Exploratory Optimal Control: Applications to Reinforcement Learning

David Hyland

Supervised by: Dr. Zhou Zhou

An essay presented in partial fulfillment of
the requirements for the degree of
Bachelor of Science (Advanced Mathematics) (Honours)

Applied Mathematics
School of Mathematics & Statistics
University of Sydney
November 2020

Contents

Abstract

Exploration is a powerful tool in learning systems. In this thesis, we investigate how methods from stochastic optimal control and concepts in reinforcement learning can be combined to study exploration in stochastic control problems. Using this new problem formulation, we analyse well-known problems in the control theory literature, namely the Linear Quadratic and Mean-Variance problems. We derive a generalised policy improvement theorem and use this result to propose a new reinforcement learning algorithm for implementing Stochastic LQ control. We also investigate the performance of a reinforcement learning algorithm known as the EMV algorithm. A key theoretical result underpinning these two problems is that the optimal policy an agent should use is a Gaussian distribution, which forms the basis for implementing an effective, interpretable, and adaptive reinforcement learning algorithm.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Zhou Zhou for his invaluable insights, patience and encouragement throughout this endeavour. It has been an honour and pleasure to learn from him. I would also like to express my deepest gratitude toward my parents for making this opportunity possible for me and their unending love and guidance. Finally, I would like to thank my brother for his continued support and feedback throughout the year.

Notation

The following notational shorthand will be used throughout this essay, where the use of A denotes an arbitrary 1-dimensional function:

$A_t(x_1, x_2, ...) \coloneqq A(t, x_1, x_2, ...)$

$\dot{A} \coloneqq \dfrac{dA}{dt}$

$A_x \coloneqq \dfrac{\partial A}{\partial x}$

$u_t \coloneqq u(t, X_t)$ (Value of a control function u at time t and state X_t)

$A^u(t, x) \coloneqq A(t, x, u_t)$

$A^u_t \coloneqq A(t, X^u_t, u_t)$

$C^u(t, x) \coloneqq \sigma^u(t, x)\sigma^u(t, x)^\intercal$ (Volatility Matrix)

$\mathcal{A}^u \coloneqq \sum_{i=1}^{n} \mu^u_i(t, x)\dfrac{\partial}{\partial x_i} + \dfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} C^u_{ij}(t, x)\dfrac{\partial^2}{\partial x_i \partial x_j}$ (Infinitesimal Generator)

$\mathcal{S}^n \coloneqq \{\text{Symmetric } n \times n \text{ matrices}\}$

$C^{a,b}(A \times B; \mathbb{R}^n) \coloneqq \{f : A \times B \to \mathbb{R}^n \text{ which are } C^a \text{ on } A \text{ and } C^b \text{ on } B\}$

$\mathcal{L}^p(0, T; \mathbb{R}^n) \coloneqq \left\{\text{Lebesgue measurable functions } \varphi : [0, T] \to \mathbb{R}^n \text{ such that } \int_0^T |\varphi(t)|^p dt < \infty\right\}$

$\mathcal{L}^p_{\mathcal{F}}(\Omega; \mathbb{R}^n) \coloneqq \{\mathbb{R}^n\text{-valued } \mathcal{F}\text{-measurable random variables with finite absolute } p^{th}\text{-moment}\}$

$\mathcal{L}^p_{\mathcal{F}}([0, T]; \mathbb{R}^n) \coloneqq \left\{\mathbb{R}^n\text{-valued } \{\mathcal{F}_t\}_{t\geq 0}\text{-adapted processes } X(\cdot) \text{ such that } \mathbb{E}\left[\int_0^T |X(t)|^p dt\right] < \infty\right\}$

1 Introduction

In the recent development of control theory, the gap between optimal control and learning has been shrinking. In particular, the field of reinforcement learning (RL) has seen a rapid expansion of interest and research due to its growing effectiveness in real-world applications such as autonomous driving and robotics [1, 2]. However, the theoretical underpinnings which prove *why* a certain algorithm design choice improves performance are often left behind in the pursuit of impressive benchmark performance. While intuition can be useful in explaining the reason for the efficacy of certain algorithms, it is more difficult in general to specify the *conditions* under which certain choices will lead to improved performance without casting the problem in a mathematical framework and carrying out the requisite analysis to identify the assumptions required for performance improvement guarantees.

1.1 Reinforcement Learning and Optimal Control

One feature of both reinforcement learning and optimal control which render them well suited to practical applications is the fact that the notion of controls and the rewards they induce in a specified environment are at the core of the problem formulations, that is, their primary concern is in selecting the best actions to take according to some specified reward function.

While optimal control theory is concerned with *deriving* the optimal set of actions to take in order to maximise a certain reward function under a set of assumptions, reinforcement learning starts with the assumption that the model describing the underlying problem is *not fully known*. On one hand, optimal control assumes knowledge of the state dynamics. On the other hand, reinforcement learning takes the stance that such knowledge has to be acquired through interaction between the agent and its environment. One seemingly paradoxical question which we may ask is this:

How can we use optimal control in reinforcement learning?

The aim of this essay is to consider this question through the lens of one of the central pillars in reinforcement learning: exploration.

1.2 Exploration in Reinforcement Learning

In any situation where an agent is trying to maximise a specified reward with imperfect knowledge of the underlying environment, there are two sensible strategies they may consider:

1. **Exploit** their current knowledge of the environment and select the action which, to the best of their knowledge, will maximise the expected rewards gained.
2. **Explore** by selecting an action in a non-deterministic way to hopefully gain more information about the environment or which actions lead to higher rewards.

A good example of this which is often presented as an illustration of this trade-off in RL is the multi-armed bandit problem [3]: in a simplified formulation of this problem, consider an agent who enters a casino containing various slot machines whose payoffs are unknown to the agent upon arrival. Each of the slot machines may provide different expected payoffs and a problem the agent immediately faces captures the essence of the multi-armed bandit problem: *at each moment in time, what rule should they adopt in selecting which machine to play at?* Moreover, the trade-off between *exploring* different machines whose expected payoffs are more uncertain and *exploiting* machines that have performed well based on their past experience becomes apparent quite quickly.

The reason why exploration is important in reinforcement learning is that by assumption, agents do not have perfect knowledge about the environment in which they are operating so at best, they will be operating according to *estimates* of how the environment works. These current estimates are commonly referred to as the agent's *model* of the environment. Thus, until an agent is confident that their model accurately describes the full environment in the present moment and into the future, it is always useful to employ some level of exploration to avoid becoming stuck in local optima. Moreover, in many complex environments, the underlying environment is often *changing* with time and exploration gives a way of better maintaining an up-to-date description of the environment than in a purely exploitative strategy. In the field of optimal control however, emphasis is always placed on exploitation; if the full model is known, there is no need to perform exploration to gain knowledge about the system.

To study exploration under the optimal control framework then, we will need a method for incorporating the notion of selecting exploratory actions. In reinforcement learning, exploration has mostly been treated as an exogenous component to be selected [4, 5] until more recently, where the benefits of methods which *intrinsically reward exploration* have become more widespread [6, 7]. This idea was then analysed in the context of stochastic optimal control [8, 9], which has formed much of the basis of this thesis. The unique power of theoretical analysis in this context allows results derived using control theory to prescribe the optimal method of performing exploration in an environment. Thus, the bridge between control theory and reinforcement learning lies in the powerful simplification of the problem from learning among *all possible policies* to a much smaller set of policies, namely those that have the same *form* as the theoretically optimal one.

1.3 Roadmap

In **Chapter 2**, we derive and present fundamental results in stochastic optimal control which will be used several times throughout the remaining chapters, including the Hamilton-Jacobi-Bellman equation and the verification theorem. We will then analyse two well-known problems in stochastic optimal control: the Stochastic Linear Quadratic (SLQ) and continuous-time Mean-Variance (MV) problems [24, 21]. The former is considered one of the hallmark results in optimal control due to its widespread applicability and power in approximating nonlinear dynamics [11], while the latter is a widely studied problem in financial mathematics due to its intuitive appeal and relative ease in solving.

In **Chapter 3**, we will formulate the exploratory optimal control problem, which draws on key concepts from both the control theory literature [25] and reinforcement learning [7] known as control relaxation and entropy-regularisation. We also present a general framework for rewarding exploration and a policy improvement theorem which is used in designing reinforcement learning algorithms in subsequent chapters.

In **Chapter 4**, we will look at an application of the exploratory control framework to the Stochastic Linear Quadratic (SLQ) problem. The exploratory version of this problem is solved, and several interesting results are presented. We then present the Exploratory Linear Quadratic (ELQ) algorithm, which is a reinforcement learning algorithm based on the theoretical results derived and show some brief numerical results on the performance of our algorithm.

In **Chapter 5**, we look at a special case of stochastic LQ problem known as the Mean-Variance (MV) problem. Again, a solution is presented along with parallel results to those in chapter 4. Here, we present a detailed derivation for the Exploratory Mean-Variance (EMV) algorithm first described in Wang et al. [9] based once again on the theoretical results and present a numerical analysis showcasing the effectiveness of the EMV algorithm in practice. For these numerical results, we also implement a maximum likelihood estimation based algorithm for comparison purposes.

Finally in **Chapter 6**, we conclude with discussions on future areas of research and an evaluation of the results presented herein.

2 Stochastic Optimal Control

Stochastic optimal control is a mathematical framework optimising a given **reward function** whose value depends on some combination of:

1. The **state and dynamics** of a controlled stochastic system. This is a system whose state evolves according to both a deterministic and random component. At least one of these is influenced by the value of a function known as a control function.
2. The values of the **control function** themselves. Controls are usually constrained to some set known as an *admissible/feasible* set, but are otherwise free variables which can be *chosen* in such a way that the reward function is maximised.

Thus, one of the main practical concerns in optimal control can be summarised as:

Given a controlled system and a reward function, what control should be used to maximise the reward?

Several methods have been developed to address this particular question, one of which is known as Dynamic Programming (DP) which was developed in the early 1950's by Richard Bellman [14, 15]. In its essence, Dynamic Programming encapsulates the idea of expressing a problem recursively in terms of smaller sub-problems. One profound insight which Bellman proposed is known as **Bellman's 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.*'[16]

In the continuous time setting, one may wonder what happens as we take the time difference between the initial and resulting state to zero in the statement of this principle. The answer to this question is the **Hamilton-Jacobi-Bellman (HJB) equation**, which is a nonlinear 1^{st}-order PDE in deterministic problems which provides necessary and sufficient conditions for optimality, and a 2^{nd}-order PDE in stochastic problems providing only necessary conditions for optimality. We shall now present a derivation of the HJB equation and a theorem which gives sufficient conditions for optimality in this stochastic setting.

2.1 Deriving the Hamilton-Jacobi-Bellman equation

Throughout this derivation, we will be working with a filtered, complete probability space $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\}_{t\geq 0})$ with an associated $\mathcal{F}_t$-adapted, d-dimensional standard Brownian Motion $W = \{W_t\}_{t\geq 0}$ whose natural filtration $\mathcal{F}_t^W$ is augmented by all $\mathbb{P}$-null sets in $\mathcal{F}$.

We begin by defining general concepts that will be referred to throughout the rest of the thesis.

Definition 2.1. *A **controlled Stochastic Differential Equation (SDE)** is given by the system:*

$$\begin{cases} dX_t = \mu(t, X_t, u_t)dt + \sigma(t, X_t, u_t)dW_t \\ X_0 = x_0 \in \mathbb{R}^n \end{cases}, \tag{1}$$

for some functions $\mu : [0,\infty) \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ and $\sigma : [0,\infty) \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^{n\times d}$.

Definition 2.2. *A control law/function $u : [0,\infty) \times \mathbb{R}^n$ is **admissible** if*

- *$u(t,x) \in U \subseteq \mathbb{R}^k$ for all $(t,x) \in [0,\infty) \times \mathbb{R}^n$, that is, u takes values in some specified set $U \subseteq \mathbb{R}^k$*
- *For any initial point $(s,x) \in [0,\infty) \times \mathbb{R}^n$, the following controlled SDE has a unique solution:*

$$\begin{cases} dX_t = \mu(t, X_t, u_t)dt + \sigma(t, X_t, u_t)dW_t, \ s \leq t \\ X_s = x_s \end{cases}$$

Here, the class of admissible controls is denoted by $\mathcal{U}$. For the formulation of the control problem, we consider some fixed but arbitrary time $t \in [0, T]$ in a finite time horizon and a point $x \in \mathbb{R}^n$.

Definition 2.3. *If we let $f : [0, \infty) \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ be in $\mathcal{L}^1_{\mathcal{F}}(0, T; \mathbb{R})$ and $h : \mathbb{R}^n \to \mathbb{R}$ be in $\mathcal{L}^1_{\mathcal{F}_T}(\Omega; \mathbb{R})$, the **expected utility/value function** $V : [0, T] \times \mathbb{R}^n \times \mathcal{U} \to \mathbb{R}$ for an admissible control u is defined as*

$$V^u(t, x) \coloneqq \mathbb{E}\left[\int_t^T f(s, X_s^u, u_s)ds + h(X_T)\right].$$

Definition 2.4. *The **stochastic optimal control problem** is to find the optimal value function:*

$$\begin{aligned} &\textit{maximise } \mathbb{E}\left[\int_0^T f(s, X_s^u, u_s)ds + h(X_T)\right], \\ &\textit{subject to } u(s, y) \in U \quad \forall (s, y) \in [0, T] \times \mathbb{R}^n. \end{aligned}$$

Definition 2.5. *The **optimal value function** is defined by the greatest expected utility an agent can achieve from a given time and state (t, x):*

$$V(t, x) \coloneqq \sup_{u \in \mathcal{U}} V^u(t, x).$$

Definition 2.6. *The **optimal control** $\hat{u}$, if it exists, is the control under which the optimal value function is achieved:*

$$\hat{u} \coloneqq \arg\max_{u \in \mathcal{U}} V^u(t, x).$$

Here, we make an important distinction between a control's *value function* V^u (with the superscript) and the problem's *optimal value function* V (without the superscript), where the former is interchangeable with 'expected utility function' and denotes the expected accumulated utility of the agent associated with an *arbitrary admissible control*, whereas the latter denotes the value function under the *optimal control*. For this derivation, we introduce some non-specific assumptions so as not to detract from the key idea underlying the derivation. In later sections, we will be more specific about the technical conditions we require.

Assumptions:

1. There exists an optimal control $\hat{u}$.
2. The optimal value function $V \in C^{1,2}([0, T] \times \mathbb{R}^n; \mathbb{R})$ (C^1 in time and C^2 in space).

Approach: The Dynamic Programming approach to solving the control problem can be broken into three steps:

1. Given a starting point (t, x), we consider two possible strategies on $[t, T]$:
   (a) Use the optimal control $\hat{u}$.
   (b) Use a fixed, arbitrary control function u over a small interval $[t, t + h]$, and then switch to $\hat{u}$ for the remaining time interval $(t + h, T]$. We will call this strategy u^*.
2. Compute the expected utility for the two strategies and compare them to obtain an inequality.
3. Take the limit as $h \to 0$ and obtain a PDE for the optimal value function $V(t, x)$.

Step 1. Defining our strategies:

We fix $(t, x) \in (0, T) \times \mathbb{R}^n$, $h \in (0, T - t)$, and $u \in \mathcal{U}$, and define u^* in strategy (b) as:

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

Step 2.1 Computing expected utilities:

For **strategy (a)**, we can see that the expected utility starting from time t is simply the optimal value function at time t:

$$V^{\hat{u}}(t,x) = V(t,x).$$

For **strategy (b)**, we will express the total expected utility as the sum of the expected utility on the two periods $[t, t+h]$ and $(t+h, T]$. On $[t, t+h]$, we have:

$$\mathbb{E}_{t,x}\left[\int_t^{t+h} f(s, X_s^u, u_s)ds\right],$$

where we use the subscript t, x to denote that the expectation is taken given the initial condition $X_t = x$. We observe that the state at time $t+h$ will be X_{t+h}^u and since we are using the optimal control on the interval $(t+h, T]$, the expected utility on this interval is simply

$$\mathbb{E}_{t,x}\left[V(t+h, X_{t+h}^u)\right].$$

Adding these, expectations, we get the expected utility of strategy (b):

$$V^{u^*}(t,x) = \mathbb{E}_{t,x}\left[\int_t^{t+h} f(s, X_s^u, u_s)ds + V(t+h, X_{t+h}^u)\right].$$

Step 2.2 Comparing expected utilities:

We first note that since strategy (a) is optimal, it must be at least as good as strategy (b):

$$V^{\hat{u}}(t,x) = V(t,x) \geq \mathbb{E}_{t,x}\left[\int_t^{t+h} f(s, X_s^u, u_s)ds + V(t+h, X_{t+h}^u)\right] = V^{u^*}(t,x). \tag{2}$$

To write this inequality in terms of known states, we can expand the $V(t+h, X_{t+h}^u)$ term using the integral form of Itô's formula:

$$V(t+h, X_{t+h}^u) = V(t,x) + \int_t^{t+h}\left(\frac{\partial V}{\partial t}(s, X_s^u) + \mathcal{A}^u V(s, X_s^u)\right)ds + \int_t^{t+h} \nabla_x V(s, X_s^u)\sigma_s^u dW_s.$$

Taking the expectation of both sides of this equation and assuming enough integrability, the stochastic integral vanishes, leaving the expectation of the first two terms:

$$\begin{aligned}\mathbb{E}_{t,x}\left[V(t+h, X_{t+h}^u)\right] &= \mathbb{E}_{t,x}\left[V(t,x) + \int_t^{t+h}\left(\frac{\partial V}{\partial t}(s, X_s^u) + \mathcal{A}^u V(s, X_s^u)\right)ds + \int_t^{t+h} \nabla_x V(s, X_s^u)\sigma_s^u dW_s\right] \\ &= V(t,x) + \mathbb{E}_{t,x}\left[\int_t^{t+h}\left(\frac{\partial V}{\partial t}(s, X_s^u) + \mathcal{A}^u V(s, X_s^u)\right)ds\right].\end{aligned}$$

Plugging the right hand side of this equation into our inequality (2), we get

$$\begin{aligned}V(t,x) &\geq \mathbb{E}_{t,x}\left[\int_t^{t+h}\left(f(s, X_s^u, u_s) + \frac{\partial V}{\partial t}(s, X_s^u) + \mathcal{A}^u V(s, X_s^u)\right)ds\right] + V(t,x) \\ \implies \mathbb{E}_{t,x}&\left[\int_t^{t+h}\left(f(s, X_s^u, u_s) + \frac{\partial V}{\partial t}(s, X_s^u) + \mathcal{A}^u V(s, X_s^u)\right)ds\right] \leq 0.\end{aligned}$$

Step 3 Taking the limit:
For the final step, we divide the last inequality obtained above by h:

$$\mathbb{E}_{t,x}\left[\int_t^{t+h} \frac{f(s, X_s^u, u_s) + \frac{\partial V}{\partial t}(s, X_s^u) + \mathcal{A}^u V(s, X_s^u)}{h} ds\right] \leq 0.$$

Now, we want to take the limit as $h \to 0$. If we assume that the integrand is bounded by a constant or a random variable, we can use the bounded or dominated convergence theorems respectively to take the limit inside the expectation. To actually compute this limit, we can apply the fundamental theorem of integral calculus:

$$f(x) = \lim_{h \to 0} \frac{A(x+h) - A(x)}{h},$$

where $A(\cdot)$ is an area function. Applying this to the expectation term above and recalling that $X_t = x$, we get

$$\begin{aligned}\mathbb{E}_{t,x}\left[\lim_{h\to 0}\int_t^{t+h} \frac{f(s, X_s^u, u_s) + \frac{\partial V}{\partial t}(s, X_s^u) + \mathcal{A}^u V(s, X_s^u)}{h} ds\right] &= \mathbb{E}_{t,x}\left[f(t,x,u) + \frac{\partial V}{\partial t}(t,x) + \mathcal{A}^u V(t,x)\right] \\ &= f(t,x,u) + \frac{\partial V}{\partial t}(t,x) + \mathcal{A}^u V(t,x).\end{aligned}$$

Thus, we get

$$f(t,x,u) + \frac{\partial V}{\partial t}(t,x) + \mathcal{A}^u V(t,x) \leq 0,$$

for all $u \in \mathcal{U}$ and all $(t,x) \in (0,T) \times \mathbb{R}^n$ since we chose an arbitrary control and starting point to begin with. Moreover, we have equality if and only if

$$\frac{\partial V}{\partial t}(t,x) + \sup_{u \in \mathcal{U}} \left(f(t,x,u) + \mathcal{A}^u V(t,x)\right) = 0.$$

To complete this PDE, we apply the boundary condition $V_T(x) = h(x)$ which gives us the Hamilton-Jacobi-Bellman (HJB) Equation:

$$\begin{cases} \frac{\partial V}{\partial t}(t,x) + \sup_{u\in\mathcal{U}} \left(f(t,x,u) + \mathcal{A}^u V(t,x)\right) = 0 & \forall (t,x) \in (0,T) \times \mathbb{R}^n, \\ V_T(x) = h(x) & \forall x \in \mathbb{R}^n \end{cases} \tag{3}$$

2.1.1 Verification Theorem

To conclude this brief review of results in classical stochastic optimal control, we present an important result in optimal control:

Theorem 2.1 (Verification Theorem)**.** *Suppose that for the **stochastic control problem**,*

1. *A function $v(t,x)$ and a control $\hat{u}(t,x)$ satisfy the problem's HJB equation.*
2. *v is sufficiently integrable so that the Itô integral of $\nabla_x v(s, X_s^u)\sigma_s^u$ is a local martingale.*
3. *$\hat{u}(t,x) \in \mathcal{U}$: the control which maximises the HJB equation is admissible.*

Then, v is the optimal value function of the control problem and $\hat{u}$ is the optimal control at which the supremum in the HJB equation is obtained.

Proof: See page 291 of [18]. For the sake of brevity, we omit the proof here.

With these general results laid down, we will spend the rest of this chapter examining two important applications of Dynamic Programming to well known-problems in optimal control and financial mathematics.

2.2 Stochastic Linear Quadratic Regulator

The Linear Quadratic Regulator (LQR) is a special type of optimal control problem where the state equations are linear in both the state and control functions, and the cost functions are quadratic. This is a widely studied problem, with applications in astrodynamics [11], algorithmic trading [12], and even recent areas such as mean-field theory [13]. Here, we will look at a formulation of a stochastic LQ (SLQ) problem.

2.2.1 Stochastic LQ Problem Setup

Definition 2.7. *For any $(s, y) \in [0, T] \times \mathbb{R}^n$, the **SLQ state equation** is:*

$$\begin{cases} dx_t = [A_t x_t + B_t u_t + b_t]dt + \sum_{j=1}^{m} [C_j(t)x_t + D_j(t)u_t + \sigma_j(t)]\, dW_j(t), \quad t \in [s, T] \\ x_s = y \end{cases}$$

where $A, B, C_j, D_j, b, \sigma_j$ are deterministic matrix valued functions of the appropriate sizes.

Definition 2.8. *For a given admissible control u and any $(s, y) \in [0, T] \times \mathbb{R}^n$, the **SLQ Cost function** is:*

$$V^u(s, y) = E\left[\frac{1}{2}\int_s^T \left(\langle\, Q_t x_t, x_t\rangle + 2\langle\, S_t x_t, u_t\rangle + \langle\, R_t u_t, u_t\rangle\right) dt + \frac{1}{2}\langle\, G x_T, x_T\rangle\right],$$

where Q, S, and R are $\mathcal{S}^n, \mathbb{R}^{k\times n}, \mathcal{S}^k$.

Here, we note that we are using the term *cost function* because this problem is posed as a minimisation problem. In this context, the cost function may represent some combination of a 'tracking error' and a 'resource usage' term. The implications of this formulation on the solution will be discussed after the formulation of the problem.

Definition 2.9. *The **SLQ class of admissible controls** denoted $\mathcal{U}^w[s, T]$ for any $s \in [0, T)$ is defined as the set of all 5-tuples $(\Omega, \mathcal{F}, \mathbb{P}, W, u(\cdot))$ where:*

1. *$(\Omega, \mathcal{F}, \mathbb{P})$ is a complete probability space.*
2. *W is an m-dimensional standard Brownian Motion (BM) on the probability space whose natural filtration $\mathcal{F}_{s,t}^W := \sigma\{W(r) : r \in [s, t]\}$ is $\mathbb{P}$-complete.*
3. *Under $u(\cdot) \in \mathcal{L}^2_{\mathcal{F}}(s, T; \mathbb{R}^k)$, for any $y \in \mathbb{R}^n$, the SLQ state equation admits a unique solution to the SLQ state equation on $(\Omega, \mathcal{F}, \{F_{s,t}^W, \mathbb{P}\}$.*
4. *The SLQ cost function is well defined under u.*

Definition 2.10. *For any $(s, y) \in [0, T] \times \mathbb{R}^n$, the **SLQ problem** is to find $\hat{u} \in \mathcal{U}^w[s, T]$ such that*

$$V^{\hat{u}}(s, y) = \inf_{u \in \mathcal{U}^w[s,T]} V^u(s, y) \;\; (= V(s, y)).$$

If we are to apply the HJB equation as derived in section (2.1), we can simply convert the cost functional into a utility functional by taking the negative of the cost functional. This introduces minus signs in the HJB equation in front of the $\frac{\partial V}{\partial t}$ term and in the expression which is maximised over. Before we apply the technique of Dynamic Programming to solve this problem, we introduce one assumption so that the following arguments are justified and the SLQ problem is solvable.

SLQ Assumption (L1): $A, B, C_j, D_j, Q, R \in \mathcal{L}^\infty(0, T)$, $G \in \mathcal{S}^n$, and $b, \sigma_j \in \mathcal{L}^2(0, T)$ all of the appropriate sizes, for $j = 1, .., m$.

With the problem defined, we can now use Dynamic Programming to solve the SLQ problem.

