ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D MATH

On a reinforcement learning approach to an exploratory version of mean-variance portfolio selection

Master Thesis

written by

Tim Gyger

Department of Mathematics
ETH Zürich

Supervisor:
Prof. Dr. Martin Schweizer

July 25, 2022

Abstract

This thesis studies an exploratory version of the continuous-time mean–variance (MV) portfolio selection in a reinforcement learning (RL) framework. The problem is formulated as an entropy-regularized, relaxed stochastic control problem. We compare its solution and solvability to the corresponding classical version of the MV problem. Furthermore, based on a policy improvement theorem, we present an RL algorithm introduced by H. Wang & X. Y. Zhou in [25] to learn an optimal strategy. Finally, we analyze its performance in a simulation and an empirical study.

Contents

Chapter 1

Introduction

Portfolio optimization is one of the main challenges in asset management, which mainly focuses on allocating the investor's wealth among different assets according to the investor's preference, indicated by some optimization objective. The mean-variance (MV) preference is one of the most popular criteria in portfolio optimization. The classical MV portfolio selection theory, introduced by H. Markowitz [19] in 1952, marks the beginning of modern finance. Ever since, various extensions and versions have been presented but all capturing a trade-off between risk (variance) and reward (mean).

Despite the MV criterion's intuitive nature, applying MV analysis has major challenges. One issue is the time-inconsistency inherent in the underlying stochastic optimal control problem. It prevents the direct use of the dynamic programming technique, and the concept of optimality becomes problematic. X. Y. Zhou & D. Li in [27] restricted the goal upon finding the precommitted optimal strategies and showed that under these circumstances, the continuous-time MV problem is a special instance of a stochastic linear-quadratic (LQ) control problem, which can be solved analytically by dynamic programming methods.

Another frequent criticism addressed to the MV model is its sensitivity to the perturbations of the market model parameters. Since, in practice, the accurate identification of these investment opportunity parameters is a difficult task, incorrect values are adopted regularly, and the optimal portfolio generated by the MV problem is not very reliable, which in practice results in poor out-of-sample performance.

Reinforcement learning (RL) techniques can often avoid estimating model parameters. Instead, RL algorithms output optimal allocations by direct interactions with the unknown investment environment in a learning (exploring) while optimizing (exploiting) fashion. To transfer the MV problem in an RL setting, H. Wang, T. Zariphopoulou & X. Y. Zhou in [26] proposed and developed an exploratory formulation of the LQ problem as a general entropy-regularized, relaxed stochastic control problem. In this formulation, the agent emphasizes exploring

the black-box environment by using distributions of controls to randomize her actions. Based on this, H. Wang & X. Y. Zhou in [25] formulated an exploratory version of the MV problem in a finite time horizon. They showed that the optimal feedback control policy for this problem is a Gaussian distribution with a time-decaying variance and a perfect separation between exploitation and exploration in the mean and variance of the distribution. Furthermore, based on a policy improvement theorem (PIT), they provide an implementable RL algorithm to learn the optimal solution of the exploratory MV (EMV) problem.

In this thesis, we present in Chapter 2 a continuous-time MV problem with one risky asset that yields an asset allocation strategy that minimizes the variance of the final payoff while targeting some prespecified mean return. To use dynamic programming methods, we restrict ourselves to precommitted strategies. Then, we derive the optimal feedback control of the resulting LQ problem. In Chapter 3, we state the exploratory MV problem, and similar to the previous chapter, we derive the optimal feedback control law. Further, in Chapter 4, we establish ties between the MV and the EMV problem. For this purpose, we state the solvability equivalence between the two problems and analyze the cost of exploration. In Chapter 5, we state a PIT that provides an explicit updating scheme for the feedback control law. Moreover, we show that if we use a suitable initial control law, this updating rule converges to the optimal feedback control law in a finite number of iterations. Based on these results, we derive the EMV algorithm introduced by H. Wang & X. Y. Zhou in [25]. In the next step, we analyze in Chapter 6 the training performance of this RL algorithm in a simulation study. Furthermore, to backtest the EMV algorithm in an empirical study, we first consider the multi-dimensional setting in Chapter 7 and present all necessary statements and derive the multi-dimensional EMV algorithm. Lastly, we analyze the out-of-sample performance of the EMV algorithm on random sets of multiple S&P 500 stocks in Chapter 8 and compare the results to other baseline approaches.

Chapter 2

Continuous-time Mean-Variance Portfolio Selection

This chapter proposes and solves a constrained version of the continuous-time MV portfolio selection problem. First, we introduce our basic market setup and some fundamental concepts of mathematical finance. Based on these, we state a formulation of the MV problem in continuous time that yields an investment strategy that minimizes the variance of the final payoff while targeting some prespecified mean return. Next, we transfer the continuous-time MV portfolio selection problem into an unconstrained stochastic linear-quadratic (LQ) control problem. Further, we present a simple general LQ problem and discuss roughly the procedure for solving it. In the last step, we solve our problem extensively with the same procedure. This chapter follows mainly the detailed explanations in H. Wang & X. Y. Zhou [25], X. Li et al. [17] and T. Björk et al. [7] Chapter 11 & 15.

2.1 Financial Market Model

For the sake of simplicity, throughout this work (except for Chapter 7), we consider an investment market with only one risky asset and one risk-free asset. The procedure and results in the case of multiple risky assets pose no essential differences or difficulties other than notational complexity.

We fix the time horizon $T \in (0, \infty)$ and consider a filtered probability space $\big(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P}\big)$ with the filtration $\mathbb{F} := (\mathcal{F}_t)_{t\in[0,T]}$ and which satisfies the usual conditions (see Appendix A.1). One should interpret the σ-algebra $\mathcal{F}_t$ as the information available to investors at time t.

Besides, we define a standard, one-dimensional Brownian motion $B = (B_t)_{t\in[0,T]}$ on our filtered probability space (see Appendix B.2). We assume that the price of the risky asset follows a real-valued, continuous, strictly positive, and ($\mathbb{F}$-)adapted stochastic process $S = (S_t)_{t\in[0,T]}$ satisfying the following stochastic differential

equation (SDE):

$$\begin{aligned} dS_t &= S_t(\mu dt + \sigma dB_t), \quad 0 \leq t \leq T \\ S_0 &= s_0 > 0, \end{aligned}$$

where S_0 is the initial price at $t = 0$, $\mu \in \mathbb{R}$ and $\sigma > 0$. Therefore, the price process S follows a geometric Brownian motion with drift parameter μ and volatility parameter σ.

Furthermore, we assume that the price of the risk-free asset $S^0 = (S^0_t)_{t\in[0,T]}$ with an initial price $S^0_0 = 1$ is continuous, strictly positive, and modeled through a constant interest rate $r > 0$:

$$\begin{aligned} dS^0_t &= S^0_t r dt, \quad 0 \leq t \leq T \\ S^0_0 &= 1. \end{aligned}$$

The discounted price of the risky asset is denoted by $\tilde{S} = (\tilde{S})_{t\in[0,T]}$ and defined as:

$$\tilde{S}_t := \frac{S_t}{S^0_t}, \quad 0 \leq t \leq T.$$

An application of Itô's formula (see Appendix B.3) shows that:

$$\begin{aligned} d\tilde{S}_t &= \tilde{S}_t\big((\mu - r)dt + \sigma dB_t\big), \quad 0 \leq t \leq T \\ \tilde{S}_0 &= s_0 > 0. \end{aligned}$$

Moreover, we can define the Sharpe ratio of the risky asset by $\rho = \frac{\mu - r}{\sigma}$ which represents the average return earned in excess of the risk-free rate per unit of volatility. Therefore, we can write the dynamics of the discounted price in the following way:

$$d\tilde{S}_t = \sigma\tilde{S}_t\big(\rho dt + dB_t\big), \quad 0 \leq t \leq T. \tag{2.1}$$

We consider an investor (or agent) who decides to hold $\vartheta_t \in \mathbb{R}$ shares of the non-risky asset and $\Delta_t \in \mathbb{R}$ shares of the risky asset at time $t \geq 0$. We assume that both processes $\vartheta = (\vartheta_t)_{t\in[0,T]}$ and $\Delta = (\Delta_t)_{t\in[0,T]}$ are ($\mathbb{F}$-)predictable, meaning that investors can and do use all available information to construct their trading strategies. The associated value of the portfolio (wealth process) $x^{\vartheta,\Delta} = (x^{\vartheta,\Delta}_t)_{t\in[0,T]}$ is given by:

$$x^{\vartheta,\Delta}_t := \vartheta_t S^0_t + \Delta_t S_t, \quad t \in [0,T] \tag{2.2}$$

and we say that the portfolio is self-financing if:

$$dx^{\vartheta,\Delta}_t = r\vartheta_t S^0_t dt + \Delta_t dS_t, \quad 0 \leq t \leq T. \tag{2.3}$$

It means that the instantaneous variation of the wealth process depends only on the returns generated by the investment in the risky assets and at the risk-free rate r. By using (2.2), we can rewrite the dynamics in (2.3) only in terms of the process $(\Delta_t)_{t\in[0,T]}$ and an initial endowment (capital) $x_0^{\vartheta,\Delta} = x_0 \in \mathbb{R}$:

$$dx_t^{\vartheta,\Delta} = (rx_t^{\vartheta,\Delta} - r\Delta_t S_t)dt + \Delta_t dS_t, \quad 0 \le t \le T.$$

Therefore, we use from now on $(x_t^{\Delta})_{t\in[0,T]}$ instead of $(x_t^{\vartheta,\Delta})_{t\in[0,T]}$.

By using Itô's formula and (2.1), we compute the dynamics of a discounted wealth process:

$$d\tilde{x}_t^{\Delta} = \sigma\Delta_t\tilde{S}_t\big(\rho dt + dB_t\big) = \Delta_t d\tilde{S}_t, \quad 0 \le t \le T.$$

We define the (trading) strategy as the discounted dollar value put in the risky asset and denote it by $\boldsymbol{u} = (u_t)_{t\in[0,T]}$ where $u_t := \Delta_t\tilde{S}_t \in \mathbb{R}$. We assume that $\boldsymbol{u}$ is also predictable. In this way, the discounted wealth process with respect to the strategy $\boldsymbol{u}$ which we denote from now on by $x^{\boldsymbol{u}} = (x_t^{\boldsymbol{u}})_{t\in[0,T]}$ satisfies:

$$\begin{aligned} dx_t^{\boldsymbol{u}} &= \sigma u_t(\rho dt + dB_t), \quad 0 \le t \le T \\ x_0^{\boldsymbol{u}} &= x_0. \end{aligned} \tag{2.4}$$

For the exact definitions of an adapted and predictable stochastic process we refer to Appendix B.1.

2.2 Formulation of the MV Problem

On the base of the wealth processes (2.4) controlled by a strategy $\boldsymbol{u}$, we can now state the continuous-time MV problem as in [25]:

$$\begin{aligned} &\min_{\boldsymbol{u}} \operatorname{Var}(x_T^{\boldsymbol{u}}) \\ &\text{subject to } \mathbb{E}[x_T^{\boldsymbol{u}}] = z, \end{aligned} \tag{2.5}$$

where the wealth process $(x_t^{\boldsymbol{u}})_{t\in[0,T]}$ satisfies (2.4) under the trading strategy $\boldsymbol{u}$ and $z \in \mathbb{R}$ is an investment target set at time $t = 0$ as the desired mean payoff at the end of the investment horizon $[0, T]$.

Due to the term $\mathbb{E}[x_T^{\boldsymbol{u}}]^2$ appearing in the variance $\operatorname{Var}(x_T^{\boldsymbol{u}}) = \mathbb{E}[(x_T^{\boldsymbol{u}})^2] - \mathbb{E}[x_T^{\boldsymbol{u}}]^2$ in the objective function of (2.5), the MV problem is so-called time-inconsistent. In a decision theory sense, that means the problem does not allow Bellman's optimality principle and therefore, dynamic programming is not applicable for solving the problem (see Appendix C.1 for more details about Bellman's optimality principle, time-inconsistency and dynamic programming).

More generally, the objective function contains a term of the form $U(\mathbb{E}[x_T^{\boldsymbol{u}}])$ where U is a nonlinear utility function. This leads to the problem that the objective is not separable in the sense of dynamic programming. To be precise, for a term of the form $\mathbb{E}[U(x_T^{\boldsymbol{u}})]$, dynamic programming is applicable due to the so-called tower property (see Appendix C.2). However, for terms of the form $U(\mathbb{E}[x_T^{\boldsymbol{u}}])$ there is no such relation available and Bellman's optimality principle is not present. Consequently, the very concept of optimality becomes problematic since a strategy that is optimal given a specific starting point in time and space may be non-optimal when viewed from a later date and a different state. More intuitively, in the time-inconsistent case, an optimal strategy today may not be optimal tomorrow.

Since the usual dynamic programming approach fails, we have to find other ways to approach the MV problem (2.5).

There are different procedures for handling a family of time-inconsistent problems (see Section 1.4 of [7]). In this work, we focus ourselves, as in [25] on the precommitted strategies, which are optimal at $t = 0$ only. We disregard the fact that the strategy will not be optimal at a later point in time and view the problem as static.

If the goal is to find the precommitted optimal strategies, problem (2.5) becomes a special instance of a stochastic linear-quadratic (LQ) control problem (see X. Y. Zhou & D. Li [27]). In this case, the system dynamics are described by a linear SDE, and a quadratic function describes the cost or value. Moreover, methods of dynamic programming are applicable.

To transform (2.5) into an unconstrained LQ problem, we deal with the equality constraint $\mathbb{E}[x_T^{\boldsymbol{u}}] = z$ by applying a Lagrange multiplier similar as done by X. Li et al. in [17]. The method of Lagrange multipliers is a strategy for finding the local maxima or minima of a function subject to equality constraints (for more details about this method, we refer to D. P. Bertsekas [5]).

For this purpose, we form the Lagrangian function for problem (2.5):

$$\begin{aligned}\mathcal{L}(x_T^{\boldsymbol{u}}, \lambda) &= \operatorname{Var}(x_T^{\boldsymbol{u}}) - \lambda(\mathbb{E}[x_T^{\boldsymbol{u}}] - z) \\ &= \mathbb{E}[(x_T^{\boldsymbol{u}})^2] - \mathbb{E}[x_T^{\boldsymbol{u}}]^2 - \lambda(\mathbb{E}[x_T^{\boldsymbol{u}}] - z) \\ &= \mathbb{E}[(x_T^{\boldsymbol{u}})^2] - z^2 - \lambda(\mathbb{E}[x_T^{\boldsymbol{u}}] - z),\end{aligned}$$

where $\lambda \in \mathbb{R}$ is the so-called Lagrange multiplier.

By the change of variable $\lambda = 2w$, we obtain:

$$\begin{aligned}\mathcal{L}(x_T^{\boldsymbol{u}}, w) &= \mathbb{E}[(x_T^{\boldsymbol{u}})^2] - z^2 - 2w(\mathbb{E}[x_T^{\boldsymbol{u}}] - z) \\ &= \mathbb{E}[(x_T^{\boldsymbol{u}} - w)^2] - (w - z)^2.\end{aligned}$$

Further, we define the set of admissible strategies associated with (2.4):

Definition 2.1. For $x_0 \in \mathbb{R}$, we define the set of admissible strategies as:

$$\mathcal{A}(0, x_0) := \left\{ \boldsymbol{u} = (u_t)_{t\in[0,T]} : \begin{array}{c} \boldsymbol{u} \text{ is } \mathbb{F}\text{-progressively measurable} \\ \text{and } \mathbb{E}\left[\int_0^T (u_t)^2\, dt\right] < \infty \end{array} \right\}.$$

These properties ensure that the stochastic integral for $\boldsymbol{u}$ and the associated wealth process $x^{\boldsymbol{u}}$ is well-defined (see Section 3.2 in I. Karatzas & S. E. Shreve [15]). For the exact definitions of a ($\mathbb{F}$-)progressively measurable stochastic process and some relations to adapted and predictable processes, we refer to Appendix B.1.

Therefore, problem (2.5) transforms to the unconstrained problem:

$$\min_{\boldsymbol{u}\in\mathcal{A}(0,x_0)} \mathbb{E}[(x_T^{\boldsymbol{u}} - w)^2] - (w - z)^2, \tag{2.6}$$

where $w \in \mathbb{R}$ is the Lagrange multiplier and the portfolio wealth process $(x_t^{\boldsymbol{u}})_{t\in[0,T]}$ satisfies (2.4) under the investment strategy $\boldsymbol{u} \in \mathcal{A}(0, x_0)$.

We denote the optimal strategy for (2.6) by $\hat{\boldsymbol{u}} = (\hat{u}_t)_{t\in[0,T]}$, which depends on w. Then, the Lagrange multiplier w can be determined by using the original constraint $\mathbb{E}[x_T^{\hat{\boldsymbol{u}}}] = z$.

And to come full circle, the optimal strategy $\hat{\boldsymbol{u}}$ for problem (2.6) is optimal at $t = 0$ for our original MV problem (2.5) but not necessarily for any $t > 0$.

2.3 A general stochastic linear-quadratic Problem

The problem (2.6) is a special case of a stochastic LQ control problem. In this section, we present a more general stochastic LQ problem and discuss roughly the approach for solving it.

The basic setup is, that we have a Brownian motion $W = (W_t)_{t\in[0,T]}$ over $[0, T]$ on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, where $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$.

The problem to be solved is to choose an adapted $\mathbb{R}$-valued control process $\boldsymbol{u} = (u_t)_{t\in[0,T]}$ that minimizes:

$$\mathbb{E}\left[\int_0^T QX_t^2 + Ru_t^2 dt + HX_T^2\right], \tag{2.7}$$

given the state dynamics:

$$\begin{aligned} dX_t &= (AX_t + Bu_t + F)\,dt + Cu_t dW_t, \quad t \in [0,T] \\ X_0 &= x_0 \end{aligned} \tag{2.8}$$

and the constraint:

$$u_t \in U, \quad t \in [0,T],$$

where Q, R, H, A, B, F and $C \in \mathbb{R}$ are scalars, $x_0 \in \mathbb{R}$ is the initial state and $U \subseteq \mathbb{R}$.

The interpretation of the quadratic function (2.7) is that we use the terms Qx^2 and Hx^2 to bring the state close to the origin. The additional term Ru penalizes our inputs and thus allows us to factor in input constraints. The integral in (2.7) represents the running cost of the problem.

We are able to partly control the state process $X = (X_t)_{t\in[0,T]}$ in (2.8) by choosing the control process $\boldsymbol{u}$ in a suitable way. For this purpose, we only consider processes satisfying certain requirements, so-called admissible control processes. These requirements can vary according to the considered problem. In most cases, it is natural that at time t, the value u_t of the control process should only depend on observed values of the state process X up to time t.

Therefore, we consider the so-called feedback (or closed-loop) control laws. These are deterministic functions of the form $u : [0,T] \times \mathbb{R} \longrightarrow \mathbb{R}$ which take the output of the state process X one wants to control as input. More precise, the agent chooses at time t a control function $u(t,\cdot)$ such that the action taken at time t with state X_t is given by $u(t, X_t)$. Hence, the feedback control law u defines a control process $\boldsymbol{u} = (u_t)_{t\in[0,T]}$ by $u_t := u(t, X_t)$. We call a control process $\boldsymbol{u}$ also open-loop control law or just open-loop control.

To distinguish the open-loop and feedback controls in this work, we have used boldfaced $\boldsymbol{u}$ to denote open-loop controls, and the normal style u to denote feedback controls.

By inserting a fixed feedback control law $u = u(\cdot,\cdot)$ into (2.8), we obtain a standard SDE:

$$\begin{aligned} dX_t &= (AX_t + Bu(t,X_t) + F)\,dt + Cu(t,X_t)dW_t, \quad t \in [0,T] \\ X_0 &= x_0 \in \mathbb{R}. \end{aligned} \tag{2.9}$$

The process X satisfying (2.9) is called the state process induced by the feedback control law u.

To ensure that the constraint $u(t,x) \in U$ for each $(t,x) \in [0,T] \times \mathbb{R}$ is satisfied and that the SDE (2.9) has a unique solution, we define the class of admissible feedback controls:

Definition 2.2. A feedback control law u is called admissible if the following conditions hold:

1. $u(t,x) \in U$ for all $(t,x) \in [0,T] \times \mathbb{R}$.
2. For any given initial point $(s,y) \in [0,T] \times \mathbb{R}$, the SDE:

$$\begin{aligned} dX_t &= (AX_t + Bu(t,X_t) + F)\,dt + Cu(t,X_t)dW_t, \quad t \in [s,T] \\ X_s &= y \end{aligned}$$

   has a unique solution.
3. The value functional (2.7) is well-defined and finite.

We denote the class of admissible feedback control laws by $\mathcal{A}^{\mathrm{LQ}}$.

By considering feedback instead of open-loop control laws, we can rewrite our original problem (2.7)-(2.8) to that of minimizing the value given by:

$$J^{\mathrm{LQ}}(0,x_0;u) := \mathbb{E}\left[\int_0^T QX_t^2 + R\big(u(t,X_t)\big)^2 dt + HX_T^2 \Big| X_0 = x_0\right]$$

over all admissible feedback control laws $u \in \mathcal{A}^{\mathrm{LQ}}$. Hence, we define the optimal value by:

$$V^{\mathrm{LQ}}(0,x_0) := \inf_{u \in \mathcal{A}^{\mathrm{LQ}}} J^{\mathrm{LQ}}(0,x_0;u). \tag{2.10}$$

If a solution exists, a standard way of finding the optimal feedback control in (2.10) is by using dynamic programming. For this purpose, one defines the value function with respect to a given admissible feedback control $u \in \mathcal{A}^{\mathrm{LQ}}$ and the optimal value function for any initial point $(s,y) \in [0,T) \times \mathbb{R}$:

$$\begin{aligned} J^{\mathrm{LQ}}(s,y;u) &:= \mathbb{E}\left[\int_s^T QX_t^2 + R\big(u(t,X_t)\big)^2 dt + HX_T^2 \Big| X_s = y\right] \\ V^{\mathrm{LQ}}(s,y) &:= \inf_{u \in \mathcal{A}^{\mathrm{LQ}}} J^{\mathrm{LQ}}(s,y;u). \end{aligned}$$

Then, by using Bellman's optimality principle (see Appendix C.1), one is able to derive the so-called Hamilton-Jacobi-Bellman (HJB) equation which gives a

necessary and sufficient condition for optimality of a control with respect to a value function. It is, in general, a nonlinear partial differential equation satisfied by the value function.

The HJB equation for the LQ problem (2.7)-(2.8) is given by:

$$\begin{aligned}\frac{\partial v}{\partial t}(t,x) + \inf_{u\in\mathbb{R}}\left(\frac{1}{2}C^2u^2\frac{\partial^2 v}{\partial x\partial x}(t,x) + Ru^2 + \frac{\partial v}{\partial x}(t,x)\Big(Ax+Bu+F\Big)\right)\\ + Qx^2 = 0, \quad (t,x)\in[0,T)\times\mathbb{R}\\ v(T,x) = Hx^2.\end{aligned} \tag{2.11}$$

Since this equation is solved by the value function $V^{\mathrm{LQ}}(\cdot,\cdot)$ we end up with a static optimization problem, for each $(t,x)\in[0,T]\times\mathbb{R}$, of minimizing:

$$\frac{1}{2}C^2u^2\frac{\partial^2 V^{\mathrm{LQ}}}{\partial x\partial x}(t,x) + Ru^2 + \frac{\partial V^{\mathrm{LQ}}}{\partial x}(t,x)Bu. \tag{2.12}$$

Therefore, we obtain for each pair $(t,x)\in[0,T)\times\mathbb{R}$ a $u\in\mathbb{R}$ which results in a feedback control law $u(\cdot,\cdot)$ dependent on the derivatives of the value function V^{LQ}.

The HJB equation also acts as a sufficient condition for the optimal control problem (2.7)-(2.8). This result is known as the Verification theorem for dynamic programming. It says that if we find a function v that satisfies certain conditions including solving the HJB equation (2.11) and a function g that minimizes (2.12) for every $(t,x)\in[0,T)\times\mathbb{R}$ then v is equal to the optimal value function V^{LQ} and g is the optimal feedback control law.

For more information about the continuous-time LQ problem, the HJB equation, and the Verification theorem, we refer to Chapters 11 & 12 in T. Björk et al. [7].

By comparing the problem (2.6) to the general stochastic LQ problem (2.7)-(2.8), we observe that (2.6) is a particular case in the family of stochastic LQ problems, where the running cost is identically zero. However, it can be solved in the same manner as discussed.

2.4 Solution of the MV Problem

In this section, we present the solution to the problem (2.6) step-by-step. To do so, we proceed similarly as discussed in Section 2.3. This section is mainly based on the detailed explanations in Sections 11.5 - 11.7 in [7].

2.4.1 Value Functions

The idea of dynamic programming is to embed a problem in a family of problems indexed by time and space. For this purpose, we define the admissible strategies and the value function for arbitrary initial time and state.

We consider the restricted wealth process $x^{\boldsymbol{u}} = (x_t^{\boldsymbol{u}})_{t\in[s,T]}$ controlled by the strategy $\boldsymbol{u} = (u_t)_{t\in[s,T]}$ as in (2.4) but for an initial time $s \in [0,T)$ and an initial wealth $y \in \mathbb{R}$:

$$\begin{aligned} dx_t^{\boldsymbol{u}} &= \sigma u_t(\rho dt + dB_t), \quad s \le t \le T \\ x_s^{\boldsymbol{u}} &= y. \end{aligned} \tag{2.13}$$

Therefore, we can define the set of admissible strategies for arbitrary initial time and state:

Definition 2.3. For $(s,y) \in [0,T)\times\mathbb{R}$, we define the set of admissible strategies as:

$$\mathcal{A}(s,y) := \left\{ \boldsymbol{u} = (u_t)_{t\in[s,T]} : \begin{array}{l} \boldsymbol{u} \text{ is } \mathbb{F}\text{-progressively measurable} \\ \text{and } \mathbb{E}\big[\int_s^T (u_t)^2 dt\big] < \infty \end{array} \right\}.$$

As in the LQ problem setup, we call the admissible strategies in $\mathcal{A}(s,y)$ also admissible open-loop control laws or shorter admissible open-loop controls.

Furthermore, as mentioned in Section 2.3, we consider feedback controls, deterministic functions of the form $u : [0,T]\times\mathbb{R} \longrightarrow \mathbb{R}$ that can define strategies $\boldsymbol{u}$:

Definition 2.4. We denote the set of admissible feedback controls by $\mathcal{A}$. A deterministic mapping u is called an admissible feedback control if:

1. $u(t,x) \in \mathbb{R}$ for each $(t,x) \in [0,T]\times\mathbb{R}$.
2. For each $(s,y) \in [0,T)\times\mathbb{R}$ the following SDE:

$$\begin{aligned} dx_t^u &= \sigma u(t,x_t^u)(\rho dt + dB_t), \quad t \in [s,T] \\ x_s^u &= y \end{aligned}$$

has a unique strong solution $x^u = (x_t^u)_{t\in[s,T]}$ and the open-loop control $\boldsymbol{u} = (u_t)_{t\in[s,T]} \in \mathcal{A}(s,y)$ is admissible where $u_t := u(t,x_t^u)$.

In this case, the open-loop control $\boldsymbol{u}$ is said to be generated from the feedback control u with respect to the initial time and wealth (s,y) and x^u is the wealth

process induced by u.

Repeating Section 2.3, in a feedback control u, the control action at time t depends on feedback from the process x^u, in the form of the value of the process variable $x_t^u = x$, and not on some initial state as in the open-loop control. Therefore, a feedback control can generate an open-loop control for any initial $(s, y) \in [0, T) \times \mathbb{R}$ and hence, it is in itself independent of any initial values.

Now, we define the value and optimal value function for problem (2.6), which specify the value attained by the objective function under a given feedback control and under the optimal strategy, respectively:

Definition 2.5. For a fixed $w \in \mathbb{R}$ we define the value function under any given admissible feedback control $u(\cdot, \cdot; w) \in \mathcal{A}$ as:

$$J(s, y; w, u) := \mathbb{E}[(x_T^u - w)^2 | x_s^u = y] - (w - z)^2, \tag{2.14}$$

and the optimal value function as:

$$V(s, y; w) := \inf_{\boldsymbol{u} \in \mathcal{A}(s,y)} \mathbb{E}[(x_T^{\boldsymbol{u}} - w)^2 | x_s^{\boldsymbol{u}} = y] - (w - z)^2, \tag{2.15}$$

for $(s, y) \in [0, T) \times \mathbb{R}$.

We call the open-loop control in $\mathcal{A}(s, y)$ that minimizes (2.15) the optimal strategy for the initial (s, y) and denote it by $\hat{\boldsymbol{u}}^{s,y}$. It is generated from the optimal feedback control $\hat{u}^{s,y}$ with respect to the initial (s, y) such that $J(s, y; w, \hat{u}^{s,y}) = V(s, y; w)$.

However, by Bellman's optimality principle (see Appendix C.1), it turns out that the optimal feedback control law $\hat{u}^{s,y}$ is also optimal for any subinterval of the form $[t, T]$ where $s \leq t \leq T$ and hence $\hat{u}^{s,y}$ and $\hat{u}^{t,x_t^{\hat{u}}}$ coincide on $[t, T] \times \mathbb{R}$.

In particular, the optimal feedback control law for the initial point $t = 0$ will be optimal for all subintervals. This law will be denoted by $\hat{u}$, and we note that it does not have to be unique.

By using this line of thought, we can derive the following result:

Lemma 2.6. *For an arbitrary but fixed $w \in \mathbb{R}$ and any initial $(t, x) \in [0, T) \times \mathbb{R}$ and $h > 0$ such that $t + h \leq T$, it holds that for any $u \in \mathcal{A}$:*

$$V(t, x; w) \leq \mathbb{E}[V(t + h, x_{t+h}^u; w) | x_t^u = x].$$

Furthermore, we have equality if and only if the admissible feedback control u is the optimal control $\hat{u}$.

Proof. We choose an arbitrary pair $(t,x) \in [0,T) \times \mathbb{R}$ and $h > 0$ such that $t+h \leq T$. Thus, we divide the time interval $[t,T]$ into two parts, the interval $[t,t+h]$ and $(t+h,T]$, respectively. Furthermore, we consider a fixed but arbitrary feedback control $u \in \mathcal{A}$, and define the control law $u^* \in \mathcal{A}$ by:

$$u^*(s,y;w) = \begin{cases} u(s,y;w) & (s,y) \in [t,t+h] \times \mathbb{R} \\ \hat{u}(s,y;w) & (s,y) \in (t+h,T] \times \mathbb{R}, \end{cases}$$

where $\hat{u} \in \mathcal{A}$ is the optimal feedback control. The interpretation of employing the feedback control u^* is using the arbitrary control u during the time interval $[t,t+h]$, and then switching to the optimal control law during the rest of the time period.

By applying the tower property (see Appendix C.2) to the value function (2.14) under u^*, we obtain:

$$\begin{aligned} J(t,x;w,u^*) &= \mathbb{E}[(x_T^{u^*} - w)^2 | x_t^{u^*} = x] - (w-z)^2 \\ &= \mathbb{E}\big[\mathbb{E}[(x_T^{u^*} - w)^2 | x_{t+h}^{u^*}] \big| x_t^{u^*} = x\big] - (w-z)^2 \\ &= \mathbb{E}\big[\mathbb{E}[(x_T^{\hat{u}} - w)^2 | x_{t+h}^{u}] \big| x_t^{u} = x\big] - (w-z)^2 \\ &= \mathbb{E}\big[\mathbb{E}[(x_T^{\hat{u}} - w)^2 | x_{t+h}^{u}] - (w-z)^2 \big| x_t^{u} = x\big] \\ &\overset{(2.14)}{=} \mathbb{E}\big[J(t+h, x_{t+h}^u; w, \hat{u}) \big| x_t^u = x\big]. \end{aligned}$$

It follows from Bellman's optimality principle that we have:

$$\begin{aligned} J(t,x;w,u^*) &= \mathbb{E}\big[J(t+h, x_{t+h}^u; w, \hat{u}) \big| x_t^u = x\big] \\ &= \mathbb{E}\big[V(t+h, x_{t+h}^u; w) \big| x_t^u = x\big] \end{aligned}$$

and therefore, we can conclude that:

$$V(t,x;w) \leq J(t,x;w,u^*) = \mathbb{E}\big[V(t+h, x_{t+h}^u; w) \big| x_t^u = x\big].$$

Further, since $V(t,x;w) = J(t,x;w,\hat{u})$ we note that equality holds if and only if $u^* = \hat{u}$ and therefore if and only if $u = \hat{u}$. □

Lemma 2.6 is a valuable result for deriving the HJB equation. However, beforehand, we introduce (locally) square-integrable processes and show an applicable property which we will need in several proofs during this work.

2.4.2 Square-integrable stochastic Processes

For the following definitions and results, we still consider the filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ and the standard Brownian motion B on $[0,T]$ introduced in Section 2.1.

Definition 2.7. We define the set of square-integrable processes on $[0,T]$ with respect to the Brownian motion B as:

$$\mathcal{L}^2([0,T],B) := \left\{ H = (H_t)_{t\in[0,T]} \text{ predictable} : \mathbb{E}\left[\int_0^T H_t^2 dt\right] < \infty \right\}.$$

There are stochastic processes which do not satisfy the properties in Definition 2.7 on the whole interval $[0,T]$ but in a local sense. For this purpose, we introduce the so-called localising sequence of stopping times:

Definition 2.8. Let H be a stochastic process. Then, a sequence $(\tau_n)_{n\in\mathbb{N}}$ of stopping times is called localising for H in $\mathcal{L}^2([0,T],B)$ if:

1. $\forall n \geq 1 : \tau_n \leq \tau_{n+1}$
2. $\forall n \geq 1 : \mathbb{1}_{[0,\tau_n]} H \in \mathcal{L}^2([0,T],B)$
3. $\mathbb{P}(\bigcup_{n\geq 1}\{\omega : \tau_n(\omega) = T\}) = 1,$

where $\mathbb{1}_A$ is the indicator function for a set $A \subset \mathbb{R}$. The classical definition requires that τ_n diverge to infinity for $n \to \infty$, but since our time horizon is T, we restrict ourselves to stopping times whose range is in $[0,T]$.

Now, we are able to define the set of locally square-integrable processes:

Definition 2.9. We define the set of locally square-integrable processes on $[0,T]$ as:

$$\begin{aligned}\mathcal{L}^2_{\text{loc}}([0,T],B) := &\Big\{ H = (H_t)_{t\in[0,T]} \text{ predictable} : \exists (\tau_n)_{n\in\mathbb{N}} \text{ a sequence of} \\ &\qquad \text{stopping-times that is localising for } H \text{ in } \mathcal{L}^2([0,T],B) \Big\} \\ = &\left\{ H = (H_t)_{t\in[0,T]} \text{ predictable} : \int_0^T H_t^2 dt < \infty, \quad \mathbb{P}\text{-a.s.} \right\}.\end{aligned}$$

Remark 2.10. We state two useful remarks about the relation between the two spaces and how to show that a process is locally square-integrable:

1. The space $\mathcal{L}^2_{\text{loc}}([0,T],B)$ contains $\mathcal{L}^2([0,T],B)$, since if $\mathbb{E}[\int_0^T H_t^2 dt] < \infty$, then almost all integrals $\int_0^T H_t^2 dt$ are finite.
2. To show that $(F(t,X_t))_{t\in[0,T]} \in \mathcal{L}^2_{\text{loc}}([0,T],B)$, where F is a function and X an adapted stochastic process, it is sufficient to require that X and F are continuous since then the path $t \mapsto F(t,X_t)$ is bounded on $[0,T]$ and so:

$$\int_0^T \big(F(t,X_t)\big)^2 dt < \infty.$$

Furthermore, we can state a localising sequence of stopping times for every locally square-integrable process:

Lemma 2.11. *If $H \in \mathcal{L}^2_{loc}([0,T],B)$, then the sequence of stopping times defined for $n \geq 1$ by:*

$$\tau_n = \inf\left\{t \geq 0 : \int_0^t H_v^2 dv \geq n\right\},$$

with the convention that $\inf\{\emptyset\} = T$, is a localising sequence for H in $\mathcal{L}^2([0,T],B)$.

Remark 2.12. The convention about the infimum of the empty set $\inf\{\emptyset\} = T$ applies if for some paths $\int_0^T H_t^2 dt < n$, as on these paths we do not have to stop at all.

Proof. To prove this statement, we have to verify the three conditions in Definition 2.8.

First, the sequence $(\tau_n)_{n\in\mathbb{N}}$ is clearly non-decreasing.

Next, we have for each $\omega \in \Omega$ and each $n \geq 1$:

$$\int_0^T H_v^2(\omega)\mathbb{1}_{[0,\tau_n(\omega)]}dv = \int_0^{\tau_n(\omega)} H_v^2(\omega)dv \leq n.$$

Therefore, we obtain:

$$\mathbb{E}\left[\int_0^T H_v^2 \mathbb{1}_{[0,\tau_n]}dv\right] \leq n$$

and we can conclude that $\mathbb{1}_{[0,\tau_n]}H \in \mathcal{L}^2([0,T],B)$.

If, for some $\omega \in \Omega$ we have $\int_0^T H_v^2(\omega)dv < \infty$ then for some $n \in \mathbb{N}$ we have $\int_0^T H_v^2(\omega)dv \leq n$ and therefore $\tau_n(\omega) = T$. Hence:

$$\left\{\omega : \int_0^T H_v^2(\omega)dv < \infty\right\} \subset \bigcup_{n\geq 1}\{\omega : \tau_n(\omega) = T\}$$

and since $\mathbb{P}(\{\omega : \int_0^T H_v^2(\omega)dv < \infty\}) = 1$ we can conclude the third condition. □

The next result will be used in various proofs during this work:

Lemma 2.13. *If $H \in \mathcal{L}^2([0,T],B)$, we have for every $t \in [0,T]$:*

$$\mathbb{E}\left[\int_0^t H_v dB_v\right] = 0.$$

Furthermore, for any $s \in [0,t)$ we also have $\mathbb{E}[\int_s^t H_v dB_v|\mathcal{F}_s] = 0$.

Proof. For the proof we refer to Section 5.1.1 in J. Le Gall [13]. In a nutshell, one can show that if $H \in \mathcal{L}^2([0,T],B)$ then the process $(\int_0^t H_v dB_v)_{t\in[0,T]}$ is a continuous martingale and therefore one can conclude that $\mathbb{E}[\int_0^t H_v dB_v] = 0$ for all $t \in [0,T]$. □

We note that one can define $\mathcal{L}^2([s,t],B)$ and $\mathcal{L}^2_{\text{loc}}([s,t],B)$ and derive all results in this section analog for a subset $[s,t] \subseteq [0,T]$.

2.4.3 The Hamilton-Jacobi-Bellman Equation

In this section, we state the HJB equation for the problem (2.6) and prove that under certain assumptions, the optimal value function V given by (2.15) satisfies this equation with $\hat{u}$ attaining the minimum:

Theorem 2.14. ***(Hamilton-Jacobi-Bellman Equation I)**:*
We assume that there exists an admissible optimal feedback control $\hat{u}$ to problem (2.6) and that the optimal value function is regular in the sense that $V \in C^{1,2}([0,T]\times\mathbb{R})$. Then, for a fixed $w \in \mathbb{R}$, the following holds:

1. *V satisfies the HJB equation for $(t,x) \in [0,T)\times\mathbb{R}$:*

$$v_t(t,x;w) + \min_{u\in\mathbb{R}} \left(\frac{1}{2}\sigma^2 u^2 v_{xx}(t,x;w) + \rho\sigma u v_x(t,x;w) \right) = 0 \tag{2.16}$$

   with terminal condition:

$$v(T,x;w) = (x-w)^2 - (w-z)^2, \quad x \in \mathbb{R},$$

   where we denoted the partial derivatives by $v_x := \frac{\partial v}{\partial x}$, $v_t := \frac{\partial v}{\partial t}$ and $v_{xx} := \frac{\partial^2 v}{\partial x \partial x}$.

2. *For each $(t,x) \in [0,T]\times\mathbb{R}$ the minimum in the HJB equation above is attained by the optimal feedback control $\hat{u}(\cdot,\cdot;w)$ such that $u = \hat{u}(t,x;w)$.*

Proof. Through the whole proof, we fix $w \in \mathbb{R}$. Since $V \in C^{1,2}([0,T]\times\mathbb{R})$ by assumption, we can conclude that for any $x \in \mathbb{R}$:

$$\lim_{t\to T} V(t,x;w) = V(T,x;w).$$

Therefore, the terminal condition in (2.16) follows directly from the definition of the optimal value function (2.15):

$$\begin{aligned} V(T,x;w) &= \inf_{\boldsymbol{u}\in\mathcal{A}(T,x)} \mathbb{E}[(x_T^{\boldsymbol{u}} - w)^2 | x_T^{\boldsymbol{u}} = x] - (w-z)^2 \\ &= \inf_{\boldsymbol{u}\in\mathcal{A}(T,x)} \left((x-w)^2 - (w-z)^2 \right) \\ &= (x-w)^2 - (w-z)^2, \quad x \in \mathbb{R}. \end{aligned}$$

Furthermore, we choose $(t,x) \in [0,T) \times \mathbb{R}$ arbitrary and divide the interval $[t,T]$ into two parts $[t,t+h]$ and $(t+h,T]$ respectively for $h > 0$. By Lemma 2.6, we know that for any admissible feedback control $u \in \mathcal{A}$, it holds that:

$$\mathbb{E}[V(t+h, x^u_{t+h}; w) | x^u_t = x] \geq V(t,x;w), \tag{2.17}$$

where $x^u = (x^u_s)_{s\in[t,T]}$ is the wealth process induced by u.

We note that Lemma 2.6 implies equality in (2.17) if and only if $u \in \mathcal{A}$ is the optimal feedback control $\hat{u} \in \mathcal{A}$.

Since $V \in C^{1,2}([0,T] \times \mathbb{R})$ and $u \in \mathcal{A}$ we are able to use Itô's formula (see Appendix B.3) and write the optimal value function in the following way:

$$\begin{aligned} V(t+h, x^u_{t+h}; w) = V(t, x^u_t; w) \\ + \int_t^{t+h} V_t(s, x^u_s; w) + \sigma u_s \rho V_x(s, x^u_s; w) + \frac{1}{2}\sigma^2 u_s^2 V_{xx}(s, x^u_s; w) ds \\ + \int_t^{t+h} \sigma u_s V_x(s, x^u_s; w) dB_s, \end{aligned} \tag{2.18}$$

where $\boldsymbol{u} = (u_s)_{s\in[t,T]} \in \mathcal{A}(t,x)$ is the admissible open-loop control generated from u with respect to the initial (t,x) such that $u_s = u(s, x^u_s; w)$.

Further, we show that $(u_s V_x(s, x^u_s; w))_{s\in[t,T]} \in \mathcal{L}^2_{\text{loc}}([t,T], B)$ by proceeding as explained in the second point in Remark 2.10.

Since $V \in C^{1,2}([0,T] \times \mathbb{R})$ and the wealth process x^u has continuous paths, the path $s \to V_x(s, x^u_s; w)$ is bounded on $[t,T]$ and we can conclude that:

$$\int_t^T u_s^2 \big(V_x(s, x^u_s; w)\big)^2 ds \leq C \cdot \int_t^T u_s^2 ds < \infty, \quad \mathbb{P}\text{-a.s.},$$

where $0 \leq C < \infty$ is a constant and where we used in the last step that $\boldsymbol{u}$ is admissible. Therefore, we have $(u_s V_x(s, x^u_s; w))_{s\in[t,T]} \in \mathcal{L}^2_{\text{loc}}([t,T], B)$.

We define the stopping times for $n \geq 1$:

$$\tau_n := \inf\left\{\tau \geq t : \int_t^\tau u_s^2 \big(V_x(s, x^u_s; w)\big)^2 ds \geq n\right\}.$$

By Lemma 2.11, this sequence of stopping times is a localising sequence for $(u_s V_x(s, x^u_s; w))_{s\in[t,T]}$ in $\mathcal{L}^2([t,T], B)$.

