}$, $t>0$
Let $\pi_{k|i}$ be some fixed policy of the CTMDP
Choose randomly an action $a(t)=a_k$ from $\pi_{k|i}$,
Get next state $s_j$ from $\hat{p}_{j|ik}(t)$
Update the values of $\eta_{ijk}(t)$
Let $Y_{ik}(t) = -\frac{\ln(\varrho)}{\lambda_{i|ik}}$ where $\varrho$ =random
Based on the update rule given in Eq. (17) compute $\hat{\lambda}_{(j|ik)}(t)$
Update $\hat{\Lambda}_{j|ik}(t)$ according to the update rule given in Eq. (18)
Based on $\hat{\Lambda}_{j|ik}(t)$ compute $\hat{p}_{j|ik}(t) = e^{\hat{\Lambda}(t)}$
Compute $\hat{U}_{j|ik}(t)$ using the update rule given in Eq. (20)
Set $s_i = s_j$ and $t \leftarrow t+1$

**Algorithm 2** Actor/Critic.

Let $t=0$ and so $s(0)=s_i$ be the initial state.
Let $\lambda_{j|ik}$ be the matrix of the transition rates
Choose a generator $\Lambda$ and a time lag $t$ and compute the corresponding transition matrix $p_{j|ik} = e^{\Lambda t}$, $t>0$
Let $\epsilon_0 > 0$ be the initial estimation error
**while** $\epsilon_p > \epsilon_0$ and $\epsilon_U > \epsilon_0$ **do**
Based on Eq. (23) compute the optimal portfolio $\pi^*_{k|i}$
Based on Eq. (24) compute the Lagrange multipliers $\mu$.
Choose randomly an action $a(t)=a_k$ from $\pi_{k|i}$,
Get next state $s_j$ from $\hat{p}_{j|ik}(t)$
Update the values of $\eta_{ijk}(t)$
Let $Y_{ik}(t) = -\frac{\ln(\varrho)}{\lambda_{i|ik}}$ where $\varrho$ =random
Compute $\hat{\lambda}_{(j|ik)}(t)$ based on the rule given in Eq. (17)
Update $\hat{\Lambda}_{j|ik}(t)$ according to the update rule given in Eq. (18)
Based on $\hat{\Lambda}_{j|ik}(t)$ compute $\hat{p}_{j|ik}(t) = e^{\hat{\Lambda}(t)}$
Compute $\hat{U}_{j|ik}(t)$ using the update rule given in Eq. (20)
For both methods, compute and compare the errors
$\epsilon_p(t) = \sum_{k=1}^{M} tr\big((\hat{p}_k(t-1)-\hat{p}_k(t))^{\intercal}(\hat{p}_k(t-1)-\hat{p}_k(t))\big)$
$\epsilon_U(t) = \sum_{k=1}^{M} tr\big((\hat{U}_k(t-1)-\hat{U}_k(t))^{\intercal}(\hat{U}_k(t-1)-\hat{U}_k(t))\big)$
Set $s_i = s_j$ and $t \leftarrow t+1$
**end**

## 5. Numerical example

Portfolio management is the decision-making process for continuously reallocating an amount of funding into a number of different financial investment products, with the goal to maximize the reward while restraining the risk. The investor's fundamental goal is to optimize the profit. We employ the RL method proposed to optimize a portfolio of securities considering transition cost. Portfolio profits are determined by sequences of interdependent decisions. Optimal trading decisions affected by transaction costs require knowledge of the current system state. The proposed RL approach suggests an effective framework in continuous-time portfolio management with time penalization for computing scenarios in which the transition kernel and the rewards are known a priori when state-dependent transaction costs are included.

