Pure Mathematics 理论数学, 2026, 16(4), 184-194
Published Online April 2026 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2026.164103

Hans汉斯

多种返利机制下做市商问题的强化学习求解

林雪琪

广东工业大学数学与统计学院，广东 广州

收稿日期：2026年3月13日；录用日期：2026年4月11日；发布日期：2026年4月23日

摘 要

本文基于Avellaneda-Stoikov做市商模型，构建了一个同时包含常值返利、时间依赖返利和状态依赖返利的统一分析框架，并通过离散化将模型转化为马尔可夫决策过程，采用DDQN、PPO和A2C三种强化学习算法求解。数值结果表明，在无返利的简单环境下，PPO与A2C的收益水平与Avellaneda-Stoikov解析基线较为接近，而DDQN的波动和库存风险相对更高；在常值返利环境中，PPO的净收益整体较高，但启发式策略同样具有较强竞争力；在状态依赖返利环境中，A2C相比启发式策略表现出更强的自适应能力。进一步的敏感性分析说明，返利机制不仅影响做市商的利润来源，也改变了报价、成交与库存控制之间的动态权衡。

关键词

做市商模型，返利机制，强化学习，状态依赖奖励

Reinforcement Learning Solution to the Market Maker Problem under Multiple Rebate Mechanisms

Xueqi Lin

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

Received: March 13, 2026; accepted: April 11, 2026; published: April 23, 2026

Abstract

This paper develops a unified market-making framework based on the Avellaneda-Stoikov model by incorporating constant, time-dependent, and state-dependent rebate mechanisms. After discretization, the problem is formulated as a Markov decision process and solved with DDQN, PPO, and A2C. The numerical results indicate that, in simple no-rebate settings, PPO and A2C achieve

文章引用: 林雪琪. 多种返利机制下做市商问题的强化学习求解[J]. 理论数学, 2026, 16(4): 184-194.
DOI: 10.12677/pm.2026.164103

profit levels close to the Avellaneda-Stoikov analytical benchmark, while DDQN exhibits relatively higher volatility and inventory risk. Under constant rebate mechanisms, PPO attains the highest net profit on average, although heuristic policies remain competitive. Under state-dependent rebates, A2C shows a clearer adaptive advantage over heuristic baselines. The sensitivity analysis further suggests that rebate mechanisms affect not only the source of profit, but also the dynamic trade-off among quoting behavior, execution, and inventory control.

Keywords

Market Making, Rrebate Mechanism, Reinforcement Learning, State-Dependent Incentive

1. 引言

在金融市场中，做市商通过提供连续的买卖报价来赚取价差并增强市场流动性，其策略对于市场稳定性和效率至关重要。Harold Demsetz [1]最早对做市商问题进行研究，提出了买卖价差模型。Garman [2]将库存风险纳入做市商模型中，为后续 Avellaneda 和 Stoikov [3]提出同时考虑价格和流动性风险的经典模型，并利用 Hamilton-Jacobi-Bellman (HJB)方程来推导最优报价策略的解析解奠定基础。然而，这些模型往往难以捕捉市场的高度动态性和复杂性，尤其是在高频交易环境下。

近年来，随着算法交易的兴起，研究者开始探索更为先进的方法来改进做市策略。例如，Guéant [4]等人通过解决常微分方程来推导解析解，而 Ait-Sahalia 和 Sağlam [5]则引入了跳跃过程来模拟价格变动。这些研究在 Avellaneda-Stoikov 模型的基础上进行了有意义的扩展，为做市策略的优化提供了新的视角。

强化学习作为一种数据驱动的决策方法，为做市策略的研究带来了革命性的变化。它能够直接从市场数据中学习最优策略，无需依赖复杂的数学建模。Bruno Gašperov 和 Zvonko Kostanj [6] [7]等学者运用强化学习技术在以随机过程为假设的订单流环境中训练他们的做市策略模型；Spooner 和 Thoma [8]使用 TD 算法，说明强化学习方法在订单流由随机过程驱动的做市环境中具有可行性，并能够在一定程度上适应高频交易中的动态特征。

在本文中，我们考虑市场中多种返利机制对做市商行为的影响，构建一个同时包含常值返利、时间依赖返利和状态依赖返利的统一做市模型，并采用 DDQN、PPO 与 A2C 三种强化学习方法进行求解。本文通过数值实验说明在不同激励机制和市场条件下，强化学习方法给出了有效策略。

