arXiv:2611.32790v1 [q-fin.CP] 24 Nov 2026

运用深度强化学习进行Delta-Vega对冲：从模拟走向现实*

Michele Colombi†[1]、Marco Di Francesco‡[2,3]和Pietro Rossi§[2,4]

[1]比萨高等师范学校，意大利比萨[2]博洛尼亚大学统计科学系，意大利[3]*Leith`a*（*Unipol*集团），意大利博洛尼亚[4]*Prometeia*股份有限公司，意大利博洛尼亚

2026年11月24日

摘要

仅使用标的资产对期权组合进行对冲，会使头寸暴露于波动率变化之下；通过交易另一种衍生品可以管理这种风险，但会产生更高的交易成本。我们提出了一种深度强化学习方法，通过交易标的资产和第二个期限更长的期权，在离散时间交易和交易成本下同时对冲价格风险与波动率风险。我们使用随机波动率模型生成的合成数据，通过近端策略优化（PPO）训练智能体，并将其状态限制为市场中可观测的量，使得在模拟中学习到的策略能够直接应用于市场数据。在模拟中，我们将该策略与半解析Delta和Delta-Vega对冲基准进行比较验证。在基准交易成本下，PPO达到了与Delta-Vega相同的目标；随着成本上升，其表现逐渐超过Delta-Vega，同时，两种双工具策略始终优于单工具Delta对冲。在历史标普500指数期权数据上，我们将该方法与实务中的Black-Scholes Delta和Delta-Vega对冲策略进行基准比较。两种双工具策略使Delta对冲的损益离散度大致降低了一半；在交易成本相当的情况下，PPO实现了低于Delta-Vega的损益波动率。与Delta-Vega相比，PPO持有更多对冲期权，这与随机波动率模型对冲相一致；其优势主要集中在高波动时期。

关键词：深度强化学习；期权对冲；Vega；随机波动率；交易成本；仿真到现实迁移。

*我们感谢比萨高等师范学校提供市场数据。我们感谢参加第十三届巴舍利耶世界大会（博洛尼亚大学，2026年）、第二十七届量化金融研讨会（贝加莫大学，2026年）和国际金融科技研究会议（帕维亚大学，2026年）的与会者提出的宝贵意见并参与讨论。

†通讯作者。电子邮箱：michele.colombi@sns.it

‡电子邮箱：marco.difrancesco3@unibo.it, marco.difrancesco@leitha.eu§电子邮箱：pietro.rossi3@unibo.it,pietro.rossi@prometeia.it

1 引言

动态期权对冲是金融理论与实践中的长期问题。在经典的 Black-Scholes（BS）框架下（Black 和 Scholes，1973年；Merton，1973年），通过持续调整由标的资产和无风险账户组成的投资组合，可以实现最优对冲策略。然而，该策略依赖于若干假设，例如不存在交易成本、进行连续时间交易以及波动率保持恒定。事实上，这些条件在金融市场中并不成立。交易成本和离散交易频率在成本与对冲准确性之间引入了权衡。此外，随着波动率随时间变化，仅使用标的资产进行对冲的投资组合仍会面临波动率风险，而这一风险可以通过交易另一种对冲工具来管理。

近年来，深度强化学习（RL）已成为金融领域动态优化问题的有力工具，因为与经典随机控制相比，它依赖的模型假设更少，并且在现实市场模型变得难以处理时仍然适用（Hambly等人2023年）。在期权对冲中，已有研究表明，RL能够有效学习考虑市场摩擦的对冲策略，其表现优于BS Delta对冲等经典基准策略（Buehler等人 2019年；Kolm和Ritter（2019）；Du等人 2020年；Cao、Chen、Hull等人 2021年）。其中，Delta是期权价格相对于标的资产价格的一阶偏导数。

尽管具有这种灵活性，基于RL的方法大多仍然针对单一工具的Delta对冲策略（Pickard and Lawryshyn 2023年）。然而，Delta对冲会使投资组合暴露于波动率变化之下（Bakshi and Kapadia 2003年）。市场参与者通过实务中的Black-Scholes Vega来衡量这一暴露，该指标定义为期权价格相对于其隐含波动率的一阶偏导数。通过使用基于Vega的调整修正BS Delta，仅依靠标的资产便可以对冲部分波动率风险（Hull 和 White 2017年）。然而，在随机波动率条件下，完备对冲需要第二种工具，即一种可交易衍生品，因为标的资产无法使市场完备（Romano and Touzi 1997年）。在实践中，期权交易者通过交易流动性较高的普通香草期权来管理其头寸的Vega。由于即使是流动性最高的期权，其买卖价差也明显大于标的资产，因此，在这一环境下，考虑成本的对冲策略尤为重要。然而，只有少数基于RL的研究在交易成本下使用第二种对冲工具：有些研究以Gamma对冲为基准（Armstrong and Tatlow 2024年；Fraņcois et al. 2025年），其中Gamma是期权价格相对于标的资产价格的二阶偏导数；另一些研究在处理波动率风险时，则依赖模拟环境作为概念验证（Buehler等人 2019年；Cao、Chen、Farghadani 等（2023））。

我们在这一研究方向的基础上，提出一种深度强化学习方法，利用标的资产和一种对冲期权共同对冲价格风险和波动率风险，且二者均需承担交易成本；我们在受控的模拟环境中对该方法进行验证，并将其直接迁移至真实市场数据。我们采用近端策略优化（PPO）（Schulman、Wolski 等，2017年），并利用由 Heston（1993年）随机波动率模型生成的数据训练智能体。遵循 Kolm and Ritter（2019年）的做法，我们假设智能体采用均值-方差目标，以刻画成本与复制之间的权衡。重要的是，我们将智能体可获得的信息集限制为市场可观测变量，即价格、持仓和合约具体信息。我们不向智能体提供驱动数据生成过程的模型特定潜变量，因此，在模拟中学习到的策略无需任何进一步估计即可直接应用于真实市场数据。

因此，我们的方法是模型无关的：Heston模型仅用于生成训练数据，智能体从模拟价格路径中学习对冲策略，而不是根据模型计算敏感度。我们将学习到的策略称为*Delta-Vega*类策略，因为它交易的两种工具与模型的 Delta-Vega对冲相同；但一旦引入交易成本，或将该策略应用于市场数据，它便不再受模型约束。

我们分三个阶段阐述我们的方法。首先，我们在一个无摩擦的

环境中验证智能体，在该环境下，Heston模型给出最优的半解析Delta-Vega对冲。其次，我们引入交易成本；在这种情况下，模型对冲不再是最优的，智能体可以为了实现其目标而偏离该对冲。我们考虑反映标的资产与对冲期权不同价差特征的各种交易成本情景，并研究智能体在一系列风险厌恶水平下的表现。最后，我们将模拟中学习到的策略迁移至2019年至2022年的每日标准普尔500指数期权数据。这一步着眼于将我们的强化学习对冲策略应用于真实市场数据；与基于模拟的研究相比，这一方面受到的关注仍然有限（Mikkil¨a and Kanniainen 2023）。一项在模型内表现良好的策略，即在用于推导该策略的假设下表现良好的策略，并不一定能在模型外保持稳健，即无法应对模型未能捕捉的市场动态。因此，对在模拟中优化的策略进行实证评估——这一做法在更广泛的机器学习文献中被称为仿真到现实迁移 （Zhao等（2020））——是评估强化学习对冲策略在实践中适用性的关键步骤。

我们的工作有两项主要贡献。首先，我们将强化学习对冲扩展为一种联合的 Delta-Vega类策略，该策略交易标的资产和第二个期限更长的期权，且二者均需承担交易成本。智能体基于Heston数据进行训练，且仅观测市场变量。在没有成本的情况下，该策略能够复现半解析的Heston Delta-Vega对冲。在中等和较高成本下，该策略偏离模型对冲，并取得更高的均值-方差目标。我们通过在成本和风险厌恶水平构成的网格上训练智能体，开展敏感性分析，发现随着成本和风险厌恶程度变化，智能体会在两种工具之间自适应地分配对冲。尽管期权腿的交易成本更高，智能体和 Delta-Vega基准策略的表现均优于仅Delta对冲。其次，我们通过在多个校准参数集和到期日下生成的 Heston数据上训练智能体，并将学得的策略迁移至2019年至2022年的标准普尔500指数期权数据，展示了我们的 Delta-Vega类强化学习策略的适用性。我们采用实务中的 Black-Scholes Delta 和 Delta-Vega对冲作为基准；这些策略在市场参与者中得到广泛采用。结果发现，在交易成本相当的情况下，近端策略优化智能体实现了更低的损益波动率。我们观察到，与实务Delta-Vega策略相比，强化学习策略系统性地对冲过度配置了对冲期权，这与智能体所接受训练的 Heston模型对冲一致；我们还发现，智能体相对于实务基准策略的优势主要集中在高波动时期。这凸显了训练环境对学得策略的塑造作用，并有助于深入了解类似深度强化学习方法在现实场景中的适用性。

本文其余部分的组织如下。第2节概述相关工作。第3节介绍强化学习的相关记号和PPO算法。第4节概述期权对冲设置，并展示如何将其嵌入强化学习框架。第5节给出数值结果，首先研究包含和不包含交易成本的合成环境情形，随后研究将PPO策略实证迁移至标准普尔500指数市场数据的表现。最后，第6节总结全文。

2 相关工作

Halperin（2020年）是最早将期权对冲视为强化学习问题的研究者之一，建立了BS定价框架与Q-learning（一种基础强化学习算法）之间的正式联系。与本文更相关的是，Kolm and Ritter（2019年）表明，在交易成本下、且不了解市场模型的情况下，为优化均值-方差目标而训练的基于强化学习的对冲策略，其表现优于BS-Delta策略。Buehler等人（2019年）提出了深度对冲，即在凸风险度量和比例交易成本下优化神经网络对冲策略。在此基础上，越来越多的文献通过采用最先进的深度强化学习算法（Du等人（2020年）；Cao、Chen、Hull等人（2021年）；Cannelli等人（2023年））来改进方法论设定，这些算法在强化学习架构中采用了

神经网络 网络，或使用风险感知目标（Vittori 等（2020））。

上述大多数研究都集中于基于模拟的实验，其中使用由无套利模型生成的合成数据进行训练和测试。只有部分文献还研究了实证环境下的基于强化学习的对冲策略，而这是该方法能否实际应用的关键因素。一个早期例子是Buehler等人（2019）的研究：他们仅使用指数对平值标普500看涨期权进行对冲，在根据历史收益率拟合的波动率（GARCH）模型上进行训练，并在随后的一段历史路径上进行测试；结果表明，在经过滤波的波动率条件下，深度对冲的表现优于BS对冲。Giurca和Borovkova（2021）表明，在由扩散模型（Black-Scholes和Heston）集成生成的合成数据上训练的深度强化学习智能体，能够成功地将Delta对冲策略迁移至标普500期权数据。Xiao等（2021）同时使用合成的和实证的标普500指数期权路径进行训练，发现总体表现良好，但没有系统比较两种训练机制下的结果。Xu和Dai（2022）完全基于实证数据进行训练和测试，涵盖标普500、标普100和道琼斯工业平均指数。值得注意的是，作者除Delta外还针对Gamma风险，但仅依赖标的资产作为对冲证券，并在奖励函数中加入了与Gamma相关的惩罚项。Zernikov（2026）同样采用纯实证训练机制，使用2015年至2023年的标普500日度数据。作者设计了滚动前向测试：使用一年的数据进行训练，并在下一年进行测试；结果发现，与BS Delta相比，强化学习对冲存在系统性折减，同时在不同年份之间表现出机制脆弱性。Mikkil¨a和Kanniainen（2023）迄今为止开展了最系统的实证研究，明确比较了在标普500日内数据上分别经过模拟训练和实证训练的智能体，并发现后者表现更优。尽管这一结果印证了强化学习以纯数据驱动方式应对动态对冲的潜力，但我们在此主张采用仿真到现实方法，原因主要有两点。第一，真实数据稀缺且具有偶然性。相比之下，合成环境能够根据所选模型的参数化，在任意机制下生成任意数量的训练回合。第二，模型提供了可控性：在将策略迁移至真实数据之前，可以在完全理解且可控的合成环境中评估算法的准确性和可靠性。

尽管对强化学习策略进行实证评估正日益受到关注，但使用标的资产以外工具的对冲策略研究仍鲜有涉及，且大多局限于模拟环境。在介绍深度对冲的奠基性论文中，Buehler等人（2019年）在 Heston 框架下验证了该方法，其中使用方差互换与标的资产一同作为第二种对冲工具。如上文所预期，作者还在历史标准普尔500指数数据上测试了该方法，但在实证评估中依赖单工具 Delta 对冲。在深度对冲框架下，Armstrong 和 Tatlow（2024年）使用标的资产和另一份看涨期权对一份看涨期权进行对冲，同时考虑标的资产漂移率和波动率的不确定性。在 Black-Scholes 模拟环境中，当损失函数奖励对模型不确定性的稳健性时，所学习的策略趋近于 Gamma对冲。Cao、Chen、Farghadani 等（2023年）利用强化学习，在一份平值对冲期权下对 Gamma风险和 Vega风险进行管理。智能体仅决定期权交易，而 Delta 头寸则在每一步精确地、无成本地进行再平衡。训练和测试数据均基于 SABR型随机波动率模型生成。与我们的实证应用更为接近的是，Fran¸cois 等（2025年）在标普500指数期权数据上测试了深度强化学习代理，使用标的资产和一份短期期权对跨式期权进行对冲，并针对 Gamma风险，以 Delta-Gamma基准进行比较。作者在物理测度下，基于隐含波动率曲面的计量经济学模型进行训练，并以拟合得到的模型因子作为智能体的状态输入。尽管该研究表明隐含波动率信息能够提升表现，但其状态依赖于特定模型的量；此外，由于模拟器并非风险中性，因此必须检验智能体是否获取了风险溢价。相比之下，我们使用一份长期限对冲期权来管理 Vega风险，仅将可观测市场量提供给智能体，并在校准后的风险中性动态下进行训练，在这种情况下，智能体只能通过降低方差和成本来改善其

其目标仅是降低方差和成本。

数据驱动对冲的另一种路径是学习市场动态，而非对冲策略。这一方向的少数研究聚焦于使用期权进行对冲。Cohen 等人（2022年）将神经随机微分方程市场模型拟合到期权价格，并根据拟合模型推导对冲比率。在市场数据上，由此得到的对冲策略比 BS Delta 和 Delta-Vega对冲的误差更小，表现与基于 Heston模型的对冲策略相当。Cont 和 Vuleti´c（2025年）使用条件生成模型模拟标的资产与隐含波动率曲面的联合情景，并计算在交易成本下使各情景风险最小化的对冲比率。在样本外的 SPX 期权数据上，作者报告称，该方法的表现与 Delta 和 Delta-Vega 对冲相当，甚至更优。与我们一样，这两项研究都在市场数据上将结果与 Delta-Vega 对冲进行比较，但它们是基于拟合模型计算每个日期的对冲，而非学习对冲策略。

除强化学习之外，我们的工作还涉及数学金融领域的两类经典文献：交易成本下的对冲，以及波动率风险对冲。第一类研究源于 Leland（1985），其将 Black-Scholes 对冲调整为适用于离散交易和比例交易成本的情形，通常通过将再平衡限制在无摩擦对冲附近的无交易区间内来实现（Whalley 和 Wilmott 1997）。这类结果已适用于 Delta对冲；Herrmann 和 Muhle-Karbe（2017）将其扩展至 Delta-Vega对冲，并称这是一个重大挑战。

在第二类文献中，研究者提出了不同模型，以克服 BS 框架中恒定波动率设定的错误设定问题。Cr´epey（2004）将其分为三类：对每个期权使用其隐含波动率的 BS 模型，如实务对冲中所采用的模型（第4.2节）；局部波动率模型（Dupire 1994）；以及随机波动率模型，例如 Heston（1993）模型，我们在模拟实验中采用的正是该模型。在这些模型中，只有随机波动率模型能够一致地处理波动率风险，但代价是模型复杂性增加和市场不完备性（Cr´epey 2004）。如前所述，通过引入普通香草期权作为可交易工具，可以解决这一问题（Romano 和 Touzi 1997； Davis 和 Ob-l´oj 2008 ）。仅使用标的资产时，只能对冲与标的资产相关的那部分波动率风险，通常的做法是根据资产价格变动条件下波动率的预期变化，对 BS Delta 进行调整（Hull 和 White 2017）。实证数据表明，资产收益率与波动率之间的（负）相关性只能解释方差风险溢价的一部分（Carr 和 Wu 2009），这意味着存在一个独立的波动率风险因子。

借助深度强化学习，我们直接从数据中优化对冲策略，从而在统一框架下兼顾离散交易、交易成本和波动率风险。

3 强化学习：背景与记号

本节介绍本文其余部分所使用的强化学习记号，以及我们用于优化对冲策略的近端策略优化（PPO）算法。完整论述请参阅 Sutton 和 Barto（2018年）以及 Albrecht 等人（2024年）。

