Kristian Jakobsen
Christian Brennhovd Bang

How can Reinforcement Learning Algorithms be Used for Capital Allocation and Consumption for a Sovereign Wealth Fund?

Master's thesis in Economics and Business Administration
Supervisor: Denis Mike Becker

May 2024

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
NTNU Business School

NTNU
Norwegian University of
Science and Technology

Kristian Jakobsen
Christian Brennhovd Bang

How can Reinforcement Learning Algorithms be Used for Capital Allocation and Consumption for a Sovereign Wealth Fund?

Master's thesis in Economics and Business Administration
Supervisor: Denis Mike Becker
May 2024

Norwegian University of Science and Technology
Faculty of Economics and Management
NTNU Business School

NTNU
Norwegian University of
Science and Technology

Preface

This work is the result of a master's program in Business Analytics at the Department of Economics at the Norwegian University of Science and Technology (NTNU). We want to give a sincere gratitude to our supervisor Denis Mike Becker for his invaluable assistance with programming and his economic knowledge. His guidance has been crucial in making this thesis to a higher standard. The authors take full responsibility for the content of this thesis.

Trondheim, 2024

Kristian Jakobsen

Christian Brennhovd Bang

Sammendrag

Denne oppgaven ser på hvordan forsterkningslæring kan benyttes til porteføljeforvalting for et statlig investeringsfond. Oppgaven bygger på formålet til *Statens Pensjonsfond Utland*, som er å gi fremtidige generasjoner samme velferd og like muligheter som dagens generasjon. Dette gjøres gjennom å teste 5 ulike algoritmer fra PyTorch bliblioteket *Stable Baseline 3*. Hensikten med undersøkelsene er å se hvor godt tilpasset disse maskinlæringsalgoritmene er for å kunne ta investeringsbeslutninger på vegne av et investeringsfond. Maskinlæringsprosessen består av ulike agenter som skal lære å ta beslutninger i et miljø som representerer markedet fondet investerer i. Miljøet agentene samhandler med er dannet gjennom en blanding av investeringsbegrensninger og matematiske formuleringer. Særlig med inspirasjon fra tidligere studier av Knut Anton Mork (Mork et al., 2023) samt Markovs beslutningsprosess, som er et matematisk rammeverk for beslutningstaking i en tidsdiskret stokastisk kontrollprosess. Gjennom straff og belønning skal agentene trene med den hensikt å finne optimal investeringsbeslutning, basert på ulike kriterier. De agentene som presterer best er de som tar beslutninger i tråd med miljøets spesifikke kriterier. Dette inkluderer blant annet porteføljesammensetning, årlig konsum og fremtidig nytte. For å undersøke ytelsen til forsterkningslæringsagentene vil de bli testet for ulike nivåer av straffeparametere, og bli målt opp mot hverandre basert på nytte og belønning. Undersøkelsene i denne oppgaven gir varierende resultater, som indikerer at forsterkningslæringsalgoritmene har potensiealet til å lage gode modeller for investeringsbeslutninger til et statlig investeringsfond. Det trengs likevel ytterligere arbeid for å forbedre disse algoritmene slik at de passer bedre til å løse dette spesifikke problemet.

Abstract

This thesis examines how reinforcement learning can be used for portfolio management for a sovereign wealth fund (SWF). Inspired by the *Norwegian Pension Fund Global*, and their aim to provide future generations with the same welfare and opportunities as the current generation. This study tests five different algorithms from the PyTorch library *Stable Baseline 3* which is DDPG, PPO, A2C, SAC and TD3. The purpose of the research is to investigate how well these reinforcement learning algorithms can make investment decisions on behalf of an investment fund. The machine learning process involves various agents that learn to make decisions in an environment representing the market. A well-balanced mix of investment constraints and mathematical formulations creates the foundation for the environment. Drawing on previous studies by Knut Anton Mork (Mork et al., 2023) and the Markov decision process. The Reinforcement learning model in this thesis have agents that learn through a system of penalties and rewards with the aim of finding optimal investment decisions. The best-performing agents are those that find solutions within the specific criteria of the environment. The output is optimal solutions for capital allocation and consumption rates with the aim of maximizing the utility and reward. The algorithms will be tested with different levels of penalty parameters, specifically two different values of smoothing and wealth penalty. The findings of this thesis suggest that reinforcement learning algorithms have the potential to develop effective models for making investment decisions for a sovereign wealth fund. However, further work is needed to refine these algorithms to better fit this specific problem.

Contents

List of Figures

List of Tables

Abbrevations

RL	Reinforcement Learning
GPFG	Norwegian Government Pension Fund Global
MDP	Markov Desicion Process
TAA	Tactical Asset Allocatian
SAA	Strategic Asset Allocation
NOK	Norwegian Kroner
NMoF	Norwegian Ministry of Finance
NBIM	Norwegian Bank Investment Management
GBM	Geometric Brownian motion
DDPG	Deep Deterministic Policy Gradient
PPO	Proximal Policy Optimization
A2C	Advantage Actor-Critic
SAC	Soft Actor-Critic
TD3	Twin Delayed DDPG
TRPO	Trust Region Policy Optimization
NAREIT	National Association of Real Estate Investment Trusts
CRRA	Constant Relative Risk Aversion

1. Introduction

1.1. Motivation

As citizens in Norway, our motivation is based on the social benefits that arise from a well-managed sovereign wealth fund. We aim to investigate the opportunities that reinforcement learning offers as a tool for sovereign wealth fund managers to efficiently allocate capital and resources. Drawing inspiration from previous research conducted by (Mork et al., 2023), our master's thesis will depart from traditional mathematical approaches and instead leverage reinforcement learning. The lack of prior studies on this specific theme makes it an important area for investigation.

1.2. Background

Sovereign wealth funds (SWFs) play a significant role for nations in effectively allocating and managing their wealth. The Norwegian Government Pension Fund Global stands out as the biggest SWF in the world with a market value in 2024 of over 17.000 billion NOK (NBIM, 2024b). Global financial markets are complex and frequently unpredictable. Therefore, well managed and long-term investment strategies are important to maximize wealth for SWFs. Traditional portfolio management approaches often fall short when navigating such complex environments. Machine learning has emerged as a potential solution due to its capacity to comprehend complex interrelationships. Reinforcement learning stands out as one such method by its ability to learn as a continuous process by receiving rewards and punishments on actions (Russell and Norvig, 2010).

Excessive political influence is a significant challenge in the management of Sovereign Wealth

Funds (SWFs), as it can lead to decisions that prioritize short-term returns (Bernstein et al., 2013). This issue mainly takes place within the context of the agency problem involving political leaders and the management process. Politicians may prioritize achieving short-term political goals over ensuring sustainable fund management. Additionaly, SWFs encounter transparency issues, resulting from the potentially copying of investment strategies (Bahoo et al., 2020). Despite aiming to diversify their investments away from industries prevalent in their home country, these funds often invest in countries sharing similar cultures with their originating nation (Chhaochharia and Laeven, 2008). Investment decisions are therefore not solely driven by profit maximization.

Traditionally, portfolio management introduces the mean-variance optimal portfolio, where the investor should select the portfolio where the capital allocation line intersects the efficient frontier. (Markowitz, 1952). Fama and French, trough their later work demonstrated that market prices reflect all available information, implying that consistently outperforming the market through actively managed portfolios is not possible over time (Fama, 1970). This critique extends to the earlier Capital Asset Pricing Model (CAPM) (Sharpe, 1964).

Classical stochastic control theory and other analytical approaches in financial decision-making heavily rely on model assumptions. Reinforcement learning (RL) leverage extensive financial data with minimal assumptions to enhance decision-making in complex financial environments (Hambly et al., 2023). The stock market is extremely unpredictable and vulnerable to influence from various sources such as social, political, economic, and demographic factors. Therefore, different models will give different results on future data, making predictions, even with deep RL highly challenging (Sahu et al., 2023).

1.3. The Aim of the Thesis

The study will build upon Knut Anders Mork's previous work, which aimed to identify the optimal trade-off between near-future consumption and the preservation of a fund value to enable consumption by future generations (Mork et al., 2023). The article attempts to solve the Merton problem from 1971, which combines the optimal consumption and portfolio management in a continuous-time framework (Merton, 1975). This thesis extends Mork's work by using Reinforcement Learning (RL) as a tool instead of mathematical continuous series. Specifically, by identifying the optimal balance between short-term consumption and at the same time preserving the fund's value for future generations. Based on this, we have formulated the following problem statement: How can Reinforcement Learning algorithms be used to optimize the trade-off between consumption and capital allocation for the long-term preservation of a Sovereign Wealth Fund? Specifically, we will focus on the five algorithms DDPG, PPO, A2C, SAC and TD3 from the *Stable Baseline 3* library.

1.4. The Structure of the Thesis

The structure of this thesis is organized into several chapters, beginning with the literature review and theoretical background (Chapter 2). Additionally, the thesis will explore the Markov Decision Problem in chapter 3, which is the theoretical foundation for the reinforcement learning framework implemented in this study. The reinforcement learning algorithms will be presented in chapter 4, which is DDPG, PPO, A2C, SAC, and TD3. Chapter 5 will focus on the data, which serves as the inputs for the code. Furthermore, results and plots from the code execution will be provided in chapter 6. Chapter 7 will function as the discussion section providing critical reflections regarding the algorithms and the economics. Finally, the thesis will have a conclusion of the findings in chapter 8, before looking into potential further work.

2. Literature Review and Theoretical Background.

2.1. Sovereign Wealth Fund - SWF

The official definition of a Sovereign Wealth Fund (SWF) was published in Appendix 1 of the Santiago Principles, highlighting three key points that describe an SWF. Firstly, the fund should be owned by the government. Secondly, it should invest in foreign financial assets. Thirdly, the fund should invest with financial objectives in mind (International Working Group of Sovereign Wealth Funds, 2008). Over the past two decades, Sovereign Wealth Funds (SWFs) have increasingly emerged as a method for countries to invest their resources and wealth. World wide SWF currently have over 6 trillion dollar worth of asset under management (Wagner, 2013). Norway have as mentioned the world's largest SWF. China follows closely, with the second and third largest SWFs represented by the China Investment Corporation and the SAFE Investment Company, valued at 1,350 billion dollars and 1.090 billion dollars respectively (ISWF, 2024).

2.2. SWF Objectives

According to the International Forum of Sovereign Wealth Funds, different funds have varying objectives to fulfill through their existence. SWFs have diverse objectives, which can be categorized into four main parts: saving funds, stabilization funds, strategic funds, and funds with multiple objectives (IFSWF, 2024b).

2.2.1. Saving Funds

Saving funds are SWFs earmarked for the long term, typically for decades, and are frequently established by nations with commodity resources. Their primary objective is to secure wealth even after these resources are used up. Such funds mainly originate from countries with significant oil and gas reserves. On the top ten list of the largest SWFs are 70% oil and gas producers (ISWF, 2024). Certain saving funds are designed to address future liabilities, often referred to as pension funds, examples of which include Chile's Pension Reserve Fund and Australia's Future Fund. These funds accumulate wealth intended for meeting forthcoming government pension obligations (IFSWF, 2024b).

2.2.2. Stabilisation Funds

Stabilization funds are established for countries with fluctuating income streams, typically derived from commodities. In such instances, they build wealth during periods of high commodity income and expend them when market conditions are unfavorable (IFSWF, 2024b). The Ghana Stabilisation Fund is an example of such a fund, with a total withdrawal of 714 million dollars in funds. Despite this substantial expenditure, a significant portion of the spending has been allocated to debt payments rather than stabilization policies. This indicates that not all stabilization funds fulfill their intended purpose (Gyeyir, 2019).

2.2.3. Strategic Funds

Strategic funds distinguish themselves from other traditional SWFs by also investing in the development of their own countries. The Irish Strategic Investment Fund is an example of such a fund, as it is specifically designed to invest in supporting economic activity in Ireland (IFSWF, 2024b). This approach contrasts with the majority of SWFs, which typically aim to invest in other countries and diversify away from their own industries (Chhaochharia and Laeven, 2008).

2.2.4. Multiple Objectives

Many funds are incorporating two or more objectives, such as the China Investment Corporation and the State Oil Fund of Azerbaijan. A significant number of these funds originate from nations below the Sahara, which have experienced substantial revenue increases due to high commodity prices over the past two decades (IFSWF, 2024b). These countries often struggle with poverty, making it challenging to prioritize future wealth accumulation over present needs. An example is Ethiopia, which has the 34th largest SWF in the world with a market value of about 35.5 trillion dollars (ISWF, 2024). At the same time, Ethiopia is one of the poorest countries in the world, with over 25 million people living in poverty. (Utviklingsfondet, 2024).

2.3. The Santiago Principle

In the early 2000s, politicians in the U.S. and Europe became worried about sovereign wealth funds because they thought there might be financial misconduct. These concerns were rooted in a belief that global capital markets was mainly driven by private entities. The existence of state-owned funds also raised concerns about market distortion arising from non-commercial considerations (Barbary et al., 2023). In 2008, the 26 founding members of the International Forum of Sovereign Wealth Funds (IFSWF) introduced the ”Santiago Principles,” consisting of 24 generally accepted principles and practices. These principles is made to direct the governance, investment, and risk management practices of sovereign wealth funds. Their objective is to encourage good governance, accountability, transparency, and responsible investment practices. IFSWF’s full members voluntarily pledge to follow the Santiago Principles, ensuring their implementation within the rules of local laws (IFSWF, 2024a).

2.4. Optimal Asset Allocation for SWF

Sovereign wealth funds faces challenges when it comes to asset allocation decisions. Both strategic asset allocation (SAA) and tactical asset allocation (TAA) is important to spread their wealth across diverse asset classes, thereby minimizing risk and ensuring profitable returns. SAA, guided by long-term objectives and policies, optimizes the allocation proportions for assets such as bonds, equities, and alternative investments (Rasmussen, 2003). This initial decision has most importance in effectively managing overall investment risk and managing return objectives. TAA complements SAA by fine-tuning the strategic portfolio to either boost excess returns, or help against short and medium-term losses, aligning with long-term investment goals. This type of investment management process incorporates both qualitative and quantitative analyses (Yu et al., 2010).

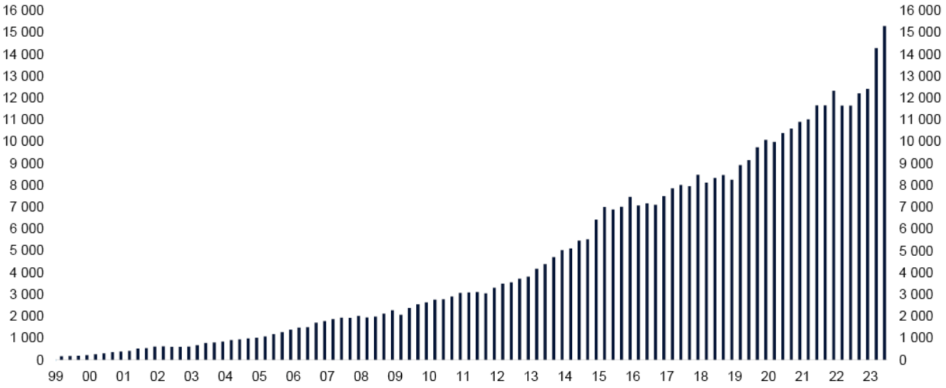
Sovereign wealth funds now have investments in almost every type of financial asset. To achieve diversity, these funds have mainly used four asset classes, which is cash, fixed income, equities and alternative investments (Yu et al., 2010).

- **Cash assets** provide short-term safety as there are long-term risks associated with them. Sovereign wealth funds, which hold cash assets, face uncertainty when renewing investments due to unknown future real interest rates.
- **Public debt securities**, despite being highly secure, offer lower returns and are highly sensitive to changes in interest rates, particularly for middle-to-long-term securities that can be traded. Even slight changes in interest rates can lead to volatility in the traded prices of these securities.
- **Equities** play a crucial role in Sovereign wealth funds' strategic asset allocation due to their risk-return characteristics for long-term investors. Recent studies show that stocks have never had negative returns over a twenty-year period, making them more secure for long-term investment than bonds. While short-term stock returns may be more volatile as they grow faster over an extended period.
- **Alternative investments** is investments in alternatives like real estate, private equity, and commodities, known for high returns and risks. These assets, unlike traditional ones, offer low correlation.

2.5. Government Pension Fund of Norway

The Norwegian Government Pension Fund Global (GPFG) is the world's largest single owner in the global stock market, with ownership of nearly 1.5% of shares in 9000 listed companies worldwide. The purpose of GPFG is to ensure the long-term management of revenues from Norway's oil and gas resources, so that the wealth benefits both current and future generations. In this way, the GPFG are both a stabilization fund (fiscal policy) and a savings fund. As of may 2024, the fund's wealth exceeds 17,749 billion NOK. The fund got its first capital injunction in 1996, and 27 years later, in 2023, the total supply to the fund have been 4 382 billion Norwegian Kroner (NOK) (NBIM, 2024). Since its inception, the fund have experienced a annual nominal return of 5.99 percent. After accounting for management costs and inflation, the real return has been 3.72 percent (NBIM, 2024).

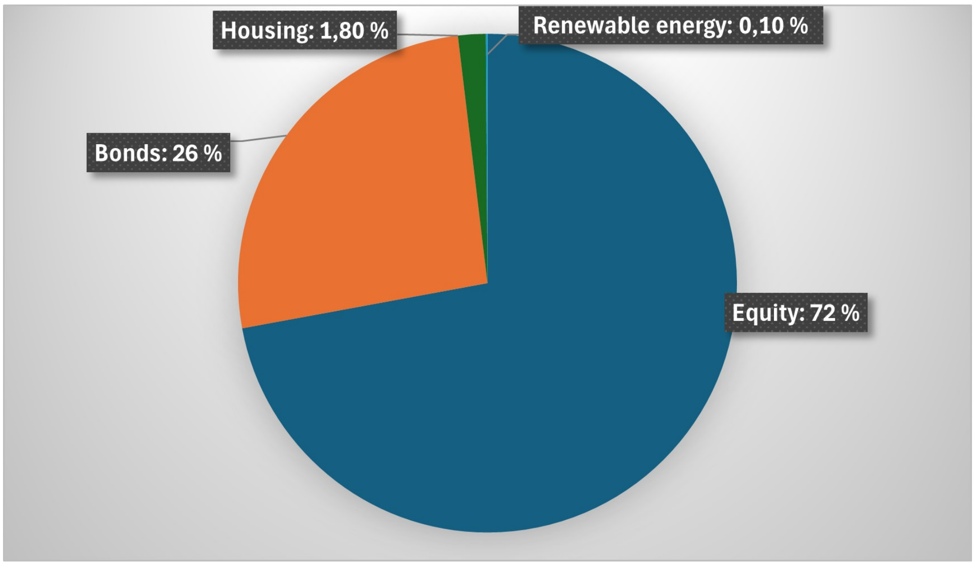
Figure 2.1.: The quarterly developement of GPFG



The figure illustrate the quarterly development of the fund, in billion NOK since 1999 until late 2023.

Norges Bank is responsible for the management of the fund, and the main board of the bank has delegated the execution of the management mandate to Norges Bank Investment Management. The majority of the fund is invested in stocks (72%), which represent ownership stakes in companies. The second biggest portion is invested in bonds (26%) , which involve lending to governments and companies. A smaller portion is invested in housing (1.8%) and infrastructure for renewable energy (0.10%) . (NBIM, 2024b). Figure 2.2 shows the capital allocation for GPFG by 1st quarter 2024 (NBIM, 2024a).

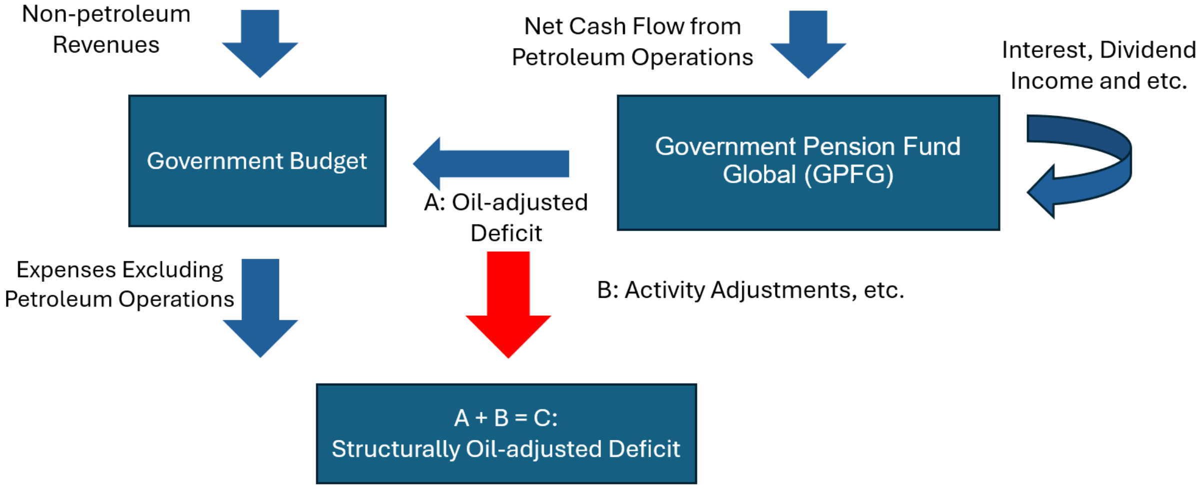
Figure 2.2.: Capital Allocation for GPFG by 1st quarter 2024



2.5.1. The Fiscal Rule

The fiscal rule is designed to ensure that the consumption out of the GPFG is sustainable over time. An important objective of the fiscal policy framework is to transform temporary revenue from the petroleum resources into a sustainable income source. This is achieved by allocating revenues from petroleum activities to the Government Pension Fund Global. Since 2001, the The Budgetary Rule has guided the use of fund resources to align with the expected real return of the fund. The goal is to smooth economic fluctuations to ensure effective capacity utilization and low unemployment. The budgetary rule limits fund usage to align with the fund's anticipated 3% real return. To adhere to this rule, withdrawals in typical years should be below this threshold (Regjeringen, 2024).

Figure 2.3.: The cash flows between the national budget and the GPFG


Non-petroleum Revenues
Net Cash Flow from Petroleum Operations
Interest, Dividend Income and etc.
Government Budget
Government Pension Fund Global (GPFG)
A: Oil-adjusted Deficit
Expenses Excluding Petroleum Operations
B: Activity Adjustments, etc.
A + B = C:
Structurally Oil-adjusted Deficit

The figure illustrates the flow of revenues from the petroleum industry into the GPFG, where returns through dividends and interest are transferred back to the national budget. The combined revenues from non-petroleum-related activities, along with returns from the GPFG, form the total national budget (Regjeringen, 2024).