2. 构造返利机制下的做市商模型

Avellaneda-Stoikov 做市商模型[3]给出了有限时域 T 内做市商最优报价的解析解；在真实市场中，为了促进做市商交易为市场提供流动性，市场为做市商提供奖励机制以鼓励做市商报价，这种返利机制与做市商的交易量呈正相关。基于此，我们考虑单一资产在有限时域中含返利机制的做市商模型。在交易时域为$[0,T]$；我们假设短期股票价格 S_t 服从零漂移布朗运动：

$$dS_t = \sigma dW_t \tag{1}$$

其中 W_t 为标准布朗运动，$\sigma > 0$ 为波动率参数。做市商的目标是最大化一段时间内的效用，这里用 CARA 效用函数来度量：

$$u(C,S,Q,t) = -\exp(-\gamma C)\exp(-\gamma QS)\exp\left(\frac{\gamma^2 Q^2 \sigma^2 (T-t)}{2}\right) \tag{2}$$

C 是做市商的现金，S 是股票的价格，Q 是做市商拥有的库存，t 是做市商已交易的时间，γ 是做市商的风险偏好系数，σ^2 是股票交易的波动率。

市场外生市场订单流 O_t^a, O_t^b 满足泊松过程，其强度由流动性参数 λ 决定。在离散步长 Δt 下有：

$$O_t^a \sim \text{Poisson}(\lambda \Delta t), \quad O_t^b \sim \text{Poisson}(\lambda \Delta t) \tag{3}$$

做市商在时刻 t 选择买卖两侧相对中间价的报价偏移 $\delta_t^a, \delta_t^b \geq 0$，成交概率随报价距离呈指数衰减：

$$p_t^a = \mathrm{e}^{-k\delta_t^a}, \quad p_t^b = \mathrm{e}^{-k\delta_t^b} \tag{4}$$

其中 $k > 0$ 为成交敏感度参数。做市商在 t 时刻的账面财富定义为：

$$W_t = C_t + Q_t S_t \tag{5}$$

当做市商不断在电子簿上给出报价 p_t^a, p_t^b 并被市场订单击中时，做市商库存 Q_t 相应发生变化，变化过程为：

$$dQ_t = q_t^b + q_t^a \tag{6}$$

由于做市商参与市场为市场提供流动性，市场根据做市商的市场交易量 $q_t^a + a_t^b$ 给出返利。在我们的模型中，返利与成交量成正比且非单一。我们给出三种不同地返利单价情况：首先是常值返利，

$$r_t^a = r^a, \quad r_t^b = r^b \tag{7}$$

其次，我们考虑交易所激励在交易时段中逐步减弱的情形，将返利设为线性衰减形式：

$$r_t^a = \overline{r}^a\left(1 - \frac{t}{T}\right), \quad r_t^b = \overline{r}^b\left(1 - \frac{t}{T}\right) \tag{8}$$

最后，我们考虑交易所根据市场压力和波动水平动态调整激励，设订单流不平衡信号 $\iota_t \in [-1,1]$，波动修正项为：

$$\text{vol}_t = \frac{\sigma}{\sigma_{\text{ref}}} - 1 \tag{9}$$

此时返利函数写为：

$$r_t^b = \overline{r}^b\left(1 + \alpha\iota_t + \beta \text{vol}_t\right), \quad r_t^a = \overline{r}^a\left(1 - \alpha\iota_t + \beta \text{vol}_t\right) \tag{10}$$

由此，我们将奖励由常值奖励扩展到更复杂的市场背景下。当做市商在市场进行交易后，其现金流也相应发生变化。现金变化由做市商在买卖市场上的价差以及为市场提供流动性带来的返利共同构成。令 C_t 为现金头寸，

$$dC_t = \delta_t^b q_t^b + \delta_t^a q_t^a + r_t\left(q_t^b + q_t^a\right) \tag{11}$$

则账面财富动态 dW_t 可写为：

$$dW_t = Q_t\, dS_t + \left(\delta_t^a + r_t^a\right) dN_t^a + \left(\delta_t^b + r_t^b\right) dN_t^b \tag{12}$$

交易时段结束时，利用 W_T^{liq} 表示终端时刻财富，于是式(2)所表示的最大化期望指数效用函数可以表示为：

$$u(S,Q,t)=\max_{\delta^a,\delta^b}\mathbb{E}_t\left[-\exp\left(-\gamma\left(C_T+QS_T\right)\right)\right] \tag{13}$$

为了避免过度持仓带来的库存风险，我们在优化目标中引入二次库存惩罚项：

$$\ell\left(Q_t\right)=\eta Q_t^2,\quad \eta>0 \tag{14}$$

于是，做市商的目标是在有限时域内最大化终端清算财富减去库存运行成本：

$$\max_{\pi}\mathbb{E}^{\pi}\left[W_T^{\mathrm{liq}}-\int_0^T\eta Q_t^2\mathrm{d}t\right] \tag{15}$$

