分类号：(F224.0，O211.6)　　单位代码：10363
密　级：公开　　学　　号：2221020125

题　　目　基于强化学习连续时间均方投资组合研究

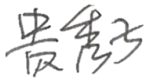
题　　目　Research on Continuous Time Mean-Variance Portfolio Based on Reinforcement Learning

学生姓名：　贵秀子
指导教师：　费为银　教授
专　　业：　数　学
研究方向：　智能金融

论文答辩日期：　2025　年　5　月　14　日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名： 

日期：2025 年 5 月 14 日

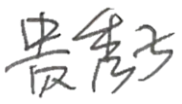
安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：  　　指导教师签名： 

日期：2025 年 5 月 14 日　　日期：2025 年 5 月 14 日

基于强化学习连续时间均方投资组合研究

摘 要

随着金融市场的日益复杂化和全球化，投资组合管理面临着前所未有的挑战。强化学习（Reinforceing Learning，RL）作为一种先进的机器学习方法，在解决复杂决策问题方面展现出巨大的潜力。强化学习能够根据不同市场环境状态寻找最优策略，通过不断学习和调整，优化投资组合的资产配置，以达到最佳收益，本文通过定义奖励函数，利用均值方差模型将给定预期收益作为奖励信号，指导策略优化，以实现风险最小化，随后将风险管理嵌入策略中，通过香农熵达到平衡风险与收益的目的，最后通过平衡探索新策略和利用已知策略，在不断变化的市场中捕捉新的机会，降低投资风险。

一方面，随着全球对气候变化的日益关注，“双碳”目标已成为各国政府和企业的重要议题。我国政府为了顺应人民对美好生活的新期待，提出了绿色发展的时代主题。在这一背景下，本文将环境、社会和公司治理（ESG）因素融入强化学习框架下的投资组合研究中。首先，本文将企业 ESG 评级和投资者 ESG 偏好引入到连续时间均值-方差投资组合的模型当中，推导出了满足自融资条件下基于 ESG 评级的最优值函数的 HJB 方程；然后，我们证明了一个策略改进定理，并针对不同的 ESG 评级水平设计了一个可实现的 RL 算法；最后通

过数值模拟，讨论了 ESG 评级以及 ESG 偏好的变化对投资者终端财富的影响。

另一方面，通货膨胀问题历来是经济金融领域的核心议题之一，对社会发展和民众生活均产生深远影响。当前，为了应对全球经济衰退的挑战，众多国家纷纷采取了宽松的财政与货币政策。在我国更加开放的政策背景下，不仅需要应对国内经济环境不确定带来的的各种挑战，还需高度关注全球化趋势所带来的影响。鉴于此，本文在通胀不确定这一复杂环境下，构建强化学习驱动的均方投资组合模型，利用随机分析理论推导通胀不确定环境下的财富演化动力学方程。在此基础上通过数值模拟，讨论了通胀不确定对投资者终端财富、值函数以及最优策略的影响。

最后，对全文进行了总结和展望。

关键词：强化学习，均值方差模型，投资组合，动态规划原理，ESG，通货膨胀

RESEARCH ON CONTINUOUS TIME MEAN-VARIANCE PORTFOLIO BASED ON REINFORCEMENT LEARNING

ABSTRACT

With the increasing complexity and globalization of financial markets, portfolio management is facing unprecedented challenges. Reinforcement learning (RL), as an advanced machine learning method, shows great potential in solving complex decision problems. Reinforcement learning can find the optimal strategy according to different market environment states, optimize the asset allocation of the portfolio through continuous learning and adjustment, and achieve the best return. This paper defines the reward function and takes the given expected return as the reward signal through the mean-variance model to guide the strategy optimization to minimize the risk. Then, risk management is embedded in the strategy. Shannon entropy is used to achieve the purpose of balancing risk and return. Finally, by balancing exploration of new strategies and utilization of known strategies, new opportunities are captured in the changing market and investment risks are reduced.

On the one hand, with the increasing global concern about climate change, the "dual carbon" goal has become an important issue for governments and enterprises. In order to meet the people's new expectations for a better life, the Chinese government has put forward the theme of green development. In this context, environmental, social and corporate governance (ESG) factors are integrated into portfolio research under the reinforcement learning framework. Firstly, this paper introduces the enterprise ESG rating and investor ESG preference into the continuous time mean-variance portfolio model, and deduces the HJB equation satisfying the optimal value function based on ESG rating under the condition of self-financing. Then, we prove a policy improvement theorem and design a realizable RL algorithm for different ESG rating levels. Finally, through numerical simulation, the influence of ESG rating and ESG preference on the end wealth of investors is discussed.

On the other hand, inflation has always been one of the core issues in the economic and financial field, which has a profound impact on social development and people's life. At present, in order to cope with the challenges of the global economic recession, many countries have adopted loose fiscal and monetary policies. In the context of Chinese more open policy, it is necessary not only to deal with the challenges brought by the uncertain domestic economic environment, but also to pay close attention to the impact of the trend of globalization. In view of this, this paper

constructs a mean square portfolio model driven by reinforcement learning under the complex environment of inflation uncertainty, and uses stochastic analysis theory to derive the dynamic equation of wealth evolution under the environment of inflation uncertainty. On this basis, through numerical simulation, the effects of inflation uncertainty on investors' terminal wealth, value function and optimal strategy are discussed.

Finally, the paper summarizes and looks forward to the research.

Gui Xiuzi (Applied Mathematics)

Surpervised by Fei Weiyin

KEYWORDS: Reinforcement learning, ESG, inflation, Mean-variance model, Dynamic programming principle , Investment portfolio

目 录

第 1 章 绪论

1.1 研究背景和意义

投资组合是将一定资金分配给多个金融资产的决策过程，通过不断改变分配权重来提高收益，减小风险。马科维茨均值-方差投资组合选择理论提供了一种通过权衡风险和回报来优化投资组合的方法，是现代金融理论发展的里程碑。然而，由于金融资产价格难以预测，交易数据体量庞大，而且具有严重的非线性，在实践中仅靠人脑难以做出最优决策。

传统的投资组合理论是对过去投资经验的总结，偏重于质的分析，而缺乏量的分析。在选择投资工具的种类及其数量构成时，主要依赖投资决策者的直觉和先前经验，缺乏一套系统的、精密计算的科学方法。随着人工智能技术的发展，人们已经可以利用计算机进行学习和大量计算并实现优化投资组合管理。近年来，在电子竞技与棋类竞技领域，深度强化学习技术大放异彩，其通过不断试错在复杂多变的环境中展现出卓越的决策能力。受此启发，若将金融市场视为一个持续演进的动态环境，强化学习技术便拥有了得天独厚的优势。当前，许多强化学习方法聚焦于预测金融市场的价格波动轨迹或趋势，它们以历史价格数据为基石，通过模型训练输出预测值，进而指导投资决策。

另外，在当今全球气候变化的大背景下，绿色金融和脱碳减排理念逐渐深入人心。党的二十大报告提出：“积极稳妥推进碳达峰碳中和”，这是以习近平同志为核心的党中央统筹国内国际两个大局作出的重大决策部署。然而，推动企业进行 ESG（Environmental, Social and Governance） 实践和发展对于实现“双碳”目标的国家战略具有重要意义。在当前国家政策指导和投资者具有 ESG 投资偏好的背景下，投资者也将关注点逐渐从短期的财务收益转向企业的长期价值和潜在风险。ESG 因素作为评估企业长期价值和风险的重要指标，越来越被投资者所重视。ESG 理念也开始嵌入到投资者的投资决策中。投资者通过投资 ESG 表现优异的企业，可以推动企业贯彻并践行 ESG 理念，从而促进经济、社会和环境的可持续发展。这不仅有助于塑造企业良好的社会形象，还能显著提升企业

的竞争优势和品牌价值。鉴于此，本课题考虑基于强化学习方法研究 ESG 因素对投资组合优化问题的影响。

同时，全球通货膨胀已成为当今社会一个备受瞩目的热点议题，它深切地影响着投资者与公司的决策与运营。从金融危机到新冠肺炎疫情，世界各国为了应对经济下行压力，普遍采取了宽松的货币和财政政策。然而，这些措施在刺激经济增长的同时，也产生了通胀加剧的副作用，使得物价水平持续攀升，并且未来通胀的走向仍充满不确定性，因此可能会对金融经济市场造成持续不断的不确定影响。然而，通胀不仅影响着金融市场，同样影响着投资者对风险资产的投资比率和投资组合选择。在此背景下，通胀预期成为了投资者在制定策略时的关键因素，尤其是在复杂多变的市场环境中，如何灵活调整并更新投资策略和经营策略，以应对通货膨胀带来的风险和挑战，保持资产价值，是投资者和公司亟需解决的重要问题，也是本文研究的焦点。故本课题拓展性地构建了通胀环境下基于强化学习方法的投资组合优化模型，着重分析了通胀波动率如何影响投资组合优化，从而向投资者提供一些更加符合现实状况的决策建议。

1.2 国内外研究现状

均值-方差（Mean-Variance，MV）是投资组合选择的标准之一，马科维茨[1]针对单期投资组合选择的开创性工作中提出了一种资产配置策略，该策略在以预定的平均收益为目标的同时最小化最终收益的方差。在此基础上，很多学者针对 MV 问题以及其推广形式进行了充分的研究。但在解决经典的基于模型的均值方差问题时往往需要从资产价格的历史时间序列中估计市场参数，可是金融资产价格难以预测，交易数据体量庞大，并且具有严重非线性，在实践中很难依靠人脑做出最优决策，RL 技术基本上不需要任何模型参数的估计，其通过与未知的投资环境直接互动就可以直接输出最优（或接近最优）分配的。在离散时间情形下，Li 和 Ng[2]提出了一种针对最优动态投资组合选择的多期均值-方差公式的解析解，扩展了马科维茨的单期结果。Nevmyvaka 等[3]通过 Q 学习方法提出了强化学习在现代金融市场优化交易执行这一重要问题上的首次大规模实证应用。Jiang 等[4]基于 DQN 展现了金融投资组合管理框架，以实现最大化收益和最小化风险的目标。

但是经典的强化学习方法往往无法解决状态和动作空间维度很高的问题，这时的深度学习技术依然解决的是离散动作的问题，如果直接应用于连续的投资组合管理可能容易陷入维数灾难。针对这样的问题，DeepMind 团队于 2021 年提出了深度确定性策略梯度方法（Deep Deterministic Policy Gradient，DDPG），这种算法使用策略网络输出确定性动作，成功地解决了连续动作空间下的强化学习问题。Lilicrap 等[5]将深度 Q-Learning 的基础思想应用于连续动作领域，提出了一种基于确定性策略梯度的可以在连续的动作空间上运行的无模型（model-free）算法。Zhou 和 Li[6]将连续时间均值-方差投资组合选择模型公式化为一最大化预期终值收益和最小化终值财富的方差的双准则优化问题，并将该问题嵌入辅助随机线性二次（LQ）问题中。Munos[7]采用了动态的方法规划（DP），满足非线性一阶的值函数并推导出了 HJB 方程 Sato 和 Kobayashi[8]提出了一种风险规避强化学习算法，并在单个期货交易系统上进行测试，改进了传统的投资组合管理系统的资产配置策略。Doya[9]在没有进行时间、状态和动作的先验离散化的前提下提出了一个连续时间动态系统的强化学习框架，而且针对策略改进提出了连续评价算法以及基于价值梯度的贪婪策略（ε-greedy）两种算法。Prashanth 和 Ghavamzadeh[10]提出了用于折扣和平均奖励马尔可夫决策过程（MDP）的方差约束的评价（Actor-Critic）算法，而且证明了算法收敛到局部风险敏感的最优策略，并在交通信号控制应用中展示了其有用性。

多数强化学习算法由两个迭代过程组成: 策略评估和策略改进，前者为当前策略提供了一个估计值函数，而后者则向正确的方向更新当前策略以改进价值函数。Haarnoja 等[11]证明了 PITS 在无限视界中的离散时间熵正则化 RL 问题，Jacka 和 Mijatoviki[12]证明了在连续时间的经典随机控制问题。Sutton 和 Barto[13]发现 PIT 是可解释 RL 算法的关键先决条件，它确保迭代值函数不增加（在最小化问题的情况下），并最终收敛到最优值函数。Liang 等[14]将连续强化学习算法应用在投资组合管理中，并在中国股票市场进行了大量实验发现基于策略梯度（Policy Gradient ，PG）的智能体优于均匀常数再平衡投资组合策略。Filos[15]探讨了强化学习技术在投资组合管理中的应用，而且他们发现即使考虑交易成本，它在年化累计收益和夏普比率方面也表现优于传统模型。Wang[16]设计了一个可扩展的数据高效的强化学习算法，并利用标准普尔 500 指数股票的数据进行实证测试发现很大程度上优于传统经济学方法 Wang 等[17]使用随机控制理论

在线性二次情况下提出了一种连续时间和空间的 RL 问题的公式化方法。Wang 和 Zhou[18]提出了一个熵正则化、松弛的随机控制模型，以实现探索（学习，Exploration）和利用（优化，Exploitation）之间的最佳权衡并表明它在模拟和实证研究中优于传统的深度神经网络的算法。Calabuig 等[19]提出了一种基于 Lipschitz 型扩展的强化学习模型，用于金融市场中的投资策略，并讨论了它与金融中使用的其他强化学习和时间序列预测方法的关系。Ye 等[20]提出了一种新颖的状态增强强化学习（SARL）框架用于投资组合管理。Wang 和 Yu[21]提出了一个全周期数据驱动的投资机器人咨询框架，该智能体结合了逆优化和深度强化学习方法来管理多种资产并优化投资组合。Jiang 等[22]建议使用强化学习算法来实现 Kelly 策略，并得出了 RL Kelly 策略优于传统的 Kelly 策略的结论。

以上内容梳理了部分强化学习方法与金融经济学问题研究相结合的文献，但他们没有考虑绿色金融和 ESG 问题。实际上绿色金融和 ESG 理念逐渐成为人们的共识，也是国内外学者关注和研究的焦点之一，ESG 投资理念受到了广泛关注。张桂品和田增瑞[23]的研究发现，ESG 赋予了企业新的发展理念，是企业可持续发展的必由之路。Xu 等[24]基于随机系统理论和契约理论，构建了 ESG 评级 Knight 不确定性下具有模糊厌恶的连续时间委托代理模型。费为银等[25]发现 ESG 因素（包括用户偏好、利用成本和政府支持）会影响平台利用，并确定了 ESG 影响下代币安全特征的最优水平。Li 等[26]创新性地从 ESG 的角度探讨基于区块链技术的代币平台中 ESG 因素对参与者决策的影响和持续冲击环境对代币平台运行的影响，并且推导出了利用者价值函数满足的 HJB 方程。Li 等[27]将 ESG 理念引入到动态金融契约模型的设计中，而且研究结果表明，考虑绿色技术投资和可持续发展的企业可以有效地提升企业价值。

与此同时，有些学者探究 ESG 因素对金融资产和投资组合选择的影响。He 等[28]研究了公司的 ESG 综合得分对股票流动性的影响，研究发现，更好的 ESG 绩效会导致更高的股票流动性。Friede 等[29]基于 2000 多项实证研究的分布概括出 ESG 评级较高的公司比 ESG 评分较低的公司表现出更好的财务绩效。Fatemi 等[30]研究表明 ESG 投资者认为他们可以从 ESG 信息中受益，从而指导他们的投资决策。Liagkouras 等[31]通过引入一种算法在符合 ESG 约束的领域进行投资组合优化。Pedersen 等[32]提出了一些投资者偏向 ESG 分数高的公司，而且当市场由这类投资者主导时，ESG 得分高的公司组合可以产生更高的财务绩效。Chen

等[33]将 ESG 评分与财务指标相结合构建了投资组合优化模型，其实证结果表明，他们构建的投资组合在各个方面都优于传统投资策略。Pástor 等[34]证明了代理人对绿色资产持有的偏好会影响资产价格。徐凤敏等[35]的研究表明，积极的环境观点、治理观点和综合 ESG 观点与未来股票收益率显著正相关，而且 ESG 理念对股票收益率的作用在“双碳”目标提出后有所增强。Fereydooni[36]利用了一个在线投资组合选择策略，强调了在投资组合选择中考虑 ESG 因素的重要性，并建议投资者可以从 ESG 投资中受益。Díaz 等[37]评估了能源行业 ESG 标签公司的月度四分位数投资组合，认为高 ESG 评级的投资组合在投资回报表现更好。

回溯至 2008 年金融危机的冲击，全球经济陷入低谷，为促进经济回暖，多国政府纷纷采取了量化宽松的货币与财政政策作为应对之策，这一举措虽旨在提振经济，却不经意间在全球范围内引发了通胀的连锁反应，我国同样面临挑战，通胀因素对资产配置所产生的显著影响已经引起国内外学者的高度关注和广泛研究，Brennan[38]在只有名义资产可用的情况下，针对有限期投资者的资产配置问题提出了框架。并分析了通胀、利率动态和投资期限对最优投资组合选择的影响。Siu[39]在一个存在通胀风险的完备市场中建立了一个长期战略资产配置的定量模型。费为银等[40]将资产价格扩展到跳扩散环境下，并利用 HJB 方程推导出最优投资策略，得出了最优动态资产配置策略的近似解

在最优投资方面，姚海祥等[41]基于连续时间均值-方差框架，研究了通胀影响下投资终止时间不确定的最优投资组合选择问题。董迎辉等[42]在一个连续时间金融市场的背景下，考虑通胀指数债券、股票和无风险债券三类资产，并分析了通胀风险和两种薪酬方案对最优投资策略的影响。

众多学者们还将研究扩展到最优消费的层面，例如，Bensoussan 等[44]在完全观测到的消费篮价格和部分观测到消费篮价格两种情形下研究了在随机通胀下的最优消费和投资组合决策问题。Chou 等[45]建立了通胀影响及两因素随机利率下的幂效用最大化模型，并利用随机动态规划方法求得了相应的最优投资和消费策略。Lim[46]应用鞅方法研究了通胀风险和生存约束下的最优投资组合问题，并且得到了最优消费率和最优投资组合。Fei[47]在一个具有马尔科夫切换和通货膨胀的金融市场中，利用马尔可夫切换过程的广义 Ito 公式，相对避险效用为常数的最优消费策略和投资组合策略。在通胀率和利率是随机的，预期通胀率是不可观察，而且投资者对消费价格水平的正确过程存在不确定性的情形下，Munk

和 Rubtsov[48]采用鲁棒控制方法推导出最优投资策略的封闭形式并说明了最优投资组合的特性。Fei 和 Fei[49]提出了一个马尔可夫转换系统的最优控制框架，将其应用于通胀下的最优投资组合和消费决策问题，并刻画了满足带有马尔可夫转换的广义 HJB 方程的最优控制策略。费为银等[50]通过考虑模型误定和通胀的随机波动对消费和投资组合带来的影响，建立投资者决策的值函数所满足的 HJB 方程，并推导了最优消费和投资决策的显示解。费晨等[51]利用动态规划原理分别在允许和不允许风险资产对冲的两种情形下得出了最优消费以及最优投资组合选择策略，并且分析了通胀如何影响企业家的最优消费和投资组合策略。

Kwak 和 Lim[52]研究了通胀风险下一个家庭的连续时间最优消费、投资和人寿保险决策问题。用一个流动性很强的通胀挂钩指数债券市场来对冲通货膨胀风险，并利用鞅方法导出了恒定相对风险规避效用情况下的显式解。Han 和 Huang[53]求解了在利率和通胀风险下工薪阶层退休前的最优人寿保险、消费和投资组合决策，他们的结果表明，工薪阶层的收入与通胀挂钩，并且人寿保险需求将随着通胀水平的提高而增加。费为银等[54]分别研究了在通胀下退休前后的代理人的退休选择、最优消费和投资组合决策问题。费为银等[55]在通胀环境下通过动态规划原理构建了以最优工作努力策略和股权激励为控制变量的满足公司和高管利益最大化的 HJB 方程，并分析了通胀风险对高管的股权激励和工作努力策略的影响。费为银等[56]在通货膨胀背景下，构建了政府最优税收和债务模型，分析通胀对政府最优税收和债务的影响，以及对达到债务容量的时间影响，并预测我国到达债务容量的时间。

1.3 论文内容及安排

基于上述分析，本文考虑在 Wang 和 Zhou[18]的模型中分别引入 ESG 因素和通胀，得到通胀不确定环境和兼顾 ESG 表现的情形下满足代理人自融资假设的最优投资组合。首先，在理论上给出 ESG 环境下投资者的价值函数，在导出投资者的价值函数所满足的哈密尔顿-雅可比-贝尔曼（HJB）方程，然后在给定边界条件下，通过数值模拟，分析企业 ESG 得分以及投资者 ESG 偏好对投资者终端财富以及价值函数的影响，并给出经济学解释。本文还考虑了现实世界的复杂多变性对投资者决策的影响，引入通胀因素，并在此基础上分析通胀参数对投资

者最优短期和长期主义投资决策的影响，以及各种参数之间的相关性对投资者短期、长期的影响，从而得出一些具有一定实际意义的经济学结论，以向投资者提供合理的管理和决策建议。

