Operations Research and Fuzziology 运筹与模糊学, 2025, 15(2), 227-236
Published Online April 2025 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2025.152079

Hans汉斯

基于Actor-Critic强化学习的投资与消费问题

刘峻均，徐海燕，卢相刚

广东工业大学数学与统计学院，广东 广州

收稿日期：2025年2月24日；录用日期：2025年3月26日；发布日期：2025年4月3日

摘　要

本文研究了基于Actor-Critic强化学习的最优资产与消费问题。为了描述个体对退休后实际消费水平较低的现象，我们假设个体在退休后的最低消费水平和养老金水平较低，金融市场的资产价格由马尔可夫链调节，考虑通胀因素、习惯消费水平，建立状态转换的财富模型。利用动态规划原理得到了Hamilton-Jacobi-Bellman(HJB)方程。由于扩散过程和状态切换，几乎不可能得到一个封闭形式的解。我们设计出一种基于Actor-Critic强化学习框架下的数值算法来解决最优控制问题，通过对财富过程、优化函数的离散化和对值函数、控制函数的神经网络参数化，采用策略梯度下降算法来改进控制函数，而对于值函数，采用一种TD误差方法来更新。最后是对该优化问题的数值结果展示。

关键词

消费，投资，制度转换，强化学习，梯度下降

Investment and Consumption Problems Based on Actor-Critic Reinforcement Learning

Junjun Liu, Haiyan Xu, Xianggang Lu

School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou Guangdong

Received: Feb. 24th, 2025; accepted: Mar. 26th, 2025; published: Apr. 3rd, 2025

Abstract

This paper investigates the optimal asset and consumption problem based on Actor-Critic reinforcement learning. To describe the phenomenon of relatively low actual consumption levels after retirement, we assume that individuals have lower minimum consumption levels and pension levels after retirement. The asset prices in the financial market are regulated by a Markov chain, and

文章引用: 刘峻均, 徐海燕, 卢相刚. 基于 Actor-Critic 强化学习的投资与消费问题[J]. 运筹与模糊学, 2025, 15(2): 227-236. DOI: 10.12677/orf.2025.152079

we consider inflation factors and habitual consumption levels to establish a wealth model with state transitions. By applying the principle of dynamic programming, we derive the Hamilton-Jacobi-Bellman (HJB) equation. Due to the diffusion process and state switching, it is nearly impossible to obtain a closed-form solution. We design a numerical algorithm based on the Actor-Critic reinforcement learning framework to solve the optimal control problem. By discretizing the wealth process and the optimization function, and parameterizing the value function and control function using neural networks, we use a gradient descent algorithm to improve the control function. For the value function, we use a TD error method to update it. Finally, we present the numerical results of the optimization problem.

Keywords

Consumption, Investment, Regime-Switching, Reinforcement Learning, Gradient Descent

1. 引言

资产配置和消费控制是金融学中的经典问题，其目标是在不确定的市场环境中，通过合理分配资金于不同资产类别并制定最优消费计划，实现投资者财富的长期稳健增长和消费效用的增大。传统方法通常基于效用最大化理论，假设市场环境已知且投资者风险偏好固定。经典的投资消费模型可以追溯到Merton [1]，具有 CRRA 效用的 Merton 模型提出了消费–总财富比和生命周期投资策略。直到近年来投资与消费的问题还在以模型不断更新优化的形式在被各个领域研究着。在 2023 年，Ferreira 等人[2]研究了稳健公用事业条件下的最佳消费、投资和人寿保险选择；如 Tao 等人[3]提出了一个非均匀状态切换的随机控制问题，并将其应用于消费投资模型；如 Wang 等人[4]研究了在年龄依赖性风险偏好下的家庭投资–消费–保险政策，即在不确定的生命周期范围内，从跨期消费、遗产和最终财富中最大化预期的贴现效用。虽然这些作者考虑了跨期消费加上终端财富的预期效用最大化，但并没有考虑到个体的习惯参考水平的内源性更新会更加贴近现实意义。

习惯的持久性最初是由 Pollak [5]、Ryder [6]研究的，它们认为消费具有习惯性参考依赖，习惯消费水平应该由过去的实际消费的加权值得到。根据实际消费是否高于习惯消费，可以把效用函数分成两类，第一类是当实际消费超过习惯消费时，个体可以获得正效用，反之，第二类，个体将获得负效用。文献[7] [8]在这基础上定义了 S 型效用函数，认为个体一生获得的总效用是正效用和负效用之和。从而探索了习惯性持久性和 S 型效用来描述个人对消费的偏好。本文目的是引入习惯消费过程来更好地研究个体参与社会养老保险模型下的资产与消费问题。本文研究内容是假设个体参与各种风险投资，拥有工资过程，以及退休金过程，考虑通胀因素和随机市场环境，寻找最优控制，实现最大化整体超额消费效用，如文献[9]，虽然该文章使用鞅方法来讨论个体退休前后的投资与消费问题，但投资模型较为简单，以及没有考虑随机市场环境、通胀因素的影响，也就是缺乏更复杂的模型场景。因此，本文不仅把模型更加丰富化，还引入一种新的强化学习算法来进行研究。