强化学习（RL）训练一种计算机模型，该模型称为智能体，在环境 中选择一系列动作，以最大化累积奖励信号。RL 问题被形式化为离散时间马尔可夫决策过程（MDP）。由于对冲期次包含有限个决策日期 n，我们考虑有限时域。MDP 是一个元组 $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{P}_0 \rangle$，其中 $\mathcal{S}$ 和 $\mathcal{A}$ 分别为状态空间和动作空间，$\mathcal{P}$（$s' \mid s, a$）表示在状态 $s_i = s \in \mathcal{S}$ 下采取动作 $a_i = a \in \mathcal{A}$ 后转移到状态 $s_{i+1} = s' \in \mathcal{S}$ 的概率，$\mathcal{R}$： $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ 为奖励函数，$\mathcal{P}_0$ 为初始状态 s_0 的分布。我们将 $r_{i+1} := \mathcal{R}(s_i, a_i, s_{i+1})$ 记为从 s_i 转移到 s_{i+1} 后获得的奖励。这一时序符合对冲的情形，即在时间 t_i 选择的头寸所产生的损益只有在 t_{i+1} 才为人所知。智能体不知道 $\mathcal{P}$，只能从采样转移中学习。这是 RL 与动态规划之间的主要

区别，后者直接求解完全确定的 MDP。智能体的行为由随机策略 $\pi(a \mid s)$描述，即在给定当前状态 s 的条件下，对动作 $a \in \mathcal{A}$ 的概率分布。我们将 $\mathbb{E}_\pi[\cdot]$ 记为在 $s_0 \sim \mathcal{P}_0$、 $a_i \sim \pi\left(\cdot \mid s_i\right)$ 和 $s_{i+1} \sim \mathcal{P}\left(\cdot \mid s_i, a_i\right)$ 下生成的轨迹 $(s_0, a_0, s_1, \ldots, a_{n-1}, s_n)$ 上的期望。给定折扣因子 $\gamma \in [0,1]$，从时间 t_i 起的折扣回报为

$$G_i := \sum_{j=i}^{n-1} \gamma^{j-i} r_{j+1}, \tag{1}$$

智能体寻求能够使回合的期望折扣回报最大化的策略

$$J(\pi) := \mathbb{E}_\pi\left[G_0\right] = \mathbb{E}_\pi\left[\sum_{i=0}^{n-1} \gamma^i r_{i+1}\right]. \tag{2}$$

π 的价值函数和动作价值函数分别表示从时间 t_i起，在给定当前状态和当前状态—动作对的条件下的期望回报，

$$V_\pi(s) := \mathbb{E}_\pi\left[G_i \mid s_i = s\right], \tag{3}$$

$$Q_\pi(s,a) := \mathbb{E}_\pi\left[G_i \mid s_i = s,\, a_i = a\right]. \tag{4}$$

这两个函数仅通过状态依赖于时间 t_i，而状态包含到期时间（见第 4.4 节）。二者之差

$$A_\pi(s,a) := Q_\pi(s,a) - V_\pi(s) \tag{5}$$

即优势函数，它表示在状态 s下，动作 a 相对于从 π 中抽取的平均动作，预期表现会好多少或差多少。

策略梯度方法基于价值的方法（例如 Halperin（2020）的 Q-learning 方法和 Kolm 与 Ritter（2019）的 SARSA 算法）通过估计动作价值函数，并在每个状态下对 $\mathcal{A}$ 取最大值来恢复策略。当动作是连续变量时，这种做法会变得不切实际，而对冲比率正是如此。策略梯度方法则直接将策略参数化为 π_θ，其参数为 $\theta \in \Theta$，并沿着 $J(\pi_\theta)$的梯度上升。根据策略梯度定理（Sutton、McAllester 等人，1999），该梯度可写为[1]

$$\nabla_\theta J(\pi_\theta) \propto \mathbb{E}_{\pi_\theta}\left[A_{\pi_\theta}(s_i, a_i)\, \nabla_\theta \log \pi_\theta(a_i \mid s_i)\right], \tag{6}$$

其中，期望是针对在 π_θ下访问到的状态-动作对取得的。式（6）提高具有正优势的动作的概率，并降低具有负优势的动作的概率。

演员-评论家架构 由于 $\mathcal{P}$ 未知，必须根据采样转移估计优势。此外，当 $\mathcal{S}$ 为连续变量时，存储每个状态对应一个值的表格表示并不可行，而且无法推广到训练期间未访问过的状态。深度 强化学习通过使用神经网络对相关函数进行参数化来解决这一问题。在演员-评论家 架构中，行动者 网络对策略 π_θ进行参数化，而评论家 网络V_ϕ（参数为 $\phi \in \Phi$）则近似价值函数 V_{π_θ}。评论家通过时序差分误差 δ_i 参与优势估计：$= r_{i+1} + \gamma V_\phi(s_{i+1}) - V_\phi(s_i)$，其中 $V_\phi(s_n)$：$= 0$。

[1]该定理通常用 Q_{π_θ}替代 A_{π_θ}来表述。减去基线 $V_{\pi_\theta}(s)$不会改变梯度，因为对于每个 s都有$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a \mid s)\right] = 0$，并且会降低其样本估计的方差（Sutton 和 Barto 2018年）。由于期望是对一个回合中的各个步骤取平均，因此该关系只在相差一个常数的意义下成立。对于 $\gamma = 1$，该常数就是回合长度 n。

遵循广义优势估计（GAE）方法（Schulman、Moritz 等，2018年），我们设定

$$\hat{A}_i := \sum_{j=i}^{n-1} (\gamma \lambda_{\mathrm{GAE}})^{j-i}\, \delta_j. \tag{7}$$

参数 $\lambda_{\mathrm{GAE}} \in [0,1]$ 在偏差与方差之间进行权衡，因为 $\lambda_{\mathrm{GAE}} = 0$ 给出一步估计 δ_i，其完全依赖评论家，而 $\lambda_{\mathrm{GAE}} = 1$ 给出 $G_i - V_\phi(s_i)$，除基线外不再依赖评论家。

近端策略优化 策略梯度方法使用 $J(\pi_\theta)$ 的局部一阶近似来更新 θ。由于 π_θ 可能以高度非线性的方式依赖于 θ，参数空间中的一小步变化可能会导致所得策略发生巨大改变。由于数据是在当前策略下收集的，错误的更新还会降低下一次更新所使用数据的质量，并且难以轻易纠正。近端策略优化（Schulman、Wolski 等，2017）遵循 Schulman、Levine 等人（2017）的信赖域思想，使每次更新都接近收集数据时所使用的策略，但将约束优化替换为标准的基于梯度的优化器可以直接求解的无约束优化。令 $\pi_{\theta_{\mathrm{old}}}$ 表示当前训练迭代开始时的策略，并令 $\mathcal{D} := \{(s_i^k,\ a_i^k,\ r_{i+1}^k,\ s_{i+1}^k) :\ i = 0,\ \ldots\ldots,\ n-1,\quad k = 1,\ \ldots\ldots,\ N_{\mathrm{batch}}\}$ 表示在该策略下收集的、来自 N_{batch} 个回合的状态转移批次，其中 k 为回合索引。随后，通过对 $\mathcal{D}$ 进行 N_{epochs} 次梯度上升迭代来更新参数。定义概率比

$$\omega_i(\theta) := \frac{\pi_\theta(a_i \mid s_i)}{\pi_{\theta_{\mathrm{old}}}(a_i \mid s_i)}. \tag{8}$$

策略通过裁剪目标函数引入

$$\mathcal{L}^{\mathrm{CLIP}}(\theta) := \hat{\mathbb{E}}_{\mathcal{D}} \left[\min\Big(\omega_i(\theta)\, \hat{A}_i,\ \mathrm{clip}\big(\omega_i(\theta), 1-\varepsilon, 1+\varepsilon\big)\, \hat{A}_i \Big) \right], \tag{9}$$

$\hat{}$ $\hat{}$ 其中，$\mathbb{E}_{\mathcal{D}}$ 表示 $\mathcal{D}$中状态转移的经验平均，A_i 是在$\pi_{\theta_{\mathrm{old}}}$下根据公式（7）计算得到的广义优势估计，$\varepsilon > 0$ 控制允许的策略更新幅度。对 $\omega_i(\theta)$进行裁剪会消除将比率推到 $[1-\varepsilon, 1+\varepsilon]$之外的动机：一旦更新使策略偏离$\pi_{\theta_{\mathrm{old}}}$过远，它就不再改善目标。由于取最小值时还要与未裁剪项比较，$\mathcal{L}^{\mathrm{CLIP}}$ 是未裁剪目标的下界：任何会使目标恶化的比率变化都不会被裁剪。

最后，通过最大化以下目标，行动者和评论家会联合训练

$$\mathcal{L}(\theta, \phi) := \mathcal{L}^{\mathrm{CLIP}}(\theta) - c_V \mathcal{L}^{V}(\phi), \tag{10}$$

其中，$\mathcal{L}^{V}(\phi) := \hat{}\mathbb{E}_{\mathcal{D}} \left[\big(V_\phi(s_i) - G_i\big)^2 \right]$ 是评论家损失，其回报目标 G_i 来自公式（1），$c_V > 0$ 是其权重。Schulman、Wolski等人（2017年）还加入了鼓励探索的熵奖励，但我们不使用该奖励。

近端策略优化已成为各领域连续控制问题中的标准选择。在期权对冲文献中，Du等人（2020年）和Neagu等人（2025年）对多种深度强化学习算法进行了基准评估，分别将近端策略优化列为表现最佳和第二佳的算法。因此，我们在Delta-Vega对冲框架中采用该算法。

4 期权对冲设置

我们逐步介绍期权对冲设定，旨在构建双工具对冲策略的强化学习框架。我们从Black-Scholes框架出发，逐步放宽其关键假设。首先，我们采用随机波动率模型来处理非恒定波动率问题，而这又要求引入第二种对冲证券。其次，我们纳入离散再平衡和交易成本，这使完美复制无法实现，并促使我们采用强化学习作为解决框架。

4.1 Delta-Vega对冲

Black-Scholes框架假设标的资产的动态由几何布朗运动所支配,

$$dS_t = \mu S_t\, dt + \sigma S_t\, dW_t, \tag{11}$$

其中， μ 和 $\sigma > 0$ 分别为常数漂移率和波动率， W_t 是在实际概率测度 $\mathbb{P}$下的标准布朗运动。在风险中性测度 $\mathbb{Q}$下，执行价为 K、到期日为 T的欧式看涨期权在时间 $t \le T$的价格可表示为以下闭式表达式

$$\begin{aligned} C_t(S_t,\, T-t,\, K) &= S_t\, \Phi(d_{1,t}) - e^{-r(T-t)} K\, \Phi(d_{2,t}), \\ d_{1,t} &= \frac{1}{\sigma\sqrt{T-t}} \left(\ln \frac{S_t}{K} + \left(r + \frac{\sigma^2}{2} \right)(T-t) \right), \\ d_{2,t} &= d_{1,t} - \sigma\sqrt{T-t}, \end{aligned} \tag{12}$$

其中， $\Phi(\cdot)$是标准正态累积分布函数， $r \ge 0$ 是无风险利率。[2] 现在考虑一位持有看涨期权空头头寸的交易者。在一个离散时间区间（$t-\delta t,\ t$]内，期权腿产生的损益（P&L）为 $C_{t-\delta t} - C_t$。为抵消这一风险敞口，交易者在该区间内持有 $\mathcal{H}_t^S$标的资产头寸，从而使对冲组合的损益为

$$\mathrm{P\&L}_t = (C_{t-\delta t} - C_t) + \mathcal{H}_t^S (S_t - S_{t-\delta t}). \tag{13}$$

在 连续时间极限 $\delta t \to 0$下，BS模型存在完美对冲，因为令 g

$$\mathcal{H}_t^S = \frac{\partial C_t}{\partial S_t} =: \Delta_t \tag{14}$$

可使损益$_t = 0$ 在一阶近似下为零。量 Δ_t 即为期权的Delta。

上述完美对冲依赖于波动率 σ 为常数（或至少是时间的确定性函数）这一假设。现实中，通过反演 BS 公式从市场期权价格中提取的隐含波动率会随执行价和到期日而变化，由此形成众所周知的波动率微笑和期限结构；随着市场状况演变，隐含波动率也会随时间变化。因此，关注波动率风险的对冲策略必须在能够刻画波动率变化的模型下进行训练。

具体而言，我们用随机波动率模型描述标的资产的风险中性动态，其中方差本身是一个随机过程。我们采用 Heston（1993）的开创性设定。定义 $X_t := \ln S_t$，令 ν_t 表示 t时刻现货价格的瞬时方差，则 SDE 系统为

$$dX_t = -\frac{1}{2}\nu_t dt + \sqrt{\nu_t} dW_t^S, \quad X_0 = 0, \tag{15}$$

$$d\nu_t = \kappa(\bar{\nu} - \nu_t)dt + \eta\sqrt{\nu_t} dW_t^\nu, \quad \nu_0 \ge 0, \tag{16}$$

其中 W_t^S， W_t^ν是风险中性概率测度$\mathbb{Q}$下两个相关的布朗运动，其相关性为 $\rho \in [-1, 1]$，且 κ， $\bar{\nu}$， $\eta > 0$。

当价格动态如公式（15）和（16）所示时，现货价格的布朗运动并不是随机性的唯一来源。为了使市场完备，即完美对冲该期权，还必须考虑波动率风险（公式（16）中的布朗增量 dW_t^ν）。为此，对冲组合必须包括标的资产和另一个基于同一标的资产的期权（下文称为对冲期权），且该期权不同于

[2]不失一般性，本文其余部分均假设 $r = 0$ 。

-目标期权[3]。令 $\tilde{C}_t$ 表示第二个期权的价格，并将 $\mathcal{H}_t^C$称为时刻 t的~衍生品头寸。根据复制投资组合论证，在每个时刻 t 实现完美对冲的 C_t和 S_t 的投资组合持仓为

$$\mathcal{H}_t^{\tilde{C}} = \frac{\mathcal{V}_t}{\tilde{\mathcal{V}}_t}, \tag{17}$$

$$\mathcal{H}_t^S = \Delta_t - \mathcal{H}_t^{\tilde{C}}\,\tilde{\Delta}_t, \tag{18}$$

其中 Δ_t 现在是目标期权的 Heston Delta，且

$$\mathcal{V}_t := \frac{\partial C_t}{\partial \nu_t} \tag{19}$$

~ ~是该期权的 Vega（Δ_t，$\mathcal{V}_t$ 是相应的对冲期权敏感度）。同时调整标的资产和对冲期权，以同时覆盖价格风险和波动率风险，这种做法称为$Delta\text{-}Vega$ 对冲[4]。在 Heston模型下，期权价格及其偏导数可通过对模型特征函数求逆，以半解析形式获得。因此，我们将这种对冲作为基准，与近端策略优化在模型内合成实验中的表现进行比较 [5]。

4.2 实务对冲

公式（17）和（18）描述了 Heston模型下的最优对冲，但需要知道瞬时方差 ν_t，而其过程无法直接观测。在实践中，交易员通常采用所谓的实务 Black-Scholes对冲策略，将市场期权价格视为由市场即期价格S_t 和隐含波动率 σ_t^{imp} （K/S_t，$T-t$）代入 BS 公式（12）计算所得的结果，其中还需使用合约的执行价 K 和到期时间$T-t$。因此，对冲所需的敏感度可以通过闭式表达式获得，

$$\Delta_t^{\text{BS}} = \left.\frac{\partial C_t}{\partial S_t}\right|_{\sigma=\sigma_t^{\text{imp}}} = \Phi(d_{1,t}), \tag{20}$$

$$\mathcal{V}_t^{\text{BS}} = \left.\frac{\partial C_t}{\partial \sigma}\right|_{\sigma=\sigma_t^{\text{imp}}} = S_t\sqrt{T-t}\,\varphi(d_{1,t}), \tag{21}$$

其中，$\varphi(\cdot) = \Phi'(\cdot)$为标准正态密度函数。尽管从业者策略并非基于内部一致的模型，但其优势在于仅依赖市场可观测量，并且是期权交易台采用的标准方法（Hull 和 White 2017年）。因此，我们将其作为仿真到现实评估练习的基准，在该练习中使用标准普尔500指数市场数据测试近端策略优化。与 Cao、Chen、Farghadani 等（2023年）的 Gamma-Vega 研究以及大多数相关强化学习工作一样，我们的基准策略不考虑交易成本。考虑成本的 Delta-Vega 对冲版本，例如 Delta对冲中围绕无摩擦对冲设置的不交易区域（Whalley 和 Wilmott 1997年），已被描述为一项重大的开放性挑战（Herrmann 和 Muhle-Karbe 2017年）。

