

银行货币储备博弈的强化学习方法

李 策

(中国科学技术大学数学科学学院, 安徽 合肥 230026)

摘要: 在大规模银行交互系统中, 各银行可通过控制与中央银行的借贷率来使自身对数字货币储备尽可能地接近样本均值, 从而降低系统性风险发生的概率. 然而当状态过程与目标函数的参数未知时, 无法直接求解随机微分博弈问题得到纳什均衡. 本文结合平均场博弈理论与连续时间强化学习的相关方法, 构造了一组大规模银行借贷网络中的近似纳什均衡. 首先通过求解向前向后耦合 HJB-FPK 方程, 得到代表银行的平均场均衡策略; 再通过所得策略的形式, 设计出迭代参数的方法用以刻画参数未知时的近似最优策略; 最后通过学到的参数, 构造银行数量较大时的近似纳什均衡.

关键词: 系统性风险; 强化学习; 近似纳什均衡; 平均场博弈

MR(2010) 主题分类号: 60H10; 91A15 中图分类号: O211.9

文献标识码: A 文章编号: 0255-7797(2025)01-0081-14

1 引言

2008 年美国次贷危机引发了全球金融危机, 这对世界经济和金融体系造成了巨大冲击的同时也揭示了银行业系统性风险的严重性. 此后, 人们普遍认识到银行间的网络结构在解释风险在整个系统中的扩散方面发挥着至关重要的作用. 通过银行间的联系, 一家金融机构的违约或倒闭会严重影响其交易对手的状况, 进而使整个金融系统受到波及. 这种由依赖网络传播产生的风险通常被称为系统性风险, 其可能会导致金融机构在金融危机期间大量受困. 因此对银行间系统性风险的研究变得越发重要, 而对银行间系统性风险的模型也在逐步发展. Fouque[1] 中从多个角度介绍了系统风险理论的发展, 为系统性风险的研究提供了参考. Zhang 等 [2] 研究并评估了近年来中国银行业的系统性风险, 并且给出了预测系统性风险的概率方法.

由于认识到系统性风险理论的重要性, 人们关心如何用严格框架描述和解释银行间的相互作用和系统性风险. 最早做这方面尝试的有 Fouque and Ichiba[3], 他们提出将银行的货币储备建模为一个相互作用的系统, 并将银行的储备金价值达到一个临界值的情况定义为违约. Fouque and Sun[4] 提出了一个更简单的模型, 在这个模型下, 他们给出了系统的平均场极限和储备金状态过程达到临界值的大偏差估计. 而 Bo and Capponi[5] 在此基础上进行了进一步的研究, 通过引入跳描述银行可能受到的影响其货币储备水平的突然冲击, 构建了新的模型, 并给出了对应的测度值过程的弱极限.

为了避免系统性风险的发生, 金融实体用各种方法对自身进行调整使得系统相对稳定是非常必要的. 实际上, 各个金融实体间通常会通过借贷来避免自身的现金流出现危机. Carmona 等 [6] 在 Fouque and Sun[4] 的基础上加入了从中央银行借款或者是向中央银行放

*收稿日期: 2024-03-16 接收日期: 2024-05-21

作者简介: 李策 (1999-), 男, 河北邯郸, 研究生, 主要研究方向: 随机分析.

E-mail: lc9916@mail.ustc.edu.cn.

贷形式的控制用来研究系统性风险. 文中得到了开环 - 闭环纳什均衡策略的解析形式, 并证明了当各银行的对数货币储备越接近样本均值, 系统性风险发生的概率越小. Bo 等 [7] 从该模型出发, 考虑了中央银行如何通过调整中心化策略来使所有银行对数货币储备尽可能地接近设定的目标值, 并证明了当银行数量趋于无穷时, 最优控制会收敛至极限时平均场控制问题的最优解.

由于市场中存在大量商业银行, 通常过多的参与者使得直接求解博弈问题较为困难. Huang 等 [8] 以及 Lasry and Lions [9] 提出的平均场博弈理论给出了一种求解大规模博弈问题的较为简单方法. 该理论基于参与者数量趋于无穷时代表性参与者的控制问题, 可以构造一组逼近纳什均衡策略, 从而极大地降低求解纳什均衡的难度. 因此该理论一经问世就在各个领域也引起了广泛关注, 得到了迅速的发展. Carmona [10] 用概率的方法得到了平均场问题的解并且证明了该解确实是有限个体博弈问题的逼近纳什均衡. Cardaliaguet 等 [11] 中对近年平均场博弈理论的发展进行了简单的介绍, 并给出了通过主方程得到平均场均衡策略的方法.

另一方面, 金融系统模型中的参数一般是未知的, 因此无法直接求解最优策略. Wang 等 [12] 将强化学习的方法引入到了连续模型中, 给出了一种通过探索来求解连续时间控制问题的方法. 该篇文章将原问题中的控制过程转换为控制取值空间的测度值过程, 并在目标函数中引入香农熵来鼓励探索. Guo 等 [13] 将连续时间强化学习的方法引入了平均场博弈问题, 并通过数值实验分析了熵正则项对学习结果的影响.

本文结合平均场博弈理论与连续时间强化学习的相关方法, 首先对极限时代表银行的随机控制问题进行分析, 通过求解向前向后耦合 HJB-FPK 方程得到了模型系数已知时的平均场均衡策略. 结合所得最优策略的形式, 选取适当参数进行迭代, 得到了系数未知时的近似最优策略, 并通过数值实验说明方法的有效性. 最后, 结合算法得到的输出参数构造一组近似最优策略, 证明其是银行数量较大时的较好近似.

2 数学模型

Carmona 等 [6] 在 Fouque and Sun [4] 中提出的银行间交互系统的基础上, 假设每家银行可以向中央银行借贷, 以期银行的对数货币储备尽可能地靠近所有银行对数货币储备的样本均值. 具体地, 设 $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ 是一个完备的过滤概率空间, 其中域流 $\mathbb{F} = \{\mathcal{F}_t\}_{t \in [0, T]}$ 满足通常性条件, $T > 0$ 为确定的时间终点, $W^i = \{W_t^i\}_{t \in [0, T]}, i = 1, \dots, N$ 是 N 个相互独立的 \mathbb{F} -布朗运动. 考虑具有 N 家银行的系统, 其中对任意 $i = 1, \dots, N$, X_t^i 表示第 i 家银行 t 时刻的对数现金储备. 进一步, 设 $X^i = \{X_t^i\}_{t \in [0, T]}$ 满足如下 SDE (随机微分方程): $i = 1, \dots, N$,

$$dX_t^i = \frac{a}{N} \sum_{j=1}^N (X_t^j - X_t^i) dt + \alpha_t^i dt + \sigma dW_t^i, \quad X_0^i \sim \nu, \quad (2.1)$$

其中, 交互项 $\bar{X}_t^N - X_t^i$ 刻画了银行货币储备出现的“群集”效应, $a > 0$ 表示银行间的常值借贷率, 该项良好地刻画了系统性风险发生概率较低, 但是一旦发生会使得大量银行违约的特征. σ 表示独立噪声项的波动率, 对应的布朗运动项则为银行间不进行借贷情况下对货币储备的描述. 初始现金储备由 $X_0^i, i = 1, \dots, N$ 来描述, 它们为独立同分布并二阶矩有限的随机变量 (记其分布为 ν), 且与布朗运动序列相互独立. 控制过程 $\alpha^i = \{\alpha_t^i\}_{t \in [0, T]}$ 表示第 i 家

银行向中央银行的借贷策略. 记策略集合 $\alpha = (\alpha^1, \dots, \alpha^N)$, 则每家银行的目标是在可容许策略集中选取最优借贷率过程, 以使如下代价泛函最小化:

$$J^{i,N}(\alpha) := \mathbb{E} \left[\int_0^T \left((X_t^i - \bar{X}_t^N)^2 + b(\alpha_t^i)^2 \right) dt + \gamma (X_T^i - \bar{X}_T^N)^2 \right],$$

其中对任意 $t \in [0, T]$, $\bar{X}_t^N := \frac{1}{N} \sum_{j=1}^N X_t^j$, 且 $b(\alpha_t^i)^2$ 表示第 i 家银行 t 时刻借贷的成本率.