本文结构安排如下：

第 1 章　绪论，主要说明了本课题在经济学领域的研究背景和意义、国内外关于该课题的研究现状和进展、文章的研究方法以及论文的总体安排。

第 2 章　基于强化学习框架下，在价格过程中引入 ESG 因素，构建一个松弛随机控制问题，通过引入拉格朗日乘子方法将原始约束问题纳入目标函数来解决风险最小化问题；接着，在满足自融资假设的条件下引入动态方程的探索性版本，利用动态规划原理和伊藤公式推导出满足代理人预期收益风险最小化的 HJB 方程，然后通过对 HJB 方程取一阶条件得到最优控制分布，证明策略改进定理是高斯的，设计探索性均值方差投资组合选择算法，通过数值模拟绘制出不同的 ESG 参数对投资者终端财富的影响。并结合经济学原理给出经济学解释以及具有一定实际意义的建议。

第 3 章　在通胀环境下建立动态模型。利用通胀过程对投资组合的名义财富过程进行折现，得到通胀不确定环境下的实际财富演化的随机微分方程，再通过应用动态规划的方法得到问题的解析解, 进而求解出原均值-方差模型的有效投资策略的解析表达。最后，通过与经典均方问题的可解性等价，确定成本函数并将其参数化。设定不同的通胀波动率参数分析通胀不确定性对投资选择的影响。

第 4 章　对全文进行总结和展望。

第 2 章 ESG 对基于强化学习的连续时间均值-方差投资组合的影响

2.1 基本模型框架

本章考虑 ESG 因素的基于强化学习框架下的连续时间均方模型的投资组合问题研究，首先假定在 0 时刻构建一个松弛随机控制问题，由于投资者对于 ESG 偏好的不同，考虑在价格过程中引入 ESG 评级，构建包含 ESG 评级对于投资组合影响的模型，通过引入拉格朗日乘子方法将原始约束问题纳入目标函数来解决风险最小化问题；接着，在满足自融资假设的条件下引入动态方程的探索性版本，利用动态规划原理和伊藤公式推导出满足代理人预期收益风险最小化的 HJB 方程；最后，通过对 HJB 方程取一阶条件得到最优控制分布，然后将最优反馈控制带入到 HJB 方程中根据终端条件求出最优值函数。证明策略改进定理是高斯的并且通过数值模拟得到探索性均方问题的全局最优解。

2.1.1 经典 MV 问题

在这一节，在 RL 背景下，我们考虑连续时间的探索性、熵正则化马科维茨的 MV 投资组合选择问题，其中 MV 问题是一个特殊的情况。我们首先回顾连续时间中的经典 MV 问题，固定投资规划期$\mathrm{T}>0$，B_t是定义在满足常规条件的带流概率空间$\left(\Omega,\mathcal{F},\{\mathcal{F}_t\}_{0\le t\le T},\mathbb{P}\right)$上的标准一维布朗运动，在不考虑 ESG 因素，同样不考虑 RL 时，风险资产的价格过程是一个几何布朗运动：

$$dP_t = P_t\,(\mu dt+\sigma dB_t),\quad 0\le \mathrm{t}\le \mathrm{T}.$$

其中$P_0=p_0>0$是$\mathrm{t}=0$时的初始价格并且μ和σ分别为风险资产的预期收益率和波动率。

在 Wang 和 Zhou[18]的研究框架基础上，本文考虑 ESG 因素。Pástor 等[34]研究表明，ESG 偏好会影响资产价格，费为银等[25]研究发现高 ESG 评级会导致平台代币价格更高，所以考虑 ESG 因素时，股票价格过程假设为：

$$dP_t = P_t\left((\mu+\gamma g)dt+\sigma dB_t\right), g\in[-1,1]. \tag{2.1}$$

$\gamma(\gamma \geq 0)$ 是 ESG 强度参数，也就是投资者 ESG 偏好水平，这是投资者根据他们的偏好而选择的值，风险资产所属公司的 ESG 评分为 $g\left(g \in [-1,1]\right)$，无风险资产的固定利率$r$> 0，夏普比率 $\rho = \dfrac{\mu - r}{\sigma}$，用 $\left\{w_t^u, 0 \leq t \leq T\right\}$ 表示代理人投资于风险股票的贴现财富过程，该代理人以策略 $u = \left\{u_t, 0 \leq t \leq T\right\}$ 重新平衡其在风险和无风险资产中的投资组合的贴现财富过程。同时满足标准的自融资假设（标准自负盈亏假设）和其他技术条件，这些条件将在下面详细说明，由（2.1）式得到下列财富过程：

$$dw_t^u = u_t \frac{dP_t}{P_t} - ru_t dt = u_t\left(\hat{\mu}dt + \sigma dB_t\right). \tag{2.2}$$

其中 $\hat{\mu} = \rho\sigma + \gamma g$，$w_0^u = w_0 \in \mathbb{R}$，$\gamma g dt$ 为投资组合的 ESG 价值，如果投资者对 ESG 因素非常关注且风险资产的 ESG 得分较高的同时，那么 $\gamma g dt$ 将是一个正值，对股票价格的影响是正向的，这意味着当投资者更关注 ESG 因素并且风险资产的 ESG 得分较高时，股票价格有可能上涨，当股票有良好的 ESG 表现时，具有 ESG 偏好的投资者就会认为此时的股票更具有价值，从而在投资中获得额外的效用。相应地，当 $g < 0$、$\gamma > 0$ 时，$\gamma g dt < 0$，此时具有正 ESG 偏好的投资者获得的效用就会减少，也就是说当投资者对 ESG 因素的关注较低或者风险资产的 ESG 得分较低时，股票价格可能下跌。股票价格的增幅不仅取决于风险资产的 ESG 得分，还取决于投资者的 ESG 偏好程度。所以 $\gamma g dt$ 可以解释为投资组合的 ESG 价值，其中股票价格与投资者对 ESG 的偏好水平 γ（$\gamma \geq 0$），风险资产的 ESG 得分 $g\left(g \in [-1,1]\right)$ 正相关。

经典的连续时间 MV 模型旨在解决以下约束优化问题：

$$\min_u Var\left[w_T^u\right].$$

$$\text{s.t. } \mathbb{E}\left[w_T^u\right] = z \tag{2.3}$$

其中 $\left\{w_t^u, 0 \leq t \leq T\right\}$ 在投资策略（组合）u 下满足动态方程（2.2），且 $z \in \mathbb{R}$ 是在 $t = 0$ 时刻设置的作为在投资期结束时的期望收益。

由于目标的差异，（2.3）是时间不一致的, 问题就变成了描述性 Descriptive）的而不是规范性（Normative）的，因为对于时间不一致的优化问题通常没有动

态的最优解决方案，而且代理人对同一时间不一致问题的反应是不同的，研究的目标之一就是描述面对这种时间不一致性时的不同行为。在本文中，我们将重点放在所谓的 MV 问题的预承诺策略上，该策略仅在$t=0$时最优。

为了求解（2.3），通过拉格朗日乘子κ将其转化为无约束优化问题：

$$\min_u \mathbb{E}\left[\left(w_T^u\right)^2\right]-z^2-2\kappa\left(\mathbb{E}\left[w_T^u\right]-z\right)=\min_u \mathbb{E}\left[\left(w_T^u-\kappa\right)^2\right]-(\kappa-z)^2. \quad (2.4)$$

这个优化问题可以通过解析求解，其解$\left\{u^*=u_t^*, 0\le t\le T\right\}$依赖于$\kappa$，可以通过原约束条件$\mathbb{E}\left[w_T^u\right]=z$最后确定$\kappa$的值。

2.1.2 经典 MV 问题的解

经典的 MV 问题（2.4），应用了动态规划原理，我们考虑容许控制集$\mathcal{A}^{cl}(s,y):=\left\{u=\{u_t, t\in[s,T]\}:\ \mathbb{E}\left[\int_s^T (u_t)^2 dt<\infty\right]\right\}$，其中$u$是$\mathcal{F}_t$循序可测的。

（最优）值函数定义为：

$$V^{cl}(s,y;\kappa):=\inf_{u\in\mathcal{A}^{cl}(s,y)} \mathbb{E}\left[(w_T^u-\kappa)^2\middle| w_s^u=y\right]-(\kappa-z)^2. \quad (2.5)$$

固定$\kappa\in\mathbb{R}$，$(s,y)\in[0,T]\times\mathbb{R}$，一旦解出方程（2.5），$\kappa$就可以通过约束$\mathbb{E}\left[w_T^*\right]=z$来确定，$\left\{w_T^*, t\in[0,T]\right\}$是在最优投资组合$u^*$下的最优财富过程。其 HJB 方程为：

$$\omega_t(t,w;\kappa)+\min_{u\in\mathbb{R}}\left(\frac{1}{2}(\sigma^2u^2\omega_{ww}(t,w;\kappa)+\rho\sigma u\omega_w(t,w;\kappa)\right)=0.$$

终端条件：

$$\omega_t(T,w;\kappa)=\left(w-\kappa\right)^2-\left(\kappa-z\right)^2.$$

推导出最优值函数为：

$$V^{cl}(t,w;\kappa)=(w-\kappa)^2e^{-\frac{\hat{\mu}^2}{\sigma^2}(T-t)}-(\kappa-z)^2. \quad (2.6)$$

最优反馈策略：

$$u^*(t,w;\kappa)=-\frac{\hat{\mu}}{\sigma^2}(w-\kappa). \quad (2.7)$$

以及相应的最优财富过程是 SDE 唯一的强解：

$$dw_t^*=-\frac{\hat{\mu}^2}{\sigma^2}(w_t^*-\kappa)dt+\sqrt{\frac{\hat{\mu}^2}{\sigma^2}(w_t^*-\kappa)^2}dB_t, w_0^*=w_0. \quad (2.8)$$

在约束条件$\mathbb{E}\left[W_t^*\right]=z$下可以得到拉格朗日乘子$\kappa=\frac{ze^{\frac{\hat{\mu}^2}{\sigma^2}T}-w_0}{e^{\frac{\hat{\mu}^2}{\sigma^2}T}-1}$。

2.2 探索性 MV 问题

给定模型参数后，经典的基于模型的 MV 问题（2.3）及其许多变体已经相当完全地解决。在实现这些解决方案时，需要从资产价格的历史时间序列中估计市场参数，这一过程在经典自适应控制中称为辨识。然而，在实践中很难估计投资机会数，尤其是具有容许精度的平均收益。此外，经典的最优 MV 策略通常对这些参数非常敏感，这主要是由于对病态方差-协方差矩阵求逆以获得最优分配权重的过程。鉴于这两个问题，马科维茨解决方案可能与潜在的投资目标不相关。

另一方面，强化学习技术并不需要对模型参数做任何估计，因为 RL 算法由历史数据驱动，直接输出最优（或接近最优）的分配。可以通过与未知投资环境的直接互动，以学习（探索）和优化（利用）的方式实现。Wang 等[17]在无限时域的背景下提出了探索性 RL 随机控制问题的一般理论框架，并对 LQ 这种特殊情况进行了详细研究。我们在这里采用了相同的框架，注意到 LQ 结构的固有特征和 MV 问题的有限时间范围。事实上，尽管探索性形式的动机基本相同，但随着从无限时域到有限时域的转变，产生了一些新的见解。

2.2.1 EMV 问题

首先，我们介绍状态动力学（2.2）的“探索性”版本。动机是强化学习中的重复学习，在这个方程中，控制（投资组合）过程$u=\{u_t,0\le t\le T\}$是随机化的，就产生了分布控制过程，其密度函数用$\pi=\{\pi_t,0\le t\le T\}$表示。动态方程（2.2）式改为：

$$dW_t^{\pi}=\tilde{b}(\pi_t)dt+\tilde{\sigma}(\pi_t)dB_t. \tag{2.9}$$

其中$0\le t\le T$, $w_0^u=w_0$。

$$\tilde{b}(\pi):=\int_{\mathbb{R}}(\rho\sigma+\gamma g)u\pi(u)du,\pi\in\mathcal{P}\left(\mathbb{R}\right). \tag{2.10}$$

$$\tilde{\sigma}(\pi) \coloneqq \sqrt{\int_{\mathbb{R}} \sigma^2 u^2 \pi(u) du}, \pi \in \mathcal{P}(\mathbb{R}). \tag{2.11}$$

$\mathcal{P}(\mathbb{R})$是$\mathbb{R}$上关于勒贝格测度绝对连续的概率密度函数的集合，μ_t和σ_t^2，$0 \le t \le T$表示与分布控制过程π相关的均值和方差过程（假设存在），即

$$\mu_t \coloneqq \int_{\mathbb{R}} u \pi_t(u) du \text{ 和 } \sigma_t^2 \coloneqq \int_{\mathbb{R}} u^2 \pi_t(u) du - \mu_t^2. \tag{2.12}$$

然后，将（2.8）~（2.10）代入到（2.7）式中得到：

$$dW_t^{\pi} = \hat{\mu} \mu_t dt + \sigma \sqrt{\left(\sigma_t^2 + \mu_t^2\right)} dB_t. \tag{2.13}$$

（2.13）式中$0 \le t \le T$，并且$W_0^{\pi} = w_0$。随机分布控制过程$\pi = \{\pi_t, 0 \le t \le T\}$用于模型探索，其总体水平由其累积微熵刻画：

$$\mathcal{H}(\pi) \coloneqq -\int_0^T \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du dt.$$

进一步，引入一个温度参数（即探索权值）$\lambda > 0$来反映利用和探索之间的权衡，接下来求解熵正则化的 EMV 问题。对于任意固定的$\kappa \in \mathbb{R}$：

$$\min_{\pi \in \mathcal{A}(0, w_0)} \mathbb{E}\left[(W_T^{\pi} - \kappa)^2 + \lambda \int_0^T \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du dt \right] - (\kappa - z)^2. \tag{2.14}$$

因为代理人寻求优化 MV 目标的同时保持规定的整体探索水平。因此（2.12）式中的累积微熵项可以被解释为对没有进行足够探索的惩罚。其中$\mathcal{A}(0, x_0)$是$[0,T]$上可允许的分布控制集，将在下面精确定义。在得到最小化问题解$\pi^* = \{\pi_t^*, 0 \le t \le T\}$之后，拉格朗日乘子$\kappa$就可以通过约束 $\mathbb{E}\left[W_T^{\pi^*}\right] = z$来确定。

优化目标（2.12）式明确鼓励探索，而经典问题（2.4）式只关注利用。我们将通过动态规划原理求解（2.10）。为此，我们需要定义值函数。

命题 1 对于$(s, y) \in [0,T) \times \mathbb{R}$，考虑状态方程（2.9）式中在$[s,T]$上$W_s^{\pi} = y$，定义一族可允许的控制集合$\mathcal{A}(s, y)$如下所示。设$\mathcal{B}(\mathbb{R})$是在 R 上的 Borel 代数。控制分布（或投资组合/策略）过程$\pi = \{\pi_t, s \le t \le T\}$属于$\mathcal{A}(s, y)$，如果满足下列条件：

(i) 对所有的$s \le t \le T, \pi_t \in \mathcal{P}(\mathbb{R})$ a.s.；

(ii) 对所有的$A \in \mathcal{B}(\mathbb{R}), \left\{\int_A \pi_t(u) du, s \le t \le T\right\}$是$\mathcal{F}_t$循序可测的；

(iii) $\mathbb{E}\left[\int_s^T \mu_t^2 + \sigma_t^2 dt\right] < \infty$；

(iv) $\mathbb{E}\left[\left|(W_T^{\pi} - \kappa)^2 + \lambda \int_0^T \int_{\mathbb{R}} \pi_t(u) \ln \pi_t(u) du dt \middle| W_s^{\pi} = y\right|\right] < \infty$。

显然，由条件（iii）可知，随机微分方程（2.9）对于$s \le t \le T$有一个唯一的强解，且满足$W_s^{\pi} = y$。$\mathcal{A}(s, y)$中的控制是可测（准确地说，是密度函数值）随

机过程，在控制术语中也称为开环控制。在经典控制理论中，区分开环控制和反馈（或闭环）控制是很重要的。

命题 2 一个确定性映射 $\pi(\cdot;\cdot;\cdot)$ 被称为容许的反馈控制，如果：

a) $\pi(\cdot;t;x)$ 是一个密度函数，对于 $(t,x)\in[0,T]\times\mathbb{R}$；

b) 对于 $(s,y)\in[0,T]\times\mathbb{R}$，下列随机微分方程（即将反馈策略 $\pi(\cdot;\cdot;\cdot)$ 代入后的系统动力学方程）为：

$$dW_t^\pi=\tilde{b}(\pi(\cdot;t,W_t^\pi))dt+\tilde{\sigma}(\pi(\cdot;t,W_t^\pi))dB_t,t\in[s,T];W_s^\pi=y.$$

该随机微分方程有一个唯一的强解 $\{W_t^\pi,t\in[s,T]\}$，并且开环控制 $\pi=\{\pi_t,t\in[s,T]\}\in\mathcal{A}(s,y)$ 其中 $\pi_t:=\pi(\cdot;t,W_t^\pi)$，在这种情形下，开环控制 π 可以说是由反馈控制 $\pi(\cdot;\cdot;\cdot)$ 相对于初始时间和状态 (s,y) 生成的。值得注意的是，开环控制及其可容许性依赖于初始值 (s,y)，而反馈控制可以对任意 $(s,y)\in[0,T]\times\mathbb{R}$ 生成开环控制，因此它本身独立于 (s,y)。

现在固定 $\kappa\in\mathbb{R}$，定义值函数：

$$V(s,y;\kappa):=\inf_{\pi\in\mathcal{A}(s,y)}\mathbb{E}\left[(W_T^\pi-\kappa)^2+\lambda\int_0^T\int_{\mathbb{R}}\pi_t(u)\ln\pi_t(u)dudt\middle|W_s^\pi=y\right]-(\kappa-z)^2.$$

对于 $(s,y)\in[0,T]\times\mathbb{R}$，值函数 $V(\cdot,\cdot;\kappa)$ 被认为是该问题的最优值函数，下面定义在给定的反馈控制 π 下的值函数为：

$$V^\pi(s,y;\kappa)=\mathbb{E}\left[(W_T^\pi-\kappa)^2+\lambda\int_s^T\int_{\mathbb{R}}\pi_t(u)\ln\pi_t(u)dudt\middle|W_s^\pi=y\right]-(\kappa-z)^2.$$

对于 $(s,y)\in[0,T]\times\mathbb{R}$，$\pi=\{\pi_t,0\le t\le T\}$ 是由 π 生成的关于 (s,y) 的开环控制，$\{W_t^\pi,t\in[s,T]\}$ 是相对应的财富过程。

为了解决 EMV 问题（2.10），我们应用经典的 Bellman 最优性原理可以得到：

$$V(t,w;\kappa)=\inf_{\pi\in\mathcal{A}(t,w)}\mathbb{E}\left[V(s,W_s^\pi;w)+\lambda\int_t^s\int_{\mathbb{R}}\pi_v(u)\ln\pi_v(u)dudv\middle|W_t^\pi=w\right].$$

利用随机最优控制理论，值函数 $V\ (0\le t\le s\le T,w\in\mathbb{R})$ 满足哈密尔顿-雅可比-贝尔曼（HJB）方程：

$$v_t(t,w;\kappa)+\min_{\pi\in\mathcal{P}(\mathbb{R})}\begin{pmatrix}\frac{1}{2}\tilde{\sigma}^2(\pi)v_{ww}(t,w;\kappa)+\tilde{b}(\pi)v_w(t,w;\kappa)\\+\lambda\int_{\mathbb{R}}\pi(u)\ln\pi(u)du\end{pmatrix}=0. \qquad (2.15)$$

将 $\tilde{b}(\pi):=\int_{\mathbb{R}}\hat{\mu}u\pi(u)du$ 和 $\tilde{\sigma}(\pi):=\sqrt{\int_{\mathbb{R}}\sigma^2u^2\pi(u)du}$ 代入（2.12）式得到：

$$v_t(t,w;\kappa)+\min_{\pi\in\mathcal{P}(\mathbb{R})}\int_{\mathbb{R}}\left(\begin{array}{l}\frac{1}{2}(\sigma^2u^2v_{ww}(t,w;\kappa)+\hat{\mu}uv_w(t,w;\kappa)\\+\lambda\int_{\mathbb{R}}\ln\pi(u)\end{array}\right)\pi(u)du=0. \quad (2.16)$$

终端条件 $v(T,w;\kappa)=(w-\kappa)^2-(\kappa-z)^2$，这里，$v$表示 HJB 方程的一般未知解，利用 $\pi\in\mathcal{P}(\mathbb{R})$ 当且仅当：

$$\int_{\mathbb{R}}\pi(u)du=1,\pi(u)\geq 0 \text{ a.e.}$$

求解 HJB 方程（2.12）中的优化问题，我们可以得到一个反馈（分布）控制，其密度函数由下式给出：

$$\begin{aligned}\pi^*(u;t,w,\kappa)&=\frac{\exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}(u^2\sigma^2v_{ww}(t,w;\kappa)+\hat{\mu}uv_w(t,w;\kappa)\right)\right)}{\int_{\mathbb{R}}\exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}(u^2\sigma^2v_{ww}(t,w;\kappa)+\hat{\mu}uv_w(t,w;\kappa)\right)\right)du}. \quad (2.17)\\&=\mathcal{N}\left(u\middle|-\frac{\hat{\mu}}{\sigma^2}\frac{v_w(t,w;\kappa)}{v_{ww}(t,w;\kappa)},\frac{\lambda}{\sigma^2v_{ww}(t,w;\kappa)}\right)\end{aligned}$$

