École Polytechnique

BACHELOR THESIS IN COMPUTER SCIENCE

Reinforcement learning for continuous-time mean-variance portfolio selection

Author:

Oscar Peyron, École Polytechnique

Advisor:

Prof. Dr. Rudi Zagst, Technical University of Munich, Chair of Mathematical Finance

Academic year 2024/2025

Abstract

In this thesis, we approach the continuous-time mean variance problem using a reinforcement algorithm. The concept of this method is to add exploration using entropy-regularization to derive a mathematical framework for our algorithm. This thesis builds upon the paper "Continuous-time mean-variance portfolio selection: A reinforcement learning framework" (2020) [5], and aims to further develop the mathematical theory of this algorithm as well as understanding it.

Contents

1 Introduction

Minimizing risk while maximizing returns of investment strategies has always been one of the key areas of research in financial mathematics. Markowitz introduced methods for computing optimal portfolio selection in a one period model [12]. However, while his research was a breakthrough for financial sciences at the time, this method relies on market parameter estimation which is notoriously hard. Noways with the increasing computational power of computers, researchers are able to derive algorithms that dynamically manage portfolio beyond what a professional can achieve. With electronic markets prevailing it is possible to gather enough data on the market microstructure in order to produce unsupervised learning methods with good performance. Recently, methods relying on reinforcement learning have particularly attracted attention in research. Notably, Moody & Safell (2001) [14], Munos & Bourgine (1998) [15] , or Kearns (2006) [16] have enhanced existing optimal solutions for trade execution and portfolio optimization using reinforcement learning. However, as explained in Wang (2020) [5] those methods only rely on optimization problems with expected utility of discounted rewards, which does not fully characterize the uncertainty of the decision making process. In this thesis, we are focusing on applying portfolio selection using the mean-variance criterion since it is one of the most natural approaches for evaluating the performance of a portfolio. In particular, the approach relies on adding exploration to the continuous-time mean-variance investment problem to transform it into an exploration versus exploitation problem; which fits the reinforcement learning framework. While Wang (2020) [5] focuses on laying out the theory behind the exploratory mean-variance problem, we are going to dive into the technicality of the theory in this thesis. We first develop the mathematical baseline of the classical as well as the exploratory continuous-time mean-variance problem. Following this, the focus is on solving this exploratory version in order to derive an algorithm. The last section of this thesis focuses on implementing the algorithm in python and analyzing its tuning and performance.

2 Mathematical foundations for the continuous-time mean-variance problem

In this chapter we lay out the principles for understanding the theory behind continuous-time mean variance problem in financial mathematics.

2.1 Definitions

This subsection gives general mathematical definitions to concepts used in later sections.
Defintion 2.1 (Stochastic Process) [18] A real-valued stochastic process $X = (X_t)_{t\geq 0}$ is a collection of random variables $X_t : \Omega \rightarrow \mathbb{R}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{Q})$
Definition 2.2 (Filtration) [18] A filtration $\mathbb{F} = \{\mathcal{F}_{t\geq 0}\}$ on a measurable space $(\Omega, \mathcal{F})$ is a family of $\sigma-$algebras $\mathcal{F}_t \subseteq \mathcal{F}$ which is increasing in the sense that $\mathcal{F}_s \subseteq F_t$ for $s \leq t$
Definition 2.3 (Usual conditions) [13] A filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{Q})$ is said to satisfy the usual conditions if :

1. The filtration $\{\mathcal{F}_t\}$ is assumed to be right-continuous, meaning that
$$\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s.$$
2. $\mathcal{F}_0$ is trivial. This means that the only events that are measurable at time 0 are the empty set $\emptyset$ and the whole sample space Ω and should also contain all $\mathcal{F}$-measurable sets of measure 0.

Definition 2.4 (Wiener Process) [13] A *standard (one-dimensional) Wiener process* (also called *Brownian motion*) is a stochastic process $\{W_t\}_{t\geq 0}$ defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{Q})$ with the following properties:

1. $W_0 = 0$
2. The function $t \mapsto W_t$ is almost surely continuous in t.
3. The process $\{W_t\}_{t\geq 0}$ has stationary, independent increments.
4. The increments $W_{t+s} - W_s$ are $\mathcal{N}(0, t)$ distributed.

Definition 2.5 (Differential Entropy) [3] For a continuous random variable X with density function $f_X : \mathbb{R} \rightarrow \mathbb{R}$ the differential entropy is defined as:

$$H(X) = -\int_{-\infty}^{\infty} f_X(x) \log(f_X(x)) dx \tag{1}$$

Example 2.6: The entropy of the random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by :

$$
\begin{aligned}
H(X) &= -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx \\
&= -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x-\mu)^2}{2\sigma^2}\right] dx \\
&= -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[-\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{(x-\mu)^2}{2\sigma^2}\right] dx \\
&= \frac{1}{2}\ln\left(2\pi\sigma^2\right) \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx + \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{(x-\mu)^2}{2\sigma^2} dx \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}[(X-\mu)^2] \\
&= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sigma^2 \\
&= \frac{1}{2}\ln(2\pi e\sigma^2)
\end{aligned}
$$

Remark 2.7 The entropy of a random variable quantifies the average level of uncertainty or information associated to the variable's potential states or outcomes. In other words, the higher the associated differential entropy, the higher the average level of uncertainty is going to be. In example 2.7, we see therefore that for the normal distribution, high σ gives high differential entropy.

Definition 2.8 (Weak convergence) [10] The sequence of probability measures μ_n on $\mathbb{R}$ is said to converge weakly to a probability measure μ if for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$ it holds that :

$$\int_{\mathbb{R}} f d\mu_n \to \int_{\mathbb{R}} f d\mu, \quad \text{as } n \to \infty$$

Definition 2.9 (Admissibility of controls) [5] For each $(s, v) \in [0, T) \times \mathbb{R}$, consider a state process defined as a Stochastic Differential Equation (SDE) :

$$dV^{\boldsymbol{f}}(t) = \tilde{b}(t, V^{\boldsymbol{f}}(t), f)dt + \tilde{\sigma}(t, V^{\boldsymbol{f}}(t), f)dW_t \tag{2}$$

- $V^{\boldsymbol{f}} = \left(V^{\boldsymbol{f}}(t)\right)_{t\in[0,T]}$ is the state process (with values in $\mathbb{R}^n$ where $n \geq 1$ at time t under the feedback (closed-loop) control $\boldsymbol{f}$ generating the distribution of controls (open-loop control) $f = \{f_t, 0 \leq t \leq T\}$
- f_t are distributions over the space of actions $U = \mathbb{R}$.
- $\tilde{b}$ the drift term, $\tilde{\sigma}$ the diffusion term, deterministic functions.
- W is a standard Wiener process.

on $[s, T]$ with $V^{\boldsymbol{f}}(s) = v$.

Define the set of admissible controls, $\Lambda(s, v)$, as follows. Let $\mathcal{B}(\mathbb{R})$ be the Borel algebra on $\mathbb{R}$. A (distributional) control (or portfolio/strategy) process $f = \{f_t, s \leq t \leq T\}$ belongs to $\Lambda(s, v)$, if

(i) for each $s \leq t \leq T$, $f_t \in \mathcal{P}(\mathbb{R})$. This is equivalent to:

$$\int_{\mathbb{R}} f_t(\theta)d\theta = 1, \quad f_t(\theta) \geq 0, \quad \forall\theta \in \mathbb{R};$$

(ii) for each $A \in \mathcal{B}(\mathbb{R})$, $\{\int_A f_t(\theta)d\theta, s \leq t \leq T\}$ is $\mathbb{F}$-progressively measurable;

(iii) $\mathbb{E}\left[\int_s^T (\hat{\mu}^2(t,f) + \hat{\sigma}^2(t,f))dt\right] < \infty;$

(iv) $\mathbb{E}\left[\left(V^{\boldsymbol{f}}(t) - w\right)^2 + \lambda \int_s^T \int_{\mathbb{R}} f_t(\theta) \ln f_t(\theta) d\theta dt \,\middle|\, V^{\boldsymbol{f}}(s) = v\right] < \infty.$

Definition 2.10 (Feedback control and open-loop control) [5] The deterministic mapping $\boldsymbol{f}(\cdot;\cdot,\cdot)$ is called an admissible feedback control (or closed-loop control) if:

(a) $\boldsymbol{f}(\cdot,t,x)$ is a density function for each $(t,x) \in [0,T] \times \mathbb{R}$;

(b) for each $(s,v) \in [0,T) \times \mathbb{R}$, the following SDE:

$$dV^{\boldsymbol{f}}(t) = \dot{b}(\boldsymbol{f}(\cdot;t,V^{\boldsymbol{f}}(t)))dt + \dot{\sigma}(\boldsymbol{f}(\cdot;t,V^{\boldsymbol{f}}(t)))dW_t, \quad t \in [s,T]; \quad V^{\boldsymbol{f}}(s) = v \tag{3}$$

has a unique strong solution $\{V^{\boldsymbol{f}}(t), t \in [s,T]\}$, and the open-loop control $f = \{f_t, t \in [s,T]\} \in \Lambda(s,v)$, where $f_t := \boldsymbol{f}(\cdot;t,V^{\boldsymbol{f}}(t))$.

An open-loop control system is a type of control system where the output has no influence on the control action. It operates purely based on predefined inputs without using feedback to adjust its behavior. In our environment, the open-loop control f is said to be generated from the feedback control $\boldsymbol{f}(\cdot;\cdot,\cdot)$ with respect to the initial time and state (s,v).

Remark 2.11 In this definition we use the following result :

$$\dot{b}(\boldsymbol{f}(\cdot;t,V^{\boldsymbol{f}}(t))) = \tilde{b}(t,V^{\boldsymbol{f}}(t),f)$$
$$\dot{\sigma}(\boldsymbol{f}(\cdot;t,V^{\boldsymbol{f}}(t))) = \tilde{\sigma}(t,V^{\boldsymbol{f}}(t),f)$$

Theorem 2.12 (Strong solution of SDE) [23] Let f be an open-loop control generated by a feedback control $\boldsymbol{f}$ with respect to some initial time and state. Let $\tilde{b}(\cdot,\cdot,f)$ and $\tilde{\sigma}(\cdot,\cdot,f)$ be the coefficients of the stochastic differential equation (2) be continuous functions in $\mathbb{R}_{\geq 0} \times \mathbb{R}^n$ that for all $t \geq 0$, $x,y \in \mathbb{R}^n$ and for some $K > 0$ the following conditions hold:

$$\|\tilde{b}(t,x,f) - \tilde{b}(t,y,f)\| + \|\tilde{\sigma}(t,x,f) - \tilde{\sigma}(t,y,f)\| \leq K\|x-y\| \tag{4}$$

$$\|\tilde{b}(t,x,f)\|^2 + \|\tilde{\sigma}(t,x,f)\|^2 \leq K^2(1+\|x\|^2) \tag{5}$$

Then there exist a unique strong solution $V^{\boldsymbol{f}}$ of the SDE (2) and a constant C, depending only on K and T > 0, such that

$$\mathbb{E}\left[\|V^{\boldsymbol{f}}(t)\|^2\right] \leq C \cdot (1+\|x\|^2) \cdot e^{C \cdot t} \text{ for all } t \in [0,T] \tag{6}$$

Moreover,

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \|V^{\boldsymbol{f}}(t)\|^2\right] < \infty. \tag{7}$$

Definition 2.13 (Dynamic Programming) Consider the state process defined in Definition 2.8. At each time t and state $V^{\boldsymbol{f}}(t)$, the decision maker selects a distribution f_t over $U = \mathbb{R}$. The goal is to optimize an expected cost for $v_0 > 0$:

$$\min_{f\in\Lambda} \mathbb{E}\left[\int_t^T \int_U g(s,V^{\boldsymbol{f}}(s),u)f_s(u)du\, dt + h(V^{\boldsymbol{f}}(T)) \mid V^{\boldsymbol{f}}(0) = v_0\right], \tag{8}$$

- $g(t,V^{\boldsymbol{f}}(t),u)$: instantaneous cost associated with state $V^{\boldsymbol{f}}(t)$ and realized control $u \in U$,
- $h(V^{\boldsymbol{f}}(T))$: terminal cost at the final time T,

- $\Lambda = \Lambda(0, v_0)$: set of admissible measure-valued control processes as defined above.

Definition 2.14 (Bellman's Principle of Optimality) [1] An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

Definition 2.15 (Value function) [1] Let $(t, v) \in [0, T] \times \mathbb{R}$. The *value function* $J^{\boldsymbol{f}}(t, v)$ represents the expected cost of executing feedback control $\boldsymbol{f}$ starting at time t with state v:

$$J^{\boldsymbol{f}}(t, v) = \mathbb{E}\left[\int_t^T \int_U g(s, V^{\boldsymbol{f}}(s), u) f_s(u) du\, ds + h(V^{\boldsymbol{f}}(T)) \mid V^{\boldsymbol{f}}(t) = v\right]$$

Definition 2.16 (Optimal value function) [1] Let $(t, v) \in [0, T] \times \mathbb{R}$. The *optimal value function* $J(t, v)$ represents the minimal value function starting at time t with state v with respect to the set of open-loop control $f = \{f_s, s \in [t, T]\}$ generated by a feedback controls $\boldsymbol{f} = \boldsymbol{f}(\cdot; \cdot, \cdot)$ (with respect to the initial time and state (t, v)):

$$J(t, v) = \inf_{f \in \Lambda} \mathbb{E}\left[\int_t^T \int_{\mathbb{R}} g(s, V^{\boldsymbol{f}}(s), u) f_s(u) du\, ds + h(V^{\boldsymbol{f}}(T)) \mid V^{\boldsymbol{f}}(t) = v\right]$$

Definition 2.17 (Bellman's Principle of Optimality) [1]. Let $(t, v) \in [0, T] \times \mathbb{R}$. We can define the optimal value function $J(t, v)$ in a recursive manner for any intermediate time s such that $t \leq t_1 \leq T$, thanks to Bellman's principle of optimality:

$$J(t, v) = \inf_{f \in \Lambda} \mathbb{E}\left[\int_t^{t_1} \int_{\mathbb{R}} g(s, V^{\boldsymbol{f}}(s), u) f_s(u) du\, ds + J(t_1, V^{\boldsymbol{f}}(t_1)) \mid V^{\boldsymbol{f}}(t) = v\right].$$

Proposition 2.18 (Hamilton-Jacobi-Bellman equation) [22]. Let $(t, v) \in [0, T] \times \mathbb{R}$. Let $J(t, v)$ be the optimal value function as defined in Definition 2.16. Then, under suitable regularity conditions, $J(t, v)$ satisfies the Hamilton-Jacobi-Bellman (HJB) equation:

$$J_t(t, v) + \inf_{f_t \in \mathcal{P}(\mathbb{R})} \left\{ \int_{\mathbb{R}} \left[g(t, v, u) + J_v(t, v) \tilde{b}(t, v, u) + \frac{1}{2} \tilde{\sigma}^2(t, v, u) J_{vv}(t, v) \right] f_t(u)\, du \right\} = 0.$$

with the terminal condition:

$$J(T, v) = h(v).$$

Theorem 2.19 (Tonelli's Theorem in Optimal Control) [4]. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let X be a set of admissible functions. Suppose $F : \mathbb{R} \times X \to \mathbb{R} \cup \{+\infty\}$ satisfies the following conditions:

1. **Non-negativity**:

$$F(\theta, f) \geq 0, \quad \text{for all } \theta \in \mathbb{R}, f \in X.$$

2. **Measurability**: The function

$$\theta \mapsto \inf_{f \in X} F(\theta, f)$$

is measurable.

3. **Pointwise minimization**: For each θ, the infimum

$$\inf_{f \in X} F(\theta, f)$$

is well-defined and finite for at least one choice of f.

4. **Integrability**: The integral

$$\int_{\mathbb{R}} \inf_{f \in X} F(\theta, f) d\theta$$

is finite.

Then, we can exchange the infimum and the integral:

$$\inf_{f \in X} \int_{\mathbb{R}} F(\theta, f) d\theta = \int_{\mathbb{R}} \inf_{f \in X} F(\theta, f) d\theta.$$

we will apply this theorem with $X = \mathcal{P}(\mathbb{R})$

Theorem 2.20 [7] Let W_t be a standard Brownian motion on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{Q})$. Let $H(t)$ be an adapted, square-integrable process, meaning that

$$\mathbb{E}\left[\int_0^t H^2(s)\, ds\right] < \infty.$$

Then the Itô integral

$$M_t = \int_0^t H(s)\, dW_s$$

is a square-integrable martingale with expectation

$$\mathbb{E}[M_t] = 0$$

and variance

$$\mathbb{E}[M_t^2] = \mathbb{E}\left[\int_0^t H^2(s)\, ds\right]. \quad \text{(Itô's Isometry)}$$

Theorem 2.21 (Feynman-Kac formula) [7] Let $s, v \in \mathbb{R}_{\geq 0} \times \mathbb{R}$ and take $J^{\boldsymbol{f}}(s, v)$ the value function under the feedback policy $\boldsymbol{f}(\cdot; \cdot, \cdot)$ with a generated open-loop policy $f = \{\boldsymbol{f}(\cdot; t, V^{\boldsymbol{f}}(t)), t \in [s, T]\}$ with the state process $(V^{\boldsymbol{f}}(t))_{s \leq t \leq T}$ with initial state (s, v). If the value function and the state process satisfy are such that

$$J^{\boldsymbol{f}}(s, v) = \mathbb{E}\left[g\left(V^{\boldsymbol{f}}(T)\right) + \int_s^T f(t, V^{\boldsymbol{f}}(t))dt \mid V^{\boldsymbol{f}}(s) = v\right]$$
$$dV^{\boldsymbol{f}}(t) = \tilde{b}(t, V^{\boldsymbol{f}}(t), f)dt + \tilde{\sigma}(t, V^{\boldsymbol{f}}(t), f)dW_t$$

then, the value function satisfies:

$$J_s^{\boldsymbol{f}}(s, v) + \tilde{b}(s, v, f)J_v^{\boldsymbol{f}}(t, v) + \frac{1}{2}\tilde{\sigma}^2(s, v, f)J_{vv}^{\boldsymbol{f}}(s, v) + f(s, v) = 0, \quad t \in [0, T)$$
$$J^{\boldsymbol{f}}(T, v) = g(v)$$

Theorem 2.22 (Lebesgue's Dominated Convergence Theorem) [9] Let $(X_n)_{n \geq 1}$ be a sequence of random variables such that

- $X_n \to X$ almost surely (a.s.).
- There exists an integrable function $Y \in L^1$ such that $|X_n| \leq Y$ almost surely for all n.

