清華大學

Market Making via Reinforcement Learning

Spooner T, Fearnley J, Savani R, et al.

AAMAS 2018（智能体领域顶会）

2021/10/14

目录

CONTENTS

背景介绍
Background

做市
Market making

报价驱动市场→做市商报价，与投资者交易

订单驱动市场→连续的双边拍卖

限价单

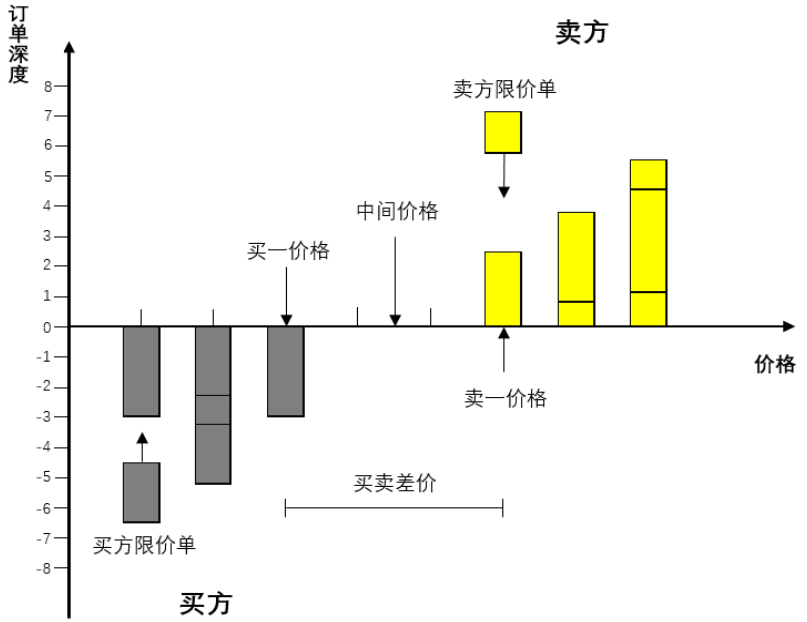
给定价格与数量的报单，此价格为限定成交价格，对于买单来说，是**最高买入价格**，对于卖单来说，是**最低出卖价格**，限价单被提交到交易所电子撮合系统之后，按照**价格优先**和**时间优先**原则在各个队列中排队等待成交

市价单

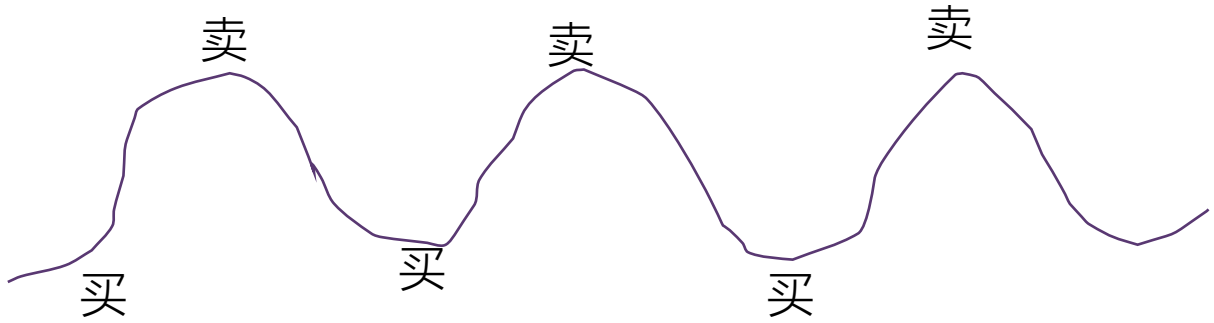
以市场价格进行买入或卖出的委托单，市价单进入交易所电子撮合系统后会立即与限价单进行撮合成交或者进入买一/卖一队列