其中我们用 $\mathcal{N}\left(u\middle|\theta,\varphi\right)$表示高斯密度函数，其均值为$\theta\in\mathbb{R}$，方差$\varphi>0$。在上述表示中，我们假设$v_{ww}(t,w;\kappa)>0$，下面将对此进行验证。

将候选的最优高斯反馈控制$\pi^*(u;t,w,\kappa)$代回 HJB 方程（2.12）式化为：

$$v_t(t,w;\kappa)-\frac{\hat{\mu}^2}{2\sigma^2}\frac{v_{ww}^2(t,w;\kappa)}{v_{ww}(t,w;\kappa)}+\frac{\lambda}{2}\left(1-\ln\frac{2\pi e\lambda}{\sigma^2v_{ww}(t,w;\kappa)}\right)=0. \quad (2.18)$$

终端条件 $v(T,w;\kappa)=\left(w-\kappa\right)^2-\left(\kappa-z\right)^2$，通过计算可以直接得到这个方程有一个经典的解：

$$\begin{aligned}v(t,w;\kappa)&=\left(w-\kappa\right)^2e^{-\frac{\hat{\mu}^2}{\sigma^2}(T-t)}-\frac{\lambda}{2}(T-t)\left(\frac{\hat{\mu}^2}{\sigma^2}T+\ln\frac{\pi\lambda}{\sigma^2}\right). \quad (2.19)\\&+\frac{\lambda}{4}\frac{\hat{\mu}^2}{\sigma^2}(T^2-t^2)-(\kappa-z)^2\end{aligned}$$

对于任意的$(t,w)\in[0,T]\times\mathbb{R}$，显然满足$v_{ww}(t,w;\kappa)>0$。因此，候选最优反馈控制（2.13）式可以化为：

$$\pi^*(u;t,w,\kappa)=\mathcal{N}\left(u\middle|-\frac{\hat{\mu}}{\sigma^2}(w-\kappa),\frac{\lambda}{2\sigma^2}e^{\frac{\hat{\mu}^2}{\sigma^2}(T-t)}\right),(t,w)\in[0,T]\times\mathbb{R}. \quad (2.20)$$

最后，最优财富过程（2.9）式在π^*条件下变为

$$dW_t^* = -\frac{\hat{\mu}^2}{\sigma^2}(W_t^* - \kappa)dt + \sqrt{\frac{\hat{\mu}^2}{\sigma^2}(W_t^* - \kappa) + \frac{\lambda}{2}e^{\frac{\hat{\mu}^2}{\sigma^2}(T-t)}}\,dB_t, W_0^* = w_0. \qquad (2.21)$$

容易验证（2.17）式在0≤t≤T上有一个唯一的强解，现在我们总结以上结果为定理1。

定理 1 熵正则化 EMV 问题（2.10）的最优值函数为：

$$V(t,w;\kappa) = (w-\kappa)^2 e^{-\frac{\hat{\mu}^2}{\sigma^2}(T-t)} - \frac{\lambda}{2}(T-t)\left(\frac{\hat{\mu}^2}{\sigma^2}T + \ln\frac{\pi\lambda}{\sigma^2}\right) + \frac{\lambda}{4}\frac{\hat{\mu}^2}{\sigma^2}(T^2-t^2) - (\kappa-z)^2.$$

满足$(t,w)\in[0,T]\times\mathbb{R}$，其最优反馈控制是高斯的，由下式给出：

$$\pi^*(u;t,w,\kappa) = \mathcal{N}\left(u\middle|-\frac{\hat{\mu}}{\sigma^2}(w-\kappa), \frac{\lambda}{2\sigma^2}e^{\frac{\hat{\mu}^2}{\sigma^2}(T-t)}\right), (t,w)\in[0,T]\times\mathbb{R}.$$

在π^*对应的最优财富过程为：

$$dW_t^* = -\frac{\hat{\mu}^2}{\sigma^2}(W_t^* - \kappa)dt + \sqrt{\frac{\hat{\mu}^2}{\sigma^2}(W_t^* - \kappa) + \frac{\lambda}{2}e^{\frac{\hat{\mu}^2}{\sigma^2}(T-t)}}\,dB_t, W_0^* = w_0.$$

最后通过初始约束条件得到拉格朗日乘子$\kappa = \dfrac{ze^{\frac{\hat{\mu}}{\sigma^2}T} - w_0}{e^{\frac{\hat{\mu}}{\sigma^2}T} - 1}$。

拉格朗日乘子κ的计算见附录1。现在讨论定理1的经济学解释，通过引入ESG强度参数和ESG评分这两个外生变量，有助于引导资本流向那些在环境、社会和治理方面表现良好的公司，促进可持续发展。最优财富过程（2.17）式，ESG评分g与ESG强度参数γ相乘，共同影响调整后的预期超额收益率$\hat{\mu}$，因此，具有一定ESG偏好的投资者在选择ESG评分较高的公司可能会获得更高的预期超额收益率，从而吸引更多对ESG因素感兴趣的投资者。而且ESG表现良好的公司面临更低的市场风险和不确定性，即ESG因素通过影响市场预期和投资者行为来间接影响市场噪声。该模型允许投资者在不同时间点根据财富变化和ESG的市场情况动态调整投资策略，以适应不断变化的市场环境。

这个结果中有几个地方值得注意。经典问题和EMV问题具有相同的拉格朗日乘子值，这是由于两个问题在各自的最优反馈控制下的最优终端财富具有相同的均值。第二，最优高斯策略的方差（用于度量探索水平）在时刻t上为$\dfrac{\lambda}{2\sigma^2}e^{\frac{\hat{\mu}^2}{\sigma^2}(T-t)}$，因此，探索随着时间的推移而衰减，也就是说代理人最初以最大水平进行探索，并随着时间的推移和接近投资期限的结束而逐渐减少探索（尽管永

远不会减少到零）。因此，与 Wang 等[17]研究的无限时域不同，探索的程度不再是恒定的，而是退火的。因为随着时间的推移，RL 代理（RL agent）对随机环境的了解越来越多，利用变得越来越重要。

第三，在任意给定$t\in[0,T]$时，在其他参数固定不变的情况下，探索性高斯分布的方差随着风险资产波动性的增加而减小，风险资产的波动性反映了投资领域的随机性水平。这意味着一个更随机的环境包含了更多的学习机会，RL 代理可以利用这些机会来减少她自己的探索努力，毕竟探索成本是高昂的。

最后，高斯分布的均值与探索权重λ无关，而其方差与状态变量w无关。这突出了利用和探索之间的完美分离，因为前者由均值刻画，后者由最优高斯探索的方差刻画。

2.2.2 策略改进定理

在前两节奠定了理论基础之后，我们现在设计一个 RL 算法来学习熵正则化 MV 问题的解并输出可实现的投资组合分配策略，而不需要假设任何关于基本参数的知识。为此，我们首先建立一个所谓的策略改进定理（policy improvement theorem，PIT）以及相应的收敛结果。同时，我们将提供一种自校正方案来学习真正的拉格朗日乘子κ。我们的 RL 算法绕过了估计任何模型参数的阶段，包括平均返回向量和方差-协方差矩阵。它还避免了在高维中对典型的病态方差-协方差矩阵求逆，因为这可能会产生非稳健的投资组合策略。下面的结果为我们的 EMV 投资组合选择问题提供了一个 PIT。

定理 2　$\kappa\in\mathbb{R}$，$\pi=\pi(\cdot;\cdot,\kappa)$是任意给定容许的反馈控制，假设相应的值函数$V^{\pi}(\cdot,\cdot;\kappa)\in C^{1,2}\left([0,T]\times\mathbb{R}\right)\cap C^{0}\left([0,T]\times\mathbb{R}\right)$满足对任意的$(t,w)\in[0,T]\times\mathbb{R}$都有$V^{\pi}_{ww}(t,w;\kappa)>0$，进一步假设反馈控制$\tilde{\pi}$：

$$\tilde{\pi}(u;t,w,\kappa)=\mathcal{N}\left(u\middle|-\frac{\hat{\mu}}{\sigma^{2}}\frac{v_{w}(t,w;\kappa)}{v_{ww}(t,w;\kappa)},\frac{\lambda}{\sigma^{2}v_{ww}(t,w;\kappa)}\right). \tag{2.22}$$

那么

$$V^{\tilde{\pi}}(t,w;\kappa)\le V^{\pi}(t,w;\kappa),(t,w)\in[0,T]\times\mathbb{R}.$$

证明见附录 1。

上述定理表明，在高斯族中总有策略可以改善任何给定策略（不一定是高斯策略）的值函数。因此，在不失一般性的情况下，我们可以在选择初始解时简单

地关注高斯策略。最优高斯策略（2.16）表明，候选初始反馈策略可能采用如下形式：

$$\pi_0(u;t,w,\kappa)=\mathcal{N}\left(u\middle|a(w-\kappa),c_1e^{c_2(T-t)}\right).$$

这样的选择价值函数和策略可能在有限次迭代中都收敛，下面给出定理。

定理 3 $\pi_0(u;t,w,\kappa)=\mathcal{N}\left(u\middle|a(w-\kappa),c_1e^{c_2(T-t)}\right)$满足$a,c_2\in\mathbb{R}$，$c_1>0$，由策略改进定理中（2.22）式给定的反馈策略序列为$\left\{\pi_n(u;t,w,\kappa),(t,w)\in[0,T]\times\mathbb{R},n\geq1\right\}$，对应的值函数序列定义为$\left\{V^{\pi_n}(t,w;\kappa),(t,w)\in[0,T]\times\mathbb{R},n\geq1\right\}$，则：

$$\lim_{n\to\infty}\pi_n(\cdot;t,w,\kappa)=\pi^*(\cdot;t,w,\kappa)$$

$$\lim_{n\to\infty}V^{\pi_n}(t,w;\kappa)=V(t,w;\kappa).$$

π^*和$V\ (t,w;\kappa)\in[0,T]\times\mathbb{R}\times\mathbb{R}$分别是对应的最优高斯策略（2.16）和最优值函数（2.15）。证明见附录 1。

上述收敛结果表明，如果我们明智地选择初始策略，学习方案在有限次（实际上是两次）迭代后理论上会收敛。当然，在实践中实现此方案时，每个策略的值函数只能近似，因此，学习过程通常需要更多的迭代才能收敛。定理 3 为更新当前策略提供了理论基础，而定理 2 则为策略空间提供了一个良好的开端。

2.3 数值模拟和经济学解释

为了验证前面得到的结论，我们在本节中进行模拟检验，来验证 EMV 算法的性能。在数值模拟中，投资者 ESG 偏好 分别取 0、0.04、0.08 三个数值是为了能够覆盖无偏好、弱偏好、强偏好三种情形。

2.3.1 参数设置

首先在一个稳定的市场环境中进行数值模拟，其中价格过程是根据几何布朗运动（2.1）式来模拟的，其中μ和σ是常数，考虑终端财富年化目标收益率为 1.10%的 MV 问题，常数化初始财富$w_0=1$，即投资周期结束时的期望平均收益$z=1.1$，利用 Python 进行数值模拟，分别讨论在不同情形下终端财富期望 的变化情况。具体参数的符号和含义详见附录 2 中的表 2-1。

2.3.2 数值模拟

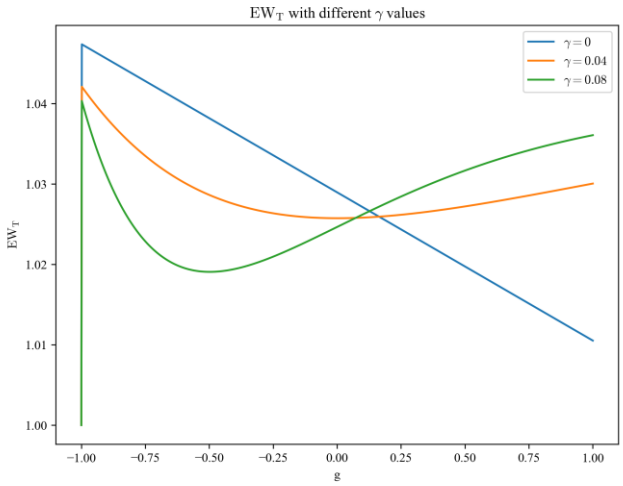
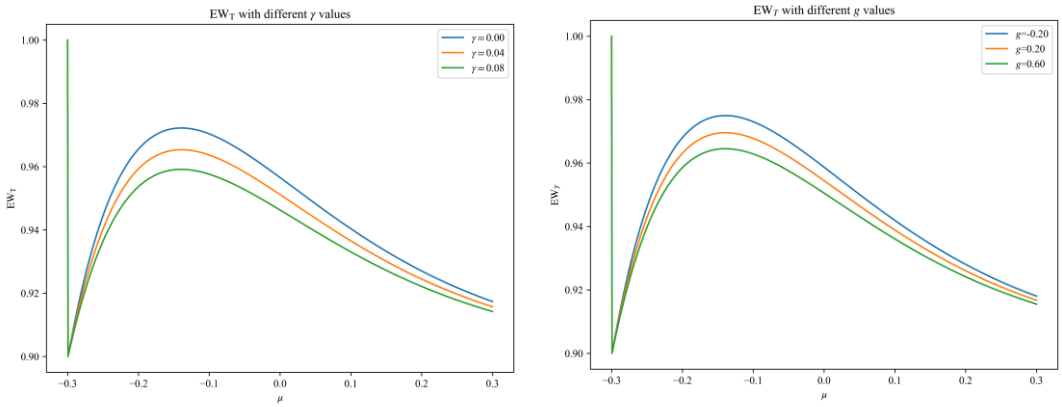


图 2-1 不同的 ESG 偏好水平 γ 以及 ESG 得分 g 对应的终端财富期望

图 2-1 展示了在不同的 ESG 强度 γ 下，终端财富的数学期望 $\mathbb{E}[W_T]$ 随着 g 的变化情况，当 γ 为 0 时，即投资者忽略 ESG 因素，曲线直线下降，这是因为忽略 ESG 因素使投资者错过了与气候变化相关的投资机会，例如投资于可再生能源或环保技术公司，这些领域可能在未来取得较高的增长和回报。当 γ 增加时，即 $\gamma=0.04$、$\gamma=0.08$，$\mathbb{E}[W_T]$ 随着 g 的增大先下降后上升，因为当具有一定 ESG 偏好的投资者在进行投资组合时会注重更环保的投资组合，因为这些资产具有长期可持续性，风险资产的 ESG 得分的提升减少了风险的冲击，从而带来更高的收益。在 $\gamma=0.04$ 时曲线上升和下降的趋势相较 $\gamma=0.08$ 而言都较为缓慢，随着股票 ESG 得分 g 增加的情况下，相对于传统投资策略而言，高 ESG 评级的投资组合在投资回报表现更好。在这一特定时期，ESG 标准作为一种异常稳健的投资方法出现。他们将近年来绿色资产的高回报归因于对这些资产的需求增加，这反映了投资者对环境的关注日益增加，投资者对绿色资产的兴趣也越来越浓厚，反过来也会进一步提高企业的经济回报。从另一个角度来看，当 g 进一步增加时。曲线增长的越来越缓慢，因为在投资者高度重视 ESG 标准的情况下，此时他们的投资选择会受到一定的限制，此时他们更难找到既符合高 ESG 标准又具有相对较高回报的资产。



(a)　　（b）

图 2-2 几种 ESG 偏好水平和 ESG 得分下不同风险预期收益率对应的终端财富期望

图 2-2 展示了终端财富的期望随着风险资产的预期收益率变化而变化的情况，其中 $\mu \in [-0.3, 0.3]$，其他基本参数遵循表 2-1，从图中可以看出 $\mathbb{E}[W_T]$ 呈现出先上升再下降的趋势。图 2（a）展示了不同的 ESG 偏好下投资者终端财富的变化情况，首先，当投资者对 ESG 因素不感兴趣时，曲线波动是最大的，这意味着这类投资者更多地受到市场的影响，而不是基于 ESG 因素的长期投资考量，此时投资者的决策完全基于风险资产的预期收益率，而不受 ESG 标准的影响。而随着 γ 的增大，曲线峰值变的越来越低，投资者在进行投资组合时，他们开始对预期收益率和符合 ESG 标准的资产之间进行权衡。导致他们降低对预期收益率较高但不符合 ESG 标准的资产的兴趣，这时他们的投资会受到他们 ESG 偏好的限制，认为这些资产在长期内会带来更好的结果，即使短期内可能会面临损失，因为他们愿意为绿色发展牺牲部分经济利益，承担社会责任，因为他们不仅关注经济效益，也包括非经济效益。随着 μ 的增加，三条曲线几乎重合，这表明当风险资产的预期收益率足够高时，投资者更倾向于投资于风险资产此时，投资者一定程度上会忽略 ESG 标准，更多地投资于高风险高回报的资产。图 2（b）展示了不同的 ESG 得分下投资者终端财富的变化情况，首先，当股票的 ESG 得分为负值时（ $g = -0.2$ ），曲线峰值最高，这表明如果投资者投资 ESG 表现不好的资产时会提供更高的股票回报。而股票 ESG 表现良好时，具有一定 ESG 偏好的投资者会更加积极地调整其投资组合以反映这些变化。随着 ESG 得分的增加，投资者会更加确信高 ESG 得分公司的长期价值，从而减少了对 ESG 波动的敏感度，导致波动减小。

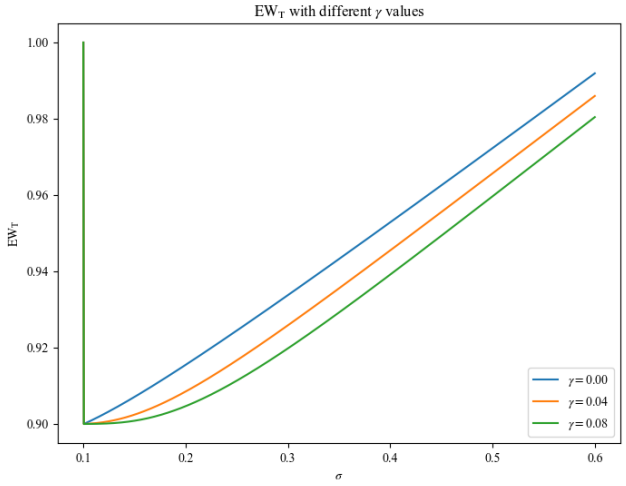


图 2-3 几种 ESG 偏好水平下不同波动率对应的终端财富期望

图 2-3 展示了在不同的投资者 ESG 偏好 γ 下，终端财富的数学期望 $\mathbb{E}W_T$ 随着 σ 变化的情况。首先，当 $\gamma=0$ 时，表示投资者更注重市场波动 σ 对终端财富的影响，他们更愿意承担更高的风险以换取更高的收益。然而当随着 γ 的增大，投资者会更关注 ESG 因素。从图中可以看出，投资者对 ESG 因素的高度重视即 $\gamma=0.08$，曲线在最下方，这是因为他们在进行选择投资组合时受到了一定的限制，比如某些行业或公司可能不符合其 ESG 标准，投资者就会不考虑这些公司，从而降低了投资组合的潜在收益。还有一个原因是由于 ESG 投资通常与更加稳健的投资策略相关联，对 ESG 因素的高度重视会导致投资者选择了较低的风险投资，这也会牺牲一定的潜在回报。这反映了投资者的道德和社会责任感，他们更愿意为了可持续发展和社会责任而放弃一定的收益。

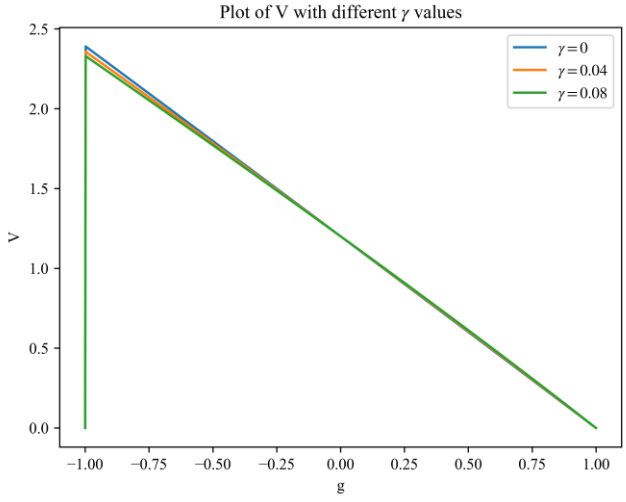


图 2-4 几种 ESG 偏好水平下不同 ESG 得分对应的值函数

图 2-4 展示了 EMV 问题的值函数V 随着g 的变化情况，从图中可以观察到，随着迭代的进行，值函数V 的值越来越小，这反映了投资者对于特定投资组合的期望收益降低的情况，即便投资者对 ESG 因素的重视程度不同，但投资组合的值函数随着g 的增加呈现相似的直线下降，这表明投资者对 ESG 因素的偏好对投资组合的值函数并没有过于直接的影响。即使γ 对于值函数的解释能力较弱，但是投资者 ESG 偏好可以量化投资者面临的权衡，这些投资者愿意牺牲一部分收益来改善其投资组合的 ESG 配置。