The mean-variance portfolio optimization problem is given by

$$\Phi(c) = E(c) - \frac{\xi}{2}Var(c) \rightarrow \max_{c\in C_{adm}}.$$

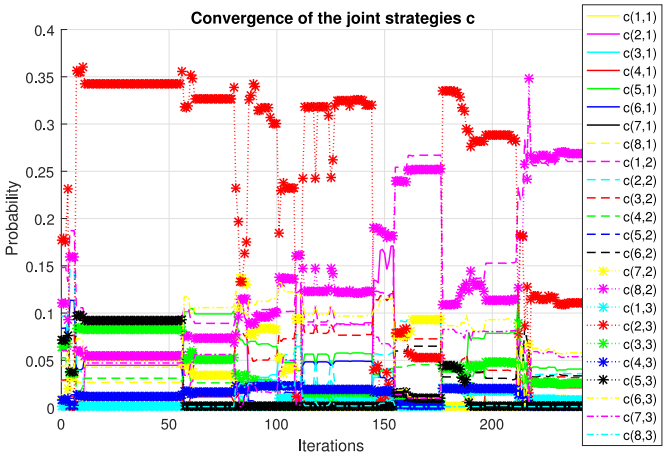


**Fig. 2.** Convergence of the portfolio strategies c with time penalization.


**Fig. 3.** Reward value with time penalization.

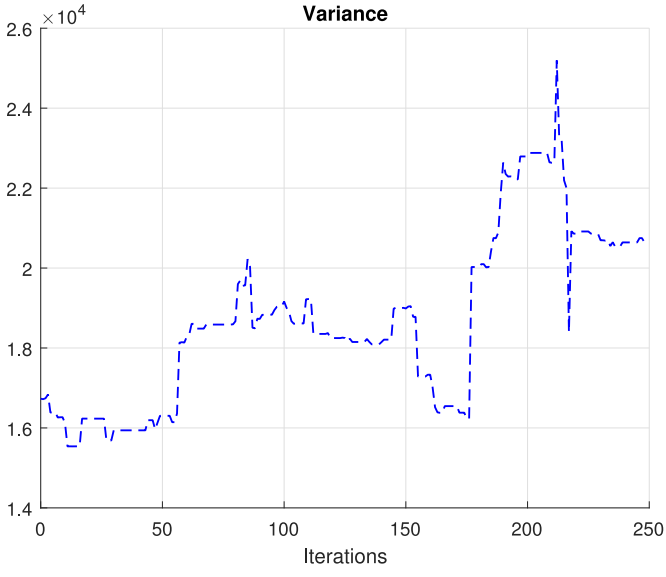


**Fig. 4.** Variance value with time penalization.

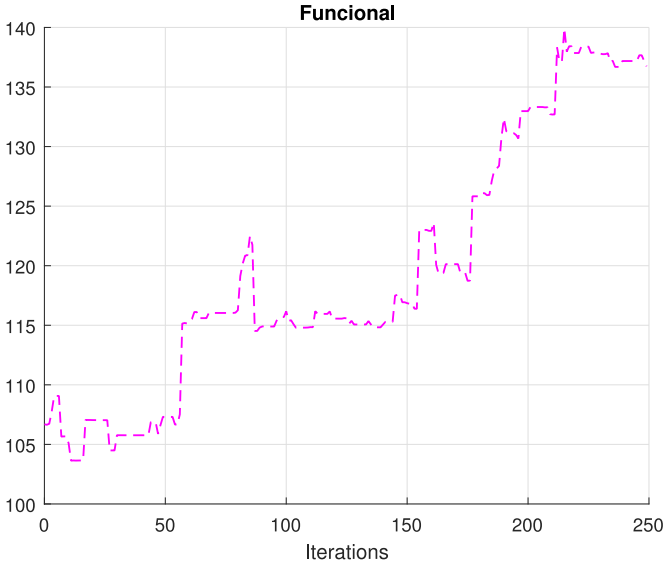


**Fig. 5.** Functional value with time penalization.

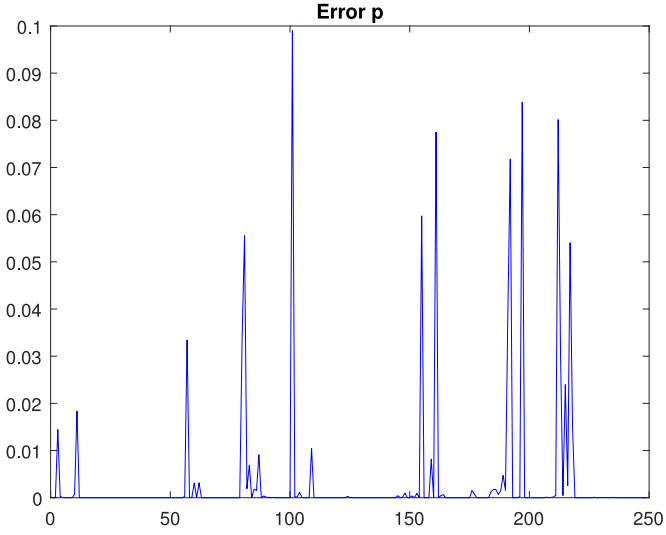


**Fig. 6.** Mean square error of the transition matrix *p*.

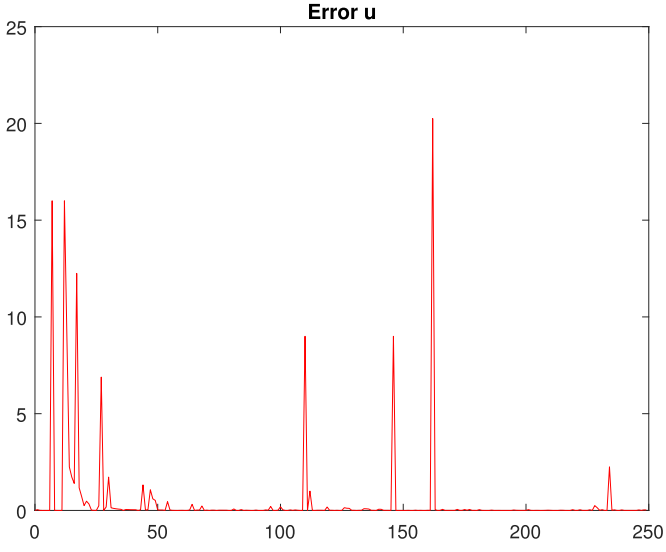


**Fig. 7.** Mean square error of the utility matrix *u*.

The model considers that the number of states is $N = 8$ and the number of actions is $M = 3$. It is employed a continuous-time and finite state space Markov chain with a state space $S = \{s_1, \ldots, s_N\}$ and the transition rate matrix $\Lambda_{j|ik}$ (a generator of the transition matrix $p_{j|ik}$). Then, applying