限价订单簿

	Price	Asks
	101.00	12
	100.50	13
	100.25	
35	100.00	
3	99.75	
11	99.50	
Bids		



高频交易
High-Frequency Trading



做市策略：低库存，低风险

均值回复原理，利用**市场波动**获利，低价挂买单，高价挂卖单，或者缩小买卖价差提供流动性

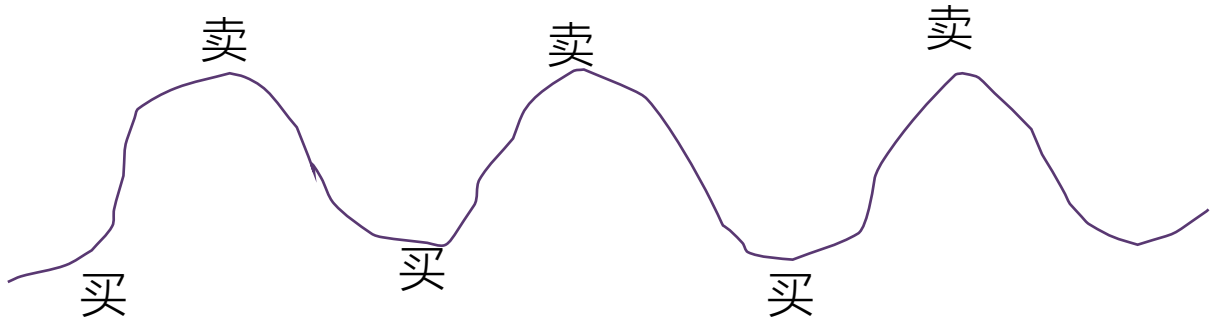
方向性策略：信息差、时间差，快速反应

对订单流信息或具体事件加以研判后得出短期大概率价格波动方向，以**速度优势**提前建仓，等到价格波动到预想点位后平仓

套利策略

跨市场或跨资产类型套利，快速自动化发掘价差

高频交易
High-Frequency Trading



做市策略：低库存，低风险

均值回复原理，利用**市场波动**获利，低价挂买单，高价挂卖单，或者缩小买卖价差提供流动性

订单处理成本（45%）

指的是交易过程中的印花税、过户费等成本，一般会由交易所返还一定比例

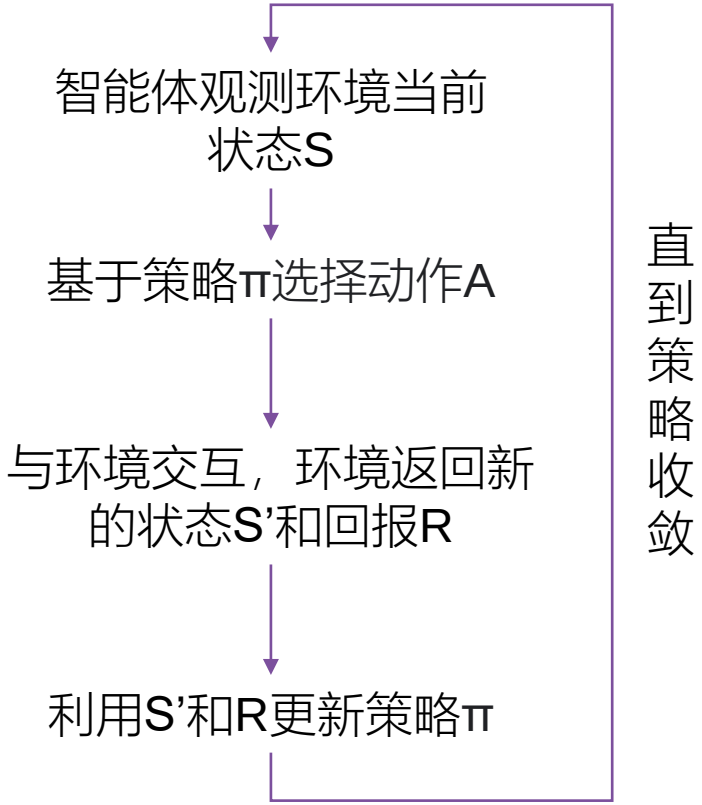
库存成本（10%）

做市商期望通过买卖价差来获取利润。做市商如果不同时买卖，就会产生库存成本。做市的核心问题就是避免积累大量的多空头寸

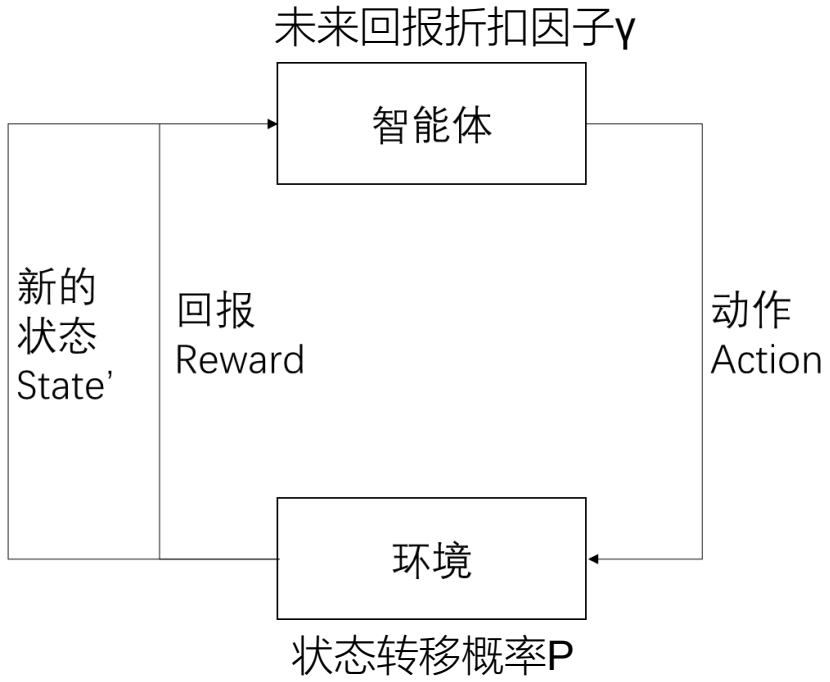
信息不对称成本（45%）

信息成本是未知情交易者对知情交易者付出的成本。如果在市场上存在信息知情者，那么在这种信息不对称的场景下，做市商如果选择和信息知情者交易，将会承担一定的损失

强化学习
Reinforcement Learning



强化学习数学模型：马尔科夫决策过程



为什么用强化学习
