深度强化学习在高频主动交易中的应用

Antonio Briola*, Jeremy Turiel*, Riccardo Marcaccioli†, Alvaro Cauderan§, Tomaso Aste*‡
伦敦大学学院计算机科学系

英国伦敦
{a.briola, jeremy.turiel.18, t.aste}@ucl.ac.uk（此邮件地址无需翻译，直接保留原样即可。）
†巴黎综合理工学院经济物理学与复杂系统教席,法国

{riccardo.marcaccioli} @ladhyx.polytechnique.fr
‡系统性*风险中心,英国伦敦经济学院*,伦敦,英国

{acauderan} @ethz.ch
§计算机科学系,*苏黎世联邦理工学院*

瑞士苏黎世

*摘要 - 我们*引入了首个基于深度强化学习（DRL）的端到端框架，用于股票市场的高频主动交易。我们采用近端策略优化算法训练 DRL 代理，使其交易英特尔公司的一股股票。训练是在连续三个月的高频限价订单簿数据上进行的，其中最后一个月的数据作为验证数据。为了在训练数据中最大化信噪比，我们仅选择价格变化最大的训练样本组成训练数据。然后在接下来一个月的数据上进行测试。使用基于序列模型的优化技术对超参数进行调整。我们考虑了三种不同的状态表征，它们在基于限价订单簿的元特征方面有所不同。通过分析代理在测试数据上的表现，我们认为代理能够动态地表征底层环境。它们能够识别数据中偶尔出现的规律，并利用这些规律制定长期盈利的交易策略。实际上，代理能够学习到在高度随机且非平稳的环境中仍能产生稳定正收益的交易策略。源代码可在 https://github.com/FinancialComputingUCL/DRL 获取，用于主动高频交易。

*索引词 - 人工*智能、深度强化学习、高频交易、市场微观结构

I. 简介

机器能够学会自动交易吗？它们要完成这一任务需要哪些数据？如果它们成功做到了，我们能否利用它们来了解价格形成机制？在本文中，我们尝试通过深度强化学习（DRL）方法来回答这些问题。

在过去的十年中，深度强化学习（DRL）算法已被应用于从机器人技术[1]、[2]到医疗保健[3]等众多领域和情境。关于主要应用领域的完整概述，我们建议感兴趣的读者参考李[4]的工作。金融市场是众所周知的随机环境，其信号与噪声比极低，受显著的非平稳动态支配，并以强烈的反馈回路和非线性效应为特征[5]。这在最高粒度的数据中尤其明显[5]、[6]，即所谓的微观结构层面。限价订单簿（LOB）是市场参与者通过在特定价格水平提交订单来表达其购买或出售特定数量（成交量）证券意图的场所。将深度强化学习算法应用于这种环境的兴趣无疑是强烈的。首先，所有利用人工神经网络（ANN）能力的算法都是众所周知的数据饥渴型模型，而且由于电子交易的普及，资产限价订单簿中发生的一切都可以以高达纳秒级的分辨率进行记录和存储。其次，可以基于此类数据定义多个不同的问题，这些问题有可能通过强化学习来解决。

有几个具有直接实际意义的研究问题亟待解决。智能体能否学会如何在一组资产中分配给定数量的资金？如果我们需要大量买入或卖出资产，能否训练智能体以最优方式完成操作？能否训练智能体直接学习不同频率和不同数量资产的交易策略?

在本研究中，我们选择解决后一个问题。我们展示了如何利用限价订单簿（LOB）数据训练深度强化学习（DRL）代理来买卖给定资产的单个单位，并实现长期盈利。我们选择交易单个单位，是因为我们的目标并非最大化代理的虚拟利润，而是展示首个端到端框架，成功部署深度强化学习算法用于主动高频交易。这在第四节 B 部分有进一步讨论。具体而言，我们应用近端策略优化（PPO）算法[7]来训练我们的代理。我们采用持续学习范式，并通过顺序调优模型的超参数。

TA and JT acknowledge the EC Horizon 2020 FIN-Tech project for partial support and useful opportunities for discussion. JT acknowledges support from EPSRC (EP/L015129/1). TA acknowledges support from ESRC (ES/K002309/1), EPSRC (EP/P031730/1) and EC (H2020-ICT-2018-2 825215).

基于模型的优化（SMBO）[8] 技术。受 Briola 等人[9]工作的启发，我们使用 Stable Baselines [10] 提供的最简单的策略对象，该对象通过使用多层感知机实现了演员 - 批评者算法。最后，为了对损益（P&L）函数的动态进行正则化，我们通过开发一种抽样策略来构建训练集，该策略选择价格活动频繁的时期，这些时期通常具有更高的信噪比。这为智能体提供了一个更具信息量的训练集。

II. 背景

A. 限价订单簿

如今，为了促进交易，大多数大大小小的金融市场交易所都采用了一种名为限价订单簿（LOB）的电子交易机制。在限价订单簿中，资产价格的形成是一个由订单提交和取消驱动的自组织过程。订单可以被视为市场参与者以特定价格买卖特定数量资产的公开声明。订单会被放入队列，直到被其所有者取消或与相反方向的订单匹配执行。一对订单的执行意味着订单所有者以约定的价格交易约定数量的资产。订单执行遵循先进先出(*FIFO*)机制。一个订单 x 由四个属性定义：方向 $\boxtimes_x$（表示所有者是想买还是卖给定资产）、提交价格 p_x、所有者想交易的股数（或交易量）V_x 以及提交时间 τ_x。因此，它由一个元组 $x = (\boxtimes x,\ p_x,\ V_x,\ \tau_x)$ 定义。每当交易员提交买入（或卖出）订单时，自动交易匹配算法会检查是否有可能以小于或等于（或大于或等于）价格 p_x 的价格与一个或多个现有反向订单执行 x 交易。通常，卖出订单的符号为负 $\boxtimes$= -1，而买入订单的符号为正 $\boxtimes$= 1。如果整个交易量 V_x 可以部分匹配，则立即执行匹配。相反，未能立即匹配的任何部分 V_x 以价格 p_x 成为活跃订单，并一直保持活跃状态，直到与新来的卖出（或买入）订单匹配或被所有者取消。

