# 信息不对称下带验证的机器学习模型交易

Xiang Li, Jianwei Huang*, *IEEE* 会士，Kai Yang*, *IEEE* 高级会员，和 Chenyou Fa n.

arXiv:2601.07510v1 [cs.GT] 12 Jan 2026

摘要——**机器学习（ML）模型交易以保护数据隐私著称，但也面临一项重大挑战：信息不对称。这一问题可能导致模型欺骗，即卖方虚报模型性能以获取更多收益，而现有文献尚未充分解决这一问题。我们提出一种博弈论方法，在机器学习模型市场中加入验证环节，使买方能够在购买前检查模型质量。然而，这种方法成本较高，且提供的信息并不完备，因而增加了买方决策的难度。我们的分析表明，卖方可能会根据模型被验证的概率，以一定概率实施模型欺骗。欺骗概率随验证准确率提高而降低，随验证成本增加而上升。为使卖方收益最大化，我们进一步设计了考虑异质买家策略行为的最优定价方案。有趣的是，我们发现，减少信息不对称有利于卖方和买方。与此同时，保护买方订单信息并不能提高买方或卖方的收益。这些发现凸显了减少机器学习模型交易中信息不对称的重要性，也为未来研究开辟了新方向。**

索引词—**机器学习模型交易、信息不对称、博弈论、机制设计**

## 一、引言

机器学习在许多数据驱动应用中变得日益重要。因此，机器学习模型交易也越来越受欢迎，各类模型市场便是明证，例如 Modzy [2] 和 StreamML [3]。尤其是，与数据交易不同，模型交易通过避免直接访问数据，能够有效缓解隐私问题。此外，模型买方可以直接部署所购的机器学习模型，无需进行额外的数据处理。

尽管近期研究（例如 [4]–[7]）已初步尝试分析机器学习模型市场，但通常假设信息完全，即模型质量对买方和卖方都是透明的。然而在实践中，买方在购买和部署之前往往不了解模型的实际质量，从而导致信息不对称。追求自身利益的卖方可能会歪曲模型质量，以便即使考虑声誉成本等未来后果，也会最大化收益 [8]，只要收益超过这些成本。这可能导致买方为低质量模型支付过高价格并遭受不利影响。例如，在桥梁安全监测中，低质量检测模型产生错误预测可能造成灾难性后果。为避免潜在损失，买方可能因此选择只购买最低质量的模型 [9]，或者完全不参与交易，从而阻碍模型市场的发展。

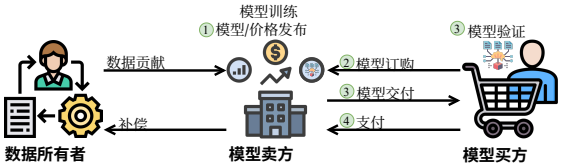


图 1. 带验证的机器学习模型交易（MTV）框架

为应对信息不对称这一挑战，我们提出了带验证的机器学习模型交易（MTV）框架，如图 1 所示。MTV 框架使买方能够在购买前使用自己的测试数据验证模型质量（例如，通过 Aimodelplace [10] 上的测试接口），从而减少模型质量方面的信息不对称。然而，数据收集所产生的验证成本，以及测试数据抽样偏差导致的不完美验证 [11]，可能会阻碍买方进行验证。这促使我们研究第一个关键问题：

**问题 1**。 买方应如何进行模型验证？

机器学习模型的独特特征，以及模型交易中相互交织的因素，为解决问题 1 带来了挑战。具体而言，由于模型预测具有概率性 [12]，买方很难根据验证结果评估模型质量并确定相应的验收标准。此外，验证的成本效益不仅取决于买方和卖方双方的属性（例如测试数据规模和模型质量版本），还取决于双方之间的互动，这进一步增加了买方制定战略性验证决策的复杂性。

鉴于上述问题，我们从理论上量化模型验证的结果，以解决问题 1。具体而言，我们以概率方式评估买方的验证准确率，同时考虑测试数据规模和所购买模型的性能。通过分析不同条件下买方的期望收益，我们推导出有关模型验证的最优决策。在此基础上，我们进一步探讨本文的第二个关键问题：

**问题 2**。验证机制如何影响机器学习模型交易中买方和卖方的收益？

*Jianwei Huang 和 Kai Yang 是本文的共同通讯作者。Xiang Li 任职于深圳市人工智能与机器人研究院（AIRS）和香港中文大学（深圳）理工学院（SSE）（电子邮箱：xiangli2@link.cuhk.edu.cn）。Jianwei Huang 任职于香港中文大学（深圳）理工学院、深圳市人工智能与机器人研究院、深圳市群体智能赋能低碳能源网络重点实验室和 CSIJRI 智能储能联合研究中心，地址为中国广东省深圳市，邮编 518172（电子邮箱：jianwei-huang@cuhk.edu.cn）。Kai Yang 任职于同济大学计算机科学与技术系、教育部嵌入式系统与服务计算重点实验室和上海智能自主系统研究院（电子邮箱：kaiyang@tongji.edu.cn）。Chenyou Fan 任职于华南师范大学（电子邮箱：fanchenyou@gmail.com）。部分分析和结果曾发表于 IEEE ICC 2023会议 [1]。

买卖双方在交易策略方面的信息不完全，使得评估个体收益变得困难。同时，随着模型版本（具有不同质量等级）的数量增加，推导最优决策和均衡收益也会变得愈发复杂。鉴于机器学习模型泛化误差（质量）的改进与训练数据规模（成本）之间呈现陡峭的幂律缩放关系，例如，[13]–[15], ，我们简化了模型交易中的均衡分析。具体而言，利用学习曲线的幂律特征，我们将卖方的策略空间简化为取决于买方订单的二元选项，从而有效地回答了问题2。

随着人们对市场中基于数据分析的歧视行为日益关注（例如，[16]–[18]），我们的研究进一步扩展至探讨买方订单信息保护的影响。在机器学习模型交易中，买方可能面临订单信息泄露的风险，从而使卖方能够采取歧视性欺骗策略。考虑到旨在防止卖方采取“个性化”交易策略的潜在法规（如《罗宾逊-帕特曼法》[19]和《数据治理法案》[20]），我们深入探讨第三个核心问题：

**问题 3.** 保护买方订单信息会如何影响买方和卖方的策略与收益？

订单信息保护进一步加剧了模型交易中的策略互动复杂性。具体而言，卖方在做出模型交付决策时，需要预判买方可能采取的各种订单和验证策略，因此我们研究了卖方和买方的反应对应关系 [21]。

本文的主要结果和贡献如下：

- 一种新颖的问题表述：据我们所知，这是首个研究信息不对称下机器学习模型交易的研究，其中追求自身利益的卖方可能会歪曲模型质量以实现收益最大化。此外，我们还探讨了模型市场中订单信息保护的影响，这种保护可以防止卖方采取潜在的歧视性交易策略。
- 模型验证的有效性：我们通过剔除劣势策略并融入学习曲线的幂律特征，改进了交易分析，从而能够在没有订单信息保护的情况下确定模型交易的均衡。我们的研究揭示了一项关键发现：模型验证能够通过减少信息不对称，同时使买方和卖方受益。具体而言，具有成本效益的验证可以缓解信息不对称所带来的策略张力，从而提高双方的收益。
- 订单信息保护的影响：我们对模型交易中订单信息保护的研究表明，这种保护并不能提高买方或卖方的收益。值得注意的是，尽管保护订单信息可以防止卖方采取歧视性交易策略，但也会抑制买方验证交付的模型。缺少模型验证反过来会促使卖方采取模型欺骗，导致买方和卖方双方的收益都低于未实施订单信息保护时的收益。
- 模型定价方案设计：我们针对有订单信息保护和无订单信息保护的模型交易，提出了实现卖方收益最大化的最优定价方案。具体而言，考虑到机器学习模型市场中的异质买家，我们设计了一种能够适应买家模型效用差异的定价算法。大量数值实验表明，在信息不对称的情况下，我们的定价方案所实现的卖方收益接近完全信息基准。

## II. 相关工作

本节回顾数据市场和机器学习模型市场的相关研究。

正如综述 [22]–[24],所总结的，数据市场研究已取得令人瞩目的进展，尤其是在隐私保护和数据定价方面（例如，[25]–[36]）。具体而言，Ghosh 等人在 [25] 从差分隐私的角度研究了数据市场。Hynes 等人 在 [26]利用隐私保护智能合约构建了去中心化数据市场。Niu 等人在[29] 提出了基于成本的数据定价框架，以补偿数据所有者的隐私损失。Koutris 等人 在 [32]–[34] 提出了基于查询的数据市场框架，从而实现灵活的数据评估和定价。相比之下，本文聚焦于数据驱动的模型市场。鉴于机器学习模型的独特特性，这一领域的研究相对不足，且问题更为复杂。

对机器学习模型市场的研究直到近期才开始兴起，详见 [4]–[7]。Chen 等人在 [4] 设计了一种基于模型的定价框架，直接为机器学习模型定价，而非为原始数据定价。Agarwal 等人 在 [5] 研究了机器学习模型市场所需的一系列算法。Liu 等人 在 [6]探讨了数据使用限制对模型交易中卖方的影响。Sun 等人 在 [7]提出了一种差分隐私联邦学习模型市场。然而，这些研究忽视了模型交易中的信息不对称问题：卖方可能为了实现收益最大化而虚报模型质量。