Then, from (2.18), we obtain:

$$
\begin{aligned}
&V\big((t+h)\wedge\tau_n, x^u_{(t+h)\wedge\tau_n}; w\big)\\
&= V(t, x^u_t; w) + \int_t^{(t+h)\wedge\tau_n} V_t(s, x^u_s; w) + \sigma u_s \rho V_x(s, x^u_s; w) + \frac{1}{2}\sigma^2 u_s^2 V_{xx}(s, x^u_s; w) ds\\
&\quad + \int_t^{(t+h)\wedge\tau_n} \sigma u_s V_x(s, x^u_s; w) dB_s.
\end{aligned}
$$

By taking the conditional expectation given $x^u_t = x$ and using Lemma 2.13, we have that:

$$
\begin{aligned}
&\mathbb{E}\Big[V\big((t+h)\wedge\tau_n, x^u_{(t+h)\wedge\tau_n}; w\big)\Big|x^u_t = x\Big]\\
&= \mathbb{E}\Big[V(t, x^u_t; w)\Big|x^u_t = x\Big]\\
&\quad + \mathbb{E}\bigg[\int_t^{(t+h)\wedge\tau_n} V_t(s, x^u_s; w) + \sigma u_s \rho V_x(s, x^u_s; w) + \frac{1}{2}\sigma^2 u_s^2 V_{xx}(s, x^u_s; w) ds\bigg|x^u_t = x\bigg]\\
&\quad + \mathbb{E}\bigg[\int_t^{(t+h)\wedge\tau_n} \sigma u_s V_x(s, x^u_s; w) dB_s\bigg|x^u_t = x\bigg]\\
&= V(t, x; w)\\
&\quad + \mathbb{E}\bigg[\int_t^{(t+h)\wedge\tau_n} V_t(s, x^u_s; w) + \sigma u_s \rho V_x(s, x^u_s; w) + \frac{1}{2}\sigma^2 u_s^2 V_{xx}(s, x^u_s; w) ds\bigg|x^u_t = x\bigg].
\end{aligned}
$$

Since $(\tau_n)_{n\geq 1}$ is a localising sequence and by using the dominated convergence theorem (see Appendix B.4), we obtain for $n \to \infty$:

$$
\begin{aligned}
\mathbb{E}\Big[V(t+h, x^u_{t+h}; w)\Big|x^u_t = x\Big] &= V(t, x; w)\\
&\quad + \mathbb{E}\bigg[\int_t^{t+h} V_t(s, x^u_s; w) + \sigma u_s \rho V_x(s, x^u_s; w)\\
&\quad + \frac{1}{2}\sigma^2 u_s^2 V_{xx}(s, x^u_s; w) ds\bigg|x^u_t = x\bigg].
\end{aligned}
$$

By substituting this formulation into (2.17), we obtain:

$$
\begin{aligned}
V(t, x; w) \leq V(t, x; w)&\\
&+ \mathbb{E}\bigg[\int_t^{t+h} V_t(s, x^u_s; w) + \sigma u_s \rho V_x(s, x^u_s; w)\\
&+ \frac{1}{2}\sigma^2 u_s^2 V_{xx}(s, x^u_s; w) ds\bigg|x^u_t = x\bigg].
\end{aligned}
$$

By subtracting V on both sides, we have:

$$
\begin{aligned}
0 \leq \mathbb{E}\bigg[&\int_t^{t+h} V_t(s, x^u_s; w) + \sigma u_s \rho V_x(s, x^u_s; w)\\
&+ \frac{1}{2}\sigma^2 u_s^2 V_{xx}(s, x^u_s; w) ds\bigg|x^u_t = x\bigg].
\end{aligned}
$$

Further, we divide by h and let $h \to 0$ to obtain:

$$\begin{aligned}
0 &\leq \mathbb{E}[V_t(t, x_t^u; w) + \sigma u_t \rho V_x(t, x_t^u; w) \\
&\quad + \frac{1}{2}\sigma^2 u_t^2 V_{xx}(t, x_t^u; w) | x_t^u = x] \\
&= \mathbb{E}[V_t(t, x_t^u; w) + \sigma u(t, x_t^u; w) \rho V_x(t, x_t^u; w) \\
&\quad + \frac{1}{2}\sigma^2 \big(u(t, x_t^u; w)\big)^2 V_{xx}(t, x_t^u; w) | x_t^u = x] \\
&= V_t(t, x; w) + \sigma u(t, x; w) \rho V_x(t, x; w) \\
&\quad + \frac{1}{2}\sigma^2 \big(u(t, x; w)\big)^2 V_{xx}(t, x; w).
\end{aligned}$$

Considering that we have equality for $u(t, x; w) = \hat{u}(t, x; w)$ and $u \in \mathcal{A}$ was arbitrary chosen, we can conclude that:

$$\begin{aligned}
0 &= V_t(t, x; w) + \sigma \hat{u}(t, x; w) \rho V_x(t, x; w) + \frac{1}{2}\sigma^2 \big(\hat{u}(t, x; w)\big)^2 V_{xx}(t, x; w) \\
&= V_t(t, x; w) + \min_{u \in \mathbb{R}} \big(\sigma u \rho V_x(t, x; w) + \frac{1}{2}\sigma^2 u^2 V_{xx}(t, x; w)\big).
\end{aligned}$$

Furthermore, since the choice of $(t, x) \in [0, T) \times \mathbb{R}$ was arbitrary, we conclude that the value function V, given that $V \in C^{1,2}([0, T] \times \mathbb{R})$, satisfies the HJB equation (2.16) and that the minimum is attained by $u = \hat{u}(t, x; w)$. □

2.4.4 The Verification Theorem

The HJB equation also acts as a sufficient condition for the optimal control problem. This result is known as the Verification theorem for dynamic programming and states that functions v and g, satisfying certain conditions related to the HJB equation, coincide with the optimal value function V and the optimal feedback control $\hat{u}$:

Theorem 2.15. ***(Verification Theorem 1)****:*
Suppose that $w \in \mathbb{R}$ is fixed and that we have two functions $v \in C^{1,2}([0, T] \times \mathbb{R})$ and $g : [0, T] \times \mathbb{R} \longrightarrow \mathbb{R}$ such that:

1. *v is sufficiently integrable (see Remark 2.16 below) and solves the HJB equation (2.16).*
2. *The function g is an admissible feedback control in the sense of Definition 2.4.*
3. *For each fixed $(t, x) \in [0, T) \times \mathbb{R}$ the minimum in the HJB equation (2.16):*

$$\min_{u \in \mathbb{R}} \left(\frac{1}{2}\sigma^2 u^2 v_{xx}(t, x; w) + \rho\sigma u v_x(t, x; w)\right)$$

is attained by the choice $u = g(t, x; w)$.

Then the following hold:

1. *The optimal value function V is given by:*

$$V(t,x;w) = v(t,x;w), \quad (t,x) \in [0,T] \times \mathbb{R}.$$

2. *There exists an optimal feedback control $\hat{u}$, and in fact:*

$$\hat{u}(t,x;w) = g(t,x;w), \quad (t,x) \in [0,T] \times \mathbb{R}.$$

Proof. We assume that v and g are given as above and we choose an arbitrary point $(t,x) \in [0,T) \times \mathbb{R}$ and fix $w \in \mathbb{R}$. Further, we choose an arbitrary admissible feedback control $u = u(\cdot,\cdot;w) \in \mathcal{A}$ and consider the induced wealth process $x^u = (x_s^u)_{s\in[t,T]}$ with following dynamics as in Definition 2.4:

$$\begin{aligned} dx_s^u &= \sigma u(s,x_s^u;w)(\rho ds + dB_s), \quad t \le s \le T \\ x_t^u &= x. \end{aligned}$$

By inserting the process x^u into the function v and using Itô's formula (see Appendix B.3), we obtain:

$$\begin{aligned} v(T,x_T^u;w) = v(t,x_t^u;w) &+ \int_t^T \big(v_t(s,x_s^u;w) + \rho u(s,x_s^u;w)\sigma v_x(s,x_s^u;w) \\ &+ \frac{1}{2}\big(u(s,x_s^u;w)\big)^2\sigma^2 v_{xx}(s,x_s^u;w)\big)ds \\ &+ \int_t^T u(s,x_s^u;w)\sigma v_x(s,x_s^u;w)dB_s. \end{aligned} \tag{2.19}$$

Analog to the proof of Theorem 2.14, we define the stopping times for $n \ge 1$:

$$\tau_n := \inf\left\{\tau \ge t : \int_t^\tau \big(u(s,x_s^u;w)\big)^2\big(v_x(s,x_s^u;w)\big)^2 ds \ge n\right\}$$

and use (2.19) to obtain:

$$\begin{aligned} v(T\wedge\tau_n, x_{T\wedge\tau_n}^u;w) = v(t,x_t^u;w) &\\ &+ \int_t^{T\wedge\tau_n} \big(v_t(s,x_s^u;w) + \rho u(s,x_s^u;w)\sigma v_x(s,x_s^u;w) \\ &+ \frac{1}{2}\big(u(s,x_s^u;w)\big)^2\sigma^2 v_{xx}(s,x_s^u;w)\big)ds \\ &+ \int_t^{T\wedge\tau_n} u(s,x_s^u;w)\sigma v_x(s,x_s^u;w)dB_s. \end{aligned}$$

Further, we take the conditional expectation given $x_t^u = x$:

$$\begin{aligned}
&\mathbb{E}[v(T \wedge \tau_n, x^u_{T\wedge\tau_n}; w)|x_t^u = x] \\
&= v(t,x;w) + \mathbb{E}\bigg[\int_t^{T\wedge\tau_n} \big(v_t(s,x_s^u;w) + \rho u(s,x_s^u;w)\sigma v_x(s,x_s^u;w) \\
&\quad + \frac{1}{2}\big(u(s,x_s^u;w)\big)^2\sigma^2 v_{xx}(s,x_s^u;w)\big)ds\bigg|x_t^u = x\bigg] \\
&\quad + \sigma\cdot\mathbb{E}\bigg[\int_t^{T\wedge\tau_n} u(s,x_s^u;w)v_x(s,x_s^u;w)dB_s\bigg|x_t^u = x\bigg].
\end{aligned} \tag{2.20}$$

By Remark 2.16 and Lemma 2.11, we can conclude that the sequence of stopping times $(\tau_n)_{n\geq 1}$ is a localising sequence for $(u(s,x_s^u;w)v_x(s,x_s^u;w))_{s\in[t,T]}$ in $\mathcal{L}^2([t,T],B)$ and therefore, the expectation of the stochastic integral vanishes due to Lemma 2.13:

$$\begin{aligned}
&\mathbb{E}[v(T \wedge \tau_n, x^u_{T\wedge\tau_n}; w)|x_t^u = x] \\
&= v(t,x;w) + \mathbb{E}\bigg[\int_t^{T\wedge\tau_n} \big(v_t(s,x_s^u;w) + \rho u(s,x_s^u;w)\sigma v_x(s,x_s^u;w) \\
&\quad + \frac{1}{2}\big(u(s,x_s^u;w)\big)^2\sigma^2 v_{xx}(s,x_s^u;w)\big)ds\bigg|x_t^u = x\bigg].
\end{aligned}$$

By the properties of the localising sequence in Definition 2.8 and by using the dominated convergence theorem (see Appendix B.4), we obtain for $n \to \infty$:

$$\begin{aligned}
\mathbb{E}[v(T,x_T^u;w)|x_t^u = x] &= v(t,x;w) \\
&\quad + \mathbb{E}\bigg[\int_t^T \big(v_t(s,x_s^u;w) + \rho u(s,x_s^u;w)\sigma v_x(s,x_s^u;w) \\
&\quad + \frac{1}{2}\big(u(s,x_s^u;w)\big)^2\sigma^2 v_{xx}(s,x_s^u;w)\big)ds\bigg|x_t^u = x\bigg].
\end{aligned} \tag{2.21}$$

Since v solves the HJB equation (2.16), we obtain for any $u \in \mathbb{R}$:

$$v_t(s,y;w) + \frac{1}{2}\sigma^2u^2v_{xx}(s,y;w) + \rho\sigma u v_x(s,y;w) \geq 0, \quad (s,y) \in [0,T)\times\mathbb{R}.$$

By integrating from t to T, we have that:

$$\int_t^T \bigg(v_t(s,y;w) + \frac{1}{2}\sigma^2u^2v_{xx}(s,y;w) + \rho\sigma u v_x(s,y;w)\bigg)ds \geq 0 \tag{2.22}$$

and by substitute (2.22) with $u = u(s,x_s^u;w)$ and $y = x_s^u$ into (2.21), we obtain following inequality:

$$\mathbb{E}[v(T,x_T^u;w)|x_t^u = x] \geq v(t,x;w).$$

Furthermore, we use the terminal condition of the HJB equation (2.16) and obtain:

$$\begin{aligned} v(t,x;w) &\leq \mathbb{E}[v(T,x_T^u;w)|x_t^u = x] = \mathbb{E}[(x_T^u - w)^2 - (w-z)^2|x_t^u = x] \\ &= E[(x_T^u - w)^2|x_t^u = x] - (w-z)^2 \overset{(2.14)}{=} J(t,x;w,u). \end{aligned}$$

Since the feedback control $u \in \mathcal{A}$ was chosen arbitrary, we obtain:

$$v(t,x;w) \leq V(t,x;w). \tag{2.23}$$

To obtain the reverse inequality we choose $u = g$. Since this choice minimizes the second term in the HJB equation and v solves (2.16) by assumption, we have for $(s,y) \in [0,T) \times \mathbb{R}$:

$$v_t(s,y;w) + \frac{1}{2}\sigma^2(g(s,y;w)^2 v_{xx}(s,y;w) + \rho\sigma g(s,y;w)v_x(s,y;w) = 0.$$

Again, by proceeding as above, we obtain:

$$v(t,x;w) = J(t,x;w,g) \geq V(t,x;w). \tag{2.24}$$

By combining (2.23) and (2.24), we can finally conclude that:

$$v(t,x;w) = V(t,x;w) = J(t,x;w,g).$$

Since $(t,x) \in [0,T) \times \mathbb{R}$ was chosen arbitrary we have $v(t,x;w) = V(t,x;w)$ and $\hat{u}(t,x;w) = g(t,x;w)$ for all $(t,x) \in [0,T) \times \mathbb{R}$. □

Remark 2.16. The assumption that v is sufficiently integrable in Theorem 2.15 above is made in order for the expectation of the stochastic integral in (2.20) to vanish. By Lemma 2.13, this will be the case if v satisfies the condition:

$$\big(u(s,x_s^u;w)v_x(s,x_s^u;w)\big)_{s\in[t,T]} \in \mathcal{L}^2_{\text{loc}}([t,T],B).$$

2.4.5 Solution

Finally, we show that for a fixed $w \in \mathbb{R}$ the optimal value function V is given by:

$$V(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} - (w-z)^2, \quad (t,x) \in [0,T) \times \mathbb{R} \tag{2.25}$$

and the optimal feedback control $\hat{u}(\cdot,\cdot;w)$ is given by

$$\hat{u}(t,x;w) = -\frac{\rho}{\sigma}(x-w), \quad (t,x) \in [0,T) \times \mathbb{R}. \tag{2.26}$$

However, first, we investigate the wealth process induced by $\hat{u}$.

For this purpose, we consider the following SDE for any $(s, y) \in [0, T) \times \mathbb{R}$:

$$\begin{aligned} dx_t^{\hat{u}} &= \sigma \hat{u}(t, x_t^{\hat{u}}; w)(\rho dt + dB_t) \\ &= -\rho^2(x_t^{\hat{u}} - w)dt - \rho(x_t^{\hat{u}} - w)dB_t, \quad t \in [s, T] \\ x_s^{\hat{u}} &= y. \end{aligned} \tag{2.27}$$

This is a linear SDE for which we can compute an explicit solution $(x_t^{\hat{u}})_{t\in[s,T]}$ that is adapted and has continuous paths (see Appendix B.7).

Furthermore, we compute the first two moments of the wealth process $x^{\hat{u}}$:

It follows from (2.27), Lemma 2.13 and Fubini's theorem (Theorem B.7) that:

$$\begin{aligned} \mathbb{E}[x_t^{\hat{u}} | x_s^{\hat{u}} = y] &= \mathbb{E}\Big[x_s^{\hat{u}} - \int_s^t \rho^2(x_v^{\hat{u}} - w)dv - \int_s^t \rho(x_v^{\hat{u}} - w)dB_v \Big| x_s^{\hat{u}} = y\Big] \\ &= y - \mathbb{E}\Big[\int_s^t \rho^2(x_v^{\hat{u}} - w)dv \Big| x_s^{\hat{u}} = y\Big] - \mathbb{E}\Big[\int_s^t \rho(x_v^{\hat{u}} - w)dB_v \Big| x_s^{\hat{u}} = y\Big] \\ &\overset{\text{Lem. 2.13}}{=} y - \mathbb{E}\Big[\int_s^t \rho^2(x_v^{\hat{u}} - w)dv \Big| x_s^{\hat{u}} = y\Big] \\ &\overset{\text{Theo. B.7}}{=} y - \int_s^t \rho^2(\mathbb{E}\big[x_v^{\hat{u}} \big| x_s^{\hat{u}} = y\big] - w)dv, \quad t \in [s, T]. \end{aligned}$$

This yields the ordinary differential equation (ODE) for $n(t) := \mathbb{E}[x_t^{\hat{u}} | x_s^{\hat{u}} = y]$:

$$\begin{aligned} dn(t) &= -\rho^2(n(t) - w)dt, \quad t \in [s, T] \\ n(s) &= y, \end{aligned}$$

which can be solved using integrating factors:

$$\mathbb{E}[x_t^{\hat{u}} | x_s^{\hat{u}} = y] = (y - w)e^{-\rho^2(t-s)} + w, \quad t \in [s, T]. \tag{2.28}$$

For the second moment, we first use Itô's formula (see Appendix B.3) and (2.27), to obtain:

$$\begin{aligned} \big(x_t^{\hat{u}}\big)^2 &= \big(x_s^{\hat{u}}\big)^2 - \int_s^t 2x_v^{\hat{u}}\rho^2(x_v^{\hat{u}} - w) - \rho^2(x_v^{\hat{u}} - w)^2 dv - \int_s^t 2x_v^{\hat{u}}\rho(x_v^{\hat{u}} - w)dB_v \\ &= y^2 - \int_s^t \rho^2\Big(\big(x_v^{\hat{u}}\big)^2 - w^2\Big)dv - \int_s^t 2x_v^{\hat{u}}\rho(x_v^{\hat{u}} - w)dB_v, \quad t \in [s, T]. \end{aligned}$$

And again, it follows from Lemma 2.13 and Fubini's theorem (Theorem B.7) that:

$$\mathbb{E}\Big[\big(x_t^{\hat{u}}\big)^2 \Big| x_s^{\hat{u}} = y\Big] = y^2 - \int_s^t \rho^2\Big(\mathbb{E}\Big[\big(x_v^{\hat{u}}\big)^2 \Big| x_s^{\hat{u}} = y\Big] - w^2\Big)dv, \quad t \in [s, T].$$

This yields the ODE for $m(t) := \mathbb{E}[(x_t^{\hat{u}})^2 | x_s^{\hat{u}} = y]$:

$$\begin{aligned} dm(t) &= -\rho^2(m(t) - w^2)dt, \quad t \in [s, T] \\ m(s) &= y^2, \end{aligned}$$

which again can be solved using integrating factors:

$$\mathbb{E}\Big[(x_t^{\hat{u}})^2 \Big| x_s^{\hat{u}} = y\Big] = (y^2 - w^2)e^{-\rho^2(t-s)} + w^2, \quad t \in [s, T]. \tag{2.29}$$

By combining (2.28) and (2.29), we obtain a result that we use in the next lemma but that will also be important in Chapter 4:

$$\begin{aligned} \mathbb{E}\Big[(x_t^{\hat{u}} - w)^2 \Big| x_s^{\hat{u}} = y\Big] &= \mathbb{E}\Big[(x_t^{\hat{u}})^2 - 2w x_t^{\hat{u}} + w^2 \Big| x_s^{\hat{u}} = y\Big] \\ &= \mathbb{E}\Big[(x_t^{\hat{u}})^2 \Big| x_s^{\hat{u}} = y\Big] - 2w\mathbb{E}[x_t^{\hat{u}} | x_s^{\hat{u}} = y] + w^2 \\ &= (y^2 - w^2)e^{-\rho^2(t-s)} + w^2 \\ &\quad - 2w\big((y - w)e^{-\rho^2(t-s)} + w\big) + w^2 \\ &= (y - w)^2 e^{-\rho^2(t-s)}, \quad t \in [s, T]. \end{aligned} \tag{2.30}$$

In the next step, we show that the functions given by (2.25) and (2.26) satisfy the conditions in the Verification theorem (Theorem 2.15):

Lemma 2.17. *For a fixed $w \in \mathbb{R}$, the feedback control $\hat{u}(\cdot, \cdot; w)$ given in (2.26) is admissible in the sense of Definition 2.4.*

Proof. We verify the conditions in Definition 2.4 for a fixed $w \in \mathbb{R}$.

We have clearly $\hat{u}(t, x; w) \in \mathbb{R}$ for $(t, x) \in [0, T] \times \mathbb{R}$ and therefore, the first condition is satisfied.

Next, we choose $(s, y) \in [0, T) \times \mathbb{R}$ arbitrary. In the explanations before this lemma, we derived an explicit solution $(x_t^{\hat{u}})_{t \in [s,T]}$ to the linear SDE (2.27) that is adapted and has continuous paths (see Appendix B.7). Hence, the first part in the second condition in Definition 2.4 is satisfied.

To prove that the strategy $\hat{\boldsymbol{u}} = (\hat{u}_t)_{t \in [s,T]}$ generated from $\hat{u}$ with respect to the initial (s, y) is admissible, we have to show that $\hat{\boldsymbol{u}}$ is progressively measurable and $\mathbb{E}[\int_s^T (\hat{u}_t)^2 dt] < \infty$.

Since the wealth process $(x_t^{\hat{u}})_{t \in [s,T]}$ is adapted with continuous paths it is progressively measurable (see Appendix B.1). Therefore, the stochastic process $\hat{\boldsymbol{u}}$ with $\hat{u}_t = -\frac{\rho}{\sigma}(x_t^{\hat{u}} - w)$ is also progressively measurable.

At last, we show:

$$\mathbb{E}\Big[\int_s^T (\hat{u}_t)^2 dt\Big] = \mathbb{E}\Big[\int_s^T \big(\hat{u}(t, x_t^{\hat{u}}; w)\big)^2 dt\Big] = \frac{\rho^2}{\sigma^2}\mathbb{E}\Big[\int_s^T (x_t^{\hat{u}} - w)^2 dt\Big] < \infty. \tag{2.31}$$

By (2.30), we can conclude that $\mathbb{E}[(x_t^{\hat{u}} - w)^2] < \infty$ for all $t \in [s, T]$.

Hence, we can show (2.31) by applying Fubini's theorem (Theorem B.7):

$$\mathbb{E}\Big[\int_s^T (\hat{u}_t)^2 dt\Big] = \frac{\rho^2}{\sigma^2}\mathbb{E}\Big[\int_s^T (x_t^{\hat{u}} - w)^2 dt\Big] = \frac{\rho^2}{\sigma^2}\int_s^T \mathbb{E}\Big[(x_t^{\hat{u}} - w)^2\Big] dt < \infty.$$

And we can conclude that the strategy $\hat{\boldsymbol{u}}$ is admissible and therefore, $\hat{u} \in \mathcal{A}$. □

Now, we are able to deduce the optimal value function V and the optimal feedback control $\hat{u}$ of the problem (2.6):

Theorem 2.18. *The optimal value function V is given by (2.25) and the optimal feedback control $\hat{u}$ is given by (2.26). Furthermore, the induced optimal wealth process $x^{\hat{u}} = (x_t^{\hat{u}})_{t\in[0,T]}$ is the unique solution to the linear SDE:*

$$\begin{aligned} dx_t^{\hat{u}} &= -\rho^2(x_t^{\hat{u}} - w)dt - \rho(x_t^{\hat{u}} - w)dB_t, \quad 0 \le t \le T \\ x_0^{\hat{u}} &= x_0 \end{aligned}$$

and the Lagrange multiplier w is given by $w = \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}$.

Proof. By the Verification Theorem 2.15, we have to show that the functions v and g given by $v(t, x; w) = (x-w)^2 e^{-\rho^2(T-t)} - (w-z)^2$ and $g(t, x; w) = -\frac{\rho}{\sigma}(x-w)$ for $(t, x) \in [0, T] \times \mathbb{R}$ satisfy the three assumptions in Theorem 2.15.

By Lemma 2.17, we can conclude that g is admissible and the second assumption of Theorem 2.15 holds.

Obviously, the function v is in $C^{1,2}([0, T] \times \mathbb{R})$. And since v_x is continuous and for an admissible feedback control $u \in \mathcal{A}$ the associated wealth process x^u has continuous paths, the path $t \to v_x(t, x_t^u; w)$ is bounded on $[0, T]$ and therefore:

$$\int_0^T \big(u(t, x_t^u; w)\big)^2 \big(v_x(t, x_t^u; w)\big)^2 dt \le C \cdot \int_0^T \big(u(t, x_t^u; w)\big)^2 dt < \infty, \quad \mathbb{P}\text{-a.s.},$$

where $0 \le C < \infty$ is a constant. Hence, v is sufficiently integrable by Remark 2.16.

Furthermore, v satisfies the terminal condition of the HJB equation (2.16) and by substituting v into (2.16), we get for $(t,x) \in [0,T) \times \mathbb{R}$:

$$\begin{aligned} &v_t(t,x;w) + \min_{u\in\mathbb{R}} \left(\frac{1}{2}\sigma^2 u^2 v_{xx}(t,x;w) + \rho\sigma u v_x(t,x;w) \right) \\ &= \rho^2(x-w)^2 e^{-\rho^2(T-t)} + e^{-\rho^2(T-t)} \min_{u\in\mathbb{R}} \left(\sigma^2 u^2 + 2\rho\sigma u(x-w) \right). \end{aligned} \tag{2.32}$$

If we minimize the corresponding term by using standard methods from Analysis, we obtain:

$$\arg\min_{u\in\mathbb{R}} \left(\sigma^2 u^2 + 2\rho\sigma u(x-w) \right) = -\frac{\rho}{\sigma}(x-w) = \hat{u}(t,x;w). \tag{2.33}$$

Therefore the third condition in Theorem 2.15 is satisfied.

Furthermore, if we substitute (2.33) into (2.32), we obtain for $(t,x) \in [0,T) \times \mathbb{R}$:

$$\begin{aligned} &\rho^2(x-w)^2 e^{-\rho^2(T-t)} + e^{-\rho^2(T-t)} \min_{u\in\mathbb{R}} \left(\sigma^2 u^2 + 2\rho\sigma u(x-w) \right) \\ &= \rho^2(x-w)^2 e^{-\rho^2(T-t)} + e^{-\rho^2(T-t)} (\rho^2(x-w)^2 - 2\rho^2(x-w)^2) \\ &= \rho^2(x-w)^2 e^{-\rho^2(T-t)} - \rho^2(x-w)^2 e^{-\rho^2(T-t)} \\ &= 0. \end{aligned}$$

Therefore, v satisfies the HJB equation (2.16) with $u = g(t,x;w)$ and the first assumption in Theorem 2.15 is satisfied.

Finally, we can apply the Verification theorem and conclude that the optimal value function $V = v$ and that the optimal admissible feedback control $\hat{u} = g$ on $[0,T] \times \mathbb{R}$.

Furthermore, since $\hat{u}$ is admissible by Lemma 2.17 the process $x^{\hat{u}} = (x_t^{\hat{u}})_{t\in[0,T]}$ is the unique strong solution of the SDE:

$$\begin{aligned} dx_t^{\hat{u}} &= \sigma\hat{u}(t, x_t^{\hat{u}}; w)(\rho dt + dB_t) \\ &= -\rho^2(x_t^{\hat{u}} - w)dt - \rho(x_t^{\hat{u}} - w)dB_t, \quad 0 \le t \le T \\ x_0^{\hat{u}} &= x_0. \end{aligned}$$

By (2.28), we have:

$$\mathbb{E}[x_T^{\hat{u}}] = (x_0 - w)e^{-\rho^2 T} + w.$$

Finally, we can compute by using the constraint $\mathbb{E}[x_T^{\hat{u}}] = z$ of problem (2.6) that:

$$w = \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}.$$

□

Theorem 2.18 implies that in the classical setting, the optimal control can be derived if the model is fully known. That means the parameters r, σ and μ (or σ and ρ) are specified. Unfortunately, it is challenging to determine these market parameters, and one usually has to estimate them by using historical time series of asset prices. In reality, these estimations are often not accurate enough to derive reliable strategies (see, for example, M. J. Best & R. R. Grauer [6]). Moreover, the optimal strategies are sensitive to these parameters, especially due to inverting ill-conditioned covariance matrices, which are not reliably invertible by a numerical method. These issues lead to solutions that are not satisfying enough to solve the MV problem. Therefore, we are interested in other approaches to solve (2.6). A promising one that is discussed in the next chapter is the application of Reinforcement Learning (RL) techniques since they do not require any estimation of model parameters.

Chapter 3

Exploratory continuous-time Mean-Variance Portfolio Selection

In this chapter, we discuss an entropy-regularized, exploratory mean-variance portfolio selection problem in continuous time in the setup of RL. First, we formulate the strategy and the associated wealth process in an exploratory sense. Based on these, we state the exploratory MV (EMV) problem. Moreover, in the last step, we present its solution. This chapter mainly follows the detailed explanations in H. Wang & X. Y. Zhou [25] and H. Wang et al. [26].

3.1 Exploratory Formulation

As mentioned at the end of Chapter 2, RL methods can skip the estimation of model parameters and output directly (near) optimal strategies. An agent interacts with the unknown environment and collects information (exploring) while optimizing (exploiting) a policy that defines how an agent acts in a specific state. In our case, this environment is a financial market, the information is asset prices, and the policy is a trading strategy.

In order to represent exploration and exploitation, we randomize policies by extending them to distributions of policies. In this way, a randomized policy specifies the probability of taking an action given a specific state. We call such a policy also a distributional control process and denote it by:

$$\boldsymbol{\pi} = (\pi_t)_{t\in[0,T]}.$$

This process gives for each time $t \in [0,T]$ a density function π_t and is so-called measure-valued or more precise density-function-valued.

The agent executes a strategy $\boldsymbol{u}$ over the time horizon $[0,T]$ by sampling it

first from the distributions given by $\boldsymbol{\pi}$ and obtaining for each time $t \in [0,T]$ a real-valued action $u_t \sim \pi_t du$. Therefore, if we repeat this procedure over a large enough number $N \in \mathbb{N}$, the value function $J(\cdot,\cdot;w,u)$ under a feedback control $u \in \mathcal{A}$ given in (2.14) can be estimated accurately. Nevertheless, since this function depends on the wealth process $x^{\boldsymbol{u}}$, we first have to define the exploratory version of a wealth process $x^{\boldsymbol{\pi}}$ by adapting the dynamics (2.4). Here, $x^{\boldsymbol{\pi}}$ indicates a wealth process $x^{\boldsymbol{u}}$ following the dynamics (2.4) with corresponding strategy $\boldsymbol{u}$ sampled independently under $\boldsymbol{\pi}$. Therefore, $x^{\boldsymbol{u}}$ can be viewed as an independent sample from $x^{\boldsymbol{\pi}}$.

To motivate an exploratory formulation of the wealth dynamics, we investigate how the increments of the wealth process and its squares are affected by repetitive learning under a distributional control process $\boldsymbol{\pi}$.

Let, for $i = 1,2,...,N$, B^i be N independent sample paths of the Brownian motion B and x^i be the copies of the wealth process respectively under the strategy $\boldsymbol{u}^i$ each sampled from $\boldsymbol{\pi}$. The increments of the copies of the wealth process for $t \in [0,T)$ are given by:

$$\begin{aligned}\Delta x_t^i &= x_{t+\Delta t}^i - x_t^i \\ &\overset{(2.4)}{=} \int_t^{t+\Delta t} \rho\sigma u_s^i ds + \int_t^{t+\Delta t} \sigma u_s^i dB_s \\ &\approx \rho\sigma u_t^i \Delta t + \sigma u_t^i (B_{t+\Delta t}^i - B_t^i), \quad t \in [0,T],\end{aligned}$$

where we choose Δt small enough. By using the law of large numbers (see Appendix A.2), we obtain as $N \to \infty$:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \Delta x_t^i &\approx \frac{1}{N}\sum_{i=1}^N \left(\rho\sigma u_t^i \Delta t + \sigma u_t^i (B_{t+\Delta t}^i - B_t^i)\right) \\ &\xrightarrow{\mathbb{P}\text{-a.s.}} \mathbb{E}\left[\int_{\mathbb{R}} \rho\sigma u \pi_t(u) du \Delta t + \int_{\mathbb{R}} \sigma u \pi_t(u) du (B_{t+\Delta t} - B_t)\right].\end{aligned}$$

Furthermore, we assume that π_t and $x^{\boldsymbol{\pi}}$ are independent of the increments of the Brownian motion sample paths. Hence, we obtain by using the linearity of the expectation and the properties of a standard Brownian motion (see Appendix B.2) that:

$$\begin{aligned}&\mathbb{E}\left[\int_{\mathbb{R}} \rho\sigma u \pi_t(u) du \Delta t + \int_{\mathbb{R}} \sigma u \pi_t(u) du (B_{t+\Delta t} - B_t)\right] \\ &= \mathbb{E}\left[\int_{\mathbb{R}} \rho\sigma u \pi_t(u) du \Delta t\right] + \mathbb{E}\left[\int_{\mathbb{R}} \sigma u \pi_t(u) du\right] \mathbb{E}\left[(B_{t+\Delta t} - B_t)\right] \\ &= \mathbb{E}\left[\int_{\mathbb{R}} \rho\sigma u \pi_t(u) du \Delta t\right].\end{aligned}$$

Therefore, we can conclude that:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\Delta x_t^i &\approx \frac{1}{N}\sum_{i=1}^{N}\left(\rho\sigma u_t^i\Delta t+\sigma u_t^i(B_{t+\Delta t}^i-B_t^i)\right)\\ &\xrightarrow{\mathbb{P}\text{-a.s.}}\mathbb{E}\bigg[\int_{\mathbb{R}}\rho\sigma u\pi_t(u)du\Delta t\bigg].\end{aligned}\tag{3.1}$$

For the squares of the increments, we proceed similarly:

$$\begin{aligned}(\Delta x_t^i)^2=(x_{t+\Delta t}^i-x_t^i)^2 &\overset{(2.4)}{=}\left(\int_t^{t+\Delta t}\rho\sigma u_s^i ds+\int_t^{t+\Delta t}\sigma u_s^i dB_s\right)^2\\ &=\left(\int_t^{t+\Delta t}\rho\sigma u_s^i ds\right)^2+\left(\int_t^{t+\Delta t}\sigma u_s^i dB_s\right)^2\\ &\quad+2\left(\int_t^{t+\Delta t}\rho\sigma u_s^i ds\right)\left(\int_t^{t+\Delta t}\sigma u_s^i dB_s\right)\\ &\approx(\rho\sigma u_t^i\Delta t)^2+\left(\sigma u_t^i(B_{t+\Delta t}^i-B_t^i)\right)^2+2(\rho\sigma u_t^i\Delta t)\left(\sigma u_t^i(B_{t+\Delta t}^i-B_t^i)\right)\\ &=\rho^2\sigma^2(u_t^i)^2(\Delta t)^2+\sigma^2(u_t^i)^2(B_{t+\Delta t}^i-B_t^i)^2+2\rho\sigma^2(u_t^i)^2\Delta t(B_{t+\Delta t}^i-B_t^i)\\ &\approx\sigma^2(u_t^i)^2(B_{t+\Delta t}^i-B_t^i)^2+2\rho\sigma^2(u_t^i)^2\Delta t(B_{t+\Delta t}^i-B_t^i),\quad t\in[0,T],\end{aligned}$$

where in the last estimation, we used that Δt is small and therefore $(\Delta t)^2\approx 0$.

We use again the law of large numbers and obtain as $N\to\infty$:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}(\Delta x_t^i)^2 &\approx \frac{1}{N}\sum_{i=1}^{N}\sigma^2(u_t^i)^2(B_{t+\Delta t}^i-B_t^i)^2+2\rho\sigma^2(u_t^i)^2\Delta t(B_{t+\Delta t}^i-B_t^i)\\ &\xrightarrow{\mathbb{P}\text{-a.s.}}\mathbb{E}\bigg[\int_{\mathbb{R}}\sigma^2u^2\pi_t(u)du(B_{t+\Delta t}-B_t)^2\\ &\qquad+\int_{\mathbb{R}}2\rho\sigma^2u^2\pi_t(u)du\Delta t(B_{t+\Delta t}-B_t)\bigg].\end{aligned}$$

Furthermore, we obtain by the assumption that π_t and $x^{\boldsymbol{\pi}}$ are independent of the increments of the Brownian motion sample paths and by using the properties of a standard Brownian motion (see Appendix B.2) that:

$$\begin{aligned}&\mathbb{E}\bigg[\int_{\mathbb{R}}\sigma^2u^2\pi_t(u)du(B_{t+\Delta t}-B_t)^2+\int_{\mathbb{R}}2\rho\sigma^2u^2\pi_t(u)du(B_{t+\Delta t}-B_t)\Delta t\bigg]\\ &=\mathbb{E}\bigg[\int_{\mathbb{R}}\sigma^2u^2\pi_t(u)du\bigg]\mathbb{E}\Big[(B_{t+\Delta t}-B_t)^2\Big]\\ &\quad+\mathbb{E}\bigg[\int_{\mathbb{R}}2\rho\sigma^2u^2\pi_t(u)du\Delta t\bigg]\mathbb{E}\Big[(B_{t+\Delta t}-B_t)\Big]\\ &=\mathbb{E}\bigg[\int_{\mathbb{R}}\sigma^2u^2\pi_t(u)du\bigg]\mathrm{Var}(B_{t+\Delta t}-B_t)=\mathbb{E}\bigg[\int_{\mathbb{R}}\sigma^2u^2\pi_t(u)du\Delta t\bigg].\end{aligned}$$

Therefore, we can conclude that:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}(\Delta x_t^i)^2 &\approx \frac{1}{N}\sum_{i=1}^{N}\sigma^2(u_t^i)^2(B_{t+\Delta t}^i - B_t^i)^2 + 2\rho\sigma^2(u_t^i)^2\Delta t(B_{t+\Delta t}^i - B_t^i) \\ &\xrightarrow{\mathbb{P}\text{-a.s.}} \mathbb{E}\left[\int_{\mathbb{R}}\sigma^2u^2\pi_t(u)du\Delta t\right].\end{aligned} \tag{3.2}$$

Finally, as the individual wealth x_t^i, for $i = 1, 2, ..., N$, is an independent sample from $x_t^{\boldsymbol{\pi}}$, we have that Δx_t^i and $(\Delta x_t^i)^2$ are the independent samples from $\Delta x_t^{\boldsymbol{\pi}}$ and $(\Delta x_t^{\boldsymbol{\pi}})^2$, respectively. Then, the law of large numbers gives that as $N \to \infty$:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\Delta x_t^i &\xrightarrow{\mathbb{P}\text{-a.s.}} \mathbb{E}\left[\Delta x_t^{\boldsymbol{\pi}}\right] \\ \frac{1}{N}\sum_{i=1}^{N}(\Delta x_t^i)^2 &\xrightarrow{\mathbb{P}\text{-a.s.}} \mathbb{E}\left[(\Delta x_t^{\boldsymbol{\pi}})^2\right], \quad t \in [0,T].\end{aligned}$$

By combining this interpretation with (3.1) and (3.2), it appears reasonable to propose the following exploratory version of the state dynamics:

$$\begin{aligned}dx_t^{\boldsymbol{\pi}} &= \tilde{b}(\pi_t)dt + \tilde{\sigma}(\pi_t)dB_t, \quad 0 \le t \le T \\ x_0^{\boldsymbol{\pi}} &= x_0,\end{aligned} \tag{3.3}$$

where:

$$\tilde{b}(\pi) := \int_{\mathbb{R}}\rho\sigma u\pi(u)du, \quad \pi \in \mathcal{P}(\mathbb{R}) \tag{3.4}$$

and

$$\tilde{\sigma}(\pi) := \sqrt{\int_{\mathbb{R}}\sigma^2u^2\pi(u)du}, \quad \pi \in \mathcal{P}(\mathbb{R}), \tag{3.5}$$

where $\mathcal{P}(\mathbb{R})$ is the set of density functions of probability measures on $\mathbb{R}$ that are absolutely continuous with respect to the Lebesgue measure. We will call (3.3) the exploratory formulation of the controlled wealth dynamics, and $\tilde{b}$ and $\tilde{\sigma}$ in (3.4) and (3.5), respectively, the exploratory drift and the exploratory volatility.

The motivation for (3.3), based on the derivation above, is supported by the fact that it coincides with the relaxed control formulation in classical control theory (for more details, see W. H. Fleming & M. Nisio [11], and X. Y. Zhou [28]).

Furthermore, we can define the mean and variance processes (assuming they exist for now) associated with the distributional control process $\boldsymbol{\pi}$:

$$\begin{aligned}\mu_t &:= \int_{\mathbb{R}}u\pi_t(u)du, \\ \sigma_t^2 &:= \int_{\mathbb{R}}u^2\pi_t(u)du - \mu_t^2, \quad 0 \le t \le T.\end{aligned}$$

Therefore, we can write the exploratory dynamics (3.3) as:

$$\begin{aligned} dx_t^{\boldsymbol{\pi}} &= \rho\sigma\mu_t dt + \sigma\sqrt{\mu_t^2 + \sigma_t^2}dB_t, \quad 0 \le t \le T \\ x_0^{\boldsymbol{\pi}} &= x_0. \end{aligned} \tag{3.6}$$

3.2 Entropy Regularization

A common problem in RL is that during the learning process, an agent rather uses an action that was promising in the past than explores the behavior of other actions. Consequently, the agent can get stuck in a local optimum. There are several approaches to this problem. We use the entropy regularization method to encourage exploration and thus, include a so-called regularization term in the EMV problem. We refer to Z. Ahmed et al. [2] for detailed information about entropy regularization.

So far, in Section 3.1, we have introduced a relaxed stochastic control formulation to model exploration and learning in RL through the randomized, distributional control process $\boldsymbol{\pi}$. However, if exploration and learning were not needed because the model is fully known, the control distributions would all degenerate to the Dirac measures (see Appendix A.3). Therefore, we would find ourselves in the setup of the last chapter. Thus, in the RL framework, we need to add a regularization term to account for model uncertainty and to encourage exploration.

We use Shannon's differential entropy to measure the level of exploration:

$$\mathcal{H}(\pi) := -\int_{\mathbb{R}} \pi(u) \ln\big(\pi(u)\big) du, \quad \pi \in \mathcal{P}(\mathbb{R}).$$

This concept from information theory extends the entropy, a measure for the average uncertainty of a random variable, to continuous probability distributions. For detailed information about Shannon's differential entropy, we refer to C. E. Shannon [22].

In this work, we only encounter Gaussian distributions whose density is denoted by:

$$\mathcal{N}(u|\mu,\sigma^2) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(u-\mu)^2}{\sigma^2}}, \quad u \in \mathbb{R},$$

with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Therefore, we state the differential entropy of a Gaussian density:

$$\mathcal{H}\big(\mathcal{N}(\cdot|\mu,\sigma^2)\big) = \ln\Big(\sigma\sqrt{2\pi e}\Big) = \ln\Big(\sigma\sqrt{2\pi}\Big) + \frac{1}{2}, \tag{3.7}$$

for the detailed derivation, see Appendix A.4. We note that the differential entropy of a Gaussian distribution (3.7) depends only on the variance. This makes total sense since the variance is another measure of uncertainty.

To measure the uncertainty of a distributional control process $\boldsymbol{\pi} = (\pi_t)_{t\in[0,T]}$, we further integrate the differential entropy over time and obtain the accumulative differential entropy:

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\pi}) := \int_0^T \mathcal{H}(\pi_t)dt = -\int_0^T \int_{\mathbb{R}} \pi_t(u) \ln\big(\pi_t(u)\big)dudt.$$

With this expression, we can regulate exploration in the problem formulation.

3.3 Formulation of the EMV Problem

We have all the tools introduced to formulate the entropy-regularized EMV problem. Nevertheless, first, we define the set of admissible distributional control processes associated with (3.6):

Definition 3.1. For $x_0 \in \mathbb{R}$ we denote the set of admissible control distributions by $\mathcal{A}^{\mathrm{E}}(0, x_0)$ and the distributional control process $\boldsymbol{\pi} = (\pi_t)_{t\in[0,T]}$ belongs to $\mathcal{A}^{\mathrm{E}}(0, x_0)$ if:

1. $\forall t \in [0, T] : \pi_t \in \mathcal{P}(\mathbb{R})$, $\mathbb{P}$-a.s.
2. $\forall A \in \mathcal{B}(\mathbb{R}) : \big(\int_A \pi_t(u)du\big)_{t\in[0,T]}$ is $\mathbb{F}$-progressively measurable
3. $\mathbb{E}\Big[\int_0^T (\mu_t^2 + \sigma_t^2)dt\Big] < \infty$
4. $\mathbb{E}\Big[\big|(x_T^{\boldsymbol{\pi}} - w)^2 + \lambda \int_0^T \int_{\mathbb{R}} \pi_t(u) \ln\big(\pi_t(u)\big)dudt\big| \Big| x_0^{\boldsymbol{\pi}} = x_0\Big] < \infty,$

where $\lambda > 0$ is a so-called temperature parameter (or exploration weight) reflecting the trade-off between exploitation and exploration.

The second and third points ensure that the stochastic integral for $\boldsymbol{\pi}$ and the induced wealth process $x^{\boldsymbol{\pi}}$ are well-defined. We can interpret the expectation in the fourth point as the expected absolute reward under the control process $\boldsymbol{\pi}$ which is by this condition well-defined and finite.

We note that the superscript $^{\mathrm{E}}$ refers to the exploratory MV problem and should prevent confusion with the definitions for the classical MV problem in Chapter 2.