3. MDP 构造与强化学习求解

在引入时间依赖与状态依赖返利机制下，返利随阶段或市场状态变化，利用传统随机控制理论难以得到其解析解。而强化学习能够通过与环境互动，根据市场状态给出最优决策。本文将连续时间随机控制问题离散化为有限时域马尔可夫决策过程。将区间$[0,T]$划分为N个等长时段，记$t_n=n\Delta t$。离散状态定义为：

$$s_n=\left(t_n,S_n,Q_n,W_n,O_n^a,O_n^b\right), \tag{16}$$

其中包含当前时间、中间价、库存、账面财富以及买卖两侧订单流。

中间价在离散网格上用 Euler-Maruyama 采样：

$$S_{n+1}=S_n+\sigma\sqrt{\Delta t}\varepsilon_n,\quad \varepsilon_n\sim\mathcal{N}(0,1) \tag{17}$$

外生订单流满足：

$$O_n^a\sim\mathrm{Poisson}\left(\lambda\Delta t\right),\quad O_n^b\sim\mathrm{Poisson}\left(\lambda\Delta t\right) \tag{18}$$

给定报价偏移δ_n^a,δ_n^b，成交概率为：

$$p_n^a=\mathrm{e}^{-k\delta_n^a},\quad p_n^b=\mathrm{e}^{-k\delta_n^b} \tag{19}$$

实际成交量由二项抽样给出，因此在给定(s_n,a_n)的条件下，可以通过采样得到(s_{n+1},r_n)。

我们通过半价差与偏斜组合的参数化动作a_n：

$$a_n=\left(h_n,\mathrm{skew}_n\right). \tag{20}$$

进一步定义每个时间步t_n的价差$\delta_n\{a,b\}$：

$$\delta_n^b=h_n-\mathrm{skew}_n,\quad \delta_n^a=h_n+\mathrm{skew}_n \tag{21}$$

并要求$\delta_n^a,\delta_n^b\geq 0$。其中$h_n$控制整体报价宽度，$\mathrm{skew}_n$控制买卖两侧报价的非对称性。

在返利机制离散化后，三类返利分别写为：常值返利$r_n^{a,b}=\bar{r}^{a,b}$，时间衰减返利$r_n^{a,b}=\bar{r}^{a,b}\left(1-t_n/T\right)$，以及将式(10)中$t$替换成对应时间步$t_n$的状态依赖返利。执行动作$\delta_n^{a,b}$后，现金、库存与财富依次更新，本文定义一步奖励$r_n$为：

$$r_n=\left(W_{n+1}-W_n\right)-\eta Q_n^2\Delta t. \tag{22}$$

算法选择与统一训练框架

本文采用 DDQN、A2C 与 PPO 三种代表性方法。DDQN 适用于有限动作集合下的$Q(s,a)$学习，能够直接学习“状态–动作”价值排序，并通过经验回放提高样本效率。A2C 采用 Actor-critic 架构并利用优势函数降低策略梯度方差，更适合在高噪声环境下学习平滑决策。PPO 通过截断目标限制策略更新幅

度，当返利为时间依赖或状态依赖时，回报分布会随阶段或状态变化而更不稳定，PPO 的保守更新有助于降低训练震荡与策略退化风险，因此更适合跨机制对比实验。

在离散时刻 t_n 下，智能体首先观察状态 s_n，随后从离散动作集合中选择 (h_n, skew_n)，并由此确定双边报价偏移。环境根据价格扩散、泊松订单流和二项成交机制采样买卖市场的实际成交量，再结合返利函数计算现金更新、库存更新与账面财富更新，最终返回一步奖励与下一状态 (s_{n+1}, r_n)。对于 DDQN，本文利用经验回放与目标网络提升样本复用效率并缓解自举偏差。对于 A2C 与 PPO，则采用 Actor-Critic 结构，通过优势函数估计降低策略梯度方差。为确保跨机制比较具有一致性，我们设置三类算法共享同一环境接口与指标体系。

4. 实验设计

本节首先给出统一的环境设定与算法参数，然后说明市场环境与返利机制的分组方式，最后介绍评价指标及不同实验模块的比较对象。整体实验设计分为三个层面：其一，在无返利的简单常值环境下，将强化学习方法与 Avellaneda-Stoikov 解析基线进行比较，用以检验强化学习框架在经典 setting 下的合理性；其二，在常值返利、时间依赖返利和状态依赖返利环境中，引入经典非学习型启发式策略，与强化学习方法进行并列比较；其三，在此基础上进一步对关键市场参数进行敏感性分析，以考察不同方法对环境变化的响应特征。

