Operations Research and Fuzziology 运筹与模糊学, 2023, 13(5), 5177-5194
Published Online October 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.135520

Hans汉斯

基于生存函数的连续时间强化学习问题

张　娅

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2023年8月9日；录用日期：2023年10月11日；发布日期：2023年10月20日

摘　要

在智能体具有时间不一致性偏好的条件下，用生存分析中的生存函数和风险率刻画折现函数，使用跳跃－扩散过程对强化学习中的环境状态建模，探讨跳跃过程对值函数和偏好逆转时间的影响。研究表明：1) 跳跃－扩散过程下值函数与时间高度相关，且发生偏好逆转的时间会提前。2) 跳跃幅度服从同一分布时，随着泊松过程的强度增大，值函数的值也会越大，发生偏好逆转的时间会越早。3) 在同一泊松过程强度下，当跳跃幅度服从正态分布时，值函数的值最大，发生偏好逆转的时间最早。这些结论为决策者预测识别潜在的突发事件，从而采取相应的预防措施提供重要参考。

关键词

强化学习，跳跃－扩散过程，时间不一致性，HJB方程

Continuous-Time Reinforcement Learning Problem Based on Survival Functions

Ya Zhang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Aug. 9th, 2023; accepted: Oct. 11th, 2023; published: Oct. 20th, 2023

Abstract

Under the condition of agents having time-inconsistent preferences, this study characterizes the discount function using survival functions and hazard rates from survival analysis. It models the environmental state in reinforcement learning using jump-diffusion processes and explores the impact of jump processes on value functions and the time of preference reversal. The study indi-

文章引用: 张娅. 基于生存函数的连续时间强化学习问题[J]. 运筹与模糊学, 2023, 13(5): 5177-5194.
DOI: 10.12677/orf.2023.135520

cates that: 1) Under the jump-diffusion process, value functions are highly correlated with time, and the time of preference reversal is advanced. 2) When jump amplitudes follow the same distribution, with an increase in the intensity of the Poisson process, the value of the value function also increases, and the time of preference reversal occurs earlier. 3) Under the same intensity of the Poisson process, when jump amplitudes follow a normal distribution, the value of the value function is maximized, and the time of preference reversal is earliest. These findings provide important references for decision-makers to predict and identify potential sudden events and take corresponding preventive measures.

Keywords

Reinforcement Learning, Jump-Diffusion Process, Time-Inconsistency, Hamilton-Jacobi-Bellman Equation

1. 引言

强化学习方法起源于动物心理学的相关原理，模仿人类和动物学习的试错机制，是一种通过与环境交互，以获得最大累积期望回报的方法。强化学习通常使用指数函数对未来的奖励进行折现，以实现理论上的收敛。然而由于强化学习中奖励的延迟机制，通常会导致时间不一致性。来自神经科学、心理学和经济学的研究也表明，双曲折现模型可以更好地捕捉人类和动物的时间偏好。

作为双曲时间偏好的一个例子，考虑以下情境：假设一个陌生人向你提出一个交易。他立即给你 100 美元，没有风险，但如果你可以等到明天，他承诺给你 110 美元。由于没有进一步的信息。许多人对这位潜在的捐助者持怀疑态度，选择立即获得 100 美元，因为未来的承诺会有风险。然而，在另一个交易中，他承诺 365 天内给你 100 美元，366 天内给你 110 美元。在这个交易中，许多人选择 110 美元。实际上贴现率已经进一步下降了，这表明人们相信，如果在 365 天没有违背承诺，那么在 366 天不太可能被违背[1]。

指数折现在这些选择之间总是保持一致，而双曲折现可以显示出时间偏好逆转。这种时间偏好之间的差异可能被认为是不合理的。然而，这种行为在数学上与智能体对环境中的风险率保持一定不确定性是一致的，奖励根据主体可能会遭受风险而被折现的可能性来进行计算，因此主体可能无法幸存以获得奖励[2]。强化学习环境也具有风险的特征，同时风险可能会随着值函数和策略的变化而变化。

纵观该领域的研究，已有不少研究成果。Alexander 等[3]定义了一个学习算法，双曲贴现时间差分学习，它构成了双曲模型的递归公式。Alia 等[4]研究了具有随机跳跃的随机微分方程驱动的时间不一致随机控制模型，在博弈论框架中表述时间不一致的问题，并寻找在时间不一致性下的时间一致的纳什均衡解。Schultheis 等[5]研究了非指数折现下的强化学习，将环境的状态过程建模为扩散过程，并提出了一种任意折现函数下的强化学习方法。Nafi 等[6]研究了双曲折现对泛化任务的影响，并提出了强化学习中的泛化双曲贴现。Ali 等[7]论述了非指数折现函数对强化学习中智能体学习的影响，并研究其对多智能体系统和泛化任务的影响，同时指出使用非指数折现对强化学习的必要性。Kwiatkowski 等[8]提出广义优势估计，允许以任意折扣计算优势值直接应用于现代策略行为批评算法。这些研究结果都表明，利用强化学习进行决策优化时，不能简单认为智能体是时间一致性的。

本文的主要贡献和创新体现为以下两点：第一，从研究方法上看，使用生存分析中的生存函数和风险率刻画折现函数，这种方法可以表示一般的折现函数，包括指数折现和非指数折现两种情况。从研究内容上看，使用跳跃－扩散过程对强化学习中的环境状态建模，研究跳跃－扩散过程下值函数和发生偏好逆转时间的变化。通常，使用跳跃－扩散过程建模具有不连续变化的变量，比如突发事件的冲击，这为决策者预测识别潜在的突发事件，从而采取相应的预防措施，也为开发出更具稳定性和适用性的算法提供有益的参考。

2. 理论分析

生存分析中涉及两个重要的概念生存函数和风险率。使用生存函数表示折现函数，风险率则表示未来的奖励无法实现的风险[2]。当假设风险率 $\alpha(t)=\lambda$ 为常数时，主体以指数函数的方式进行折。当假设风险率是一个恒定但未知的量 λ 时，其概率分布假设为 $p(\lambda)=Gamma(\lambda;\alpha_0,\beta_0)$，就可以得到关于预期生存函数的双曲形式。使用生存函数和风险率可以准确刻画指数折现和双曲折现的情况，能满足智能体具有时间不一致性的要求。

强化学习是在马尔科夫决策过程框架下进行的，跳跃－扩散过程是马尔科夫决策过程。使用跳跃－扩散过程建模强化学习中的环境状态是合理的。

3. 模型构建

3.1. 环境模型

考虑一个具有跳跃的状态空间 R^n 和有限的行动集 U 的随机控制系统[9]，使用生存函数 $S(t)$ 和基于时间依赖的风险率 $\alpha(t)$ 刻画折现函数。这样生成的值函数是依赖于时间的。在强化学习中，将环境状态建模为由泊松过程驱动的具有跳跃的几何布朗运动，环境状态模型可表示如下：

$$\mathrm{d}X(t)=f(X(t),u(t),t)\mathrm{d}t+G(X(t),u(t),t)\mathrm{d}W(t)+h(X(t),u(t),t)\mathrm{d}P(t) \tag{1}$$

其中，$t\in R_0^+$，$f:X\times U\to X$ 是漂移函数，$G:X\times U\to X\times R^m$ 是方差矩阵，$W(t)$ 是 n 维布朗运动，$h:X\times U\to R^n$ 是泊松过程跳跃的幅度，$P(t)$ 是泊松过程。

强化学习的目标是最大化长期奖励，因此将总的期望折现奖励作为智能体在执行策略时的预期的累积奖励：

$$J\left(u_{[0,\infty)}\right)=E\left[\int_0^{\infty}S(\tau)R(X(\tau),u(\tau),\tau)\mathrm{d}\tau\right] \tag{2}$$