Now, we finally state the entropy-regularized EMV problem for a fixed $w \in \mathbb{R}$:

$$\min_{\boldsymbol{\pi} \in \mathcal{A}^{\mathrm{E}}(0,x_0)} \mathbb{E}\left[(x_T^{\boldsymbol{\pi}} - w)^2 + \lambda \int_0^T \int_{\mathbb{R}} \pi_t(u) \ln\big(\pi_t(u)\big) du dt\right] - (w - z)^2, \tag{3.8}$$

where $\mathcal{A}^{\mathrm{E}}(0, x_0)$ is the set of admissible distributional control processes on $[0, T]$ and $\lambda > 0$ is the exploration weight.

If we compare the problem (3.8) with the formulation (2.6), the term with the double integral stands out. This expression is equal to $-\lambda \boldsymbol{\mathcal{H}}(\boldsymbol{\pi})$ and regularizes the entropy by negatively weighting the uncertainty of the distributional control process. Therefore, the weight λ enables emphasis on exploration.

We denote the optimal control process for problem (3.8) by $\hat{\boldsymbol{\pi}} = (\hat{\pi}_t)_{t\in[0,T]}$, which depends on w. As in the MV problem the Lagrange multiplier w can be determined by the constraint $\mathbb{E}[x_T^{\hat{\boldsymbol{\pi}}}] = z$.

3.4 Solution of the EMV Problem

We solve the problem (3.8) by using the same dynamic programming methods as in Section 2.4.

3.4.1 Value Functions

For this purpose, we define the admissible strategies and the value function for arbitrary initial time and state.

We consider the restricted wealth process $x^{\boldsymbol{\pi}} = (x_t^{\boldsymbol{\pi}})_{t\in[s,T]}$ controlled by the strategy $\boldsymbol{\pi} = (\pi_t)_{t\in[s,T]}$ as in (3.6) but for an initial time $s \in [0, T)$ and an initial wealth $y \in \mathbb{R}$:

$$\begin{aligned} dx_t^{\boldsymbol{\pi}} &= \rho\sigma\mu_t dt + \sigma\sqrt{\mu_t^2 + \sigma_t^2} dB_t, \quad s \le t \le T \\ x_s^{\boldsymbol{\pi}} &= y. \end{aligned}$$

Thus, we can define the set of admissible strategies for arbitrary initial time and state:

Definition 3.2. For $(s, y) \in [0, T) \times \mathbb{R}$ we denote the set of admissible control distributions as $\mathcal{A}^{\mathrm{E}}(s, y)$ and the distributional control process $\boldsymbol{\pi} = (\pi_t)_{t\in[s,T]}$ belongs to $\mathcal{A}^{\mathrm{E}}(s, y)$ if:

1. $\forall t \in [s, T] : \pi_t \in \mathcal{P}(\mathbb{R})$, $\mathbb{P}$-a.s.

2. $\forall A \in \mathcal{B}(\mathbb{R}) : \big(\int_A \pi_t(u)du\big)_{t\in[s,T]}$ is $\mathbb{F}$-progressively measurable
3. $\mathbb{E}\Big[\int_s^T (\mu_t^2 + \sigma_t^2)dt\Big] < \infty$
4. $\mathbb{E}\Big[\big|(x_T^{\boldsymbol{\pi}} - w)^2 + \lambda \int_s^T \int_{\mathbb{R}} \pi_t(u) \ln\big(\pi_t(u)\big)dudt\big|\Big|x_s^{\boldsymbol{\pi}} = y\Big] < \infty,$

where $\lambda > 0$ is the exploration weight.

As in Chapter 2, we call control processes also open-loop controls and represent them with deterministic functions, the feedback (or closed-loop) controls:

Definition 3.3. We denote the set of admissible feedback controls by $\mathcal{A}^{\mathrm{E}}$. A deterministic mapping $\pi(\cdot;\cdot,\cdot)$ is called an admissible feedback control if:

1. $\pi(\cdot;t,x)$ is a density for each $(t,x) \in [0,T] \times \mathbb{R}$.
2. For each $(s,y) \in [0,T) \times \mathbb{R}$ the following SDE:

$$\begin{aligned} dx_t^\pi &= \tilde{b}(\pi(\cdot;t,x_t^\pi))dt + \tilde{\sigma}(\pi(\cdot;t,x_t^\pi))dB_t, \quad t \in [s,T] \\ x_s^\pi &= y \end{aligned}$$

has a unique strong solution $(x_t^\pi)_{t\in[s,T]}$ and the open-loop control $\boldsymbol{\pi} = (\pi_t)_{t\in[s,T]} \in \mathcal{A}^{\mathrm{E}}(s,y)$ where $\pi_t := \pi(\cdot;t,x_t^\pi)$.

In this case, the open-loop control $\boldsymbol{\pi}$ is said to be generated from the feedback control $\pi(\cdot;\cdot,\cdot)$ with respect to the initial time and wealth (s,y).

We recapitulate Section 2.4.1 and note that in a feedback control $\pi(\cdot;t,x)$, the randomized control action depends on feedback from the process x^π, in the form of the value of the process variable $x_t^\pi = x$, and not on some initial state as in the open-loop control. Therefore, a feedback control can generate an open-loop control for any initial $(s,y) \in [0,T) \times \mathbb{R}$ and hence, it is in itself independent of any starting point.

Now, we define the value and optimal value function for problem (3.8):

Definition 3.4. For a fixed $w \in \mathbb{R}$, we define the value function under any given admissible feedback control $\pi(\cdot;\cdot,\cdot,w) \in \mathcal{A}^{\mathrm{E}}$:

$$J^{\mathrm{E}}(s,y;w,\pi) := \mathbb{E}\Big[(x_T^\pi - w)^2 + \lambda \int_s^T \int_{\mathbb{R}} \pi_t(u) \ln\big(\pi_t(u)\big)dudt\Big|x_s^\pi = y\Big] - (w-z)^2, \tag{3.9}$$

for $(s,y) \in [0,T) \times \mathbb{R}$, where $\boldsymbol{\pi} = (\pi_t)_{t\in[s,T]}$ is the admissible open-loop control generated from $\pi(\cdot;\cdot,\cdot,w)$ with respect to the initial (s,y) such that $\pi_t(\cdot) = \pi(\cdot;t,x_t^\pi,w)$. Further, we define the optimal value function as:

$$V^{\mathrm{E}}(s,y;w) := \inf_{\boldsymbol{\pi}\in\mathcal{A}^{\mathrm{E}}(s,y)} \mathbb{E}\Big[(x_T^{\boldsymbol{\pi}} - w)^2 + \lambda \int_s^T \int_{\mathbb{R}} \pi_t(u) \ln\big(\pi_t(u)\big) du dt \Big| x_s^{\boldsymbol{\pi}} = y\Big] - (w-z)^2, \tag{3.10}$$

for $(s,y) \in [0,T) \times \mathbb{R}$.

We note that the optimal feedback control law for the initial point $t = 0$ will be optimal for all subintervals and will be denoted by $\hat{\pi} = \hat{\pi}(\cdot;\cdot,\cdot,w)$.

By applying Bellman's optimality principle (see Appendix C.1), we can derive the following result:

Lemma 3.5. *For an arbitrary but fixed $w \in \mathbb{R}$ and any $(t,x) \in [0,T) \times \mathbb{R}$ and $h > 0$ such that $t + h \le T$, it holds for any $\pi \in \mathcal{A}^E$:*

$$V^E(t,x;w) \le \mathbb{E}\Big[V^E(t+h, x_{t+h}^\pi; w) + \lambda \int_t^{t+h} \int_{\mathbb{R}} \pi_s(u) \ln(\pi_s(u)) du ds \Big| x_t^\pi = x\Big],$$

where $\boldsymbol{\pi} = (\pi_t)_{t\in[s,T]} \in \mathcal{A}^E(t,x)$ is the open-loop control generated from π with respect to the initial (t,x) such that $\pi_s(u) = \pi(u;s,x_s^\pi,w)$ for $s \in [t,T]$.

Furthermore, we have equality if and only if the admissible feedback control π is the optimal control $\hat{\pi}$.

Proof. We choose an arbitrary pair $(t,x) \in [0,T) \times \mathbb{R}$ and $h > 0$ such that $t + h \le T$. Hence, we divide the time interval $[t,T]$ into two parts, the interval $[t,t+h]$ and $(t+h,T]$ respectively. Furthermore, we consider a fixed but arbitrary feedback control $\pi \in \mathcal{A}^{\mathrm{E}}$, and define the control law $\pi^* \in \mathcal{A}^{\mathrm{E}}$ by:

$$\pi^*(\cdot;s,y,w) = \begin{cases} \pi(\cdot;s,y,w) & (s,y) \in [t,t+h] \times \mathbb{R} \\ \hat{\pi}(\cdot;s,y,w) & (s,y) \in (t+h,T] \times \mathbb{R}, \end{cases}$$

where $\hat{\pi} \in \mathcal{A}^{\mathrm{E}}$ is the optimal feedback control. The interpretation of employing the feedback control $\hat{\pi}$ is using the arbitrary control π during the time interval $[t,t+h]$, and then switching to the optimal control law during the rest of the time period.

By applying the tower property (see Appendix C.2) to the value function under the feedback control π^* defined in (3.9), we obtain:

$$\begin{aligned}
J^{\mathrm{E}}(t,x;w,\pi^*) &= \mathbb{E}\bigg[(x_T^{\pi^*}-w)^2+\lambda\int_t^T\int_{\mathbb{R}}\pi_s^*(u)\ln\big(\pi_s^*(u)\big)duds\bigg|x_t^{\pi^*}=x\bigg]\\
&\quad-(w-z)^2\\
&= \mathbb{E}\bigg[\mathbb{E}\Big[(x_T^{\hat{\pi}}-w)^2+\lambda\int_{t+h}^T\int_{\mathbb{R}}\hat{\pi}_s(u)\ln\big(\hat{\pi}_s(u)\big)duds\Big|x_{t+h}^{\pi}\Big]\\
&\quad+\lambda\int_t^{t+h}\int_{\mathbb{R}}\pi_s(u)\ln\big(\pi_s(u)\big)duds\bigg|x_t^{\pi}=x\bigg]-(w-z)^2\\
&= \mathbb{E}\bigg[\mathbb{E}\Big[(x_T^{\hat{\pi}}-w)^2+\lambda\int_{t+h}^T\int_{\mathbb{R}}\hat{\pi}_s(u)\ln\big(\hat{\pi}_s(u)\big)duds\Big|x_{t+h}^{\pi}\Big]\\
&\quad-(w-z)^2+\lambda\int_t^{t+h}\int_{\mathbb{R}}\pi_s(u)\ln\big(\pi_s(u)\big)duds\bigg|x_t^{\pi}=x\bigg]\\
&\overset{(3.9)}{=} \mathbb{E}\bigg[J^{\mathrm{E}}(t+h,x_{t+h}^{\pi};w,\hat{\pi})\\
&\quad+\lambda\int_t^{t+h}\int_{\mathbb{R}}\pi_s(u)\ln\big(\pi_s(u)\big)duds\bigg|x_t^{\pi}=x\bigg],
\end{aligned}$$

where $(\pi_s^*)_{s\in[t,T]}$ is the open-loop control process generated from π^* with respect to the initial (t,x), $(\pi_s)_{s\in[t,T]}$ is the open-loop control process generated from π with respect to the initial (t,x) and $(\hat{\pi}_s)_{s\in[t+h,T]}$ is the open-loop control process generated from $\hat{\pi}$ with respect to the initial $(t+h,x_{t+h}^{\pi})$.

It follows from Bellman's optimality principle that we have:

$$\begin{aligned}
J^{\mathrm{E}}(t,x;w,\pi^*) &= \mathbb{E}\bigg[J^{\mathrm{E}}(t+h,x_{t+h}^{\pi};w,\hat{\pi})\\
&\qquad+\lambda\int_t^{t+h}\int_{\mathbb{R}}\pi_s(u)\ln\big(\pi_s(u)\big)duds\bigg|x_t^{\pi}=x\bigg]\\
&= \mathbb{E}\bigg[V^{\mathrm{E}}(t+h,x_{t+h}^{\pi};w)\\
&\qquad+\lambda\int_t^{t+h}\int_{\mathbb{R}}\pi_s(u)\ln\big(\pi_s(u)\big)duds\bigg|x_t^{\pi}=x\bigg]
\end{aligned}$$

and therefore, we can conclude that:

$$\begin{aligned}
V^{\mathrm{E}}(t,x;w) &\le J^{\mathrm{E}}(t,x;w,\pi^*)\\
&= \mathbb{E}\bigg[V^{\mathrm{E}}(t+h,x_{t+h}^{\pi};w)+\lambda\int_t^{t+h}\int_{\mathbb{R}}\pi_s(u)\ln\big(\pi_s(u)\big)duds\bigg|x_t^{\pi}=x\bigg].
\end{aligned}$$

Further, since $V^{\mathrm{E}}(t,x;w)=J^{\mathrm{E}}(t,x;w,\hat{\pi})$ we note that equality holds if and only if $\pi^*=\hat{\pi}$ and therefore if and only if $\pi=\hat{\pi}$. □

This lemma is valuable for the derivation of the HJB equation.

3.4.2 The Hamilton-Jacobi-Bellman Equation

Now, we state the HJB equation for the problem (3.8). The proof of this theorem is more or less analog to the proof of Theorem 2.14:

Theorem 3.6. ***(Hamilton-Jacobi-Bellman Equation II)**:*
We assume that there exists an admissible optimal feedback control $\hat{\pi}$ to problem (3.8) and that the optimal value function is regular in the sense that $V^E \in C^{1,2}([0,T]\times\mathbb{R})$. Then, for a fixed $w\in\mathbb{R}$, the following holds:

1. *V^E satisfies the HJB equation:*

$$\begin{aligned} v_t(t,x;w) + \min_{\pi\in\mathcal{P}(\mathbb{R})} \Big(&\frac{1}{2}\tilde{\sigma}^2(\pi)v_{xx}(t,x;w) + \tilde{b}(\pi)v_x(t,x;w) \\ &+ \lambda\int_{\mathbb{R}} \pi(u)\ln\big(\pi(u)\big)du\Big) = 0, \quad (t,x)\in[0,T)\times\mathbb{R} \\ v(T,x;w) &= (x-w)^2 - (w-z)^2, \quad x\in\mathbb{R}. \end{aligned} \tag{3.11}$$

2. *For each $(t,x)\in[0,T]\times\mathbb{R}$ the minimum in the HJB equation above is attained by the optimal feedback control $\hat{\pi}(\cdot;\cdot,\cdot,w)$ such that $\pi = \hat{\pi}(\cdot;t,x,w)$.*

Proof. Through the whole proof we fix $w\in\mathbb{R}$. Since $V^{\mathrm{E}}\in C^{1,2}([0,T]\times\mathbb{R})$ by assumption we can conclude for any $x\in\mathbb{R}$:

$$\lim_{t\to T} V^{\mathrm{E}}(t,x;w) = V^{\mathrm{E}}(T,x;w).$$

Therefore, the terminal condition in (3.11) follows directly from the definition of the value function (3.9):

$$\begin{aligned} V^{\mathrm{E}}(T,x;w) &= \inf_{\boldsymbol{\pi}\in\mathcal{A}(T,x)} \mathbb{E}\Big[(x_T^{\boldsymbol{\pi}}-w)^2 + \lambda\int_T^T\int_{\mathbb{R}} \pi_s(u)\ln\big(\pi_s(u)\big)duds\Big|x_T^{\boldsymbol{\pi}} = x\Big] \\ &\quad - (w-z)^2 \\ &= \inf_{\boldsymbol{\pi}\in\mathcal{A}(T,x)} \big((x-w)^2 - (w-z)^2\big) \\ &= (x-w)^2 - (w-z)^2, \quad x\in\mathbb{R}. \end{aligned}$$

Furthermore, we choose $(t,x)\in[0,T)\times\mathbb{R}$ arbitrary and divide the interval $[t,T]$ into two parts $[t,t+h]$ and $(t+h,T]$ respectively for $h>0$. By Lemma 3.5, we know that for any admissible feedback control $\pi\in\mathcal{A}^{\mathrm{E}}$, it holds:

$$\mathbb{E}\Big[V^{\mathrm{E}}(t+h,x_{t+h}^{\pi};w) + \lambda\int_t^{t+h}\int_{\mathbb{R}} \pi_s(u)\ln\big(\pi_s(u)\big)duds\Big|x_t^{\pi} = x\Big] \geq V^{\mathrm{E}}(t,x;w), \tag{3.12}$$

where $\boldsymbol{\pi} = (\pi_s)_{s\in[t,T]}$ is the admissible open-loop control generated from π with respect to the initial (t,x) such that $\pi_s(u) = \pi(u; s, x_s^\pi, w)$.

We note that Lemma 3.5 implies equality in (3.12) if and only if $\pi \in \mathcal{A}^{\mathrm{E}}$ is the optimal feedback control $\hat{\pi} \in \mathcal{A}^{\mathrm{E}}$.

Since $V^{\mathrm{E}} \in C^{1,2}([0,T]\times\mathbb{R})$ and $\pi \in \mathcal{A}^{\mathrm{E}}(t,x)$ we are able to use Itô's formula (see Appendix B.3) and write the optimal value function in the following way:

$$\begin{aligned} V^{\mathrm{E}}(t+h, x_{t+h}^\pi; w) = V^{\mathrm{E}}(t, x_t^\pi; w) &+ \int_t^{t+h} V_t^{\mathrm{E}}(s, x_s^\pi; w) \\ &+ \tilde{b}(\pi_s) V_x^{\mathrm{E}}(s, x_s^\pi; w) + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w) ds \\ &+ \int_t^{t+h} \tilde{\sigma}(\pi_s) V_x^{\mathrm{E}}(s, x_s^\pi; w) dB_s, \end{aligned} \tag{3.13}$$

where $\tilde{b}(\cdot)$ and $\tilde{\sigma}(\cdot)$ are given by (3.4) and (3.5), respectively.

Further, we show that $(\tilde{\sigma}(\hat{\pi}_s) V_x^{\mathrm{E}}(s, x_s^\pi; w))_{s\in[t,T]} \in \mathcal{L}^2_{\mathrm{loc}}([t,T], B)$ by proceeding as explained in the second point in Remark 2.10.

Since $V^{\mathrm{E}} \in C^{1,2}([0,T]\times\mathbb{R})$ and the wealth process x^π has continuous paths, the pasth $s \to V_x^{\mathrm{E}}(s, x_s^\pi; w)$ is bounded on $[t,T]$ and we can conclude that:

$$\begin{aligned} \int_t^T \big(\tilde{\sigma}(\pi_s)\big)^2 \big(V_x^{\mathrm{E}}(s, x_s^\pi; w)\big)^2 ds &\le C \cdot \int_t^T \big(\tilde{\sigma}(\pi_s)\big)^2 ds \\ &= C\sigma^2 \cdot \int_t^T \mu_s^2 + \sigma_s^2 ds < \infty, \quad \mathbb{P}\text{-a.s.}, \end{aligned}$$

where $0 \le C < \infty$ is a constant and where we used in the last step that $\boldsymbol{\pi}$ is admissible. Therefore, we have $(\tilde{\sigma}(\pi_s) V_x^{\mathrm{E}}(s, x_s^\pi; w))_{s\in[t,T]} \in \mathcal{L}^2_{\mathrm{loc}}([t,T], B)$. We define the stopping times for $n \ge 1$:

$$\tau_n := \inf\left\{\tau \ge t : \int_t^\tau \big(\tilde{\sigma}(\pi_s)\big)^2 \big(V_x^{\mathrm{E}}(s, x_s^\pi; w)\big)^2 ds \ge n\right\}.$$

By Lemma 2.11, this sequence of stopping times is a localising sequence for $(\tilde{\sigma}(\pi_s) V_x^{\mathrm{E}}(s, x_s^\pi; w))_{s\in[t,T]}$ in $\mathcal{L}^2([t,T], B)$.

Then, from (3.13), we obtain:

$$\begin{aligned} &V^{\mathrm{E}}\big((t+h)\wedge\tau_n, x^\pi_{(t+h)\wedge\tau_n}; w\big) \\ &= V(t, x_t^\pi; w) + \int_t^{(t+h)\wedge\tau_n} V_t^{\mathrm{E}}(s, x_s^\pi; w) + \tilde{b}(\pi_s) V_x^{\mathrm{E}}(s, x_s^\pi; w) \\ &\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w) ds + \int_t^{(t+h)\wedge\tau_n} \tilde{\sigma}(\pi_s) V_x^{\mathrm{E}}(s, x_s^\pi; w) dB_s. \end{aligned}$$

By taking the conditional expectation given $x_t^\pi = x$ and using Lemma 2.13, we have that:

$$\begin{aligned}
&\mathbb{E}\Big[V^{\mathrm{E}}\big((t+h)\wedge\tau_n, x^\pi_{(t+h)\wedge\tau_n}; w\big)\Big|x_t^\pi = x\Big] \\
&= \mathbb{E}\Big[V^{\mathrm{E}}(t, x_t^\pi; w)\Big|x_t^\pi = x\Big] + \mathbb{E}\Big[\int_t^{(t+h)\wedge\tau_n} V_t^{\mathrm{E}}(s, x_s^\pi; w) + \tilde{b}(\pi_s)V_x^{\mathrm{E}}(s, x_s^\pi; w) \\
&\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w)ds\Big|x_t^\pi = x\Big] \\
&\quad + \mathbb{E}\Big[\int_t^{(t+h)\wedge\tau_n} \tilde{\sigma}(\pi_s)V_x^{\mathrm{E}}(s, x_s^\pi; w)dB_s\Big|x_t^\pi = x\Big] \\
&= V^{\mathrm{E}}(t, x; w) + \mathbb{E}\Big[\int_t^{(t+h)\wedge\tau_n} V_t^{\mathrm{E}}(s, x_s^\pi; w) + \tilde{b}(\pi_s)V_x^{\mathrm{E}}(s, x_s^\pi; w) \\
&\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w)ds\Big|x_t^\pi = x\Big].
\end{aligned}$$

Since $(\tau_n)_{n\geq 1}$ is a localising sequence and by using the dominated convergence theorem (see Appendix B.4), we obtain for $n \to \infty$:

$$\begin{aligned}
\mathbb{E}\Big[V^{\mathrm{E}}(t+h, x_{t+h}^\pi; w)\Big|x_t^\pi = x\Big] &= V^{\mathrm{E}}(t, x; w) \\
&\quad + \mathbb{E}\Big[\int_t^{t+h} V_t^{\mathrm{E}}(s, x_s^\pi; w) + \tilde{b}(\pi_s)V_x^{\mathrm{E}}(s, x_s^\pi; w) \\
&\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w)ds\Big|x_t^\pi = x\Big].
\end{aligned}$$

By substituting this formulation into (3.12), we obtain:

$$\begin{aligned}
V^{\mathrm{E}}(t, x; w) \leq\ & V^{\mathrm{E}}(t, x; w) \\
&+ \mathbb{E}\Big[\int_t^{t+h} V_t^{\mathrm{E}}(s, x_s^\pi; w) + \tilde{b}(\pi_s)V_x^{\mathrm{E}}(s, x_s^\pi; w) \\
&+ \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w)ds \\
&+ \lambda\int_t^{t+h}\int_{\mathbb{R}} \pi_s(u)\ln\big(\pi_s(u)\big)duds\Big|x_t^\pi = x\Big].
\end{aligned}$$

By subtracting V^{E} on both sides, we have:

$$\begin{aligned}
0 \leq\ & \mathbb{E}\Big[\int_t^{t+h} V_t^{\mathrm{E}}(s, x_s^\pi; w) + \tilde{b}(\pi_s)V_x^{\mathrm{E}}(s, x_s^\pi; w) \\
&+ \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w)ds \\
&+ \lambda\int_t^{t+h}\int_{\mathbb{R}} \pi_s(u)\ln\big(\pi_s(u)\big)duds\Big|x_t^\pi = x\Big].
\end{aligned}$$

Further, we divide by h and let $h \to 0$ to obtain:

$$\begin{aligned}
0 \leq \mathbb{E}\Bigg[& V_t^{\mathrm{E}}(s, x_t^\pi; w) + \tilde{b}(\pi_t) V_x^{\mathrm{E}}(s, x_t^\pi; w) \\
& + \frac{1}{2}\big(\tilde{\sigma}(\pi_t)\big)^2 V_{xx}^{\mathrm{E}}(s, x_t^\pi; w) + \lambda \int_{\mathbb{R}} \pi_t(u) \ln\big(\pi_t(u)\big) du \Bigg| x_t^\pi = x \Bigg] \\
= & V_t^{\mathrm{E}}(s, x; w) + \tilde{b}\big(\pi(\cdot; t, x, w)\big) V_x^{\mathrm{E}}(s, x; w) \\
& + \frac{1}{2}\Big(\tilde{\sigma}\big(\pi(\cdot; t, x, w)\big)\Big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w) + \lambda \int_{\mathbb{R}} \pi(u; t, x, w) \ln\big(\pi(u; t, x, w)\big) du.
\end{aligned}$$

Considering that we have equality for $\pi(\cdot; t, x, w) = \hat{\pi}(\cdot; t, x, w)$ and $\pi \in \mathcal{A}^{\mathrm{E}}$ was arbitrary chosen, we can conclude that:

$$\begin{aligned}
0 = & V_t^{\mathrm{E}}(s, x_s^\pi; w) + \tilde{b}\big(\hat{\pi}(\cdot; t, x, w)\big) V_x^{\mathrm{E}}(s, x_s^\pi; w) \\
& + \frac{1}{2}\Big(\tilde{\sigma}\big(\hat{\pi}(\cdot; t, x, w)\big)\Big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w) + \lambda \int_{\mathbb{R}} \hat{\pi}(u; t, x, w) \ln\big(\hat{\pi}(u; t, x, w)\big) du \\
= & V_t^{\mathrm{E}}(s, x_s^\pi; w) + \min_{\pi \in \mathcal{P}(\mathbb{R})} \Big(\tilde{b}(\pi) V_x^{\mathrm{E}}(s, x_s^\pi; w) \\
& + \frac{1}{2}\big(\tilde{\sigma}(\pi)\big)^2 V_{xx}^{\mathrm{E}}(s, x_s^\pi; w) + \lambda \int_{\mathbb{R}} \pi(u) \ln\big(\pi(u)\big) du \Big).
\end{aligned}$$

Since the choice of $(t, x) \in [0, T) \times \mathbb{R}$ was arbitrary, we conclude that the optimal value function V^{E}, given that $V^{\mathrm{E}} \in C^{1,2}([0, T] \times \mathbb{R})$, satisfies the HJB equation (3.11) and that the minimum is attained by $\pi = \hat{\pi}(\cdot; t, x, w)$. □

Remark 3.7. By using (3.4) and (3.5), we obtain the following equivalent HJB equation:

$$\begin{aligned}
& v_t(t, x; w) + \min_{\pi \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \Big(\frac{1}{2}\sigma^2 u^2 v_{xx}(t, x; w) + \rho\sigma u v_x(t, x; w) \\
& \qquad + \lambda \ln\big(\pi(u)\big) \Big) \pi(u) du = 0, \quad (t, x) \in [0, T) \times \mathbb{R} \\
& v(T, x; w) = (x - w)^2 - (w - z)^2, \quad x \in \mathbb{R}.
\end{aligned} \tag{3.14}$$

3.4.3 The Verification Theorem

As in Section 2.4.4, we state the Verification theorem for dynamic programming which will help us to determine the value function and the optimal control. The proof is very similar to the proof of Theorem 2.15:

Theorem 3.8. ***(Verification Theorem II)**:*
Suppose that $w \in \mathbb{R}$ is fixed and that we have two functions $v \in C^{1,2}([0, T] \times \mathbb{R})$ and g such that:

1. *v is sufficiently integrable (see Remark 3.9), and solves the HJB equation (3.11) (or equivalently (3.14)).*
2. *The function g is admissible in the sense of Definition 3.3.*
3. *For each fixed $(t,x) \in [0,T) \times \mathbb{R}$ the minimum in the HJB equation (3.11) (or equivalently (3.14)):*

$$\min_{\pi \in \mathcal{P}(\mathbb{R})} \left(\frac{1}{2}\tilde{\sigma}^2(\pi) v_{xx}(t,x;w) + \tilde{b}(\pi) v_x(t,x;w) + \lambda \int_{\mathbb{R}} \pi(u) \ln\big(\pi(u)\big) du \right)$$

is attained by the choice $\pi = g(\cdot;t,x,w)$.

Then the following hold:

1. *The optimal value function V^E is given by:*

$$V^E(t,x;w) = v(t,x;w), \quad (t,x) \in [0,T) \times \mathbb{R}.$$

2. *There exists an optimal feedback control $\hat{\pi}$, and in fact $\hat{\pi}(\cdot;t,x,w) = g(\cdot;t,x,w)$.*

Proof. We assume that v and g are given as above and we fix an arbitrary point $(t,x) \in [0,T) \times \mathbb{R}$ and $w \in \mathbb{R}$. Further, we choose an arbitrary admissible feedback control $\pi = \pi(\cdot;\cdot,\cdot,w) \in \mathcal{A}^{\mathrm{E}}$ and consider the process $x^\pi = (x_s^\pi)_{s\in[t,T]}$ with following dynamics as in Definition 3.3:

$$\begin{aligned} dx_s^\pi &= \tilde{b}\big(\pi(\cdot;s,x_s^\pi,w)\big)ds + \tilde{\sigma}\big(\pi(\cdot;s,x_s^\pi,w)\big)dB_s, \quad t \le s \le T \\ x_t^\pi &= x. \end{aligned}$$

By inserting the process x^π into the function v and using Itô's formula (see Appendix B.3), we obtain:

$$\begin{aligned} v(T,x_T^\pi;w) = v(t,x_t^\pi;w) &+ \int_t^T \Big(v_t(s,x_s^\pi;w) + \tilde{b}\big(\pi(\cdot;s,x_s^\pi,w)\big) v_x(s,x_s^\pi;w) \\ &+ \frac{1}{2}\Big(\tilde{\sigma}\big(\pi(\cdot;s,x_s^\pi,w)\big)\Big)^2 v_{xx}(s,x_s^\pi;w)\Big) ds \\ &+ \int_t^T \tilde{\sigma}\big(\pi(\cdot;s,x_s^\pi,w)\big) v_x(s,x_s^\pi;w) dB_s \end{aligned} \tag{3.15}$$

Analog to the proof of Theorem 3.6, we define the stopping times for $n \ge 1$:

$$\tau_n := \inf\left\{ \tau \ge t : \int_t^\tau \Big(\tilde{\sigma}\big(\pi(\cdot;s,x_s^\pi,w)\big)\Big)^2 \big(v_x(s,x_s^\pi;w)\big)^2 ds \ge n \right\}$$

and use (3.15) to obtain:

$$v(T\wedge\tau_n,x_T^\pi;w)=v(t,x_t^\pi;w)+\int_t^{T\wedge\tau_n}\Big(v_t(s,x_s^\pi;w)+\tilde b\big(\pi(\cdot;s,x_s^\pi,w)\big)v_x(s,x_s^\pi;w)\\+\frac{1}{2}\Big(\tilde\sigma\big(\pi(\cdot;s,x_s^\pi,w)\big)\Big)^2v_{xx}(s,x_s^\pi;w)\Big)ds\\+\int_t^{T\wedge\tau_n}\tilde\sigma\big(\pi(\cdot;s,x_s^\pi,w)\big)v_x(s,x_s^\pi;w)dB_s$$

We take the conditional expectation given $x_t^\pi=x$:

$$\begin{aligned}&\mathbb{E}[v(T\wedge\tau_n,x_T^\pi;w)|x_t^\pi=x]\\&=v(t,x;w)+\mathbb{E}\bigg[\int_t^{T\wedge\tau_n}\Big(v_t(s,x_s^\pi;w)+\tilde b\big(\pi(\cdot;s,x_s^\pi,w)\big)v_x(s,x_s^\pi;w)\\&\quad+\frac{1}{2}\Big(\tilde\sigma\big(\pi(\cdot;s,x_s^\pi,w)\big)\Big)^2v_{xx}(s,x_s^\pi;w)\Big)ds\bigg|x_t^\pi=x\bigg]\\&\quad+\mathbb{E}\bigg[\int_t^{T\wedge\tau_n}\tilde\sigma\big(\pi(\cdot;s,x_s^\pi,w)\big)v_x(s,x_s^\pi;w)dB_s\bigg|x_t^\pi=x\bigg].\end{aligned}\tag{3.16}$$

By Remark 3.9 and Lemma 2.11, we can conclude that the sequence of stopping times $(\tau_n)_{n\geq1}$ is a localising sequence for $(\tilde\sigma(\pi(\cdot;s,x_s^\pi,w))v_x(s,x_s^\pi;w))_{s\in[t,T]}$ in $\mathcal{L}^2([t,T],B)$ and therefore, the expectation of the stochastic integral vanishes due to Lemma 2.13:

$$\begin{aligned}&\mathbb{E}[v(T\wedge\tau_n,x_T^\pi;w)|x_t^\pi=x]\\&=v(t,x;w)+\mathbb{E}\bigg[\int_t^{T\wedge\tau_n}\Big(v_t(s,x_s^\pi;w)+\tilde b\big(\pi(\cdot;s,x_s^\pi,w)\big)v_x(s,x_s^\pi;w)\\&\quad+\frac{1}{2}\Big(\tilde\sigma\big(\pi(\cdot;s,x_s^\pi,w)\big)\Big)^2v_{xx}(s,x_s^\pi;w)\Big)ds\bigg|x_t^\pi=x\bigg].\end{aligned}$$

By the properties of the localising sequence in Definition 2.8 and by using the dominated convergence theorem (see Appendix B.4), we obtain for $n\to\infty$:

$$\begin{aligned}&\mathbb{E}[v(T,x_T^\pi;w)|x_t^\pi=x]\\&=v(t,x;w)+\mathbb{E}\bigg[\int_t^{T}\Big(v_t(s,x_s^\pi;w)+\tilde b\big(\pi(\cdot;s,x_s^\pi,w)\big)v_x(s,x_s^\pi;w)\\&\quad+\frac{1}{2}\Big(\tilde\sigma\big(\pi(\cdot;s,x_s^\pi,w)\big)\Big)^2v_{xx}(s,x_s^\pi;w)\Big)ds\bigg|x_t^\pi=x\bigg].\end{aligned}\tag{3.17}$$

Since v solves the HJB equation (3.11), we obtain for any density function $\pi(\cdot)\in\mathcal{P}(\mathbb{R})$:

$$v_t(s,y;w)+\frac{1}{2}\Big(\tilde\sigma\big(\pi(\cdot)\big)\Big)^2v_{xx}(s,y;w)+\tilde b\big(\pi(\cdot)\big)v_x(s,y;w)\\+\lambda\int_{\mathbb{R}}\pi(u)\ln\big(\pi(u)\big)du\geq0,\quad(s,y)\in[0,T)\times\mathbb{R}.$$

By integrating from t to T, we see that:

$$\int_t^T \Big(v_t(s,y;w) + \frac{1}{2}\big(\tilde{\sigma}\big(\pi(\cdot)\big)\big)^2 v_{xx}(s,y;w) + \tilde{b}\big(\pi(\cdot)\big) v_x(s,y;w) + \lambda \int_{\mathbb{R}} \pi(u) \ln\big(\pi(u)\big) du \Big) ds \geq 0 \tag{3.18}$$

and by substitute (3.18) with $\pi(\cdot) = \pi(\cdot; s, x_s^\pi, w)$ and $y = x_s^\pi$ into (3.17), we obtain following inequality:

$$\begin{aligned} v(t,x;w) \leq\ & \mathbb{E}[v(T, x_T^\pi; w) | x_t^\pi = x] \\ & + \mathbb{E}\bigg[\lambda \int_t^T \int_{\mathbb{R}} \pi(u; s, x_s^\pi, w) \ln\big(\pi(u; s, x_s^\pi, w)\big) du ds \bigg| x_t^\pi = x \bigg]. \end{aligned}$$

Furthermore, we use the terminal condition in (3.11) and obtain:

$$\begin{aligned} v(t,x;w) &\leq \mathbb{E}[(x_T^\pi - w)^2 - (w-z)^2 | x_t^\pi = x] \\ &\quad + \mathbb{E}\bigg[\lambda \int_t^T \int_{\mathbb{R}} \pi(u; s, x_s^\pi, w) \ln\big(\pi(u; s, x_s^\pi, w)\big) du ds \bigg| x_t^\pi = x \bigg] \\ &= E\bigg[(x_T^\pi - w)^2 + \lambda \int_t^T \int_{\mathbb{R}} \pi_s(u) \ln\big(\pi_s(u)\big) du ds \bigg| x_t^\pi = x \bigg] - (w-z)^2 \\ &\overset{(3.9)}{=} J(t,x;w,\pi), \end{aligned}$$

where $\boldsymbol{\pi} = (\pi_s)_{s\in[t,T]}$ is the open-loop control generated from the feedback control π with respect to the initial (t,x). Since the feedback control $\pi \in \mathcal{A}^{\mathrm{E}}$ is chosen arbitrary, we obtain:

$$v(t,x;w) \leq V^{\mathrm{E}}(t,x;w). \tag{3.19}$$

To obtain the reverse inequality we choose $\pi(\cdot; t, x, w) = g(\cdot; t, x, w)$. Since this choice minimizes the second term in the HJB equation and v solves (3.11) by assumption, we have for $(s,y) \in [0,T) \times \mathbb{R}$:

$$\begin{aligned} v_t(s,y;w) + \frac{1}{2}\big(\tilde{\sigma}(g(\cdot;s,y,w))\big)^2 v_{xx}(s,y;w) + \tilde{b}\big(g(\cdot;s,y,w)\big) v_x(s,y;w) \\ + \lambda \int_{\mathbb{R}} g(u;s,y,w) \ln\big(g(u;s,y,w)\big) du = 0. \end{aligned}$$

Again, by proceeding as above, we obtain:

$$v(t,x;w) = J^{\mathrm{E}}(t,x;w,g) \geq V^{\mathrm{E}}(t,x;w). \tag{3.20}$$

By combining (3.19) and (3.20), we can finally conclude that:

$$v(t,x;w) = V^{\mathrm{E}}(t,x;w) = J^{\mathrm{E}}(t,x;w,g).$$

Since $(t,x) \in [0,T) \times \mathbb{R}$ was chosen arbitrary we have $v(t,x;w) = V^{\mathrm{E}}(t,x;w)$ and $\hat{\pi}(\cdot;t,x,w) = g(\cdot;t,x,w)$ for all $(t,x) \in [0,T) \times \mathbb{R}$. □

Remark 3.9. The assumption that v is sufficiently integrable in Theorem 3.8 above is made in order for the expectation of the stochastic integral in (3.16) to vanish. By Lemma 2.13, this will be the case if v satisfies the condition:

$$\Big(\tilde{\sigma}\big(\pi(\cdot;s,x_s^\pi,w)\big)v_x(s,x_s^\pi;w)\Big)_{s\in[t,T]} \in \mathcal{L}^2_{\text{loc}}([t,T],B).$$

3.4.4 Solution

Finally we can show that for a fixed $w \in \mathbb{R}$ the optimal value function V^{E} is given by:

$$\begin{aligned} V^{\text{E}}(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda\rho^2}{4}(T^2-t^2) \\ - \frac{\lambda}{2}\Big(\rho^2 T - \ln\big(\frac{\sigma^2}{\pi\lambda}\big)\Big)(T-t) - (w-z)^2, \quad (t,x)\in[0,T]\times\mathbb{R} \end{aligned} \tag{3.21}$$

and the optimal feedback control $\hat{\pi}(\cdot;\cdot,\cdot,w)$ is Gaussian, with its density function given by:

$$\hat{\pi}(u;t,x,w) = \mathcal{N}\Big(u\Big| -\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}\Big), \quad (t,x)\in[0,T]\times\mathbb{R}. \tag{3.22}$$

However, first, we investigate the wealth process induced by $\hat{\pi}$.

For this purpose, we consider the following dynamics for any $(s,y)\in[0,T)\times\mathbb{R}$:

$$\begin{aligned} dx_t^{\hat{\pi}} &= \tilde{b}\big(\hat{\pi}(\cdot;t,x_t^{\hat{\pi}},w)\big)dt + \tilde{\sigma}\big(\hat{\pi}(\cdot;t,x_t^{\hat{\pi}},w)\big)dB_t, \quad t\in[s,T] \\ x_s^{\hat{\pi}} &= y. \end{aligned} \tag{3.23}$$

We can compute the exploratory drift and volatility from (3.4) and (3.5):

$$\begin{aligned} \tilde{b}\big(\hat{\pi}(\cdot;t,x_t^{\hat{\pi}},w)\big) &= \int_{\mathbb{R}} \rho\sigma u\hat{\pi}(u;t,x_t^{\hat{\pi}},w)du = \rho\sigma\int_{\mathbb{R}} u\hat{\pi}(u;t,x_t^{\hat{\pi}},w)du \\ &= \rho\sigma\mathbb{E}_{\hat{\pi}}[U] = -\rho\sigma\frac{\rho}{\sigma}(x_t^{\hat{\pi}}-w) = -\rho^2(x_t^{\hat{\pi}}-w) \end{aligned}$$

$$\begin{aligned} \tilde{\sigma}\big(\hat{\pi}(\cdot;t,x_t^{\hat{\pi}},w)\big) &= \sqrt{\int_{\mathbb{R}} \sigma^2u^2\hat{\pi}(u;t,x_t^{\hat{\pi}},w)du} = \sigma\sqrt{\int_{\mathbb{R}} u^2\hat{\pi}(u;t,x_t^{\hat{\pi}},w)du} \\ &= \sigma\sqrt{\mathbb{E}_{\hat{\pi}_t}[U^2]} = \sigma\sqrt{\mathbb{E}_{\hat{\pi}_t}[U]^2 + \text{Var}_{\hat{\pi}_t}(U)} \\ &= \sigma\sqrt{\frac{\rho^2}{\sigma^2}(x_t^{\hat{\pi}}-w)^2 + \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}} \\ &= \sqrt{\rho^2(x_t^{\hat{\pi}}-w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}}, \end{aligned}$$

where we denote by $\mathbb{E}_\pi[U]$ the expectation and by $\mathrm{Var}_\pi(U)$ the variance with respect to a density $\pi \in \mathcal{P}(\mathbb{R})$:

$$\mathbb{E}_\pi[U] = \int_{\mathbb{R}} u\pi_t(u)du$$
$$\mathrm{Var}_\pi(U) = \int_{\mathbb{R}} (u - \mathbb{E}_\pi[U])^2\pi(u)du.$$

Consequently, we obtain the following SDE for (3.23):

$$\begin{aligned} dx_t^{\hat{\pi}} &= -\rho^2(x_t^{\hat{\pi}} - w)dt + \sqrt{\rho^2(x_t^{\hat{\pi}} - w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}}dB_t, \quad s \le t \le T \\ x_s^{\hat{\pi}} &= y. \end{aligned} \tag{3.24}$$

By using standard procedure, one can verify that (3.24) has a unique solution with continuous paths that is adapted and:

$$\mathbb{E}\left[\int_s^T \left(x_t^{\hat{\pi}}\right)^2 dt\right] < \infty. \tag{3.25}$$

For the detailed proof, we refer to Lemma B.9 in Appendix B.6.

Furthermore, we compute the first two moments of the wealth process $x^{\hat{\pi}}$:

It follows from (3.24) that:

$$x_t^{\hat{\pi}} = x_s^{\hat{\pi}} + \int_s^t -\rho^2(x_v^{\hat{\pi}} - w)dv + \int_s^t \sqrt{\rho^2(x_v^{\hat{\pi}} - w)^2 + \frac{\lambda}{2}e^{\rho^2(T-v)}}dB_v, \quad t \in [s,T].$$

Since $x^{\hat{\pi}}$ has continuous paths and $f(v,x) := \sqrt{\rho^2(x-w)^2 + \frac{\lambda}{2}e^{\rho^2(T-v)}}$ is continuous, the function $v \mapsto f(v, x_v^{\hat{\pi}})$ is bounded on $[s,T]$ and hence, by Remark 2.10, the corresponding stochastic process is in $\mathcal{L}^2_{\mathrm{loc}}([s,T],B)$. Further, by Lemma 2.11, we can define a localising sequence $(\tau_n)_{n\ge1}$ for $(f(t,x_t^{\hat{\pi}}))_{t\in[s,T]}$ in $\mathcal{L}^2([s,T],B)$.