2.5 本章小结

在本章中，我们使用探索性随机控制形式并融入 ESG 因素，在强化学习的框架下研究了一个连续时间 MV 投资组合选择问题，在在满足标准自融资的假设下，利用动态规划原理，导出了在满足一定的探索水平下的 HJB 方程，然后得到了一个已证明的 PIT 以及值函数和最优策略的明确函数，将其参数化并采用随机梯度下降算法进行模拟检验。最后通过数值模拟验证了前面得到的结论，并分析了资产 ESG 评级以及投资者 ESG 偏好等对终端财富的影响，并给出了相关结果的经济学含义。

研究结果表明，在考虑 ESG 因素的投资组合中，投资者的 ESG 偏好以及企业 ESG 得分会对最优终端财富产生不可忽视的影响。在同时考虑投资者 ESG 偏好和资产 ESG 评级的可以有效地提高投资组合的期望终端财富。此外，如果投资者对 ESG 因素高度重视，那么他们愿意牺牲一部分的收益来承担社会责任。投资者在面对风险和预期收益时，会考虑到社会责任投资的因素，从而影响他们的投资决策，因为社会责任投资通常关注的不仅仅是经济回报，投资者会更倾向于选择那些在社会责任层面表现良好的企业或行业从而降低整体投资组合的风险水平。同样地可以发现，在随着企业 ESG 得分越来越高的情况下，投资组合的终端财富并不会很明显的增加，因为这类投资者进行投资时，会受到一定范围的限制。最后，在投资组合的资产保持一定的 ESG 评级的情况下，投资者如果具有一定的 ESG 偏好，那么他们将会更加重视长期投资，不会过多的进行投机或短期交易。

第 3 章 通胀对连续时间均值-方差投资组合的影响

3.1 通胀下的基本模型

在这一节，在 RL 背景下，我们考虑连续时间的探索性、熵正则化马科维茨的均值方差（MV）投资组合选择问题，其中 MV 问题是一个特殊的情况。我们首先回顾连续时间中的经典 MV 问题，固定投资规划期$\mathrm{T}>0$，B_t^P，B_t^L是定义在满足常规条件的带流概率空间$\left(\Omega,\mathcal{F},\{\mathcal{F}_\mathrm{t}\}_{0\le \mathrm{t}\le \mathrm{T}},\mathbb{P}\right)$上的标准一维布朗运动，$B_t^P$和$B_t^L$分别为风险股票价格过程和通胀过程的噪声。这里$B_t^P$和$B_t^L$的相关系数$\rho_{pl}$满足：

$$dB_t^P dB_t^L = \rho_{pl} dt .$$

在不考虑 RL 时，风险资产的名义价格过程是一个几何布朗运动：

$$dP_t = P_t(\mu_p dt + \sigma_p dB_t^P), 0 \le \mathrm{t} \le \mathrm{T} .$$

其中$P_0 = p_0 > 0$是 0 时刻的初始价格，μ_p和σ_p为常数，分别为风险资产的预期收益率和波动率。无风险资产的固定利率$r>0$。

消费篮子价格服从几何布朗运动：

$$dL_t = b_l L_t dt + \sigma_l L_t dB_t^L, L_0 = l .$$

b_l，σ_l为常数，分别表示预期通胀率和通胀波动率。

现假设金融市场有实际回报率为r的通胀指数债券，与通胀有着相同风险源，假设通胀指数债券名义价格$\{I_t, 0 \le t \le T\}$满足下列方程：

$$dI_t / I_t = rdt + dL_t / L_t = \left(r + b_l\right)dt + \sigma_l dB_t^L, I_0 = i .$$

由上式可知，通胀指数债券的名义预期收益率是实际收益率与预期通胀率之和。用价格水平L_t将风险资产的名义价值折算成真实的价值，令$\tilde{P}_t = P_t / L_t$表示t时刻投资风险资产的真实价值，利用 Itô 公式，可得$\tilde{P}_t$满足随机微分方程：

$$d\tilde{P}_t = \left(\mu_p - b_l + \sigma_l^2 - \rho_{pl}\sigma_p\sigma_l\right)\tilde{P}_t dt + \sigma_p \tilde{P}_t dB_t^P - \sigma_l \tilde{P}_t dB_t^L .$$

其中$\mu_p - b_l + \sigma_l^2 - \rho_{pl}\sigma_p\sigma_l$表示风险资产真实的预期收益率。

根据 Siu[39]令 $\tilde{I}_t = I_t / L_t$ 表示 t 时刻通胀指数债券的真实价格，可得 $\tilde{I}_t$ 满足 $d\tilde{I}_t = r\tilde{I}_t dt$。

设风险股票上的投资比例是 θ_t，投资组合的名义财富过程为：

$$\begin{aligned} dG_t^N &= G_t^N \left[(1-\theta_t)\frac{dI_t}{I_t} + \theta_t \frac{dP_t}{P_t} \right] \\ &= G_t^N \left((r+b_l)(1-\theta_t) + \mu_p \theta_t \right) dt + \sigma_p \theta_t G_t^N dB_t^P + \sigma_l \left(1-\theta_t\right) G_t^N dB_t^L \end{aligned} \quad ,(3.1)$$

其中初始财富 G_0^N 是已知的。

由式（3.1）表示的投资组合的名义财富过程 G_t^N 经通胀调整后的真实财富过程 G_t 满足如下动态方程

$$dG_t = (rG_t + a_l \theta_t G_t)dt + \theta_t G_t \left(\sigma_p dB_t^P - \sigma_l dB_t^L \right).$$

其中 $a_l = \mu_p - b_l + \sigma_l^2 - \rho_{pl}\sigma_p \sigma_l - r$，表示在考虑通胀对风险资产价格影响的情况下调整后的风险资产的预期收益率。

引入布朗运动 Z_t 满足：

$$dB_t^P := \sqrt{1-\rho_{pl}^2}\, dZ_t + \rho_{pl} dB_t^L, \quad \left\langle dZ_t, dB_t^L \right\rangle = 0.$$

$$\sigma_{ZL}^2 := \sigma_p^2 (1-\rho_{pl}^2) + (\sigma_p \rho_{pl} - \sigma_l)^2.$$

则：

$$\sigma_p dB_t^P - \sigma_l dB_t^L = \sigma_p \sqrt{1-\rho_{pl}}\, dZ_t + (\sigma_p \rho_{pl} - \sigma_l) dB_t^L.$$

$$= \sigma_{ZL} \left(\frac{\sigma_p \sqrt{1-\rho_{pl}}}{\sigma_{ZL}} dZ_t + \frac{\sigma_p \rho_{pl} - \sigma_l}{\sigma_{ZL}} dB_t^L \right) := \sigma_{ZL} B_t^{ZL}.$$

令财富过程 $x_t^u = e^{-rt} G_t$，其中 $u_t = e^{-rt}\zeta_t G_t$ 表示经过通胀调整后投资在风险资产的折现价值。可以得到：

$$dx_t^u = -r x_t^u dt + e^{-rt} dG_t.$$

将 dG_t 代入上式得：

$$\begin{aligned} dx_t^u &= (-r x_t^u + r x_t^u + a_l e^{-rt} \theta_t G_t)dt + \sigma_{ZL} \theta_t G_t e^{-rt} dB_t^{ZL} \\ &= u_t \left(a_l dt + \sigma_{ZL} dB_t^{ZL} \right) \end{aligned} \quad (3.2)$$

其中 $x_0^u = x_0 \in \mathbb{R}$。

考虑通胀因素下，连续时间均值-方差投资组合选择问题是：对于给定真实终端财富的期望 z，寻找最优的可行策略，使得真实终端财富的风险达最小，其中风险用方差来度量。

然后介绍状态动力学的探索版本，在这个方程中，控制（投资组合）过程 $u=\{u_t, 0\le t\le T\}$ 是随机化的，于是就产生了分布控制过程，其密度函数用 $\theta=\{\theta_t, 0\le t\le T\}$ 表示。动态方程（3.2）式改为：

$$dX_t^\theta = \hat{K}(\theta_t)dt + \hat{N}(\theta_t)dB_t^{ZL}. \tag{3.3}$$

其中 $0\le t\le T$，$X_0^\theta = x_0$，

$$\hat{K}(\theta) := \int_{\mathbb{R}} a_l u\theta(u)du, \theta\in\mathcal{P}(\mathbb{R}). \tag{3.4}$$

$$\hat{N}(\theta) := \sqrt{\int_{\mathbb{R}} \sigma_{ZL}^2 u^2\theta(u)du}, \theta\in\mathcal{P}(\mathbb{R}). \tag{3.5}$$

$\mathcal{P}(\mathbb{R})$ 是 $\mathbb{R}$ 上关于勒贝格测度绝对连续的概率密度函数的集合，ζ_t 和 ω_t^2，$0\le t\le T$ 表示与分布控制过程 θ 相关的均值和方差过程（假设存在），即

$$\zeta_t := \int_{\mathbb{R}} u\theta_t(u)du \quad 和 \quad \omega_t^2 := \int_{\mathbb{R}} u^2\theta_t(u)du - \zeta_t^2. \tag{3.6}$$

然后，将（3.4）~（3.6）式代入到（3.3）式中得到下列式子

$$dX_t^\theta = a_l\zeta_t dt + \sigma_{ZL}\sqrt{(\omega_t^2+\zeta_t^2)}dB_t^{ZL}. \tag{3.7}$$

（3.7）式中 $0\le t\le T$ 并且 $X_0^\theta = x_0$。随机分布控制过程 $\theta=\{\theta_t, 0\le t\le T\}$ 用于模型探索，其总体水平由其累积微熵刻画

$$\mathcal{H}(\theta) := -\int_0^T\int_{\mathbb{R}} \theta_t(u)\ln\theta_t(u)dudt.$$

进一步，引入一个探索权值 $\lambda>0$ 来反映利用和探索之间的权衡，接下来求解熵正则化的 EMV 问题。对于任意固定的 $\delta\in\mathbb{R}$：

$$\min_{\theta\in\mathcal{A}(0,x_0)} \mathbb{E}\left[(X_T^\theta-\delta)^2 + \lambda\int_0^T\int_{\mathbb{R}} \theta_t(u)\ln\theta_t(u)dudt\right] - (\delta-z)^2 \tag{3.8}$$

因为代理人寻求优化 MV 目标的同时保持规定的整体探索水平。因此（3.8）式中的累积微熵项可以被解释为没有进行足够探索的惩罚。其中 $\mathcal{A}(0,x_0)$ 是 $[0,T]$ 上可允许的分布控制集，在得到最小化问题解 $\theta^*=\{\theta_t^*, 0\le t\le T\}$ 之后，拉格朗日乘子 δ 就可以通过约束 $\mathbb{E}\left[X_T^{\theta^*}\right]=z$ 来确定。

对于$(s,y)\in[0,T)\times\mathbb{R}$，考虑状态方程（3.7）式中在$[s,T]$上$X_s^\theta=y$，设$\mathcal{B}(\mathbb{R})$是在$\mathbb{R}$上的 Borel 代数。控制分布（或投资组合策略）过程$\theta=\{\theta_t,s\le t\le T\}$属于$\mathcal{A}(s,y)$，随机微分方程（3.7）式对于$s\le t\le T$有一个唯一的强解，且满足$X_s^\theta=y$。$\mathcal{A}(s,y)$中的控制是可测密度函数是可测随机过程，在控制术语中也称为开环控制。

具体地说，一个确定性映射$\theta(\cdot;\cdot,\cdot)$被称为容许的反馈控制，再将反馈策略$\theta(\cdot;\cdot,\cdot)$代入后的系统动力学方程）为：

$$dX_t^\theta=\widehat{K}(\theta(\cdot;t,X_t^\theta))dt+\widehat{N}(\theta(\cdot;t,X_t^\theta))dB_t^{ZL},t\in[s,T];X_s^\theta=y.$$

该随机微分方程有一个唯一的强解$\{X_t^\theta,t\in[s,T]\}$，并且开环控制$\theta=\{\theta_t,t\in[s,T]\}\in\mathcal{A}(s,y)$，其中$\theta_t:=\theta(\cdot;t,X_t^\theta)$，在这种情形下，开环控制$\theta$可以说是由反馈控制$\theta(\cdot;\cdot;\cdot)$相对于初始时间和状态$(s,y)$生成的。值得注意的是，开环控制及其可容许性依赖于初始值(s,y)，而反馈控制可以对任意$(s,y)\in[0,T]\times\mathbb{R}$生成开环控制，因此它本身独立于$(s,y)$。

现在固定$\delta\in\mathbb{R}$，定义值函数：

$$V(s,y;\delta):=\inf_{\theta\in\mathcal{A}(s,y)}\mathbb{E}\left[(X_T^\theta-\kappa)^2+\lambda\int_0^T\int_{\mathbb{R}}\theta_t(u)\ln\theta_t(u)dudt\middle|X_s^\theta=y\right]-(\delta-z)^2.$$

对于$(s,y)\in[0,T]\times\mathbb{R}$，值函数$V(\cdot,\cdot;\delta)$被认为是该问题的最优值函数，下面定义在给定的反馈控制下的值函数为：

$$V^\theta(s,y;\kappa)=\mathbb{E}\left[(X_T^\theta-\delta)^2+\lambda\int_s^T\int_{\mathbb{R}}\theta_t(u)\ln\theta_t(u)dudt\middle|X_s^\theta=y\right]-(\delta-z)^2.$$

对于$(s,y)\in[0,T]\times\mathbb{R}$，$\theta=\{\theta_t,0\le t\le T\}$是由$\theta$生成的关于$(s,y)$的开环控制，$\{X_t^\theta,t\in[s,T]\}$是相对应的财富过程。

为了解决 EMV 问题（3.8）式，我们应用经典的 Bellman 最优性原理可以得到：

$$V(t,x;\delta)=\inf_{\theta\in\mathcal{A}(t,x)}\mathbb{E}\left[V(s,X_s^\theta;\kappa)+\lambda\int_t^s\int_{\mathbb{R}}\theta_v(u)\ln\theta_v(u)dudv\middle|X_t^\theta=x\right].\qquad(3.9)$$

利用随机最优控制理论，值函数V $(0\le t\le s\le T,x\in\mathbb{R})$满足哈密尔顿-雅可比-贝尔曼（HJB）方程，再将（3.4）、（3.5）式代入（3.7）式得到

$$v_t(t,x;\delta)+\min_{\theta\in\mathcal{P}(\mathbb{R})}\int_{\mathbb{R}}\left(\frac{1}{2}(\sigma_{ZL}^2u^2v_{xx}(t,x;\delta)+a_luv_x(t,x;\delta)+\lambda\int_{\mathbb{R}}\ln\theta(u)\right)\theta(u)du=0.(3.9)$$

终端条件 $v(T,x;\kappa)=(x-\kappa)^2-(\kappa-z)^2$，这里，$v$表示 HJB 方程的一般未知解。

求解 HJB 方程（3.10）式中的优化问题，我们可以得到一个反馈（分布）控制，其密度函数由下式给出：

$$\begin{aligned}\theta^*(u;t,x,\delta)&=\frac{\exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}\sigma_{ZL}^2u^2v_{xx}(t,x;\delta)+a_luv_x(t,x;\delta)\right)\right)}{\int_{\mathbb{R}}\exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}\sigma_{ZL}^2u^2v_{xx}(t,x;\delta)+a_luv_x(t,x;\delta)\right)\right)du}\\&=\mathcal{N}\left(u\middle|-\frac{a_l}{\sigma_{ZL}^2}\frac{v_x(t,x;\delta)}{v_{xx}(t,x;\delta)},\frac{\lambda}{\sigma_{ZL}^2v_{xx}(t,x;\delta)}\right)\end{aligned}. \quad (3.10)$$

其中我们用 $\mathcal{N}\left(u|\varsigma,\varphi\right)$表示高斯密度函数，其均值为$\varsigma\in\mathbb{R}$，方差$\varphi>0$。在上述表示中，我们假设 $v_{xx}(t,x;\delta)>0$，下面将对此进行验证。

将候选的最优高斯反馈控制$\theta^*(u;t,x,\delta)$代回 HJB 方程（3.10）式，（3.10）式转化为

$$v_t(t,w;\delta)-\frac{a_l^2}{2\sigma_{ZL}^2}\frac{v_x^2(t,x;\delta)}{v_{xx}(t,x;\delta)}+\frac{\lambda}{2}\left(1-\ln\frac{2\pi e\lambda}{\sigma_{ZL}^2v_{xx}(t,x;\delta)}\right)=0. \quad [1]$$

通过终端条件 $v(T,x;\kappa)=\left(x-\kappa\right)^2-\left(\kappa-z\right)^2$ 计算可以直接得到这个方程有一个经典的解：

$$v(t,x;\kappa)=\left(x-\delta\right)^2e^{-\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}-\frac{\lambda}{2}(T-t)\left(\frac{a_l^2}{\sigma_{ZL}^2}T+\ln\frac{\theta\lambda}{\sigma_{ZL}^2}\right)+\frac{\lambda}{4}\frac{a_l^2}{\sigma_{ZL}^2}(T^2-t^2)-(\delta-z)^2.$$

上式中第一项乘以一个考虑了通胀影响的折现因子$e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}$，通胀影响通过$\frac{a_l^2}{\sigma_{ZL}^2}$的比例来调节。第二项包含了熵正则化的惩罚，权重$\lambda$控制了熵项的影响程度。较大的 λ值将更多地鼓励探索，因为它会增加熵项的奖励，从而促使决策者更频繁地尝试新的决策选择，第三项类似于第二项，它表示了随时间变化的通胀调整影响。上式表明在优化过程中，我们不仅要考虑资产的期望收益，还要考虑熵的贡献以及通胀的影响。

[1] 文中用黑体 $\boldsymbol{\pi}$ 表示圆周率

对于任意的$(t,x)\in[0,T]\times\mathbb{R}$，因为$v_{xx}=2e^{-\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}$，显然大于 0。因此，候选最优反馈控制（3.11）式可以化为：

$$\theta^*(u;t,x,\delta)=\mathcal{N}\left(u\Big|-\frac{a_l}{\sigma_{ZL}^2}(x-\delta),\frac{\lambda}{2\sigma_{ZL}^2}e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}\right),(t,x)\in[0,T]\times\mathbb{R}.$$

最后，最优财富过程（3.9）式在θ^*条件下变为：

$$dX_t^*=-\frac{a_l^2}{\sigma_{ZL}^2}(X_t^*-\delta)dt+\sqrt{\frac{a_l^2}{\sigma_{ZL}^2}(X_t^*-\delta)^2+\frac{\lambda}{2}e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}}dB_t^{ZL},X_0^*=x_0. \quad (3.11)$$

上式中$\sqrt{\frac{a_l^2}{\sigma_{ZL}^2}(X_t^*-\kappa)+\frac{\lambda}{2}e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}}dB_t^{ZL}$是一个随机项，反映了由于通胀率随机变动而引起的财富过程 X_t^* 的波动，其波动受风险股票价格波动风险、通胀波动风险以及他们之间相关系数的影响，并随时间t的增加而衰减。
最后通过初始约束条件得到拉格朗日乘子：

$$\delta=\left(ze^{\frac{a_l^2}{\sigma_{ZL}^2}T}-x_0\right)\left(e^{\frac{a_l^2}{\sigma_{ZL}^2}T}-1\right)^{-1}.$$

3.2 模型的解

在下一节中可以发现经典问题和 EMV 问题在各自的最优反馈控制下的最优终端财富具有相同的均值，从而两者具有同样的拉格朗日乘子值δ，其次，最优高斯策略的方差项$\frac{\lambda}{2\sigma_{ZL}^2}e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}$，作为衡量探索强度的指标，随时间的推移呈现递减趋势，表明代理人的探索行为从初期的全力投入逐渐减弱，直至投资期限尾声，但始终保持着一定的探索水平以避免完全停止，因为随着对随机环境认知的加深，利用（exploitation）的重要性日益凸显。

再者，在任意给定$t\in[0,T]$时，在其他参数固定不变的情况下，探索性高斯分布的方差随着风险资产波动性的增加而减小，风险资产的波动性反映了投资领域的随机性水平。这一现象揭示了投资环境随机性的增加实际上为学习（利用）提供了更多机会，使得 RL 代理能够更有效地利用这些机会，从而减少对高

成本探索的依赖。换句话说，在一个更不稳定的环境中，RL 代理倾向于更加精准地利用已知信息，减少不必要的探索努力。

3.2.1 解的等价性

在本节中，我们建立了经典的和探索性的、熵正则化的 MV 问题之间的可解等价性。请注意，这两个问题可以而且确实已经分别独立地解决了。这里“可解等价”的意思是，一个问题的解将直接导致另一个问题的解，而不需要单独求解。