the method presented by Yin and Zhang (1998) it is obtained that $\Lambda_{j|ik} = \frac{1}{\varepsilon}\tilde{\Lambda}_{j|ik} + \hat{\Lambda}_{j|ik}$, $\varepsilon > 0$ where $\tilde{\Lambda} = diag(\tilde{\Lambda}^1, \ldots, \tilde{\Lambda}^N)$, i.e. the transition rates $\lambda_{j|ik} \geq 0$ for $i \neq j$ and $\sum_j^N \lambda_{j|ik} = 0$.

Fixing $\gamma = 5 \times 10^{-4}$, $\kappa = 0.001$ and $\delta = 0.5$ we have that the resulting optimal portfolio in Fig. 2. Using Eq. (12) we compute the strategies $\pi_{i|k}^{*1}$ involving time penalization generated by the RL algorithm as follows:

$$\pi_{i|k}^{*1} = \begin{bmatrix} 0.0037 & 0.9654 & 0.0308 \\ 0.0068 & 0.2426 & 0.7506 \\ 0.0171 & 0.5413 & 0.4416 \\ 0.4687 & 0.5090 & 0.0224 \\ 0.7727 & 0.2083 & 0.0190 \\ 0.0108 & 0.3626 & 0.6266 \\ 0.0160 & 0.1251 & 0.8589 \\ 0.0060 & 0.9903 & 0.0037 \end{bmatrix}$$

and the distribution vector is given by

$$\zeta_i = \begin{bmatrix} 0.2696 & 0.1477 & 0.0584 & 0.0447 & 0.0527 & 0.0924 & 0.0626 & 0.2717 \end{bmatrix}$$