2.2.2 Dynamic Programming Solution to SLQ

We recall that the optimal value function of our SLQ problem satisfies the HJB equation for our minimisation problem:

$$\begin{cases} -\dfrac{\partial V}{\partial t}(t,x) + \sup_{u\in\mathcal{U}} G(t,x,u,V_x,V_{xx}) = 0 & \forall (t,x) \in (0,T)\times\mathbb{R}^n \\ V_T(x) = \frac{1}{2}\langle Gx, x\rangle & \forall x \in \mathbb{R}^n, \end{cases} \tag{4}$$

where $G(t,x,u,V_x,V_{xx})$ is known as the generalised Hamiltonian of this system:

$$\begin{aligned} G(t,x,u,V_x,V_{xx}) = &-\frac{1}{2}\langle Qx,x\rangle - \langle Sx,u\rangle - \frac{1}{2}\langle Ru,u\rangle \\ &- \langle V_x, Ax+Bu+b\rangle - \frac{1}{2}\langle V_{xx}(Cu+Dx+\sigma), Cu+Dx+\sigma\rangle. \end{aligned}$$

Observing that the instantaneous contribution to the cost function is quadratic in x and that the terminal value of the optimal value function is quadratic in x, we make the ansatz that the optimal value function $V(t,x)$ is quadratic in x, having the form

$$\begin{aligned} V(t,x) &= \frac{1}{2}\langle P_t x, x\rangle + \langle \phi_t, x\rangle + f_t, \\ P_T = G, \phi_T = 0, f_T &= 0, \end{aligned}$$

for some suitable functions $P(t), \phi(t), f(t)$ where $P(t)$ is symmetric. Therefore, differentiating with respect to x and dropping the explicit time dependence, we get

$$\begin{aligned} V_x &= Px + \phi \\ V_{xx} &= P. \end{aligned}$$

We now substitute this ansatz for the optimal value function into our equation for the generalised Hamiltonian G:

$$\begin{aligned} G(t,x,u,V_x,V_{xx}) = &-\frac{1}{2}\langle Qx,x\rangle - \langle Sx,u\rangle - \frac{1}{2}\langle Ru,u\rangle \\ &- \langle Px+\phi, Ax+Bu+b\rangle - \frac{1}{2}\langle P(Cu+Dx+\sigma), Cu+Dx+\sigma\rangle. \end{aligned}$$

To find the control function that maximises the generalised Hamiltonian, we can complete the square on the control function u. To this end, we rewrite G in terms of powers of u:

$$\begin{aligned} G(t,x,u,V_x,V_{xx}) = &-\frac{1}{2}u^\intercal\left(R + D^\intercal P D\right)u - \frac{1}{2}\left[2Sx + 2B^\intercal(Px+\phi) + C^\intercal P(Dx+\sigma) + (Dx+\sigma)^\intercal PC\right]^\intercal u \\ &- \frac{1}{2}\left(x^\intercal Qx + 2(Px+\phi)^\intercal(Ax+b) + (Dx+\sigma)^\intercal P(Dx+\sigma)\right). \end{aligned}$$

For notational purposes, we define

$$\begin{cases} \hat{R} \coloneqq R + D^\intercal P D \\ \hat{S} \coloneqq B^\intercal P + S + D^\intercal P C \\ \Psi \coloneqq \hat{R}^{-1}\hat{S} = (R + D^\intercal P D)^{-1}(B^\intercal P + S + D^\intercal P C) \\ \psi \coloneqq \hat{R}^{-1}(B^\intercal\phi + D^\intercal P\sigma) = (R + D^\intercal P D)^{-1}(B^\intercal\phi + D^\intercal P\sigma) \end{cases} \tag{5}$$

We can now complete the square several times to obtain a factored form for the generalised Hamiltonian in terms of these newly defined matrices:

$$\begin{aligned} G(t,x,u,V_x,V_{xx}) &= -\frac{1}{2}|\hat{R}^{1/2}[u+\Psi x+\psi]|^2 \text{ (quadratic in } u) \\ &+ \frac{1}{2}\left\{\langle\, \hat{S}^\intercal \hat{R}^{-1}\hat{S}x, x\rangle + \langle\, (-Q - C^\intercal PC - P^\intercal A - A^\intercal P)x, x\rangle\right\} \text{ (quadratic in } x) \\ &- \langle\, A^\intercal\phi + C^\intercal P\sigma - \Psi^\intercal \hat{R}\psi + Pb, x\rangle \text{ (linear in } x) \\ &- \frac{1}{2}|\hat{R}^{1/2}\psi|^2 - \langle\, \phi, b\rangle - \frac{1}{2}\sigma^\intercal P\sigma \text{ (constant)}. \end{aligned}$$

Now that we have isolated the control function in the generalised Hamiltonian, we can use the Verification Theorem 2.1 which says that the optimal control $\hat{u}$ has the property that for $t \in [0,T]$, the optimal control should maximise the generalised Hamiltonian. Thus, since the term which is quadratic in u is strictly ≤ 0, G is maximised precisely when this term is equal to 0:

$$\hat{u} = -\Psi x - \psi,$$

provided that $\hat{R} = R + D^\intercal PD > 0,\ \forall t \in [0,T]$. Thus, the HJB equation (4) can be written as

$$\begin{aligned} \frac{\partial V}{\partial t} &= \frac{1}{2}\langle\, \dot{P}x, x\rangle + \langle\, \dot{\phi}, x\rangle + \dot{f} = \sup_{u\in\mathcal{U}} G(t,x,u,V_x,V_{xx}) \\ &= +\frac{1}{2}\left\{\langle\, \hat{S}^\intercal \hat{R}^{-1}\hat{S}x, x\rangle + \langle\, (-Q - C^\intercal PC - P^\intercal A - A^\intercal P)x, x\rangle\right\} \\ &- \langle\, A^\intercal\phi + C^\intercal P\sigma - \Psi^\intercal \hat{R}\psi + Pb, x\rangle \\ &- \frac{1}{2}|\hat{R}^{1/2}\psi|^2 - \langle\, \phi, b\rangle - \frac{1}{2}\sigma^\intercal P\sigma. \end{aligned}$$

Comparing coefficients of the terms quadratic in x, we get the **stochastic Riccati equation** associated with SLQ:

$$\begin{cases} \dot{P} - \hat{S}^\intercal \hat{R}^{-1}\hat{S} + P^\intercal A + A^\intercal P + C^\intercal PC + Q = 0 \quad a.e.\ t \in [s,T] \\ P_T = G \\ R_t + D_t^\intercal P_t D_t > 0 \qquad\qquad\qquad\qquad\qquad\qquad a.e.\ t \in [s,T], \end{cases} \tag{6}$$

where the last condition is required to have a well-posed, real-valued solution. Comparing coefficients of the terms linear in x, we get an ODE for ϕ which can be solved directly by integration:

$$\begin{cases} \dot{\phi} + \left[A - B\hat{R}^{-1}\hat{S}\right]^\intercal \phi + \left[C - D\hat{R}^{-1}\hat{S}\right] P\sigma + Pb = 0 \quad a.e.\ t \in [s,T] \\ \phi_T = 0. \end{cases} \tag{7}$$

Finally, comparing the constant terms, we arrive at a representation for the value function $V(s,y)$ by integrating:

$$V(s,y) = \frac{1}{2}\langle\, P_s y, y\rangle + \langle\, \phi_s, y\rangle + \frac{1}{2}E\left[\int_s^T \left\{2\langle\, \phi, b\rangle + \langle\, P\sigma, \sigma\rangle - |\hat{R}^{1/2}\psi|^2\right\} dt\right] \ \ \forall y \in \mathbb{R}^n. \tag{8}$$

Thus, the solvability of our SLQ is reduced to the solvability of the stochastic Riccati equation, which can be numerically solved if solutions exist. We conclude this derivation with a Theorem on the solvability:

Theorem 2.2 (Solvability of SLQ)**.** *Let SLQ assumption (L1) hold. Let $P \in C([s,T];\mathcal{S}^n)$ and $\phi \in C([s,T];\mathbb{R}^n)$ be solutions of systems* (6) *and* (7) *respectively for some $s \in [0,T]$ such that*

$$\begin{cases} B\Psi, D\Psi \in \mathcal{L}^\infty(s,T;\mathbb{R}^{n\times n}) \\ B\psi, D\psi \in \mathcal{L}^2(s,T;\mathbb{R}^n), \end{cases}$$

where Ψ, ψ are as defined in (5).
Then SLQ is solvable at s, that is, there exists an optimal control $\hat{u}$ given by the feedback form:

$$\hat{u}_t = -\Psi_t x_t - \psi_t \quad a.e.\ t \in [s, T],$$

and the optimal value function $V(s, y)$ is the SLQ cost functional evaluated at the optimal control, given by (8).

Proof: See page 315 in [24]. We omit the proof here for brevity since similar verification arguments will be outlined in later sections.

2.3 Classical Mean-Variance Problem

We will now examine a formulation of the classical continuous-time MV problem, which is a well-known problem in financial mathematics and will serve as a useful problem which we can analyse in the exploratory setting later on.

2.3.1 Classical MV Problem Setup

For this problem, we will consider a market consisting of one risky asset and one riskless asset for simplicity, although this can be extended to the multi-dimensional case. We fix an investment planning horizon $T > 0$ and let $\{W_t, 0 \leq t \leq T\}$ be a standard one-dimensional Brownian motion defined on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ satisfying the usual conditions ($\mathbb{P}$-completeness and right continuity).

Definition 2.11. *The risky asset's **price process** is a stochastic process governed by a **Geometric Brownian Motion**, which has the form*

$$dS_t = S_t(\mu dt + \sigma dW_t), \ \ 0 \leq t \leq T, \tag{9}$$

where $S_0 = s_0 > 0, \mu \in \mathbb{R}$, and $\sigma > 0$.

Definition 2.12. *The **Sharpe ratio** of the risky asset is given by*

$$\rho \coloneqq \frac{\mu - r}{\sigma}. \tag{10}$$

We note that for the Mean-Variance problem and its exploratory counterpart, ρ will denote the Sharpe ratio of the risky asset as opposed to the discount rate seen in section (3). One way to think about this quantity is that it represents the average excess return of choosing the risky asset per unit of risk.

Definition 2.13. *The **riskless asset**'s price is a continuous deterministic process following exponential growth with a constant interest rate $r > 0$.*

Definition 2.14. *The **(discounted) wealth process** of an agent is denoted $\{x_t^u, 0 \leq t \leq T\}$ with **asset allocation strategy** $u = \{u_t, 0 \leq t \leq T\}$ where u_t denotes the discounted dollar value put in the risky asset at time t.*

In essence, discounting allows us to ignore the effect of the growth accrued by investing in the riskless asset for notational convenience, so that if wealth is allocated to the riskless asset, its discounted value does not grow over time. Furthermore, the asset allocation strategy must satisfy the *self-financing assumption* which states that purchase of new assets must be financed by the sale of currently held assets, and there is no injection or withdrawal of funds during the process.

To ensure existence of solutions to this problem, we define the set of admissible controls:

Definition 2.15. *The **set of admissible controls** $\mathcal{A}^{cl}(s, y)$ for $(s, y) \in [0, T) \times \mathbb{R}$ for the classical MV problem* (17) *is:*

$$\mathcal{A}^{cl}(s, y) \coloneqq \left\{ u = \{u_t, t \in [s, T]\} : u \text{ is } \mathcal{F}_t\text{-progressively measurable and } \mathbb{E}\left[\int_s^T (u_s)^2 ds\right] < \infty \right\}. \tag{11}$$

Under this definition, the set of admissible controls $\mathcal{A}^{cl}$ (and admissible wealth processes generated by these controls) are convex sets or intervals.

Definition 2.16. *Following an admissible allocation strategy $u \in A^{cl}(0,x)$, the **wealth dynamics** of the agent is*

$$dx_t^u = u_t(\mu dt + \sigma dW_t) - u_t r dt = \sigma u_t(\rho dt + dW_t), \tag{12}$$

with initial endowment $x_0^u = x_0 \in \mathbb{R}$.

Definition 2.17. *The **classical continuous-time MV problem** is given by the following constrained optimisation problem:*

$$\begin{aligned} &\min_{u \in \mathcal{A}^{cl}(0,x)} Var(x_T^u) \\ &\textit{subject to } \mathbb{E}\left[x_T^u\right] = z. \end{aligned} \tag{13}$$

The intuition behind this problem is that an agent wishes to achieve a target mean wealth of z by the end of the planning horizon ($t = T$) with minimal variance.

To solve this problem, we note that in its given form, the classical MV problem is a **constrained** optimisation problem, which was not considered in the Dynamic Programming approach outlined in section 2.1. Therefore, we need to formulate an equivalent **unconstrained** problem whose solution coincides with the solution to the constrained one above before we can apply Dynamic Programming. Specifically, we can solve the classical MV problem by:

1. Introducing a Lagrange multiplier w to incorporate the constraint into the problem so that we have an *unconstrained* optimisation problem.
2. Solving the unconstrained problem using Dynamic Programming.
3. Searching for a Lagrange multiplier w such that under the value of w, the expected terminal wealth constraint $\mathbb{E}\left[x_T^u\right] = z$ is satisfied.

The advantage of such an approach is that using the Lagrangian formulation allows us to apply the powerful technique of Dynamic Programming which we have seen earlier to this control problem.

2.3.2 Dynamic Programming Solution To Classical MV Problem

We begin with step 1 outlined above. Writing out the variance in full and using the constraint, the Lagrangian function of this problem is given by

$$\begin{aligned} \mathcal{L}(u,\lambda) &= \text{Var}(x_T^u) - 2w(\mathbb{E}\left[x_T^u\right] - z) && (14)\\ &= \mathbb{E}\left[x_T^u\right] - z^2 - 2w(\mathbb{E}\left[x_T^u\right] - z) + w^2 - w^2 && (15)\\ &= \mathbb{E}\left[(x_T^u - w)^2\right] - (w-z)^2. && (16) \end{aligned}$$

Due to the convexity of the objective function $\text{Var}(x_T^u)$ under the technical assumptions on our asset allocation strategy, it is sufficient to minimise this Lagrangian function over the set of admissible controls to find the optimal strategy, so we can write (16) as a minimization problem:

$$\min_{u \in \mathcal{A}^{cl}(0,x)} \mathbb{E}\left[(x_T^u - w)^2\right] - (w-z)^2, \tag{17}$$

which can be solved analytically using the tools of Dynamic Programming. With this set of admissible classical controls defined, we can state the **optimal value function** of the classical MV problem: For $(s,y) \in [0,T) \times \mathbb{R}$ and $w \in \mathbb{R}$ fixed:

$$V^{cl}(s,y;w) \coloneqq \inf_{u \in \mathcal{A}^{cl}(s,y)} \mathbb{E}\left[(x_T^u - w)^2 \middle| x_s^u = y\right] - (w-z)^2. \tag{18}$$

Now that we have posed the unconstrained problem we wish to optimise, we can apply the technique of Dynamic Programming. The HJB equation for this problem is given by

$$\frac{\partial v}{\partial t}(s,y;w) + \min_{u\in\mathbb{R}}\left(\frac{1}{2}\sigma^2 u^2 v_{xx}(s,y;w) + \rho\sigma u v_x(s,y;w)\right) = 0,\ (s,y)\in[0,T]\times\mathbb{R}. \tag{19}$$

Solving the simple quadratic minimisation problem on the left hand side, we get a candidate for the optimal feedback control in terms of the solution to the HJB equation v:

$$\hat{\boldsymbol{u}}(t,x;w) = -\frac{\rho}{\sigma}\frac{v_x(t,x;w)}{v_{xx}(t,x;w)}. \tag{20}$$

We can then substitute this optimal control into the HJB equation to yield a PDE for v which can be solved using the same ansatz technique as we have seen in the solution to the SLQ problem to solve for v. Carrying this out, the solution can be written as

$$v(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} - (w-z)^2.$$

Taking the necessary derivatives of this solution and substituting them into the minimising feedback control (20), we get a candidate optimal feedback control in terms of the problem parameters ρ,σ. It can be checked that the problem as formulated and the functions $v,\hat{\boldsymbol{u}}$ satisfy the assumptions of the Verification Theorem 2.1, so the optimal value function, optimal control, and state dynamics under the optimal control are indeed given by:

$$V^{cl}(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} - (w-z)^2 \tag{21}$$

$$\hat{\boldsymbol{u}}(u;t,x,w) = -\frac{\rho}{\sigma}(x-w) \tag{22}$$

$$\begin{cases} d\hat{x}_t = -\rho^2(\hat{x}_t - w)dt - \rho(\hat{x}_t - w)dW_t, \\ \hat{x}_0 = x_0 \end{cases} \tag{23}$$

Finally, we carry out step 3 by **choosing** the Lagrange multiplier such that the solutions given satisfy the terminal expected wealth constraint:

$$w = \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}. \tag{24}$$

The derivation of this form for the Lagrange multiplier will be given in the proof of Theorem 5.1.

3 Introduction to Exploratory Optimal Control

This section will explore the application of ideas in stochastic optimal control to a key problem in reinforcement learning: determining the trade-off between exploration of the control space and exploitation of an agent's current knowledge about the best control to select.

As mentioned in the introduction, the classical solution to a stochastic control problem forces an agent to select the optimal control at any given time. In the context of Reinforcement Learning (RL) however, the agent is not assumed to have an accurate model of the state dynamics, so it is useful to perform exploration. To model exploration in the stochastic control setting, we introduce two ideas which, when coupled together, form a new problem which allows us to formalise questions such as 'what is the optimal trade-off between exploration and exploitation?'

The first of these ideas is known as **control relaxation**, a technique originating from the stochastic control literature which was first posed by Fleming and Nisio in 1984 [25]. However, this concept was originally introduced to answer questions about the existence of optimal controls, and had not been viewed under the exploratory lens until recent work by Wang et al. in 2019 [8].

The second idea is **entropy-regularisation** which comes from the RL community and has seen growing adoption as one of the mainstream techniques for performing exploration in RL settings [6, 7]. The key idea behind entropy-regularisation is that the rewards an agent receives should also be linked to the 'amount of exploration' that their policy induces, that is, exploration should be intrinsically rewarding for the agent. Naturally, there are many possible ways to quantify the amount of exploration that an agent carries out, so we will first formulate a general version of this idea known as **exploration rewards**, which allows for the specification of any sufficiently smooth function to be used in rewarding exploration. After some results have been presented for this general case, we will introduce a specific choice of exploration reward which leads to the standard entropy-regularisation technique. Entropy-regularisation will be used as our exploration reward functional for the rest of the applications throughout this thesis due to its well-known properties and intuitive link to exploration. We will see that these two techniques complement each other very well to produce a new class of problems known as **exploratory optimal control** problems.

3.1 Exploratory Stochastic Control Problem

We begin by considering a general classical stochastic optimal control problem whose coefficients are time-independent for simplicity.

3.1.1 Classical Stochastic Control Problem

We will work with a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\}_{t\geq 0})$ with an associated $\{\mathcal{F}_t\}$-adapted Brownian Motion $W = \{W_t\}_{t\geq 0}$. As before, the set of possible values for an admissible control is denoted U, the set of admissible controls taking values in U is denoted $\mathcal{U}^{cl}$, and the **classical state dynamics** is given by

$$dx_t^u = b(x_t^u, u_t)dt + \sigma(x_t^u, u_t)dW_t,\ t > 0;\ x_0 = x\ \in\ \mathbb{R}. \tag{25}$$

The associated **classical optimal value function** which we are interested in, denoted V^{cl}, is the expected total discounted reward:

$$V^{cl}(x) \coloneqq \sup_{u\in\mathcal{U}^{cl}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t} r(x_t^u, u_t)dt \,\middle|\, x_0 = x\right]. \tag{26}$$

Where ρ is known as a **discount rate** on later rewards and $r(x_t^u, u_t)$ is known as the **instantaneous reward function**. In this case, the optimal control which attains the supremum in the value function is a deterministic mapping from the current state x_t to U and is denoted $\hat{u}_t \equiv \hat{u}(\hat{x}_t)$.

3.1.2 Relaxed Stochastic Control

We can model exploration of the control space by using control relaxation as mentioned earlier. The key idea of this framework is to use a **distribution of controls** $\pi = \{\pi_t(u), t \geq 0\}$ over U from which individual controls are sampled. In the continuous-time context, controls are continuously being sampled from a certain distribution π chosen by the agent and applied on infinitesimal time intervals. From this perspective, we can view the agent's next state after applying a sampled control as being itself sampled from a distribution of possible states *induced* by the distribution of controls π.

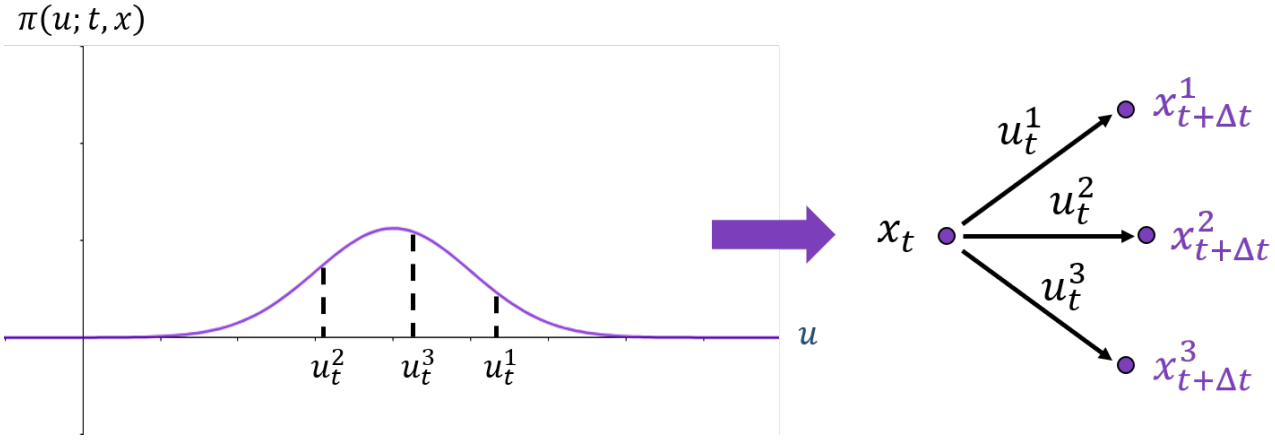

Figure 1: Illustration of the idea behind control sampling. Sampling and applying controls from a policy/control distribution π in state x_t over a time interval Δt is equivalent to sampling a next state $x_{t+\Delta t}$ from a *distribution* of possible states X_t^π.

As the agent is not allowed to specify the control they select but rather the distribution which they sample from, the control problem is concerned with selecting an optimal *distribution of controls* rather than an optimal deterministic control. We will henceforth use the term **policy** to refer to an agent's control distribution.

However, it is not immediately clear in this new setting how to quantify the notion of a 'good' policy. Indeed, the classical state dynamics (25) and optimal value function (26) were defined in terms of the value of the state process and a deterministic control, which are no longer valid under the control sampling regime. To define the relaxed analogue of these two equations, we need a method called **policy evaluation** which can be described by the following procedure:

1. Given a starting point x_t, sample N controls $u_t^i,\ i = 1, ..., N$ from the policy π.
2. Apply each control u_t^i from the state x_t using the classical state dynamics (25) over a small time period Δt.
3. Take the sample mean of the discounted reward over all the trials to obtain an estimate for the '*average reward under the policy*'.
4. Send $\Delta t \to 0$ to obtain a stochastic differential equation.

To carry out steps 1 and 2, we note that since the Brownian motion generates random paths for each trial, we can express them as $W_t^i,\ i = 1, ..., N$, so the resulting state dynamics for each trial are approximated by

$$\Delta x_t^i := x_{t+\Delta t}^i - x_t^i \approx b(x_t^i, u_t^i)\Delta t + \sigma(x_t^i, u_t^i)(\Delta W_t^i),$$

where $\Delta W_t^i := W_{t+\Delta t}^i - W_t^i$.

We let X^π be the *distribution of states* from which each x^i is independently sampled as a result of sampling controls u^i from π. If we assume that the distributions π_t, X_t^π are independent of the increments of

the Brownian motion, we can average out the Δx^i's and use the law of large numbers to take the limit as $N \to \infty$:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \Delta x_t^i &\approx \frac{1}{N}\sum_{i=1}^{N} \left\{ b(x_t^i, u_t^i)\Delta t + \sigma(x_t^i, u_t^i)(\Delta W_t^i) \right\} \\
&\overset{a.s.}{\to} \mathbb{E}\left[b(X_t^\pi, u)\Delta t \right] + \mathbb{E}\left[\sigma(X_t^\pi, u)(\Delta W_t^i) \right] \\
&= \int_U b(X_t^\pi, u)\pi_t(u)du\Delta t + \mathbb{E}\left[\sigma(X_t^\pi, u)\right]\mathbb{E}\left[\Delta W_t^i\right] \quad \text{(by independence)} \\
&= \int_U b(X_t^\pi, u)\pi_t(u)du\Delta t \quad \text{(by property of BM's)} \\
&= \mathbb{E}[\Delta X_t^\pi] \quad \text{(since each } x_t^i \text{ is independently sampled from } X_t^\pi).
\end{aligned}$$

In a similar manner, we can compute the second moment of ΔX_t^π. Ignoring terms involving $(\Delta t)^2$, we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} (\Delta x_t^i)^2 &\approx \frac{1}{N}\sum_{i=1}^{N} \left\{ b(x_t^i, u_t^i)(\Delta t)(\Delta W_t^i) + \sigma^2(x_t^i, u_t^i)(\Delta W_t^i)^2 \right\} \\
&\overset{a.s.}{\to} \mathbb{E}\left[b(X_t^\pi, u)\right]\mathbb{E}\left[\Delta W_t^i\right]\Delta t + \mathbb{E}\left[\sigma^2(x_t^i, u_t^i)\right]\mathbb{E}\left[(\Delta W_t^i)^2\right] \\
&= 0 + \int_U \sigma^2(X_t^\pi, u)\pi_t(u)du\Delta t \\
&= \mathbb{E}[(\Delta X_t^\pi)^2],
\end{aligned}$$

where we use that $\mathbb{E}[(\Delta W_t^i)^2] = \Delta t$ for the second-last equality.

Using the computed first and second moments, we propose the exploratory formulation of the controlled state dynamics

Definition 3.1. *The **exploratory SDE** is*

$$dX_t^\pi = \tilde{b}(X_t^\pi, \pi_t)dt + \tilde{\sigma}(X_t^\pi, \pi_t)dW_t; \;\; X_0^\pi = x \in \mathbb{R}, \tag{27}$$

*where the **exploratory drift** $\tilde{b}(\cdot,\cdot)$ is given by*

$$\tilde{b}(y, \pi) \coloneqq \int_U b(y, u)\pi(u)du, \;\; y \in \mathbb{R}, \; \pi \in \mathcal{P}(U), \tag{28}$$

*the **exploratory volatility** $\tilde{\sigma}(\cdot,\cdot)$ is given by*

$$\tilde{\sigma}(y, \pi) \coloneqq \sqrt{\int_U \sigma^2(y, u)\pi(u)du}, \;\; y \in \mathbb{R}, \; \pi \in \mathcal{P}(U), \tag{29}$$

and $\mathcal{P}(U)$ is the set of probability density functions $\pi(\cdot)$ such that

$$\int_U \pi(u)du = 1 \text{ and } \pi(u) \geq 0 \;\; \textit{a.e. on } U. \tag{30}$$

Intuitively, we can think of the exploratory state dynamics as the 'expected' state increment that we would get when sampling controls from the policy π, which is why the drift and diffusion terms are governed by the first and second moments of the increment under the distribution π.