强化学习作为机器学习领域的重要分支，近年来在解决复杂决策问题方面取得了显著进展。近年来，基于 Actor-Critic 框架的强化学习方法在连续时间金融决策问题中得到了广泛关注。Actor-Critic 框架结合了策略梯度(Policy Gradient)算法和价值函数逼近的优点，能够有效处理高维状态空间和连续动作空间。

例如，Wang 和 Zhou 等人[10]在强化学习框架下研究了连续时间和空间下的探索性随机控制问题，为金融决策问题提供了新的解决思路。Jia 和 Zhou [11]进一步在 Actor-Critic 框架下研究了连续时间和空间下的策略梯度算法，显著提升了算法的稳定性和效率。此外，Zhou 等人[12]提出了一种求解高维哈密顿－雅可比－贝尔曼(HJB)方程的数值方法，通过学习确定性策略和策略梯度，有效提高了控制效率和样本利用率，这有助于平衡探索和开发，避免局部最优。基于强化学习中的 Actor-Critic 框架，本文目的是设计一种数值方法来解决上述最优控制问题。该框架包括两部分：第一部分是用于评论家的策略评估(Policy Evaluation, PE)，PE 能够求解价值函数，另一部分是用于演员的策略迭代(Policy Iteration, PI)，PI 能够求解策略函数。演员根据策略做出相应的动作，而评论家对其进行价值评估，演员根据得分重新改变策略，调整动作。它们能够相互评价，促进策略和价值函数的更新与优化。目前关于研究消费问题的强化学习方法还尚少，本文为连续时间下具有复杂模型的优化问题提供一种新的数值方法，与传统方法相比，具有超强的学习能力，能够更快地收敛。

本文的其余部分如下，第二节是最优控制问题的制定；在第三节中，构造基于策略迭代的演员－评论家方法来近似值函数和控制函数；在第四节中，给出一个数值例子作为优化问题的数值结果。

2. 数学模型

我们设$(\Omega,\mathcal{F},\mathbb{F},\mathbb{P})$是一个完全滤波的概率空间，其中$\mathbb{F}=\{\mathcal{F}_t\}_{t\in[0,T]}$是一个滤子，满足通常的条件$\mathcal{F}_t$表示直到时间$t$所有可获得的市场信息，$T$是最大生存时刻，下面所有的随机过程都适应于$\mathbb{F}$。

2.1. 财富过程

为了描述随机市场环境对我们的投资过程、工资过程和养老金过程的影响，我们使用了一个连续时间的马尔科夫链$\alpha(t)$去表示市场状态，它具有有限状态并取值在一个有限空间$\mathcal{M}=\{1,\cdots,m\}$中。转移概率矩阵$Q=(q_{ij})\in\mathbb{R}^{m\times m}$生成连续时间的马尔科夫链$\alpha(t)$，也就是说

$$P\{\alpha(t+\delta)=j|\alpha(t)=i,\alpha(u),u\le t\}=\begin{cases}q_{ij}\delta+o(\delta), & j\neq i,\\ 1+q_{ii}\delta+o(\delta), & j=i.\end{cases}$$

假设投资者有三种可供投资的金融资产：无风险资产、风险资产和通胀指数债券。无风险资产价格$S_0=\{S_0(t),t\ge0\}$被定义为：

$$\frac{\mathrm{d}S_0(t)}{S_0(t)}=R(\alpha(t))\mathrm{d}t.$$

风险资产价格$S_1=\{S_1(t),t\ge0\}$满足以下随机微分方程：

$$\frac{\mathrm{d}S_1(t)}{S_1(t)}=\mu_S(\alpha(t))\mathrm{d}t+\sigma_S(\alpha(t))\mathrm{d}W_S(t).$$

通货膨胀水平$I=\{I(t),t\ge0\}$表示为：

$$\frac{\mathrm{d}I(t)}{I(t)}=\mu_I(\alpha(t))\mathrm{d}t+\sigma_I(\alpha(t))\mathrm{d}W_I(t).$$

此外，由于通货膨胀的影响，个人可以考虑通过购买通货膨胀指数债券进行对冲，因此通货膨胀指数债券被纳入模型。假设通货膨胀指数债券的价格水平$B=\{B(t),t\ge0\}$服从如下式子：

$$\frac{\mathrm{d}B(t)}{B(t)}=r(\alpha(t))\mathrm{d}t+\frac{\mathrm{d}I(t)}{I(t)}=(r(\alpha(t))+\mu_I(\alpha(t)))\mathrm{d}t+\sigma_I(\alpha(t))\mathrm{d}W_I(t).$$