The matrices $\tilde{\Lambda}_{j|ik}$ and $\hat{\Lambda}_{j|ik}$ are generators of the estimated stationary transition probabilities $\hat{p}_{j|ik}$ given by

$$\hat{p}_{j|i1} = \begin{bmatrix} 0.9730 & 0.0054 & 0.0053 & 0.0021 & 0.0031 & 0.0075 & 0.0024 & 0.0013 \\ 0.0245 & 0.8480 & 0.0577 & 0.0018 & 0.0508 & 0.0042 & 0.0070 & 0.0060 \\ 0.1025 & 0.0169 & 0.7124 & 0.0439 & 0.0077 & 0.0052 & 0.0027 & 0.1086 \\ 0.2117 & 0.0168 & 0.0122 & 0.7003 & 0.0069 & 0.0034 & 0.0297 & 0.0190 \\ 0.0129 & 0.1283 & 0.0482 & 0.0033 & 0.7939 & 0.0055 & 0.0025 & 0.0055 \\ 0.0115 & 0.0441 & 0.0202 & 0.0050 & 0.0104 & 0.8144 & 0.0751 & 0.0193 \\ 0.0099 & 0.0095 & 0.0055 & 0.0076 & 0.0311 & 0.0081 & 0.9007 & 0.0275 \\ 0.0344 & 0.0109 & 0.0562 & 0.1928 & 0.0027 & 0.0162 & 0.0170 & 0.6699 \end{bmatrix}$$

$$\hat{p}_{j|i2} = \begin{bmatrix} 0.8682 & 0.0296 & 0.0138 & 0.0062 & 0.0046 & 0.0186 & 0.0122 & 0.0468 \\ 0.0227 & 0.6163 & 0.0495 & 0.0049 & 0.0163 & 0.0194 & 0.0074 & 0.2635 \\ 0.0278 & 0.0018 & 0.8415 & 0.0047 & 0.0006 & 0.1094 & 0.0108 & 0.0034 \\ 0.0197 & 0.0255 & 0.0021 & 0.7567 & 0.0311 & 0.0230 & 0.0172 & 0.1247 \\ 0.0660 & 0.0184 & 0.0019 & 0.0092 & 0.0348 & 0.0022 & 0.0020 & 0.8655 \\ 0.1740 & 0.0144 & 0.0044 & 0.0212 & 0.0030 & 0.7491 & 0.0129 & 0.0209 \\ 0.0039 & 0.0017 & 0.0152 & 0.0099 & 0.0041 & 0.0062 & 0.9052 & 0.0540 \\ 0.0798 & 0.0176 & 0.0020 & 0.0104 & 0.0294 & 0.0016 & 0.0010 & 0.8582 \end{bmatrix}$$

$$\hat{p}_{j|i3} = \begin{bmatrix} 0.5295 & 0.0186 & 0.0140 & 0.0842 & 0.0215 & 0.1021 & 0.2144 & 0.0157 \\ 0.0185 & 0.9107 & 0.0163 & 0.0034 & 0.0089 & 0.0177 & 0.0197 & 0.0047 \\ 0.0055 & 0.0276 & 0.7123 & 0.0139 & 0.1946 & 0.0349 & 0.0085 & 0.0027 \\ 0.0038 & 0.0285 & 0.0066 & 0.7139 & 0.2029 & 0.0364 & 0.0042 & 0.0037 \\ 0.0111 & 0.1882 & 0.0090 & 0.0085 & 0.7463 & 0.0123 & 0.0204 & 0.0042 \\ 0.0090 & 0.0477 & 0.0305 & 0.0678 & 0.0272 & 0.8016 & 0.0094 & 0.0068 \\ 0.0030 & 0.0086 & 0.0069 & 0.0109 & 0.0084 & 0.1073 & 0.8205 & 0.0345 \\ 0.0074 & 0.0666 & 0.0044 & 0.0124 & 0.0070 & 0.1705 & 0.0052 & 0.7265 \end{bmatrix}$$

The resulting estimated utility matrices are given by

$$\hat{u}_{ij1} = \begin{bmatrix} 8.2514 & 1.4866 & 3.3379 & 7.3472 & 1.3722 & 5.7335 & 0.5776 & 8.2549 \\ 1.2967 & 4.0000 & 2.2846 & 8.6927 & 5.7512 & 9.0436 & 2.2471 & 14.2552 \\ 1.7854 & 3.8954 & 1.1250 & 12.7116 & 10.2917 & 1.4350 & 0.4215 & 2.8647 \\ 7.6241 & 6.4732 & 2.3321 & 1.5000 & 0.7620 & 1.7819 & 4.9330 & 3.2113 \\ 2.5924 & 4.3193 & 3.0355 & 3.0552 & 9.0000 & 3.3276 & 0.9278 & 11.0789 \\ 1.9868 & 6.3119 & 0.7806 & 7.0341 & 7.5571 & 5.0000 & 15.0613 & 1.0241 \\ 6.6504 & 1.1753 & 0.2436 & 1.6831 & 2.1933 & 0.8036 & 1.0000 & 1.5452 \\ 10.0213 & 1.2081 & 1.4997 & 2.6649 & 0.5341 & 0.4247 & 9.2025 & 9.0000 \end{bmatrix}$$