本文其余部分安排如下。第三节介绍系统模型，并给出三阶段博弈建模。第四节对机器学习模型交易流程的阶段3进行均衡分析。第五节和第六节分别推导阶段2中买方的最优订购决策，以及阶段1中卖方的最优定价方案。第七节将研究扩展至具有订单信息保护的机器学习模型交易。第八节讨论异质买方效用下的定价方案。第九节给出数值结果，并在第十节总结全文。由于篇幅限制，结果的详细证明留在附录中。

## III. 系统模型

本节介绍不具有订单信息保护的机器学习模型交易流程，并刻画买方和卖方双方的决策与收益。

### *A.* 机器学习模型交易流程

我们首先简要描述机器学习模型交易流程，并将其建模为以下三阶段顺序博弈：

- **阶段1：** 卖方发布一组机器学习模型，每个模型都标有价格和特定的公告质量。

• **第2阶段：** 买方决定是否购买，以及从已发布的机器学习模型中购买哪一个。
• **阶段3：** 如果买方购买机器学习模型，卖方将决定交付策略，即为买方订购的模型交付何种真实质量（这可能不同于阶段1中公告的质量）。与此同时，买方 将决定验证策略，即是否使用本地测试数据验证交付的机器学习模型，这会产生相应的成本。此外，买方还会设定相应的接受标准（即如果经过验证，何时接受交付的模型）。

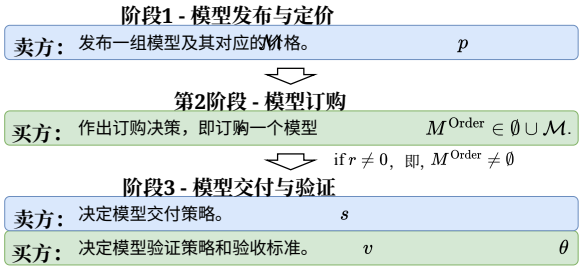


图2. 机器学习模型交易流程

接下来，我们将详细阐述每个阶g段，如图 g2所示。

*1）阶段 1 –* 模型发布：在阶段1中，卖方训练一组质量等级不同的机器学习模型，然后公布这些模型及其对应的价格：

• 已发布模型集合：令 $\mathcal{N} = \{1, 2, \cdots, N\}$ 表示机器学习模型索引集，令 $\mathcal{M} = \{M_n, n \in \mathcal{N}\}$表示已发布待售模型的集合。遵循文献中的常见惯例，例如 [5], [35], [37],我们将模型质量定义为符合买方标准的结果所占的比例，例如准确预测或损失低于预定阈值的结果所占的比例。令 $\alpha_n$ 表示模型 $M_n$的这一质量指标，其中 $\alpha_n$ 的取值范围为 $[0, 1]$，对每个 $n \in \mathcal{N}$均如此，$m \in \mathcal{N}$。买方使用模型 $M_n$ 所获得的效用 $U_n$ 随模型质量 $\alpha_n$的提高而增加。因此，对于所有 $n$， $m \in \mathcal{N}$，只要 $n \leq m$，就有 $0 \leq U_n \leq U_m$。

• 模型价格：我们用 $p_n$ 表示每个 $n \in \mathcal{N}$中模型 $M_n$的标价。模型价格与质量相匹配：当所有 $n$， $m \in \mathbb{N}$时， $0 \leq p_n \leq p_m$ $n \leq m$ 。

*2）阶段 2 –*模型订购：卖方在阶段1发布机器学习模型后，买方决定是否购买，以及（如果购买）从模型集合 $\mathcal{M}$中订购哪种机器学习模型 $M^{\text{Order}}$ 。我们将买方的订购决策表示为 $r \in \{0\} \cup \mathcal{N}$，其中包括不订购这一选项。因此，

$$M^{\text{Order}} = \begin{cases} \emptyset, & \text{if } r = 0, \\ M_r, & \text{if } r \in \mathcal{N}. \end{cases} \tag{1}$$

*3）阶段 3 –* 模型交付与模型验证：在阶段3，卖方决定模型交付策略，买方决定验证策略和接受标准，具体如下：

• 模型交付策略：我们将卖方的交付策略定义为 $s \in \mathcal{N}$，表示交付的机器学习模型的实际质量为 $M_{\text{交付}} \in \mathcal{M}$，即 $M_{\text{交付}} = M_s$。

• 模型验证策略：在按照阶段1公布的价格向卖方付款之前，买方可以决定是否使用其[1] 本地测试数据验证交付的机器学习模型质量。我们将买方的测试数据规模记为 $T \in \mathbb{N}^+$，并用 $v \in \{NV, V\}$ 表示其验证策略：

$$v = \begin{cases} NV, & \text{not verify}, \\ V, & \text{verify}. \end{cases} \tag{2}$$

然而，模型验证往往并不完美。例如，同一机器学习模型在不同测试数据集 [38],上的表现可能有所不同，测试数据 [11] 的抽样偏差会影响验证结果。

• 接受标准：采用模型验证时，买方必须设定一个接受标准，以确定在验证结果并不完美的情况下，哪些验证结果足以接受交付的模型。我们将 $\theta \in \mathbb{N}$记为买方的接受标准，即买方愿意接受的、由验证结果证明的最低模型质量。为简单起见，我们考虑 $\theta \in [0, T+1]$，其中 $\theta = 0$ 表示无论验证结果如何，买方都会接受任何模型，而 $\theta = T + 1$ 则相反，即拒绝所有模型。给定标准 $\theta$ 和交付策略 $s \in \mathcal{N}$，我们可以确定交付的模型 $M_{\text{交付}} = M_s$的买方接受概率 $\delta(\theta, s)$。具体而言，在标准 $\theta$ 下，当买方收到质量指标为 $\alpha_s$的模型 $M_s$ 时，我们根据式（3）计算接受概率 $\delta(\theta, s)$，其取值介于 0 和 1 之间。

$$\delta(\theta, s) = \sum_{i=\theta}^{T} \binom{T}{i} \alpha_s^i (1 - \alpha_s)^{T-i}. \tag{3}$$

根据式（3），给定 $T$ 个测试数据样本时， $\delta(\theta, s)$ 是模型 $M_s$ 在所有可能验证结果下满足买方标准 $\theta$ 的累积概率。图 3 直观展示了接受标准、概率与机器学习模型质量之间的关系。带有斜线的蓝色阴影区域表示在标准 $\theta$ 下，交付的模型 $M^{\text{Deliver}} = M_r$时买方的接受概率 $\delta(\theta, r)$。

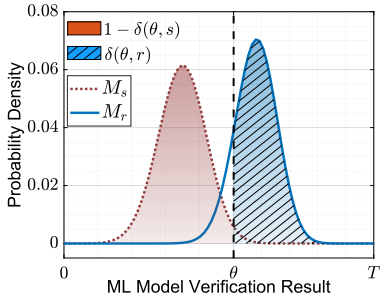


图 3. 接受标准与概率示意图

## *B.* 买方预期收益

买方收益包括使用机器学习模型的期望效用、模型的预期付款以及验证成本。接下来，我们将详细介绍各个部分。

[1]为便于讨论，我们用“她”指代卖方，用“他”指代买方。这种称谓选择并不意味着存在任何性别偏见。

*1*）期望模型效用：模型效用取决于交付的机器学习模型的实际质量。注意 $U_n$ 是买方使用模型 $M_n$时的效用，其中，如果 $n \le m$，则 $0 \le U_n \le U_m$，对于 $n,\ m \in \mathcal{N}$。然而，由于模型质量可能并非卖方所声称的那样，当买方订购模型 $M_r$，而卖方交付模型 $M_s$时，买方期望效用为：

$$\mathbb{E}[U(r,s,v,\theta)] = \begin{cases} 0, & \text{if } r = 0, \quad (4a) \\ U_s, & \text{if } r \in \mathcal{N},\ v = NV, \quad (4b) \\ \delta(\theta,s)\cdot U_s, & \text{if } r \in \mathcal{N},\ v = V. \quad (4c) \end{cases}$$

根据 (4a)，当买方不订购任何模型时，效用为零。如果买方订购模型 $M_r$ 但不进行验证（$v = NV$），如 (4b) 所示，他会直接接受该模型，效用为 $U_s$，具体取决于交付策略 $s$。最后，在 (4c) 中，如果买方订购 $M_r$，并在实际付款前验证交付的模型，则验证将以概率 $\delta(\theta, s)$ 成功（对交付的模型 $M_s$ 进行测试，确认其质量不低于买方的接受标准 $\theta$），因此期望效用为 $\delta(\theta,\ s)U_s$。

*2*）模型期望支付：机器学习模型的期望支付取决于以下交易决策：

$$\mathbb{E}[P(r,s,v,\theta)] = \begin{cases} 0, & \text{if } r = 0, \quad (5a) \\ p_r, & \text{if } r \in \mathcal{N},\ v = NV, \quad (5b) \\ \delta(\theta,s)\cdot p_r, & \text{if } r \in \mathcal{N},\ v = V. \quad (5c) \end{cases}$$