2.6. Portfolio Optimization

Modern portfolio theory are based on the principle's of the American economist, Harry Markowitz (Markowitz, 1959). His theories explain the best method of balancing risk and return in an investment portfolio. The composition of investors' portfolios depends on their willingness to take on risk. Individuals have different risk profiles, with risk-averse investors aiming to preserve capital over the potential of getting a higher return (Samuelson, 1967). Investing in real estate, stocks, and other assets carries risks, as market conditions are influenced by various factors such as interest rates, inflation, geopolitical tensions, and company performance. To minimize market risk, diversification and spreading investments across different assets and industries help reduce the impact of poor performance in a single investment (Steigum, 2012). A fund consisting of a wide selection of assets offers several advantages compared to a less diversified fund. It helps reduce overall risk exposure and provides more stable returns over time (Markowitz, 1959).

The expected return of a portfolio, is the weighted sum of each security´s expected return, determined as follows:

$$\mu_p = \sum_{i=1}^{n} x_i \mu_i \tag{2.1}$$

where $x_1, x_2, \ldots, x_n$ is the number of assets, μ is the expected return to each asset, x_i represent the i-th asset in the portfolio and μ_p is the expected return to the portfolio.

The formula for the variance of a portfolio σ_p^2 is given by:

$$\sigma_p^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \cdot \operatorname{cov}(i, j) \tag{2.2}$$

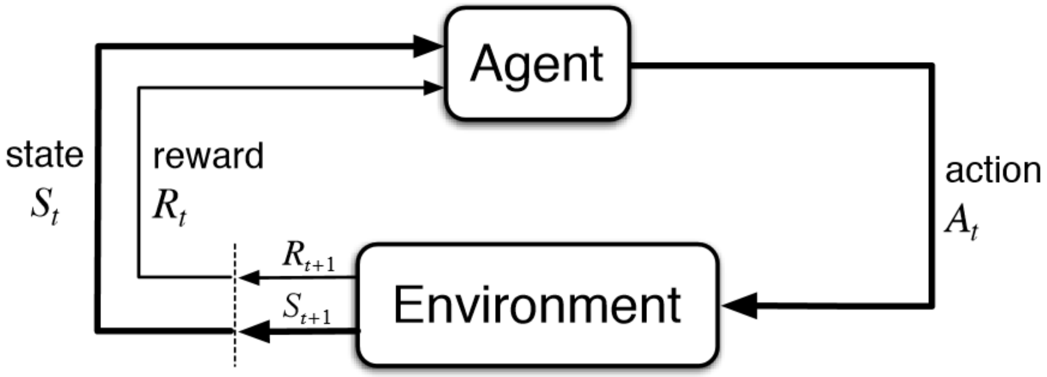
where Cov(i,j) is the covariance between the i-th and the j-th security.

2.7. Reinforcement Learning

In contrast to Markowitz, Reinforcement Learning approaches portfolio optimization by learning through a policy that interacts with an environment (Sutton and Barto, 1998). Similar to traditional portfolio theory, the RL model presented in this thesis aims to find the optimal weights. While portfolio theory seeks to identify the mean-variance optimal portfolio, RL maximizes the reward by receiving feedback through penalties. Additionally, the model accounts for consumption, an aspect not considered in Markowitz's portfolio theory.

Reinforcement learning is a topic within machine learning, differing from both supervised and unsupervised learning methods. As illustrated in Figure 2.4, reinforcement learning focuses on reaching the optimal behavior for an agent interacting with an environment. The state represents the starting point of the model. Through iterations within the environment, the agent learns to do the best actions by receiving feedback in the form of penalties and rewards. The objective for the agent is to maximize the accumulated reward. In order to function effectively within a complex environment, an agent needs to be capable of managing high-dimensional and uncertain states and actions (Tizhoosh, 2005).

Figure 2.4.: The Environment and Agent



2.7.1. Action

An action is a decision made by an agent within an environment, aiming to maximize the reward. The action taken depends on the policy, which is a strategy that maps the states of the environment to the actions the agent take. Each action influences the following state and, consequently, the next action. By continuously updating the policy, the agent can make better decisions. The goal is for the agent to develop improved strategies by learning from past feedback, and thereby aiming to maximize the reward.

2.7.2. State and Observation

State and observation represent the situation for the environment, providing the foundation for the agent to understand the environment. The agent learns by interacting with the environment and adapting its policy based on states and observations. While the state of the environment is used to determine the next observation and reward or penalty, the agent's state involves the agent's position within the environment (Sutton and Barto, 1998). Containing all the information needed for decision-making, the state is important, since the observation refers to what the agent observe or receives from the environment. In fully observable environments, the observation is equivalent to the state.

2.7.3. Reward and Penalty

The agent in RL needs to have the right feedback to know what is a good or bad move. Without proper feedback, the agent encounters difficulty in determining its actions. Employing a reward system to guide the agent towards achieving its goals stands as one of the fundamental principles in reinforcement learning (Sutton and Barto, 1998). Understanding the positive outcome of, for instance, accidentally making a good investment, or the negative outcome of a bad one, is crucial. This type of feedback is termed a reward or penalty. In games like chess, this feedback is typically received only at the end of the game, whereas in other environments, feedback comes more frequent. In the agent framework, the reward is viewed as part of the input perception, but the agent must be programmed to recognize it as a reward rather than just another sensory input (Russell and Norvig, 2010) .

3. Modeling SWF as Markov Decision Problem

In this chapter, the Markov Decision problem will be introduced and the extensions proposed by Knut Anton Mork (Mork et al., 2023). The Markov Decision Problem is a mathematical framework for decision-making in situations where the outcome is random or controlled by an agent. It is modeled as a discrete-time stochastic control process (Puterman, 1990). The chapter will start by giving an introduction to Markov Decision Problem. An introduction to Geometric Brownian motion which is a mathematical model used to describe the random movement of prices in financial markets, will also be given.

3.1. Markov Decision Process

The Bellman equation is central in the Markov Decision Processes (MDP). It offers a structured approach for determining the optimal reward for an agent within a specific state. The agent's objective is to maximize this reward by selecting the most advantageous actions in all future decisions (Barron and Ishii, 1989) (Leite et al., 2020). The Bellman equation can be described as follows:

$$V(s) = \max[R(s, a) + \gamma \cdot V(s')] \tag{3.1}$$

Where *V* is the expected value or return at the current state *s* and *R(s, a)* is the expected reward by taking an action *a* at state *s*. The last term in the equation is the discount factor γ multiplied by the next state.

Furthermore, the mathematical formulations are inspired by earlier studies. Mainly by (Mork et al., 2023), and (Garcia and Rachelson, 2013). The model want to find the best trade-off between near-future consumption and the maintenance of a funds value that allows consumption by future generations.

The MDP is a sequential model that take decisions in discrete and stochastic environments. The model formulate an environment that changes the state as a response on the agents actions. Discrete and equidistant points in time are considered at $t = 0, 1, \ldots, T$.

The Markov Decision problem can be described by:

$$MDP = (S, A, F, R, \rho) \tag{3.2}$$

Where,

- S is a set of states
- A is a set of actions
- F s a state-transition model
- R is a reward function
- And ρ is a discount factor, where $0 < \rho \leq 1$

States $S_t \in S$ can be described by different quantities, that influence the environment. Like amount of wealth, habit, macroeconomic variables etc.

The states evolve according to some transition model F: $S \times A \times [0, 1] \to S, \quad$ i.e.:

$$S_{t+1} = F(S_t, A, u_t) \tag{3.3}$$

where u_t is a uniformly distributed random variable in the interval [0,1]. In the transition model there can be some mechanism that transforms this variable into a random variable of other distribution.

The reward is generated by some function R: S $\times A \times [0, 1] \to \mathbb{R}$

$$R_{t+1} = R : S \times A \times [0,1] \rightarrow \mathbb{R} \tag{3.4}$$

Actions $a \in A$ are determined by policies π, and are generally differentiate deterministic and stochastic actions, i.e. π(s) = a (given some state s we take action a), or π(a | $s) = P(A_t = a \mid S_t = s$ (given some state s, we take action a with probability π is a distribution over actions).

3.1.1. Emulation of the Model by Mork

In the following section the model introduced by Mork will be replicated as a Markov decision process. The goal is to analyze if a formulation as an MDP can be solved using suitable RL algorithms and whether comparable results can be obtained. The actions taken in each time step are the rate of consumption $\alpha_t \in [0,1]$ to be invested into the risky asset classes. These actions are determined by some policy $(\mathrm{c}_t, \alpha_t) = \pi(\mathrm{s}_t)$, which tells which actions c_t and α_t to take in some time step after observing a state $\mathrm{s}_t \in S$.

Following the analysis of Mork, we assume that the state is defined solely by the accumulated wealth w_t at point in time t: $\mathrm{s}_t = w_t$.

The transition of wealth is modeled as follows:

$$w_{t+1} = w_t + (1 - \alpha_t) \cdot r_f \cdot w_t + \alpha_t \cdot \mu \cdot w_t + \alpha_t \cdot \sigma \cdot \varepsilon_t \cdot w_t - c_t \tag{3.5}$$

Where r_f is the risk free rate, μ is the determines the trend of the risky asset (expected risky return over the time interval from t to t + 1, ρ drives the volatility (standard deviation), and $_t$ is a standard-normally distributed random variable.

We can retrieve this transition with respect to the equity premium π = μ - r_f:

$$w_{t+1} = w_t + (r_f + \alpha_t \cdot \pi) \cdot w_t + \alpha_t \cdot \sigma \cdot \varepsilon_t \cdot w_t - c_t \tag{3.6}$$

Expression 3.5 and 3.6 have some shortcomings. The first shortcoming is that consumption is not reducing the wealth before reinvesting; Basically, consumption appears one period later before the next decision. The second shortcoming is that consumption appears as an absolute number. Together, with the first shortcoming this implies that wealth can become negative, after the return of the risky asset is observed.

To avoid the first shortcoming:

$$w_{t+1} = (w_t - c_t) + \alpha_t \cdot r_f \cdot (w_t - c_t) + \alpha_t \cdot \mu \cdot (w_t - c_t) + \alpha_t \cdot \sigma \cdot \varepsilon_t \cdot (w_t - c_t) \tag{3.7}$$

There is still missing some constraint $w_t \geq \mathrm{c}_t$, to implement this, the new equation will be:

$$w_{t+1} = (1 - \eta_t) \cdot w_t + \alpha_t \cdot r_f \cdot (1 - \eta_t) \cdot w_t + \alpha_t \cdot \mu \cdot (1 - \eta_t) \cdot w_t + \alpha_t \cdot \sigma \cdot \varepsilon_t \cdot (1 - \eta_t) \cdot w_t \tag{3.8}$$

Expressed in a shorter way:

$$w_{t+1} = (1 - \eta_t) \cdot w_t \cdot (1 + \alpha_t \cdot r_f + \alpha_t \cdot \mu + \alpha_t \cdot \sigma \cdot \varepsilon_t) \tag{3.9}$$

3.1.2. Utility Function by Mork

A disadvantage of this formulation can be the multiplicative connection of the decision variables. To remedy these shortcomings, there is a better way of expression. In this MDP the reward equals utility (i.e. the reward function is a utility function) which depends on consumption and habit. It is modeled the following way:

$$u_t = \begin{cases} \frac{1}{1-\gamma_1}\left[\left(\frac{c_t}{x_t}\right)^{1-\gamma_1} - 1\right] & \text{if } c_t < x_t \\ \frac{1}{1-\gamma_2}\left[\left(\frac{c_t}{x_t}\right)^{1-\gamma_2} - 1\right] & \text{if } c_t \geq x_t \end{cases} \tag{3.10}$$

In this expression the parameters are: $\gamma_1 > \gamma_2, \quad$ and $\quad \gamma_1, \gamma_2 \neq 1$.
For $\gamma_1, \gamma_2 > 0$, for the utility function to have the property $u_t'' < 0$.

If $\gamma_1, \gamma_2 > 1$, then $\left(\frac{c_t}{x_t}\right)^{1-\gamma_1} = \left(\frac{x_t}{c_t}\right)^{\gamma_1-1}$ tends to infinity as c_t tends to zero. For numerical solution methods, it is therefore wise to set a lower bound on c_t.

The habit is modeled as follows:

$$x_{t+1} = x_t \cdot (1 + g) \tag{3.11}$$

Where g represents an externally given growth-rate.

The total return (an expression used in reinforcement learning or MDP, use the term ´payoff´) is:

$$P = \sum_{t=0}^{\infty} (1+\theta)^{-t} \cdot u_t \tag{3.12}$$

Where θ is a subjective discount rate. This is the expression the model wants to maximize.

3.2. Geometric Brownian Motion

The Merton model is characterized by Geometric Brownian motion (GBM). GBM implies that stock prices follow a stationary pattern that is log-normally distributed, indicating that price changes are unpredictable and random (Merton, 1975). In financial markets, a commonly accepted hypothesis concerning stock price movements is that they adhere to a random walk of returns, or in continuous-time formulation, GBM. Mork's methodology can be described by two Constant Relative Risk Aversion (CRRA) utility functions. These utility functions are identical except for the curvature parameter . This suggests that the utility function with lower consumption rates is more robust than the one with higher levels of consumption rates.

The idea behind the model is a level of consumption, denoted by $x > 0$, and the utility function that flatter even more when the consumption rate exceed the norm. The utility function is described as follows:

$$u'(c) = c^{-\gamma(c)} \tag{3.13}$$

In the equation above, y is a decreasing function that depends on the norm of x.

3.3. Implementation of new Parameters into the Model

For this thesis, it is necessary to do some changes and adjustments to make the model even more realistic for a SWF. The mathematical formulations are inspired by (Becker, 2024). The changes and adjustments are presented below.

3.3.1. Soft Habit

The model to Mork is missing a central part within the domain of habit, which can be named ”soft habit”. The only difference to Morks habit (Equation 3.11) is that the habit is now adjusted to the consumption. This is modelled in the following way:

$$x_{t+1} = \theta \cdot c_t + (1 - \theta) \cdot x_t \tag{3.14}$$

This correspond to: $x_{t+1} = x_t + \Delta x_t, \quad \text{where } \Delta x_t = \theta \cdot (c_t - x_t)$

3.3.2. New Reward Function

The model consist of a developed reward function that have some elements:

- The utility u_t, obtained from the consumption of wealth.
- A penalty $P_S,_t$ for smoothing of consumption.
- A penalty $P_W,_t$ if the agent does not meet the target wealth.
- A penalty P_D, t for deviation from decision variables.

The new reward function is therefore formulated as the utility subtracted by penalties:

$$R_t = u_t - P_{St} - P_{Wt} - P_{Dt} \tag{3.15}$$

3.3.3. New Utility Function

The isoelastic function is designed to express utility as a function of consumption or other economic variables, depending on the specific economic issues faced by the decision-maker. A utility function similar to that described by Mork uses pronounced gradients that challenges the learning process. An example of this can be fined in the Appendix B.19. The model consist, therefore of a new utility function. The isoelastic function represents a unique form of hyperbolic risk aversion, where risk aversion increases as wealth increases (Ingersoll, 1987).

$$u(t) = \frac{(C_t + 1)^{\tau} - 1}{1 - \tau} \tag{3.16}$$

where $\tau > 0$ and $\tau \neq 1$ are constants that describe the degree of relative risk aversion (Becker, 2024).

The function involves two inputs: τ (eta) and C (consumption). Here, τ represents the elasticity of marginal utility with respect to income.

3.3.4. Further Explanation of Penalties

The penalty $P_S,_t$ is incurred when the agent experiences a significant negative deviation from previous consumption levels. This penalty serves the purpose of maintaining or increasing the level of wealth.

$$P_{St} = \lambda_S \cdot \max(C_{t-1} - C_t, 0)^2 \tag{3.17}$$

In this equation, λ is the parameter to decide. Squaring the equation ensures that larger deviations have a greater impact compared to small deviations.

The target value penalty P_{Wt}, where P_{Wt} = 0 for all t to t = T-1, and $P_{Wt} \geq 0$ for t = T, have the purpose to ensure that the agent does not use up the funds wealth.

$$P_{Wt} = \lambda_W \cdot \max(\hat{W} - W_t, 0)^2 \tag{3.18}$$

Where $\hat{w}$ represents the target wealth at the end of the period, and λ_W determines the extent of the penalty. The square function serves the same purpose as in the previous equation.

Finally, a penalty P_{Dt} is applied for violating the constraints on the decision variables. Similar to the constraints for the Government Pension Fund Global (GPFG), there are specific limits on the maximum percentages that can be invested in each asset class. The precise details of these constraints will be outlined in the data section. The penalty is determined as follows:

$$PD_t = \max(x^- - x_t, 0)^2 + \max(x_t - x^+, 0)^2 \cdot W_{t-1} \cdot \lambda_D \tag{3.19}$$

In this equation x^- and x^+ represent the limitations for the shares of the assets, and λ_D is the penalty parameter.

3.3.5. Action Space

Within the action space, there exists a consumption rate $c_t \geq 0$, where the percentage of wealth at the beginning of each period is allocated for consumption. Additionally, the model have proportions of different assets, with each asset in the portfolio having a share denoted by $0 \leq x_t \leq 1$. The Sovereign Wealth Fund (SWF) will be invested in four asset classes, with the decision vector represented as $x_t = [x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}]$.

3.4. Optimal Asset Allocation for SWF

A Sovereign Wealth Fund (SWF) can be defined as a portfolio consisting of diverse assets. The model is a discrete time model where u_t is the funds value at time t, at the beginning of a period [t, t+1] and Y_t is the income allocated to the fund at time t. Let C_t denote consumption of the fund over the interval [t, t+1]. Financial markets consists of a theoretical risk free asset and *n* risky assets. The return on the risk free and risky asset *I, i, . . . , n* over a given interval [t, t+1], are denoted $R_{0,t}$ and $R_{i,t}$.

The funds progress from time t to time t+1 is given by

$$F_{t+1} = ((F_t + Y_t - C_t)R_{F,t+1} \tag{3.20}$$

Where $R_{F,t+1}$ is the return on the funds portfolio. The return for a portfolio is determined in the following way:

$$R_{F,t+1} = \pi_{0,t}R_{0,t+1} + \pi_t R_{t+1} = R_{0,t+1} + \sum_{i=1}^{n} \pi_{i,t}(R_{i,t+1} - R_{0,t+1}) \tag{3.21}$$

In Equation 3.21, $\pi_{0,t}$ is the weight of the risky asset, where $\pi_{i,t}$ is the weight of the i-th risky asset. $R_{i,t+1} - R_{0,t+1}$ represents the excess return of asset i. The excess return on an asset is the investments return minus the risk free rate (Moutanabbir and Noureldin, 2020).

4. Reinforcement Learning Algorithms

In this chapter, a brief introduction to the five algorithms used in this experiment will be provided. The model are based on algorithm's from the PyTorch library *Stable-baselines3*, which is a framework for machine learning, and commonly used for reinforcement learning. Stable-baseline3 includes many different algorithms. However, the model can not run all of them, because the models uses continuous real numbers. The model only runs the algorithm's that have a N-dimensional box that contains every point in the action space. Traditional methods like Deep Q Network (DQN) work well for discrete actions, but not for continuous actions. The algorithms contains of an actor and a critic where the actor decides the actions, and the critic evaluates and give the actor rewards, based on their performance (Raffin et al., 2021).

The algorithm´s that will be tested are:

- DDPG
- PPO
- A2C
- SAC
- TD3

4.1. Deep Deterministic Policy Gradient (DDPG)

DDPG stands for *Deep Deterministic Policy Gradient* which learns a Q-function and a policy based on the Bellman equation 3.1. It uses the Bellman equation to improve its Q-function. With the Q-function DDPG estimates a reward and actions for each states, and the agent make decisions based on the values in the Q-function (Lillicrap et al., 2015).

DDPG is an actor-critic algorithm that is model free and off-policy which use deep function approximators to learn policies in a high-dimensional action space. Consider a agent interacting with an environment E in discrete time-steps. The agent receive an observation x_t, take an action a_t and receive a reward r_t (Lillicrap et al., 2015). The action value function can be described as follow:

$$Q\pi(s_t, a_t) = \mathbb{E}_{r_{i \geq t}, s_i > t \sim E, a_i > t \sim \pi}[R_t | s_t, a_t] \tag{4.1}$$

Where $\mathbb{E}$ is expectation, π is the policy and γ is the discount factor, for all real numbers R.

Further more there is possible to decompose the Bellman equation into more parts.

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{r_t, s_{t+1} \sim E}[r(s_t, a_t) + \gamma \mathbb{E}_{a_{t+1} \sim \pi}[Q_{\pi}(s_{t+1}, a_{t+1})]] \tag{4.2}$$

Since the target policy is deterministic the function can be described as: $\mu : S \leftarrow A$. This is to avoid the inner expectation.

$$Q^{\mu}(s_t, a_t) = \mathbb{E}_{r_t, s_{t+1} \sim E}[r(s_t, a_t) + \gamma Q_{\mu}(s_{t+1}, \mu(s_{t+1}))] \tag{4.3}$$

The equation above shows that the expectation only depends on the environment. In this way there is possible to learn a off-policy (Q^{μ}) using a different stochastic behavior policy (β). A often used off-policy algorithm uses a greedy policy $\mu(s) = \text{argmax}_a Q(s, a)$.

$$L(\theta^Q) = \mathbb{E}_{s_t \sim \rho^{\beta}, a_t \sim \beta, r_t \sim E}\left[\left(Q(s_t, a_t | \theta^Q) - y_t\right)^2\right] \tag{4.4}$$

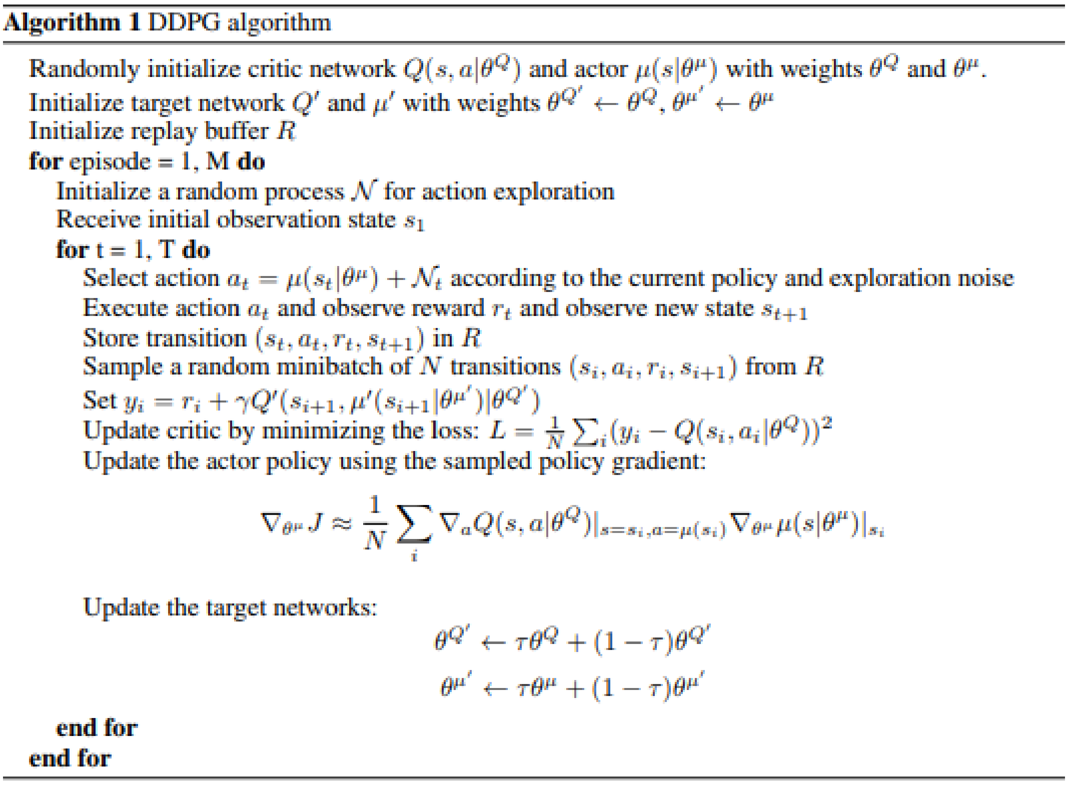
where

$$y_t = r(s_t, a_t) + \gamma Q(s_{t+1}, \mu(s_{t+1}) | \theta^Q)$$

To make it possible to use DDPG in continuous action spaces:

$$\nabla_{\theta \mu} \approx \mathbb{E}_{s_t \sim \rho^{\beta}}\left[\nabla_{\theta \mu} Q(s, a | \theta^Q) | s = s, a = \mu(s_t | \theta^{\mu})\right] \tag{4.5}$$

Figure 4.1.: DDPG - Pseudo-code



```
Algorithm 1 DDPG algorithm
Randomly initialize critic network $Q(s,a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights $\theta^Q$ and $\theta^\mu$.
Initialize target network $Q'$ and $\mu'$ with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer $R$
for episode = 1, M do
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state $s_1$
  for t = 1, T do
    Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise
    Execute action $a_t$ and observe reward $r_t$ and observe new state $s_{t+1}$
    Store transition $(s_t, a_t, r_t, s_{t+1})$ in $R$
    Sample a random minibatch of $N$ transitions $(s_i, a_i, r_i, s_{i+1})$ from $R$
    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
    Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$
    Update the actor policy using the sampled policy gradient:
        $\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s,a|\theta^Q)|_{s=s_i,a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_i}$
    Update the target networks:
        $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$
        $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$
  end for
end for
```

Figure shows an example of a DDPG pseudo-code (Lillicrap et al., 2015)

4.2. Proximal Policy Optimization (PPO)

Proximal Policy Optimization (PPO) is a new and simpler edition of the TRPO algorithm, better known as *Trust Region Policy Optimization*. It combines ideas from both A2C and DDPG, by having multiple workers and using trust solution to improve the actor. PPO are using a first-order optimization, while TRPO are using a second-order optimization. This make PPO more simple to implement than TRPO, and having at least as good performance. (Schulman et al., 2017).