Then,

$$X_n \in L^1, \quad X \in L^1, \quad \text{and} \quad \mathbb{E}[X_n] \to \mathbb{E}[X] \text{ as } n \to \infty$$

2.2 Classical mean-variance problem

In this paper, we consider an investment universe consisting of one risky asset and one riskless asset. Given an investment planning horizon $T > 0$, we consider a standard one-dimensional Wiener process $\{W_t\}_{t\geq 0}$ defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{Q})$ which satisfies the usual conditions. The price process of the risky asset is given by :

$$dP(t) = P(t)(\mu \, dt + \sigma \, dW_t), \quad 0 \leq t \leq T \tag{9}$$

where

- $P(t)$ is the price of the risky asset at time t,
- $\mu \in \mathbb{R}$ is the drift of the risky asset,
- $\sigma > 0$ is the volatility of the asset's returns,
- W_t is the Wiener process.

The riskless asset satisfies

$$dB(t) = rB(t)dt$$

, where $r > 0$ is the riskless interest rate. Hence $B(t) = e^{rt}$. Additionally we have

$$\gamma = \sigma^{-1}(\mu - r)$$

is the market price of risk. Denote by $\{V(\varphi, t), 0 \leq t \leq T\}$ the wealth process of an agent who rebalances her portfolio by investing in the risky and riskless assets with a self-financing strategy $\varphi = \{\varphi(t), 0 \leq t \leq T\}$. Here, $\varphi(t)$ is the number of shares of the risky asset held at time t. This gives for the discounted wealth $\tilde{V}(\varphi, t)$

$$\begin{align} \tilde{V}(\varphi, t) &= \frac{V(\varphi, t)}{B(t)} \tag{10} \\ &= e^{-rt}V(\varphi, t) \tag{11} \end{align}$$

Due to self-financing condition we have $dV(\varphi, t) = \varphi_0(t)dB(t) + \varphi_1(t)dP(t)$ [24] where $\varphi_0(t)$ and $\varphi_1(t)$ represent the allocation in the riskless asset and the risky asset respectively at time $t \in [0, T]$. Applying Itô's lemma on the above result and using the expression of $dB(t)$ and $dP(t)$ we have:

$$\begin{aligned} d\tilde{V}(\varphi, t) &= e^{-rt}dV(\varphi, t) - re^{-rt}V(\varphi, t)dt \\ &= e^{-rt}\left(\varphi_0(t)dB(t) + \varphi_1(t)dP(t)\right) - re^{-rt}V(\varphi, t)dt \\ &= \left(e^{-rt}\varphi_0(t)rB(t) - e^{-rt}\varphi_0(t)rB(t) - e^{-rt}\varphi_1(t)rP(t)\right)dt \\ &+ e^{-rt}\varphi_1(t)P(t)(\mu dt + \sigma dW_t) \\ &= e^{-rt}\varphi_1(t)P(t)(\mu - r)dt + e^{-rt}\varphi_1(t)P(t)\sigma dW_t \end{aligned}$$

We have therefore that setting $\theta(t) = e^{-rt}\varphi_1(t)P(t)$:

$$d\tilde{V}(\varphi, t) = \theta(t)((\mu - r)dt + \sigma dW_t) \tag{12}$$

Hence the wealth process satisfies the following equation.

$$d\tilde{V}(\theta, t) = \sigma\theta(t)(\gamma \, dt + dW_t), \quad 0 \leq t \leq T, \tag{13}$$

with an initial endowment $\tilde{V}(\theta, 0) = v_0 > 0$.

Our first goal is understanding the classical continuous-time mean-variance problem. First, we define the set of admissible controls for the classical setting:

$$\Lambda^{cl}(s,v) := \left\{\theta = \{\theta(t) \colon t \in [s,T]\} \colon \theta \text{ is } \mathbb{F}\text{-progressively measurable and } \mathbb{E}\left[\int_s^T (\theta(t))^2 dt\right] < \infty\right\}$$

for all $(s,v) \in [0,T) \times \mathbb{R}_{\geq 0}$. As stated above in the definition of admissible controls we have therefore that for $\theta \in \Lambda^{cl}(s,v)$, $\tilde{V}(\theta,0) = v$. This problem aims to solve the following constrained optimization problem.

$$\begin{cases} \min\limits_{\theta \in \Lambda^{cl}(0,v_0)} \quad \text{Var}\left[\tilde{V}(\theta,T)\right] \\ \qquad\qquad \mathbb{E}\left[\tilde{V}(\theta,T)\right] = \bar{v} \end{cases} \tag{14}$$

where $\bar{v}$ is a given target wealth level. This problem is a generalization of the single period portfolio selection problem. Set $g(\tilde{V}(\theta,T)) = \text{Var}\left[\tilde{V}(\theta,T)\right]$ and $h(\tilde{V}(\theta,T)) = \mathbb{E}\left[\tilde{V}(\theta,T)\right] - \bar{v}$, and we get:

$$\begin{cases} \min\limits_{\theta \in \Lambda^{cl}(0,v_0)} \; g(\tilde{V}(\theta,T)) \\ \qquad\qquad h(\tilde{V}(\theta,T)) = 0 \end{cases}$$

We apply the Lagrange multiplier to transform the constrained optimization above into an unstrained one. We define on $U(0,T) \times \mathbb{R}$ the following Lagrange function with $2w$ being the Lagrange multiplier

$$\mathcal{L}(\theta,w) = g(\tilde{V}(\theta,T)) - 2wh(\tilde{V}(\theta,T))$$

Assuming that the constrained $\mathbb{E}[\tilde{V}(\theta,T)] = \bar{v}$ is satisfied we compute:

$$\begin{aligned}
\mathcal{L}(\theta,w) &= \text{Var}(\tilde{V}(\theta,T)) - 2w(\mathbb{E}\left[\tilde{V}(\theta,T)\right] - \bar{v}) \\
&= \mathbb{E}\left[\tilde{V}(\theta,T)^2\right] - \mathbb{E}\left[\tilde{V}(\theta,T)\right]^2 - 2w(\mathbb{E}\left[\tilde{V}(\theta,T)\right] - \bar{v}) \\
&= \mathbb{E}\left[\tilde{V}(\theta,T)^2\right] - \bar{v}^2 - 2w(\mathbb{E}\left[\tilde{V}(\theta,T)\right] - \bar{v}) \\
&= \mathbb{E}\left[\tilde{V}(\theta,T)^2\right] - 2w\mathbb{E}\left[\tilde{V}(\theta,T)\right] + w^2 - w^2 + 2w\bar{v} - \bar{v}^2 \\
&= \mathbb{E}\left[(\tilde{V}(\theta,T) - w)^2\right] - (w - \bar{v})^2
\end{aligned}$$

This leads to the unconstrained optimization problem.

$$\begin{aligned} \min_{\theta \in \Lambda^{cl}(0,v_0)} \; & \mathbb{E}\left[(\tilde{V}(\theta,T) - w)^2\right] - (w - \bar{v})^2 \\ & \mathbb{E}\left[\tilde{V}(\theta,T)\right] = \bar{v} \end{aligned} \tag{15}$$

For this purpose we apply dynamic programming. By equation (15), the optimal value function is defined here for $(s,v) \in [0,T) \times \mathbb{R}$ as:

$$J^{cl}(s,v;w) := \inf_{\theta \in \Lambda^{cl}(s,v)} \mathbb{E}\left[(\tilde{V}(\theta,T) - w)^2 \mid \tilde{V}(\theta,s) = v\right] - (w - \bar{v})^2 \tag{16}$$

If assume that J^{cl} is twice continously differentiable with respect to v and continously differentiable with respect to t then it satisfies the HJB equation by Proposition 2.12:

$$J_t^{cl}(s,v;w) + \min_{\theta \in \mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}^{cl}(s,v;w) + \gamma\sigma\theta J_v^{cl}(s,v;w)\right) = 0 \tag{17}$$

Let us find J^{cl}. We have that the minimization term is a quadratic optimization. We have that the optimal $\theta^* = \theta^*(s, v; w)$ satisfies for a fixed $(s, v) \in [0, T) \times \mathbb{R}$ the first order condition:

$$\sigma^2 \theta^* J^{cl}_{vv}(s, v; w) + \gamma\sigma J^{cl}_v(s, v; w) = 0$$

Hence,

$$\theta^* = \frac{-\gamma}{\sigma} \frac{J^{cl}_v(s, v; w)}{J^{cl}_{vv}(s, v; w)} \tag{18}$$

Plugging (18) into (17) we get:

$$J^{cl}_t(s, v; w) + \frac{1}{2}\sigma^2 \left(\frac{-\gamma}{\sigma} \frac{J^{cl}_v(s, v; w)}{J^{cl}_{vv}(s, v; w)}\right)^2 J^{cl}_{vv}(s, v; w) + \gamma\sigma \frac{-\gamma}{\sigma} \frac{J^{cl}_v(s, v; w)}{J^{cl}_{vv}(s, v; w)} J^{cl}_v(s, v; w) = 0 \tag{19}$$

$$J^{cl}_t(s, v; w) - \frac{\gamma^2}{2} \frac{(J^{cl}_v(s, v; w))^2}{J^{cl}_{vv}(s, v; w)} = 0 \tag{20}$$

Similarly as in the solution for the EMV we assume

$$J^{cl}(s, v; w) = a(t)(v - w)^2 + K$$

where a is functions of t which is differentiable and K is a constant in $\mathbb{R}$ By (20) we get that using the above formula:

$$a'(t)(v - w)^2 = \frac{\gamma^2}{2} \frac{4a^2(t)(v - w)^2}{2a(t)}$$
$$a'(t)(v - w)^2 = \gamma^2 a(t)(v - w)^2$$

We have therefore that $a'(t) = \gamma^2 a(t)$. Hence we find that $a(t) = Ce^{\gamma^2 t}$ and since we have that $a(T) = 1$ as $J^{cl}(T, v; w) = (v - w)^2 - (w - \bar{v})^2$, $C = e^{-\gamma^2 T}$ giving us :

$$a(t) = e^{-\gamma^2(T-t)} \tag{21}$$

Immediately, following the condition at T, we have that $K = -(w - \bar{v})^2$. Hence giving us with (21):

$$J^{cl}(t, v; w) = (v - w)^2 e^{-\gamma^2(T-t)} - (w - \bar{v})^2 \tag{22}$$

and we have from equation (18) that the optimal feeback control polciy satisfies:

$$\theta^*(t, v; w) = \frac{-\gamma}{\sigma} \frac{2(v - w)e^{-\gamma^2(T-t)}}{2e^{-\gamma^2(T-t)}}$$
$$= \frac{-\gamma}{\sigma}(v - w)$$

We have in equation (13) the wealth process of a classical MV problem :

$$d\tilde{V}(\theta, t) = \sigma\theta(t)(\gamma\, dt + dW_t), \quad 0 \le t \le T,$$

Hence we have that $\tilde{V}^*(t) = \tilde{V}(\theta^*, t)$ satisfies for all $0 \le t \le T$:

$$d\tilde{V}^*(t) = \sigma\gamma\theta^*(t, \tilde{V}^*(t), w))dt + \sigma\theta^*(t, \tilde{V}^*(t), w))dW_t \tag{23}$$
$$= -\gamma^2(\tilde{V}^*(t) - w)dt - \gamma(\tilde{V}^*(t) - w)dW_t \tag{24}$$

We have in addition the terminal wealth $\mathbb{E}[\tilde{V}^*(T)] = \bar{v}$. We get by (24) the following ODE:

$$d\mathbb{E}\left[\tilde{V}^*(t)\right] = -\gamma^2(\mathbb{E}\left[\tilde{V}^*(t)\right] - w)dt \tag{25}$$

And therefore we find that $w = \frac{\bar{v}e^{\gamma^2 T} - v_0}{e^{\gamma^2 T} - 1}$.

2.3 Exploratory mean variance problem

In this subsection we consider a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{Q})$ with an $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ adapted Brownian motion $W = \{W_t\}_{t\geq 0}$. For simplicity we now denote $\tilde{V} = V$

Before we had an action space $U = \mathbb{R}$ which corresponds to the discounted amount of money invested into the risky asset and an (open loop) control (or strategy) $\theta = \{\theta(t), 0 \leq t \leq T\}$ taking values in $U = \mathbb{R}$. To solve the classic mean-variance problem, in a real-world application the knowledge of the model parameters is required (the volatility σ, the drift parameter μ) as we can see it in (13). Unfortunately, those model parameters are relatively difficult to estimate. For instance, the mean-return estimation is known to be particularly challenging as described in the mean-blur problem [11]. However, reinforcement learning techniques do not require estimations of market parameters [5], which makes them convenient in this setting. This is also the reason why methods using reinforcement learning are model-free methods. They learn the parameters or the market using sampling and without having already a pre-conceived model of the market.

The motivation of an exploratory continuous-time mean-variance problem instead of a classical one can be found in [6]. In the classical setting where the model is fully known (namely when μ, σ, are fully specified (13) and dynamic programming is applicable, the optimal control can be derived and represented as a deterministic mapping from the current state to an action space U (in our case modeled by θ). However, in the reinforcement learning setting, the underlying model is not known and, therefore, dynamic learning is applied to avoid estimation. The agent employs exploration to interact with and learn the undiscovered environment. This exploration can be modeled by a distribution of controls (open-loop control) $f = \{f_t, 0 \leq t \leq T\}$ over the control space U. The basis of the reinforcement learning algorithm is the step of policy evaluation in which the agent executes a control repeatedly N-times over the same time horizon, while at each round, a classical control $\theta = \{\theta(t), 0 \leq t \leq T\}$ is sampled from the control distribution f (generated by a feedback policy $\boldsymbol{f}$). We discretize equation (13), sample and execute N paths of the wealth process. Using then the law of large numbers we get as $N \to \infty$ [6]:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \Delta V^i(t) \approx \frac{1}{N}\sum_{i=1}^{N} \gamma\sigma\theta^i(t)\Delta t + \sigma\theta^i(t)(W^i(t+\Delta t) - W^i(t)), \quad t \geq 0. \\
&\to \mathbb{E}\left[\int_U \gamma\sigma\theta f_t(\theta)d\theta\Delta t + \int_U \sigma\theta f_t(\theta)d\theta\,(W_{t+\Delta t} - W_t)\right] \\
&= \mathbb{E}\left[\Delta V^{\boldsymbol{f}}(t)\right]
\end{aligned}$$

where

- θ^i, $i = 1, 2, \ldots, N$ the controls/wealth allocations sampled from f
- V^i, $i = 1, 2, \ldots, N$ be the N copies of the discretized state process under the controls θ_i
- W^i , $i = 1, 2, \ldots, N$ be N independent sample paths of the Brownion motion W
- σ and γ given as above (13)

The rest of the computations leading to the exploratory mean-variance mathematical framework is described in the reinforcement learning literature [6]. Eventually, one can rewrite the the exploratory version of the state dynamics given in (13) as:

$$dV^{\boldsymbol{f}}(t) = \tilde{b}(t, V^{\boldsymbol{f}}(t), f)dt + \tilde{\sigma}(t, V^{\boldsymbol{f}}(t), f)dW_t \tag{26}$$

where :

$$\tilde{b}(t, V^{\boldsymbol{f}}(t), f) = \int_{\mathbb{R}} \gamma\sigma\theta f_t(\theta)\,d\theta, \quad \tilde{\sigma}(t, V^{\boldsymbol{f}}(t), f) = \sqrt{\int_{\mathbb{R}} \sigma^2\theta^2 f_t(\theta)\,d\theta}, \quad \text{with } f_t \in \mathcal{P}(\mathbb{R}). \tag{27}$$

$\mathcal{P}(\mathbb{R})$ is the set of density functions of probability measures on $\mathbb{R}$ that are absolutely continuous with respect to the Lebesgue measure. As given above $\theta = \{\theta(t), 0 \leq t \leq T\}$ (representing exploration and learning) is a measure valued or distributional control process, associated to an open loop $f = \{f_t, 0 \leq t \leq T\}$.

We now define the mean and variance processes, $\hat{\mu}(t, f))$, $\hat{\sigma}^2(t, \boldsymbol{f}), 0 \leq t \leq T$ and f_t associated with the distributional control process f.

$$\hat{\mu}(t,f) = \int_{\mathbb{R}} \theta f_t(\theta) d\theta, \quad \hat{\sigma}^2(t,f) = \int_{\mathbb{R}} \theta^2 f_t(\theta) d\theta - \hat{\mu}^2(t,f)) \tag{28}$$

We therefore have the following.

$$dV^{\boldsymbol{f}}(t) = \int_{\mathbb{R}} \gamma\sigma\theta f_t(\theta)\, d\theta dt + \sqrt{\int_{\mathbb{R}} \sigma^2\theta^2 f_t(\theta)\, d\theta dW_t} \tag{29}$$

$$= \gamma\sigma \int_{\mathbb{R}} \theta f_t(\theta)\, d\theta dt + \sqrt{\int_{\mathbb{R}} \sigma^2\theta^2 f_t(\theta)\, d\theta - \hat{\mu}^2(t,f) + \hat{\mu}^2(t,f))dW_t} \tag{30}$$

$$= \gamma\sigma\hat{\mu}(t,f)dt + \sigma\sqrt{\hat{\mu}^2(t,f) + \hat{\sigma}^2(t,f)}dW_t \tag{31}$$

With the exploratory-mean variance problem defined we want to create an RL algorithm which works through "learning (exploring) while optimizing" instead of estimating the market parameters. Therefore we need a quantity to capture the level of exploration.