根据 (5a)，当买方不向卖方订购任何模型时，支付为零。如果买方订购机器学习模型 $M_r$ 但不进行验证，则如 (5b) 所述，他将支付 $p_r$。如果买方决定验证交付的模型，则支付取决于验证结果，接受概率为 $\delta(\theta,\ s)$，如 (5c) 所示。

*3*）验证成本：为简单起见，我们假设买方需要支付成本 $C_T > 0$ 来收集用于模型验证的测试数据。这样可以灵活地刻画验证成本 $C_T$ 与测试样本量 $T$之间的不同关系。

*4*）买方期望收益：总之，机器学习模型交易中的买方期望收益$\pi_{buyer}(r,\ s,\ v,\ \theta)$为：

$$\begin{aligned} \pi_{buyer}(r,s,v,\theta) =& \mathbb{E}[U(r,s,v,\theta) - P(r,s,v,\theta)] \\ & - C_T \cdot \mathbf{1}_{v=V}, \end{aligned} \quad (6)$$

其中，$\mathbb{E}[U(r,\ s,\ v,\ \theta)]$和 $\mathbb{E}[P(r,\ s,\ v,\ \theta)]$分别在式（4）和式（5）中定义。指示变量 $\mathbf{1}_{v=V}$ 表示买方是否会验证交付的模型。买方旨在对 $r$、$v$和 $\theta$作出最优决策，以最大化$\pi_{buyer}(r,\ s,\ v,\ \theta)$。

### C. 卖方期望收益

卖方收益不仅包括买方的预期付款（即卖方收入），还包括销售机器学习模型的预期成本，下面将对此进行讨论。

*1*）模型期望成本：对卖方而言，机器学习模型交易会产生向数据所有者收集训练数据的成本以及处理数据的成本。由于模型是由数据衍生而来的数字资产，我们重点关注补偿数据所有者敏感隐私损失的成本，如 [6]所示。也就是说，我们将所有一次性成本[2]（例如计算成本）分摊到每个交易，并将其归一化为零。令 $C_n$ 表示交易模型 $M_n$ 的成本，其中 $0 < C_n < C_m$，当 $n < m$ 时，对于每个 $n \in \mathcal{N}$（即，模型质量改进通常需要更大的训练数据规模，因此成本也更高）。我们将销售模型的预期成本总结如下：

$$\mathbb{E}[C_M(r,s,v,\theta)] = \begin{cases} 0, & \text{if } r = 0, \quad (7a) \\ C_s, & \text{if } r \in \mathcal{N},\ v = NV, \quad (7b) \\ \delta(\theta,s)\cdot C_s, & \text{if } r \in \mathcal{N},\ v = V. \quad (7c) \end{cases}$$

请注意，当买方在实践中应用机器学习模型时会产生隐私补偿，即当买方接受交付的模型时，卖方会承担模型成本。因此，对（7）的解释与对（4）的解释类似，后者同样取决于买方是否接受交付的机器学习模型。

*2*）卖方期望收益：总而言之，卖方期望收益是收入与模型成本之差：

$$\pi_{seller}(r,s,v,\theta) = \mathbb{E}[P(r,s,v,\theta) - C_M(r,s,v,\theta)]. \quad (8)$$

接下来，我们通过逆向归纳法分析三阶段序贯博弈并推导交易均衡 [21]。

## IV. 阶段3中的模型交付与验证

本节首先对模型交易过程阶段3中买方与卖方之间的同时博弈进行建模，然后推导该博弈的纳什均衡。

### A. 交易博弈建模

给定卖方在阶段1公布的模型信息以及买方在阶段2的订单决策[3] $r$，我们将阶段3的交易博弈表述如下：

**博弈1** (阶段3的交易博弈)。 阶段*3*的交易博弈如下：

- 参与者：卖方和买方。
- 策略：卖方决定交付策略 $s \in \mathcal{N}$。买方决定验证策略 $v \in \{NV, V\}$和接受标准 $\theta \in [0, T+1] \cup \{NA\}$，其中 $NA$ 在 $v = NV$ 时表示不适用。
- 收益：卖方在(8)中最大化其期望收益。买方在(6)中最大化其期望收益。

### B. 交易博弈*1*的均衡

在交易博弈1中，卖方和买方决定各自的策略，以最大化各自的收益。接下来，我们将推导纳什均衡，在该均衡中，任何参与者都无法通过单方面偏离来提高个人收益。我们首先关注纯策略纳什均衡（PNE），在这种均衡中，每个参与者只选择一个策略，如定义1所述。

**定义1** (博弈1的纯策略纳什均衡)。 如果纯策略组合 $(s^*,(v^*,\theta^*))$ 中，卖方和买方在给定彼此均衡策略的情况下所作的决策都能使各自收益最大化，那么该组合就是博弈*1*的纯策略纳什均衡，即：

$$\pi_{seller}(s^*,(v^*,\theta^*)) \ge \pi_{seller}(s,(v^*,\theta^*)), \quad (9a)$$
$$\pi_{buyer}(s^*,(v^*,\theta^*)) \ge \pi_{buyer}(s^*,(v,\theta)), \quad (9b)$$

[2]作为一种数字产品，机器学习模型的复制成本可以忽略不计。因此，与每次机器学习模型交易中的隐私补偿相比，一次性成本造成的机器学习模型边际成本微不足道。

[3]请注意，只有在买方订购模型（即 $r \neq 0$）后，才会进入阶段3。

对于每个 $s \in \mathcal{N}$、 $v \in \{NV, V\}$ 和 $\theta \in [0, T+1] \cup \{NA\}$。

根据买方在第2阶段的订购决策 $r$，我们分两种情况分析均衡。

*1)* 买方下单 $M_1$ 在第*2*阶段：当买方在第2阶段订购最低质量模型（$r = 1$）时，无需对其进行验证。原因是，对于任意 $s \in \mathcal{N}$，集合$\mathcal{M}$ 中的任何交付的模型 $M_s$ 都满足买方订单的质量要求，即 $\alpha_1 \leq \alpha_s$。同时，卖方无法通过向买方交付优于$M_1$的模型来提高其收益，因此命题1中存在唯一的纯策略纳什均衡。

**命题1**。给定 $r = 1$ 在第*2*阶段，博弈*1*的唯一纯策略纳什均衡为$(s^*, (v^*, \theta^*)) = (1, (NV, NA))$。

我们在附录中给出命题1的详细证明。

*2*）买方在第*2*阶段订购 $M_r$ （$r \in \mathcal{N} \setminus \{1\}$）：当买方订购的模型优于 $M_1$，即 $r > 1$时，均衡分析需要推导出相应条件，以确定验证模型是否在经济上可行且对买方有利。为此，我们先从[4] 中展示的基本关系入手：机器学习模型质量（以泛化误差量化，泛化误差衡量模型在未见数据上的表现）与训练数据规模之间的关系。[13]–[15], 中确立的学习曲线幂律特征表明，每提升一个单位的模型质量，都需要以对数级增长的训练数据量。这种关系直接影响模型开发中的成本—质量权衡。

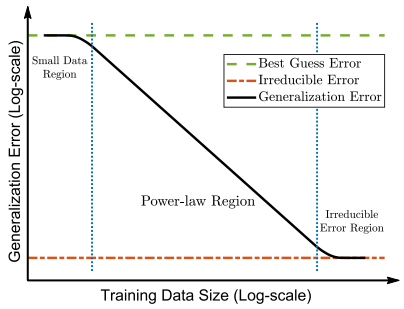


图4. 幂律学习曲线示意图

利用学习曲线的幂律特征，我们采用严格被支配策略的迭代剔除法，从而得到命题2。

**命题2**。给定 $r \in \mathcal{N} \setminus \{1\}$ 在第*2*阶段，博弈*1*的均衡中，卖方只会交付 $M_r$ 或 $M_1$ 。

我们在附录中给出了命题2的详细证明。借助命题2，我们随后深入分析排序 $M_r$ ($r \in \mathcal{N} \setminus \{1\}$)的纳什均衡。该分析考虑了 $p_r \leq U_r$ 和 $p_r > U_r$两种情况，但我们重点讨论前一种情况，因为它会带来更有趣的策略互动。在后一种情况下，由于卖方从不交付质量高于买方支付水平的模型，买方将获得负收益，因此不会购买该模型。

[4]当数据量非常小时，机器学习模型的表现与随机猜测一样好，达到最佳猜测泛化误差。相比之下，当数据量非常大时，模型会达到不可约泛化误差，即信息论误差下界 [13], , 如图4所示。

*a)* 不经济的模型验证：当针对模型欺骗进行验证不经济时，即 $C_T > p_r - U_1$，存在唯一的纯策略纳什均衡，如命题 3 所示。

**命题 3**。考虑第*2*阶段中 $r \in \mathcal{N} \setminus \{1\}$ 且 $p_r \leq U_r$的情形。当模型验证不具有经济性，即 $C_T > p_r - U_1$时，博弈*1*的唯一纯策略纳什均衡为（$s^*$、（$v^*$，$\theta^*$））=（1、（$NV$，$NA$））。