A general policy optimization model starts by defining the policy gradient loss as the expected value of the logarithm of the policy action, multiplied by an estimate of the advantage function.

$$L^{PG}(\theta) = \hat{E}_t \left[\log \pi_\theta (a_t | s_t) \hat{A}_t \right] \tag{4.6}$$

where pi_θ is the policy, which is a neural network that takes the observed states from the environment as an input, and suggest actions to take as an output. The second term is the advantage function A which tries to estimate what the relative value of the selected action in the current state.

In order to compute the advantage A, two things is required; the discounted sum of the rewards and a baseline estimate. The discounted sum of reward (return) is the weighted sum of all the rewards the agent got during each time step.

$$\sum_{k=0}^{\infty} \gamma^k \cdot r_t + k \tag{4.7}$$

Where γ is the discounted value of future rewards, and the value of receiving a reward R after *k+1* time-step is $\gamma^k \cdot r$. The value of γ is between 0.9 and 0.99 which means that the agent cares more about the reward it get quickly compared to rewards it gets many time-steps in the future.

The second part of the advantage function A is the value function (baseline). The value function tries to give an estimate of the discounted sum of rewards from time-step k and onwards. Because the value function is an estimate, the function will not always predict the exact value of the current state, which give some noisy estimate. In sum, the advantage function is the discounted reward subtracted by the baseline estimate.

$$\hat{A}_t \ = \text{Discounted reward - Baseline estimate} \tag{4.8}$$

When the advantage estimate $\hat{A}_t$ is positive, it indicates that the actions taken by the agent led to outcomes better than the average return. This increases the likelihood of selecting those actions again in the same state. On the other hand, if the advantage estimate is negative, the likelihood of selecting those action again decreases (Schulman et al., 2017).

A challenge associated with running gradient descent on a single batch of collected experience is that the parameter updates within the neural network can extend far beyond the expected range. This occurs due to the inherent noise in estimating the actual advantage, and in the end destroy the policy.

To prevent the issue when updating the policy and never move to far from the old policy, the solution is TRPO. To avoid that the updated policy does not move to far from the old policy, TRPO adds a KL constraints to the objective function.

$$\begin{aligned} &\text{maximize}\, \theta = \hat{E}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t \right] \\ &\textit{subject to}\, \hat{E}_t \left[KL \left[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t) \right] \right] \leq \delta. \end{aligned}$$

With PPO it is possible to add the new constraint directly into the optimization objective. Let $r_t\theta$ denote the probability ratio between the new updated policy outputs and the previous old version of the policy network.

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \,\text{so}\, r(\theta_{old}) = 1 \tag{4.9}$$

For a sample of actions and states the $r_t(\theta)$ will be larger than 1 if the current action is more likely now than in the old policy, and between 0 and 1 if the current action is less likely than previous. By multiplying the advantage function with $r_t(\theta)$ the output will be:

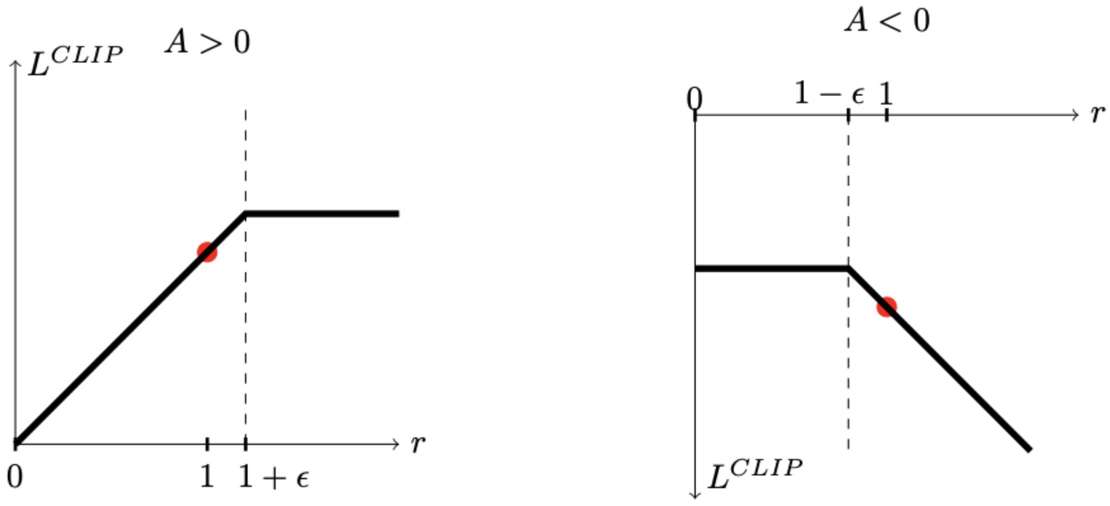
$$L^{CPI}(\theta) = \hat{E}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t \right] = \hat{E}_t \left[r_t(\theta) \hat{A}_t \right] \tag{4.10}$$

From this equation the main objective function for PPO can be deduced.

$$L^{CLIP}(\theta) = \hat{E}_t \left[\min(r_t(\theta)\hat{A}_t\, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t) \right] \tag{4.11}$$

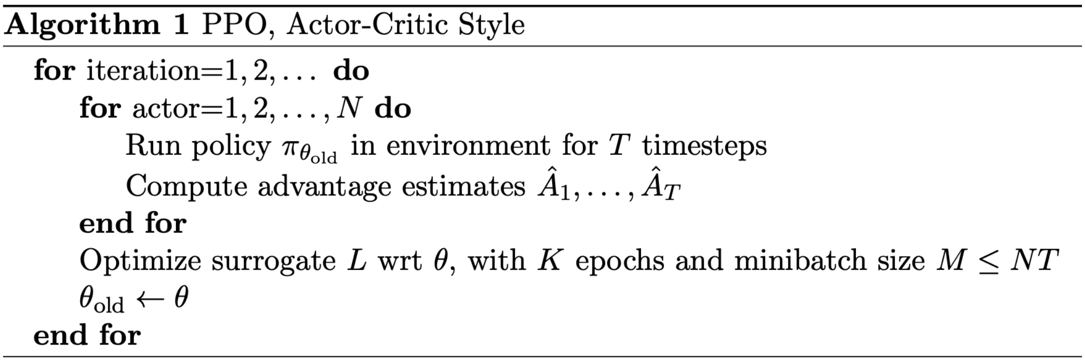
As the equation above shows, PPO try to optimize an expectation operator over batches of trajectories. Where the expectation operator is taken over the minimum of two terms. The first term $r_t(\theta)\hat{A}_t$ is the default objective for normal policy gradient. The second term is a clipping operation between 1 - ϵ and 1 + ϵ, where epsilon is a hyper parameter, often with the value of 0.2. The advantage estimate $\hat{A}_t$ can both be positive and negative which changes the effect on the main operator (Schulman et al., 2017).

Figure 4.2.: Effect on objective function



Plots showing the effect on the objective function for both positive and negative values. In the left plot the actions have better than expected return. In the right plot actions have worse than expected return. When r gets to high, the function flattens, which happens when the action is more likely in the new policy than the old policy. Therefore the function get ”clipped” to limit the effect of the gradient update. The opposite occurs when the actions had a negative estimated value. The function flattens when r gets near zero. This applies to actions that are less likely to happen in the new policy than in the old one.

Figure 4.3.: PPO - Pseudo-code



```
Algorithm 1 PPO, Actor-Critic Style
  for iteration=$1, 2, \ldots$ do
      for actor=$1, 2, \ldots, N$ do
          Run policy $\pi_{\theta_{\text{old}}}$ in environment for $T$ timesteps
          Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
      end for
      Optimize surrogate $L$ wrt $\theta$, with $K$ epochs and minibatch size $M \leq NT$
      $\theta_{\text{old}} \leftarrow \theta$
  end for
```

Figure shows an example of a PPO pseudo-code (Schulman et al., 2017)

4.3. Advantage Actor-Critic (A2C)

Advantage Actor-Critic (A2C) is an deterministic variant of Asynchronous Advantage Actor Critic (A3C). It updates the model parameters simultaneously using data gathered from multiple agents, which are running in parallel. In contrast, A3C is an asynchronous algorithm where each agent interacts with its own instance of the environment independently and updates the model network asynchronously, without waiting for the other agents to finish their updates. Because of this, A3C has in general a faster running speed, but A2C tends to be more stable due to their synchronicity (Mnih et al., 2016).

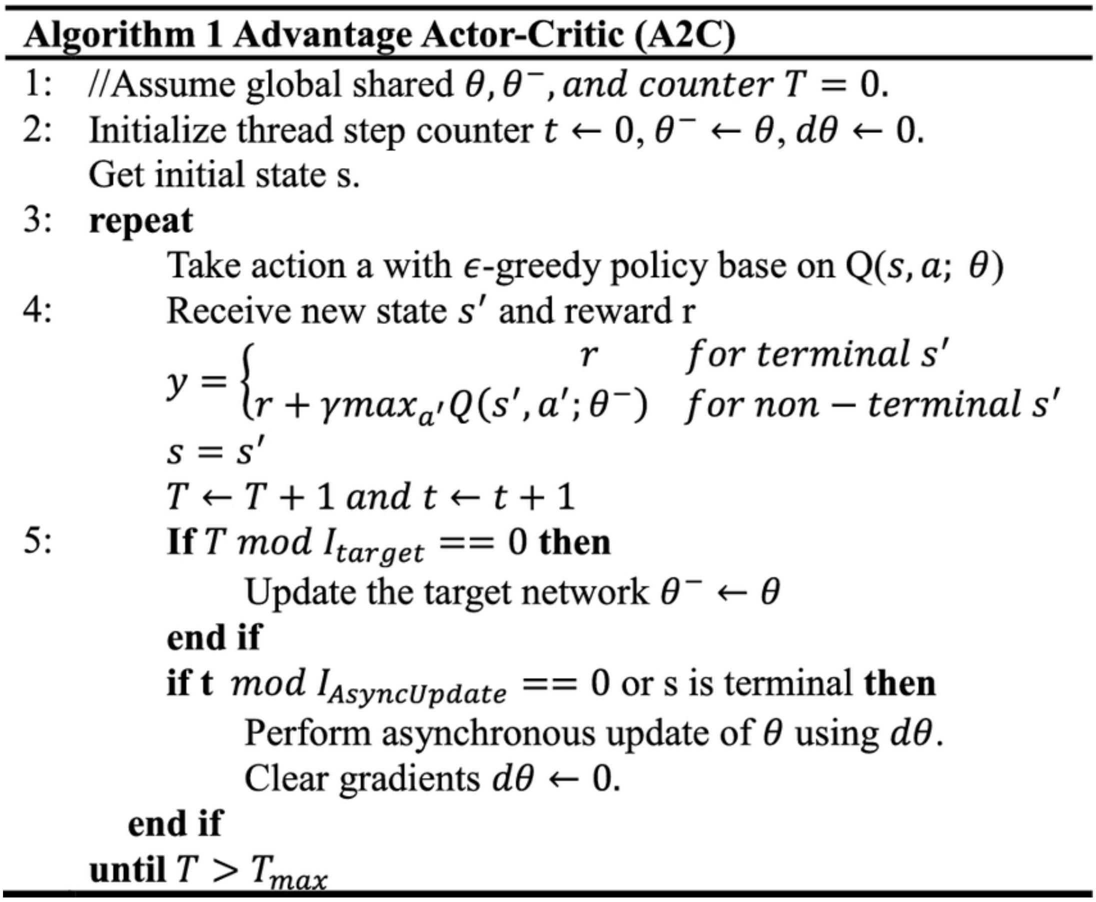
The theoretical background for A2C are based on A3C which maintains the policy $\pi(a_t|s_t;\theta)$ and an estimate for the value function $V(s_t;\theta_v)$. To update the policy and the value function the algorithm uses a mix of n-step returns. The updates persists until each *t_max* or until the terminal state are reached.

The performed update can be explained as $\nabla_{\theta'}\log\pi(a_t|s_t;\theta)A(s_t,s_t;\theta,\theta_v)$ where the last term is an estimate of the advantage function:

$$\sum_{i=0}^{k-1}\gamma^i\cdot r_{t+i}+\gamma^k V(s_{t+k};\theta_v)-V(s_t;\theta v) \tag{4.12}$$

Where the function is limited by the upper-bound *t_max* and k vary from state to state.

Figure 4.4.: A2C - Pseudo-code



```
Algorithm 1 Advantage Actor-Critic (A2C)
//Assume global shared $\theta, \theta^-, and\ counter\ T = 0$.
Initialize thread step counter $t \leftarrow 0, \theta^- \leftarrow \theta, d\theta \leftarrow 0$.
Get initial state s.
repeat
    Take action a with $\epsilon$-greedy policy base on Q($s, a;\ \theta$)
    Receive new state $s'$ and reward r
    $y = \begin{cases} r & for\ terminal\ s' \\ r + \gamma max_{a'} Q(s', a'; \theta^-) & for\ non - terminal\ s' \end{cases}$
    $s = s'$
    $T \leftarrow T + 1\ and\ t \leftarrow t + 1$
    If $T\ mod\ I_{target} == 0$ then
        Update the target network $\theta^- \leftarrow \theta$
    end if
    if t $mod\ I_{AsyncUpdate} == 0$ or s is terminal then
        Perform asynchronous update of $\theta$ using $d\theta$.
        Clear gradients $d\theta \leftarrow 0$.
  end if
until $T > T_{max}$
```

Figure shows an example of a A2C pseudo-code (Day et al., 2023)

4.4. Soft Actor Critic (SAC)

The Soft Actor-Critic algorithm (SAC) is a deep reinforcement learning (RL) algorithm that extends the maximum entropy reinforcement learning framework. In SAC, the agent aims to maximize both rewards and entropy simultaneously. SAC, unlike other RL algorithms such as PPO and A3C, does not require new samples for each gradient update giving it a low sample efficiency. SAC addresses the challenge of increasing computational costs with task complexity by reducing the need for numerous gradient steps and samples per step, to learn an effective policy (Haarnoja et al., 2018).

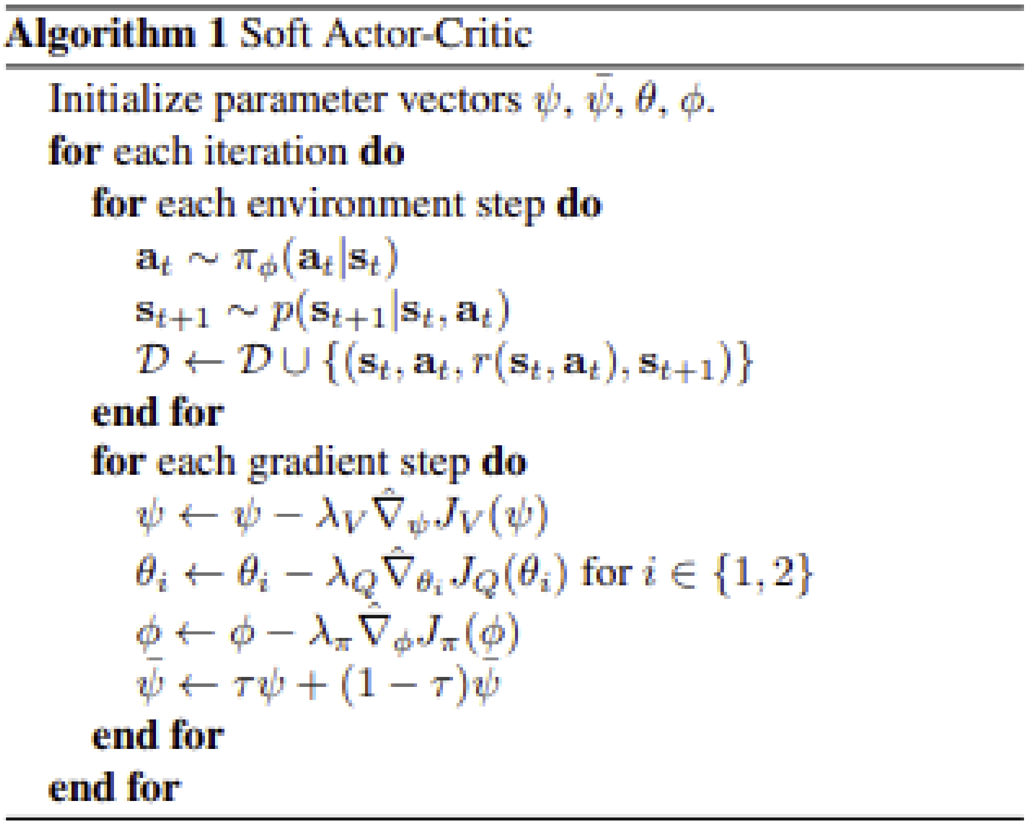
SAC diverges from other RL methods, particularly those formulated as Q-learning, by striving to achieve task success with as much randomness in decision-making as possible. SAC employs a stochastic actor, which contributes to more stable results across various random seeds, surpassing earlier on-policy and off-policy approaches. This algorithm has demonstrated superior performance and robustness compared to previous methods, particularly in scenarios requiring continuous control tasks (Haarnoja et al., 2018).

SAC learns a stochastic policiy that are maximizing both rewards and entropy where the expected entropy of the policy over $\rho_\pi(s_t)$):

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))] \tag{4.13}$$

α is the temperature parameter that represents the importance of the entropy term relative to the reward.

Figure 4.5.: SAC - Pseudo-code



```
Algorithm 1 Soft Actor-Critic
  Initialize parameter vectors $\psi$, $\bar{\psi}$, $\theta$, $\phi$.
  for each iteration do
    for each environment step do
      $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
      $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
    end for
    for each gradient step do
      $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
      $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
      $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
      $\bar{\psi} \leftarrow \tau\psi + (1 - \tau)\bar{\psi}$
    end for
  end for
```

Figure shows an example of a SAC pseudo-code (Haarnoja et al., 2018)

4.5. Twin Delayed DDPG (TD3)

Twin Delayed DDPG (TD3) addresses several issues associated with the original DDPG algorithm. One significant problem with the DDPG algorithm is that its Q-function frequently overestimates Q-values, which can result in policy degradation. Twin Delayed DDPG employs three strategies to tackle this issue effectively (Fujimoto et al., 2018).

- **Trick One: Clipped Doubble Q Learning:** The algorithm uses two instead of one Q-function (DDPG) and uses the smallest Q-value as target in the Belleman error loss function.
- **Trick Two: "Delayed" Policy Updates:** The Q function updates way faster than the policy. For every two Q-function updates, there should be one policy update. (Fujimoto et al., 2018)
- **Trick Three: Target Policy Smoothing:** Twin delayed DDPG adds noise to the target action. To make it more challenging for the policy to exploit Q-function errors, a technique involves smoothing out the Q-function along changes in action.