$$\hat{u}_{ij2} = \begin{bmatrix} 10.2166 & 7.2494 & 1.2277 & 16.9037 & 1.4017 & 17.4177 & 2.6716 & 2.2014 \\ 0.3132 & 1.0000 & 10.8393 & 12.8445 & 4.9220 & 6.1505 & 1.4293 & 1.0374 \\ 7.5757 & 8.9688 & 9.0000 & 1.0159 & 9.3265 & 7.1240 & 6.3279 & 15.3052 \\ 2.6502 & 5.7062 & 16.6307 & 9.0000 & 4.5502 & 7.0389 & 4.1775 & 1.8941 \\ 13.6171 & 2.5847 & 5.5158 & 2.1866 & 4.0000 & 8.7667 & 1.4226 & 0.7144 \\ 8.1213 & 0.8705 & 17.0950 & 5.3906 & 25.4750 & 4.0000 & 1.2252 & 2.4823 \\ 2.0559 & 7.9203 & 4.1161 & 16.2815 & 11.5381 & 2.6876 & 4.5000 & 2.6353 \\ 5.6226 & 5.2069 & 9.7505 & 3.2675 & 9.6576 & 4.8126 & 1.5996 & 7.0000 \end{bmatrix}$$

$$\hat{u}_{ij3} = \begin{bmatrix} 0.2363 & 4.8282 & 2.1204 & 0.3135 & 3.7217 & 0.5656 & 9.8648 & 1.4295 \\ 12.5073 & 10.0000 & 3.2152 & 0.5134 & 11.6724 & 3.3801 & 3.3482 & 2.0895 \\ 1.1511 & 1.2170 & 10.0000 & 17.1223 & 6.4109 & 4.6978 & 4.7885 & 0.2915 \\ 3.8932 & 3.5008 & 1.3221 & 9.0000 & 12.4898 & 1.3630 & 2.9501 & 1.5928 \\ 4.0219 & 1.0274 & 1.2404 & 0.5726 & 6.0000 & 0.9190 & 4.7934 & 4.0488 \\ 3.0159 & 3.4095 & 2.0965 & 6.5712 & 6.2574 & 10.0000 & 5.3794 & 3.0873 \\ 2.1889 & 1.6067 & 9.1598 & 10.3542 & 2.3929 & 2.6519 & 5.0000 & 2.1398 \\ 0.8823 & 3.6541 & 1.5199 & 9.7898 & 10.6048 & 11.3669 & 9.2735 & 3.0000 \end{bmatrix}$$

The rational investor wants to earn a certain return and tries to identify a portfolio of minimal risk which satisfies this goal. Following this purpose, we plot the portfolio and the risky asset in a variance diagram. Fig. 3 shows the convergence of the expected returns and Fig. 4 shows the convergence of the variance of the portfolio. Fig. 5 shows the convergence of the functional $\Phi(c) = E(c) - \frac{\xi}{2}Var(c)$. This equation shows that the investor is directly rewarded for taking risk. The mean square error of the transition matrix and the utilities are shown in Figs. 6 and 7.

## 6. Conclusions and future work

This paper proposes a solution to the portfolio problem in which the discrete state of the system evolves according to a continuous-time process, characterized by probability transition rate matrices and rewards matrices. We suppose that the market is arbitrage-free with transaction costs, and the underlying asset price process is assumed to possess a Markov-chain structure. In addition, all investors have homogeneous expectations, they target the portfolio with the lowest volatility and have the same one-period horizon. We enhance the continuous-time Markov model for the portfolio problem considering transaction costs. We develop a RL algorithm for the estimation of probability transition matrix and the rewards matrices from observed data. We consider one of the fundamental problems in the estimation of transition matrices, which is that transition probabilities usually correspond to a long-term frame and, in practice, a short-time frame is needed, for which the probabilities are lower than those in the initial transition matrices. Sometimes, transition matrices are extremely sensitive to small shifts in probabilities, and accurate and consistent estimation is essential in the computation. We investigate and include relevant aspects in the analysis, to provide rewards-effectiveness estimates. For computing optimal policy we employ a proximal optimization approach, which involves time penalization in the transaction costs and the rewards. The usefulness and effectiveness of all points outlined above has been studied, in particular within an application example related to banks.