图 1 展示了限价订单簿（LOB）动态变化的示意图。所有活跃的卖出限价订单的集合构成了限价订单簿的卖方报价部分，而所有活跃的买入限价订单的集合则构成了买方报价部分。价格差 $\sigma\tau$ = pbest

$_{a,\ \tau}$ 减去
pb, τ 最好的

在卖单和买单之间

分别以它们的最佳可得价格（即价格更低的卖出价/更高的买入价）来定义买卖价差，而某一证券的最佳买入价和最佳卖出价的平均值则称为中间价 $m\tau$ = (pbest

$_{a,\ \tau}$ + 最佳值 pbest

$_{(b+\tau)}$) / 2

以最优买价/卖价成交的订单被称为市价单（MO）。市价单获得时间优先权（因为其订单会立即执行），但代价是每单位给定资产需额外支付至少等于价差 σ_τ 的金额。这就是为什么价差也被称作市价单的成本。未在到达时完全执行的订单

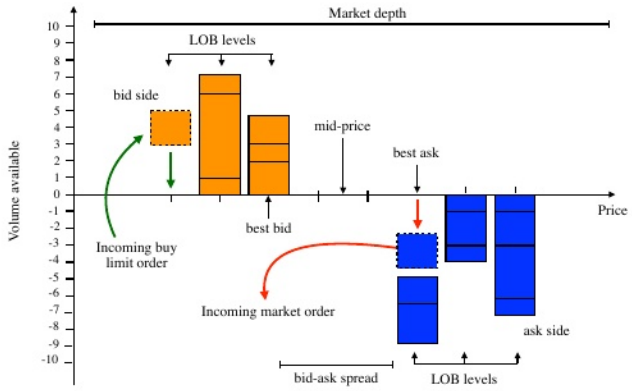

图 1.限价订单簿的图形说明，包括其组成部分、相关数量及动态变化。

（因此，这些订单全部或部分处于限价订单簿中，直到匹配或取消）被称为限价订单。处于最优买价或卖价的限价订单会通过可能因市场而异的机制与新来的市价订单匹配。最常见的优先级机制（也是本研究中所考虑的限价订单簿动态的规则）称为价格时间优先级：处于相同价格的买入/卖出限价订单通过选择提交时间最早的限价订单来执行 τ_x。因此，我们可以将限价订单簿定义为在给定平台上的给定资产在时间 τ 时未匹配的限价订单的集合。时间参数 τ 自然是一个连续变量，可以单独研究，例如通过表征两种或多种类型订单之间的等待时间分布，或者研究一个或多个限价订单簿可观测值的可能季节性。然而，当有人想要表征某个特定事件对限价订单簿的影响（例如市价订单后的平均价格变动）时，将时间标准化并以交易时间来工作会更容易。在这个系统/参考框架中，时间作为一个离散变量，每当限价订单簿（LOB）上记录一个事件时就会更新一次。为了便于阐述，我们在此结束对限价订单簿的描述。然而，已经可以看出，这样的场所同时是强化学习算法极具吸引力却又极具挑战性的试验场。规则简单，却能产生复杂的模式，环境的动态是各种间接交互的结果，其中代理尽可能掩盖其真实意图，且不同代理根据各自的特定策略和目标采取行动。*感兴趣的读者可参考 Bouchaud* 等人所著的书籍[11]，以获取关于该主题的详细介绍以及对限价订单簿上可定义的各种量的动态的深入描述。

B. 近端策略优化算法

强化学习（RL）可以被描述为学习者所面临的问题，即学习者试图通过与动态环境进行反复试验和错误的交互来实现特定目标，同时将相关成本降至最低。在这个学习过程中，智能体经历不同的状态（即环境所涵盖的相空间中的不同点）。

智能体处于特定环境状态中，并且能够执行多种动作。然后通过一个评分函数和一个数值奖励来评估与特定状态-动作对相关的效果，该数值奖励用于确定所选动作序列的好坏。在这种情况下，智能体需要在利用已有的状态-动作对经验和探索未知的状态-动作对之间找到最佳平衡。找到这种平衡绝非易事。解决这种最大化问题所涉及的内在随机性，使强化学习有别于其他学习范式，因为它并非试图逼近任何特定的底层函数或最大化一个明确且确定的分数。

累积的这种对应于在时间 T ∈ [0, ∞) 内找到最优的从状态集 S 到动作集 A 的概率分布的映射方法，能够最大化折扣值，解决强化学习问题可以归为不同的类别。我们将重点关注基于策略搜索的方法，但有兴趣的读者可参考理查德·S·萨顿和安德鲁·G·巴托所著的书籍[12]，以获取对该主题的全面综述。

尽管在基本配置下，强化学习算法对各种各样的问题都十分有效，但事实证明，它们会受到维度灾难的影响[13]。用于描述智能体状态的特征数量越多，智能体学习最优策略就越困难。实际上，它需要在一个其规模随特征数量呈组合式增长的空间中进行搜索。为了将传统的强化学习算法扩展到高维问题，引入了深度强化学习（DRL）。利用人工神经网络（ANN）从高维输入中自动生成低维特征表示，可以克服维度灾难。然而，深度强化学习算法可能需要大量数据才能实现良好的泛化，而且训练任务也众所周知地困难[4]。

除其他问题外，对超参数调整和初始化的高度敏感性会影响深度强化学习（DRL）智能体的泛化能力。例如，错误的学习率可能会使策略网络进入参数空间的一个区域，在该区域中，它将根据非常糟糕的策略收集下一批数据，从而导致其无法恢复。为了解决这个问题，引入了近端策略优化（PPO）算法[7]。PPO 以可扩展、数据高效、稳健[14]而著称，并且是应用于高度随机领域最合适的算法之一。因此，它是这里所考虑的应用领域的有效候选算法。