Target policy smoothing involves adjusting actions for the Q-learning target based on the target policy $\mu_{\theta_{\text{targ}}}$ with clipped noise added to each action dimension. The resulting target action is clipped to the valid action range defined by $a_{\text{Low}} \leq a \leq a_{\text{High}}$:

$$a'(s') = \text{clip}\left(\mu_{\theta_{\text{targ}}}(s') + \text{clip}(\epsilon, -c, c), a_{\text{Low}}, a_{\text{High}}\right), \quad \epsilon \sim \mathcal{N}(0, \sigma) \tag{4.14}$$

Clipped double-Q learning utilizes a single target $y(r, s', d)$ computed using the smaller value between two Q-functions $Q_{\phi_{1,\text{targ}}}$ and $Q_{\phi_{2,\text{targ}}}$:

$$y(r, s', d) = r + \gamma(1 - d) \min_{i=1,2} Q_{\phi_i}(s', a'(s')) \tag{4.15}$$

and then both are learning by regressing to this target:

$$L(\phi_1, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}}\left[\left(Q_{\phi_1}(s, a) - y(r, s', d)\right)^2\right] \tag{4.16}$$

$$L(\phi_2, \mathcal{D}) = \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}}\left[\left(Q_{\phi_2}(s,a) - y(r,s',d)\right)^2\right] \tag{4.17}$$

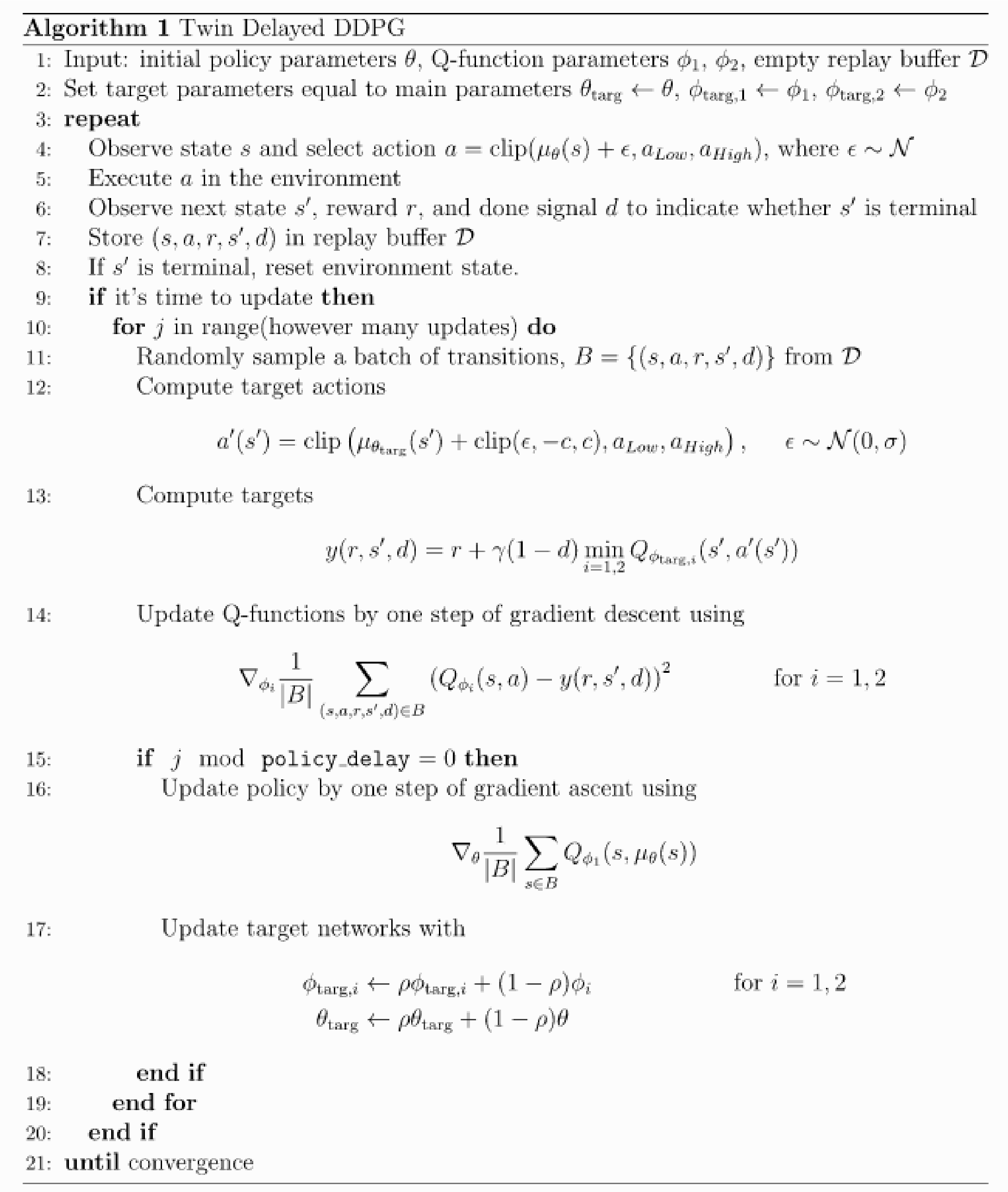
By using the smaller Q-value for the target, this approach avoid overestimation in the Q-function.

The policy is learned by maximizing Q_{ϕ_1} with respect to the policy parameters θ:

$$\max_{\theta} \mathbb{E}_{s\sim\mathcal{D}}\left[Q_{\phi_1}(s, \mu_\theta(s))\right] \tag{4.18}$$

In Twin delayed DDPG, the policy is updated less frequently than the Q-functions to reduce volatility resulting from policy updates (Fujimoto et al., 2018).

Figure 4.6.: TD3 - Pseudo-code



Algorithm 1 Twin Delayed DDPG

```
Input: initial policy parameters $\theta$, Q-function parameters $\phi_1$, $\phi_2$, empty replay buffer $\mathcal{D}$
Set target parameters equal to main parameters $\theta_{\text{targ}} \leftarrow \theta$, $\phi_{\text{targ},1} \leftarrow \phi_1$, $\phi_{\text{targ},2} \leftarrow \phi_2$
repeat
  Observe state $s$ and select action $a = \text{clip}(\mu_\theta(s) + \epsilon, a_{Low}, a_{High})$, where $\epsilon \sim \mathcal{N}$
  Execute $a$ in the environment
  Observe next state $s'$, reward $r$, and done signal $d$ to indicate whether $s'$ is terminal
  Store $(s, a, r, s', d)$ in replay buffer $\mathcal{D}$
  If $s'$ is terminal, reset environment state.
  if it's time to update then
    for $j$ in range(however many updates) do
      Randomly sample a batch of transitions, $B = \{(s, a, r, s', d)\}$ from $\mathcal{D}$
      Compute target actions
          $a'(s') = \text{clip}\left(\mu_{\theta_{\text{targ}}}(s') + \text{clip}(\epsilon, -c, c), a_{Low}, a_{High}\right), \quad \epsilon \sim \mathcal{N}(0, \sigma)$
      Compute targets
          $y(r, s', d) = r + \gamma(1 - d) \min_{i=1,2} Q_{\phi_{\text{targ},i}}(s', a'(s'))$
      Update Q-functions by one step of gradient descent using
          $\nabla_{\phi_i} \frac{1}{|B|} \sum_{(s,a,r,s',d) \in B} \left(Q_{\phi_i}(s, a) - y(r, s', d)\right)^2 \qquad \text{for } i = 1, 2$
      if $j$ mod policy_delay $= 0$ then
        Update policy by one step of gradient ascent using
            $\nabla_\theta \frac{1}{|B|} \sum_{s \in B} Q_{\phi_1}(s, \mu_\theta(s))$
        Update target networks with
            $\phi_{\text{targ},i} \leftarrow \rho \phi_{\text{targ},i} + (1 - \rho)\phi_i \qquad \text{for } i = 1, 2$
            $\theta_{\text{targ}} \leftarrow \rho \theta_{\text{targ}} + (1 - \rho)\theta$
      end if
    end for
  end if
until convergence
```

Figure shows an example of a Twin Delayed DDPG pseudo-code (Fujimoto et al., 2018)

5. Data

This chapter will present the inputs used in the code. This includes the returns and standard deviations for the four assets: bills, bonds, equity, and housing. Additionally, the model has several other input parameters and constraints that will be described.

5.1. The Four Asset Model Inputs

Most SWFs allocate their capital across various asset classes. These assets are often diversified across multiple markets to capture the overall returns within the respective asset classes. In our RL environment, we will use historical standard deviations and returns for the different asset classes. Calculating the return and standard deviation for the entire market within an asset class can be challenging, as it involves weighing all markets globally against each other. The thesis will therefore rely on the article 'The Rate of Return on Everything, 1870–2015' to find standard deviations and returns for the different markets (Jordà et al., 2019).

Table 5.1.: Nominal Standard Deviation and Returns

	Bills	Equity	Bonds	Real Estate
Returns	4.58%	10.65%	6.06%	11.00%
St. Deviation	3.32%	22.55%	8.88%	10.64%

The table illustrate the nominal returns and standard deviation for the four asset classes taken from the article 'The Rate of Return on Everything, 1870–2015' (Jordà et al., 2019)

5.1.1. Equity

The global stock market has witnessed significant returns, with nominal returns exceeding 12.97% after 1950 (Jordà et al., 2019). The U.S. market stands as the biggest market, with a market capitalization of over 52,000 trillion dollars in 2024 (CompaniesMarketcap.com, 2024).The mechanisms driving the expansion of global equity markets are not fully understood. While some attribute this growth to improved macroeconomic and financial fundamentals, skeptics, who doubt the efficiency of capital markets, suggest that positive investors may be purchasing stocks without actually considering the historical relationships between market fundamentals and equity valuation (Li, 2002).

5.1.2. Housing

The housing market is complex, characterized by variations in size, location, and numerous other attributes. Transactions involving houses are relatively infrequent and often escape notice by both governments and centralized market-makers (Malpezzi, 2014). Additionally, the housing market serves as a significant indicator of a country's economic development (Yan, 2007). It plays an important role in the banking system by providing substantial amounts of money through loans. This underscores the importance of maintaining a stable housing market, as historical instances of bubbles have precipitated economic crises, exemplified by the financial crisis in 2008. The start of this crisis occurred when house prices peaked around the middle of 2006, initiating a downward trajectory and resulting in an increase in default rates (Baker, 2008).

5.1.3. Bonds

Bonds serve as fixed-income instruments representing loans given by investors to either governmental or corporate entities. The investor faces the risk of the borrower defaulting the loan, while the potential reward comes in the form of interest rates. Unlike stocks, bonds have a limited upside, but a complete downside risk. As an individual becomes more risk-averse, the emphasis on short-term returns diminishes, but the strategy to safeguard against risks remains intact. Consequently, a long-term, risk-averse investor will allocate some wealth to a portfolio of long-term bonds (Campbell and Viceira, 2001). Global bond markets have low correlation, indicating that investors can reduce their risk by diversifying their bond investments internationally. A U.S.-based investor has the potential to double their profit at the same risk level by diversifying their portfolio globally. The same principle applies to achieving diversification by combining international stocks and bonds (Levy and Lerman, 1988).

5.1.4. Risk Free Asset

The risk-free rate represents the potential return for an investor without assuming any risk. However, it is largely a theoretical concept, as no asset in the world could provide a zero percent risk. U.S. Treasury bills are often employed as a asset for the risk-free rate, yet even these assets have a nominal standard deviation of 3.3% (Jordà et al., 2019). Treasury bills are issued at a discount to their face value, have no coupon rate, and reach maturity at their face value. These short-term loans are regularly issued with maturities of 91, 182, and 364 days, also known as 3-month, 6-month, and 1-year Treasury bills. Investors profit from Treasury bills by purchasing them at a discount to their maturity value.

5.2. Additional Input Parameters

5.2.1. Constraints for Asset Allocation

The investment constraints for the four-asset model are primarily made by the investing limitations to the *Norwegian Pension Fund Global*. These constraints outline that the fund's allocation should approximately consist of 70% in equities and 30% in fixed income. Additionally, the portion allocated to unlisted real estate should not exceed 7%, while investment in renewable energy should not exceed 2% (NBIM, 2024c). Given the limited contribution and absence of historical data, investments in renewable energy are excluded from the analysis. Instead, a risk-free asset (T-bills) is included to provide flexibility to the fund and allowing it to avoid being fully exposed to the risky market at all times. In this context, the following limitations have been set for the various asset classes:

- 70% of the fund can be invested in equity.

- 30% of the fund can be invested in fixed income (Bonds).
- 7% of the fund can be invested in unlisted real estate (Housing).
- 30% of the fund can be invested in the US Treasury bill (Bills).

5.2.2. Time Horizon

Drawing inspiration from the Government Pension Fund of Norway, where the objective is to support the welfare state and serve as a safeguard for future generations (NBIM, 2024b). The duration of the experiment is set to 100 years, equivalent to 600 two-month periods. To ensure the fund's longevity and sustained growth for the benefit of future generations, fund management must adopt a perspective that extends over a considerable time frame. A shorter time frame, such as 50 years, would likely be insufficient as it only covers a single generation.

5.2.3. Wealth Target

The wealth target is set to 1000 adjusted for inflation on 2% according to the inflation goal for Norges Bank (Norges Bank, 2020). As the life time is 100 years, The wealth target will be calculated as following:

$$\text{wealth\_target} = 1000 \times \left(1 + \frac{0.02}{\text{lifetime}}\right)^{\text{lifetime}} \tag{5.1}$$

By adjusting for inflation, the fund will maintain its real value, ensuring that future generations in 100 years will have the same amount of wealth as today.

5.2.4. Correlation Matrix

For this thesis, we are using the asset correlation map estimated by (Becker, 2024) from the dataset of (Jordà et al., 2019). Correlations are estimated by excluding any rows containing null values, and by removing outliers (Becker, 2024).

Table 5.2.: Correlation Matrix for different asset classes

	Bills	**Equity**	**Bonds**	**Real Estate**
Bills	100.00%			
Equity	5.24%	100.00%		
Bonds	23.41%	22.58%	100.00%	
Real Estate	22.20%	9.51%	-2.57%	100.00%

This table shows the correlation between different asset classes estimated by (Becker, 2024) using data sources from (Jordà et al., 2019).

5.2.5. Number of Trials

In order to analyze whether the different algorithms learn similar or optimal policies, we run several trials of the algorithms. Another term for ”trials” is ”agents.” It refers to the number of agents that gather experience from interacting with the environment. By increasing the number of trials, the algorithms will improve their exploration and learning in the environment, particularly in environments with many possible states and actions. The number of trials used in the model is 5, due to limitations in data power and time (Brownlee, 2018).

5.2.6. Number of Epochs

Increased epochs enable the agents to learn more from its accumulated experiences. An epoch is described as a single pass through the entire dataset during the training phase of a model. Increasing the number of epochs could be beneficial when the agent's learning rate is slow or when the policy or value function needs more iterations to converge. The duration on one epoch depends on which algorithm that is runned, and the size and complexity of the data (Yu et al., 2019).

Time and computational resources are limitations when it comes to number of epochs. This thesis use a approach that involves having 100 epochs for each step and then evaluating the output. If the output proves satisfactory, further epochs running the algorithm's will not take place. However, it is anticipated that certain algorithms may not produce satisfactory results within the initial 100 epochs. The number of epochs may need to be extended beyond 3000 in such cases.

5.2.7. Penalties

This thesis will mainly use two sets of penalties, one for smoothing and one for wealth. The first set applies a smoothing penalty of 0.1 and a wealth penalty of 0.4, while the second set uses a smoothing penalty of 0.3 and a wealth penalty of 0.8. The purpose of the smoothing penalty is to make larger deviations have more impact compared to smaller deviations. The wealth penalty give a penalty on agents for failing to meet the target wealth value. The smoothing penalty is detailed in equation 3.17, and the wealth penalty is detailed in equation 3.18.

In addition to smoothing and the wealth penalty, this thesis incorporates a share penalty of 0.2, as described in equation 3.19.

5.2.8. Risk Aversion

Risk aversion can be described as the behavior of an agent that prefer options with lower levels of uncertainty or potential loss, often choosing safety over potentially higher gains (Werner, 2008). Within the model, we adapt the utility function illustrated in equation 3.16. The parameter τ is set at 1.26, as determined by (Layard et al., 2008), who analyzed data from over 50 countries in the period from 1972 to 2005.

5.2.9. Consumption Rate Constraints

The upper and lower constraints for the consumption rate is sat to 0.5% and 10% yearly. This approach provides trials with greater flexibility to select the optimal consumption rate rather than using an existing fiscal rule, such as the 3% rule established for the GPFG (Regjeringen, 2024).

- Lower bound 0.5% / intervals per year = 0.0833% per 2 month period
- Upper bound 10% / intervals per year = 1.667% per 2 month period

5.2.10. Visualizations

The visualization of the algorithms scatter plots and graphs in the appendix is based on 50 random runs. This implies that the visualization shows only a sample of the agents solutions, and 50 other random runs would have performed differently. However, we assume that the sample provides a representative view of the algorithms performance, The variation between runs is considered to not be significant, provided that the agents have been sufficiently trained.

6. Presentation of the Results

This thesis implement three risky assets alongside an additional risk-free asset (T-bills) to formulate an optimal portfolio for Sovereign Wealth Funds (SWFs). In the modeling phase, different scenarios involving only one risky asset were investigated, both with and without correlation. Further research has developed to the four-asset model, which provides the results to be presented in this chapter. These results originate from the application of five algorithms: PPO, DDPG, A2C, SAC, and TD3. These algorithms were employed within a model using correlations among three risky asset classes, aiming to achieve realistic outcomes.

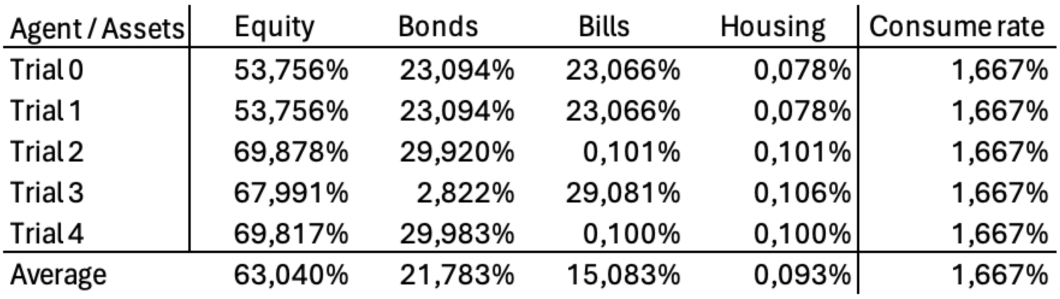
6.1. Capital Allocation and Consumption

In this section, the outputs from the five different algorithms will be presented. The outputs consist of optimal capital allocation and consumption rates for 5 trials for each of the five algorithms, tested for different values of smoothing and wealth penalty. At the end of the section, the utility for each of the algorithms will be presented, providing insights into their performance. The results for capital allocation for each algorithm is constructed from weighted average. The reason is that certain algorithms do not converge to a linear pattern, resulting in inconsistent capital allocation across varying levels of wealth. Appendix B includes several example of this. The scatter plot produced by SAC (Figure B.12) have a curved graph. A2C (Figure B.9) on the other hand have a horizontal graph with zero growth, suggesting that the optimal capital allocation is constant across different wealth levels.

6.1.1. DDPG

Table 6.1 illustrates an unstable model, as most agents have different capital allocation weights. With a stable consumption rate of 1.667%, a large portion of the funds is directed towards withdrawals, allocating capital at the upper bound constraint for consumption. DDPG allocates the majority of the wealth to equity, with some higher weights in bonds than in T-bills. Housing carries a low weight, remaining close to zero, far under the investment constraint of 7%. Trials 2 and 4 stand out, with over 99% of the capital allocated to equity and bonds.

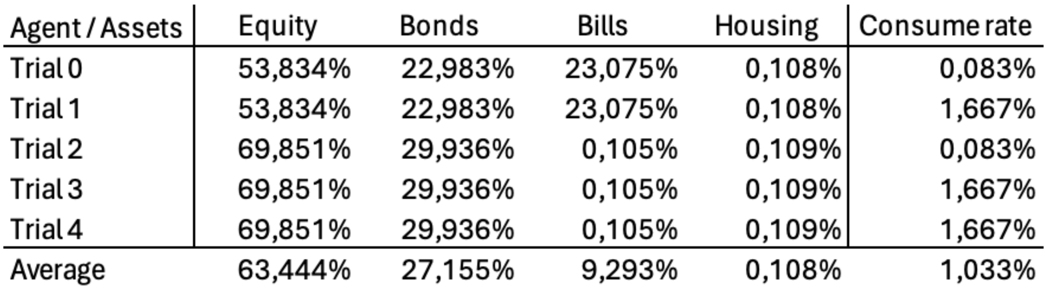
Table 6.1.: DDPG: Wealth Penalty at 0.8 and Smoothing Penalty at 0.3



Agent / Assets	Equity	Bonds	Bills	Housing	Consume rate
Trial 0	53,756%	23,094%	23,066%	0,078%	1,667%
Trial 1	53,756%	23,094%	23,066%	0,078%	1,667%
Trial 2	69,878%	29,920%	0,101%	0,101%	1,667%
Trial 3	67,991%	2,822%	29,081%	0,106%	1,667%
Trial 4	69,817%	29,983%	0,100%	0,100%	1,667%
Average	63,040%	21,783%	15,083%	0,093%	1,667%

Looking at Table 6.2, where lower wealth and smoothing penalties are applied. The asset weights have less variation as the trials cluster around two solutions, trial 0 and 1 have the same allocation, and 2, 3 and 4 are allocating at the same levels. The consumption rate is at lower bound for some agents at 0.083%, compared to the higher penalty rate Table (6.1), where all the trials had a consumption rate at 1.667%. Agents allocate more capital towards equity and bonds, with an average of 63,44% in equity and and an average of 27,155% invested in bonds. It's noteworthy that two agents try to allocate a larger portion of their wealth to bills, considered at lower risk, but the three other agents have almost all the capital allocated into bonds and equity. All agents allocate relatively low weights to housing, almost the same as the model with higher penalties (Table 6.1).

Table 6.2.: DDPG: Wealth Penalty at 0.4 and Smoothing Penalty at 0.1

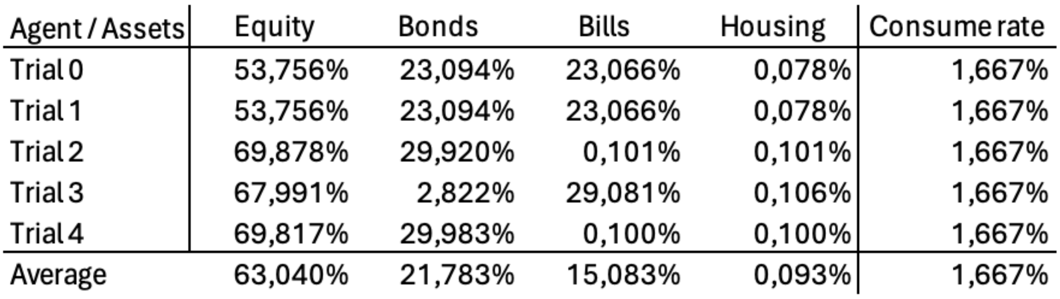


Agent / Assets	Equity	Bonds	Bills	Housing	Consume rate
Trial 0	53,834%	22,983%	23,075%	0,108%	0,083%
Trial 1	53,834%	22,983%	23,075%	0,108%	1,667%
Trial 2	69,851%	29,936%	0,105%	0,109%	0,083%
Trial 3	69,851%	29,936%	0,105%	0,109%	1,667%
Trial 4	69,851%	29,936%	0,105%	0,109%	1,667%
Average	63,444%	27,155%	9,293%	0,108%	1,033%

6.1.2. PPO

Table 6.3 illustrates the capital allocation and consumption rate for the PPO algorithm with a wealth penalty of 0.8 and a smoothing penalty of 0.3. The various trials converge on the same optimal consumption rate of 1.667%, but their capital allocations vary across trials. While all trials aim to allocate a majority of funds to equity, their distribution among bills and bonds differs significantly. Trial 2 and 4 have a minimal allocation to bills, opting for a higher proportion in bonds and equity. This could indicate different risk evaluations among trials, as bills have lower risk inputs (standard deviation) compared to bonds and equity.