Reinforcement Learning

无监督

强化学习方法通过定义奖惩函数来训练智能体，因为高频交易中不可能存在由专家标记的最佳买卖点，所以很难应用监督学习方法

端到端

强化学习方法可以从数据中直接学习到决策并执行，而不是像传统算法预测股价或趋势，后续还需要人工干预来进行投资

自适应

金融市场具有暂时性，市场并不是一成不变的，而传统的方法在遇到市场变化时，需要重新训练并测试模型，而强化学习方法可以自适应连续地优化模型从而使智能体适应市场环境的变化

方法介绍
Proposed method

模拟环境&动作空间
Simulator & Action space

数据

重建的LOB最优五档数据 + 成交/撤单数据

模拟环境

智能体的挂单进入队列排队，环境反馈成交信号。
智能体下单数量少，忽略对市场的影响

动作空间

Action ID	0	1	2	3	4	5	6	7	8
Ask (θ_a)	1	2	3	4	5	1	3	2	5
Bid (θ_b)	1	2	3	4	5	3	1	5	2
Action 9	MO with $\text{Size}_m = -\text{Inv}(t_i)$								

$$p_{a,b}(t_i) = \text{Ref}(t_i) + \text{Dist}_{a,b}(t_i), \quad (1)$$

$$\text{Dist}_{a,b}(t_i) = \theta_{a,b}(t_i) \cdot \text{Spread}(t_i), \quad (2)$$

$\text{Spread}(t_i)$ is calculated by taking a moving average of the market half-spread: $s(t_i)/2$.