## References


Anderson, W. J. (1991). *Continuous-time Markov chains*. New York: Springer.

Asiain, E., Clempner, J. B., & Poznyak, A. S. (2018). Controller exploitation-exploration: A reinforcement learning architecture. *Soft Computing*. To be published.

Bladt, M., & Sørensen, M. (2005). Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67*(7), 395–410.

Bladt, M., & Sørensen, M. (2009). Efficient estimation of transition rates between credit ratings from observations at discrete time points. *Quantitative Finance, 9*(2), 147–160.

Carrillo, L., Escobar, J., Clempner, J. B., & Poznyak, A. S. (2016). Solving optimization problems in chemical reactions using continuous-time Markov chains. *Journal of Mathematical Chemistry, 54*, 1233–1254.

Chen, W. (2015). Artificial bee colony algorithm for constrained possibilistic portfolio optimization problem. *Physica A, 429*, 125–139.

Clempner, J. B., & Poznyak, A. S. (2014). Simple computing of the customer lifetime value: A fixed local-optimal policy approach. *Journal of Systems Scienceand Systems Engineering, 23*(4), 439–459.

Clempner, J. B., & Poznyak, A. S. (2018). Sparse mean-variance customer Markowitz portfolio selection for Markov chains: A tikhonovs regularization penalty approach. *Optimization and Engineering, 19*(2), 383–417.

Costa, O., & Araujo, M. (2008). A generalized multi-period portfolio optimization with Markov switching parameters. *Automatica, 44*(10), 2487–2497.

Crommelin, D., & Vanden-Eijnden, E. (2006). Fitting timeseries by continuous-time Markov chains: a quadratic programming approach. *Journal of Computational Physics, 217*, 782–805.

Dai, M., Xu, Z., & Zhou, X. (2010). Continuous-time Markowitzs model with transaction costs. *SIAM Journal on Financial Mathematics, 1*, 96–125.

Dempster, M., & Leemans, V. (2006). An automated fx trading system using adaptive reinforcement learning. *Expert Systems with Applications, 30*(3), 543–552.

Deng, Y., Bao, F., Kong, Y., Ren, Z., & Dai, Q. (2017). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems, 28*(3), 653–664.

Dominguez, F., & Clempner, J. B. (2019). Multiperiod mean-variance customer constrained portfolio optimization for finite discrete-time Markov chains. *Economic Computation And Economic Cybernetics Studies And Research, 1*(53), 39–56.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer series in statistics.

Garcia-Galicia, M., Carsteanu, A., & Clempner, J. B. (2018). Continuous-time mean variance portfolio with transaction costs: A proximal approach involving time penalization. *International Journal of General Systems*. To be published.

Guo, X., & Hernandez-Lerma, O. (2009). *Continuous-time Markov decision processes*. Berlin Heidelberg: Springer-Verlag.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2016). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*. URL http://www.ssrn.com/abstract=2838013.

Huang, X. (2008). Expected model for portfolio selection with random fuzzy returns. *International Journal of General Systems, 37*(3), 319–328.

Iorio, C., Frasso, G., D'Ambrosio, A., & Siciliano, R. (2018). A p-spline based clustering approach for portfolio selection. *Expert Systems with Applications, 95*, 88–103.

Cumming, J. (2015). An investigation into the use of reinforcement learning techniques within the algorithmic trading domain. Master's thesis. Imperial College London. http://www.doc.ic.ac.uk/teaching/distinguished-projects/2015/j.cumming.pdf.

Leccadito, A., Lozza, S. O., & Russo, E. (2007). Portfolio selection and risk management with Markov chains. *International Journal of Computer Science and Network Security, 7*(6), 115–123.

Li, D., & Ng, W. (2000). Optimal dynamic portfolio selection: Multi-period mean–variance formulation. *Mathematical Finance, 10*, 387–406.