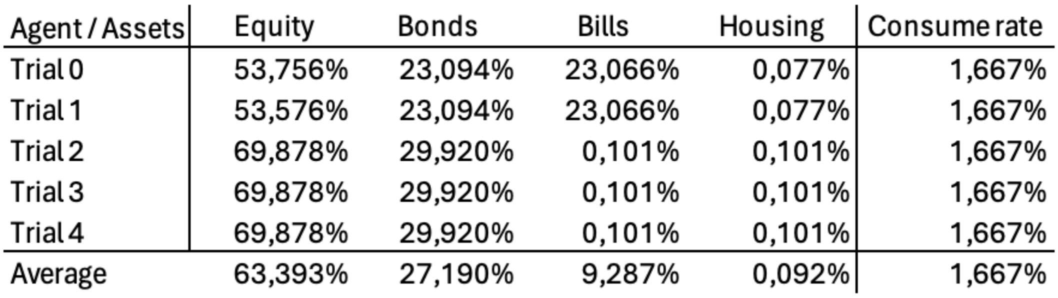
Table 6.3.: PPO: Wealth Penalty at 0.8 and Smoothing Penalty at 0.3



Agent / Assets	Equity	Bonds	Bills	Housing	Consume rate
Trial 0	53,756%	23,094%	23,066%	0,078%	1,667%
Trial 1	53,756%	23,094%	23,066%	0,078%	1,667%
Trial 2	69,878%	29,920%	0,101%	0,101%	1,667%
Trial 3	67,991%	2,822%	29,081%	0,106%	1,667%
Trial 4	69,817%	29,983%	0,100%	0,100%	1,667%
Average	63,040%	21,783%	15,083%	0,093%	1,667%

The decrease in wealth and smoothing penalty illustrated in Table 6.4, results in a larger proportion of capital allocated to equity and bonds compared to the allocation shown in Table 6.3. This may suggest that the trials are more willing to take risk as both smoothing and wealth penalties decrease. The consumption rate remains consistent across both high (Table 6.3) and low (Table 6.4) levels of wealth and smoothing penalties. Additionally, it’s a significant difference in the allocations of capital between the various trials, regardless of the differences in penalties applied. There is remarkably that trial 0 and 1 invests almost 25% in bills. This indicates that the models are unstable, as all trials are not converging to the same capital allocation.

Table 6.4.: PPO: Wealth Penalty at 0.4 and Smoothing Penalty at 0.1

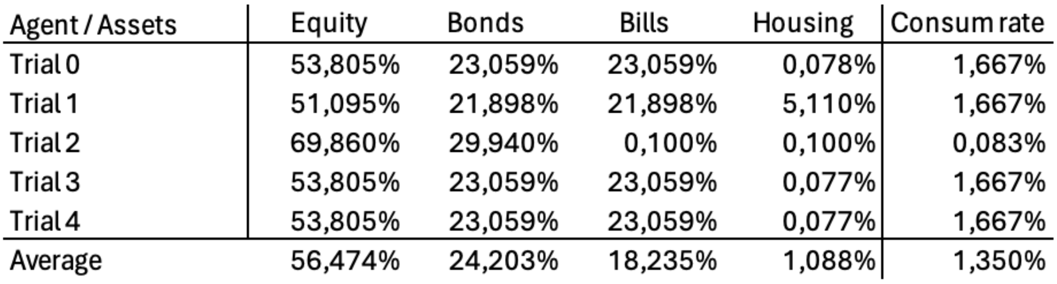


Agent / Assets	Equity	Bonds	Bills	Housing	Consume rate
Trial 0	53,756%	23,094%	23,066%	0,077%	1,667%
Trial 1	53,576%	23,094%	23,066%	0,077%	1,667%
Trial 2	69,878%	29,920%	0,101%	0,101%	1,667%
Trial 3	69,878%	29,920%	0,101%	0,101%	1,667%
Trial 4	69,878%	29,920%	0,101%	0,101%	1,667%
Average	63,393%	27,190%	9,287%	0,092%	1,667%

6.1.3. A2C

Table 6.5 illustrates the capital allocation and consumption rate for A2C with a wealth penalty of 0.8 and a smoothing penalty of 0.3. Trial 2 stands out with a large proportion of wealth invested in equity and bonds (99.8%), and a significantly smaller portion allocated to bills and housing (0.2%). Additionally, this agent consume at a much lower rate compared to the other four agents. It's noteworthy that in trial 1, the allocation in housing is at 5%, which is substantially higher than the allocations in the other trials, all of which are under 0.10%.

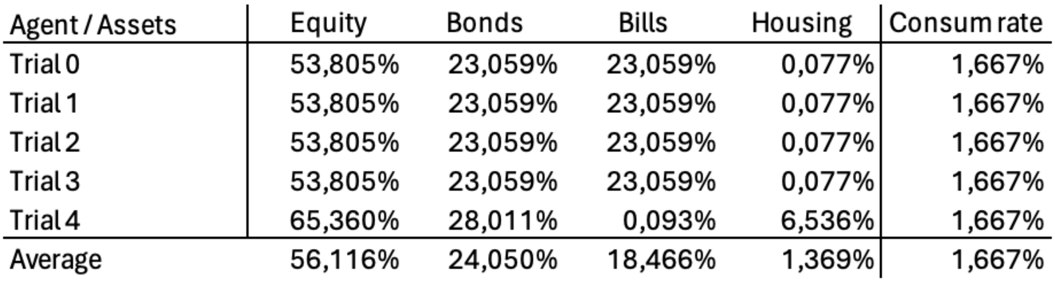
Table 6.5.: A2C: Wealth Penalty at 0.8 and Smoothing Penalty at 0.3



Agent / Assets	Equity	Bonds	Bills	Housing	Consum rate
Trial 0	53,805%	23,059%	23,059%	0,078%	1,667%
Trial 1	51,095%	21,898%	21,898%	5,110%	1,667%
Trial 2	69,860%	29,940%	0,100%	0,100%	0,083%
Trial 3	53,805%	23,059%	23,059%	0,077%	1,667%
Trial 4	53,805%	23,059%	23,059%	0,077%	1,667%
Average	56,474%	24,203%	18,235%	1,088%	1,350%

The A2C algorithm with lower penalties, as illustrated in Table 6.6, shows that 4 out of 5 trials result in the same capital allocation. The consumption rate is similar across all 5 trials and are similar to 4 of the trials in Table 6.5. Additionally, Table 6.6 includes one trial that stands out in terms of housing allocation, with trial 4 allocating 6.5% to housing. This trial also allocates significantly more to riskier assets like bonds and equity, rather than in bills.

Table 6.6.: A2C: Wealth Penalty at 0.4 and Smoothing Penalty at 0.1

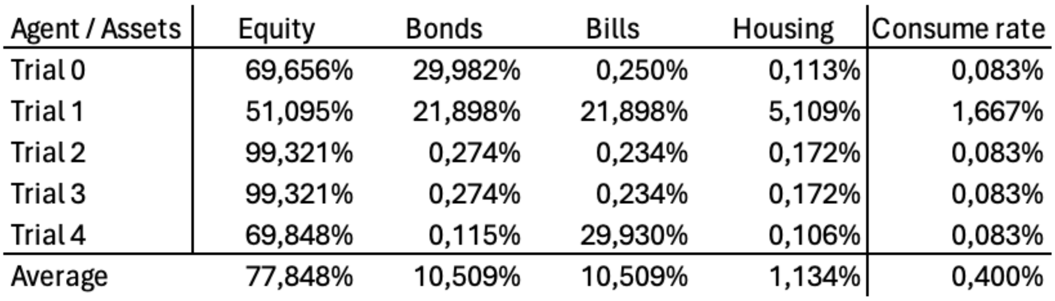


Agent / Assets	Equity	Bonds	Bills	Housing	Consum rate
Trial 0	53,805%	23,059%	23,059%	0,077%	1,667%
Trial 1	53,805%	23,059%	23,059%	0,077%	1,667%
Trial 2	53,805%	23,059%	23,059%	0,077%	1,667%
Trial 3	53,805%	23,059%	23,059%	0,077%	1,667%
Trial 4	65,360%	28,011%	0,093%	6,536%	1,667%
Average	56,116%	24,050%	18,466%	1,369%	1,667%

6.1.4. SAC

Table 6.7 illustrates the capital allocation and consumption rate for the SAC algorithm with a wealth penalty at 0.8 and a smoothing penalty at 0.3. This algorithm indicate instability in the results, as the trials are not converging to one capital allocation. The consumption rate remains at the lower bound constraint across all trials except for trial 1 (1.67%). Trials 2 and 3 allocate a large portion (99%) of the capital to equity, thereby violating the constraint at a maximum of 70% allocation in equity. On average, there is a significant portion of wealth allocated to equity, with a fifty-fifty allocation to bills and bonds. However, the allocation in housing is considerably low, at 1.13% on average.

Table 6.7.: SAC: Wealth Penalty at 0.8 and Smoothing Penalty at 0.3

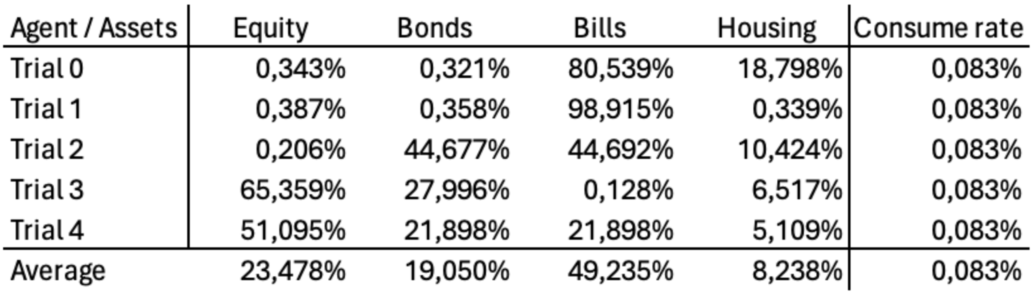


Agent / Assets	Equity	Bonds	Bills	Housing	Consume rate
Trial 0	69,656%	29,982%	0,250%	0,113%	0,083%
Trial 1	51,095%	21,898%	21,898%	5,109%	1,667%
Trial 2	99,321%	0,274%	0,234%	0,172%	0,083%
Trial 3	99,321%	0,274%	0,234%	0,172%	0,083%
Trial 4	69,848%	0,115%	29,930%	0,106%	0,083%
Average	77,848%	10,509%	10,509%	1,134%	0,400%

Table 6.8 illustrates the capital allocation and consumption rate for SAC with a wealth penalty of 0.4 and a smoothing penalty of 0.1. Lower wealth and smoothing penalties results in a significantly lower allocation to equity (23.48%) compared to Table 6.7. Additionally, this algorithm violates the maximum constraints on several occasions. In 3 out of the 5 trials, allocations exceed the constraints on bills (30%) while Trial 1, allocates as high as 98.9% to

this asset class. Housing also has two trials violating the constraints, and bonds have one trial. The consumption rate remains stable at the lower bound constraint at 0.083% for all trials. In summary, the SAC algorithm with low penalties produces highly unstable results with frequent violations of constraints.

Table 6.8.: SAC: Wealth Penalty at 0.4 and Smoothing Penalty at 0.1

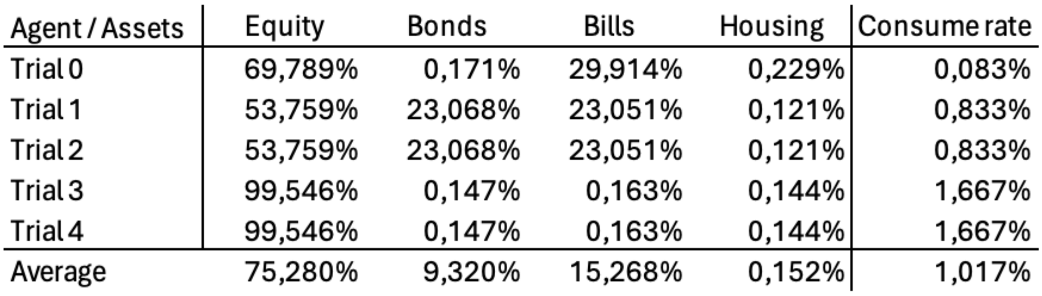


Agent / Assets	Equity	Bonds	Bills	Housing	Consume rate
Trial 0	0,343%	0,321%	80,539%	18,798%	0,083%
Trial 1	0,387%	0,358%	98,915%	0,339%	0,083%
Trial 2	0,206%	44,677%	44,692%	10,424%	0,083%
Trial 3	65,359%	27,996%	0,128%	6,517%	0,083%
Trial 4	51,095%	21,898%	21,898%	5,109%	0,083%
Average	23,478%	19,050%	49,235%	8,238%	0,083%

6.1.5. TD3

Table 6.9 illustrate the performance of the TD3 algorithm with a smoothing penalty of 0.3 and a wealth penalty of 0.8. Trials 3 and 4 allocate 99.5% of the portfolio in equity, violating the equity constraint, which is that no more than 70% of the capital should be invested in this asset class. All trials allocate a small part of the wealth to housing. Trials 0, 1, and 2 demonstrate lower consumption rates compared to trials 3 and 4. The model have varying results, with allocations and consumption rates not converging to any specific values, indicating instability within the model.

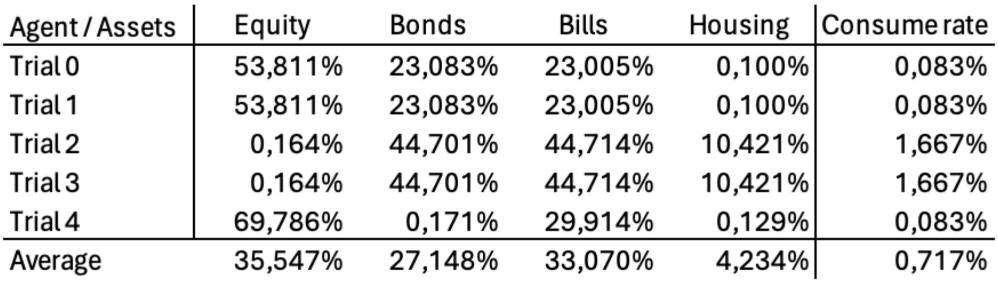
Table 6.9.: TD3: Wealth Penalty at 0.8 and Smoothing Penalty at 0.3



Agent / Assets	Equity	Bonds	Bills	Housing	Consume rate
Trial 0	69,789%	0,171%	29,914%	0,229%	0,083%
Trial 1	53,759%	23,068%	23,051%	0,121%	0,833%
Trial 2	53,759%	23,068%	23,051%	0,121%	0,833%
Trial 3	99,546%	0,147%	0,163%	0,144%	1,667%
Trial 4	99,546%	0,147%	0,163%	0,144%	1,667%
Average	75,280%	9,320%	15,268%	0,152%	1,017%

In contrast to Table 6.9, the model illustrated in Table 6.10 does not violate the constraint for maximum capital allocation to equity. However, it does violate the constraints for housing, bills, and bonds in trials 2 and 3. These trials stand out as they allocate the smallest portion to equity. The model have variation in results, indicating that the model with lower smoothing and wealth penalty also are giving unstable results.

Table 6.10.: TD3: Wealth Penalty at 0.4 and Smoothing Penalty at 0.1



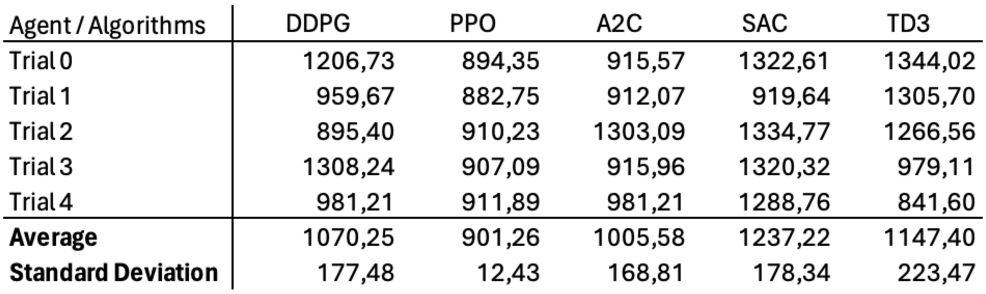
Agent / Assets	Equity	Bonds	Bills	Housing	Consume rate
Trial 0	53,811%	23,083%	23,005%	0,100%	0,083%
Trial 1	53,811%	23,083%	23,005%	0,100%	0,083%
Trial 2	0,164%	44,701%	44,714%	10,421%	1,667%
Trial 3	0,164%	44,701%	44,714%	10,421%	1,667%
Trial 4	69,786%	0,171%	29,914%	0,129%	0,083%
Average	35,547%	27,148%	33,070%	4,234%	0,717%

6.2. Average Utility

The two tables in this section illustrate the utility for the five trials across various algorithms for two values of smoothing and wealth penalty. The utility is computed according to the function outlined in Equation 3.16 and is an average of 50 random runs. The utilities represent the total utility in the time period for each trial. An optimal solution should have approximately the same utility across all trials.

Table 6.11 illustrates the utility when the wealth penalty is 0.8 and the smoothing penalty is 0.3. SAC has the highest average utility at 1237.22 and demonstrates stability, but has one trial with significantly lower utility (Trial 1). A2C also has one trial that deviates from the rest (Trial 2), significantly inflating the average. PPO also demonstrates consistent utility results, but has the lowest average utility among the algorithms. DDPG and TD3 attain utilities above 1300, but have significant variations between trials, suggesting challenges in converging to an optimal solution.

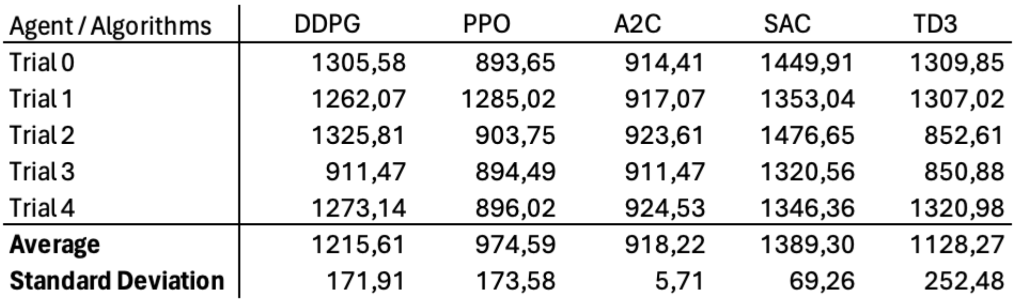
Table 6.11.: Utility with wealth penalty of 0.8 and smoothing penalty of 0.3



Agent / Algorithms	DDPG	PPO	A2C	SAC	TD3
Trial 0	1206,73	894,35	915,57	1322,61	1344,02
Trial 1	959,67	882,75	912,07	919,64	1305,70
Trial 2	895,40	910,23	1303,09	1334,77	1266,56
Trial 3	1308,24	907,09	915,96	1320,32	979,11
Trial 4	981,21	911,89	981,21	1288,76	841,60
Average	1070,25	901,26	1005,58	1237,22	1147,40
Standard Deviation	177,48	12,43	168,81	178,34	223,47

Table 6.12 illustrates the utility when the wealth penalty is 0.4 and the smoothing penalty is 0.1. With the exception of A2C and PPO, lower penalties result in higher average utility across all algorithms. A2C have lower utility, but all utilities fall within a similar range. SAC also demonstrates more stability with all 5 trials falling within the same range, giving higher average utility. In Table 6.12, SAC emerges with the highest average utility at 1389.30. This also leads to a lower standard deviation, with 69.26 in Table 6.11 compared to 178.34 in Table 6.12. TD3 have some instability in the results, as utility variate by several hundred points for each trial. However, these results appear to cluster around two optimal solutions, as three values are closely grouped, as are two others.

Table 6.12.: Utility with wealth penalty of 0.4 and smoothing penalty of 0.1.

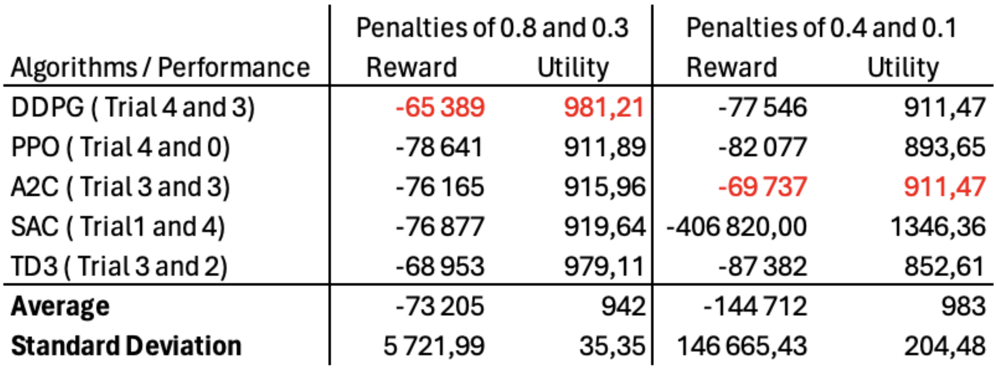


Agent / Algorithms	DDPG	PPO	A2C	SAC	TD3
Trial 0	1305,58	893,65	914,41	1449,91	1309,85
Trial 1	1262,07	1285,02	917,07	1353,04	1307,02
Trial 2	1325,81	903,75	923,61	1476,65	852,61
Trial 3	911,47	894,49	911,47	1320,56	850,88
Trial 4	1273,14	896,02	924,53	1346,36	1320,98
Average	1215,61	974,59	918,22	1389,30	1128,27
Standard Deviation	171,91	173,58	5,71	69,26	252,48

6.3. Average Rewards

In the appendix, Table A.1 and A.2 illustrate the rewards generated by the algorithms. The variability in the results are following the same pattern as for utility, where trials having high utility also have the lowest rewards. Table 6.13 illustrate the top-performing trials for the algorithms. The reward is a measure on the agents performance when interacting with the environment. DDPG, Trial 4, with penalties at 0.8 and 0.3 have the highest rewards among all trials and penalties, at -65,389. A2C trial 3 achieve the highest reward (-69,737) for the models with penalties of 0.4 and 0.1. On average, the higher penalty models outperform the lower penalty model in terms of rewards. In the model with penalties of 0.4 and 0.1, SAC stands out with the lowest reward even tough this model had the highest average utility, as illustrated in Table 6.11 and 6.12.