By using this localising sequence, the conditional expectation given $x_s^{\hat{\pi}} = y$ and Lemma 2.13, we obtain for $t \in [s,T]$:

$$\begin{aligned} \mathbb{E}\left[x_{t\wedge\tau_n}^{\hat{\pi}}\middle| x_s^{\hat{\pi}} = y\right] &= \mathbb{E}\Bigg[x_s^{\hat{\pi}} + \int_s^{t\wedge\tau_n} -\rho^2(x_v^{\hat{\pi}} - w)dv \\ &\quad + \int_s^{t\wedge\tau_n} \sqrt{\rho^2(x_v^{\hat{\pi}} - w)^2 + \frac{\lambda}{2}e^{\rho^2(T-v)}}dB_v\middle| x_s^{\hat{\pi}} = y\Bigg] \\ &= y + \mathbb{E}\left[\int_s^{t\wedge\tau_n} -\rho^2(x_v^{\hat{\pi}} - w)dv\middle| x_s^{\hat{\pi}} = y\right] \\ &\quad + \mathbb{E}\left[\int_s^{t\wedge\tau_n} \sqrt{\rho^2(x_v^{\hat{\pi}} - w)^2 + \frac{\lambda}{2}e^{\rho^2(T-v)}}dB_v\middle| x_s^{\hat{\pi}} = y\right] \\ &= y + \mathbb{E}\left[\int_s^{t\wedge\tau_n} -\rho^2(x_v^{\hat{\pi}} - w)dv\middle| x_s^{\hat{\pi}} = y\right]. \end{aligned}$$

Furthermore, we apply the dominated convergence theorem (see Appendix B.4) for $n \to \infty$ and obtain:

$$\mathbb{E}\Big[x_t^{\hat{\pi}}\Big|x_s^{\hat{\pi}} = y\Big] = y + \mathbb{E}\bigg[\int_s^t -\rho^2(x_v^{\hat{\pi}} - w)dv\bigg|x_s^{\hat{\pi}} = y\bigg].$$

By Fubini's theorem (Theorem B.7), we have:

$$\mathbb{E}\Big[x_t^{\hat{\pi}}\Big|x_s^{\hat{\pi}} = y\Big] = y + \int_s^t -\rho^2(\mathbb{E}\Big[x_v^{\hat{\pi}}\Big|x_s^{\hat{\pi}} = y\Big] - w)dv, \quad t \in [s,T].$$

This yields the ordinary differential equation (ODE) for $N(t) := \mathbb{E}[x_t^{\hat{\pi}}|x_s^{\hat{\pi}} = y]$:

$$\begin{aligned} dN(t) &= -\rho^2(N(t) - w)dt, \quad t \in [s,T] \\ N(s) &= y, \end{aligned}$$

which can be solved using integrating factors:

$$\mathbb{E}[x_t^{\hat{\pi}}|x_s^{\hat{\pi}} = y] = (y - w)e^{-\rho^2(t-s)} + w, \quad t \in [s,T]. \tag{3.26}$$

For the second moment, we first use Itô's formula (see Appendix B.3) and (3.24), to obtain for the squared process $((x_t^{\hat{\pi}})^2)_{t\in[s,T]}$:

$$\begin{aligned} \big(x_t^{\hat{\pi}}\big)^2 &= y^2 + \int_s^t -2x_v^{\hat{\pi}}\rho^2(x_v^{\hat{\pi}} - w) + \rho^2(x_v^{\hat{\pi}} - w)^2 + \frac{\lambda}{2}e^{\rho^2(T-v)}dv \\ &\quad + \int_s^t 2x_v^{\hat{\pi}}\sqrt{\rho^2(x_v^{\hat{\pi}} - w)^2 + \frac{\lambda}{2}e^{\rho^2(T-v)}}dB_v \\ &= y^2 + \int_s^t \rho^2(w^2 - \big(x_v^{\hat{\pi}}\big)^2) + \frac{\lambda}{2}e^{\rho^2(T-v)}dv \\ &\quad + \int_s^t 2x_v^{\hat{\pi}}\sqrt{\rho^2(x_v^{\hat{\pi}} - w)^2 + \frac{\lambda}{2}e^{\rho^2(T-v)}}dB_v, \quad t \in [s,T]. \end{aligned}$$

And again, we proceed as above and obtain:

$$\begin{aligned} \mathbb{E}\Big[\big(x_t^{\hat{\pi}}\big)^2|x_s^{\hat{\pi}} = y\Big] &= y^2 + \int_s^t \rho^2\Big(w^2 - \big(\mathbb{E}\big[x_v^{\hat{\pi}}|x_s^{\hat{\pi}} = y\big]\big)^2\Big) \\ &\quad + \frac{\lambda}{2}e^{\rho^2(T-v)}dv, \quad t \in [s,T]. \end{aligned}$$

This yields the ODE for $M(t) := \mathbb{E}[(x_t^{\hat{\pi}})^2|x_s^{\hat{\pi}} = y]$:

$$\begin{aligned} dM(t) &= \Big(-\rho^2(M(t) - w^2) + \frac{\lambda}{2}e^{\rho^2(T-t)}\Big)dt, \quad t \in [s,T] \\ M(s) &= y^2, \end{aligned}$$

which again can be solved using integrating factors:

$$\mathbb{E}\Big[\big(x_t^{\hat{\pi}}\big)^2 \big| x_s^{\hat{\pi}} = y\Big] = (y^2 - w^2)e^{-\rho^2(t-s)} + w^2 + \frac{\lambda}{2}(t-s)e^{\rho^2(T-t)}, \quad t \in [s,T]. \tag{3.27}$$

By combining (3.26) and (3.27), we obtain a result that we use in proofs in this section but that will also be important in Chapter 4:

$$\begin{aligned}\mathbb{E}\Big[(x_t^{\hat{\pi}} - w)^2 \big| x_s^{\hat{\pi}} = y\Big] &= \mathbb{E}\Big[(x_t^{\hat{\pi}})^2 \big| x_s^{\hat{\pi}} = y\Big] - 2w\mathbb{E}\Big[x_t^{\hat{\pi}} \big| x_s^{\hat{\pi}} = y\Big] + w^2 \\ &= (y^2 - w^2)e^{-\rho^2(t-s)} + w^2 + \frac{\lambda}{2}(t-s)e^{\rho^2(T-t)} \\ &\quad - 2w\big((y-w)e^{-\rho^2(t-s)} + w\big) + w^2 \\ &= (y-w)^2 e^{-\rho^2(t-s)} + \frac{\lambda}{2}(t-s)e^{\rho^2(T-t)}, \quad t \in [s,T].\end{aligned} \tag{3.28}$$

Next, we show that the functions given by (3.21) and (3.22) satisfy the conditions in the Verification theorem (Theorem 3.8). First, we determine the density function that attains the minimum in the HJB equation (3.14). For this purpose, we state some useful results:

Remark 3.10. The following points are used in ensuing proofs:

1. $\pi \in \mathcal{P}(\mathbb{R})$ if and only if:
$$\int_{\mathbb{R}} \pi(u)du = 1$$
$$\pi(u) \geq 0, \quad \text{a.e. on } \mathbb{R}.$$

2. We recapitulate the Gaussian density function with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$:
$$\mathcal{N}(u|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(u-\mu)^2}{\sigma^2}}, \quad u \in \mathbb{R}.$$

3. Since the Gaussian density function is in $\mathcal{P}(\mathbb{R})$, we can combine the first and second point to state the formula for the normalising constant:
$$\int_{\mathbb{R}} e^{-\frac{1}{2}\frac{(u-\mu)^2}{\sigma^2}} du = \sqrt{2\pi}\sigma.$$

Lemma 3.11. *For a fixed $w \in \mathbb{R}$, we assume that $v \in C^{1,2}([0,T] \times \mathbb{R})$ and $v_{xx}(t,x;w) > 0$ for $(t,x) \in [0,T) \times \mathbb{R}$. Then, the density $\pi \in \mathcal{P}(\mathbb{R})$ that minimizes the corresponding term in the HJB equation (3.14) for $(t,x) \in [0,T) \times \mathbb{R}$:*

$$\min_{\pi \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2 u^2 v_{xx}(t,x;w) + \rho\sigma u v_x(t,x;w) + \lambda \ln\big(\pi(u)\big)\right)\pi(u)du, \tag{3.29}$$

is given by the density function:

$$\mathcal{N}\Big(\cdot\,\Big|-\frac{\rho v_x(t,x;w)}{\sigma v_{xx}(t,x;w)},\frac{\lambda}{\sigma^2 v_{xx}(t,x;w)}\Big). \tag{3.30}$$

Proof. By using calculus of variations and the first point of Remark 3.10, we obtain that (3.29) is minimized for any $(t,x)\in[0,T)\times\mathbb{R}$ by:

$$\pi(u;t,x,w)=\frac{\exp\big(-\frac{1}{\lambda}(\frac{1}{2}\sigma^2u^2v_{xx}(t,x;w)+\rho\sigma uv_x(t,x;w))\big)}{\int_{\mathbb{R}}\exp\big(-\frac{1}{\lambda}(\frac{1}{2}\sigma^2u^2v_{xx}(t,x;w)+\rho\sigma uv_x(t,x;w))\big)du},\quad u\in\mathbb{R}.$$

We can rewrite the term in the nominator as:

$$\begin{aligned}&\exp\Big(-\frac{1}{\lambda}(\frac{1}{2}\sigma^2u^2v_{xx}(t,x;w)+\rho\sigma uv_x(t,x;w))\Big)\\&=\exp\Big(-\frac{\sigma^2v_{xx}(t,x;w)}{2\lambda}(u+\frac{\rho v_x(t,x;w)}{\sigma v_{xx}(t,x;w)})^2\Big)\cdot\exp\Big(\frac{\rho^2\big(v_x(t,x;w)\big)^2}{2\lambda v_{xx}(t,x;w)}\Big).\end{aligned}$$

If we do the same with the denominator and use the formula for the normalising constant in Remark 3.10, we obtain:

$$\begin{aligned}&\int_{\mathbb{R}}\exp\Big(-\frac{1}{\lambda}(\frac{1}{2}\sigma^2u^2v_{xx}(t,x;w)+\rho\sigma uv_x(t,x;w))\Big)du\\&=\exp\Big(\frac{\rho^2\big(v_x(t,x;w)\big)^2}{2\lambda v_{xx}(t,x;w)}\Big)\int_{\mathbb{R}}\exp\Big(-\frac{\sigma^2v_{xx}(t,x;w)}{2\lambda}(u+\frac{\rho v_x(t,x;w)}{\sigma v_{xx}(t,x;w)})^2\Big)du\\&=\exp\Big(\frac{\rho^2\big(v_x(t,x;w)\big)^2}{2\lambda v_{xx}(t,x;w)}\Big)\sqrt{2\pi\frac{\lambda}{\sigma^2v_{xx}(t,x;w)}}.\end{aligned}$$

Therefore, we can conclude that:

$$\begin{aligned}\pi(u;t,x,w)&=\frac{\exp\Big(-\frac{\sigma^2v_{xx}(t,x;w)}{2\lambda}\big(u+\frac{\rho v_x(t,x;w)}{\sigma v_{xx}(t,x;w)}\big)^2\Big)\cdot\exp\Big(\frac{\rho^2\big(v_x(t,x;w)\big)^2}{2\lambda v_{xx}(t,x;w)}\Big)}{\exp\Big(\frac{\rho^2\big(v_x(t,x;w)\big)^2}{2\lambda v_{xx}(t,x;w)}\Big)\sqrt{2\pi\frac{\lambda}{\sigma^2v_{xx}(t,x;w)}}}\\&=\mathcal{N}\Big(u\Big|-\frac{\rho v_x(t,x;w)}{\sigma v_{xx}(t,x;w)},\frac{\lambda}{\sigma^2v_{xx}(t,x;w)}\Big),\quad u\in\mathbb{R}.\end{aligned}$$

□

In the next step, we verify the admissibility of $\hat{\pi}$:

Lemma 3.12. *For a fixed $w\in\mathbb{R}$, the feedback control $\hat{\pi}(\cdot;\cdot,\cdot,w)$ given in (3.22) is admissible in the sense of Definition 3.3.*

Proof. We verify the conditions in Definition 3.3 for a fixed $w \in \mathbb{R}$.

By (3.22) we have clearly that $\hat{\pi}(\cdot; t, x, w)$ is a density for each $(t, x) \in [0, T] \times \mathbb{R}$ and therefore, the first condition holds.

To prove the second condition, we choose $(s, y) \in [0, T) \times \mathbb{R}$ arbitrary and consider the SDE (3.24). We saw that, by using standard procedure, one could verify that (3.24) has a unique solution that is adapted with continuous paths. For the detailed proof, we refer to Lemma B.9 in Appendix B.6.

To prove that the strategy $\hat{\boldsymbol{\pi}} = (\hat{\pi}_t)_{t\in[s,T]}$ generated from $\hat{\pi}$ with respect to the initial (s, y) is admissible, we have to verify the four conditions in Definition 3.2.

The first one is clear since by (3.22), $\hat{\pi}_t = \hat{\pi}(\cdot; t, x_t^{\hat{\pi}}, w)$ is a density for $t \in [s, T]$.

Further, we show that the process $\big(\int_A \hat{\pi}_t(u)du\big)_{t\in[s,T]}$ is progressivley measurable for all $A \in \mathcal{B}(\mathbb{R})$. For this purpose we choose an arbitrary $A \in \mathcal{B}(\mathbb{R})$. We note that the process $x^{\hat{\pi}}$ has continuous paths and that $\hat{\pi}$ is continuous on $[s, T] \times \mathbb{R}$. Therefore, the path $t \to \int_A \hat{\pi}_t(u)du$ is continuous on $[s, T]$. Furthermore, we know that $x^{\hat{\pi}}$ is adapted and so is also $(\int_A \hat{\pi}_t(u)du)_{t\in[s,T]}$. Hence, by Lemma B.2 in Appendix B.1, we can conclude that the process $(\int_A \hat{\pi}_t(u)du)_{t\in[s,T]}$ is progressivley measurable.

For the third condition, we note that by (3.28) we have $\mathbb{E}\big[(x_t^{\hat{\pi}} - w)^2\big] < \infty$ for every $t \in [s, T]$. Therefore, by using Fubini's theorem (see Appendix B.5), we can conclude that:

$$\begin{aligned}
\mathbb{E}\bigg[\int_s^T (\hat{\mu}_t)^2 + (\hat{\sigma}_t)^2 dt\bigg] &= \mathbb{E}\bigg[\int_s^T \frac{\rho^2}{\sigma^2}(x_t^{\hat{\pi}} - w)^2 + \frac{\lambda}{2\sigma^2} e^{\rho^2(T-t)} dt\bigg] \\
&\leq C + \frac{\rho^2}{\sigma^2} \cdot \mathbb{E}\bigg[\int_s^T (x_t^{\hat{\pi}} - w)^2 dt\bigg] \\
&\leq C + \frac{\rho^2}{\sigma^2} \cdot \int_s^T \mathbb{E}\big[(x_t^{\hat{\pi}} - w)^2\big] dt < \infty,
\end{aligned}$$

where $0 \leq C < \infty$ is a constant.

For the last conditions, we observe by using the triangle inequality and (3.28)

over again that:

$$\begin{aligned}
&\mathbb{E}\Big[\big|(x_T^{\hat{\boldsymbol{\pi}}}-w)^2+\lambda\int_s^T\int_{\mathbb{R}}\hat{\pi}_t(u)\ln\big(\hat{\pi}_t(u)\big)dudt\big|\Big|x_s^{\hat{\boldsymbol{\pi}}}=y\Big]\\
&\leq\mathbb{E}\Big[(x_T^{\hat{\boldsymbol{\pi}}}-w)^2|x_s^{\hat{\boldsymbol{\pi}}}=y\Big]+\mathbb{E}\Big[\big|\lambda\int_s^T\int_{\mathbb{R}}\hat{\pi}_t(u)\ln\big(\hat{\pi}_t(u)\big)dudt\big|\Big|x_s^{\hat{\boldsymbol{\pi}}}=y\Big]\\
&\leq C+\mathbb{E}\Big[\big|\lambda\int_s^T\int_{\mathbb{R}}\hat{\pi}_t(u)\ln\big(\hat{\pi}_t(u)\big)dudt\big|\Big|x_s^{\hat{\boldsymbol{\pi}}}=y\Big],
\end{aligned}$$

where $0\leq C<\infty$ is a constant.

The second term can be computed by using (3.7):

$$\begin{aligned}
&\mathbb{E}\Big[\big|\lambda\int_s^T\int_{\mathbb{R}}\hat{\pi}_t(u)\ln\big(\hat{\pi}_t(u)\big)dudt\big|\Big|x_s^{\boldsymbol{\pi}^*}=y\Big]\\
&=\lambda\cdot\mathbb{E}\Big[\Big|\int_s^T-\frac{1}{2}\ln\Big(\pi\frac{\lambda}{\sigma^2}e^{\rho^2(T-t)}\Big)-\frac{1}{2}dt\Big|\Big|x_s^{\boldsymbol{\pi}^*}=y\Big]\\
&\leq\lambda\Big(\frac{1}{2}(T-s)\Big(\Big|\ln\Big(\pi\frac{\lambda}{\sigma^2}\Big)\Big|+1\Big)+\frac{\rho^2}{2}(T-s)^2\Big)<\infty.
\end{aligned}$$

Therefore, we verified all necessary conditions and can conclude that $\hat{\boldsymbol{\pi}}$ is admissible, hence, also the feedback control $\hat{\pi}$ is admissible. □

With the help of these lemmas, we can finally determine the optimal value function and the optimal feedback control of problem (3.8):

Theorem 3.13. *The optimal value function V^E is given by (3.21) and the optimal feedback control is given by (3.22).*

Proof. By the Verification theorem 3.8, we have to show that the functions v and g given by

$$\begin{aligned}
v(t,x;w)&=(x-w)^2e^{-\rho^2(T-t)}+\frac{\lambda\rho^2}{4}(T^2-t^2)\\
&\quad-\frac{\lambda}{2}\Big(\rho^2T-\ln\big(\frac{\sigma^2}{2\pi\lambda}\big)\Big)(T-t)-(w-z)^2\\
g(u;t,x,w)&=\mathcal{N}\Big(u\Big|-\frac{\rho}{\sigma}(x-w),\frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}\Big)
\end{aligned}$$

for $(t,x)\in[0,T]\times\mathbb{R}$ satisfy the three assumptions in Theorem 3.8.

By Lemma 3.12, we know that the feedback control g is admissible and the second condition holds.

Moreover, we can compute the derivatives and the terminal condition of v:

$$v_t(t,x;w) = \rho^2(x-w)^2 e^{-\rho^2(T-t)} - \frac{\lambda\rho^2 t}{2} + \frac{\lambda\rho^2 T}{2} - \frac{\lambda}{2}\ln\left(\frac{\sigma^2}{\pi\lambda}\right) \tag{3.31}$$

$$v_x(t,x;w) = 2(x-w)e^{-\rho^2(T-t)} \tag{3.32}$$

$$v_{xx}(t,x;w) = 2e^{-\rho^2(T-t)}, \quad (t,x) \in [0,T]\times\mathbb{R}, \tag{3.33}$$

$$v(T,x;w) = (x-w)^2 - (w-z)^2. \tag{3.34}$$

Hence, by (3.34), we can conclude that v satisfies the terminal condition of the HJB equation (3.14).

Besides, since v_x is continuous by (3.32) and for any admissible feedback control π the associated process x^π has continuous paths, the path $t \to v_x\left(t, x_t^\pi; w\right)$ is bounded on $[0,T]$ and therefore:

$$\begin{aligned}\int_s^T \left(\tilde{\sigma}(\pi_t)\right)^2 \left(v_x\left(t,x_t^\pi;w\right)\right)^2 dt &\le C\cdot\int_s^T \left(\tilde{\sigma}(\pi_t)\right)^2 dt \\ &= \sigma^2 C\cdot\int_s^T \mu_t^2 + \sigma_t^2 dt < \infty, \quad \mathbb{P}\text{-a.s.},\end{aligned}$$

where $0 \le C < \infty$ is a constant. Hence, v is sufficiently integrable by Remark 3.9.

Furthermore, we see by (3.31)-(3.33) that $v \in C^{1,2}([0,T]\times\mathbb{R})$ and $v_{xx} > 0$. Hence, Lemma 3.11 provides us with the feedback control whose density function minimizes the corresponding term in (3.14):

$$g^*(u;t,x,w) = \mathcal{N}\Big(u\Big| -\frac{\rho v_x(t,x;w)}{\sigma v_{xx}(t,x;w)}, \frac{\lambda}{\sigma^2 v_{xx}(t,x;w)}\Big). \tag{3.35}$$

By substituting the Gaussian feedback control process g^* given in (3.35) into the minimum term of the HJB equation (3.14), we obtain for $(t,x)\in[0,T)\times\mathbb{R}$:

$$\begin{aligned}&\min_{\pi\in\mathcal{P}(\mathbb{R})}\int_{\mathbb{R}}\left(\frac{1}{2}\sigma^2u^2v_{xx}(t,x;w)+\rho\sigma uv_x(t,x;w)+\lambda\ln\left(\pi(u)\right)\right)\pi(u)du\\ &=\int_{\mathbb{R}}\left(\frac{1}{2}\sigma^2u^2v_{xx}(t,x;w)+\rho\sigma uv_x(t,x;w)+\lambda\ln\left(g^*(u;t,x,w)\right)\right)g^*(u;t,x,w)du\\ &=\frac{1}{2}\sigma^2v_{xx}(t,x;w)\mathbb{E}_{g^*}[U^2]+\rho\sigma v_x(t,x;w)\mathbb{E}_{g^*}[U]-\lambda\mathcal{H}(g^*(\cdot;t,x,w))\\ &=\frac{1}{2}\sigma^2v_{xx}(t,x;w)\left(\mathbb{E}_{g^*}[U]^2+\mathrm{Var}_{g^*}(U)\right)+\rho\sigma v_x(t,x;w)\mathbb{E}_{g^*}[U]-\lambda\mathcal{H}(g^*(\cdot;t,x,w)).\end{aligned}$$

Since $g^*(\cdot;t,x,w)$ is Gaussian we can compute the differential entropy term with

(3.7) and insert the mean and variance of our Gaussian feedback control:

$$\begin{aligned}
&\frac{1}{2}\sigma^2 v_{xx}(t,x;w)\big(\mathbb{E}_{g^*}[U]^2+\mathrm{Var}_{g^*}(U)\big)+\rho\sigma v_x(t,x;w)\mathbb{E}_{g^*}[U]-\lambda\mathcal{H}(g^*(\cdot;t,x,w))\\
&=\frac{1}{2}\sigma^2 v_{xx}(t,x;w)\big(\mathbb{E}_{g^*}[U]^2+\mathrm{Var}_{g^*}(U)\big)+\rho\sigma v_x(t,x;w)\mathbb{E}_{g^*}[U]\\
&\quad-\lambda\Big(\ln\big(\sqrt{2\pi\mathrm{Var}_{g^*}(u)}\big)+\frac{1}{2}\Big)\\
&=\frac{1}{2}\Big(\frac{\rho^2(v_x(t,x;w))^2}{v_{xx}(t,x;w)}+\lambda\Big)-\frac{\rho^2(v_x(t,x;w))^2}{v_{xx}(t,x;w)}-\frac{\lambda}{2}\Big(\ln\big(\frac{2\pi\lambda}{\sigma^2 v_{xx}(t,x;w)}\big)+1\Big)\\
&=-\frac{1}{2}\frac{\rho^2(v_x(t,x;w))^2}{v_{xx}(t,x;w)}-\frac{\lambda}{2}\ln\big(\frac{2\pi\lambda}{\sigma^2 v_{xx}(t,x;w)}\big).
\end{aligned}$$

By adding this to $v_t(t,x;w)$ and using the derivatives of v, we obtain for $(t,x)\in[0,T)\times\mathbb{R}$:

$$\begin{aligned}
&v_t(t,x;w)-\frac{\rho^2(v_x(t,x;w))^2}{2v_{xx}(t,x;w)}-\frac{\lambda}{2}\ln\Big(2\pi\frac{\lambda}{\sigma^2 v_{xx}(t,x;w)}\Big)\\
&=\rho^2(x-w)^2e^{-\rho^2(T-t)}-\frac{\lambda\rho^2 t}{2}+\frac{\lambda\rho^2 T}{2}-\frac{\lambda}{2}\ln\Big(\frac{\sigma^2}{\pi\lambda}\Big)\\
&\quad-\frac{\rho^2}{2}\frac{4(x-w)^2e^{-2\rho^2(T-t)}}{2e^{-\rho^2(T-t)}}-\frac{\lambda}{2}\ln\Big(\frac{2\pi\lambda}{\sigma^2 2e^{-\rho^2(T-t)}}\Big)\\
&=\rho^2(x-w)^2e^{-\rho^2(T-t)}+\frac{\lambda\rho^2}{2}(T-t)-\frac{\lambda}{2}\ln\Big(\frac{\sigma^2}{\pi\lambda}\Big)\\
&\quad-\rho^2(x-w)^2e^{-\rho^2(T-t)}-\frac{\lambda\rho^2}{2}(T-t)+\frac{\lambda}{2}\ln\Big(\frac{\sigma^2}{\pi\lambda}\Big)\\
&=0.
\end{aligned}$$

Hence, v satisfies the HJB equation (3.14) and the first condition in Theorem 3.8 holds.

Further, we can write the density of the feedback control as:

$$\begin{aligned}
g^*(u;t,x,w)&=\mathcal{N}\Big(u\Big|-\frac{\rho v_x(t,x;w)}{\sigma v_{xx}(t,x;w)},\frac{\lambda}{\sigma^2 v_{xx}(t,x;w)}\Big)\\
&=\mathcal{N}\Big(u\Big|-\frac{\rho(x-w)}{\sigma},\frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}\Big)=g(u;t,x,w).
\end{aligned}$$

Therefore, g satisfies the minimum in the HJB equation and the last condition in Theorem 3.8 also holds.

Finally, we can apply the Verification theorem. Hence, we conclude that:

$$V^{\mathrm{E}}(t,x;w)=v(t,x;w),\quad (t,x)\in[0,T]\times\mathbb{R}$$

and that there exists an optimal feedback control $\hat{\pi}$ with $\hat{\pi}(\cdot;t,x,w)=g(\cdot;t,x,w)$ for $(t,x)\in[0,T]\times\mathbb{R}$. □

Further, we can determine the optimal wealth process and the Lagrange multiplier w:

Theorem 3.14. *We consider the wealth process $x^{\hat{\pi}} = (x_t^{\hat{\pi}})_{t\in[0,T]}$ induced by the optimal feedback control $\hat{\pi}(\cdot;\cdot,\cdot)$ (3.22). This optimal wealth process $x^{\hat{\pi}} = (x_t^{\hat{\pi}})_{t\in[0,T]}$ is the unique solution to the SDE:*

$$\begin{aligned} dx_t^{\hat{\pi}} &= -\rho^2(x_t^{\hat{\pi}} - w)dt + \sqrt{\rho^2(x_t^{\hat{\pi}} - w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}}dB_t, \quad 0 \le t \le T \\ x_0^{\hat{\pi}} &= x_0 \end{aligned}$$

and the Lagrange multiplier w is given by $w = \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}$.

Proof. Furthermore, since $\hat{\pi}$ is admissible by Lemma 3.12 the process $x^{\hat{\pi}} = (x_t^{\hat{\pi}})_{t\in[0,T]}$ is the unique strong solution of the SDE:

$$\begin{aligned} dx_t^{\hat{\pi}} &= -\rho^2(x_t^{\hat{\pi}} - w)dt + \sqrt{\rho^2(x_t^{\hat{\pi}} - w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}}dB_t, \quad 0 \le t \le T \\ x_0^{\hat{\pi}} &= x_0. \end{aligned} \tag{3.36}$$

Besides, by (3.26), we obtain:

$$\mathbb{E}[x_T^{\hat{\pi}}] = (x_0 - w)e^{-\rho^2 T} + w.$$

Finally, we can show by using the constraint $\mathbb{E}[x_T^{\hat{\pi}}] = z$ that:

$$w = \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}.$$

□

Moreover, we note that by comparing (3.36) to (3.6) we can conclude that the mean and variance process associated with $\hat{\boldsymbol{\pi}} = (\hat{\pi}_t)_{t\in[0,T]}$ denoted by $(\hat{\mu}_t)_{t\in[0,T]}$ and $(\hat{\sigma}_t^2)_{t\in[0,T]}$, respectively, are given by the mean and the variance of the density functions $(\hat{\pi}_t)_{t\in[0,T]}$:

$$\begin{aligned} \hat{\mu}_t &= -\frac{\rho}{\sigma}(x_t^{\hat{\pi}} - w) = \mathbb{E}_{\hat{\pi}_t}[U] \\ \hat{\sigma}_t^2 &= \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)} = \mathrm{Var}_{\hat{\pi}_t}[U], \quad t \in [0,T]. \end{aligned}$$

Remark 3.15. The results in Theorem 3.13. and Theorem 3.14. lead to some interesting conclusions:

1. The variance of the optimal Gaussian feedback control $\frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}$ measures the exploration. Therefore, it is obvious that the exploration decays in time to its final value $\frac{\lambda}{2\sigma^2}$ at time T. In other words, the agent reduces the level of exploration over the investment horizon by learning more about the initially unknown environment, and the exploitation becomes more important over time.

2. For fixed μ and r and at any given time $t \in [0, T]$, we further see that the variance of the optimal Gaussian feedback control decreases as the volatility of the risky asset σ increases, which represents the level of randomness of the environment. That means a high volatility of the risky asset reduces the need for exploration through the feedback control since the randomness of the asset ensures exploration of the state space.

3. The mean of the optimal Gaussian feedback control is independent of the exploration weight λ, and the variance is independent of the state x. This indicates a separation between exploration and exploitation since the former is captured by the variance and the latter by the mean. The implication is that the agent should concentrate on the most promising region in the action space while randomly selecting actions to interact with the unknown environment.

Chapter 4

Comparison of the MV and EMV problem

In this chapter, we point out some interesting connections between the MV and EMV problem. For this purpose, we prove the solvability equivalence between the two problems. That is the implication of the solution of one problem to that of the other, without needing to solve it separately. Further, we introduce the cost of exploration, which is a measure for the loss in the objective of the MV problem due to exploration. This chapter is mainly based on the detailed explanations in H. Wang & X. Y. Zhou [25] and H. Wang et al. [26].

A first interesting point to note is that by Theorem 2.18 and Theorem 3.14 the MV and the EMV problems have the same Lagrange multiplier value. This is caused because the optimal terminal wealth under the respective optimal feedback controls of the two problems has the same mean.

4.1 Solvability equivalence between MV and EMV problems

In this section, we present the solvability equivalence between the two problems (2.6) and (3.8):

$$\text{(MV:)} \qquad \min_{\boldsymbol{u}\in\mathcal{A}(0,x_0)} \mathbb{E}[(x_T^{\boldsymbol{u}} - w)^2] - (w-z)^2,$$

$$\text{(EMV:)} \qquad \min_{\boldsymbol{\pi}\in\mathcal{A}^{\mathrm{E}}(0,x_0)} \mathbb{E}\left[(x_T^{\boldsymbol{\pi}} - w)^2 + \lambda \int_0^T \int_{\mathbb{R}} \pi_t(u) \ln\big(\pi_t(u)\big) du dt\right] - (w-z)^2.$$

We first note that we solved both problems in Section 2.4 and Section 3.4 separately and independently.

We recap some crucial statements from the last two chapters for a better overview.

We derived the following HJB equations for the two problems respectively:

$$\begin{aligned} &w_t(t,x;w) + \min_{u\in\mathbb{R}} \Big(\frac{1}{2}\sigma^2 u^2 w_{xx}(t,x;w) \\ &\qquad + \rho\sigma u w_x(t,x;w) \Big) = 0, \quad (t,x) \in [0,T)\times\mathbb{R} \\ &w(T,x;w) = (x-w)^2 - (w-z)^2, \quad x\in\mathbb{R}, \end{aligned} \tag{4.1}$$

that is solved by:

$$w(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} - (w-z)^2, \quad (t,x)\in[0,T]\times\mathbb{R}, \tag{4.2}$$

and:

$$\begin{aligned} &v_t(t,x;w) + \min_{\pi\in\mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \Big(\frac{1}{2}\sigma^2 u^2 v_{xx}(t,x;w) + \rho\sigma u v_x(t,x;w) \\ &\qquad + \lambda \ln\big(\pi(u)\big) \Big) \pi(u) du = 0, \quad (t,x)\in[0,T)\times\mathbb{R} \\ &v(T,x;w) = (x-w)^2 - (w-z)^2, \quad x\in\mathbb{R}, \end{aligned} \tag{4.3}$$

that is solved by:

$$\begin{aligned} v(t,x;w) =& (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda\rho^2}{4}(T^2-t^2) \\ &- \frac{\lambda}{2}\Big(\rho^2 T - \ln\big(\frac{\sigma^2}{\pi\lambda}\big)\Big)(T-t) - (w-z)^2, \quad (t,x)\in[0,T]\times\mathbb{R}. \end{aligned} \tag{4.4}$$

The optimal admissible feedback control for the MV problem (2.6) is given by:

$$\hat{u}(t,x;w) = -\frac{\rho}{\sigma}(x-w), \quad (t,x)\in[0,T]\times\mathbb{R}. \tag{4.5}$$

The optimal admissible feedback control for the EMV problem (3.8) is given by:

$$\hat{\pi}(u;t,x,w) = \mathcal{N}\Big(u\Big| -\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}\Big), \quad (t,x)\in[0,T]\times\mathbb{R}. \tag{4.6}$$

Furthermore, we know from (2.30) and (3.28) that:

$$\mathbb{E}\Big[(x_t^{\hat{\boldsymbol{u}}} - w)^2 \Big| x_0^{\hat{\boldsymbol{u}}} = x_0\Big] = (x_0-w)^2 e^{-\rho^2 t}, \quad t\in[0,T] \tag{4.7}$$

$$\mathbb{E}\Big[(x_t^{\hat{\boldsymbol{\pi}}} - w)^2 \Big| x_0^{\hat{\boldsymbol{\pi}}} = x_0\Big] = (x_0-w)^2 e^{-\rho^2 t} + \frac{\lambda}{2} t e^{\rho^2(T-t)}, \quad t\in[0,T], \tag{4.8}$$

where $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\pi}}$ are the open-loop controls generated from the optimal feedback controls $\hat{u}$ and $\hat{\pi}$, respectively, with respect to the initial $(0,x_0)$.

As a first step for the solvability equivalence, we show the equivalence between the admissibility of the optimal feedback controls of the two problems:

Lemma 4.1. *For an arbitrary but fixed $w \in \mathbb{R}$, the following two statements are equivalent:*

1. *The open-loop control $\hat{\boldsymbol{\pi}} = (\hat{\pi}_t)_{t\in[0,T]}$ generated from the feedback control $\hat{\pi}(\cdot;\cdot,\cdot,w)$, given by (4.6), is admissible in the sense of Definition 3.2.*
2. *The open-loop control $\hat{\boldsymbol{u}} = (\hat{u}_t)_{t\in[0,T]}$ generated from the feedback control $\hat{u}(\cdot,\cdot;w)$, given by (4.5), is admissible in the sense of Definition 2.3.*

Proof. We denote by $(\hat{\mu}_t)_{t\in[0,T]}$ the mean and by $(\hat{\sigma}_t^2)_{t\in[0,T]}$ the variance processes associated to the distributional control process $\hat{\boldsymbol{\pi}}$. Further, we note that $\hat{\sigma}_t^2 = \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)} < \infty$ for every $t \in [0,T]$ and that the value of $\hat{u}(t,x;w)$ coincides with the mean of the density $\hat{\pi}(\cdot;t,x,w)$ for every $(t,x) \in [0,T]\times\mathbb{R}$.

Moreover, by (4.7) and (4.8), we know that:

$$\mathbb{E}\Big[(x_t^{\hat{\boldsymbol{\pi}}} - w)^2 \Big| x_0^{\hat{\boldsymbol{\pi}}} = x_0\Big] = \mathbb{E}\Big[(x_t^{\hat{\boldsymbol{u}}} - w)^2 \Big| x_0^{\hat{\boldsymbol{u}}} = x_0\Big] + \frac{\lambda}{2}te^{\rho^2(T-t)}, \quad t \in [0,T].$$

Therefore, we can conclude that:

$$\mathbb{E}\Big[\int_0^T (\hat{u}_t)^2 dt\Big] < \infty \iff \mathbb{E}\Big[\int_0^T (\hat{\mu}_t)^2 + (\hat{\sigma}_t)^2 dt\Big] < \infty.$$

Furthermore, the equivalence between the progressive measurability follows by the continuity of $(\hat{\sigma}_t^2)_{t\in[0,T]}$.

By the formula for the differential entropy of a Gaussian density (3.7), we obtain:

$$\mathbb{E}\Big[\Big|\lambda\int_0^T\int_{\mathbb{R}} \hat{\pi}_t(u)\ln\big(\hat{\pi}_t(u)\big)duds\Big|\Big| x_0^{\hat{\boldsymbol{\pi}}} = x_0\Big] = \Big|\frac{\lambda}{2}T\Big(1 + \ln\Big(\frac{\pi\lambda}{\sigma^2}\Big) + \frac{\rho^2}{2}T\Big)\Big| < \infty.$$

Hence, we can conclude that:

$$\mathbb{E}\Big[\Big|(x_T^{\hat{\boldsymbol{\pi}}} - w)^2 + \lambda\int_0^T\int_{\mathbb{R}} \hat{\pi}_t(u)\ln\big(\hat{\pi}_t(u)\big)dudt\Big|\Big| x_0^{\hat{\boldsymbol{\pi}}} = x_0\Big] < \infty$$
$$\iff$$
$$\mathbb{E}\Big[(x_T^{\hat{\boldsymbol{u}}} - w)^2\Big| x_0^{\hat{\boldsymbol{u}}} = x_0\Big] < \infty.$$

□

Now, we can prove the solvability equivalence between the MV and the EMV problem:

Theorem 4.2. ***(Solvability Equivalence)**:*

For an arbitrary but fixed $w \in \mathbb{R}$, the following two statements are equivalent:

1. *The function v given by (4.4) is the optimal value function of the EMV problem (3.8), and the optimal corresponding feedback control is given by (4.6).*
2. *The function w given by (4.2) is the optimal value function of the MV problem (2.6), and the optimal corresponding feedback control is given by (4.5).*

Proof. We first note that when 1 holds, the function v solves the HJB equation (4.3) of the exploratory MV problem and when 2 holds, the function w solves the HJB equation (4.1) of the MV problem.

Furthermore, Lemma 4.1 establishes the equivalence between the admissibility of the optimal open-loop control $\hat{\boldsymbol{\pi}}$ generated from the feedback control $\hat{\pi}$ and the admissibility of the optimal open-loop control $\hat{\boldsymbol{u}}$ generated from the feedback control $\hat{u}$ with respect to the initial $(0, x_0)$.

A comparison between the two HJB equations (4.1) and (4.3) yields that if v in 1 solves the former, then w in 2 solves the latter, and vice versa:

For the first direction, we assume that v is the optimal value function of the EMV problem (3.8) and therefore solves the HJB equation (4.3). Further, we assume that the optimal corresponding feedback control is $\hat{\pi}$, such that the density function $\hat{\pi}(\cdot; t, x, w)$ minimizes the minimum term in the HJB equation (4.3).

First, we observe that the terminal value $v(T, x; w) = w(T, x; w)$ for all $x \in \mathbb{R}$.

Further, we choose an arbitrary $(t, x) \in [0, T) \times \mathbb{R}$ and denote the corresponding density $\hat{\pi}(\cdot; t, x, w)$ by $\hat{\pi}_{t,x}$ with variance and mean:

$$\begin{aligned}\hat{\mu}_{t,x} &:= \mathbb{E}_{\hat{\pi}_{t,x}}[U] = \int_{\mathbb{R}} u\hat{\pi}_{t,x}(u)du = -\frac{\rho}{\sigma}(x - w)\\ \hat{\sigma}^2_{t,x} &:= \mathrm{Var}_{\hat{\pi}_{t,x}}(U) = \int_{\mathbb{R}} (x - \hat{\mu}_{t,x})^2\hat{\pi}_{t,x}(u)du = \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}.\end{aligned}$$

We note that by (3.7), we have:

$$\begin{aligned}\lambda\mathcal{H}(\hat{\pi}_{t,x}) &= \lambda \ln\left(\sqrt{2\pi e\hat{\sigma}^2_{t,x}}\right) = \lambda \ln\left(\sqrt{\frac{\lambda\pi}{\sigma^2}}\sqrt{e^{\rho^2(T-t)}}\sqrt{e}\right)\\ &= \frac{\lambda}{2}\left(\ln\left(\frac{\lambda\pi}{\sigma^2}\right) + \rho^2(T - t) + 1\right).\end{aligned} \tag{4.9}$$

Besides, we observe the following relationship between the functions v and w:

$$v(t, x; w) = w(t, x; w) + \frac{\lambda\rho^2}{4}(T^2 - t^2) - \frac{\lambda}{2}\left(\rho^2 T - \ln\left(\frac{\sigma^2}{\pi\lambda}\right)\right)(T - t)$$

and we can compute the derivatives:

$$\begin{aligned} v_t(t,x;w) &= \rho^2(x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda}{2}\left(\rho^2(T-t) + \ln\left(\frac{\pi\lambda}{\sigma^2}\right)\right) \\ &= w_t(t,x;w) + \frac{\lambda}{2}\left(\rho^2(T-t) + \ln\left(\frac{\pi\lambda}{\sigma^2}\right)\right) \\ &\overset{(4.9)}{=} w_t(t,x;w) + \lambda\mathcal{H}(\hat{\pi}_{t,x}) - \frac{\lambda}{2} \end{aligned} \tag{4.10}$$

$$v_x(t,x;w) = 2(x-w)e^{-\rho^2(T-t)} = w_x(t,x;w) \tag{4.11}$$

$$v_{xx}(t,x;w) = 2e^{-\rho^2(T-t)} = \frac{\lambda}{\sigma^2}(\hat{\sigma}^2_{t,x})^{-1} = w_{xx}(t,x;w). \tag{4.12}$$

By using (4.11) - (4.12), we can write the minimum term:

$$\begin{aligned} &\min_{\pi\in\mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2u^2v_{xx}(t,x;w) + \rho\sigma u v_x(t,x;w) + \lambda\ln\big(\pi(u)\big)\right)\pi(u)du \\ &= \min_{\pi\in\mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2u^2w_{xx}(t,x;w) + \rho\sigma u w_x(t,x;w) + \lambda\ln\big(\pi(u)\big)\right)\pi(u)du \\ &= \min_{\pi\in\mathcal{P}(\mathbb{R})} \left(\frac{1}{2}\sigma^2 w_{xx}(t,x;w)\mathbb{E}_\pi[U^2] + \rho\sigma w_x(t,x;w)\mathbb{E}_\pi[U] - \lambda\mathcal{H}(\pi)\right) \\ &= \frac{1}{2}\sigma^2 w_{xx}(t,x;w)\mathbb{E}_{\hat{\pi}_{t,x}}[U^2] + \rho\sigma w_x(t,x;w)\mathbb{E}_{\hat{\pi}_{t,x}}[U] - \lambda\mathcal{H}(\hat{\pi}_{t,x}) \\ &= \frac{1}{2}\sigma^2 w_{xx}(t,x;w)\Big(\mathbb{E}_{\hat{\pi}_{t,x}}[U]^2 + \mathrm{Var}_{\hat{\pi}_{t,x}}(U)\Big) + \rho\sigma w_x(t,x;w)\mathbb{E}_{\hat{\pi}_{t,x}}[U] - \lambda\mathcal{H}(\hat{\pi}_{t,x}) \\ &= \frac{1}{2}\sigma^2 w_{xx}(t,x;w)\Big(\hat{\mu}^2_{t,x} + \hat{\sigma}^2_{t,x}\Big) + \rho\sigma w_x(t,x;w)\hat{\mu}_{t,x} - \lambda\mathcal{H}(\hat{\pi}_{t,x}) \\ &\overset{(4.12)}{=} \frac{1}{2}\sigma^2 w_{xx}(t,x;w)\hat{\mu}^2_{t,x} + \frac{\lambda}{2} + \rho\sigma w_x(t,x;w)\hat{\mu}_{t,x} - \lambda\mathcal{H}(\hat{\pi}_{t,x}), \end{aligned} \tag{4.13}$$

where we used $\mathrm{Var}(U) = \mathbb{E}[U^2] - \mathbb{E}[U]^2$.

Since v solves the HJB equation (4.3) and by substituting (4.10) and (4.13), we obtain:

$$\begin{aligned} &v_t(t,x;w) + \min_{\pi\in\mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2u^2v_{xx}(t,x;w) + \rho\sigma u v_x(t,x;w) + \lambda\ln\big(\pi(u)\big)\right)\pi(u)du \\ &= w_t(t,x;w) + \lambda\mathcal{H}(\hat{\pi}_{t,x}) - \frac{\lambda}{2} + \frac{1}{2}\sigma^2 w_{xx}(t,x;w)\hat{\mu}^2_{t,x} + \frac{\lambda}{2} + \rho\sigma w_x(t,x;w)\hat{\mu}_{t,x} \\ &\quad - \lambda\mathcal{H}(\hat{\pi}_{t,x}) \\ &= w_t(t,x;w) + \frac{1}{2}\sigma^2 w_{xx}(t,x;w)\hat{\mu}^2_{t,x} + \rho\sigma w_x(t,x;w)\hat{\mu}_{t,x} = 0. \end{aligned}$$

Furthermore, if we use $\hat{\mu}_{t,x} = \mathbb{E}_{\hat{\pi}_{t,x}}[U] = \hat{u}(t,x;w)$, we have:

$$w_t(t,x;w) + \frac{1}{2}\sigma^2\big(\hat{u}(t,x;w)\big)^2 w_{xx}(t,x;w) + \rho\sigma\hat{u}(t,x;w)w_x(t,x;w) = 0.$$

For an arbitrary $u \in \mathbb{R}$, we obtain:

$$\begin{aligned} &w_t(t,x;w) + \frac{1}{2}\sigma^2 u^2 w_{xx}(t,x;w) + \rho\sigma u w_x(t,x;w) \\ &= \rho^2(x-w)^2 e^{-\rho^2(T-t)} + \sigma^2 u^2 e^{-\rho^2(T-t)} + 2\sigma\rho u(x-w)e^{-\rho^2(T-t)} \\ &= e^{-\rho^2(T-t)}\big(\rho(x-w) + \sigma u\big)^2 \geq 0. \end{aligned}$$

And since $(t,x) \in [0,T) \times \mathbb{R}$ was chosen arbitrary, we can conclude that w solves the HJB equation (4.1) and the optimal feedback control $\hat{u}$ attains the minimization term.