尽管 Carmona 等 [6] 给出了一组纳什均衡策略, 然而在现实金融系统中, 模型参数往往是未知的, 故所构造的最优策略无法直接使用. 本文结合平均场博弈以及 Wang 等 [12] 中连续时间强化学习的相关方法, 构造模型中参数未知时的近似最优控制. 当模型中参数未知时, 参与者可以通过试错来探索和学习未知的环境. 此时, 控制过程可看作控制策略取值空间 $U = \mathbb{R}$ 上的概率测度流 $\pi = (\pi_t)_{t \in [0, T]}$. 具体地, 对任意 $i = 1, \dots, N$, 第 i 家银行的受控对数现金储备满足如下 SDE (参见 Wang 等 [12] 中第 2 节):

$$dX_t^i = a(\bar{X}_t^N - X_t^i) dt + \left[\int_U u \pi_t^i(du) \right] dt + \sigma dW_t^i, \quad X_0^i \sim \nu. \quad (2.2)$$

不同于控制过程 α^i , 测度值控制过程 $\pi^i = \{\pi_t^i\}_{t \in [0, T]} \in \mathcal{D}$, 其中可允许控制策略集合 \mathcal{D} 表示 \mathbb{F} -适应并关于 t 右连左极的测度值过程, 且对任意 $t \in [0, T]$, $\pi_t^i \in \mathcal{P}_2(U)$. 这里, $\mathcal{P}_2(U)$ 是 U 上全体二阶矩有限且具有概率密度函数的分布构成的集合. 进一步, 还需在目标函数中引入香农熵作为正则项来鼓励对环境信息的探索, 故相应的目标泛函定义如下:

$$J^{i,N}(\pi^N) := \mathbb{E} \left[\int_0^T \left((\bar{X}_t^N - X_t^i)^2 + b \int_U u^2 \pi_t^i(du) - \lambda H(\pi_t^i) \right) dt + \gamma (\bar{X}_T^N - X_T^i)^2 \right], \quad (2.3)$$

其中 $\lambda > 0$ 是调节探索信息重要性所占比重的参数, 策略集合 $\pi^N := (\pi^1, \dots, \pi^N)$, $H(\pi_t^i)$ 表示概率测度 π_t^i 的香农熵. 具体地,

$$H(\pi_t^i) := - \int_U p_t^i(u) \log [p_t^i(u)] du, \quad (2.4)$$

其中 p_t^i 是 π_t^i 的密度函数. 基于各银行的目标函数 (2.3), 首先给出纳什均衡的定义如下:

定义 2.1 对上述具有 N 个参与者的博弈问题, 称策略 $\pi^{*,N} = (\pi^{*,1}, \dots, \pi^{*,N}) \in \mathcal{D}^N$ 为纳什均衡, 如果

$$J^{i,N}(\pi^{*,N}) = \inf_{\pi^i \in \mathcal{D}} J^{i,N}(\pi^i, \pi^{*,-i}), \quad \forall i = 1, \dots, N, \quad (2.5)$$

其中 $(\pi^i, \pi^{*,-i})$ 表示策略 $(\pi^{*,1}, \dots, \pi^{*,i-1}, \pi^i, \pi^{*,i+1}, \dots, \pi^{*,N})$. 对给定的 $\varepsilon > 0$, 称策略 $\pi^* = (\pi^{*,1}, \dots, \pi^{*,N}) \in \mathcal{D}^N$ 为 ε 纳什均衡, 如果

$$J^{i,N}(\pi^{*,N}) \leq \inf_{\pi^i \in \mathcal{D}} J^{i,N}(\pi^i, \pi^{*,-i}) + \varepsilon, \quad \forall i = 1, \dots, N. \quad (2.6)$$

称 $\{\pi^{*,N}\}_{N \geq 1}$ 为逼近纳什均衡, 如果对每一个 $N > 0$, 存在 $\{\pi^{*,N}\}_{N \geq 1}$ 满足 $\pi^{*,N}$ 为 ε_N -纳什均衡且 $\lim_{N \rightarrow \infty} \varepsilon_N = 0$.

由于模型中参数未知且所考虑的是具有多个参与者的博弈问题, 故直接使用 Wang[12] 中强化学习的框架难以直接进行求解, 尤其当银行数量 N 较大时, 算法的收敛速度较为缓慢. 因此, 本文结合平均场博弈的相关思想, 首先求解参数已知的条件下, 银行数量 N 趋于无穷时代表性银行的平均场均衡, 再通过强化学习的方法学到参数未知时的均衡策略, 最后通过上述两个步骤构造一组银行数量 N 有限时的近似最优借贷策略.

3 代表银行的平均场均衡策略

本节首先求解银行数量 $N \rightarrow \infty$ 时代表银行的平均场均衡策略. 具体地, 对给定的确定性函数 $m = \{m_t\}_{t \in [0, T]} \in C([0, T]; \mathbb{R})$, 代表银行的最优控制问题可表示为如下形式:

$$\begin{cases} \inf_{\pi \in \mathcal{D}} J(\pi; m) \\ := \inf_{\pi \in \mathcal{D}} \mathbb{E} \left[\int_0^T \left((m_t - X_t^\pi)^2 + b \int_U u^2 \pi_t(du) - \lambda H(\pi_t) \right) dt + \gamma (m_T - X_T^\pi)^2 \right], \\ \text{s.t. } dX_t^\pi = a(m_t - X_t^\pi) dt + \left[\int_U u \pi_t(du) \right] dt + \sigma dW_t, X_0^\pi = X_0 \sim \nu. \end{cases} \quad (3.1)$$

上述控制问题的平均场均衡 (也称为平均场博弈问题的纳什均衡) 策略的定义如下:

定义 3.2 策略 $(\pi^* = \{\pi_t^*\}_{t \in [0, T]}, m^* = \{m_t^*\}_{t \in [0, T]})$ 称为 (3.1) 的平均场均衡策略, 如果以下条件满足:

- 对任意 $\pi \in \mathcal{D}$, 都有

$$J(\pi^*; m^*) \leq J(\pi; m^*).$$

- 对任意 $t \in [0, T]$,

$$\mathbb{E}[X_t^*] = m_t^*,$$

其中 $X^* = (X_t^*)_{t \in [0, T]}$ 为 (π^*, m^*) 下由 (3.1) 中状态方程所确定的唯一强解.