假设 $\tau \in [0,T)$ 是个人的退休时间，并且以个人名义工资水平过程 $L_I = \{L_I(t), 0 \le t \le \tau\}$ 表示为：

$$\frac{\mathrm{d}L_I(t)}{L_I(t)} = \mu_L(\alpha(t))\mathrm{d}t + \sigma_I(\alpha(t))\mathrm{d}W_I(t) + \sigma_L(\alpha(t))\mathrm{d}W_L(t).$$

在个人退休后，假设个人名义养老金水平过程 $D_I = \{D_I(t), \tau \le t \le T\}$ 被描述为：

$$\frac{\mathrm{d}D_I(t)}{D_I(t)} = \mu_D(\alpha(t))\mathrm{d}t + \sigma_I(\alpha(t))\mathrm{d}W_I(t) + \sigma_D(\alpha(t))\mathrm{d}W_D(t).$$

在工作期间，个人有长期稳定的工作以获得工资收入，并向社会养老保险基金支付一定比例 k。在退休后，个人将获得养老金收入。因此，个人的名义财富过程满足以下 SDE：

$$\begin{aligned}\mathrm{d}X_I(t) = X_I(t)&\left[(1-u_1(t)-u_2(t))\frac{\mathrm{d}S_0(t)}{S_0(t)} + u_1(t)\frac{\mathrm{d}S_1(t)}{S_1(t)} + u_2(t)\frac{\mathrm{d}B(t)}{B(t)}\right]\\ &+(1-k)L_I(t)\mathrm{d}t - C_I(t)\mathrm{d}t, \quad \text{for } 0 \le t \le \tau,\end{aligned}$$

和

$$\begin{aligned}\mathrm{d}X_I(t) = X_I(t)&\left[(1-u_1(t)-u_2(t))\frac{\mathrm{d}S_0(t)}{S_0(t)} + u_1(t)\frac{\mathrm{d}S_1(t)}{S_1(t)} + u_2(t)\frac{\mathrm{d}B(t)}{B(t)}\right]\\ &+D_I(t)\mathrm{d}t - C_I(t)\mathrm{d}t, \quad \text{for } \tau \le t \le T.\end{aligned}$$

其中 $u_1(t)$ 和 $u_2(t)$ 分别表示个人名义财富投资于金融市场风险资产和通胀指数债券的金额比例，$1-u_1(t)-u_2(t)$ 代表投资于无风险资产的名义财富比例，$C_I(t)$ 代表名义实际消费量。

我们设

$$L(t) = \frac{L_I(t)}{I(t)}, D(t) = \frac{D_I(t)}{I(t)}, X(t) = \frac{X_I(t)}{I(t)}, C(t) = \frac{C_I(t)}{I(t)},$$

上述表示剔除通胀因素后的过程。因此，根据伊藤公式，我们有：对于 $0 \le t \le \tau$

$$\begin{aligned}\mathrm{d}X(t) = &\Big\{X(t)\Big[R(\alpha(t)) - \mu_I(\alpha(t)) + \sigma_I^2(\alpha(t)) + u_1(t)\big(\mu_S(\alpha(t)) - R(\alpha(t))\big)\\ &+u_2(t)\big(r(\alpha(t)) + \mu_I(\alpha(t)) - R(\alpha(t)) - \sigma_I^2(\alpha(t))\big)\Big] + (1-k)L(t) - C(t)\Big\}\mathrm{d}t\\ &+X(t)(u_2(t)-1)\sigma_I(\alpha(t))\mathrm{d}W_I(t) + X(t)u_1(t)\sigma_S(\alpha(t))\mathrm{d}W_S(t),\\ \mathrm{d}L(t) = &L(t)\Big[\mu_L(\alpha(t)) - \mu_I(\alpha(t))\Big]\mathrm{d}t + L(t)\sigma_L(\alpha(t))\mathrm{d}W_L(t),\end{aligned}$$

和对于 $\tau \le t \le T$

$$\begin{aligned}dX(t) = &\Big\{X(t)\Big[R(\alpha(t)) - \mu_I(\alpha(t)) + \sigma_I^2(\alpha(t)) + u_1(t)\big(\mu_S(\alpha(t)) - R(\alpha(t))\big)\\ &+u_2(t)\big(r(\alpha(t)) + \mu_I(\alpha(t)) - R(\alpha(t)) - \sigma_I^2(\alpha(t))\big)\Big] + D(t) - C(t)\Big\}\mathrm{d}t\\ &+X(t)(u_2(t)-1)\sigma_I(\alpha(t))\mathrm{d}W_I(t) + X(t)u_1(t)\sigma_S(\alpha(t))\mathrm{d}W_S(t),\\ \mathrm{d}D(t) = &D(t)\Big[\mu_D(\alpha(t)) - \mu_I(\alpha(t))\Big]\mathrm{d}t + D(t)\sigma_D(\alpha(t))\mathrm{d}W_D(t).\end{aligned}$$