For the other direction we can proceed reversely. □

Next, we show that the EMV problem converges to its classical counterpart as the exploration weight λ decreases to 0.

Theorem 4.3. *Assume that statement 1 (or equivalently, 2) of Theorem 4.2 holds. Then, for each $(t,x,w) \in [0,T] \times \mathbb{R} \times \mathbb{R}$:*

$$\lim_{\lambda \to 0} \hat{\pi}(\cdot; t, x, w) = \delta_{\hat{u}(t,x;w)}(\cdot), \quad \textit{weakly},$$

where $\delta_{\hat{u}(t,x;w)}(\cdot)$ is a Dirac measure (see Appendix A.3). Moreover,

$$\lim_{\lambda \to 0} \Big| V^{\mathrm{E}}(t,x;w) - V(t,x;w) \Big| = 0$$

Proof. The weak convergence of the feedback controls follows from the fact that for all $(t,x,w) \in [0,T] \times \mathbb{R} \times \mathbb{R}$, the variance of $\hat{\pi}(\cdot; t, x, w)$ converges to 0 as $\lambda \to 0$ and that the mean is equal to $\hat{u}(t,x;w)$.

The pointwise convergence of the value functions follows from the explicit forms in Theorem 4.2:

$$\begin{aligned} \lim_{\lambda \to 0} \Big| V^{\mathrm{E}}(t,x;w) - V(t,x;w) \Big| &= \lim_{\lambda \to 0} \Big| \frac{\lambda\rho^2}{4}(T^2 - t^2) - \frac{\lambda}{2}\Big(\rho^2 T - \ln\big(\frac{\sigma^2}{\pi\lambda}\big)\Big)(T-t) \Big| \\ &= (T-t) \cdot \lim_{\lambda \to 0} \frac{\lambda}{2} \ln\left(\frac{\sigma^2}{\pi\lambda}\right). \end{aligned}$$

Further, we see that:

$$\lim_{\lambda \to 0} \frac{\lambda}{2} \ln\left(\frac{\sigma^2}{\pi\lambda}\right) = \lim_{\lambda \to 0} \Big(\frac{\lambda}{2}\ln(\sigma^2) - \frac{\lambda}{2}\ln(\pi\lambda)\Big) = -\lim_{\lambda \to 0} \frac{\lambda}{2}\ln(\pi\lambda) = 0,$$

where we used l'Hôpital's rule in the last step. And consequently, the proof is completed. □

4.2 Cost of Exploration

In H. Wang et al. [26] the exploration cost for a general RL problem is defined as the difference between the discounted accumulated rewards following the corresponding optimal open-loop controls under the classical optimal value function V and the exploratory optimal value function V^{E}, net of the value of the entropy. For our clear case, the cost is the difference between the two optimal value functions V^{E} and V defined respectively in (3.10) and (2.15), adjusting for the additional contribution due to the entropy value of the optimal exploratory strategy. Note that the solvability equivalence established in the previous subsection is essential for this definition, not least because the cost is well-defined only if both the classical and the exploratory problems are solvable:

Definition 4.4. The cost associated with the MV problem due to the explicit inclusion of exploration in the objective (3.10) is defined by:

$$\mathcal{C}^{\hat{\boldsymbol{u}},\hat{\boldsymbol{\pi}}}(0,x_0;w) := \left(V^{\mathrm{E}}(0,x_0;w) - \lambda\mathbb{E}\left[\int_0^T \int_{\mathbb{R}} \hat{\pi}_t(u)\ln(\hat{\pi}_t(u))dudt \mid x_0^{\hat{\boldsymbol{\pi}}} = x_0\right]\right) - V(0,x_0;w),$$

for $x_0 \in \mathbb{R}$, where $\hat{\boldsymbol{\pi}} = (\hat{\pi}_t)_{t\in[0,T]}$ is the optimal (open-loop) control generated from the optimal feedback law $\hat{\pi}$ with respect to the initial $(0,x_0)$.

The term $V(0,x_0;w)$ is the optimal value of the classical objective without exploration, while the first term in brackets is the value of the classical objective under the solution that maximizes the regularized objective. Hence, the exploration cost measures the loss in the non-regularized objective due to exploration.

We next compute the exploration cost. We show that it depends only on the exploration weight $\lambda > 0$ and the investment horizon $T > 0$, which are both parameters that the agent specifies:

Theorem 4.5. *Assume that statement 1 (or equivalently, 2) of Theorem 4.2 holds. Then, the exploration cost for the MV problem is:*

$$\mathcal{C}^{\hat{\boldsymbol{u}},\hat{\boldsymbol{\pi}}}(0,x_0;w) = \frac{\lambda T}{2}, \quad x_0 \in \mathbb{R}, \quad w \in \mathbb{R}.$$

Proof. Let $(\hat{\pi}_t)_{t\in[0,T]}$ be the open-loop control generated by the feedback control $\hat{\pi}$ given by (4.6) with respect to the initial state x_0 at $t = 0$:

$$\hat{\pi}_t(u) = \mathcal{N}\left(u \mid -\frac{\rho}{\sigma}\left(x_t^{\hat{\boldsymbol{\pi}}} - w\right), \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}\right),$$

where $(x_t^{\hat{\boldsymbol{\pi}}})_{t\in[0,T]}$ is the corresponding optimal wealth process of the EMV problem, starting from the state x_0. By using (3.7) we obtain:

$$\begin{aligned}
&-\lambda\mathbb{E}\left[\int_0^T\int_{\mathbb{R}}\hat{\pi}_t(u)\ln(\hat{\pi}_t(u))dudt \mid x_0^{\hat{\boldsymbol{\pi}}}=x_0\right]\\
&=\frac{\lambda}{2}\mathbb{E}\left[\int_0^T\left(\ln\left(\frac{\pi\lambda}{\sigma^2}\right)+\rho^2(T-t)+1\right)dt \mid x_0^{\hat{\boldsymbol{\pi}}}=x_0\right]\\
&=\frac{\lambda}{2}\mathbb{E}\left[\ln\left(\frac{\pi\lambda}{\sigma^2}\right)T+\frac{\rho^2}{2}T^2+T \mid x_0^{\hat{\boldsymbol{\pi}}}=x_0\right]\\
&=\frac{\lambda\rho^2}{4}T^2+\frac{\lambda T}{2}\ln\left(\frac{\pi\lambda}{\sigma^2}\right)+\frac{\lambda T}{2}.
\end{aligned}\tag{4.14}$$

By using the expressions of V given by (4.2) and V^{E} given by (4.4), we can compute that:

$$\begin{aligned}
V^{\mathrm{E}}(0,x_0;w)-V(0,x_0;w)&=\frac{\lambda\rho^2}{4}T^2-\frac{\lambda}{2}\left(\rho^2T-\ln\left(\frac{\sigma^2}{\pi\lambda}\right)\right)T\\
&=-\frac{\lambda\rho^2}{4}T^2-\frac{\lambda T}{2}\ln\left(\frac{\pi\lambda}{\sigma^2}\right).
\end{aligned}\tag{4.15}$$

Adding (4.14) and (4.15) results in the desired equality:

$$\mathcal{C}^{\hat{\boldsymbol{u}},\hat{\boldsymbol{\pi}}}(0,x_0;w)=\frac{\lambda T}{2}.$$

□

Intuitively, the exploration cost increases with the exploration weight and horizon. Moreover, Theorem 4.5 shows that the dependence is linear with respect to each of the two exploration parameters, λ and T, and that the cost $\mathcal{C}^{\hat{\boldsymbol{u}},\hat{\boldsymbol{\pi}}}$ is independent of the Lagrange multiplier w. Therefore, the exploration cost will not increase when the agent increases its expected investment target reflected by z or equivalently by the Lagrange multiplier w.

Chapter 5

The Reinforcement Learning Algorithm

In the previous chapters, we introduced various concepts and proved different results, which form the theoretical basis for an RL algorithm learning the solution of the EMV problem without assuming any knowledge about the underlying parameters. For this purpose, in this chapter, we will first state and prove a so-called policy improvement theorem (PIT) and the corresponding convergence results. Moreover, we will introduce an RL algorithm to solve (3.8). This chapter follows mainly the detailed explanations by H. Wang & X. Y. Zhou in [25].

5.1 Policy Improvement Theorem

Most RL algorithms consist of two iterative procedures: one evaluates, and the other improves the policy. The policy evaluation provides an estimated value function for the current policy. In contrast, the policy improvement updates the current policy to improve the value function (for more detailed explanations, we refer to Sections 4.1 - 4.2 in R. S. Sutton & A. G. Barto [23]). To ensure that the policy improvement iteration generates policies that lead to non-increasing (in the case of a minimization problem) iterated value functions, a PIT is crucial for interpretable RL algorithms (for more detailed explanations about PITs, we refer to Section 4.2 in R. S. Sutton & A. G. Barto [23]).

We first state some valuable results to prove the PIT for the EMV problem.

Lemma 5.1. *Let $w \in \mathbb{R}$ be fixed and $\pi = \pi(\cdot;\cdot,\cdot,w) \in \mathcal{A}^E$ be an arbitrarily admissible feedback control. Suppose that the corresponding value function $J^E(\cdot,\cdot;w,\pi) \in C^{1,2}([0,T)\times\mathbb{R}) \cap C^0([0,T]\times\mathbb{R})$ and satisfies $J^E_{xx}(t,x;w,\pi) > 0$,*

for any $(t,x) \in [0,T) \times \mathbb{R}$. *Then, the following equality holds:*

$$J_t^E(t,x;w,\pi) + \int_{\mathbb{R}} \Big(\frac{1}{2}\sigma^2 u^2 J_{xx}^E(t,x;w,\pi) + \rho\sigma u J_x^E(t,x;w,\pi) + \lambda \ln(\pi(u;t,x,w)) \Big) \pi(u;t,x,w) du = 0, \quad (t,x) \in [0,T) \times \mathbb{R}.$$

Proof. We choose $(t,x) \in [0,T) \times \mathbb{R}$ arbitrary and divide the interval $[t,T)$ into two parts $[t,t+h]$ and $(t+h,T)$ respectively for a suitable $h > 0$. By the proof of Lemma 3.5, we know that for any admissible feedback control $\pi(\cdot;\cdot,\cdot,\cdot)$ it holds:

$$\mathbb{E}\Big[J^{\mathrm{E}}(t+h, x_{t+h}^{\pi}; w, \pi) + \lambda \int_t^{t+h} \int_{\mathbb{R}} \pi_s(u) \ln\big(\pi_s(u)\big) du ds \Big| x_t^{\pi} = x \Big] = J^{\mathrm{E}}(t,x;w,\pi), \tag{5.1}$$

where $\boldsymbol{\pi} = (\pi_s)_{s\in[t,T]}$ is the open-loop control generated from the feedback control π with respect to the initial (t,x).

Since $J^{\mathrm{E}}(\cdot,\cdot;w,\pi) \in C^{1,2}([0,T) \times \mathbb{R}) \cap C^0([0,T] \times \mathbb{R})$ we can apply Itô's formula:

$$\begin{aligned} J^{\mathrm{E}}(t+h, x_{t+h}^{\pi}; w, \pi) = J^{\mathrm{E}}(t, x_t^{\pi}; w, \pi) &+ \int_t^{t+h} J_t^{\mathrm{E}}(s, x_s^{\pi}; w, \pi) \\ &+ \tilde{b}(\pi_s) J_x^{\mathrm{E}}(s, x_s^{\pi}; w, \pi) + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 J_{xx}^{\mathrm{E}}(s, x_s^{\pi}; w, \pi) ds \\ &+ \int_t^{t+h} \tilde{\sigma}(\pi_s) J_x^{\mathrm{E}}(s, x_s^{\pi}; w, \pi) dB_s. \end{aligned} \tag{5.2}$$

Moreover, since π is admissible we have $(\tilde{\sigma}(\pi_s) J_x^{\mathrm{E}}(s, x_s^{\pi}; w, \pi))_{s\in[t,T]} \in \mathcal{L}^2_{\mathrm{loc}}([t,T], B)$.

We define the stopping times for $n \geq 1$:

$$\tau_n := \inf\Big\{ \tau \geq t : \int_t^{\tau} \big(\tilde{\sigma}(\pi_s)\big)^2 \Big(J_x^{\mathrm{E}}(s, x_s^{\pi}; w, \pi) \Big)^2 ds \geq n \Big\}.$$

By Lemma 2.11, this sequence of stopping times is a localising sequence for $(\tilde{\sigma}(\pi_s) J_x^{\mathrm{E}}(s, x_s^{\pi}; w, \pi))_{s\in[t,T]}$ in $\mathcal{L}^2([t,T], B)$.

Then, from (5.2), we obtain:

$$\begin{aligned} &J^{\mathrm{E}}\big((t+h) \wedge \tau_n, x_{(t+h)\wedge\tau_n}^{\pi}; w, \pi\big) \\ &= J^{\mathrm{E}}(t, x_t^{\pi}; w, \pi) + \int_t^{(t+h)\wedge\tau_n} J_t^{\mathrm{E}}(s, x_s^{\pi}; w, \pi) + \tilde{b}(\pi_s) J_x^{\mathrm{E}}(s, x_s^{\pi}; w, \pi) \\ &\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 J_{xx}^{\mathrm{E}}(s, x_s^{\pi}; w, \pi) ds + \int_t^{(t+h)\wedge\tau_n} \tilde{\sigma}(\pi_s) J_x^{\mathrm{E}}(s, x_s^{\pi}; w, \pi) dB_s. \end{aligned}$$

By taking the conditional expectation given $x_t^\pi = x$ and using Lemma 2.13, we have that:

$$\begin{aligned}
&\mathbb{E}\Big[J^{\mathrm{E}}\big((t+h)\wedge\tau_n, x^\pi_{(t+h)\wedge\tau_n}; w, \pi\big)\Big|x_t^{\mathbf{u}} = x\Big] \\
&= \mathbb{E}\Big[J^{\mathrm{E}}(t, x_t^\pi; w, \pi)\Big|x_t^\pi = x\Big] + \mathbb{E}\Bigg[\int_t^{(t+h)\wedge\tau_n} J_t^{\mathrm{E}}(s, x_s^\pi; w, \pi) + \tilde{b}(\pi_s) J_x^{\mathrm{E}}(s, x_s^\pi; w, \pi) \\
&\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 J_{xx}^{\mathrm{E}}(s, x_s^\pi; w, \pi) ds\Big|x_t^\pi = x\Bigg] \\
&\quad + \mathbb{E}\Bigg[\int_t^{(t+h)\wedge\tau_n} \tilde{\sigma}(\pi_s) J_x^{\mathrm{E}}(s, x_s^\pi; w, \pi) dB_s\Big|x_t^{\mathbf{u}} = x\Bigg] \\
&= J^{\mathrm{E}}(t, x; w, \pi) + \mathbb{E}\Bigg[\int_t^{(t+h)\wedge\tau_n} J_t^{\mathrm{E}}(s, x_s^\pi; w, \pi) + \tilde{b}(\pi_s) J_x^{\mathrm{E}}(s, x_s^\pi; w, \pi) \\
&\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 J_{xx}^{\mathrm{E}}(s, x_s^\pi; w, \pi) ds\Big|x_t^\pi = x\Bigg].
\end{aligned}$$

Since $(\tau_n)_{n\geq 1}$ is a localising sequence and by using the dominated convergence theorem (see Appendix B.4), we obtain for $n \to \infty$:

$$\begin{aligned}
\mathbb{E}\Big[J^{\mathrm{E}}(t+h, x_{t+h}^\pi; w, \pi)\Big|x_t^\pi = x\Big] &= J^{\mathrm{E}}(t, x; w, \pi) \\
&\quad + \mathbb{E}\Bigg[\int_t^{t+h} J_t^{\mathrm{E}}(s, x_s^\pi; w, \pi) + \tilde{b}(\pi_s) J_x^{\mathrm{E}}(s, x_s^\pi; w, \pi) \\
&\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 J_{xx}^{\mathrm{E}}(s, x_s^\pi; w, \pi) ds\Big|x_t^\pi = x\Bigg].
\end{aligned}$$

By substituting this formulation into (5.1), we obtain:

$$\begin{aligned}
J^{\mathrm{E}}(t, x; w, \pi) &= J^{\mathrm{E}}(t, x; w, \pi) \\
&\quad + \mathbb{E}\Bigg[\int_t^{t+h} J^{\mathrm{E}}(s, x_s^\pi; w, \pi) + \tilde{b}(\pi_s) J_x^{\mathrm{E}}(s, x_s^\pi; w, \pi) \\
&\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 J_{xx}^{\mathrm{E}}(s, x_s^\pi; w, \pi) ds \\
&\quad + \lambda \int_t^{t+h} \int_{\mathbb{R}} \pi_s(u) \ln\big(\pi_s(u)\big) du ds\Big|x_t^\pi = x\Bigg].
\end{aligned}$$

And by subtracting J^{E} on both sides, we have:

$$\begin{aligned}
0 = \mathbb{E}\Bigg[&\int_t^{t+h} J_t^{\mathrm{E}}(s, x_s^\pi; w, \pi) + \tilde{b}(\pi_s) J_x^{\mathrm{E}}(s, x_s^\pi; w, \pi) \\
&+ \frac{1}{2}\big(\tilde{\sigma}(\pi_s)\big)^2 J_{xx}^{\mathrm{E}}(s, x_s^\pi; w, \pi) ds + \lambda \int_t^{t+h} \int_{\mathbb{R}} \pi_s(u) \ln\big(\pi_s(u)\big) du ds\Big|x_t^\pi = x\Bigg].
\end{aligned}$$

Further, if we divide by h and let $h \to 0$, we obtain:

$$\begin{aligned}
0 &= \mathbb{E}\Bigg[J_t^{\mathrm{E}}(t, x_t^\pi; w, \pi) + \tilde{b}(\pi_t) J_x^{\mathrm{E}}(t, x_t^\pi; w, \pi) \\
&\quad + \frac{1}{2}\big(\tilde{\sigma}(\pi_t)\big)^2 J_{xx}^{\mathrm{E}}(t, x_t^\pi; w, \pi) + \lambda \int_{\mathbb{R}} \pi_t(u) \ln\big(\pi_t(u)\big) du \Big| x_t^\pi = x \Bigg] \\
&= J_t^{\mathrm{E}}(t, x; w, \pi) + \tilde{b}\big(\pi(\cdot; t, x, w)\big) J_x^{\mathrm{E}}(t, x; w, \pi) \\
&\quad + \frac{1}{2}\Big(\tilde{\sigma}\big(\pi(\cdot; t, x, w)\big)\Big)^2 J_{xx}^{\mathrm{E}}(t, x; w, \pi) + \lambda \int_{\mathbb{R}} \pi(u; t, x, w) \ln\big(\pi(u; t, x, w)\big) du.
\end{aligned}$$

By the definition of $\tilde{b}(\cdot)$ in (3.4) and $\tilde{\sigma}(\cdot)$ in (3.5), we can conclude that:

$$\begin{aligned}
0 &= J_t^{\mathrm{E}}(t, x; w, \pi) + \tilde{b}\big(\pi(\cdot; t, x, w)\big) J_x^{\mathrm{E}}(t, x; w, \pi) \\
&\quad + \frac{1}{2}\Big(\tilde{\sigma}\big(\pi(\cdot; t, x, w)\big)\Big)^2 J_{xx}^{\mathrm{E}}(t, x; w, \pi) + \lambda \int_{\mathbb{R}} \pi(u; t, x, w) \ln\big(\pi(u; t, x, w)\big) du \\
&= J_t^{\mathrm{E}}(t, x; w, \pi) + \int_{\mathbb{R}} \Bigg(\frac{1}{2}\sigma^2 u^2 J_{xx}^{\mathrm{E}}(t, x; w, \pi) + \rho\sigma u J_x^{\mathrm{E}}(t, x; w, \pi) \\
&\quad + \lambda \ln(\pi(u; t, x, w)) \Bigg) \pi(u; t, x, w) du.
\end{aligned}$$

Since the choice of (t, x) and the admissible feedback control $\pi(\cdot; \cdot, \cdot, \cdot)$ was arbitrary, we can conclude the proof. □

The following result provides a PIT for our EMV portfolio selection problem:

Theorem 5.2. ***(Policy Improvement Theorem)****:*
Let $w \in \mathbb{R}$ be fixed and $\pi = \pi(\cdot; \cdot, \cdot, w)$ be an arbitrarily admissible feedback control. Suppose that the corresponding value function $J^E(\cdot, \cdot; w, \pi) \in C^{1,2}([0, T) \times \mathbb{R}) \cap C^0([0, T] \times \mathbb{R})$ and satisfies $J_{xx}^E(t, x; w, \pi) > 0$, for any $(t, x) \in [0, T) \times \mathbb{R}$. Suppose further that the feedback policy $\tilde{\pi}$ defined by

$$\tilde{\pi}(u; t, x, w) = \mathcal{N}\left(u \mid -\frac{\rho}{\sigma} \frac{J_x^E(t, x; w, \pi)}{J_{xx}^E(t, x; w, \pi)}, \frac{\lambda}{\sigma^2 J_{xx}^E(t, x; w, \pi)}\right) \tag{5.3}$$

is admissible. Then,

$$J^E(t, x; w, \tilde{\pi}) \leq J^E(t, x; w, \pi), \quad (t, x) \in [0, T] \times \mathbb{R}.$$

Proof. We fix $(t, x) \in [0, T] \times \mathbb{R}$ and choose an arbitrary feedback control $\pi \in \mathcal{A}^{\mathrm{E}}$. By assumption, the feedback policy $\tilde{\pi}$ is admissible. Let $(x_s^{\tilde{\pi}})_{s \in [t, T]}$ be the induced wealth process.

By applying Itô's formula and using the definitions of the exploratory drift and diffusion in (3.4) and (3.5), respectively, we obtain for $s \in [t, T]$:

$$\begin{aligned} J^{\mathrm{E}}(s, x_s^{\tilde{\pi}}; w, \pi) &= J^{\mathrm{E}}(t, x_t^{\tilde{\pi}}; w, \pi) + \int_t^s J_t^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) + \tilde{b}(\tilde{\pi}_v) J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \\ &\quad + \frac{1}{2}\big(\tilde{\sigma}(\tilde{\pi}_v)\big)^2 J_{xx}^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) dv + \int_t^s \tilde{\sigma}(\tilde{\pi}_v) J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) dB_v \\ &= J^{\mathrm{E}}(t, x_t^{\tilde{\pi}}; w, \pi) \\ &\quad + \int_t^s \bigg(J_t^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) + \int_{\mathbb{R}} \rho\sigma u \tilde{\pi}_v(u) du J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \\ &\quad + \frac{1}{2} \int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du J_{xx}^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \bigg) dv \\ &\quad + \int_t^s \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du} J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) dB_v, \end{aligned} \tag{5.4}$$

where $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_v)_{v \in [t,T]}$ is the open-loop control generated from $\tilde{\pi}$ with respect to the initial (t, x).

Next, we define the stopping times for $n \geq 1$:

$$\tau_n := \inf \left\{ s \geq t : \int_t^s \int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du \left(J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \right)^2 dv \geq n \right\},$$

which constitute a localising sequence in $\mathcal{L}^2([t, T], B)$ by Lemma 2.11.

Then, from (5.4), we obtain:

$$\begin{aligned} J^{\mathrm{E}}(s \wedge \tau_n, x_{s \wedge \tau_n}^{\tilde{\pi}}; w, \pi) &= J^{\mathrm{E}}(t, x_t^{\tilde{\pi}}; w, \pi) + \int_t^{s \wedge \tau_n} \bigg(J_t^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \\ &\quad + \int_{\mathbb{R}} \rho\sigma u \tilde{\pi}_v(u) du J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \\ &\quad + \frac{1}{2} \int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du J_{xx}^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \bigg) dv \\ &\quad + \int_t^{s \wedge \tau_n} \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du} J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) dB_v. \end{aligned}$$

Now, we take the conditional expectation given $x_t^{\tilde{\pi}} = x$ and use Lemma 2.13:

$$\begin{aligned}
&\mathbb{E}\big[J^{\mathrm{E}}(s \wedge \tau_n, x_{s\wedge\tau_n}^{\tilde{\pi}}; w, \pi)|x_t^{\tilde{\pi}} = x\big] \\
&= J^{\mathrm{E}}(t, x; w, \pi) + \mathbb{E}\bigg[\int_t^{s\wedge\tau_n} \bigg(J_t^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \quad + \int_{\mathbb{R}} \rho\sigma u \tilde{\pi}_v(u) du J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \\
&\quad + \frac{1}{2}\int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du J_{xx}^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi)\bigg) dv \\
&\quad + \int_t^{s\wedge\tau_n} \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du} J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) dB_v \Big| x_t^{\tilde{\pi}} = x\bigg] \\
&= J^{\mathrm{E}}(t, x; w, \pi) + \mathbb{E}\bigg[\int_t^{s\wedge\tau_n} \bigg(J_t^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \quad + \int_{\mathbb{R}} \rho\sigma u \tilde{\pi}_v(u) du J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \\
&\quad + \frac{1}{2}\int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du J_{xx}^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi)\bigg) dv \Big| x_t^{\tilde{\pi}} = x\bigg]
\end{aligned}$$

and therefore:

$$\begin{aligned}
J^{\mathrm{E}}(t, x; w, \pi) = \mathbb{E}\big[J^{\mathrm{E}}(s \wedge \tau_n, x_s^{\tilde{\pi}}; w, \pi) - \int_t^{s\wedge\tau_n} \bigg(J_t^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \\
+ \int_{\mathbb{R}} \rho\sigma u \tilde{\pi}_v(u) du J_x^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi) \\
+ \frac{1}{2}\int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du J_{xx}^{\mathrm{E}}(v, x_v^{\tilde{\pi}}; w, \pi)\bigg) dv \Big| x_t^{\tilde{\pi}} = x\bigg].
\end{aligned} \tag{5.5}$$

By Lemma 5.1, we have:

$$\begin{aligned}
J_t^{\mathrm{E}}(t, x; w, \pi) + \int_{\mathbb{R}} \bigg(\frac{1}{2}\sigma^2 u^2 J_{xx}^{\mathrm{E}}(t, x; w, \pi) + \rho\sigma u J_x^{\mathrm{E}}(t, x; w, \pi) \\
+ \lambda \ln(\pi_t(u))\bigg)\pi_t(u) du = 0,
\end{aligned}$$

for any $(t, x) \in [0, T) \times \mathbb{R}$. It follows that:

$$\begin{aligned}
J_t^{\mathrm{E}}(t, x; w, \pi) + \min_{\pi\in\mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \bigg(\frac{1}{2}\sigma^2 u^2 J_{xx}^{\mathrm{E}}(t, x; w, \pi) + \rho\sigma u J_x^{\mathrm{E}}(t, x; w, \pi) \\
+ \lambda \ln \pi(u)\bigg)\pi(u) du \leq 0,
\end{aligned} \tag{5.6}$$

for any $(t, x) \in [0, T) \times \mathbb{R}$.

By Lemma 3.11, we know that the minimizer in this inequality is given by the feedback policy $\tilde{\pi}(\cdot; t, x, w)$. It then follows that if we integrate (5.6) over the interval $[t, s \wedge \tau_n]$ with $\pi = \tilde{\pi}$ and substitute in equation (5.5), we obtain:

$$J^{\mathrm{E}}(t, x; w, \pi) \geq \mathbb{E}\big[J^{\mathrm{E}}(s \wedge \tau_n, x_{s\wedge\tau_n}^{\tilde{\pi}}; w, \pi) + \lambda \int_t^{s\wedge\tau_n} \int_{\mathbb{R}} \ln \tilde{\pi}_v(u) \tilde{\pi}_v(u) du dv \Big| x_t^{\tilde{\pi}} = x\bigg],$$

for $(t,x) \in [0,T] \times \mathbb{R}$ and $s \in [t,T]$.

We note that $J^{\mathrm{E}}(T,x;w,\pi) = J^{\mathrm{E}}(T,x;w,\tilde{\pi}) = (x-w)^2 - (w-z)^2$ for all $x \in \mathbb{R}$.

Then, taking $s = T$ and using the assumption that $\tilde{\pi}$ is admissible, we obtain by sending $n \to \infty$ and applying the dominated convergence theorem (see Appendix B.4), that:

$$\begin{aligned} J^{\mathrm{E}}(t,x;w,\pi) &\geq \mathbb{E}\Big[J^{\mathrm{E}}(T,x_T^{\tilde{\pi}};w,\tilde{\pi}) + \lambda \int_t^T \int_{\mathbb{R}} \ln \tilde{\pi}_v(u)\tilde{\pi}_v(u)dudv \Big| x_t^{\tilde{\pi}} = x\Big] \\ &= \mathbb{E}\Big[(x_T^{\tilde{\pi}} - w)^2 + \lambda \int_t^T \int_{\mathbb{R}} \ln \tilde{\pi}_v(u)\tilde{\pi}_v(u)dudv \Big| x_t^{\tilde{\pi}} = x\Big] - (w-z)^2 \\ &\overset{(3.9)}{=} J^{\mathrm{E}}(t,x;w,\tilde{\pi}) \end{aligned}$$

for any $(t,x) \in [0,T] \times \mathbb{R}$. □

Theorem 5.2 not only validates a policy improvement step but also suggests that Gaussian policies can continually improve the value function of any given, not necessarily Gaussian, policy. Hence, we can simplify the policy improvement step by focusing only on Gaussian policies without loss of generality.

5.2 Convergence Results

The optimal Gaussian policy (3.22) for problem (3.8) proposes that a suitable initial feedback policy may take the form:

$$\pi_0(u;t,x,w) = \mathcal{N}\Big(u \Big| a(x-w), c_1 e^{c_2(T-t)}\Big), \tag{5.7}$$

with $a, c_2 \in \mathbb{R}$ and $c_1 > 0$. It turns out that, theoretically, such a choice leads to the convergence of both the value functions and the policies in a finite number of iterations:

Theorem 5.3. *Consider $\pi_0(u;t,x,w) = \mathcal{N}(u \mid a(x-w), c_1 e^{c_2(T-t)})$, with $a, c_2 \in \mathbb{R}$ and $c_1 > 0$. Denote by $(\pi_n(u;t,x,w))_{n\geq 0}$ for $(t,x) \in [0,T] \times \mathbb{R}$ the sequence of feedback policies updated by the policy improvement scheme (5.3), and by $(J^E(t,x;w,\pi_n))_{n\geq 0}$ the sequence of the corresponding value functions. Then,*

$$\lim_{n\to\infty} \pi_n(\cdot;t,x,w) = \hat{\pi}(\cdot;t,x,w) \text{ weakly}$$

and

$$\lim_{n\to\infty} J^E(t,x;w,\pi_n) = V^E(t,x;w)$$

for any $(t,x,w) \in [0,T]\times\mathbb{R}\times\mathbb{R}$, where $\hat{\pi}$ and V^E are the optimal Gaussian policy (3.22) and the optimal value function (3.21), respectively.

Proof. To show that the feedback policy $\pi_0(\cdot;\cdot,\cdot,\cdot)$ generates an admissible open-loop policy $\boldsymbol{\pi}^0$ with respect to the initial (t,x), we refer to Lemma 3.12. The proof is analog but with a, c_2 and c_1 instead of $-\frac{\rho}{\sigma}$, $\frac{\lambda}{2\sigma^2}$ and ρ^2.

From the Feynman–Kac formula (see Section 4.4 in I. Karatzas & S. E. Shreve [15]), it follows that the corresponding value function $J^{\mathrm{E}}(\cdot,\cdot;w,\pi_0)$ satisfies the PDE:

$$J_t^{\mathrm{E}}(t,x;w,\pi_0) + \int_{\mathbb{R}} \Big(\frac{1}{2}\sigma^2 u^2 J_{xx}^{\mathrm{E}}(t,x;w,\pi_0) + \rho\sigma u J_x^{\mathrm{E}}(t,x;w,\pi_0) + \lambda \ln \pi_0(u;t,x,w) \Big) \pi_0(u;t,x,w)du = 0, \quad (t,x) \in [0,T) \times \mathbb{R}$$

with terminal condition $J^{\mathrm{E}}(T,x;w,\pi_0) = (x-w)^2 - (w-z)^2$. Simplifying this equation, we obtain for $(t,x) \in [0,T) \times \mathbb{R}$:

$$J_t^{\mathrm{E}}(t,x;w,\pi_0) + \Big(a^2(x-w)^2 + c_1 e^{c_2(T-t)} \Big) \frac{\sigma^2}{2} J_{xx}^{\mathrm{E}}(t,x;w,\pi_0) + a\rho\sigma(x-w) J_x^{\mathrm{E}}(t,x;w,\pi_0) - \frac{\lambda}{2}\Big(\ln(2\pi c_1) + c_2(T-t) + 1 \Big) = 0$$

A classical solution to this equation is given by:

$$\begin{aligned} J^{\mathrm{E}}(t,x;w,\pi_0) &= (x-w)^2 e^{\left(2\rho\sigma a + \sigma^2 a^2\right)(T-t)} + \frac{\lambda}{2}\Big(\ln(2\pi c_1)t - \frac{c_2}{2}(T-t)^2 + t \Big) + \frac{c_1}{c_2} e^{c_2(T-t)} \\ &= (x-w)^2 e^{\left(2\rho\sigma a + \sigma^2 a^2\right)(T-t)} + F_0(t), \end{aligned}$$

where $F_0(t)$ is obviously a smooth function that only depends on t.

It is clear that $J_{xx}^{\mathrm{E}}(t,x;w,\pi^0) > 0$ and $J^{\mathrm{E}}(\cdot,\cdot;w,\pi_0) \in C^{1,2}([0,T)\times\mathbb{R}) \cap C([0,T]\times\mathbb{R})$. Therefore, $J^{\mathrm{E}}(\cdot,\cdot;w,\pi^0)$ satisfies the conditions in Theorem 5.2 and we can apply the PIT. The improved policy is given by (5.3):

$$\begin{aligned} \pi_1(u;t,x,w) &= \mathcal{N}\Big(u \Big| -\frac{\rho}{\sigma}\frac{J_x^{\mathrm{E}}(t,x;w,\pi_0)}{J_{xx}^{\mathrm{E}}(t,x;w,\pi_0)}, \frac{\lambda}{\sigma^2 J_{xx}^{\mathrm{E}}(t,x;w,\pi_0)}\Big) \\ &= \mathcal{N}\Big(u \Big| -\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2 \exp\left((2\rho\sigma a + \sigma^2 a^2)(T-t)\right)}\Big). \end{aligned}$$

Again, we can use the Feynman-Kac formula and proceed as above, to obtain the corresponding value function:

$$J^{\mathrm{E}}(t,x;w,\pi_1) = (x-w)^2 e^{-\rho^2(T-t)} + F_1(t),$$

where $F_1(t)$ is again a smooth function that only depends on t. It is clear that $J_{xx}^{\mathrm{E}}(t,x;w,\pi_1) > 0$ and $J^{\mathrm{E}}(\cdot,\cdot;w,\pi_1) \in C^{1,2}([0,T)\times\mathbb{R}) \cap C([0,T]\times\mathbb{R})$. Therefore,

$J^{\mathrm{E}}(\cdot,\cdot;w,\pi_1)$ satisfies the conditions in Theorem 5.2 and we can apply the PIT again. The improved policy is given by (5.3):

$$\begin{aligned}\pi_2(u;t,x,w) &= \mathcal{N}\Big(u\Big| -\frac{\rho}{\sigma}\frac{J_x^{\mathrm{E}}(t,x;w,\pi_1)}{J_{xx}^{\mathrm{E}}(t,x;w,\pi_1)}, \frac{\lambda}{\sigma^2 J_{xx}^{\mathrm{E}}(t,x;w,\pi_1)}\Big)\\ &= \mathcal{N}\Big(u\Big| -\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}\Big).\end{aligned}$$

We observe that this policy coincides with the optimal feedback control $\hat{\pi}(\cdot;\cdot,\cdot,\cdot)$ and that the corresponding value function is the optimal value function V^{E}. Therefore, the policy and the value function will no longer strictly improve under the policy improvement scheme (5.3) for $n \geq 2$. Hence, the desired convergence results hold. □

These convergence results show that if we choose the initial policy as in (5.7), the learning scheme will, theoretically, converge after a finite number of iterations. Since, in practice, the value function for each policy can only be approximated, the process needs more iterations to converge. However, Theorem 5.2 suggests a theoretically well-founded policy improvement scheme, and Theorem 5.3 provides a beneficial starting policy. These results form the base for designing an implementable RL algorithm for the EMV problem.

5.3 The EMV Algorithm

In this section, we present the EMV algorithm. H. Wang and X. Y. Zhou first introduced this RL algorithm in [25] to solve the EMV problem (3.8). It consists of three simultaneously ongoing procedures: the policy evaluation, the policy improvement, and a self-correcting scheme for learning the Lagrange multiplier w.

5.3.1 Policy Evaluation

For the policy evaluation, we minimize the so-called Bellman's error. This error measures how much the value function fits the Bellman equation (for more details see K. Doya [10]).

By the proof of Lemma 3.5, we have for any admissible feedback control $\pi \in \mathcal{A}$ and any initial pair $(t,x) \in [0,T] \times \mathbb{R}$ and $s \in [t,T]$:

$$J^{\mathrm{E}}(t,x;w,\pi) = \mathbb{E}\Big[J^{\mathrm{E}}(s,x_s^{\pi};w,\pi) + \lambda\int_t^s\int_{\mathbb{R}} \pi_v(u)\ln(\pi_v(u))dudv\Big|x_t^{\pi}=x\Big],$$

where $\boldsymbol{\pi}$ is the admissible open-loop control generated from π with respect to the initial (t,x). Rearranging this equation and dividing both sides by $s-t$, we

obtain:

$$\mathbb{E}\left[\frac{J^{\mathrm{E}}(s,x_s^\pi;w,\pi)-J^{\mathrm{E}}(t,x_t^\pi;w,\pi)}{s-t}+\frac{\lambda}{s-t}\int_t^s\int_{\mathbb{R}}\pi_v(u)\ln(\pi_v(u))dudv\middle|x_t^\pi=x\right]=0.$$

Taking $s\to t$ gives rise to the Bellman's error:

Definition 5.4. The continuous-time Bellman's error of an admissible feedback control $\pi\in\mathcal{A}^{\mathrm{E}}$ is given for the initial pair $(t,x)\in[0,T]\times\mathbb{R}$ by:

$$\delta_t := \dot{J}_t^{\mathrm{E}}(t,x;w,\pi)+\lambda\int_{\mathbb{R}}\pi_t(u)\ln\pi_t(u)du, \tag{5.8}$$

where $\dot{J}_t^{\mathrm{E}}(t,x;w,\pi)=\frac{J^{\mathrm{E}}\left(t+\Delta t,x_{t+\Delta t}^\pi;w,\pi\right)-J^{\mathrm{E}}(t,x_t^\pi;w,\pi)}{\Delta t}$ is the total derivative and Δt is the discretization step for the learning algorithm.

A common practice is representing value functions and policies using (deep) neural networks for continuous RL problems. Since we obtained the parametric expressions in Theorem 3.13 and Theorem 5.3, we can avoid this.

We parameterize the value functions and the admissible feedback controls. We can rearrange the optimal value function V^{E} given by (3.21) in the following way for $(t,x)\in[0,T]\times\mathbb{R}$:

$$\begin{aligned}V^{\mathrm{E}}(t,x;w)&=(x-w)^2e^{-\rho^2(T-t)}+\frac{\lambda\rho^2}{4}\left(T^2-t^2\right)\\&\quad-\frac{\lambda}{2}\Big(\rho^2T-\ln\big(\frac{\sigma^2}{\pi\lambda}\big)\Big)(T-t)-(w-z)^2\\&=(x-w)^2e^{-\rho^2(T-t)}-\frac{\lambda\rho^2}{4}t^2+\frac{\lambda}{2}\Big(\rho^2T-\ln\big(\frac{\sigma^2}{\pi\lambda}\big)\Big)t+C,\end{aligned} \tag{5.9}$$

where C is a constant dependent on T, ρ, σ, w, z and λ. Hence, we consider the parameterized value functions denoted by V^θ, where $\theta=(\theta_0,\theta_1,\theta_2,\theta_3)^\top$:

$$V^\theta(t,x)=(x-w)^2e^{-\theta_3(T-t)}+\theta_2t^2+\theta_1t+\theta_0,\quad(t,x)\in[0,T]\times\mathbb{R}. \tag{5.10}$$

By comparing (5.9) and (5.10), we observe the following relation between θ_2 and θ_3:

$$\theta_2=-\frac{\lambda\theta_3}{4}. \tag{5.11}$$

Furthermore, by Theorem 5.3, we will focus on Gaussian distributions π with variance taking the form $c_1e^{c_2(T-t)}$, where $c_1>0$ and $c_2\in\mathbb{R}$. This in turn leads by (3.7) to the entropy:

$$\mathcal{H}(\pi)=\ln\big(\sqrt{2\pi c_1}\big)+\frac{1}{2}+\frac{1}{2}c_2(T-t).$$

Therefore, we consider the parameterized feedback controls π^{φ}, where $\varphi = (\varphi_1, \varphi_2)^\top$ with entropy:

$$\mathcal{H}\left(\pi_t^{\varphi}\right) = \varphi_1 + \varphi_2(T-t) \tag{5.12}$$

with $\varphi_1 \in \mathbb{R}$ and $\varphi_2 > 0$.

The objective of the policy evaluation procedure is to minimize the Bellman's error (5.8). In general, this can be carried out as follows. We first define the cost function:

$$C(\theta,\varphi) := \frac{1}{2}\mathbb{E}\left[\int_0^T |\delta_t|^2\, dt\right] = \frac{1}{2}\mathbb{E}\left[\int_0^T \left|\dot{V}_t^{\theta} + \lambda \int_{\mathbb{R}} \pi_t^{\varphi}(u)\ln \pi_t^{\varphi}(u) du\right|^2 dt\right], \tag{5.13}$$

where $\boldsymbol{\pi}^{\varphi} = (\pi_t^{\varphi})_{0\le t\le T}$ is generated from π^{φ} with respect to a given initial $(0, x_0)$ and $\dot{V}_t^{\theta}$ is the total derivative of the parametrized value function.

To approximate $C(\theta,\varphi)$ in an implementable algorithm, we first discretize $[0,T]$ into small intervals $[t_i, t_{i+1}]$ with equal length Δt, for $i = 0, 1, \ldots, l$, where $t_0 = 0$ and $t_{l+1} = T$. Then we collect a set of samples $\mathcal{D} = \{(t_i, x_i),\, i = 0, 1, \ldots, l+1\}$ in the following way:

The initial sample is $(0, x_0)$ for $i = 0$. Then we sample, at each time step t_i, $i = 0, 1, \ldots, l$, from $\pi_{t_i}^{\varphi}$ to obtain an allocation $u_i \in \mathbb{R}$ in the risky asset. Hence, we can then observe the wealth x_{i+1} at the next time instant t_{i+1}. We can now approximate (5.13) by:

$$C(\theta,\varphi) = \frac{1}{2}\sum_{(t_i,x_i)\in\mathcal{D}} \left(\dot{V}_t^{\theta}(t_i,x_i) + \lambda \int_{\mathbb{R}} \pi_{t_i}^{\varphi}(u)\ln \pi_{t_i}^{\varphi}(u) du\right)^2 \Delta t, \tag{5.14}$$

where $\dot{V}_t^{\theta}(t_i,x_i) = \frac{V^{\theta}(t_{i+1},x_{i+1}) - V^{\theta}(t_i,x_i)}{\Delta t}$.

Due to the policy improvement updating scheme (5.3), we are able to relate the parameters φ and θ to each other. Since the variance of the improved policy is given by $\frac{\lambda}{\sigma^2 V_{xx}^{\theta}}$ it follows that the variance of the policy π_t^{φ} is $\frac{\lambda}{2\sigma^2} e^{\theta_3(T-t)}$, resulting in the entropy $\frac{1}{2}\ln\frac{\pi e\lambda}{\sigma^2} + \frac{\theta_3}{2}(T-t)$. Comparing this to (5.12), we deduce:

$$\begin{aligned} \sigma^2 &= \lambda\pi e^{1-2\varphi_1} \\ \theta_3 &= 2\varphi_2 = \rho^2 \\ \theta_2 &= -\frac{\lambda\varphi_2}{2}. \end{aligned} \tag{5.15}$$

The improved policy, in turn, becomes, according to (5.3):

$$\begin{aligned}\pi(u;t,x,w) &= \mathcal{N}\left(u \mid -\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2}e^{\theta_3(T-t)}\right) \\ &= \mathcal{N}\left(u \mid -\sqrt{\frac{2\varphi_2}{\lambda\pi}}e^{\frac{2\varphi_1-1}{2}}(x-w), \frac{1}{2\pi}e^{2\varphi_2(T-t)+2\varphi_1-1}\right),\end{aligned} \tag{5.16}$$

where we have assumed that the true (unknown) Sharpe ratio $\rho > 0$.