由于智能体是在 Heston模型生成的数据上训练的，因此将 Heston 与从业者对冲联系起来是有益的。对于固定执行价和到期日，Heston 价格是

[3]参见 Romano 和 Touzi 1997年；Davis 和 Ob-l´oj 2008年关于理论结果的研究。[4]基于对瞬时方差（或即期波动率）敏感性的对冲有时被称为 Delta-Sigma 对冲，以区别于使用对隐含波动率敏感性的 Delta-Vega对冲（Herrmann 和 Muhle-Karbe 2017年）。为便于记述，我们将 Vega 一词交替使用，这在文献中也很常见（Hull 和 White 1987年）；我们将在第4.2节阐明二者之间的关系。[5]毋庸置疑，对于看涨期权，根据看涨—看跌平价关系，可以构造一个（静态）完美对冲。不过，这要求存在一个合约具体信息相同的期权。

在模型隐含波动率下计算的 BS 价格，$C_t = C^{\mathrm{BS}}(S_t,\ \sigma_t^{\mathrm{imp}})$，其中 σ_t^{imp} 是 S_t 和 ν_t 的函数。链式法则给出

$$\mathcal{V}_t = \mathcal{V}_t^{\mathrm{BS}} \frac{\partial \sigma_t^{\mathrm{imp}}}{\partial \nu_t}, \tag{22}$$

$$\Delta_t = \Delta_t^{\mathrm{BS}} + \mathcal{V}_t^{\mathrm{BS}} \frac{\partial \sigma_t^{\mathrm{imp}}}{\partial S_t}, \tag{23}$$

而（17）、（18）中的对冲比率变为

$$\mathcal{H}_t^{\tilde{C}} = \frac{\mathcal{V}_t^{\mathrm{BS}}}{\tilde{\mathcal{V}}_t^{\mathrm{BS}}} \chi_t^{\nu}, \tag{24}$$

$$\mathcal{H}_t^{S} = \Delta_t^{\mathrm{BS}} - \mathcal{H}_t^{\tilde{C}} \tilde{\Delta}_t^{\mathrm{BS}} + \chi_t^{S}, \tag{25}$$

with

$$\chi_t^{\nu} = \frac{\partial \sigma_t^{\mathrm{imp}} / \partial \nu_t}{\partial \tilde{\sigma}_t^{\mathrm{imp}} / \partial \nu_t}, \tag{26}$$

$$\chi_t^{S} = \mathcal{V}_t^{\mathrm{BS}} \left(\frac{\partial \sigma_t^{\mathrm{imp}}}{\partial S_t} - \chi_t^{\nu} \frac{\partial \tilde{\sigma}_t^{\mathrm{imp}}}{\partial S_t} \right). \tag{27}$$

从业者对冲对应于 $\chi_t^S = 0$ 和 $\chi_t^{\nu} = 1$，也就是说，对应于不随现货价格变动的隐含波动率（由此产生所谓的固定执行价波动率微笑），以及在不同到期日之间平行移动的隐含波动率。在 Heston 框架下，隐含波动率同时取决于 S_t 和 ν_t，并带有修正因子 χ_t^S和 χ_t^{ν}，这些因子取决于校准参数和合约规格。

4.3 交易成本下的对冲

BS 框架和 Heston 框架尽管对波动率的处理方式不同，但都采用连续时间交易和无摩擦市场的假设。在现实中，交易只能在离散的时间点进行，投资组合经理每次改变市场中的头寸时都会面临交易成本。因此，二者之间存在权衡。一方面，完美复制要求通过更加频繁地交易，逐步逼近连续再平衡的极限。另一方面，连续交易会使对冲成本变得无限高。目标是以一种有原则的方式处理这一权衡——而强化学习为此提供了一个自然的框架。

我们考虑一位智能体，其持有一份执行价为 $K = S_0$ 、到期日为 T 的平值（ATM）欧式看涨期权空头头寸。该智能体于 $t_0 = 0$ 建立头寸，并在等间隔的交易日期 $0 = t_0 < t_1 < \ldots < t_n = T$[6] 动态调整头寸。在本文余下部分，我们用 $S_i := S_{t_i}$ 表示日期 t_i，$i = 0$、.....、n 的价格，用 $\tau_i := T - t_i$ 表示到期时间。为对冲价格风险和波动率风险，该智能体交易两种工具：标的资产，以及一份以与目标看涨期权相同标的资产为标的的对冲期权。依照惯例，我们假设对冲期权为平值看涨期权，且~到期时间为 $T > T$；较长期期权对波动率风险更敏感（BS Vega（22）会随到期时间增加），因此更常用于 Vega 对冲（Cao、Chen、Farghadani 等（2023））。

令 $\mathcal{H}_i^S$和 $\mathcal{H}_i^{\tilde{C}}$表示区间 $(t_{i-1},\ t_i]$ 中标的资产和对冲期权所持有的头寸，分别为 $\mathcal{H}_1^S = h \in \mathbb{R}$ 和 $\mathcal{H}_1^C = \tilde{h} \in \mathbb{R}$。单期投资组合

[6] 在实证应用中，对冲期限不必与期权的到期日一致，即 $t_n < T$. 我们将在第5.3节讨论这一问题。

在（t_{i-1}， t_i] 上的损益为

$$\begin{aligned}\mathrm{P\&L}_i \;=\; &\underbrace{C_{i-1}-C_i}_{\text{option P\&L}}+\underbrace{\mathcal{H}_i^S(S_i-S_{i-1})}_{\text{asset P\&L}}+\underbrace{\mathcal{H}_i^{\tilde{C}}(\tilde{C}_i-\tilde{C}_{i-1})}_{\text{hedging option P\&L}}\\ &-\underbrace{c\,\big|S_{i-1}\big(\mathcal{H}_i^S-\mathcal{H}_{i-1}^S\big)\big|}_{\text{asset costs}}-\underbrace{\tilde{c}\,\big|\tilde{C}_{i-1}\big(\mathcal{H}_i^{\tilde{C}}-\mathcal{H}_{i-1}^{\tilde{C}}\big)\big|}_{\text{hedging option costs}},\end{aligned} \tag{28}$$

其中，参数 c 和 $\tilde{c}$ 分别表示相对于标的资产和期权价值的比例交易成本。交易在再平衡日期 t_{i-1} 执行，此时建立头寸 $\mathcal{H}_i$ ，因此成本与该日期的现行价格成比例。我们的方法类似于 Mikkil¨a 和 Kanniainen（2023年）的方法，他们同样根据交易资产价格对比例成本进行建模；而 Kolm and Ritter（2019年）则考虑每单位交易的固定最小变动价位成本（当 $c\,S_i \approx$ 为 Tick Size 时，两者具有可比性）。等价地， c 可以解释为相对于中间价的买卖价差的一半，因此 $c=1$ 基点表示跨越相当于该工具中间价一个基点的半个价差所产生的成本。与相关文献中的通常做法一致，我们舍弃佣金和费用等交易成本的任何显式组成部分，仅对价差进行建模，因为在高流动性指数产品中，价差通常占主导地位。

根据上述定义的单步损益，可以将终端财富写为 as

$$w_n = w_0 + \sum_{i=1}^{n} \mathrm{P\&L}_i, \tag{29}$$

其中 $w_0 = 0$ 是在自融资策略的假设下。遵循 Kolm and Ritter (2019) 的做法，我们通过终端财富上的二次效用来刻画智能体的偏好。因此，对冲问题可表述为均值—方差优化问题

$$\max_{\{\mathcal{H}_i^S,\,\mathcal{H}_i^{\tilde{C}}\}_{i=1}^{n}} \Big(\mathbb{E}[w_n] - \lambda\,\mathbb{V}[w_n]\Big), \tag{30}$$

其中 $\lambda > 0$ 是决定成本—方差权衡的风险厌恶系数。期望$\mathbb{E}[\cdot]$ 和方差 $\mathbb{V}[\cdot]$ 均根据标的资产和期权价格的分布（来自市场模型或实证数据），以及策略所选择的头寸来计算。

4.4 将 Delta-Vega对冲嵌入强化学习框架

下文将定义状态空间、动作空间和奖励函数，从而将公式 (30) 中的问题表述为强化学习问题，并通过 PPO算法求解。

状态空间

对于每个时间 t_i， $i = 0$，.....， n，状态向量定义为

$$s_i := \big(S_i, C_i, \tilde{C}_i, \mathcal{H}_i^S, \mathcal{H}_i^{\tilde{C}}, \tau_i, \tilde{\tau}_i\big) \in \mathcal{S}, \tag{31}$$

$$\mathcal{S} := \mathbb{R}_+^3 \times [h_l, h_u]^2 \times [0, T] \times [0, \tilde{T}], \tag{32}$$

其中， S_i 表示标的现货价格， C_i 表示待对冲目标看涨期权的价格，˜和 $\tilde{C}_i$ 表示对冲看涨期权的价格； $\mathcal{H}_i^S$和 $\mathcal{H}_i^C$分别表示在（t_{i-1}， t_i]期间持有的标的资产头寸和对冲期权头寸，其边界为 h_l， $h_u \in \mathbb{R}$， $h_l < h_u$；˜$\tau_i = T - t_i$和 $\tilde{\tau}_i = T - t_i$分别表示主期权和对冲˜ ˜期权的到期时间，其中 T， $T \in \mathbb{R}_+$、 $T < T$分别为以年份表示的到期日。

方程（32）中的状态空间定义完全不依赖于任何特定模型，仅包含可从市场直接观测到的变量。具体而言，我们

强调，状态空间中不包含任何潜在的或特定于模型的波动率变量。这种选择不同于使用Heston模型作为市场模拟器的相关研究。例如，Giurca和Borovkova（2021年）在合成Heston环境上进行训练时，将对数收益过程的瞬时方差纳入可观测量集合；而在将策略部署到标准普尔500指数市场数据上时，则依赖VIX指数作为波动率代理指标。为避免对波动率代理指标作出任何假设，我们不使用瞬时方差扩展状态空间 ν_t。此外，两个期权价格及其各自的到期时间，隐含地携带了有关平值隐含波动率及其期限结构的信息。Fran˛cois等人（2025年）的研究发现，此类信息具有相关性；他们将用于训练的隐含波动率模型的拟合因子纳入状态空间。在本文中，智能体以模型无关的方式从市场价格中推断波动率信息。

动作空间

在每个日期 t_i、 $i = 0 \ldots\ldots n-1$，智能体选择投资组合持仓

$$a_i := \left(\mathcal{H}_{i+1}^{S}, \mathcal{H}_{i+1}^{\tilde{C}}\right) \in \mathcal{A}, \tag{33}$$

$$\mathcal{A} := [h_l, h_u]^2 \subseteq \mathbb{R}^2. \tag{34}$$

这些持仓是指在区间（t_i、 t_{i+1}]内持有的标的资产和对冲期权头寸。我们始终假设，交易者账簿由单位名义规模的单个˛目标期权构成，因此， $\mathcal{H}_i^S$和 $\mathcal{H}_i^C$可解释为对冲比率，即每单位目标期权所持有的标的资产和对冲期权单位数。界限 h_l和 h_u 将比率限制在经济上合理的范围内，同时也提高了样本效率和收敛性（Cao、Chen、Farghadani 等（2023））。由于智能体持有欧式看涨期权的空头头寸，该头寸的Delta为 $-\Delta_i \in [-1, 0]$，因此标的资产的Delta对冲位于 $[0, 1]$内。类似地，由于对冲期权具有相同的执行价，而 $T > T$，实务对冲比率 $\mathcal{V}_i^{\mathrm{BS}}/\tilde{\mathcal{V}}_i^{\mathrm{BS}}$ 也位于 $[0, 1]$ 内，因为对于近平值期权， $\tilde{\mathcal{V}}_i^{\mathrm{BS}} > \mathcal{V}_i^{\mathrm{BS}} \geq 0$。因此，在数值实验中，我们设定 $h_l = 0$ 和 $h_u = 1$。实证上，我们发现，在合成测试的校准参数和到期日下，界限并不会约束Heston模型对冲；而在仿真到现实研究中更长的到期日和多组校准参数下，对冲期权的模型对冲比率可能超过1（见第5.3.3节）。

奖励函数

为了将优化问题（30）化简为逐步强化学习目标，我们遵循 Kolm and Ritter（2019），将即时奖励定义为一个周期内方程（28）中损益$_{i+1}$ 的线性二次函数，即

$$\mathcal{R}(s_i, a_i, s_{i+1}) := \mathrm{P\&L}_{i+1} - \lambda\, \mathrm{P\&L}_{i+1}^2. \tag{35}$$

Ritter（2017）表明，最大化此类奖励的期望总和可以近似求解均值-方差问题（30），Kolm and Ritter（2019）将这一结果应用于期权对冲。当逐步损益增量在各期之间不相关时，该近似成立，此时 $\mathbb{V}[w_n]$ 是逐步方差之和；并且当预期逐步损益较小时，该近似也成立，此时 $\mathbb{E}[$（损益$_{i+1}$）$^2]$ 接近于 $\mathbb{V}[$损益$_{i+1}]$。因此，系数为 λ 的二次惩罚可通过奖励信号直接控制成本—方差权衡。

5 结果

我们开展了三项主要数值实验，以评估我们的 PPO Delta-Vega 类方法。在第 5.1 节中，我们将 PPO 与现有最稳健的基准进行验证，即

基于 Heston模型半解析偏导数、且不考虑交易成本的 Delta-Vega对冲策略。我们还将考虑仅Delta对冲策略（其中仅交易标的资产这一对冲工具）作为额外基准，因为这是研究人员和从业者最熟悉的对冲方案；更重要的是，它揭示了持有第二种对冲工具的成本与收益。

在第5.2节中，我们引入交易成本，并研究PPO智能体相对于同样两种基准策略的均值-方差优化表现。我们针对三种不同的交易成本情景，结合不同的风险厌恶程度设计了一组实验网格，从而拆分考察各类对冲证券所发挥的作用。

最后，在第5.3节中，我们研究将在合成数据上训练的策略迁移至标准普尔500指数市场数据时的仿真到现实迁移效果。在这一设置下，我们采用业界实践中的Delta和Delta-Vega对冲策略作为基准，这些策略基于数据供应商提供的Black-Scholes希腊字母指标。

5.1 无摩擦环境

第一项测试旨在受控环境中验证近端策略优化。具体而言，我们考虑一个在无成本假设（$c = \tilde{c} = 0$ 方程（28））下模拟的Heston市场；对于该市场，我们知道，除时间离散化影响外，Delta-Vega对冲策略确实是最优策略。基准投资组合头寸由方程（17）和（18）给出，其中偏导数通过算法进行计算，即对根据Heston模型的特征函数计算得到的期权价格应用自动微分。

在以下示例中，我们考虑一个在 Heston动力学下生成的单一标的资产（例如股票）。我们使用单一参数化来模拟所有训练和测试回合，$\kappa = 1.03, \bar{\nu} = 0.08, \eta = 1.0, \nu_0 = 0.04$, $\rho = -0.65$。[7] 时间是离散的，智能体每天观测 5 个价格，因此各日期等距分布，即 $t_i = i\,\delta t$, $i = 0, \ldots, n$，其中$\delta t = 1/(5 \cdot 252)$年，每个回合包含 $n = 300$ 个步骤。初始时刻的标的资产价格为$S_0 = 100$。待对冲期权为一份平值（ATM）看涨期权，即 $K = 100$，到期日为$T = 60$天。第二份对冲期权为一份到期日为 $T = 120$ 天的平值看涨期权。每个回合在主期权到期时终止。回合长度大致为一个财务季度，这与文献中的其他应用（Cao、Chen、Farghadani 等（2023）；Vittori 等（2020））一致，也是我们将在实证研究中采用的长度。在每个回合开始时，将智能体的持仓设定为等于 Delta-Vega基准的持仓。我们注意到，在不存在交易成本时，这不会产生任何影响，因为智能体会在单个步骤内学会直接转向最优头寸；但在存在交易摩擦时，这一点十分重要，因为连续步骤之间的头寸跳变会导致较高的交易成本。

我们在单个中央处理器上使用包含 $N_{\text{train}} = 40{,}000$个回合的数据集进行训练，并在特意设置的大型测试集（ $N_{\text{test}} = 10{,}000$个样本）上进行测试，以获得可靠的统计结果。在这一不涉及交易成本的特定设置中，我们将折扣因子设为 $\gamma = 0$。由于在方程（28）中使用 $c = \tilde{c} = 0$ ，日期 t_i 所选择的头寸会进入损益$_{i+1}$ ，而不会出现在奖励序列的其他项中，因此每个动作只会影响紧随其后的步骤。由此，对冲问题可分解为 n 个相互独立的单步问题，即针对下一价格增量进行静态均值-方差优化。在其他所有涉及交易成本的设置中，我们将 $\gamma = 0.99$设为相应值。尽管对冲问题具有有限时域，理论上在 $\gamma = 1$的情况下不会受到折扣因子的影响，但我们遵循 PPO（Schulman、Wolski 等，2017年）的原始实现，将 γ 设为略低于一，以确保价值函数估计的数值稳定性。其他 PPO 超参数和训练性能详见附录A。