我们将环境基准设置为：交易时域 $T = 1.0$，初始中间价 $S_0 = 100$，波动率参数 $\sigma = 2.0$，成交敏感度参数 $k = 1.5$，离散时间步数 $n_{\text{steps}} = 200$，库存惩罚系数 $\eta = 0.02$。动作空间采用离散网格构造，其中，

$$h \in [0,1], \quad \text{skew} \in [-1,1]$$

网格步长均为 0.05，在可行性约束 $\delta^a \geq 0, \delta^b \geq 0$ 下，共得到 441 个可行动作。

训练部分中，基准实验统一使用 1000 个训练回合与 50 个评估回合，敏感性分析使用 300 个训练回合，并复用与基准实验相同的 50 个评估种子。DDQN 采用经验回放与目标网络结构，学习率为 10^{-3}，折扣因子为 0.99，批大小为 128，经验池容量为 20,000，ε-greedy 探索从 1.0 衰减至 0.05。PPO 与 A2C 采用共享编码层的 actor-critic 网络，学习率均为 5×10^{-4}，折扣因子为 0.99。其中 PPO 的截断参数为 0.2，每回合执行 4 个策略更新轮次，而 A2C 每回合执行 1 轮更新。为保证实验可复现，全文统一使用全局随机种子 42。

在比较对象的设置上，本文对不同场景采用分层设计。对于无返利的简单常值环境，只将 DDQN、PPO、A2C 与 Avellaneda-Stoikov 解析基线进行比较，以考察强化学习框架在经典 setting 下的有效性。对于有返利的常值环境以及时间依赖、状态依赖返利环境，则进一步引入两类非学习型启发式策略：固定半价差策略 ConstantPolicy_h08 和线性偏斜策略 LinearSkewPolicy_h08。其中前者固定取 $h = 0.8$、$\text{skew} = 0$；后者设定 $h_0 = 0.8$、$\phi = 0.05$，并按

$$\text{skew}_t = \text{clip}(-\phi Q_t, -1, 1), \quad h_t = \max(h_0, |\text{skew}_t|)$$

生成状态相关的报价偏斜。这样的分层比较方式，一方面保留了解析基线在 simple 场景中的校准作用，另一方面也能够在更复杂的激励环境中评估强化学习方法相对于规则型策略的实际价值。

4.1. 市场环境与返利机制分组

基准实验中的常值返利机制设置了四类市场环境：(1) 无返利低流动性市场 NoRebate_lam120：$\lambda = 120$，返利为 0；(2) 低流动性高返利市场 LowLiq：$\lambda = 120$，基础返利为 0.4；(3) 无返利高流动性市场 NoRebate：$\lambda = 140$，返利为 0；(4) 高流动性低返利市场 HighLiq：$\lambda = 140$，基础返利为 0.2。

在非常值返利机制中，时间依赖返利与状态依赖返利均以 LowLiq 市场为基础展开，时间依赖返利采用线性衰减形式：

$$r_t = \bar{r}\left(1-\frac{t}{T}\right),$$

其中基础返利 $\bar{r}=0.4$。状态依赖返利则设定：

$$r_t^b = \bar{r}\left(1+\alpha_{\text{imb}} I_t + \beta_{\text{vol}}\text{vol}_t\right),\quad r_t^a = \bar{r}\left(1-\alpha_{\text{imb}} I_t + \beta_{\text{vol}}\text{vol}_t\right)$$

其中 $\alpha_{\text{imb}}=0.25$，$\beta_{\text{vol}}=0.15$，$\sigma_{\text{ref}}=2.0$，返利下界与上界分别为 0 与 0.6。

4.2. 评价指标

本文使用以下指标综合评价策略表现：

(1) **NetPnL**：考虑库存惩罚后的净收益，用于衡量策略的总体盈利水平；

(2) **NetPnL_std** 与 **NetPnL_q05**：分别衡量收益波动与左尾风险；

(3) **InvStd**：库存标准差，用于刻画库存控制效果；

(4) **AvgHalfSpread**：平均半价差，用于反映报价激进程度；

(5) **E [skew|Q]拟合斜率**：用于识别策略是否能够随着库存状态变化进行系统性偏斜调整，从而反映库存回归行为。

在 simple 无返利场景中，本文主要关注 NetPnL、库存波动与解析基线之间的相对位置；在有返利和动态激励场景中，则更加关注收益、风险与策略行为之间的权衡关系。这样的指标设置能够更全面地刻画不同方法在不同市场环境下的相对优劣。