Table 6.13.: Best overall performance



	Penalties of 0.8 and 0.3		Penalties of 0.4 and 0.1	
Algorithms / Performance	Reward	Utility	Reward	Utility
DDPG (Trial 4 and 3)	-65 389	981,21	-77 546	911,47
PPO (Trial 4 and 0)	-78 641	911,89	-82 077	893,65
A2C (Trial 3 and 3)	-76 165	915,96	-69 737	911,47
SAC (Trial1 and 4)	-76 877	919,64	-406 820,00	1346,36
TD3 (Trial 3 and 2)	-68 953	979,11	-87 382	852,61
Average	-73 205	942	-144 712	983
Standard Deviation	5 721,99	35,35	146 665,43	204,48

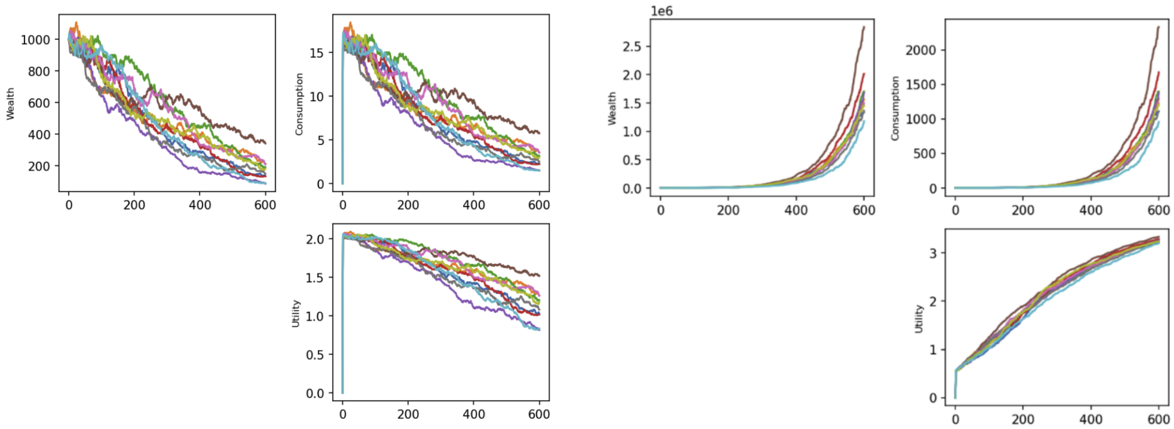
6.4. Evaluation of Plots

In Section 6.1, the average capital allocation and consumption were presented, and in section 6.2, the average utility was presented. However, the average may not always accurately reflect the actual situation, as utility, consumption, and capital allocation could vary throughout the lifetime of the experiment. This section will look into the different plots for some of the trials for the algorithms. Each trial represents a single plot, and therefore, some of the results presented here only serve to illustrate the outcome of one trial and may not be representative of the entire algorithm. Many of the plots from each trial have similar patterns that will be addressed in the following section. One plot are included for each algorithm in the appendix.

6.4.1. Wealth, Consumption and Utility

Figure 6.1 shows the graphs for 50 random runs on how the wealth, consumption and utility is evolving trough the 600 time steps for two different trials for A2C and PPO. The main issue with the wealth generated by different algorithms is that some algorithms cause the wealth to decline from the initial value, while others lead to an increase in wealth. This effects the consumption as the consumption rate is constant for all time periods and for all trials. The consumption will therefore always follow the wealth graph as illustrated in the 6.1. Because some algorithms and trials are maximizing the utility by having a decline wealth graph, will these agents also witness a high utility in the beginning of the life time and facing a decrease in utility from there. The trials with a high consumption rate at 1.667%, which is similar to a 10% yearly consumption, are experiencing a downward trend for utility, approaching to zero. The trials with a 0.083% consumption rate are having an upward trend in utility and wealth. A2C trial 1 (6.1a) is an example of this, and PPO (6.1b) on the other hand, are having a low utility on the start of the life time and then increase as the wealth of the fund increases.

Figure 6.1.: Graphical plots A2C and PPO with penalty of 0.8 and 0.3



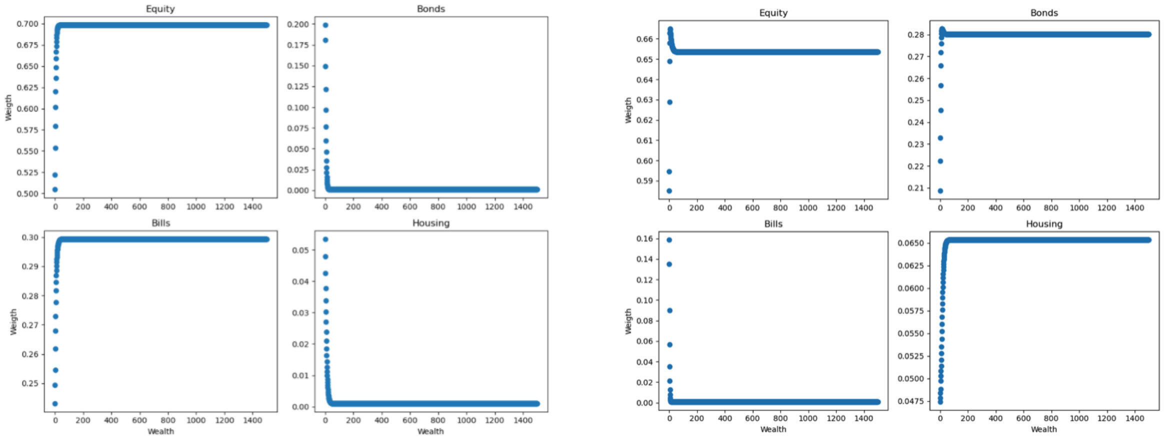
(a) Results for A2C trial 1

(b) Results for PPO trial 1

6.4.2. Capital Allocation

In Figure 6.2, examples of capital allocation at different wealth levels are illustrated. Both scatter plots shows how the algorithms converge to a specific capital allocation at relatively low values of wealth. The Figure 6.2a shows how the agent is less inclined to invest in equity and bills at lower wealth levels compared to bonds and housing. This is evident in the scatter plot's shape, with initially low allocations to equity in wealth step one, but which then increase rapidly to 69.8%. A similar phenomenon is observed in the scatter plot of one of the trials for SAC (Figure 6.2b). However, in this case, the shape of the scatter plot is different and it takes more time to converge to a specific value. All the plots from the various trials and algorithms described in this thesis have the same pattern of rapid convergence to a certain value. Therefore, as long as the fund's wealth remains above a value of 100 to 200, it will not significantly affect the capital allocation.

Figure 6.2.: Capital allocation for TD3 and SAC



(a) TD3 penalty 0.4 and 0.1 for trial 0

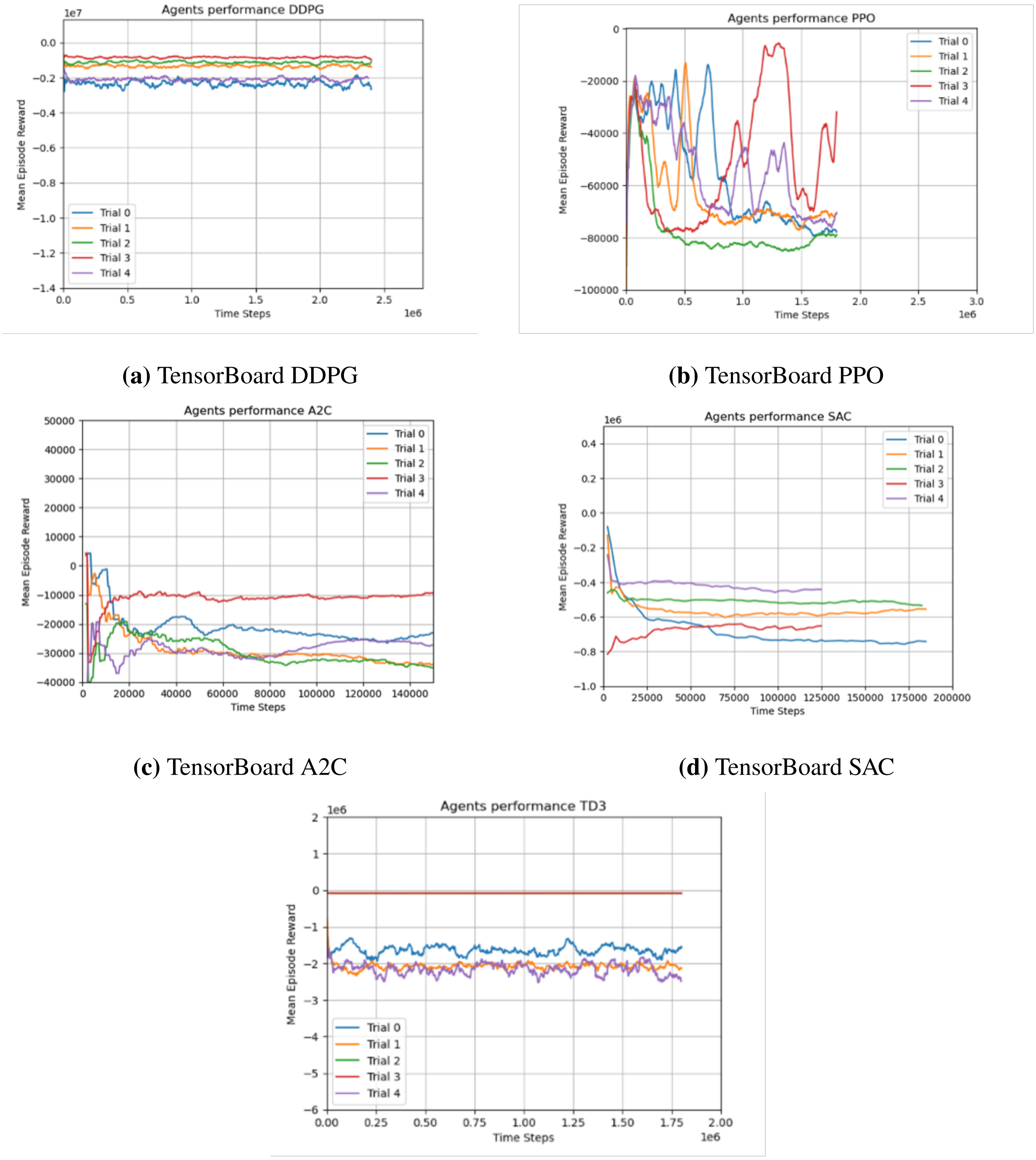
(b) SAC penalty 0.4 and 0.1 trial 3

6.4.3. TensorBoards

A selection of TensorBoard visualizations for the different algorithms can be found in the end of this section. These illustrate the training process of the various agents. The x-axis represents the number of time steps, calculated as the product of the lifetime and the number of episodes the algorithm runs. The y-axis shows the average reward the agents receive per episode.

The TensorBoard visualizations for DDPG (Figure 6.3a) and A2C (Figure 6.3c) illustrate very stable performance over the time period, with all trials having low volatility. This indicates that these trials learn and adapt to the environment effectively. In contrast, PPO (Figure 6.3b) demonstrates significantly more instability for trial 3 and 4 throughout the training period, while agents 0, 1, and 2 stabilize halfway through. The TensorBoard visualization for TD3 is illustrated in Figure 6.3e, where trial 2 and 3 show a constant zero reward during the training period, suggesting that they might not be adapting to the environment optimally. This is supported by the utility values in Figure 6.12, where Trials 2 and 3 achieve a utility of approximately 850, while the other three agents exceed a utility of 1300. Finally, the trials for SAC perform with high stability and low rewards throughout the training process. However, the algorithm stops unexpectedly, which has been an issue during the training. As shown in Figure 6.3d, agents 3 and 4 stop at 125,000 time steps. Using more powerful computers may resolve this problem.

Figure 6.3.: TensorBoards for the different algorithms



(a) TensorBoard DDPG

(b) TensorBoard PPO

(c) TensorBoard A2C

(d) TensorBoard SAC

(e) TensorBoard TD3

7. Discussion and Critical Reflections

7.1. Optimizing Asset Diversification

According to Markowitz, diversification minimizes risk and maximizes returns by creating an optimal portfolio aligned with the efficient frontier (Markowitz, 1952). There are no direct risk management in our reinforcement learning model like in Markowitz. However the smoothing parameter control the fluctuations in consumption and wealth violations parameter will force less risky strategies. The experiments in this thesis diversify assets in various ways, with some trials allocating over 99% of the assets to a single asset class. Both SAC (98.9% in bills) and TD3 (99.55% in equity) have trials with this level of allocation. This indicates sub optimal solutions due to the risk-return relationship. The trial with 99% allocated to equity is the best performing trial for TD3 for the model with penalties of 0.8 and 0.3 (Table 6.13). This could indicate that the best solution in this case not align with the diversification theory to Markowitz. However, the equity asset class in this model is already diversified, as it includes data from 16 equity markets (Jordà et al., 2019).

As previously mentioned, the agent's feedback and reward are fundamental to reinforcement learning, as the agent learns by receiving rewards and penalties that guide its future actions (Russell and Norvig, 2010). The agent's lifespan can influence model diversification, as short-term rewards might lead to sub optimal solutions. RL algorithms seek for the best trade-off between immediate reward and future rewards. However, this trade-off can be influenced by the discount rate and the choice of the utility function (particularly parameter τ) . As illustrated in Figure 6.1a, the utility approaches zero for some trials, indicating an investment and consumption strategy that fails to maintain the fund's value. The reward for each action affects the utility, with a higher utility value indicating greater success both in the past and the future. Actions with higher utility are more likely to be chosen in future decisions (Janssen and Gray, 2012).

It is also important to note that the model is based on a stochastic control process that involves uncertainty and randomness. This may lead agents to make risky choices to achieve the highest rewards.

7.2. Sustainable Consumption Levels

As described in Chapter 6, the consumption rate fluctuates between 0.083% and 1.667% for each 2 month periods. Figures B.8 and B.5 in the appendix illustrate that different algorithms and trials have varying optimal strategies, with some converging towards zero while others escalate to higher levels of wealth. Given the experiment's 100-year time frame, it suggests that this duration may be insufficient, as wealth nearing zero provides no utility for the generations after the life time. The PPO graph in Figure B.5 illustrate a sharp increase in wealth after 400 time steps, resulting in significant wealth accumulation over the 100-year period. It is reasonable to conclude that such wealth expansion would provide greater benefits to future generations compared to a fund that goes to zero after 100 years.

It could be argued that if the consumption rate is lower than the rate of return, the value of the fund will increase over time. *Government Pension Fund Global* (GPFG) had a rate of return on 6.09% since its first deposit in 1997 (NBIM, 2024). This suggests that if the model adapted in this thesis could achieve a similar rate of return, it would continue to grow in value as long as the consumption rate is at a lower level. However, the model have a yearly consumption rate that is either 0.05% or 10%, which means that the trials is either having a consumption rate at the lower or upper constraints for consumption. A constant yearly consumption rate at 10% for some trials could describe why some experiments is going to zero. The realism of maintaining a constant consumption rate is still in question. The need for additional funds by governments varies from year to year, particularly during economic downturns when governments are more likely to demand more from their Sovereign Wealth Fund (SWF) compared to periods of economic growth (Baldacci, Emanuele and Gupta, Sanjeev, 2009).

TD3 have two trials that have a consumption rate at 0.833% with a wealth and smoothing penalty level of 0.8 and 0.3. Figure B.18 illustrates the graphical results for trials 1 and 2. Trial 1 shows an increase in wealth, consumption, and utility, where trial 2 on the other hand have a decrease in these metrics, despite both trials suggesting the exact same portfolio (Table 6.9). In contrast, almost all SAC trials result in a consumption rate of 1.667%, except for trial 1 in the experiment with a wealth and smoothing penalty of 0.8 and 0.3. Figure B.14 illustrate how the consumption rate impacts the development of wealth, consumption, and utility.

Overall, the trials with higher consumption result gives a lower level of utility. A higher consumption rate indicates that a larger portion of the fund is withdrawn and not reinvested for the future. This reduces the fund's ability to provide the same utility for future generations as it does for the current generation. As mentioned, the consumption rate is 1.667% every second month, equivalent to 10% annually and this is consuming the funds value over the life time. Consequently, none of the trials deliver an annual return higher than 10%, while the upper bound of 10% is too high. With this level of consumption, utility in the future decreases. A smaller gap between the upper and lower bound constraints may be appropriate, as 1.667% seems too high. All of the trials, except two of TD3's trials, find their solution at a consumption level of 0.833% and 1.667%. Therefore, it should not be excluded that the trials would have discovered the optimal solution at both the upper and lower bounds, regardless of the level of the restrictions.

7.3. Equity Limitation at 70%

The model have an equity investment constraint at 70%. Equity have higher expected return and standard deviation than bonds and treasury-bills. Housing have the highest expected return among the assets and have lower standard deviation than equity. but the investment constraints placed to only 7%. By comparing the constraints for equity and real estate with the standard deviation, the model is benefiting investments to equity even tough housing have a better reward to risk. Equity have more than double the standard deviation as real estate. Some of best preforming trials in terms of reward and utility are allocating less than 70% to equity. Since the investment horizon is set to 100 years, higher equity allocation might be appropriate, given high historical returns.

As mentioned, some of the trials allocate more than 70% of the capital to equity. This may suggest that the penalties for violating the constraints are insufficient. In both the SAC and TD3 models, it is the models with higher penalties that have trials allocating over 99% of wealth to equity. A heightened penalty for constraint violations could incentive the trials to make other capital allocation decisions. This is particularly evident in the case of the high penalty model for TD3, where the most rewarding model (Trial 3) allocates 99.5% of its resources to equity.

7.4. Assessing Asset Classes

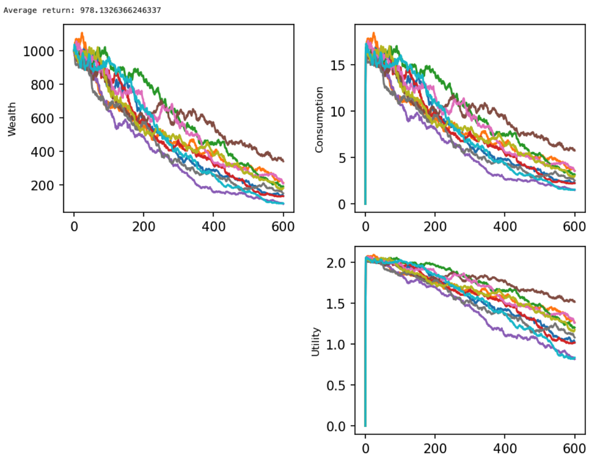
Historically (1870-2015), investments in housing have provided the best returns, with nominal returns of 11%, compared to 10.65% in equity, while also having a far lower standard deviation (21.27% against 9.93%) (Jordà et al., 2019). This raises questions about the relatively low allocation of housing within the fund. Across all trials, approximately 10% or less is allocated to housing, with many trials allocating close to 0%. This allocation is also constrained by our limitation of 7% on housing investment and a 70% allocation limit on equity that gives the agent penalties for exceeding these limits.

The high investment in T-bills may be questionable for long-term investments as they historically give low returns, although they also have a low standard deviation. The trials with a low consumption rate at 0.083%, could have a significant allocation to risky assets as this classes might give better long-term returns. This is because the capital could remain invested in these assets for extended periods, and therefore avoid potential losses due to market volatility. As illustrated in Figure 6.2, over 21% of the wealth for this DDPG experiment is invested in bills, resulting in a total allocation to fixed income (Bills and Bonds) of over 40%. In this context, it could be discussed whether this is the optimal allocation for maximizing wealth.

7.5. The Cost of Smoothing Penalty for Consume

The degree of smoothing penalty serves varied purposes as it makes low deviations less countable than big deviations. A low smoothing penalty offers agents greater flexibility. In a dynamic macroeconomic environment, agents may make decisions based on short-term fluctuations. In such a scenario, agents gets penalties for negatively deviating from consumption compared to a scenario where the agent are facing positive economic growth. A low penalty level may also incentives riskier behavior among agents. There might be increased variability in the evolution of wealth over time. The Figure 7.1 illustrate, that there is no significant variance in wealth development across different runs.

Figure 7.1.: Graphical plot A2C



Graphical plot for A2C trial 1.

Increased levels of smoothing penalty secures stability over the consumption rate. This is an important objective as it is crucial for SWFs to have stability in fund management and consumption as it should guarantee consistent welfare, utility, and benefits for future generations. However, imposing a too high smoothing penalty may result in sub optimal decisions, restricting the agent from fully exploring all available investment opportunities. This could potentially sacrifice optimal decisions for higher stability.

7.6. Impact of Future Target Wealth

Constructing a reinforcement learning algorithm presents a challenge, particularly concerning the future value of today's currency. Considering a 100-year time-frame, a dollar's worth today will significantly diminish over time. Preserving the wealth of future generations is about ensuring that the fund's real value in the future remains equivalent to its present value. It seems natural to expect that a sovereign wealth fund, designed to secure wealth for generations, should at least maintain its real value over the long term. From this standpoint, challenges arise, notably the potential for variations in governmental spending. Factors such as wars or pandemics have historically had a dramatic influence on expenditure. This raises questions about the allowing of overspending during crises compared to more stable periods (Norges Bank, 2020).

The environment has a future target value at 1000 adjusted for inflation sat at 2% for the next

100 years. The formula is describes in equation 5.1, with an initial value of 1000 and a 2% inflation rate the calculated future wealth target is 7243. Even with this high future target value the wealth could go infinitely high or go to zero depending on the size of the wealth penalty. Figure 6.11 and 6.12 shows that the algorithms suggest both significantly increased wealth over the lifespan and, in some instances, wealth going to zero. This might suggest that the wealth penalties are too low, as the agent does not receive sufficient punishment for a decrease in wealth, especially when the wealth is going close to zero, which significantly deviates from the future target value. The relatively short lifespan of only 100 years might also explain why certain trials and algorithms culminate in near-zero wealth towards the end. A longer life time could have changed the behavior of the algorithms as the sums of penalties would be higher.

7.7. Risk Aversion Parameter Testing

In this model, the value of risk aversion (τ) has been set as a constant at 1.26. Testing different values of τ could provide a more accurate representation of actual risk aversion in the market. Various models use different values of τ; for instance, (Evans, 2005) suggests a value at 1.4, estimated for 20 OECD countries, while (Groom and Maddison, 2019) proposes a value of 1.5, estimated for the United Kingdom. On the other hand, some models suggest much higher values for τ, such as 50 (Mehra and Prescott, 1985). This studies shows variations in the science about the right value of τ indicating relevance of testing the algorithms for different values. Higher values of τ illustrate greater risk aversion among individuals compared to lower values. In this thesis, where τ is relatively close to 1, individuals are less responsive to changes in wealth, indicating a greater willingness to take risks.

A higher value of τ could potentially impact the allocation for various trials and algorithms. Particularly considering that in the results the asset with the highest standard deviation attracted the highest allocation (Equity). Intuitively, an agent with higher risk aversion might allocate more wealth to lower-risk assets such as bills, bonds, and housing, which have lower standard deviations compared to equity. It could also be discusses whether the level of risk aversion used in a model for a Sovereign Wealth Fund (SWF) should be the same as of individuals. It's possible that a government fund should adopt a higher level of risk aversion, given that preserving the funds wealth is as crucial as higher returns.

7.8. Extension of the Model by Mork

In this section will we discuss the extension in the model by Mork. The significance of this aspect in creating a realistic model is that the habit adjusts based on consumption. A Sovereign Wealth Fund (SWF) is intended to last for many decades, where consumption and wealth may experience significant fluctuations due to economic uncertainties. There would be unrealistic to develop a model for the future solely based on current macroeconomic trends. In time of instability governments and corporations change they spending patterns to fit and adapt new trends in the market, which may lead to unpredictable shift in the utilization of a SWF. Accurately forecasting future consumption and economic trends presents a formidable challenge. Therefore, incorporating the equation 3.14 may not provide an exact visualisation of the future, but it offers a more realistic one.

The utility function proposed by Mork in equation 3.10 assigns notably low utility values to certain algorithms, as demonstrated in the example provided in Appendix B.19. While this analysis fails to explain the better performance of the isoelastic utility function, it is plausible that the function offers a more precise representation of the agent decision-making within the environment. This could be attributed to the additional penalty parameters for wealth and consumption.