图1展示了一个具有代表性的样本外对冲过程。左上图展示了进入智能体可观测空间的市场变量路径，即

[7]为避免选择任意的参数化方式，我们选取了根据历史隐含波动率曲面（2025年4月30日的欧洲斯托克 50指数期权）校准的参数，但需要强调的是，在合成实验中，我们并不关注由Heston模拟所代表的特定市场状态。这将是第5.3节实证应用的重点。

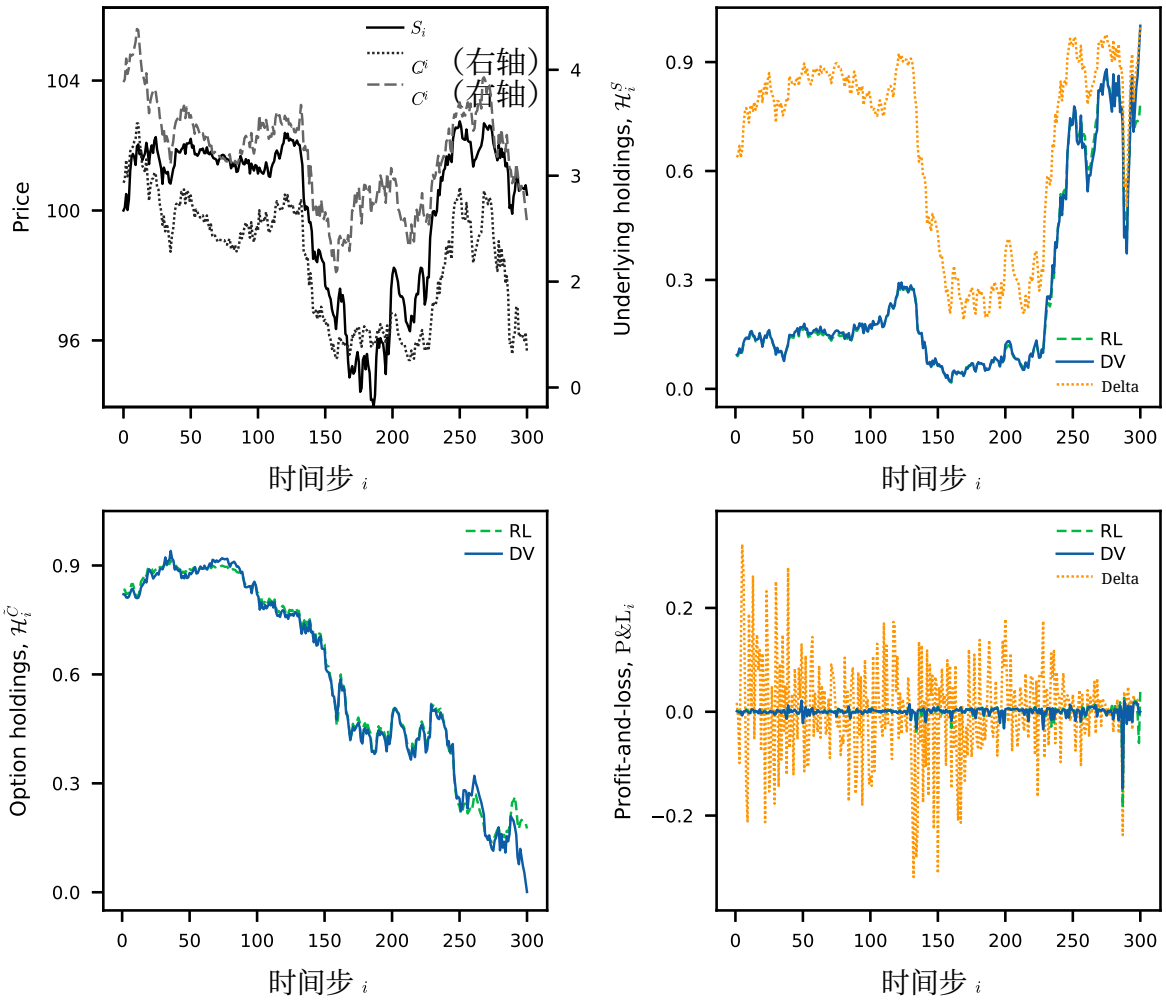

图1：无摩擦测试集中的一个样本外回合，$c = \tilde{c} = 0$。左上面板：˜标的资产价格 S_i（左轴）、主要期权价格 C_i和第二期权价格 C_i（右轴）。右上面板：所有对冲策略在标的资产上的对冲头寸 $\mathcal{H}_i^S$：_近端策略优化（强化学习）、Delta-Vega 和仅Delta。左下面板：近端策略优化和 Delta-Vega 策略在第二个期权上的对冲头寸 $\mathcal{H}_i^C$（仅Delta的风险敞口为零）。右下面板：所有策略的单步损益$_i$。

标的资产价格、主期权价格（需对冲）和第二期权价格（用于对冲）。右上图对比了智能体在每个时间步选择的标的资产对冲头寸 $a_i = \mathcal{H}_{i+1}^S$，以及 Delta-Vega对冲策略和仅Delta对冲策略规定的头寸。这里，Delta-Vega应视为Heston随机波动率环境下的最优策略，且已扣除离散时间近似误差。显而易见，近端策略优化的策略与最优策略高度一致，而Delta对冲对标的资产的配置过多，这与公式（18）相符。左下图展示了策略的第二个维度，即近端策略优化和最优策略在对冲期权上的持仓（根据定义，仅Delta对冲策略的权重为零）。同样，近端策略优化紧密跟随最优策略。我们注意到，在部分其他测试回合中，回合临近结束时也会出现这种偏差。临近到期日时，平值期权的Delta对标的资产价格的小幅变动会变得极为敏感（即该期权的Gamma较高），从而导致最优持仓急剧变化。由于每个回合中这些状态只会出现一次，因此它们可能是策略最难准确学习的区域。此外，

即使采用折扣因子 $\gamma = 0$，强化学习仍会从一批轨迹中优化平均奖励信号，而半解析基准则是在逐点意义下达到最优，也就是说，在模型生成的每条轨迹的每个步骤上均为最优。最后，右下方面板显示了三种所考察对冲策略在回合每个步骤上的损益（见方程（28）），并将其表示为时间 t_0 标的资产价格的百分比。尽管所有策略随时间推移产生的损益均值为零，但正如随机波动率环境中所预期的那样，Delta对冲明显呈现出更嘈杂的模式。

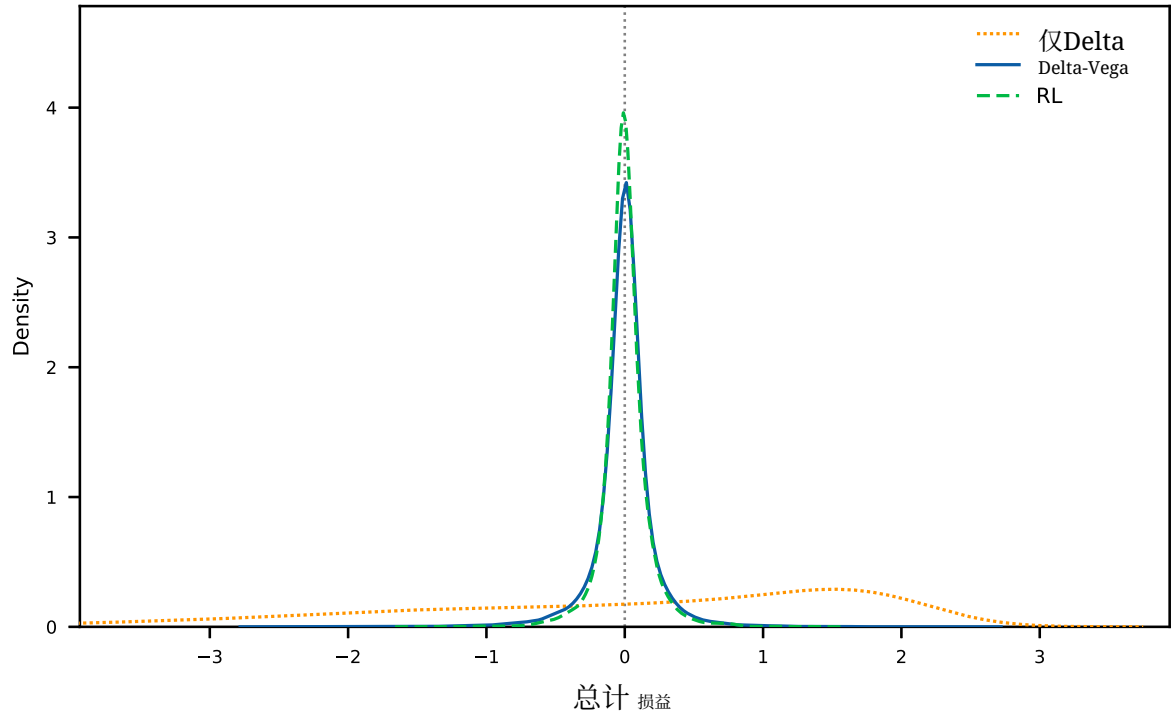

图2：无摩擦 Heston 环境中近端策略优化（强化学习）、仅Delta和Delta-Vega对冲策略的总计损益分布。曲线是基于 10,000 个回合测试集计算的核密度估计。

图2通过展示每种策略在 $N_{\text{test}} = 10{,}000$ 个测试回合中的累计损益 $\sum_{i=1}^{n}$ 损益$_i$ 分布，验证了我们从上述单回合分析中得到的见解。近端策略优化的分布与半解析Delta-Vega对冲的分布高度一致，证明智能体有效复现了最优策略。相比之下，Delta对冲表现出显著更高的离散度和明显的负偏度，这与卖出看涨期权头寸未对冲的高阶Gamma敞口相一致。

5.2 引入成本

本节将在模拟市场环境中引入交易成本。作为基线校准，我们将标的资产的成本率设为 $c = 1$ 基点，并将期权交易成本设为方程（28）中的 $\tilde{c} = 1\%$。这反映了与高流动性指数市场相一致的低摩擦环境。1%的数值与实证文献（Kaeck 等，2022年）记载的 SPX期权买卖价差的一半处于同一数量级，也接近我们数据集中（对冲）期权的报价半价差（第5.3.1节）。正如 Mikkil¨a 和 Kanniainen（2023年）所指出的，成本越低，强化学习智能体就越难以利用节约成本策略（因此，也越难以跑赢不考虑成本的策略）。风险厌恶系数设为 $\lambda = 1$。

图　图 3 展示了与图　1

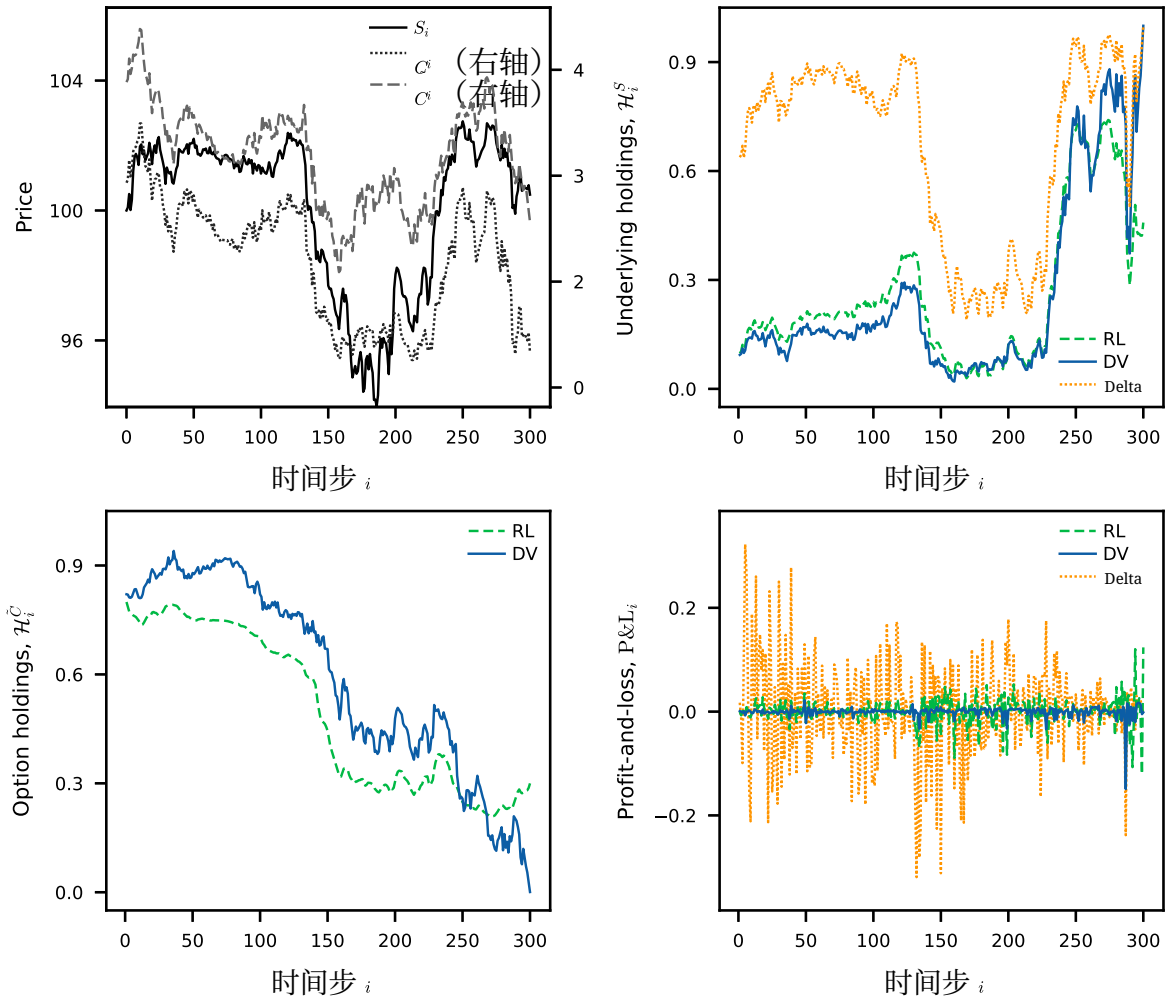

图3：一个包含交易成本（$c = 1$基点，$\tilde{c} = 1\%$）的样本外回合。左上图：标的资产价格 S_i（左轴）、主要期权价格 C_i（右轴）和第二期权价格 $\tilde{C}_i$（右轴）。右上图：所有对冲策略（近端策略优化（强化学习）、Delta-Vega 和仅Delta）在标的资产上的对冲头寸 $\mathcal{H}_i^S$。左下图：强化学习和 Delta-Vega 策略在第二个期权上的对冲头寸 $\mathcal{H}_i^C$（仅Delta的风险敞口为零）。右下图：所有策略的单步损益 P&L$_i$。

中无摩擦情景所使用的完全相同的市场模拟，但此处存在交易成本。正如右上图和左下图所示，这两幅图分别报告了标的资产和期权的对冲头寸；与零成本情景相比，PPO 对冲不再如此紧密地复现 Delta-Vega 基准。这是意料之中的，因为一旦在环境中引入交易成本，Delta-Vega 就不再是最优策略。与 Delta 对冲相关文献中的既有研究结果一致（例如 Kolm和Ritter（2019）），PPO 似乎以更为保守的方式跟随基准头寸，试图预判价格变动并进行更平滑的交易，从而减少交易成本。在这一具体示例中，我们发现，这种影响在期权腿（左下图）中比在标的资产腿（右上图）中更为明显；这与期权交易相较于资产交易的成本相对更高相一致（扣除两种证券不同的价格尺度后）。我们还注意到，在状态空间的某些区域，智能体甚至决定完全不进行再平衡；这一点从左下图可以看出，在回合的前半段，PPO 的期权持仓有若干步基本保持不变。通过查看右下图所示的逐步损益$_i$ 路径，我们可以看出，该智能体偏离半解析