为求得代表性银行的平均场均衡策略, 首先给出值函数的定义:

$$\begin{aligned} V(t, x; m) &:= \inf_{\pi \in \mathcal{D}} J(t, x, \pi; m^*) \\ &:= \inf_{\pi \in \mathcal{D}} \mathbb{E} \left[\int_t^T \left((m_s - X_s^\pi)^2 + b \int_U u^2 \pi_s(du) - \lambda H(\pi_s) \right) ds + \gamma (m_T - X_T^\pi)^2 \middle| X_t^\pi = x \right]. \end{aligned}$$

进一步, 定义平均场均衡策略 (π^*, m^*) 下的博弈值函数为

$$V(t, x) := V(t, x; m^*) = J(t, x, \pi^*; m^*) = \inf_{\pi \in \mathcal{D}} J(t, x, \pi; m^*).$$

下面的定理给出了代表银行的平均场均衡策略.

定理 3.1 代表银行控制问题 (3.1) 的博弈值函数为: 对任意 $(t, x) \in [0, T] \times \mathbb{R}$,

$$V(t, x) = V(t, x; m^*) = \frac{1}{2} \zeta_t (\bar{m} - x)^2 + \eta_t,$$

其中, $\bar{m} = \mathbb{E}[X_0]$, 且 $m^* = \{m_t^*\}_{t \in [0, T]}$ 满足对任意 $t \in [0, T]$, $m_t^* \equiv \bar{m}$, 确定性函数 $\{\zeta_t\}_{t \in [0, T]}$ 和 $\{\eta_t\}_{t \in [0, T]}$ 的定义如下: 对任意 $t \in [0, T]$,

$$\begin{cases} \zeta_t = \frac{1 + \left(\frac{2\gamma + 2ab - \sqrt{4a^2b^2 + 4b}}{2\gamma + 2ab + \sqrt{4a^2b^2 + 4b}} \right) e^{\frac{\sqrt{4a^2b^2 + 4b}}{b}(t-T)}}{1 - \left(\frac{2\gamma + 2ab - \sqrt{4a^2b^2 + 4b}}{2\gamma + 2ab + \sqrt{4a^2b^2 + 4b}} \right) e^{\frac{\sqrt{4a^2b^2 + 4b}}{b}(t-T)}} \sqrt{4a^2b^2 + 4b} - 2ab, \\ \eta_t = \left(-\frac{\lambda}{2} \log \frac{\pi\lambda}{b} \right) (T-t) + \int_t^T \frac{\sigma^2}{2} \zeta_s ds. \end{cases} \quad (3.2)$$

定义最优反馈控制函数:

$$\pi^*(t, x; m^*) \sim \mathcal{N} \left(\frac{\zeta_t}{2b} (\bar{m} - x), \frac{\lambda}{2b} \right). \quad (3.3)$$

进一步, 对任意 $t \in [0, T]$, 设 $\pi_t^* = \pi^*(t, X_t^*)$, 其中 $X^* = (X_t^*)_{t \in [0, T]}$ 是如下随机微分方程的唯一解:

$$dX_t^* = \left(a + \frac{\zeta_t}{2b} \right) (\bar{m} - X_t^*) dt + \sigma dW_t, \quad X_0^* = X_0. \quad (3.4)$$

于是, $(\pi^* = \{\pi_t^*\}_{t \in [0, T]}, m^* = \{m_t^*\}_{t \in [0, T]})$ 即为博弈问题 (3.1) 的平均场均衡策略.

证 对任意固定的确定性函数 $m = \{m_t\}_{t \in [0, T]} \in C([0, T]; \mathbb{R})$, 相应的 HJB 方程为:

$$\begin{aligned} -\frac{\partial V}{\partial t}(t, x; m) &= \frac{\sigma^2}{2} \frac{\partial^2 V}{\partial x^2}(t, x; m) + a(m_t - x) \frac{\partial V}{\partial x}(t, x; m) + (m_t - x)^2 \\ &\quad + \inf_{\pi \in \mathcal{D}} \left\{ b \int_U u^2 \pi(du) - \lambda H(\pi) + \int_U u \pi(du) \frac{\partial V}{\partial x}(t, x; m) \right\}, \\ V(T, x; m) &= \gamma (m_T - x)^2. \end{aligned} \quad (3.5)$$

于是, 最优反馈函数满足:

$$\pi^*(t, x; m) = \arg \min_{\pi \in \mathcal{D}} \left\{ b \int_U u^2 \pi(du) - \lambda H(\pi) + \int_U u \pi(du) \frac{\partial V}{\partial x}(t, x) \right\}.$$

由拉格朗日乘子法, 相应的密度函数满足:

$$p^*(u, t, x; m) = \frac{\exp \left(-\frac{1}{\lambda} \left(\frac{\partial V}{\partial x} u + bu^2 \right) \right)}{\int_{\mathbb{R}} \exp \left(-\frac{1}{\lambda} \left(\frac{\partial V}{\partial x} u + bu^2 \right) \right) du} = \frac{1}{\sqrt{\frac{\lambda}{b} \pi}} \exp \left(-\frac{\left(u + \frac{\frac{\partial V}{\partial x}}{2b} \right)^2}{\frac{\lambda}{b}} \right).$$

进一步, 设值函数 $V(t, x; m)$ 具有如下形式:

$$V(t, x; m) = \frac{1}{2} \zeta_t (m_t - x)^2 + \eta_t,$$

其中 $\zeta = \{\zeta_t\}_{t \in [0, T]}$ 和 $\eta = \{\eta_t\}_{t \in [0, T]}$ 是 $[0, T] \rightarrow \mathbb{R}$ 的连续一次可导的函数. 此时, $\pi^*(t, x; m)$ 服从如下正态分布:

$$\pi^*(t, x; m) \sim \mathcal{N} \left(\frac{\zeta_t}{2b} (m_t - x), \frac{\lambda}{2b} \right). \quad (3.6)$$

从而, 其期望和二阶矩为 $\int_R u\pi_t^*(du) = \frac{\zeta_t}{2b}(m_t - x)$, $\int_R u^2\pi_t^*(du) = \frac{\lambda}{2b} + \frac{\zeta_t^2}{4b^2}(m_t - x)^2$. 记反馈分布 $\pi^*(t, x; m)$ 下状态方程 (3.1) 的解的密度函数为 $q(t, x)$, 则其满足下述 FPK (Fokker Planck Kolmogorov) 方程:

$$\frac{\partial q(s, x)}{\partial s} = -\frac{\partial}{\partial x} \left(\left(a(m_s - x) + \frac{\zeta_s}{2b}(m_s - x) \right) q(s, x) \right) + \frac{\sigma^2}{2} \frac{\partial^2 q(s, x)}{\partial x^2}.$$