奖励函数
Reward

Profit and Loss (PnL)

$$\psi_{\mathrm{a}}(t_i) \triangleq \mathrm{Matched}_{\mathrm{a}}(t_i) \cdot [p_{\mathrm{a}}(t_i) - m(t_i)],$$
$$\psi_{\mathrm{b}}(t_i) \triangleq \mathrm{Matched}_{\mathrm{b}}(t_i) \cdot [m(t_i) - p_{\mathrm{b}}(t_i)],$$

$$\Psi(t_i) \triangleq \psi_{\mathrm{a}}(t_i) + \psi_{\mathrm{b}}(t_i) + \mathrm{Inv}(t_i)\Delta m(t_i)$$

持有库存带来的收益

做市策略追求低库存，不能让智能体有投机心理

库存惩罚项

对称惩罚
Symmetrically dampened PnL

$$r_i = \Psi(t_i) - \eta \cdot \mathrm{Inv}(t_i)\Delta m(t_i)$$

问题：不仅惩罚盈利，也惩罚了损失

不对称惩罚
Asymmetrically dampened PnL

$$r_i = \Psi(t_i) - \max[0,\ \eta \cdot \mathrm{Inv}(t_i)\Delta m(t_i)]$$

仅仅惩罚盈利

市场状态表征
States

Agent-state

(1) 库存 inv(t_i)

(2) Ask价格θa

(3) Bid价格θb

Normalized by spead(t_i)

Market-state

(1) 买卖价差 spread(s)

(2) 中间价格变动Δm

(3) 订单不平衡

(4) 成交量

(5) 波动性

(6) 相对强弱指数RSI

实验
Experiments

实验
Experiments

训练算法

Double Q-learning
Expected SARSA
On-policy R-learning
Double R-learning

评价指标

盈利：*normalised daily* PnL
每天的利润除以平均价差

库存：平均绝对位置MAP

稳定性：PnL和MAP的标准差和平均绝对偏差

Table 2: Default parameters as used by the learning algorithm and the underlying trading strategy.

	Value
Training episodes	1000 days
Training sample size	~ 120 days
Testing sample size	40 days
Memory size	10^7
Number of tilings (M)	32
Weights for linear combination of tile codings [agent, market, full] (λ_i)	(0.6, 0.1, 0.3)
Learning rate (α)	0.001
Step-size [R-learning] (β)	0.005
Discount factor (γ)	0.97
Trace parameter (λ)	0.96
Exploration rate (ε)	0.7
$\varepsilon_{\text{Floor}}$	0.0001
ε_{T}	1000
Order size (ω)	1000
Min inventory (min Inv)	-10000
Max inventory (max Inv)	10000

实验
Experiments

固定和随机策略：固定报价with库存强制清零

	ND-PnL [10^4]	MAP [units]
Fixed ($\theta_{a,b} = 1$)	−20.95 ± 17.52	3646 ± 2195
Fixed ($\theta_{a,b} = 2$)	2.97 ± 13.12	3373 ± 2181
Fixed ($\theta_{a,b} = 3$)	0.42 ± 9.62	2674 ± 1862
Fixed ($\theta_{a,b} = 4$)	1.85 ± 10.80	2580 ± 1820
Fixed ($\theta_{a,b} = 5$)	**2.80 ± 10.30**	2678 ± 1981
Random (Table 1)	−10.82 ± 5.63	135 ± 234