对冲策略可能会以损益分布更加波动为代价，尤其是在临近到期日时。

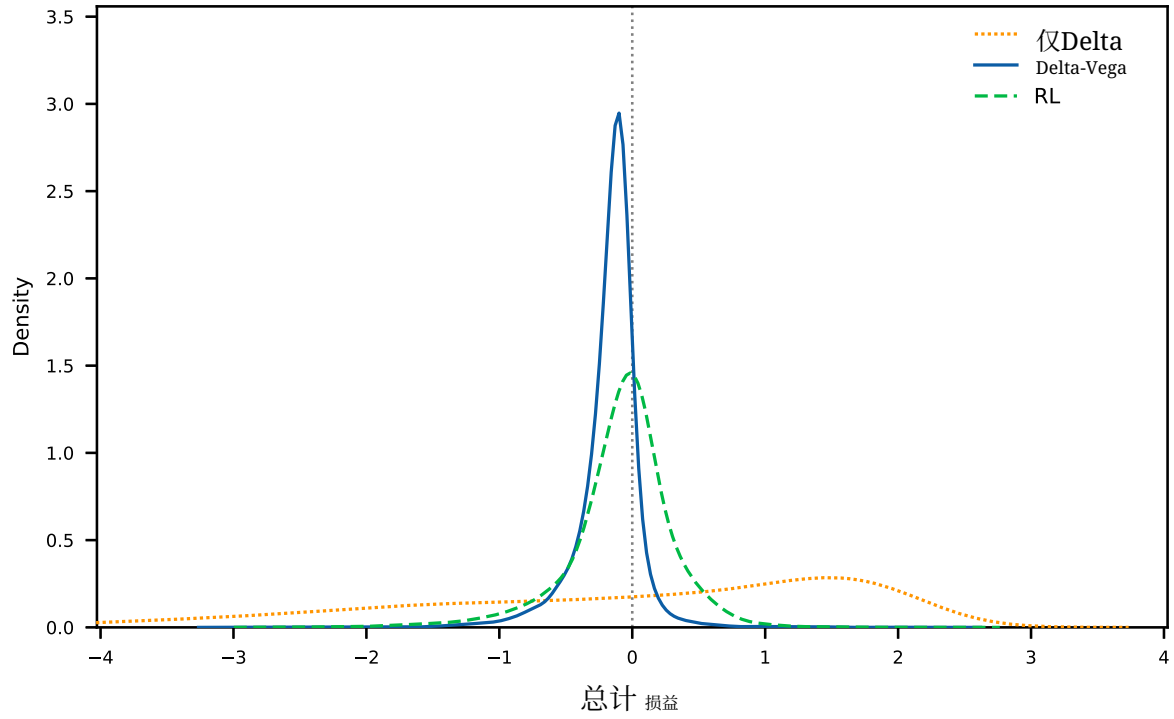

图 4：在交易成本为 $c = 1$ 基点、 $\tilde{c} = 1\%$（风险厌恶$\lambda = 1.0$）的 Heston 环境中，近端策略优化（强化学习）、仅Delta 和 Delta-Vega对冲策略的总损益分布。曲线为基于 10,000 个回合测试集的核密度估计。

图 4 明确展示了对冲成本与对冲误差之间的权衡，其中显示了每种对冲策略在完整测试集上的累计损益分布。Delta-Vega 策略在所有策略中实现了最低的总损益波动率（0.249%），但其特征变为累计损益为负值，即 -0.170%；这可以由平均累计交易成本等于 0.168% 来解释。近端策略优化以波动率换取成本节省：平均损益为 -0.085%（平均总成本为 0.079%，约为基准的一半），而标准差现为 0.396%。

下面，我们将对均值-方差目标进行更详细的考察，并在一系列交易成本和风险厌恶情景下扩展分析。

5.2.1 对交易成本的稳健性

为评估 PPO智能体在不同市场状况下的稳健性，我们在 $c = 1$ 基点和 $\tilde{c} = 1\%$的基准校准之外，考虑另外两种交易成本情景。在第一种替代情景中，我们仅将期权成本 $\tilde{c}$ 从 1% 提高到 5%，同时将资产成本维持在 $c = 1$ 基点。这代表了标的资产具有高流动性、但期权买卖价差相对较宽的市场，例如单只股票期权市场（Muravyev 和 Pearson（2020））。在第二种替代情景中，我们进一步将资产成本提高至 $c = 5$ 基点。这一水平与流动性较低的指数期货（如欧洲斯托克 50）的重新调整后的最小变动价位一致（Vittori 等（2020））。我们有意排除高资产成本、低期权成本情景，因为它不符合期权市场的典型成本结构：期权的买卖价差通常比标的资产更宽。

表 1 比较了强化学习智能体在三种摩擦情景下的对冲绩效与两种基准策略的表现。报告的数据为四次独立训练

表 1：在 Heston模型下生成的 10,000 个回合测试集上，不同交易成本情景下的对冲绩效。报告值汇总了四个独立训练轮次的结果，风险厌恶系数为 $\lambda = 1$。所有数值均以 S_0 的百分比表示。粗体标记了每种成本情景下均值-方差目标最高的结果。

策略	总损益		资产市场		期权市场		目标
	Mean	Std	Cost	换手率	Cost	换手率	
低交易成本 ($c = 1$ 基点，$\tilde{c} = 1\%$)							
Delta	-0.047	1.719	0.052	6.51	0.000	0.00	-3.00
Delta-Vega	-0.170	0.249	0.036	4.27	0.132	4.22	$\mathbf{-0.23}$
RL	-0.080	0.429	0.025	3.07	0.055	2.33	-0.26
温和的交易成本 ($c = 1$ 基点，$\tilde{c} = 5\%$)							
Delta	-0.047	1.719	0.052	6.51	0.000	0.00	-3.00
Delta-Vega	-0.697	0.419	0.036	4.27	0.659	4.22	-0.87
RL	-0.277	0.482	0.024	2.91	0.251	2.35	$\mathbf{-0.51}$
高交易成本 ($c = 5$ 基点，$\tilde{c} = 5\%$)							
Delta	-0.256	1.776	0.261	6.51	0.000	0.00	-3.41
Delta-Vega	-0.841	0.482	0.180	4.27	0.659	4.22	-1.07
RL	-0.350	0.460	0.096	2.41	0.252	2.35	$\mathbf{-0.56}$

运行结果的平均值，每次均在同一个 $N_{\text{test}} = 10{,}000$ 回合测试集上进行评估。除总损益的均值及其标准差外，我们还展示了测试集上每个回合总成本和总绝对头寸变动的平均值，后者定义为换手率 $= \sum_{i=1}^{n} \left| \mathcal{H}_i - \mathcal{H}_{i-1} \right|$。这一指标衡量一个回合内累计交易活动的程度，不受价格波动影响，因此可作为总成本的补充指标，用于概括智能体的再平衡行为。特别是，它使我们能够量化强化学习智能体相对于基准策略平滑其交易轨迹的程度（这一行为已在图3的可视化分析中得到识别）。标的资产和对冲期权的成本及换手率指标分别报告，以区分各工具的贡献。表中的第一部分对应于第5.2节讨论的低成本情形，此处取四次独立训练运行的平均值。均值-方差权衡在最后一列中进行概括，该列报告了在测试数据上根据方程（30）计算的目标函数。在这一情景下，PPO智能体取得了与Delta-Vega策略相当但略低的目标值，同时明显优于仅德尔塔基准策略。这一结果符合预期，因为当交易成本较低时，Delta-Vega 策略理论上应保持接近最优。事实上，与基准策略相比，PPO在两个市场中的成本都更低（按照定义，Delta的期权成本为零），但在当前摩擦水平下，节省的绝对金额可以忽略不计。总成本和换手率均表明，相较于标的资产市场，智能体在期权市场采用了更为保守的交易策略，这与此前对上方单个样本外回合的目视检查结果一致（见图3左下图）。在第二个成本部分中，当期权成本提高 $\tilde{c} = 5$ % 时，PPO consistently 优于两种基准策略。值得注意的是，尽管智能体在成本更高的期权腿上确实实现了低于基准策略的交易成本，但其相对于Delta-Vega策略的优势并非仅仅来自成本方面。智能体将损益波动率稳定在从低成本情景的0.429%到中等成本情景的0.482%的水平，而Delta-Vega的损益波动率则明显上升，从0.249%变为0.419%。由于成本直接计入损益（方程（28）），当成本率上升时，不调整其交易行为的策略的损益实现波动率也会随之上升。另一方面，PPO在降低成本的同时，将实现波动率维持在与Delta-Vega相当的水平，从而获得更高的目标值。当

此外，随着标的资产成本率上升（表格中的底部区块），同样的效应使近端策略优化在均值和波动率方面均优于Delta-Vega策略，从而在三个摩擦情景中实现最高的目标值优势。

有趣的是，在三种不同的成本设置下，近端策略优化在对冲期权上的换手率基本保持不变，这表明智能体策略的这一维度还受到其他因素的支配（我们将在第5.2.2节对此进行讨论）。此外，随着资产成本率上升，我们此前发现的期权腿相较于资产交易更平滑的倾向不再明显：在第三种情景中，资产（2.41 对 4.27）与期权腿（2.35 对 4.22）上强化学习与Delta-Vega之间的换手率差异相当。

5.2.2 对风险厌恶的敏感性

作为最后一项验证，我们研究智能体的对冲绩效对风险厌恶参数 λ 的敏感性，该参数直接控制公式（30）中的成本—方差权衡。图5报告了三个摩擦情景下不同风险厌恶系数取值时表 1 的对冲统计数据。每个点均按照相应情景的Delta-Vega指标进行了归一化。在顶行中，我们绘制了样本外回合的平均总损益（左）及其标准差（右）。波动率与风险厌恶系数呈现明显的单调关系，随着风险厌恶程度提高而下降。这是预期中的行为，因为根据公式（35），随着风险厌恶程度提高，智能体会受到奖励函数中二次项越来越强的惩罚。在低交易成本下，智能体的波动率在高于Delta-Vega基准水平的位置趋于平稳。而在摩擦程度最高的情景中，它在 $\lambda = 3$ 处与基准持平，之后略低于基准。正如上述分析所预期的，这一现象是由基准策略波动率上升所驱动的，而不仅仅是近端策略优化收紧所致。在左上面板中，平均损益随风险厌恶系数的变化规律性较弱。我们观察到的不是单调下降，而是存在一个转折点，其位置随成本情景而变化。

一旦转向交易成本和换手率，图像便更加清晰；资产腿和期权腿的相关结果分别显示在中排和底排中。资产腿（中排）与风险厌恶程度呈单调关系：在三个情景下，随着 λ增加，强化学习的Delta对冲成本和换手率均稳步上升，这与波动率面板所显示的风险厌恶程度越高、跟踪越紧密的证据相一致。期权腿（底排）的表现则有所不同。成本和换手率先下降后再次上升；在高摩擦情景下，最低点位于 $\lambda = 0.3$ ，而在低摩擦和中摩擦情景下，最低点位于 $\lambda = 1$。这并非学习Vega对冲特有的困难，而是与方程（28）中内含的成本不对称性相一致。交易对冲期权在结构上比交易标的资产更加昂贵。其比例成本$\tilde{c} \in \{1\%, 5\%\}$比标的资产的 $c \in \{1$基点，5基点$\}$高出两个数量级，而其价格仅低约20倍（$C_0 \approx 4.7$ 相对于 $S_0 = 100$）。按每单位交易量计算，这使低摩擦情景下的净成本差约为五倍。智能体的行为符合这一不对称性。当 λ 从最低值开始增加时，Delta对冲换手率急剧上升，而Vega对冲换手率下降，这表明额外的风险降低主要是通过收紧成本更低的资产腿来实现的。当进一步收紧Delta对冲所带来的风险降低逐渐递减后，智能体开始为额外的Vega对冲付出成本，于是期权腿的成本和换手率在更高的 λ下上升。同一机制也解释了高摩擦情景中更早出现的转折点。当交易标的资产的成本更高时（ c更高），进一步收紧Delta对冲会更早变得昂贵，因此智能体会在较低的 λ下转向Vega腿。

总体而言，在整个 λ 范围内及全部三种情景下，无论是在资产市场（最高为基准水平的79%）还是期权市场（最高为55%）中，近端策略优化的交易成本始终低于Delta-Vega。为便于可视化，本文未在该图中报告仅Delta的数值；不过需要指出的是，在所考虑的 λ 范围内，近端策略优化的资产相关成本和损益波动率始终低于仅Delta。

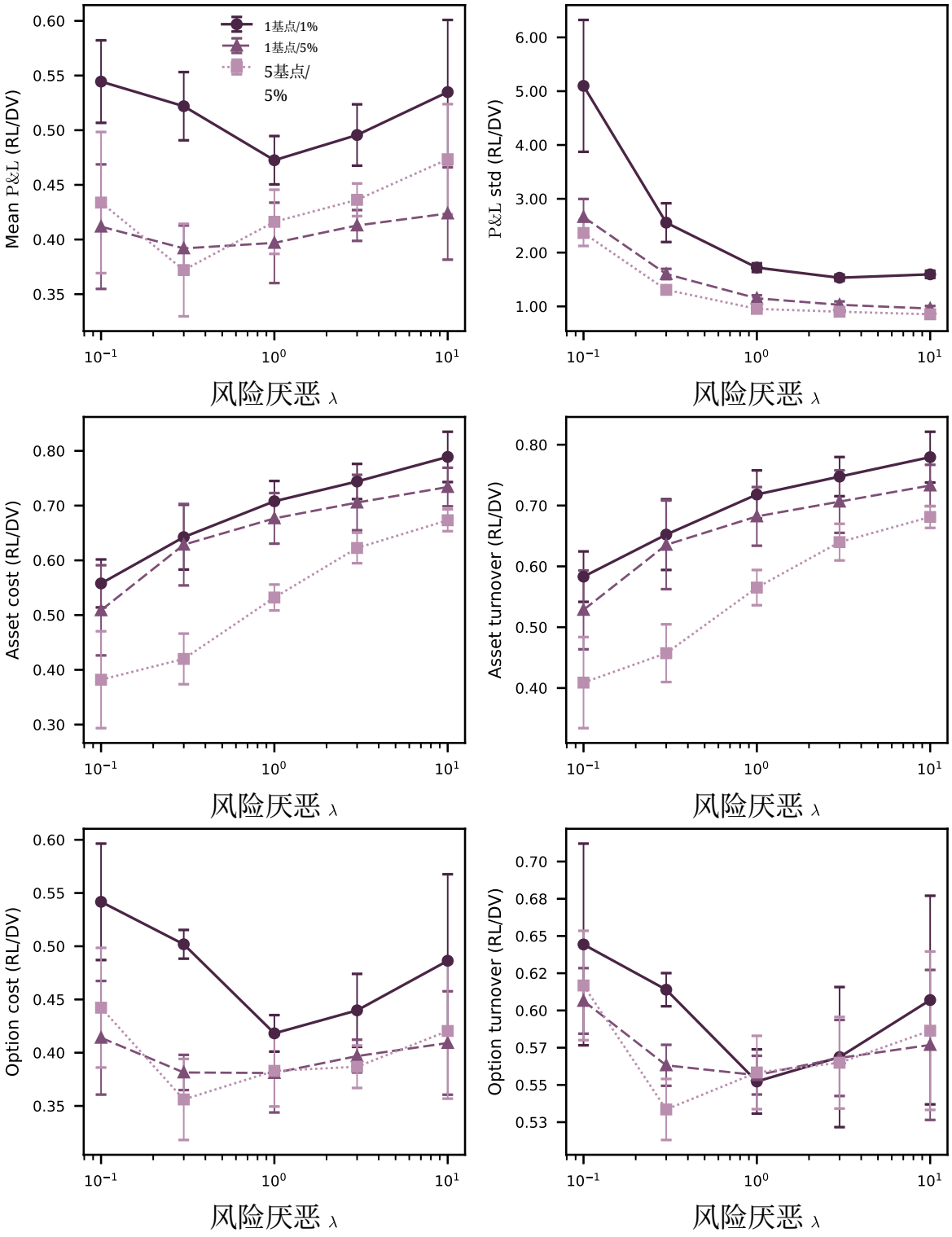

图5：风险厌恶 λ 对近端策略优化（强化学习）样本外对冲绩效的影响，相对于 Heston 的 Delta-Vega 策略。第一行：平均损益（左）和损益波动率（右）。第二行：资产腿上的交易成本（左）和换手率（右）。第三行：对冲期权腿上的交易成本（左）和换手率（右）。所有指标均在三个成本情景的对数间隔风险厌恶水平 λ 下展示（每种情景独立运行四次）。每条曲线均按其所属层级的 Delta-Vega 值进行归一化，因此 $y = 1$ 在每个面板中均对应 Delta-Vega，但每个层级的绝对水平不同（见表 1）。带状区域表示各次运行均值的一个标准差。

表2：用于评估近端策略优化对冲策略的标普500指数测试样本（2019–2022年）。 $\tau_0, \tilde{\tau}_0$ 分别表示回合开始时主期权和对冲期权的到期时间（日历日），$\log|(K/S_0)|$ 表示绝对对数价内外程度。半价差是对冲期权的相对报价半价差（$\tilde{C}^{\mathrm{Ask}} - \tilde{C}^{\mathrm{Bid}}$）/（$\tilde{C}^{\mathrm{Ask}} + \tilde{C}^{\mathrm{Bid}}$），取所有步骤的均值。报告数值为各回合的均值，括号内为一个标准差。

Year	时间窗口	回合	τ_0 （天）	$\tilde{\tau}_0$ （天）	$\lvert\log(K/S_0)\rvert$	半价差（%）
2019	4	188	263.3 (126.6)	437.7 (173.9)	0.015 (0.009)	0.56 (0.19)
2020	4	211	235.7 (124.0)	430.9 (155.7)	0.015 (0.009)	1.94 (1.80)
2021	4	445	272.7 (117.7)	450.5 (147.3)	0.015 (0.009)	0.96 (1.00)
2022	3	349	284.0 (109.9)	461.5 (143.3)	0.015 (0.009)	0.90 (0.59)
总计	15	1,193	268.0 (119.2)	448.2 (152.6)	0.015 (0.009)	1.05 (1.11)