定理 4 以下两个表述 a）和 b）是等价的。

a) EMV 问题的最优值函数为：

$$v(t,x;\kappa)=\left(x-\delta\right)^2 e^{-\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}-\frac{\lambda}{2}(T-t)\left(\frac{a_l^2}{\sigma_{ZL}^2}T+\ln\frac{\theta\lambda}{\sigma_{ZL}^2}\right)+\frac{\lambda}{4}\frac{a_l^2}{\sigma_{ZL}^2}(T^2-t^2)-(\delta-z)^2.$$

对应的最优反馈控制为：

$$\theta^*(u;t,x,\delta)=\mathcal{N}\left(u\middle|-\frac{a_l}{\sigma_{ZL}^2}(x-\delta),\frac{\lambda}{2\sigma_{ZL}^2}e^{\frac{a_l}{\sigma_{ZL}^2}(T-t)}\right).$$

b)经典 MV 问题的最优值函数为:

$$V^{cl}(t,x;\delta)=(x-\delta)^2 e^{-\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}-(\delta-z)^2.$$

最优反馈策略为:

$$u^*(t,x;\delta)=-\frac{a_l}{\sigma_{ZL}^2}(x-\delta).$$

两个问题具有相同的拉格朗日乘子：

$$\delta=\left(ze^{\frac{a_l^2}{\sigma_{ZL}^2}T}-x_0\right)\left(e^{\frac{a_l^2}{\sigma_{ZL}^2}T}-1\right)^{-1}。$$

现在讨论定理 4 的经济学解释，在通胀背景下，EMV 问题的值函数与反馈策略不仅考虑了经过通胀调整后的预期收益率和波动率的影响，而且考虑了探索成本 λ，这反映了代理人在探索与利用之间的权衡。

而经典MV问题并没有探索成本 λ，其最优值函数和策略更加简化和直接，MV 问题的策略是一个确定性的控制。而 EMV 问题的反馈策略是一个高斯分布。

$\frac{a_l}{\sigma_{ZL}^2}$是一个调节参数，可以称为通胀影响系数。当 $a_l>0$时，如果$x>\delta$，反馈策略的均值$-\frac{a_l}{\sigma_{ZL}^2}(x-\delta)$将下降，这反映了通胀对资产价值的负面影响。

反馈策略的方差部分是一个依赖时间的函数，在通胀背景下探索成本的调整，即为了获得新信息，投资者愿意支付的额外成本，而且在通胀情况下，这一成本可能会增加，因为通胀可能增加了市场的不确定性和波动性，使得获得新信息变得更加困难和昂贵。高通胀环境下的投资者可能需要更谨慎地进行信息获取成本和风险管理之间的权衡，来最大化投资组合的效益和对冲通胀带来的风险。陈述 a）和 b）它们都根据经通胀调整后的收益率a_l和波动率σ_{ZL}^2影响进行投资决策，也就是说通胀对资产价格和投资策略有长期影响。在进行投资组合时，投资者能够更好地应对通胀对资产价格和投资策略的影响。而且当探索权值减小到 0 时，可以合理地期望探索问题收敛于经典问题，下面的定理可以使结果更加精确。

定理 5 陈述 a）或 b）成立时， 对于$(t,x;\delta)\in[0,T]\times\mathbb{R}\times\mathbb{R}$，可以得到：

$$\lim_{\lambda\to 0}\theta^*(\cdot;t,x,\delta)=\zeta_{u^*(t,w;\delta)}(\cdot)\qquad weakly$$
$$\lim_{\lambda\to 0}\left|V(t,x;\delta)-V^{cl}(t,x;\delta)\right|=0$$

证明见附录 1。

对于给出的极限结果，我们可以从经济学的角度进行如下解释：首先，考虑第一个极限：当强化学习的探索强度λ趋于 0 时，最优反馈控制 弱收敛于一个以u^*为中心的狄利克雷分布。在这里，它意味着当没有探索（即$\lambda=0$）时，最优策略将确定地收敛于经典 MV 问题的最优策略u^*。这反映了在完全确定的环境下（即没有探索需求时），投资者的最优策略是明确且唯一的，与经典 MV 问题的解一致。探索强度λ的引入增加了策略的不确定性，允许投资者在追求最优解的同时进行一定的探索。然而，当探索成本趋于 0 时，这种不确定性也随之消失，最优策略回归到经典 MV 问题的解。

接下来，考虑第二个极限，当λ趋于 0 时，EMV 问题的最优值函数V与经典 MV 问题的最优值函数V^{cl}之间的差异趋于 0。这意味着在没有探索成本的情况下，EMV 问题的最优解与经典 MV 问题的最优解在数值上是相等的。这进一

步证实了当探索不是必需时（即$\lambda=0$），EMV 问题退化为经典 MV 问题。两者在最优解和最优值函数上是一致的，表明在确定性环境下，引入探索机制并不会改变问题的本质解。然而，在实际情况中，探索强度λ通常不为 0，因为它反映了投资者在面对不确定性时的行为特征。

在确定性环境下（即$\lambda=0$），两者在最优解和最优值函数上是一致的；而在不确定性环境下（即$\lambda>0$），EMV 问题通过引入探索机制来允许投资者在追求最优解的同时进行一定的策略探索。

定理 6 最后，我们通过考察探索成本来结束本节，由于在目标（3.10）中明确包含了探索，与 EMV 问题相关的成本定义为：

$$C^{u^*,\theta^*}(0,x_0;\delta):=\begin{pmatrix}V(0,x_0;\delta)-\\ \lambda\mathbb{E}\left[\int_0^T\int_{\mathbb{R}}\theta_t^*(u)\ln\theta_t^*(u)dudt\,\middle|\,X_0^{\theta^*}=x_0\right]\end{pmatrix}-V^{cl}(0,x_0;\delta).$$

对于$x_0\in\mathbb{R}$，最优策略（开环控制）$\theta^*=\left\{\theta_t^*,0\le t\le T\right\}$可以说是由反馈控制$\theta^*$相对于初始条件$X_0^{\theta^*}=x_0$生成的。这个成本就是在两个最优值函数之间的差值，调整了由于最优探索策略的熵值而产生的额外成本，它衡量的是由于探索而导致的原始（即非探索性）目标的损失。MV 问题的探索成本为：

$$C^{u^*,\theta^*}(0,x_0;\delta)=\frac{\lambda T}{2},x_0\in\mathbb{R}\ ,\delta\in\mathbb{R}\ .$$

证明见附录 1。

探索成本仅取决于两个“特定于代理人”的参数，即探索权重$\lambda>0$和投资范围$T>0$，后者也是探索范围。结果直接地表明：探索成本随探索权值和探索层数的增加而增加。事实上，这两个属性（λ和T）的相关性是线性的。值得注意的是，探索成本与拉格朗日乘数δ无关。这表明，如果代理人更激进，也就是说风险偏好更强时，探索成本并不会增加，这是因为预期目标z 或等价的拉格朗日乘数δ反映了代理人的风险偏好。

3.2.2 改进的 EMV 算法

在本节中，为了更好地说明通胀对于最优投资策略以及最优财富的影响，研究两种不同情形下（即 EMV 和 MV）通胀风险因子对最优投资组合的影响。提出经过通胀调整后的最优投资策略和财富过程的 RL 算法，即 EMV 算法来解决

问题（3.10）式。它由三个并行进行的过程组成：策略评估，策略改进和基于随机逼近的拉格朗日乘子δ的学习。对于策略评估，在任意给定的可容许反馈策略θ下学习值函数V。根据贝尔曼一致性原理：

$$V^{\theta}(t,x)=\mathbb{E}\left[V^{\theta}(s,X_s)+\lambda\int_t^s\int_{\mathbb{R}}\theta_v(u)\ln\theta_v(u)dudv\middle|X_t=x\right],s\in[t,T].$$

重新考虑这个方程，两边除以$s-t$，我们得到左边取$s\to t$会产生连续时间Bellman误差也就是时序差分（Temporal Difference，TD）误差。

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

$\frac{V^{\theta}(t+\Delta t,X_{t+\Delta t})-V^{\theta}(t,X_t)}{\Delta t}=\dot{V}_t^{\theta}$为导数，$\Delta t(\Delta t\to 0)$为学习算法的离散化步长。

策略评估程序的目标是尽量减少Bellman误差：

$$\eta_t:=\dot{V}_t^{\theta}+\lambda\int_{\mathbb{R}}\theta(u)\ln\theta(u)du.$$

首先最小化下列式子：

$$C(\alpha,\beta)=\frac{1}{2}\mathbb{E}\left[\int_0^T|\eta_t|^2\,dt\right]=\frac{1}{2}\mathbb{E}\left[\int_0^T|\dot{V}_t^{\alpha}+\lambda\int_{\mathbb{R}}\theta_t^{\beta}(u)\ln\theta_t^{\beta}(u)du|^2\,dt\right]. \quad (3.12)$$

$\theta^{\beta}=\left\{\theta_t^{\beta},t\in[0,T]\right\}$是在$t=0$时刻的初始状态 $X_0=x_0$的策略。为了在一个可实现的算法中近似$C(\alpha,\beta)$，我们首先将$[0,T]$离散成小的等长区间$[t_i,t_{i+1}],i=0,1,...,l$，其中$t_0=0,t_{l+1}=T$，然后我们按照下面的方法收集一组样本$D=\{(t_i,x_i),i=0,1,...,l+1\}$，初始样本为$(0,x_0)$。现在，我们对$\theta_{t_i}^{\beta}$进行抽样，得到风险资产的配置$u_i\in\mathbb{R}$，然后观察下一个时刻$t_{i+1}$的财富$x_{i+1}$。

我们现在可以通过下面的式子来获得（3.13）式的近似。

$$C(\alpha,\beta)=\frac{1}{2}\sum_{(t_i,w_i)\in D}\left(\dot{V}^{\alpha}(t_i,x_i)+\lambda\int_{\mathbb{R}}\dot{\theta}_{t_i}^{\beta}(u)\ln\theta_{t_i}^{\beta}(u)du\right)^2\Delta t. \quad (3.13)$$

高斯策略的方差为$q_1e^{q_2(T-t)}$，则熵的参数化表示为$\mathcal{H}(\theta_t^{\beta})=\beta_1+\beta_2(T-t)$，其中$\beta=(\beta_1,\beta_2)^{\top}$是要学习的参数向量。$\beta_1\in\mathbb{R}$，$\beta_2>0$，接下来考虑参数化的$V^{\alpha}$，$\alpha=(\alpha_0,\alpha_1,\alpha_2,\alpha_3)^{\top}$

$$V^{\alpha}(t,x)=(x-\delta)^2e^{-\alpha_3(T-t)}+\alpha_2t^2+\alpha_1t+\alpha_0,(t,x)\in[0,T]\times\mathbb{R}.$$

在参数化的V^{α}中第一项中考虑了通胀影响的折现因子参数化为α_3，通胀影响通过参数α_3来调节，其中$\alpha_3=\frac{a_l^2}{\sigma_{ZL}^2}(T-t)$，第二项和第三项表示熵正则化的惩罚项，根据参数$\alpha_1$和$\alpha_2$来调节，后面使用梯度下降算法更新参数时，表明在优化过程中，我们不仅要考虑资产的期望收益，还要考虑熵的贡献以及通胀的影响。策略θ_t^{β}的方差为$\frac{\lambda}{2\sigma_{ZL}^2}e^{\alpha_3(T-t)}$，从而得到熵的参数化为：

$$\begin{aligned}\mathcal{H}(\theta_t^{\beta})&=-\int_{\mathbb{R}}\theta_t^{\beta}(u)\ln\theta_t^{\beta}(u)du=\frac{1}{2}\ln\left(\frac{\theta_t^{\beta}e\lambda}{\sigma_{ZL}^2}e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}\right).\\&=\frac{1}{2}\ln\frac{\theta_t^{\beta}e\lambda}{\sigma_{ZL}^2}+\frac{1}{2}\alpha_3(T-t)\end{aligned}$$

将它等同于之前的推导形式$\mathcal{H}(\theta_t^{\beta})=\beta_1+\beta_2(T-t)$。于是可以得到$\sigma_{ZL}^2$和$a_l$的参数化形式：

$$\sigma_{ZL}^2=\lambda\theta e^{1-2\beta_1},\quad \alpha_3=2\beta_2=\frac{a_l^2}{\sigma_{ZL}^2},\quad a_l=\sqrt{\alpha_3\sigma_{ZL}^2}=\sqrt{2\lambda\theta\beta_2 e^{1-2\beta_1}}.$$

通过上述式子可以得到，经通胀调整后的波动率σ_{ZL}依赖参数β_1，通胀预期收益率a_l依赖参数β_1和β_2。

将a_l和 σ_{ZL}^2代入到（3.）式中，则参数化的θ^*变为

$$\begin{aligned}\theta^*(u;t,x,\delta)&=\mathcal{N}\left(u\middle|-\frac{a_l}{\sigma_{ZL}^2}(x-\delta),\frac{\lambda}{2\sigma_{ZL}^2}e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}\right).\\&=\mathcal{N}\left(u\middle|-\sqrt{\frac{2\beta_2}{\lambda\theta}}e^{\frac{2\beta_1-1}{2}}(x-\delta),\frac{1}{2\theta}e^{2\beta_2(T-t)+2\beta_1-1}\right)\end{aligned}\tag{3.14}$$

$\mathcal{H}(\theta_t^{\beta})=\beta_1+\beta_2(T-t)$代入（3.14）式中得到

$$C(\alpha,\beta)=\frac{1}{2}\sum_{(t_i,x_i)\in D}\left(\dot{V}^{\alpha}(t_i,x_i)-\lambda\left(\dot{\beta}_1+\beta_2\left(T-t_i\right)\right)\right)^2\Delta t.\tag{3.15}$$

使用随机梯度下降算法分别更新α_1、 α_2、β_1、β_2，各梯度的计算如下：

$$\frac{\partial C}{\partial \alpha_1}=\sum_{(t_i,x_i)\in D}(\dot{V}^{\alpha}(t_i,x_i)-\lambda(\beta_1+\beta_2(T-t_i)))\Delta t\,. \tag{3.16}$$

$$\frac{\partial C}{\partial \alpha_2}=\sum_{(t_i,x_i)\in D}(\dot{V}^{\alpha}(t_i,x_i)-\lambda(\beta_1+\beta_2(T-t_i)))(t_{i+1}^2-t_i^2)\,. \tag{3.17}$$

$$\frac{\partial C}{\partial \beta_1}=-\lambda\sum_{(t_i,x_i)\in D}(\dot{V}^{\alpha}(t_i,x_i)-\lambda(\beta_1+\beta_2(T-t_i)))\Delta t\,. \tag{3.18}$$

$$\begin{aligned}\frac{\partial C}{\partial \beta_2}&=\sum_{(t_i,x_i)\in D}(\dot{V}^{\alpha}(t_i,x_i)-\lambda(\beta_1+\beta_2(T-t_i)))\Delta t\\ &\times(-\frac{2(x_{i+1}-\kappa)^2e^{-2\beta_2(T-t_{i+1})}(T-t_{i+1})-2(x_i-\kappa)^2e^{-2\beta_2(T-t_i)}(T-t_i)}{\Delta t}-\lambda(T-t_i)\end{aligned}\,. \tag{3.19}$$

当使用梯度下降算法来更新参数时，我们将$[0,T]$离散成了小的等长区间$[t_i,t_{i+1}],i=0,1,...,l$，因为在大规模数据集中，计算整个数据集的梯度显然是不现实的，相比之下，我们采用随机梯度下降，每次迭代只使用一个样本的梯度来更新模型参数（α_1、α_2、β_1、β_2），因为它只考虑一个样本的信息，而不是整个数据集。它可以在不断接收新数据的情况下持续更新模型。随机梯度下降方法更新参数β_1和β_2时，σ_{ZL}和a_l都会有相应的变化，然后讨论分析通胀风险因子对投资组合中最优投资策略、最优值函数以及最优财富过程的影响。

而且参数更新为：

$$\alpha_3=2\beta_2\,.$$

α_0是根据终端条件$V^{\alpha}(T,x;\delta)=\left(x-\delta\right)^2-\left(\delta-z\right)^2$更新的：

$$\alpha_0=-\alpha_2T^2-\alpha_1T-(\delta-z)^2\,. \tag{3.20}$$

根据约束$\mathbb{E}\left[X_T^*\right]=z$更新拉格朗日乘子$\delta$：

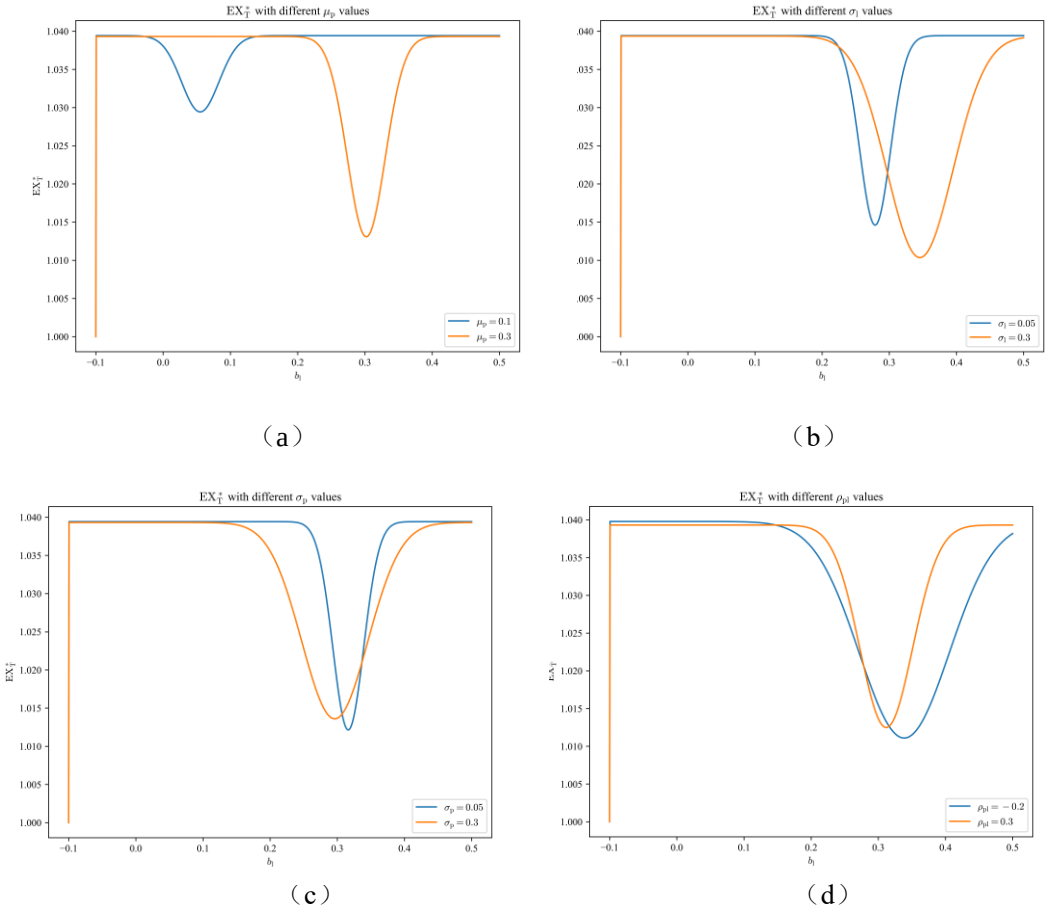
$$\delta_{n+1}=\delta_n-\varphi_n\left(X_T-z\right). \tag{3.21}$$

其中$\varphi_n>0(n\geq 1)$是学习率。在算法的实现中，我们可以用样本平均值$\frac{1}{N}\sum_j x_T^j$代替（3.22）式的X_T，以获得更稳定的学习过程。$N\geq 1$是样本容量大小，x_T^j是更新x时获得的最新的N个终端财富值，而且（3.22）式的学习方案在统计上是自我校正的。如果样本平均终端财富高于目标值z，则更新规则（3.22）

式将减少δ，考虑到（3.13）式，这反过来又降低了探索性高斯策略θ的均值。这意味着，在下一步的学习和优化过程中，风险资产的分配会更小，也就是说代理人会采取更加保守的资产配置策略，平均而言，最终财富会减少。

3.3 数值模拟与经济学解释

接下来研究通胀对最优值函数、策略以及财富过程的影响。对于两种不同的情形，我们利用 jupyter 软件对定理 4、定理 5 以及（3.16）式-（3.22）式进行模拟运算和定量分析，并就结果给出对应的经济意义上的解释，具体初始参数值详见附录 2。



（a）　　（b）

（c）　　（d）

图 3-1 在不同的风险资产预期收益率μ_p，通胀波动率σ_I，风险资产波动率σ_p以及布朗运动相关系数下预期通胀率b_I对终端财富的数学期望 $\mathbb{E}\left[X_T^*\right]$的影响

图 3-1（a）展示了不同的风险资产预期收益率μ_p下，预期通胀率对投资者终端财富的影响，首先当$\mu_p = 0.05$时，曲线随着b_I的增加而上升，这时适度的通胀刺激了经济增长，从而提高了资产价值，接着曲线开始下降，因为过高的通胀