稳定性都很差，可能是因为库存管理

Basic-agent：不加库存惩罚项，只用agent-state训练，训练算法为Q-learning和SARSA

	QL	SARSA
CRDI.MI	**8.14 ± 21.75**	4.25 ± 42.76
GASI.MI	−4.06 ± 48.36	**9.05 ± 37.81**
GSK.L	4.00 ± 89.44	**13.45 ± 29.91**
HSBA.L	**−12.65 ± 124.26**	−12.45 ± 155.31
ING.AS	−67.40 ± 261.91	**−11.01 ± 343.28**
LGEN.L	**5.13 ± 36.38**	2.53 ± 37.24
LSE.L	4.40 ± 16.39	**5.94 ± 18.55**
NOK1V.HE	**−7.65 ± 34.70**	−10.08 ± 52.10
SAN.MC	−4.98 ± 144.47	**39.59 ± 255.68**
VOD.L	**15.70 ± 43.55**	6.65 ± 37.26

ND-PnL

实验
Experiments

库存可能是不稳定性的主要来源

库存惩罚系数

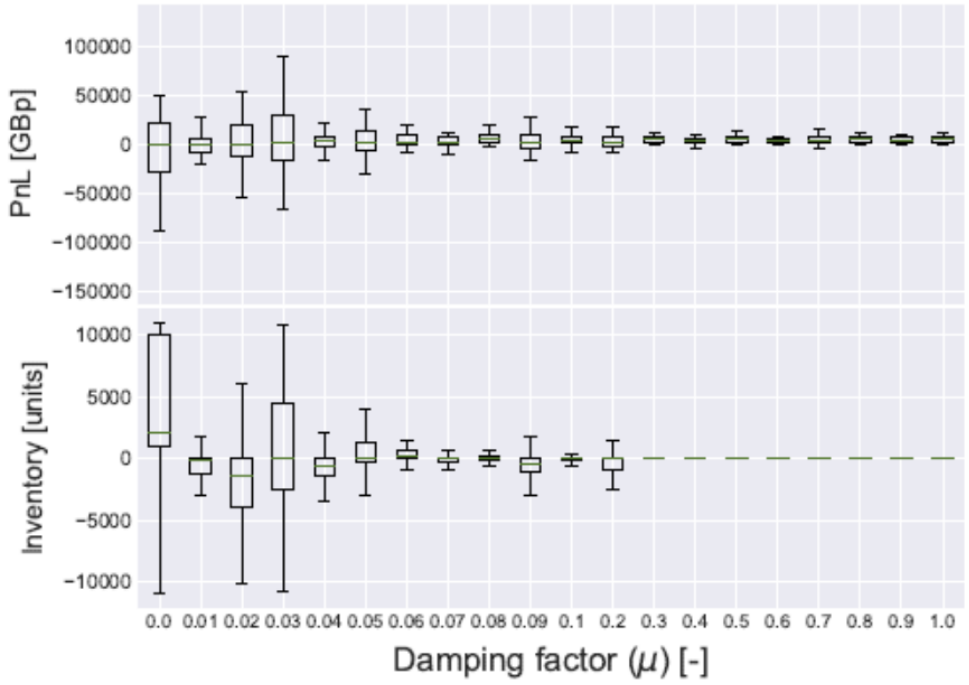


Figure 2: Distributions of daily out-of-sample PnL and mean inventory for increasing values of the damping factor, η, evaluated on HSBA.L.