Rewriting the objective (5.14) using $\mathcal{H}\left(\pi_t^{\varphi}\right) = \varphi_1 + \varphi_2(T-t)$, we obtain

$$C(\theta,\varphi) = \frac{1}{2}\sum_{(t_i,x_i)\in\mathcal{D}}\left(\dot{V}_t^{\theta}\left(t_i,x_i\right) - \lambda\big(\varphi_1 + \varphi_2\left(T-t_i\right)\big)\right)^2\Delta t.$$

Remark 5.5. We note that the assumption for the Sharpe ratio $\rho > 0$ does not limit the use of the algorithm. The case of a negative Sharpe ratio can be dealt with similarly by changing the sign of the mean in (5.16). Since, in reality, the true ρ is unknown, we use this algorithm only for simulation studies, where we choose the Sharpe ratio in advance. For the case of multiple risky assets, we will derive an algorithm that is applicable in practice with no such assumptions.

5.3.2 Policy Improvement

We are now able to devise the updating rules for the parameters. Using stochastic gradient descent algorithms (see, e.g., Goodfellow, Bengio & Courville [14]), one can minimize a function $Q : D \longrightarrow \mathbb{R}$ with $D \subset \mathbb{R}^n$ by performing the following update for $n \geq 1$:

$$z_{n+1} = z_n - \eta\nabla Q(z_n),$$

where $\eta > 0$ is a step size (sometimes called the learning rate in machine learning) and ∇ is the Nabla-operator.

Hence, to construct an updating rule for the parameters θ_1, φ_1 and φ_2 we com-

pute the respective derivatives of $C(\theta, \varphi)$:

$$\frac{\partial C}{\partial \theta_1} = \sum_{(t_i,x_i)\in\mathcal{D}} \Big(\dot{V}_t^\theta\left(t_i, x_i\right) - \lambda\big(\varphi_1 + \varphi_2\left(T - t_i\right)\big)\Big)\,\Delta t \tag{5.17}$$

$$\frac{\partial C}{\partial \varphi_1} = -\lambda \sum_{(t_i,x_i)\in\mathcal{D}} \Big(\dot{V}_t^\theta\left(t_i, x_i\right) - \lambda\big(\varphi_1 + \varphi_2\left(T - t_i\right)\big)\Big)\,\Delta t \tag{5.18}$$

$$\begin{aligned}\frac{\partial C}{\partial \varphi_2} = & \sum_{(t_i,x_i)\in\mathcal{D}} \Big(\dot{V}_t^\theta\left(t_i, x_i\right) - \lambda\big(\varphi_1 + \varphi_2\left(T - t_i\right)\big)\Big)\,\Delta t \\ & \cdot \Bigg(- \frac{2\left(x_{i+1} - w\right)^2 e^{-2\varphi_2(T-t_{i+1})}\left(T - t_{i+1}\right) - 2\left(x_i - w\right)^2 e^{-2\varphi_2(T-t_i)}\left(T - t_i\right)}{\Delta t} \\ & - \lambda\left(T - t_i\right) - \frac{\lambda}{2}\Big(\frac{t_{i+1}^2 - t_i^2}{\Delta t}\Big)\Bigg),\end{aligned} \tag{5.19}$$

where we used in the last equation that $\theta_3 = 2\varphi_2$ and $\theta_2 = -\frac{\lambda\varphi_2}{2}$.

The equations (5.17)-(5.19) lead to the following updating rules:

$$\begin{aligned}\theta_1 &\longleftarrow \theta_1 - \eta_{\theta_1}\frac{\partial C}{\partial \theta_1} \\ \varphi_1 &\longleftarrow \varphi_1 - \eta_{\varphi_1}\frac{\partial C}{\partial \varphi_1} \\ \varphi_2 &\longleftarrow \varphi_2 - \eta_{\varphi_2}\frac{\partial C}{\partial \varphi_2},\end{aligned}$$

where η_{θ_1}, η_{φ_1} and $\eta_{\varphi_2} > 0$ are learning rates.

Moreover, the parameters θ_2 and θ_3 are updated by (5.15):

$$\begin{aligned}\theta_2 &\longleftarrow -\frac{\lambda\varphi_2}{2} \\ \theta_3 &\longleftarrow 2\varphi_2\end{aligned}$$

and θ_0 is updated based on the terminal condition:

$$(x - w)^2 - (w - z)^2 = V^\theta(T, x; w) = (x - w)^2 + \theta_2 T^2 + \theta_1 T + \theta_0,$$

which yields

$$\theta_0 \longleftarrow -\theta_2 T^2 - \theta_1 T - (w - z)^2.$$

Finally, we provide a scheme for learning the underlying Lagrange multiplier w. Indeed, the constraint $\mathbb{E}\left[x_T\right] = z$ itself suggests the standard stochastic approximation update:

$$w \longleftarrow w - \alpha\left(x_T - z\right) \tag{5.20}$$

with $\alpha > 0$, being the learning rate. For the sake of stability in the learning process, we replace the terminal wealth x_T with a sample average:

$$\frac{1}{N}\sum_{j=1}^{N} x_T^j,$$

where $N \geq 1$ is the sample size and the x_T^j 's are the most recent N terminal wealth values obtained at the time when w is to be updated. Furthermore, we note that the learning scheme of w, (5.20), is self-correcting. Indeed, suppose the sample average of the terminal wealth is above the target z. In that case, the updating rule (5.20) will decrease w, which in turn decreases the mean of the exploratory Gaussian policy given in (5.16). It implies that there will be less risky allocation in the next step of actions for learning and optimizing, leading to, on average, a decreased terminal wealth.

5.3.3 Pseudo-Code

We now have all updating rules for the parameters available. Hence, we can present the pseudo-code for the EMV as in [25]:

Algorithm 1 EMV: Exploratory Mean–Variance Portfolio Selection

Input: Market Simulator, learning rates α, η_θ and η_φ, initial wealth x_0, target payoff z, investment horizon T, discretization Δ_t, exploration rate λ, number of iterations M and sample average size N.

Initialize θ, φ and w.

for $k = 1$ **to** M **do**
  for $i = 1$ **to** $\lfloor \frac{T}{\Delta t} \rfloor$ **do**
    Sample (t_i^k, x_i^k) from Market Simulator under π^φ.
    Obtain collected samples $\mathcal{D} = \{(t_j^k, x_j^k),\ 1 \leq j \leq i\}$.
    Update $\theta_1 \leftarrow \theta_1 - \eta_\theta \frac{\partial C}{\partial \theta_1}(\theta, \varphi)$ using (5.17).
    Update $\theta_3 \leftarrow 2\varphi_2$
    Update $\theta_2 \leftarrow -\frac{\lambda}{2}\varphi_2$
    Update $\theta_0 \leftarrow -\theta_2 T^2 - \theta_1 T - (w - z)^2$
    Update $\varphi \leftarrow \varphi - \eta_\varphi \nabla_\varphi C(\theta, \varphi)$ using (5.18) & (5.19).
  end for
  Update $\pi^\varphi \leftarrow \mathcal{N}\left(u \middle| -\sqrt{\frac{2\varphi_2}{\lambda\pi}} e^{\frac{2\varphi_1 - 1}{2}}(x - w), \frac{1}{2\pi} e^{2\varphi_2(T-t)+2\varphi_1-1}\right)$.
  if k mod $N == 0$ **then**
    Update $w \leftarrow w - \alpha\left(\frac{1}{N}\sum_{j=k-N+1}^{k} x^j_{\lfloor \frac{T}{\Delta t} \rfloor}\right)$.
  end if
end for

5.3.4 Implementation

In this section, we present the implementation of the EMV algorithm in Python 3.8.8. The only package we used is "numpy":

```
1    import numpy as np
```

We start with the functions that compute the derivatives of the cost function (5.17)-(5.19). By the following procedure, we are able to simplify the implementation and reduce the computational complexity.

First, we note that for the set of samples $\mathcal{D} = \{(t_i, x_i), i = 0, 1, ..., n\}$ the wealth x_{n+1} is unobserved and we observe x_n by sampling the allocation $u_{n-1} \in \mathbb{R}$ from the distribution $\pi^{\varphi}_{t_{n-1}}$. Therefore, we have that:

$$\begin{aligned} C(\theta, \varphi) &= \frac{1}{2} \sum_{(t_i, x_i) \in \mathcal{D}} \left(\dot{V}_t^{\theta} (t_i, x_i) - \lambda \mathcal{H}(\pi^{\varphi}_{t_i}) \right)^2 \Delta t \\ &= \frac{1}{2} \sum_{i=0}^{n-1} \left(\dot{V}_t^{\theta} (t_i, x_i) - \lambda \mathcal{H}(\pi^{\varphi}_{t_i}) \right)^2 \Delta t. \end{aligned}$$

Then, we assume that the time interval Δt is constant over time and we simplify the derivative of the parametrized value function $\dot{V}_t^{\theta}$ to:

$$\begin{aligned} &\dot{V}_t^{\theta} (t_i, x_i) \\ &= \frac{(x_{i+1} - w)^2 e^{-\theta_3 (T - t_{i+1})} - (x_i - w)^2 e^{-\theta_3 (T - t_i)} + \theta_2 (t_{i+1}^2 - t_i^2) + \theta_1 (t_{i+1} - t_i)}{\Delta t} \\ &= \frac{(x_{i+1} - w)^2 e^{-\theta_3 (T - t_{i+1})} - (x_i - w)^2 e^{-\theta_3 (T - t_i)} + \theta_2 \left(2 t_i \Delta t + (\Delta t)^2 \right) + \theta_1 \Delta t}{\Delta t} \end{aligned}$$

Further, we use some standard formulas for partial sums of series to obtain:

$$\sum_{i=0}^{n-1} t_i = \Delta t \sum_{i=0}^{n-1} i = \Delta t \frac{n(n-1)}{2}$$

$$\sum_{i=0}^{n-1} t_i^2 = (\Delta t)^2 \sum_{i=0}^{n-1} i^2 = (\Delta t)^2 \frac{n(n-1)(2n-1)}{6}.$$

Hence, by assuming $t_0 = 0$, we obtain for (5.17):

$$\begin{aligned} \frac{\partial C}{\partial \theta_1} (\theta, \varphi) &= \sum_{(t_i, x_i) \in \mathcal{D}} \left(\dot{V}_t^{\theta} (t_i, x_i) - \lambda \left(\varphi_1 + \varphi_2 (T - t_i) \right) \right) \Delta t \\ &= (x_n - w)^2 e^{-\theta_3 (T - t_n)} - (x_0 - w)^2 e^{-\theta_3 T} \\ &\quad + \theta_2 (\Delta t)^2 n^2 + n \theta_1 \Delta t - n \lambda \Delta t (\varphi_1 + \varphi_2 T) + \lambda \varphi_2 (\Delta t)^2 \frac{n(n-1)}{2} \end{aligned}$$

and for (5.18):

$$\frac{\partial C}{\partial \varphi_1}(\theta, \varphi) = -\lambda \frac{\partial C}{\partial \theta_1}(\theta, \varphi)$$

and for (5.19):

$$\begin{aligned}\frac{\partial C}{\partial \varphi_2}(\theta, \varphi) = -\lambda \left(T + \frac{\Delta t}{2}\right) \frac{\partial C}{\partial \theta_1}(\theta, \varphi) \\ + 2 \sum_{(t_i, x_i) \in \mathcal{D}} \left(\dot{V}_t^\theta (t_i, x_i) - \lambda (\varphi_1 + \varphi_2 (T - t_i))\right) \\ \cdot \Big((x_i - w)^2 e^{-2\varphi_2 (T - t_i)} (T - t_i) \\ - (x_{i+1} - w)^2 e^{-2\varphi_2 (T - t_{i+1})} (T - t_{i+1}) \Big).\end{aligned}$$

These computations lead to the following implementation, where we need as input the parameter vectors $\theta \in \mathbb{R}^4$ and $\varphi \in \mathbb{R}^2$, the time interval $\Delta t > 0$, the Lagrange multiplier $w \in \mathbb{R}$, the terminal time $T > 0$, the set of samples $\mathcal{D}$ and the exploration weight $\lambda > 0$:

```
def Deriv_Cost_Funct_Theta_1(theta,phi,dt,w,T,D,lam):

    ## Input:
    # theta: Parameter (numpy array with shape (4,1))
    # phi: Parameter (numpy array with shape (2,1))
    # dt: time intervall (float)
    # w: Lagrange multiplier (float)
    # T: terminal time (float or int)
    # D: Set of samples (numpy array with shape (2,k) with k>1)
    # lam: Exploration weight (float or int)
    #
    ## Output:
    # The derivative of the cost function with respect to theta_1
    # (float)

    # Size of set of samples
    n = D.shape[1]

    # Computation of first term
    Term1 = (D[1,-1]-w)**2*np.exp(-theta[3]*(T-D[0,-1]))
            -(D[1,0]-w)**2*np.exp(-theta[3]*(T-D[0,0]))

    # Computation of second term
    Term2 = theta[2]*dt**2*(n-1)*(n-1)+(n-1)*theta[1]*dt
            -lam*dt*((n-1)*(phi[0]+phi[1]*T-phi[1]*dt*(n-2)/2))

    DCT1 = Term1 + Term2

    return DCT1
```

Listing 5.1: The derivative of the cost function with respect to θ_1: $\frac{\partial C}{\partial \theta_1}(\theta,\varphi)$

```
def Deriv_Cost_Funct_Phi(theta,phi,dt,w,T,D,lam):

    ## Input:
    # theta: Parameter (numpy array with shape (4,1))
    # phi: Parameter (numpy array with shape (2,1))
    # dt: time intervall (float)
    # w: Lagrange multiplier (float)
    # T: terminal time (float or int)
    # D: Set of samples (numpy array with shape (2,k) with k>1)
    # lam: Exploration weight (float or int)
    #
    ## Output:
    # The derivative of the cost function with respect to phi
    # (numpy array with shape (2,1))

    # Size of set of samples
    n = D.shape[1]

    ## Derivative of cost function with respect to theta_1
    DCT1 = Deriv_Cost_Funct_Theta_1(theta,phi,dt,w,T,D,lam)

    ## Derivative of cost function with respect to phi_0
    DCP0 = -lam*DCT1

    ## Derivative of cost function with respect to phi_1

    # Value function
    f_x = lambda x: (x-w)**2
    f1_t = lambda t: np.exp(-theta[3]*(T-t))
    f2_t = lambda t: theta[2]*t**2+theta[1]*t+theta[0]

    V = f_x(D[1,:])*f1_t(D[0,:])+f2_t(D[0,:])

    # Derivative of value function
    Deriv_V = (V[1:]-V[:-1])/dt

    # Differential entropy term
    f3_t = lambda t: -lam*(phi[0]+phi[1]*(T-t))

    H = f3_t(D[0,:-1])

    # First factor
    C1 = (Deriv_V+H)*dt

    f2 = lambda x: 2*(x[1]-w)**2*np.exp(-2*phi[1]*(T-x[0]))
                   *(T-x[0])+lam/2*x[0]**2
    f2 = f2(D)

    # Second factor
    C2 = -(f2[1:]-f2[:-1])/dt

    Term1 = -lam*(T+dt/2)*DCT1
    Term2 = np.dot(C1,C2)

```

```
    DCP1 =  Term1 + Term2

    return np.array([DCP0,DCP1])
```

Listing 5.2: The derivative of the cost function with respect to φ: $\nabla_{\varphi} C(\theta,\varphi)$

These two functions are used in the implementation of the EMV algorithm. We follow the pseudo-code (Algorithm 1) and need the same input parameters. Moreover, we need price returns as input that represent the investment environment and allow interaction:

```
def EMV(M,N,eta_phi,eta_theta,alpha,x0,z,w,Returns,theta,phi,dt,T,
    lam):

    ## Input:
    # M: Number of iterations (int>0)
    # N: Sample average size (int>0)
    # eta_phi, eta_theta, alpha: learning rates (float)
    # x0: Initial wealth (float)
    # z: Target payoff (float)
    # w: Lagrange multiplier (float)
    # Returns: Returns of risky asset
    # (numpy array with shape (T/dt,1))
    # theta: Initail parameter (numpy array with shape (4,1))
    # phi: Initial parameter (numpy array with shape (2,1))
    # dt: time intervall (float > 0)
    # T: terminal time (float or int > 0)
    # lam: Exploration weight (float or int)

    ## Output:
    # mean_funct: The learned mean function of the control process
    # var_funct: The learned variance function of the control
    # process
    # terminal_wealth: The achieved terminal wealth in each
    # iteration
    # SR_sq: The learned \rho^2 in each iteration

    ## Initialization

    # Initialize arrays to store the terminal wealth and learned \
    rho^2 of each
    # iteration
    terminal_wealth = np.zeros([M])
    SR_sq = np.zeros([M])

    # Number of time steps
    T1 = int(np.floor(T/dt))

    # Initial mean and variance function
    mean_funct = lambda x: -np.sqrt(2*phi[1]/(lam*np.pi))
                           *np.exp((2*phi[0]-1)/2)*(x-w)
    var_funct =  lambda x: 1/(2*np.pi)*np.exp(2*phi[1]*(T-x)
                           +2*phi[0]-1)

```

```
    # Set of observations/samples
    D = np.zeros([2,T1])
    D[1,0] = np.array(x0)
    D[0,:] = np.linspace(0,T,T1)

    ## Iterations
    for k in range(1, M+1):

        # Initialization
        x = float(x0)

        # Episodes
        for i in range(1, T1):

            # Mean and Variance
            variance = var_funct((i-1)*dt)
            mean = mean_funct(x)

            # Strategy sample
            u = np.random.normal(loc=mean, scale=variance)

            # Compute the wealth at the next time point
            x = x+float(u*Returns[i-1])
            # Update the set of samples
            D[1,i] = np.array(x)

            # Compute the Derivative of the cost function with
            # respect to theta
            Cost_Derivative_theta = Deriv_Cost_Funct_Theta_1(theta,
                                    phi,dt,w,T,D[:,:(i+1)],lam)

            # Update theta
            theta[1] = theta[1] - eta_theta*Cost_Derivative_theta
            theta[3] = 2*phi[1]
            theta[2] = -lam/2*phi[1]
            theta[0] = -theta[2]*T**2-theta[1]*T-(w-z)**2

            # Compute the Derivative of the cost function with
            # respect to phi
            Cost_Derivative_phi = Deriv_Cost_Funct_Phi(theta,phi,
                                  dt,w,T,D[:,:(i+1)],lam)

            # Update phi
            phi = phi - eta_phi*Cost_Derivative_phi

        # Store terminal wealth of k-th iteration
        terminal_wealth[k-1] = x

        # Update Lagrange multiplier
        if k % N == 0:
            w = w-alpha*(1/N*(np.sum(terminal_wealth[(k-N):k]))-z)

        # Update control process
        mean_funct = lambda x: -np.sqrt(2*phi[1]/(lam*np.pi))
```

```
96                                  *np.exp((2*phi[0]-1)/2)*(x-w)
97          var_funct =  lambda x: 1/(2*np.pi)*np.exp(2*phi[1]*(T-x)
98                                  +2*phi[0]-1)
99
100         # Store learned squared Sharpe ratio
101         SR_sq[k-1] = theta[3]
102
103         # Print current learning state
104         if k % 100 == 0:
105             print(str(round(100*k/M,2)) + " %" +
106                   "  |  Average Terminal Wealth: " +
107                   str(round(np.mean(terminal_wealth[k-100:k]),2)) +
108                   "  |  Average Variance: " +
109                   str(round(np.var(terminal_wealth[k-100:k]),4)))
110
111     return [[mean_funct,var_funct,terminal_wealth,SR_sq]]
```

Listing 5.3: The EMV algorithm

Remark 5.6. We note that we assumed $\rho > 0$ in (5.16). However, as already mentioned in Remark 5.5, in the case of $\rho < 0$, we can change the sign of the mean function in lines 37 and 95 of the implementation 5.3. Hence, this code is applicable in the following Chapter 6, the simulation study, since we will know the true ρ in each simulation. Unfortunately, in practice, the parameters of the risky asset's price process are unknown. For this reason, in Chapter 7, we will derive the EMV algorithm in the multi-dimensional setup, where we do not make such assumptions on ρ.

Chapter 6

Simulation Study

In this chapter, we perform simulations to analyze the effectiveness and efficiency of the EMV algorithm. For this purpose, we carry out numerical simulations in a stationary and non-stationary market environment. Then, we analyze the learning curves of the sample mean and sample variance of the terminal wealth for different price simulations. This chapter follows mainly the procedure in Chapter 5 of H. Wang & X. Y. Zhou [25]. For the detailed implementations used in this chapter, we refer to https://github.com/TimGyger/Master-Thesis.

6.1 The stationary Market Case

In a stationary market environment the price processes are simulated according to the geometric Brownian motion:

$$\begin{aligned} dS_t &= S_t(\mu dt + \sigma dB_t), \quad 0 \leq t \leq T \\ S_0 &= s_0 > 0, \end{aligned}$$

with constant $\mu \in \mathbb{R}$ and $\sigma > 0$.

For simplicity, we consider the MV problem over a 1-year period with daily rebalancing. Hence, we take $T = 1$ and since there are about 252 trading days a year, we choose $\Delta t = \frac{1}{252}$.

Moreover, the annualized return and volatility will be taken from the two sets $\mu \in \{-50\%, -30\%, -10\%, 0\%, 10\%, 30\%, 50\%\}$ and $\sigma \in \{10\%, 20\%, 30\%, 40\%\}$ to simulate typical stock prices and also more extreme cases. The risk-free interest rate is taken to be $r = 2\%$. Therefore, we can compute the true underlying ρ of the risky asset and modify the implementation in Listing 5.3 as explained in Remark 5.6. We then sample a new price process corresponding to the chosen drift μ and volatility σ in each iteration of the algorithm to generate the returns.

Furthermore, we start from a normalized initial wealth $x_0 = 1$ and consider

the MV problem with a 40% annualized target return what corresponds to an investment target of $z = 1.4$ in the MV problem (2.5). We fix the model parameters for the EMV algorithm. We take the total training episodes $M = 20'000$, the sample size $N = 10$ for learning the Lagrange multiplier w and the exploration weight $\lambda = 2$. Further, we consider the fixed learning rates $\alpha = 0.05$ and $\eta_\theta = \eta_\varphi = 0.0005$.

In Table 6.1, we summarize the results under different market scenarios corresponding to different combinations of μ and σ. For each market scenario, we present the annualized return and standard deviation of the last $2'000$ values of the trained terminal wealth and the corresponding annualized Sharpe ratio.

Market Scenarios	Return	Standard Deviation	Sharpe Ratio
$\sigma = 10\%$, $\mu = -50\%$	39.34%	7.2%	5.46
$\sigma = 10\%$, $\mu = -30\%$	38.37%	12.31%	3.12
$\sigma = 10\%$, $\mu = -10\%$	31.99%	35.12%	0.91
$\sigma = 10\%$, $\mu = 0\%$	26.73%	36.04%	0.74
$\sigma = 10\%$, $\mu = 10\%$	29.85%	39.32%	0.76
$\sigma = 10\%$, $\mu = 30\%$	37.97%	14.44%	2.63
$\sigma = 10\%$, $\mu = 50\%$	39.24%	7.96%	4.93
$\sigma = 20\%$, $\mu = -50\%$	38.09%	15.14%	2.51
$\sigma = 20\%$, $\mu = -30\%$	38.97%	25.99%	1.42
$\sigma = 20\%$, $\mu = -10\%$	29.7%	46.91%	0.63
$\sigma = 20\%$, $\mu = 0\%$	28.41%	47.76%	0.59
$\sigma = 20\%$, $\mu = 10\%$	29.31%	46.83%	0.63
$\sigma = 20\%$, $\mu = 30\%$	35.18%	29.95%	1.17
$\sigma = 20\%$, $\mu = 50\%$	37.89%	16.66%	2.27
$\sigma = 30\%$, $\mu = -50\%$	36.23%	23.84%	1.52
$\sigma = 30\%$, $\mu = -30\%$	32.78%	40.42%	0.81
$\sigma = 30\%$, $\mu = -10\%$	29.46%	52.15%	0.56
$\sigma = 30\%$, $\mu = 0\%$	28.99%	53.12%	0.55
$\sigma = 30\%$, $\mu = 10\%$	29.91%	50.81%	0.59
$\sigma = 30\%$, $\mu = 30\%$	33.22%	43.75%	0.76
$\sigma = 30\%$, $\mu = 50\%$	36.31%	26.19%	1.39
$\sigma = 40\%$, $\mu = -50\%$	34.54%	33.91%	1.02
$\sigma = 40\%$, $\mu = -30\%$	31.69%	49.51%	0.64
$\sigma = 40\%$, $\mu = -10\%$	29.11%	56.37%	0.52
$\sigma = 40\%$, $\mu = 0\%$	29.18%	57.3%	0.51
$\sigma = 40\%$, $\mu = 10\%$	30.0%	54.05%	0.55
$\sigma = 40\%$, $\mu = 30\%$	31.7%	52.1%	0.61
$\sigma = 40\%$, $\mu = 50\%$	35.12%	38.04%	0.92

Table 6.1: The annualized return, standard deviation and Sharpe ratio for EMV.

Overall, we can conclude that the EMV algorithm performs well during training. In every market scenario, a positive terminal wealth is achieved and, in many cases, even close to the investment target. Moreover, we observe that the algorithm performs best in market scenarios of lower volatility and higher mean return. The more volatile the stock price is, the harder it gets for the EMV algorithm to reach the investment target.

In Figures 6.1 & 6.2, we present the learning curves for the EMV algorithm in the market scenario $\mu = -30\%$ and $\sigma = 20\%$. We illustrate the sample mean and sample variance of every non-overlapping 50 terminal wealth values as the learning proceeds. We observe that the EMV algorithm converges fast, achieving relatively good performance even in the early phase of the learning process. This is consistent with the convergence result in Theorem 5.3.

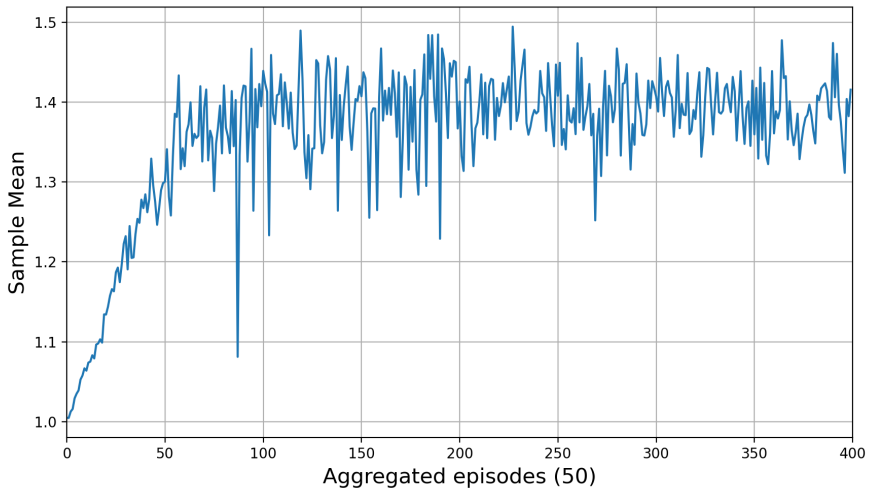

Figure 6.1: Learning curves of sample means of terminal wealth (over every 50 iterations) for EMV ($\mu = -30\%$, $\sigma = 20\%$).

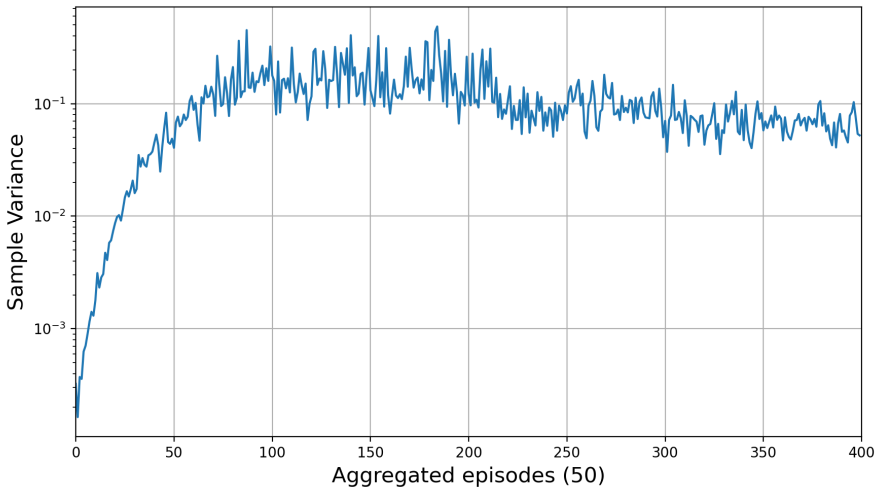

Figure 6.2: Learning curves of sample variance of terminal wealth (over every 50 iterations) for EMV ($\mu = -30\%$, $\sigma = 20\%$).

In addition to analyzing the training performance through the terminal wealth, we can also investigate the approximation accuracy of the EMV approach relative to the ground truth values for $\rho = \frac{\mu - r}{\sigma}$. By (5.15), we obtain the learned value for ρ^2 by θ_3. Therefore, we can compute the relative error between the estimated and the true ρ^2 by:

$$\mathrm{Err}_\theta := \frac{|\theta_3 - \rho^2|}{\rho^2}.$$

The approximation errors for the different market scenarios are presented in Table 6.2. As already observed by H. Wang & X. Y. Zhou in [25] these results are not satisfying.

Market Scenarios	Learned ρ^2	Ground Truth ρ^2	Relative Error Err_θ
$\sigma = 10\%$, $\mu = -50\%$	2.69	27.04	0.9
$\sigma = 10\%$, $\mu = -30\%$	3.18	10.24	0.69
$\sigma = 10\%$, $\mu = -10\%$	4.1	1.44	1.85
$\sigma = 10\%$, $\mu = 0\%$	3.78	0.04	93.41
$\sigma = 10\%$, $\mu = 10\%$	4.08	0.64	5.37
$\sigma = 10\%$, $\mu = 30\%$	3.32	7.84	0.58
$\sigma = 10\%$, $\mu = 50\%$	2.75	23.04	0.88
$\sigma = 20\%$, $\mu = -50\%$	3.91	6.76	0.42
$\sigma = 20\%$, $\mu = -30\%$	4.56	2.56	0.78
$\sigma = 20\%$, $\mu = -10\%$	5.01	0.36	12.9
$\sigma = 20\%$, $\mu = 0\%$	4.92	0.01	490.98
$\sigma = 20\%$, $\mu = 10\%$	4.96	0.16	30.02
$\sigma = 20\%$, $\mu = 30\%$	4.69	1.96	1.39
$\sigma = 20\%$, $\mu = 50\%$	4.01	5.76	0.3
$\sigma = 30\%$, $\mu = -50\%$	4.87	3	0.62
$\sigma = 30\%$, $\mu = -30\%$	5.38	1.14	3.73
$\sigma = 30\%$, $\mu = -10\%$	5.59	0.16	33.96
$\sigma = 30\%$, $\mu = 0\%$	5.63	0	1264.89
$\sigma = 30\%$, $\mu = 10\%$	5.54	0.07	76.9
$\sigma = 30\%$, $\mu = 30\%$	5.45	0.87	5.25
$\sigma = 30\%$, $\mu = 50\%$	4.96	2.56	0.94
$\sigma = 40\%$, $\mu = -50\%$	5.58	1.69	2.3
$\sigma = 40\%$, $\mu = -30\%$	5.91	0.64	8.23
$\sigma = 40\%$, $\mu = -10\%$	6.07	0.09	66.48
$\sigma = 40\%$, $\mu = 0\%$	6.09	0	2436.02
$\sigma = 40\%$, $\mu = 10\%$	6.03	0.04	149.64
$\sigma = 40\%$, $\mu = 30\%$	6.03	0.49	11.3
$\sigma = 40\%$, $\mu = 50\%$	5.63	1.44	2.91

Table 6.2: Comparison of the learned and true ρ^2.

Furthermore, in [25], it is mentioned that the error Err_θ heavily depends on the values of the hyperparameters for the EMV algorithm. To investigate this, we repeated the training phase for the market scenario $\sigma = 10\%$ and $\mu = 30\%$ multiple times with different, randomly chosen initial parameters θ, φ, and w. We fixed the other parameters as in the simulations before. To use random parameters in a reasonable range, we sampled them from a normal distribution with a mean equal to the theoretically true value given by (5.10) and (5.15) and with a variance of 2. In Appendix D in Table D.1, we present the mean return and the standard deviation of the last $2'000$ terminal wealth's for each of the 50 training processes and besides, the learned ρ^2 and the corresponding relative error Err_θ. If we take the mean of the learned parameter ρ^2 over all simulations, we obtain a value

of 7.44, corresponding to a small relative error of $\text{Err}_\theta = 0.05$. Unfortunately, this promising result goes hand in hand with an averaged standard deviation of 104.16% over the 50 simulations. This indicates that the learning curve did not converge yet. Indeed, if we repeat the 50 simulations with the doubled number of iterations than before ($M = 40'000$), we gain an averaged standard deviation of the terminal wealth's of 45.55%. However, the learned parameter ρ^2 is in average 5.24 what corresponds to a big relative error of $\text{Err}_\theta = 0.33$. We refer to Appendix D Table D.2 for the complete results. Consequently, we note that the EMV algorithm has trouble approximating the parameter ρ^2 even if we set the hyperparameters near their theoretically true values. Nevertheless, since in real-life applications, the ground truth of μ, σ, or ρ is not available, we weight the effectiveness more than the accuracy of the estimations. Moreover, we note that the EMV algorithm achieves good annualized returns and standard deviations for different hyperparameters (if the training process is long enough). Therefore, we can conclude that the EMV algorithm is not sensitive to the initial parameters of θ, φ, and w.

6.2 The nonstationary Market Case

The application of RL in quantitative finance differs from other domains in the underlying unknown investment environment that is typically time-varying. In this section, we analyze the performance of the EMV algorithm in a nonstationary market scenario where a stochastic factor model models the price process. To have a well-defined learning problem, the stochastic factor needs to change much slower than the learning process. Therefore, we assume that the factor driving the volatility is a slowly varying diffusion process (see, e.g., Section 2.1.2 in Fouque, Papanicolaou, Sircar, & Solna [12]). The price process follows:

$$\begin{aligned} dS_t &= S_t(\mu_t dt + \sigma_t dB_t), \quad 0 \leq t \leq T \\ S_0 &= s_0 > 0, \end{aligned}$$

with $(\mu_t)_{t\in[0,T]}$ and $(\sigma_t)_{t\in[0,T]}$ being, respectively, the real-valued drift and volatility processes restricted to each simulation episode $[0, T]$. Hence, these processes may vary across different episodes.

Analog to Section 2.1, we obtain the following controlled wealth dynamics over each episode $[0, T]$:

$$\begin{aligned} dx_t^u &= \sigma_t u_t(\rho_t dt + dB_t), \quad 0 \leq t \leq T \\ x_0^u &= x_0 \in \mathbb{R}, \end{aligned}$$

with the stochastic factor model:

$$\begin{aligned} d\rho_t &= \delta dt, \\ d\sigma_t &= \sigma_t(\delta dt + \sqrt{\delta} dB_t^1), \quad 0 \leq t \leq MT, \end{aligned}$$

with the initial values $\rho_0 \in \mathbb{R}$ and $\sigma_0 > 0$, a small parameter $\delta > 0$ and a Brownian motion $B^1 = (B_t^1)_{t\in[0,T]}$ satisfying $d[B, B^1]_t = \gamma dt$, where $\gamma \in [-1, 1]$ is the correlation between the two Brownian motions. Further, we note that the time-horizon for the stochastic factor model is MT indicating that ρ_t and σ_t change across all episodes.

In Figures 6.3 & 6.4, we present the learning curves for the EMV algorithm in the nonstationary market scenario with $r = 2\%$ and with initials $\rho_0 = -3.2$ and $\sigma_0 = 20\%$ corresponding to the case $\mu_0 = -30\%$. Further, we take $\delta = 0.0001$ such that the factors stay in reasonable ranges and $\gamma = 0$.

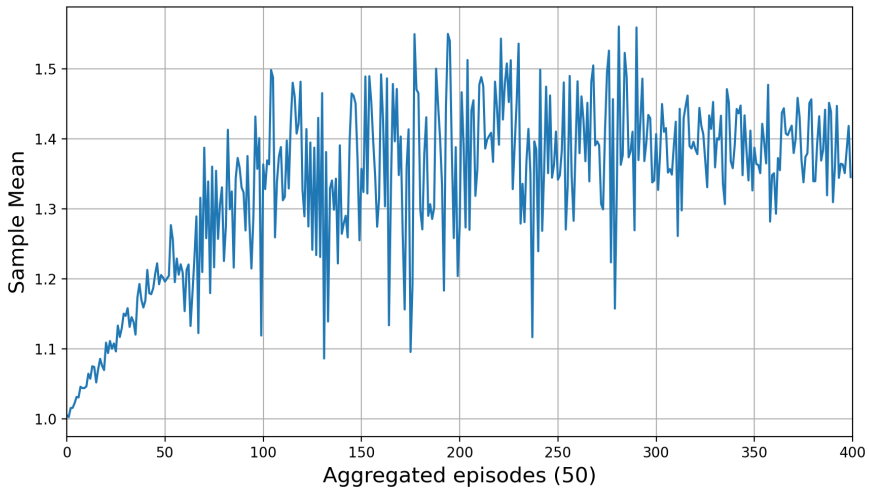

Figure 6.3: Learning curves of sample means of terminal wealth (over every 50 iterations) for EMV ($\mu_0 = -30\%$, $\sigma_0 = 20\%$, $\delta = 0.0001$, $\gamma = 0$).

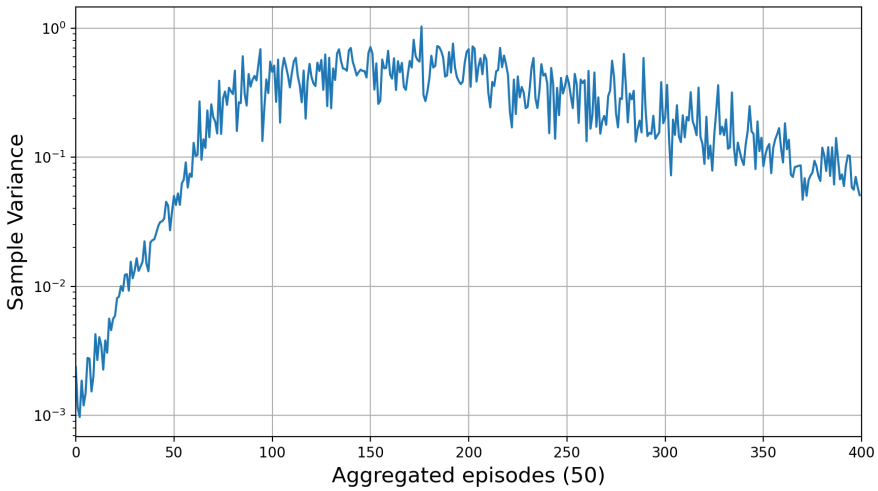

Figure 6.4: Learning curves of sample variance of terminal wealth (over every 50 iterations) for EMV ($\mu_0 = -30\%$, $\sigma_0 = 20\%$, $\delta = 0.0001$, $\gamma = 0$).

The two learning curves show that the EMV algorithm also converges in the nonstationary market case quite fast.

Besides, we report the annualized mean return and standard deviation of the last $2'000$ values of the trained terminal wealth given by 38.8% and 29.5%, respectively. This results in a Sharpe ratio of 1.25.

Hence, we can conclude that the EMV algorithm also achieves a stable learning performance in a nonstationary market scenario.

Chapter 7

Multi-dimensional Setting

To examine the efficiency of the EMV algorithm on real data and compare its performance against baseline approaches, we first introduce the EMV problem and algorithm in a multi-dimensional setting. Hence, in this chapter, we present the main results of Chapter 3 and 5 for the case of multiple risky assets. The proofs of the theorems are conceptually identical to the one risky asset case, and only a few computations are more complex in the multi-dimensional case. Therefore, we will not repeat every proof of the main statements but refer to the corresponding proofs in H. Wang [24].

To avoid any confusion with former chapters, we denote vectors and vector stochastic processes with an arrow $\vec{a}$, matrices in bold text $\boldsymbol{a}$ and repeating definitions of Chapter 3 for the multiple asset case with a superscript ^{d}.

7.1 Controlled Wealth Dynamics

We first derive the controlled wealth process for the case of d risky assets. Let $W = (W_t)_{t\in[0,T]}$ be a d-dimensional Brownian motion over $[0,T]$ with $W_t = \left(W_t^1, \dots, W_t^d\right)^\top$ defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ that satisfies the usual conditions (see Appendix A.1 & B.2). The price of the i-th risky asset follows a geometric Brownian motion:

$$\begin{aligned} dS_t^i &= S_t^i \left(\mu^i dt + \vec{\sigma}^i \cdot dW_t\right), \quad 0 \le t \le T, \quad i = 1, \dots, d, \\ S_0^i &= s_0^i > 0, \end{aligned}$$

with drift parameter $\mu^i \in \mathbb{R}$ and the volatility coefficients $\vec{\sigma}^i = \left(\sigma^{1i}, \dots, \sigma^{di}\right)^\top \in \mathbb{R}^d$ of the i-th risky asset, respectively.

We denote the drift vector by $\vec{\mu} \in \mathbb{R}^d$, and the volatility matrix by $\boldsymbol{\sigma} \in \mathbb{R}^{d\times d}$, whose i-th column represents the (co-)volatility vector $\vec{\sigma}^i$ of the i-th risky asset. As in the one-dimensional setup the riskless asset has a constant interest rate

$r > 0$. Further, we assume that $\boldsymbol{\sigma}$ is non-degenerate (invertible) and hence there exists a d-dimensional vector $\vec{\rho}$ that satisfies $\boldsymbol{\sigma}^\top \vec{\rho} = \vec{\mu} - r\vec{1}$, where $\vec{1}$ is the d-dimensional vector with all components being 1. Also in the multi-dimensional case, $\vec{\rho}$ is known as Sharpe ratio or the market price of risk.

We denote by $\vec{\boldsymbol{u}} = (\vec{u}_t)_{t\in[0,T]}$ with $\vec{u}_t = \left(u_t^1, \ldots, u_t^d\right)^\top$ the discounted dollar value put in the d risky assets, respectively, at time t. By using the self-financing condition, we can then derive similar as in Section 2.1 that the discounted wealth process $X^{\vec{\boldsymbol{u}}} = (X_t^{\vec{\boldsymbol{u}}})_{t\in[0,T]}$ satisfies:

$$\begin{aligned} dX_t^{\vec{\boldsymbol{u}}} &= \sum_{i=1}^{d} u_t^i \left(\left(\mu^i - r\right) dt + \vec{\sigma}^i \cdot dW_t \right) = \boldsymbol{\sigma}\vec{u}_t \cdot \left(\vec{\rho} dt + dW_t\right), \quad 0 \le t \le T, \\ X_0^{\vec{\boldsymbol{u}}} &= X_0, \end{aligned} \tag{7.1}$$

where $X_0 \in \mathbb{R}$ is the initial capital.

As in Section 3.1, we introduce the exploratory version of the state dynamics (7.1). In this formulation, the control process $\vec{\boldsymbol{u}} = (\vec{u}_t)_{t\in[0,T]}$ is randomized to represent exploration in RL, leading to a measure-valued or distributional control process denoted by $\boldsymbol{\pi}^d = (\pi_t^d)_{t\in[0,T]}$. The dynamics (7.1) is changed to

$$dX_t^{\boldsymbol{\pi}^d} = \left(\int_{\mathbb{R}^d} \vec{\rho}^\top \boldsymbol{\sigma} \vec{u} \pi_t^d(\vec{u}) d\vec{u} \right) dt + \left(\int_{\mathbb{R}^d} \vec{u}^\top \boldsymbol{\sigma}^\top \boldsymbol{\sigma} \vec{u} \pi_t^d(\vec{u}) d\vec{u} \right)^{\frac{1}{2}} dB_t, \tag{7.2}$$

where $(B_t)_{t\in[0,T]}$ is a one-dimensional standard Brownian motion on the filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$.

7.2 Formulation of the multi-dimensional EMV Problem

Analog to Section 3.3, we define the set of admissible distributional control processes with respect to the initial wealth and time:

Definition 7.1. For $(s, y) \in [0, T) \times \mathbb{R}$ we denote the set of admissible control distributions as $\mathcal{A}^d(s, y)$ and the distributional control process $\boldsymbol{\pi}^d = (\pi_t^d)_{t\in[s,T]}$ belongs to $\mathcal{A}^d(s, y)$ if:

1. $\forall t \in [s, T] : \pi_t^d \in \mathcal{P}(\mathbb{R}^d)$, $\mathbb{P}$-a.s.
2. $\forall A \in \mathcal{B}(\mathbb{R}^d) : \left(\int_A \pi_t^d(\vec{u}) d\vec{u} \right)_{t\in[s,T]}$ is $\mathcal{F}_t$-progressively measurable
3. $\mathbb{E}\left[\int_s^T \int_{\mathbb{R}^d} |\boldsymbol{\sigma}\vec{u}|^2 \pi_t^d(\vec{u}) d\vec{u} dt \right] < \infty$

4. $\mathbb{E}\big[|(X_T^{\boldsymbol{\pi}^d} - w)^2 + \lambda \int_s^T \int_{\mathbb{R}^d} \pi_t^d(\vec{u}) \ln\big(\pi_t^d(\vec{u})\big) d\vec{u}dt| \big| X_s^{\boldsymbol{\pi}^d} = y\big] < \infty,$

where $\lambda > 0$ is the exploration weight.