其中 $\mu_S(\alpha(t))$、$\sigma_S(\alpha(t))$ 是风险资产的预期回报和波动性，$\mu_I(\alpha(t))$、$\sigma_I(\alpha(t))$ 是通货膨胀因子的预期水平和波动性，$\mu_L(\alpha(t))$ 和 $\sigma_L(\alpha(t))$ 是个人退休前工资的预期回报和波动性，$\mu_D(t)$ 和 $\sigma_D(t)$ 是个人退休后养老金的预期回报和波动性，$R(\alpha(t))$ 是名义无风险利率，$r(\alpha(t))$ 为实际利率。

2.2. 优化目标

习惯消费过程定义如下：

$$\mathrm{d}H(t)=\left[\psi(t)C(t)-\eta(t)H(t)\right]\mathrm{d}t,\mathrm{H}(0)=h,t\in[s,T],$$

其中$\psi(t)$、$\eta(t)$是习惯消费参数。

参照文献[9]，我们建立了个体的 S 型效用函数，只有实际消费和习惯性消费之间的差异才能产生效用。此外，我们假设实际消费低于习惯水平是允许的。因此，我们需要一个定义良好的效用函数来综合表示正效用和负效用，我们自然地选择 S 型效用函数，

$$u(c(t))=u(C(t)-H(t))=\frac{(C(t)-H(t))^{1-\gamma}}{1-\gamma}1_{\{C(t)\geq H(t)\}}+(-\kappa)\frac{(H(t)-C(t))^{1-\gamma}}{1-\gamma}1_{\{C(t)<H(t)\}},$$

对于任意$t\in[s,T]$，定义$c(t)=C(t)-H(t)$表示为超额消费。其中$\gamma(0<\gamma<1)$是指个人的相对风险规避参数，κ为损失厌恶参数，$C(t)$为实际消费，$H(t)$为习惯消费。

我们从可容许控制集Π中找到最优资产配置和消费使得在有限时间内投资者的超额消费的整体效用最大化，也就是说找到最优资产配置和消费策略$\pi=\{\pi(t)=(u_1(t),u_2(t),C(t)),s\leq t\leq T\}$使下述优化准则达到最优：

$$J(s,x,l,d,h,i,\pi)=E_{s,x,l,d,h,i}\left[\int_s^T \mathrm{e}^{-\rho t}u(C(t)-H(t))\mathrm{d}t\right],t\in[s,T].$$

其中，$\rho>0$表示折扣系数。接着我们将与随机优化问题对应的值函数定义为

$$V(s,x,l,d,h,i)=\sup_{\pi\in\Pi}J(s,x,l,d,h,i,\pi).$$

2.3. HJB 方程

对于一个固定的退休时间τ，我们将详细讨论个人退休前的最优控制问题，而个人退休后的优化问题可以类似地进行分析。为了方便阐述，我们对模型进行了改写。对于$0\leq t\leq\tau$，

$$\mathrm{d}Z(t)=f(Z(t),\alpha(t),\pi(t))\mathrm{d}t+\sigma(Z(t),\alpha(t),\pi(t))\mathrm{d}W(t).$$

其中，

$$Z(t)=(X(t),L(t),H(t))^{\mathrm{T}},$$

$$W(t)=(W_I(t),W_S(t),W_L(t))^{\mathrm{T}},$$

$$f(Z(t),\alpha(t),\pi(t))=(f_X,f_L,f_H)^{\mathrm{T}},$$

$$\sigma(Z(t),\alpha(t),\pi(t))=\begin{bmatrix}\sigma_{X_1} & \sigma_{X_2} & 0\\ 0 & 0 & \sigma_{L_1}\\ 0 & 0 & 0\end{bmatrix},$$

$$\begin{aligned}f_X=&X(t)\Big[R(\alpha(t))-\mu_I(\alpha(t))+\sigma_I^2(\alpha(t))+u_1(t)(\mu_S(\alpha(t))-R(\alpha(t)))\\&+u_2(t)(r(\alpha(t))+\mu_I(\alpha(t))-R(\alpha(t))-\sigma_I^2(\alpha(t)))\Big]+(1-k)L(t)-C(t),\end{aligned}$$

$$
\begin{aligned}
f_L &= L(t)\left[\mu_L(\alpha(t)) - \mu_I(\alpha(t))\right],\\
f_H &= \psi(t)C(t) - \eta(t)H(t),\\
\sigma_{X_1} &= X(t)\left(u_2(t)-1\right)\sigma_I(\alpha(t)),\\
\sigma_{X_2} &= X(t)u_1(t)\sigma_S(\alpha(t)),\\
\sigma_{L_1} &= L(t)\sigma_L(\alpha(t)).
\end{aligned}
$$