我们在附录中给出了命题 3 的详细证明。命题 3 中的不具有经济性的条件意味着，模型验证成本高于其预期收益，其中 $p_r - U_1$ 刻画了买方通过验证最多可以避免的收益损失。因此，买方不会验证交付的模型，而卖方将始终交付最低质量模型，从而导致持续的模型欺骗。

*b)* 经济可行但无效的模型验证：当模型验证在经济上可行，但无法阻止卖方进行模型欺骗时，我们在命题 4 中总结了相应的唯一纯策略纳什均衡（PNE）。

**命题 4**。考虑满足 $p_r \leq U_r$ 的第*2*阶段 $r \in \mathcal{N} \setminus \{1\}$ 。当验证在经济上可行但无效时，即

- $C_T < p_r - U_1$, *and*
- $(p_r - C_r) \cdot \delta(\theta, r) \leq (p_r - C_1) \cdot \delta(\theta, 1)$, *for any* $\theta \in \{\theta_r \in [0, T+1] : C_T < (p_r - U_1) \cdot (1 - \delta(\theta_r, 1))\}$,

博弈*1*的唯一纯策略纳什均衡为 $(s^*, (v^*, \theta^*)) = (1, (V, T+1))$。

我们在附录中给出命题 4 的详细证明。根据命题 4，当模型验证在经济上可行但无效时，即使考虑概率验证结果，卖方也总是会通过交付最低质量模型 $M_1$ 进行欺骗，因为这样能带来最高收益。作为回应，买方会在验证后拒绝交付的模型，以将交易损失降至最低，即设置 $\theta = T + 1$。

*c)* 经济可行且有效的模型验证：当模型验证在经济上可行且能有效防范模型欺骗时，我们证明不存在纯策略纳什均衡（PNE）。如果买方不验证模型，卖方就会选择模型欺骗。反之，如果买方进行验证，卖方就会选择如实交付。因此，在纯策略中，不存在相互最佳回应。这促使我们考虑混合策略纳什均衡（MNE），允许参与者在均衡状态下以概率选择多种策略。我们先在定义 2 中介绍博弈1的混合策略，随后在定义 3 中给出混合策略纳什均衡的定义。

**定义 2** (卖方和买方的混合策略)。卖方和买方的混合策略分别是向量 $\phi_s \in \Phi_S$、 $\phi_v \in \Phi_V$:

$$\phi_s \triangleq (Prob[s = n], \forall n \in \mathcal{N}), \tag{10a}$$

$$\phi_v \triangleq (Prob[v = NV], Prob[v = V]), \tag{10b}$$

其中， $\Phi_S$和 $\Phi_V$分别是集合 $\mathcal{N}$和 $\{NV, V\}$上的单纯形，即 $\sum_{s \in \mathcal{N}} Prob[s] = 1$ 和 $\sum_{v \in \{NV, V\}} Prob[v] = 1$。

**定义 3** （博弈1的混合策略纳什均衡）。如果混合策略组合 $(\phi_s^*, (\phi_v^*, \theta^*))$ 中，卖方和买方在给定彼此均衡策略的情况下所作的决策都能使各自收益最大化，那么该组合就是博弈*1*的混合策略纳什均衡，即，

$$\pi_{seller}(\phi_s^*, (\phi_v^*, \theta^*)) \geq \pi_{seller}(\phi_s, (\phi_v^*, \theta^*))), \tag{11a}$$

$$\pi_{buyer}(\phi_s^*, (\phi_v^*, \theta^*))) \geq \pi_{buyer}(\phi_s^*, (\phi_v, \theta)), \tag{11b}$$

对于每个 $\phi_v \in \Phi_V$、 $\phi_s \in \Phi_S$和 $\theta \in [0, T+1] \cup \{NA\}$。

接下来，我们推导模型验证既经济又有效时博弈1的混合策略纳什均衡。我们确定均衡接受标准的必要条件 $\Gamma_r$ （见式（12）：

$$\Gamma_r(\theta): \begin{cases} C_T < (p_r - U_1) \cdot (1 - \delta(\theta, 1)), & (12a) \\ (p_r - C_1) \cdot \delta(\theta, 1) < (p_r - C_r) \cdot \delta(\theta, r), & (12b) \\ \sum_{i=0}^{\theta-1} \binom{T}{i} \left( (\frac{\alpha_r}{1-\alpha_r})^{-\theta+i} - (\frac{\alpha_1}{1-\alpha_1})^{-\theta+i} \right) / C_T \\ + \frac{\alpha_r^{-\theta}(1-\alpha_r)^{\theta-1}}{U_r - p_r} + \frac{\alpha_1^{-\theta}(1-\alpha_1)^{\theta-1}}{p_r - U_1} \leq 0. & (12c) \end{cases}$$

利用 $\Gamma_r$，我们建立了买方的均衡判据 $\theta^*(r)$，如下所示

$$\theta^*(r) = \min\{\theta \in [0, T+1] : \Gamma_r(\theta)\}, \quad (13)$$

并在命题5中推导出模型交付和验证策略的均衡概率分布。为简单起见，我们将均衡接受概率$\delta(\theta^*(r), 1)$和 $\delta(\theta^*(r)\ r)$分别记为 $\delta^*_{r,\ 1}$ 和 $\delta^*_{r,\ r}$。

**命题5**。 考虑处于第*2*阶段且满足 $p_r \leq U_r$ 的 $r \in \mathcal{N} \setminus \{1\}$ 。当验证具有经济性且有效时，即

$$\begin{cases} C_T < (p_r - U_1) \cdot (1 - \delta(\theta, 1)), \textit{ and} & (14a) \\ (p_r - C_1) \cdot \delta(\theta, 1) < (p_r - C_r) \cdot \delta(\theta, r), & (14b) \end{cases}$$

其中， $\theta \in [0, T+1]$博弈*1*的唯一混合策略纳什均衡为：

- 卖方：以概率 $\textit{Prob}[s^* = 1]$交付 $M_1$，如(15a)所示；以概率 $1 - \textit{Prob}[s^* = 1]$交付 $M_r$。
- 买方：设定 $\theta^*(r) = \min\{\theta \in [0, T+1] : \Gamma_r(\theta)\}$，如(12)所示；以概率 $\textit{Prob}[v^* = V]$ 进行验证，如(15b) 所示；以概率 $1 - \textit{Prob}$ $v^* = V$不进行验证。

$$\begin{cases} \textit{Prob}[s^* = 1] = \frac{(1-\delta^*_{r,r})(U_r - p_r) + C_T}{(1-\delta^*_{r,1})(p_r - U_1) + (1-\delta^*_{r,r})(U_r - p_r)}, & (15a) \\ \textit{Prob}[v^* = V] = \frac{C_r - C_1}{(1-\delta^*_{r,1})(p_r - C_1) - (1-\delta^*_{r,r})(p_r - C_r)}. & (15b) \end{cases}$$

我们在附录中给出命题5的详细证明。命题5表明，验证成本 $C_T$ 上升会导致均衡接受标准 $\theta^*(r)$提高，使买方更不愿意接受经过验证的模型。这是因为验证成本提高会加大买方承担的损失，因此需要采用更严格的接受标准，以防止采用劣质模型。然而，更严格的标准加上不断上升的验证成本，可能会激励卖方实施模型欺骗。

## V. 第2阶段的最优订购决策

在考察阶段3的均衡后，本节研究买方在第2阶段的最优订购决策。

### A. 买方收益优化问题

在第2阶段，买方旨在作出最优订购决策 $r^*$，以最大化其在（6）中的收益。为强调阶段3决策对订购决策 $r$ 的依赖关系，我们将博弈1的均衡写为（$s^*$（$r$），（$v^*$（$r$） $\theta^*$（$r$））。随后，我们将买方收益优化问题表述如下：

**问题1**。 （买方收益优化问题）。

$$\max \quad \mathbb{E}[U(r, s^*(r), v^*(r), \theta^*(r)) - P(r, s^*(r), v^*(r), \theta^*(r))] - C_T \cdot \mathbf{1}_{v^*(r)=V}, \quad (16a)$$

$$var. \quad r \in \{0\} \cup \mathcal{N}. \quad (16b)$$

### B. 最优订购决策

基于阶段3的均衡分析，我们可以推导出买方在第2阶段的最优订购决策 $r^*$ 及其相关条件。具体而言，我们将围绕 $r^*$ 的讨论分为三种情况，详述如下。

*1）* $r^* \in \mathcal{N} \setminus \{1\}$ 第*2*阶段：我们明确最优订购决策 $r^*$ 选择最低质量模型 $M_1$ 之外的任何选项所需满足的条件。根据命题3—5中的纳什均衡分析，经济上不合理或无效的验证会促使卖方始终提供模型 $M_1$，从而导致买方过度支付。因此，对于订购模型 $M_n$ 的决策，必须满足（14）中的经济有效的验证，其中 $n \in \mathcal{N} \setminus \{1\}$，并由以下条件 $\Lambda_n^{EE}$ 确定：

$$\Lambda_n^{EE}: \begin{cases} C_T < (1 - \delta^*_{n,1})(p_n - U_1), & (17a) \\ \delta^*_{n,1}(p_n - C_1) < \delta^*_{n,n}(p_n - C_n). & (17b) \end{cases}$$