其中，将预期回报定义为在某个状态或者状态－动作对下，根据当前策略和未来奖励的期望累积回报。值函数度量了智能体在达到某个状态或状态－动作对后，预期能够获得的累积奖励。智能体在特定状态或状态－动作对下采取特定策略时的长期回报期望可表示如下：

$$V^*(x,t)=\max_{u[t,\infty)}E\left[\int_t^{\infty}\frac{S(\tau)}{S(t)}R(X(\tau),u(\tau),\tau)\middle|X(t)=x\right] \tag{3}$$

其中，$S(\tau)/S(t)$ 表示在时间 τ 之后存活到时间 t 的概率，假设个体已经在时间 t 存活。等式(3)中的值函数和最优策略与时间相关，这表明时间在决策和优化过程中时间需要被考虑。

3.2. 折现函数

生存分析理论涉及两个基本概念，即生存函数和风险率。在生存分析中，对事件发生前的持续时间

非常感兴趣[10]。考虑单个事件时，可以将其持续时间描述为一个连续随机变量 T，事件持续时间的累积分布函数可以表示为 $F(t)=P(T\le t)$，其中 $t\in R_0^+$ 表示持续的时间。同时，使用概率密度函数 $f(x)$ 描述该持续时间的分布特征。一般情况下，$F(t)$ 作为失效函数已知，并且定义生存函数为 $S(t)=1-F(t)=P(T>t)$。生存函数单调递减，并且满足 $S(0)=1$ 和 $\lim_{t\to\infty}S(t)=0$。

通过条件概率准则，可以得到 $t_1>t_0$，$P(T>t_1|T>t_0)=S(t_1)/S(t_0)$。这表明，在给定事件已经持续 t_0 的情况下，事件持续到 t_1 的概率等于这两个时刻的生存函数的比值。

风险率 $\alpha(t)$ 定义为在给定 $T\ge t$ 的条件下，单位时间内事件发生的概率。具体而言，$\alpha(t)=\lim_{\Delta t\to 0}\frac{1}{\Delta t}P(t\le T<t+\Delta t|T\ge t)$，风险率和生存函数之间的关系为：

$$\begin{aligned}\alpha(t)&=\lim_{\Delta t\to 0}\frac{1}{\Delta t}\frac{S(t)-S(t+\Delta t)}{S(t)}=-\frac{S'(t)}{S(t)}\\ S(t)&=\exp\left(-\int_0^t\alpha(\tau)\mathrm{d}\tau\right)\end{aligned}\tag{4}$$

考虑智能体获得单一奖励并对环境表现出一种时间偏好，将奖励值 r 建模成一个函数 $L:R\times R_0^+\to R$，其中 $L(r,t)=S(t)\cdot r$，$S(t)$ 是随时间递减的折现函数，$t\in R_0^+$ 将 $S(t)$ 视为生存函数，存在一个风险率 $\alpha(t)$，表示未来的奖励无法实现的风险[2]。当假设风险率 $\alpha(t)=\lambda$ 为常数时，智能体以指数函数的方式进行折现，即 $S(t)=\exp(-\lambda t)$。当假设风险率是一个恒定但未知的量时，其信念 λ 的概率分布假设为 $p(\lambda)=Gamma(\lambda;\alpha_0,\beta_0)$，就可以得到关于预期生存函数的双曲形式：

$$S(t;\alpha_0,\beta_0)=\int_\lambda\exp(-\lambda t)\,p(\lambda)\mathrm{d}\lambda=\frac{1}{\left(\frac{t}{\beta_0}+1\right)^{\alpha_0}}\tag{5}$$

使用贝叶斯法则推导出关于λ的后验信念，表示为 $p(\lambda|t)=Gamma(\lambda;\alpha_0,\beta_0+t)$，根据后验分布，预期的风险率 $\alpha(t)$ 由后验均值计算得出：

$$\alpha(t)=\int_\lambda\lambda p(\lambda|t)\mathrm{d}\lambda=\frac{\alpha_0}{\beta_0+t}\tag{6}$$

证明 见附录。

3.3. 基本假设

假设 1：第 2 节中定义的随机过程具有强唯一解的性质，即漂移函数 f 和方差矩阵 G 以及跳跃幅度 h 是线性增长的，并且在相同变量上是 Lipschitz 连续[11]。即存在常数 $K>0$，对所有的 $(x,y)\in R^n\times R^n$ 和 $(u,t)\in U\times[0,T]$，有

$$|f(x,u,t)-f(y,u,t)|^2+|G(x,u,t)-G(y,u,t)|^2+|h(x,u,t)-h(y,u,t)|\le K|x-y|^2$$

$$|f(0,u,t)|^2+|G(0,u,t)|^2+|h(0,u,t)|^2\le K(1+|u|^2)$$

假设 2：值函数通常不够平滑，以满足“经典”意义上的解。因此，考虑使用粘度解，它在适当的广义意义上满足 HJB 方程。解存在的一个充分条件是 f 和 G 有界，关于 x 和 t 具有有界连续的一阶导数。函数 R 和 S 是多项式增长的[12]。

即存在常数 $K>0$，有对所有的 $x\in R^n$ 和 $(u,t)\in U\times[0,T]$，有

$$|R(u,t)| \le K\left(1+|u|^2\right)$$
$$|S(x)| \le K\left(1+|x|^2\right)$$

假设 3：等式(2)中的积分是收敛的，对具有双曲折现函数的情况时，以下定理成立：

定理 1：对于等式(5)中的双曲折现函数，如果奖励函数 $R(x,u,t)$ 对于所有的 $(x,u,t) \in X \times U \times R_0^+$ 有上界，并且 $\alpha_0 > 1$，那么方程(3)中的值函数是符合定义的。如果奖励函数 $R(x,u,t)$ 对于所有的 $(x,u,t) \in X \times U \times R_0^+$ 有下界，并且 $\alpha_0 \le 1$，那么方程(3)中的值函数不符合定义。

证明 见附录。

3.4. HJB 方程的推导

这里简要概述一般折现函数的 HJB 方程的推导。首先，将等式中的积分分割成两项，从而得到值函数的递归形式：

$$V^*(x,t) = \max_{u[t,t+\Delta t]} E\left[\int_t^{t+\Delta t} \frac{S(\tau)}{S(t)} R\left(X(\tau),u(\tau),\tau\right)\mathrm{d}\tau + \frac{S(t+\Delta t)}{S(t)} V^*\left(X(t+\Delta t),t+\Delta t\right) \middle| X(t)=x\right] \tag{7}$$

对于第二项，利用泰勒展开和伊藤引理进行推导，得到：

$$\begin{aligned} V^*\left(X(t+\Delta t),t+\Delta t\right) = {} & V^*\left(X(t),t\right) + \int_t^{t+\Delta t} V_x^*\left(X(\tau),\tau\right) f\left(X(\tau),u(\tau),\tau\right)\mathrm{d}\tau \\ & + \int_t^{t+\Delta t} V_x^*\left(X(\tau),\tau\right) G\left(X(\tau),u(\tau),\tau\right)\mathrm{d}W(\tau) + \int_t^{t+\Delta t} V_t^*\left(X(\tau),\tau\right)\mathrm{d}\tau \\ & + \int_t^{t+\Delta t} \frac{1}{2} tr\left\{V_{xx}^*\left(X(\tau),\tau\right) G\left(X(\tau),u(\tau),\tau\right) G\left(X(\tau),u(\tau),\tau\right)^T\right\}\mathrm{d}\tau \\ & + \int_t^{t+\Delta t} V^*\left(X(\tau)+h\left(X(\tau),u(\tau),\tau\right),t\right) - V^*\left(X(\tau),\tau\right)\mathrm{d}P(t) + o(\Delta t) \end{aligned} \tag{8}$$