如果价值函数 $V^{\pi}(s,z,i)$ 足够光滑，它可以被描述为通过动态规划原理的 Hamilton-Jacobi-Bellman (HJB)方程的解。相应的 HJB 方程如下：

$$
\begin{aligned}
0 = V_s^{\pi} &+ \sup_{\pi\in U}\left\{V_h^{\pi} f_h + V_x^{\pi} f_x + V_{xx}^{\pi} f_{xx} + V_l^{\pi} f_l + V_{ll}^{\pi} f_{ll}\right.\\
&\left.+\sum_{j\neq i} q_{ij}\left[V^{\pi}(s,z,j) - V^{\pi}(s,z,i)\right] - \rho V^{\pi}(s,z,i) + U(c)\right\},
\end{aligned}
$$

其中 $z\in R^3, i\in\mathcal{M}, s\in[0,\tau)$。然而，上述 HJB 方程的显式解不容易获得，因此我们采用数值求解方法。

3. 数值算法

基于强化学习中的演员－评论家(Actor-Critic)框架，可以同时求解价值函数和策略函数。该框架包括两部分：第一部分是用于评论家的策略评估(Policy Evaluation, PE)，PE 能够求解价值函数，另一部分是用于演员的策略迭代(Policy Iteration, PI)，PI 能够求解策略函数。目前，在 Actor-Critic 框架下已经开发了许多算法，例如文献[11] [13]。特别是，文献[12]设计了一种 TD 算法来学习 PE 部分的值函数，使用策略梯度下降算法来更新 PI 部分的控制函数，从而提高其收敛性，提高样本利用率，避免局部最优，并平衡开发和利用的程度。

3.1. 策略评估

在本节中，我们开发相应的 PE 程序，并使用函数逼近方法来获得价值函数的估计。对于连续最优控制问题的 TD，我们应用 Itô's 公式求解 $\mathrm{e}^{-\rho(t-s)}V^{\pi}(s,z,i)$，得到

$$
\begin{aligned}
V^{\pi}(s,z,i) = &\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}U(c(t))\mathrm{d}t - \int_s^{\tau}\mathrm{e}^{-\rho(t-s)}V_x^{\pi}\sigma_{x_1}\mathrm{d}W_I(t)\\
&-\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}V_x^{\pi}\sigma_{x_2}\mathrm{d}W_S(t) - \int_s^{\tau}\mathrm{e}^{-\rho(t-s)}V_l^{\pi}\sigma_{l_1}\mathrm{d}W_L(t) + \mathrm{e}^{-\rho(\tau-s)}V^{\pi}(\tau,z,i).
\end{aligned}
$$

现在让我们将连续设置中的 TD 误差定义为：

$$
\begin{aligned}
TD^{\pi} = &\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}U(c(t))\mathrm{d}t - \int_s^{\tau}\mathrm{e}^{-\rho(t-s)}V_x\sigma_{x_1}\mathrm{d}W_I(t)\\
&-\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}V_x\sigma_{x_2}\mathrm{d}W_S(t) - \int_s^{\tau}\mathrm{e}^{-\rho(t-s)}V_l\sigma_{l_1}\mathrm{d}W_L(t) + \mathrm{e}^{-\rho(\tau-s)}V(\tau,z,i) - V(s,z,i).
\end{aligned}
$$

为了对高维价值函数 V 和控制 π 进行参数化，我们使用神经网络来拟合它们，价值函数 V 通过 $V(\cdot,\cdot,\cdot;\theta_V) = F_V(\cdot,\cdot,\cdot;\theta_V)$ 参数化，控制 π 通过 $\pi(\cdot,\cdot,\cdot;\theta_{\pi}) = F_{\pi}(\cdot,\cdot,\cdot;\theta_{\pi})$ 参数化，其参数分别表示为 θ_V 和 θ_{π}。

此外，我们计划使用另一个神经网络 $G(\cdot,\cdot,\cdot;\theta_G)$ 来表示 V，即 $V(\cdot,\cdot,\cdot;\theta_G) = G(\cdot,\cdot,\cdot;\theta_G)$，这将用于计算 V 的梯度，可以使用有限差分算法来拟合梯度。

进一步介绍我们的神经网络结构。具有 $\mathcal{Q}$ 个隐藏层的神经网络 $F(\cdot,\cdot,\cdot;\varphi)$ 可以设置为

$$
F(\cdot,\cdot,\cdot;\varphi) = \varphi_{\mathcal{Q}}\circ\sigma_{\mathcal{Q}}\circ\varphi_{\mathcal{Q}-1}\circ\sigma_{\mathcal{Q}-1}\circ\cdots\circ\varphi_1\circ\sigma_1\circ\varphi_0(\cdot,\cdot,\cdot),
$$

其中 $\varphi = \{\varphi_i, i=1,\cdots,\mathcal{Q}\}$，$\varphi_i = \{w_i, b_i\}$ 是关于隐藏层输入和输出维度的适当维度的线性变换，而 σ_i 是作用于相邻隐藏层的激活函数：$\sigma_i(\cdot,\cdot,\cdot) = \mathrm{ReLU}(\cdot,\cdot,\cdot)$。