开始侵蚀资产的实际价值，导致终端财富的期望下降。而当$\mu_p=0.3$时，随着b_l的增加，曲线急速下降，比$\mu_p=0.05$的情形下下降区间更大，说明高风险资产对通胀的敏感性更高，当通胀水平过高时，资产价值的下降速度也更快，因为高风险资产通常伴随着更高的，也就是说适度的通胀可以刺激经济增长和消费，从而提高资产价值。

图 3-1（b）展示了在不同的通胀波动率σ_l对终端财富数学期望$\mathbb{E}\left[X_T^*\right]$的影响。具体来说，我们观察到了以下现象：两条曲线首先在$\mathbb{E}\left[X_T^*\right]$为 1.04 的位置保持平稳状态。当$\sigma_l=0.3$时相较于$\sigma_l=0.05$时下降的区域更大，这表明，第一，在通胀水平较低且相对稳定时（即较低的b_l和σ_l），风险资产会因为它们提供了相对较高的预期收益而更具吸引力，然而，随着σ_l不确定性增加，也就导致投资者对未来经济环境的预期更加不确定，所以在较高的σ_l情形下，投资者不会再购入太多的风险资产，反而选择稳健的投资策略以应对潜在的风险。第二，两条曲线出现了交点，当σ_l较高时，投资者更加关注通胀风险，并减少风险资产的持有，而当σ_l较低时，投资者可能更加关注风险资产的潜在收益，这两种效应相互抵消从而导致两条曲线相交。

图 3-1（c）展示了在不同的风险资产波动率σ_p对终端财富数学期望$\mathbb{E}\left[X_T^*\right]$的影响。从图（c）可以比较对于$\sigma_p=0.1$和$\sigma_p=0.3$两种情形下$\mathbb{E}\left[X_T^*\right]$随着$b_l$先下降后上升的趋势，首先当$\sigma_p=0.1$时，可以看出图像首先有一段平稳区域，这表明在低波动率下，b_l的变化对终端财富的影响较小，这是因为σ_p较低时，此时整体风险较低，市场对通胀预期的反应不那么敏感，b_l的微小变化不足以显著影响投资组合的表现。当b_l进一步上升时，未来现金流的购买力会下降，这导致投资者会减少对风险资产的配置，从而降低了投资组合的预期回报。此外，高通胀预期也会引发央行采取紧缩政策，提高名义利率，但这并不足以抵消通胀对投资组合实际价值的负面影响。与$\sigma_p=0.1$的情况相比，当$\sigma_p=0.3$时下降区域更加陡峭，因为高波动率的风险资产对通胀预期的变化更加敏感，导致投资者会更快地调整他们的投资组合。

图 3-1（d）展示了在不同的相关系数ρ_{pl}下对终端财富数学期望$\mathbb{E}\left[X_T^*\right]$的影响。首先两条曲线分别处于平稳状态，在较低的通胀波动率下，不同ρ_{pl}对$\mathbb{E}\left[X_T^*\right]$的影响相对较小，而且投资者处于相对稳定的投资环境中。随着b_l的增加，两条曲线都开始下降，但下降的速度和幅度有所不同。当ρ_{pl}为负时，表明风险资产

价格和通胀之间存在负相关关系，这种关系使得投资者在面临高通胀时更加倾向于减少风险资产的持有，从而降低了终端财富数学期望，相反，而当 ρ_{pl} 为正时，表明风险资产价格和通胀之间存在正相关关系，即当通胀上升时，风险资产价格也会上升，投资者在面临高通胀时更加倾向于增加风险资产的持有以获取更高的收益，从而在一定程度上抵消了通胀对终端财富数学期望的负面影响。这两种效应在相交点相互抵消，导致两条曲线相交。

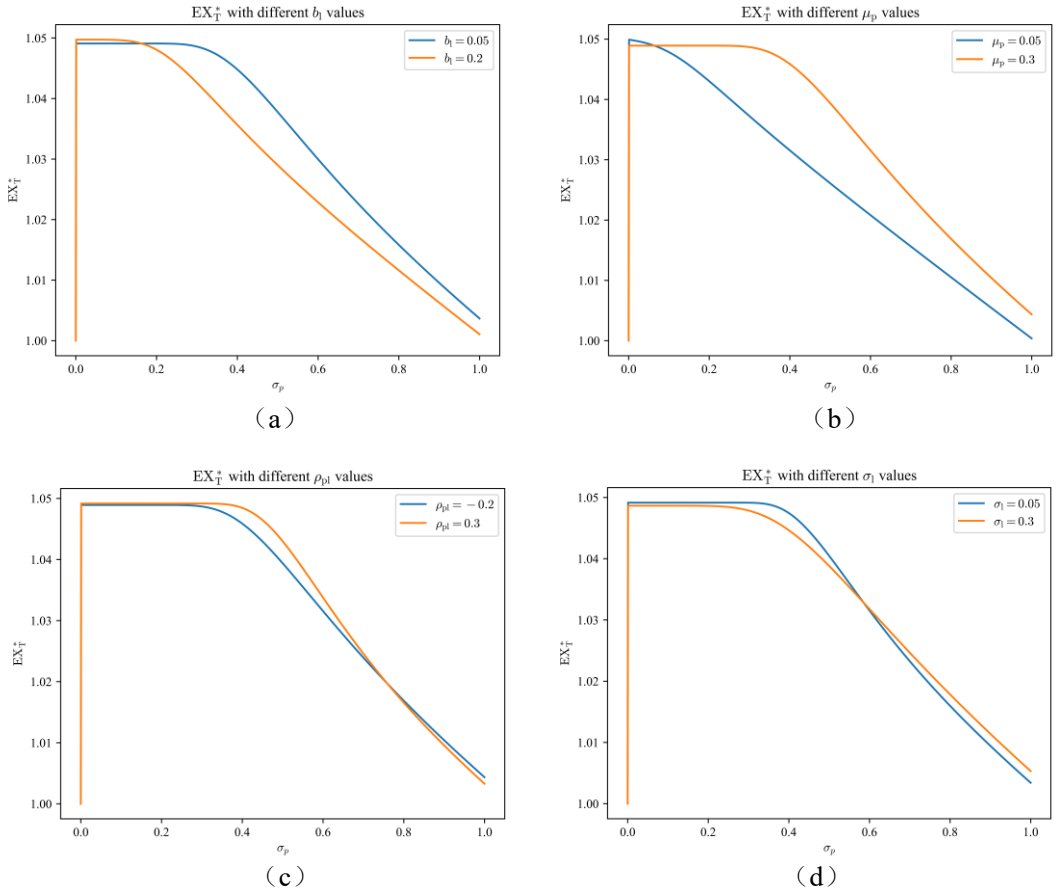


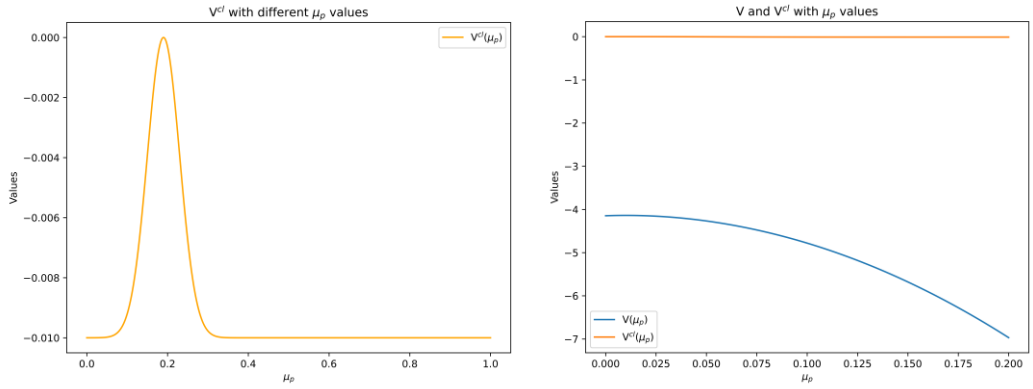
图 3-2 在不同的预期通胀率 b_l，风险资产预期收益率 μ_p，布朗运动相关系数 ρ_{pl}，以及通胀波动率 σ_l 下风险资产波动率 σ_p 对终端财富的数学期望 $\mathbb{E}\left[X_T^*\right]$ 的影响

图 3-2 展示了不同参数下 σ_p 对终端财富数学期望 $\mathbb{E}\left[X_T^*\right]$ 的影响，图（a）证明了 b_l 较高时，投资者更倾向于持有能够抵御通胀的资产（如实物资产），而不是纯粹的风险资产。这导致在相同的 σ_p 水平下，$\mathbb{E}\left[X_T^*\right]$ 较低。在 σ_p 较低时，风险资产提供了相对稳定的收益，因此投资者愿意持有这些资产。然而，随着 σ_p 的增加，风险资产的收益变得更加不确定，投资者开始寻求更安全的投资选择，这也会导致 $\mathbb{E}\left[X_T^*\right]$ 下。图（b）中 $\mu_p = 0.3$ 时，当通胀波动率 σ_p 较低时（即市场

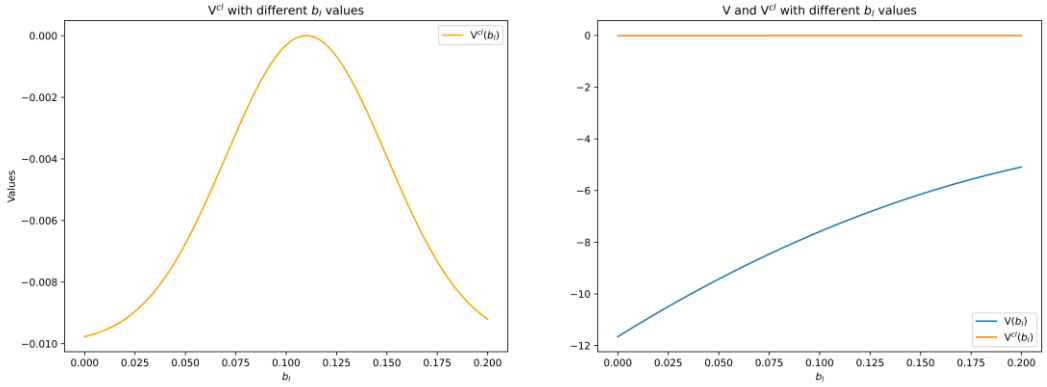
相对稳定时），投资者愿意承担更多的风险以追求高收益，导致$\mathbb{E}\left[X_T^*\right]$保持在较高的水平，当$\mu_p=0.05$时，风险资产的预期收益率较低。这意味着投资者从风险资产中获得的潜在回报也较低。因此，即使在通胀波动率σ_p较低时，投资者也不愿意承担过多的风险。这导致相较于$\mu_p=0.3$时，$\mathbb{E}\left[X_T^*\right]$更快地下降。

图（c）（d）也验证了σ_p（风险资产波动率）、σ_I（通胀波动率）以及ρ_{pI}（风险股票价格过程与通胀随机状态之间的相关系数）对终端财富数学期望$\mathbb{E}\left[X_T^*\right]$产生的影响。当$\sigma_p$较高时，投资者倾向于持有能够抵御通胀的资产而非纯粹的风险资产，导致在相同σ_I水平下$\mathbb{E}\left[X_T^*\right]$较低。随着$\sigma_p$的增加，风险资产的收益不确定性增大，促使投资者寻求更安全投资，进一步降低$\mathbb{E}\left[X_T^*\right]$。而$\sigma_I$较低时，市场相对稳定，风险资产提供相对稳定的收益，投资者愿意承担更多风险以追求高收益，使$\mathbb{E}\left[X_T^*\right]$保持较高的水平。此外，$\rho_{pI}$的正负也显著影响投资者行为，$\rho_{pI}=0.3$时，通胀与风险资产价格正相关，投资者在$\sigma_I$较低时认为市场稳定而愿意冒险，但随着$\sigma_p$增加，市场不确定性增大，投资者因为谨慎行事而导致$\mathbb{E}\left[X_T^*\right]$快速下降。$\rho_{pI}$为负时，例如图中$\rho_{pI}=-0.2$的情况，通胀与风险资产价格负相关，投资者更加关注通胀影响，即使在σ_p较低时也表现出谨慎态度导致$\mathbb{E}\left[X_T^*\right]$较早下降，然而，在高$\sigma_p$水平下，无论$\rho_{pI}$如何，投资者均会减少风险资产配置。

值得注意的是，在低σ_I环境下，通胀不确定性小，有助于稳定市场预期，σ_p对$\mathbb{E}\left[X_T^*\right]$影响更积极；随着$\sigma_I$上升，投资者对风险资产信心下降，$\mathbb{E}\left[X_T^*\right]$随$\sigma_p$增加而下降的趋势加剧。然而，在极高$\sigma_p$水平下，通胀波动率对投资者行为影响变得次要，投资者更关注风险资产本身的风险与收益平衡。



（a）　　（b）



（c）　　（d）

图 3-3 不同的风险资产预期收益率 μ_p 以及预期通胀率 b_l 对值函数 V 和 V^{cl} 的影响

图 3-3（a）和（b）反映了值函数 V 和 V^{cl} 随着 μ_p 的增大而出现的变化情况，首先值函数 V 随着 μ_p 的增大而下降，这种现象可以从以下几个方面进行解释：第一，在金融市场中，风险资产的预期收益率 μ_p 通常与其风险成正比。当 μ_p 增加时，意味着风险资产的风险也在增加。投资者在面对更高的风险时，往往要求更高的风险溢价来补偿这种风险。值函数 V 与 V^{cl} 相比多出了与 λ 相关的三项。也就是说，探索强度决定了值函数的变化大小，而且当 μ_p 变化时，a_l 和 σ_{ZL}^2 都会随之变化，进而影响整个值函数。由于 μ_p 是 a_l 的正项，所以当风险资产收益率 μ_p 增大时，a_l 会增大，但是这会导致指数项 $e^{-\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}$ 中的指数部分减小，从而使得指数项的值减小。因此，在值函数 V 中，随着 μ_p 的增大导致值函数 V 随着 μ_p 的增大而减小。而值函数 V^{cl}，当 μ_p 变化时，虽然 a_l 和 σ_{ZL}^2 会变化，但由于没有额外的与 λ 相关的项，所以值函数的变化相对较小。当 a_l^2 较小时，指数衰减项的值较大；当 a_l^2 增大到一定程度时，指数衰减项的值开始迅速减小。这解释了为什么 V^{cl} 随着 μ_p 的增大而先上升后下降，并且上升和下降的图像是一条对称的抛物线，图 3-3（d）也验证了这一说法。

图 3-3（c）展示了值函数V随着预期通胀率b_l变化的情况，在给定其他参数不变的情况下，当$b_l=0$时，a_l取得其最大值。此时，指数衰减项$e^{-\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}$的值相对较大，因为分母中的$a_l^2$较小，同时，由于 λ 项和常数项的存在，整个值函数V在$b_l=0$时取得一个初始值。随着b_l的增大，a_l的值逐渐减小。这也就会导致指数衰减项的值逐渐增大。然而，由于指数函数的衰减特性，随着b_l的进一步增大，指数衰减项的变化速度会逐渐减慢（即斜率逐渐减小）。

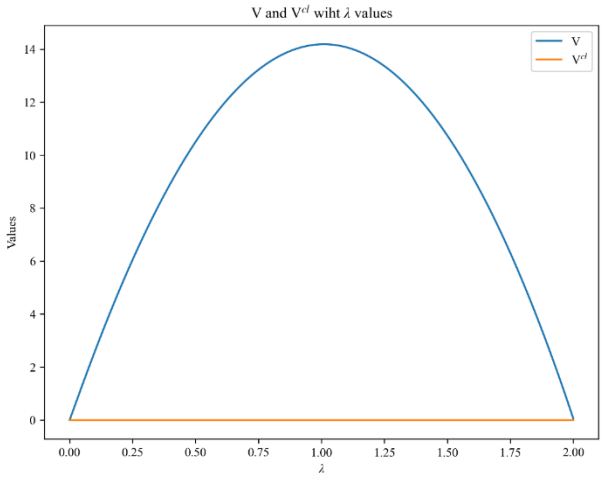
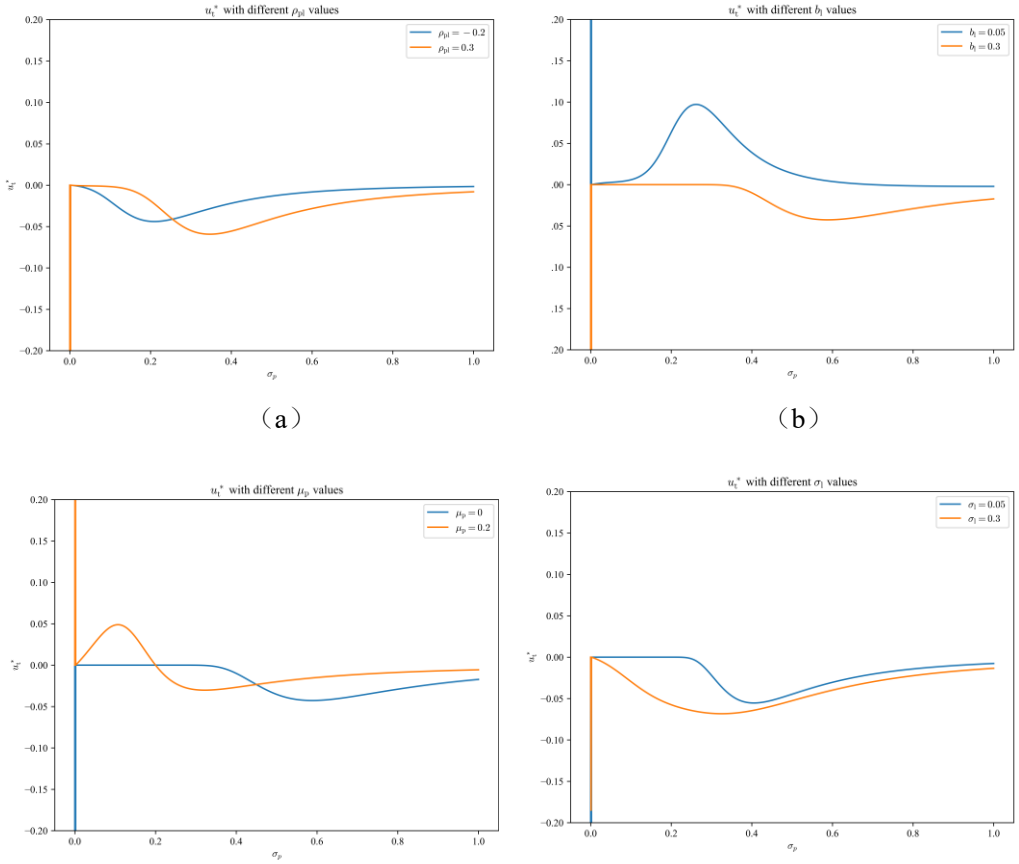


图 3-4 探索率λ对值函数V和V^{cl}的影响

图 3-4 展示了$\lambda=0$时值函数V会退化为不包含熵正则化项的V^{cl}函数，随后值函数V随着λ的增大先上升后下降，表明λ对V的影响是非线性的。具体来说，λ的增加最初是通过（3.15）式的第二项和第三项从而导致V的增加。然而，值函数V通过引入熵正则化来平衡探索与利用，使得投资者在追求最大化收益的同时，也考虑了对新策略的探索。随着λ的进一步增加，由于熵正则化的增大，（3.15）式第二项中的对数项和第三项的二次项开始主导，此时投资者开始考虑探索不同的策略，这可以解释为一种风险调整效应，在开始时，增加λ会促使投资者更加积极地寻求风险补偿，也就会导致值函数V的上升，因为探索会带来额外的收益机会。在$\lambda=1$时，探索的益处与当前策略的收益达到最佳平衡，使得V达到最大值。然而，当探索增加到一定程度时，风险补偿的边际效益递减，甚至可能由于过高的风险厌恶而导致投资者减少投资或采取更保守的策略，从而降低值函数的值。

对于值函数V^{cl}随着λ的增加基本保持在 0 左右的范围内不变，由于V^{cl}不包含与λ直接相关的项，因此探索程度的变化对V^{cl}没有影响。此外，在评估不涉及这些风险调整效应的投资决策时，V^{cl}是一个更合适的工具，因为值函数V^{cl}则更侧重于当前策略的优化，不考虑未来的探索可能性。



（a）　　（b）

（c）　　（d）

图 3-5 在不同的布朗运动相关系数ρ_{pl}，预期通胀率b_l，风险资产预期收益率μ_p以及通胀波动率σ_l下对最优投资组合策略u_t^*的影响

图 3-5（a）展示了在不同的相关性ρ_{pl}时风险资产波动率对最优策略的影响，图像中u_t^*的变化反映了投资者在不同市场条件下通过探索和策略调整来优化投资组合的过程。当$\rho_{pl} = -0.2$时，在初始阶段σ_p接近 0 时，投资组合的风险较低，此时u_t^*在 0 附近保持平稳，当σ_p开始增加时，投资者开始增加对风险资产的配置，以寻求更高的收益，曲线开始上升。随着σ_p的进一步增加，风险变得过高，投资者开始减少风险资产的配置，而且σ_p持续增加的过程中，由于风险极高，投资者不会再大幅度的改变自己的投资组合，所以此时u_t^*处于一个平稳