将等式(8)代入等式(7)，两边同时除以 Δt，并取极限 $\Delta t \to 0$，计算关于 $W(t)$ 的期望，可以得到所需的 HJB 方程：

$$\begin{aligned} \alpha(t) V^*(x,t) = \max_{u[t,t+\Delta t]} \Big\{ & R(x,u,t) + V_t^*(x,t) + V_x^*(x,t) f(x,u,t) \\ & + \frac{1}{2} tr\left(V_{xx}^*(x,t) G(x,u,t) G(x,u,t)^T\right) \\ & + \sum_{k=1}^m \lambda_k(t)\left[V^*\left(x+h_k(x,u,t),t\right) - V^*(x,t)\right]\Big\} \end{aligned} \tag{9}$$

$\alpha(t)$ 代表生存函数对应的风险率，将 HJB 方程等式(9)的右边关于动作没有最大化时定义为：

$$\begin{aligned} Q(x,u,t) = {} & R(x,u,t) + V_t^*(x,t) + V_x^*(x,t) f(x,u,t) \\ & + \frac{1}{2} tr\left(V_{xx}^*(x,t) G(x,u,t) G(x,u,t)^T\right) \\ & + \sum_{k=1}^m \lambda_k(t)\left[V^*\left(x+h_k(x,u,t),t\right) - V^*(x,t)\right] \end{aligned}$$

因此所求的值函数表示为 $V^*(x,t)$，最优策略表示为 $\pi^*(x,t) = argmax_u Q(x,u,t)$。

推导 见附录。

3.5. 求解 HJB 方程

基于配点法[13] [14] [15] [16]求解等式(9)，首先将其重新表述为：

$$E(V,x,t)=-\alpha(t)V(x,t)+\max_u\{R(x,u,t)+V_t(x,t)+V_x(x,t)f(x,u,t)+\frac{1}{2}tr\left(V_{xx}(x,t)G(x,u,t)G(x,u,t)^T\right)+\sum_{k=1}^m\lambda_k(t)\left[V\left(x+h_k(x,u,t),t\right)-V(x,t)\right]\} \tag{10}$$

使用值函数$V^\psi(x,t)$近似最优值函数$V^*(x,t)$，其中参数ψ通过对状态$\hat{x}_i$和时间$\hat{t}_i$随机采样，关于ψ最小化$\sum_i E\left(V^\psi,\hat{x}_i,\hat{t}_i\right)^2$获得。选择神经网络作为函数逼近器，偏导数$V_x^\psi(x,t)$，$V_t^\psi(x,t)$，$V_{xx}^\psi(x,t)$可以通过自动微分直接计算得到。由于$t$不是有界的，$t$需要被重新参数化，因此将所有的$t$映射到区间$[0,1)$。具体的做法如下：

假设一个有界的状态空间$X\in R$，从这个状态空间随机均匀地采样$\hat{x}_i$，时间点$\hat{t}_i\in R_0^+$从指数分布中采样，让$\hat{y}_i\sim U(0,1)$，然后计算$\hat{t}_i=-\log(1-\hat{y}_i)/\lambda$，为了将一个标准化的时间值输入网络，使用$\hat{y}_i$而不是$\hat{t}_i$作为网络的输入，用$V(x,y)$表示依赖于$y$的值函数网络。可以通过$y(t)=1-\exp(-\lambda t)$计算一个特定的时间值$t$的表示。当计算偏导数$V_t$时，必须考虑再参数化。根据链式法则，有：

$$V_t(x,t)=\tilde{V}_y(x,t)y_t(t)$$

根据所选变量的参数化，有以下结果：

$$y_t(t)=\lambda\exp(-\lambda t)$$

4. 数值例子

使用一个投资问题作为数值模拟实验的案例[5]。在这个问题中，一个主体需要决定将其的收入投资到银行账户以获得未来的奖励，或者即时消费作为即时奖励。将状态建模为银行账户的当前余额和当前利率，当主体选择消费时，主体会获得利率为0.1的奖励，但账户的余额保持不变。当主体选择投资时，银行账户的余额以0.1的利率增加，但没有额外的奖励。在这两种情况下，主体通过利息获得奖励。奖励与银行账户的当前余额成比例。假设利率根据跳跃－扩散模型随时间变化，并且为了使状态有界，对账户余额引入一个最大余额的限制。具体模型的描述如下：

状态空间$X=[0,1]\times[0,1]$，账户余额和利率建模为$x=[x_b,x_i]$

行动空间$u=\{$消费，投资$\}$

动态模型

$$f(x,u)=\begin{cases}[0,0]^T & \text{如果}u\text{是消费}\\ [0.1,0]^T & \text{如果}u\text{是投资}\end{cases}$$

$$G(x,u)=\begin{pmatrix}0 & 0\\ 0 & 0.01\end{pmatrix}$$

奖励函数

$$R(x,u)=R^x(x)+R^u(u)$$

$$R^x([x_b,x_i])=x_b\cdot x_i$$

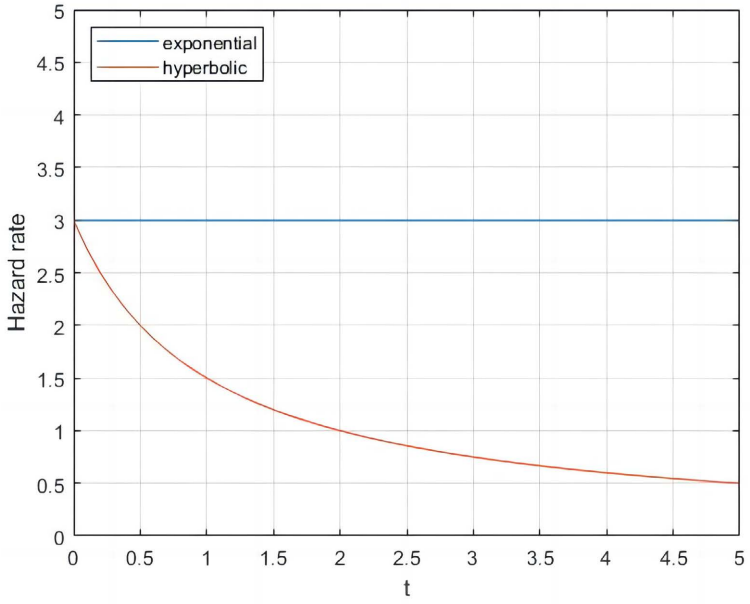
$$R^u(u)=\begin{cases}0.1 & \text{如果}u\text{是消费}\\ 0 & \text{如果}u\text{是投资}\end{cases}$$

风险率$\alpha_0=3$，$\beta_0=1$

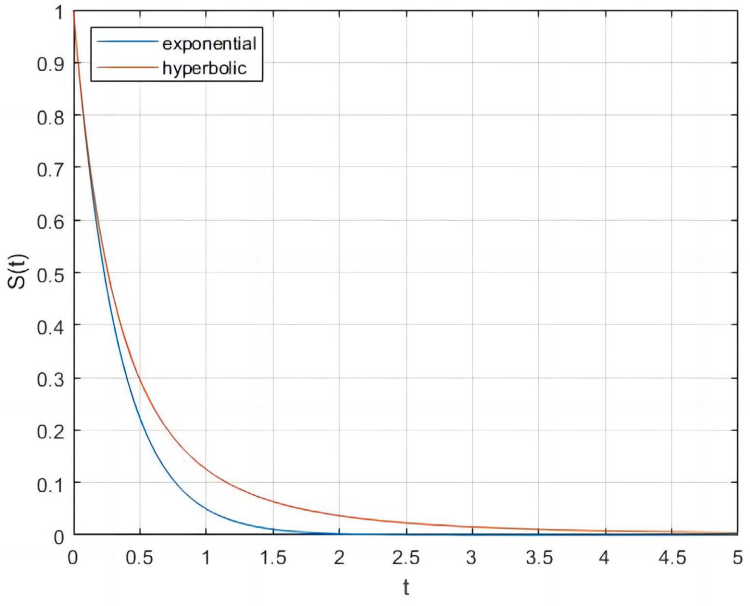
5. 数值结果及讨论

5.1. 风险率和生存函数

在投资问题中，$\alpha_0=3$，$\beta_0=1$，$\lambda=3$。即双曲折现条件下的风险率为$\alpha(t)=\frac{1}{1+t}$，生存函数为$S(t)=\frac{1}{(t+1)^3}$。指数折现下的风险率为3，生存函数为$S(t)=\exp(-3*t)$。以下是风险率和生存函数图：