5.3 仿真到现实迁移

在下文中，我们使用由Heston随机波动率模型生成的数据训练PPO智能体，并研究将该策略迁移至真实SPX期权市场数据的效果。第5.3.1节介绍实证数据集。第5.3.2节概述如何构建专用于实证迁移任务的合成训练集。最后，第5.3.3节报告样本外结果。

5.3.1 数据

我们使用覆盖2018年1月至2022年7月交易日的标准普尔500指数（SPX）期权每日报价（DeltaNeutral LLC 2026）。对于每份合约，我们观察标的资产（指数）水平、执行价、到期日、期权类型、买价和卖价、交易量、未平仓合约数，以及数据提供方计算的相应希腊字母指标（Delta和Vega）。本文报告的所有损益数据均根据报价中间价计算。作为初步处理，我们剔除买价或卖价非正，或卖价低于买价的报价，因为我们认为这表明记录有误。我们完全留出2018年的数据，用于校准生成合成训练数据所用的Heston模型（详见下文第5.3.2节），并将2019—2022年的数据用作训练后近端策略优化智能体的样本外测试集。

我们从2019年1月开始，连续选取长度为60个交易日的不重叠窗口，以构建对冲期次。因此，时间跨度与此前的合成数据研究保持一致；唯一的变化是频率降低了，从每天五个时间步变为每天一个时间步。在一个窗口内，合约必须在每个交易日都有有效报价才具备入选资格（最多两天的孤立缺口使用最近一次可用报价进行填补，而缺口更长的合约则会从该窗口中完全剔除——最终样本中，日观测总数的0.2%经过了填补）。对于每个符合条件的主合约，我们选择第二份合约用于Vega 对冲。第二份合约的到期日必须至少比主合约晚30个日历日，从而使两条腿在到期日维度上对波动率具有有意义的不同敏感度；此外，在回合开始时，其价内外程度必须与主合约相近（差异在两个百分点以内）。当有多份合约同时满足这两个条件时，我们依次优先选择与主合约价内外程度最接近、未平仓合约数更高以及买卖价差更小的合约。

标的资产、主期权和对冲期权的价格均以情景起始时点的指数值 S_0进行归一化，并乘以100进行重新缩放，因此 $S_0 = 100$ 在每个回合中都成立。这种处理保留了三种工具之间的相对尺度，并使损益能够在不同指数水平的各个回合中始终以 S_0 的百分比形式报告。重要的是，这也与训练集中的合成回合所采用的惯例相一致，因此演员网络和评论家网络在训练和测试时接收的输入处于相同尺度。

我们进一步采用了两项筛选条件，其选择旨在匹配合成训练数据分布，从而避免训练数据和测试数据之间出现任何实质性偏差。首先，我们将范围限定为主合约在起始时点处于平值的回合（$|\log(K/S_0)| \leq 0.03$），这与前文各节始终采用的假设一致，即对冲期权在每个回合开始时均为平值。其次，我们将主期权和对冲期权在起始时点的到期时间分别限制为不超过600和800个日历日，以便在足够宽泛但又不过度的到期期限范围内测试智能体。

这些筛选条件合计得到1,193个测试回合，分布于15个独立时间窗口中。显然，不应将同一时间窗口内的回合视为相互独立的抽样，因为它们共享标准普尔500指数的相同历史数据。尤其是在第5.3.3节的逐年分析中，我们承认这一局限性；在该分析中，独立窗口的数量甚至更少（每年约四个）。相关研究通过考虑更长的时间跨度（Xu和Dai（2022）），或利用更高的观测频率（Mikkil¨a和Kanniainen 2023）来纳入更多独立路径，从而应对这一问题。Xiao等（2021）同样处理日观测数据，他们在构建并行情景时采用了与我们类似的方法，并指出，将行权价 K 纳入智能体状态空间，还可以纳入虚值和实值合约——不过我们在此不探讨这一途径。

表2按年份报告了所得样本的汇总统计量。2021年和2022年是数据集中大多数回合集中的年份。尽管如此，由于采用了上述筛选，合约具体信息（即到期日和价内外程度）的统计数据在各年份之间仍具有可比性。主期权与对冲期权到期日之间的平均差异为180个日历日，即大约半年，而价内外程度大致以零为中心。总半价差计算为 $(\tilde{C}^{\mathrm{Ask}} - \tilde{C}^{\mathrm{Bid}})/(\tilde{C}^{\mathrm{Ask}} + \tilde{C}^{\mathrm{Bid}})$，平均为1.05%，这印证了我们对 $\tilde{c} = 1\%$成本率的假设。价差数据与各年份的市场状态相一致：2019年股票市场相对稳定，我们记录到最低水平（0.56%）；2020年市场受到疫情冲击，价差达到最高值（1.94%）。

5.3.2 合成训练集

我们针对仿真到现实迁移应用生成一个专用训练集。与前文一样，我们仍使用 Heston 模型作为市场模拟器，但现在调整模拟设置，以重现实证数据集的结构及其标的资产的特征。我们考虑持续 60 天、每日再平衡的回合，其中期权在 $t_0 = 0$ 时为平值。由于实证数据集中的合约并不要求在每个回合结束时到期，且实际上具有多种不同的到期日，因此我们考虑八种不同的到期日分组，$\tau_0 \in [60, 100, 145, 195, 250, 305, 360, 420]$（交易日），以大致匹配表2所示的实证数据分布范围。假设各分组中主期权与第二个期权的到期日间隔恒定，均为 130 个交易日；这同样反映了数据中观察到的平均差值 $\tilde{\tau}_0 - \tau_0$（按交易日计）。我们仅考虑一种交易成本情景，即 $c = 1$ 基点、$\tilde{c} = 1\%$，与上文表2中根据标普500指数市场数据得出的实证估计一致。我们选取风险厌恶系数 $\lambda = 0.5$，该数值与当前环境中前两个损益矩所采用的典型取值相符（训练集上 Delta Vega 的每步平均损益波动率为 0.07%，而在每天五步的回合中为 0.018%）。本文的实证部分不考察智能体对风险厌恶水平的敏感性。

表3：根据2018年每日标普500指数隐含波动率曲面校准的Heston模型参数（每个交易日进行一次校准，共251天）。我们报告了每个参数的优化边界和汇总统计量。

参数	下界	上界	Mean	Std	Min	Max
κ（均值回复）	1.00	8.00	4.10	1.52	1.00	7.37
$\bar{\nu}$ （长期方差）	0.01	0.10	0.0470	0.0054	0.0376	0.0746
η (波动率的波动率)	0.10	3.00	1.235	0.510	0.371	3.000
ν_0 (初始方差)	0.001	0.10	0.0230	0.0200	0.0012	0.1000
ρ (相关性)	-0.90	-0.10	-0.806	0.055	-0.900	-0.628

为了校准用于生成合成训练路径的 Heston 模型，我们根据相同的底层报价，为 2018 年 251 个标普500指数交易日逐日构建流动性良好的隐含波动率曲面，并保留流动性充足的执行价（未平仓合约数至少为 500，或成交量至少为 10）、相对价差不超过 15%、且价内外程度处于平值附近适中范围内的报价；每个到期日保留流动性最好的 25 个执行价。经过这些筛选后，平均每天约有 220 个可用报价。随后，通过最小化隐含波动率误差平方，将五个 Heston 参数（κ，$\bar{\nu}$，η，ν_0，ρ）校准到每日曲面；校准时使用神经网络代理模型估算 Heston 价格，而非直接进行数值积分，并对每天采用五个随机初始点，保留拟合效果最佳者。251 天均成功收敛。表3汇总了每日参数的分布。

我们生成了 $N_{\text{train}} = 360,000$个回合，训练数据在251个校准参数集之间平均分配。与之前运行中使用的数据集（$N_{\text{train}} = 40,000$个）相比，数据集规模有所扩大，主要有两个原因。首先，与之前的设置相比，每个回合包含的步数更少了（从每天五步减少到一步）；为确保稳健学习，每个训练批次的回合数从 $N_{\text{batch}} = 128$ 增加到256。其次，学习目标更加困难：训练路径现在源自一组不同的模型参数，因此，对不可观测的瞬时方差过程进行推断比之前更加不易。因此，我们也应当能够在更多次迭代中进行训练。

我们在附录A中报告近端策略优化的超参数，并对训练性能进行评论。

5.3.3 实证数据检验

在使用 Heston 模拟的回合训练 PPO智能体后，我们在标准普尔500指数市场数据上对该策略进行样本外测试。基准策略现采用第4.2节所述的业界 BS Delta 和 Delta-Vega对冲策略。

与前几节一样，我们先在图6中展示测试集中的一条代表性路径。该回合追踪了2021年2月9日至2021年5月5日期间标普500指数的走势，反映了市场状况稳定的时期，指数上涨了6.5%。主合约（SPX210618C03940000）是我们持有空头头寸的看涨期权，其对数价内外程度为 $\log(K/S_0) = 0.007$，到期时间为 $\tau_0 = 129$ 日历日。用于对冲的期权（SPX211217C03950000）是一个看涨期权，其 $\log(K/S_0) = 0.01$，且距到期时间更长，为 $\tilde{\tau}_0 = 311$。因此，当主期权距离到期日还剩 $\tau_n = 44$ 天时，该回合便终止；这再次体现了它与合成实验的一个重大区别。左上面板展示了三种证券的价格时间序列（经归一化，使得 $S_0 = 100$）。右上和左下面板分别显示了不同策略在标普500指数和标普500指数期权上的头寸。在此示例中，近端策略优化（PPO）偏离了仅Delta和Delta-Vega从业者对冲策略。具体而言，我们观察到，与Delta-Vega相比，PPO在对冲期权上过度对冲，并通过减少对标的资产的对冲来进行补偿，沿着基准策略的轨迹保持较低的敞口。如下文所示，这种模式并非该回合所独有。

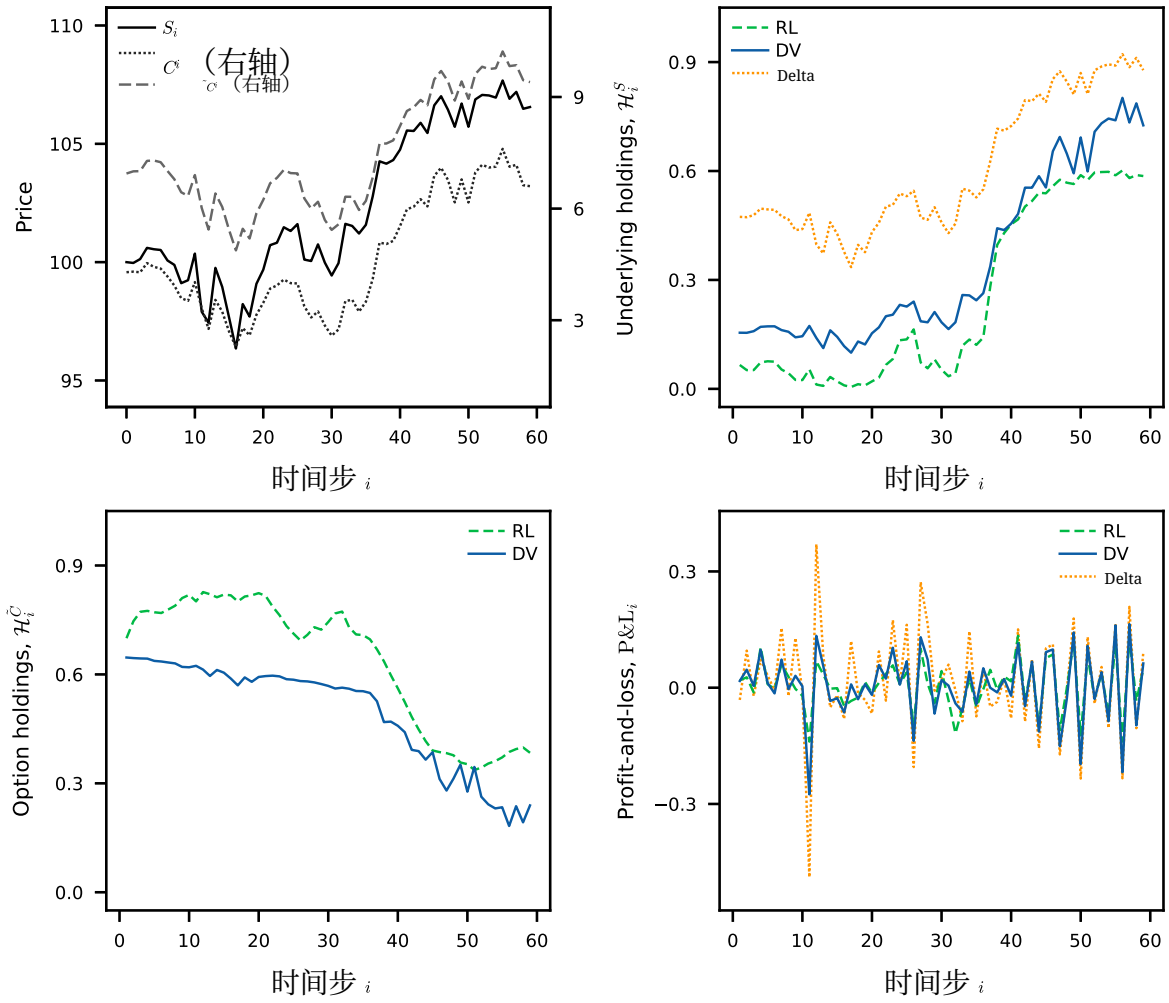

图6：基于2021年2月9日至2021年5月5日标普500指数数据的样本外对冲过程。主期权（SPX210618C03940000）：$\log(K/S_0) = 0.007$，$\tau_0 = 129$ 日历日；对冲期权（SPX211217C03950000）：$\log(K/S_0) = 0.01$，$\tilde{\tau}_0 = 311$。假设交易成本为 $c = 1$ 基点，$\tilde{c} = 1\%$。风险厌恶为 $\lambda = 0.5$。

表4汇总了仿真到现实迁移实验的结果，将PPO智能体与两种从业者基准策略在Heston测试集（上方部分）和标普500测试集（下方部分）上的表现进行比较。总体而言，结果表明迁移成功：与在模拟数据上的表现一样，该智能体在标普500数据上的均值-方差目标方面优于基准策略。我们有三点主要观察。首先，PPO相对于从业者基准策略的优势来源不同于第5.2节中的合成实验。在该实验中，基准策略是Heston模型对冲；在无摩擦极限下，它是最优的（即使方差最小化），因此智能体只有通过以更高的损益离散度换取更低的成本，才能改善目标值。而在这里，BS从业者对冲并非方差最小化策略，无论是在Heston模型下还是在市场数据上都是如此。因此，PPO主要通过降低损益离散度实现更优的均值-方差目标；在Heston路径上的总成本更高（0.041%，而Delta-Vega为0.026%），在标普500数据上的成本则相当（0.059%，而Delta-Vega为0.060%）。表4中的损益标准差数据也证实，在真实数据上使用两种工具对冲的优势依然存在：与仅Delta相比，Delta-Vega和PPO分别使总损益标准差降低了52%和59%；这一结果与留出的Heston路径上观察到的55%和63%相近。

第二 d，正如任何仿真到现实迁移实践一样，考虑两者之间的差异非常重要 s

表4：仿真到现实的对冲绩效。上部分：合成 Heston 测试集，其校准参数、到期日分组和回合结构与训练集相同。下部分：完整的2019–2022年标准普尔500指数测试集。报告值汇总自四次独立训练轮次。所有数值均以 S_0的百分比表示。粗体标示每个区块中均值-方差目标最高的结果。

策略	总损益		资产市场		期权市场		目标
	Mean	Std	Cost	换手率	Cost	换手率	
Heston（n=*10,000*）							
BS Delta	-0.018	1.286	0.010	2.10	0.000	0.00	-0.85
BS Delta-Vega	-0.032	0.582	0.006	0.92	0.020	1.73	-0.20
RL	-0.042	0.481	0.005	0.68	0.036	1.99	**-0.16**
标准普尔*500*指数，*2019–2022*年（n=*1193*）							
BS Delta	-0.339	1.174	0.015	2.63	0.000	0.00	-1.03
BS Delta-Vega	-0.047	0.560	0.010	1.40	0.050	1.95	-0.20
RL	0.051	0.484	0.007	0.90	0.052	2.08	**-0.07**

合成训练数据与用于测试的实证数据之间的差异。在训练过程中，智能体接触的是在风险中性测度 $\mathbb{Q}$ 下生成的路径，这意味着它无法学习从漂移率或风险溢价中获利。这一点体现在Heston测试中：在扣除成本之前， -0.042% 的平均损益约为零（0.041%，综合两种工具）。相比之下，标准普尔500指数价格路径反映了包含漂移率和风险溢价的真实世界动态，这可能导致市场数据上的平均损益不为零，正如我们观察到的，近端策略优化的平均总损益为正（0.051%）。