我们进一步为评论家定义以下损失函数：

$$L(\theta_V,\theta_G)=E_{s,z,i}\left[\frac{1}{2}\left(TD^{\pi}\right)^2\right]=E_{s,z,i}\left[\frac{1}{2}\left(\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}U\left(c(t)\right)\mathrm{d}t\right.\right.$$
$$-\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}\nabla_x V\left(t,z,i;\theta_G\right)\sigma_{x_1}\mathrm{d}W_I(t)-\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}\nabla_x V\left(t,z,i;\theta_G\right)\sigma_{x_2}\mathrm{d}W_S(t)$$
$$\left.\left.-\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}\nabla_l V\left(t,z,i;\theta_G\right)\sigma_{l_1}\mathrm{d}W_L(t)+\mathrm{e}^{-\rho(\tau-s)}V\left(\tau,z,i;\theta_V\right)-V\left(s,z,i;\theta_V\right)\right)^2\right].$$

我们考虑使用随机梯度下降法来最小化上述损失函数，以获得价值函数的最佳近似。损失函数的梯度近似如下：

$$\nabla_V L(\theta_V,\theta_G)\approx\nabla_V\left[\frac{1}{2}\left(TD^{\pi}\right)^2\right]=TD^{\pi}\nabla_V TD^{\pi}.$$

然后使用 Adam 优化器来更新值函数的参数 $\theta_{value}=(\theta_V,\theta_G)$：

$$\theta_{\text{value}}\to\theta_{\text{value}}+\alpha\nabla_V L.$$

其中，α 为合适的学习率。

3.2. 策略迭代

在本小节中，我们介绍演员部分，并使用策略梯度来改进演员的策略。回忆我们的最优控制问题并利用动态规划原理，我们可以使用以下目标函数来定义演员(Actor)的目标，我们有

$$J(s,z,i;\pi)=E_{s,z,i,\pi}\left[\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}U\left(c(t)\right)\mathrm{d}t+\mathrm{e}^{-\rho(\tau-s)}V(\tau,z,i)\right].$$

在数值算法中，我们将控制 π 参数化为神经网络 $\pi(\cdot;\theta_\pi)$，其中 θ_π 称为参数。使用对 $J(\cdot;\pi)$ 梯度的随机近似(即导数 $\tilde{\nabla}J\approx\nabla_{\theta_\pi}J$)来优化参数，因此参数更新为

$$\theta_\pi\to\theta_\pi+\alpha\tilde{\nabla}J.$$

其中，α 为合适的学习率。我们将功能导数近似为

$$\tilde{\nabla}J=E_{s,z,i,\pi}\left[\int_s^{\tau}\mathrm{e}^{-\rho(t-s)}\pi_t^{-\gamma}\frac{\delta\pi_t}{\delta\theta_\pi}\mathrm{d}t+\mathrm{e}^{-\rho(\tau-s)}\frac{\delta V}{\delta z}\frac{\delta z_\tau}{\delta\theta_\pi}\right].$$

4. 数值结果

在本节中，我们为数值实验选择了适当的参数并获得了一些结果。对于金融市场，金融市场的无风险利率为 $R=0.01$，对冲债券的预期回报率为 $r=0.02$，通胀因子的预期水平为 $\mu_I=0.01$，通胀的波动率为 $\sigma_I=0.2$。对于切换模型，转移率矩阵为 $Q=\begin{pmatrix}-2 & 2\\ 1 & -1\end{pmatrix}$。风险资产的预期回报率为 $\mu_S(1)=0.04$，$\mu_S(2)=0.08$，风险资产的波动率为 $\sigma_S(1)=0.1$，$\sigma_S(2)=0.2$。对于个人工资过程，工资增长的预期速率为 $\mu_L=0.4\times t+0.05$，工资过程的波动率为 $\sigma_L=0.02$，养老金增长的预期速率为 $\mu_D=0.1\times t+0.05$，养老金过程的波动率为 $\sigma_D=0.02$。

不失一般性，我们假设初始时间为 $s=0$，退休时间为 $\tau=1$，最大生存时间为 $T=2$，退休前个人养老保险的缴费率为 $k=0.2$，退休前习惯性消费的预期增长率为 $\psi=\eta=0.1\times t+0.05$，退休后习惯性消费的预期增长率为 $m\psi=m\eta$，习惯性消费的敏感度为 $m=0.6$，损失厌恶参数为 $\kappa=2.25$，相对风险厌恶系数为 $\beta=0.6$，贴现因子为 $\rho=0.5$，工资的比例为 $x_1=x_2=0.8$。假设这些初始值为 $h_0=12$，$L_0=20$，$D_0=16$，