PPO 算法属于深度强化学习（DRL）中策略梯度方法的一个分支，这些方法主要是基于策略的方法，即它们根据当前状态来寻找下一个动作。为了实现这一点，策略梯度方法采用了策略损失函数 $L^{PG}(Q)$

$$L^{PG}(Q) = E_t[\log \pi_Q(a_t|s_t)]\hat{A}_t \ , \qquad (1)$$

其中 $E_t[\log \pi_Q(a_t|s_t)]$ 表示在状态 s 下采取行动 a 的预期对数概率，而估计的优势函数 A_t 则提供了当前行动相对于当前状态 s 下所有其他可能行动的平均值的价值估计。

PPO 方法通过直接控制策略损失函数所隐含的梯度上升步长的大小来确保策略更新过程的稳定性。为了实现这一点，目标函数按照公式 2 进行编写。

$$L^{CPI}(Q) = E_t\left[\frac{\pi_Q(a_t|s_t)}{\pi_{Q_{old}}(a_t|s_t)}\hat{A}_t\right] = E_t[r_t(Q)\hat{A}_t] \ , \qquad (2)$$

其中，将术语 $\log \pi_Q(a_t|s_t)$ 替换为当前策略下采取该行动的概率与前一策略下采取该行动的概率之比，即 $r_t(Q) = \frac{\pi_Q(a_t|s_t)}{\pi(a_t|s_t)}$。这意味着对于 $r_t(Q) > 1$ 的情况

在当前策略中采取行动的可能性比在前一策略中更大，而如果 $0 < r_t(Q) < 1$，则在前一策略中采取行动的可能性更大。然后我们可以注意到，当 $r_t(Q)$ 偏离 1 时，策略更新会过大。为了避免这种情况，根据方程 3 对目标函数进行修改，以惩罚超出 [1−ε， 1+ε] 范围的 $r_t(Q)$ 值。这里我们取 ε = 0.2，与原论文 [7] 中一致。

$$L^{CLIP}(Q) = E_t\Big[min\Big[\Big(r_t(Q)\hat{A}_t\Big), clip\Big(r_t(Q), [1-\epsilon, 1+\epsilon]\Big)\hat{A}_t\Big]\Big] \ . \qquad (3)$$

方程 3 中的裁剪代理目标函数取原始的 $L^{CPI}(Q)$ 与裁剪范围 clip $r_t(Q)$，[1 - ε， 1 + ε] A_t 之间的最小值，从而消除了对于超出 [1 − ε， 1 + ε] 范围的激励项。最小值构成了未裁剪目标的下限（即悲观下限）。

III. 相关工作

正如第 I 节所述，电子交易的增长引发了人们对基于深度学习的应用在限价订单簿（LOB）数据中的兴趣。利用监督学习模型对限价订单簿数量进行预测的相关文献正在迅速增加。Kearns 和 Nevmyvaka [15] 的工作概述了基于人工智能的应用在市场微观结构数据和任务中的情况，包括基于先前限价订单簿状态的收益预测。最近，Tsantekidis 等人在 [16] 中首次尝试对基于限价订单簿的股票价格预测的深度学习方法进行了广泛分析。从对诸如支持向量机等传统机器学习方法与更结构化的深度学习方法的赛马式比较开始，Tsantekidis 等人接着考虑了将卷积神经网络（CNN）应用于检测金融市场中的异常事件，并采取有利可图的头寸。Sirignano 和 Cont [17] 的工作则提供了更理论化的研究方法。它对基于限价订单簿数据的多种深度学习架构进行收益预测进行了比较，并讨论了这些模型捕捉和推广到各种资产通用价格形成机制的能力。最近，

张等人完成了两项旨在预测中价收益的研究工作。第一项工作[18]直接在微观结构层面利用贝叶斯网络预测收益，而第二项[19]则开发了一种深度学习架构，旨在同时对买卖双方的收益分位数进行建模。后者引入了目前最先进的结合卷积神经网络（CNN）和长短期记忆网络（LSTM）的建模架构，以更深入地探究限价订单簿（LOB）的复杂结构。最后，关于上述所有方法在收益分位数预测中的应用，近期的一篇综述[9]指出，通过时间维度和空间维度对限价订单簿进行建模（如[19]中引入的 CNN-LSTM 模型）能够提供良好的近似，但并非必要且最优的方法。这些发现证明了在本研究中使用多层感知机（MLP）作为深度强化学习（DRL）代理的底层策略网络是合理的。

涉及金融市场数据的强化学习（RL）应用可追溯到 21 世纪初。这些应用通常集中在低频单资产（或指数）交易[20]、[21]、投资组合配置[22]-[24]，直至早期形式的日内交易[25]。关于该主题的全面综述，我们建议感兴趣的读者参阅[26]。深度强化学习（DRL）及其各种应用近期受到的关注激增，这自然促使研究人员将此类算法应用于金融市场数据或模拟数据[27]、[28]。在交易应用方面，研究人员主要关注使用模拟或真实数据在低频（通常是每日）领域进行单资产的主动交易[29]-[36]。其他研究则侧重于多种任务，如多资产配置[37]-[39]或一般意义上的投资组合管理[40]。然而，将 DRL 应用于高频领域的文献却出奇地少。除了近期旨在将 DRL 算法应用于模拟高频市场做市的几项工作[41]、[42]外，尚无关于 DRL 在真实高频限价订单簿（LOB）数据上应用的文献。在当前的研究中，我们试图填补这一空白，并提出首个端到端框架，用于在真实的高频市场数据上训练和部署深度强化学习代理，以完成主动高频交易任务。