4.3. 实验伪代码

根据前文的实验设计，实验流程采用分层比较结构：在无返利简单场景中，仅比较 DDQN、PPO、A2C 与 Avellaneda-Stoikov 解析基线。在常值返利、时间依赖返利和状态依赖返利场景中，则比较 DDQN、PPO、A2C 以及两类启发式策略。敏感性分析部分仍只针对三种强化学习算法展开。统一实验流程见算法 5.3。环境状态转移、成交抽样、返利计算与奖励函数仍按第 3 节模型定义执行，此处仅给出训练、比较与敏感性分析的逻辑框架。

Algorithm 1 统一返利机制下的实验流程框架

Input: reviewer 场景集合；主实验场景集合；待比较算法与评价指标

Reviewer block for 每个无返利 *simple* 场景 **do**
  训练 DDQN、PPO 与 A2C；校准 Avellaneda–Stoikov 解析基线；在相同评估种子下比较 RL 与 A–S；保存短表、逐回合结果与配对检验结果。

Main block for 每个返利机制场景 **do**
  训练 DDQN、PPO 与 A2C；在相同评估种子下比较 RL 与启发式策略；保存短表、NetPnL 对比图与行为曲线。

Sensitivity block for 每个机制 – 市场组合 **do**
  for 参数 $p \in \{\sigma, \lambda, k\}$ 的三点网格 **do**
    重新训练 DDQN、PPO 与 A2C；在固定评估种子下重新评估三种 RL；记录均值、95% 置信区间与配对检验结果。

Output: 无返利场景下 RL 与 A–S 的比较表；返利机制下 RL 与启发式策略的比较表与图；敏感性分析曲线、整体热图与统计检验结果。

6. 实验结果

6.1. 常值返利机制下的基准结果

首先考虑无返利的简单常值环境，其结果如表 1 所示。在 NoRebate_constant 场景下，Avellaneda-Stoikov 解析基线的平均净收益为 65.1261，略高于 PPO 与 A2C 的 64.7652，也高于 DDQN 的 56.0978。与此同时，解析基线的库存波动标准差仅为 2.4198，显著低于三种强化学习方法。进一步地，在 NoRebate_lam120_constant 场景下，Avellaneda-Stoikov 的平均净收益为 55.9232，A2C 以 54.4043 接近，而 DDQN 与 PPO 分别降至 30.3712 与 27.3690，且对应的左尾分位数和库存波动也更差。总体而言，这一组结果表明，在结构较简单、无返利的经典环境中，解析基线仍然是较强的参照，而 PPO 与 A2C 已能够达到与解析基线相近的收益量级，说明所构建的强化学习框架在该类 simple setting 下是有效的。相较之下，DDQN 对高噪声收益信号更为敏感，因此在该场景下表现相对偏弱。

Table 1. Comparison of RL and A-S baseline in simple scenarios without rebates
表 1. 无返利简单场景下 RL 与 A-S 基线的比较

市场场景	策略	NetPnL	Std	q05	InvStd	AvgHalfSpread	ESkewGivenQ-Slope
NoRebate	A-S	65.13	8.90	50.19	2.42	0.684	−0.0297
	PPO	64.77	18.38	29.67	4.26	0.650	0.0000
	A2C	64.77	18.38	29.67	4.26	0.650	0.0000
	DDQN	56.10	28.95	16.25	5.72	0.591	−0.0103
NoRebate-lam120	A-S	55.92	10.69	38.52	2.71	0.674	−0.0141
	A2C	54.40	18.49	24.92	4.11	0.800	0.0000
	DDQN	30.37	51.92	−50.88	11.81	0.655	−0.0102
	PPO	27.37	49.57	−59.59	12.92	0.400	0.0000

注：A-S 表示 Avellaneda-Stoikov 解析基线。NetPnL、Std、q05、InvStd、AvgHalfSpread 和 ESkewGivenQ-Slope 分别表示平均净收益、净收益标准差、5%分位数、库存标准差、平均半价差以及 $E[\text{skew}\,|\,Q]$ 的线性拟合斜率。

在有返利的常值激励环境下，结果如表 2 所示。与 simple 场景不同，此时除了 RL 方法外，还引入了固定半价差策略 ConstantPolicy 与线性偏斜策略 LinearSkewPolicy 作为启发式基线。在 HighLiq_constant 场景中，PPO 的平均净收益最高，为 90.4265；两类启发式策略分别达到 84.8672 和 82.0828，同时库存波动明显低于 PPO；DDQN 的净收益为 81.9014，A2C 则在该场景下表现相对较弱。在 LowLiq_constant 场景中，PPO 仍以 89.9195 位居首位，A2C 与两类启发式策略的收益水平较为接近，而 DDQN 相对较低。由此可见，在常值返利机制下，启发式策略已经具备较强竞争力，尤其是在控制库存波动方面表现得更为稳健；PPO 在收益层面更具优势，但这一优势并非在所有风险维度上都同时成立。更合适的理解是，返利增强了积极做市的边际收益，从而重新塑造了报价宽度、库存风险与成交机会之间的平衡，不同算法在这一平衡上的取舍存在差异。