(a) 风险率



(b) 生存函数

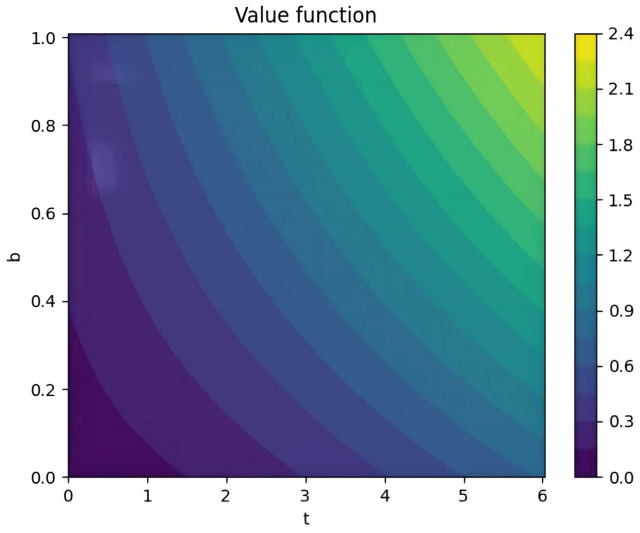
Figure 1. Hazard rate and survival function
图 1. 风险率和生存函数

从图 1 中可以看出，双曲折现函数的风险率呈递减的趋势，随着时间的变化，风险率越来越低。指数折现的风险率保持不变。用生存函数刻画双曲折现和指数折现函数时，两种折现函数的递减趋势相同。

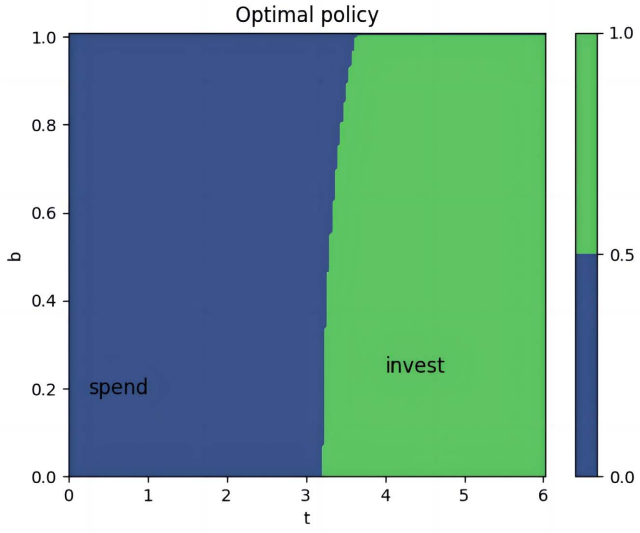
5.2. 跳跃过程对值函数和发生偏好逆转时间的影响

使用双指数分布建模跳跃幅度，其中双指数分布的参数为 $x \sim Laplace(-0.13, 0.14)$，泊松过程的强度 $\lambda = 0$ 和 $\lambda = 0.5$。

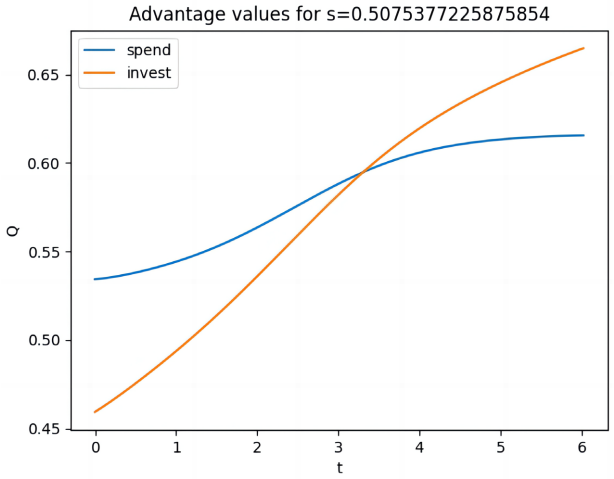
图 2 给出了当 $\lambda = 0$，即环境被建模为扩散过程时，双曲折现和指数折现下的结果。从图(a)可以看到随着账户余额的增加，值也会增加。因为余额增加，通过利息获得的预期回报会增加。随着时间的推移，值会进一步增加，因为当风险率比较低时，人们预期能够长期获得回报。图(b)和图(c)展示了学习到的策略和 Q 函数。在开始阶段，消费是有利的，但当风险相对较低时，投资变得更加吸引人，从而引发偏好逆转的现象。图(d)和图(e)展示了指数折现下的值函数和 Q 函数，由于风险率保持不变，可以看到值函数和 Q 函数不显示偏好逆转，并且随着时间保持不变，因此人们的偏好逆转不发生改变。