Given the distributional control process $f = \{f_t, 0 \leq t \leq T\}$, we can capture its level of exploration through its accumulative differential entropy. [20]. When considering the probability density f_t, its differential entropy measures the level of unpredictability of the action for a given state t and the accumulative entropy for a given distributional control f measures how much randomness the agent seeks in the whole time interval [2]. Therefore using accumulative differential entropy serves in the reinforcement learning framework as a penalty for not carrying enough exploration. [5].

$$\mathcal{H}(f) := -\int_0^T \int_{\mathbb{R}} f_t(\theta) \ln f_t(\theta) d\theta dt \tag{32}$$

As a next step we want to solve the exploratory mean-variance problem for any $w \in \mathbb{R}$. We can reformulate this problem as a minimization of the objective function (as in the classical sense) over the set of admissie distributional control (open-loop controls) $\Lambda(0, x_0)$ as defined above (definition 2.9) in which we impose a certain level of exploration $l \geq 0$. Mathematically we have :

$$\min_{f \in \Lambda(0,x_0)} \mathbb{E}\left[\left(V^{\boldsymbol{f}(T)} - w\right)^2\right] - (w - \bar{v})^2$$
$$-\int_0^T \int_{\mathbb{R}} f_t(\theta) \ln f_t(\theta) d\theta dt \geq l$$

This constrained optimization can then be reformulated using a Lagrange multiplier λ which represents the trade-off between exploitation and exploration. In sum, the entropy-regularized EMV problem is to solve

$$\begin{aligned} &\min_{f \in \Lambda(0,x_0)} \mathbb{E}\left[\left(V^{\boldsymbol{f}}(T) - w\right)^2 + \lambda \int_0^T \int_{\mathbb{R}} f_t(\theta) \ln f_t(\theta) d\theta dt\right] - (w - \bar{v})^2 \\ &\mathbb{E}\left[V^{\boldsymbol{f}}(T)\right] = \bar{v}. \end{aligned} \tag{33}$$

for $w \in \mathbb{R}$ with f the distributional control process generated by a feeback control $\boldsymbol{f}$ with respect to 0 and v_0. As in the classical setting w is determined by the terminal condition. We can then solve (33) using dynamic

programming.
Due to condition (iii) of definition 2.9 for admissible controls, we have that the stochastic integrand in (31) is well defined. We want to find if the SDE in (13) admits a strong solution. For $V \in \mathbb{R}$, $t \in [0,T]$, $f \in \Lambda(s,v)$, the quantities $\tilde{b}(t,V,f)$ and $\tilde{\sigma}(t,V,f)$ do not depend on V. Therefore for all $V_1, V_2 \in \mathbb{R}$ we get for $K > 0$:

$$0 = |\tilde{b}(t,V_1,f) - \tilde{b}(t,V_2,f)| + |\tilde{\sigma}(t,V_1,f) - \tilde{\sigma}(t,V_2,f)| \leq K|V_1 - V_2| \tag{34}$$

Moreover, we have that for $V \in \mathbb{R}$ by (31) and by condition (iii):

$$|\tilde{b}(t,V,f)|^2 = \gamma^2\sigma^2|\hat{\mu}(t,f))|^2 < \infty$$
$$|\tilde{\sigma}(t,V,f)|^2 = \sigma^2\left|\hat{\mu}^2(t,f) + \hat{\sigma}^2(t,f)\right| < \infty$$

Hence there exist $\tilde{K} > 0$ such that:

$$|\tilde{b}(t,V_1,f) - \tilde{b}(t,V_2,f)| + |\tilde{\sigma}(t,V_1,f) - \tilde{\sigma}(t,V_2,f)| \leq \tilde{K}|V_1 - V_2| \tag{35}$$
$$|\tilde{b}(t,V,f)|^2 + |\tilde{\sigma}(t,V,f)|^2 \leq \tilde{K}(1 + |V|^2) \tag{36}$$

Hence by Thereorem 2.12 the SDE (13) has a unique strong solution $V^{\boldsymbol{f}} = (V^{\boldsymbol{f}}(t))_{t\in[s,T]}$ for $s \leq t \leq T$ which satisfies $V^{\boldsymbol{f}}(s) = v$. Moreover, we have that:

$$\mathbb{E}\left[\sup_{s\leq t\leq T}\left|V^{\boldsymbol{f}}(t)\right|^2\right] < \infty. \tag{37}$$

For a fixed $w \in \mathbb{R}$:

$$J(s,v;w) = \inf_{f\in\Lambda(s,v)} \mathbb{E}\left[\left(V^{\boldsymbol{f}}(T) - w\right)^2 + \lambda\int_s^T\int_{\mathbb{R}} f_t(\theta)\ln f_t(\theta)d\theta dt \,\middle|\, V^{\boldsymbol{f}}(s) = v\right] - (w - \bar{v})^2 \tag{38}$$

for $(s,v) \in [0,T) \times \mathbb{R}$. The function $J(\cdot,\cdot;w)$ is the optimal cost function for the EMV problem in (33) Moreover, we define the value function under a given feedback control f.

$$J^{\boldsymbol{f}}(s,v;w) = \mathbb{E}\left[\left(V^{\boldsymbol{f}}(T) - w\right)^2 + \lambda\int_s^T\int_{\mathbb{R}} f_t(\theta)\ln f_t(\theta)d\theta dt \,\middle|\, V^f(s) = v\right] - (w - \bar{v})^2 \tag{39}$$

Remark 24: The value functions defined above conicide with the expression given in definition 2.15 and definition 2.16, where $h(V^{\boldsymbol{f}}(T)) = \left(V^{\boldsymbol{f}}(T) - w\right)^2 - (w - v)^2$ and $g(t, V^{\boldsymbol{f}}(t), u) = \ln f_t(u)$ for $u \in \mathbb{R}$ and $t \in [s,T]$.t

3 Solving the exploratory mean variance (EMV) problem

3.1 Optimal Gaussian policy

In this section we solve the entropy-regularized mean variance problem. For this purpose we use the dynamic programming approch given above.
Theorem 3.1 [5] *The optimal value of the entropy-regularized mean-variance (EMV) problem* (33) *is given by :*

$$J(t,v;w) = e^{-\gamma^2(T-t)}(v-w)^2 + \frac{\lambda}{4}\gamma^2(T^2 - t^2) - \frac{\lambda}{2}\left(\gamma^2 T - \ln\frac{\sigma^2}{\pi\lambda}\right)(T-t) - (w - \bar{v})^2 \tag{40}$$

for $(t,v) \in [0,T] \times \mathbb{R}$*. Moreover, the optimal feedback control is a Gaussian and its density function is given by :*

$$\boldsymbol{f}^*(\theta;t,v,w) = f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}(v-w), \frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}\right)}(\theta) \tag{41}$$

and the associated optimal wealth under the generated open-loop control f^* is the unique solution of the SDE:

$$dV^*(t) = -\gamma^2(V^*(t) - w)dt + \sqrt{\gamma^2 \left(V^*(t) - w\right)^2 + \frac{\lambda}{2}e^{\gamma^2(T-t)}}dW_t \quad V^*(0) = v_0 \tag{42}$$

Finally the Lagrange multiplier w is given by $w = \frac{ze^{p^2T} - v_0}{e^{p^2T} - 1}$
Proof: We apply Bellman's principle of optimality to the optimal value function in (38) and for $v \in \mathbb{R}$ and $0 \leq t < s \leq T$ we get:

$$J(t, v; w) = \inf_{f \in \Lambda(t,v)} \mathbb{E}\left[J(s, V^f(s); w) + \lambda \int_t^s \int_{\mathbb{R}} f_u(\theta) \ln f_u(\theta) d\theta du \,\middle|\, V^{\boldsymbol{f}}(t) = v\right]$$

If assume that J is twice continously differentiable with respect to v and continously differentiable with respect to t then it satisfies the HJB equation by Proposition 2.12.

$$J_t(t, v; w) + \min_{f_t \in \mathcal{P}(\mathbb{R})} \left(\frac{1}{2}J_{vv}(t, v, w)\tilde{\sigma}^2(t, V^{\boldsymbol{f}}(t), f) + J_v(t, v, w)\tilde{b}(t, V^{\boldsymbol{f}}(t), f) + \lambda \int_{\mathbb{R}} f_t(\theta) \ln f_t(\theta) d\theta\right) = 0$$

we have the terminal condition which is $G(T, v; w) = (v - w)^2 - (w - \bar{v})^2$.
and since we have that $\tilde{\sigma}^2(t, V^{\boldsymbol{f}}(t), f) = \int_{\mathbb{R}} \sigma^2\theta^2 f_t(\theta)\, d\theta$, and $\tilde{b}(t, V^{\boldsymbol{f}}(t), f) = \int_{\mathbb{R}} \gamma\sigma\theta f_t(\theta)\, d\theta$. Hence we get the following equation :

$$J_t(t, v; w) + \min_{f_t \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}(t, v; w) + \gamma\sigma\theta J_v(t, v; w) + \lambda \ln f_t(\theta)\right) f_t(\theta) d\theta = 0 \tag{43}$$

We can reformulate this problem into a more general framework with $q = f_t$ and therefore :

$$J_t(t, v; w) + \min_{q \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}(t, v; w) + \gamma\sigma\theta J_v(t, v; w) + \lambda \ln q(\theta)\right) q(\theta) d\theta = 0 \tag{44}$$

We define $F(\theta, q)$ on $\mathbb{R} \times \mathcal{P}(\mathbb{R})$ as

$$F(\theta, q) = \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}(t, v; w) + \gamma\sigma\theta J_v(t, v; w) + \lambda \ln q(\theta)\right) q(\theta). \tag{45}$$

Since $q(\theta) \geq 0$ and the quadratic term is always non-negative, we have $F(\theta, q) \geq 0$. Additionally, $F(\theta, q)$ is convex in q due to the entropy term $\lambda \ln q(\theta)$, ensuring that pointwise minimization is well-defined. Indeed, we have that :

$$\frac{\partial^2 F(\theta, q)}{\partial q(\theta)^2} = \frac{\lambda}{q(\theta)} > 0 \tag{46}$$

since $q(\theta) > 0$ for all $\theta \in \mathbb{R}$. Given that $\theta \mapsto \inf_{q \in \mathcal{P}(\mathbb{R})} F(\theta, q)$ is measurable and integrable, we can apply Tonelli's theorem :

$$\min_{q \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} F(\theta, q) d\theta = \int_{\mathbb{R}} \min_{q \in \mathcal{P}(\mathbb{R})} F(\theta, q) d\theta. \tag{47}$$

Thus, we minimize $F(\theta, q)$ pointwise with respect to $q(\theta)$, leading to the first-order condition :

$$\frac{\partial F(\theta, q)}{\partial q(\theta)} = \frac{1}{2}\sigma^2\theta^2 J_{vv}(t, v; w) + \gamma\sigma\theta J_v(t, v; w) + \lambda(\ln q(\theta) + 1) = 0. \tag{48}$$

We get therefore:

$$q(\theta) = \exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}\sigma^2\theta^2 J_{vv}(t,v;w) + \gamma\sigma\theta J_v(t,v;w)\right) - 1\right)$$

q is measurable as we have that it is the exponential of a polynomial function. Composition of a measurable function with a continuous function gives a measurable and hence q is measurable. Normalizing the density function of the feedback distributional control result so that $\int_{\mathbb{R}} q(\theta)d\theta = 1$ we get the following result :

$$\boldsymbol{f}^*(\theta;t,v,w) = \frac{\exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}\sigma^2\theta^2 J_{vv}(t,v;w) + \gamma\sigma\theta J_v(t,v;w)\right)\right)}{\int_{\mathbb{R}}\exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}\sigma^2\theta^2 J_{vv}(t,v;w) + \gamma\sigma\theta J_v(t,v;w)\right)\right)d\theta} \tag{49}$$

Moreover we have, assuming that $J_{vv}(t,v;w) > 0$, that :

$$\begin{aligned}
&-\frac{1}{\lambda}\left(\frac{1}{2}\sigma^2\theta^2 J_{vv}(t,v;w) + \gamma\sigma\theta J_v(t,v;w)\right)\\
&= -\frac{1}{2}\frac{\sigma^2 J_{vv}(t,v;w)}{\lambda}\left(\theta^2 + 2\frac{\gamma}{\sigma}\frac{\theta J_v(t,v;w)}{J_{vv}(t,v;w)}\right)\\
&= -\frac{1}{2}\frac{\sigma^2 J_{vv}(t,v;w)}{\lambda}\left(\left(\theta + \frac{\gamma}{\sigma}\frac{J_v(t,v;w)}{J_{vv}(t,v;w)}\right)^2 - \left(\frac{\gamma}{\sigma}\frac{J_v(t,v;w)}{J_{vv}(t,v;w)}\right)^2\right).
\end{aligned}$$

It follows:

$$\begin{aligned}
\boldsymbol{f}^*(\theta;t,v,w) &= \frac{\exp\left(-\frac{1}{2}\frac{1}{\frac{\lambda}{\sigma^2 J_{vv}(t,v;w)}}\left(\left(\theta + \frac{\gamma}{\sigma}\frac{J_v(t,v;w)}{J_{vv}(t,v;w)}\right)^2\right)\right)}{\int_{\mathbb{R}}\exp\left(-\frac{1}{2}\frac{1}{\frac{\lambda}{\sigma^2 J_{vv}(t,v;w)}}\left(\left(\theta + \frac{\gamma}{\sigma}\frac{J_v(t,v;w)}{J_{vv}(t,v;w)}\right)^2\right)\right)d\theta} && (50)\\
&= f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}\frac{J_v(t,v;w)}{J_{vv}(t,v;w)},\frac{\lambda}{\sigma^2 J_{vv}(t,v;w)}\right)}(\theta) && (51)
\end{aligned}$$

Plugging the result in (51) into the HJB equation in (44) and setting $f_t^*(\theta) = \boldsymbol{f}^*(\theta : t,v,w)$ we find:

$$\begin{aligned}
0 = J_t(t,v;w) &+ \frac{1}{2}\sigma^2 J_{vv}(t,v;w)\int_{\mathbb{R}}\theta^2 f_t^*(\theta)\,d\theta\\
&+ \gamma\sigma J_v(t,v;w)\int_{\mathbb{R}}\theta f_t^*(\theta)\,d\theta\\
&+ \lambda\int_{\mathbb{R}} f_t^*(\theta)\ln f_t^*(\theta)\,d\theta.
\end{aligned}$$

Take now $X \sim \mathcal{N}\left(-\frac{\gamma}{\sigma}\frac{J_v(t,v;w)}{J_{vv}(t,v;w)},\frac{\lambda}{\sigma^2 J_{vv}(t,v;w)}\right)$. From example 2.6 we find:

$$-H(X) = \int_{\mathbb{R}} f_t^*(\theta)\ln f_t^*(\theta)d\theta = -\frac{1}{2}\ln\left(\frac{2\pi e\lambda}{\sigma^2 J_{vv}(t,v;w)}\right)$$

and from the first and second moment of Gaussian distributions :

$$\begin{aligned}
\int_{\mathbb{R}}\theta f_t^*(\theta)d\theta &= -\frac{\gamma}{\sigma}\frac{J_v(t,v;w)}{J_{vv}(t,v;w)}\\
\int_{\mathbb{R}}\theta^2 f_t^*(\theta)d\theta &= Var\left[X\right] + \mathbb{E}\left[X\right]^2\\
&= \frac{\lambda}{\sigma^2 J_{vv}(t,v;w)} + \frac{\gamma^2}{\sigma^2}\frac{J_v^2(t,v;w)}{J_{vv}^2(t,v,;w)}
\end{aligned}$$

Combining everything we get:

$$\begin{aligned}0 =& J_t(t,v;w)\\ &+\frac{1}{2}\sigma^2 J_{vv}(t,v;w)\left(\frac{\lambda}{\sigma^2 J_{vv}(t,v;w)}+\frac{\gamma^2}{\sigma^2}\frac{J_v^2(t,v;w)}{J_{vv}^2(t,v,;w)}\right)\\ &+\gamma\sigma J_v(t,v;w)\left(-\frac{\gamma}{\sigma}\frac{J_v(t,v;w)}{J_{vv}(t,v;w)}\right)\\ &-\frac{\lambda}{2}\ln\left(\frac{2\pi e\lambda}{\sigma^2 J_{vv}(t,v;w)}\right)\end{aligned}$$

This gives the following partial differential equation (PDE):

$$J_t(t,v;w)-\frac{\gamma^2}{2}\frac{J_v^2(t,v;w)}{J_{vv}(t,v;w)}+\frac{\lambda}{2}\left(1-\ln\frac{2\pi e\lambda}{\sigma^2 J_{vv}(t,v;w)}\right)=0 \tag{52}$$

with boundary condition:

$$J(T,v;w)=(v-w)^2-(w-\bar{v})^2 \tag{53}$$

In order to find the optimal solution we assume:

$$J(t,v;w)=a(t)(v-w)^2+c(t) \tag{54}$$

with $a:\mathbb{R}_{\geq 0}\to\mathbb{R}$, $c:\mathbb{R}_{\geq 0}\to\mathbb{R}$ it follows:

$$J_t(t,v;w)=a'(t)(v-w)^2+c'(t), \tag{55}$$

$$J_v(t,v;w)=2a(t)(v-w), \tag{56}$$

$$J_{vv}(t,v;w)=2a(t), \tag{57}$$

$$\frac{J_v^2(t,v;w)}{J_{vv}(t,v;w)}=2a(t)(v-w)^2 \tag{58}$$

We have that from (52) and (55):

$$\begin{aligned}c'(t)&=-\frac{\lambda}{2}\left(1-\ln\frac{2\pi e\lambda}{\sigma^2 J_{vv}(t,v;w)}\right)\\ &=-\frac{\lambda}{2}\left(1-\ln\frac{2\pi e\lambda}{\sigma^2 2a(t)}\right)\\ &=-\frac{\lambda}{2}\left(\ln e-\ln\frac{2\pi e\lambda}{\sigma^2 2a(t)}\right)\\ &=-\frac{\lambda}{2}\left(\ln\frac{\sigma^2 a(t)}{\pi\lambda}\right)\end{aligned}$$

and using (58) and the expression of c'(t) we get :

$$J_t(t,v;w)=\gamma^2 a(t)(v-w)^2-\frac{\lambda}{2}\ln\frac{\sigma^2 a(t)}{\pi\lambda} \tag{59}$$