方程两边同时乘以 x 并对 x 积分可得 $\frac{d\mathbb{E}[X_s]}{ds} = \left(\int x \partial_s q(s, dx) \right) = 0$. 注意到平均场均衡策略 m^* 与最优控制下状态过程的期望相等, 上式表明其关于时间不变, 因此对任意 $t \in [0, T]$, $m_t^* \equiv \mathbb{E}[X_0] =: \bar{m}$. 将其代入最优反馈函数 (3.6) 和 HJB 方程 (3.5), 比较方程两边关于 x 的系数, 可得 ζ 和 η 满足以下常微分方程组:

$$\begin{cases} \frac{d\zeta_t}{dt} = 2a\zeta_t + \frac{\zeta_t^2}{2b} - 2, & \zeta_T = 2\gamma, \\ \frac{d\eta_t}{dt} = -\frac{\sigma^2}{2}\zeta_t + \frac{\lambda}{2} \log \frac{\pi\lambda}{b}, & \eta_T = 0. \end{cases}$$

该方程组存在唯一解, 并有如下显示表达式:

$$\begin{cases} \zeta_t = \frac{1 + \left(\frac{2\gamma + 2ab - \sqrt{4a^2b^2 + 4b}}{2\gamma + 2ab + \sqrt{4a^2b^2 + 4b}} \right) e^{\frac{\sqrt{4a^2b^2 + 4b}}{b}(t-T)}}{1 - \left(\frac{2\gamma + 2ab - \sqrt{4a^2b^2 + 4b}}{2\gamma + 2ab + \sqrt{4a^2b^2 + 4b}} \right) e^{\frac{\sqrt{4a^2b^2 + 4b}}{b}(t-T)}} \sqrt{4a^2b^2 + 4b} - 2ab, \\ \eta_t = \left(-\frac{\lambda}{2} \log \frac{\pi\lambda}{b} \right) (T - t) + \int_t^T \frac{\sigma^2}{2} \zeta_s ds. \end{cases}$$

容易验证, 最优反馈函数 $\pi^*(t, x; m)$ 下状态过程 $X^* = (X_t^*)_{t \in [0, T]}$ 满足方程 (3.4), 其具有如下解析表达式 (参阅 Protter[14] 中定理 V.52): 对任意 $t \in [0, T]$, $X_t^* = (X_0 - \bar{m})e^{-\int_0^t (a + \frac{\zeta_s}{2b}) ds} + \bar{m} + \int_0^t \sigma e^{-\int_s^t (a + \frac{\zeta_u}{2b}) du} dW_s$. 故有 $\mathbb{E}[X_t^*] \equiv \bar{m}$. 从而定理得证.

定理 3.1 给出了最优反馈控制函数 $\pi^*(t, x; \bar{m})$ 具有形式 (3.3), 其中 $\bar{m} = \mathbb{E}[X_0]$. 因此, 为求得参数未知时的最优控制, 只需考虑如下形式的策略来进行探索优化:

$$\pi_t \sim \mathcal{N} \left(\widehat{M}_t(\bar{m} - X_t), \widehat{\theta}^2 \right), \quad t \in [0, T],$$

其中 $\widehat{M} = \{\widehat{M}_t\}_{t \in [0, T]} \in C([0, T]; \mathbb{R})$, 方差参数 $\widehat{\theta} \in \mathbb{R}^+$ 是常数. 于是, 极限时的目标函数 $J(\cdot)$ 可由参数 $(\widehat{M}, \widehat{\theta}^2)$ 完全确定, 下面的结论给出了目标函数取值的具体形式, 其在下节中将被用于计算给定参数下目标函数的值, 从而刻画相对误差.

推论 3.2 对任意 $\widehat{M} = \{\widehat{M}_t\}_{t \in [0, T]} \in C([0, T]; \mathbb{R})$ 和 $\widehat{\theta} \in \mathbb{R}^+$, 若控制策略 $\pi = \{\pi_t\}_{t \in [0, T]}$ 满足:

$$\pi_t \sim \mathcal{N} \left(\widehat{M}_t(\bar{m} - X_t^\pi), \widehat{\theta}^2 \right), \quad t \in [0, T], \quad (3.7)$$

其中 $X^\pi = \{X_t^\pi\}_{t \in [0, T]}$ 是 (π, \bar{m}) 下状态方程 (3.1) 的解. 于是, 目标函数 (3.1) 具有以下形式:

$$J(\widehat{M}, \widehat{\theta}^2) = \int_0^T \left(\left(1 + b\widehat{M}_t^2 \right) \left(f(t; \widehat{M}) - \bar{m}^2 \right) + b\widehat{\theta}^2 - \frac{\lambda}{2} \log \left(2\pi e\widehat{\theta}^2 \right) \right) dt + \gamma \left(f(T; \widehat{M}) - \bar{m}^2 \right) \quad (3.8)$$

这里, 对任意 $t \in [0, T]$, $f(t; \widehat{M}) := \mathbb{E}[(X_t^\pi)^2]$, 且具有以下形式:

$$f(t; \widehat{M}) = (\mathbb{E}[X_0^2] - \bar{m}^2)e^{-2g(t)} + \bar{m}^2 + \sigma^2 e^{-2g(t)} \int_0^t e^{2g(s)} ds,$$

其中 $g(t) = \int_0^t (a + \widehat{M}_s) ds$.

证 令状态方程 (3.1) 中的参数 $m = (m_t)_{t \in [0, T]}$ 恒等于 \bar{m} , 并代入控制策略 (3.7), 从而 (π, \bar{m}) 下的状态过程 $X^\pi = \{X_t^\pi\}_{t \in [0, T]}$ 满足 $dX_t^\pi = (a + \widehat{M}_t)(\bar{m} - X_t^\pi) dt + \sigma dW_t$, $X_t^\pi = X_0$. 于是, 由线性 SDE 解的解析表达式 (参阅 Protter[14] 中定理 V.52), 对任意 $t \in [0, T]$,

$$X_t^\pi = (X_0 - \bar{m})e^{-g(t)} + \bar{m} + \sigma \int_0^t e^{g(s)-g(t)} dW_s. \quad (3.9)$$

等式两边同时取期望, 发现在选取具有形式 (3.7) 的控制下, 状态过程的期望 $\mathbb{E}[X_t^\pi] \equiv \bar{m}$. 进一步, 对方程 (3.9) 两边同时平方再取期望, 则有

$$f(t; \widehat{M}) := \mathbb{E}[X_t^2] = (\mathbb{E}[X_0^2] - \bar{m}^2)e^{-2g(t)} + \bar{m}^2 + \sigma^2 e^{-2g(t)} \int_0^t e^{2g(s)} ds.$$

将 $f(t; \widehat{M})$ 和 $\mathbb{E}[X_t^\pi]$ 代入目标函数 (3.1) 即得 (3.8), 推论得证.