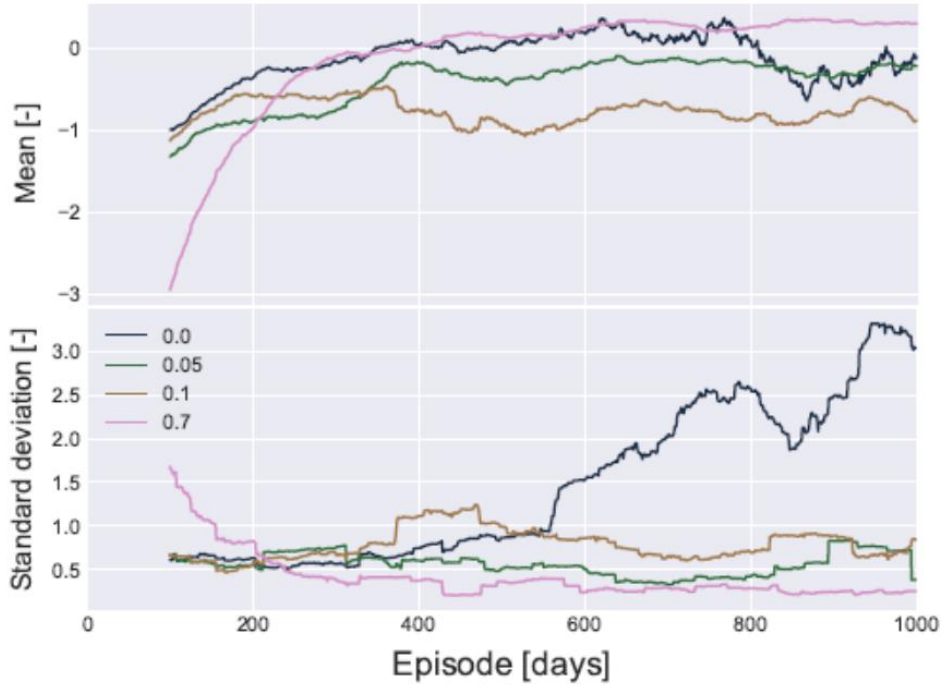


Figure 3: Rolling mean and standard deviation on the average episodic reward during training for increasing values of the damping factor, η, evaluated on HSBA.L.

实验
Experiments

消融实验：从训练方法/回报函数/状态表示三个层面对basic-agent进行修改

Table 5: Mean and standard deviation on the normalised daily PnL (given in units of 10^4) for various extensions to the basic agent, each evaluated over the basket of stocks.

	CRDI.MI	GASI.MI	GSK.L	HSBA.L	ING.AS	LGEN.L	LSE.L	NOK1V.HE	SAN.MC	VOD.L
Double Q-learning	−5.04 ± 83.90	5.46 ± 59.03	6.22 ± 59.17	5.59 ± 159.38	58.75 ± 394.15	2.26 ± 66.53	16.49 ± 43.10	−2.68 ± 19.35	5.65 ± 259.06	7.50 ± 42.50
Expected SARSA	0.09 ± 0.58	3.79 ± 35.64	−9.96 ± 102.85	25.20 ± 209.33	6.07 ± 432.89	2.92 ± 37.01	6.79 ± 27.46	−3.26 ± 25.60	32.28 ± 272.88	15.18 ± 84.86
R-learning	5.48 ± 25.73	−3.57 ± 54.79	12.45 ± 33.95	−22.97 ± 211.88	−244.20 ± 306.05	−3.59 ± 137.44	8.31 ± 23.50	−0.51 ± 3.22	8.31 ± 273.47	32.94 ± 109.84
Double R-learning	19.79 ± 85.46	−1.17 ± 29.49	21.07 ± 112.17	−14.80 ± 108.74	5.33 ± 209.34	−1.40 ± 55.59	6.06 ± 25.19	2.70 ± 15.40	32.21 ± 238.29	25.28 ± 92.46
On-policy R-learning	0.00 ± 0.00	4.59 ± 17.27	14.18 ± 32.30	9.56 ± 30.40	18.91 ± 84.43	−1.14 ± 40.68	5.46 ± 12.54	0.18 ± 5.52	25.14 ± 143.25	16.30 ± 32.69
Symm. Damp. ($\eta = 0.6$)	12.41 ± 143.46	9.07 ± 68.39	30.04 ± 135.89	−11.80 ± 214.15	90.05 ± 446.09	5.54 ± 119.86	8.62 ± 27.23	−4.40 ± 84.93	27.38 ± 155.93	8.87 ± 93.14
Asymm. Damp. ($\eta = 0.6$)	0.08 ± 2.21	−0.10 ± 1.04	9.59 ± 10.72	13.88 ± 10.60	−6.74 ± 68.80	4.08 ± 7.73	1.23 ± 1.80	0.52 ± 3.29	5.79 ± 13.24	9.63 ± 6.94
Full-state	−31.29 ± 27.97	−35.83 ± 13.96	−31.29 ± 27.97	−84.78 ± 31.71	−189.81 ± 68.31	−14.39 ± 9.38	−6.76 ± 11.52	−9.30 ± 23.17	−144.70 ± 104.64	−21.76 ± 17.71
LCTC-state	−5.32 ± 52.34	5.92 ± 40.65	5.45 ± 40.79	−0.79 ± 68.59	9.00 ± 159.91	6.73 ± 22.88	3.04 ± 5.83	−2.72 ± 19.23	52.55 ± 81.70	7.02 ± 48.80