IV. 方法

A. 数据

来自 LOBSTER 数据集 [43] 的高质量限价订单簿（LOB）数据是本研究工作的基础。该数据集在市场微观结构文献中被广泛采用 [11]、[44]-[46]，它为纳斯达克交易所上市的每只股票提供了极其详尽的逐笔市场活动记录。该数据集列出了每个交易日 09:30 至 16:00 期间纳斯达克平台上发生的每笔市场订单到达、限价订单到达和取消情况。周末和公共假日不进行交易，因此这些日期未纳入本研究的所有分析。纳斯达克交易所所有股票的最小价格变动单位为 $\theta = 0.01$ 美元。订单流及其规模和方向均记录在事件文件中。LOBSTER 还提供了衍生的限价订单簿聚合数据，这是本研究工作的输入数据。对于每个交易日，限价订单簿的演变情况最多记录到 10 个层级。值得注意的是，每笔订单的美元价格乘以 10000 记录。接下来几节中所描述的实验是使用英特尔公司（INTC）股票的重建限价订单簿（LOB）进行的，该公司股票是纳斯达克交易所交易最活跃的大价差股票之一[11]。虽然高市场活跃度可能有助于深度强化学习（DRL）训练（智能体将探索更广泛的相空间），也可能无助于训练（由于价格过程更有效，导致信号与噪声比降低），但我们选择考虑大价差股票这一事实无疑是有益的。实际上，大价差股票的价格，即价格水平间隔较大的股票，相较于小价差股票，其随机性更低，且更依赖于底层可用的限价订单簿信息[11]。此外，由于每个价差都具有实际意义，限价订单簿较为密集，这使得仅将成交量传递给 DRL 智能体也不会造成信息损失。

整个训练数据集由 60 个文件组成（2019 年 2 月 4 日至 2019 年 4 月 30 日期间的每个交易日各一个）。验证数据集包含 22 个文件（2019 年 5 月 1 日至 2019 年 5 月 31 日期间的每个交易日各一个）。测试数据集包含 20 个文件（2019 年 6 月 3 日至 2019 年 6 月 28 日期间的每个交易日各一个）。当前工作中展示的所有实验均基于限价订单簿（LOB）快照进行，其深度（订单簿每侧按最小价格变动单位划分的限价订单层级数量）为 10。由于市场开盘和收盘期间的动态变化差异较大且波动性更高，因此每天的训练集和测试集都排除了这两个时段（开盘和收盘前后的 2×10^{5} 个最小价格变动单位）。这样做是为了让智能体在*正常*的市场条件下进行学习。

B. 模型

在当前的研究中，我们考虑了三种不同的训练和测试场景 $ci\in[201,202,203]$，它们仅在对深度强化学习（DRL）智能体的状态 S 的描述方式上有所不同：

- S_{c201}：在时间 τ 时，代理的状态由买卖双方前十个层级的限价订单簿（LOB）的成交量来定义。此外，代理还获得了过去 9 个时间间隔的限价订单簿状态以及代理当前的持仓情况（多头、空头、中性）。
- S_{c202}：代理人在时间τ的状态定义如S_{c201}所述，另外还需加上代理人当前未平仓头寸的市值（若未持仓，则按定义为零）。市值定义为若代理人决定在当前时间平仓所能获得的利润。这使得代理人能够根据其交易的表现以及平仓可能带来的收益（从而回报）来评估行动。此外，相较于S_{c201}有所改善的回报情况表明，交易的最佳退出时机并非独立于其当前的回报。
- S_{c203}：代理在时间 τ 的状态定义如 S_{c202} 所示，另外增加了当前的买卖价差。这使得代理能够在进行交易前评估交易成本。

开仓/平仓。这些会影响潜在交易的盈利能力，从而影响对一笔交易潜在净收益的估算能力。

对于所有不同的状态定义，在每个时间步长 τ 时，代理必须选择执行特定的动作 $a \in A$。代理可用的所有动作集合为 $A = \{a_i\}_{i=0}^{3} = \{$卖出，持有，*买入*，*每日止损*$\}$。每个动作的结果取决于代理的持仓状态 $P = \{p_i\}_{i \in N,L,S} = \{$中性，多头，空头$\}$。一个持仓-动作对 (p_i, a_i) 完全定义了代理如何与限价订单簿环境进行交互。因此，我们列出所有可能的对及其相应的结果。

- (p_N, a_0): 代理决定卖出（即 a_0)），即便其当前并不持有该股票（即 N）。因此，它决定建立一个针对基础股票单位的空头头寸 $p : N \rightarrow S$。建立空头头寸即决定以当前最优买价卖出股票，并在未来某个时间以届时的市场价格买回。这一行动基于对未来价格下跌的预期。
- (p_N, a_2): 代理人决定买入（即 a_2)），而目前并不持有该股票（即 N）。因此，它决定建立一个单位标的股票的多头头寸 $p : N \rightarrow L$。建立多头头寸即决定以当前最优卖价买入股票，并在未来某个时间以届时的市场价格卖出。这一行动很可能是基于对未来价格上涨的预期。
- $(\cdot, a_1)$: 无论其状态如何，如果智能体决定处于 A_1 状态，它将不会执行任何操作。
- (p_S, a_0): 代理决定卖出（即 a_0)，即便其已持有空头头寸（即 S）。头寸保持不变，不再进行新的交易（无变化）。这源于本研究的基本方法，即代理只能持有单位股票头寸。
- (p_L, a_2): 代理决定买入（即 a_2)，即便其已持有多头头寸（即 L）。头寸保持不变，不再进行新的交易（无变化）。这源于本研究的基本方法，即代理只能持有单位股票头寸。
- (p_L, a_0): 代理决定卖出（即 a_0) 在持有多头头寸（即 L）时。该头寸被平仓并恢复为零 $p : L \rightarrow N$。与该交易相关的利润因此被计算出来，并作为代理行动的反馈提供给代理。
- (p_S, a_2): 代理决定买入（即 a_2) 在持有空头头寸（即 S）时）。头寸被平仓并恢复为零 $p : S \rightarrow N$。与该交易相关的利润因此被计算出来，并作为代理行动的反馈提供给代理。
- $(\cdot, a_3)$: 代理会检查当天的累计收益 Rday；若 Rday < 0，则当前仓位将被平仓（执行日止损操作 a_3) 并且当天的进一步交易将被禁止以避免过度损失，否则将执行持仓操作（即 a_1) ）。