在条件 $\Lambda_n^{EE}$ 下，买方订购 $M_n$ 的收益为 $U_n - p_n - \Delta\pi_n$，其中 (18) 中的 $\Delta\pi_n$ 表示信息不对称造成的收益损失：

$$\Delta\pi_n = \frac{(U_n - U_1)C_T + (1 - \delta^*_{n,n})(U_n - U_1)(U_n - p_n)}{(1 - \delta^*_{n,1})(p_n - U_1) + (1 - \delta^*_{n,n})(U_n - p_n)}. \quad (18)$$

要得到最优决策 $r^* = n \in \mathcal{N} \setminus \{1\}$，模型 $M_n$ 必须为买方带来高于所有其他决策的最高收益。总体而言，我们将订购 $M_n$ 的条件总结如下： $\Lambda_n$

$$\Lambda_n: \begin{cases} \Lambda_n^{EE}, & (19a) \\ \max\{0, U_1 - p_1\} \leq U_n - p_n - \Delta\pi_n, & (19b) \\ n = \operatorname{argmax}_{m \in \mathcal{N} \setminus \{1\}} \ U_m - p_m - \Delta\pi_m. & (19c) \end{cases}$$

*2）* $r^* = 1$ 在第*2*阶段：我们考察买方何时因订购最低质量模型 $M_1$ 而获得最大收益，即 $r^* = 1$。如第五节-B1 所述， $r^*$ 等于 1，当且仅当 $U_1 - p_1$ 非负，且对于任何 $n \in \mathcal{N} \setminus \{1\}$，其他条件 $\Lambda_n$ 均不成立。为便于说明，我们将此类条件记为 $\Lambda_1$：$\{\neg\Lambda_n, \forall n \in \mathcal{N} \setminus \{1\}\} \wedge \{0 \leq U_1 - p_1\}$。

*3）* $r^* = 0$ 在第*2*阶段：当订购任何模型的收益为负时，最优订购决策是不订购。

总结上述三种可能决策的讨论，定理1总结了机器学习模型交易第2阶段买方的最优订购决策 $r^*$ 。

**定理1**。 买方的最优订购决策 $r^*$ 如下：

$$r^* = \begin{cases} n \in \mathcal{N}, & \textit{if } \Lambda_n \textit{ holds}, & (20a) \\ 0, & \textit{otherwise}. & (20b) \end{cases}$$

定理1的详细证明见附录。定理1指出了（17）中概述的关键条件 $\Lambda_n^{EE}$，以实现经济有效的验证，使买方能够避开卖方的交易陷阱。

**例 1** （交易陷阱）。 在有两个模型待售、其中$U_1 = 11$ 和 $U_2 = 1000$，对应价格分别为 $p_1 = 9$和 $p_2 = 10$的情况下， $C_T = 0.0001$的微小验证成本会使买方选择 $r^* = 1$，从而获得最大收益 $U_1 - p_1 = 2$。

例 1 似乎有悖直觉，因为 $U_2 - p_2 = 990$ 远高于 $U_1 - p_1 = 2$，也就是说，高质量模型 $M_2$对买方更具吸引力，而 $C_T$ 微乎其微。然而，如果卖方在订单为 $r = 2$ 时提供低质量模型 $M_1$，买方仍会接受该模型并支付价格 $p_2$，因为 $U_1 - p_2 = 1 > 0$。也就是说，买方永远不会验证交付的模型，因为这样做不会提高其收益（即验证不具经济性）。因此，模型验证就变成了买方从不实施的空洞威胁，卖方据此总是通过交付 $M_1$来进行模型欺骗。为避免这一陷阱，定理1在（17）式中引入了条件 $\Lambda_n^{EE}$，确保经济有效的验证，从而最大化买方收益。

### C. 买方最优收益

根据定理1中的最优订购决策 $r^*$，我们可以计算出由此产生的最优买方收益 $\pi_{buyer}^*$ 如下：

$$\pi_{buyer}^* = \begin{cases} 0, & \text{if } r^* = 0, \quad (21a) \\ U_1 - p_1, & \text{if } r^* = 1, \quad (21b) \\ U_{r^*} - p_{r^*} - \Delta\pi_{r^*}, & \text{if } r^* \in \mathcal{N} \setminus \{1\}. \quad (21c) \end{cases}$$

为研究模型验证在应对信息不对称方面的有效性，我们将买方的最优收益 $\pi_{buyer}^*$ 与完全信息基准下的收益进行比较。具体而言，在完全信息下，卖方和买方都能清楚了解机器学习模型的质量，这意味着买方购买机器学习模型所获得的收益等于模型效用与价格之差，如式（22）所示。

$$\pi_{buyer}^{benchmark} = \begin{cases} 0, & \text{if } r^* = 0, \quad (22a) \\ U_{r^*} - p_{r^*}, & \text{if } r^* \in \mathcal{N}. \quad (22b) \end{cases}$$

通过比较式（21）和式（22）中的最优收益可以发现，尽管买方在机器学习模型交易中可以采用验证策略，信息不对称仍会使其最优收益减少 $\Delta\pi_{r^*}$，如式（18）所示。这种收益损失有以下两个原因：

- 验证成本：根据命题5，当购买$M_{r^*}$ 且 $r^* \in \mathcal{N} \setminus \{1\}$ 时，买方将采用概率验证策略，以应对卖方可能实施的模型欺骗。因此，模型验证成本会降低买方的最优收益。
- 不完美验证：在不完美的模型验证下，买方可能高估或低估交付模型的质量，从而使买方的最优收益低于完全信息基准。

## VI. 第一阶段的最优定价方案

在分析第3阶段的纳什均衡和第2阶段买方的最优订购决策后，我们接下来确定卖方在阶段1的最优定价方案。

### A. 卖方收益优化问题

在阶段1，卖方决定模型定价方案$\boldsymbol{p} = \{p_n, \forall n \in \mathcal{N}\}$，以最大化（8）中的收益 $\pi_{seller}$。考虑到买方的最优订购决策 $r^*$ 取决于模型价格 $\boldsymbol{p}$，我们将其表示为 $r^*(\boldsymbol{p})$。由此，我们可以将卖方的收益优化问题表述为

问题2，其中包含一个关于问题1中买方收益优化的子问题。

**问题2**。卖方收益优化问题。

$$\max \quad \mathbb{E}[P(r^*(\boldsymbol{p})) - C_M(r^*(\boldsymbol{p}))], \quad (23a)$$
$$s.t. \quad r^* = \underset{r \in \{0\} \cup \mathcal{N}}{\operatorname{argmax}} \pi_{buyer}(r, \boldsymbol{p}), \quad (23b)$$
$$0 \le p_i \le p_j, \quad \forall i \le j, \ \forall i, j \in \mathcal{N}, \quad (23c)$$
$$var. \quad \boldsymbol{p} = \{p_n, \forall n \in \mathcal{N}\}. \quad (23d)$$

### B. 最优定价方案

为解决问题2，我们首先讨论每个机器学习模型价格的上界，并据此确定最优定价方案 $\boldsymbol{p}^*$ 及其相关条件。

*1*）模型价格上界：可行的模型价格必须满足买方的个体理性（IR）约束 [21]，即带来非负收益，从而确保其愿意购买其中一个模型。因此，我们可以推导出每个机器学习模型 $M_n$ 的价格上界，记作 $p_n$，对任意 $n \in \mathcal{N}$ 均成立，即

$$\overline{p}_n = \begin{cases} U_1, & \text{if } n = 1, \quad (24a) \\ \frac{U_n + U_1}{2} + \sqrt{\frac{(U_n - U_1)^2}{4} - \frac{(U_n - U_1)C_T}{\delta_{n,n}^* - \delta_{n,1}^*}}, & \text{if } n > 1. \quad (24b) \end{cases}$$

请注意，当且仅当$4C_T \le (\delta_{n,n}^* - \delta_{n,1}^*)(U_n - U_1)$时，买方的个体理性（IR）约束才成立；否则，（24b）中的价格上界$\overline{p}_n$ 不存在。

*2*）最优模型价格：如（24）所示，将模型价格设为上界 $\overline{p}_n$，可使卖方针对每个模型的收益最大化。为使卖方的总体收益最大化，并考虑买方的策略性选择，我们确定了在任意给定模型下实现利润最大化的条件。该分析确定了模型交易阶段1中卖方的最优定价方案 $\boldsymbol{p}^*$。