7.9. Black Box Problem

The reinforcement learning algorithms discussed in this thesis can be considered a ”black box” problem. This means that users can observe the input and output of these algorithms without understanding the internal processes. In contrast, ”white box” models allow for human interpretation of how the algorithm generates its output. However, white box algorithms often struggle to model complex relationships, potentially resulting in lower accuracy (Loyola-Gonzalez, 2019).

The lack of transparency in these algorithms means that humans cannot explain how the output was reached; only the algorithm itself knows how the decision was made. All five algorithms used in this thesis are stochastic, meaning they make decisions based on probability. This makes it even more challenging for managers of a Sovereign Wealth Fund (SWF) to understand the decision-making process. They must make their own assumptions based on the output, considering risk and their own policies (Christin, 2020). The advantages of black box algorithms include high predictive accuracy and effectiveness in sorting large amounts of data (Tsang, Daisy, 2023). The question then becomes whether it is better for fund managers to achieve a more accurate result at the expense of understanding the process that leads to that result.

8. Conclusion

The results indicate that all five algorithms have different performance in identifying the optimal portfolio construction. However, they share a common feature, a consumption level of either 0.083% or 1.667% for every 2-month period. This translates to an optimal annual consumption rate of either 0.5% or 10%, representing the lower and upper bound constraints. The trials with a consumption level of 1.667% have a wealth that goes to zero in contrast to the trials with 0.0833% have a increase in wealth over the time period. Given the number of training iterations the algorithms have undergone it appears that there may exist more than one optimal solution. Certain trials fail to allocate capital within the the portfolio constraints. A potential reason for this could be too low penalties or not optimal training.

DDPG is the top-performing algorithm in terms of both reward and utility when the penalties were set at 0.8 (Wealth) and 0.3 (Smoothing). However, at lower values of wealth and smoothing penalties, A2C was the best performer when it comes to reward. SAC stood out as the poorest performer under these penalty levels as it had a reward far lower than the other algorithms. Apart from SAC, the differences between the performance of the other algorithms were relatively small but still of a certain size. The results from the capital allocations have unstable results as many of the trials finds different optimal solutions.

The results from the model indicate that the algorithms within the *Stable Baseline3* framework may not be ideally suited for determining the optimal capital allocation and consumption for sovereign wealth funds. The variation in results between the algorithms could indicate that further development of the algorithms is needed.

9. Further Work and Possible Extensions

Further work could expand the RL environment by adding more asset classes. This might include dividing some asset classes into smaller segments. For example, equities could be categorized into sectors like shipping, technology or industry to explore detailed portfolio composition. Additionally, commodities such as oil and gold could be introduced to diversify the asset classes further. Additionally, the available data and computational power have been issues for training the model. Future work could extend the number of iterations as well as increase the model's complexity.

Testing other algorithms and libraries such like *RL_Coach*, *Pyqlearning*, and *MushroomRL* could be beneficial. Further work could adjust the existing algorithms to be better suitable for solving the capital allocation and consumption problem. Algorithms could also specifically be designed to optimize this problem from the ground up, rather than relying on standard algorithms from libraries.

Further work could involve testing additional parameters for smoothing and wealth penalty. For instance, agents could be penalized more for exceeding capital allocation and wealth constraints. Future research could also explore different utility functions and varying values of the risk aversion parameter.

Referanser

Bahoo, S., Alon, I., and Paltrinieri, A. (2020). Sovereign wealth funds: Past, present and future. *International Review of Financial Analysis*, 67:101418.

Baker, D. (2008). The housing bubble and the financial crisis.

Baldacci, Emanuele and Gupta, Sanjeev (2009). Fiscal expansions: What works. Accessed: May 18, 2024.

Barbary, V., Dixon, A. D., and Schena, P. J. (2023). *The Evolving Landscape of Sovereign Wealth Funds in a Changing World Economy: How Resilient Are the Santiago Principles?* Springer, 1st edition.

Barron, E. and Ishii, H. (1989). The bellman equation for minimizing the maximum cost. *NONLINEAR ANAL. THEORY METHODS APPLIC.*, 13(9):1067–1090.

Becker, M. D. B. (2024). Optimal investment consumption of swf by menas of reinforcement learning, incomplete excerpt!! *Extended Abstract for FBA Conference, Athens/Greece, 12th-14th June 2024*, pages 1–7.

Bernstein, S., Lerner, J., and Schoar, A. (2013). The investment strategies of sovereign wealth funds. *Journal of Economic Perspectives*, 27(2):219–238.

Brownlee, J. (2018). What is the difference between a batch and an epoch in a neural network. *Machine learning mastery*, 20.

Campbell, J. Y. and Viceira, L. M. (2001). Who should buy long-term bonds? *American Economic Review*, 91(1):99–127.

Chhaochharia, V. and Laeven, L. A. (2008). Sovereign wealth funds: Their investment strategies and performance. CEPR Discussion Paper No. DP6959.

Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. *Theory and Society*, 49(5):897–918.

CompaniesMarketcap.com (2024). Countries ranked by market cap. https://companiesmarketcap.com/all-countries/. Accessed on 29.02.2024.

Day, M.-Y., Yang, C.-Y., and Ni, Y. (2023). Portfolio dynamic trading strategies using deep reinforcement learning. *Soft Computing*.

Evans, D. (2005). The elasticity of marginal utility of consumption: Estimates for 20 oecd countries. *Fiscal Studies*, 26(2):197–224.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Fujimoto, S., van Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. Version: v3.

Garcia, F. and Rachelson, E. (2013). Markov decision processes. *Markov Decision Processes in Artificial Intelligence*, pages 1–38.

Groom, B. and Maddison, D. (2019). New estimates of the elasticity of marginal utility for the uk. *Environmental and Resource Economics*, 72(4):1155–1182.

Gyeyir, D. M. (2019). The Ghana Stabilisation Fund: Relevance and Impact so far. *Energy Policy*, 135:110989.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*.

Hambly, B., Xu, R., and Yang, H. (2023). Recent advances in reinforcement learning in finance. *Journal of Financial Engineering*, 7(1):1–15.

IFSWF (2024a). Santiago principles. https://www.ifswf.org/santiago-principles. Accessed on 29.02.2024.

IFSWF (2024b). What is a sovereign wealth fund? https://www.ifswf.org/what-is-a-sovereign-wealth-fund. Accessed on 29.02.2024.

Ingersoll, J. E. (1987). *Theory of Financial Decision Making*. Rowman & Littlefield, Totowa, NJ.

International Working Group of Sovereign Wealth Funds (2008). Santiago principles. urlhttps://www.iwg-swf.org/pubs/eng/santiago$_{p}rinciples.pdf.Accessed : March$20, 2024.

ISWF (2024). Countries ranked by market cap. `https://www.swfinstitute.org/fund-rankings/sovereign-wealth-fund`. Accessed: March 20, 2024.

Janssen, C. P. and Gray, W. D. (2012). When, what, and how much to reward in reinforcement learning-based models of cognition. *Cognitive science*, 36(2):333–358.

Jordà, O., Knoll, K., Kuvshinov, D., Schularick, M., and Taylor, A. M. (2019). The rate of return on everything, 1870–2015. *The Quarterly Journal of Economics*, 134(3):1225–1298.

Layard, R., Mayraz, G., and Nickell, S. (2008). The marginal utility of income. *Journal of Public Economics*, 92(8-9):1846–1857.

Leite, A., Candadai, M., and Izquierdo, E. J. (2020). Reinforcement learning beyond the bellman equation: Exploring critic objectives using evolution. In *Artificial Life Conference Proceedings 32*, pages 441–449. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . .

Levy, H. and Lerman, Z. (1988). The benefits of international diversification in bonds. *Financial Analysts Journal*, pages 56–64.

Li, K. (2002). What explains the growth of global equity markets? *Forthcoming in Canadian Investment Review*.

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

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access*, 7:154096–154113.

Malpezzi, S. (2014). Global perspectives on housing markets and policy. *Marron Institute of Urban Management Working Paper*, 3.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*.

Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. Yale University Press.

Mehra, R. and Prescott, E. (1985). The equity premium puzzle. *Journal of Monetary Economics*, 15:145–161.

Merton, R. C. (1975). Optimum consumption and portfolio rules in a continuous-time model. In *Stochastic optimization models in finance*, pages 621–661. Elsevier.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR.

Mork, K. A., Harang, F. A., Trønnes, H. A., and Bjerketvedt, V. S. (2023). Dynamic spending and portfolio decisions with a soft social norm. *Journal of Economic Dynamics & Control*.

Moutanabbir, K. and Noureldin, D. (2020). Optimal asset allocation and consumption rules for commodity-based sovereign wealth funds. *International Review of Economics & Finance*, 69:708–730.

NBIM (2024). Avkastningi. `https://www.nbim.no/no/oljefondet/avkastning/`. Accessed: May 10, 2024.

NBIM (2024a). Markedsverdi. `https://www.nbim.no/no/oljefondet/markedsverdi/`. Accessed: May 08, 2024.

NBIM (2024b). Norges bank investment management. `https://www.nbim.no/no`. Accessed on 24.01.2024.

NBIM (2024c). Slik er fondet investert. `https://www.nbim.no/no/oljefondet/slik-er-fondet-investert/`. Accessed: April 26, 2024.

Norges Bank (2020). Inflasjon. `https://www.norges-bank.no/tema/pengepolitikk/Inflasjon/`. Accessed: May 11, 2024.

Puterman, M. L. (1990). Markov decision processes. *Handbooks in operations research and management science*, 2:331–434.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8.

Rasmussen, M. (2003). Strategic and tactical asset allocation. In *Quantitative Portfolio Optimisation, Asset Allocation and Risk Management*, pages 273–290. Springer.

Regjeringen (2024). Handlingsregelen. *Government Publication*. Accessed on: 24.01.2024.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Upper Saddle River, 3rd edition.

Sahu, S. K., Mokhade, A., and Bokde, N. D. (2023). An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: Recent progress and challenges. *Appl. Sci.*, 13:1956.

Samuelson, P. A. (1967). General proof that diversification pays. *Journal of Financial and Quantitative Analysis*, 2(1):1–13.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.

Steigum, E. (2012). Sovereign wealth funds for macroeconomic purposes.

Sutton, R. S. and Barto, A. G. (1998). The reinforcement learning problem. *Reinforcement learning: An introduction*, pages 51–85.

Tizhoosh, H. R. (2005). Reinforcement learning based on actions and opposite actions. In *International conference on artificial intelligence and machine learning*, volume 414.

Tsang, Daisy (2023). White box vs. black box algorithms in machine learning. https://www.activestate.com/blog/white-box-vs-black-box-algorithms-in-machine-learning/. Accessed: May 16, 2024.

Utviklingsfondet (2024). Etiopia. https://www.utviklingsfondet.no/her-jobber-vi/etiopia. Accessed: March 20, 2024.

Wagner, D. (2013). Sovereign wealth funds: Investment objectives and asset allocation strategies. *SSRN*.

Werner, J. (2008). Risk aversion. In *The New Palgrave Dictionary of Economics*, pages 1–6.

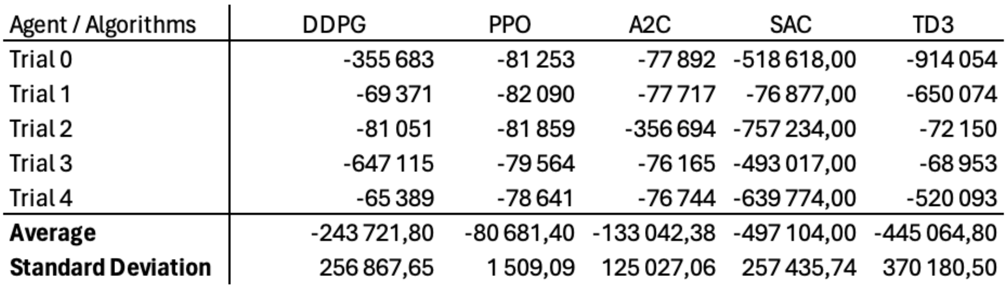
Yan, G. (2007). *Research on the Influencing Factors of Real Estate Price in Shanghai - An Example of Residential Market*. PhD thesis, Tongji University, Shanghai.

Yu, J., Xu, B., Yang, H., and Shi, Y. (2010). The strategic asset allocation optimization model of sovereign wealth funds based on maximum crra utility & minimum var. *International Conference on Computational Science (ICCS)*, 1(1):2433–2440.

Yu, P., Lee, J. S., Kulyatin, I., Shi, Z., and Dasgupta, S. (2019). Model-based deep reinforcement learning for dynamic portfolio optimization. *arXiv preprint arXiv:1901.08740*.

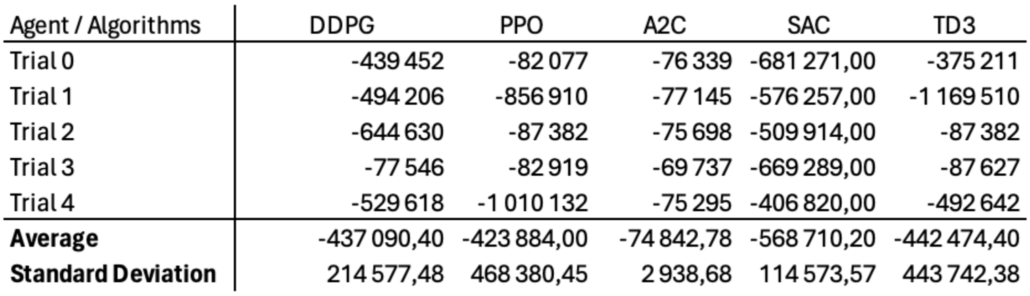
A. Results from the algorithms

Table A.1.: Reward with penalty of 0.4 and smoothing penalty of 0.1.



Agent / Algorithms	DDPG	PPO	A2C	SAC	TD3
Trial 0	-355 683	-81 253	-77 892	-518 618,00	-914 054
Trial 1	-69 371	-82 090	-77 717	-76 877,00	-650 074
Trial 2	-81 051	-81 859	-356 694	-757 234,00	-72 150
Trial 3	-647 115	-79 564	-76 165	-493 017,00	-68 953
Trial 4	-65 389	-78 641	-76 744	-639 774,00	-520 093
Average	-243 721,80	-80 681,40	-133 042,38	-497 104,00	-445 064,80
Standard Deviation	256 867,65	1 509,09	125 027,06	257 435,74	370 180,50

Table A.2.: Reward with penalty of 0.4 and smoothing penalty of 0.1.

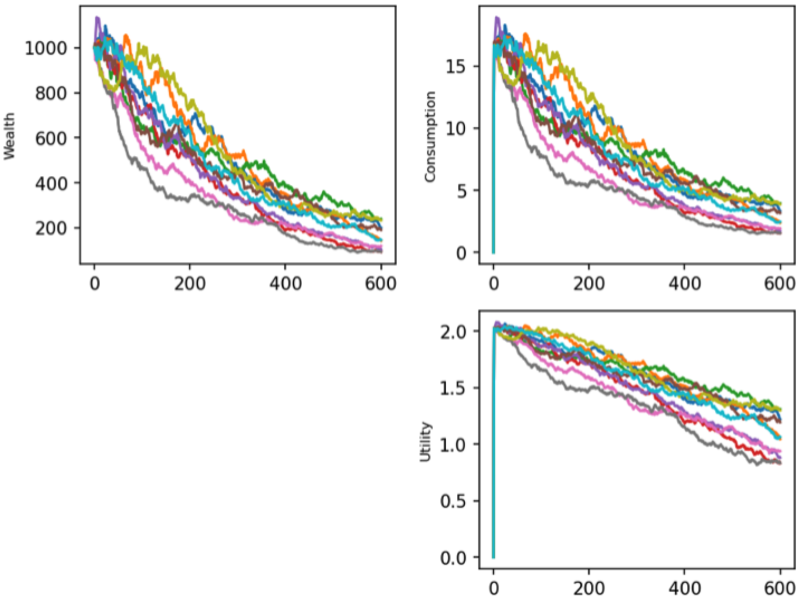


Agent / Algorithms	DDPG	PPO	A2C	SAC	TD3
Trial 0	-439 452	-82 077	-76 339	-681 271,00	-375 211
Trial 1	-494 206	-856 910	-77 145	-576 257,00	-1 169 510
Trial 2	-644 630	-87 382	-75 698	-509 914,00	-87 382
Trial 3	-77 546	-82 919	-69 737	-669 289,00	-87 627
Trial 4	-529 618	-1 010 132	-75 295	-406 820,00	-492 642
Average	-437 090,40	-423 884,00	-74 842,78	-568 710,20	-442 474,40
Standard Deviation	214 577,48	468 380,45	2 938,68	114 573,57	443 742,38

B. Graphical Plots

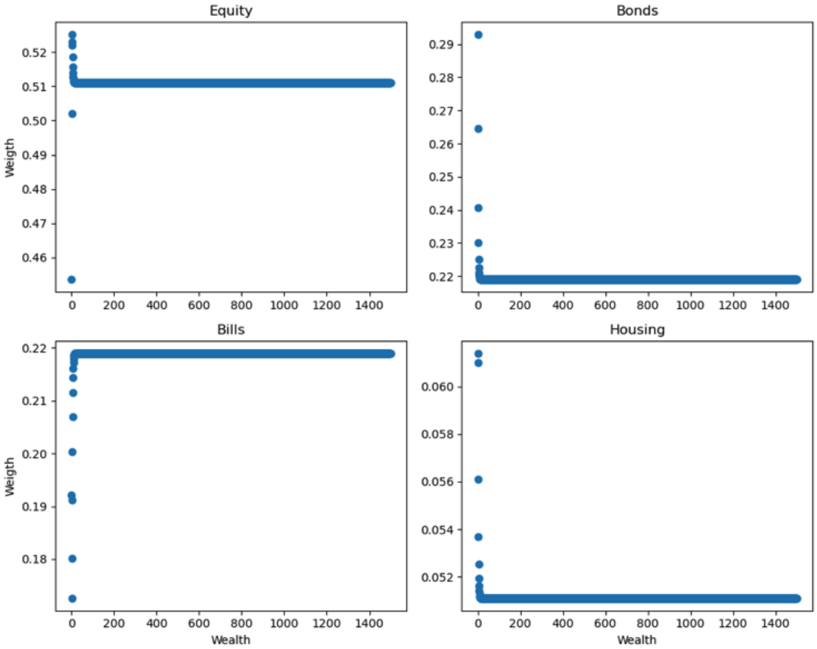
B.1. DDPG

Figure B.1.: Graphical plot DDPG



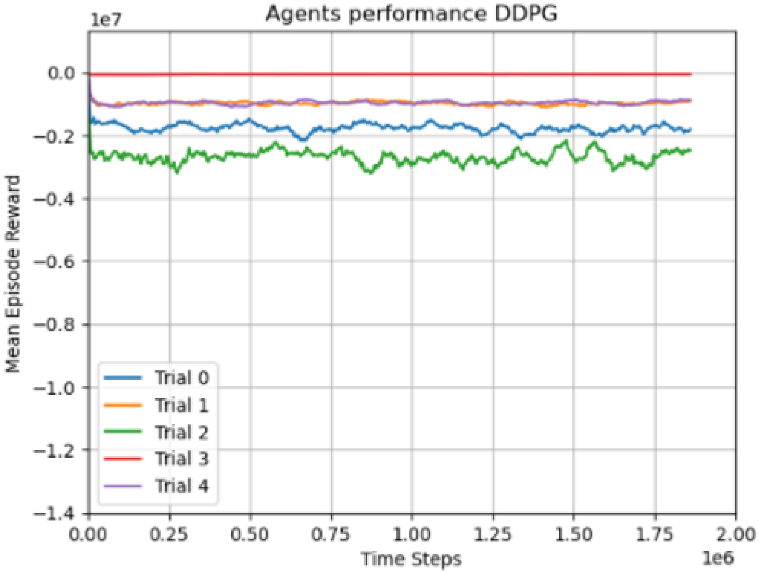
Trial 2 with wealth penalty of 0.8 and smoothing penalty of 0.3

Figure B.2.: Capital Allocation DDPG



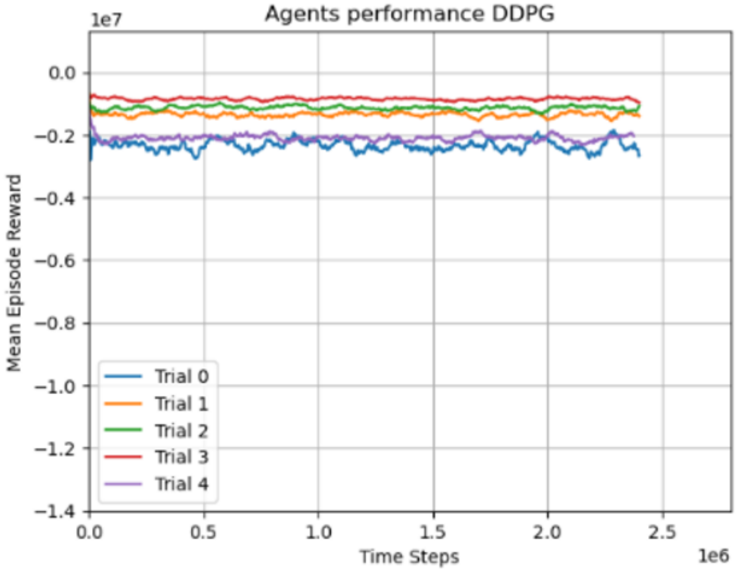
Trial 1 with wealth penalty of 0.8 and smoothing penalty of 0.3