Lim, A. E. B. (2004). Quadratic hedging and mean-variance portfolio selection with random parameters in an incomplete market. *Mathematical Methods of Operations Research, 29*, 132–161.

Lim, A. E. B., & Zhou, X. (2002). Quadratic hedging and mean-variance portfolio selection with random parameters in a complete market. *Mathematical Methods of Operations Research, 27*(1), 101–120.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*, 77–98.

Metzner, P., Dittmer, E., Jahnke, T., & Schütte, C. (2007). Generator estimation of Markov jump processes. *Journal of Computational Physics, 227*, 353–375.

Moody, J., & Saffell, M. (2001). Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks, 12*(4), 875–889.

Pendharkar, P. C., & Cusatis, P. (2018). Trading financial indices with reinforcement learning agents. *Expert Systems with Applications, 103*, 1–13.

Petropoulos, A., Chatzis, S. P., Siakoulis, V., & Vlachogiannakis, N. (2017). A stacked generalization system for automated forex portfolio trading. *Expert Systems with Applications, 90*, 290–302.

Petropoulos, A., Chatzis, S. P., & Xanthopoulos, S. (2016). A novel corporate credit rating system based on students-t hidden Markov models. *Expert Systems with Applications, 53*, 87–105.

Poznyak, A. S., Najim, K., & Gomez-Ramirez, E. (2000). *Self-learning control of finite markov chains*. New York: Marcel Dekker.

Rolski, T., Schmidli, H., Schmidt, V., & Teuggels, J. (1998). *Stochastic processes for insurance and finance*. Chichester: J. Wiley.

Sánchez, E. M., Clempner, J. B., & Poznyak, A. S. (2015a). A priori-knowledge/actor-critic reinforcement learning architecture for computing the mean-

variance customer portfolio: the case of bank marketing campaigns. *Engineering Applicationsof Artificial Intelligence, 46*(Part A), 82–92.
Sánchez, E. M., Clempner, J. B., & Poznyak, A. S. (2015b). Solving the mean-variance customer portfolio in Markov chains using iterated quadratic/lagrange programming: A credit-card customer-credit limits approach. *Expert Systemswith Applications, 42*(12), 5315–5327.
Sotomayor, L. R., & Cadenillas, A. (2009). Explicit solutions of consumption-investment problems in financial markets with regime-switching. *Mathematical Finance, 19*(2), 251–279.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press Cambridge.
Takano, Y., & Gotoh, J.-Y. (2014). Multi-period portfolio selection using kernel-based control policy with dimensionality reduction. *Expert Systems with Applications, 41*(8), 3901–3914.
Trejo, K., Clempner, J. B., & Poznyak, A. (2018). Proximal constrained optimization approach with time penalization. *Engineering Optimization*. To be published.
Wu, H. (2013). Mean-variance portfolio selection with a stochastic cash flow in a Markov-switching jump-diffusion market. *Journal of Optimization Theory and Applications, 158*, 918–934.
Xia, J. M. (2005). Mean-variance portfolio choice: Quadratic partial hedging. *Mathematical Finance, 15*(3), 533–538.
Yin, G., & Zhang, Q. (1998). *Continuous-time Markov chains and applications: A singular perturbations approach*. New York: Springer-Verlag.
Yin, G., & Zhou, X. Y. (2004a). Markowitzs mean-variance portfolio selection with regime switching: From discrete-time models to their continuous-time limits. *IEEE Transactionson Automatic Control, 49*(3), 349–360.
Yin, G., & Zhou, X. Y. (2004b). Markowitzs mean-variance portfolio selection with regime switching: From discrete-time models to their continuous-time limits. *IEEE Transactions on Automatic Control, 49*(3), 349–360.
Yiu, K., Liu, J., Siu, T., & Ching, W. (2010). Optimal portfolios with regime switching and value-at-risk constraint. *Automatica, 46*, 979–989.
Zhou, X., & Li, D. (2000). Continuous time mean variance portfolio selection: A stochastic lq framework. *Applied Mathematics and Optimization, 42*, 19–33.
Zhou, X. Y., & Yin, G. (2003). Markowitzs mean-variance portfolio selection with regime switching: A continuous-time model. *SIAM Journal on Control Optimization, 42*, 1466–1482.