我们首先确立模型 $M_n$、$n \in \mathcal{N} \setminus \{1\}$ 的销售能够为卖方带来最高收益的条件，记为条件 $\Psi_n$，其中模型验证必须兼具经济性和有效性，具体如定理1所述。更具体地说，

$$\Psi_n : \begin{cases} 4C_T \le (\delta_{n,n}^* - \delta_{n,1}^*)(U_n - U_1), & (25a) \\ \delta_{n,1}^*(\overline{p}_n - C_1) < \delta_{n,n}^*(\overline{p}_n - C_n), & (25b) \\ 0 \le \frac{(\overline{p}_n - C_n)(\delta_{n,n}^* - \delta_{n,1}^*)(\overline{p}_n - C_1)}{(1 - \delta_{n,1}^*)(\overline{p}_n - C_1) - (1 - \delta_{n,n}^*)(\overline{p}_n - C_n)}, & (25c) \\ U_1 - C_1 \le \frac{(\overline{p}_n - C_n)(\delta_{n,n}^* - \delta_{n,1}^*)(\overline{p}_n - C_1)}{(1 - \delta_{n,1}^*)(\overline{p}_n - C_1) - (1 - \delta_{n,n}^*)(\overline{p}_n - C_n)}, & (25d) \\ n = \underset{i \in \mathcal{N} \setminus \{1\}}{\operatorname{argmax}} \frac{(\overline{p}_i - C_i)(\delta_{i,i}^* - \delta_{i,1}^*)(\overline{p}_i - C_1)}{(1 - \delta_{i,1}^*)(\overline{p}_i - C_1) - (1 - \delta_{i,i}^*)(\overline{p}_i - C_i)}. & (25e) \end{cases}$$

类似地，对于满足 (25a) 和 (25b) 的所有 $i \in \mathcal{N} \setminus \{1\}$，我们可以总结出条件 $\Psi_1$，即在该条件下，卖方销售模型 $M_1$ 所获得的收益在买方可能订购的所有模型中最高，即 $M_i$。

$$\Psi_1 : \begin{cases} 0 \le U_1 - C_1, & (26a) \\ \max_i \frac{(\overline{p}_i - C_i)(\delta_{i,i}^* - \delta_{i,1}^*)(\overline{p}_i - C_1)}{(1 - \delta_{i,1}^*)(\overline{p}_i - C_1) - (1 - \delta_{i,1}^*)(\overline{p}_i - C_i)} < U_1 - C_1, & (26b) \end{cases}$$

其中 $i \in \mathcal{N} \setminus \{1\}$：满足 (25a) 和 (25b)。

基于上述讨论，我们得出定理2中的卖方最优定价方案 $\boldsymbol{p}^*$：对于所有 $n \in \mathcal{N}$，条件$\Psi_n$互斥。

**定理2.** 卖方的最优定价方案为：

- 若 $\Psi_n$ 成立：$p_n^* = \bar{p}_n, p_m^* > \bar{p_m}, \forall n \in \mathcal{N}, \forall m \in \mathcal{N}\backslash\{n\}$。
- 若 $\Psi_n$ 不成立，则对所有 $n \in \mathcal{N}$：$p_n^* > \bar{p}_n, \forall n \in \mathcal{N}$。

我们在附录中给出了定理2的详细证明。定理2揭示了信息不对称如何影响卖方的最优定价方案。具体而言，由于信息不对称，买方会遭受收益损失 $\Delta\pi_n$，从而降低其在交易中能够承受的最高模型价格。只有在购买最低质量的模型 $M_1$时，这种负面影响才不存在，因为该模型不受模型欺骗影响，因此不会受到信息不对称的影响。因此，与完全信息 基准情形（其中最优模型价格等于买家效用）相比，信息不对称会降低卖方对机器学习模型所能收取的最优价格 $M_n$, $n \in \mathcal{N} \setminus \{1\}$，该价格低于买方相应的模型效用 $U_n$，如式（24b）所示。

### C. 卖方最优收益

根据定理2中的最优定价方案，我们可以计算出相应的卖方最优收益 $\pi^*_{seller}$ 如下：

$$\pi^*_{seller} = \begin{cases} 0, & \text{if } r^* = 0, \\ p_1^* - C_1, & \text{if } r^* = 1, \\ \frac{(p^*_{r^*} - C_{r^*})(\delta^*_{r^*,r^*} - \delta^*_{r^*,1})(p^*_{r^*} - C_1)}{(1-\delta^*_{r^*,1})(p^*_{r^*} - C_1) - (1-\delta^*_{r^*,r^*})(p^*_{r^*} - C_{r^*})}, & \text{if } r^* > 1, \end{cases} \tag{27}$$

其中 $r^* \in \{0\} \cup \mathcal{N}$。为说明信息不对称对卖方的影响，我们进一步推导完全信息基准下卖方的最优收益，即

$$\pi^{benchmark}_{seller} = \begin{cases} 0, & \text{if } r^* = 0, \\ U_{r^*} - C_{r^*}, & \text{if } r^* \in \mathcal{N}. \end{cases} \tag{28}$$

比较式(27)和式(28)中的卖方最优收益后，令人惊讶的是，尽管卖方有机会在信息不对称条件下实施模型欺骗（当 $r^* \in \mathcal{N}\backslash\{1\}$ 时），其最优收益仍低于完全信息基准下的收益。我们对这一结果背后的原因解释如下：

- 策略冲突：根据命题5，信息不对称引发的不信任促使卖方和买方都策略性地参与模型欺骗和模型验证（并采用相应的接受标准）。因此，这种策略冲突会对卖方的最优收益产生负面影响。

- 买方收益降低：信息不对称会导致买方收益降低 $\Delta\pi_n$ ，进而降低其可负担的最高模型价格。根据买方的个体理性（IR）约束，这种收益损失 $\Delta\pi_n$ 会相应降低机器学习模型交易中卖方的最优价格和收益。

总体而言，通过与完全信息基准进行比较可以发现，信息不对称导致机器学习模型交易中的卖方和买方陷入双输局面。为减轻这种负面影响，我们可以降低买方的验证成本或提高验证准确率，例如生成合成测试数据用于模型验证，从而提高买方和卖方的收益。

到目前为止，我们的分析主要集中于不包含订单信息保护（OIP）的模型交易。在观察到买方订单后，卖方可以制定模型交付策略，如图2所示。这种订单信息泄露使卖方能够采取“个性化”（即歧视性）的交易策略，从而引发了广泛关注 [18]。受此推动，下一节将探讨另一种交易情景：买方的模型订单在卖方最终确定交付策略之前处于隐藏状态。我们的目标是了解潜在的信息保护法规，例如《罗宾逊-帕特曼法案》 [19] 和《数据治理法案》 [20]，将如何影响模型交易中的买方和卖方收益。

## VII. 订单信息保护下的模型交易

本节深入探讨订单信息保护（OIP）下的模型交易。我们首先在第七节A部分正式介绍OIP下的交易流程，其中卖方在不知道买方订购决策的情况下确定模型交付策略，这不同于没有OIP时第III节A部分所述的流程。随后，我们在第七节B部分通过逆向归纳法研究交易纳什均衡，并在第七节C部分总结OIP对买方和卖方的影响。

### A. OIP下的机器学习模型交易流程

在带 OIP 的机器学习模型交易中，我们可以将交易流程描述为一个两阶段博弈，具体如下：

- **阶段1：** 卖方发布一组机器学习模型，每个模型都有相应的价格和公告质量。
- **第2阶段：** 卖方为每个模型决定交付策略。同时，买方根据接受标准做出订购决策并选择验证策略。

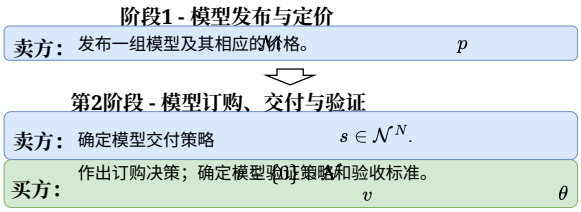



图 5. 带有订单信息保护的机器学习模型交易流程

具体而言，对于第2阶段的交易互动，我们可以进一步将其表述为博弈2中的同时博弈。

**博弈 2**（第2阶段带有订单信息保护的交易博弈）。 第2阶段带有 *OIP* 的交易博弈如下：

- 参与者：卖方和买方。
- 策略：卖方决定交付策略 $\boldsymbol{s} \in \mathcal{N}^N$。买方作出订购决策 $r \in \{0\} \cup \mathcal{N}$ 并决定验证策略 $v \in \{NV, V, NA\}$，其中$NA$ 表示在其不订购任何模型的情况下不适用。此外，买方决定接受标准 $\theta \in [0, T+1] \cup \{NA\}$，其中 $NA$表示如果 $v = NV$则不适用。

- 收益：卖方最大化(8)中的期望收益。买方最大化(6)中的期望收益。

请注意，在第2阶段，由于买方的订购决策未知，卖方针对所有模型制定交付策略 $\boldsymbol{s} \in \mathcal{N}^N$。具体而言，$\boldsymbol{s}(n) \in \mathcal{N}$ 表示针对订购模型 $M_n$实际交付的模型 $M_{\boldsymbol{s}(n)}$，其中 $n \in \mathcal{N}$。

### *B.* 带 *OIP* 的机器学习模型交易分析

接下来，我们通过逆向归纳法，研究带 OIP 的模型交易在第2阶段的均衡策略，以及第1阶段的最优定价方案。

*1)* 第*2*阶段的均衡策略：针对第2阶段的交易博弈 2，我们首先分析买方的严格劣势策略，并以此进一步推导买方和卖方的均衡策略。具体而言，我们将买方的最优订购决策 $r^*$ 分为以下几种情况，讨论相应的验证策略。