It follows from (55) by identification that $a'(t)=\gamma^2 a(t)$ with $a(T)=1$. This is a differential equation with solution: $a(t)=e^{-\gamma^2(T-t)}$. We have therefore that:

$$J_t(t,v;w)=\gamma^2 e^{-\gamma^2(T-t)}(v-w)^2-\frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda}+\frac{\lambda}{2}\gamma^2(T-t)$$

Again from (55) we find:

$$c'(t) = -\frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda} + \frac{\lambda}{2}\gamma^2(T-t)$$

and hence : $c(t) = -\frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda}t - \frac{\lambda}{4}\gamma^2 t^2 + \frac{\lambda}{2}\gamma^2 Tt + K$ with $K \in \mathbb{R}$. Due to the boundary condition (59)

$$c(T) = -(w-\bar{v})^2$$

it has to hold:

$$\begin{aligned} -\frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda}T - \frac{\lambda}{4}\gamma^2T^2 + \frac{\lambda}{2}\gamma^2T^2 + K = -(w-\bar{v})^2 \\ \implies K = \frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda}T + \frac{\lambda}{4}\gamma^2T^2 - \frac{\lambda}{2}\gamma^2T^2 - (w-\bar{v})^2 \end{aligned}$$

Hence we get:

$$\begin{aligned} c(t) &= \frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda}(T-t) + \frac{\lambda}{4}\gamma^2(T^2-t^2) - \frac{\lambda}{2}\gamma^2T(T-t) - (w-\bar{v})^2 \\ &= \frac{\lambda}{4}\gamma^2(T^2-t^2) - \frac{\lambda}{2}\left(\gamma^2T - \ln\frac{\sigma^2}{\pi\lambda}\right)(T-t) - (w-\bar{v})^2 \end{aligned}$$

Finally combining the results:

$$J(t,v,w) = e^{-\gamma^2(T-t)}(v-w)^2 + \frac{\lambda}{4}\gamma^2(T^2-t^2) - \frac{\lambda}{2}\left(\gamma^2T - \ln\frac{\sigma^2}{\pi\lambda}\right)(T-t) - (w-\bar{v})^2 \tag{60}$$

In particular we get for any $(t,v) \in [0,T] \times \mathbb{R}$:

$$J_v(t,v;w) = 2e^{-\gamma^2(T-t)}(v-w) \tag{61}$$

$$J_{vv}(t,v;w) = 2e^{-\gamma^2(T-t)} > 0 \tag{62}$$

and

$$-\frac{\gamma}{\sigma}\frac{J_v(t,v;w)}{J_{vv}(t,v;w)} = -\frac{\gamma}{\sigma}\frac{2e^{-\gamma^2(T-t)}(v-w)}{2e^{-\gamma^2(T-t)}} = -\frac{\gamma}{\sigma}(v-w) \tag{63}$$

$$\frac{\lambda}{\sigma^2 J_{vv}(t,v;w)} = \frac{\lambda}{2\sigma^2 e^{-\gamma^2(T-t)}} = \frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)} \tag{64}$$

By (63) and (64) we find the optimal feedback control:

$$\boldsymbol{f}^*(\theta;t,v,w) = f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}(v-w),\frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}\right)}(\theta) \tag{65}$$

for all $(t,v) \in [0,T] \times \mathbb{R}$.
In particular we have:

$$\hat{\mu}(t,\boldsymbol{f}) = -\frac{\gamma}{\sigma}(V(t)-w) \quad \hat{\sigma}^2(t,\boldsymbol{f}) = \frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}$$

Combining those results we get :

$$\tilde{b}(t,V^{\boldsymbol{f}}(t),f) = \gamma\sigma\hat{\mu}(t,\boldsymbol{f}) = -\gamma^2(V^{\boldsymbol{f}}(t)-w) \tag{66}$$

$$\tilde{\sigma}(t,V^{\boldsymbol{f}}(t),f) = \sigma\sqrt{\hat{\mu}^2(t,\boldsymbol{f}) + \hat{\sigma}^2(t,\boldsymbol{f})} = \sqrt{\gamma^2\left(V^{\boldsymbol{f}}(t)-w\right)^2 + \frac{\lambda}{2}e^{\gamma^2(T-t)}} \tag{67}$$

and plugging into the SDE in (26)

$$dV^*(t) = -\gamma^2(V^*(t) - w)dt + \sqrt{\gamma^2\left(V^*(t) - w\right)^2 + \frac{\lambda}{2}e^{\gamma^2(T-t)}}dW_t \quad V^*(0) = v_0 \tag{68}$$

We can determine the Lagrange multiplier w by using the condition $\mathbb{E}\left[V^*(T)\right] = \bar{v}$. One integrate the expression in (68) from 0 to t and compute its expectation giving:

$$\begin{aligned}\mathbb{E}\left[V^*(t)\right] &= \mathbb{E}\left[v_0 + \int_0^t -\gamma^2(V^*(s) - w)ds + \int_0^t \sqrt{\gamma^2\left(V^*(s) - w\right)^2 + \frac{\lambda}{2}e^{\gamma^2(T-s)}}dW_s\right] \\ &= \mathbb{E}\left[v_0 + \int_0^t -\gamma^2(V^*(s) - w)ds\right] + \mathbb{E}\left[\int_0^t \sqrt{\gamma^2\left(V^*(s) - w\right)^2 + \frac{\lambda}{2}e^{\gamma^2(T-s)}}dW_s\right]\end{aligned}$$

Define:

$$M_t = \int_0^t \sqrt{\gamma^2\left(V^*(s) - w\right)^2 + \frac{\lambda}{2}e^{\gamma^2(T-s)}}dW_s$$

. We have that M_t is an Itô integral with respect to the Brownian motion and is therefore adapted to the filtration $\mathbb{F}$. Let us show that M_t is a martingale. To do so we need to prove that the function $g(t, V^*(t)) = \sqrt{\gamma^2\left(V^*(t) - w\right)^2 + \frac{\lambda}{2}e^{\gamma^2(T-t)}}$ is square integrable in expectation. From (37) we have shown that $\mathbb{E}\left[\sup_{s\leq t\leq T}|V^*(t)|^2\right] < \infty$. In particular we have that for $w \in \mathbb{R}$,

$$\mathbb{E}\left[(V^*(t) - w)^2\right] < \infty \quad \forall t \in [0, T] \tag{69}$$

Moreover, $\mathbb{E}\left[\int_0^{+\infty} \frac{\lambda}{2}e^{\gamma^2(T-t)}dt\right] < \infty$. Hence we can conclude that for all $t \in [0, T]$

$$\mathbb{E}\left[\int_0^t \gamma^2\left(V^*(s) - w\right)^2 + \frac{\lambda}{2}e^{\gamma^2(T-s)}ds\right] < \infty \tag{70}$$

M_t is therefore a martingale by theorem 2.20 and:

$$\mathbb{E}\left[M_t\right] = \mathbb{E}\left[M_0\right] = 0 \tag{71}$$

This yields the following:

$$\begin{aligned}\mathbb{E}\left[V^*(t)\right] &= \mathbb{E}\left[v_0 + \int_0^t -\gamma^2(V^*(s) - w)ds\right] \\ &= v_0 + \mathbb{E}\left[\int_0^t -\gamma^2(V^*(s) - w)ds\right] \\ &= v_0 + \int_0^t -\gamma^2(\mathbb{E}\left[V^*(s)\right] - w)ds\end{aligned}$$

since $\mathbb{E}\left[V^*(s)\right] < \infty$. We therefore solve the following ODE:

$$\begin{aligned} & \frac{d\mathbb{E}[V^*(t)]}{dt} = -\gamma^2(\mathbb{E}[V^*(t)] - w) \\ \Leftrightarrow \quad & \frac{d\mathbb{E}[V^*(t)]}{dt} + \gamma^2\mathbb{E}[V^*(t)] = \gamma^2 w \\ \Leftrightarrow \quad & e^{\gamma^2 t}\frac{d\mathbb{E}[V^*(t)]}{dt} + \gamma^2\mathbb{E}[V^*(t)]e^{\gamma^2 t} = \gamma^2 w e^{\gamma^2 t}\end{aligned}$$

Integrating both sides from 0 to T we get :

$$\begin{aligned} & \int_0^T e^{\gamma^2 s}\frac{d\mathbb{E}[V^*(s)]}{ds} + \gamma^2\mathbb{E}[V^*(s)]e^{\gamma^2 s}ds = \int_0^T \gamma^2 w e^{\gamma^2 s}ds \\ \Leftrightarrow \quad & \int_0^T \frac{d}{ds}(\mathbb{E}[V^*(s)]e^{\gamma^2 s})ds = \int_0^T \frac{d}{ds}(we^{\gamma^2 s})ds \\ \Leftrightarrow \quad & \mathbb{E}[V^*(t)]e^{\gamma^2 T} - \mathbb{E}[V^*(0)] = w(e^{\gamma^2 T} - 1) \\ \Leftrightarrow \quad & \frac{\mathbb{E}[V^*(t)]e^{\gamma^2 T} - \mathbb{E}[V^*(0)]}{e^{\gamma^2 T} - 1} = w \end{aligned}$$

as $\mathbb{E}[V^*(T)] = \bar{v}$ and $\mathbb{E}[V^*(0)] = v_0$ we get :

$$w = \frac{\bar{v}e^{\gamma^2 T} - v_0}{e^{\gamma^2 T} - 1}$$

□

Remark 3.2 The variance of the optimal Gaussian policy is a decreasing function of t ($t \mapsto \frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}$). Since this policy measures the level of exploration of the agent, one can interpret that as a stronger exploration at the beginning and a lower exploration towards the end of the investment period. Intuitively, as the RL agent learns more about the environment and as it progressively reaches the investment horizon it seeks to exploit more than explore to increase chances to reach the given target return.

3.2 Equivalence of solvability of classical mean-variance and EMV

In this subsection we investigate how solving one of the problems (classical or EMV problem) can lead directly to the solution of the other. This relationship has been first discovered for the infinite horizon Linear Quadratic case [6] and has been stated in [5]. Let us first consider the following lemma.

Lemma 3.3 Define by $V^* = (V^*(t))_{0\leq t\leq T}$ the optimal wealth process in the EMV problem and $(V^*_{cl}(t))_{0\leq t\leq T}$ the optimal wealth process in the classical mean-variance problem. Then (i) and (ii) hold

1. (i) $\inf_{T\to\infty}\mathbb{E}\left[(V^*(t))^2\right] = 0$ if and only if $\inf_{T\to\infty}\mathbb{E}\left[(V^*_{cl}(t))^2\right] = 0$
2. (ii) $\mathbb{E}\left[\int_0^\infty (V^*(t))^2\right] < \infty$ if and only if $\mathbb{E}\left[\int_0^\infty (V^*_{cl}(t))^2\right] < \infty$

Proof: Proven in [6]

Theorem 3.4 (*Solvability equivalence between classical and EMV problems*) [6] *If the exploratory mean-variance is solvable* (33) *with solution given in statement (i) then the classical mean-variance is solvable* (15) with solution given in (ii) and vice-versa.

(i) *The function*

$$J(t,v;w) = (v-w)^2 e^{-\gamma^2(T-t)} + \frac{\lambda\gamma^2}{4}(T^2 - t^2) - \frac{\lambda}{2}(\gamma^2 T - \ln\frac{\sigma^2}{\pi\lambda})(T-t) - (w-\bar{v})^2$$

for $(t,v) \in [0,T]\times\mathbb{R}$, *is the optimal value function of the EMV problem in* (33), *and the corresponding optimal feedback control is:*

$$\boldsymbol{f}^*(\theta;t,v,w) = f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}(v-w),\, \frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}\right)}(\theta).$$

(ii) *The function*

$$J^{cl}(t,v;w) = (v-w)^2 e^{-\gamma^2(T-t)} - (w-\bar{v})^2$$

for $(t,v) \in [0,T] \times \mathbb{R}$, *is the optimal value function of the classical MV problem in* (15), *and the corresponding optimal feedback control is:*

$$\theta^*(t,v;w) = -\frac{\gamma}{\sigma}(v-w).$$

The Lagrange multiplier in both settings (i) and (ii) is

$$w = \frac{\bar{v}e^{\gamma^2 T} - v_0}{e^{\gamma^2 T} - 1}.$$

Proof: We have shown in the above parts that the solution stated in (i) and (ii) are indeed the solutions of the EMV and the classical mean-variance problem respectively. Now we examine the equivalence relation between (i) and (ii). We see that if J solves the HJB equation in (43) then J^{cl} solves the HJB equation in (17) and vice versa. Hence, we want to prove that the admissibility of one problem implies the admissibility of the other. Let us consider $V^* = (V^*(t))_{0\leq t\leq T}$ the optimal wealth process in the EMV problem and $(V^*_{cl}(t))_{0\leq t\leq T}$ the optimal wealth process in the classical mean-variance problem. The EMV and the classical MV are only defined if respectively $\mathbb{E}\left[|V^*(t)|^2\right] < \infty$ and $\mathbb{E}\left[|V^*_{cl}(t)|^2\right] < \infty$ as demonstrated in (37) and in (16). One has that $\mathbb{E}\left[|V^*(t)|^2\right] < \infty$ if and only if $\mathbb{E}\left[|V^*_{cl}(t)|^2\right] < \infty$ thanks to Lemma 3.23and the definition of $\mathbb{E}\left[V^*(t)\right]$ in (25). □

Remark: The fact that the Lagrange multiplier is the same for both problems follows from the fact that the drift term in (24) of the classical MV is identical to the EMV counterpart in (68). However, one can observe that the diffusion coefficient is different in (24) and (68). It is also interesting to note that the Lagrange multiplier w being equal in both of those problems means that the expectation of the optimal wealth is identical. This is because the expectation of the optimal wealth follows the same ODE (25) in both classical and exploratory mean variance problems.

We now want to investigate if and how the solution of the EMV problem converges as the exploration rate λ tends to 0. In particular, the variance of $\boldsymbol{f}^*(\theta;t,v,w)$ tends to 0.

Theorem 3.5 [5] *Assume that (i) or equivalently (ii) of Theorem 3.4 hold. Denote the optimal feedback control of the EMV problem as:*

$$\boldsymbol{f}^*_\lambda(\theta;t,v,w) = f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}(v-w),\frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}\right)}(\theta).$$

Then for all $(t,v;w) \in [0,T] \times \mathbb{R}_{\geq 0} \times \mathbb{R}$ *we have the following weak-convergence of probability measures.*

$$\lim_{\lambda\to 0} \boldsymbol{f}^*_\lambda(\cdot,t,v;w) = \delta_{\theta^*(t,v;w)}(\cdot) \quad \textit{weakly}$$

Additionally,

$$\lim_{\lambda\to 0} \left| J(t,v;w) - J^{cl}(t,v;w) \right| = 0$$

Proof: Let $(t,v,w) \in [0,T] \times \mathbb{R}_{\geq 0} \times \mathbb{R}$, $\sigma_\lambda = \frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}$ and $X \sim \mathcal{N}(\theta^*(t,v;w),\sigma_\lambda)$ (with $\theta^*(t,v;w)$ as defined in theorem 3.4 or equivalently $X \sim \mathcal{N}(-\frac{\gamma}{\sigma}(v-w),\sigma_\lambda)$. Furthermore, let ζ an arbitrary bounded continuous function from $\mathbb{R}$ to $\mathbb{R}$. For $\lambda > 0$ consider the following :

$$\begin{aligned} I_\lambda &= \mathbb{E}\left[\zeta(X)\right] \\ &= \int_{\mathbb{R}} \zeta(\theta)\boldsymbol{f}^*_\lambda(\theta;t,v,w)d\theta \end{aligned}$$

The above equality can be stated as a result of Theorem 3.4 and the expression of the optimal feedback control. We want to prove that :

$$\lim_{\lambda \to 0} I_\lambda = \int_{\mathbb{R}} \zeta(\theta)\delta_{\theta^*(t,v,w)}(\theta)d\theta = \zeta(\theta^*(t,v;w)) = \int_{\mathbb{R}} \zeta(\theta)d\delta_{\theta^*(t,v;w)}(\theta) \tag{72}$$

in order to prove weak convergence of the optimal feedback control measure to the associated dirac probability measure. Let $Z \sim \mathcal{N}(0,1)$ and f_Z its probability density function. We have that :

$$Z = \frac{X - \theta^*(t,v,w)}{\sigma_\lambda}$$

. Hence,

$$\begin{aligned} I_\lambda = \mathbb{E}[\zeta(X)] &= \mathbb{E}[\zeta(\theta^*(t,v,w) + \sigma_\lambda Z)] \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}}\zeta(\theta^*(t,v,w) + \sigma_\lambda z)e^{-\frac{z^2}{2}}dz \end{aligned}$$

Since ζ is bounded on $\mathbb{R}$ we have that there exist $M > 0$ such that:

$$|\zeta(\theta^*(t,v,w) + \sigma_\lambda Z)| \leq M \quad \forall z \in \mathbb{R} \tag{73}$$

. Hence:

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}|\zeta(\theta^*(t,v,w) + \sigma_\lambda Z)| \leq M\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \quad \forall z \in \mathbb{R} \tag{74}$$

where $z \mapsto M\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ is integrable on $\mathbb{R}$. By the dominated convergence theorem we have that :

$$\lim_{\lambda \to 0} I_\lambda = \lim_{\lambda \to 0}\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}}\zeta(\theta^*(t,v,w) + \sigma_\lambda z)e^{-\frac{z^2}{2}}dz = \int_{\mathbb{R}} \lim_{\lambda \to 0}\frac{1}{\sqrt{2\pi}}\zeta(\theta^*(t,v,w) + \sigma_\lambda z)e^{-\frac{z^2}{2}}dz \tag{75}$$

and therefore we have since $\sigma_\lambda \to 0$ as $\lambda \to 0$:

$$\begin{aligned} \int_{\mathbb{R}} \lim_{\lambda \to 0}\frac{1}{\sqrt{2\pi}}\zeta(\theta^*(t,v,w) + \sigma_\lambda z)e^{-\frac{z^2}{2}}dz &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}}\zeta(\theta^*(t,v,w))e^{-\frac{z^2}{2}}dz \\ &= \zeta(\theta^*(t,v,w))\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz \\ &= \zeta(\theta^*(t,v,w)) \end{aligned}$$

The last line is a result of $z \to \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ being the density function of the random variable $Z \sim \mathcal{N}(0,1)$. Hence $\lim_{\lambda\to 0} I_\lambda = \zeta(\theta^*(t,v;w))$ for all $\zeta : \mathbb{R} \to \mathbb{R}$, bounded and continuous. This proves the weak-convergence of the optimal feedback control $\boldsymbol{f}^*_\lambda(\cdot,t,v,w)$ to its associated Dirac probability measure.