4 数值模拟

本节应用上一节中的结论来设计求得最优控制的算法, 并进行模拟验证. 首先, 将时间 T 离散为 K 份, 即步长为 $\delta = \frac{T}{K}$. 进一步, 注意到定理 3.1 给出了平均场均衡策略中 $m_t^* \equiv \bar{m} = \mathbb{E}[X_0]$, 故只需考虑 $m^* = \{m_t^*\}_{t \in [0, T]}$ 下代表性企业的状态方程 (3.1). 于是, 相应的时间离散后的方程为: 对 $t = 0, 1, \dots, K-1$,

$$X_{t+1}^\pi = X_t^\pi + \left(a(\bar{m} - X_t^\pi) + \int_U u \pi_t(du) \right) \delta + \sigma \Delta W_t, \quad X_0^\pi = X_0 \sim \nu, \quad (4.1)$$

其中 ΔW_t , $t = 0, \dots, K-1$ 相互独立且都服从正态分布 $\mathcal{N}(0, \delta)$, 离散控制策略 $\pi = \{\pi_t\}_{t=0}^{K-1}$ 满足 $\pi_t \in \mathcal{P}(U)$. 另一方面, 定理 3.1 和推论 3.2 表明, 仅需考虑具有形如 (3.7) 的控制策略, 从而时间离散后的目标函数可以看作只依赖于参数 $\widehat{M} = \{\widehat{M}_t\}_{t=0}^{K-1}$ 和 $\widehat{\theta}^2$ 的函数. 定义参数 $\widehat{R} := (\widehat{M}, \widehat{\theta}^2)$, 则离散情形下的目标函数可表示为

$$J(\widehat{R}) = \mathbb{E} \left[\sum_{t=0}^{K-1} \left((X_t^\pi - \bar{m})^2 + b \int_U u^2 \pi_t(du) - \lambda H(\pi_t) \right) \delta + \gamma (X_K^\pi - \bar{m})^2 \right]. \quad (4.2)$$

然而, 在强化学习的设定下, 模型参数 a, σ, b, γ 和目标函数 (4.2) 的取值均为未知, 决策者只能观测到自身的样本轨道 $\{x_t\}_{t=0}^K$ 和相应的模拟损失函数:

$$\widehat{j}(\widehat{R}) = \sum_{t=0}^{K-1} \left((x_t - \bar{m})^2 + b \int_U u^2 \pi_s(du) - \lambda H(\pi_s) \right) \delta + \gamma (x_K - \bar{m})^2. \quad (4.3)$$

基于此, 本节给出一个通过迭代寻求最优参数 \widehat{R} 的算法.

算法

- 1: 输入: 给定初始分布 ν , 初始控制参数 \widehat{R}^0 , 样本数量 n , 光滑参数 r , 学习率 η 以及优化轮数 I .
- 2: 对 $i \in \{0, \dots, I-1\}$ 做
- 3: 对 $j \in \{1, \dots, n\}$ 做
- 4: 选取控制参数 $\widehat{R}^{i,j} = \widehat{R}^i + U^{i,j}$, 这里 $U^{i,j} \in \mathbb{R}^{K+1}$ 是在所有满足 $\|U\|_F = r$ 的向量中均匀选取的向量, 其中 $\|\cdot\|_F$ 为 Frobenius 范数.
- 5: 求得参数 $\widehat{R}^{i,j}$ 下状态过程的一条样本所对应的模拟损失函数 $\widehat{j}(\widehat{R}^{i,j})$.
- 6: 对 j 循环结束.
- 7: 求得目标函数 (关于参数) 梯度 $\nabla J(\widehat{R}^i)$ 的估计量 $\widehat{\nabla J}(\widehat{R}^i) = \frac{1}{n} \sum_{j=1}^n \frac{1}{r^2} \widehat{j}(\widehat{R}^{i,j}) U^{i,j}$.
- 8: 使用梯度下降法通过估计量 $\widehat{\nabla J}(\widehat{R}^i)$ 更新参数: $\widehat{R}^{i+1} = \widehat{R}^i - \eta \widehat{\nabla J}(\widehat{R}^i)$.
- 9: 对 i 循环结束.
- 10: 得到 \widehat{R}^I .

上述算法的核心部分就是决策者对参数 $\widehat{R} := (\widehat{M}, \widehat{\theta}^2)$ 的控制下的样本进行关于 \widehat{j} 反馈的零阶优化. 由定理 3.1, 真实的最优策略参数 $R^* = (\zeta/2b, \lambda/2b)$, 其中 $\zeta = \{\zeta_t\}_{t \in [0, T]}$ 的定义见 (3.2). 则策略 \widehat{R} 下的相对误差定义为

$$\text{Err}(\widehat{R}) := \frac{|J(\widehat{R}) - J(R^*)|}{|J(R^*)|}. \quad (4.4)$$

于是, 通过 $\text{Err}(\widehat{R})$ 可以检验算法得到结果的有效性.

接下来, 这里给出一个算法实验的结果. 设模型参数 $a = 2$, $b = 1$, $\sigma = 1$, $T = 0.1$, $\delta = 0.02$, $\gamma = 1$, $\lambda = 2$, $X_0 \sim \mathcal{N}(0.2, 1)$, 则 $\bar{m} = 0.2$. 取算法中的参数 $r = 2$, $\eta = 0.1$, $\widehat{R}^0 = (0.3, 0.3, 0.3, 0.3, 0.3, 0.3)$, $n = 500$, $I = 4000$, 所得结果如下:

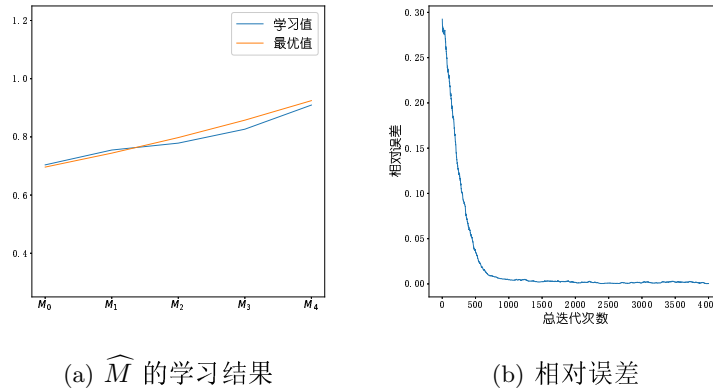


图 1 实验结果

其中图 1(a) 给出了输出的参数值 \widehat{M} 与最优值 $\zeta/2b$, 而参数 $\widehat{\theta} = 1.012$, 最优值 $\theta = 1$, 表明学习到的参数与真实的最优值较为接近. 图 1(b) 给出了相对误差 $\text{Err}(\widehat{R})$ 在迭代中的变化情况. 在本实验中, 随着策略 \widehat{R}^i 的迭代, 相对误差会骤降且最终会稳定在 0.2% 以下, 表明所

给出的算法较为有效. 在连续时间线性二次模型中, Jia and Zhou[15] 提出了 Actor-Critic 算法, 相较于该篇文章中的模拟结果, 本文所提出算法得到结果的相对误差显著较低, 且收敛速度更快. 而强化学习在平均场博弈问题的应用方面, Guo 等 [13] 在算法的最外层套了一层对平均场均衡策略 m 的循环以期收敛至 MFE (平均场均衡) m^* . 但本文严格证明了均衡策略 m^* 恒为常数 \bar{m} (具体参见定理 3.1), 故本文所提出算法的时间复杂度显然较小.

5 近似纳什均衡

本节通过前两节中的结论, 构造模型中参数未知时银行数量 $N \rightarrow \infty$ 的近似纳什均衡.