$X_0 = 200$。

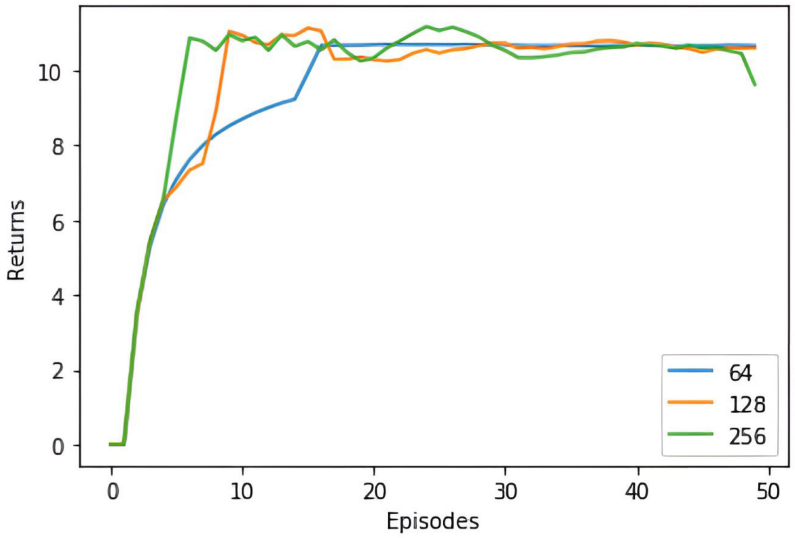


Figure 1. Selection of the number of neurons
图 1. 神经元个数的选择

在神经网络的架构方面，如图 1，隐藏层神经元个数不宜过大，选择 128 个能减少数据量，同时也有不错的收敛性，策略网络中的隐藏层数量为 2，价值网络中的隐藏层数量为 3，学习率设置为 0.01。我们还使用了经验回放技术，就是每次抽取一定数量的样本进行训练，可以重复抽取，达到有效利用样本的效果，经验证，将每次批量大小设置为 128 为宜，并将基本步长 Δt 设置为 0.001。

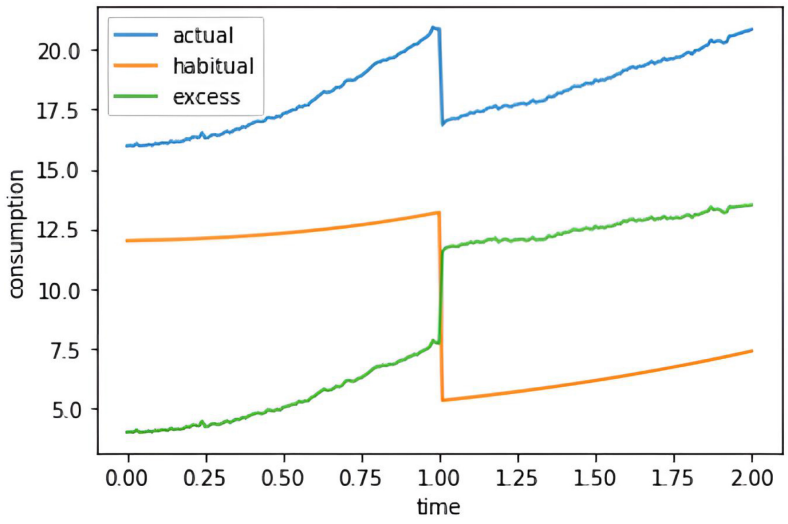
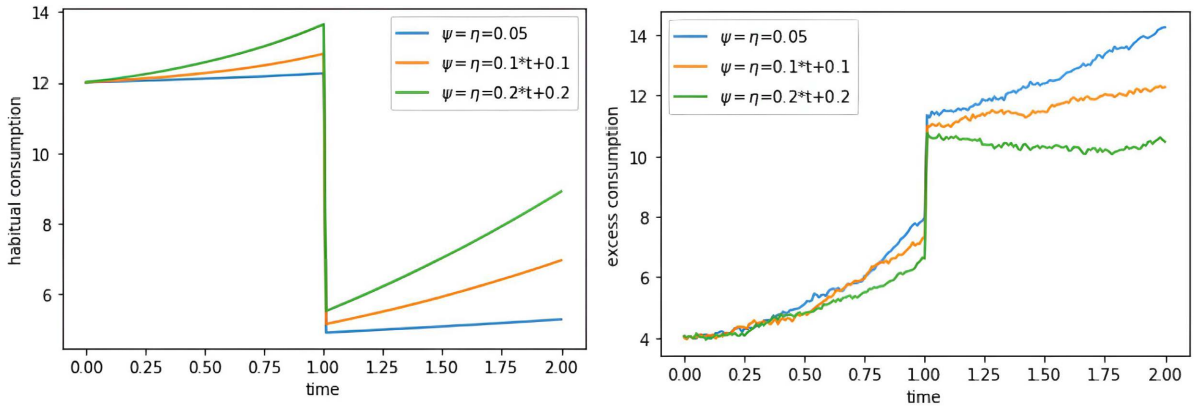


Figure 2. Optimal actual consumption $C(t)$, habitual consumption $H(t)$, excess consumption $c(t)$
图 2. 最优实际消费 $C(t)$，习惯消费 $H(t)$，超额消费 $c(t)$