如上所述，对于某些持仓-操作组合，代理体会经历非零的收益（或损失）。$_{a/b}$，$\tau -$ 与在时间 $\tau_{a\ 结束\ 在时间\ \tau-t\ 开立的多头/空头头寸相关的收益}\tau-t$，直观地定义为 $R_{l/s}$ = pbest best，即交易的净美元收益（价格差减去交易成本）。直接纳入价差穿越动态（在买价/卖价而非中间价进行交易）直接将交易成本纳入考量，使实验更贴近现实。

此外，引入*每日止损*操作的理由在于，如果根据当前策略，实施盈利交易策略难度过大，代理方可能会决定跳过整个环境（即一天）。

在第 d 个交易日的第 E 个周期的第 e 个时间步长结束时，智能体所体验到的奖励函数 $r_{e,E,d}$ 是动作-状态对的函数，对于除止损之外的所有动作，该奖励函数都与收益 $R_{l/c}$ 相同。

在此需要作一个最终的说明。细心的读者会注意到，我们所定义的代理人在每个时点只能交易并持有基础资产的一个单位。尽管这看似是一个不切实际的假设，会影响我们方法的适用性，但我们恰恰认为并非如此。

代理人在限价订单簿（LOB）上执行的每一项操作都会改变 LOB 本身，并影响其他市场参与者的未来行动和反应。每次我们买入（卖出）一个或多个单位的资产时，我们就在卖价（买价）处消耗了流动性。这一简单行为会创造需求过剩，从而导致价格按订单规模平方根的比例上涨（下跌）[11]。通过仅交易一个单位，我们能够确保有足够的流动性（因为最小单位为一个）让代理人在给定的最佳买价/卖价处执行交易，而无需消耗限价订单簿中更高层级的流动性（这会恶化执行价格并影响收益）。此外，较大规模的交易通常会被分解为较小的单位分别执行。可变的订单规模意味着需要更复杂的执行策略，并且代理人在下单时需要考虑对执行价格的自我影响。因此，我们选择单单位交易，这支持了我们的框架及其结果在实际场景中的简单性和适用性。

C. 训练-测试流程

所有实验都分为三个独立的阶段。

1) 训练阶段采用持续学习的方法进行。对于 82 个可用交易日中的每一个，选取 5 个连续 10^{-4} 个报价的限价订单簿快照。这些快照包含了当日中间价的最大绝对差值。从中仅抽取 25 个快照，并将其转换为向量化的 DRL 环境 [10]。这种抽样程序能够选择包含高信噪比和更有可能采取行动的价格变化的数据。然后，每个环境运行 30 个周期。底层代理的策略网络由一个简单的多层感知机表示，该感知机由两个各含 64 个神经元的隐藏层组成，这两个隐藏层在动作网络和价值网络之间共享。

最后的隐藏层直接映射到网络的两个最终单值输出（动作和价值）上。

2) 对特定模型超参数（即学习率和熵系数）在验证集上执行贝叶斯优化（BO）[47]-[50]，该验证集的环境（即限价订单簿快照）与训练集的采样方式类似。使用具有平方指数核函数的高斯过程（GP）回归器[51]来拟合将超参数值映射到样本外奖励的未知函数。基于所有验证交易日累计的利润，利用预期改进函数[52]、[53]迭代选择下一个待评估的实现。拟合后的函数进而给出用于训练的最优超参数值。

3) 在最后阶段，智能体所学习到的策略会在测试集的所有日期上独立进行测试。每当平仓时，累计日收益 R_{day} 会更新一次，并记录相应的奖励，以定义损益轨迹。因此，所获得的结果用于比较不同状态表征对智能体样本外表现的影响。

V. 结果与讨论

我们报告了根据第 IV-B 节中状态定义的 DRL 模型集成的样本外（测试）结果。为了比较性能，我们从以下角度考虑了代理生成的回报：测试天数的累计损益、不同迭代中的日损益均值和标准差以及代理在测试天数和迭代中的交易回报分布。

图 2、3、4 展示了上述三种分析以及第 IV-B 节中三种状态特征的结果。

我们首先注意到，对于每种状态特征，代理在它们的集成中平均能够产生盈利策略（扣除交易成本）。从图 2(a)、3(a)、4(a) 可以看出，在 S_{c201} 和 S_{c202} 之间，累积利润的规模以及随时间持续上升的趋势都有显著改善。S_{c203} 的表现优于 S_{c201}，但逊于 S_{c202}。这表明在 S_{c202} 状态定义中加入的市值对代理的预期收益具有高度的指示性，并且极大地提高了收敛性和性能。另一方面，包含在 S_{c203} 中的价差并未改善性能，这可能是由于在大单位股票中价差基本保持不变。与 S_{c201} 相比，S_{c203} 的表现更优，这可能是由于状态定义中包含了市值。未来的工作应研究更大范围的环境定义，因为这超出了本基础性和初步性工作的范围。

图 2(b)、3(b)、4(b) 中的曲线展示了每个时间点上所有集成元素的累计日均利润及其一个标准差的置信区间。我们观察到均值和置信区间都呈正偏态，并且存在时间趋势，某些特定日期的波动性比其他日期更大。这些图表能够让我们看到代理在不同环境（即市场状况）下的表现，并且清楚地表明利润并非具有日季节性特征，也不是仅集中在极少数的几天。实际上，我们注意到在整个测试集中，大多数日期都呈现出正收益，而且平均而言，正收益大于负收益。

我们注意到，S_{c202}所达到的损益水平比其他状态特征高出一个数量级。这一结果表明，该代理利用了更多的交易机会。这可以理解为代理掌握了其当前未实现利润的信息，从而能够更平稳地进行交易操作。实际上，在日收益图中我们观察到更平滑的曲线，而且交易次数比 S_{c201} 高出两个数量级（比 S_{c203} 高出一个数量级）。