Carrying out steps 3 and 4 to achieve a relaxed stochastic control problem, we define the analogue for the reward function by averaging the reward function over the trials and taking the limit as $N \to \infty$:

$$\frac{1}{N}\sum_{i=1}^{N} e^{-\rho t} r(x_t^i, u_t^i)\Delta t \xrightarrow{a.s.} e^{-\rho t}\int_U r(X_t^\pi, u)\pi_t(u)du\Delta t,$$

which leads to the **exploratory reward**:

$$\tilde{r}(y,\pi) \coloneqq \int_U r(y,u)\pi(u)du,\ y \in \mathbb{R},\ \pi \in \mathcal{P}(U). \tag{31}$$

3.1.3 Exploration Rewards

The second key component of our formulation is known as **exploration rewards**. To motivate the necessity of this feature, consider the effect of control relaxation outlined above. By itself, the relaxed problem formulated above has simply *embedded* the classical control problem into a stochastic decision making framework, but has not provided any **incentive** for an agent to select a control other than the optimal one. Indeed, the optimal policy for such a formulation corresponds to the Dirac measure centered at the optimal control for the classical problem, denoted $\delta_{\hat{\boldsymbol{u}}}(\cdot)$.

To incentivise exploratory behaviour which corresponds to selecting possibly sub-optimal actions, we can include a **reward for exploration** in the problem's reward function. This is the key idea behind exploration rewards.

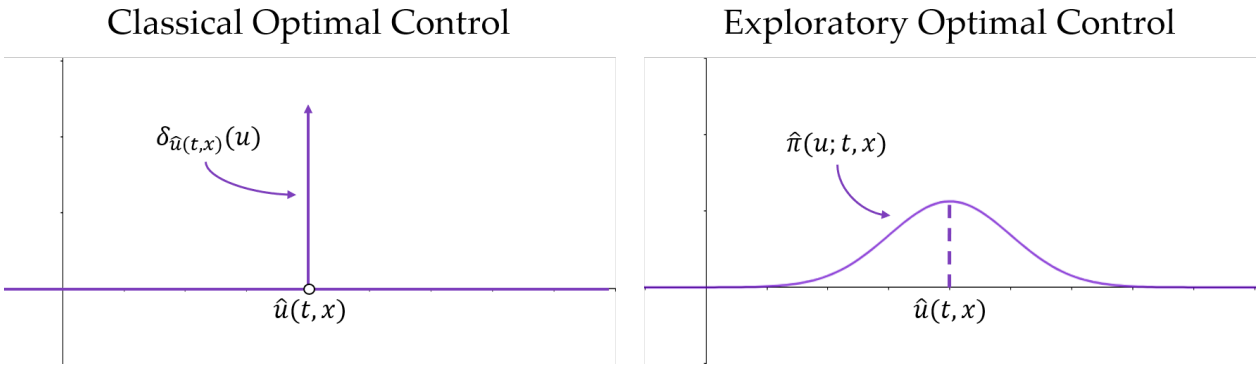

Figure 2: Desirable policies in the classical vs exploratory stochastic control problems. Left: The optimal *classical control* can be viewed as a Dirac measure centered at the optimal control $\hat{\boldsymbol{u}}$. Right: The optimal *exploratory policy* is a distribution which allows the agent to sample sub-optimal controls.

However, by including an exploration reward, there are now two possibly competing terms in the problem's reward function: the original objective whose value we would ultimately like to maximise through the exploitation of 'good' controls, and the exploration reward term which incentivises exploration to promote learning. This introduces a trade-off in the objective and it is not immediately clear how one should balance the two terms. Thus, we will vary this level of trade-off between exploration and exploitation by multiplying the exploration reward by a parameter λ known as the **exploration rate**. Combining the ideas of relaxed controls and exploration rewards, we define the following:

Definition 3.2. *The **exploration reward** for a given policy/control distribution π is defined as:*

$$\mathcal{X}[\mathcal{R},\pi] \coloneqq \int_U \mathcal{R}[\pi(u)]du,\ \pi \in \mathcal{P}(U). \tag{32}$$

Definition 3.3. *The **exploration-rewarded expected utility/value function** for a given policy π under the exploratory state dynamics* (27) *is:*

$$V^{\pi}[\mathcal{R}, x] \coloneqq \mathbb{E}\left[\int_0^{\infty} e^{-\rho t}(\tilde{r}(X_t^{\pi}, \pi_t) + \lambda \mathcal{X}[\mathcal{R}, \pi_t])dt \;\middle|\; X_0^{\pi} = x\right]. \tag{33}$$

Definition 3.4. *The **Exploration-Rewarded, Relaxed Stochastic Control (XRSC) Problem** is to find the optimal value function and policy:*

$$V[\mathcal{R}, x] \coloneqq \sup_{\pi \in \mathcal{U}(x)} V^{\pi}[\mathcal{R}, x] = \sup_{\pi \in \mathcal{U}(x)} \mathbb{E}\left[\int_0^{\infty} e^{-\rho t}(\tilde{r}(X_t^{\pi}, \pi_t) + \lambda \mathcal{X}[\mathcal{R}, \pi_t])dt \;\middle|\; X_0^{\pi} = x\right]. \tag{34}$$

We now prescribe a minimal but non-exhaustive set of conditions for a policy to be admissible. A policy π is said to be **admissible for the XRSC problem** ($\pi \in \mathcal{U}(x)$) if:

1. $\pi_t \in \mathcal{P}(U)\ a.s.\ \forall\ t \geq 0$;
2. $\left\{\int_A \pi_t(u)du, t \geq 0\right\}$ is $\mathcal{F}_t$-progressively measurable for each A in the Borel sigma algebra on U denoted $\mathcal{B}(U)$;
3. The drift and volatility coefficients $\tilde{b}, \tilde{\sigma}$ are Lipschitz and grow at most linearly in the state variable, so that the exploratory SDE (27) has a unique strong solution X^{π} under the policy π;
4. The expectation in the optimal value function (34) is finite.

Before we derive an expression for candidate optimal policies, it will be useful to distinguish between **open-loop** and **closed-loop** (or **feedback**) policies (control distributions):

Definition 3.5. *$\boldsymbol{\pi}(\cdot; \cdot, \cdot)$ is called an **admissible feedback policy** if*

1. *$\boldsymbol{\pi}(\cdot; t, x)$ is a probability density function on U for each $(t, x) \in [0, \infty) \times \mathbb{R}$.*
2. *For each $(s, y) \in [0, \infty) \times \mathbb{R}$, the SDE*

$$dX_t^{\boldsymbol{\pi}} = \tilde{b}(\boldsymbol{\pi}(\cdot; t, X_t^{\boldsymbol{\pi}}))dt + \tilde{\sigma}(\boldsymbol{\pi}(\cdot; t, X_t^{\boldsymbol{\pi}}))dW_t,\ t \in [s, \infty);\ X_s^{\boldsymbol{\pi}} = y,$$

has a unique strong solution $\{X_t^{\boldsymbol{\pi}}, s \leq t\}$ and the open-loop policy $\pi = \{\pi_t, s \leq t\} = \{\boldsymbol{\pi}(\cdot; t, X_t^{\boldsymbol{\pi}}), s \leq t\}$ is admissible under the four conditions stated above.

In other words, the open-loop policy $\pi_t \coloneqq \boldsymbol{\pi}(\cdot; t, X_t^{\boldsymbol{\pi}})$ is the process generated from the feedback policy with respect to the initial conditions (s, y); it is an *adapted process* independent of the current state. In contrast, the feedback policy is dependent on both the time and the current state.

3.1.4 Entropy-Regularised Relaxed Stochastic Control

We will now consider a specific choice of functional $\mathcal{R}$ for our problem. Our choice for this ultimately comes down to the question: *what is a good method for rewarding exploration?* One intuitive method is to incentivise *uncertainty* about the value of the control which is sampled, so that the likelihood of always exploiting by selecting the optimal control is lowered. In the context of policies/density functions, pure exploitation would correspond to selecting the optimal control with probability 1 (see figure 2), giving us minimal uncertainty about the value of our sampled control. Fortunately, information theory gives us a useful measure of such uncertainty via Shannon's **differential entropy** [10]:

$$\mathcal{H}(\pi) \coloneqq -\int_U \pi(u) \ln \pi(u) du,\ \pi \in \mathcal{P}(U). \tag{35}$$

This definition aligns with our intuition, as the differential entropy of a distribution gives a measure of how much 'uncertainty' there is in the value of a sample from drawn from the distribution (see figure 3).

Using this as our exploration reward, we choose $\mathcal{R}[\pi(u)] = -\pi(u)\ln\pi(u)$ which corresponds to the widely used technique known as **entropy-regularisation** [6, 7, 32].

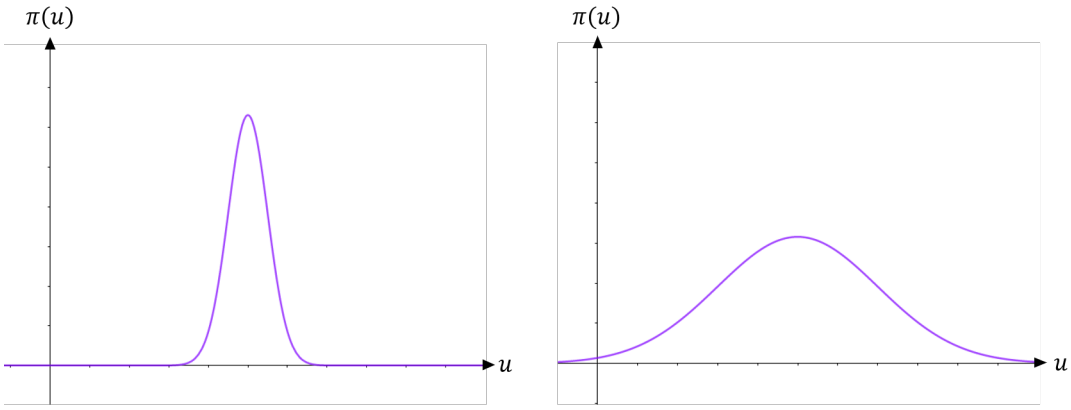

Figure 3: Illustration of the intuition behind differential entropy. Left: Narrower density function with lower uncertainty in sampled controls. Right: Wider density functions with higher uncertainty in sampled controls.

Selecting the differential entropy as our exploration reward functional, we thus define the central problem which we wish to investigate in this essay:

Definition 3.6. *The **entropy-regularised value function** for a given policy π under the exploratory state dynamics* (27) *is:*

$$V^\pi(x) \coloneqq \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\tilde{r}(X_t^\pi, \pi_t) + \lambda\mathcal{H}(\pi_t)\right) dt \,\middle|\, X_0^\pi = x\right]. \tag{36}$$

Definition 3.7. *The **Entropy-Regularised, Relaxed Stochastic Control** (ERSC) problem is to find the optimal value function and optimal policy:*

$$V(x) \coloneqq \sup_{\pi\in\mathcal{U}(x)} V^\pi(x) = \sup_{\pi\in\mathcal{U}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\tilde{r}(X_t^\pi, \pi_t) + \lambda\mathcal{H}(\pi_t)\right) dt \,\middle|\, X_0^\pi = x\right]. \tag{37}$$

3.2 Candidate Optimal Policies for Exploratory Control Problems

Here, we present brief derivations of candidate optimal policies using Dynamic Programming for the general exploratory control problems outlined in this chapter .

3.2.1 Candidate Optimal Policy for the XRSC Problem

For the exploration-rewarded, relaxed stochastic control (XSRC) problem (34), we can specify the form that a candidate optimal policy must take using the technique of Dynamic Programming we saw earlier. To this end, the HJB equation for this problem can be written in integral form as

$$\rho v(x) = \max_{\pi\in\mathcal{P}(U)} \int_U \left[\left(r(x,u) + \frac{1}{2}\sigma^2(x,u)v''(x) + b(x,u)v'(x)\right)\pi(u) + \lambda\mathcal{R}[\pi(u)]\right] du. \tag{38}$$

Applying the method of Lagrange multipliers in the variational setting to the constraint $\int_U \pi(u)du = 1$, the Euler-Lagrange equation gives a necessary condition for a maximising density function:

$$\frac{\delta L}{\delta \pi} = \alpha(x,u) - \lambda\left(\frac{\delta \mathcal{R}}{\delta \pi}[\hat{\pi}(u;x)]\right) + \mu = 0. \tag{39}$$

Supposing that the inverse of such a functional derivative exists, we get a candidate for an optimal feedback policy:

$$\hat{\boldsymbol{\pi}}(u;x) = \left(\frac{\delta \mathcal{R}}{\delta \pi}\right)^{-1}\left[\frac{\alpha(x,u)+\mu}{\lambda}\right]. \tag{40}$$

We emphasise that in general, the HJB equation for stochastic problems and the Euler-Lagrange equation only give a *necessary* set of conditions for optimality, and thus we refer to the policy (40) as a **candidate** optimal policy. For specific applications, it must be verified that the feedback policy (40) given above is indeed optimal and leads to the optimal value function.

3.2.2 Candidate Optimal Policy for the ERSC Problem

In a similar way, we can find a candidate optimal control for the ERSC problem by carrying out the same procedure as seen above. Due to the specification of our exploration reward function $\mathcal{R}$ as the differential entropy though, we can be more explicit about the form of the optimal policy.

To find a candidate optimal policy, we can apply Dynamic Programming to find the maximiser of the ERSC problem's HJB equation. Bellman's principle of optimality states that for any $s > 0$, the optimal value function satisfies

$$V(x) = \sup_{\pi \in \mathcal{U}(x)} \mathbb{E}\left[\int_0^s e^{-\rho t}\left(\tilde{r}(X_t^\pi, \pi_t) + \lambda \mathcal{H}(\pi_t)\right) dt + e^{-\rho s} V(X_s^\pi) \,\middle|\, X_0^\pi = x\right].$$

The associated HJB equation is given by

$$\rho v(x) = \max_{\pi \in \mathcal{P}(U)}\left(\tilde{r}(x,\pi) + \lambda \mathcal{H}(\pi) + \frac{1}{2}\tilde{\sigma}^2(x,\pi)v''(x) + \tilde{b}(x,\pi)v'(x)\right),\ x \in \mathbb{R}, \tag{41}$$

which can be written in integral form as

$$\rho v(x) = \max_{\pi \in \mathcal{P}(U)} \int_U \left(r(x,u) - \lambda \ln \pi(u) + \frac{1}{2}\sigma^2(x,u)v''(x) + b(x,u)v'(x)\right)\pi(u)\, du. \tag{42}$$

We can solve the right hand side of equation (42) with the constraint $\int_U \pi(u;x)du = 1$ for a maximising function using the variational calculus with a Lagrange multiplier μ. Firstly, we define

$$\alpha(x,u) \coloneqq r(x,u) + \frac{1}{2}\sigma^2(x,u)v''(x) + b(x,u)v'(x).$$

Carrying out the variation, we get the following necessary condition for the maximising function:

$$\frac{\delta L}{\delta \pi(u;x)} = \alpha(x,u) - \lambda(\ln(\hat{\boldsymbol{\pi}}(u;x)) + 1) + \mu = 0.$$

Solving for $\hat{\pi}$ gives a general solution of the feedback form

$$\hat{\boldsymbol{\pi}}(u;x) = \frac{\exp\left(\frac{\alpha(x,u)}{\lambda}\right)}{\exp\left(1 - \frac{\mu}{\lambda}\right)}.$$

Using the constraint that the integral of $\hat{\pi}$ over U be equal to 1, we get that

$$\exp\left(1-\frac{\mu}{\lambda}\right)=\int_U \exp\left(\frac{\alpha(x,u)}{\lambda}\right)du,$$

which gives our optimal feedback distribution

$$\hat{\boldsymbol{\pi}}(u;x)=\frac{\exp\left(\frac{\alpha(x,u)}{\lambda}\right)}{\int_U \exp\left(\frac{\alpha(x,u)}{\lambda}\right)du}. \tag{43}$$

An open-loop control distribution $\hat{\pi}_t$ can be generated from this feedback control distribution by simply substituting in $\hat{X}_t$ where $\{\hat{X}_t, t\geq 0\}$ solves the exploratory SDE (27) when $\hat{\pi}$ is applied.

3.3 Exploratory Policy Improvement Theorem

We conclude this introduction to exploratory optimal control with a theorem which will be essential in designing reinforcement learning algorithms for specific applications of the theory. In its essence, the result states that given an arbitrary admissible feedback policy $\boldsymbol{\pi}$ and under mild conditions on the exploration reward functional $\mathcal{R}$ and the associated value function $V^{\boldsymbol{\pi}}[\mathcal{R},\cdot]$, we can generate a new policy whose value function is *at least as good as the given policy*:

Theorem 3.1 (Exploration-Rewarded Policy Improvement Theorem)**.** *Let $\boldsymbol{\pi}=\boldsymbol{\pi}(\cdot;\cdot)$ be an arbitrary admissible feedback control policy for the XRSC problem* (34)*. Suppose that:*

1. *For the given functional $\mathcal{R}[\cdot]$, the value function satisfies $V^{\boldsymbol{\pi}}[\mathcal{R},\cdot]\in C^2(\mathbb{R})$;*
2. *The functional derivative $\left(\frac{\delta\mathcal{R}}{\delta\pi}\right)[\pi(u)]$ is continuously differentiable with respect to π for all $\pi\in\mathcal{P}(U), u\in U$;*
3. *The feedback policy $\tilde{\pi}$ defined by*

$$\tilde{\boldsymbol{\pi}}(u;x)\coloneqq\left(\frac{\delta\mathcal{R}}{\delta\pi}\right)^{-1}\left[\frac{\alpha(x,u)+\mu}{\lambda}\right], \tag{44}$$

   where $\alpha(x,u)=r(x,u)+\frac{1}{2}\sigma^2(x,u)V^{\boldsymbol{\pi}}_{xx}[\mathcal{R},x]+b(x,u)V^{\boldsymbol{\pi}}_x[\mathcal{R},x]$, is admissible, and μ is the Lagrange multiplier for the right hand side of (38)*;*
4. *The value function applied to the state process $X^{\tilde{\pi}}_t$ induced by applying the open-loop strategy generated from the feedback policy $\tilde{\pi}(u;x)$ satisfies $\lim_{t\to\infty}e^{-\rho t}\mathbb{E}(V^{\boldsymbol{\pi}}[\mathcal{R},X^{\tilde{\pi}}_t])=0$.*

Then,

$$V^{\tilde{\boldsymbol{\pi}}}[\mathcal{R},x]\geq V^{\boldsymbol{\pi}}[\mathcal{R},x],\ \textit{for all } x\in\mathbb{R}. \tag{45}$$

Proof: See Appendix A.

4 Exploratory Linear Quadratic (ELQ) Problem

We now consider an application of the developed theory so far to a well-known problem in optimal control: the **Linear Quadratic (LQ) control problem**. For this class of problems, the drift and volatility terms are linear in the state and control variables while the reward function is quadratic in these variables. Thus, we let

$$b(x,u) = Ax + Bu, \;\; \sigma(x,u) = Cx + Du, \tag{46}$$

$$r(x,u) = -\left(\frac{M}{2}x^2 + Rxu + \frac{N}{2}u^2 + Px + Qu\right) \quad x \in \mathbb{R},\; u \in U, \tag{47}$$

for $A, B, C, D, R, P, Q \in \mathbb{R},\; M \geq 0,\; N > 0$. In this scenario we let our controls take on any real values and so we set $U = \mathbb{R}$.

4.1 ELQ Problem Setup

For a fixed initial state $x \in \mathbb{R}$, we perform control relaxation following the example in section (3.2) and denote the **mean and variance processes** $\mu_t,\; \sigma_t^2, t \geq 0$ for an open loop control $\pi \in \mathcal{U}(x)$ respectively by

$$\mu_t \coloneqq \int_{\mathbb{R}} u\pi_t(u)du, \;\; \sigma_t^2 \coloneqq \int_{\mathbb{R}} u^2\pi_t(u)du - \mu_t^2. \tag{48}$$

Using our definitions of exploratory drift (28) and exploratory volatility (29), we can express the exploratory SDE (27) in terms of the mean and variance processes:

$$\begin{aligned}
dX_t^\pi &= \left(\int_{\mathbb{R}} b(X_t^\pi, u)\pi_t(u)du\right)dt + \left(\sqrt{\int_{\mathbb{R}} \sigma^2(X_t^\pi, u)\pi_t(u)du}\right)dW_t \\
&= \left(\int_{\mathbb{R}} (AX_t^\pi + Bu)\pi_t(u)du\right)dt + \left(\sqrt{\int_{\mathbb{R}} (CX_t^\pi + Du)^2\pi_t(u)du}\right)dW_t \\
&= (AX_t^\pi + B\mu_t)dt + \sqrt{C^2(X_t^\pi)^2 + 2CDX_t^\pi\mu_t + D^2(\sigma_t^2 + \mu_t^2)}dW_t,
\end{aligned}$$

using the fact that $\pi_t \in \mathcal{P}(U)$ so its integral over $\mathbb{R}$ is 1. We can complete the square on the expression under the square root to give the **LQ exploratory state dynamics**:

$$dX_t^\pi = (AX_t^\pi + B\mu_t)dt + \sqrt{(CX_t^\pi + D\mu_t)^2 + D^2\sigma_t^2}dW_t, \;\; X_0^\pi = x, t > 0. \tag{49}$$

For this exploratory LQ problem, we will explicitly state the set of admissible controls. Let us denote

$$L(X_t^\pi, \pi_t) \coloneqq \int_{\mathbb{R}} r(X_t^\pi, u)\pi_t(u)du + \lambda\mathcal{H}(\pi_t).$$

A policy is **admissible for the Exploratory LQ problem** ($\pi \in \mathcal{A}(x)$), if

1. $\pi_t \in \mathcal{P}(\mathbb{R})\; a.s.\; \forall\, t \geq 0$;
2. $\{\int_A \pi_t(u)du, t \geq 0\}$ is $\mathcal{F}_t$-progressively measurable for each $A \in \mathcal{B}(\mathbb{R})$;
3. $\mathbb{E}\left[\int_0^t (\mu_s^2 + \sigma_s^2)ds\right] < \infty\; \forall t \geq 0$;
4. $\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}[(X_T^\pi)^2] = 0$;
5. $\mathbb{E}[\int_0^\infty e^{-\rho t}|L(X_t^\pi, \pi_t)|dt] < \infty$,

where $\{X_t^\pi, t \geq 0\}$ solves the exploratory SDE (49). This is to ensure that a unique, strong solution to the exploratory SDE exists under any admissible policy and that Dynamic Programming and the verification theorem are applicable to the problem.

Definition 4.1. *The **Exploratory LQ problem** is to find the optimal value function and policy:*

$$V(x) = \sup_{\pi \in \mathcal{A}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\tilde{r}(X_t^\pi, \pi_t) + \lambda \mathcal{H}(\pi_t)\right) dt \;\middle|\; X_0^\pi = x\right], \tag{50}$$

where

$$\tilde{r}(X_t^\pi, \pi_t) = \int_{\mathbb{R}} r(X_t^\pi, u)\pi_t(u)du, \tag{51}$$

$$\mathcal{H}(\pi_t) = -\int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du. \tag{52}$$

4.2 Solution to the ELQ Problem

4.2.1 Solving for the optimal policy

Since we are working with a *stronger* set of assumptions than in section (3.2), we can apply the results derived there to this problem to give an optimal feedback distribution from (43):

$$\hat{\boldsymbol{\pi}}(u; x) = \frac{\exp\left(\frac{\alpha(x,u)}{\lambda}\right)}{\int_{\mathbb{R}} \exp\left(\frac{\alpha(x,u)}{\lambda}\right) du},$$

where

$$\alpha(x,u) = -\left(\frac{M}{2}x^2 + Rxu + \frac{N}{2}u^2 + Px + Qu\right) + \frac{1}{2}(Cx + Du)^2 v''(x) + (Ax + Bu)v'(x).$$

It turns out that this expression factorises nicely to give

$$\alpha(x,u) = -\left(u - \frac{CDxv''(x) + Bv'(x) - Rx - Q}{N - D^2 v''(x)}\right)^2 \Bigg/ \frac{2}{N - D^2 v''(x)}, \tag{53}$$

and therefore, our optimal feedback distribution is

$$\hat{\pi}(u; x) = \frac{\exp\left(-\left(u - \frac{CDxv''(x)+Bv'(x)-Rx-Q}{N-D^2v''(x)}\right)^2 \Big/ \frac{2\lambda}{N-D^2v''(x)}\right)}{\int_{\mathbb{R}} \exp\left(\left(u - \frac{CDxv''(x)+Bv'(x)-Rx-Q}{N-D^2v''(x)}\right)^2 \Big/ \frac{2\lambda}{N-D^2v''(x)}\right) du}. \tag{54}$$

We observe that $\hat{\pi}(u; x)$ satisfies all the conditions for a probability density function and that it takes the form of a Gaussian distribution. We can thus write, using standard notation for Gaussian distributions,

$$\hat{\pi}(u; x) = \mathcal{N}\left(u \mid \hat{\mu}(x), \hat{\sigma}^2(x)\right), \tag{55}$$

$$\hat{\mu}(x) = \frac{CDxv''(x) + Bv'(x) - Rx - Q}{N - D^2 v''(x)}, \tag{56}$$

$$\hat{\sigma}^2(x) = \frac{\lambda}{N - D^2 v''(x)}. \tag{57}$$

4.2.2 Solving for the value function

Now that we have a candidate for the optimal distribution (55), we can substitute it into the integral form of the HJB equation (42) to get

$$\rho v(x) = \int_{\mathbb{R}} \left(r(x,u) - \lambda \ln \hat{\pi}(u;x) + \frac{1}{2}\sigma^2(x,u)v''(x) + b(x,u)v'(x) \right) \hat{\pi}(u;x)\, du.$$

After evaluation of the individual integrals and simplification, we get

$$\begin{aligned} \rho v(x) =& \frac{1}{2}(C^2 v''(x) - M)x^2 + (Av'(x) - P)x \\ &+ \frac{(CDxv''(x) + Bv'(x) - Rx - Q)^2}{2(N - D^2 v''(x))} + \frac{\lambda}{2}\left(\ln\left(\frac{2\pi e \lambda}{N - D^2 v''(x)} \right) - 1 \right). \end{aligned} \tag{58}$$

Following the example from the general discussion of SLQ problems in section (2.2), we make the ansatz that the value function $v(x)$ is quadratic. In order, to make this ansatz work, we require two additional assumptions to ensure that the HJB equation yields a solution which is optimal:

$$\rho > 2A + C^2 + \max\left(\frac{D^2R^2 - 2NR(B+CD)}{N}, 0 \right), \tag{59}$$

$$R^2 < MN. \tag{60}$$

Under these assumptions, we propose a smooth value function which solves the HJB equation (58):

$$v(x) = \frac{1}{2}k_2 x^2 + k_1 x + k_0,$$

for some constants k_2, k_1, k_0. Substituting this ansatz into (58) and comparing coefficients of quadratic, linear, and constant terms in x yields the system of algebraic equations

$$\rho k_2 = \frac{(k_2(B+CD) - R)^2}{N - k_2 D^2} + k_2(2A + C^2) - M, \tag{61}$$

$$\rho k_1 = \frac{(k_1 B - Q)(k_2(B+CD) - R)}{N - k_2 D^2} + k_1 A - P, \tag{62}$$

$$\rho k_0 = \frac{(k_1 B - Q)^2}{2(N - k_2 D^2)} + \frac{\lambda}{2}\left(\ln\left(\frac{2\pi e \lambda}{N - k_2 D^2} \right) - 1 \right). \tag{63}$$

This system of equations has two solutions since the equation for k_2 is a quadratic. Noting that the reward function r is concave, we want to look for concave solutions $v(x)$. Under assumptions (59) and (60), there is a solution such that $k_2 < 0$ which leads to the desired concave $v(x)$. If we define

$$E = \rho - (2A + C^2),\ F = B + CD,$$

we get the following solutions for k_2, k_1, k_0:

$$k_2 = \frac{1}{2}\left(\frac{EN + 2FR - D^2M - \sqrt{(EN + 2FR - D^2M)^2 - 4(F^2 + ED^2)(R^2 - MN)}}{F^2 + ED^2} \right), \tag{64}$$

$$k_1 = \frac{P(N - k_2 D^2) - QR}{k_2 BF + (A - \rho)(N - k_2 D^2) - BR}, \tag{65}$$

$$k_0 = \frac{(k_1 B - Q)^2}{2\rho(N - k_2 D^2)} + \frac{\lambda}{2\rho}\left(\ln\left(\frac{2\pi e \lambda}{N - k_2 D^2} \right) - 1 \right). \tag{66}$$

With this, we have all the pieces to fully state the solution to the ELQ problem.