在更复杂的动态激励环境中，强化学习方法与启发式策略的差异更加具有解释意义。表 3 给出了 LowLiq_time_decay 与 LowLiq_state_dependent 两个场景下的结果。在 LowLiq_time_decay 场景中，DDQN 的平均净收益为 73.0494，略高于 LinearSkewPolicy 的 71.5734 和 ConstantPolicy 的 71.3276；A2C 的收益下降至 64.7886，而 PPO 仅为 32.4160，同时具有最大的收益波动和最差的左尾分位数。这表明，在时间递减返利环境下，值函数方法和简单启发式策略都能够较好地利用“早期激励较高”的结构信息，而 PPO

在该场景中的策略波动相对更大。

Table 2. Comparison of RL and heuristic strategies under a constant rebate mechanism
表 2. 常值返利机制下 RL 与启发式策略的比较

市场场景	策略	NetPnL	Std	q05	InvStd	AvgHalfSpread	ESkewGivenQ-Slope
HighLiq	PPO	90.43	19.81	66.25	5.63	0.300	0.0000
	LinearSkewPolicy	84.87	9.40	72.34	2.09	0.800	−0.0500
	ConstantPolicy	82.08	16.73	56.10	3.58	0.800	0.0000
	DDQN	81.90	22.23	43.69	4.32	0.818	−0.0212
	A2C	41.62	75.71	−86.31	18.88	0.350	0.0000
LowLiq	PPO	89.92	33.26	25.23	8.11	0.450	0.0000
	A2C	86.95	58.05	7.39	14.92	0.500	0.0000
	LinearSkewPolicy	86.08	8.14	71.49	2.00	0.800	−0.0500
	ConstantPolicy	85.75	13.40	66.59	3.18	0.800	0.0000
	DDQN	74.15	15.70	56.47	2.60	0.864	−0.0767

注：ConstantPolicy 表示固定半价差启发式策略，取 $h=0.8$， $\text{skew}=0$；LinearSkewPolicy 表示线性偏斜启发式策略，取 $h_0=0.8$， $\phi=0.05$。

Table 3. Comparison of RL and heuristic strategies under time-varying and state-dependent rebate mechanisms
表 3. 时变与状态依赖返利机制下 RL 与启发式策略的比较

市场场景	返利机制	策略	NetPnL	Std	q05	InvStd	AvgHalfSpread	ESkewGivenQ-Slope
LowLiq	Time-decay	DDQN	73.05	10.89	58.44	1.87	0.650	−0.0972
		LinSkew	71.57	7.12	59.30	2.00	0.800	−0.0500
		Const	71.33	12.31	54.24	3.18	0.800	0.0000
		A2C	64.79	22.45	36.80	6.39	0.150	0.0000
		PPO	32.42	61.27	−64.45	15.84	0.900	0.0000
LowLiq	State-dependent	A2C	110.27	25.37	64.09	5.73	0.250	0.0000
		LinSkew	90.44	8.42	75.52	2.00	0.800	−0.0500
		Const	90.08	13.62	70.81	3.18	0.800	0.0000
		PPO	85.42	31.10	38.03	7.92	0.850	0.0000
		DDQN	83.69	36.23	26.21	6.52	0.767	−0.0361

注：LinSkew 表示线性偏斜启发式策略 LinearSkewPolicy，其参数设为 $h_0=0.8$， $\phi=0.05$；Const 表示固定半价差启发式策略 ConstantPolicy，其参数设为 $h=0.8$， $\text{skew}=0$。Time-decay 表示时间依赖返利机制，State-dependent 表示状态依赖返利机制。

相比之下，在 LowLiq_state_dependent 场景中，A2C 的平均净收益达到 110.2703，明显高于两类启发式策略约 90 的水平，也高于 PPO 和 DDQN。与此同时，LinearSkewPolicy 的库存波动标准差仅为 2.0022，依然表现出较强的库存控制能力。由此可见，在状态依赖返利机制下，策略梯度方法更能够根据

订单不平衡和波动水平的变化，自适应地调整报价与偏斜，从而取得更高的收益；而启发式策略虽然在风险控制层面仍具有可解释性优势，但其固定规则形式限制了收益上限。综合两类动态激励环境的结果，更审慎的结论是：RL 方法并非在所有场景中都一致优于启发式策略，但在激励结构更复杂、状态依赖性更强的环境下，A2C 等策略梯度方法的优势更为明显。