- 买方在第*2*阶段不订购任何模型（$r^* = 0$）：当买方的最优订购决策是不购买任何模型时，验证和接受标准决策均不适用，即 $(v^*, \theta^*) = (NA, NA)$。
- 买方在第*2*阶段订购 $M_r$ （$r^* \in \mathcal{N}$）：如果买方的最优订购决策是向卖方购买机器学习模型，那么相应的买方收益非负。因此，相应的验证策略是不验证模型：$(v^*, \theta^*) = (NV, NA)$，从而避免因不完美验证产生额外成本和损失。

根据上述分析，我们可以得出买方的劣势验证决策，概括如命题 6 所示。

**命题 6.** 在交易博弈 *2* 的均衡中，买方不会验证交付的机器学习模型。

命题 6 的详细证明见附录。根据命题 6 中买方的严格劣势策略，对每个 $r \in \mathcal{N}$，卖方的最佳响应是 $\boldsymbol{s}^*(r) = 1$，即始终交付最低质量模型 $M_1$。

总而言之，买方将在第2阶段专注于购买最低质量模型 $M_1$。如果 $U_1 - p_1 \geq 0$，即买方在不验证的情况下购买 $M_1$ 所获得的收益非负，那么博弈 2 的唯一纯策略纳什均衡是 $(\boldsymbol{s}^*(r^*), (r^*, v^*, \theta^*)) = (1, (1, NV, NA))$。相反，如果 $U_1 - p_1 < 0$，买方将不会从卖方处购买任何模型。我们在定理 3 中总结交易博弈 2 的均衡。

**定理 3.** 在带 *OIP* 的机器学习模型交易的第*2*阶段，若$U_1 \geq p_1$，博弈 *2* 的唯一纯策略纳什均衡为 $(\boldsymbol{s}^*(r^*), (r^*, v^*, \theta^*)) = (1, (1, NV, NA))$。否则，买方不会向卖方购买任何机器学习模型。

我们在附录中给出定理 3 的详细证明。定理 3 表明，OIP 会导致机器学习模型交易仅在最低质量模型 $M_1$ 上成功。具体而言，由于 OIP，买方可以通过预测卖方的交付策略来避免信息不对称造成的损失，而无需进行模型验证。然而，这种做法适得其反，使模型验证无法有效威慑卖方。因此，模型欺骗必然发生，模型 $M_1$ 成为买方唯一可行的选择。

*2*）阶段*1*中的最优定价方案：鉴于第2阶段的纳什均衡，我们确定卖方在阶段1的最优定价方案，如命题 7 所示。

**命题 7.** 在带 *OIP* 的机器学习模型交易的阶段*1*，卖方的最优定价方案为 $p_1^* =$ 最大值$\{U_1, C_1\}$ 和$p_n^* = C_n$，对所有 $n \in \mathcal{N} \setminus \{1\}$ 均成立。

我们在附录中给出命题 7 的详细证明。根据定理 3 和命题 7，我们可以推导出买方在带 OIP 的模型交易中的最优收益 $\pi_{buyer}^*$。

$$\pi_{buyer}^* = \begin{cases} U_1 - p_1^*, & \text{if } U_1 \geq p_1^*, \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

同样，在 OIP 下，卖方的最优收益 $\pi_{seller}^*$ 仅与最低质量模型 $M_1$有关，如式 (30) 所示。

$$\pi_{seller}^* = \begin{cases} p_1^* - C_1, & \text{if } U_1 \geq p_1^*, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

### *C.* 订单信息保护的影响

在分析带 OIP 的机器学习模型交易后，我们探讨 OIP 对买方和卖方策略及收益的影响。

*1*）对买方的影响：我们首先考察 OIP 如何影响买方的收益和均衡交易策略。

**观察 1**。*OIP* 不会增加买方的收益。

鉴于在没有 OIP 的模型交易中，卖方对买方采取歧视性交付策略，观察 1 似乎有些反直觉。然而，尽管在没有 OIP 的情况下，买方不可避免地会因信息不对称而遭受损失，但策略性模型验证可以缓解这种损失，并促进卖方诚信。相反，在采用 OIP 的情况下，买方节省了验证成本，并避免了不完美的验证结果所造成的损失。不过，这会使模型验证失效，并持续导致模型欺骗。因此，OIP 会降低买方在机器学习模型交易中的收益。

*2*）对卖方的影响：接下来，我们研究 OIP 对模型交易中卖方策略和收益的影响。

**观察 2.** *OIP* 不会提高卖方的收益。

关于观察 2，详细解释如下：没有 OIP 时，买方会以一定概率验证交付的模型，从而促使卖方采取概率性欺骗策略。相比之下，有 OIP 时，买方从不进行模型验证，导致卖方始终实施模型欺骗。在这种情况下，买方有 OIP 时只对最低质量模型感兴趣，这使卖方无法成功出售质量更高的模型，而这些模型本可带来更高的收益。

总之，OIP 对买方和卖方都没有任何好处。尽管 OIP 能保护买方免受卖方歧视性定价和策略制定的影响，但最终却损害了模型交易中各方的利益。

在分析了有 OIP 和没有 OIP 的机器学习模型交易后，我们接下来将目光转向一个更加多元化的市场，在这个市场中，不同买方可能会对同一模型赋予不同的效用。例如，商业公司通常比业余人士从机器学习模型中获得更高的效用，因为它们可以部署模型

以提升服务质量和收益 [39]。卖方可以针对不同买方调整交付策略，以实现收益最大化，但模型价格必须对所有买方保持一致，以避免消费者不满 [40]。这一情况促使我们设计一种定价方案，以优化在信息不对称和异质买方效用背景下模型市场中卖方的收益。

## VIII. 异质买家的定价方案

本节首先介绍异质买方效用下卖方的收益优化问题。随后，我们分别推导 ML 模型市场中有 OIP 和无 OIP 时的最优定价方案。

### A. 异质买方效用下的卖方收益优化问题

鉴于买家模型效用存在异质性，卖方旨在设计一种统一的定价方案，以最大化其期望收益。为避免维数灾难，并在这一初步尝试中获得更清晰的见解，我们采用数字产品中常见的二元选择设置 [41]–[43]，例如 Adobe [44]和 ChatGPT [45] 的基础版与高级版分级产品。具体而言，卖方将提供两种机器学习模型供销售：低质量模型 $M_1$ 和高质量模型 $M_2$。我们进一步假设，卖方能够获取买方对这两种模型的效用分布（例如通过市场调查 [4], [6], [46]），即模型 $M_1$的效用 $U_1$ 和模型 $M_2$ 的效用 $U_2$ 。我们采用联合概率密度函数 $f(U_1, U_2)$ 来表示买方对这两种模型的异质效用。这个简化的设定有助于深入理解现实中的机器学习模型市场。

在每次模型交易中，买方都会优化订购决策 $r$ ，以最大化其收益；该收益取决于模型效用 $U_1$ 和 $U_2$，以及卖方的定价方案 $\boldsymbol{p}$。因此，我们将买方的最优订购决策表示为 $r^*(\boldsymbol{p}, U_1, U_2)$，并将异质买方效用下的卖方收益优化问题表述为问题3。

**问题3**。 异质买方效用下的卖方收益优化问题。

$$\max \quad \int_{U_1}^{\infty}\int_{0}^{\infty} f(U_1,U_2)\cdot \mathbb{E}[P(r^*(\boldsymbol{p},U_1,U_2)) - C_M(r^*(\boldsymbol{p},U_1,U_2))]\,dU_1dU_2, \tag{31a}$$

$$s.t. \quad r^* = \underset{r\in\{0,1,2\}}{\operatorname{argmax}}\ \pi_{buyer}(r,\boldsymbol{p},U_1,U_2), \tag{31b}$$

$$0 \le p_i \le p_j, \quad \forall i \le j,\ \forall i,j \in \{1,2\}, \tag{31c}$$

$$var. \quad \boldsymbol{p} = \{p_1, p_2\}. \tag{31d}$$

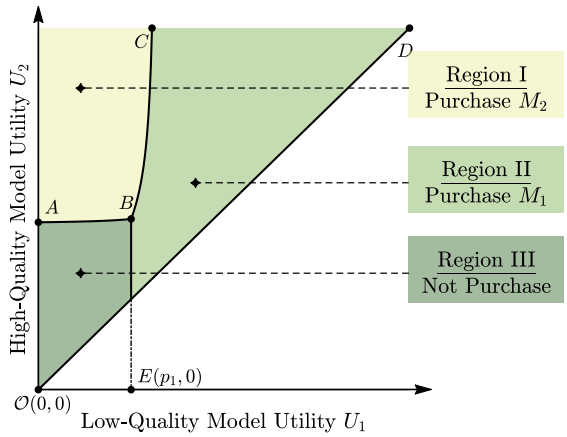
### B. 无 OIP 的机器学习模型市场

引入问题3后，我们如下探讨不含 OIP 的模型市场。具体而言，我们首先对不含 OIP 的问题3进行理论分析。该问题本质上是一个具有挑战性的双层多变量优化问题（即使在线性实例中也属于强 NP难问题 [47]）。随后，我们利用第 IV-V 节中建立的机器学习模型价格与买方最优订购决策之间的关系，设计一种用于求解问题3的高效算法。

*1*）理论分析：在不含 OIP 的模型市场中，卖方针对异质买方制定差别化交付策略。根据定理1，图6展示了买方的最优订购决策随模型效用异质性变化的分布情况。

在图6中，直线 $OD$ 表示买方对高质量模型 $M_2$的最低效用值，其中 $U_2 = U_1$。曲线$AB$ 和 $BC$ 表示 $U_1$ 和 $U_2$ 的效用阈值，买方的最优决策将在此发生转变：

- 曲线 $AB$：从购买模型 $M_2$ 转变为不购买。
- 曲线 $BC$：从购买模型 $M_2$ 转变为购买模型 $M_1$。




AB： $U_2 = (p_2^2 - (\frac{C_T}{\delta^*_{2,2}-\delta^*_{2,1}} + p_2)U_1)/(p_2 - U_1 - \frac{C_T}{\delta^*_{2,2}-\delta^*_{2,1}})$.