具体地, 对任意 $i = 1, \dots, N$, 定义只与第 i 个银行自身状态有关的分散策略 $\pi^{i,*} = \{\pi_t^{i,*}\}_{t \in [0, T]}$: 对任意 $t \in [0, T]$,

$$\pi_t^{i,*} \sim \mathcal{N}(\widehat{M}_t(\bar{m} - X_t^{i,*}), \widehat{\theta}^2), \quad (5.1)$$

其中参数 $\widehat{R} := (\widehat{M}, \widehat{\theta}^2)$ 为上一节中通过强化学习算法得到的参数未知的情况下, 代表银行的近似最优策略. $X^{i,*} = \{X_t^{i,*}\}$ 为策略 $\pi^{*,N} = (\pi^{1,*}, \dots, \pi^{N,*})$ 下状态方程 (2.1) 的解, 即其满足: 对任意 $i = 1, \dots, N$,

$$dX_t^{i,*} = a(\bar{X}_t^{*,N} - X_t^{i,*}) dt + \left[\int_U u \pi_t^{i,*}(du) \right] dt + \sigma dW_t^i, \quad X_0^{i,*} = X_0^i, \quad (5.2)$$

其中样本均值 $\bar{X}_t^{*,N} = \frac{1}{N} \sum_{i=1}^N X_t^{i,*}$. 注意, 这里为了符号简便, 我们省略了所构造策略 $\pi^{i,*}$ 与相应的状态过程 $X^{i,*}$ 以及市场中的银行数量 N 的相关性. 于是, 基于目标函数 (2.3), 本文的主要结论如下:

定理 5.3 对任意 $i = 1, \dots, N$, 由 (5.1) 定义的策略 $\pi^{*,N}$ 满足:

$$J^{i,N}(\pi^{*,N}) - \inf_{\pi^i \in \mathcal{D}} J^{i,N}(\pi^i, \pi^{*, -i}) \leq \varepsilon_N + \text{Err}(\widehat{R})J(R^*), \quad (5.3)$$

其中 $\lim_{N \rightarrow \infty} \varepsilon_N = 0$. $\text{Err}(\widehat{R})$ 是由 (4.4) 定义的参数 \widehat{R} 下的相对误差, $J(R^*)$ 是代表银行控制问题 (3.1) 的真实最优值.

不同于经典的平均场博弈问题, 由于模型中参数未知, 故无法通过平均场均衡策略构造逼近纳什均衡, 所构造策略对应的目标函数值比参数已知时的逼近策略稍大. 然而由于上一节中输出参数 \widehat{R} 对应的相对误差 $\text{Err}(\widehat{R})$ 稳定在 0.2% 以下, 故上述定理表明, 本节所构造的策略 $\pi^{i,*}$ 在参数未知时是纳什均衡的较好近似.

为证明本文的主要定理, 我们首先给出几个技术性引理. 下面的结论表明, 当每家银行采用由 (5.1) 确定的策略 $\pi^{i,*}$ 时, 样本均值 $\bar{X}_t^{*,N}$ 在 L^2 意义下关于 $t \in [0, T]$ 一致地收敛于 $\bar{m} := \mathbb{E}[X_0^i]$.

引理 5.4 对任意 $i = 1, \dots, N$, 设第 i 家银行采用形如 (5.1) 的控制策略 $\pi^{i,*} = \{\pi_t^{i,*}\}_{t \in [0, T]}$. 于是,

$$\lim_{N \rightarrow \infty} \mathbb{E} \left[\sup_{t \in [0, T]} |\bar{X}_t^{*,N} - \bar{m}|^2 \right] = 0,$$

其中 $\bar{X}_t^{*,N} = \frac{1}{N} \sum_{i=1}^N X_t^{i,*}$, $\{X_t^{i,*}\}_{1 \leq i \leq N}$ 满足 SDE (5.2).

证 方程 (5.2) 两边同时对 $i = 1, \dots, N$ 求和并除以 N , 则有

$$d\bar{X}_t^{*,N} = \widehat{M}_t(\bar{m} - \bar{X}_t^{*,N})dt + \frac{\sigma}{N} \sum_{i=1}^N dW_t^i. \quad (5.4)$$

上述线性方程的解有如下解析形式: 对任意 $t \in [0, T]$,

$$\bar{X}_t^{*,N} = \bar{m} + (\bar{X}_0^{*,N} - \bar{m})e^{-\int_0^t \widehat{M}_s ds} + \frac{\sigma}{N} \sum_{i=1}^N \int_0^t e^{-\int_s^t \widehat{M}_u du} dW_s^i.$$

注意到平方可积的初始随机变量列 X_0^i 相互独立且均值为 \bar{m} , 参数 $\widehat{M} = \{\widehat{M}_t\}_{t \in [0, T]}$ 是 $[0, T]$ 上的连续函数. 由 Doob 不等式即可得到

$$\mathbb{E} \left[\sup_{t \in [0, T]} |\bar{X}_t^{*,N} - \bar{m}|^2 \right] \leq C \mathbb{E} |\bar{X}_0^{*,N} - \bar{m}|^2 + C \frac{1}{N} \int_0^T \sigma^2 e^{-2\int_s^T \widehat{M}_u du} ds,$$

C 只依赖 \widehat{M} , 从而引理得证.

为证明 (5.3), 考虑目标泛函在可允许策略集合 \mathcal{D} 上的最小值等价于只考虑在以下集合上的最小值:

$$\mathcal{A}^{i,N} := \{\pi \in \mathcal{D} : J^{i,N}(\pi, \boldsymbol{\pi}^{*,N,-i}) \leq J^{i,N}(\boldsymbol{\pi}^{*,N})\}. \quad (5.5)$$

这里首先给出 $\mathcal{A}^{i,N}$ 中元素的一个估计结果.

引理 5.5 对任意 $\pi \in \mathcal{A}^{i,N}$, 存在与 i, N 无关的常数 $D > 0$, 使得:

$$\sup_{i,N} \sup_{\pi \in \mathcal{A}^{i,N}} \int_0^T \left(\int_U u^2 \pi_t(du) \right) dt \leq D.$$

证 对任意 $\pi \in \mathcal{A}^{i,N}$, 和 $\boldsymbol{\pi}^{*,N,-i}$, 由目标函数 (2.3) 的形式, 首先有

$$J^{i,N}(\pi, \boldsymbol{\pi}^{*,N,-i}) \geq \int_0^T \left(b \int_U u^2 \pi_t(du) - \lambda H(\pi_t) \right) dt,$$

其中 $H(\pi_t)$ 表示概率测度 π_t 的香农熵, 具体定义参见 (2.4). 又因为相同均值和方差下, 正态分布的香农熵最大, 故有

$$\begin{aligned} & J^{i,N}(\pi, \boldsymbol{\pi}^{*,N,-i}) \\ & \geq \int_0^T \left[b \int_U u^2 \pi_t(du) - \lambda \frac{1}{2} \log \left(2\pi e \left(\int_U u^2 \pi_t(du) - \left(\int_U u \pi_t(du) \right)^2 \right) \right) \right] dt \\ & \geq \int_0^T \left(b \int_U u^2 \pi_t(du) - \lambda \frac{1}{2} \log \left(\int_U u^2 \pi_t(du) \right) \right) dt. \end{aligned}$$

进一步, 再由琴生不等式可得

$$J^{i,N}(\pi, \boldsymbol{\pi}^{*,N,-i}) \geq b \int_0^T \int_U u^2 \pi_t(du) dt - \lambda \frac{T}{2} \log \left(\frac{1}{T} \int_0^T \int_U u^2 \pi_t(du) dt \right).$$

而 $J^{i,N}(\pi, \pi^{*,N,-i}) \leq J^{i,N}(\pi^{*,N})$, 由 SDE 解的矩估计 (参见 Krylov[16] 中第 2 章第 5 节), 可得 $J^{i,N}(\pi^{*,N})$ 有与 i, N 无关的上界 L . 这样 $\int_0^T (\int_U u^2 \pi_t(du)) dt$ 有与 i, N 无关的上界 D , 否则 $\int_0^T (\int_U u^2 \pi_t(du)) dt$ 足够大时会使上式右边大于 L .