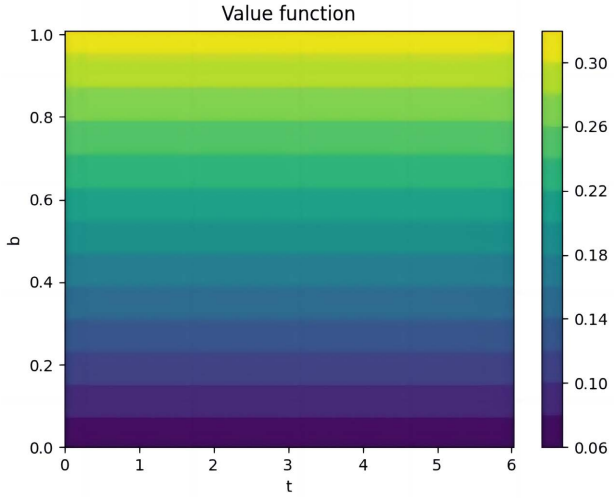
(a) 值函数



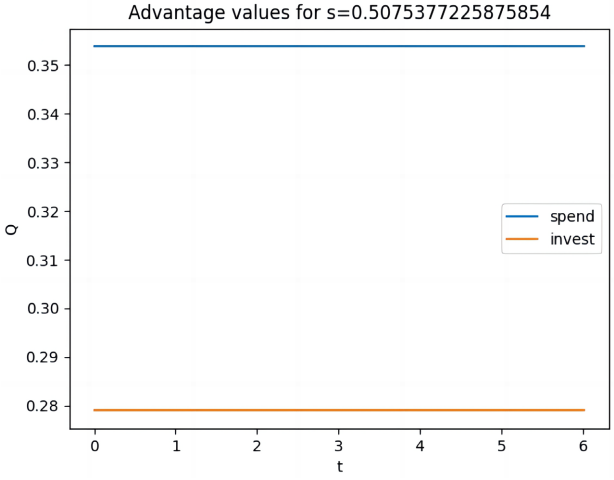
(b) 最优策略



(c) Q 函数



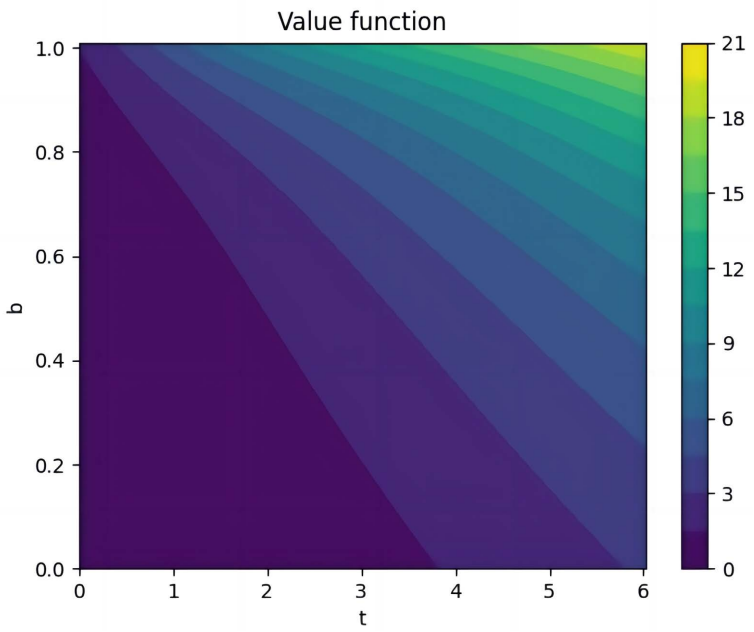
(d) 值函数



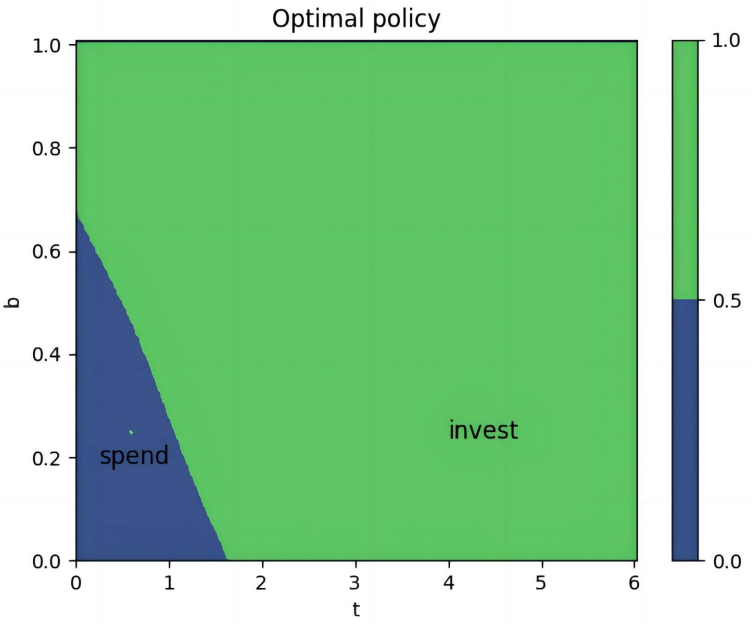
(e) Q 函数

Figure 2. Value Function, Optimal Policy, and Q-function under Hyperbolic and Exponential Discounting when $\lambda = 0$
图 2. $\lambda = 0$ 时，双曲折现和指数折现下的值函数、最优策略、Q 函数

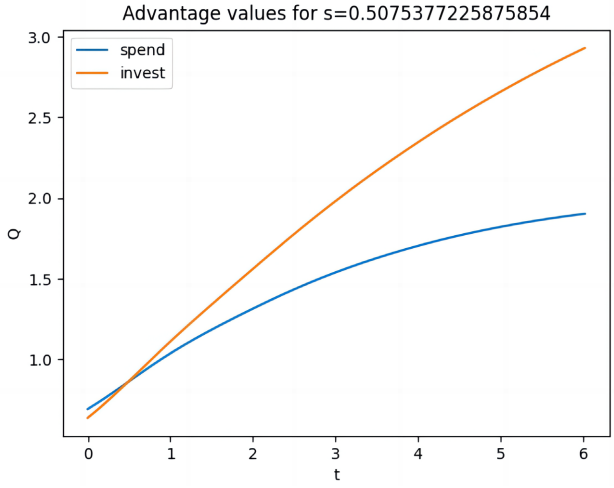
图 3 给出了 $\lambda = 0.5$，即环境被建模为跳跃－扩散过程时，跳跃幅度为双指数分布时，双曲折现和指数折现下的结果。从图(a)可以看到环境被建模为跳跃－扩散过程时，账户余额基本保持不变，值函数与时间高度相关，并且在一定的时间范围内，值函数大小不发生变化。图(b)和图(c)展示了学习到的策略和 Q 函数。在开始阶段且账户余额小于 0.7 时消费是有利的，但当风险相对较低时，投资变得更加吸引人，从而引发偏好逆转的现象，并且与环境被建模为扩散过程相比，偏好逆转发生的时间较早。图(d)和图(e)展示了指数折现下的值函数和 Q 函数，可以看到值函数和 Q 函数同样不显示偏好逆转，并且随时间保持不变。



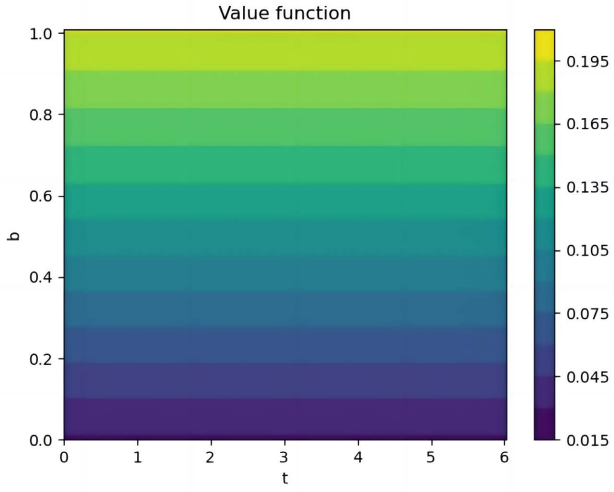
(a) 值函数



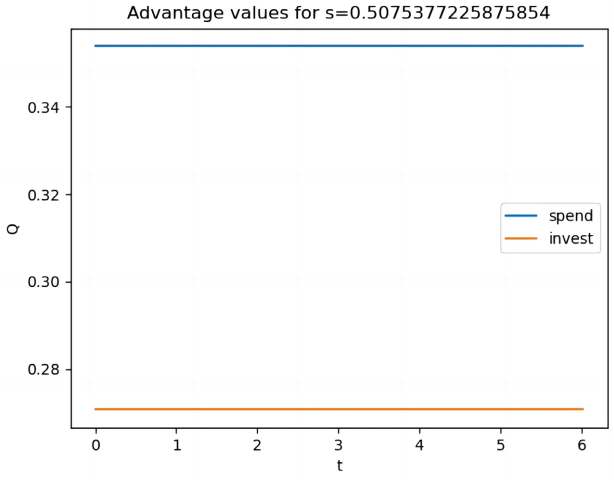
(b) 最优策略



(c) Q 函数



(d) 值函数



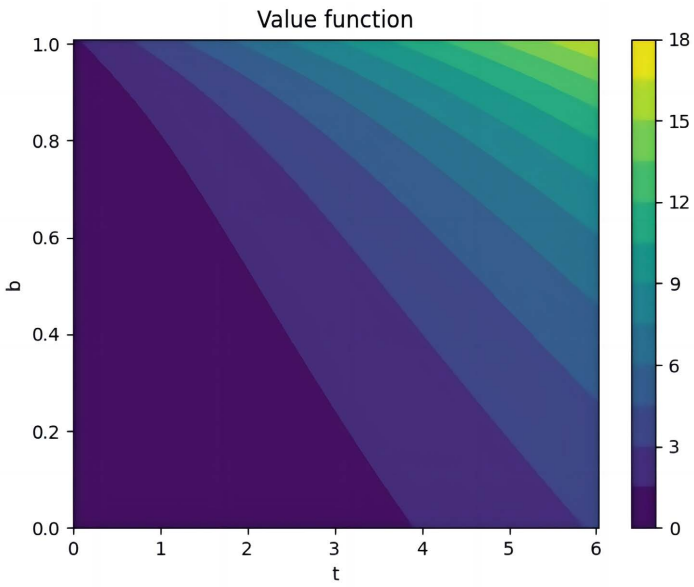
(e) Q 函数

Figure 3. Value Function, Optimal Policy, and Q-function under Hyperbolic and Exponential Discounting when $\lambda = 0.5$
图 3. $\lambda = 0.5$ 时，双曲折现和指数折现下的值函数、最优策略、Q 函数