第三，与前一点相关的是，测试集的规模不足以支持有力的统计结论，因为将15个独立时间窗口视为有效样本量时，平均损益估计值无法与零区分。因此，我们会谨慎解读均值-方差目标数据，并重点关注损益离散度以及各策略如何配置其对冲头寸。

对冲头寸呈现出一致的模式。在模拟测试集和实证数据测试集中，与从业者采用的Delta-Vega基准相比，近端策略优化都表现出系统性地超配对冲期权、低配标的资产。具体而言，在标普500测试集中，我们发现近端策略优化平均持有0.77单位的对冲期权，而Delta-Vega为0.70；同时持有0.12单位的标的资产，而基准为0.18。在模拟数据中也观察到相同现象（对冲期权为0.76对0.70，标的资产为0.11对0.16）。

这一模式与智能体接受训练所依据的Heston模型一致。如第4.2节所示，模型对冲与从业者对冲的差异体现在公式（24）中的修正因子 χ_t^{ν}， χ_t^{S}。在我们的Heston测试集中，模型对冲平均持有0.96单位的对冲期权（裁剪至智能体的边界范围内），而从业者对冲为0.70，因此表明平均而言存在一个修正因子 $\chi_t^{\nu} > 1$ 。这与Heston模型中瞬时方差遵循CIR均值回复过程（公式（16））是一致的，即当前方差的变化对较短期限期权隐含波动率的影响大于对较长期限对冲期权的影响，即 $\partial\sigma_t^{\mathrm{imp}}/\partial\nu_t \geq \partial\tilde{\sigma}_t^{\mathrm{imp}}/\partial\nu_t$。因此，近端策略优化学习到了一种与其训练集数据生成过程一致的策略（如第5.2节所示，该策略脱离纯模型对冲，以计入交易成本），并将其应用于市场数据。

图7展示了三种对冲策略在标普500测试集上的总损益分布。近端策略优化与Delta-Vega的分布大体重叠，而仅Delta的分布更为分散。这些

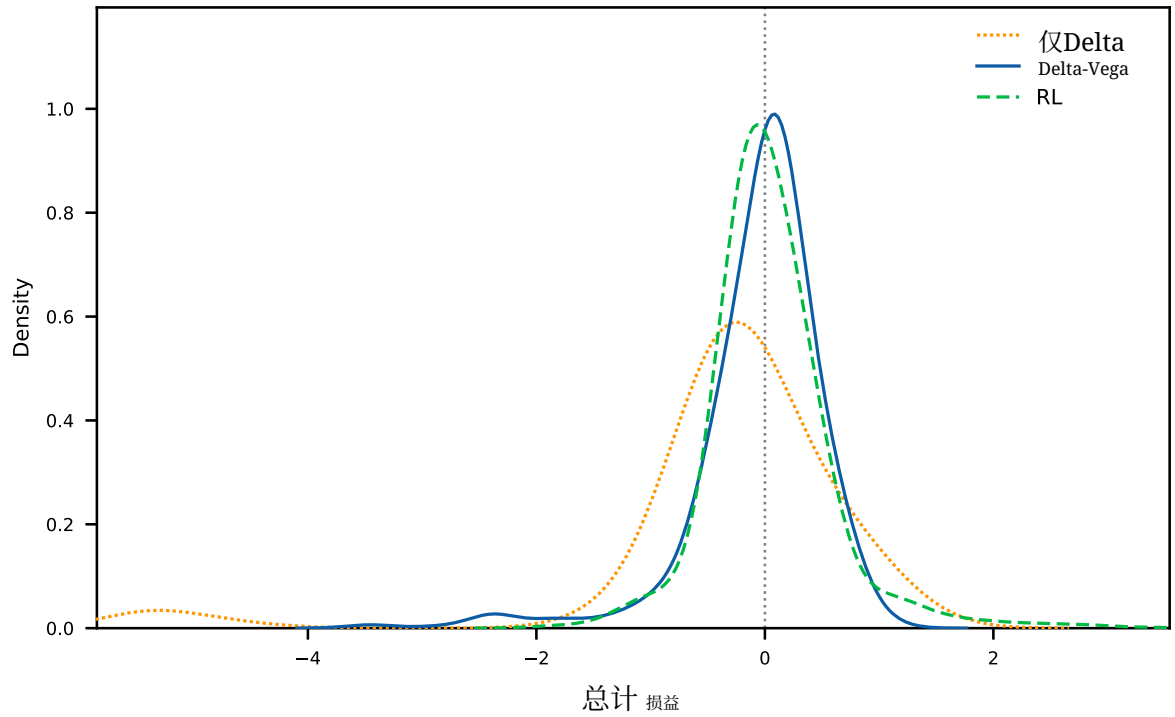

图7：在交易成本为 $c = 1$ 基点、 $\tilde{c} = 1\%$（风险厌恶为 $\lambda = 0.5$）的情况下，基于2019年-2022年标普500指数数据，强化学习（近端策略优化）、仅Delta和Delta-Vega对冲策略的总损益（P&L）分布。曲线为包含1，193个回合的测试集上的核密度估计。

策略在不对称性和尾部风险方面则有所不同。两种基准策略的累积损益分布均呈现负偏度（Delta-Vega -2.31，仅Delta -3.05），其95%条件风险价值（CVaR_{95}）分别为1.84%和4.78%。另一方面，近端策略优化产生的损益呈正偏度（偏度为0.91），且尾部风险较低（$\text{CVaR}_{95} = 0.97\%$）。导致这些统计结果的回合大多集中在2020年，当时标准普尔500指数因疫情冲击而经历了大幅回撤和高波动率。具体而言，仅Delta的总损益最差的5%回合全部发生在2020年；而Delta-Vega最差的60个回合中，有45个发生在该年。另一方面，近端策略优化的尾部集中度较低，其最差的5%回合中只有约三分之一发生在2020年。

我们通过在表5中报告结果按年份的细分情况，考察尾部分析所揭示的状态依赖性。在各年份中，完整样本所呈现的模式依然一致：我们发现，平均而言，在样本涵盖的所有年份中，智能体持有的对冲期权多于Delta-Vega策略，而持有的标的资产少于Delta-Vega策略。此外，相较于从业者策略的超额敞口规模，与该年份的实现波动率呈同向变化。2020年，标的资产日对数收益率的年化标准差为29%（完整样本的数值为19%），而近端策略优化平均持有的对冲期权比从业者策略多20%。同年，近端策略优化的净Vega敞口最大（约为主期权Vega的18%），其相对于基准策略的优势也最大。这也是双工具对冲优势最为明显的一年。仅Delta对冲的总损益标准差为2.18%，超过近端策略优化（0.72%）的三倍，也超过Delta-Vega（1.03%）的两倍。

除2020年外，与仅Delta相比，离散度的降幅较小，Delta-Vega约为28%，近端策略优化约为19%。因此，其他年份的结果更加均衡。2022年是实现波动率较高的一年（24%），近端策略优化排名第一，但其损益标准差高于Delta-Vega。2021年是较为平静的一年（实现波动率为13%），近端策略优化表现不如

表5：按年份列出的标准普尔500指数测试集上的仿真到现实的对冲绩效。约定与表4相同。

策略	总损益		资产市场		期权市场		目标
	Mean	Std	Cost	换手率	Cost	换手率	
			*2019*年（*n=188*）				
BS Delta	−0.178	0.377	0.019	3.01	0.000	0.00	−0.25
BS Delta-Vega	−0.292	0.381	0.015	1.89	0.064	2.21	−0.37
RL	−0.151	0.306	0.008	1.02	0.041	2.01	**−0.20**
			*2020*年（*n=211*）				
BS Delta	−1.565	2.182	0.018	3.04	0.000	0.00	−3.94
BS Delta-Vega	−0.321	1.032	0.013	1.80	0.061	1.94	−0.85
RL	0.072	0.718	0.008	1.00	0.061	2.21	**−0.19**
			*2021*年 *(n=445)*				
BS Delta	0.149	0.581	0.012	2.32	0.000	0.00	−0.02
BS Delta-Vega	0.165	0.255	0.009	1.25	0.036	1.78	**0.13**
RL	0.087	0.378	0.007	0.93	0.046	1.99	0.02
			*2022*年 *(n=349)*				
BS Delta	−0.307	0.319	0.016	2.57	0.000	0.00	−0.36
BS Delta-Vega	−0.018	0.365	0.009	1.08	0.054	2.02	−0.08
RL	0.101	0.458	0.006	0.73	0.062	2.15	**−0.00**

Delta-Vega。2019年在实现波动率方面同样较为平稳，近端策略优化对冲期权的超额敞口最小（平均仅比Delta-Vega高6%），且智能体实现了更低的总成本和更低的损益标准差。

我们认为，我们的结果有助于说明，在模拟环境中训练并迁移至真实市场数据的深度强化学习对冲策略，在实践中可能达到怎样的效果。一方面，在存在市场摩擦的情况下，基于风险中性价格路径训练的强化学习智能体可以成功部署到真实市场数据上，因为近端策略优化仍具有与传统基准策略相当的竞争力。另一方面，实证数据实验凸显了对冲绩效对训练所使用的市场模拟器的依赖性：我们发现，智能体在真实数据上的行为与其在模拟环境中的训练方式高度一致。因此，智能体很可能面临与其训练模型相同的缺陷，这些缺陷会反映在剩余风险敞口中，进而影响实现损益。尽管如此，与基于模型的对冲相比，该智能体仍保留了自然处理离散时间以及交易成本等非线性因素的优势，而这可以说正是深度强化学习对冲方法的主要优点。最后，我们承认，我们每年的样本规模不足以支持有力的统计结论。更高频的数据集，或包含不同标的资产的面板数据，将有助于更严格地评估智能体在不同市场状况下的稳健性。

第6节 结论与未来研究

我们展示了如何采用近端策略优化（PPO）智能体，在不同交易成本情形下，同时使用两种工具对冲欧式期权：使用标的资产管理价格（Delta）风险，使用对冲期权管理波动率（Vega）风险。我们在根据Heston模型模拟的随机波动率环境中训练该智能体，并将策略迁移至真实市场数据。

我们 如今，在无摩擦Heston环境中，近端策略优化能够有效复现理论上的 ly

最优对冲策略。智能体仅根据精简且可由市场观测的状态，在无法观测潜在瞬时方差的情况下，恢复出半解析基准。

引入交易成本后，PPO 学会在成本与复制误差之间进行均值-方差权衡。在低摩擦和高摩擦情景下，智能体的表现均优于半解析Delta对冲和Delta-Vega基准策略；在基准低成本校准下，其结果则具有可比性。该机制在不同成本水平下保持一致：PPO 的交易策略比 Delta-Vega 更为保守，尤其是在期权市场中，从而实现更低的交易成本，同时避免 P&L 实现波动率大幅上升。

针对风险偏好进行的敏感性分析进一步证实了这一发现的稳健性。在我们考察的整个范围内，PPO 的交易成本始终低于 Delta-Vega；其 P&L 波动率也始终低于仅Delta策略，并且在较高摩擦、较高风险厌恶的设定下，偶尔还会低于 Delta-Vega 自身的水平。

随后，我们开展了仿真到现实迁移实验：在一组经过校准的 Heston 参数分布下训练 PPO 智能体，并利用 2019年至2022年的每日标准普尔500指数期权数据，将其与实践者采用的 Black-Scholes Delta 和 Delta-Vega 对冲进行比较。与单工具 Delta 对冲相比，PPO 和 Delta-Vega 大致将 P&L 离散度降低了一半；在交易成本相当的情况下，PPO 达到的 P&L 波动率低于 Delta-Vega。在完整实证样本以及四个单独年份中的三个年份里，PPO 还取得了最佳的均值-方差目标，尽管平均损益的差异并不具有统计显著性。我们发现，相较于基准策略，PPO 的优势主要集中在波动率显著的回合中。这与以下情况一致：智能体提出的对冲策略受到其训练所采用的 Heston模型的塑造。与实践者采用的 Delta-Vega 相比，PPO 持有更多较长期期权，这与 Heston模型隐含的对冲比率一致，凸显了训练环境在塑造所学习策略方面的重要性。

我们的工作可以从多个方向拓展，这些方向也有助于弥补其主要局限。当前实证数据集的日频采样和有限时间跨度，限制了可用于测试的独立回合数量，因此削弱了按年份得出的结论的统计说服力。更长时间跨度、更高频率的数据集将有助于更严格地评估不同市场状态下的样本外表现。

在训练方面，Heston模型虽然易于处理且广为人知，但只是众多候选模拟器之一。更丰富的环境，例如粗糙波动率模型或路径依赖波动率模型，或许能够提高仿真到现实迁移的保真度。同样，文献中已有研究在单工具Delta对冲中采用混合方法，在训练期间结合合成数据与实证数据；这类方法也可拓展至本研究所探讨的多工具场景。

我们的分析还局限于普通欧式期权和单一的均值—方差奖励函数。下一步自然是将该框架拓展至奇异收益；对于这类收益，使用多种工具进行对冲尤为重要，而且尾部敏感目标可能更为合适。

最后，我们采用的比例成本模型虽然以实证数据为基础，但忽略了市场冲击和时变价差。引入更现实的市场微观结构，或直接考虑可观测的买卖价差，将使该设定更接近从业者所面临的实际条件。

参考文献

Albrecht, Stefano V.、Filippos Christianos 和 Lukas Sch¨afer（2024年）。《多智能体强化学习：基础与现代方法》。马萨诸塞州剑桥：麻省理工学院出版社。Armstrong, John 和 George Tatlow（2024年）。《深度伽马对冲》。arXiv：2409 . 13567[q-fin.CP]。Bakshi, Gurdip 和 Nikunj Kapadia（2003年）。《德尔塔对冲收益与负市场波动率风险溢价》。载于：《金融研究评论》16.2，第527–566页。 DOI：10.1093/rfs/hhg002。Black, Fischer 和 Myron Scholes（1973年）。《期权与公司负债定价》。载于：《政治经济学杂志》81.3，第637–654页。 DOI：10.1086/260062。Buehler, H.、L. Gonon、J. Teichmann 和 B. Wood（2019年）。《深度对冲》。载于：《定量金融》19.8，第1271–1291页。 DOI：10.1080/14697688.2019.1571683。Cannelli, L.、G. Nuti、M. Sala 和 O. Szehr（2023年）。《使用强化学习进行对冲：上下文K臂老虎机与Q学习》。载于：《金融与数据科学杂志》9，第100101页。 DOI：10.1016/j.jfds.2023.100101。Cao, J.、J. Chen、S. Farghadani、J. Hull、Z. Poulos、Z. Wang 和 J. Yuan（2023年）。《使用深度分布式强化学习进行伽马与织女星对冲》。载于：《人工智能前沿》6。 DOI：10.3389/frai.2023.1129370。Cao, J.、J. Chen、J. Hull 和 Z. Poulos（2021年）。《使用强化学习进行衍生品深度对冲》。载于：《金融数据科学杂志》3.1，第10–27页。 DOI：10.3905/jfds. 2020.1.052。Carr, Peter 和 Liuren Wu（2009年）。《方差风险溢价》。载于：《金融研究评论》22.3，第1311–1341页。 DOI：10.1093/rfs/hhn038。Cohen, Samuel N.、Christoph Reisinger 和 Sheng Wang（2022年）。《使用神经随机微分方程市场模型对冲期权组合》。载于：《应用数学金融》29.5，第366–401页。 DOI：10.1080/1350486X.2023.2221448。Cont, Rama 和 Milena Vuleti´c（2025年）。《利用生成模型进行数据驱动对冲》。载于：《运筹学年鉴》366.1，第281–312页。 DOI：10.1007/s10479-025-06867-3。Cr´epey, St´ephane（2004年）。《德尔塔对冲织女星风险？》。载于：《定量金融》4.5，第559–579页。DOI：10.1080/14697680400008718。Davis, Mark 和 Jan Ob̷l´oj（2008年）。《使用期权完成市场》。载于：《巴拿赫中心出版物》83，第49–60页。arXiv：0710.2792。DeltaNeutral LLC（2026）。《历史期权数据–二级》。日终历史期权数据。URL：https://historicaloptiondata.com/data-structure-greeks/。Du, Jiayi、Muyang Jin、Petter Kolm、Gordon Ritter、Yixuan Wang 和 Bofei Zhang（2020年）。《用于期权复制与对冲的深度强化学习》。载于：《金融数据科学杂志》2.4，第44–57页。 DOI：10.3905/jfds.2020.1.045。Dupire, Bruno（1994年）。《微笑曲线定价》。载于：《Risk》7.1，第18–20页。Fraņcois, Pascal、Genevi`eve Gauthier、Fr´ed´eric Godin 和 Carlos O. P´erez-Mendoza（2025年）。《使用隐含波动率曲面的期权深度对冲》。arXiv：2504.06208 [q-fin.RM]。Giurca, Alexandru 和 Svetlana Borovkova（2021年）。《使用深度强化学习进行衍生品德尔塔对冲》。可见于SSRN。 DOI：10.2139/ssrn.3847272。Halperin, Igor（2020年）。《QLBS：Black-Scholes（-Merton）世界中的Q学习器》。载于：《〈衍生品杂志》28.1，第99–122页。 DOI：10.3905/jod.2020.1.108。Hambly, B.、R. Xu 和 H. Yang（2023年）。《金融强化学习最新进展》。载于：《数学金融》33.3，第437–503页。 DOI：10.1111/mafi.12382。Herrmann, Sebastian 和 Johannes Muhle-Karbe（2017年）。《模型不确定性、重新校准与德尔塔–织女星对冲的出现》。载于：《金融与随机学》21.4，第873–930页。DOI：10.1007/s00780-017-0342-6。