下面开始证明本文的主要结论.

定理 5.3 的证明 由 $\mathcal{A}^{i,N}$ 的定义 (5.5), 我们有

$$\begin{aligned} & J^{i,N}(\pi^{*,N}) - \inf_{\pi^i \in \mathcal{D}} J^{i,N}(\pi^i, \pi^{*, -i}) = J^{i,N}(\pi^{*,N}) - \inf_{\pi^i \in \mathcal{A}^{i,N}} J^{i,N}(\pi^i, \pi^{*, -i}) \\ & \leq |J^{i,N}(\pi^{*,N}) - J(\widehat{R})| + |J(\widehat{R}) - J(R^*)| + |J(R^*) - \inf_{\pi^i \in \mathcal{D}} J(\pi^i; \bar{m})| \\ & \quad + (\inf_{\pi^i \in \mathcal{D}} J(\pi^i; \bar{m}) - \inf_{\pi^i \in \mathcal{A}^{i,N}} J^{i,N}(\pi^i, \pi^{*, -i})) \vee 0 \\ & := I_1^{(N)} + I_2 + I_3 + I_4^{(N)}, \end{aligned}$$

其中 $J(\pi^i; \bar{m})$ 是代表银行的目标函数 (3.1), $J(\widehat{R})$ 是当控制过程具有形式 (3.7) 时, 代表银行目标函数的取值, 其可看作只与参数 $\widehat{R} = (\widehat{M}, \widehat{\theta}^2)$ 有关的函数 (参见推论 3.2), $R^* = (\zeta/2b, \lambda/2b)$ 表示由定理 3.1 得到的最优控制的相应参数.

定义辅助过程 $Y^{i,*} = \{Y_t^{i,*}\}_{t \in [0, T]}$, 其满足如下 SDE: $i = 1, \dots, N$

$$dY_t^{i,*} = a(\bar{m} - Y_t^{i,*}) dt + \left[\int_U u \widehat{\pi}_t^{i,*}(du) \right] dt + \sigma dW_t^i, \quad Y_0^{i,*} = X_0^i \sim \nu, \quad (5.6)$$

控制策略 $\widehat{\pi}^{i,*} = \{\widehat{\pi}_t^{i,*}\}_{t \in [0, T]}$, $\widehat{\pi}^{i,*}$ 与所构造策略 (5.1) 具有相同的反馈控制函数, 即对任意 $t \in [0, T]$,

$$\widehat{\pi}_t^{i,*} \sim \mathcal{N}(\widehat{M}_t(\bar{m} - Y_t^{i,*}), \widehat{\theta}^2),$$

其中, 参数 $(\widehat{M}, \widehat{\theta}^2)$ 与 (5.1) 中的定义相同, 是上一节中通过强化学习算法得到的参数. 于是, 由引理 5.4 和线性 SDE 解的稳定性定理 (参见 [14] 中定理 V.9), 我们有

$$\lim_{N \rightarrow \infty} \mathbb{E} \left[\sup_{t \in [0, T]} |X_t^{i,*} - Y_t^{i,*}|^2 \right] = 0. \quad \text{结合引理 5.4, 即有}$$

$$\lim_{N \rightarrow \infty} I_1^{(N)} = 0. \quad (5.7)$$

另一方面, 由相对误差的定义 (4.4) 和定理 3.1,

$$I_2 = \text{Err}(\widehat{R})J(R^*), \quad I_3 = 0. \quad (5.8)$$

对于最后一项 $I_4^{(N)}$, 首先注意到:

$$I_4^{(N)} = (\inf_{\pi^i \in \mathcal{D}} J(\pi^i; \bar{m}) - \inf_{\pi^i \in \mathcal{A}^{i,N}} J^{i,N}(\pi^i, \pi^{*, -i})) \vee 0 \leq \sup_{\pi^i \in \mathcal{A}^{i,N}} |J(\pi^i; \bar{m}) - J^{i,N}(\pi^i, \pi^{*, -i})|.$$

对任意 $\pi^i \in \mathcal{A}^{i,N}$, 为符号简便, 我们仍用 $X^j = \{X_t^j\}, j = 1, \dots, N$ 表示策略 $(\pi^i, \pi^{*, -i})$ 下状态方程 (2.1) 的解, 并记该策略下的样本均值为 $\bar{X}_t^{*, -i} := \frac{1}{N} \sum_{j=1}^N X_t^j$. 辅助过程 $Y^i = \{Y_t^i\}_{t \in [0, T]}$ 满足如下 SDE:

$$dY_t^i = a(\bar{m} - Y_t^i) dt + \left[\int_U u \pi_t^i(du) \right] dt + \sigma dW_t^i, \quad Y_0^i = X_0^i \sim \nu,$$

令 $\bar{Y}_t^{*, -i} := \frac{1}{N} Y_t^i + \frac{1}{N} \sum_{j \neq i} Y_t^{j, *}$ 这样有

$$X_t^i - Y_t^i = \int_0^t (a(\bar{X}_s^{*, -i} - \bar{m}) + a(Y_s^i - X_s^i)) ds.$$

由 Hölder 不等式, 均值不等式可得

$$\mathbb{E} \left[\sup_{0 \leq s \leq t} |X_s^i - Y_s^i|^2 \right] \leq 2aT \mathbb{E} \left[\int_0^t (\bar{X}_s^{*, -i} - \bar{m})^2 ds \right] + 2aT \int_0^t \mathbb{E} \left[\sup_{0 \leq u \leq s} (Y_u^i - X_u^i)^2 \right] ds.$$

由 Gronwall 不等式:

$$\mathbb{E} \left[\sup_{0 \leq s \leq t} |X_s^i - Y_s^i|^2 \right] \leq C \mathbb{E} \left[\int_0^t (\bar{X}_s^{*, -i} - \bar{m})^2 ds \right].$$

这里及之后的 $C > 0$ 表示不依赖于 i 和 N 的常数, 在不同位置的取值可能不同. 当 $j \neq i$ 时, 类似可得

$$\mathbb{E} \left[\sup_{0 \leq s \leq t} |X_s^j - Y_s^{j, *}|^2 \right] \leq C \mathbb{E} \left[\int_0^t (\bar{X}_s^{*, -i} - \bar{m})^2 ds \right].$$