在图 2 中，我们可以看到我们可以首先，财富随着时间增加，而较高的财富会导致较高的个人消费，从而导致实际消费随时间增加。其次，习惯性消费是个体过去实际消费的加权平均值，因此习惯性消费会随着实际消费的增长而增加。此外，个人消费受到财富的影响，而在退休后，他们的工资收入减少，

导致实际消费迅速下降。最后，由于个体在退休后减少了许多不必要的开支，即习惯性水平和敏感度的降低，导致超额消费增加。

总之，在退休后，尽管个人的实际消费有所下降，但超额消费却有所增加，这表明个体在退休后仍然对较低的实际消费水平感到满意。



(a) (b)

Figure 3. Impacts of the parameters ψ and η on the habitual consumption and the optimal excess consumption
图 3. 不同习惯参数 ψ 和 η 对习惯消费和最优超额消费的影响

在图 3 中，展示了习惯性消费和超额消费随习惯性参数 ψ 和 η 的演变。我们可以看到，习惯性消费随着习惯性参数的增加而增加，而超额消费则随着习惯性参数的增加而减少。这是因为当个体工资水平保持不变时，实际消费也保持不变，而习惯性参数的增加会导致习惯性消费水平的上升，从而导致个体超额消费水平的下降。

总之，习惯性参数 ψ 和 η 影响个体习惯水平增长的幅度。如果个体在退休前的习惯性增长速度较快，那么他们在退休后也会经历过快的习惯性增长速度。如果实际消费保持不变，这将不可避免地导致超额消费水平较低，个体可能会感到不满意。因此，最好让个体保持适度的习惯性增长速度。

参考文献

[1] Merton, R.C. (1969) Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case. *The Review of Economics and Statistics*, **51**, 247-257. https://doi.org/10.2307/1926560

[2] Ferreira, M., Pinheiro, D. and Pinheiro, S. (2023) Optimal Consumption, Investment and Life Insurance Selection under Robust Utilities. *International Journal of Financial Engineering*, **10**, Article ID: 2350016. https://doi.org/10.1142/s2424786323500160

[3] Tao, C., Rong, X. and Zhao, H. (2023) Stochastic Control with Inhomogeneous Regime Switching: Application to Consumption and Investment with Unemployment and Reemployment. *Journal of Mathematical Economics*, **107**, Article ID: 102849. https://doi.org/10.1016/j.jmateco.2023.102849

[4] Wang, H., Wang, N., Xu, L., Hu, S. and Yan, X. (2022) Household Investment-Consumption-Insurance Policies under the Age-Dependent Risk Preferences. *International Journal of Control*, **96**, 2542-2554. https://doi.org/10.1080/00207179.2022.2100278

[5] Pollak, R.A. (1970) Habit Formation and Dynamic Demand Functions. *Journal of Political Economy*, **78**, 745-763. https://doi.org/10.1086/259667

[6] Ryder, H.E. and Heal, G.M. (1973) Optimal Growth with Intertemporally Dependent Preferences. *The Review of Economic Studies*, **40**, 1-31. https://doi.org/10.2307/2296736

[7] Curatola, G. (2017) Optimal Portfolio Choice with Loss Aversion over Consumption. *The Quarterly Review of Economics and Finance*, **66**, 345-358. https://doi.org/10.1016/j.qref.2017.04.003

[8] van Bilsen, S., Laeven, R.J.A. and Nijman, T.E. (2020) Consumption and Portfolio Choice under Loss Aversion and

Endogenous Updating of the Reference Level. *Management Science*, **66**, 3927-3955. https://doi.org/10.1287/mnsc.2019.3393

[9] He, L., Liang, Z., Song, Y. and Ye, Q. (2022) Optimal Asset Allocation, Consumption and Retirement Time with the Variation in Habitual Persistence. *Insurance*: *Mathematics and Economics*, **102**, 188-202. https://doi.org/10.1016/j.insmatheco.2021.10.004

[10] Wang, H., Zariphopoulou, T. and Zhou, X.Y. (2020) Reinforcement Learning in Continuous Time and Space: A Stochastic Control Approach. *Journal of Machine Learning Research*, **21**, 1-34.

[11] Jia, Y. and Zhou, X. (2021) Policy Gradient and Actor-Critic Learning in Continuous Time and Space: Theory and Algorithms. *Journal of Machine Learning Research*, **23**, 1-50.

[12] Zhou, M., Han, J. and Lu, J. (2021) Actor-Critic Method for High Dimensional Static Hamilton-Jacobi-Bellman Partial Differential Equations Based on Neural Networks. *SIAM Journal on Scientific Computing*, **43**, A4043-A4066. https://doi.org/10.1137/21m1402303

[13] Wang, Z., Bapst, V., Heess, N., *et al*. (2016) Sample Efficient Actor-Critic with Experience Replay. arXiv: 1611.01224.