Theorem 4.1 (Solution to the ELQ Problem)**.** *Suppose the reward function for the LQ case is given by* (47) *and assumptions* (59) *and* (60) *hold. Then, the value function of the LQ problem is*

$$V(x) = \frac{1}{2}k_2x^2 + k_1x + k_0,$$

where k_2, k_1, k_0 are given by (64),(65)*, and* (66) *respectively.*
Moreover, the optimal feedback control distribution is given by

$$\hat{\boldsymbol{\pi}}(u;x) = \mathcal{N}\left(u \mid \frac{(k_2(B+CD)-R)x + k_1B - Q}{N - k_2D^2}, \frac{\lambda}{N-k_2D^2}\right).$$

Finally, the optimal state process $\{\hat{X}_t, t \geq 0\}$ under $\hat{\pi}(u;x)$ is the unique solution of the SDE

$$\begin{cases} d\hat{X}_t = & \left(\left(A + \frac{B(k_2F-R)}{N-k_2D^2}\right)\hat{X}_t + \frac{B(k_1B-Q)}{N-k_2D^2}\right)dt + \\ & \sqrt{\left(\left(C + \frac{D(k_2F-R)}{N-k_2D^2}\right)\hat{X}_t + \frac{D(k_1B-Q)}{N-k_2D^2}\right)^2 + \frac{\lambda D^2}{N-k_2D^2}}dW_t \\ \hat{X}_0 = x. \end{cases}$$

Proof: See Appendix B.

4.3 Example: Case of the State-Independent Reward

Here, we will look at a specific case of the general exploratory LQ problem formulated above, where the reward function is independent of the state. Mathematically, the reward function takes on the form

$$r(u) = -\left(\frac{N}{2}u^2 + Qu\right).$$

We can treat this derivation as a simplified version of the general case with but with $M = R = P = 0$. The optimal value function for the state equation solves the HJB equation

$$\rho v(x) = \frac{(CDxv''(x) + Bv'(x) - Q)^2}{2(N - D^2v''(x))} + \frac{\lambda}{2}\left(\ln\left(\frac{2\pi e\lambda}{N - D^2v''(x)}\right) - 1\right) + \frac{1}{2}x^2C^2v''(x) + Axv'(x). \tag{67}$$

This equation has many possible solutions. Unlike the general state-dependent case however, one possible solution is when $v(x)$ is a constant:

$$v(x) = \frac{Q^2}{2\rho N} + \frac{\lambda}{2\rho}\left(\ln\frac{2\pi e\lambda}{N} - 1\right). \tag{68}$$

Using this form, all of the derivatives of v in the optimal policy are eliminated, giving simply

$$\hat{\boldsymbol{\pi}}(u;x) = \frac{\exp\left(-\left(u + \frac{Q}{N}\right)^2 / \frac{2\lambda}{N}\right)}{\int_{\mathbb{R}}\exp\left(-\left(u + \frac{Q}{N}\right)^2 / \frac{2\lambda}{N}\right)du}.$$

It turns out that this $v(x)$ is indeed the value function for the case of the state-independent reward and this is summarised in the following theorem:

Theorem 4.2 (Solution to the State-Independent ELQ Problem)**.** *If $r(x,u) = -(\frac{N}{2}u^2 + Qu)$, then the value function of the ERSC is*

$$V(x) = \frac{Q^2}{2\rho N} + \frac{\lambda}{2\rho}\left(\ln\frac{2\pi e\lambda}{N} - 1\right), \tag{69}$$

the optimal feedback policy is

$$\hat{\boldsymbol{\pi}}(u;x) = \mathcal{N}\left(u \,\middle|\, -\frac{Q}{N}, \frac{\lambda}{N}\right), \tag{70}$$

and the optimal state process $\{\hat{X}_t\}, t \geq 0$ under $\hat{\pi}$ is the unique solution of

$$d\hat{X}_t = \left(A\hat{X}_t - \frac{BQ}{N}\right)dt + \sqrt{\left(C\hat{X}_t - \frac{DQ}{N}\right)^2 + \frac{\lambda D^2}{N}}dW_t,\ \hat{X}_0 = x. \tag{71}$$

4.3.1 Proof (sketch):

As before in the state-dependent case, we need to verify that the ansatz $v(x) =$ constant as defined in (68) is indeed the value function of the problem and that the optimal feedback control distribution is admissible.

Firstly, we note that by the way we derived the optimal control distribution,

$$\rho v = \int_{\mathbb{R}} -\left(\frac{N}{2}u^2 + Qu + \lambda\ln\hat{\pi}_t(u)\right)\hat{\pi}_t(u)du.$$

Now, for any $\pi \in \mathcal{A}(x)$ and $T \geq 0$, the HJB equation (41) and the sub-optimality of our distribution π in general tells us that

$$\begin{aligned}
e^{-\rho T}v &= v - \int_0^T e^{-\rho t}\rho v dt \\
&= v - \mathbb{E}\left[\int_0^T e^{-\rho t}\left(\int_{\mathbb{R}} -\left(\frac{N}{2}u^2 + Qu + \lambda\ln\hat{\pi}_t(u)\right)\hat{\pi}_t(u)du\right)dt\right] \\
&\leq v - \mathbb{E}\left[\int_0^T e^{-\rho t}\left(\int_{\mathbb{R}} -\left(\frac{N}{2}u^2 + Qu + \lambda\ln\pi_t(u)\right)\pi_t(u)du\right)dt\right].
\end{aligned}$$

Using condition 5 (since $\pi \in \mathcal{A}(x)$) to apply the dominated convergence theorem to send the left hand side to 0 as $T \to \infty$ and rearranging yields

$$v = V(x)\ \forall x \in \mathbb{R}.$$

The admissibility of the open-loop control generated by $\hat{\pi}$ can be verified by observing that it satisfies the conditions stated previously. Specifically, conditions 1-3 are straightforward to verify, condition 4 is not necessary to verify as the state process does not appear in $\hat{\pi}$ and condition 5 can be verified by evaluating the integrals using the properties of distribution functions to deduce that $L(X_t^\pi, \pi_t)$ is constant and so $\mathbb{E}\left[\int_0^\infty e^{-\rho t}|L(X_t^\pi,\pi_t)|dt\right] < \infty$. The exploratory state dynamics (71) then follow by substituting $\hat{\mu}_t = \frac{-Q}{N}$, $(\hat{\sigma}_t)^2 = \frac{\lambda}{N}$ into the general exploratory SDE (49).

□

4.3.2 Deriving explicit solutions to the optimal state SDEs

Here, we derive explicit solutions for the optimal state SDE in the simpler state-independent case. To find explicit solutions to (71), we can consider certain cases depending on the values of the coefficients and the initial state:

1. $D = 0, x \geq 0, BQ \leq 0$: This simplifies the coefficient of the local martingale term so we can solve this directly to give

$$\hat{X}_t = xe^{\left(A-\frac{C^2}{2}\right)t+|C|W_t} - \frac{BQ}{N}\int_0^t e^x e^{\left(A-\frac{C^2}{2}\right)(t-s)+|C|(W_t-W_s)}ds.$$

2. $D = 0, x \leq 0, BQ \geq 0$: As above, we can solve the simplified SDE to give

$$\hat{X}_t = xe^{\left(A-\frac{C^2}{2}\right)t-|C|W_t} - \frac{BQ}{N}\int_0^t e^x e^{\left(A-\frac{C^2}{2}\right)(t-s)-|C|(W_t-W_s)}ds.$$

3. $C = 0, A \neq 0$: This removes the dependence of the local martingale term on the state so in a similar manner, we can solve this to give

$$\hat{X}_t = xe^{At} - \frac{BQ}{AN}(1-e^{At}) + \frac{|D|}{N}\sqrt{Q^2+\lambda N}\int_0^t e^{A(t-s)}dW_s,\ t \geq 0.$$

4. $C = 0, A = 0$: This removes the dependence of the dynamics on the state completely which simply gives

$$\hat{X}_t = x - \frac{BQ}{N}t + \frac{|D|}{N}\sqrt{Q^2+\lambda N}W_t,\ t \geq 0.$$

5. $C \neq 0, D \neq 0$: This case requires a bit more work. Noting that the coefficient of the dW_t term in the SDE (71) is C^2 and both derivatives are bounded, we can apply a technique known as the Doss-Sussman transformation which uses the ansatz $\hat{X}(\omega) = F(W_t(\omega), Y_t(\omega)),\ t \geq 0,\ \omega \in \Omega$, for a deterministic function F and an adapted deterministic process $Y(t)$ with random initial condition. Expanding this ansatz using the multi-dimensional version of Itô's formula and comparing it with the SDE (71), we can derive an ODE for F which can be solved to give

$$\hat{X}_t(\omega) = F(z,y) = \sqrt{\frac{\lambda}{N}}\left|\frac{D}{C}\right|\sinh\left(|C|z + \sinh^{-1}\left(\sqrt{\frac{N}{\lambda}}\left|\frac{C}{D}\right|\left(y - \frac{DQ}{CN}\right)\right)\right) + \frac{DQ}{CN},$$
$$W_t(\omega) = z,\ Y_t(\omega) = y,\ \omega \in \Omega.$$

4.3.3 Interpretation of the mean and variance of the optimal policy

In both of these cases, the mean of the optimal policy does not depend on the exploration rate λ. Due to the fact that Gaussian distributions have the highest density at their mean values, this suggests that the agent should be more likely to sample controls closer to its estimate of the optimal control than others, regardless of its propensity towards exploration. Furthermore, the fact that the exploration rate λ appears in the variance provides a good way to interpret the optimal policy, where the mean is centered around the optimal classical control and the variance is characterised by the amount of exploration the agent wishes to undertake.

4.4 The Cost and Effect of Exploration

In this section, we quantify the cost and effect of modifying the classical stochastic control problem to include exploration. To achieve this, we need a reference frame to compare the model against. Naturally, we will use

the classical stochastic LQ control problem as this reference frame. Here, we will carry out the derivation of the solution using Dynamic Programming directly rather than using the results in the general treatment of the SLQ problem from 2.2 due to the relative ease of solving this problem in 1-dimension and to highlight the similarities between the exploratory and classical formulations, which will be useful in formulating some results later on.

4.4.1 Classical LQ problem

We begin by formulating the classical LQ problem in this context. For a standard Brownian motion $\{W_t, t \geq 0\}$ defined on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$, the controlled state process $\{x_t^u, t \geq 0\}$ solves:

$$dx_t^u = (Ax_t^u + Bu_t)dt + (Cx_t^u + Du_t)dW_t,\ t \geq 0,\ x_0^u = x, \tag{72}$$

with $A, B, C, D \in \mathbb{R}$. The classical optimal value function as in equation (26) is

$$V^{cl}(x) \coloneqq \sup_{u \in \mathcal{U}^{cl}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\frac{M}{2}x^2 + Rxu + \frac{N}{2}u^2 + Px + Qu\right)dt \,\middle|\, x_0 = x\right]. \tag{73}$$

The HJB equation for this problem is

$$\begin{aligned}\rho w(x) = \max_{u\in\mathbb{R}}\Big(-\frac{1}{2}(N - D^2)w''(x)u^2 + (CDxw''(x) + Bw'(x) - Rx - Q)u\Big) \\ + \frac{1}{2}(C^2w''(x) - M)x^2 + (Aw'(x) - P)x.\end{aligned} \tag{74}$$

Setting the u-derivative of the term in brackets to 0, we get a candidate optimal control

$$\hat{\boldsymbol{u}}(x) = \frac{CDxw''(x) + Bw'(x) - Rx - Q}{N - D^2w''(x)},\ x \in \mathbb{R}. \tag{75}$$

Substituting this into the HJB equation yields a PDE for the value function for the classical LQ problem

$$\rho w(x) = \frac{(CDxw''(x) + Bw'(x) - Rx - Q)^2}{2(N - D^2w''(x))} + \frac{1}{2}(C^2w''(x) - M)x^2 + (Aw'(x) - P)x. \tag{76}$$

To solve this, we apply the same ansatz technique used in solving the exploratory LQ problem by letting $w(x) = \frac{1}{2}\alpha_2 x^2 + \alpha_1 x + \alpha_0$. Substituting the ansatz into the HJB equation yields the exact same algebraic equations for α_2, α_1 as was seen in the exploratory case, and the equation for the constant term is given by

$$\rho\alpha_0 = \frac{(\alpha_1 B - Q)^2}{2(N - \alpha_2 D^2)}.$$

Solving this system for the coefficients, we get

$$\alpha_2 = k_2,\ \alpha_1 = k_1,\ \alpha_0 = \frac{(\alpha_1 B - Q)^2}{2\rho(N - \alpha_2 D^2)}. \tag{77}$$

Finally, we take the necessary derivatives of the solution to the HJB equation w and substitute them into the feedback form of our minimising control (75) to give a candidate optimal control:

$$\hat{\boldsymbol{u}}(x) = \frac{(\alpha_2 F - R)x + \alpha_1 B - Q}{N - \alpha_2 D^2}. \tag{78}$$

It turns out that the verification theorem 2.1 is applicable in this case, so the optimal control is indeed given by (78) and the optimal value function is the solution w to the HJB equation (74) under the optimal control.

4.4.2 Solvability equivalence in the ELQ problem

Here we establish an important link between the classical and exploratory LQ problems by the result that if one problem is solvable, then so is the other. To this end, we fix x as the initial state of both the classical and exploratory LQ problems (73) and (37) respectively. Denoting the optimal open-loop controls $\hat{\pi}, \hat{u}$ and corresponding optimal state trajectories $\hat{X}, \hat{x}$, we have the state equations:

$$d\hat{X}_t = (A_1\hat{X}_t + A_2)dt + \sqrt{(B_1\hat{X}_t + B_2)^2 + C_1}dW_t, \tag{79}$$

$$d\hat{x}_t = (A_1\hat{x}_t + A_2)dt + (B_1\hat{x}_t + B_2)dW_t, \tag{80}$$

where

$$A_1 \coloneqq A + \frac{B(\alpha_2 F - R)}{N - \alpha_2 D^2},\ A_2 \coloneqq \frac{B(\alpha_1 B - Q)}{N - \alpha_2 D^2},$$
$$B_1 \coloneqq C + \frac{D(\alpha_2 F - R)}{N - \alpha_2 D^2},\ B_2 \coloneqq \frac{D(\alpha_1 B - Q)}{N - \alpha_2 D^2}, C_1 \coloneqq \frac{\lambda D^2}{N - \alpha_2 D^2}.$$

The following lemma will be useful in the proof of the theorem on solvability equivalence:

Lemma 4.3 (Solvability Equivalence)**.** *The following statements hold:*

(i) $\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}[(\hat{X}_t)^2] = 0$ *if and only if* $\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}[(\hat{x}_t)^2] = 0$.

(ii) $\liminf_{T\to\infty} \mathbb{E}[\int_0^\infty e^{-\rho t}(\hat{X}_t)^2 dt] < \infty$ *if and only if* $\liminf_{T\to\infty} \mathbb{E}[\int_0^\infty e^{-\rho t}(\hat{x}_t)^2 dt] < \infty$.

Proof of Lemma 4.3 (sketch): If we denote $n(t) \coloneqq \mathbb{E}[\hat{X}_t],\ t \geq 0$, we can integrate (79) using a similar stopping time argument as before, take expectations, and then take the time derivative to yield the ODE

$$\frac{dn(t)}{dt} = A_1 n(t) + A_2 \implies n(t) = \begin{cases} \left(x + \frac{A_2}{A_1}\right)e^{A_1 t} - \frac{A_2}{A_1}, & A_1 \neq 0 \\ x + A_2 t, & A_1 = 0. \end{cases}$$

If we employ the same argument for the classical case, we get $\mathbb{E}[\hat{x}_t] = n(t)$. Now, we compute the second moment of the state variables $m(t) \coloneqq \mathbb{E}[(\hat{X}_t)^2], \tilde{m}(t) \coloneqq \mathbb{E}[(\hat{x}_t)^2]$ in a similar fashion to yield the ODE's

$$\frac{dm(t)}{dt} = (2A_1 + B_1^2)m(t) + 2(A_2 + B_1 B_2)n(t) + B_2^2 + C_1,\ m(0) = x^2,$$
$$\frac{d\tilde{m}}{dt} = (2A_1 + B_1^2)\tilde{m}(t) + 2(A_2 + B_1 B_2)n(t) + B_2^2,\ \tilde{m}(0) = x^2.$$

Depending on the values of A_1 and B_1, we end up with 6 different possible combinations of solutions for $n(t), m(t), \tilde{m(t)}$. In all cases however, m and $\tilde{m}$ contain the *same* combination of the terms $e^{A_1 t}, e^{B_1^2 t}, e^{(2A_1 + B_1^2)t}, t, x^2$ with different coefficients due to the C_1 term. Since these two solutions only differ by their (constant) coefficients, we can conclude that the relations in (i) and (ii) both hold.

□

Theorem 4.4 (ELQ Solvability Equivalence)**.** *The following statements are equivalent:*

1. *For the exploratory LQ problem* (37)*, the value function and optimal feedback policy are given respectively by*

$$V(x) = v(x) = \frac{1}{2}\alpha_2 x^2 + \alpha_1 x + \alpha_0 + \frac{\lambda}{2\rho}\left(\ln\left(\frac{2\pi e\lambda}{N - \alpha_2 D^2}\right) - 1\right)m \tag{81}$$

$$\hat{\boldsymbol{\pi}}(u; x) = \mathcal{N}\left(u \,\middle|\, \frac{(\alpha_2 F - R)x + \alpha_1 B - Q}{N - \alpha_2 D^2},\ \frac{\lambda}{N - \alpha_2 D^2}\right). \tag{82}$$

For $\alpha_0, \alpha_1 \in \mathbb{R}$ *and* $\alpha_2 < 0$.

2. For the classical LQ problem (73)*, the value function and optimal feedback control are given respectively by*

$$V^{cl}(x) = w(x) = \frac{1}{2}\alpha_2 x^2 + \alpha_1 x + \alpha_0, \tag{83}$$

$$\hat{\boldsymbol{u}}(x) = \frac{(\alpha_2 F - R)x + \alpha_1 B - Q}{N - \alpha_2 D^2}. \tag{84}$$

Proof (sketch): We first note that when statement 1 holds, $v(x)$ satisfies the HJB equation (58) in the exploratory case. When statement 2 holds, $w(x)$ satisfies the HJB equation (74). Secondly, we can see from a comparison of the HJB equations (58) and (74) that if v in statement 1 solves (58), then w in statement 2 solves (74).

To complete the proof, we need to show that $\hat{\pi}$ is admissible in the exploratory problem if and only if $\hat{u}$ is admissible in the classical problem. Admissibility conditions 1-3 for the admissible set of exploratory LQ policies are straightforward to verify and Lemma 4.3 gives us equivalent conditions for the admissibility of $\hat{\pi}$ and $\hat{u}$ by showing that conditions 4 and 5 hold in one case if and only if they hold in the other. Therefore, statements 1 and 2 of Theorem 4.4 are equivalent.

□

4.4.3 The Cost of Exploration

The objective of this comparison is to examine the effect of formulating the problem in an exploratory manner, so the most direct way to quantify this effect is by taking the difference of the *accumulated rewards* in each problem. Since the exploratory formulation includes an entropy-regularisation term however, we subtract this term from the optimal value function in the exploratory case to get a direct comparison between the two accumulated rewards. Thus, the **exploration cost** is defined as

$$C^{\hat{u},\hat{\pi}}(x) \coloneqq V^{cl}(x) - \left(V(x) + \lambda \mathbb{E}\left[\int_0^\infty e^{-\rho t} \left(\int_U \hat{\pi}_t(u) \ln \hat{\pi}_t(u) du \right) dt \;\middle|\; \hat{X}_0^\pi = x \right] \right), \tag{85}$$

where $V(x), \hat{\pi}(u;x), V^{cl}(x), \hat{u}(x)$ are the optimal value function and optimal control of the exploratory and classical stochastic control problems respectively. This exploration cost measures the optimal accumulated reward as a result of including exploration in the entropy-regularised objective relative to the classical objective.

Theorem 4.5 (ELQ Exploration Cost)**.** *Assume that statement 1 or 2 of Theorem 4.4 holds. Then, the exploration cost for the ELQ problem is*

$$C^{\hat{u},\hat{\pi}}(x) = \frac{\lambda}{2\rho}, \;\; x \in \mathbb{R}.$$

Proof: We can evaluate the cost of exploration for the ELQ problem directly, where $V(x), \hat{\pi}(u;x), V^{cl}(x), \hat{u}(x)$ are given by (81)-(84) respectively. Firstly, we calculate the entropy term:

$$\begin{aligned}\int_0^\infty e^{-\rho t} \left(\int_{\mathbb{R}} \hat{\pi}_t(u) \ln \hat{\pi}_t(u) du \right) dt &= \int_0^\infty e^{-\rho t} \left(-\frac{1}{2} \ln \left(\frac{2\pi e \lambda}{N - \alpha_2 D^2} \right) \right) dt \\ &= -\frac{1}{2\rho} \ln \left(\frac{2\pi e \lambda}{N - \alpha_2 D^2} \right).\end{aligned}$$

Substituting this into (85), we get the exploration cost as given above.

□

This result confirms some intuitions that we may have, which is that the cost of exploration is proportional to the emphasis placed on exploration given by λ and is inversely proportional to the discount rate ρ which means that exploration is more costly in scenarios with longer effective time horizons. This makes sense because the longer the process goes on for, the more time there is for the cost of sub-optimal decisions to accumulate.

4.4.4 Vanishing exploration

If we compare (82) with (84), we notice that the optimal policy's mean in the exploratory case coincides exactly with the optimal control in the classical case (which are both independent of λ), and the variance is a multiple of the exogenous exploration parameter λ. Furthermore, as we have just seen, the exploration cost is also a multiple of λ. Thus, if we send the exploration parameter $\lambda \to 0$, we observe that the entropy-regularised LQ problem converges to the classical one.

Theorem 4.6 (Convergence of ELQ Solution to SLQ Solution)**.** *If the exploratory LQ problem is solvable, then for each $x \in \mathbb{R}$,*

$$\lim_{\lambda \to 0} \hat{\pi}(\cdot; x) = \delta_{\hat{u}(x)}(\cdot) \ \textit{weakly}, \tag{86}$$

where $\delta_{\hat{u}(x)}(\cdot)$ is the Dirac distribution with mean $\hat{u}(x)$. Moreover,

$$\lim_{\lambda \to 0} |V(x) - V^{cl}(x)| = 0. \tag{87}$$

Proof (sketch): The weak convergence of the optimal control distribution can be observed from the fact that the mean of $\hat{\pi}(u; x)$ in (82) coincides with the optimal control $\hat{u}(x)$ and the fact that α_2, α_1 are independent of λ, along with the definition of the Dirac distribution.

The pointwise convergence of $V(\cdot)$ and $V^{cl}(\cdot)$ follows from the fact that these two value functions differ by a term which has the form $\beta\lambda(\ln(\gamma\lambda) - 1)$ where γ is a constant. This converges to 0 as $\lambda \to 0$, thus concluding the proof.

□

4.5 Exploratory Linear Quadratic (ELQ) Algorithm

We can use these results along with the exploration-rewarded policy improvement theorem 3.1, which applies in this problem to design a reinforcement learning algorithm to implement an **Exploratory Linear Quadratic Regulator** (see Appendix E for a link to the implementation). We will not present a full derivation here for each design choice of this algorithm as this will be done in the design of the Exploratory Mean-Variance (EMV) algorithm, but some important details will be mentioned here:

1. Due to the exponential decay of the instantaneous rewards, we run simulations over a sufficiently large time horizon $[0, T]$ so that rewards become negligible well before time T. This time interval is then subdivided into K episodes so that the agent updates its parameters (learns) at the end of each episode.
2. To remove the assumption of knowing the full state dynamics (the coefficients A, B, C, D), we parameterise the optimal value function by a vector $\theta = (\theta_0, \theta_1, \theta_2)'$ such that $V^\theta(t, x) = \theta_2 x^2 + \theta_1 x + \theta_0$, and the policy by θ and a scalar $\phi > 0$ such that $\mathcal{H}(\pi^{\theta,\phi}) = \phi$.
3. We use the feedback policy form given by (55) to update our policy after each training episode so that we remain within the class of Gaussian policies over the whole period.
4. We update the parameters θ, ϕ using Stochastic Gradient Descent on a discretised version of the Bellman/TD-error denoted $\tilde{C}(\theta, \phi)$ (see Chapter 5.2 for details).
5. Due to the fact that there are only 3 equations available to link the model parameters A, B, C, D with our learned parameters θ, we must select a subset of the model parameters to be *directly estimated*. It is well-documented that in general, accurately estimating the drift parameters is significantly more difficult than estimating the volatility parameters [19], so for our implementation, we **simulate** an asymptotically normal estimation procedure for the volatility parameters C, D by sampling 'estimates' directly from the normal distributions $\mathcal{N}(C, |C|)$ and $\mathcal{N}(D, |D|)$ respectively.