在查看代理人在各天和各集合实现中进行的每笔交易的回报分布（图 2(c)、3(c)、4(c) 中的直方图）时，我们立即注意到不同状态特征下的交易频率存在差异。如上所述，有关当前未实现利润的信息使代理人能够更准确、更频繁地管理其交易活动。然而，就收益分布的主体形状以及存在正偏态尾部和零附近峰值而言，实际收益分布大致相似。负收益分布尾部的截断可能归因于代理人可用的当日止损操作。正如预期的那样，图 3(c)、4(c) 中几乎对称的分布主体表明价格可预测性有限。实际上，许多交易在早期就被放弃，其结果为正或为负的可能性相等。对于较大的回报，这种差异可能源于以下几种可能的解释：价差（交易成本）的影响较小，较大的价格波动源于可利用的市场低效，而且在缺乏有意义的价格可预测性的情况下，良好的头寸管理能够实现正收益。后一种情况可能适用于 S_{c202}、S_{c203}，而罕见的低效可能是 S_{c201} 盈利的来源。然而，这直接表明，至少在微观结构领域，市场在短期内是低效的（并非仅仅指价格可预测性，而是指可被利用以获取利润）。这一发现可能是由于该策略基于创新的深度强化学习技术，这些技术尚未在该领域广泛实施，因此尚未利用并消除这些低效因素。

在观察各分布的主体部分及其尾部时，并没有强有力的证据能够证明不同交易状态特征之间存在差异巨大的收益情况。真正导致不同收益情况的因素似乎是交易次数，这在直方图中有所体现。实际上，知晓其市值的交易者确切地了解其潜在收益，只需评估进一步上涨的可能性是否值得继续持有头寸。这使得他们能够利用更多的交易机会。

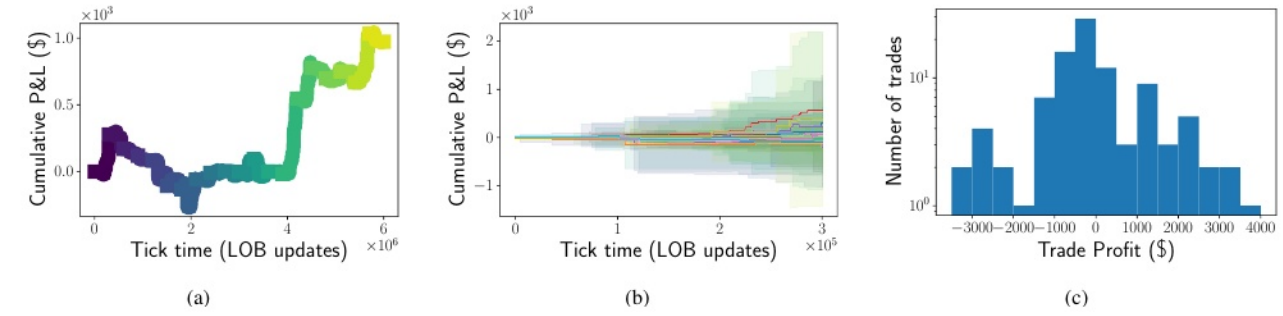

图 2.在完整测试（20 天）和代理状态定义 Sc_{201} 的情况下，30 个集成元素的累计平均回报（a）、每日平均累计回报和标准差（b）以及交易回报分布（c）。请注意，盈亏值直接取决于这样一个事实，即在 LOBSTER 数据中，每个订单的美元价格乘以 10000（见第 IV-A 节）。

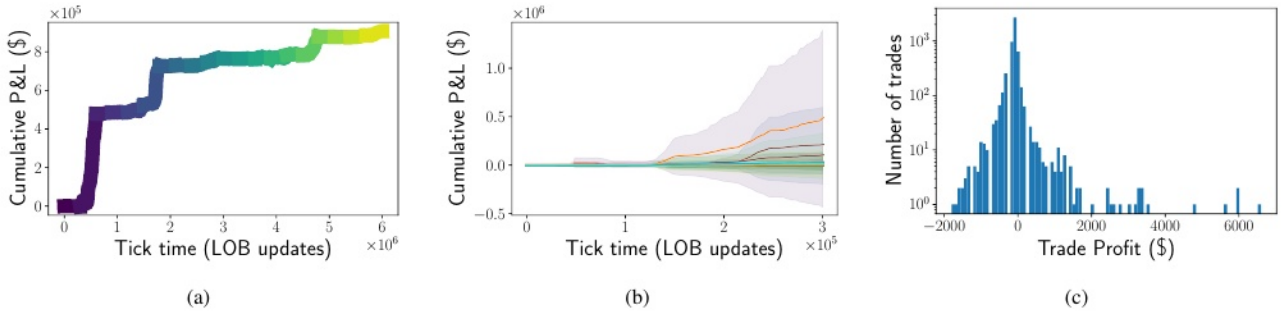

图 3. 对于完整测试（20 天）和代理状态定义 Sc_{202}，30 个集成元素的累积平均回报（a）、每日平均累积回报和标准差（b）以及交易回报分布（c）。请注意，盈亏值直接取决于这样一个事实，即在 LOBSTER 数据中，每个订单的美元价格乘以 10000（见第 IV-A 节）。

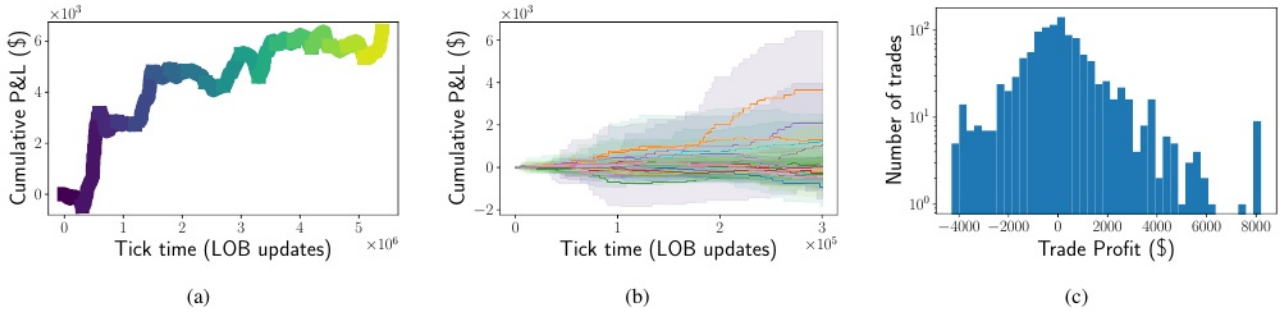

图 4. 对于完整测试（20 天）和代理状态定义 Sc_{203}，30 个集成元素的累计平均回报（a）、每日平均累计回报和标准差（b）以及交易回报分布（c）。请注意，盈亏值直接取决于这样一个事实，即在 LOBSTER 数据中，每个订单的美元价格乘以 10000（见第 IV-A 节）。