于是, 由均值不等式:

$$\mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{X}_s^{*, -i} - \bar{Y}_s^{*, -i}|^2 \right] \leq C \mathbb{E} \left[\int_0^t (\bar{X}_s^{*, -i} - \bar{m})^2 ds \right],$$

从而有

$$\begin{aligned} \mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{X}_s^{*, -i} - \bar{m}|^2 \right] &\leq 2\mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{X}_s^{*, -i} - \bar{Y}_s^{*, -i}|^2 \right] + 2\mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{Y}_s^{*, -i} - \bar{m}|^2 \right] \\ &\leq C \mathbb{E} \left[\int_0^t (\bar{X}_s^{*, -i} - \bar{m})^2 ds \right] + 2\mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{Y}_s^{*, -i} - \bar{m}|^2 \right] \\ &\leq C \int_0^t \mathbb{E} \left[\sup_{0 \leq u \leq s} |\bar{X}_u^{*, -i} - \bar{m}|^2 \right] ds + 2\mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{Y}_s^{*, -i} - \bar{m}|^2 \right]. \end{aligned}$$

由 Gronwall 不等式:

$$\mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{X}_s^{*, -i} - \bar{m}|^2 \right] \leq C \mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{Y}_s^{*, -i} - \bar{m}|^2 \right].$$

考虑 $\bar{Y}_s^{*, -i}$, 有

$$\mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{Y}_s^{*, -i} - \bar{m}|^2 \right] \leq \mathbb{E} \left[\frac{2}{N^2} \sup_{0 \leq s \leq t} |Y_s^i - \bar{m}|^2 \right] + \mathbb{E} \left[\frac{2}{N^2} \sup_{0 \leq s \leq t} \left| \sum_{j \neq i}^N (Y_s^{j, *} - \bar{m}) \right|^2 \right],$$

由 SDE 解的矩估计以及引理 5.5 可以得到 $\sup_{i, N} \sup_{\pi^i \in \mathcal{A}^{i, N}} \mathbb{E} \left[\sup_{0 \leq s \leq t} |Y_s^i - \bar{m}|^2 \right] < \infty$. 因此,

$$\lim_{N \rightarrow \infty} \sup_{1 \leq i \leq N} \sup_{\pi^i \in \mathcal{A}^{i, N}} \mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{Y}_s^{*, -i} - \bar{m}|^2 \right] = 0.$$

记

$$G_N := C \sup_{1 \leq i \leq N} \sup_{\pi^i \in A^{i,N}} \mathbb{E} \left[\sup_{0 \leq s \leq T} |\bar{Y}_s^{*, -i} - \bar{m}|^2 \right],$$

则 $\lim_{N \rightarrow \infty} G_N = 0$, 因此, 前面所述 $\mathbb{E} \left[\sup_{0 \leq s \leq t} |\bar{X}_s^{*, -i} - \bar{m}|^2 \right] \leq G_N$ 且 $\mathbb{E} \left[\sup_{0 \leq s \leq t} |X_s^i - Y_s^i|^2 \right] \leq CG_N$. 因此可得

$$|J^{i,N}(\pi^i, \pi^{*, -i}) - J(\pi^i)| \leq CG_N, \quad \forall \pi^i \in A^{i,N}.$$

此即 $\lim_{N \rightarrow \infty} I_4^{(N)} = 0$. 令 $\varepsilon_N = I_1^{(N)} + I_4^{(N)}$, 结合 (5.7), (5.8) 即得 (5.3). 定理得证.

参 考 文 献

- [1] Fouque J P, Langsam J A. Handbook on systemic risk[M]. Cambridge: Cambridge University Press, 2013.
- [2] Zhang Xiaoming, Zhang Xinsong, Lee Chien-Chiang, et al. Measurement and prediction of systemic risk in China's banking industry[J]. Research in International Business and Finance, 2023, 64: 101874.
- [3] Fouque J P, Ichiba T. Stability in a model of interbank lending[J]. SIAM Journal on Financial Mathematics, 2013, 4(1): 784–803.
- [4] Sun Li-Hsien. Systemic risk illustrated[A]. Fouque J P. Handbook on systemic risk[M]. Cambridge: Cambridge University Press, 2013: 444–452.
- [5] Bo Lijun, Capponi A. Systemic risk in interbanking networks[J]. SIAM Journal on Financial Mathematics, 2015, 6(1): 386–424.
- [6] Carmona R, Fouque J P, Sun Li-Hsien. Mean field games and systemic risk[J]. Communications in Mathematical Sciences, 2015, 13(4): 911–933.
- [7] Bo Lijun, Li Tongqing, Yu Xiang. Centralized systemic risk control in the interbank system: Weak formulation and gamma-convergence[J]. Stochastic Processes and their Applications, 2022, 150: 622–654.
- [8] Huang Minyi, Caines P E, Malham é R P. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and thenash certainty equivalence principle[J]. Communications in Information and Systems, 2006, 6: 221–252.
- [9] Lasry J M, Lions P L. Mean field games[J]. Japanese Journal of Mathematics, 2007, 2: 229–260.
- [10] Carmona R, Delarue F. Probabilistic analysis of mean-field games[J]. SIAM Journal on Control and Optimization, 2013, 51(4): 2705–2734.
- [11] Cardaliaguet P, Delarue F, Lasry J M, et al. The master equation and the convergence problem in mean field games[M]. Princeton: Princeton University Press, 2019.
- [12] Wang Haoran, Zariphopoulou T, Zhou Xunyu. Reinforcement learning in continuous time and space: A stochastic control approach[J]. Journal of Machine Learning Research, 2020, 21(1): 8145–8178.
- [13] Guo Xin, Xu Renyuan, Zariphopoulou T. Entropy regularization for mean field games with learning[J]. Mathematics of Operations Research, 2022, 47(4): 3239–3260.
- [14] Protter E P. Stochastic integration and differential equations[M]. Springer Berlin, Heidelberg, 2005.
- [15] Jia Yanwei, Zhou Xunyu. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms[J]. Journal of Machine Learning Research, 2022, 23(275): 1–50.

[16] Krylov N V. Controlled diffusion processes[M]. Springer Berlin, Heidelberg, 2008.

REINFORCEMENT LEARNING METHODS FOR BANK CURRENCY RESERVE GAMES

LI Ce

(*School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, China*)

Abstract: In large-scale bank interaction systems, individual banks can adjust their borrowing and lending rates with the central bank to bring their currency reserves as close as possible to the sample mean, thereby reducing the probability of systemic risk. However, when the state process and parameters of the objective function are unknown, it is not directly possible to solve the stochastic differential game problem to obtain a Nash equilibrium. In this study, we combined mean-field game theory with relevant methods from continuous-time reinforcement learning to construct an approximate Nash equilibrium in a large-scale bank lending network. First, by solving the forward-backward coupled HJB-FPK equation, we obtained the mean-field equilibrium strategy representing the banks. Next, based on the form of the obtained strategy, we designed an iterative parameter method to characterize the approximate optimal strategy when parameters are unknown. Finally, using the learned parameters, we constructed an approximate Nash equilibrium for a large number of banks.

Keywords: systemic risk; Reinforcement learning; approximate Nash equilibrium; Mean field games

2010 MR Subject Classification: 60H10; 91A15