Heston, Steven L.（1993年）。“具有随机波动率的期权的闭式解及其在债券和货币期权中的应用”。载于：《金融研究评论》6.2，第327–343页。 DOI：10.1093/rfs/6.2.327。Hull, John 和 Alan White（1987年）。“对冲外币期权卖方所面临的风险”。载于：《国际货币与金融杂志》6.2，第131–152页。 DOI：10.1016/0261- 5606(87)90029-5。——（2017年）。“期权的最优Delta对冲”。载于：《银行与金融杂志》82，第180– 190页。 DOI：10.1016/j.jbankfin.2017.05.006。Kaeck, Andreas、Vincent Van Kervel 和 Norman J. Seeger（2022年）。“指数期权市场中的价格冲击与买卖价差”。载于：《金融市场杂志》59，第100675页。 DOI：10.1016/j.finmar.2021.100675。Kolm, Petter N. 和 Gordon Ritter（2019年）。“动态复制与对冲：一种强化学习方法”。载于：《金融数据科学杂志》1.1，第159–171页。 DOI：10.3905/jfds.2019.1.1.159。Leland, Hayne E.（1985年）。“含交易成本的期权定价与复制”。载于：《金融杂志》40.5，第1283–1301页。 DOI：10.1111/j.1540-6261.1985.tb02383.x。Merton, Robert Cox（1973年）。“理性期权定价理论”。载于：《贝尔经济学与管理科学杂志》4.1，第141–183页。 DOI：10.2307/3003143。Mikkil¨a, Oskari 和 Juho Kanniainen（2023年）。“实证深度对冲”。载于：《定量金融》23.1，第111–122页。 DOI：10.1080/14697688.2022.2136037。Muravyev, Dmitriy 和 Neil D. Pearson（2020年）。“期权交易成本低于你的想象”。载于：《金融研究评论》33.11，第4973–5014页。 DOI：10.1093/rfs/ hhaa010。Neagu, Andrei、Fr´ed´eric Godin 和 Leila Kosseim（2025年）。《期权对冲的深度强化学习算法》。arXiv：2504.05521。Pickard, Reilly 和 Yuri Lawryshyn（2023年）。“动态股票期权对冲的深度强化学习：综述”。载于：《数学》11.24，第4943页。 DOI：10.3390/math11244943。Ritter, Gordon（2017年）。“用于交易的机器学习”。载于：Risk 30.10，第84–89页。Romano, Marc 和 Nizar Touzi（1997年）。“随机波动率模型中的或有索取权与市场完备性”。载于：《数学金融》7.4，第399–412页。 DOI：10.1111/1467- 9965.00038。Schulman, John、Sergey Levine、Philipp Moritz、Michael I. Jordan 和 Pieter Abbeel（2017年）。《信赖域策略优化》。arXiv：1502.05477 [cs.LG]。Schulman, John、Philipp Moritz、Sergey Levine、Michael Jordan 和 Pieter Abbeel（2018年）。《使用广义优势估计的高维连续控制》。arXiv：1506.02438 [cs.LG]。Schulman, John、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov（2017年）。《近端策略优化算法》。arXiv：1707.06347 [cs.LG]。Sutton, R. S. 和 A. G. Barto（2018年）。《强化学习：导论》。第2版。麻省理工学院出版社。Sutton, Richard S.、David McAllester、Satinder Singh 和 Yishay Mansour（1999年）。“使用函数逼近的强化学习策略梯度方法”。载于：《神经信息处理系统进展》第12卷。麻省理工学院出版社。Vittori, E.、M. Trapletti 和 M. Restelli（2020年）。“风险厌恶强化学习下的期权对冲”。载于：《首届ACM金融人工智能国际会议论文集》。ACM，第1–8页。 DOI：10.1145/3383455.3422532。Whalley, A. E. 和 P. Wilmott（1997年）。“含交易成本的期权定价最优对冲模型的渐近分析”。载于：《数学金融》7.3，第307–324页。 DOI：10.1111/1467-9965.00034。

Xiao, Bo、Wuguannan Yao 和 Xiang Zhou（2021年）。“利用策略梯度进行最优期权对冲”。载于：*2021*年国际数据挖掘研讨会会议（*ICDMW*），第1112– 1119页。 DOI: 10.1109/ICDMW53433.2021.00145。Xu, Wei 和 Bing Dai（2022年）。“强化学习技术下考虑交易成本的类Delta-Gamma对冲”。载于：《衍生品杂志》29.5，第60–82页。 DOI: 10.3905/jod.2022.1.156。Zernikov, Kirill（2026）。深度对冲究竟学到了什么？*Delta*修正、机制脆弱性与符号蒸馏。arXiv：2605.21696 [q-fin.RM]。Zhao, Wenshuai、Jorge Pe˜na Queralta 和 Tomi Westerlund（2020年）。“机器人学深度强化学习中的仿真到现实迁移：综述”。载于：*2020*年*IEEE*计算智能研讨会系列（*SSCI*）。IEEE，第737–744页。 DOI: 10.1109/SSCI47803.2020. 9308468。

表6：合成数据和仿真到现实训练运行的主要近端策略优化及环境超参数。

	合成数据	仿真到现实
网络架构		
演员隐藏层	128×128	256×256
评论家隐藏层	$128 \times 128 \times 128 \times 128$	$256 \times 256 \times 256 \times 256$
优化		
学习率	3×10^{-4}	3×10^{-4}
PPO裁剪比率 ε	0.1	0.1
评论家损失权重 c_V	0.5	0.5
折扣因子 γ	0.99（如果 $c = \tilde{c} = 0$ 则为 0）	0.99
λ_{GAE}	0.99	0.99
每次迭代的训练轮数 N_{epochs}	25	25
批量大小 N_{batch}	128	256
训练迭代次数 N_{iter}	300	1,200

训练设置与超参数

表6列出了合成数据（第5.2节）和实证数据（第5.3节）训练运行的主要超参数。两种设置采用相同的近端策略优化（PPO）优化超参数，但在网络宽度、训练时长和批量大小方面有所不同。正如第5.3.2节所述，这样做是为了适应为仿真到现实实验构建的规模更大、更多样化的数据集。

图 8 展示了无摩擦 Heston 情形下四次独立训练运行的平均训练曲线。在训练期间的不同检查点，我们通过计算留出验证集上的平均回合奖励和累计盈亏标准差来评估当前策略，该验证集的大小等于 N_{batch}。在这种设定下，收敛速度非常快，近端策略优化在最初几次训练迭代中就达到了 Delta-Vega 最优策略的奖励水平。相应地，到训练结束时，验证集各回合的总损益波动率已达到基准水平。

图 9 展示了包含交易成本的 Heston 环境下的训练指标（采用第5.2节的基准校准）。与无摩擦情形不同，近端策略优化虽然起初呈现出较慢的学习模式，但最终超过了基准奖励水平，而此时的基准已不再是最优的。需要指出的是，这里展示的平均奖励（近端策略优化高于基准）与表 1 中展示的目标值并不矛盾；两者实际上几乎重合。如第 4.4 节所述，奖励函数是对目标值的近似；此外，表中报告的测试结果是在一个包含 10，000 个回合的数据集上计算得出的，而当前图中报告的验证指标则基于一个包含 $N_{\text{batch}} = 128$ 个回合的数据集。

图10展示了仿真到现实运行的训练曲线。不同于上文所述的合成环境（这些环境针对单一期权到期日和单一 Heston 参数化进行校准），当前训练集包含多个到期日和校准参数（第5.3.2节），因此学习问题更为困难，且噪声明显更大。我们注意到，个别随机种子在训练期间偶尔会出现验证奖励的短暂下降。由于整个过程中我们都采用早停并恢复最佳检查点，因此我们仅报告每个随机种子截至其最佳验证奖励时的曲线，即实际保留并在第5.3.3节中进行评估的策略的训练轨迹，而不是完整且噪声更大的运行曲线。近端策略优化最终达到并超过了 BS从业者Delta-Vega 策略所取得的验证奖励。正如上文所述，我们强调，图中所示曲线表示在一个包含 $N_{\text{batch}} = 256$ 个回合的验证集上的奖励，该验证集用于评估训练状态，而第5.3.3节中的最终结果展示了

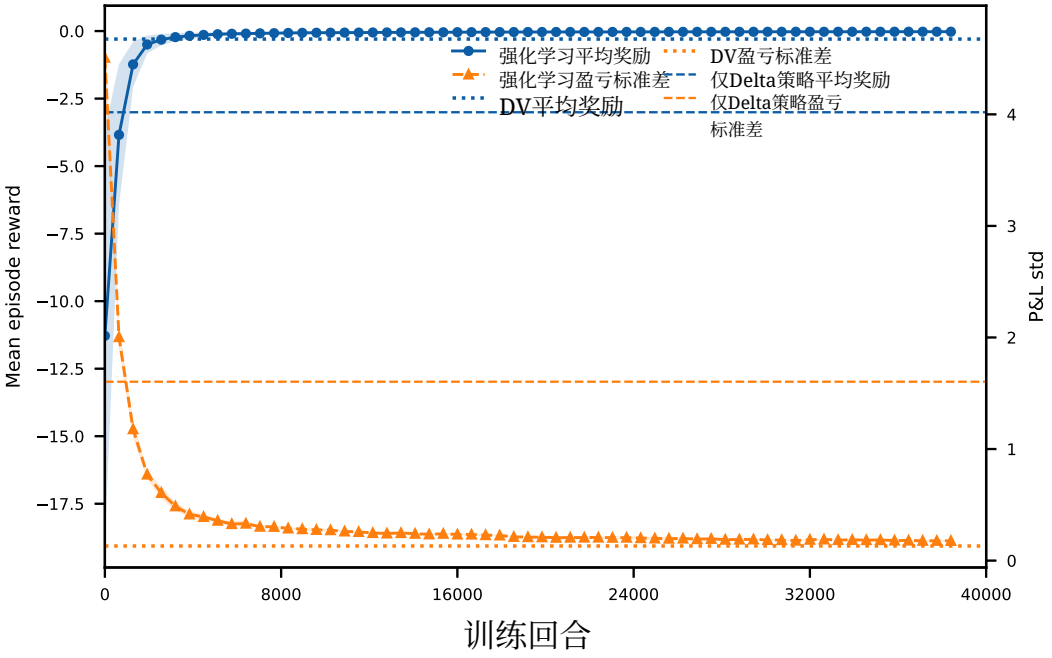

图 8：在无摩擦 Heston 环境中训练期间，于每个检查点进行评估时，留出验证集上的平均回合奖励（左轴）和损益标准差（右轴）。阴影带表示四个独立随机种子下的一个标准差。点线和虚线水平线分别标示 Heston模型的 Delta-Vega 和仅Delta基准策略的指标。

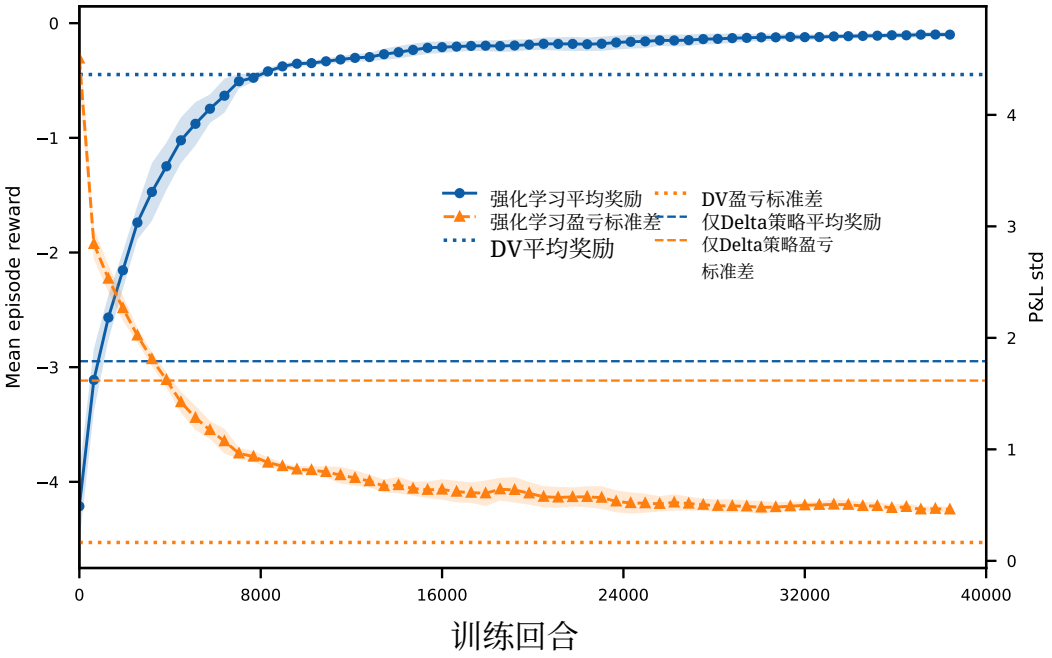

图 9：在包含交易成本的 Heston 环境中训练期间，于每个检查点进行评估时，留出验证集上的平均回合奖励（左轴）和损益标准差（右轴）。阴影带表示四个独立随机种子下的一个标准差。点线和虚线水平线分别标示 Heston模型的 Delta-Vega 和仅Delta基准策略的指标。

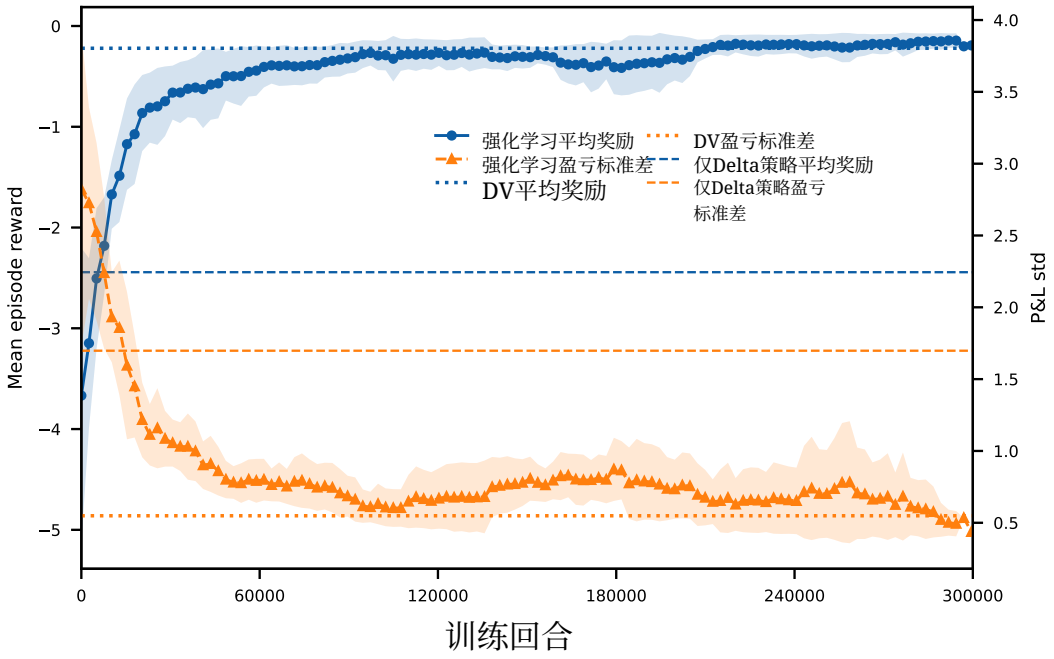

图10：在为仿真到现实设计的多到期日、多校准Heston数据集上进行训练期间，不同检查点在留出验证集上的平均回合奖励（左轴）和损益标准差（右轴）。阴影带表示四个独立随机种子下的一个标准差。点线和虚线水平线分别标示BS从业者Delta-Vega和仅Delta基准策略的指标。

在 $N_{\text{test}} = 10{,}000$个合成回合以及标准普尔500指数和SPX期权的1,193个历史回合上计算的实际均值-方差目标。