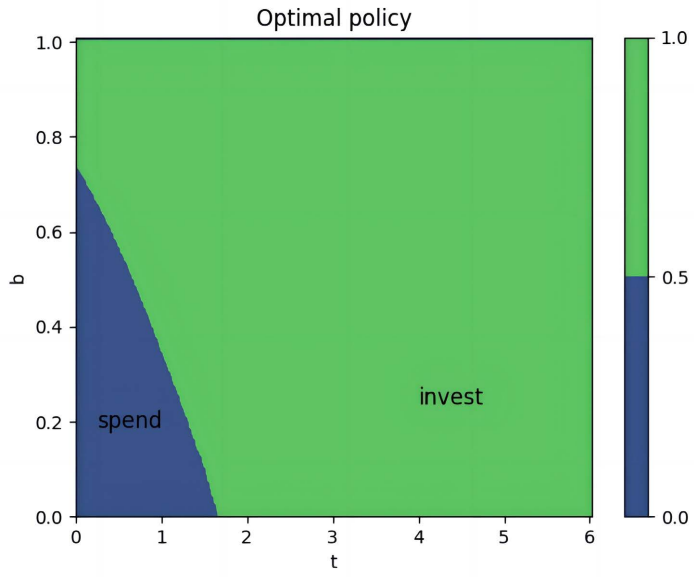
5.3. 不同强度的泊松过程对值函数和发生偏好逆转时间的影响

使用双指数分布建模突发事件的跳跃幅度，其中双指数分布的参数为 $x \sim Laplace(-0.13, 0.14)$，泊松过程的强度 $\lambda = 1.5$ 和 $\lambda = 3$。

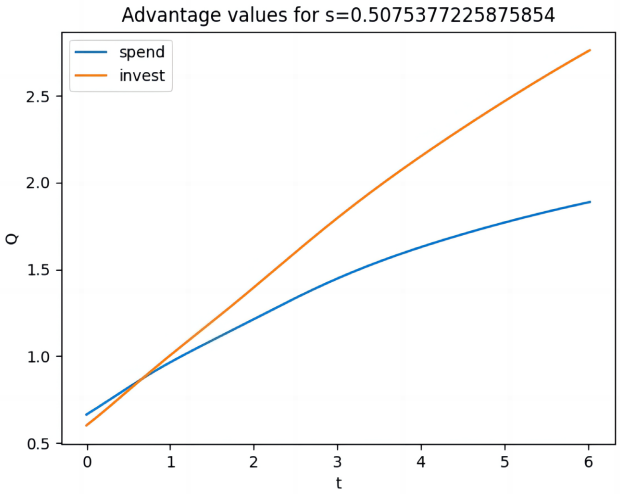
图 4 给出了跳跃幅度为双指数分布时，$\lambda = 1.5$ 和 $\lambda = 3$，双曲折现下的结果。从图(a)看出，值函数与时间高度相关，账户余额基本保持不变。从图(d)看出，值与时间高度相关，在时间区间[3.5, 4.1]上，随着账户余额的增加，值函数的值也增加。可以看出，泊松过程的强度越大，值函数的范围越大。图(b)、图(c)和图(e)、图(f)展示了两种强度下的策略和 Q 函数。可以看出在开始阶段且账户余额小于 0.7 时消费是有利的，当风险相对较低时，投资变得更加吸引人，从而引发偏好逆转的现象，并且泊松过程的强度越大，偏好逆转发生的时间越早。



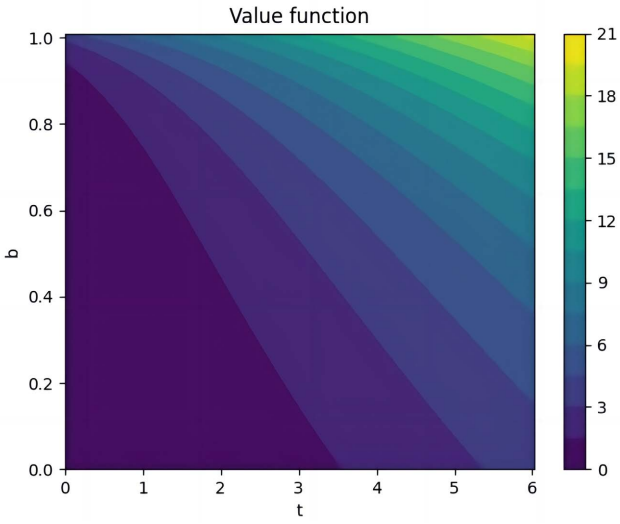
(a) 值函数



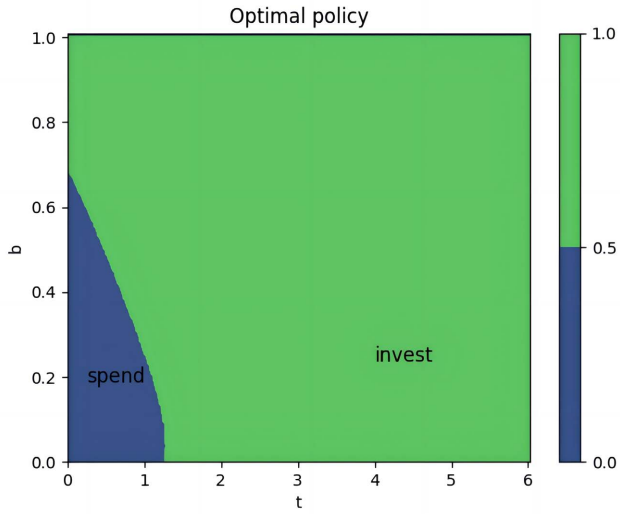
(b) 最优策略



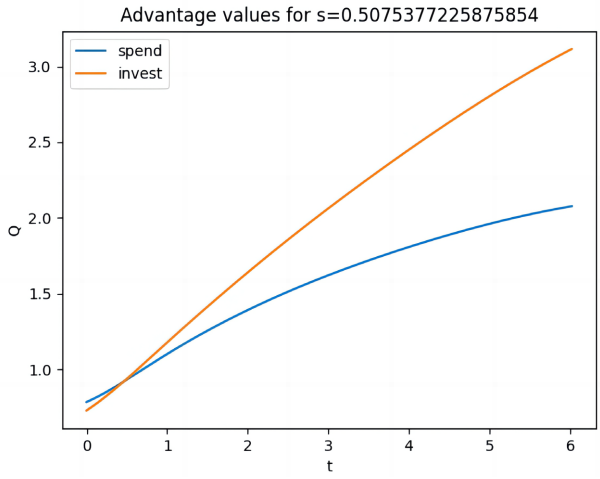
(c) Q 函数



(d) 值函数



(e) 最优策略



(f) Q 函数

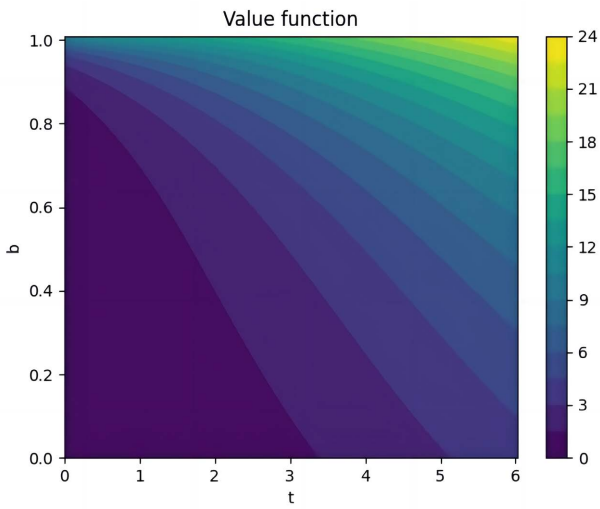
Figure 4. Value Function, Optimal Policy, and Q-function under Hyperbolic Discounting for $\lambda = 1.5$ and $\lambda = 3$
图 4. $\lambda = 1.5$ 和 $\lambda = 3$ 时，双曲折现下的值函数、最优策略、Q 函数