阶段。$\rho_{pl}=-0.2$ 的情形中，在 σ_p 较小时由于通胀与风险资产的正相关性，风险资产的表现会受到通胀的负面影响，投资者此时会快速减少风险资产的配置，导致 u_t^* 快速下降。然而，由于通胀与风险资产的正相关性仍然存在，投资者在增加风险资产配置时会更加谨慎，u_t^* 上升的速度越来越缓慢。

图 3-5（b）当预期通胀率 $b_l=0.05$ 时，风险资产的预期收益率低于无风险利率 r 加上通胀调整后的值时，投资者会迅速增加风险资产的配置，导致 u_t^* 从 0 快速上升，也反映了投资者会对不利市场条件做出迅速反应从而增加风险资产的配置，而相对于 $b_l=0.05$ 的曲线而言，预期通胀率对投资策略的影响相对微小。图 3-5（c）当风险资产预期收益率 $\mu_p=0$ 时，通胀调整后的预期收益率 a_l 仍为正），投资者对风险资产的配置持乐观态度，市场的短期波动率激发短期投资者的风险偏好，一定的波动率意味着市场的不确定性增加，但同时也带来了更多的投资机会，此时他们更愿意承担一定的风险来追求更高的收益，u_t^* 会开始缓慢上升。这反映了投资者在观察到有利市场条件或风险容忍度变化后，逐渐增加风险资产配置的行为。图 3-5（d）随着 σ_p 的增加，风险资产的价格波动性增大，但在这个阶段，风险资产的预期收益率逐渐改善，u_t^* 下降到最低点意味着此时预期收益率仍然不利，导致投资者对风险资产的配置达到最低水平。当通胀波动率 $\sigma_l=0.3$ 时，由于通胀波动率较高，投资者对通胀不确定性的担忧加剧，投资者也会减少他的风险资产配置。

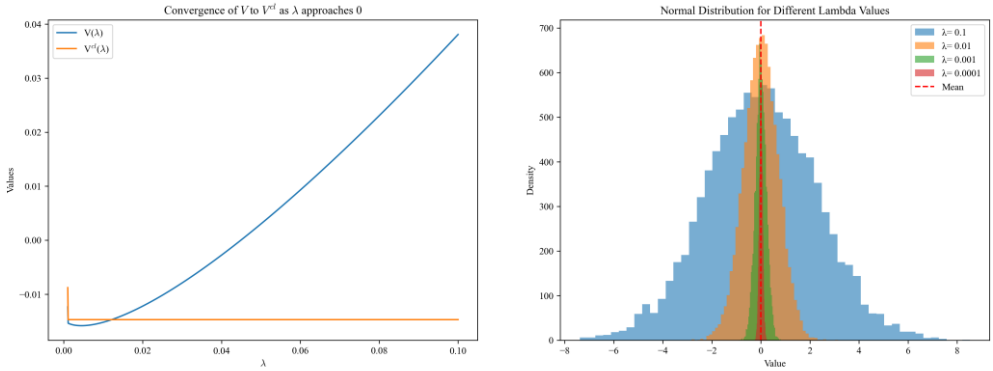


图 3-6　最优策略和值函数的收敛性

图 3-6 验证了定理 5，从图中可以看出，当 λ 逐渐减小时，EMV 问题的策略 θ^* 对应的正态分布趋向于单点分布，由此验证了两个问题的策略的弱收敛性。当 λ 为 0 时，EMV 问题的值函数与经典问题的值函数相等，当 λ 从 0 逐渐增加

时，两个值函数相差越来越大，由此通过模拟可以看出当$\lambda \to 0$时，定理 5 成立。

3.4 本章小结

本章建立了通胀背景下强化学习技术驱动的均值方差投资组合模型框架，研究了一个连续时间 MV 投资组合选择问题。通过将 MV 投资组合选择重新定义为探索与利用问题，探索部分为松弛的随机控制形式并由此推导出了探索性状态动力学方程，通过随机分析、动态规划原理等推导出 EMV 问题的 HJB 方程并求解。随后，我们设计了一个可实现的 EMV 算法，得到了一个已证明的 PIT 以及值函数和最优策略的明确函数，将其参数化并采用随机梯度下降算法进行模拟检验。最后模拟结果验证了前面得到的结论，并分析了通胀波动率等对终端财富以及最优策略的影响，并给出了相关结果的经济学含义。

研究结果表明，在考虑通胀的投资组合中，通胀波动率对最优投资组合策略有显著影响，因为它改变了投资者对风险和收益的预期。首先，在通胀波动率较低时，投资者可能倾向于保守的投资策略；而在通胀波动率较高时，他们可能更愿意承担风险以寻求更高的收益，熵正则化来平衡探索与利用，使得投资者在追求最大化收益的同时，也考虑了对新策略的探索。其次，投资者在调整投资组合时也会考虑风险与收益的平衡，因此最优投资组合策略并不是随着通胀波动率的增加而单调变化的。投资者应该根据市场情况和自身风险承受能力来制定合适的投资策略，并随着市场条件的变化进行动态调整。本章将强化学习框架下的均值方差模型推广到通胀环境中，为投资者在通胀环境中制定投资计划提供了参考依据，所得结论更加具有现实指导意义。

第 4 章 总结和展望

4.1 研究结论

首先通过将 MV 投资组合选择重新定义为探索与利用问题，探索部分为松弛的随机控制形式并由此推导出了探索性状态动力学方程。其次，在满足标准自融资的假设下，利用动态规划原理，导出了在满足一定的探索水平下的 HJB 方程，然后将最优策略代入到 HJB 方程，转化为偏微分方程进行求解，进而推导出最优财富过程。随后我们证明了探索和利用的最优平衡，反馈控制分布是一个高斯分布，其方差随时间衰减，又建立了 EMV 与经典 MV 问题之间的紧密联系，包括可解性等价和探索权重衰减到零的收敛性。再次，我们设计了一个可实现的 EMV 算法，得到了一个已证明的 PIT 以及值函数和最优策略的明确函数，将其参数化并采用随机梯度下降算法进行模拟检验。最后通过数值模拟验证了前面得到的结论，并分析了资产 ESG 评级以及投资者 ESG 偏好等对终端财富的影响，并给出了相关结果的经济学含义。研究结果表明：

（1）当投资者忽略 ESG 因素时，他们会错过与气候变化相关的投资机会，导致财富期望下降。而随着 ESG 偏好的增加，投资者更注重环保的投资组合，这些资产具有长期可持续性，能减少风险冲击，带来更高收益。但 ESG 偏好过高时，投资者的选择会受到限制，难以找到既符合高 ESG 标准又具有相对较高回报的资产。

（2）不具有 ESG 偏好的投资者更多地受到市场影响，决策完全基于预期收益率。而随着 ESG 偏好的增加，投资者会在预期收益率和 ESG 标准之间进行权衡，更倾向于投资符合 ESG 标准的资产，即使短期内可能面临损失。但当风险资产的预期收益率足够高时，投资者会一定程度上忽略 ESG 标准，更多地投资于高风险高回报的资产。

（3）投资者会在风险和收益之间做出平衡的考量，不会过于追求 ESG 而忽视其他方面。对 ESG 因素的高度重视会导致投资者选择较低风险的投资，牺牲一定的潜在回报，体现了投资者的道德和社会责任感。

（4）尽管投资者对 ESG 因素的偏好对投资组合的值函数没有过于直接的影响，但 ESG 偏好可以量化投资者面临的权衡，他们愿意牺牲一部分收益来改善其投资组合的 ESG 配置。

另一方面，引入新的布朗运动以刻画市场上连续性的通胀冲击构建投资者的的期望价值函数。接着，在满足自融资条件下，以财富作为状态变量，利用动态规划原理推导出投资者满足预期收益的条件下最小化其方差的的 HJB 方程，并求解出最优策略及财富过程。最后，利用 Jupyter 软件进行数值模拟，结合图像分析投资者在通胀情形下的最优投资决策，以及不同冲击之间的相关性对最优决策的影响。研究结果表明：

（1）风险资产预期收益率、通胀率与终端财富以及风险资产预期收益率对终端财富有显著影响。适度的通胀可以刺激经济增长，提高资产价值，但过高的通胀会侵蚀资产的实际价值。高风险资产对通胀的敏感性更高，当通胀水平过高时，资产价值的下降速度更快。通胀波动率增加时，投资者对未来经济环境的预期更加不确定，可能会减少风险资产的持有，选择稳健的投资策略。

（2）风险资产波动率较低时，市场对通胀预期的反应不敏感，微小的通胀变化不足以显著影响投资组合的表现。随着风险资产波动率的增加，未来现金流的购买力下降，投资者减少风险资产配置，降低了投资组合的预期回报。高通胀和高波动率的组合加剧了市场的不确定性，进一步降低了投资组合的预期回报。在极高通胀波动率下，央行采取积极的紧缩政策对抗高通胀，提高了名义利率和风险资产的预期收益率，终端财富的数学期望开始上升。

（3）风险资产价格和通胀之间的相关系数影响投资者的行为。负相关关系使投资者在高通胀时减少风险资产持有，正相关关系则使投资者在高通胀时增加风险资产持有以获取更高收益。

（4）投资者在不同市场条件下通过探索和策略调整来优化投资组合。风险资产波动率增加时，投资者根据风险与收益的平衡来调整风险资产的配置。通胀波动率、预期通胀率和风险资产预期收益率也影响最优投资组合策略。投资者会根据这些参数的变化来快速调整他们的投资组合，以应对市场的不确定性。

4.2 研究展望

基于本文的研究基础上，可以在未来拓展以下几个方面工作：

（1）深化 ESG 因素在投资组合选择中的影响研究：本文考虑了 ESG 评级和投资者 ESG 偏好，二者是衡量企业 ESG 表现的重要指标，投资者可以根据评级结果筛选符合标准的资产。除此之外，还可以考虑将更多的 ESG 指标纳入投资组合优化模型，比如碳足迹，以更全面地反映投资者的价值观和道德偏好。

（2）考虑到投资者的非理性行为会直接影响其投资决策，投资者情绪的变化也可能导致其行为偏差。因此，可以将投资者情绪因素引入到模型构建中，有助于投资者在制定投资组合方案时更加理性且灵活地进行资产配置，以适应市场和个人情绪的变化，从而实现长期稳定的收益，并更好地实现财富增值和风险管理。

（3）在当前通胀冲击的大背景下，未来还可以考虑将利率变动和政策风险因素引入到我们的模型中，讨论更加全面的多因素投资组合优化模型。

（4）实证分析可以揭示投资组合中资产之间的相关性以及市场趋势的变化，因此，未来可以立足于实际的金融证券市场，搜集实际数据开展实证研究，将实证研究结果与理论分析相结合，从而更深入地揭示投资组合选择的复杂性和多样性，帮助投资者建立更加合理的投资组合方案。

参考文献

[1] Markowitz H M . Portfolio selection[J]. The Journal of Finance, 1952, 7(1): 77.

[2] Li D, Ng W L. Optimal dynamic portfolio selection: Multiperiod mean-variance formulation[J]. Mathematical Finance, 2000, 10(3): 387-406.

[3] Nevmyvaka Y, Feng Y, Kearns M. Reinforcement learning for optimized trade execution[C]//Proceedings of the 23rd international conference on Machine learning. 2006: 673-680.

[4] Jiang Z, Xu D, Liang J. A deep reinforcement learning framework for the financial portfolio management problem[J]. arxiv preprint arxiv:1706.10059, 2017.

[5] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. arxiv preprint arxiv:1509.02971, 2015.

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

[7] Munos R. A study of reinforcement learning in the continuous case by the means of viscosity solutions[J]. Machine Learning, 2000, 40: 265-299.

[8] Sato M, Kobayashi S. Variance-penalized reinforcement learning for risk-averse asset allocation[C]//International Conference on Intelligent Data Enginee ring and Automated Learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 2 000: 244-249.

[9] Doya K. Reinforcement learning in continuous time and space[J]. Neural Comp utation, 2000, 12(1): 219-245.

[10] Prashanth L A, Ghavamzadeh M. Variance-constrained actor-critic algorithms for discounted and average reward MDPs[J]. Machine Learning, 2016, 105: 367-417.

[11] Haarnoja T, Tang H, Abbeel P, et al. Reinforcement learning with deep energy-based policies[C]//International Conference on Machine Learning. PMLR, 2017: 1352-1361

[12] Jacka S D, Mijatović A. On the policy improvement algorithm in continuous time[J]. Stochastics, 2017, 89(1): 348-359.

[13] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. MIT press, 2018.

[14] Liang Z, Chen H, Zhu J, et al. Adversarial deep reinforcement learning in portfolio management[R]. arxiv preprint arxiv:1808.09940, 2018.

[15] Filos A. Reinforcement learning for portfolio management[R]. arxiv preprint arxiv:1909.09571, 2019.

[16] Wang H. Large scale continuous-time mean-variance portfolio allocation via reinforcement learning[J]. arxiv preprint arxiv:1907.11718, 2019.

[17] Wang H, Zariphopoulou T, Zhou X Y. Reinforcement learning in continuous time and space: A stochastic control approach[J]. The Journal of Machine Learning Research, 2020, 21(1): 8145-8178.

[18] Wang H , Zhou X Y. Continuous-time mean-variance portfolio selection: A reinforcement learning framework[J]. Mathematical Finance, 2020, 30(4): 1273-1308.

[19] Calabuig J M, Falciani H, Sánchez-Pérez E A. Dreaming machine learning: Lipschitz extensions for reinforcement learning on financial markets[J]. Neurocomputing, 2020, 398: 172-184.

[20] Ye Y, Pei H, Wang B, et al. Reinforcement-learning based portfolio management with augmented asset movement prediction states[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(01): 1112-1119.

[21] Wang H, Yu S. Robo-advising: Enhancing investment with inverse optimization and deep reinforcement learning[C]//2021 20th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2021: 365-372.

[22] Jiang R, Saunders D, Weng C. The reinforcement learning Kelly strategy[J]. Quantitative Finance, 2022, 22(8): 1445-1464.

[23] 张桂品，田增瑞. 企业 ESG 表现对股票流动性的影响研究——来自 A 股上市公司的经验证据[J]. 现代金融, 2023(10): 30-39+21.

[24] Xu Y, Fei C, Fei W. Optimal dynamic contract with Knightian uncertainty of ESG rating[J]. International Journal of General Systems，2024, 53(3): 381-402.

[25] 费为银，谢成成，费晨. ESG 因素对平台代币价格的影响[J]. 系统工程理论

与实践, 2023, 43(8): 2304-2323.

[26] Li R, Fei C, Pan H, et al. Analysis of investment and decision-making based on ESG token platform under jump-diffusion[J]. International Journal of Systems Science, 2023, 54(4): 907-928.

[27] Li L, Fei C, Fei W. Research on investment incorporating both environmental performance and long (short) term financial performance of firms[J]. International Journal of Systems Science, 2024, 55(2): 273-299.

[28] He F, Feng Y, Hao J. Corporate ESG rating and stock market liquidity: Evidence from China[J]. Economic Modelling, 2023, 129: 106511.

[29] Friede G, Busch T, Bassen A. ESG and financial performance: aggregated evidence from more than 2000 empirical studies[J]. Journal of Sustainable Finance & Investment, 2015, 5(4): 210-233.

[30] Fatemi A, Glaum M, Kaiser S. ESG performance and firm value: The moderating role of disclosure[J]. Global Finance Journal, 2018, 38: 45-64.

[31] Liagkouras K, Metaxiotis K, Tsihrintzis G. Incorporating environmental and social considerations into the portfolio optimization process[J]. Annals of Operations Research, 2020: 1-26.

[32] Pedersen L H, Fitzgibbons S, Pomorski L. Responsible investing: The ESG-efficient frontier[J]. Journal of Financial Economics, 2021, 142(2): 572-597.

[33] Chen L, Zhang L, Huang J, et al. Social responsibility portfolio optimization incorporating ESG criteria[J]. Journal of Management Science and Engineering, 2021, 6(1): 75-85.

[34] Pástor Ľ, Stambaugh R F, Taylor L A. Sustainable investing in equilibrium[J]. Journal of Financial Economics, 2021, 142(2): 550-571.

[35] 徐凤敏, 马杰傲, 景奎. ESG 观点与股票市场定价——来自 AI 语言模型和新闻文本的证据[J]. 当代经济科学, 2023, 45(6): 29-43.

[36] Fereydooni A, Barak S, Sajadi S M A. A novel online portfolio selection approach based on pattern matching and ESG factors[J]. Omega, 2024, 123: 102975.

[37] Díaz A, Esparcia C, Alonso D, et al. Portfolio management of ESG-labeled energy companies based on PTV and ESG factors[J]. Energy Economics, 2024: 107545.

[38] Brennan M J, Xia Y. Dynamic asset allocation under inflation[J]. Journal of Finance, 2002, 57(3): 1201-1238.

[39] Siu T K. Long-term strategic asset allocation with inflation risk and regime switching[J]. Quantitative Finance, 2011, 11(10): 1565-1580.

[40] 费为银, 蔡振球, 夏登峰. 跳扩散环境下带通胀的最优动态资产配置[J]. 管理科学学报, 2015, 18(8): 83-94.

[41] Munk C, Rubtsov A. Portfolio management with stochastic interest rates and inflation ambiguity[J]. Annals of Finance, 2014, 10: 419-455.

[42] 姚海祥, 伍慧玲, 曾燕. 不确定终止时间和通货膨胀影响下风险资产的最优投资策略[J]. 系统工程理论与实践, 2014, 34(5): 1089-1099.

[43] 董迎辉, 魏思媛, 殷子涵等. 通胀风险下基于相对业绩的资产管理人的最优投资[J]. 工程数学学报, 2023, 40(1): 20-40.

[44] Bensoussan A, Keppo J, Sethi S P. Optimal consumption and portfolio decisions with partially observed real prices[J]. Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics, 2009, 19(2): 215-236.

[45] Chou Y Y, Han N W, Hung M W. Optimal portfolio-consumption choice under stochastic inflation with nominal and indexed bonds[J]. Applied Stochastic Models in Business and Industry, 2011, 27(6): 691-706.

[46] Lim B H. The effect of inflation risk and subsistence constraints on portfolio choice[J]. Journal of the Korean Society for Industrial and Applied Mathematics, 2013, 17(2): 115-128.

[47] Fei W. Optimal consumption and portfolio under inflation and Markovian switching[J]. Stochastics An International Journal of Probability and Stochastic Processes, 2013, 85(2): 272-285.

[48] Munk C, Rubtsov A. Portfolio management with stochastic interest rates and inflation ambiguity[J]. Annals of Finance, 2014, 10: 419-455.

[49] Fei C, Fei W Y. Optimal control of Markovian switching systems with applications to portfolio decisions under inflation. Acta Mathematica Scientia, 2015, 35B(2): 439–458.

[50] 费为银, 费晨, 夏登峰, 等. 模型不确定下带通胀的最优消费和投资组合问题研究[J]. 管理工程学报, 2017, 31(2): 177-184.

[51] 费晨, 杜宏俊, 费为银, 等. 通胀风险下的企业家投资——消费和对冲问题[J]. 系统工程学报, 2019, 34(3): 383-394.

[52] Kwak M, Lim B H. Optimal portfolio selection with life insurance under inflation risk[J]. Journal of Banking & Finance, 2014, 46: 59-71.

[53] Han N W, Hung M W. Optimal consumption, portfolio, and life insurance policies under interest rate and inflation risks[J]. Insurance: Mathematics and Economics, 2017, 73: 54-67.

[54] 费为银, 陈雅豪, 费晨. 通胀与生存约束下最优投资和自愿退休选择[J]. 系统工程学报, 2020, 35(1): 60-72.

[55] 费为银,张繁红,杨晓光.通胀对高管的股权激励和工作努力策略的影响[J]. 2021(2020-11):1-11.

[56] 费为银，张桂银，费晨. 通胀不确定环境下政府最优税收和债务问题研究[J]. 计量经济学报，2024，4(4): 1064-1090.