5.2. 敏感性分析

为进一步检验模型与策略的稳健性，本文在 LowLiq 的 constant 机制下，对波动率参数 σ、流动性参数 λ 和成交敏感度参数 k 进行了三点敏感性分析。敏感性实验统一复用基准实验的评估种子，以减少由随机样本差异带来的额外噪声。

图 1 表明，在 LowLiq 常值返利环境中，策略表现对环境参数变化具有明确的结构性响应。整体来看，

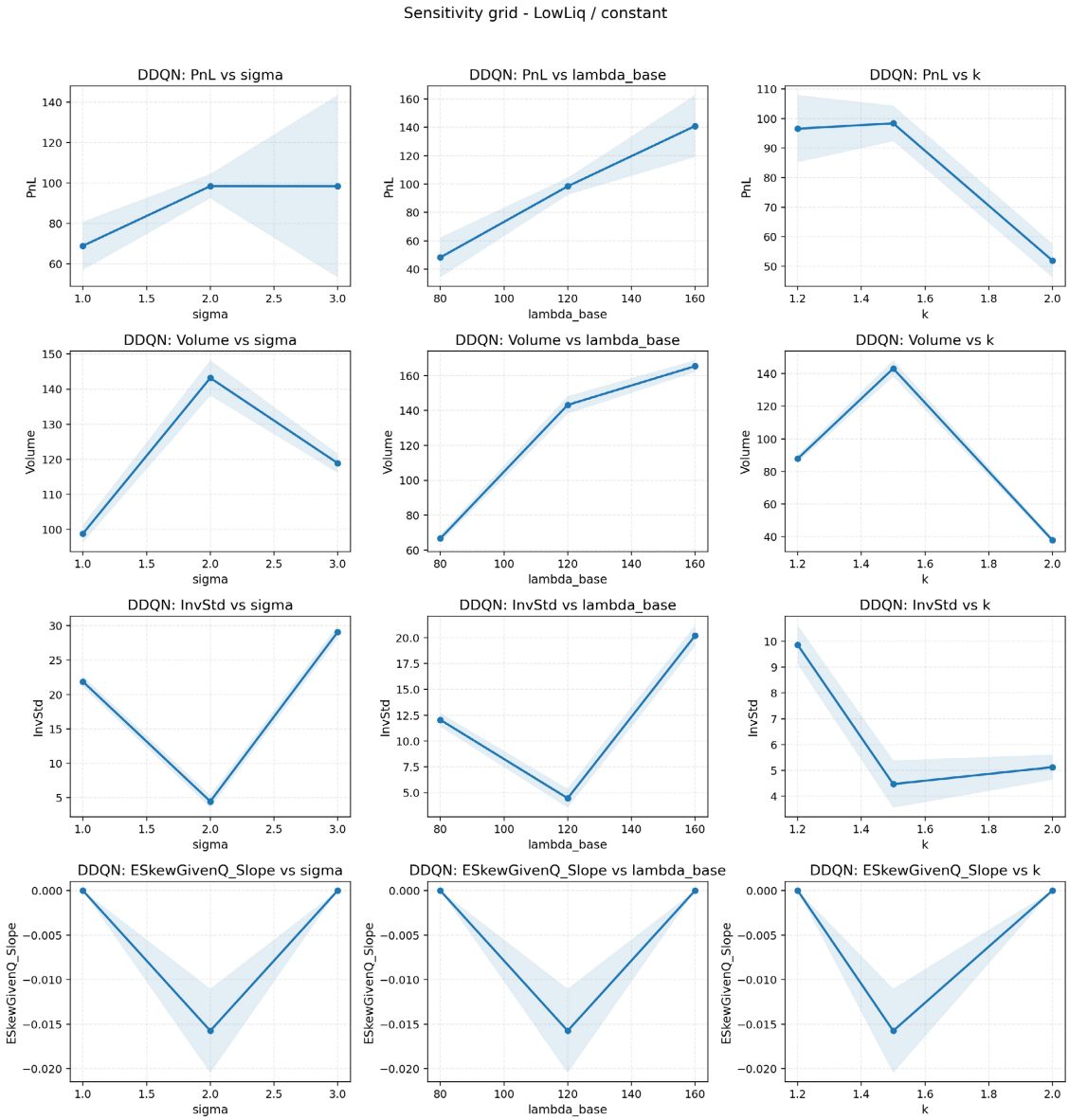

Figure 1. Parameter sensitivity analysis of DDQN in LowLiq constant rebate environment
图 1. LowLiq 常值返利环境下 DDQN 的参数敏感性分析