这会导致损益增加，但对单笔交易的质量和盈利能力似乎没有明显影响。实际上，这表明在各种策略中存在一个典型的交易和回报周期，这可能是该资产及其价格动态的特征。

VI. 限制条件

本研究工作所提出的方法存在一些局限性。其中之一是样本效率低下，这是强化学习中的常见问题。由于智能体仅在平仓时获得奖励，这意味着反馈的频率仅取决于智能体的交易频率。在我们的案例中，持仓时间可能长达数十、数百甚至数千个时间步长，这阻碍了收敛速度，并导致奖励函数变得稀疏。

此外，鉴于任务的挑战性以及奖励的稀疏性，我们发现，偶尔地，智能体需要正确初始化（即收敛到次优策略，比如完全不进行交易）。这一发现是意料之中的，源于 PPO 算法的探索策略以及智能体在探索阶段初期所经历的负预期回报。有一些解决方案可以缓解这些问题，比如对输入的时间序列进行时间聚合以减少奖励的稀疏性，或者使用具有更合适探索策略的不同强化学习算法，但我们将其留作未来的工作。

VII. 结论

在本研究中，我们展示了如何在高频交易的背景下成功训练和部署深度强化学习模型。我们研究了三种不同的状态定义如何影响代理的样本外表现，并发现了解其当前持仓的市值回报率对整体损益函数和单笔交易奖励都有极大的益处。然而，值得注意的是，无论采用何种状态定义，代理始终能够“跑赢市场”，在所考虑的整个训练样本中以及其中的大多数单个交易日中都能实现净正收益。

尽管我们考虑的是在最优卖价买入、最优买价卖出的交易者（因此需要支付价差来平仓），但此处所做的尝试只是在限价订单簿和金融市场这种明显非平稳环境中实际应用深度强化学习算法的初步尝试。首先，我们考虑的是只能交易单一单位基础资产的交易者。因此，我们合理地认为其自身交易对市场的影响为零。然而，众所周知，在现实环境中，可以买卖多个单位的基础资产，自身交易的影响会严重地影响策略的最终收益。不过，即便在如此简化的环境中，深度强化学习代理也能在高度随机和非平稳的环境中表现出色这一事实本身就值得重视。此外，能够生成样本外盈利策略表明，市场存在暂时的低效，可以加以利用以获取长期利润，从而为长期以来认为市场本质上是有效的这一观点提供了反驳的证据。

参考文献

[1] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: A survey,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[2] M. Deisenroth and C. E. Rasmussen, “Pilco: A model-based and data-efficient approach to policy search,” in *Proceedings of the 28th International Conference on machine learning (ICML-11)*, 2011, pp. 465–472.

[3] Y. Ling, S. A. Hasan, V. Datla, A. Qadir, K. Lee, J. Liu, and O. Farri, “Diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning: A preliminary study,” in *Machine Learning for Healthcare Conference*, 2017, pp. 271–285.

[4] Y. Li, “Deep reinforcement learning: An overview,” *arXiv preprint arXiv:1701.07274*, 2017.

[5] C. Comerton-Forde and T. J. Putnin、s˅, “Dark trading and price discovery,” *Journal of Financial Economics*, vol. 118, no. 1, pp. 70–92, 2015.

[6] M. O’Hara, “High frequency market microstructure,” *Journal of Financial Economics*, vol. 116, no. 2, pp. 257–270, 2015.

[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.

[8] F. Hutter, H. H. Hoos, and K. Leyton-Brown, “Sequential model-based optimization for general algorithm configuration,” in *International conference on learning and intelligent optimization*. Springer, 2011, pp. 507–523.

[9] A. Briola, J. Turiel, and T. Aste, “Deep learning modeling of limit order book: a comparative perspective,” *arXiv preprint arXiv:2007.07319*, 2020.

[10] A. Hill, A. Raffin, M. Ernestus, A. Gleave, A. Kanervisto, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, “Stable baselines,” https://github.com/hill-a/stable-baselines, 2018.

[11] J.-P. Bouchaud, J. Bonart, J. Donier, and M. Gould, *Trades, Quotes and Prices: Financial Markets Under the Microscope*. Cambridge University Press, 2018.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[13] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[14] P. Abbeel and J. Schulman, “Deep reinforcement learning through policy optimization,” *Tutorial at Neural Information Processing Systems*, 2016.

[15] M. Kearns and Y. Nevmyvaka, “Machine learning for market microstructure and high frequency trading,” *High Frequency Trading: New Realities for Traders, Markets, and Regulators*, 2013.

[16] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, “Forecasting stock prices from the limit order book using convolutional neural networks,” in *2017 IEEE 19th Conference on Business Informatics (CBI)*, vol. 1. IEEE, 2017, pp. 7–12.

[17] J. Sirignano and R. Cont, “Universal features of price formation in financial markets: perspectives from deep learning,” *Quantitative Finance*, vol. 19, no. 9, pp. 1449–1459, 2019.

[18] Z. Zhang, S. Zohren, and S. Roberts, “Bdlob: Bayesian deep convolutional neural networks for limit order books,” *arXiv preprint arXiv: 1811.10041*, 2018.