Algorithm 1 ELQ: Exploratory Linear Quadratic Regulator

Input: Model parameters $A, B, C, D, M, R, N, P, Q, \rho$, exploration rate λ, learning rates η_θ, η_ϕ, initial state x_0, effective horizon T, discretisation Δt, number of episodes K

Initialise θ, ϕ
$n \leftarrow \lfloor \frac{T}{K\Delta t} \rfloor$
for $k \leftarrow 1$ *to* K **do**
  for $i \leftarrow 1$ *to* n **do**
    Sample and store (t_i^k, x_i^k) in $\mathcal{D}$ from state dynamics under $\pi^{\theta,\phi}$
  end for
  $\theta_i \leftarrow \theta_i - \eta_\theta \frac{\partial \tilde{C}(\theta,\phi)}{\partial \theta_i}$
  $\phi \leftarrow \phi - \eta_\phi \frac{\partial \tilde{C}(\theta,\phi)}{\partial \phi}$
  $C, D \leftarrow$ estimate from trajectory data $\mathcal{D}$
  $A, B \leftarrow$ solve algebraic equations (61),(62) for A,B using estimates C, D and parameter θ
  $\pi^{\theta,\phi} \leftarrow \mathcal{N}\left(u \,\middle|\, \frac{(2\theta_2 F - R)x + \theta_1 B - Q}{N - 2\theta_2 D^2}, \frac{\lambda}{N - 2\theta_2 D^2}\right)$
end for

4.5.1 Empirical Results

We investigated the performance of the ELQ algorithm in comparison to an agent employing the optimal classical control and exploratory policy, both of which assume full knowledge of the state dynamics. The following model parameters were used in the simulations:

A	B	C	D	M	R	N	N	Q	ρ	Δt	T	K	n
1	4	-1	-1	2	-1	1	-1	-1	32	1/30000	1/5	600	10

Figure 4: Parameter values used in ELQ simulations. These were chosen to satisfy conditions under which the algebraic equations (61),(62) have real solutions for the parameters A, B.

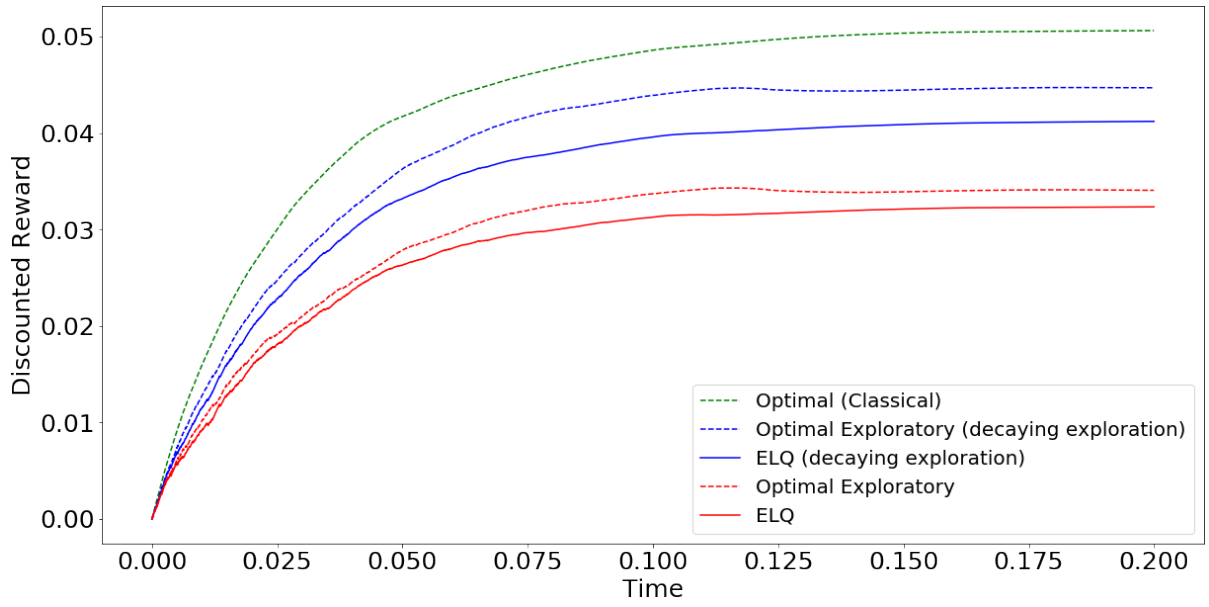

Figure 5: Performance of the ELQ algorithm with and without decaying exploration. Dotted lines indicate theoretically optimal controls under different decision making schemes and solid lines indicate ELQ performance where the agent does not know the true values of A, B, C, D.

In figure 5, we plotted the **accumulated discounted rewards** of five different agents across the time horizon $[0, T]$, given by

$$\sum_{k=1}^{nK} e^{-\rho k \Delta t} \left(\frac{M}{2} X_k^2 + R X_k u_k + \frac{N}{2} u_k^2 + P x_k + Q u_k \right) \Delta t,$$

where X_k denotes the agent's state at iteration k and u_k denotes the agent's control (whether selected or sampled from a policy) at iteration k. We note that in the exploratory case, even though an agent is selecting policies to maximise the *tradeoff* between the original quadratic reward and the exploration reward, we do not take this exploration reward into account when plotting performance as the original reward function is the one we ultimately want to maximise.

We investigated three main types of agent:

1. An optimal classical agent (green line) who executes the optimal feedback control (75) for the whole period. Since this classical control maximises the total expected discounted reward, their performance is expected to be better than any agent who performs exploration and incurs an exploration cost.

2. An exploratory agent with *constant exploration rate* λ (red lines). This agent samples controls from its feedback policy at each moment in time.

3. An exploratory agent with *decaying exploration rate* (blue lines) given by multiplying the initial exploration rate λ by $\exp(-k/n)$ where k denotes the index of the current training episode and n denotes the total number of training episodes. This agent also samples controls from its feedback policy at each moment.

In figure 5, dotted lines represent the performance of agents who **know** the full model parameters A, B, C, D. In the case of the optimal exploratory agents (blue dotted and red dotted lines), controls are sampled from the optimal feedback policy for the ELQ problem (55). These curves should be an upper bound to the performance of their learning counterparts on average, as they are operating under the best exploratory policy for their given exploration rates. In contrast, solid lines represent ELQ agents who **do not know** the full model parameters, but use Algorithm 1 to continually learn by updating their policies and estimates.

As depicted in figure 5, performance of an exploratory agent is in general bounded by an agent executing the optimal controls with perfect knowledge, confirming the fact that there is a cost associated with exploration. Interestingly, the ELQ algorithm's performance is not too different from an exploratory algorithm assuming perfect knowledge of the environment parameters A, B, C, and D. This suggests that even with crude estimation procedures and no knowledge of the underlying dynamics, our reinforcement learning agent's performance is fairly close to a theoretically optimal exploratory policy.

Surprisingly, the results depicted also suggest that adopting a decaying exploration rate λ across training episodes leads to significantly better performance than using a fixed λ, with our ELQ agent outperforming a theoretically optimal exploratory agent with constant exploration rate. This highlights the important fact that always performing a high level of exploration can incur quite a significant cost on the rewards an agent gains.

5 Exploratory Mean-Variance (EMV) Portfolio Selection

We shall now revisit the Mean-Variance problem outlined in Chapter 2.3 under the exploratory lens with a view to designing a reinforcement learning agent which is able to automatically trade risky assets in real time.

5.1 Exploratory Mean-Variance (EMV) Problem

For this formulation, we denote the control process $u = \{u_t,\ 0 \leq t \leq T\}$ which is sampled from a distribution of controls $\pi = \{\pi_t, 0 \leq t \leq T\}$. Denoting X_t^π as the state process in the relaxed formulation, we have the state dynamics:

Definition 5.1. *The **Exploratory Mean-Variance (EMV) state dynamics** is:*

$$dX_t^\pi = \tilde{b}(\pi_t)dt + \tilde{\sigma}(\pi_t)dW_t,\ \ 0 < t \leq T \text{ and } X_0^\pi = x_0, \tag{88}$$

where

$$\tilde{b}(\pi) \coloneqq \int_{\mathbb{R}} \rho\sigma u \pi(u)du,\ \pi \in \mathcal{P}(\mathbb{R}), \tag{89}$$

$$\tilde{\sigma}(\pi) \coloneqq \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \pi(u)du},\ \pi \in \mathcal{P}(\mathbb{R}). \tag{90}$$

As before, $\mathcal{P}(\mathbb{R})$ denotes the set of probability density functions on $\mathbb{R}$ that are absolutely continuous with respect to the Lebesgue measure.

We can define the **mean and variance processes** μ_t, σ_t^2 similarly by

$$\mu_t \coloneqq \int_{\mathbb{R}} u\pi_t(u)du \tag{91}$$

$$\sigma_t^2 \coloneqq \int_{\mathbb{R}} u^2\pi_t(u)du - \mu_t^2. \tag{92}$$

When we substitute these definitions into (88), we get

$$dX_t^\pi = \rho\sigma\mu_t dt + \sigma\sqrt{\mu_t^2 + \sigma_t^2}dW_t. \tag{93}$$

Using the differential entropy $\mathcal{H}(\pi)$ as defined in section (3.1), we can define the **accumulative differential entropy** for a time horizon t as

$$\tilde{\mathcal{H}}(\pi, t) \coloneqq \int_0^t \mathcal{H}(\pi_s)ds = -\int_0^t \left(\int_{\mathbb{R}} \pi_s(u)\ln\pi_s(u)du\right)ds. \tag{94}$$

Finally, using an exploration rate λ we can pose the exploratory MV problem:

Definition 5.2. *The **Exploratory Mean-Variance (EMV) problem** is:*

$$\min_{\pi\in\mathcal{A}(x_0,0)} Var(X_T^\pi) - \lambda\tilde{\mathcal{H}}(\pi, T) \tag{95}$$

$$\text{subject to } \mathbb{E}[X_T^\pi] = z. \tag{96}$$

As before, we use a Lagrange multiplier w so that we can apply the technique of Dynamic Programming. Under this formulation, the problem can be stated as

$$\min_{\pi\in\mathcal{A}(x_0,0)} \mathbb{E}\left[(X_T^\pi - w)^2 - \lambda\tilde{\mathcal{H}}(\pi, T)\right] - (w - z)^2,\ w \in \mathbb{R}, \tag{97}$$

where $\mathcal{A}(x_0, 0)$ is the set of admissible policies on $[0, T]$ obeying a set of technical assumptions.

Definition 5.3. *For each $(s, y) \in [0, T) \times \mathbb{R}$, we consider the state equation (93) on $[s, T]$ with $X_s^\pi = y$. A policy is **admissible for the EMV problem** ($\pi \in \mathcal{A}(s, y)$) if:*

1. $\pi_t \in \mathcal{P}(\mathbb{R})\ \forall\ s \leq t \leq T$, $\mathbb{P}$*-a.s;*
2. $\{\int_A \pi_t(u)du,\ s \leq t \leq T\}$ *is $\mathcal{F}_t$-progressively measurable, for each $A \in \mathcal{B}(\mathbb{R})$ (the Borel σ-algebra on $\mathbb{R}$);*
3. $\mathbb{E}\left[\int_s^T (\mu_t^2 + \sigma_t^2)dt\right] < \infty$;
4. $\mathbb{E}\left[\left|(X_T^\pi - w)^2 - \lambda \int_s^T \mathcal{H}(\pi_t)dt\right| \middle| X_s^\pi = y\right] < \infty$.

Under this definition, the exploratory SDE (93) under an admissible policy has a unique strong solution on $[s, T]$ satisfying $X_s^\pi = y$. Once this problem is solved with a minimising policy $\hat{\pi}$, the Lagrange multiplier w can be chosen to satisfy the constraint (96). Finally, we can define the value and optimal value functions for this problem:

Definition 5.4. *For a given admissible policy π, any $(s, y) \in [0, T) \times \mathbb{R}$, and a fixed $w \in \mathbb{R}$, the **value function** $V^{\boldsymbol{\pi}}$ is:*

$$V^{\boldsymbol{\pi}}(s, y; w) \coloneqq \mathbb{E}\left[(X_T^{\boldsymbol{\pi}} - w)^2 - \lambda\tilde{\mathcal{H}}\left(\pi, T\right) \,\middle|\, X_s^{\boldsymbol{\pi}} = y\right] - (w - z)^2. \tag{98}$$

Definition 5.5. *For $(s, y) \in [0, T) \times \mathbb{R}$ and a fixed $w \in \mathbb{R}$, the **optimal value function** for the EMV problem is:*

$$V(s, y; w) \coloneqq \inf_{\pi \in \mathcal{A}(s,y)} \mathbb{E}\left[(X_T^\pi - w)^2 - \lambda\tilde{\mathcal{H}}\left(\pi, T\right) \,\middle|\, X_s^\pi = y\right] - (w - z)^2. \tag{99}$$

With the key concepts defined for this problem, we are ready to derive its solution using Dynamic Programming.

5.1.1 Solution To Exploratory MV Problem

We approach the solution of this EMV problem in the exact same way as in section (3.2). Applying Bellman's principle of optimality, we have

$$V(t, x; w) = \inf_{\pi \in \mathcal{A}(t,x)} \mathbb{E}\left[V(s, X_s^\pi; w) - \lambda \int_t^s \mathcal{H}(\pi_v)dv \,\middle|\, X_t^\pi = x\right], \tag{100}$$

for $x \in \mathbb{R},\ 0 \leq t < s \leq T$. As before, we know that V satisfies the HJB equation, which for this problem is

$$\frac{\partial v}{\partial t}(t, x; w) + \min_{\pi \in \mathcal{P}(\mathbb{R})} \left(\frac{1}{2}\tilde{\sigma}^2(\pi)v_{xx}(t, x; w) + \tilde{b}(\pi)v_x(t, x; w) - \lambda\mathcal{H}(\pi)\right) = 0, \tag{101}$$

or equivalently in integral form

$$\frac{\partial v}{\partial t}(t, x; w) + \min_{\pi \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2 u^2 v_{xx}(t, x; w) + \rho\sigma u v_x(t, x; w) + \lambda \ln \pi(u)\right) \pi(u)du = 0, \tag{102}$$

with terminal condition $v(T, x; w) = (x - w)^2 - (w - z)^2$.

Applying the admissibility constraints and solving the HJB equation using the variational method described in section (3.2), the optimal feedback control distribution $\hat{\boldsymbol{\pi}}(u; t, x, w)$ is given by

$$\hat{\boldsymbol{\pi}}(u; t, x, w) = \frac{\exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}\sigma^2 u^2 v_{xx}(t, x; w) + \rho\sigma v_x(t, x; w)\right)\right)}{\int_{\mathbb{R}} \exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}\sigma^2 u^2 v_{xx}(t, x; w) + \rho\sigma v_x(t, x; w)\right)\right) du} \tag{103}$$

$$= \mathcal{N}\left(u \,\middle|\, -\frac{\rho}{\sigma}\frac{v_x(t, x; w)}{v_{xx}(t, x; w)}, \frac{\lambda}{\sigma^2 v_{xx}(t, x; w)}\right), \tag{104}$$

where we assume $v_{xx}(t,x;w) > 0$. Substituting this optimal feedback policy (abbreviated to $\hat{\boldsymbol{\pi}}(u)$) back into the HJB equation (101) and using the definitions in (89) and (90), we obtain

$$\frac{\partial v}{\partial t}(t,x;w) + \frac{\sigma^2 v_{xx}(t,x;w)}{2}\int_{\mathbb{R}} u^2 \hat{\boldsymbol{\pi}}(u)du + v_x(t,x;w)\rho\sigma\int_{\mathbb{R}} u\hat{\boldsymbol{\pi}}(u)du + \lambda\int_{\mathbb{R}}\hat{\boldsymbol{\pi}}(u)\ln\hat{\boldsymbol{\pi}}(u)du = 0. \tag{105}$$

It is a well known result that the differential entropy of a normal distribution with mean μ and standard deviation σ is given by

$$\mathcal{H}(u \mid \mathcal{N}(\mu,\sigma^2)) = \ln(\sigma\sqrt{2\pi e}). \tag{106}$$

Using this to evaluate the entropy term and noting that $\int_{\mathbb{R}} u\hat{\boldsymbol{\pi}}(u)du$ is simply the mean of the optimal feedback policy and $\int_{\mathbb{R}} u^2\hat{\boldsymbol{\pi}}(u)du$ is simply the sum of the squared mean and variance, this simplifies to

$$\frac{\partial v}{\partial t}(t,x;w) - \frac{\rho^2}{2}\frac{v_x^2(t,x;w)}{v_{xx}(t,x;w)} + \frac{\lambda}{2}\left(1 - \ln\frac{2\pi e\lambda}{\sigma^2 v_{xx}(t,x;w)}\right) = 0. \tag{107}$$

Noting that the value function at the terminal time T is quadratic in x, we can employ a similar method used in solving the general SLQ problem in section 2.2 and make the ansatz that the solution to this PDE takes the form

$$v(t,x;w) = \frac{1}{2}P(t)x^2 + \phi(t)x + f(t).$$

Taking the necessary derivatives, we get

$$\begin{aligned} v_x(t,x;w) &= P(t)x + \phi_t, \\ v_{xx}(t,x;w) &= P(t), \\ \frac{\partial v}{\partial t}(t,x;w) &= \frac{1}{2}P'(t)x^2 + \phi'(t)x + f'(t). \end{aligned}$$

Substituting these into equation (107), we get

$$\frac{1}{2}P'(t)x^2 + \phi'(t)x + f'(t) = \frac{\rho^2}{2}\left(P(t)x^2 + 2\phi(t)x + \frac{\phi^2(t)}{P(t)}\right) - \frac{\lambda}{2}\left(1 - \ln\frac{2\pi e\lambda}{\sigma^2 P(t)}\right).$$

Equating coefficients of x^2, x and the constants, we get simple ODE's for $P(t), \phi(t), f(t)$ which can be solved directly and coefficients compared with the terminal condition $v(T,x;w) = x^2 - 2wx + 2wz - z^2$ to yield

$$P(t) = e^{-\rho^2(T-t)} \tag{108}$$

$$\phi(t) = -2we^{\rho^2(T-t)} \tag{109}$$

$$f(t) = w^2 e^{-\rho^2(T-t)} + \frac{\lambda\rho^2}{4}(T^2 - t^2) - \frac{\lambda}{2}\left(\rho^2 T - \ln\frac{\sigma^2}{2\pi\lambda}\right)(T-t) - (w-z)^2. \tag{110}$$

Finally, substituting these back into our ansatz and factorising, we are left with

$$v(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda\rho^2}{4}(T^2 - t^2) - \frac{\lambda}{2}\left(\rho^2 T - \ln\frac{\sigma^2}{2\pi\lambda}\right)(T-t) - (w-z)^2. \tag{111}$$

We can then take the necessary partial derivatives of $v(t,x;w)$ and then substitute them into (104) to yield an explicit form for the optimal feedback policy. This can then be substituted into the state dynamics (88) to yield the optimal wealth process. The results are summarised in the following theorem:

Theorem 5.1 (Solution to EMV Problem)**.** *The optimal value function of the Lagrangian formulation of the EMV problem* (97)*, which is*

$$\min_{\pi\in\mathcal{A}(x_0,0)} \mathbb{E}\left[(X_t^\pi - w)^2 - \lambda\tilde{\mathcal{H}}(\pi)\right] - (w-z)^2,\ w\in\mathbb{R}, \tag{112}$$

is given by

$$V(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda\rho^2}{4}(T^2-t^2) - \frac{\lambda}{2}\left(\rho^2 T - \ln\frac{\sigma^2}{\pi\lambda}\right)(T-t) - (w-z)^2, \tag{113}$$

for $(t,x)\in[0,T]\times\mathbb{R}$*. Moreover, the optimal feedback policy is given by*

$$\hat{\boldsymbol{\pi}}(u;t,x,w) = \mathcal{N}\left(u \,\middle|\, -\frac{\rho}{\sigma}(x-w),\ \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}\right), \tag{114}$$

and the associated optimal wealth process under $\hat{\boldsymbol{\pi}}$ *is the unique strong solution of*

$$d\hat{X}_t = -\rho^2(\hat{X}_t - w)dt + \sqrt{\rho^2(\hat{X}_t-w)^2 + \frac{\lambda}{2}e^{\rho^2(T-t)}}dW_t,\ \hat{X}_0 = x_0. \tag{115}$$

Finally, the Lagrange multiplier w *is given by*

$$w = \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}. \tag{116}$$

Proof (sketch): The same arguments as outlined in proofs of theorems 4.1 and 4.2 for verification of the optimal value function (113) and policy (114) can be applied here. Similarly, the admissibility of the optimal policy can be verified in the same way and the optimal state equation (115) can be obtained by substitution of the optimal policy into the SDE (88). Therefore, it remains to show that the Lagrange multiplier does indeed take on the form in (116).

We can determine the value of w by using the constraint $\mathbb{E}[\hat{X}_T] = z$. Using standard arguments involving the Burkholder-Davis-Gundy inequality and Gronwall's lemma, we can bound $\mathbb{E}[\sup_{0\le t\le T}|\hat{X}_t|^2] < \infty$. Using this, we can integrate the optimal SDE (115) and then take expectations to yield

$$\mathbb{E}\left[\hat{X}_t\right] = x_0 + \mathbb{E}\left[\int_0^t -\rho^2(\hat{X}_s - w)ds\right].$$

Using the finiteness of $\mathbb{E}[\sup_{0\le t\le T}\hat{X}_t^2]$ on $[0,T]$, we can apply Fubini's theorem to give

$$\mathbb{E}\left[\hat{X}_t\right] = x_0 + \int_0^t -\rho^2(\mathbb{E}\left[\hat{X}_s\right] - w)ds.$$

We can convert this equation into an ODE and solve it to yield

$$\mathbb{E}\left[\hat{X}_t\right] = (x_0 - w)e^{-\rho^2 t} + w. \tag{117}$$

Finally, using the constraint on the optimal state at time T, we get

$$\begin{aligned}\mathbb{E}\left[\hat{X}_T\right] &= (x_0-w)e^{-\rho^2 T} + w = z\\ \implies w &= \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T}-1}.\end{aligned}$$

□

5.1.2 Solvability Equivalence in the EMV problem

Here, we once again establish the solvability equivalence between a classical formulation of the MV problem and the EMV problem as seen earlier. From the solutions to the classical and exploratory MV problems, we can state the result on solvability equivalence between the two cases.

Theorem 5.2 (MV Solvability Equivalence)**.** *The following two statements are equivalent:*

1. *The optimal value function of the exploratory MV problem* (97) *and the corresponding optimal feedback control distribution are given respectively by*

$$V(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} + \frac{\lambda\rho^2}{4}(T^2-t^2) - \frac{\lambda}{2}\left(\rho^2 T - \ln\frac{\sigma^2}{\pi\lambda}\right)(T-t) - (w-z)^2, \tag{118}$$

$$\hat{\boldsymbol{\pi}}(u;t,x,w) = \mathcal{N}\left(u \,\middle|\, -\frac{\rho}{\sigma}(x-w),\ \frac{\lambda}{2\sigma^2}e^{\rho^2(T-t)}\right). \tag{119}$$

2. *The optimal value function of the classical MV problem* (17) *and the corresponding optimal feedback control are given respectively by*

$$V^{cl}(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} - (w-z)^2, \tag{120}$$

$$\hat{\boldsymbol{u}}(t,x,w) = -\frac{\rho}{\sigma}(x-w). \tag{121}$$

Moreover, the two problems have the same Lagrange multiplier:

$$w = \frac{ze^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}. \tag{122}$$

The proof of this follows the same line of reasoning as in the proof of Theorem 4.4 so we will omit it here.

5.1.3 Convergence as $\lambda \to 0$ in the EMV problem

The result on convergence of the controls and value functions from the LQ case also applies here:

Theorem 5.3 (Convergence of EMV Solution to Classical MV Solution)**.** *Assume that statement 1 or 2 of Theorem 5.2 holds. Then, for each* $(t,x,w) \in [0,T]\times\mathbb{R}\times\mathbb{R}$*,*

$$\lim_{\lambda\to 0} \hat{\boldsymbol{\pi}}(\cdot;t,x;w) = \delta_{\hat{\boldsymbol{u}}(t,x;w)}(\cdot) \ \textit{weakly}, \tag{123}$$

$$\lim_{\lambda\to 0} |V(t,x;w) - V^{cl}(t,x;w)| = 0. \tag{124}$$

The proof of this again follows by observing that the variance of $\hat{\pi}$ is a multiple of λ and taking a simple limit of the difference between the two value functions.

5.1.4 Cost of exploration in the EMV problem

Our final result in the theoretical analysis of the EMV problem will be a computation of the cost of exploration for this application. We define the cost of exploration associated with the EMV problem again to be the difference between the two optimal value functions adjusted for the contribution due to the entropy term in the EMV formulation:

$$C^{\hat{u},\hat{\pi}}(0,x_0;w) \coloneqq \left(V(0,x_0;w) + \lambda\mathbb{E}\left[\tilde{\mathcal{H}}(\hat{\pi})\middle| X_0^{\hat{\pi}} = x_0\right]\right) - V^{cl}(0,x_0;w), \tag{125}$$

for $x_0 \in \mathbb{R}$, where $\hat{\pi}, \hat{u}$ are the *open-loop* controls generated by the optimal feedback controls (119) and (121) respectively.

By direct computation, we get the following result:

Theorem 5.4 (EMV Exploration Cost). *Assume that statement 1 or 2 of Theorem 5.2 holds. Then the exploration cost for the MV problem is*

$$C^{\hat{u},\hat{\pi}}(0, x_0; w) = \frac{\lambda T}{2}, \; x_0 \in \mathbb{R}, \; w \in \mathbb{R}. \tag{126}$$

Proof (sketch): Substitute the definitions of the optimal value functions and explicitly evaluate the differential entropy of the optimal policy $\hat{\pi}$.

5.2 Exploratory Mean-Variance (EMV) Algorithm

5.2.1 Designing Reinforcement Learning Algorithms

Using the results derived in this chapter, we work towards a complete RL algorithm called the Exploratory Mean Variance (EMV) algorithm. This algorithm consists of two main procedures:

1. A policy evaluation procedure for computing and updating the value function for a given policy. For this, we will use a method which involves minimising a measure of the error known as the continuous-time Bellman's/Temporal Difference (TD) error as introduced by Doya [26].
2. A policy improvement procedure. This will make use of a policy improvement theorem for the EMV problem.