实验
Experiments

消融实验

最终的模型：asymmetrically dampened reward function with a LCTC state-space, trained using SARSA.

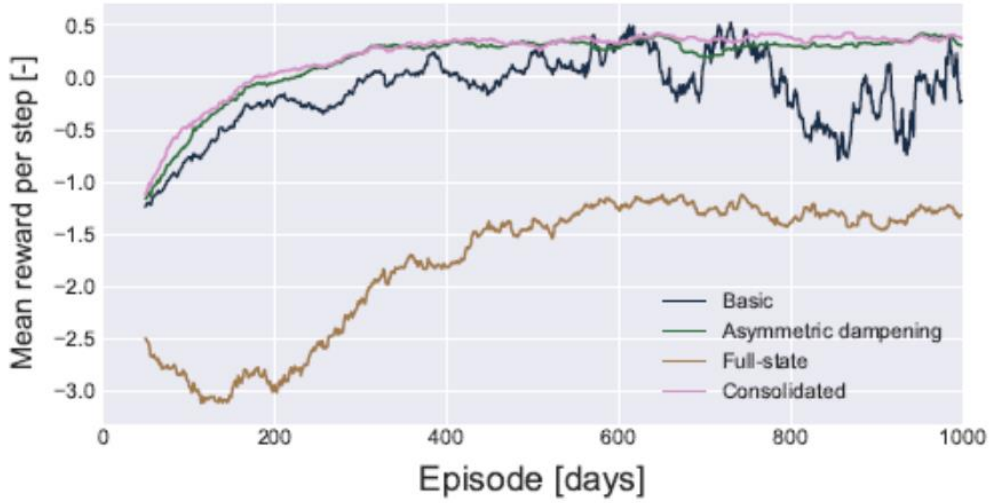


Figure 4: Rolling mean of the average episodic reward for different agent variants during training on HSBA.L.

实验
Experiments

与其他方法对比

Table 6: Comparison of the out-of-sample normalised daily PnL (ND-PnL) and mean absolute positions (MAP) of the benchmark strategies against the final presented reinforcement learning agent.

	Abernethy and Kale (MMMW) Benchmark		Fixed ($\theta_{a,b} = 5$) Benchmark		Consolidated Agent	
	ND-PnL [10^4]	MAP [units]	ND-PnL [10^4]	MAP [units]	ND-PnL [10^4]	MAP [units]
CRDI.MI	−1.44 ± 22.78	7814 ± 1012	−0.14 ± 1.63	205 ± 351	**0.15 ± 0.59**	1 ± 2
GASI.MI	−1.86 ± 9.22	5743 ± 1333	**0.01 ± 1.36**	352 ± 523	0.00 ± 1.01	33 ± 65
GSK.L	−3.36 ± 13.75	8181 ± 1041	0.95 ± 2.86	1342 ± 1210	**7.32 ± 7.23**	57 ± 105
HSBA.L	1.66 ± 22.48	7330 ± 1059	2.80 ± 10.30	2678 ± 1981	**15.43 ± 13.01**	104 ± 179
ING.AS	−6.53 ± 41.85	7997 ± 1265	**3.44 ± 23.24**	2508 ± 1915	−3.21 ± 29.05	10 ± 20
LGEN.L	−0.03 ± 11.42	5386 ± 1297	0.84 ± 2.45	986 ± 949	**4.52 ± 8.29**	229 ± 361
LSE.L	−2.54 ± 4.50	4684 ± 1507	0.20 ± 0.63	382 ± 553	**1.83 ± 3.32**	72 ± 139
NOK1V.HE	−0.97 ± 8.20	5991 ± 1304	**−0.52 ± 4.16**	274 ± 497	−5.28 ± 33.42	31 ± 62
SAN.MC	−2.53 ± 26.51	8865 ± 671	1.52 ± 11.64	3021 ± 2194	**5.67 ± 13.41**	4 ± 9
VOD.L	1.80 ± 22.83	7283 ± 1579	1.26 ± 4.60	1906 ± 1553	**5.02 ± 6.35**	46 ± 87