Let us now prove the second result. We have that for all $(t,v;w) \in [0,T] \times \mathbb{R}_{\geq 0} \times \mathbb{R}$:

$$\begin{aligned} \Big|J(t,v;w) - J^{cl}(t,v;w)\Big| &= \left|\frac{\lambda\gamma^2}{4}(T^2 - t^2) - \frac{\lambda}{2}(\gamma^2 T - \ln\frac{\sigma^2}{\pi\lambda})(T-t)\right| \\ &= \left|\frac{\lambda\gamma^2}{4}(T^2 - t^2) - \frac{\lambda}{2}\gamma^2 T(T-t) + \frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda}(T-t)\right| \end{aligned}$$

Furthermore, we set $h(\lambda) = \ln\left(\frac{\sigma^2}{\pi\lambda}\right)$ and $g(\lambda) = \frac{2}{\lambda}$. Note that h and g ar differentiable on $(0,\infty)$, which gives for $\lambda > 0$ $h'(\lambda) = -\frac{1}{\lambda}$ and $g'(\lambda) = \frac{-2}{\lambda^2}$ with $\lim_{\lambda\to 0} h'(\lambda) = \lim_{\lambda\to 0} g'(\lambda) = -\infty$. Using l'Hopital's rule:

$$\lim_{\lambda \to 0}\frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda} = \lim_{\lambda \to 0}\frac{h(\lambda)}{g(\lambda)} = \lim_{\lambda \to 0}\frac{h'(\lambda)}{g'(\lambda)} = \lim_{\lambda \to 0}\frac{\lambda}{2} = 0$$

. In addition, the terms $\frac{\lambda\gamma^2}{4}(T^2-t^2)$ and $\frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda}(T-t)$ also tend to 0 as $\lambda\to 0$. In conclusion we get:

$$\begin{aligned}\lim_{\lambda\to 0}\left|J(t,v;w)-J^{cl}(t,v;w)\right| &= \lim_{\lambda\to 0}\left|\frac{\lambda\gamma^2}{4}(T^2-t^2)-\frac{\lambda}{2}\gamma^2T(T-t)+\frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda}(T-t)\right|\\ &\le \lim_{\lambda\to 0}\left|\frac{\lambda\gamma^2}{4}\right|(T^2-t^2)+\lim_{\lambda\to 0}\left|\frac{\lambda}{2}\gamma^2\right|T(T-t)+\lim_{\lambda\to 0}\left|\frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda}\right|(T-t)\\ &= \lim_{\lambda\to 0}\left|\frac{\lambda\gamma^2}{4}\right|(T^2-t^2)+\lim_{\lambda\to 0}\left|\frac{\lambda}{2}\gamma^2\right|T(T-t)+\lim_{\lambda\to 0}\left|\frac{h'(\lambda)}{g'(\lambda)}\right|(T-t)\\ &= 0\end{aligned}$$

□

By this result a classical control $\theta=\{\theta(t),t\ge 0\}$ can be regarded as a random variable with Dirac distribution with $f=\{f(\theta,t)),t\ge 0\}$ where $f(\cdot,t)=\delta_{\theta(t)}(\cdot)$ [6].

3.3 Cost of exploration

Here we want to investigate the cost of exploration of the reinforcement learning algorithm. We define in the cost similarly to the infinite horizon algorithm developped in [6].

$$C^{\theta^*,\boldsymbol{f}^*}(0,v_0;w):=\left(J(0,v_0;w)-\lambda\mathbb{E}\left[\int_0^T\int_{\mathbb{R}}\boldsymbol{f}^*(\theta;t)\ln\boldsymbol{f}^*(\theta;t)\,d\theta\,dt\mid V^{f^*}(0)=v_0\right]\right)\\ -J^{cl}(0,v_0;w) \tag{76}$$

Equation (76) measures the loss in the original (i.e non-exploratory strategy) objective due to exploration.[5].
Theorem 3.6[5] *Assume that (i) or (ii) in Theorem 3.5 holds then we have that the exploration cost for the mean-variance problem is*

$$C^{\theta^*,\boldsymbol{f}^*}(0,v_0;w)=\frac{\lambda T}{2},\quad x_0\in\mathbb{R},\quad w\in\mathbb{R} \tag{77}$$

Proof: Let $f^*=\{\boldsymbol{f}^*(\cdot;t,v_0),t\in[0,T]\}$ be the open-loop control generated by the optimal feedback control $\boldsymbol{f}^*(\cdot;\cdot,\cdot)$ from the EMV problem given the initial state v_0 at $t=0$. Then we get for all $t\in[0,T]$:

$$\boldsymbol{f}^*(\theta;t,v_0)=f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}(V^*(t)-w),\frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}\right)}(\theta) \tag{78}$$

where $\{V^*(t),t\in[0,T]\}$ is the corresponding optimal wealth process of the EMV problem. Since $\boldsymbol{f}^*(\theta;t,v_0)$ is the density function of a Gaussian distribution with mean $-\frac{\gamma}{\sigma}(V^*(t)-w)$ and variance $\frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}$, the formula in example 2.6 gives :

$$\int_{\mathbb{R}}\boldsymbol{f}^*(\theta;t,v_0)\ln\boldsymbol{f}^*(\theta;t,v_0)d\theta=-\frac{1}{2}\ln\left(2\pi e\frac{\lambda}{2\sigma^2}e^{\gamma^2(T-t)}\right). \tag{79}$$

We have therefore that:

$$\begin{aligned}\int_0^T\int_{\mathbb{R}}\boldsymbol{f}^*(\theta;t,v_0)\ln\boldsymbol{f}^*(\theta;t,v_0)d\theta dt &= \int_0^T-\frac{1}{2}\ln\left(2\pi e\frac{\lambda}{\sigma^2}\right)-\frac{1}{2}\ln\left(e^{\gamma^2(T-t)}\right)dt\\ &= -\frac{1}{2}\left(\ln\left(2\pi e\frac{\lambda}{2\sigma^2}\right)T+\int_0^T\gamma^2T-\gamma^2tdt\right)\\ &= -\frac{1}{2}\left(1+\ln\frac{\pi\lambda}{\sigma^2}\right)T-\frac{1}{2}\left[\gamma^2Tt-\gamma^2\frac{t^2}{2}\right]_0^T\\ &= -\frac{1}{2}\left(1+\ln\frac{\pi\lambda}{\sigma^2}\right)T-\frac{\gamma^2T^2}{4}\end{aligned}$$

Hence we get:

$$
\begin{aligned}
C^{\theta^*,\boldsymbol{f}^*}(0,v_0;w) &= J(0,v_0;w) - \lambda\mathbb{E}\left[-\frac{1}{2}\left(1+\ln\frac{\pi\lambda}{\sigma^2}\right)T - \frac{\gamma^2T^2}{4}\right] - J^{cl}(0,v_0;w) \\
&= e^{-\gamma^2T}(v_0-w)^2 + \frac{\lambda}{4}\gamma^2T^2 - \frac{\lambda}{2}\left(\gamma^2T - \ln\frac{\sigma^2}{\pi\lambda}\right)T - (w-\bar{v})^2 \\
&+ \frac{\lambda}{2}\left(1+\ln\frac{\pi\lambda}{\sigma^2}\right)T + \lambda\frac{\gamma^2T^2}{4} - (v_0-w)^2e^{-\gamma^2T} + (w-\bar{v})^2 \\
&= \frac{\lambda}{2}\gamma^2T^2 - \frac{\lambda}{2}\gamma^2T^2 + \frac{\lambda}{2}\ln\frac{\pi\lambda}{\sigma^2}T + \frac{\lambda}{2}\ln\frac{\sigma^2}{\pi\lambda} + \frac{\lambda}{2}T \\
&= \frac{\lambda}{2}T.
\end{aligned}
$$

□

The exploration cost here depends linearly on only two parameters, which are the exploration weight $\lambda > 0$ and the investment horizon $T > 0$. The higher the exploration weight, the higher the exploration cost. However, we can see that the cost does not depend on the Lagrange multiplier $w = \frac{\bar{v}e^{\gamma^2T}-v_0}{e^{\gamma^2T}-1}$. Hence with higher target return $\bar{v}$, and therefore higher Lagrange multiplier, the cost is not increased.

4 Reinforcement learning algorithm

In this subsection we rigorously derive the reinforcement learning algorithm described in [5] from the theoretical foundations developed above. This algorithm learns the optimal value function without estimating of model parameters in advance. First, we state the "Policy Improvement Theorem", essential for estimating the optimal value function and the associated optimal Gaussian policy. The Lagrange multiplier is estimated using a self-correcting scheme based on stochastic approximation. [5]. This algorithm does not use any classical discrete-time MDP process by discretizing time and space since following that approach is unfortunately poorly generalizing to higher dimension with more risky assets due to the curse of dimensionality. Our algorithm is also not using any deep reinforcement learning as it has been proven that such algorithms for continuous-time and space are highly sensitive to hyper parameter tuning [21] and therefore suboptimal for our framework as we seek to easily be able to modify target returns.

4.1 Policy Improvement

In this subsection we develop the mathematical framework for estimating the value function as described above. In general, RL algorithms are built on a policy evaluation and policy improvement steps, which allow to optimize the value function while staying within the set of admissible policies and updating the associated policy properly. The Policy Improvement Theorem, developed here, is therefore guaranteeing convergence to an optimal value function.

Theorem 4.1 (Policy Improvement Theorem) [5]
Let $w \in \mathbb{R}$, $i \in \mathbb{N}$ and $\boldsymbol{f}^i = \boldsymbol{f}^i(\cdot;\cdot,\cdot,w)$ be an arbitrarily given admissible feedback control. Suppose that the corresponding value function $J^{\boldsymbol{f}^i}(\cdot,\cdot;w)$ is $C^{1,2}([0,T)\times\mathbb{R})\cap C^0([0,T]\times\mathbb{R})$ and satisfies $J^{\boldsymbol{f}^i}_{vv}(t,v;w) > 0$ for any $(t,v)\in[0,T)\times\mathbb{R}$. Suppose that the feedback control $\boldsymbol{f}^{i+1}$ defined by:

$$
\boldsymbol{f}^{i+1}(\theta;t,v,w) = f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}\frac{J^{\boldsymbol{f}^i}_v(t,v;w)}{J^{\boldsymbol{f}^i}_{vv}(t,v;w)},\,\frac{\lambda}{\sigma^2 J^{\boldsymbol{f}^i}_{vv}(t,v;w)}\right)}(\theta). \tag{80}
$$

is admissible. Then,

$$J^{\boldsymbol{f}^{i+1}}(t,v;w) \leq J^{\boldsymbol{f}^{i}}(t,v;w) \quad (t,v) \in [0,T] \times \mathbb{R} \tag{81}$$

Proof: Let us take some fixed $(t,v) \in [0,T] \times \mathbb{R}_{\geq 0}$. By assumption, the feedback control $\boldsymbol{f}^{i+1}$ is admissible. We consider the open-loop control strategy $f^{i+1} = \{f_s^{i+1}, s \in [t,T]\} = \{\boldsymbol{f}^{i+1}(\cdot; s, v), s \in [t,T]\}$ generated from $\boldsymbol{f}^{i+1}$ with the initial condition $V^{\boldsymbol{f}^{i+1}}(t) = v$. We denote the corresponding wealth process by $\{V^{\boldsymbol{f}^{i+1}}(s), s \in [t,T]\}$. We have by (31):

$$dV^{\boldsymbol{f}^{i+1}}(t) = \gamma\sigma\hat{\mu}(t, f^{i+1})dt + \sigma\sqrt{\hat{\mu}^2(t, f^{i+1}) + \hat{\sigma}^2(t, f^{i+1})}dW_t \tag{82}$$

Since we have that $J^{\boldsymbol{f}}(t,v;w)$ is twice differentiable with respect to v as well as continuous we get by Itô's lemma that:

$$\begin{aligned} dJ^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) = \Bigg[& J_t^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) + \gamma\sigma\hat{\mu}(s, f^{i+1}) J_v^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) \\ & + \frac{1}{2}\sigma^2 \big(\hat{\mu}^2(s, f^{i+1})) + \hat{\sigma}(s, f^{i+1})\big) J_{vv}^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) \Bigg] ds \\ & + \sigma\sqrt{\hat{\mu}^2(s, f^{i+1}) + \hat{\sigma}^2(s, f^{i+1})} J_v^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) dW_s. \end{aligned}$$

Integrating on both sides from t to T we get the following :

$$\begin{aligned} J^{\boldsymbol{f}^{i}}(T, V^{\boldsymbol{f}^{i+1}}(T)) - J^{\boldsymbol{f}^{i}}(t, V^{\boldsymbol{f}^{i+1}}(t)) = & \int_t^T J_t^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) + \gamma\sigma\hat{\mu}(s, f^{i}) J_v^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) \\ & + \frac{1}{2}\sigma^2 \left(\hat{\mu}^2(s, f^{i}) + \hat{\sigma}^2(s, f^{i})\right) J_{vv}^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) ds \\ & + \int_t^T \sigma\sqrt{\hat{\mu}^2(s, f^{i}) + \hat{\sigma}^2(t, f^{i})} J_v^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) dW_s \end{aligned}$$

Due to

$$\hat{\mu}^2(s, f^{i+1}) + \hat{\sigma}^2(s, f^{i+1}) = \int_{\mathbb{R}} \theta^2 \boldsymbol{f}^{i+1}(\theta, s, V^{\boldsymbol{f}^{i+1}}(s)) d\theta$$

it follows

$$\begin{aligned} J^{\boldsymbol{f}^{i}}(T, V^{\boldsymbol{f}^{i+1}}(T)) - J^{\boldsymbol{f}^{i}}(t, V^{\boldsymbol{f}^{i+1}}(t)) = & \int_t^T J_t^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) ds \\ & + \gamma\sigma \int_t^T \int_{\mathbb{R}} \theta \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s)) J_v^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) d\theta ds \\ & + \int_t^T \int_{\mathbb{R}} \frac{1}{2}\sigma^2\theta^2 \boldsymbol{f}^{i+1}(\theta, s, V^{\boldsymbol{f}^{i+1}}(s)) J_{vv}^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) d\theta ds \\ & + \int_t^T \sigma \left(\int_{\mathbb{R}} \theta^2 \boldsymbol{f}^{i+1}(\theta, s, V^{\boldsymbol{f}^{i+1}}(s)) d\theta \right)^{\frac{1}{2}} J_v^{\boldsymbol{f}^{i}}(s, V^{\boldsymbol{f}^{i+1}}(s)) dW_s \end{aligned}$$

and since $V^{\boldsymbol{f}^{i+1}}(t) = v$:

$$\begin{aligned} J^{\boldsymbol{f}^i}(T, V^{\boldsymbol{f}^{i+1}}(T)) =& J^{\boldsymbol{f}^i}(t,v) \\ &+ \int_t^T J_t^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^{i+1}}(s))ds + \gamma\sigma \int_t^T \int_{\mathbb{R}} \theta \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s)) J_v^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^{i+1}}(s)) d\theta ds \\ &+ \int_t^T \int_{\mathbb{R}} \frac{1}{2}\sigma^2\theta^2 \boldsymbol{f}^{i+1}(\theta, s, V^{\boldsymbol{f}^{i+1}}(s)) J_{vv}^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^{i+1}}(s)) d\theta ds \\ &+ \int_t^T \sigma \left(\int_{\mathbb{R}} \theta^2 \boldsymbol{f}^{i+1}(\theta, s, V^{\boldsymbol{f}^{i+1}}(s)) d\theta \right)^{\frac{1}{2}} J_v^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^{i+1}}(s)) dW_s \end{aligned} \tag{83}$$

We define the stopping time by the following stopping times for $n \geq 1$

$$\tau_n := \inf\{s \geq t : \int_t^s \sigma^2 \int_{\mathbb{R}} \theta^2(\theta, u, V^{\boldsymbol{f}^{i+1}}(u)) d\theta (J^f(u, V^{\boldsymbol{f}^{i+1}}(u)))^2 du \geq n\}$$

. With this stopping time, the stochastic process:

$$\hat{M}_s = \int_t^{\min(\tau_n, s)} \left(\int_{\mathbb{R}} \theta^2 \boldsymbol{f}^{i+1}(\theta, u, V^{\boldsymbol{f}^{i+1}}(u)) d\theta \right)^{\frac{1}{2}} J^{\boldsymbol{f}^i}(u, V^{\boldsymbol{f}^{i+1}}(u)) dW_u$$

is adapted to the filtration $\mathbb{F}$ and is square integrable in expectation as :

$$\int_t^{\min(\tau_n, s)} \sigma^2 \int_{\mathbb{R}} \theta^2(\theta, u, V^{\boldsymbol{f}^{i+1}}(u)) d\theta (J^f(u, V^{\boldsymbol{f}^{i+1}}(u)))^2 du < \infty$$

by definition. Hence, $\hat{M}_s$ is a Martingale and taking the expectation in (83) we find :

$$\begin{aligned} J^{\boldsymbol{f}^i}(t,v) =& \mathbb{E}\Bigg[J^{\boldsymbol{f}^i}\left(\min(s,\tau_n), V^{\boldsymbol{f}^{i+1}}(\min(s,\tau_n))\right) \\ &- \int_t^{\min(s,\tau_n)} J_t^{\boldsymbol{f}^i}\left(u, V^{\boldsymbol{f}^{i+1}}(u)\right) du \\ &- \int_t^{\min(s,\tau_n)} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}^{\boldsymbol{f}^i}(u, V^{\boldsymbol{f}^{i+1}}(u)) + \gamma\sigma\theta J_v^{\boldsymbol{f}^i}(u, V^{\boldsymbol{f}^{i+1}}(u)) \right) \boldsymbol{f}^{i+1}(\theta, u, V^{\boldsymbol{f}^{i+1}}(u)) d\theta\, du \\ &\mid V^{\boldsymbol{f}^{i+1}}(t)) = v \Bigg] \end{aligned} \tag{84}$$

since $\mathbb{E}[\hat{M}_s \mid V^{\boldsymbol{f}^{i+1}}(t) = v] = 0$. Since we have that $J^{\boldsymbol{f}^i}$ and $V^{\boldsymbol{f}^i}$ satisfy by Definition (39) and (82), we can apply the Feynman-Kac formula on $J^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^i}(u))$:

$$\begin{aligned} J_t^{\boldsymbol{f}^i}(s,v) + \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}^{\boldsymbol{f}^i}(s,v) + \gamma\sigma\theta J_v^{\boldsymbol{f}^i}(u,v) + \lambda \ln \boldsymbol{f}^i(\theta; s, v) \right) \\ \times \boldsymbol{f}^i(\theta; s, v) d\theta = 0 \end{aligned}$$

for all $(s,v) \in [0,T) \times \mathbb{R}$. Taking the minimum over all probability density functions f' over $\mathbb{R}$ we get:

$$J_t^{\boldsymbol{f}^i}(s,v) + \min_{f' \in \mathcal{P}(\mathbb{R})} \left(\int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}^{\boldsymbol{f}^i}(s,v) + \gamma\sigma\theta J_v^{\boldsymbol{f}^i}(s,v) + \lambda \ln f'(\theta) \right) f'(\theta) d\theta \right) \leq 0 \tag{85}$$