[19] ——, “Extending deep learning models for limit order books to quantile regression,” *arXiv preprint arXiv:1906.04404*, 2019.

[20] J. Moody and M. Saffell, “Learning to trade via direct reinforcement,” *IEEE transactions on neural Networks*, vol. 12, no. 4, pp. 875–889, 2001.

[21] J. E. Moody and M. Saffell, “Reinforcement learning for trading,” in *Advances in Neural Information Processing Systems*, 1999, pp. 917– 923.

[22] J. Moody, L. Wu, Y. Liao, and M. Saffell, “Performance functions and reinforcement learning for trading systems and portfolios,” *Journal of Forecasting*, vol. 17, no. 5-6, pp. 441–470, 1998.

[23] R. Neuneier, “Optimal asset allocation using adaptive dynamic programming,” in *Advances in Neural Information Processing Systems*, 1996, pp. 952–958.

[24] O. Mihatsch and R. Neuneier, “Risk-sensitive reinforcement learning,” *Machine learning*, vol. 49, no. 2-3, pp. 267–290, 2002.

[25] J. Moody, M. Saffell, W. L. Andrew, Y. S. Abu-Mostafa, B. LeBaraon, and A. S. Weigend, “Minimizing downside risk via stochastic dynamic programming,” *Computational Finance*, pp. 403–415, 1999.

[26] T. L. Meng and M. Khushi, “Reinforcement learning in financial markets,” *Data*, vol. 4, no. 3, p. 110, 2019.

[27] D. Byrd, M. Hybinette, and T. H. Balch, “Abides: Towards high-fidelity market simulation for ai research,” *arXiv preprint arXiv:1904.12066*, 2019.

[28] M. Karpe, J. Fang, Z. Ma, and C. Wang, “Multi-agent reinforcement learning in a realistic limit order book market simulation,” *arXiv preprint arXiv:2006.05574*, 2020.

[29] F. Bertoluzzo and M. Corazza, “Testing different reinforcement learn-ing configurations for financial trading: Introduction and applications,” *Procedia Economics and Finance*, vol. 3, pp. 68–77, 2012.

[30] L. Chen and Q. Gao, “Application of deep reinforcement learning on automated stock trading,” in *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2019, pp. 29–33.

[31] Q.-V. Dang, “Reinforcement learning in stock trading,” in *International Conference on Computer Science, Applied Mathematics and Applica-tions*. Springer, 2019, pp. 311–322.

[32] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, “Deep direct reinforcement learning for financial signal representation and trading,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 3, pp. 653–664, 2016.

[33] G. Jeong and H. Y. Kim, “Improving financial trading decisions using deep q-learning: Predicting the number of shares, action strategies, and transfer learning,” *Expert Systems with Applications*, vol. 117, pp. 125–138, 2019.

[34] Y. Kim, W. Ahn, K. J. Oh, and D. Enke, “An intelligent hybrid trading system for discovering trading rules for the futures market using rough

sets and genetic algorithms," *Applied Soft Computing*, vol. 55, pp. 127– 140, 2017.

[35] Z. Zhang, S. Zohren, and S. Roberts, "Deep reinforcement learning for trading," *The Journal of Financial Data Science*, vol. 2, no. 2, pp. 25–40, 2020.

[36] T. Th´eate and D. Ernst, "An application of deep reinforcement learning to algorithmic trading," *arXiv preprint arXiv:2004.06627*, 2020.

[37] H. Yang, X.-Y. Liu, S. Zhong, and A. Walid, "Deep reinforcement learning for automated stock trading: An ensemble strategy," *Available at SSRN*, 2020.

[38] Z. Jiang and J. Liang, "Cryptocurrency portfolio management with deep reinforcement learning," in *2017 Intelligent Systems Conference (IntelliSys)*. IEEE, 2017, pp. 905–913.

[39] Z. Zhang, S. Zohren, and S. Roberts, "Deep learning for portfolio optimisation," *arXiv preprint arXiv:2005.13665*, 2020.

[40] Y.-J. Hu and S.-J. Lin, "Deep reinforcement learning for optimizing finance portfolio management," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*. IEEE, 2019, pp. 14–20.

[41] Y.-S. Lim and D. Gorse, "Reinforcement learning for high-frequency market making." in *ESANN*, 2018.

[42] Y. Wang, "Electronic market making on large tick assets," Ph.D. dissertation, The Chinese University of Hong Kong (Hong Kong), 2019.

[43] R. Huang and T. Polak, "Lobster: Limit order book reconstruction system," *Available at SSRN 1977207*, 2011.

[44] T. H. Balch, M. Mahfouz, J. Lockhart, M. Hybinette, and D. Byrd, "How to evaluate trading strategies: Single agent market replay or multiple agent interactive simulation?" *arXiv preprint arXiv:1906.12010*, 2019.

[45] M. Bibinger, C. Neely, and L. Winkelmann, "Estimation of the discontinuous leverage effect: Evidence from the nasdaq order book," *Journal of Econometrics*, vol. 209, no. 2, pp. 158–184, 2019.

[46] M. Bibinger, M. Jirak, M. Reiss *et al.*, "Volatility estimation under one-sided errors with applications to limit order books," *The Annals of Applied Probability*, vol. 26, no. 5, pp. 2754–2790, 2016.

[47] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2015.

[48] P. I. Frazier, "Bayesian optimization," in *Recent Advances in Optimiza-tion and Modeling of Contemporary Problems*. INFORMS, 2018, pp. 255–278.

[49] F. Archetti and A. Candelieri, *Bayesian Optimization and Data Science*. Springer, 2019.

[50] A. Candelieri and F. Archetti, "Global optimization in machine learning: the design of a predictive analytics application," *Soft Computing*, vol. 23, no. 9, pp. 2969–2977, 2019.

[51] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Springer, 2003, pp. 63–71.

[52] J. Mo˘ckus, "On bayesian methods for seeking the extremum," in *Optimization techniques IFIP technical conference*. Springer, 1975, pp. 400–404.

[53] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimiza-tion of expensive black-box functions," *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.