5.4. 同一强度下不同分布的跳跃幅度对值函数和偏好逆转时间的影响

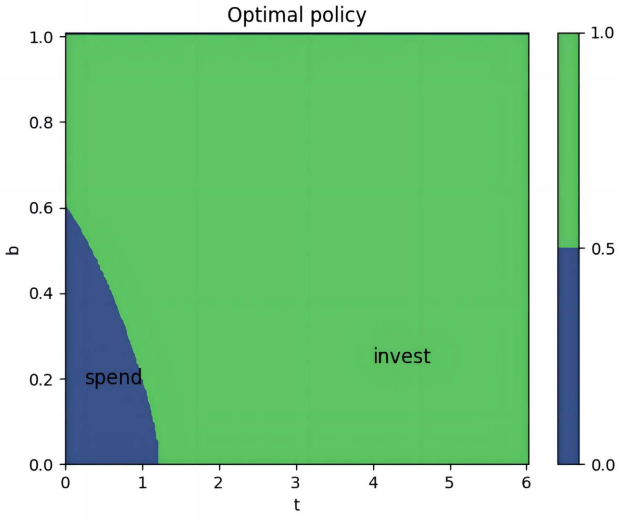
使用正态分布、指数分布、双指数分布建模跳跃幅度，其中跳跃幅度 $h(\cdot)$ 服从正态分布的参数为 $x \sim N(-0.13,0.14)$，跳跃幅度 $h(\cdot)$ 服从指数分布的参数为 $x \sim \exp(0.13)$，跳跃幅度 $h(\cdot)$ 服从双指数分布的参数为 $x \sim Laplace(-0.13,0.14)$。

图 5(a)、图 5(d)、图 5(g)展示了跳跃幅度的分布分别为正态分布、指数分布、双指数分布下的值函数，可以看出值函数与时间高度相关，在时间区间[3.2, 6]上，随着账户余额的增加，值函数的值也增加。

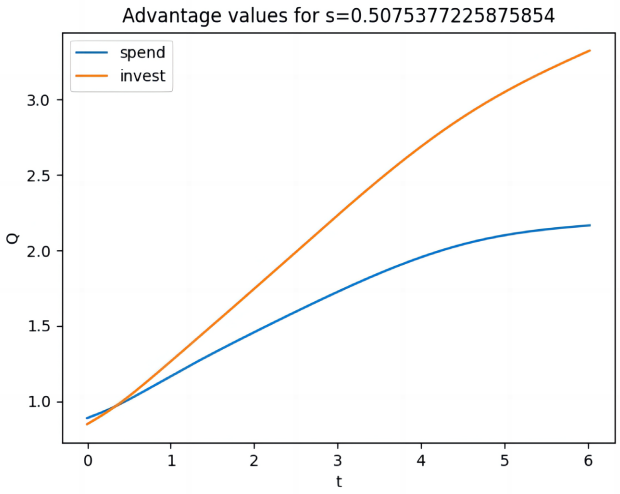
图 5(b)、图 5(c)和图 5(e)、图 5(f)和图 5(h)、图 5(g)展示了三种分布下的策略和 Q 函数。可以看出跳跃幅度为正态分布时，在开始阶段且账户余额小于 0.6 时消费是有利的，但当风险相对较低时，投资变得更加吸引人，从而引发偏好逆转的现象。当跳跃幅度为指数分布和双指数分布时，在开始阶段且账户余额小于 0.64 时消费是有利的，但当风险相对较低时，投资变得更加吸引人，从而引发偏好逆转的现象。同一强度下的跳跃幅度，当跳跃幅度为正态分布时，偏好逆转发生的时间更早。