As we see in the proof of Theorem 3.1, the density function minimizing the expression above is exactly the density function $f' = \boldsymbol{f}^{i+1}(\cdot; u, V^{\boldsymbol{f}^{i+1}}(u), w)$. Hence we get:

$$-\int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}^{\boldsymbol{f}^i}(s,v) + \gamma\sigma\theta J_v^{\boldsymbol{f}^i}(s,v) + \lambda \ln \boldsymbol{f}^{i+1}(\theta; s, v, w)\right) \boldsymbol{f}^{i+1}(\theta; s, v, w) d\theta \\ \geq J_t^{\boldsymbol{f}^i}(s,v).$$

Integrating from t to $\min(\tau_n, T)$ and taking the expectation with initial condition $V^{\boldsymbol{f}^{i+1}}(t) = v$ we have that:

$$\mathbb{E}\Bigg[-\int_t^{\min(\tau_n,T)} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^{i+1}}(s)) + \gamma\sigma\theta J_v^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^{i+1}}(s)) + \lambda \ln \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w)\right) \\ \times \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w) d\theta ds \mid V^{\boldsymbol{f}^{i+1}}(t) = v \Bigg] \\ \geq \\ \int_t^{\min(\tau_n,T)} J_t^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^{i+1}}(s)) ds$$

for $(t, v) \in [0, T] \times \mathbb{R}$. We rearrange terms to get:

$$\mathbb{E}\Bigg[-\int_t^{\min(\tau_n,T)} J_t^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^{i+1}}(s)) ds - \int_t^{\min(\tau_n,T)} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J_{vv}^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^i}(s)) + \gamma\sigma\theta J_v^{\boldsymbol{f}^i}(s, V^{\boldsymbol{f}^i}(s))\right) \\ \times \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w) d\theta ds \mid V^{\boldsymbol{f}^{i+1}}(t) = v \Bigg] \\ \geq \\ \mathbb{E}\left[\lambda \int_t^{\min(\tau_n,T)} \int_{\mathbb{R}} \ln \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w) \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w) d\theta ds \mid V^{\boldsymbol{f}^{i+1}}(t) = v \right].$$

Hence adding $J^{\boldsymbol{f}^i}\left(\min(T, \tau_n), V^{\boldsymbol{f}^{i+1}}(\min(T, \tau_n))\right)$ on both sides we find by the result in (84):

$$J^{\boldsymbol{f}^i}(t, v) \\ \geq \\ \mathbb{E}\Bigg[J^{\boldsymbol{f}^i}\left(\min(T, \tau_n), V^{\boldsymbol{f}^{i+1}}(\min(T, \tau_n))\right) \\ + \lambda \int_t^{\min(\tau_n,T)} \int_{\mathbb{R}} \ln \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w) \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w) d\theta ds \mid V^{\boldsymbol{f}^{i+1}}(t) = v \Bigg] \tag{86}$$

Taking $n \to \infty$ and since $|J(t,v)| < \infty$ we can apply the dominated convergence theorem by the above inequality and we get that:

$$\begin{aligned}&\mathbb{E}\Bigg[J^{\boldsymbol{f}^{\boldsymbol{i}}}\Big(\min(T,\tau_n), V^{\boldsymbol{f}^{i+1}}(\min(T,\tau_n))\Big)\\ &\qquad + \lambda \int_t^{\min(\tau_n,T)} \int_{\mathbb{R}} \ln \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w)\boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w) d\theta ds \mid V^{\boldsymbol{f}^{i+1}}(t) = v\Bigg]\\ &\qquad \to\\ &\mathbb{E}\Bigg[J^{\boldsymbol{f}^{\boldsymbol{i}}}\Big(T, V^{\boldsymbol{f}^{i+1}}(T)\Big)\\ &\qquad + \lambda \int_t^{T} \int_{\mathbb{R}} \ln \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w)\boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w) d\theta ds \mid V^{\boldsymbol{f}^{i+1}}(t) = v\Bigg]\end{aligned}$$

which gives in conclusion for all $(t,v) \in [0,T] \times \mathbb{R}$

$$\begin{aligned}&J^{\boldsymbol{f}^{i}}(t,v)\\ &\quad \geq\\ &\quad \mathbb{E}\Bigg[J^{\boldsymbol{f}^{\boldsymbol{i}}}\Big(T, V^{\boldsymbol{f}^{i+1}}(T)\Big)\\ &\quad + \lambda \int_t^{T} \int_{\mathbb{R}} \ln \boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w)\boldsymbol{f}^{i+1}(\theta; s, V^{\boldsymbol{f}^{i+1}}(s), w) d\theta ds \mid V^{\boldsymbol{f}^{i+1}}(t) = v\Bigg]\\ &\quad = J^{\boldsymbol{f}^{i+1}}(t,v)\end{aligned} \quad . \tag{87}$$

□

Theorem 4.2 [5] *Let $\boldsymbol{f}_0(\theta; t, v, w) = f_{\mathcal{N}\left(a(v-w), c_1 e^{c_2(T-t)}\right)}(\theta)$ with $a, c_2 \in \mathbb{R}$ and $c_1 > 0$. Denote $\{\boldsymbol{f}_n(\theta; t, , v, w), (t,v) \in [0;T] \times \mathbb{R}, n \geq 1\}$ the sequence of feedback controls (or policies) generated by the policy improvement in (80) and by $\{J^{\boldsymbol{f}_n}(t, v; w), (t,v) \in [0,T] \times \mathbb{R} n \geq 1\}$ the sequence of value functions corresponding those feeback controls. We then have:*

$$\lim_{n\to\infty} \boldsymbol{f}_n(\cdot, t, v, w) = f^*(\cdot, t, v, w) \textit{ weakly} \tag{88}$$

and

$$\lim_{n\to\infty} J^{\boldsymbol{f}_n}(t, v; w) = J(t, v; w) \tag{89}$$

for any $(t,v,w) \in [0,T] \times \mathbb{R} \times \mathbb{R}$, where $\boldsymbol{f}^*$ and J are the optimal Gaussian policy and the optimal value function respectively.

Proof: The feedback control $\boldsymbol{f}_0(\theta; t, v, w) = f_{\mathcal{N}(a(v-w), c_1 e^{c_2(T-t)})}(\theta)$ is generating an open-loop policy $f_0 = \{\boldsymbol{f}_0(\cdot; t, v, w), t \in [0,T]\}$ that is admissible with respect to the initial condition (t,v) as we have that $\boldsymbol{f}_0(\cdot; t, v, w) \in \mathcal{P}(\mathbb{R})$ for all $t \in [0,T]$ by definition. It also satisfies that for $\mathcal{A} \in \mathcal{B}(\mathbb{R})$, $\int_{\mathcal{A}} \boldsymbol{f}_0(\theta; t, v, w) d\theta$ is progressively measurable with respect to the filtration $\mathbb{F}$. Conditions (iii) and (iv) from Definition 2.9 (Admissibility of controls) are satisfied by the properties of a probability density function of Gaussian distributions. We have therefore that by the Feynman-Kac formula the value function $J^{\boldsymbol{f}_0}$ satisfies the Hamilton-Jacobi-Bellman equation and it follows that:

$$J^{\boldsymbol{f}_0}(t, v; w) + \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\theta^2 J^{\boldsymbol{f}_0}_{vv}(t, v; w) + \gamma\sigma\theta J^{\boldsymbol{f}_0}(t, v; w) + (\lambda \ln \boldsymbol{f}_0(\theta; t, v, w))\boldsymbol{f}_0(\theta; t, v, w)\right) d\theta = 0 \tag{90}$$

with a terminal condition $J^{\boldsymbol{f}_0}(T,v,w) = (v-w)^2 - (w-\bar{v})^2$. Similarly to (54) we assume the value function to satisfy :

$$J^{\boldsymbol{f}_0}(t,v;w) = l_0(t)(v-w)^2 + c_0(t)$$

with $l_0 : [0,T] \to \mathbb{R}$ and $c_0 : [0,T] \to \mathbb{R}$ continuous and differentiable functions. In the same way as in the proof of Theorem 3.1 we have:

$$J^{\boldsymbol{f}_0}(t,v;w) = (v-w)^2 e^{(2\gamma\sigma a+\sigma^2a^2)(T-t)} + c_0(t)$$

We have $J^{\boldsymbol{f}_0}_{vv}(t,v;w) = 2e^{(2\gamma\sigma a+\sigma^2 a^2)(T-t)} > 0$.
Moreover, we see that the function $J^{\boldsymbol{f}_0} \in C^{1,2}([0,T) \times \mathbb{R}) \cap C^0([0,T] \times \mathbb{R})$. Since $\boldsymbol{f}_0$ is admissible then Theorem 4.1 is applicable and we have that :

$$\boldsymbol{f}_1(\theta;t,v,w) = f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}\frac{J^{\boldsymbol{f}_0}_v(t,v;w)}{J^{\boldsymbol{f}_0}_{vv}(t,v;w)}, \frac{\lambda}{\sigma^2 J^{\boldsymbol{f}_0}_{vv}(t,v;w)}\right)}(\theta),$$

is a suitable policy improvement. Moreover, since we have that $J^{\boldsymbol{f}_0}_v(t,v;w) = 2(v-w)e^{(2\gamma\sigma a+\sigma^2a^2)(T-t)}$ and $J^{\boldsymbol{f}_0}_{vv}(t,v;w) = 2e^{(2\gamma\sigma a+\sigma^2a^2)(T-t)}$, we get :

$$\boldsymbol{f}_1(\theta;t,v,w) = f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}(v-w), \frac{\lambda}{2\sigma^2 e^{(2\gamma\sigma a+\sigma^2a^2)(T-t)}}\right)}(\theta),$$

Repeating the argument from above we have that the associated value function is of the form:

$$J^{\boldsymbol{f}_1}(t,v;w) = (v-w)^2 e^{-\gamma^2(T-t)} + c_1(t)$$

with c_1 a smooth function with respect to t. We get for $(t,v) \in [0,T] \times \mathbb{R}$:

$$J^{\boldsymbol{f}_1}_v(t,v;w) = 2(v-w)e^{-\gamma^2(T-t)} \tag{91}$$

$$J^{\boldsymbol{f}_1}_{vv}(t,v;w) = 2e^{-\gamma^2(T-t)} \tag{92}$$

We apply Theorem 4.1 again and find that the feedback control $\boldsymbol{f}_2(\theta,t,v,w)$ is given by :

$$\boldsymbol{f}_2(\theta;t,v,w) = f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}(v-w), \frac{\lambda}{2\sigma^2 e^{-\gamma^2(T-t)}}\right)}(\theta)$$

for all $\theta \in \mathbb{R}$. By Theorem 3.1, $\boldsymbol{f}_2$ is the optimal Gaussian policy and the associated value function is directly:

$$J^{\boldsymbol{f}_2}(t,v,w) = J(t,v,w) \tag{93}$$

for all $(t,v) \in [0,T] \times \mathbb{R}$. We have therefore that $\boldsymbol{f}_2(\theta;t,v,w) = \boldsymbol{f}^*(\theta;t,v,w)$ for all $\theta \in \mathbb{R}$ because there is no strict improvement anymore for the Gaussian for $n \geq 2$. We therefore have that for any $\varphi : \mathbb{R} \to \mathbb{R}$ continuous and bounded, we can apply the dominated convergence theorem (using the same argument as in the proof of Theorem 3.4):

$$\lim_{n\to\infty} \int_{\mathbb{R}} \varphi(\theta) f_n(\theta;t,w)d\theta = \int_{\mathbb{R}} \lim_{n\to\infty} \varphi(\theta) f_n(\theta;t,w)d\theta = \int_{\mathbb{R}} \varphi(\theta) f_2(\theta;t,w)d\theta \tag{94}$$

. This proves weak convergence of the probability measure given by $f_n(\cdot;t,v,w)$ to optimal Gaussian probability measure $f^*(\cdot;t,v,w)$ for all $(t,v) \in [0,T] \times \mathbb{R}$. This also means that the associated value function is optimal for $n \geq 2$. Hence, the two desired convergences are proven. □

4.2 The EMV algorithm

The general structure of the EMV alogrithm consists of three steps : Policy Evaluation, Policy Improvement, and finally estimation of the Lagrange multiplier w (which is called a self-correcting scheme [5]).

Policy Evaluation Step The objective is to learn the value function $J^{\boldsymbol{f}}$ under any given admissible feedback control $\boldsymbol{f}$ for $s \in [t,T]$ and for all $(t,v) \in [0,T] \times \mathbb{R}$

$$J^{\boldsymbol{f}}(t,v) = \mathbb{E}\left[J^{\boldsymbol{f}}(s,V(s)) + \lambda \int_t^s \int_{\mathbb{R}} \boldsymbol{f}(\theta,\tau)\ln \boldsymbol{f}(\theta,\tau)d\theta d\tau \mid V(t)=v\right] \tag{95}$$

Rearranging the parts this is equivalent to:

$$\mathbb{E}\left[\frac{J^{\boldsymbol{f}}(s,V(s)) - J^{\boldsymbol{f}}(t,V(t))}{s-t} + \frac{\lambda}{s-t}\int_t^s \int_{\mathbb{R}} \boldsymbol{f}(\theta,\tau)\ln \boldsymbol{f}(\theta,\tau)d\theta d\tau \mid V(t)=v\right] = 0$$

and we define:

$$\delta_t := \frac{J^f(t+\Delta t, V(t+\Delta t)) - J^f(t,V(t))}{\Delta t} + \lambda \int_{\mathbb{R}} f(\theta,t)\ln f(\theta,t)d\theta$$

and

$$\dot{J}^{\boldsymbol{f}}(t) = \frac{J^{\boldsymbol{f}}(t+\Delta t, V(t+\Delta t)) - J^{\boldsymbol{f}}(t,V(t))}{\Delta t}$$

This is the quotient of differences in the value function for discretizations Δt (note that the quantity δ_t is often described as temporal difference error or TD error in RL literature) [19]. We want to minimize this value in order to approximate the function $J^{\boldsymbol{f}}$ as accurately as possible. In order to minimize this quantity, we parametrize J and $\boldsymbol{f}$ using a vector of weights $\kappa = (\kappa_0,\kappa_1,\kappa_2,\kappa_3)'$ and $\psi = (\psi_0,\psi_1)'$ to be learned. The accumulated Bellman error on $[0,T]$ is then given by:

$$C(\kappa,\psi) = \frac{1}{2}\mathbb{E}\left[\int_0^T |\delta_t|^2 dt\right] = \frac{1}{2}\mathbb{E}\left[\int_0^T \left|\dot{J}^{\kappa}(t) + \lambda\int_{\mathbb{R}} \boldsymbol{f}^{\psi}(\theta,t)\ln \boldsymbol{f}^{\psi}(\theta,t)d\theta\right|^2 dt\right] \tag{96}$$

Here $f^{\psi} = \{\boldsymbol{f}^{\psi}(\cdot,t), t \in [0,T]\}$ is generated with respect to a given initial state $V(0) = \bar{v}$ In order to approximate $C(\kappa,\psi)$ we discretize the continuous time space $[0,T]$ into l intervals $[t_i;t_{i+1}]$ with $t_0 = 0$ and $t_{l+1} = T$ and collect the samples $\mathcal{D} = \{(t_i,v_i), i = 0,1,..,l+1\}$ where v_i is the wealth at time t_i. To generate the $(v_i)_i$, we take an initial sample $(0,v_0)$ and then iterate by applying for each t_i the policy $f^{\psi}(t_i)$ and sample an allocation θ_i which enables us to generate v_{i+1}. This can be done by using the return of the market simulator at t_i alongside with the allocation θ_i in the market. In mathematical terms we have as in section 2 using the self-financing argument [24] that (with the notation from section 2) and $\theta(t) = \varphi_1(t)\tilde{P}(t)$:

$$\begin{aligned}
dV(\varphi,t) &= \varphi_0(t)dB(t) + \varphi_1(t)dP(t) \\
&= \varphi_0(t)B(t)\frac{dB(t)}{B(t)} + \varphi_1(t)P(t)\frac{dP(t)}{P(t)} \\
&= \varphi_0(t)B(t)rdt + \theta(t)B(t)\frac{dP(t)}{P(t)} \\
&= (V(\varphi,t) - B(t)\theta(t))rdt + \theta(t)B(t)\frac{dP(t)}{P(t)}
\end{aligned}$$

Moreover, we have seen above that the discounted wealth $\tilde{V}(\varphi, t)$ satisfies:

$$\begin{aligned} d\tilde{V}(\varphi,t) &= -re^{-rt}V(\varphi,t)dt + e^{-rt}dV(\varphi,t) \\ &= -r\frac{V(\varphi,t)}{B(t)}dt + \frac{1}{B(t)}\left((V(\varphi,t) - B(t)\theta(t))rdt + \theta(t)B(t)\frac{dP(t)}{P(t)}\right) \\ &= -r\tilde{V}(\varphi,t)dt + (\tilde{V}(\varphi,t) - \theta(t))rdt + \theta(t)\frac{dP(t)}{P(t)} \\ &= \theta(t)(\frac{dP(t)}{P(t)} - rdt). \end{aligned}$$

Discretizing the above expression yields :

$$\tilde{V}(\varphi, t+\Delta t) \approx \tilde{V}(\varphi,t) + \theta(t)\left(\frac{P(t+\Delta t) - P(t)}{\Delta t} - r\Delta t\right) \tag{97}$$

and this therefore yields the following iterative expression with $v_i = \tilde{V}(\varphi, t_i)$:

$$v_{i+1} \leftarrow v_i + \theta_i \times (\text{market simulation return} \times -r\Delta t) \tag{98}$$

We get therefore the following approximation similar to a Monte Carlo method:

$$C(\kappa,\psi) = \frac{1}{2}\sum_{(t_i,v_i)\in\mathcal{D}}\left(\dot{J}^{\kappa}(t_i,v_i) + \lambda\int_{\mathbb{R}} f^{\psi}(\theta,t_i)\ln f^{\psi}(\theta,t_i)d\theta\right)^2 \Delta t \tag{99}$$

Motivated by the proof of theorem 4.2 we want to look at Gaussian feedback policies $\boldsymbol{f}^{\psi}(\cdot;\cdot,\cdot)$ with associated distribution having a variance of the form $c_0e^{c_1(T-t)}$. We compute the entropy $\mathcal{H}(\boldsymbol{f}^{\psi}(\cdot;t))$ of such a Gaussian policy as:

$$\begin{aligned} H(\boldsymbol{f}^{\psi}(\cdot;t)) &= \frac{1}{2}\ln\left(2\pi ec_0e^{c_1(T-t)}\right) \\ &= \frac{1}{2}\ln(2\pi ec_0) + \frac{1}{2}c_1(T-t) \\ &= \psi_0 + \psi_1(T-t), \end{aligned}$$

where $\psi_0 := \frac{1}{2}\ln(2\pi ec_0)$ and $\psi_1 := \frac{1}{2}c_1 > 0$.