实验
Experiments

表现最稳定的一只股票

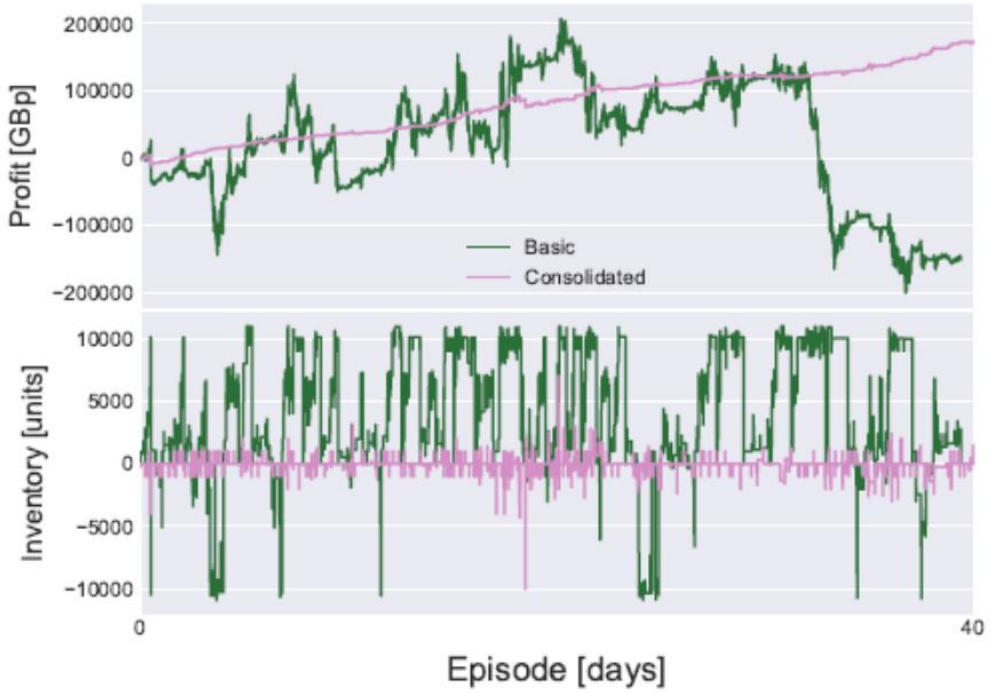


Figure 5: Out-of-sample equity curve and inventory process for the basic and consolidated agents, evaluated on HSBA.L.

最终的模型与basic-agent相比，证明了方法的有效性

优点&不足
Conclusion

优点

1. 提出的库存惩罚系数，显著控制了风险
2. 实验完善，分析了各种方法的实际效果
3. C++开源代码

不足

1. Agent的稳定性太差了，泛化性和鲁棒性有待提高
2. 使用的是基于值的强化学习，有研究指出基于策略的强化学习方法鲁棒性会更强
3. 没有使用深度学习方法抽取市场特征，没有依赖时序信息（RNN/注意力/WaveNet）
4. 订单簿信息不全，没有逐个订单量，无法判断交易情绪和不对称信息等
5. 动作空间离散化之后对智能体的限制较大

THANKS

2021/10/14