随着流动性参数λ增大，PnL 与成交量通常呈上升趋势，这与更高订单到达强度带来更多成交机会的机制是一致的。当成交敏感度参数k提高时，报价远离中间价所带来的成交衰减更为明显，策略收益与做量能力受到抑制。而波动率σ的上升则主要体现在风险侧指标的放大，例如库存波动、库存惩罚和回撤水平上升。该结果说明，本文所构建的返利做市环境与强化学习策略在比较静态意义上具有较好的经济一致性。

6. 结论

本文研究了统一返利机制下的做市商问题，并将其从连续时间随机控制模型离散化为有限时域马尔可夫决策过程，再采用 DDQN、A2C 与 PPO 三种强化学习算法进行求解。与仅研究常值返利的做法不同，本文在同一框架中同时考虑了常值返利、时间依赖返利和状态依赖返利三类机制，使返利能够作为可变化的外生激励变量进入做市商优化问题。

数值结果表明，返利机制会显著影响做市商的收益—风险权衡及报价行为。在无返利的简单常值环境下，Avellaneda-Stoikov 解析基线仍然是一个较强的参照，而 PPO 与 A2C 已能够达到与其接近的收益水平，说明强化学习框架在经典 simple setting 下具有合理性。进一步地，在有返利的常值环境中，PPO 在收益层面整体较强，但两类启发式策略同样展现出较高竞争力，尤其在库存风险控制上更为稳健；这说明在结构相对简单的固定激励环境中，强化学习方法与启发式策略之间并不存在单边的绝对优势，而是体现为不同收益–风险权衡。对于时间依赖返利环境，DDQN 与启发式策略表现较为接近；而在状态依赖返利环境中，A2C 的净收益明显高于启发式基线，显示出策略梯度方法在更复杂、状态依赖的激励结构下具有更强的自适应能力。

因此，本文的结果更适合被理解为：强化学习方法在复杂返利机制做市问题中具有可行性，并且其优势主要体现在激励结构更复杂、状态依赖性更强的环境下；与此同时，在结构较简单的环境中，解析方法和启发式策略仍然是必要且有价值的比较基线。本文的主要贡献在于：一是构建了一个统一的返利机制做市模型，将常值、时变和状态依赖返利纳入同一分析框架；二是从数值实验上表明，强化学习方法能够在缺乏一般解析解的情形下对复杂返利机制下的做市问题给出有效近似策略；三是从收益、风险与行为三个层面刻画了不同返利结构如何重塑报价、成交、库存与收益之间的关系。

未来研究可进一步引入真实订单簿数据、多资产场景以及连续动作算法，以提升模型的现实适用性和外部有效性；同时，也可进一步讨论不同启发式策略在动态激励环境中的适用边界，从而更系统地理解强化学习方法相对于经典规则策略的优势来源。

参考文献

[1] Demsetz, H. (1968) The Cost of Transacting. *The Quarterly Journal of Economics*, **82**, 33-53. https://doi.org/10.2307/1882244

[2] Garman, M.B. (1976) Market Microstructure. *Journal of Financial Economics*, **3**, 257-275. https://doi.org/10.1016/0304-405x(76)90006-4

[3] Avellaneda, M. and Stoikov, S. (2008) High-Frequency Trading in a Limit Order Book. *Quantitative Finance*, **8**, 217-224. https://doi.org/10.1080/14697680701381228

[4] Guéant, O., Lehalle, C. and Fernandez-Tapia, J. (2013) Dealing with the Inventory Risk: A Solution to the Market Making Problem. *Mathematics and Financial Economics*, **7**, 477-507. https://doi.org/10.1007/s11579-012-0087-0

[5] Aït-Sahalia, Y. and Brunetti, C. (2020) High Frequency Traders and the Price Process. *Journal of Econometrics*, **217**, 20-45. https://doi.org/10.1016/j.jeconom.2019.11.005

[6] Gasperov, B. and Kostanjcar, Z. (2022) Deep Reinforcement Learning for Market Making under a Hawkes Process-Based Limit Order Book Model. *IEEE Control Systems Letters*, **6**, 2485-2490. https://doi.org/10.1109/lcsys.2022.3166446

[7] Gong, S.Q., Liu, S.Q. and Sun, D.D. (2023) Optimal Market Making in the Chinese Stock Market: A Stochastic Control and Scenario Analysis. https://doi.org/10.48550/arXiv.2306.02764

[8] Spooner, T. and Savani, R. (2020) Robust Market Making via Adversarial Reinforcement Learning. https://arxiv.org/abs/2003.01820