Policy Improvement Step Moreover, we approximate the optimal value function in Theorem 3.1:

$$\begin{aligned} J(t,v;w) &= e^{-\gamma^2(T-t)}(v-w)^2 + \frac{\lambda}{4}\gamma^2(T^2-t^2) - \frac{\lambda}{2}\left(\gamma^2T - \ln\frac{\sigma^2}{\pi\lambda}\right)(T-t) - (w-\bar{v})^2 \\ &= J^{\kappa}(t,v) = (v-w)^2e^{-\kappa_3(T-t)} + \kappa_2t^2 + \kappa_1t + \kappa_0 \end{aligned}$$

with the vector $\kappa = (\kappa_0, \kappa_1, \kappa_2, \kappa_3)'$ to be learned. We have that J^{κ} is twice differentiable with respect to v.

$$\begin{aligned} J^{\kappa}_v(t,v) &= 2(v-w)e^{-\kappa_3(T-t)} \\ J^{\kappa}_{vv}(t,v) &= 2e^{-\kappa_3(T-t)} \end{aligned}$$

We have therefore by the Policy Improvement Theorem :

$$\boldsymbol{f}^{\psi}(\theta;t,v,w) = f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}\frac{J^{\kappa}_v(t,v;w)}{J^{\kappa}_{vv}(t,v;w)},\frac{\lambda}{\sigma^2J^{\kappa}_{vv}(t,v;w)}\right)}(\theta) \tag{100}$$

$$= f_{\mathcal{N}\left(-\frac{\gamma}{\sigma}(v-w),\frac{\lambda}{2\sigma^2}e^{\kappa_3(T-t)}\right)}(\theta) \tag{101}$$

and computing the entropy we get:

$$\begin{aligned}\mathcal{H}(\boldsymbol{f}^{\psi}(\cdot,t,w)) &= \frac{1}{2}\ln\left(2\pi e\frac{\lambda}{2\sigma^2}e^{\kappa_3(T-t)}\right)\\ &= \frac{1}{2}\ln\frac{\pi e\lambda}{\sigma^2}+\frac{\kappa_3}{2}(T-t).\end{aligned}$$

Hence we get:

$$\begin{aligned}\psi_0 &= \frac{1}{2}\ln\frac{\pi e\lambda}{\sigma^2},\\ \psi_1 &= \frac{\kappa_3}{2}.\end{aligned}$$

This yields that :

$$\sigma^2 = \lambda\pi e^{1-2\psi_0} \text{ and } \kappa_3 = 2\psi_1 = \gamma^2 \tag{102}$$

Due to $\gamma > 0$ and $\sigma > 0$ we can rewrite the mean and variance in equation (101) as:

$$\begin{aligned}-\frac{\gamma}{\sigma}(v-w) &= -\sqrt{\frac{\gamma^2}{\sigma^2}}(v-w)\\ &= -\sqrt{\frac{2\psi_1}{\lambda\pi e^{1-2\psi_0}}}(v-w)\\ &= -\sqrt{\frac{2\psi_1}{\lambda\pi}}e^{\frac{1-2\psi_0}{2}}(v-w)\end{aligned}$$

and:

$$\begin{aligned}\frac{\lambda}{2\sigma^2}e^{\kappa_3(T-t)} &= \frac{\lambda}{2\lambda\pi e^{1-2\psi_0}}e^{2\psi_1(T-t)}\\ &= \frac{1}{2\pi}e^{2\psi_1(T-t)+2\psi_0-1}\end{aligned}$$

Collecting the results we find:

$$f^{\psi}(\theta;t,v,w) = f_{\mathcal{N}\left(-\sqrt{\frac{2\psi_1}{\lambda\pi}}e^{\frac{1-2\psi_0}{2}}(v-w),\frac{1}{2\pi}e^{2\psi_1(T-t)+2\psi_0-1}\right)}(\theta) \tag{103}$$

Determining the Bellman error in this parametrization yields:

$$C(\kappa,\psi) = \frac{1}{2}\sum_{(t_i,v_i)\in\mathcal{D}}\left(\dot{J}^{\kappa}(t_i,v_i)-\lambda\mathcal{H}(f^{\psi}(t_i))\right)^2\Delta t \tag{104}$$

$$= \frac{1}{2}\sum_{(t_i,v_i)\in\mathcal{D}}\left(\dot{J}^{\kappa}(t_i,v_i)-\lambda(\psi_0+\psi_1(T-t_i))\right)^2\Delta t \tag{105}$$

We now use a form of gradient descent to minimize $C(\gamma, \psi)$[5]:

$$\frac{\partial C}{\partial \kappa_1} = \sum_{(t_i,v_i)\in\mathcal{D}} \Big(\dot{J}^\kappa(t_i, v_i) - \lambda(\psi_0 + \psi_1(T - t_i)) \Big) \Delta t, \tag{106}$$

$$\frac{\partial C}{\partial \kappa_2} = \sum_{(t_i,v_i)\in\mathcal{D}} \Big(\dot{J}^\kappa(t_i, v_i) - \lambda(\psi_0 + \psi_1(T - t_i)) \Big) (t_{i+1}^2 - t_i^2), \tag{107}$$

$$\frac{\partial C}{\partial \psi_0} = -\lambda \sum_{(t_i,v_i)\in\mathcal{D}} \Big(\dot{J}^\kappa(t_i, v_i) - \lambda(\psi_0 + \psi_1(T - t_i)) \Big) \Delta t, \tag{108}$$

$$\frac{\partial C}{\partial \psi_1} = \sum_{(t_i,v_i)\in\mathcal{D}} \Big(\dot{J}^\kappa(t_i, v_i) - \lambda(\psi_0 + \psi_1(T - t_i)) \Big) \Delta t \tag{109}$$

$$\times \left(-\frac{2(v_{i+1} - w)e^{-2\psi_0(T - t_{i+1})}(T - t_{i+1}) - 2(v_i - w)^2 e^{-2\psi_1(T - t_i)}(T - t_i)}{\Delta t} - \lambda(T - t_i) \right). \tag{110}$$

Moreover, we have that since $J^\kappa(T, v; w) = (v - w)^2 - (w - \bar{v})^2$ we get therefore:

$$(v - w)^2 - (w - \bar{v})^2 = (v - w)^2 + \kappa_2 T^2 + \kappa_1 T + \kappa_0$$
$$\kappa_0 = -\kappa_2 T^2 - \kappa_1 T - (w - \bar{v})^2$$

We get therefore the following update rules for the parameters $\kappa = (\kappa_0, \kappa_1, \kappa_2, \kappa_3)'$ and $\psi = (\psi_0, \psi_1)'$:

$$\psi_0 \leftarrow \psi_0 - \eta_\psi \frac{\partial C}{\partial \psi_0}, \tag{111}$$

$$\psi_1 \leftarrow \psi_1 - \eta_\psi \frac{\partial C}{\partial \psi_1}, \tag{112}$$

$$\kappa_1 \leftarrow \kappa_1 - \eta_\kappa \frac{\partial C}{\partial \kappa_1}, \tag{113}$$

$$\kappa_2 \leftarrow \kappa_2 - \eta_\kappa \frac{\partial C}{\partial \kappa_2}, \tag{114}$$

$$\kappa_3 \leftarrow 2\psi_1, \tag{115}$$

$$\kappa_0 \leftarrow -\kappa_2 T^2 - \kappa_1 T - (w - \bar{v})^2 \tag{116}$$

Here we define the hyper parameters η_κ and η_ψ as the learning rates for the gradient descent algorithm where $C(\kappa, \psi)$ is differentiated with respect to κ and ψ.

Estimation of the Lagrange multiplier In order to find an update rule for the Lagrange multiplier w, we look at the original constraint of the problem: $\mathbb{E}[V(T)] = \bar{v}$. This clearly hints at the stochastic approximation problem in which we want to find the root of $l(w) = \mathbb{E}[V(T)] - \bar{v}$. We therefore have that we can use this value for:

$$w \leftarrow w - \alpha(V(T) - \bar{v}) \tag{117}$$

In our algorithm we follow the same scheme as in the paper from Wang 2020 [5] and use instead of V(T) the sample average $\frac{1}{N}\sum_j v_T^j$ where v_T^j are the last sample values of the wealth.

The python implementation of the EMV alogrithm can be seen in the Appendix (2).

5 Simulation and important results

In this section we are going to test the EMV algorithm on a market simulation and analyze its convergence.

5.1 Market parameters and hyperparameter tuning

We are conducting the market simulation using the log-return of a stock price which follows a geometric Brownian motion as in (9) with annual drift $\mu = 10\%$ and volatility $\sigma = 20\%$ (stationary market scneario). Using Itô's lemma on (9) we can derive that :

$$P(t) = P(0)e^{(\mu-\frac{1}{2}\sigma^2)t+\sigma dW_t} \quad \forall t \in [0,T]x \tag{118}$$

We implement this in Python (Appendix (1)) where this provides the data to train our algorithm. We test this algorithm during 100 trading days which means $\Delta t = \frac{1}{252}$ and $T = \frac{100}{252}$. In addition we allow the agent to invest in a riskless asset with interest rate at 2% annually. In order to produce interpretable results, the initial wealth is set to $v_0 = 1$ and the annualized target return is set to 80% of the annual drift μ. In particular the target return is adapted to the time horizon. In our case this gives a target return of approximately 3%:

$$\begin{aligned}\bar{v} &= 1 + 0.8\mu T \\ &= 1 + 0.8 \times 0.1 \times \frac{100}{252} \\ &\approx 1.03\end{aligned}$$

This target return is higher than the riskless interest on the 100 days trading period $1 + \frac{100}{252} \times 0.02 \approx 1,008$, and therefore corresponds to a reasonable objective as an agent who wants to maximize the return of its investments. We initialize the values of ψ and κ as $(0.05, 0.1)'$ and $\big(-(v_0 - \bar{v})^2, -0.1, -0.05, 0.2\big)'$ respectively after testing for the stock market modeled by the above Geometric Brownian Motion. Additionally, we set $\alpha = 0.05$, $\eta_\kappa = 0.0000005$ and $\eta_\psi = 0.0000005$. We fix also the number of iterations to learn the parameters κ and ψ to $M = 4000$ with sample size 25. This mean that the algorithm will repeat 4000 times the trading of the asset during those 100 days to learn the optimal Gaussian. After every 25 iterations, the algorithm will update the Lagrange multiplier w and will store the variance as well as the mean of the payoff over the latest 25 iterations (Appendix (2)). The tuning of the EMV-algorithm is not trivial as it has to be adapted to different market scenarios even when different stock prices are modeled with Brownian Motions having identical annual drift and volatility. In the meantime, the parameters η_κ and η_ψ are set relatively low in order to keep ψ_1 positive in equation (103). This is because, the gradient descent algorithm (113) in many market-scenarios tend to decrease the value of ψ_1 leading to a negative parameter and making the EMV algorithm obsolete. This issue can be avoided by resampling the allocation θ_i randomly until the generated ψ_1 stays positive. However, this adds in the meantime additional time complexity. Keeping the parameters η_κ and η_ψ small also avoids entries all entries in the vectors κ and ψ to explode (in absolute value) to large numbers and alter the efficiency of the algorithm. Additionally, the tuning has to be made so that learning of parameters (κ, ψ and w) enables fast "enough" convergence to the target return while maintaining "low enough" variance before the computer reaches machine precision ($\approx 10^{-30}$) for those parameters [17]. Hence, reducing η_κ and η_ψ towards "too" low values can alter satisfying convergence even though theoretical convergence is guaranteed (Theorem 4.2).

5.2 Influence of the exploration rate on the EMV algorithm

The main interest of this thesis is to investigate the idea of adding exploration to a classical mean-variance optimization problem. It is therefore particularly interesting to investigate the influence of different exploration rates on the behavior of the EMV algorithm under the parameters given above and grasp the value of adding exploration. The EMV algorithm is tested here on two different exploration rates $\lambda = 1$ and $\lambda = 3$ (which we denote as "low" and "high" respectively). We observe that with a low exploration rate $\lambda = 1$ on 100 days of trading, the target return is achieved (Figure (1)), while the variance stays low and descreases slightly throughout the learning process (Figure (2)). This can be seen, as the value-function J which represents the trade-off between mean, variance and exploration also stays low and decreases slightly throughout the learning of the parameters κ and ψ (Figure (3)). Hence, for a low exploration rate the mean-variance problem is solved as it reaches the target while minimizing variance. However, with a higher value of exploration rate ($\lambda = 3$), we observe that the variance is not as low as in the previous parametrization, leading to higher values of J and suboptimal solution for the mean-variance problem. In the meantime the frequency of the terminal payoff for the low exploration is tightly distributed around the target payoff(Figure (4)). However, for $\lambda = 3$ exploration rate, allows better results as the mean terminal payoffs is achieved with higher frequency (Figure (5)) even though extreme values have also higher frequencies. Hence, balancing high and low exploration allows to keep variance under a certain threshold while optimizing allocation to get the target return with a higher frequency.