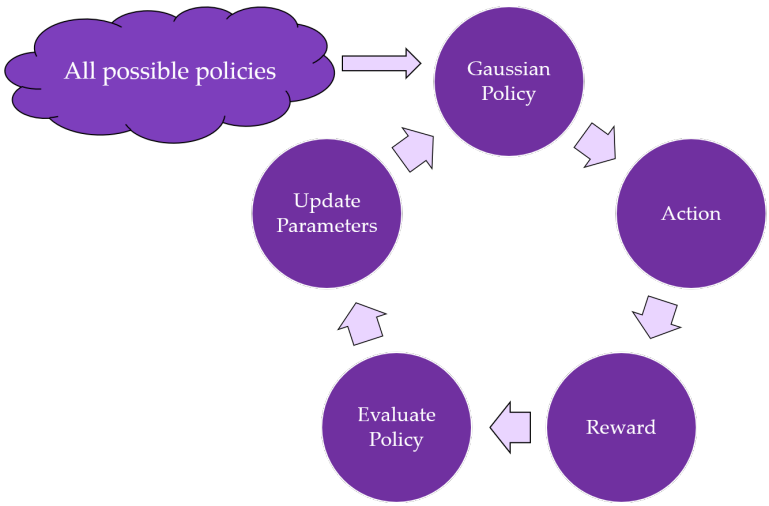

Figure 6: Schematic diagram for the EMV algorithm: we initialise a Gaussian policy, sample allocations in the risky asset to generate a wealth trajectory, and then use the trajectory to perform policy evaluation and improvement.

5.2.2 Policy Improvement

To decide on an updating scheme for our feedback policy $\boldsymbol{\pi}$, we can use the form of the optimal feedback distribution found in equation (104) to come up with a Policy Improvement Theorem (PIT) that expresses this desirable property:

Theorem 5.5 (EMV Policy Improvement Theorem)**.** *Let $w \in \mathbb{R}$ be fixed and $\boldsymbol{\pi} = \boldsymbol{\pi}(\cdot;\cdot,\cdot,w)$ be an arbitrary admissible feedback policy. Suppose that:*

1. *The value function satisfies $V^{\boldsymbol{\pi}}(\cdot,\cdot;w) \in C^{1,2}([0,T]\times\mathbb{R}) \cap C^0([0,T]\times\mathbb{R})$*
2. *The value function satisfies $V^{\boldsymbol{\pi}}_{xx}(t,x;w) > 0$, for any $(t,x) \in [0,T)\times\mathbb{R}$.*
3. *The feedback policy $\tilde{\pi}$ defined by*

$$\tilde{\pi}(u;t,x,w) \coloneqq \mathcal{N}\left(u \,\middle|\, -\frac{\rho}{\sigma}\frac{V^{\boldsymbol{\pi}}_x(t,x)}{V^{\boldsymbol{\pi}}_{xx}(t,x;w)}, \frac{\lambda}{\sigma^2 V^{\boldsymbol{\pi}}_{xx}(t,x;w)}\right), \tag{127}$$

   is admissible.

Then,

$$V^{\tilde{\boldsymbol{\pi}}}(t,x;w) \leq V^{\boldsymbol{\pi}}(t,x;w),\ (t,x) \in [0,T]\times\mathbb{R}. \tag{128}$$

Proof: See Appendix C. The proof is quite similar to that of the Exploration-Rewarded Policy Improvement Theorem but does not have a discounting term and uses a finite horizon.

5.2.3 Eventual convergence to optimal solution

The Policy Improvement Theorem (PIT) above tells us that starting from any given policy $\boldsymbol{\pi}$, we can select a Gaussian policy $\tilde{\boldsymbol{\pi}}$ based on $\boldsymbol{\pi}$ which will never increase (worsen) the value function of the agent's policy from a given state. From this theorem, we get the remarkable insight that the policies we select for policy improvement can remain within the family of Gaussian distributions. Thus, we may ask what happens when we *begin* with a randomly initialised Gaussian policy of the form $\boldsymbol{\pi}_0(u;t,x,w) = \mathcal{N}(u|a(x-w), c_1 e^{c_2(T-t)})$ and update our policy according to (127). The result of such a procedure can be summarised by the following theorem:

Theorem 5.6 (Eventual Convergence Under the Policy Improvement Theorem)**.** *Let*

- $\boldsymbol{\pi}_0(u;t,x,w) = \mathcal{N}(u \mid a(x-w), c_1 e^{c_2(T-t)})$*, with* $a, c_2 \in \mathbb{R},\ c_1 > 0$*;*
- $\{\boldsymbol{\pi}_n(u;t,x,w), (t,x) \in [0,T]\times\mathbb{R}, n \geq 1\}$ *be the sequence of feedback policies updated by the policy improvement scheme* (127)*;*
- $\{V^{\boldsymbol{\pi}_n}(t,x;w), (t,x) \in [0,T]\times\mathbb{R}, n \geq 1\}$ *be the sequence of corresponding value functions.*

Then,

$$\lim_{n\to\infty} \boldsymbol{\pi}_n(\cdot;t,x,w) = \hat{\boldsymbol{\pi}}(\cdot;t,x,w) \text{ weakly,} \tag{129}$$

and

$$\lim_{n\to\infty} V^{\boldsymbol{\pi}_n}(t,x;w) = V(t,w;w), \tag{130}$$

for any $(t,x,w) \in [0,T]\times\mathbb{R}\times\mathbb{R}$ where $\hat{\boldsymbol{\pi}}, V$ are the optimal feedback control distribution (114) *and optimal value function* (118) *respectively.*

Proof: See Appendix D.

In practice however, we do not know the precise values of the parameters ρ, σ. Therefore we will have to use approximations for the value function in designing an implementable policy updating algorithm.

5.2.4 Policy evaluation

We begin with Bellman's principle of optimality, which for this problem states that for any admissible feedback policy $\boldsymbol{\pi}$:

$$V^{\boldsymbol{\pi}}(t,x) = \mathbb{E}\left[V^{\boldsymbol{\pi}}(s,X_s) + \lambda\int_t^s\int_{\mathbb{R}}\pi_v(u)\ln\pi_v(u)dudv \;\middle|\; X_t = x\right],\ s\in[t,T], \tag{131}$$

for $(t,x)\in[0,T]\times\mathbb{R}$. Rearranging this equation and dividing by $s-t$, we get

$$\mathbb{E}\left[\frac{V^{\boldsymbol{\pi}}(s,X_s)-V^{\boldsymbol{\pi}}(t,X_t)}{s-t} + \frac{\lambda}{s-t}\int_t^s\int_{\mathbb{R}}\pi_v(u)\ln\pi_v(u)dudv \;\middle|\; X_t = x\right] = 0. \tag{132}$$

Taking the limit as $s\to t$ and using the fundamental theorem of calculus, the left hand side of the equation above becomes the continuous-time temporal difference (TD) error [26]:

$$\delta_t \coloneqq \dot{V}_t^{\boldsymbol{\pi}} + \lambda\int_{\mathbb{R}}\pi_t(u)\ln\pi_t(u)du. \tag{133}$$

For implementation purposes, we need to discretise the time domain. If each time step has increment Δt, we can define

$$\dot{V}_t^{\boldsymbol{\pi}} \coloneqq \frac{V^{\boldsymbol{\pi}}(t+\Delta t, X_{t+\Delta t}) - V^{\boldsymbol{\pi}}(t,X_t)}{\Delta t}. \tag{134}$$

For our policy evaluation procedure, we wish to minimise the accumulated TD error δ_t over the whole period. The approach taken for this procedure is to write the value function and policy as functions of *parameters* which we can estimate from information we obtain. If we parameterise the value function and policy by V^θ and $\boldsymbol{\pi}^\phi$ respectively, then the problem can be stated as

$$\min_{\theta,\phi} C(\theta,\phi) = \frac{1}{2}\mathbb{E}\left[\int_0^T|\delta_t|^2dt\right] = \frac{1}{2}\mathbb{E}\left[\int_0^T|\dot{V}_t^\theta + \lambda\int_{\mathbb{R}}\pi_t^\phi(u)\ln\pi_t^\phi(u)du|^2dt\right], \tag{135}$$

where $\pi^\phi = \{\pi_t^\phi, t\in[0,T]\}$ is the open-loop policy generated from $\boldsymbol{\pi}^\phi$ from $X_0 = x_0$ as usual.

To approximate the outer integral on the right hand side of (135), we can discretise $[0,T]$ into equal-length intervals of size Δt with $[t_i, t_{i+1}], i = 0,1,...,l$ where $t_0 = 0, t_{l+1} = T$. After this, we can obtain a sample trajectory (collection of states) $\mathcal{D} = \{(t_i,x_i), i = 0,1,...,l+1\}$ by executing the following procedure:

1. Set the initial sample $(t_0,x_0) = (0,x_0)$, and set $i = 0$.
2. Sample from the control distribution $\pi_{t_i}^\phi$ to obtain a control/allocation u_i.
3. Apply this control on the interval $[i,i+1]$ to obtain the next state (t_{i+1},x_{i+1}).
4. Set $i = i+1$.
5. Repeat steps 2-4 until $i = l+1$

With this discretisation, our approximation for the function $C(\theta,\phi)$ becomes:

$$\tilde{C}(\theta,\phi) = \frac{1}{2}\sum_{(t_i,x_i)\in\mathcal{D}}\left(\dot{V}^\theta(t_i,x_i) + \lambda\int_{\mathbb{R}}\pi_{t_i}^\phi(u)\ln\pi_{t_i}^\phi(u)du\right)^2\Delta t. \tag{136}$$

To evaluate this TD error term, a natural question is *what parameterisation should we use for the value function and policy?*. To answer this, we will make use of explicit parametric expressions obtained in Theorems 5.1 and 5.6. Furthermore, due to the theoretical convergence of Gaussian policies to the optimal policy under Theorem 5.6, we will stick to the family of Gaussian policies. These will take the form $\mathcal{N}(u\mid a(x-w), c_1e^{c_2(T-t)})$. Again

using the fact that the differential entropy of a Gaussian distribution with mean μ and variance σ^2 is given by

$$\mathcal{H}(u \mid \mathcal{N}(\mu, \sigma^2)) = \ln(\sigma\sqrt{2\pi e}),$$

and since the variance of our family of Gaussian policies is of the form $c_1 e^{c_2(T-t)}$, their differential entropy will take the form of a linear function in $(T-t)$. Therefore, we can parameterise the differential entropy of our policies by

$$\mathcal{H}(\pi_t^\phi) = \phi_1 + \phi_2(T-t),\ \phi = (\phi_1, \phi_2)',\ \phi_1 \in \mathbb{R},\ \phi_2 > 0. \tag{137}$$

Furthermore, the explicit form of the optimal value function (113) suggests that we parameterise the value function by

$$V^\theta(t,x) = (x-w)^2 e^{-\theta_3(T-t)} + \theta_2 t^2 + \theta_1 t + \theta_0,\ \theta = (\theta_0, \theta_1, \theta_2, \theta_3)', \tag{138}$$

for $(t,x) \in [0,T] \times \mathbb{R}$. From the policy improvement formula (127), we can define a relation between this parameterised value function (138) and the parameterised policy π_t^ϕ through the differential entropy. Taking the necessary partial derivatives of the parameterised value function (138) and substituting this into our PIT formula (127), we get that the *variance* of the policy π_t^ϕ is given by $\frac{\lambda}{2\sigma^2} e^{\theta_3(T-t)}$.

Now, we can calculate the differential entropy of the policy π_t^ϕ using this variance, giving the following relation:

$$\mathcal{H}(\pi_t^\phi) = \frac{1}{2}\ln\frac{\pi e\lambda}{\sigma^2} + \frac{\theta_3}{2}(T-t) = \phi_1 + \phi_2(T-t). \tag{139}$$

From this and the explicit form of the optimal value function (113), we can deduce formulae for the unknown σ, ρ in terms of the parameters which we will be learning:

$$\sigma^2 = \lambda\pi e^{1-2\phi_1}, \tag{140}$$

$$\rho^2 = \theta_3 = 2\phi_2. \tag{141}$$

Using these, we can write the improved policy (127) in terms of the parameters rather than the unknowns ρ, σ:

$$\boldsymbol{\pi}^\phi(u;t,x,w) = \mathcal{N}\left(u \,\middle|\, -\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2} e^{\rho^2(T-t)}\right) \tag{142}$$

$$= \mathcal{N}\left(u \,\middle|\, -\sqrt{\frac{2\phi_2}{\lambda\pi}} e^{\frac{2\phi_1-1}{2}(x-w)},\ \frac{1}{2\pi} e^{2\phi_2(T-t)+2\phi_1-1}\right), \tag{143}$$

assuming that $\rho > 0$. Under this, our task is reduced to finding a method of updating the parameters θ, ϕ to minimise the accumulated TD error. Rewriting this objective using the parameterised form of the differential entropy, we have

$$\tilde{C}(\theta,\phi) = \frac{1}{2}\sum_{(t_i,x_i)\in\mathcal{D}} \left(\dot{V}^\theta(t_i,x_i) - \lambda(\phi_1 + \phi_2(T-t_i))\right)^2 \Delta t. \tag{144}$$

To update these parameters, we use a method known as stochastic gradient descent (SGD) where the update rule for our parameters θ_i, ϕ_i are:

$$\theta_i \coloneqq \theta_i - \eta_{\theta_i}\frac{\partial\tilde{C}(\theta,\phi)}{\partial\theta_i} \tag{145}$$

$$\phi_i \coloneqq \phi_i - \eta_{\phi_i}\frac{\partial\tilde{C}(\theta,\phi)}{\partial\phi_i}, \tag{146}$$

where $\eta_\theta = (\eta_{\theta_0}, \eta_{\theta_1}, \eta_{\theta_2}, \eta_{\theta_3})$ and $\eta_\phi = (\eta_{\phi_1}, \eta_{\phi_2})$ are vectors of learning rates for each of the parameters to be estimated. In our implementation and for many applications, η_θ and η_ϕ are taken to be scalars. Instead of learning all six parameters however, we note that we can use the relation $\theta_3 = 2\phi_2$ as an update rule to remove one of the approximations using SGD. Furthermore, using the terminal wealth condition $V^\theta(T, x; w) = (x-w)^2-(w-z)^2$, we can express θ_0 in terms of $\theta_{1,2}$ by substituting the terminal condition into the parameterised equation for V^θ (138) to give

$$\theta_0 = -\theta_2 T^2 - \theta_1 T - (w-z)^2. \tag{147}$$

Thus, instead of approximating all six parameters using SGD, we can approximate four $(\theta_1, \theta_2, \phi_1, \phi_2)$ and then use the relations derived to compute the other two (θ_0, θ_3). Before we compute the partial derivatives of $\tilde{C}(\theta, \phi)$ with respect to the parameters, we define for notational convenience the discrete TD error $\tilde{\delta}_t$ to be

$$\tilde{\delta}_t(t_i, x_i) \coloneqq \dot{V}^\theta(t_i, x_i) - \lambda(\phi_1 + \phi_2(T - t_i)). \tag{148}$$

This is simply the term which is squared in the discretised form of the accumulated TD error (144).

Taking the partial derivative of $C(\theta, \phi)$ with respect to these four parameters and using the discretised version of $\dot{V}_t^{\boldsymbol{\pi}}$ given in (134) yields

$$\frac{\partial C}{\partial \theta_1} = \sum_{(t_i, x_i) \in \mathcal{D}} \left(\tilde{\delta}_t\right) \Delta t, \tag{149}$$

$$\frac{\partial C}{\partial \theta_2} = \sum_{(t_i, x_i) \in \mathcal{D}} \left(\tilde{\delta}_t\right) (t_{i+1}^2 - t_i^2), \tag{150}$$

$$\frac{\partial C}{\partial \phi_1} = -\lambda \cdot \sum_{(t_i, x_i) \in \mathcal{D}} \left(\tilde{\delta}_t\right) \Delta t, \tag{151}$$

$$\frac{\partial C}{\partial \phi_2} = \left(\frac{\partial \tilde{\delta}_t}{\partial \phi_2}\right) \cdot \sum_{(t_i, x_i) \in \mathcal{D}} \left(\tilde{\delta}_t\right) \Delta t, \tag{152}$$

where

$$\frac{\partial \tilde{\delta}_t}{\partial \phi_2} = -\frac{2e^{-2\phi_2 T}}{\Delta t} \cdot \Big((x_{i+1} - w)^2 e^{2\phi_2 t_{i+1}}(T - t_{i+1}) - (x_i - w)^2 e^{2\phi_2 t_i}(T - t_i)\Big) - \lambda(T - t_i), \tag{153}$$

using the substitution $\theta_3 = 2\phi_2$.

Finally, what remains is to determine a procedure for **learning the the Lagrange multiplier** w from the data. Noting that the Lagrange multiplier is chosen to satisfy the expected terminal wealth constraint $\mathbb{E}\left[\hat{X}_T^\pi\right] = z$, we can use a stochastic approximation update to update the Lagrange multiplier to satisfy this constraint:

$$w_{n+1} = w_n + \alpha_n(z - X_T),$$

with $\alpha_n > 0, n \geq 1$ being the learning rate for the Lagrange multiplier.

In practice, we will not update this on every single iteration of training, but only once every $N \geq 1$ iterations to obtain a more stable learning process for this parameter. This is because using a sample mean over the N most recent terminal wealths gives a better approximation for the expected terminal wealth than just a single value. This leads to the update formula

$$w_{n+1} = w_n + \alpha_n \left(z - \frac{1}{N}\sum_{j=1}^{N} X_T^j\right). \tag{154}$$

A desirable property of this simple learning scheme is that it is statistically self-correcting, which means that when the terminal wealth tends to over-shoot the target wealth z, w will decrease, leading to a decreased mean in the Gaussian policy (143) which will tend to lower the terminal wealth, and vice versa when it under-shoots.

5.2.5 EMV Algorithm Pseudocode

With all the update rules specified, we can state the EMV algorithm:

Algorithm 2 EMV: Exploratory Mean-Variance Portfolio Selection

Input: Market Simulator *Market*, learning rates $\alpha, \eta_\theta, \eta_\phi$, initial wealth x_0, target payoff z, investment horizon T, discretisation Δt, exploration rate λ, number of episodes K, sample mean size N

Initialise θ, ϕ, w
$n \leftarrow \lfloor \frac{T}{\Delta t} \rfloor$
for $k \leftarrow 1$ *to* K **do**
  for $i \leftarrow 1$ *to* n **do**
    Sample and store (t_i^k, x_i^k) in $\mathcal{D}$ from *Market* under π^ϕ
  end for
  $\theta_i \leftarrow \theta_i - \eta_\theta \dfrac{\partial \tilde{C}(\theta, \phi)}{\partial \theta_i},\ i = 1, 2$ from (149) and (150)
  $\theta_0 \leftarrow -\theta_2 T^2 - \theta_1 T - (w - z)^2$ from (147)
  $\theta_3 \leftarrow 2\phi_2$ from (141)
  $\phi_i \leftarrow \phi_i - \eta_\phi \dfrac{\partial \tilde{C}(\theta, \phi)}{\partial \phi_i},\ i = 1, 2$ using (151) and (152)
  $\pi^\phi \leftarrow \mathcal{N}\left(u \,\middle|\, -\sqrt{\dfrac{2\phi_2}{\lambda\pi}} e^{\frac{2\phi_1 - 2}{2}} (x - w),\ \dfrac{1}{2\pi} e^{2\phi_2(T-t) + 2\phi_1 - 1}\right)$
  if $k \bmod N = 0$ **then**
    $w \leftarrow w + \alpha \left(z - \dfrac{1}{N} \sum_{j=k-N+1}^{k} x_n^j \right)$
  end if
end for

5.3 Maximum Likelihood Estimation (MLE) Mean-Variance Algorithm

One way to assess the performance of this RL algorithm is to compare its performance on several metrics to other well established algorithms that have been developed to compute allocation strategies for the classical continuous-time MV problem (17). Here, we will briefly outline the popular Maximum Likelihood Estimation (MLE) method which is used to estimate the parameters μ, σ in the geometric brownian motion model (9). In our case, MLE calculates the values of μ, σ which maximise the log-likelihood of the risky asset's price data being observed (see page 361-363 of [20]).

5.3.1 Mathematical Description of MLE

We begin by assuming that the price process P_t of a risky asset is known up to a vector of unknown parameters $\vartheta := (\alpha, \beta) \in \mathbb{R} \times \mathbb{R}$. Additionally, we assume this price process satisfies the SDE:

$$dP_t = a(P, t; \alpha)dt + b(P, t; \beta)dW_t,\ t \in [0, T]. \tag{155}$$

In such a form, the functions $a(P, t; \alpha)$ and $b(P, t; \beta)$ are known as the drift and diffusion functions respectively. In the case of geometric brownian motion, the drift and diffusion functions are given by $a(P, t; \alpha) = \mu P_t,\ b(P, t; \beta) = \sigma P_t$.

Supposing we have a sequence of $n+1$ historical observations of P_t sampled at times $t_0 < t_1 < ... < t_n$, we would like to find an expression for the joint density function f of the sequence. Observing that the price differential of the SDE for geometric brownian motion only depends on the current price, increments of the price are independent of previous increments on non-overlapping intervals. Thus, we may write the joint density function as a product of conditional density functions on the intervals constructed by the samples

$$f(P_0, ..., P_n; \vartheta) = f_0(P_0; \vartheta) \prod_{k=1}^{n} f(P_k, t_k \mid P_{k-1}, t_{k-1}; \vartheta), \tag{156}$$

where $P_k \coloneqq P_{t_k}$, $f_0(P_0)$ is the marginal density function of P_0, and $f_k \coloneqq f(P_k, t_k \mid P_{k-1}, t_{k-1}; \vartheta)$ is the conditional density function of P_k given P_{k-1} which is also known as a *transition density function*. Given the observations $P_0, ..., P_n$, we would like to find the parameter vector $\hat{\vartheta}$ which maximises the *log-likelihood function* $\ell(\vartheta)$ of observing the data. The function $\ell(\vartheta)$ is defined as the natural logarithm of the joint density function (156):

$$\ell(\vartheta) \coloneqq \log(f(P_0, ..., P_n; \vartheta)) = \sum_{k=1}^{n} \log f_k. \tag{157}$$

The maximum likelihood estimator is then

$$\hat{\vartheta} \coloneqq \underset{\vartheta \in \mathbb{R}\times\mathbb{R}}{\arg\max}\, \ell(\vartheta). \tag{158}$$

5.3.2 MLE Mean-Variance Algorithm

For the MV problem, we can begin by applying Itô's formula to $\log(P)$, which yields

$$d\log(P) = \frac{\partial \log(P)}{\partial t} dt + \frac{\partial \log(P)}{\partial P} dP + \frac{1}{2} \frac{\partial^2 \log(P)}{\partial P^2} (dP)^2 \tag{159}$$

$$= \frac{dP}{P} - \frac{1}{2P^2} (dP)^2 \tag{160}$$

$$= \alpha dt + \sigma dW_t, \tag{161}$$

where $\alpha \coloneqq \left(\mu - \frac{1}{2}\sigma^2\right)$. As before in the case of the EMV algorithm, we will discretise the interval $[0, T]$ into $n+1$ equally spaced time intervals of size Δt. Therefore, for $k = 0, ..., n$, we can write $P_k \coloneqq P_{k\Delta t}$. Noticing that (161) describes a scaled Brownian motion increment σdW_t plus a drift term αdt, the continuously compounded returns $r_k(\Delta t) \coloneqq \log \dfrac{P_k}{P_{k-1}}$ are i.i.d normal random variables with mean $\alpha \Delta t$ and variance $\sigma^2 \Delta t$. This Brownian Motion process has a well known transition density function. Under samples of continuously compounded returns $r_1(\Delta t), ..., r_n(\Delta t)$, the log-likelihood function is

$$\ell(\alpha, \sigma) = -\frac{n}{2} \log(2\pi\sigma^2 \Delta t) - \frac{1}{2\sigma^2 \Delta t} \sum_{k=1}^{n} (r_k(\Delta t) - \alpha \Delta t)^2. \tag{162}$$

Finding the critical points by setting partial derivatives of this function with respect to the parameters to zero, we get the maximum likelihood estimators

$$\hat{\alpha} = \frac{1}{n\Delta t} \sum_{k=1}^{n} r_k(\Delta t) \tag{163}$$

$$\hat{\sigma}^2 = \frac{1}{n\Delta t} \sum_{k=1}^{n} (r_k(\Delta t) - \hat{\alpha} \Delta t)^2. \tag{164}$$

From the closed form expressions of the maximum likelihood estimators, we can make use of the theoretical results shown in the MV problem to devise a portfolio allocation algorithm utilising MLE. For this, we use the most recent 100 data points to continually update the parameter estimates and then substitute these into the equations for the optimal Lagrange multiplier (116) to obtain the optimal allocation given by (22). This leads to the MLE MV Portfolio Selection algorithm:

Algorithm 3 MLE: Mean-Variance Portfolio Selection

Input: Market Simulator *Market*, initial wealth x_0, target payoff z, investment horizon T, discretisation Δt, number of iterations K, sample average size N, risk-free rate r

$n \leftarrow \left\lfloor \frac{T}{\Delta t} \right\rfloor$
for $k \leftarrow 1$ *to* K **do**
  for $i \leftarrow 1$ *to* n **do**
    if $i > N$ **then**
      $\hat{\alpha} \leftarrow \frac{1}{n\Delta t} \sum_{j=i-N}^{i} r_j(\Delta t)$
      $\hat{\sigma}^2 \leftarrow \frac{1}{n\Delta t} \sum_{j=i-N}^{i} \left(r_j(\Delta t) - \hat{\alpha}\Delta t\right)^2$
      $\hat{\mu} \leftarrow \hat{\alpha} + \frac{\hat{\sigma}^2}{2}$
      $\rho \leftarrow \frac{\hat{\mu} - r}{\hat{\sigma}}$
      $w \leftarrow \frac{z e^{\rho^2 T} - x_0}{e^{\rho^2 T} - 1}$
    end if
    $u \leftarrow -\frac{\rho}{\hat{\sigma}}(x - w)$
    Sample next price P_{i+1}^k and wealth x_{i+1}^k from *Market* under allocation u according to (9) and (12)
  end for
end for

5.4 Empirical Results

The EMV and MLE Mean-Variance algorithms outlined previously were implemented in Python (see Appendix E for links to implementations) and the following questions were investigated numerically:

1. How do the EMV and MLE algorithms perform in a **stationary** market?
2. How do the EMV and MLE algorithms perform in a **non-stationary** market?
3. How does the **magnitude of the exploration rate** λ affect the performance of the EMV algorithm?