Furthermore, we define the corresponding admissible feedback controls:

Definition 7.2. We denote the set of admissible feedback controls by $\mathcal{A}^d$. A deterministic mapping $\pi^d(\cdot;\cdot,\cdot)$ is called an admissible feedback control if:

1. $\pi^d(\cdot;t,x)$ is a density for each $(t,x) \in [0,T] \times \mathbb{R}$.
2. For each $(s,y) \in [0,T) \times \mathbb{R}$ the following SDE:

$$\begin{aligned} dX_t^{\pi^d} &= \left(\int_{\mathbb{R}^d} \vec{\rho}^\top \boldsymbol{\sigma} \vec{u} \pi^d \left(\vec{u}; t, X_t^{\pi^d}\right) d\vec{u} \right) dt \\ &\quad + \left(\int_{\mathbb{R}^d} \vec{u}^\top \boldsymbol{\sigma}^\top \boldsymbol{\sigma} \vec{u} \pi^d \left(\vec{u}; t, X_t^{\pi^d}\right) d\vec{u} \right)^{\frac{1}{2}} dW_t, \quad t \in [s,T] \\ X_s^{\pi^d} &= y \end{aligned}$$

has a unique strong solution $(X_t^{\pi^d})_{t\in[s,T]}$ and the open-loop control $\boldsymbol{\pi}^d = (\pi_t^d)_{t\in[s,T]} \in \mathcal{A}^d(s,y)$ where $\pi_t^d := \pi^d(\cdot;t,X_t^{\pi^d})$.

In this case, the open-loop control $\boldsymbol{\pi}^d$ is said to be generated from the feedback control $\pi^d(\cdot;\cdot,\cdot)$ with respect to the initial time and wealth (s,y).

Now, we state the multi-dimensional, entropy-regularized EMV problem for a fixed $w \in \mathbb{R}$:

$$\min_{\boldsymbol{\pi}^d \in \mathcal{A}^d(0,X_0)} \mathbb{E}\left[\left(X_T^{\boldsymbol{\pi}^d} - w \right)^2 + \lambda \int_0^T \int_{\mathbb{R}^d} \pi_t^d(\vec{u}) \ln\left(\pi_t^d(\vec{u})\right) d\vec{u}dt \right] - (w-z)^2, \tag{7.3}$$

where $\mathcal{A}^d(0,X_0)$ is the set of admissible distributional control processes on $[0,T]$ and $\lambda > 0$ is the exploration weight.

Furthermore, we define the value and optimal value function:

Definition 7.3. For a fixed $w \in \mathbb{R}$, we define the value function under any given admissible feedback control $\pi(\cdot;\cdot,\cdot,w) \in \mathcal{A}^d$:

$$\begin{aligned} J^d(s,y;w,\pi^d) := \mathbb{E}\bigg[(X_T^{\pi^d} - w)^2 + \lambda \int_s^T \int_{\mathbb{R}^d} \pi_t^d(\vec{u}) \ln\big(\pi_t^d(\vec{u})\big) d\vec{u}dt \Big| x_s^{\pi^d} = y \bigg] \\ - (w-z)^2, \end{aligned} \tag{7.4}$$

for $(s,y) \in [0,T) \times \mathbb{R}$, where $\boldsymbol{\pi}^d = (\pi_t^d)_{t\in[s,T]}$ is the admissible open-loop control generated from $\pi^d(\cdot;\cdot,\cdot,w)$ with respect to the initial (s,y) such that $\pi_t^d = \pi^d(\cdot;t,X_t^\pi,w)$. Further, we define the optimal value function as:

$$V^d(s,y;w) := \inf_{\boldsymbol{\pi}^d\in\mathcal{A}^d(s,y)} \mathbb{E}\left[(X_T^{\boldsymbol{\pi}^d} - w)^2 + \lambda\int_s^T\int_{\mathbb{R}^d} \pi_t^d(\vec{u})\ln\left(\pi_t^d(\vec{u})\right)d\vec{u}dt \middle| X_s^{\boldsymbol{\pi}^d} = y\right] - (w-z)^2, \tag{7.5}$$

for $(s,y) \in [0,T)\times\mathbb{R}$.

We denote the optimal control process for problem (7.3) by $\hat{\boldsymbol{\pi}}^d = (\hat{\pi}_t^d)_{t\in[0,T]}$. This strategy is generated from the optimal feedback control denoted by $\hat{\pi}^d = \hat{\pi}^d(\cdot;\cdot,\cdot)$ with respect to the initial $(0,X_0)$.

7.3 Solution of the multi-dimensional EMV Problem

To solve the exploratory MV problem (7.3), we apply the classical Bellman's principle of optimality for the optimal value function V^d as in Lemma 3.5:

$$V^d(t,x;w) = \inf_{\boldsymbol{\pi}^d\in\mathcal{A}^d(t,x)} \mathbb{E}\left[V^d\left(s, X_s^{\boldsymbol{\pi}^d};w\right) + \lambda\int_t^s\int_{\mathbb{R}^d}\pi_v^d(\vec{u})\ln\pi_v^d(\vec{u})d\vec{u}dv \mid X_t^{\boldsymbol{\pi}^d} = x\right],$$

for $x\in\mathbb{R}$ and $0\le t<s\le T$. Following the standard arguments as in Theorem 3.6, we deduce that V^d satisfies the Hamilton-Jacobi-Bellman (HJB) equation:

$$\begin{aligned} &v_t(t,x;w) + \min_{\pi\in\mathcal{P}(\mathbb{R}^d)}\int_{\mathbb{R}^d}\left(\frac{1}{2}\vec{u}^\top\boldsymbol{\sigma}^\top\boldsymbol{\sigma}\vec{u}v_{xx}(t,x;w)\right.\\ &\qquad\left.+\vec{\rho}^\top\boldsymbol{\sigma}\vec{u}v_x(t,x;w) + \lambda\ln\pi(\vec{u})\right)\pi(\vec{u})d\vec{u} = 0, \quad (t,x)\in[0,T)\times\mathbb{R}\\ &v(T,x;w) = (x-w)^2 - (w-z)^2. \end{aligned}$$

Following the concepts in the proof of Theorem 3.13, we obtain the following results:

Theorem 7.4. *The optimal value function of the multi-dimensional, entropy-regularized, exploratory MV problem (7.3) is given by:*

$$\begin{aligned} V^d(t,x;w) = (x-w)^2e^{-\vec{\rho}^\top\vec{\rho}(T-t)} + \frac{\lambda d}{4}\vec{\rho}^\top\vec{\rho}\left(T^2-t^2\right)\\ -\frac{\lambda d}{2}\left(\vec{\rho}^\top\vec{\rho}T - \frac{1}{d}\ln\frac{\left|\boldsymbol{\sigma}^\top\boldsymbol{\sigma}\right|}{\pi\lambda}\right)(T-t) - (w-z)^2, \end{aligned} \tag{7.6}$$

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

$$\hat{\pi}^d(\vec{u};t,x,w) = \mathcal{N}\Big(\vec{u}\Big| -\boldsymbol{\sigma}^{-1}\vec{\rho}(x-w), \Big(\boldsymbol{\sigma}^\top\boldsymbol{\sigma}\Big)^{-1}\frac{\lambda}{2}e^{\vec{\rho}^\top\vec{\rho}(T-t)}\Big). \tag{7.7}$$

The associated optimal wealth process under $\hat{\pi}$ *is the unique solution of the stochastic differential equation*

$$dX_t^{\hat{\pi}^d} = -\vec{\rho}^\top\vec{\rho}\Big(X_t^{\hat{\pi}^d} - w\Big)\,dt + \left(\vec{\rho}^\top\vec{\rho}\Big(X_t^{\hat{\pi}^d} - w\Big)^2 + \frac{\lambda}{2}e^{\vec{\rho}^\top\vec{\rho}(T-t)}\right)^{\frac{1}{2}} dB_t, \quad t \in [0,T]$$
$$X_0^{\hat{\pi}^d} = X_0.$$

Finally, the Lagrange multiplier w *is given by* $w = \frac{ze^{\vec{\rho}^\top\vec{\rho}T} - X_0}{e^{\vec{\rho}^\top\vec{\rho}T} - 1}$.

We refer to Theorem 1 in H. Wang [24] for the detailed proof.

7.4 The RL Algorithm

Similar to Section 5.1, we can state a policy improvement theorem (PIT):

Theorem 7.5. *Let* $w \in \mathbb{R}$ *be fixed and* $\pi^d = \pi^d(\cdot;\cdot,\cdot,w)$ *be an arbitrarily given admissible feedback control. Suppose that the corresponding value function* $J^d(\cdot,\cdot;w,\pi^d) \in C^{1,2}([0,T)\times\mathbb{R}) \cap C^0([0,T]\times\mathbb{R})$ *and satisfies* $J^d_{xx}(t,x;w,\pi^d) > 0$, *for any* $(t,x) \in [0,T)\times\mathbb{R}$. *Suppose further that the feedback policy* $\tilde{\pi}^d$ *defined by:*

$$\tilde{\pi}^d(\vec{u};t,x,w) = \mathcal{N}\Big(\vec{u}\Big| -\boldsymbol{\sigma}^{-1}\vec{\rho}\frac{J^d_x(t,x;w,\pi^d)}{J^d_{xx}(t,x;w,\pi^d)}, \Big(\boldsymbol{\sigma}^\top\boldsymbol{\sigma}\Big)^{-1}\frac{\lambda}{J^d_{xx}(t,x;w,\pi^d)}\Big) \tag{7.8}$$

is admissible. Then,

$$J^d(t,x;w,\tilde{\pi}^d) \le J^d(t,x;w,\pi^d), \quad (t,x) \in [0,T]\times\mathbb{R}.$$

We refer to Theorem 3 in H. Wang [24] for the detailed proof of this PIT.

The following result shows the convergence of both the value functions and the policies from a specifically parameterized Gaussian policy:

Theorem 7.6. *Let* $\pi_0^d(\vec{u};t,x,w) = \mathcal{N}\left(\vec{u} \mid \vec{\alpha}(x-w), \boldsymbol{\Sigma}e^{\beta(T-t)}\right)$, *with* $\vec{\alpha} \in \mathbb{R}^d, \beta \in \mathbb{R}$ *and* $\boldsymbol{\Sigma}$ *being a* $d\times d$ *positive definite matrix. Denote by* $\left(\pi_n^d(\vec{u};t,x,w)\right)_{n\ge1}$ *the sequence of feedback policies updated by the policy improvement scheme (7.8), and* $\left(J^d(t,x;w,\pi_n^d)\right)_{n\ge1}$ *the sequence of the corresponding value functions. Then,*

$$\lim_{n\to\infty}\pi_n^d(\cdot;t,x,w) = \hat{\pi}^d(\cdot;t,x,w) \text{ weakly,}$$

and

$$\lim_{n\to\infty} J^d(t,x;w,\pi_n^d) = V^d(t,x;w)$$

for any $(t,x,w) \in [0,T]\times\mathbb{R}\times\mathbb{R}$, *where* $\hat{\pi}^d$ *and* V^d *are the optimal Gaussian policy (7.7) and the optimal value function (7.6), respectively.*

We refer to Theorem 4 in H. Wang [24] for the detailed proof of this theorem.

With these results, we can derive the EMV algorithm in a similar way as in Section 5.3.

First of all, we parameterize the value function (7.6) to:

$$V^\theta(t,x) = (x-w)^2 e^{-\theta_3(T-t)} + \theta_2 t^2 + \theta_1 t + \theta_0, \quad (t,x)\in[0,T]\times\mathbb{R}, \tag{7.9}$$

where $\theta = (\theta_0,\theta_1,\theta_2,\theta_3)^\top$. We have a similar relation between θ_2 and θ_3 as in the one-dimensional case:

$$\theta_2 = -\frac{\lambda\theta_3 d}{4}. \tag{7.10}$$

Furthermore, by Theorem 7.5, we will focus on multivariate Gaussian policies π^d with variance taking the form $\boldsymbol{\Sigma}e^{\beta(T-t)}$. By using the formula of the entropy for multivariate Gaussian distributions (see Lemma A.4 in Appendix A.4), we obtain:

$$\begin{aligned}\mathcal{H}(\pi^d) &= \frac{1}{2}\ln\Big((2\pi e)^d \det\big(\boldsymbol{\Sigma}e^{\beta(T-t)}\big)\Big)\\ &= \frac{d}{2}(1+\ln(2\pi)) + \frac{1}{2}\ln\big(\det(\boldsymbol{\Sigma})\big) + \frac{d}{2}\beta(T-t).\end{aligned}$$

Due to the improvement updating scheme (7.8) the variance of the improved policy $\tilde{\pi}^d$ is given by $(\boldsymbol{\sigma}^\top\boldsymbol{\sigma})^{-1}\frac{\lambda}{J^d_{xx}(t,x;w,\pi^d)}$. By substituting the parameterized value function (7.9), we obtain the variance $(\boldsymbol{\sigma}^\top\boldsymbol{\sigma})^{-1}\frac{\lambda}{2}e^{\theta_3(T-t)}$, resulting in the entropy:

$$\mathcal{H}(\tilde{\pi}^d) = \frac{d}{2}\big(1+\ln(\lambda\pi)\big) + \frac{1}{2}\ln\Big(\det\big((\boldsymbol{\sigma}^\top\boldsymbol{\sigma})^{-1}\big)\Big) + \frac{d}{2}\theta_3(T-t).$$

Therefore, we consider the parameterized feedback controls $\pi^{\boldsymbol{\varphi}}$, where $\boldsymbol{\varphi}$ represents the two parameters $\boldsymbol{\varphi}_1\in\mathbb{R}^{d\times d}$ and $\vec{\varphi}_2\in\mathbb{R}^d$, with entropy:

$$\mathcal{H}(\pi_t^{\boldsymbol{\varphi}}) = \frac{d}{2}\big(1+\ln(\lambda\pi)\big) + \frac{1}{2}\ln\big(\det\big((\boldsymbol{\varphi}_1^\top\boldsymbol{\varphi}_1)^{-1}\big)\big) + \vec{\varphi}_2^\top\vec{\varphi}_2(T-t).$$

Hence, we obtain the following relations:

$$\begin{aligned}\theta_3 &= \frac{2}{d}\vec{\varphi}_2^\top\vec{\varphi}_2 = \vec{\rho}^\top\vec{\rho}\\ \theta_2 &= -\frac{\lambda}{2}\vec{\varphi}_2^\top\vec{\varphi}_2\\ \boldsymbol{\sigma} &= \boldsymbol{\varphi}_1\end{aligned} \tag{7.11}$$

and the improved policy, in turn, becomes, according to (7.8):

$$\begin{aligned}\pi^d(\vec{u};t,x,w) &= \mathcal{N}\Big(\vec{u}\Big| -\boldsymbol{\sigma}^{-1}\vec{\rho}(x-w), (\boldsymbol{\sigma}^\top\boldsymbol{\sigma})^{-1}\frac{\lambda}{2}e^{\theta_3(T-t)}\Big)\\ &= \mathcal{N}\Big(\vec{u}\Big| -\boldsymbol{\varphi}_1^{-1}\vec{\varphi}_2\sqrt{\frac{2}{d}}(x-w), \frac{\lambda}{2}(\boldsymbol{\varphi}_1^\top\boldsymbol{\varphi}_1)^{-1}e^{\frac{2}{d}\vec{\varphi}_2^\top\vec{\varphi}_2(T-t)}\Big).\end{aligned}$$

We note that with these parametrizations an assumption similar to Remark 5.5 is not necessary.

For the policy evaluation step, we minimize the following discretized cost function:

$$C_d(\theta,\boldsymbol{\varphi}) = \frac{1}{2}\sum_{(t_i,x_i)\in\mathcal{D}}\left(\dot{V}_t^\theta(t_i,x_i) + \lambda\int_{\mathbb{R}^d}\pi_{t_i}^{\boldsymbol{\varphi}}(\vec{u})\ln\pi_{t_i}^{\boldsymbol{\varphi}}(\vec{u})d\vec{u}\right)^2\Delta t,$$

where $\mathcal{D}$ is the set of samples. Rewriting this cost function using $\mathcal{H}(\pi_t^{\boldsymbol{\varphi}})$, we obtain:

$$\begin{aligned}C_d(\theta,\boldsymbol{\varphi}) &= \frac{1}{2}\sum_{(t_i,x_i)\in\mathcal{D}}\Bigg(\dot{V}_t^\theta(t_i,x_i)\\ &\quad -\lambda\Big(\frac{d}{2}\big(1+\ln(\lambda\pi)\big)+\frac{1}{2}\ln\Big(\det\big((\boldsymbol{\varphi}_1^\top\boldsymbol{\varphi}_1)^{-1}\big)\Big)+\vec{\varphi}_2^\top\vec{\varphi}_2(T-t_i)\Big)\Bigg)^2\Delta t\\ &= \frac{1}{2}\sum_{(t_i,x_i)\in\mathcal{D}}\Bigg(\dot{V}_t^\theta(t_i,x_i)\\ &\quad -\lambda\Big(\frac{d}{2}\big(1+\ln(\lambda\pi)\big)-\ln\big(\det(\boldsymbol{\varphi}_1)\big)+\vec{\varphi}_2^\top\vec{\varphi}_2(T-t_i)\Big)\Bigg)^2\Delta t,\end{aligned}$$

where we used the following properties of the determinant (see Chapter 2 in H. Anton & C. Rorres [1]):

$$\begin{aligned}\det(\boldsymbol{A}^{-1}) &= \frac{1}{\det(\boldsymbol{A})}\\ \det(\boldsymbol{BC}) &= \det(\boldsymbol{B})\cdot\det(\boldsymbol{C})\\ \det(\boldsymbol{B}^\top) &= \det(\boldsymbol{B}),\end{aligned}$$

for $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ invertible and $\boldsymbol{B},\boldsymbol{C}\in\mathbb{R}^{n\times n}$.

For the policy improvement step, we construct an updating rule for the parameters θ_1, θ_2, $\boldsymbol{\varphi}_1$ and $\vec{\varphi}_2$ by computing the respective derivatives of $C_d(\theta,\boldsymbol{\varphi})$:

$$\begin{aligned}\frac{\partial C_d}{\partial\theta_1} = \sum_{(t_i,x_i)\in\mathcal{D}}\Bigg(&\dot{V}_t^\theta(t_i,x_i)\\ &-\lambda\Big(\frac{d}{2}\big(1+\ln(\lambda\pi)\big)-\ln\big(\det(\boldsymbol{\varphi}_1)\big)+\vec{\varphi}_2^\top\vec{\varphi}_2(T-t_i)\Big)\Bigg)\Delta t\end{aligned} \tag{7.12}$$

$$\begin{aligned}
\frac{\partial C_d}{\partial \boldsymbol{\varphi}_1} &= \sum_{(t_i,x_i)\in\mathcal{D}} \Big(\dot{V}_t^\theta\,(t_i,x_i) \\
&\quad - \lambda\Big(\frac{d}{2}\big(1+\ln(\lambda\pi)\big) - \ln\big(\det(\boldsymbol{\varphi}_1)\big) + \vec{\varphi}_2^\top\vec{\varphi}_2(T-t_i)\Big)\Big)\Delta t\lambda(\boldsymbol{\varphi}_1^{-1})^\top, \\
\frac{\partial C_d}{\partial \vec{\varphi}_2} &= \sum_{(t_i,x_i)\in\mathcal{D}} \Big(\dot{V}_t^\theta\,(t_i,x_i) \\
&\quad - \lambda\Big(\frac{d}{2}\big(1+\ln(\lambda\pi)\big) - \ln\big(\det(\boldsymbol{\varphi}_1)\big) + \vec{\varphi}_2^\top\vec{\varphi}_2(T-t_i)\Big)\Big)\Delta t \\
&\quad \cdot \Big(- \frac{4\vec{\varphi}_2}{d\Delta t}\Big((x_{i+1}-w)^2\, e^{-\frac{2}{d}\vec{\varphi}_2^\top\vec{\varphi}_2(T-t_{i+1})}\,(T-t_{i+1}) \\
&\quad - (x_i-w)^2\, e^{-\frac{2}{d}\vec{\varphi}_2^\top\vec{\varphi}_2(T-t_i)}\,(T-t_i)\Big) - 2\vec{\varphi}_2\lambda\,(T-t_i) - \frac{\lambda\vec{\varphi}_2(t_{i+1}^2-t_i^2)}{\Delta t}\Big),
\end{aligned} \tag{7.13}$$

where we used in the last equation that $\theta_3 = \frac{2}{d}\vec{\varphi}_2^\top\vec{\varphi}_2$ and $\theta_2 = -\frac{\lambda}{2}\vec{\varphi}_2^\top\vec{\varphi}_2$. The equations (7.12) & (7.13) leads to the following updating rules:

$$\begin{aligned}
\theta_1 &\longleftarrow \theta_1 - \eta_{\theta_1}\frac{\partial C_d}{\partial \theta_1} \\
\boldsymbol{\varphi}_1 &\longleftarrow \boldsymbol{\varphi}_1 - \eta_{\boldsymbol{\varphi}_1}\frac{\partial C_d}{\partial \boldsymbol{\varphi}_1} \\
\vec{\varphi}_2 &\longleftarrow \vec{\varphi}_2 - \eta_{\vec{\varphi}_2}\frac{\partial C_d}{\partial \varphi_2},
\end{aligned}$$

where η_{θ_1}, $\eta_{\boldsymbol{\varphi}_1}$ and $\eta_{\vec{\varphi}_2} > 0$ are learning rates.

Moreover, the parameters θ_2 and θ_3 are updated by (7.11):

$$\begin{aligned}
\theta_2 &\longleftarrow -\frac{\lambda}{2}\vec{\varphi}_2^\top\vec{\varphi}_2 \\
\theta_3 &\longleftarrow \frac{2}{d}\vec{\varphi}_2^\top\vec{\varphi}_2
\end{aligned}$$

and θ_0 is updated based on the terminal condition:

$$(x-w)^2 - (w-z)^2 = V^\theta(T,x;w) = (x-w)^2 + \theta_2 T^2 + \theta_1 T + \theta_0,$$

which yields

$$\theta_0 \longleftarrow -\theta_2 T^2 - \theta_1 T - (w-z)^2.$$

Finally, we provide the same scheme as in the one-dimensional case for learning the underlying Lagrange multiplier w:

$$w \longleftarrow w - \alpha\,(x_T - z) \tag{7.14}$$

with $\alpha > 0$, being the learning rate. In implementation, we can replace x_T in (7.14) by a sample average $\frac{1}{N}\sum_j x_T^j$ to have a more stable learning process.

We now have all updating rules for the parameters available. Hence, we can present the pseudo-code for the multi-dimensional EMV:

Algorithm 2 d-EMV: Multi-dimensional Exploratory MV Portfolio Selection

Input: Market Simulator, learning rates α, η_θ and $\eta_{\boldsymbol{\varphi}}$, initial wealth X_0, target payoff z, investment horizon T, discretization Δt, exploration rate λ, number of iterations M and sample average size N.

Initialize θ, $\boldsymbol{\varphi}$ and w.

for $k = 1$ **to** M **do**
  for $i = 1$ **to** $\lfloor \frac{T}{\Delta t} \rfloor$ **do**
    Sample (t_i^k, x_i^k) from Market Simulator under π^φ.
    Obtain collected samples $\mathcal{D} = \{\left(t_j^k, x_j^k\right), 1 \le j \le i\}$.
    Update $\theta_1 \leftarrow \theta_1 - \eta_\theta \frac{\partial C_d}{\partial \theta_1}(\theta, \boldsymbol{\varphi})$ using (7.12).
    Update $\theta_3 \leftarrow \frac{2}{d}\vec{\varphi}_2^\top \vec{\varphi}_2$
    Update $\theta_2 \leftarrow -\frac{\lambda}{2}\vec{\varphi}_2^\top \vec{\varphi}_2$
    Update $\theta_0 \leftarrow -\theta_2 T^2 - \theta_1 T - (w - z)^2$
    Update $\boldsymbol{\varphi} \leftarrow \boldsymbol{\varphi} - \eta_{\boldsymbol{\varphi}} \nabla_{\boldsymbol{\varphi}} C_d(\theta, \boldsymbol{\varphi})$ using (7.13).
  end for
  Update $\pi^{\boldsymbol{\varphi}} \leftarrow \mathcal{N}\left(\vec{u} \mid -\boldsymbol{\varphi}_1^{-1}\vec{\varphi}_2\sqrt{\frac{2}{d}}(x - w), \frac{\lambda}{2}(\boldsymbol{\varphi}_1^\top \boldsymbol{\varphi}_1)^{-1} e^{\frac{2}{d}\vec{\varphi}_2^\top \vec{\varphi}_2 (T-t)}\right)$.
  if k mod $N == 0$ **then**
    Update $w \leftarrow w - \alpha\left(\frac{1}{N}\sum_{j=k-N+1}^{k} x^j_{\lfloor \frac{T}{\Delta t} \rfloor}\right)$.
  end if
end for

7.5 Implementation

In this section, we present the implementation of the d-EMV algorithm (Algorithm 2).

We start with the functions that compute the derivatives of the cost function (7.12) & (7.13). The implementation is similar to the one-dimensional case presented in 5.1 and 5.2.

```
def Deriv_Cost_Funct_Theta_1(theta,phi,dt,w,T,D,lam,d):

    ## Input:
    # theta: Parameter (numpy array with shape (4,1))
```

```
    # phi: Parameter (list of numpy arrays with shape (d,d)
    # and (d,1))
    # dt: time intervall (float)
    # w: Lagrange multiplier (float)
    # T: terminal time (float or int)
    # D: Set of samples (numpy array with shape (2,k) with k>1)
    # lam: Exploration weight (float or int)
    # d: number of risky assets (int)
    #
    ## Output:
    # The derivative of the cost function with respect to theta_1
    # (float)

    # Size of set of samples
    n = D.shape[1]

    ## Derivative of cost function with respect to theta_1

    # Computation of first term
    Term1 = (D[1,-1]-w)**2*np.exp(-theta[3]*(T-D[0,-1]))
            -(D[1,0]-w)**2*np.exp(-theta[3]*(T-D[0,0]))

    # Computation of second term
    Term2 = theta[2]*dt**2*(n-1)*(n-1)+(n-1)*theta[1]*dt
            -lam*dt*((n-1)*(d/2*(1+np.log(np.pi*lam))
            -np.log(det(phi[0]))+np.dot(phi[1],phi[1])*T
            -np.dot(phi[1],phi[1])*dt*(n-2)/2))

    DCT1 = Term1 + Term2

    return DCT1
```

Listing 7.1: The derivative of the cost function with respect to θ_1: $\frac{\partial C_d}{\partial \theta_1}(\theta,\varphi)$

```
def Deriv_Cost_Funct_Phi(theta,phi,dt,w,T,D,lam,d):

    ## Input:
    # theta: Parameter (numpy array with shape (4,1))
    # phi: Parameter (list of numpy arrays with shape (d,d)
    # and (d,1))
    # dt: time intervall (float)
    # w: Lagrange multiplier (float)
    # T: terminal time (float or int)
    # D: Set of samples (numpy array with shape (2,k) with k>1)
    # lam: Exploration weight (float or int)
    # d: number of risky assets (int)
    #
    ## Output:
    # The derivative of the cost function with respect to phi
    # (list of numpy arrays with shape (d,d) and (d,1))

    # Size of set of samples
    n = D.shape[1]

```

```
    ## Derivative of cost function with respect to theta_1
    DCT1 = Deriv_Cost_Funct_Theta_1(theta,phi,dt,w,T,D,lam,d)

    ## Derivative of cost function with respect to phi_0
    DCP0 = lam*inv(phi[0]).T*DCT1

    ## Derivative of cost function with respect to phi_1

    # Value function
    f_x = lambda x: (x-w)**2
    f1_t = lambda t: np.exp(-theta[3]*(T-t))
    f2_t = lambda t: theta[2]*t**2+theta[1]*t+theta[0]

    V = f_x(D[1,:])*f1_t(D[0,:])+f2_t(D[0,:])

    # Derivative of value function
    Deriv_V = (V[1:]-V[:-1])/dt

    # Differential entropy term
    H = list(map(lambda x: -lam*(d/2*(1+np.log(lam*np.pi))
                           -np.log(det(phi[0]))
                           +np.dot(phi[1],phi[1])*(T-x)),D[0,:-1]))

    # First factor
    C1 = (Deriv_V+H)*dt

    f2 = list(map(lambda x: 4*phi[1]/d*(x[1]-w)**2
                            *np.exp(-2*np.dot(phi[1],phi[1])/d
                            *(T-x[0]))*(T-x[0])+lam*phi[1]*x[0]**2,
                            D.T))
    f3 = list(map(lambda x: -2*phi[1]*lam*(T-x),D[0,:-1]))

    # Second factor
    C2 = -(np.asarray(f2[1:])-np.asarray(f2[:-1]))/dt+f3
    C2 = np.reshape(np.asarray(C2),(n-1,d))

    DCP1 = sum(np.multiply(C2, C1[:, np.newaxis]))

    return [DCP0,DCP1]
```

Listing 7.2: The derivative of the cost function with respect to $\boldsymbol{\varphi}$: $\nabla_{\boldsymbol{\varphi}} C_d(\theta,\boldsymbol{\varphi})$

Finally, we present the implementation of the d-EMV algorithm (Algorithm 2):

```
def EMV(M,N,eta_phi,eta_theta,alpha,x0,z,w,Returns,theta,phi,dt,T,
    lam,d):

    ## Input:
    # M: Number of iterations (int>0)
    # N: Sample average size (int > 0)
    # eta_phi, eta_theta, alpha: learning rates (float)
    # x0: Initial wealth (float)
    # z: Target payoff (float)
```

```
    # w: Lagrange multiplier (float)
    # Returns: Returns of risky assets (numpy array
    # with shape (T/dt,d))
    # theta: Initail parameter (numpy array with shape (4,1))
    # phi: Initial parameter (list of numpy arrays with shape (d,d)
    # and (d,1))
    # dt: time intervall (float > 0)
    # T: terminal time (float or int > 0)
    # lam: Exploration weight (float or int > 0)
    # d: number of risky assets (int > 0)
    #
    ## Output:
    # mean_funct: The learned mean function of the control process
    # var_funct: The learned variance function of the control
    # process
    # terminal_wealth: The achieved terminal wealth in each
    # iteration

    ## Initialization

    # Initialize arrays to store the terminal wealth of each
    # iteration
    terminal_wealth = np.zeros([M])

    # Number of time steps
    T1 = int(np.floor(T/dt))

    # Initial mean and variance function
    mean_funct = lambda x: -np.dot(inv(phi[0]),phi[1]*np.sqrt(2/d))
                           *(x-w)
    var_funct =  lambda x: (lam/2)*np.exp(2/d*np.dot(phi[1],phi[1])
                           *(T-x))*inv(np.dot(phi[0].T,phi[0]))

    ## Iterations
    for k in range(1, M+1):

        # Initialization
        x = x0
        D = np.zeros([2,T1])
        D[1,0] = np.array(x)
        D[0,:] = np.linspace(0,T,T1)

        # Episodes
        for i in range(1, T1):

            # Mean and Variance
            variance = var_funct((i-1)*dt)
            mean = mean_funct(x)

            # Strategy sample
            u = np.random.multivariate_normal(mean,variance)

            # Compute the wealth at the next time point
            x = x+float(np.dot(u,Returns[i-1,:]))
```

```
63
64              # Update the set of samples
65              D[1,i] = np.array(x)
66
67              # Compute the Derivative of the cost function with
68              # respect to theta_1
69              Cost_Derivative_theta = Deriv_Cost_Funct_Theta_1(theta,
70                                      phi,dt,w,T,D[:,:(i+1)],lam,d)
71
72              # Update theta
73              theta[1] = theta[1] - eta_theta*Cost_Derivative_theta
74              theta[3] = 2*np.dot(phi[1],phi[1])/d
75              theta[2] = -lam/2*np.dot(phi[1],phi[1])
76              theta[0] = -theta[2]*T**2-theta[1]*T-(w-z)**2
77
78              # Compute the Derivative of the cost function with
79              # respect to phi
80              Cost_Derivative_phi = Deriv_Cost_Funct_Phi(theta,phi,
81                                    dt,w,T,D[:,:(i+1)],lam,d)
82
83              # Update phi
84              phi[0] = phi[0] - eta_phi*Cost_Derivative_phi[0]
85              phi[1] = phi[1] - eta_phi*Cost_Derivative_phi[1]
86
87          # Store terminal wealth
88          terminal_wealth[k-1] = x
89
90          # Update Lagrange multiplier
91          if k % N == 0:
92              w = w-alpha*(1/N*(np.sum(terminal_wealth[(k-N):k]))-z)
93
94          # Update control process
95          mean_funct = lambda x: -np.dot(inv(phi[0]),phi[1]
96                                 *np.sqrt(2/d))*(x-w)
97          var_funct =  lambda x: (lam/2)*np.exp(2/d*np.dot(phi[1],
98                                 phi[1])*(T-x))*inv(np.dot(phi[0].T,
99                                 phi[0]))
100
101         # Print current state
102         if k % 100 == 0:
103             print(str(round(100*k/M,2)) + " %" +
104                   "  |  Average Terminal Wealth: " +
105                   str(round(np.mean(terminal_wealth[k-100:k]),2)) +
106                   "  |  Average Variance: " +
107                   str(round(np.var(terminal_wealth[k-100:k]),4)))
108
109     return [[mean_funct,var_funct,terminal_wealth]]
```

Listing 7.3: The d-EMV algorithm

Chapter 8

Empirical Analysis

In this chapter, we analyze and compare the performance of the multi-dimensional EMV algorithm on real data. To do so, we use two different rebalancing frequencies and investment horizons. Furthermore, we modify the EMV algorithm with an additional leverage constraint. For all implementations, figures and results in this chapter, we refer to https://github.com/TimGyger/Master-Thesis.

We face a problem in an empirical study: we cannot generate unlimited data to train the EMV algorithm. Different reasons cause this. The nonstationarity of financial data implies that only the relatively recent data periods should be used for training an RL algorithm. Moreover, in some situations, especially for long-term, low-frequency rebalancing strategies, only a few thousand data points are available for algorithm training.

We artificially generate randomness during the training process to handle the limited data availability by selecting d stocks by chance from a pool of stocks for each training episode. Although this method differs from the classical batch RL training method, it has been shown in H. Wang [24] to be robust and to accommodate a volatile market.

The used stock pool comprises the current S&P 500 companies (see https://github.com/TimGyger/Master-Thesis for the actual list of companies and the data used). Further, all data is downloaded from *Yahoo Finance* using the Python-library *yfinance* (we refer to https://pypi.org/project/yfinance/ for more information about this library).

We compare the EMV algorithm against two baseline approaches. The first one is the classical Markowitz method (Markowitz, [19]). It uses a rolling window for estimating the mean return vector and the variance-covariance matrix to compute the allocation strategy. At the same time, the most recent data point is added to the rolling window at each time step, and the most obsolete data point is deleted from the window. For the implementation, we used the library *pymarkowitz* (we refer to https://pypi.org/project/pymarkowitz/ for more infor-

mation about this library). The second method is the equally weighted strategy (DeMiguel, Garlappi & Uppal, [9]), where we uniformly diversify by giving each stock a weight of $\frac{1}{d}$. This method does not need a training period.

Furthermore, we introduce a constraint variant of the EMV algorithm by only allowing a certain leverage level. For this purpose, we modify each allocation vector $\vec{u}_t \in \mathbb{R}^d$ generated by the EMV algorithm to $\frac{\vec{u}_t}{||\vec{u}_t||_1} L x_t^{\boldsymbol{u}}$, where $||\cdot||_1$ denotes the l_1-norm and L is the leverage level. For example, $L = 200\%$ means that at each time $t \in [0, T]$ the invested value can not exceed $2 \cdot x_t^{\boldsymbol{u}}$.

For the first experiments, we consider a time horizon of 10 years and monthly rebalancing, that is $T = 10$ and $\Delta t = \frac{1}{12}$. Since the transactions are low-frequent, the transaction costs are not included in the first part of this analysis. In our empirical study, we take 100 sets of S&P 500 stocks to test, each set containing $d = 20$ randomly selected stocks. We choose the initial wealth to be normalized as $x_0 = 1$ and set the 10-year target to be $z = 8$ (corresponding to a 23% annualized return or a 1.74% monthly return) except for the equally weighted strategy.

We train the EMV algorithm on a 10-year window starting from July 1992 - June 2002 and testing for the ensuing ten years (July 2002 - June 2012). We continue this procedure for the next ten 10-year horizons until June 2022. Otherwise, the Markowitz strategy is trained on the rolling, immediate prior 10-year monthly data.

In Figure 8.1, we present the corresponding averaged wealth processes and the 95% confidence interval of all strategies for the first testing period (July 2002 - July 2012).

We observe that in this experiment, the EMV algorithm approximately satisfies the investment target of $z = 8$. The leverage constraint variants fail to reach that target but still generate good profits over the 10-year horizon and outperform the baseline approaches in the sense of average return. However, if we analyze the used capital for the EMV strategy, we can see why we introduced the leverage constraint variants in the first place. To illustrate this, we present in Figure 8.2 the mean and the 95% confidence interval of the invested value as percentage of the current portfolio wealth $\frac{||\vec{u}_t||_1}{x_t^{\boldsymbol{u}}}$ over time.

We observe that the used leverage, especially at the beginning of the investment period, is not realistic ($> 1'000\%$). Therefore, introducing the leverage constraint variants is necessary. This is undoubtedly a downside to the practical use of the EMV algorithm.

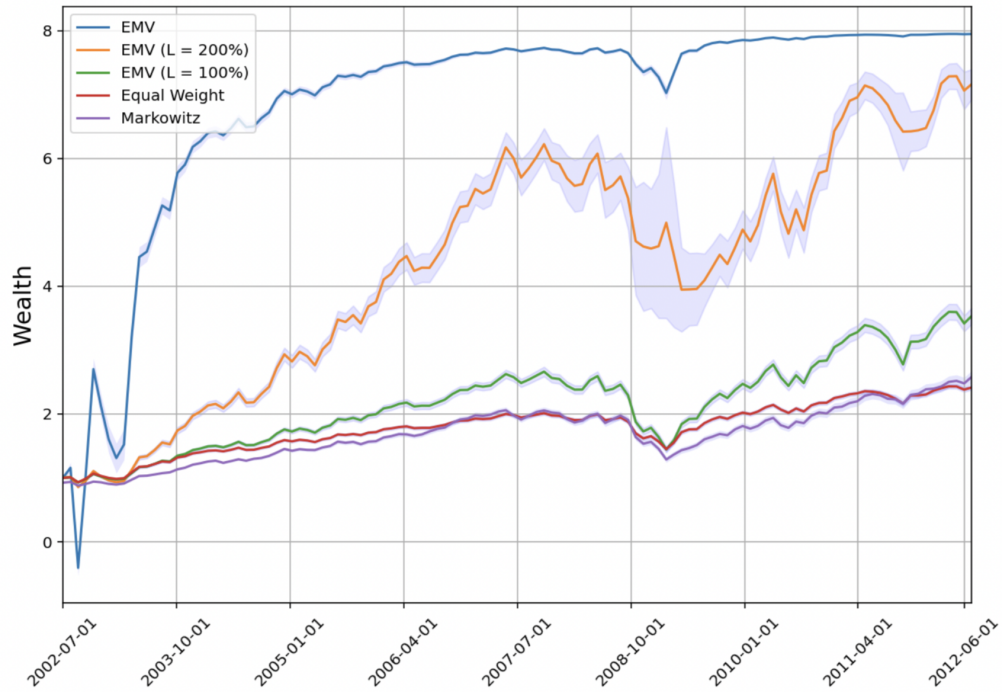

Figure 8.1: Investment performance comparison of the EMV method, the leverage constraint variants with $L = 200\%$, $L = 100\%$, the equally weighted method and the Markowitz method with monthly rebalancing over the 1 year horizon: July 2002 - July 2012.

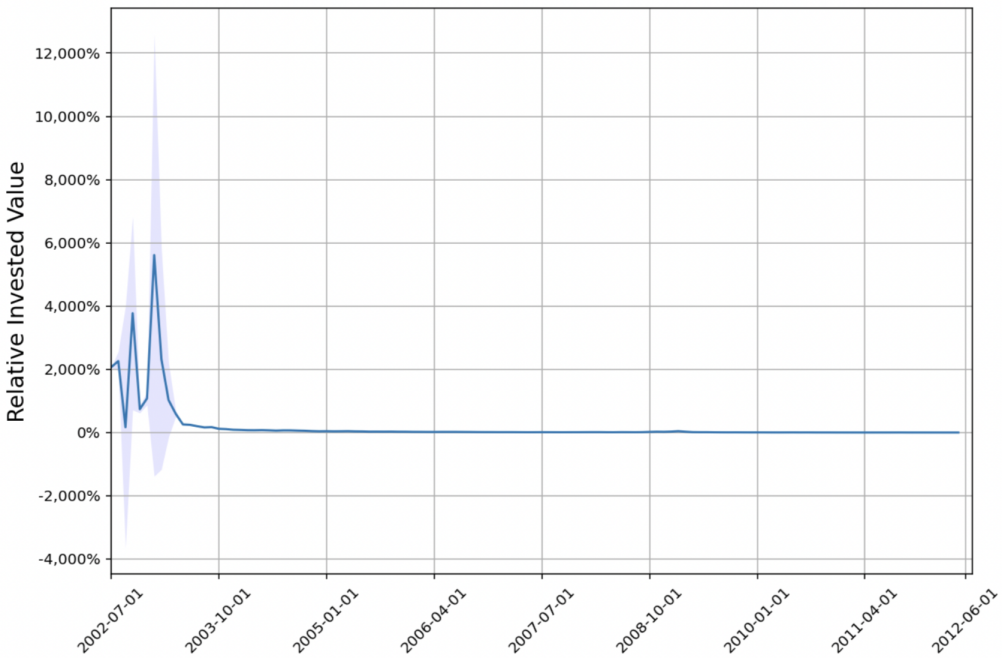

Figure 8.2: Invested value relative to current portfolio wealth of the EMV method with monthly rebalancing over the 1 year horizon: July 2002 - July 2012.

In Table 8.1, we report each method's annualized mean returns, standard deviations, and the corresponding Sharpe ratios separately for all eleven experiments (from top to bottom for each testing period).

Testing period	EMV	EMV (200%)	EMV (100%)	Equally	Markowitz
1. July 2002 - 1. July 2012	23.03% 0.66% 34.95	21.69% 38.37% 0.57	13.46% 22.27% 0.60	9.22% 5.79% 1.59	9.95% 17.14% 0.58
1. July 2003 - 1. July 2013	23.20% 0.54% 42.68	22.66% 30.02% 0.75	14.03% 21.82% 0.64	9.38% 5.22% 1.80	11.48% 18.31% 0.63
1. July 2004 - 1. July 2014	23.49% 0.47% 50.23	23.06% 27.41% 0.84	14.39% 19.80% 0.73	9.53% 4.71% 2.02	10.85% 15.68% 0.69
1. July 2005 - 1. July 2015	23.26% 0.67% 34.54	21.81% 38.34% 0.57	12.83% 15.36% 0.84	8.92% 4.11% 2.17	10.14% 17.26% 0.59
1. July 2006 - 1. July 2016	22.78% 0.57% 40.15	21.41% 34.46% 0.62	12.60% 13.62% 0.93	8.86% 3.63% 2.44	11.06% 17.99% 0.62
1. July 2007 - 1. July 2017	22.80% 0.58% 39.20	21.88% 29.02% 0.75	12.94% 13.34% 0.97	8.98% 3.67% 2.45	9.92% 16.22% 0.61
1. July 2008 - 1. July 2018	22.95% 0.52% 43.90	22.87% 32.59% 0.70	14.72% 18.62% 0.79	9.63% 4.61% 2.09	11.49% 18.74% 0.61
1. July 2009 - 1. July 2019	23.25% 0.39% 59.94	22.86% 12.35% 1.85	17.41% 23.18% 0.75	10.44% 4.38% 2.38	15.97% 26.60% 0.60
1. July 2010 - 1. July 2020	21.88% 0.53% 41.55	20.63% 45.20% 0.46	14.81% 21.11% 0.70	9.55% 4.93% 1.94	14.56% 23.91% 0.61
1. July 2011 - 1. July 2021	23.90% 0.30% 80.19	23.75% 23.43% 1.01	17.11% 18.61% 0.92	10.43% 4.55% 2.29	14.22% 23.02% 0.62
1. July 2012 - 1. July 2022	23.12% 0.42% 55.01	22.58% 86.79% 0.26	15.44% 25.47% 0.61	9.76% 4.17% 2.34	13.19% 20.28% 0.65

Table 8.1: The annualized mean return, standard deviation and the corresponding Sharpe ratio from top to bottom for each method respectively and for each 10-year testing period.