附 录

附录 1

拉格朗日乘子κ的证明：

证明：现在通过约束条件$\mathbb{E}\left[W_T^u\right]=z$确定拉格朗日乘子$\kappa$，结合（2.1）式并根据标准估计: $\mathbb{E}\left[\max_{t\in[0,T]}\left(W_t^*\right)^2\right]<\infty$和富比尼定理可以得到：

$$\mathbb{E}\left[W_t^*\right]=w_0+\mathbb{E}\left[\int_0^t\frac{\hat{\mu}^2}{\sigma^2}\left(W_s^*-\kappa\right)^2ds\right]=w_0+\int_0^t\frac{\hat{\mu}^2}{\sigma^2}\left(\mathbb{E}\left[W_s^*\right]-\kappa\right)ds$$

$$\int_0^t\frac{d\mathbb{E}\left[W_s^*\right]}{\mathbb{E}\left[W_s^*\right]-\kappa}=-\int_0^t\frac{\hat{\mu}^2}{\sigma^2}ds$$

因此，$\mathbb{E}\left[W_t^*\right]-\kappa=(w_0-\kappa)e^{-\frac{\hat{\mu}^2}{\sigma^2}t}$，约束$\mathbb{E}\left[w_T^u\right]=z$现在变为

$$(w_0-\kappa)e^{-\frac{\hat{\mu}^2}{\sigma^2}t}+\kappa=z$$

即拉格朗日乘子$\kappa=\dfrac{ze^{\frac{\hat{\mu}^2}{\sigma^2}T}-w_0}{e^{\frac{\hat{\mu}^2}{\sigma^2}T}-1}$。证毕

定理 2 的证明

证明：当$(t,w)\in[0,T]\times\mathbb{R}$，假设反馈控制$\tilde{\pi}$是可容许的，由 $\tilde{\pi}$对初始条件$W_t^{\tilde{\pi}}=w$ 生成的开环控制 $\tilde{\pi}=\left\{\tilde{\pi}_v,v\in[t,T]\right\}$， $\left\{W_s^{\tilde{\pi}},s\in[t,T]\right\}$是满足反馈控制 $\tilde{\pi}$的财富过程，运用伊藤公式可以得到：

$$\begin{aligned}&V^{\pi}(s,\tilde{W}_s)=V^{\pi}(t,w)\\&+\int_t^sV_t^{\pi}(v,W_v^{\tilde{\pi}})dv+\int_t^s\int_{\mathbb{R}}\left(\frac{1}{2}(\sigma^2u^2v_{ww}^{\pi}(v,W_v^{\tilde{\pi}})+\hat{\mu}uV_w^{\pi}(v,W_v^{\tilde{\pi}})\right)\tilde{\pi}(u)dudv\quad(1)\\&+\int_t^s\sqrt{\int_{\mathbb{R}}\sigma^2u^2\pi(u)du}V_w^{\pi}(v,W_v^{\tilde{\pi}})dB_v\end{aligned}$$

定义停时$\tau_n:=\inf\left\{s\geq t:\int_t^s\sigma^2\int_{\mathbb{R}}u^2\tilde{\pi}_v(u)du\left(V_w^{\pi}(v,W_v^{\tilde{\pi}})dv\geq n\right)\right\},n\geq1$。通过（1）式可以得到：

$$V^{\pi}(t,w)=\mathbb{E}\left[\begin{array}{l}V^{\pi}(s\wedge\tau_n,\tilde{W}_{s\wedge\tau_n})-\int_t^{s\wedge\tau_n}V_t^{\pi}(v,W_v^{\tilde{\pi}})dv-\\ \int_t^{s\wedge\tau_n}\int_{\mathbb{R}}\left(\frac{1}{2}(\sigma^2u^2v_{ww}^{\pi}(v,W_v^{\tilde{\pi}})+\hat{\mu}uV_w^{\pi}(v,W_v^{\tilde{\pi}})\right)\tilde{\pi}(u)dudv\end{array}\middle|W_t^{\tilde{\pi}}=w\right] \quad (2)$$

假设V^{π}是光滑函数，有下列式子：

$$V_t^{\pi}(t,w)+\int_{\mathbb{R}}\left(\frac{1}{2}(\sigma^2u^2V_{ww}^{\pi}(t,W)+\hat{\mu}uV_w^{\pi}(t,w)+\lambda\int_{\mathbb{R}}\ln\pi(u;t,w)\right)\pi(u;t,w)du=0$$

当$(t,w)\in[0,T]\times\mathbb{R}$，可以得到：

$$V_t(t,w)+\min_{\hat{\pi}\in\mathcal{P}(\mathbb{R})}\int_{\mathbb{R}}\left(\frac{1}{2}(\sigma^2u^2V_{ww}(t,w)+\hat{\mu}uV_w(t,w)+\lambda\int_{\mathbb{R}}\ln\hat{\pi}(u)\right)\hat{\pi}(u)du\le 0 \quad (3)$$

（3）式的最小化哈密顿量是由反馈控制（2.20）式给出的，（3）式满足

$$V^{\pi}(t,w)\ge\mathbb{E}\left[V^{\pi}(s\wedge\tau_n,\tilde{W}_{s\wedge\tau_n})+\lambda\int_t^{s\wedge\tau_n}\int_{\mathbb{R}}\tilde{\pi}(u)\ln\tilde{\pi}(u)dudv\middle|W_t^{\tilde{\pi}}=w\right]$$

对于$(t,w)\in[0,T]\times\mathbb{R}$，$s\in[t,T]$，取$s=T$，$n\to\infty$，并且根据终端条件$V^{\pi}(T,w)=V^{\tilde{\pi}}(T,w)=(w-\kappa)^2-(\kappa-z)^2$，可以得到

$$V^{\pi}(t,w)\ge\mathbb{E}\left[V^{\pi}(T,\tilde{W}_T)+\lambda\int_t^T\int_{\mathbb{R}}\tilde{\pi}(u)\ln\tilde{\pi}(u)dudv\middle|W_t^{\tilde{\pi}}=w\right]=V^{\tilde{\pi}}(t,w)$$

证毕。

定理 3 的证明

证明：在满足初始条件时，反馈策略$\pi_0(u;t,w,\kappa)=\mathcal{N}\left(u\middle|a(w-\kappa),c_1e^{c_2(T-t)}\right)$可以生成一个开环策略$\pi_0$，而且由 Feynman-Kac 定理可知，对应的值函数V^{π_0}满足以下偏微分方程：

$$V_t^{\pi_0}(t,w;\kappa)+\int_{\mathbb{R}}\left(\begin{array}{l}\frac{1}{2}(\sigma^2u^2V_{ww}^{\pi_0}(t,;\kappa)+\hat{\mu}uV_w^{\pi_0}(t,w;\kappa)\\ +\lambda\int_{\mathbb{R}}\pi_0(u;t,w,\kappa)\end{array}\right)\pi_0(u;t,w,\kappa)du=0 \quad (4)$$

满足终端条件$V^{\pi_0}(T,w;\kappa)=(w-\kappa)^2-(\kappa-z)^2$

下面解方程（4）。将（2.13）代入到（4）可以得到下列式子

$$V_t^{\pi_0}(t,w;\kappa)+\int_{\mathbb{R}}\left(\begin{array}{l}\dfrac{1}{2}\dfrac{(\sigma^2a^2\left(V_w^{\pi_0}(t,w;\kappa)\right)^2}{V_{ww}^{\pi_0}(t,w;\kappa)}+\hat{\mu}a\dfrac{\left(V_w^{\pi_0}(t,w;\kappa)\right)^2}{V_{ww}^{\pi_0}(t,w;\kappa)}\\ +\lambda\int_{\mathbb{R}}\ln\pi_0(u;t,w,\kappa)\end{array}\right)\pi_0(u;t,w,\kappa)du=0$$

化简得到：

$$V_t^{\pi_0}(t,w;\kappa)+\frac{a\left(V_w^{\pi_0}(t,w;\kappa)\right)^2(a\sigma^2+2\hat{\mu})}{2V_{ww}^{\pi_0}(t,w;\kappa)}-\frac{\lambda}{2}\ln\frac{2\pi e\lambda}{\sigma^2V_{ww}^{\pi_0}(t,w;\kappa)}=0 \qquad (5)$$

假设解的形式为：

$$v(t,w;\kappa)=(w-\kappa)^2e^{f(t)}+h(t) \qquad (6)$$

$$f(T)=0,\quad h(T)=-(\kappa-z)^2$$

对（6）式求偏导可以得到：

$$v_t(t,w;\kappa)=f'(t)(w-\kappa)^2e^{f(t)}+h'(t)$$
$$v_w(t,w;\kappa)=2(w-\kappa)e^{f(t)}$$
$$v_{ww}(t,w;\kappa)=2e^{f(t)}$$

将上述式子代入（5）式中得到：

$$(w-\kappa)^2f'(t)e^{f(t)}+h'(t)-(\frac{4a(w-\kappa)^2(a\sigma^2+2\hat{\mu})e^{2f(t)}}{4e^{f(t)}})+\frac{\lambda}{2}\left[1-\ln\frac{2\pi e\lambda}{2\sigma^2e^{f(t)}}\right]$$
$$=\left[f'(t)-a(a\sigma^2+2\hat{\mu})\right](w-\kappa)^2e^{f(t)}+h'(t)+\frac{\lambda}{2}\left[-\ln\frac{\pi\lambda}{\sigma^2}+f(t)\right]=0$$

通过上式可以看出

$$\begin{cases}h'(t)+\dfrac{\lambda}{2}\left[f(t)-\ln\dfrac{\pi\lambda}{\sigma^2}\right]=0\\ f'(t)-a(a\sigma^2+2\hat{\mu})=0\end{cases}$$

即可以得到

$$h(t)=-\frac{\lambda}{2}(T-t)\ln\frac{\pi\lambda}{\sigma^2}-\frac{\lambda}{2}\frac{\hat{\mu}^2}{\sigma^2}(T-t)T+\frac{\lambda}{4}\frac{\hat{\mu}^2}{\sigma^2}(T^2-t^2)-(w-z)^2$$
$$=-\frac{\lambda}{2}(T-t)\left(\frac{\hat{\mu}^2}{\sigma^2}T+\ln\frac{\pi\lambda}{\sigma^2}\right)+\frac{\lambda}{4}\frac{\hat{\mu}^2}{\sigma^2}(T^2-t^2)-(w-z)^2$$

$$f(t)=a(a\sigma^2+2\hat{\mu})$$

将 $f(t)$、$h(t)$ 代入（6）式可以得到

$$V^{\pi_0}(t,w;\kappa)=(w-\kappa)^2e^{(2a\hat{\mu}+\sigma^2a^2)(T-t)}+F_0(t)$$

其中 $h(t)$ $F_0(t)$（即 $h(t)$）是仅依赖于 t 的光滑函数，很容易验证满足定理 5 中的条件，因此，定理适用，根据策略改进定理（2.20）式，在当前情形下反馈策略为：

$$\pi_1(u;t,w,\kappa)=\mathcal{N}\left(u\Big|-\frac{\hat{\mu}}{\sigma^2}(w-\kappa),\frac{\lambda}{\sigma^2 e^{(2a\hat{\mu}+\sigma^2a^2)(T-t)}}\right)$$

再计算 π_1 对应的值函数应该为 $V^{\pi_1}(t,w;\kappa)=(w-\kappa)^2e^{-\frac{\hat{\mu}^2}{\sigma^2}(T-t)}+F_1(t)$，其中 $F_1(t)$ 是只依赖于 t 的光滑函数，计算类似如上，定理 4 再次适用，值函数 V^{π_1} 和策略 π_2 将恰好改进为最优值函数（2.15）和最优高斯策略（2.16），因此，当 $n\geq 2$ 时，在策略改进定理中（2.20）式下，策略和价值函数都不再严格改进。证毕。

定理 5 的证明

证明：反馈控制的弱收敛性来自于表述 a）和 b）中 u^* 和 θ^* 的显式形式。值函数的逐点收敛性很容易来自于 $V(\cdot)$ 和 $V^{cl}(\cdot)$ 的形式，并且 $\lim\limits_{\lambda\to 0}\frac{\lambda}{2}\ln\frac{\sigma^2}{\theta\lambda}=0$。进一步

$$\begin{aligned}&\lim_{\lambda\to 0}\left|V(t,x;\delta)-V^{cl}(t,x;\delta)\right|\\&=\lim_{\lambda\to 0}\left|-\frac{\lambda}{2}(T-t)\left(\frac{\hat{\mu}^2}{\sigma^2}T+\ln\frac{\theta\lambda}{\sigma^2}\right)+\frac{\lambda}{4}\frac{\hat{\mu}^2}{\sigma^2}(T^2-t^2)\right|=0\end{aligned}$$

证毕。

定理 6 的证明

证明：$\left\{\theta_t^*,0\leq t\leq T\right\}$ 是由反馈控制 θ^* 相对于在 $t=0$ 时刻的初始状态 x_0 生成的开环控制

$$\theta^*(u;t,x,\delta)=\mathcal{N}\left(u\left|-\frac{a_l}{\sigma_{ZL}^2}(x-\delta),\frac{\lambda}{2\sigma_{ZL}^2}e^{\frac{a_l}{\sigma_{ZL}^2}(T-t)}\right.\right)$$

$\left\{X_t^*,t\in[0,T]\right\}$ 是 EMV 问题相应的最优财富过程，最后的结果可以从陈述（a）中 $V(\cdot)$ 和 陈述（b）中的 $V^{cl}(\cdot)$ 表达形式中得到：

$$C^{u^*,\theta^*}(0,x_0;\delta)=-\frac{\lambda}{2}T\left(\frac{a_l^2}{\sigma_{ZL}^2}T+\ln\frac{\boldsymbol{\pi}\lambda}{\sigma_{ZL}^2}\right)+\frac{\lambda}{4}\frac{a_l^2}{\sigma_{ZL}^2}T^2-\lambda\mathbb{E}\left[\int_0^T\theta_t^*(u)\ln\theta_t^*(u)dudt\right]$$

其中 $\int_{\mathbb{R}} \theta_t^*(u)\ln v_t^*(u)du = -\frac{1}{2}\ln\left(\frac{\pi e\lambda}{\sigma_{ZL}^2}e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}\right)$，计算过程如下：

令 $\mu_1 = -\frac{a_l}{\sigma_{ZL}^2}(x-\delta)$，$\sigma_1 = \frac{\lambda}{2\sigma_{ZL}^2}e^{\frac{a_l}{\sigma_{ZL}^2}(T-t)}$

$$\int_{\mathbb{R}} \theta_t^*(u)\ln\theta_t^*(u)du$$

$$= \int_{\mathbb{R}} \frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\exp\left\{-\frac{(u-\mu_1)^2}{2\sigma_1^2}\right\}\ln\frac{1}{\left(2\pi\sigma_1^2\right)^{1/2}}\exp\left\{-\frac{(u-\mu_1)^2}{2\sigma_1^2}\right\}du$$

$$= \frac{1}{\left(2\pi\sigma_1^2\right)^{1/2}}\cdot -\ln\sqrt{2\pi}\sigma_1\int_{\mathbb{R}}\exp\left\{-\frac{(u-\mu_1)^2}{2\sigma_1^2}\right\}du + \frac{1}{\left(2\pi\sigma_1^2\right)^{1/2}}\int_{\mathbb{R}}\exp\left\{-\frac{(u-\mu_1)^2}{2\sigma_1^2}\right\}\frac{(u-\mu_1)^2}{2\sigma_1^2}du$$

$$= -\frac{\ln\sqrt{2\pi}\sigma}{\sqrt{2\pi}\sigma}\sqrt{2}\sigma\int_{\mathbb{R}}\exp\left\{\frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\right\}d\left(\frac{u-\mu_1}{\sqrt{2}\sigma_1}\right) + \frac{1}{\left(2\pi\sigma_1^2\right)^{1/2}}\sqrt{2}\sigma\int_{\mathbb{R}}\exp\left\{\left(\frac{u-\mu_1}{\sqrt{2}\sigma_1}\right)^2\right\}\frac{(u-\mu_1)^2}{2\sigma_1^2}d\left(\frac{u-\mu_1}{\sqrt{2}\sigma_1}\right)$$

$$= -\frac{\ln\sqrt{2\pi}\sigma_1}{\sqrt{\pi}}\int_{\mathbb{R}}e^{-y^2}dy + \frac{1}{\sqrt{\pi}}\cdot\int_{\mathbb{R}}e^{-y^2}y^2dy$$

$$= -\ln\sqrt{2\pi}\sigma_1 + \frac{1}{\sqrt{\pi}}\cdot -\frac{1}{2}(0-\int_{\mathbb{R}}e^{-y^2}dy)$$

$$= -\frac{1}{2}\left(1+\ln(2\pi\sigma_1^2)\right)$$

$\sigma_1 = \frac{\lambda}{2\sigma_{ZL}^2}e^{\frac{a_l}{\sigma_{ZL}^2}(T-t)}$ 代入，则可以得到 $\int_{\mathbb{R}} \theta_t^*(u)\ln\theta_t^*(u)du = -\frac{1}{2}\ln\left(\frac{\pi e\lambda}{\sigma_{ZL}^2}e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-t)}\right)$，

则：

$$C^{u^*,\theta^*}(0,x_0;\delta) = -\frac{\lambda}{2}T\left(\frac{a_l^2}{\sigma_{ZL}^2}T + \ln\frac{\pi\lambda}{\sigma_{ZL}^2}\right) + \frac{\lambda}{4}\frac{a_l^2}{\sigma_{ZL}^2}T^2 - \lambda\mathbb{E}\left[\int_0^T \pi_t^*(u)\ln\pi_t^*(u)dudt\right]$$

$$= -\frac{\lambda}{4}\frac{a_l^2}{\sigma_{ZL}^2}T^2 + \frac{\lambda}{2}T\left(\ln\frac{\sigma_{ZL}^2}{\pi\lambda}\right) - \lambda\mathbb{E}\int_0^T -\frac{1}{2}\ln\left(\frac{\pi e\lambda}{\sigma_{ZL}^2}e^{\frac{a_l^2}{\sigma_{ZL}^2}(T-s)}\right)ds$$

$$= -\frac{\lambda}{4}\frac{a_l^2}{\sigma_{ZL}^2}T^2 + \frac{\lambda}{2}T\left(\ln\frac{\sigma_{ZL}^2}{\pi\lambda} + \ln\frac{\pi e\lambda}{\sigma_{ZL}^2}\right) + \frac{\lambda}{2}\frac{a_l^2}{\sigma_{ZL}^2}\frac{1}{2}(T^2-0^2)$$

$$= \frac{\lambda}{2}T$$

证毕。

附录 2

各类表格

表 2-1　基本参数设置

符号	名称	数值
μ	预期收益率	0.1
σ	波动率	0.2
T	终端时刻	1
Δt	步长	1/2520
r	年利率	0.02
λ	探索权重	2
g	ESG 得分	0.6
θ	学习率	0.0005
M	总训练集	20000
N	样本量	10
ρ	夏普比率	0.05

表 3-1　基本参数设置

符号	名称	数值
μ_p	风险资产预期收益率	0.3
σ_p	风险资产波动率	0.2
T	终端时刻	1
Δt	步长	1/2520
r	年利率	0.02
λ	探索权重	2
α 、 β	学习的参数	0.1
θ	学习率	0.0005
M	总训练集	20000
N	样本量	10
η_α	更新 α 的学习率	0.0005
η_β	更新 β 的学习率	0.005
b_l	预期通胀率	0.02
σ_l	通胀波动率	0.2
δ	拉格朗日乘子值	1.0
z	目标收益	1.1

硕士期间发表的论文

[1] 贵秀子, 费为银, 邓寿年. ESG 对强化学习的连续时间均方投资组合的影响.[J]安徽工程大学

已录用 待发表。

致 谢

回想整个研究生学习和科研生涯，我收获了无数的帮助和支持，我想借此机会向所有关心、支持和帮助过我的人表示最诚挚的感谢。

首先，我要由衷地感谢我的恩师——费为银教授。在整个硕士研究生期间，费老师给予了我无私的指导和支持，不仅在学术上向我提供了宝贵的指导和建议，还在生活上给予了我无微不至的关心和照顾。在科研工作中费老师始终引导着我、鞭策着我、激励着我，使我的科研能力、学习能力、独立思考能力以及分析和解决问题的能力都得到了前所未有的提高。导师严谨负责的治学态度和科研精神深深影响着我，我也养成了严谨、踏实、专注等良好的学习和工作习惯。生活方面，费老师也经常主动关心和了解我的思想、身体和家庭状况，帮助我解决一些生活上的问题，让我能够更加专注于自己的学习和工作。师恩浩荡，感激不尽！

同时，我要深深地感谢各位教授我研究生课程的辛勤的园丁们。感谢带我《经济控制论》课程的夏登峰老师，感谢您的耐心讲解，激发了我做好科研的斗志；感谢教授我《金融数学》课程的沈明轩老师，感谢您帮助我夯实了金融数学和随机分析的基础；感谢辅导我数值计算和仿真的张伟老师，感谢潘海峰老师、李志民老师、张伟老师、张玥老师、刘大俊老师、魏晴晴老师等等。师恩难忘，铭记于心！

其次，我要感谢各位帮助过我的优秀的师兄师姐们。感谢李任重、谢成成、李亮、吴开强师兄，在我学习 Matlab 软件和做数值模拟的过程中，向我分享了宝贵的学习经验和技巧，帮助我快速全面地掌握了数值模拟的关键技术。感谢费晨师姐和徐玥师姐，在我推导数学模型和公式以及汇报的过程中耐心地向我答疑解惑。感谢李文瑞师兄以及吴开强师兄，在我寻找工作机会时向我提供了很多宝贵的信息和及时的帮助！

最后，我要特别感谢我的父母。感谢你们在我整个学生生涯中一直给予我无限的支持和理解，始终是我最坚强的后盾。也感谢你们教会我很多做人的道理，让我懂得感恩，明辨是非！我会继续加油，努力成为你们的骄傲。

谨以此文，献给所有支持和帮助过我的人，再次表示最诚挚的感谢。

2025 年 5 月 12 日