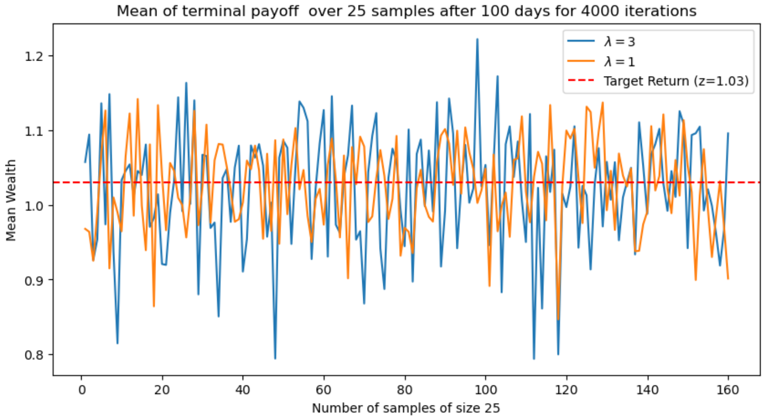

Figure 1: For stationary market scenario $\mu = 10\%$ and $\sigma = 20\%$

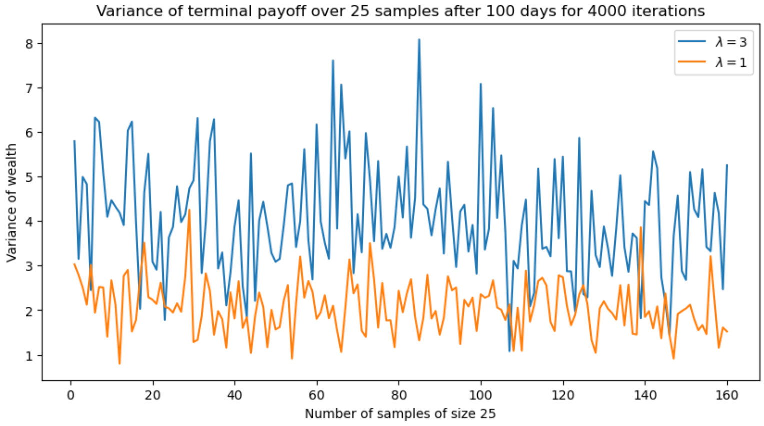

Figure 2: For non-stationary market scenario $\mu = 10\%$ and $\sigma = 20\%$

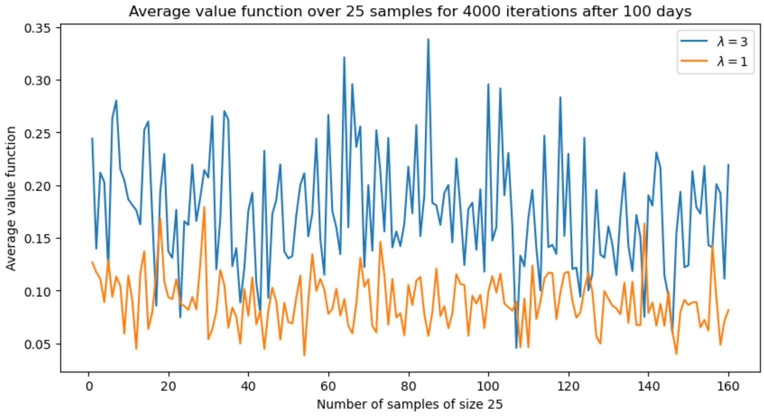

Figure 3: For stationary market scenario $\mu = 10\%$ and $\sigma = 20\%$

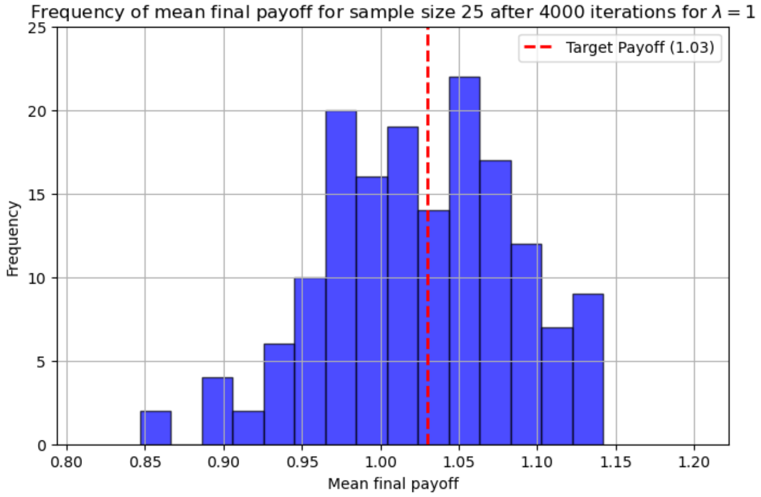

Figure 4: For stationary market scenario $\mu = 10\%$ and $\sigma = 20\%$

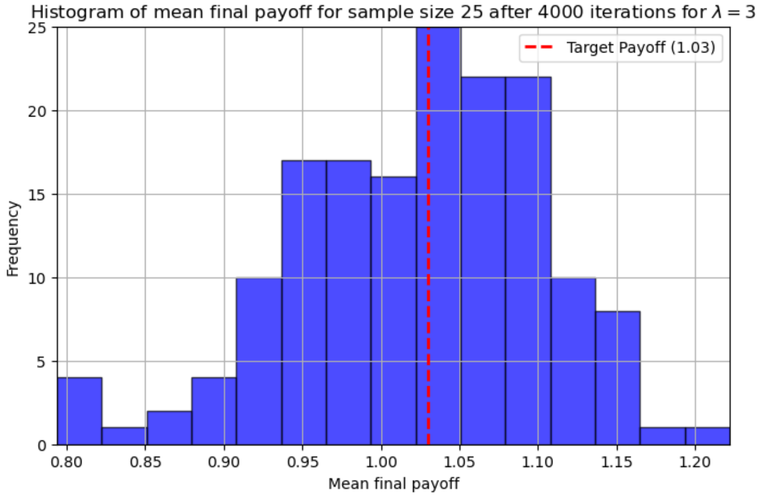

Figure 5: For stationary market scenario $\mu = 10\%$ and $\sigma = 20\%$

5.3 Advantage of the EMV algorithm

If proper tuning is achieved, the training time of this algorithm is fast with respect to other computational methods. For the learning of κ and ψ for the two exploration rates $\lambda = 1$ and $\lambda = 3$, it only takes approximately 20 minutes on a Mackbook Air Laptop M2. This is lower than MLE methods (maximum likelihood estimation) which often require hours of training time for estimating market parameters [5] with similar GPU power. Moreover, the tuning is relatively easier than models using deep-reinforcement which are highly sensitive to hyper-parameter changes [8]. For the EMV algorithm changing the values of the learning rates η_κ and η_ψ does not critically change the convergence of the algorithm but rather alters it. Another positive aspect of the EMV algorithm is its stability. As we see in figure (2), the variance of the terminal payoff remains low around the terminal payoff and decreases consistently for a well-tuned value of lambda. Even with high exploration, after 4000 iterations the frequency around the target payoff is the highest.

6 Conclusion

In this thesis, we derived carefully the mathematical foundations of the reinforcement learning algorithm for solving the continuous-time mean-variance portfolio problem. We see that by introducing entropy-regularization to the classical mean-variance problem and randomizing allocation strategy enables the mathematical derivation of an expression which can be used as value function under feedback policies. Those feedback policies are used to generate gaussian distributions which govern allocation of investment. Applying Itô calculus and probability theory, it is possible to derive convergence of feedback policies to an optimal feedback policy and, hence, to an optimal gaussian distribution. Using those results we are able to design an algorithm based on optimizing exploration against exploitation while skipping the burden of parameter estimation. Knowing the general mathematical expression of the optimal value function and its gaussian, we can use parametrized expression of them. This parametrized value function and gaussian, can be learned through minimization of the Bellman error via gradient descent. Even though hyper parameter tuning is not easy for this algorithm depending on the simulation of the market, the EMV algorithm can give convergence results within relatively short training time and stable behavior.

7 References

[1] R.E. Bellman. Dynamic programming. *Princeton University Press*, 1957.

[2] Shannon C. E. A mathematical theory of communication. *Bell System Technical Journal*, 1948.

[3] Jaynes E.T. Information theory and statistical mechanics. *Brandeis University Summer Institute Lectures in Theoretical Phycis*, 106(4):sect. 4b, 1963.

[4] Gerald B. Folland. Real analysis: Modern techniques and their applications. *Pure and Applied Mathematics: A Willey-Interscience Series of Texts, Monographs, and Tracts*, 1999.

[5] Wang H and Zhou XY. Continuous-time mean-variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30(4):1273–1308, 2020.

[6] Thaleia Zariphopoulou Haoran Wang and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34, 2020.

[7] Karatzas I. and S. Shreve. Brownian motion and stochastic calculus. *Springer*, Calculus.

[8] Duan Y. Chen X. Houthooft R. Schulman J and Abbeel P. Benchmarking deep reinforcement learning for continuous control. *International Conference on Machine Learning*, pages 1329–1338, 2016.

[9] Jean Jacod and Philip Protter. Probability essentials. *Springer-Verlag Berlin Heidelberg GmbH*, 2004.

[10] Achim Klenke. Probability theory: A comprehensive course. *Universitext*, page Chapter 13, 2006.

[11] D.G Luenberger. Investment science. *New York: Oxford Univerisity Press*, 1998.

[12] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[13] Péter Medvegyev. Stochastic processes: A very simple introduction. page 24, 2009.

[14] J. Moody and M. Safell. Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889, 2001.

[15] R. Munos and P. Bourgine. Reinforcement learning for continuous stochastic control problems. *Advances in Neural Information Processing Systems*, pages 1029–1035, 1998.

[16] Y. Nevmyvaka, Y. Feng, and M. Kearns. Reinforcement learning for optimized trade executions. *Proceedings of the 23rd International Conference on Machine Learning*, pages 673–680, 2006.

[17] Hao Shen Martin Gottwald Tianming Qiu and Stephan Rappensperger. Applied dynamic programming and reinforcement learning. *Technical University of Munich*, 2024.

[18] Martin Stefanik. Mathematical foundations for finance. *ETH Zürich*.

[19] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

[20] Pieter Abbeel Tuomas Haarnoja, Aurick Zhou and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *Cornell University*, 2018.

[21] Koray Kavukcuoglu ... Shane Legg Volodymyr Mnih and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.

[22] Zhou Yong Jiongmin and Xun Yu. Dynamic programming and hjb equations. *Stochastic Controls Hamiltonian Systems and Hamilton-Jacobi-Bellman Equations, Springer*, 1999.

[23] Rudi Zagst. Interest rate management. *Springer Finance*, pages 36–37, 2002.

[24] Rudi Zagst. Investment strategies. *Technical University of Munich*, 2021.

A Appendix

```
# Parameters
T = 2           # Total years
N = 252 * T     # Number of trading days
dt = 1 / 252    # Time step (daily)
mu = 0.10       # Annual drift (10%)
sigma = 0.20    # Annual volatility (20%)
P0 = 1          # Initial price

# Generate Brownian motion
dW = np.random.normal(0, np.sqrt(dt), N)  # Brownian increments

# Compute the price process using the GBM formula
P = np.zeros(N+1)
P[0] = P0
for t in range(1, N+1):
    P[t] = P[t-1] * np.exp((mu - 0.5 * sigma**2) * dt + sigma * dW[t-1])

# Convert to numpy array
P_array = np.array(P)
log_returns = np.diff(np.log(P_array))
market_log_returns = (log_returns - log_returns.mean()) / log_returns.std()
```

Listing 1: Simulating a stock price using Geometric Brownian Motion

```
import numpy as np
import random

def emv_portfolio_selection(
    market_simulator, learning_rates, initial_wealth, target_payoff,
    investment_horizon, discretization_dt, exploration_rate,
    num_iterations, sample_average_size, interest_rate
):
    # Extract parameters
    alpha, eta_kappa, eta_psi = learning_rates
    v_0 = initial_wealth
    z = target_payoff
    T, delta_t = investment_horizon, discretization_dt
    lamb = exploration_rate
    M = num_iterations
    N = sample_average_size
    r= interest_rate #yearly interest rate

    #initialization of parameters (tuning as best as possible)
    psi = [0.05,0.1]
    kappa = [-(v_0 - z)**2, -0.1, -0.05, 2 * psi[1]]

    w= v_0

    psi_policy = psi

   #helper lists
    D_final = []
    D_final_pi = []
    D_final_value = []

    #return lists
    pay_off_mean = []
    pay_off_mean_non_agg = []
    variance_off_sample = []
    value_sample = [] #value function
    ratio_risky_asset_sample = []
    ratio_risky_over_time = []


    for k in range(1, M + 1):
        D = [(0, v_0)]  # Collected samples as tuples (time, wealth)
        D_pi = [0] #Collected samples as tuples (time, risky_allocation)
        D_value = [compute_V(kappa, psi, v_0, w, 0, T)]
        for i in range(1, int(T / delta_t)):

            #simulate the market
            t_k, v_k, theta_k = simulate_market(market_simulator,psi_policy, D, w, T,
    lamb,delta_t,r)

            pi_k = theta_k * np.exp(r*t_k)/v_k

            value_function_k = compute_V(kappa, psi, v_k, w, t_k, T)

            #sample collection
            D.append((t_k, v_k))

            #ratio_risky_asset_collection
            D_pi.append(pi_k)

            #collection of value_function at t_k
            D_value.append(value_function_k)

            # Compute Bellman error
            delta_t_error = compute_bellman_error(kappa, psi, D, w, T, lamb, delta_t)

            #Compute gradient (Equations 106-110)
            grad_kappa_1 = compute_grad_kappa_1(kappa, psi, D, w, T, lamb, delta_t)
            grad_kappa_2 = compute_grad_kappa_2(kappa, psi, D, w, T, lamb, delta_t)
            grad_psi_0   = compute_grad_psi_0(kappa, psi, D, w, T, lamb, delta_t)
            grad_psi_1   = compute_grad_psi_1(kappa, psi, D, w, T, lamb, delta_t)

            # Update parameters (Equations 111- 116)
            kappa[1] -= eta_kappa * grad_kappa_1
            kappa[2] -= eta_kappa * grad_kappa_2
            kappa[3] = 2 * psi[1]
            kappa[0] = -kappa[2] * T**2 - kappa[1] * T - (w - z)**2
            #update psi
            psi[0]    -= eta_psi * grad_psi_0
            psi[1]    -= eta_psi * grad_psi_1

        #update the policy_psi
        psi_policy[0] = psi[0]
        psi_policy[1] = psi[1]

        #get all the values of x^j_{T/delta} for j in 1 to M
        D_final.append(D[-1][1])

        #get all the values of pi^j_{T/delta}
        D_final_pi.append(D_pi[-1])

        #get all the values of V_j(T/delta) for j  in 1 to M
        D_final_value.append(D_value[-1])


        if k == M:
            ratio_risky_over_time = D_pi

        # Update Lagrange multiplier every N iterations (Equation 52)
        if k % N == 0:
            #to plot aggregated mean
            pay_off_mean.append(np.mean([v for v in D_final]))
            recent_samples = D_final[k-N+1:k+1]
            recent_pi_samples = D_final_pi[k-N+1:k+1]
            recent_value_samples = D_final_value[k-N+1:k+1]

            #compute mean of risky allocations
            average_terminal_allocation = np.mean([v for v in recent_pi_samples])

            #compute mean of value functions
            average_terminal_value_function = np.mean([v for v in recent_value_samples])

            #compute mean and variance over sample size
            average_terminal_wealth = np.mean([v for v in recent_samples])
            variance = sum([(v-average_terminal_wealth)**2 for v in recent_samples])

            #append mean of risky allocations
            ratio_risky_asset_sample.append(average_terminal_allocation)

            #append mean of value functions
            value_sample.append(average_terminal_value_function)


            #append mean and variance over sample
            pay_off_mean_non_agg.append(average_terminal_wealth)
            variance_off_sample.append(variance)

            #update of the Lagrange mutiplier omega
            w -= alpha * (average_terminal_wealth - target_payoff)

    return kappa, psi, w, pay_off_mean, pay_off_mean_non_agg, variance_off_sample,
    ratio_risky_asset_sample, ratio_risky_over_time, value_sample

def policy_psi(psi_policy, v,t, T,w, lamb): #(checked)
    mean = -np.sqrt(2 * psi_policy[1] / (lamb * np.pi))* np.exp((2*psi_policy[0]-1)/2) *
    ((v - w))
    variance = (1 / (2 * np.pi)) * np.exp(2 * psi_policy[1]* (T-t) +  (2 * psi_policy[0]
    - 1))
    return np.random.normal(mean,variance)

def simulate_market(market_simulator, psi_policy, D, w, T, lamb, delta_t,r):
    t_k, v_k = D[-1] #get the last sample
    theta_k = policy_psi(psi_policy, v_k, t_k, T,w,lamb)
    #market return going in

    #(Equation 98)
    dv = market_simulator[int(252*t_k) + 1] #change when going to t_k+1

    #one computes the next wealth
    v_k = v_k + theta_k*(dv - r*delta_t)
    t_k  += delta_t
    return t_k, v_k, theta_k

def compute_bellman_error(kappa, psi, D, w, T, lamb, delta_t): #(checked)
```

```
    C = 0
    for i in range(len(D) - 1):
        t_i   = D[i][0]
        t_i_1 = D[i + 1][0]  # next time step
        v_i   = D[i][1]
        v_i_1 = D[i + 1][1]
        V_t   = compute_V(kappa, psi, v_i, w, t_i, T)
        V_t_1 = compute_V(kappa, psi, v_i_1, w, t_i_1, T)
        V_dot = (V_t_1 - V_t) / delta_t
        entropy = psi[0] + psi[1] * (T - t_i)
        C += (V_dot - lamb * entropy)**2 * delta_t
    return C / 2

def compute_V(kappa, psi, v, w, t, T): #(checked)
    arg = -kappa[3] * (T - t)
    V_gamma = (v - w)**2 * np.exp(arg) + kappa[2] * t**2 + kappa[1] * t + kappa[0]
    return V_gamma

def compute_grad_kappa_1(kappa, psi, D, w, T, lamb, delta_t):
    C = 0
    for i in range(len(D) - 1):
        t_i   = D[i][0]
        t_i_1 = D[i + 1][0]
        v_i   = D[i][1]
        v_i_1 = D[i + 1][1]
        V_t   = compute_V(kappa, psi, v_i, w, t_i, T)
        V_t_1 = compute_V(kappa, psi, v_i_1, w, t_i_1, T)
        V_dot = (V_t_1 - V_t) / delta_t
        entropy = psi[0] + psi[1] * (T - t_i)
        C += (V_dot - lamb * entropy) *delta_t
    return C

def compute_grad_kappa_2(kappa, psi, D, w, T, lamb, delta_t):
    C = 0
    for i in range(len(D) - 1):
        t_i   = D[i][0]
        t_i_1 = D[i + 1][0]
        v_i   = D[i][1]
        v_i_1 = D[i + 1][1]
        V_t   = compute_V(kappa, psi, v_i, w, t_i, T)
        V_t_1 = compute_V(kappa, psi, v_i_1, w, t_i_1, T)
        V_dot = (V_t_1 - V_t) / delta_t
        entropy = psi[0] + psi[1] * (T - t_i)
        C += (V_dot - lamb * entropy) * (t_i_1**2 - t_i**2)
    return C

def compute_grad_psi_0(kappa, psi, D, w, T, lamb, delta_t):
    C = 0
    for i in range(len(D) - 1):
        t_i   = D[i][0]
        t_i_1 = D[i + 1][0]
        v_i   = D[i][1]
        v_i_1 = D[i + 1][1]
        V_t   = compute_V(kappa, psi, v_i, w, t_i, T)
        V_t_1 = compute_V(kappa, psi, v_i_1, w, t_i_1, T)
        V_dot = (V_t_1 - V_t) /delta_t
        entropy = psi[0] + psi[1] * (T - t_i)
        C += (-lamb) * (V_dot - lamb * entropy) * delta_t
    return C

```

```python
def compute_grad_psi_1(kappa, psi, D, w, T, lamb, delta_t):
    C = 0
    for i in range(len(D) - 1):
        t_i   = D[i][0]
        t_i_1 = D[i + 1][0]
        v_i   = D[i][1]
        v_i_1 = D[i + 1][1]
        V_t   = compute_V(kappa, psi, v_i, w, t_i, T)
        V_t_1 = compute_V(kappa, psi, v_i_1, w, t_i_1, T)
        V_dot = (V_t_1 - V_t) / delta_t
        entropy = psi[0] + psi[1] * (T - t_i)
        arg1 = -2 * psi[1] * (T - t_i_1)
        arg2 = -2 * psi[1] * (T - t_i)
        exp_factor_1 = np.exp(arg1)
        exp_factor_2 = np.exp(arg2)

        gradient_term = (2 * (v_i_1 - w)**2 * exp_factor_1 * (T - t_i_1) -
                         2 * (v_i - w)**2 * exp_factor_2 * (T - t_i)) / delta_t

        C += (V_dot - lamb * entropy) *  (-gradient_term - lamb * (T - t_i)) * delta_t

    return C
```

Listing 2: EMV Portfolio Selection Algorithm