We note that not only the EMV algorithm but also the constraint variants achieve consistently good results. Comparing the methods, we observe that the Markowitz method is outperformed in almost every experiment. The equally weighted method achieves a very high Sharpe ratio but gains the lowest mean return in each experiment. H. Wang & X. Y. Zhou in [25] also compared the EMV algorithm to another RL approach, the Deep Deterministic Policy Gradient (DDPG) method (for detailed information, see T. Lillicrap et al. [18]). However, they observed that the DDPG method performs poorly. This was justified by the lack of data that is used to estimate the value functions by neural networks.

Next, we analyze the EMV algorithm with different leverage constraints on a 1-year horizon with daily rebalancing, that is $T = 1$ and $\Delta t = \frac{1}{252}$. For this purpose, we train the EMV algorithm on a 1-year window starting with period 28. June 2015 - 28. June 2016 and testing on the ensuing year (28. June 2016 - 28. June 2017). We continue this procedure for the next five 1-year horizons until June 2022. Again, we select randomly $d = 20$ stocks for each episode.

We choose the initial wealth to be normalized $x_0 = 1$ and set the investment target to be $z = 1.4$ (corresponding to a 40% annualized return). Since the trading frequency is relatively high, we include trading costs. For this purpose, we just subtract at each time t_i for $i \in \{1, ..., \lfloor \frac{T}{\Delta t} \rfloor\}$ the value $\frac{p}{100} \cdot ||\vec{u}_{t_i} - \vec{u}_{t_{i-1}} \circ (\vec{1} + \vec{R}_{t_{i-1}})||_1$ from the current wealth $x^{\boldsymbol{u}}_{t_i}$, where $\vec{R}_{t_{i-1}} \in \mathbb{R}^d$ is the d-dimensional vector of the returns of the d stocks at time t_{i-1}, $\vec{1}$ is the d-dimensional all-ones vector and the operation $\circ$ is the elementwise multiplication. This corresponds to a transaction cost of $p\%$. Since the level of transaction costs is nowadays a competitive subject, we choose a reasonable value of $p = 0.5$. Again, we take 100 sets of S&P 500 stocks to test on, each set containing $d = 20$ randomly selected stocks.

In Figure 8.3, we present the averaged wealth process and the 95% confidence interval of the constraint EMV algorithm with $L = 200\%$, $L = 150\%$, and $L = 100\%$ and the equally weighted method over the first testing phase (28. June 2016 - 28. June 2017).

We observe that the constraint versions with $L = 200\%$ and $L = 150\%$ perform well and approximately achieve the investment target $z = 1.4$ despite the transaction costs. The variant with $L = 100\%$ still generates a good profit but does not significantly outperform the equally weighted method in terms of mean return.

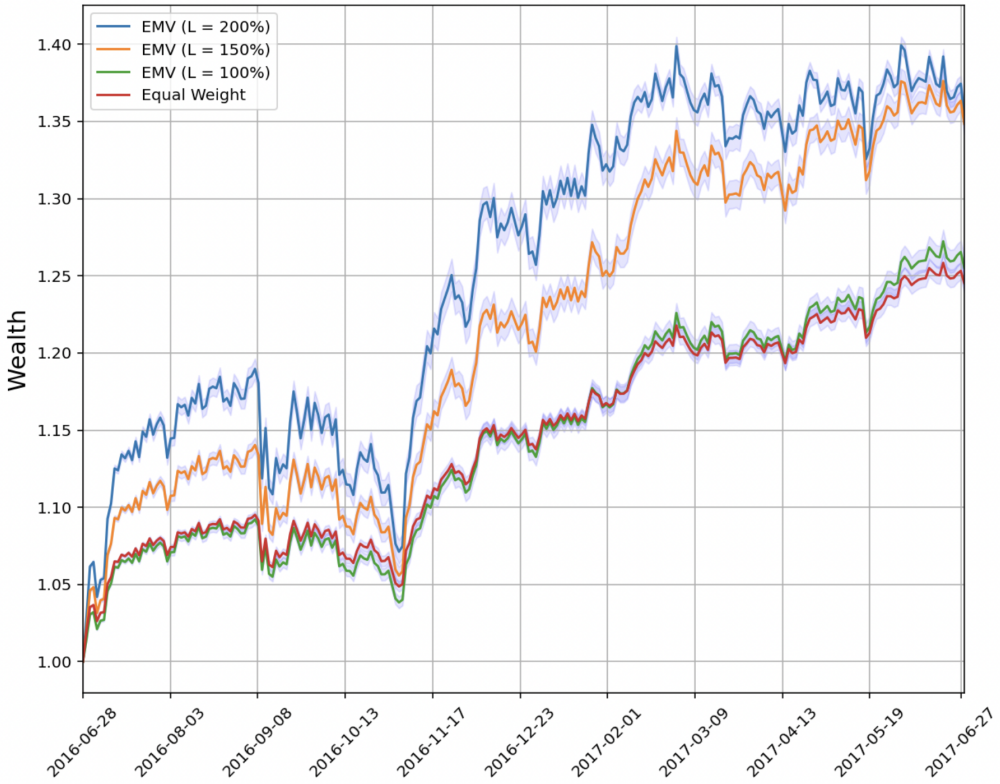

Figure 8.3: Investment performance comparison of the EMV method with $L = 200\%$, $L = 150\%$ and $L = 100\%$ with daily rebalancing and a transaction cost of 0.5% for the 1 year horizon: 28. June 2016 - 28. June 2017.

In Table 8.2, we report the annualized mean returns, standard deviations, and Sharpe ratios of the constraint variants of the EMV method and the equally weighted method separately for all six experiments (from top to bottom for each testing period).

Testing period	EMV (200%)	EMV (150%)	EMV (100%)	Equally
28. June 2016 - 28. June 2017	35.38% 4.40% 8.05	34.84% 5.35% 6.51	25.51% 5.79% 4.41	24.51% 4.47% 5.49
28. June 2017 - 28. June 2018	26.36% 10.99% 2.40	21.84% 10.28% 2.13	14.85% 7.07% 2.10	16.12% 5.48% 2.94
28. June 2018 - 28. June 2019	17.81% 11.68% 1.52	14.62% 8.97% 1.63	10.28% 5.77% 1.78	12.79% 4.85% 2.64

Continued on next page

Table 8.2 – *Continued from previous page*

Testing period	EMV (200%)	EMV (150%)	EMV (100%)	Equally
1. July 2019 - 1. July 2020	23.95% 111.96% 0.21	18.2% 86.72% 0.21	−4.32% 42.97% −0.10	8.14% 5.63% 1.44
1. July 2020 - 1. July 2021	39.53% 5.07% 7.8	41.33% 3.76% 10.98	42.60% 3.44% 12.39	45.10% 6.26% 7.20
1. July 2021 - 1. July 2022	−20.77% 9.13% -2.27	−14.33% 7.34% -1.95	−8.75% 5.17% -1.69	−5.14% 5.46% -0.94

Table 8.2: The annualized mean return, standard deviation and the corresponding Sharpe ratio from top to bottom for three constraint variants of the EMV method and the equally weighted method respectively and for multiple testing phases of 1 year and with a transaction cost of 0.5%.

We observe that the performance of all three leverage constraint EMV methods fluctuates. We even have one experiment where all mean returns are negative. We explain this by significant changes in the market environment between the training and testing phases. For example, we can see this by the extreme profit differences of the equally weighted method between 2020 - 2021 and 2021 - 2022. More precisely, during the training period from 1. July 2020 to 1. July 2021, the trend of the market was highly positive, indicated by a mean return of 45.10% of the equally weighted method, but during the corresponding testing phase, the trend of the market was rather negative, indicated by a mean return of −5.14% of the equally weighted method.

By comparing the monthly and daily rebalancing results with an investment horizon of 10 years and 1 year, we observe that the results are more robust over the longer horizon. We explain this by saying that the S&P 500 stocks perform more stable over extended periods since single market drops are outweighed over enough time. Therefore, the market behaves similarly in the training and testing period.

Chapter 9

Summary

This master thesis studied an exploratory version of the mean-variance (MV) problem formulated as an entropy-regularized, relaxed stochastic control problem. We derived the exploratory MV (EMV) algorithm by H. Wang & X. Y. Zhou in [25] to solve this problem. Our goal was to implement the EMV algorithm, analyze its learning performance in a simulation study, and compare its out-of-sample performance to other methods in an empirical study.

In the first part of the thesis, we presented the MV problem of finding an investment strategy that minimizes the variance of the final payoff while targeting a prespecified mean return. We solved this problem under some crucial assumptions.

In the next part, we motivated and stated an exploratory formulation of the EMV problem. Further, we derived the optimal feedback control law to solve this problem.

Consequently, we could compare the MV and the EMV problem and their solutions. It turned out that the optimal terminal wealth's under the respective optimal feedback controls of the two problems have the same mean. Further, we proved the solvability equivalence between the problems and showed that the EMV converges to the MV problem as the exploration weight decreases to 0. Moreover, we saw that the cost of exploration is linearly dependent with respect to the exploration weight λ and to the time horizon T.

Next, we stated a policy improvement theorem and a convergence result on whose base we derived the EMV algorithm. We presented the pseudo-code and the actual implementation in Python 3.8.8.

Then, we analyzed the learning performance for simulated stock prices in a stationary and a non-stationary market case. The results were convincing since, in most market scenarios, a strategy was learned that approximately achieved the

investment target with a relatively low standard deviation of the final payoffs. However, we observed that the EMV method has more difficulty learning successfully in market scenarios with nearly no drift or high volatility. Besides, we investigated the approximation accuracy of the EMV approach by comparing the learned ρ^2 with the ground truth value. The results showed that the algorithm has trouble approximating the parameter ρ^2.

To examine the efficiency of the EMV algorithm on real data with multiple assets, we derived the multi-dimensional EMV algorithm based on similar results as in the one-dimensional case. Again, we presented the pseudo-code and the actual implementation in Python.

Finally, in the last part of this thesis, we analyzed the out-of-sample performance of the EMV algorithm on S&P stock data over different time horizons and rebalancing frequencies. We compared the annualized mean return and the standard deviation of the final payoffs with the equally weighted method and the Markowitz approach. We observed that the EMV method performs exceptionally well in all experiments over a 10-years horizon with monthly rebalancing. However, the leverage used to achieve these results was unrealistically high. Therefore, we introduced a leverage constraint variant of the EMV method. Compared to the benchmark approaches, the constraint versions with a leverage level of 200% and 100% generated a high annualized return for all experiments.

Further, we analyzed the EMV algorithm over a one-year horizon with daily rebalancing. For the sake of reality, we included trading costs in these experiments. The results were not as good as over the 10-year horizon since there was a too big difference between the market trend during the training and testing phase. We concluded that the EMV algorithm is more successful if the fundamental behavior of the market is similar in the training and testing period. Therefore, it works better over more extended investment periods. However, the EMV algorithm and its leverage constraint variants are certainly an improvement compared to non-RL algorithms that estimate the market parameters.

Appendices

Appendix A

Probability Theory

A.1 Probability Space

A filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ with the filtration $\mathbb{F} := (\mathcal{F}_t)_{t\in[0,T]}$ satisfies the usual conditions if:

1. The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is complete, that is:
   For all $B \in \mathcal{F}$ with $P(B) = 0$ and all $A \subset B$ one has $A \in \mathcal{F}$.
2. For all $t \in [0, T]$ the σ-algebra $\mathcal{F}_t$ contains all the sets in $\mathcal{F}$ of zero probability.
3. The filtration $\mathbb{F}$ is right-continuous, that is:
   $\forall t \in [0, T] : \mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$.

A.2 Strong Law of Large Numbers

Theorem A.1. *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, each having finite mean μ and $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then*

$$\mathbb{P}\left(\lim_{n\to\infty} M_n = \mu\right) = 1.$$

That is, the averages converge with probability 1 to the common mean μ or $M_n \overset{a.s.}{\to} \mu$.

For the proof, we refer to J. S. Rosenthal [21].

A.3 Dirac Measure

Definition A.2. A Dirac measure on the set of events $\mathcal{F}$ is a probability measure δ_x defined for a given $x \in \Omega$ and any measurable set $A \subset \mathcal{F}$ by:

$$\delta_x(A) = \begin{cases} 0, & x \notin A \\ 1, & x \in A \end{cases}$$

It represents the almost sure outcome x in the sample space Ω.

A.4 Differential Entropy

Lemma A.3. *For an one-dimensional Gaussian density:*

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad x \in \mathbb{R},$$

with μ and $\sigma \in \mathbb{R}$, the differential entropy is given by:

$$\mathcal{H}(\mathcal{N}) = \ln\Big(\sigma\sqrt{2\pi e}\Big) = \ln\Big(\sigma\sqrt{2\pi}\Big) + \frac{1}{2}.$$

Proof. From the definition of differential entropy:

$$\mathcal{H}(\mathcal{N}) = -\int_{\mathbb{R}} \pi(x) \ln \pi(x) dx$$

we can compute:

$$\begin{aligned}
\mathcal{H}(\mathcal{N}) &= -\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ln\left(\sigma\sqrt{2\pi} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)\right) dx \\
&= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-t^2\right) \ln\left(\sigma\sqrt{2\pi} \exp\left(t^2\right)\right) dt \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left(\ln\Big(\sigma\sqrt{2\pi}\Big) + \ln\left(\exp\left(t^2\right)\right)\right) \exp\left(-t^2\right) dt \\
&= \frac{\ln\big(\sigma\sqrt{2\pi}\big)}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-t^2\right) dt + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^2 \exp\left(-t^2\right) dt \\
&= \frac{\sqrt{\pi}\ln\big(\sigma\sqrt{2\pi}\big)}{\sqrt{\pi}} + \frac{1}{\sqrt{\pi}} \left(\left[-\frac{t}{2}\exp\left(-t^2\right)\right]_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty} \exp\left(-t^2\right) dt\right) \\
&= \ln\Big(\sigma\sqrt{2\pi}\Big) + \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-t^2\right) dt \\
&= \ln\Big(\sigma\sqrt{2\pi}\Big) + \frac{\sqrt{\pi}}{2\sqrt{\pi}} = \ln\Big(\sigma\sqrt{2\pi}\Big) + \frac{1}{2},
\end{aligned}$$

where we substituted $t = \frac{x-\mu}{\sqrt{2}\sigma}$ and used integration by parts. □

Lemma A.4. *For a multivariate Gaussian density:*

$$\mathcal{N}(\vec{x}|\vec{\mu},\Sigma) = \frac{\exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^\top\Sigma^{-1}(\vec{x}-\vec{\mu})\right)}{(2\pi)^{N/2}|\Sigma|^{1/2}}, \quad \vec{x}\in\mathbb{R}^n,$$

with mean vector $\vec{\mu}\in\mathbb{R}^n$ and covariance matrix $\Sigma\in\mathbb{R}^{n\times n}$, the differential entropy is given by:

$$\mathcal{H}(\mathcal{N}) = \frac{1}{2}\ln((2\pi e)^n|\Sigma|) = \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}n,$$

where $|\Sigma| := \det(\Sigma)$.

Proof. The definition of differential entropy in the multi-dimensional case is given by:

$$\mathcal{H}(\mathcal{N}) = -\int_{\mathbb{R}^n}\pi(\vec{x})\ln\pi(\vec{x})d\vec{x} = -\mathbb{E}[\ln\pi(\vec{x})].$$

With the probability density function of the multivariate normal distribution, the differential entropy of $\mathcal{N}$ is:

$$\begin{aligned}\mathcal{H}(\mathcal{N}) &= -\mathrm{E}\left[\ln\left(\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\cdot\exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^\top\Sigma^{-1}(\vec{x}-\vec{\mu})\right]\right)\right]\\ &= -\mathrm{E}\left[-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\vec{x}-\vec{\mu})^\top\Sigma^{-1}(\vec{x}-\vec{\mu})\right]\\ &= \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\mathrm{E}\left[(\vec{x}-\vec{\mu})^\top\Sigma^{-1}(\vec{x}-\vec{\mu})\right].\end{aligned}$$

The last term can be evaluated as:

$$\begin{aligned}\mathrm{E}\left[(\vec{x}-\vec{\mu})^\top\Sigma^{-1}(\vec{x}-\vec{\mu})\right] &= \mathrm{E}\left[\mathrm{tr}\left((\vec{x}-\vec{\mu})^\top\Sigma^{-1}(\vec{x}-\vec{\mu})\right)\right]\\ &= \mathrm{E}\left[\mathrm{tr}\left(\Sigma^{-1}(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^\top\right)\right]\\ &= \mathrm{tr}\left(\Sigma^{-1}\mathrm{E}\left[(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^\top\right]\right)\\ &= \mathrm{tr}\left(\Sigma^{-1}\Sigma\right)\\ &= \mathrm{tr}\left(\boldsymbol{I}_n\right)\\ &= n,\end{aligned}$$

where $\boldsymbol{I}_n$ is the identity matrix of size n.

Consequently, the differential entropy is given by:

$$\mathcal{H}(\mathcal{N}) = \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}n.$$

□

Appendix B

Stochastic Calculus

B.1 Stochastic Processes

For the following definitions, we consider a filtered probability space $\big(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P}\big)$ with the filtration $\mathbb{F} := (\mathcal{F}_t)_{t\in[0,T]}$.

Definition B.1. Let (S, Σ) be a measurable space and $X = (X_t)_{t\in[0,T]}$ a stochastic process considered as a function $X : [0, T] \times \Omega \longrightarrow S$:

1. A function $f : \Omega \longrightarrow S$ is said to be $(\mathcal{F}, \Sigma)$-measurable if for every $E \in \Sigma$:
$$f^{-1}(E) := \{w \in \Omega : f(w) \in E\} \in \mathcal{F}.$$
2. The process X is said to be ($\mathbb{F}$-)adapted if the random variable $X_t : \Omega \longrightarrow S$ is a $(\mathcal{F}_t, \Sigma)$-measurable function for each $t \in [0, T]$.
3. The process X is said to be ($\mathbb{F}$-)predictable if X is a continuous-time stochastic process and measurable with respect to the σ-algebra generated by all left-continuous adapted processes (for example the Brownian motion in Appendix B.2).
4. The process X is said to be ($\mathbb{F}$-)progressivley measurable if for every time $t \in [0, T]$ the map $[0, t]\times\Omega \longrightarrow S$ defined by $(s, \omega) \mapsto X_s(\omega)$ is $\mathcal{B}([0, t])\otimes\mathcal{F}_t$-measurable.

Lemma B.2. *We can state the following relations between these definitions:*

1. *If X is progressively measurable then X is also adapted.*
2. *If X is adapted and every sample path is right-continuous or else every sample path is left-continuous, then X is also progressively measurable.*
3. *If X is predictable then X is also progressively measurable.*

Proof. The first statement follows directly from the definitions. For the second statement we refer to Proposition 1.13 in I. Karatzas & S. E. Shreve [15]. The third statement follows directly from the second one. □

B.2 Brownian Motion

We define the Brownian motion as in I. Karatzas & S. E. Shreve [15], Definition 5.1:

Definition B.3. A standard, m-dimensional Brownian motion defined on some filtered probability space $\left(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P}\right)$ is an $\mathbb{R}^m$-valued, adapted stochastic process $B = (B_t)_{t\geq 0}$, with continuous paths and the properties that $B_0 = 0$ $\mathbb{P}$-almost surely ($\mathbb{P}$-a.s.) and for any $0 \leq s < t < \infty$, the increment $B_t - B_s$ is independent of $\mathcal{F}_s$ and distributed under $\mathbb{P}$ as an m-dimensional Gaussian random variable with mean zero and covariance matrix $(t-s)\boldsymbol{I}_m$, where $\boldsymbol{I}_m$ is the identity matrix of size m.

B.3 Itô's formula

We introduce the Itô process and state a simple, one-dimensional formulation of Itô's formula:

Definition B.4. A real-valued Itô process $X = (X_t)_{t\geq 0}$ is an adapted process with $\mathbb{P}$-a.s. continuous trajectories, such that there exist μ and σ, two real-valued processes, which are measurable and adapted and satisfying with $\mathbb{P}$-probability 1 that:

$$\int_0^t \left(|\mu_s| + |\sigma_s|^2\right) ds < \infty, \quad t \geq 0$$

with

$$X_t = x_0 + \int_0^t \mu_s \, \mathrm{d}s + \int_0^t \sigma_s \cdot dB_s, \quad t \geq 0, \quad \mathbb{P}\text{-a.s.},$$

for some $x_0 \in \mathbb{R}$ and a Brownian motion B. We call μ the drift of X and σ its volatility.

Theorem B.5. *Let X be a real-valued Itô process with drift μ and volatility σ, and let $f : [0,\infty) \times \mathbb{R} \longrightarrow \mathbb{R}$ be a map which is once continuously differentiable with respect to its first variable, and twice continuously differentiable with respect to its second variable ($f \in C^{1,2}$). Then, the process $(f(t, X_t))_{t\geq 0}$ is itself an Itô process and one has:*

$$df(t, X_t) = \left(\frac{\partial f}{\partial t} + \mu_t \frac{\partial f}{\partial x} + \frac{\sigma_t^2}{2}\frac{\partial^2 f}{\partial x^2}\right) dt + \sigma_t \frac{\partial f}{\partial x}\, dB_t, \quad t \geq 0.$$

For the proof and the multi-dimensional formulation, we refer to Chapter 4 in B. Øksendal [20].

B.4 Dominated Convergence Theorem

Theorem B.6. *Let $(f_n)_{n\geq 1}$ be a sequence of integrable functions which converges almost everywhere to a real-valued measurable function f. If there exists an integrable function g such that $|f_n| \leq g$ for all n, then f is integrable and:*

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu,$$

where μ is the Lebesgue measure.

For the proof, we refer to Theorem 5.6 in Chapter 5 in R. G. Bartle [3].

B.5 Fubini's Theorem

Theorem B.7. *Let $(E, \mathcal{E}, \mu)$ and $(F, \mathcal{F}, \nu)$ be finite measure spaces and let $f : E \times F \to \mathbb{R}$ be a bounded $\mathcal{E} \otimes \mathcal{F}$-measurable function. Then,*

$$y \mapsto \int f(x, y) d\mu(x)$$

is $\mathcal{F}$-measurable,

$$x \mapsto \int f(x, y) d\nu(y)$$

is $\mathcal{E}$-measurable, and,

$$\int_F \int_E f(x, y) d\mu(x) d\nu(y) = \int_F \int_E f(x, y) d\nu(x) d\mu(y).$$

For the proof of this theorem, we refer to Section 5.2 in D. Cohn [8].

In this work, we often use Fubini's theorem for expressions of the form:

$$\mathbb{E}\Big[\int_0^T f(t, x_t) dt\Big],$$

where $x = (x_t)_{t\in[0,T]}$ is a stochastic process with continuous paths in time and f a function. In this case, we have the following two measure spaces $([0, T], \mu)$, where μ is the one-dimensional Lebesgue measure, and $(\Omega, \mathcal{F}, \mathbb{P})$ on which the Brownian motion is defined on. Then, if $\int_0^T \mathbb{E}[|f(t, x_t)|]dt < \infty$ we have:

$$\begin{aligned}\mathbb{E}\Big[\int_0^T f(t, x_t) dt\Big] &= \int_\Omega \int_0^T f(t, x_t(w)) dt d\mathbb{P}(w) \\ &= \int_0^T \int_\Omega f\big(t, x_t(w)\big) d\mathbb{P}(w) dt = \int_0^T \mathbb{E}\big[f(t, x_t)\big] dt.\end{aligned}$$

B.6 An Existence and Uniqueness Result

Theorem B.8. ***(Existence and uniqueness theorem for SDE's)****:*

Let $T > 0$, and let:

$$\mu : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$$
$$\sigma : \mathbb{R}^n \times [0,T] \to \mathbb{R}^{n\times m}$$

be measurable functions for which there exist constants C and D such that:

$$\textit{Linear growth:} \quad |\mu(x,t)| + |\sigma(x,t)| \leq C(1+|x|)$$
$$\textit{Lipschitz condition:} \quad |\mu(x,t) - \mu(y,t)| + |\sigma(x,t) - \sigma(y,t)| \leq D|x-y|$$

for all $t \in [0,T]$ and all x and $y \in \mathbb{R}^n$, where $|\sigma|^2 = \sum_{i,j=1}^n |\sigma_{ij}|^2$.

Let Z be a random variable that is independent of the σ-algebra generated by B_s, for $s \geq 0$, and with finite second moment:

$$\mathbb{E}\left[|Z|^2\right] < +\infty.$$

Then the stochastic differential equation:

$$dX_t = \mu\left(X_t, t\right) dt + \sigma\left(X_t, t\right) dB_t, \quad t \in [0,T]$$
$$X_0 = Z$$

has a $\mathbb{P}$-a.s. unique t-continuous solution $(t,\omega) \mapsto X_t(\omega)$ such that $X = (X)_{t\in[0,T]}$ is adapted and:

$$\mathbb{E}\left[\int_0^T |X_t|^2 \, dt\right] < +\infty.$$

For the proof of this theorem we refer to Section 5.2 in B. Øksendal [20].

Lemma B.9. *The stochastic differential equation:*

$$dX_t = -\rho^2(X_t - w)dt + \sqrt{\rho^2(X_t-w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}}dB_t, \quad t \in [0,T]$$
$$X_0 = x_0,$$

with ρ, w, $x_0 \in \mathbb{R}$ and $\lambda > 0$, has a $\mathbb{P}$-a.s. unique t-continuous solution $(t,\omega) \mapsto X_t(\omega)$ such that $X = (X_t)_{t\in[0,T]}$ is adapted and:

$$\mathbb{E}\left[\int_0^T |X_t|^2 \, dt\right] < +\infty.$$

Proof. We proof this statement by checking the necessary conditions in Theorem B.8. Since we have constant coefficients and continuous functions describing the drift and the volatility, the only unobvious properties are the linear growth and the Lipschitz condition. We observe for $t \in [0,T]$ and $x \in \mathbb{R}$ that:

$$\begin{aligned}\mu(x,t) &= -\rho^2(x-w)\\ \sigma(x,t) &= \sqrt{\rho^2(x-w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}}.\end{aligned}$$

Therefore, we can compute that for all $(t,x) \in [0,T] \times \mathbb{R}$:

$$|\mu(x,t)| = |-\rho^2(x-w)| = \rho^2|(x-w)| \leq \rho^2(|x|+|w|) \leq C_1(1+|x|),$$

where $C_1 = \max(\rho^2, \rho^2|w|)$, and that:

$$\begin{aligned}|\sigma(x,t)| &= \sqrt{\rho^2(x-w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}} \leq |\rho(x-w)| + \sqrt{\frac{\lambda}{2}}e^{\frac{1}{2}\rho^2(T-t)}\\ &\leq |\rho|(|x|+|w|) + \sqrt{\frac{\lambda}{2}}e^{\frac{1}{2}\rho^2 T} \leq C_2(1+|x|),\end{aligned}$$

where $C_2 = \max(|\rho|, |\rho w| + \sqrt{\frac{\lambda}{2}}e^{\frac{1}{2}\rho^2 T})$.

Hence, we can conclude that the linear growth condition is satisfied:

$$|\mu(x,t)| + |\sigma(x,t)| \leq C_1(1+|x|) + C_2(1+|x|) \leq C(1+|x|),$$

for a constant $C = \max(C_1, C_2)$ and for all $(t,x) \in [0,T] \times \mathbb{R}$.

Further, for the Lipschitz condition we find upper bounds for the derivatives:

$$|\mu_x(x,t)| = |-\rho^2| = \rho^2$$

and

$$|\sigma_x(x,t)| = \left|\frac{\rho^2(x-w)}{\sqrt{\rho^2(x-w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}}}\right| \leq \left|\frac{\rho^2(x-w)}{\sqrt{\rho^2(x-w)^2}}\right| = |\rho|,$$

for all $(t,x) \in [0,T] \times \mathbb{R}$.

By using the mean value theorem, we obtain for arbitrary x, $y \in \mathbb{R}$ and $t \in [0,T]$ that:

$$\begin{aligned}|\mu(x,t) - \mu(y,t)| &\leq \rho^2|x-y|\\ |\sigma(x,t) - \sigma(y,t)| &\leq |\rho|(|x-y|)\end{aligned}$$

and therefore:

$$|\mu(x,t) - \mu(y,t)| + |\sigma(x,t) - \sigma(y,t)| \leq \rho^2|x-y| + |\rho|(|x-y|) \leq D|x-y|,$$

where $D = \max(\rho^2, |\rho|)$ is a constant. □

B.7 Linear Stochastic Differential Equations

Definition B.10. The general form of a scalar linear stochastic differential equation is:

$$dX_t = (a_1(t)X_t + a_2(t))\,dt + (b_1(t)X_t + b_2(t))\,dW_t, \quad t \in [0,T], \tag{B.1}$$

where $W = (W_t)_{t\in[0,T]}$ is a standard Brownian motion and the coefficients a_1, a_2, b_1 and b_2 are specified functions of time t or constants. Provided they are Lebesgue measurable and bounded on an interval $[0,T]$, the existence and uniqueness theorem (Theorem B.8) applies, ensuring the existence of a strong solution X_t on $t_0 \le t \le T$ for each $0 \le t_0 < T$ and each measurable initial value X_{t_0}.

Theorem B.11. *The solution of the scalar linear stochastic differential equation (B.1) is:*

$$X_t = \Phi_{t,t_0}\left(X_{t_0} + \int_{t_0}^{t} \Phi_{s,t_0}^{-1}(a_2(s) - b_1(s)b_2(s))ds + \int_{t_0}^{t} \Phi_{s,t_0}^{-1} b_2(s) dW_s\right),$$

where

$$\Phi_{t,t_0} = \exp\left(\int_{t_0}^{t}\left(a_1(s) - \frac{b_1^2(s)}{2}\right)ds + \int_{t_0}^{t} b_1(s)dW_s\right).$$

For the proof of this statement and more detailed information about linear stochastic differential equations, we refer to Section 4.2 in P. E. Kloeden & E. Platen [16].

In our case, we have the following linear SDE with constant coefficients:

$$\begin{aligned} dX_t &= -\rho^2(X_t - w)dt - \rho(X_t - w)dB_t, \quad t \in [s,T] \\ X_s &= y. \end{aligned}$$

If we compare this to (B.1), we have for $t \in [s,T]$:

$$\begin{aligned} a_1(t) &= -\rho^2, \qquad a_2(t) = \rho^2 w, \\ b_1(t) &= -\rho, \qquad b_2(t) = \rho w. \end{aligned}$$

Therefore, we obtain:

$$\begin{aligned} \Phi_{t,s} &= \exp\left(\int_s^t \left(-\frac{3\rho^2}{2}\right)dv + \int_s^t -p dB_v\right) \\ &= \exp\left(-\frac{3\rho^2}{2}(t-s) - p(B_t - B_s)\right). \end{aligned}$$

And we can conclude that:

$$X_t = \Phi_{t,s}\left(X_s + \int_s^t 2\rho^2 w \cdot \Phi_{v,s}^{-1} dv + \int_s^t \rho w \cdot \Phi_{v,s}^{-1} dB_v\right).$$

We can see that $(X_t)_{t\in[s,T]}$ is adapted and has continuous paths in time.

Appendix C

Optimal Control Theory

C.1 Bellman's optimality principle and Dynamic Programming

Principle of Optimality: 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. (See Chapter III. 3. in R. Bellman [4])

More intuitively, an optimal strategy today remains optimal tomorrow.

For a more formal description of Bellman's optimality principle, we refer to Section 11.4 in T. Björk et al. [7].

We note that in decision theory, a problem is called time-consistent if Bellman's principle of optimality holds and time-inconsistent if it does not hold.

Methods of dynamic programming are based on this optimality principle. Dynamic programming refers to simplifying a decision by breaking it down into a sequence of decision steps over time. This is done by defining the optimal value functions for any initial state $(s, y) \in [0, T) \times \mathbb{R}$, representing sub-problems. Then one can derive a recursive relation for the optimal value function. In the discrete case, this results in the Bellman equation which can often be solved analytically or numerically. In the continuous case, we end up with a PDE for the optimal value function. In this work, the derivation of this PDE for the specific problems 2.6 and 3.8 is explained detailed in Sections 2.4.3 and 3.4.2, respectively.

For more information about dynamic programming, we refer to Chapters 2 & 11 in T. Björk et al. [7].

C.2 Tower Property

Theorem C.1. *Let X be an integrable random variable, $\mathbb{F} = (F_k)_{k\geq 1}$ a filtration and U a nonlinear function. Then for any $n \geq 1$ it holds that:*

$$\mathbb{E}\big[\mathbb{E}[U(X)|\mathcal{F}_m]\big|\mathcal{F}_n\big] = \mathbb{E}[U(X)|\mathcal{F}_n], \quad m \geq n$$

Proof. The proof follows by the definition of the conditional expectation. □

Appendix D

More Results

Return	Standard Deviation	Learned ρ^2	Err_θ
36.74%	131.09%	8.46	0.08
39.95%	59.88%	5.71	0.27
36.63%	134.22%	8.91	0.14
39.84%	39.75%	5.14	0.34
36.86%	126.93%	8.12	0.04
39.52%	84.61%	6.52	0.17
39.47%	88.91%	6.67	0.15
36.9%	123.36%	7.92	0.01
37.22%	113.6%	7.48	0.05
37.76%	106.46%	7.22	0.08
37.01%	135.85%	9.47	0.21
39.83%	67.15%	5.91	0.25
39.55%	80.46%	6.39	0.18
36.48%	134.74%	9.18	0.17
37.73%	106.61%	7.22	0.08
36.88%	123.45%	7.93	0.01
37.3%	113.03%	7.46	0.05
39.46%	78.25%	6.33	0.19
36.94%	123.08%	7.9	0.01
39.95%	59.49%	5.7	0.27
36.75%	132.05%	8.67	0.11
36.82%	119.78%	7.73	0.01
36.86%	122.3%	7.85	0
36.75%	128.73%	8.22	0.05
36.23%	135.39%	9.32	0.19
36.6%	130.63%	8.41	0.07
36.37%	135.63%	9.39	0.2
36.51%	129.67%	8.31	0.06
36.66%	131.81%	8.62	0.1
36.67%	132.49%	8.72	0.11

Continued on next page

Table D.1 – *Continued from previous page*

Return	Standard Deviation	Learned ρ^2	Err_θ
37.1%	114.41%	7.51	0.04
39.52%	75.84%	6.23	0.21
36.79%	120.55%	7.76	0.01
36.81%	128.29%	8.19	0.04
39.62%	32.9%	4.89	0.38
38.08%	102.41%	7.1	0.09
39.73%	37.24%	5.05	0.36
36.74%	132.9%	8.76	0.12
36.85%	127.02%	8.12	0.04
36.66%	135.83%	9.42	0.2
39.83%	67.4%	5.92	0.24
39.13%	92.94%	6.8	0.13
39.43%	89.55%	6.69	0.15
36.75%	131.92%	8.65	0.1
39.48%	83.67%	6.5	0.17
39.84%	47.89%	5.39	0.31
39.8%	38.46%	5.09	0.35
36.83%	127.5%	8.14	0.04
36.31%	135.55%	9.36	0.19
39.86%	56.39%	5.62	0.28
37.86%	**104.16%**	**7.44**	

Table D.1: Annualized mean, standard deviation, learned ρ^2 and relative error of 50 simulations with randomly chosen initial parameters θ, φ and w under the market scenario $\sigma = 10\%$ and $\mu = 30\%$ (what corresponds to $\rho^2 = 7.84$) and with $M = 20'000$. In the last row, we present the mean of the respective columns.

Return	Standard Deviation	Learned ρ^2	Err_θ
39.85%	51.47%	5.49	0.3
39.82%	30.17%	4.7	0.4
39.79%	61.94%	5.87	0.25
39.85%	24.75%	4.39	0.44
39.81%	48.32%	5.39	0.31
39.77%	36.64%	4.98	0.37
39.76%	37.96%	5.03	0.36
39.81%	45.69%	5.3	0.32

Continued on next page

Table D.2 – *Continued from previous page*

Return	Standard Deviation	Learned ρ^2	Err_θ
39.79%	42.52%	5.19	0.34
39.78%	41.07%	5.14	0.34
39.9%	83.2%	6.35	0.19
39.81%	31.92%	4.79	0.39
39.78%	35.33%	4.93	0.37
39.79%	69.57%	6.04	0.23
39.78%	41.09%	5.15	0.34
39.81%	45.73%	5.3	0.32
39.79%	42.45%	5.19	0.34
39.79%	34.66%	4.9	0.37
39.81%	45.53%	5.29	0.32
39.83%	30.07%	4.7	0.4
39.87%	53.36%	5.55	0.29
39.8%	44.46%	5.26	0.33
39.8%	45.14%	5.28	0.33
39.83%	49.65%	5.43	0.31
39.76%	74.21%	6.14	0.22
39.84%	50.93%	5.47	0.3
39.73%	77.2%	6.2	0.21
39.84%	50.26%	5.45	0.3
39.86%	52.44%	5.52	0.3
39.9%	54.49%	5.59	0.29
39.79%	42.63%	5.2	0.34
39.79%	33.96%	4.88	0.38
39.8%	44.76%	5.27	0.33
39.83%	49.4%	5.42	0.31
39.86%	22.52%	4.23	0.46
39.77%	40.25%	5.12	0.35
39.86%	23.95%	4.34	0.45
39.93%	56.14%	5.66	0.28
39.81%	48.4%	5.39	0.31
39.69%	79.74%	6.26	0.2
39.81%	31.99%	4.79	0.39
39.77%	39.0%	5.07	0.35
39.76%	38.16%	5.04	0.36
39.86%	52.93%	5.54	0.29
39.78%	36.41%	4.97	0.37
39.84%	27.16%	4.54	0.42
39.85%	24.32%	4.37	0.44
39.82%	48.79%	5.4	0.31
39.72%	75.92%	6.18	0.21

Continued on next page

Table D.2 – *Continued from previous page*

Return	Standard Deviation	Learned ρ^2	Err_θ
39.83%	29.3%	4.66	0.41
39.81%	**45.56%**	**5.25**	

Table D.2: Annualized mean, standard deviation, learned ρ^2 and relative error of 50 simulations with randomly chosen initial parameters θ, φ and w under the market scenario $\sigma = 10\%$ and $\mu = 30\%$ (what corresponds to $\rho^2 = 7.84$) and with $M = 40'000$. In the last row, we present the mean of the respective columns.

Bibliography

[1] H. Anton & C. Rorres. "Elementary Linear Algebra: Applications Version". 11th edition. *Wiley Interscience* (2014). ISBN: 9781118434413. URL: https://www.bibsonomy.org/bibtex/23d125adef6e8701a7dce93bdc817f3bc/ytyoun.

[2] Z. Ahmed, N. L. Roux, M. Norouzi & D. Schuurmans. "Understanding the impact of entropy on policy optimization". In: *International Conference on Machine Learning* (2018). DOI: https://doi.org/10.48550/arXiv.1811.11214.

[3] R. G. Bartle. "The Elements of Integration and Lebesgue Measure". *Wiley Interscience* (1995). ISBN: 9780471042228. DOI: https://doi.org/10.1002/9781118164471

[4] R. Bellman, "Dynamic Programming". In: *Princeton Landmarks in Mathematics and Physics Series*, Princeton: Princeton University Press (1957). ISBN: 9780691146683. URL: https://press.princeton.edu/books/paperback/9780691146683/dynamic-programming.

[5] D. P. Bertsekas. "Constrained Optimization and Lagrange Multiplier Methods". *New York: Academic Press* (1982). ISBN: 978-0-12-093480-5. DOI: https://doi.org/10.1016/C2013-0-10366-2.

[6] M. J. Best & R. R. Grauer. "On the sensitivity of mean–variance-efficient portfolios to changes in asset means: some analytical and computational results". In: *Review of Financial Studies*, Vol. 4, Iss. 2, pp. 315-342, (1991). DOI: https://doi.org/10.1093/rfs/4.2.315.

[7] T. Björk, M. Khapko & A. Murgoci. "Time-Inconsistent Control Theory with Finance Applications". In: *Springer Finance Series*, published by *Springer Cham* (2021). ISBN: 978-3-030-81842-5. ISSN: 1616-0533. DOI: https://doi.org/10.1007/978-3-030-81843-2.

[8] D. L. Cohn. "Measure Theory". 2nd edition. In: *Birkhäuser Advanced Texts Basler Lehrbücher*, published by *Birkhäuser New York, NY* (2013). ISBN: 978-1-4614-6955-1. ISSN: 1019-6242. DOI: https://doi.org/10.1007/978-1-4614-6956-8.

[9] V. DeMiguel, L. Garlappi & R. Uppal. "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?". *The Review of Financial Studies*, Vol. 22, Iss. 5, 1915–1953 (2007). DOI: https://doi.org/10.1093/rfs/hhm075

[10] K. Doya. "Reinforcement learning in continuous time and space". In: *Neural Computation*, Vol. 12, Iss. 1 published by *MIT Press* (2000). ISSN: 0899-7667. DOI: https://doi.org/10.1162/089976600300015961.

[11] W. H. Fleming & M. Nisio. "On stochastic relaxed control for partially observed diffusions". In: *Nagoya Mathematical Journal*, 93, 71-108, (1984). DOI: https://doi.org/10.1017/S0027763000020742.

[12] J.-P. Fouque, G. Papanicolaou, R. Sircar & K. Solna. "Multiscale stochastic volatility asymptotics". In: *Multiscale Modeling & Simulation*, Vol. 2, Iss. 1, 22-42, published by *The Society for industrial and applied Mathematics* (2003). DOI: https://doi.org/10.1137/030600291

[13] J. Le Gall. "Brownian Motion, Martingales, and Stochastic Calculus". In: *Graduate Texts in Mathematics*, published by *Springer Cham* (2016). ISBN: 978-3-319-31088-6. ISSN: 0072-5285. DOI: https://doi.org/10.1007/978-3-319-31089-3. Springer International Publishing Switzerland, 2016.

[14] I. Goodfellow, Y. Bengio & A. Courville. "Deep learning". *MIT Press* (2016). URL: http://www.deeplearningbook.org.

[15] I. Karatzas & S. E. Shreve. "Brownian motion and stochastic calculus". 2nd edition. In: *Graduate Texts in Mathematics, Volume 113*, published by *Springer New York, NY* (1998). ISBN: 978-0-387-97655-6. ISSN: 0072-5285. DOI: https://doi.org/10.1007/978-1-4612-0949-2.

[16] P. E. Kloeden & E. Platen, "Numerical Solution of Stochastic Differential Equations". In: *Stochastic Modelling and Applied Probability Series*, published by *Springer Berlin*, Heidelberg 6th edition (1995). ISBN: 978-3-642-08107-1. ISSN: 0172-4568. DOI: https://doi.org/10.1007/978-3-662-12616-5.

[17] X. Li, X. Y. Zhou & Andrew E. B. Lim. "Dynamic mean-variance portfolio selection with no-shorting constraints". In : *SIAM Journal on Control and Optimization*, Vol. 40, ISS. 5, 1540–1555, (2002). DOI: https://doi.org/10.1137/S0363012900378504.

[18] T. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y., Tassa & D. Wierstra. "Continuous control with deep reinforcement learning". In: *International Conference on Learning Representations, San Juan* (2016). DOI: https://doi.org/10.48550/arXiv.1509.02971

[19] H. Markowitz. "Portfolio selection". In: *The Journal of Finance*, Vol 7, Iss. 1, 77–91 (1952). DOI: https://doi.org/10.1111/j.1540-6261.1952.tb01525.x.

[20] B. Øksendal. "Stochastic Differential Equations. An Introduction with Applications". *Springer Berlin*, Heidelberg 6th edition (2003). ISBN: 978-3-540-04758-2. ISSN: 0172-5939. DOI: https://doi.org/10.1007/978-3-642-14394-6

[21] J. S. Rosenthal. "A First Look at Rigorous Probability Theory". 2nd edition. *World Scientific Publishing Co.* (2006). ISBN: 981-02-4322-7. DOI: https://doi.org/10.1142/6300.

[22] C. E. Shannon. "The mathematical theory of communication". In: *The Bell System Technical Journal* published by *Nokia Bell Labs* (1948). ISBN: 978-0-252-72548-7. DOI: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[23] R. S. Sutton & A. G. Barto. "Reinforcement learning: An introduction". 2nd edition. *Cambridge, MA: MIT Press*, (2018). ISBN: 978-0-262-03924-6. URL: https://mitpress.mit.edu/books/reinforcement-learning-second-edition.

[24] H. Wang. "Large scale continuous-time mean-variance portfolio allocation via reinforcement learning". Available at SSRN 3428125, (2019). DOI: https://doi.org/10.48550/arXiv.1907.11718.

[25] H. Wang & X. Y. Zhou. "Continuous-time mean–variance portfolio selection: A reinforcement learning framework". In: *Journal of Mathematical Finance* (2020). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3706168.

[26] H. Wang, T. Zariphopoulou & X. Y. Zhou. "Reinforcement learning in continuous time and space: A stochastic control approach". In: *Journal of Machine Learning Research* (2020). URL: https://jmlr.org/papers/volume21/19-144/19-144.pdf.

[27] X. Y. Zhou & D. Li. "Continuous-time mean-variance portfolio selection: A stochastic LQ framework". In: *Applied Mathematics and Optimization* 42, 19-33 (2000). DOI: https://doi.org/10.1007/s002450010003.

[28] X. Y. Zhou. "On the existence of optimal relaxed controls of stochastic partial differential equations". In: *SIAM Journal on Control and Optimization*, Vol. 30, ISS. 2, 247-261, (1992). DOI: https://doi.org/10.1137/0330016.