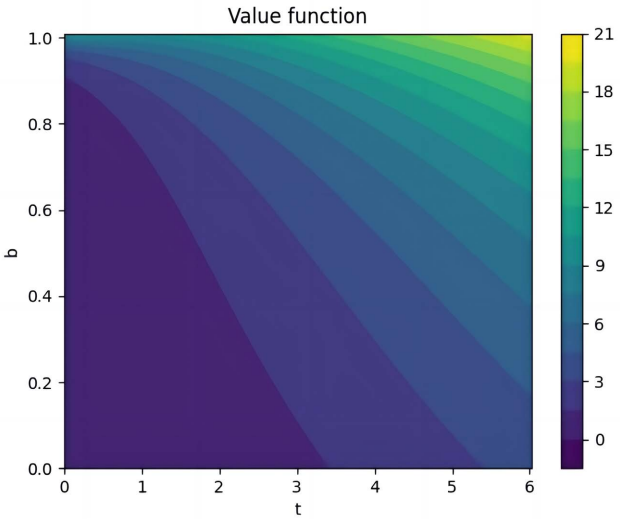
(a) 值函数



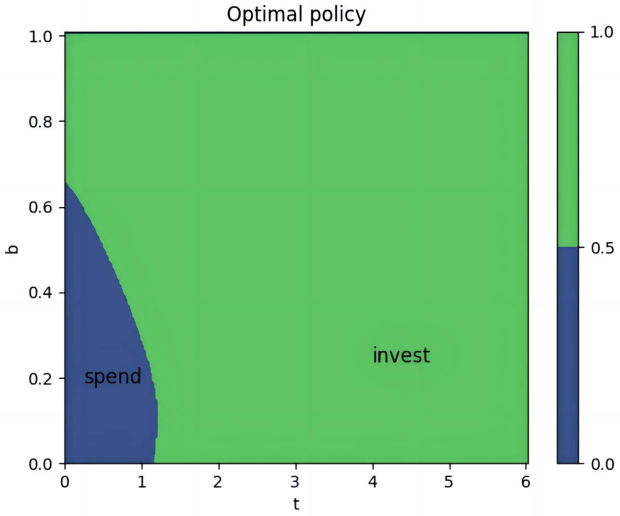
(b) 最优策略



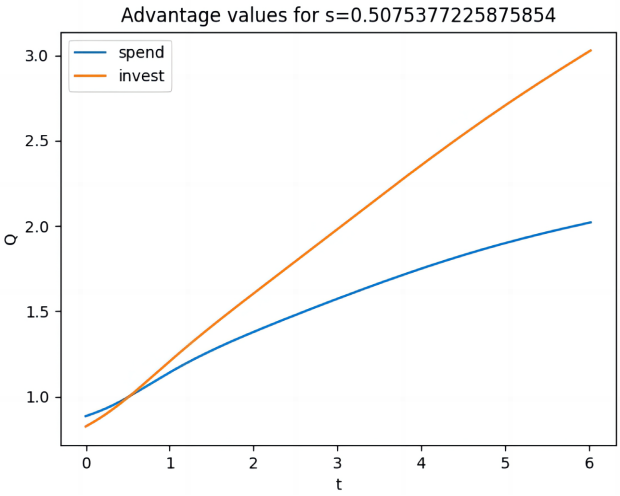
(c) Q 函数



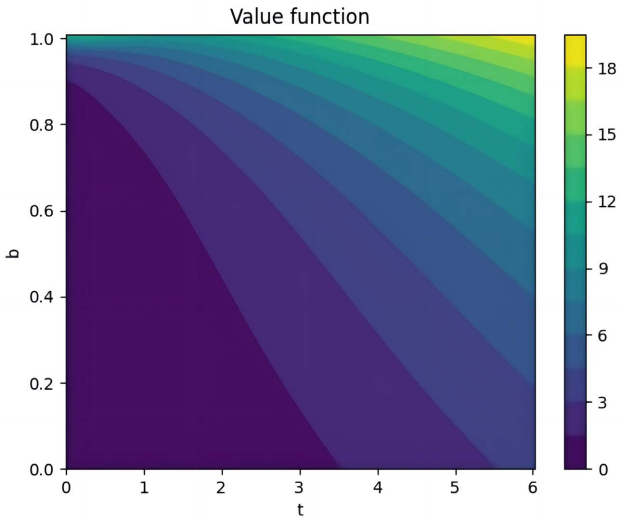
(d) 值函数



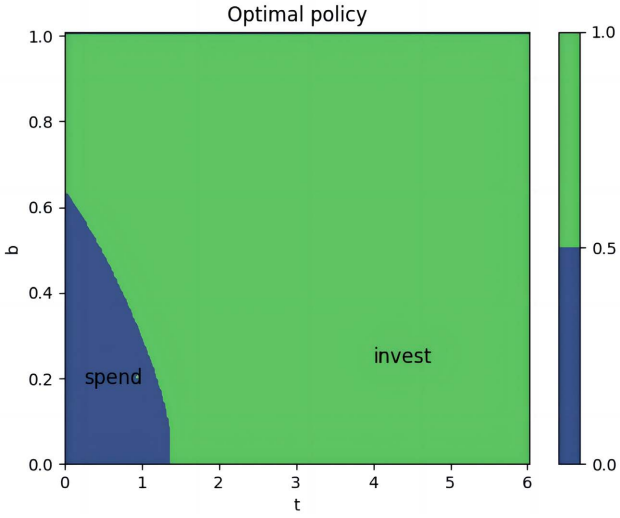
(e) 最优策略



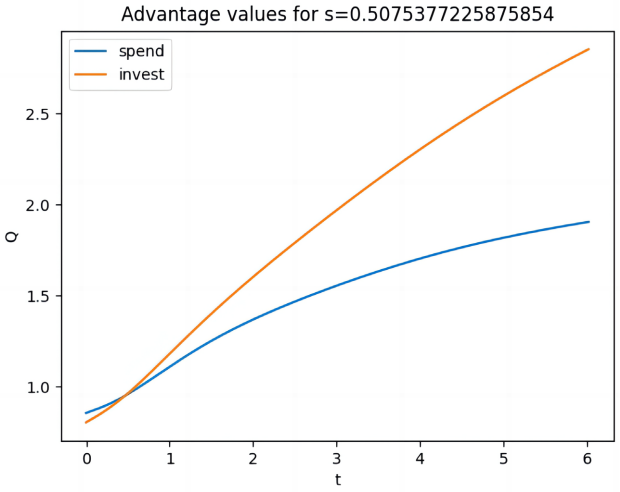
(f) Q 函数



(g) 值函数



(h) 最优策略



(i) Q 函数

Figure 5. Value Function, Optimal Policy, and Q-function under Hyperbolic Discounting for $\lambda = 5$, with Three Distributions
图 5. $\lambda = 5$，三种分布在双曲折现下的值函数、最优策略、Q 函数

6. 研究结论与启示

本文首先利用生存函数和风险率刻画折现函数，其次基于跳跃－扩散过程对强化学习中的环境状态建模。给出了使用双曲折现函数时问题的限制条件，并且推导了一个 HJB 方程，给出了要使 HJB 方程有解，方程的限制条件。探讨了跳跃－扩散过程下值函数和发生偏好逆转时间的变化。主要的结论如下：

第一，跳跃－扩散过程下值函数与时间高度相关且发生偏好逆转的时间提前。第二，同一分布下的跳跃幅度，强度越大，值函数的值越大，偏好逆转发生的时间越早。第三，同一泊松过程的强度下，当跳跃幅度的分布为正态分布时，值函数的值最大，偏好逆转发生的时间最早。最后得出，在强化学习中，使用双曲折现函数能更好的模拟智能体做决策时的时间偏好。

基于上述的研究结论，立足于利用强化学习进行决策优化，针对跳跃－扩散过程和本文的实验过程，本文提出以下建议。

第一，使用强化学习进行投资决策时，应该使用双曲折现函数，这样能更好地模拟人类的时间偏好。

利用强化学习进行决策优化时，当环境被建模为跳跃－扩散过程时，人们最好在原来时间的基础上提前调整自己的策略。

第二，在推导过程中，对跳跃项的近似处理、在数值实验时固定突发事件发生的次数。这样的处理在一定程度上会导致结果出现误差，未来的研究可以考虑更加灵活的跳跃模型，并探索使用其他方法求解 HJB 方程，以进一步提高实验结果的可靠性和准确性。

参考文献

[1] Fedus, W., Gelada, C., Bengio, Y., *et al*. (2019) Hyperbolic Discounting and Learning over Multiple Horizons. arXiv: 1902.06865.

[2] Sozou, P.D. (1998) On Hyperbolic Discounting and Uncertain Hazard Rates. *Proceedings of the Royal Society of London. Series B*: *Biological Sciences*, **265**, 2015-2020. https://doi.org/10.1098/rspb.1998.0534

[3] Alexander, W.H. and Brown, J.W. (2010) Hyperbolically Discounted Temporal Difference Learning. *Neural Computation*, **22**, 1511-1527. https://doi.org/10.1162/neco.2010.08-09-1080

[4] Alia, I. (2019) A Non-Exponential Discounting Time-Inconsistent Stochastic Optimal Control Problem for Jump-Diffusion. *Mathematical Control and Related Fields*, **9**, 541-570. https://doi.org/10.3934/mcrf.2019025

[5] Schultheis, M., Rothkopf, C.A. and Koeppl, H. (2022) Reinforcement Learning with Non-Exponential Discounting. *Advances in Neural Information Processing Systems*, **35**, 3649-3662.

[6] Nafi, N.M., Ali, R.F. and Hsu, W. (2022) Hyperbolically Discounted Advantage Estimation for Generalization in Reinforcement Learning. Decision Awareness in Reinforcement Learning Workshop at ICML 2022.

[7] Ali, R.F. (2023) Non-Exponential Reward Discounting in Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 16111-16112. https://doi.org/10.1609/aaai.v37i13.26916

[8] Kwiatkowski, A., Kalogeiton, V., Pettré, J., *et al*. (2023) UGAE: A Novel Approach to Non-exponential Discounting. arXiv: 2302.05740.

[9] Hanson, F.B. (2007) Applied Stochastic Processes and Control for Jump-Diffusions: Modeling, Analysis and Computation. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9780898718638

[10] Aalen, O., Borgan, O. and Gjessing, H. (2008) Survival and Event History Analysis: A Process Point of View. Springer, New York. https://doi.org/10.1007/978-0-387-68560-1

[11] Särkkä, S. and Solin, A. (2019) Applied Stochastic Differential Equations. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108186735

[12] Fleming, W.H. and Soner, H.M. (2006) Controlled Markov Processes and Viscosity Solutions. Springer Science & Business Media, New York.

[13] Simpkins, A. and Todorov, E. (2009) Practical Numerical Methods for Stochastic Optimal Control of Biological Systems in Continuous Time and Space. 2009 *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, Nashville, 30 March-2 April 2009, 212-218. https://doi.org/10.1109/ADPRL.2009.4927547

[14] Tassa, Y. and Erez, T. (2007) Least Squares Solutions of the HJB Equation with Neural Network Value-Function Approximators. *IEEE Transactions on Neural Networks*, **18**, 1031-1041. https://doi.org/10.1109/TNN.2007.899249

[15] Lutter, M., Belousov, B., Listmann, K., *et al*. (2020) HJB Optimal Feedback Control with Deep Differential Value Functions and Action Constraints. 3*rd Conference on Robot Learning* (*CoRL* 2019), Osaka, 640-650.

[16] Sirignano, J. and Spiliopoulos, K. (2018) DGM: A Deep Learning Algorithm for Solving Partial Differential Equations. *Journal of Computational Physics*, **375**, 1339-1364. https://doi.org/10.1016/j.jcp.2018.08.029