In this context, a stationary market refers to the risky asset having *constant* drift and diffusion coefficients μ, σ, whereas a non-stationary market refers to the drift and diffusion coefficients being *time-varying* stochastic processes themselves.

5.4.1 Comparison: Stationary Market

We begin with a performance comparison between the EMV and MLE Mean-Variance algorithms. Firstly, we examine the sample mean and variance of terminal wealths by the two algorithms in a stationary and non-stationary market scenario, recalling that the objective is to minimise the variance of terminal wealths subject to the constraint of achieving an expected terminal wealth level of z. The following parameter values were used for the simulations:

μ	σ	T	Δt	K	N	α	η_θ	η_ϕ	x_0	z	λ
-0.3	0.1	1	1/252	20000	50	0.05	0.0005	0.0005	1	1.4	2

Figure 7: Parameter values used for EMV and MLE performance comparison.

The market parameters μ, σ were chosen to represent a plausible risky asset trajectory whose value decreases by 30% over a 1 year period. Here, setting $T = 1$ and $\Delta t = 1/252$ corresponds to a 1 year trading period where trades are executed on working days. The value $z = 1.4$ represents a target expected profit of 40% on the agent's starting capital.

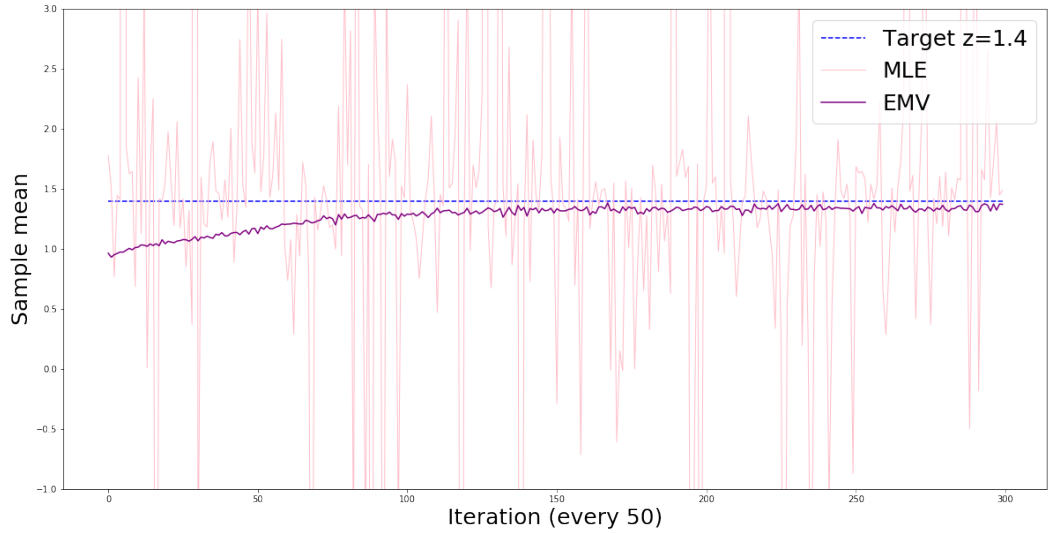

Figure 8: Sample mean of terminal wealths averaged across every 50 iterations in a stationary market scenario.

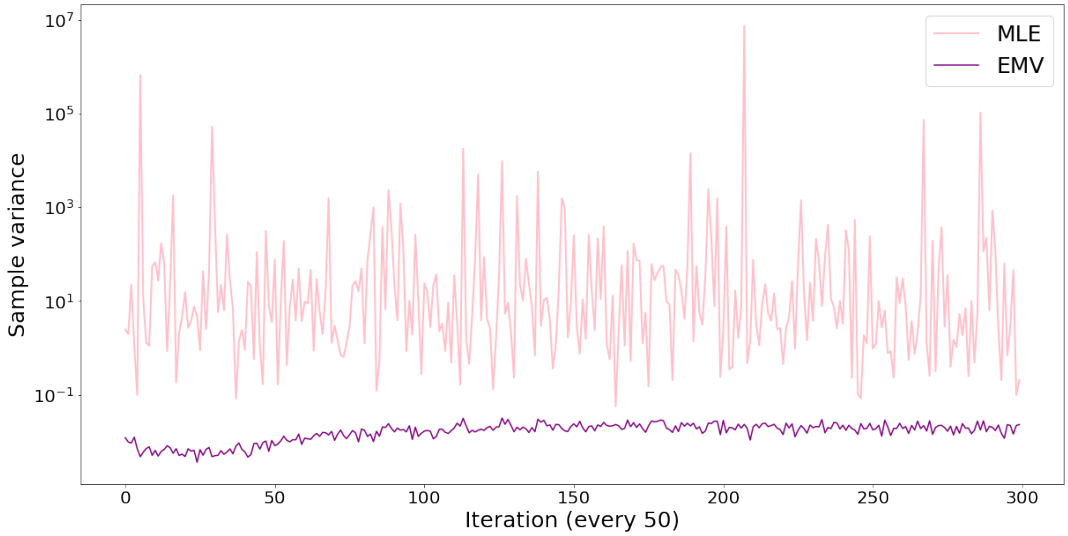

Figure 9: Sample variance of terminal wealths across every 50 iterations in a stationary market scenario.

Figures 8 and 9 replicate quite closely those of Wang et al. [9] and demonstrate a number of desirable properties of the EMV algorithm compared to the statistical estimation procedure:

1. EMV exhibits a characteristic *learning curve* seen by its improving performance in figure 8 which leads to the agent quite consistently achieving the desired terminal wealth level z on average. This indicates that the updating schemes provided in the design of the learning algorithm thanks to the

Policy Improvement Theorem 5.5 do indeed allow the agent to update its parameters to improve its policy and value function estimates.

2. EMV displays much lower variability in performance than MLE as seen by the consistently lower sample variances in figure 9 and by observing the relatively smooth learning curve in figure 8.

3. EMV runs several times more quickly than the MLE-based method, trading and learning on all 2.52×10^6 price data points in tens of seconds on average, whereas MLE takes minutes to complete. This may indicate its potential for applications to high frequency trading.

5.4.2 Comparison: Non-stationary Market

One criticism of the standard Geometric Brownian Motion model for risky asset prices is that the mean and variance of returns are typically non-constant in time. Here, we consider the performance of the two algorithms described previously (EMV and MLE) under a stochastic factor model, which assumes that the price dynamics of the risky asset satisfies the SDE

$$dS_t = S_t(\mu_t dt + \sigma_t dW_t),\ 0 < t \leq T,\ S_0 = s_0 > 0, \tag{165}$$

where $\mu_t, \sigma_t, t \in [0, T]$ vary in time and are known as the drift and volatility processes respectively. This gives rise to wealth dynamics

$$dx_t^u = \sigma_t u_t(\rho_t dt + dW_t),\ 0 < t \leq T,\ x_0^u = x_0 \in \mathbb{R}. \tag{166}$$

5.4.3 Multiscale Stochastic Volatility Models

Stochastic volatility models are a class of models used to describe a stochastic process in which the volatility process σ_t is itself stochastic, that is, the stochastic volatility is a smooth positive function of a number of stochastic processes which is bounded from above and below away from zero:

$$\sigma_t = f(Y_t^1, Y_t^2, ...Y_t^n),\ f : \mathbb{R}^n \to [a, b],\ a, b > 0. \tag{167}$$

There is significant evidence in analyses of stock market price data that points towards at least two stochastic processes (factors) which influence the volatility on different time scales: a fast scale volatility factor which can be modelled by a mean reverting diffusion process (Gaussian Ornstein-Uhlenbeck process) and a slow scale volatility factor which can be modelled by a downscaled stochastic process [28, 29].

To have a well-posed learning problem however, we require that the time change in the processes μ_t, σ_t be small over each training episode $[0, T]$. For this reason, we will consider only the slow scale volatility factor. Thus, we model the volatility process by

$$d\sigma_t = \sigma_t(\delta dt + \sqrt{\delta} dW_t^1), \tag{168}$$

$$d\rho_t = \delta dt, \tag{169}$$

for $0 < t \leq MT$ with $\rho_0 \in \mathbb{R},\ \sigma_0 > 0$ being the initial values of the processes, $\delta > 0$ being the scaling factor for the slow scale stochastic process, and W_t^1 being a Brownian Motion process which has covariation with the original Brownian Motion given by $d\langle W, W^1 \rangle_t = \gamma dt,\ |\gamma| < 1$. This can be implemented by generating another independent Brownian Motion W_t^0 and then letting

$$dW_t^1 = \gamma dW_t + \sqrt{1 - \gamma^2} dW_t^0. \tag{170}$$

Implementing this volatility model using a simple slow scale factor with $\delta = 0.0001, \gamma = 0$ so that the Brownian motions driving the price process and the volatility process are uncorrelated, we find similar results to those in figures 8 and 9:

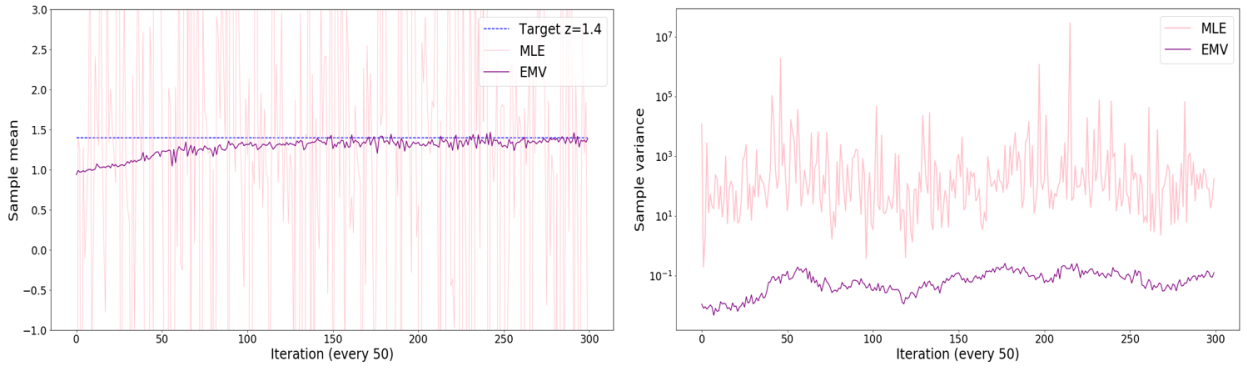

Figure 10: Performance comparison of EMV and MLE in a non-stationary market scenario driven by a slow scale volatility factor. Left: sample mean of terminal wealths across every 50 iterations. Right: sample variance of terminal wealths across every 50 iterations.

We see from figure 10 that both algorithms exhibit greater variability in performance compared to the stationary market scenario due to the non-stationary nature of the underlying dynamics. However, the EMV algorithm still displays a similar learning curve allowing it to achieve the desired returns while maintaining a low sample variance of $\approx 10^{-1}$, whereas the MLE algorithms performance appears to worsen dramatically. This points towards the EMV algorithm's ability to adapt to changing dynamics through the learning process, which is useful as many real-world systems are not stationary.

5.4.4 Exploring the Exploration Rate

Finally, we present an investigation into the effect of different exploration rates on the performance of the EMV algorithm. To begin with, we plot the sample mean of terminal wealths of the EMV algorithm for different values of the exploration rate λ, using the same stationary market scenario as described earlier.

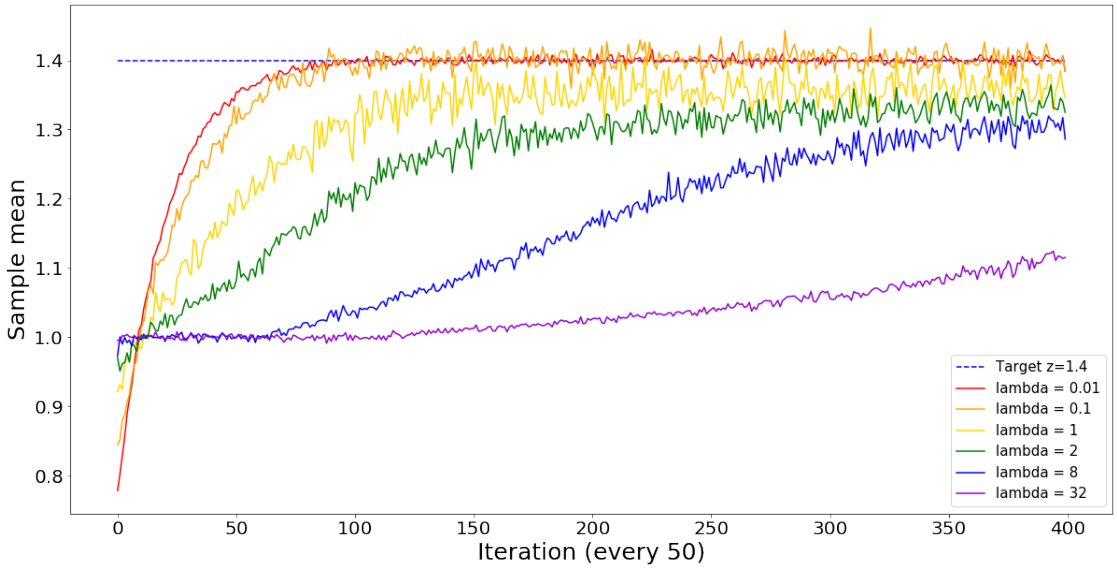

Figure 11: Sample mean of terminal wealths across every 50 iterations in a stationary market scenario for different levels of λ.

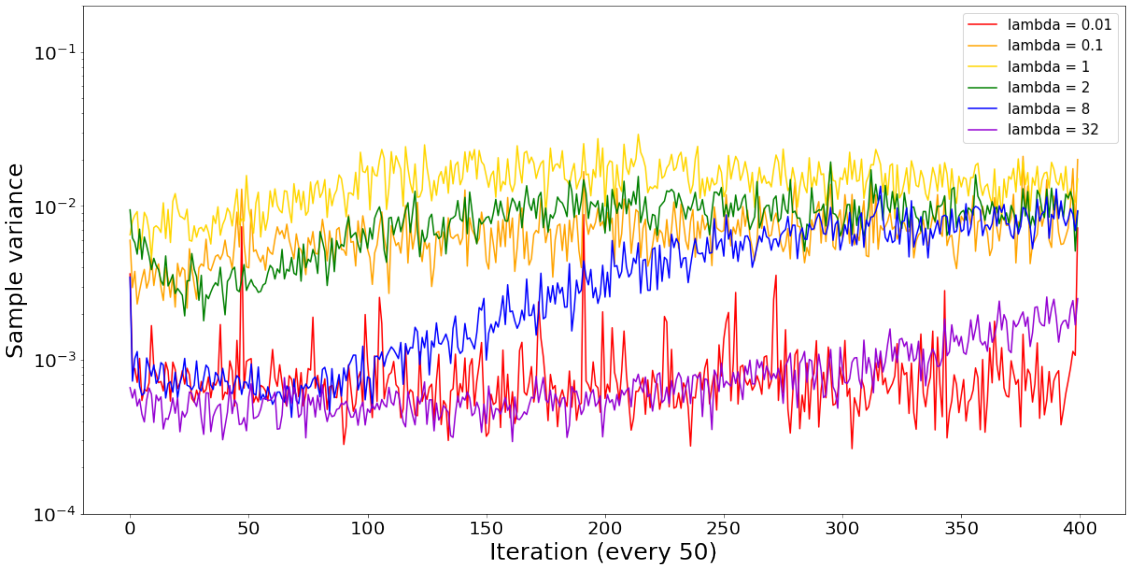

Figure 12: Sample variance of terminal wealths across every 50 iterations in a stationary market scenario for different levels of λ.

We see from figures 11 and 12 that the EMV algorithm's performance varies quite dramatically with the exploration rate across both performance metrics, highlighting the importance of further investigation into the selection of good values for this parameter.

Notably from figure 11 there seems to be a trade-off between initial performance and the rate at which the agent is able to learn how to achieve the desired mean payoff; a lower value of λ gives rise to **faster learning** but **poorer initial performance** up to approximately the 500^{th} iteration. One possible explanation for this may be that for an agent with a lower exploration rate (corresponding to higher exploitation), their initial policies which are wrong in general are over-exploited, leading to more sub-optimal allocations being made initially. A more exploratory agent on the other hand may sample better allocations in this initial stage more frequently.

In the context of asset trading, this is especially relevant as choosing a lower exploration rate to achieve faster learning may increase the probability of ruin significantly. This once again suggests that in order to have an algorithm which is stable *and* learns quickly, one might try to adopt an initially large exploration rate and decay it over time to speed up learning and reduce exploration costs.

6 Summary

6.1 Future Directions

Since it has only been relatively recently that exploration has been examined from a continuous-time stochastic control perspective, there remain many open and interesting questions which would have a significant impact on the way we design reinforcement learning algorithms. First and foremost, the question of designing optimal exploration strategies by perhaps posing a multiple-episode optimisation problem and considering **exploration functions** λ_t instead of constant exploration rates λ could yield profound insights into the nature of exploration and dramatic improvements in the algorithms' performance, as seen by the numerical results in the Exploratory LQ algorithm in figure 5.

Regarding the EMV algorithm, although many of the results seen in our analysis of the one-dimensional exploratory MV problem carry over to multiple dimensions as well, it is unclear how to *implement* a **large scale EMV algorithm** to achieve the remarkable performance results displayed by Wang et al. [9, 30]. This is primarily due to the fact that the parameterisations are expressed in terms of the determinants of various matrices rather than the full matrices themselves, so we would need an accurate method for recovering the original matrices from their determinants. Furthermore, many of the methods used in policy evaluation have been built upon by more recent work such as adaptive, momentum-based gradient descent methods [31] as opposed to the classical SGD method used here. Investigations could be made into all of these design choices in the future for the purposes of building a robust and efficient trading algorithm.

Additionally, there are many un-answered theoretical questions regarding the exploratory framework. The exploratory control framework could be applied to **optimal stopping problems** for example, which may have applicability in a wide number of decision problems currently mainly studied from a theoretical standpoint. Furthermore, although convergence has been derived in the theoretical setting (assuming perfect knowledge of the model parameters) and investigated numerically, there remains a gap in providing **practical convergence guarantees** under the policy improvement scheme. Another example of theoretical interest is that while the **cost** of exploration has been quantified in the theoretical setting, no theoretical results are known to the author which quantify the **benefits of exploration** in a setting where the model parameters are unknown. Moreover, Shannon's differential entropy was chosen as an exploration reward functional due to its intuitive appeal rather than any provable superiority it may have compared to **other possible exploration rewards**. Providing a solid mathematical basis for these questions would be a great stepping stone in laying the theoretical foundations for exploration in reinforcement learning.

Finally, the application of **entropy regularisation** in a learning context has recently found applications in the study of mean field games [32], and perhaps this may motivate investigations into the applicability of exploratory frameworks in other areas of artifical intelligence such as Multi-Agent Reinforcement Learning (MARL) problems or Swarm Intelligence models; the exploration of this field has only just begun.

6.2 Conclusion

This thesis has presented an overview of the Dynamic Programming technique for solving classical stochastic optimal control problems with some applications to the Stochastic Linear Quadratic (SLQ) and Mean-Variance (MV) problems, followed by a description of the development and applications of exploratory optimal control to the SLQ and MV problems based mainly on work by Wang et al. [8, 9], which includes solutions to the exploratory control problems, results on solvability equivalence between exploratory problems and their classical counterparts, quantification of exploration costs, and an elegant way to recover the classical optimal control and value function by sending $\lambda \to 0$ in the exploratory problem. We also build on these ideas with our contribution which includes:

1. The proposal of an Exploratory LQ (ELQ) algorithm along with a demonstrative implementation and performance comparison.
2. The development of a general theory of exploration rewards accompanied by the exploration-rewarded policy improvement theorem.
3. The first open-source implementation of the EMV algorithm in one dimension.
4. A numerical investigation into the effect of varying the exploration rate λ on performance and convergence rates.

Through numerical investigations, we have found that the ELQ and EMV algorithms both achieve promising results, and that there is considerable evidence to suggest that experimenting with decaying schemes and initial values for the exploration rate λ has benefits for performance from the ELQ demonstration in figure 5 and the EMV demonstration in figure 11.

Through a comparison between the EMV and MLE Mean-Variance algorithms, we have found that some of the strengths of EMV include the ability to perform adaptive control, computational efficiency in being able to process and learn from large amounts of data quickly, and high interpretability of the agent's policy due to the separation between exploitation and exploration in policy's mean and variance. This study has also uncovered some weaknesses of the EMV algorithm, the most problematic of which is sample inefficiency. It can be observed from the learning curves in figure 11 that even with low exploration rates, convergence required on the order of 10^6 price data points, which is difficult to obtain in many practical scenarios unless perhaps using high-frequency data. Other potential drawbacks include practical concerns such as transaction costs and leverage constraints, which were not discussed due to time constraints.

From this study, we conclude that the exploratory control framework has both powerful applications in practical Reinforcement Learning problems and useful theoretical properties which warrant further investigation and will hopefully see more development in the coming years. By working to create a bridge between learning and control through exploration, perhaps we will be one step closer to building machines whose abilities to extract valuable information from unknown domains will help to push humanity's capabilities far beyond our current limits.

Bibliography

[1] Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Sallab, A. A. A., Yogamani, S., & Perez, P. (2020). Deep reinforcement learning for autonomous driving: A survey. arXiv preprint arXiv:2002.00444 .

[2] Zhu, H., Yu, J., Gupta, A., Shah, D., Hartikainen, K., Singh, A., ... & Levine, S. (2020). The Ingredients of Real-World Robotic Reinforcement Learning. arXiv preprint arXiv:2004.12570.

[3] Audibert, J.Y., Munos, R., & Szepesvari, C. (2009). Exploration-exploitation trade-off using variance estimates in multi-armed bandits. Theoretical Computer Science, 410:1876–1902, 2009.

[4] Sutton, R.S., & Barto, A.G. (2013). Reinforcement learning: An introduction. MIT press, 2018.

[5] McFarlane, R. (2003). A Survey of Exploration Strategies in Reinforcement Learning.

[6] Neu, G., Jonsson, A., & Gómez, V. (2017). A unified view of entropy-regularized Markov decision processes. ArXiv, abs/1705.07798.

[7] Ahmed, Z., Le Roux, N., Norouzi, M., & Schuurmans, D. (2019). Understanding the impact of entropy on policy optimization. In International Conference on Machine Learning (pp. 151-160). PMLR.

[8] Wang, H., Zariphopoulou, T., & Zhou, X. Y. (2019). Exploration versus exploitation in reinforcement learning: a stochastic control approach. Available at SSRN 3316387.

[9] Wang, H, & Zhou, XY. Continuous-time mean–variance portfolio selection: A reinforcement learning framework. Mathematical Finance. (2020); 30: 1273– 1308. https://doi.org/10.1111/mafi.12281

[10] Shannon, C. E. (2001). A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review, 5(1), 3-55.

[11] Topputo, F., & Bernelli-Zazzera, F (2013). "Approximate Solutions to Nonlinear Optimal Control Problems in Astrodynamics", International Scholarly Research Notices, vol. 2013, Article ID 950912, 7 pages, 2013. https://doi.org/10.1155/2013/950912

[12] Shen, J. (2020). A Stochastic LQR Model for Child Order Placement in Algorithmic Trading. Available at SSRN 3574365.

[13] Sun, J., & Yong, J. (2020). Stochastic Linear-Quadratic Optimal Control Theory: Differential Games and Mean-Field Problems. Springer Nature.

[14] Bellman, R.E. (1954). The theory of Dynamic Programming (No. RAND-P-550). Rand corp santa monica ca.

[15] Bellman, R.E. (1952). On the theory of Dynamic Programming. Proceedings of the National Academy of Sciences of the United States of America, 38(8), 716.

[16] Bellman, R.E. (1957). Dynamic Programming. Princeton University Press.

[17] Karatzas, I., & Shreve, S.E. (1991) Brownian motion and stochastic calculus. Springer-Verlag, 2nd edition, 1991.

[18] Björk, T. (2009). Arbitrage theory in continuous time. Oxford university press.

[19] Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation (No. w0444). National Bureau of Economic Research.

[20] Campbell, J.Y., Lo, A.W., MacKinlay, A.C. & Whitelaw, R.F. (1998). The Econometrics of Financial Market. Macroeconomic Dynamics. 2. 559-562. 10.1017/S1365100598009092.

[21] Zhou, X. Y., & Li, D. (2000). Continuous-time mean-variance portfolio selection: A stochastic LQ framework. Applied Mathematics and Optimization, 42(1), 19-33.

[22] Hong Z.W., Shann, T.Y., Su, S.Y., Chang, Y.H., & Lee C.Y. (2018). Diversity-driven exploration strategy for deep reinforcement learning, arXiv preprint arXiv:1802.04564, 2018.

[23] Touzi, N. (2012). Optimal stochastic control, stochastic target problems, and backward SDE (Vol. 29). Springer Science & Business Media.

[24] Yong, J., & Zhou, X. Y. (1999). Stochastic controls: Hamiltonian systems and HJB equations (Vol. 43). Springer Science & Business Media.

[25] Fleming, W.H., & Nisio, M. (1984). On stochastic relaxed control for partially observed diffusions. Nagoya Mathematical Journal, 93:71–108, 1984.

[26] Doya, K. (2000). Reinforcement learning in continuous time and space. Neural Computation, 12(1):219–245, 2000.

[27] Fouque, J.P., Papanicolaou, G., Sircar, R., & Solna, K. (2003). Multiscale Stochastic Volatility Asymptotics. SIAM Journal on Multiscale Modeling and Simulation. 2. 10.1137/030600291.

[28] Alizadeh, S., Brandt, M. W., & Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. The Journal of Finance, 57(3), 1047-1091.

[29] Fouque, J. P., Papanicolaou, G., Sircar, R., & Solna, K. (2003). Short time-scale in S&P500 volatility. Journal of Computational Finance, 6(4), 1-24.

[30] Wang, H., & Zhou, X. Y. (2019). Large scale continuous-time mean-variance portfolio allocation via reinforcement learning. Available at SSRN 3428125.

[31] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[32] Guo, X., Xu, R., & Zariphopoulou, T. (2020). Entropy Regularization for Mean Field Games with Learning. arXiv preprint arXiv:2010.00145.

[33] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

Appendix

A. Proof sketch of Theorem 3.1

We fix $x \in \mathbb{R}, t \geq 0$ arbitrary. Since $\tilde{\boldsymbol{\pi}}$ is admissible by assumption, the open-loop strategy generated from the feedback policy $\tilde{\pi} = \{\tilde{\pi}_v, v \in [t, \infty)\}$ with initial condition $X_t^{\tilde{\pi}} = x$ is admissible. We let $\{X_s^{\tilde{\pi}}, s \geq t\}$ be the corresponding state process under the open-loop strategy $\tilde{\pi}$.