Figure B.3.: TensorBoard DDPG



TensorBoard with wealth penalty of 0.8 and smoothing penalty of 0.3

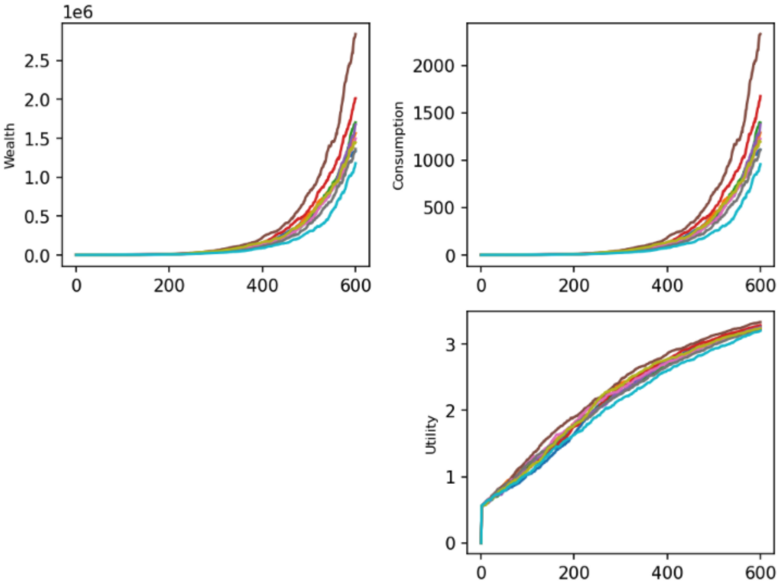
Figure B.4.: Tensorboard DDPG



TensorBoard with wealth penalty of 0.4 and smoothing penalty of 0.1

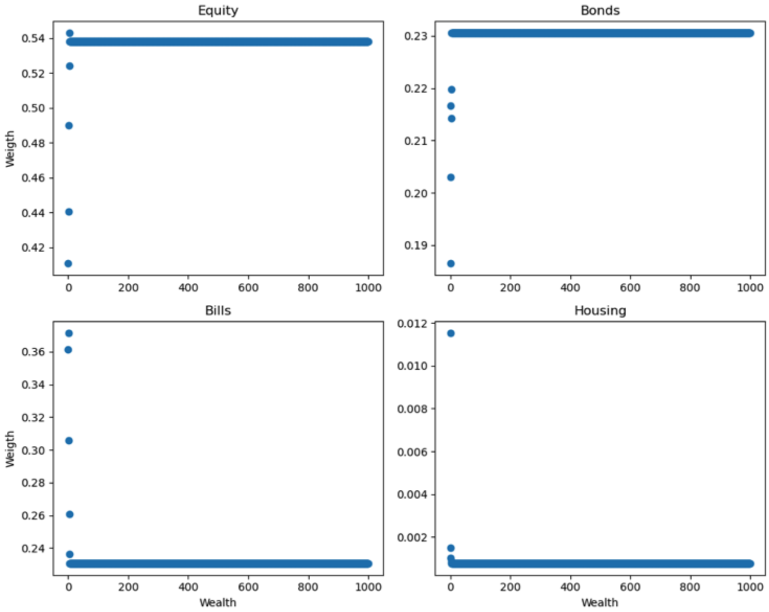
B.2. PPO

Figure B.5.: Graphical plot PPO



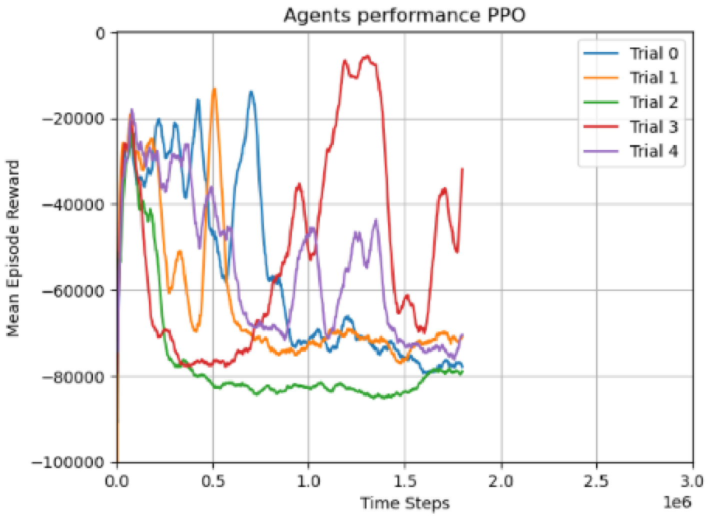
Trial 4 with wealth penalty of 0.8 and smoothing penalty of 0.3

Figure B.6.: Capital Allocation PPO



Trial 0 with wealth penalty of 0.4 and smoothing penalty of 0.1

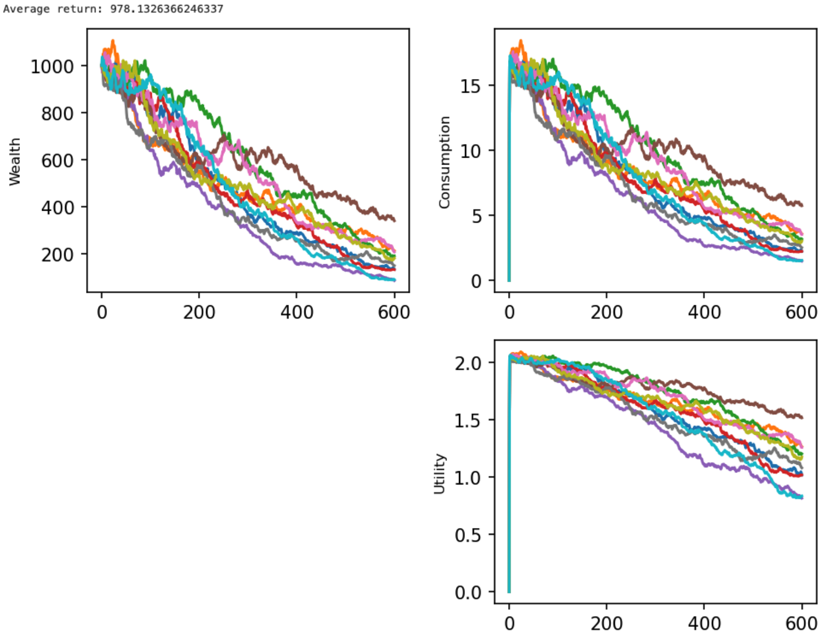
Figure B.7.: TensorBoard PPO



TensorBoard with wealth penalty of 0.8 and smoothing penalty of 0.3

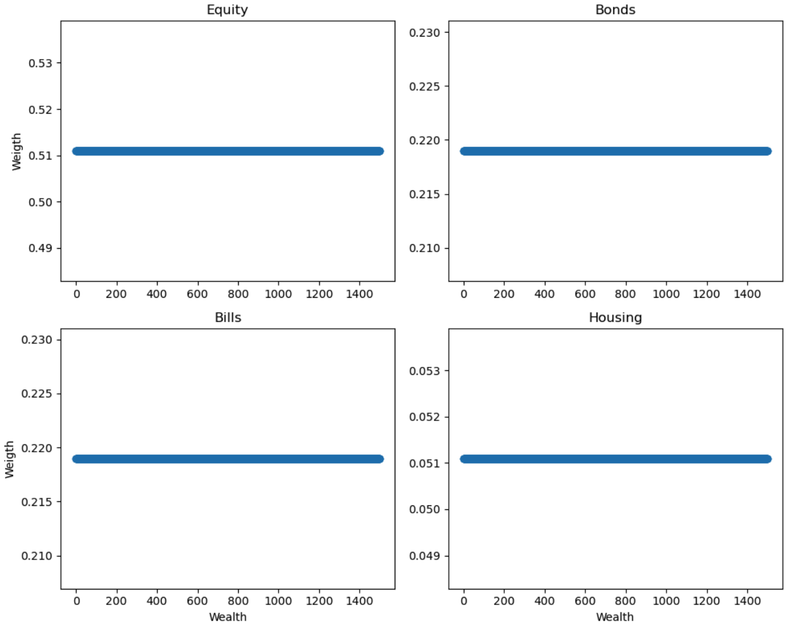
B.3. A2C

Figure B.8.: Graphical plot A2C



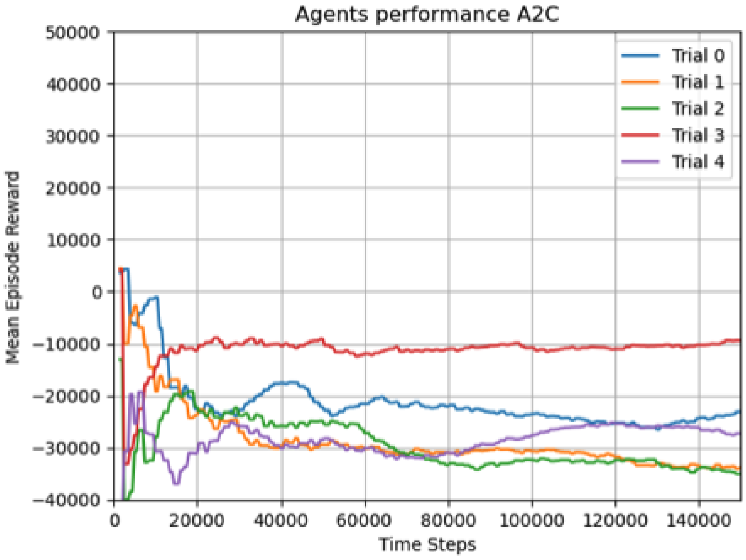
Trial 1 with wealth penalty of 0.4 and smoothing penalty of 0.1

Figure B.9.: Capital Allocation A2C



Trial 0 with wealth penalty of 0.8 and smoothing penalty of 0.3

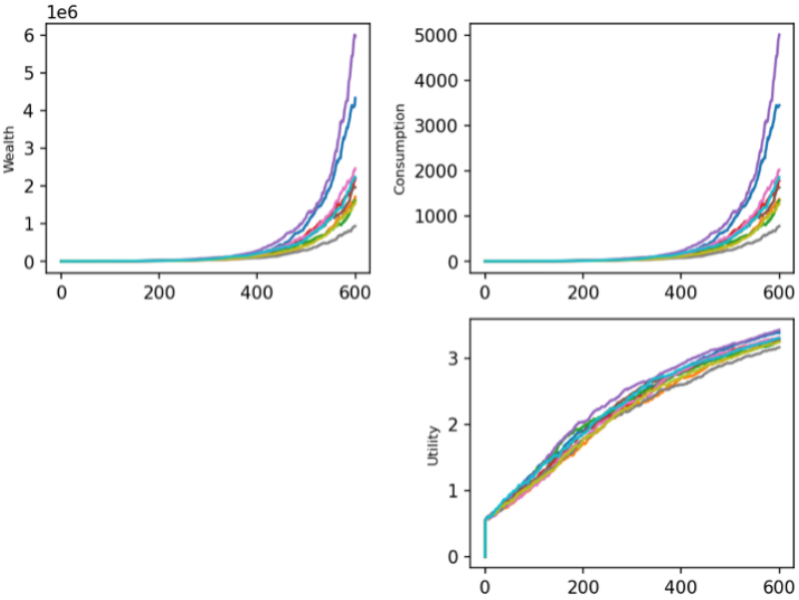
Figure B.10.: TenorBoard A2C



TensorBoard with wealth penalty of 0.8 and smoothing penalty of 0.3

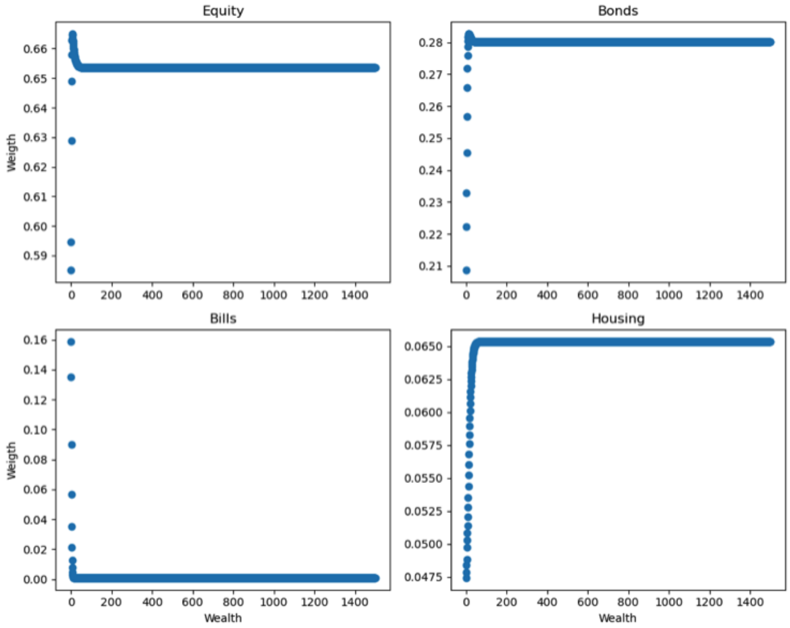
B.4. SAC

Figure B.11.: Graphical plot SAC



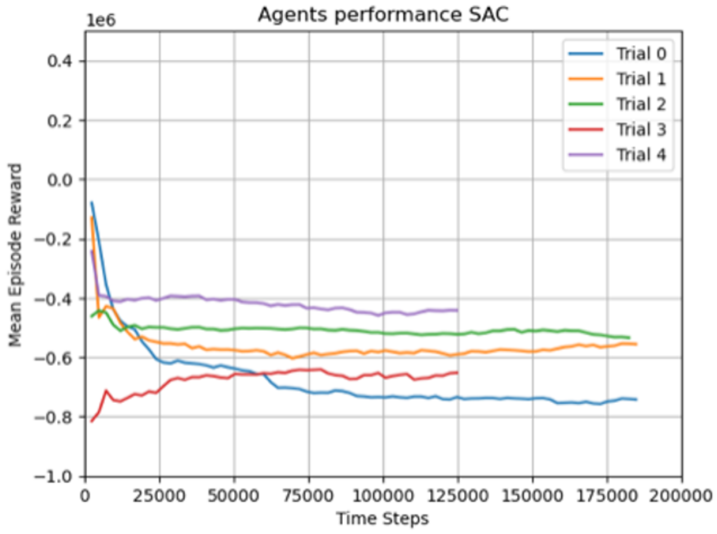
Trial 1 with wealth penalty of 0.8 and smoothing penalty of 0.3

Figure B.12.: Capital Allocation SAC



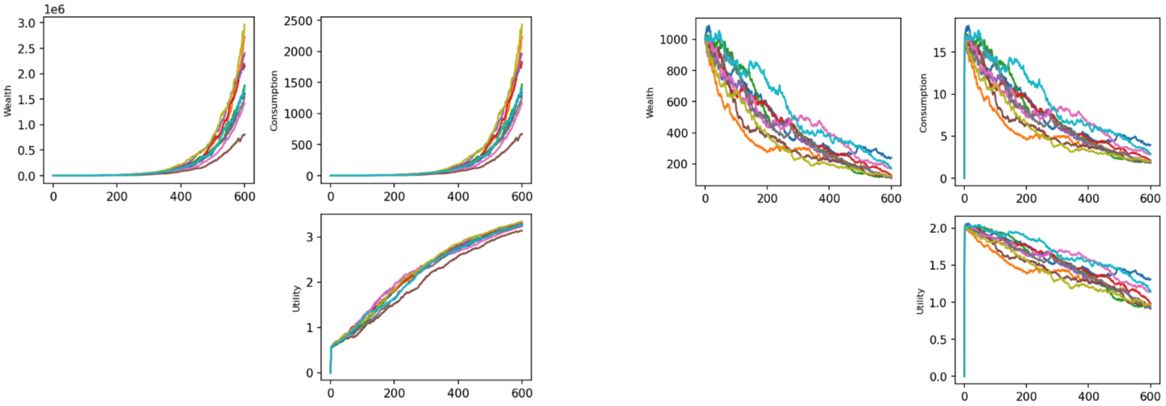
Trial 3 with wealth penalty of 0.8 and smoothing penalty of 0.3

Figure B.13.: TensorBoard SAC



TensorBoard with wealth penalty of 0.4 and smoothing penalty of 0.1

Figure B.14.: Graphical plots SAC

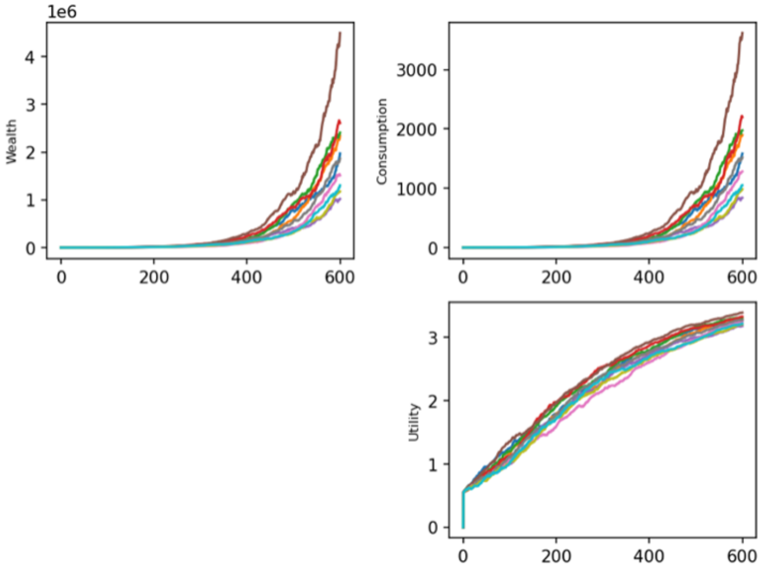


(a) Trial 0 with penalty 0.8 and 0.3

(b) Trail 1 with penalty 0.8 and 0.3

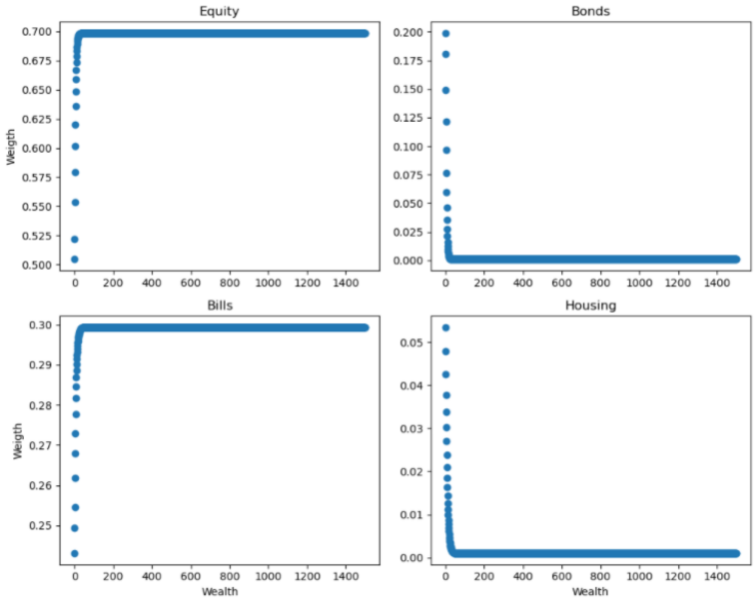
B.5. TD3

Figure B.15.: Graphical plot TD3



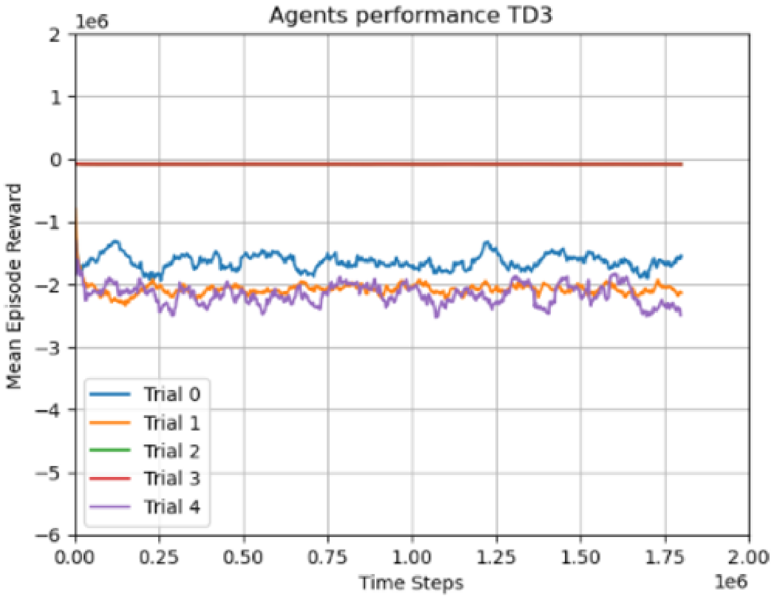
Trial 0 with wealth penalty of 0.4 and smoothing penalty of 0.1

Figure B.16.: Capital Allocation TD3



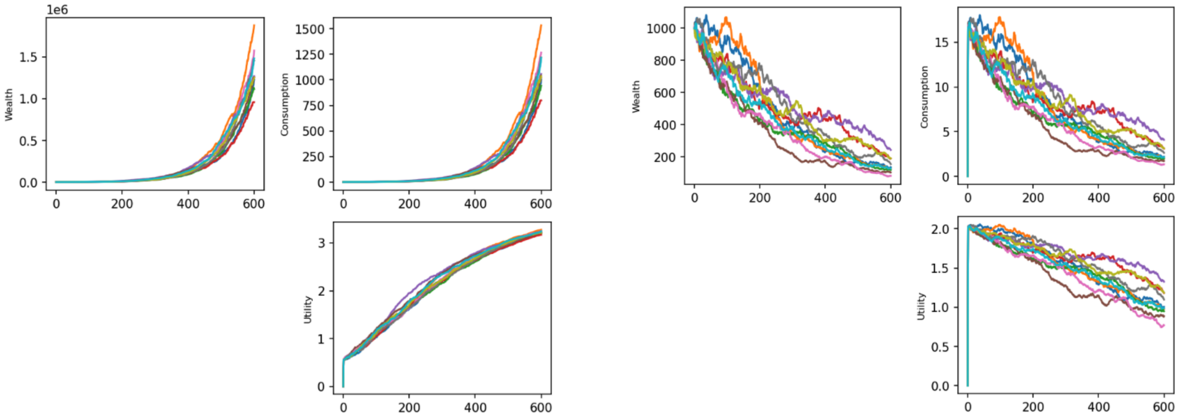
Trial 4 with wealth penalty of 0.4 and smoothing penalty of 0.1

Figure B.17.: TensorBoard TD3



Tensorboard with wealth penalty of 0.4 and smoothing penalty of 0.1

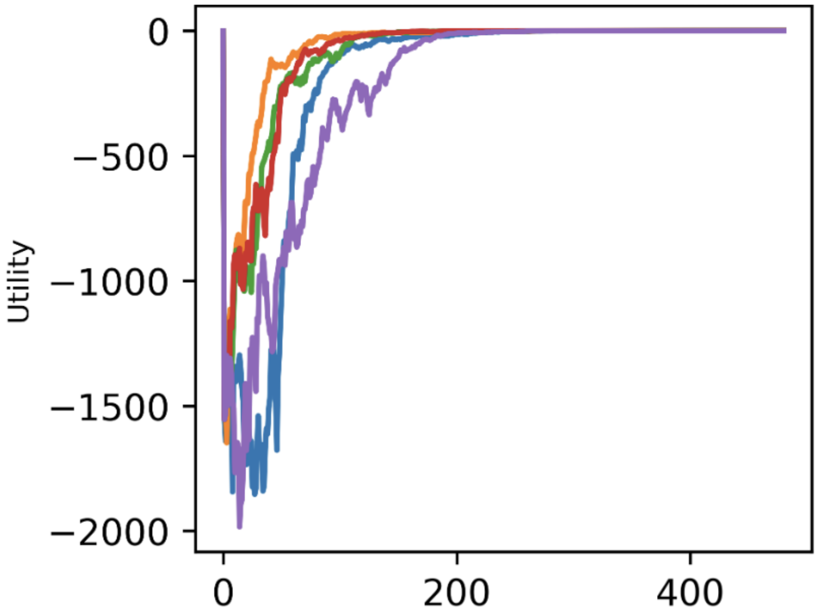
Figure B.18.: Graphical plots TD3



(a) Trial 1 with penalty 0.8 and 0.3

(b) Trail 2 with penalty 0.8 and 0.3

Figure B.19.: Utility using Morks utility function



An example of how the TD3 algorithms perform, by using the utility function to Mork.

NTNU

Norwegian University of
Science and Technology