Applying Itô's formula to the transformation $e^{-\rho t}V^{\boldsymbol{\pi}}(X_t^{\tilde{\pi}})$ as a function of the exploratory state process under the policy $\tilde{\pi}$, we have

$$\begin{aligned} e^{-\rho t}V^{\pi}(X_t^{\tilde{\pi}}) = e^{-\rho\cdot 0}V^{\pi}(X_0^{\tilde{\pi}}) + \int_0^t e^{-\rho v}\left(-\rho V^{\pi}(X_v^{\tilde{\pi}}) + \tilde{b}(X_v^{\tilde{\pi}}, \tilde{\pi}_v)V_x^{\pi}(X_v^{\tilde{\pi}}) + \frac{1}{2}\tilde{\sigma}^2(\tilde{\pi}_v)V_{xx}^{\pi}(X_v^{\tilde{\pi}})\right)dv \\ + \int_0^t e^{-\rho v}\tilde{\sigma}(\tilde{\pi}_v)V_x^{\pi}(X_v^{\tilde{\pi}})dW_v,\ t \in [0, \infty). \end{aligned} \tag{171}$$

Introducing a stopping time to bound the quadratic variation of the process, we let

$$\tau_n = \left\{\inf t \geq 0 : \int_0^t e^{-2\rho v}\tilde{\sigma}^2(\tilde{\pi}_v)(V_x^{\pi}(X_v^{\tilde{\pi}})^2)dv \geq n\right\},\ n \geq 1. \tag{172}$$

Stopping the process above at τ_n and taking expectations with respect to the initial condition, the local martingale part vanishes. Rearranging for the value function at the initial condition, we get

$$\begin{aligned} V^{\pi}(x) = \mathbb{E}\Bigg[e^{-\rho(t\wedge\tau_n)}V^{\pi}(X_{t\wedge\tau_n}^{\tilde{\pi}}) - \int_0^{t\wedge\tau_n} e^{-\rho v}\Bigg(-\rho V^{\pi}(X_v^{\tilde{\pi}}) + \tilde{b}(X_v^{\tilde{\pi}}, \tilde{\pi}_v)V_x^{\pi}(X_v^{\tilde{\pi}}) \\ + \frac{1}{2}\tilde{\sigma}^2(X_v^{\tilde{\pi}}, \tilde{\pi}_v)V_{xx}^{\pi}(X_v^{\tilde{\pi}})\Bigg)dv \ \Bigg|\ X_0^{\tilde{\pi}} = x\Bigg], \end{aligned} \tag{173}$$

for $t \in [0, \infty), n \geq 1$.

Using standard arguments to compare the definition of the exploration-rewarded value function $V^{\pi}(x)$ with its expansion using Itô's formula and the fact that an arbitrarily given policy π is sub-optimal in general, we have

$$\int_U \left[\left(-\rho V^{\pi}(x) + r(x,u) + b(x,u)V_x^{\pi}(x) + \frac{1}{2}\sigma^2(x,u)V_{xx}^{\pi}(x)\right)\pi(u) + \lambda\mathcal{R}[\pi(u)]\right]du = 0 \tag{174}$$

$$\max_{\pi'\in\mathcal{P}(U)}\int_U \left[\left(-\rho V^{\pi}(x) + r(x,u) + b(x,u)V_x^{\pi}(x) + \frac{1}{2}\sigma^2(x,u)V_{xx}^{\pi}(x)\right)\pi'(u) + \lambda\mathcal{R}[\pi'(u)]\right]du \geq 0. \tag{175}$$

However, it was derived earlier that the feedback policy $\tilde{\pi}(u; x)$ defined by (40) achieves the maximum on the left hand side of (175). Therefore, we can rewrite (175) as

$$\int_U \left[\left(-\rho V^{\pi}(x) + r(x,u) + b(x,u)V_x^{\pi}(x) + \frac{1}{2}\sigma^2(x,u)V_{xx}^{\pi}(x)\right)\tilde{\pi}(u) + \lambda\mathcal{R}[\tilde{\pi}(u)]\right]du \geq 0. \tag{176}$$

Recalling the definitions $\tilde{b}(x,\pi) = \int_U b(x,u)\pi(u)du,\ \tilde{\sigma}(x,\pi) = \sqrt{\int_U \sigma^2(x,u)\pi(u)du}$, we can substitute in the state process $X_v^{\tilde{\pi}}$, multiply both sides by $e^{-\rho v}$ and integrate from s to $t\wedge\tau_n$ with respect to v to give an inequality involving the integral in (173):

$$\begin{aligned} \int_0^{t\wedge\tau_n} e^{-\rho v}\Bigg(-\rho V^{\pi}(X_v^{\tilde{\pi}}) + \tilde{b}(X_v^{\tilde{\pi}}, \tilde{\pi}_v)V_x^{\pi}(X_v^{\tilde{\pi}}) + \frac{1}{2}\tilde{\sigma}^2(X_v^{\tilde{\pi}}, \tilde{\pi}_v)V_{xx}^{\pi}(X_v^{\tilde{\pi}})\Bigg)dv \\ \geq -\int_0^{t\wedge\tau_n} e^{-\rho v}\left(\int_U \left(r(X_v^{\tilde{\pi}}, u)\tilde{\pi}_v(u) + \lambda\mathcal{R}[\tilde{\pi}_v(u)]\right)du\right)dv. \end{aligned} \tag{177}$$

Substituting this inequality into (173) yields

$$V^{\pi}(x) \leq \mathbb{E}\left[e^{-\rho(t\wedge\tau_n)}V^{\pi}(X^{\tilde{\pi}}_{t\wedge\tau_n}) + \int_0^{t\wedge\tau_n} e^{-\rho v}\left(\int_U \left(r(X^{\tilde{\pi}}_v,u)\tilde{\pi}_v(u) + \lambda\mathcal{R}[\tilde{\pi}_v(u)]\right)du\right)dv \;\middle|\; X^{\tilde{\pi}}_0 = x\right]. \tag{178}$$

Using a standard argument involving the Burkholder-Davis-Gundy inequality and Gronwall's lemma along with the assumption that π is admissible, we can bound $\mathbb{E}[\sup_{0\leq t\leq T}|X^{\pi}_t|^2] \leq K(1+x^2)e^{KT}$ for $K>0$ large enough independent of n. Using this bound as a dominating function, we take $n\to\infty$ using the dominated convergence theorem so that $t\wedge\tau_n \to t$ almost surely to give

$$V^{\pi}(x) \leq \mathbb{E}\left[e^{-\rho t}V^{\pi}(X^{\tilde{\pi}}_t) + \int_0^{t} e^{-\rho v}\left(\int_U \left(r(X^{\tilde{\pi}}_v,u)\tilde{\pi}_v(u) + \lambda\mathcal{R}[\tilde{\pi}_v(u)]\right)du\right)dv \;\middle|\; X^{\tilde{\pi}}_0 = x\right]. \tag{179}$$

Since we chose $t\geq 0$ arbitrary, the inequality above holds for all $t\geq 0$. Thus, applying assumption 4 and taking the limit as $t\to\infty$ using the dominated convergence theorem again, the first term in the expectation vanishes and the inequality is preserved, giving

$$V^{\pi}(x) \leq \mathbb{E}\left[\int_0^{\infty} e^{-\rho v}\left(\int_U \left(r(X^{\tilde{\pi}}_v,u)\tilde{\pi}_v(u) + \lambda\mathcal{R}[\tilde{\pi}_v(u)]\right)du\right)dv \;\middle|\; X^{\tilde{\pi}}_0 = x\right] = V^{\tilde{\pi}}(x). \tag{180}$$

□

B. Proof sketch of Theorem 4.1

The essence of the proof is to show that:

1. The ansatz $v(x) = \frac{1}{2}k_2x^2 + k_1x + k_0$ is indeed the value function of the problem.
2. The derived optimal control distribution $\hat{\pi}$ is admissible under the assumptions that we have made.

Firstly, we fix $x\in\mathbb{R}$, $\pi\in\mathcal{A}(x)$ and let X^{π} be the state process solving the state equation. To show that the first statement is true, we can show that $v(x)\geq V(x)$ and $v(x)\leq V(x)$ for all states $x\in\mathbb{R}$. To show the former, we can introduce a stopping time $\tau^{\pi}_n \coloneqq \{t\geq 0 \;:\; \int_0^t \left(e^{-\rho t}v'(X^{\pi}_t)\tilde{\sigma}(X^{\pi}_t,\pi_t)\right)^2 dt \geq n\}$ for $n\geq 1$ so that the expectation of the stochastic integral will vanish. We then let $T>0$ be arbitrary, use Itô's formula on the function $v(X^{\pi}_{T\wedge\tau_n})$ applied to the process $e^{-\rho(T\wedge\tau_n)}X^{\pi}_{T\wedge\tau_n}$ to give

$$\begin{aligned} e^{-\rho(T\wedge\tau^{\pi}_n)}v(X^{\pi}_{T\wedge\tau^{\pi}_n}) =& v(x) + \int_0^{T\wedge\tau^{\pi}_n} e^{-\rho t}\left(-\rho v(X^{\pi}_t) + \frac{1}{2}v''(X^{\pi}_t)\tilde{\sigma}(X^{\pi}_t,\pi_t) + v'(X^{\pi}_t)\tilde{b}(X^{\pi}_t,\pi_t)\right)dt \\ &+ \int_0^{T\wedge\tau^{\pi}_n} e^{-\rho t}v'(X^{\pi}_t)\tilde{\sigma}(X^{\pi}_t,\pi_t)dW_t. \end{aligned}$$

Taking the expectation and bounding the left hand side using the fact that v satisifies the HJB equation (41) and π is in general sub-optimal, we get

$$\mathbb{E}\left[e^{-\rho(T\wedge\tau^{\pi}_n)}v(X^{\pi}_{T\wedge\tau^{\pi}_n})\right] \leq v(x) - \mathbb{E}\left[\int_0^{T\wedge\tau^{\pi}_n} e^{-\rho t}\left(\tilde{r}(X^{\pi}_t,\pi_t) + \lambda\mathcal{H}(\pi_t)\right)dt\right].$$

Again, we can use a standard argument involving the Burkholder-Davis-Gundy inequality and Gronwall's lemma to bound $\mathbb{E}[\sup_{0\leq t\leq T}|X^{\pi}_t|^2] \leq K(1+x^2)e^{KT}$ for $K>0$ large enough independent of n. This means that for n large enough, $\tau^{\pi}_n > T$ and so taking the limit as $n\to\infty$ using the dominated convergence theorem and the admissibility of π, we have

$$\mathbb{E}\left[e^{-\rho T}v(X^{\pi}_T)\right] \leq v(x) - \mathbb{E}\left[\int_0^{T} e^{-\rho t}\left(\tilde{r}(X^{\pi}_t,\pi_t) + \lambda\mathcal{H}(\pi_t)\right)dt\right].$$

Now, we use admissibility condition 4 ($\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}\left[(X_T^\pi)^2\right] = 0$) with the fact that $k_2 < 0$ to give $\limsup_{T\to\infty}\mathbb{E}[e^{-\rho T}v(x_T^\pi)] = 0$. Using this, we can apply the dominated convergence theorem again to send $T\to\infty$ to give the first inequality:

$$v(x) \geq \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\tilde{r}(X_t^\pi,\pi_t) + \lambda\mathcal{H}(\pi_t)\right)dt\right] \;\forall x\in\mathbb{R},\; \pi\in\mathcal{A}(x)$$
$$\implies v(x) \geq \sup_{\pi\in\mathcal{A}(x)}\mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\tilde{r}(X_t^\pi,\pi_t) + \lambda\mathcal{H}(\pi_t)\right)dt\right] = V(x)\;\forall x\in\mathbb{R}.$$

The reverse direction follows from the fact that $\hat{\pi}$ achieves the supremum in the HJB equation:

$$\mathbb{E}\left[e^{-\rho T}v(X_T^\pi)\right] = v(x) - \mathbb{E}\left[\int_0^T e^{-\rho t}\left(\tilde{r}(\hat{X}_t^\pi,\hat{\pi}_t) + \lambda\mathcal{H}(\hat{\pi}_t)\right)dt\right].$$

We can apply the dominated convergence theorem this time using that $\liminf_{T\to\infty}\mathbb{E}[e^{-\rho T}v(x_T^\pi)] \leq \limsup_{T\to\infty}\mathbb{E}[e^{-\rho T}v(x_T^\pi)] = 0$ to give

$$v(x) \leq \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\tilde{r}(X_t^{\hat{\pi}},\hat{\pi}_t) + \lambda\mathcal{H}(\hat{\pi}_t)\right)dt\right] \;\forall x\in\mathbb{R},\; \pi\in\mathcal{A}(x),$$

which gives that $v(x) = V(x)$ as required.

To show that $\hat{\pi}\in\mathcal{A}(x)$, we begin by showing admissibility condition 4:

$$\liminf_{T\to\infty} e^{-\rho T}\mathbb{E}[(\hat{X}_T)^2] = 0.$$

We can show using Itô's formula and an argument involving stopping times and the dominated convergence theorem that $\mathbb{E}[(\hat{X}_T)^2]$ contains the terms $e^{2\tilde{A}+\tilde{C}_1^2}$ and $e^{\tilde{A}T}$, where

$$\tilde{A} \coloneqq A + \frac{B(k_2F-R)}{N-k_2D^2}, \quad \tilde{C}_1 \coloneqq C + \frac{D(k_2F-R)}{N-k_2D^2}.$$

It turns out that by assumption (59) and the fact that $k_2 < 0$, condition 4 is satisfied.

Next, condition 5 which is given by $\mathbb{E}\left[\int_0^\infty e^{-\rho t}|L(\hat{X}_t,\hat{\pi}_t)|dt\right] < \infty$ can be verified by expanding our the integral to give

$$\mathbb{E}\left[\int_0^\infty e^{-\rho t}\left|\int_{\mathbb{R}} -\left(\frac{M}{2}(\hat{X}_t)^2 + R\hat{X}_t u + \frac{N}{2}u^2 + P\hat{X}_t + Qu\right)\hat{\pi}_t(u)du + \frac{\lambda}{2}\ln\left(\frac{2\pi e\lambda}{N-k_2D^2}\right)\right|dt\right].$$

For this expression to be finite, it is sufficient to show that $\mathbb{E}[\int_0^\infty e^{-\rho t}(\hat{X}_t)^2dt] < \infty$ which follows from our verification of the previous admissibility condition 4. Conditions 1-3 are straightforward to verify and thus, we have that $\hat{\pi}\in\mathcal{A}(x)$.

Finally, the optimal state process follows from substituting the mean and variance of the optimal control distribution into the general exploratory SDE (49).

□

C. Proof sketch of Theorem 5.5

This proof involves very similar arguments to that of Theorem 3.1, but does not involve an exponential term and is over a finite time horizon. We begin by fixing $(t,x)\in[0,T]\times\mathbb{R}$. Since we have assumed that $\tilde{\boldsymbol{\pi}}$ is admissible, the open-loop control $\tilde{\pi}$ on $[t,T]$ generated from $\tilde{\boldsymbol{\pi}}$ with initial condition $X_t^{\tilde{\boldsymbol{\pi}}} = x$ is admissible. If we let $\{X_s^{\tilde{\boldsymbol{\pi}}}, s\in[t,T]\}$ be the wealth process under $\tilde{\pi}$, we can apply Itô's formula to the value function $V^{\boldsymbol{\pi}}$,

which is a function of t and $\tilde{X}$:

$$V^{\boldsymbol{\pi}}(s,\tilde{X}_s) = V^{\boldsymbol{\pi}}(t,x) + \int_t^s \left(\frac{\partial V^{\boldsymbol{\pi}}}{\partial t}(v, X_v^{\tilde{\boldsymbol{\pi}}}) + \tilde{b}(\tilde{\pi}_v) V_x^{\boldsymbol{\pi}}(v, X_v^{\tilde{\boldsymbol{\pi}}}) + \frac{1}{2}\tilde{\sigma}^2(\tilde{\pi}_v) V_{xx}^{\boldsymbol{\pi}}(v, X_v^{\tilde{\boldsymbol{\pi}}}) \right) dv + \int_t^s \tilde{\sigma}(\tilde{\pi}_v) V_x^{\boldsymbol{\pi}}(v, X_v^{\tilde{\boldsymbol{\pi}}}) dW_v, \ s \in [t,T].$$

Expanding this using the definitions of $\tilde{b}$ and $\tilde{\sigma}$, we obtain

$$\begin{aligned} V^{\boldsymbol{\pi}}(s,\tilde{X}_s) = V^{\boldsymbol{\pi}}(t,x) &+ \int_t^s \left[\frac{\partial V^{\boldsymbol{\pi}}}{\partial t}(v, X_v^{\tilde{\boldsymbol{\pi}}}) + \int_{\mathbb{R}} \left(\rho\sigma u V_x^{\boldsymbol{\pi}}(v, X_v^{\tilde{\boldsymbol{\pi}}}) + \frac{1}{2}\sigma^2 u^2 V_{xx}^{\boldsymbol{\pi}}(v, X_v^{\tilde{\boldsymbol{\pi}}}) \right) \tilde{\pi}_v(u) du \right] dv \\ &+ \int_t^s \left(\sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du} \right) V_x^{\boldsymbol{\pi}}(v, X_v^{\tilde{\boldsymbol{\pi}}}) dW_v, \ s \in [t,T]. \end{aligned} \tag{181}$$

Using a similar argument as before to make expectations finite by bounding the coefficient of the Brownian motion term, we can introduce the stopping times

$$\tau_n \coloneqq \inf\left\{ s \geq t : \int_t^s \left(\int_{\mathbb{R}} \sigma^2 u^2 \tilde{\pi}_v(u) du \right) (V_x^{\boldsymbol{\pi}}(v, X_v^{\tilde{\boldsymbol{\pi}}}))^2 dv \geq n \right\}, \ n \geq 1.$$

Replacing s by $s \wedge \tau_n$ in equation (181) and taking expectations, the Brownian motion term vanishes and we can rearrange the equation to obtain

$$\begin{aligned} V^{\boldsymbol{\pi}}(t,x) = \mathbb{E}\Big[V^{\boldsymbol{\pi}}(s\wedge\tau_n, \tilde{X}_{s\wedge\tau_n}) &- \int_t^{s\wedge\tau_n} \frac{\partial V^{\boldsymbol{\pi}}}{\partial t}(v, X_v^{\tilde{\boldsymbol{\pi}}}) dv \\ &- \int_t^{s\wedge\tau_n} \int_{\mathbb{R}} \left(\rho\sigma u V_x^{\boldsymbol{\pi}}(v, X_v^{\tilde{\boldsymbol{\pi}}}) + \frac{1}{2}\sigma^2 u^2 V_{xx}^{\boldsymbol{\pi}}(v, X_v^{\tilde{\boldsymbol{\pi}}}) \right) \tilde{\pi}_v(u) du dv \ \Big| \ X_t^{\tilde{\boldsymbol{\pi}}} = x \Big]. \end{aligned} \tag{182}$$

On the other hand, we know that the optimal value function satisfies the HJB equation (102), so we can use this along with the assumption that $V^{\boldsymbol{\pi}}$ is smooth and the sub-optimality in general of arbitrary control distributions to give

$$\frac{\partial V^{\boldsymbol{\pi}}}{\partial t}(t,x) + \min_{\pi' \in \mathcal{P}(\mathbb{R})} \int_{\mathbb{R}} \left(\rho\sigma u V_x^{\boldsymbol{\pi}}(t,x) + \frac{1}{2}\sigma^2 u^2 V_{xx}^{\boldsymbol{\pi}}(t,x) + \lambda \ln \pi'(u) \right) \pi'(u) du \leq 0, \tag{183}$$

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

Due to the fact that we used the form of the minimising control distribution (104) for $\tilde{\boldsymbol{\pi}}$, we have that the distribution which minimises (183) is $\pi' = \tilde{\boldsymbol{\pi}}$. Thus, we can substitute this minimiser $\tilde{\boldsymbol{\pi}}$ into (183) remembering that $\boldsymbol{\pi}$ is fixed to give

$$\frac{\partial V^{\boldsymbol{\pi}}}{\partial t}(t,x) + \int_{\mathbb{R}} \left(\rho\sigma u V_x^{\boldsymbol{\pi}}(t,x) + \frac{1}{2}\sigma^2 u^2 V_{xx}^{\boldsymbol{\pi}}(t,x) + \lambda \ln \tilde{\boldsymbol{\pi}}(u) \right) \tilde{\boldsymbol{\pi}}(u) du \leq 0.$$

Rearranging and integrating from t to $s \wedge \tau_n$, we have

$$\int_t^{s\wedge\tau_n} \left[\frac{\partial V^{\boldsymbol{\pi}}}{\partial t}(v,x) + \int_{\mathbb{R}} \left(\rho\sigma u V_x^{\boldsymbol{\pi}}(v,x) + \frac{1}{2}\sigma^2 u^2 V_{xx}^{\boldsymbol{\pi}}(v,x) \right) \tilde{\boldsymbol{\pi}}(u) du \right] dv \geq \int_t^{s\wedge\tau_n} \int_{\mathbb{R}} \lambda \ln \tilde{\boldsymbol{\pi}}_v(u) \tilde{\boldsymbol{\pi}}_V(u) du dv.$$

We can use this inequality above to eliminate the terms involving the partial derivatives of $V^{\boldsymbol{\pi}}$ in (182) to give the inequality

$$V^{\boldsymbol{\pi}}(t,x) \geq \mathbb{E}\Big[V^{\boldsymbol{\pi}}(s\wedge\tau_n, \tilde{X}_{s\wedge\tau_n}^{\tilde{\boldsymbol{\pi}}}) + \lambda \int_t^{s\wedge\tau_n} \int_{\mathbb{R}} \tilde{\pi}_v(u) \ln \tilde{\pi}_v(u) du dv \ \Big| \ X_t^{\tilde{\boldsymbol{\pi}}} = x \Big], \tag{184}$$

for $(t,x) \in [0,T] \times \mathbb{R}, \ s \in [t,T]$. Finally, taking $s = T$, we can use the fact that the value function at time T under both policies is $V^{\boldsymbol{\pi}}(T,x) = V^{\tilde{\boldsymbol{\pi}}}(T,x) = (x-w)^2 - (w-z)^2$ and the assumption that $\tilde{\pi}$ is admissible, we can replace the $V^{\boldsymbol{\pi}}$ in the first term of the expectation above by $V^{\tilde{\boldsymbol{\pi}}}$ and then apply the dominated convergence

theorem sending $n \to \infty$ to give

$$V^{\boldsymbol{\pi}}(t,x) \geq \mathbb{E}\Big[V^{\boldsymbol{\pi}}(T,\tilde{X}_T^{\tilde{\boldsymbol{\pi}}}) + \lambda \int_t^T \int_{\mathbb{R}} \tilde{\pi}_v(u) \ln \tilde{\pi}_v(u) du dv \;\Big|\; X_t^{\tilde{\boldsymbol{\pi}}} = x\Big] \tag{185}$$

$$= \mathbb{E}\Big[V^{\tilde{\boldsymbol{\pi}}}(T,\tilde{X}_T^{\tilde{\boldsymbol{\pi}}}) + \lambda \int_t^T \int_{\mathbb{R}} \tilde{\pi}_v(u) \ln \tilde{\pi}_v(u) du dv \;\Big|\; X_t^{\tilde{\boldsymbol{\pi}}} = x\Big] = V^{\tilde{\boldsymbol{\pi}}}(t,x), \tag{186}$$

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

□

D. Proof sketch of Theorem 5.6

We can verify that the initial feedback policy $\boldsymbol{\pi}_0$ generates an admissible open-loop policy π_0 with respect to the initial state (t,x). Moreover, we note that due to the form of $\boldsymbol{\pi}_0$, the value function $V^{\boldsymbol{\pi}_0}(t,x;w)$ has the form of the Feynman-Kac formula, and therefore it must be the solution to the PDE

$$\frac{\partial V^{\boldsymbol{\pi}_0}}{\partial t}(t,x;w) + \int_{\mathbb{R}} \left(\rho\sigma u V_x^{\boldsymbol{\pi}_0}(t,x;w) + \frac{1}{2}\sigma^2 u^2 V_{xx}^{\boldsymbol{\pi}_0}(t,x;w) + \lambda \ln \boldsymbol{\pi}_0(u;t,x,w)\right) \boldsymbol{\pi}_0(u;t,x,w) du = 0, \tag{187}$$

with $V^{\boldsymbol{\pi}_0}(T,x;w) = (x-w)^2 - (w-z)^2$. Solving the PDE using the same ansatz as outlined in section 5.1, we get

$$V^{\boldsymbol{\pi}_0} = (x-w)^2 e^{(2\rho\sigma a + \sigma^2 a^2)(T-t)} + \int_t^T c_1 \sigma^2 e^{(2\rho\sigma a + \sigma^2 a^2 + c_2)(T-s)} ds \tag{188}$$

$$+ \frac{\lambda c_2}{4}(T-t)^2 - \frac{\lambda \ln(2\pi e c_1)}{2}(T-t) - (w-z)^2. \tag{189}$$

Again, we can verify that $V^{\boldsymbol{\pi}_0}$ satisfies the conditions of the Policy Improvement Theorem (PIT) 5.5: sufficient continuous differentiability, positive second order partial derivative in x, and admissibility of the open loop policy generated by (127). From this, we can apply the PIT to generate a new policy from $\boldsymbol{\pi}_0$:

$$\boldsymbol{\pi}_1(u;t,x,w) = \mathcal{N}\left(u \;\middle|\; -\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2 e^{(2\rho\sigma a + \sigma^2 a^2)(T-t)}}\right). \tag{190}$$

It turns out that we can repeat the procedure above, generating a PDE for $V^{\boldsymbol{\pi}_1}(t,x;w)$ and solving it to yield

$$V^{\boldsymbol{\pi}_1}(t,x;w) = (x-w)^2 e^{-\rho^2(T-t)} + F_1(t), \tag{191}$$

where F_1 is a function of t only. Again, the conditions for the PIT hold so we can apply it again to finally give

$$\boldsymbol{\pi}_2(u;t,x,w) = \mathcal{N}\left(u \;\middle|\; -\frac{\rho}{\sigma}(x-w), \frac{\lambda}{2\sigma^2} e^{\rho^2(T-t)}\right) = \hat{\boldsymbol{\pi}}(u;t,x,w). \tag{192}$$

We see that under $\boldsymbol{\pi}_2$, we obtain the optimal value function $V(t,x;w)$ from (113) and so we can not improve any further. Therefore, we have convergence to the optimal value function after precisely two iterations.

□

E. Link to Implementations of the ELQ, EMV, and MLE Algorithms

- Github Repository
- Exploratory LQ (ELQ) Algorithm
- Exploratory MV (EMV) Algorithm