BC： $U_2 = U_1 + \frac{(\delta^*_{2,2}-\delta^*_{2,1})(p_2-p_1)(U_1-p_2)}{C_T+(1-\delta^*_{2,1})(U_1-p_1)-(\delta^*_{2,2}-\delta^*_{2,1})(p_2-p_1)}$.

最优订购决策：
$U_2 = U_1$.

图6. 异质效用下买方的最优订购决策

买方的个体理性（IR）约束意味着需要进行经济可行的模型验证，详见第 VI-B 节。当验证有效时，图6中的区域 I 变为可行，即 $(\delta^*_{2,2}C_2 - \delta^*_{2,1}C_1) / (\delta^*_{2,2} - \delta^*_{2,1}) < p_2$。

我们的分析揭示了模型价格对买方最优订购决策的两种不同影响：

- 单调影响 $p_1$：低质量模型的价格$p_1$ 呈现出直接的影响。 $p_1$ 增加会使曲线 $BC$ 向右移动，这表明买方购买模型 $M_1$ 的可能性降低，而更倾向于选择质量更高的模型 $M_2$。
- 反直觉效应 $p_2$：高质量模型的价格会产生更复杂的影响。有趣的是，随着 $p_2$ 增加，一些买方从模型 $M_2$获得的收益更高，因此会将最优订购决策从 $M_1$ 转向 $M_2$。

通过命题5和式（18）中的买方收益损失$\Delta\pi_2$ ，我们解释了高质量模型价格 $p_2$影响买方收益损失的原因。具体而言，较高的价格$p_2$ 会缓解买卖双方之间的策略冲突（即模型验证与欺骗），从而减少信息不对称给买方带来的收益损失。因此，一旦提高价格 $p_2$所带来的收益损失改善超过价格增幅，买方的购买收益$M_2$ 就会增加。我们推导出了关于模型价格的条件

$p_2$ 以及式（32）中的买方模型效用，在此条件下，买方的购买收益 $M_2$ 将随 $p_2$增加。

$$(\delta^*_{2,2}-\delta^*_{2,1})p_2^2-2(\delta^*_{2,2}U_2-\delta^*_{2,1}U_1)p_2+(1-\delta^*_{2,1})U_1^2 \le (1-\delta^*_{2,2})U_2^2+(C_T-2p_2)(U_2-U_1). \quad (32)$$

基于上述分析，价格 $p_2$ 引发的买方最优订购决策变化，增加了问题3中双层优化问题的复杂性。这种变化导致非凸性，使得推导闭式解变得困难。因此，接下来我们将重点转向为问题3设计一种高效的算法。

*2*）算法设计：在处理此类双层非凸优化问题时，受有限元方法 [48]–[50], 的启发，我们将问题3分解为更易处理的子问题，从而设计出算法1。具体而言，我们将模型价格 $p_1$ 与最优订购决策之间的单调相关性纳入算法1。与穷举搜索相比，这种整合将搜索空间缩减至一维，从而提高了计算效率。算法1的详细步骤如下：

算法*1*的工作流程：不失一般性地，我们首先将 $U$ 定义为买方模型效用分布的效用上界。然后，对于每个可行价格 $p_2$ （第2行），只要该价格确保购买模型$M_2$的买方获得非负收益，我们便为 $p_1$ 确定两个阈值价格：

$$\begin{cases}\eta_e(p_2)=\frac{(\delta^*_{2,2}-\delta^*_{2,1})(p_2-C_2)p_2+(1-\delta^*_{2,1})(C_2-C_1)C_1}{(1-\delta^*_{2,1})(p_2-C_1)-(1-\delta^*_{2,2})(p_2-C_2)}, & (33a)\\ \eta_h(p_2)=\frac{p_2^2-(p_2-C_T/(\delta^*_{2,2}-\delta^*_{2,1}))U}{p_2-U+C_T/(\delta^*_{2,2}-\delta^*_{2,1})}. & (33b)\end{cases}$$

具体而言，我们分别从卖方和买方的角度推导出式（33）中的阈值价格 $\eta_e(p_2)$和 $\eta_h(p_2)$（第3行）。给定高质量模型的价格 $p_2$，$\eta_e(p_2)$表示 $p_1$ 的阈值价格，在该价格下，卖方销售任一模型所获得的收益相同。相比之下，$\eta_h(p_2)$是 $p_1$ 的阈值价格，对于任何购买 $M_2$并获得非负收益的买方而言，该价格会使最优订购决策落入图6中的区域 I。

根据式（33）中的阈值价格，我们确定了不同价格 $p_2$ 下卖方的最优价格 $p_1$ （第4–10行）。我们根据模型交易中的买方收益，将分析分为以下两种情景。

情景*1*：买方从两种模型中均可获得非负收益。为了优化价格 $p_1$，我们分析卖方期望收益对价格 $p_1$ 变化的响应，其中该价格相对于阈值 $\eta_e$的差值记为 $\Delta p_1=p_1-\eta_e$。通过比较阈值价格 $\eta_e(p_2)$和 $\eta_h(p_2)$，我们量化期望卖方收益随 $\Delta p_1$的变化，如式（34）所示。

$$\begin{cases}F_1=\Delta p_1(\int_{\eta_e+\Delta p_1}^{\eta}\int_{U_1}^{B} f dU_2dU_1+\int_{\eta}^{U}\int_{U_1}^{U} f dU_2dU_1)\\ \quad -(\eta_e-C_1)\int_{\eta_e}^{\eta_e+\Delta p_1}\int_{U_1}^{A} f dU_2dU_1, & (34a)\\ F_2=\Delta p_1(\int_{\eta}^{\eta_h}\int_{A}^{U} f dU_2dU_1+\int_{\eta_e}^{U}\int_{U_1}^{U} f dU_2dU_1\\ \quad +\int_{\eta_e+\Delta p_1}^{\eta}\int_{A}^{B} f dU_2dU_1)+(\eta_e+\Delta p_1-C_1)\\ \quad (\int_{\eta_e+\Delta p_1}^{\eta_h}\int_{U_1}^{A} f dU_2dU_1+\int_{\eta_h}^{\eta_e}\int_{U_1}^{U} f dU_2dU_1), & (34b)\end{cases}$$

其中，$A(U_1)$ 和 $B(U_1,\ \Delta p_1)$ 分别表示曲线函数（以 $U_2$ 为因变量） $AB$ 和 $BC$，见

**算法1：** 异质买家效用下的卖方收益优化

**输入：** 模型成本 $C_1$ 和 $C_2$，效用上界为 $U$，验证成本 $C_T$，接受概率 $\delta^*_{2,1}$ 和 $\delta^*_{2,2}$,以及步长 $\alpha$。

**输出：** $\boldsymbol{p}^*=\{p_1^*,p_2^*\}$的最优定价方案以及由此产生的期望收益 $\pi^*_{seller}$。

1 $\pi^*_{seller}=0$;

2 $p_2=\delta^*_{2,2}C_2-\delta^*_{2,1}C_1+\alpha$; **do**

3 计算（33）中的 $\eta_e(p_2)$和 $\eta_h(p_2)$;

4 **若** $\eta_e(p_2)\le\eta_h(p_2)$

5 $F(\Delta p_1)\leftarrow F_1(\Delta p_1)$ 在 (34a) 中;

6 **else**

7 $F(\Delta p_1)\leftarrow F_2(\Delta p_1)$ 在 (34b) 中;

8 $\Delta p_1=\{x\in[C_1-\eta_e,\eta_h-\eta_e]:\ F'(x)=0\}$;

9 $\hat{}p_1=\{x\in(\eta_h,\ U):\ G'(x)=0\}$;

10 $p_1=$取最大值$_{p_1\in\{\eta_e+\Delta p_1,\ \hat{}p_1\}}\pi_{seller}(p_1,p_2)$;

11 **如果** $\pi_{seller}(p_1,\ p_2)>\pi^*_{seller}$ **则**

12 $\boldsymbol{p}^*=(p_1,\ p_2)$;

13 $\pi^*_{seller}=\pi_{seller}(p_1,\ p_2)$;

14 返回 $\boldsymbol{p}^*$ 和 $\pi^*_{seller}$;

图6。此外，$\eta$ 对应图6中的点 $C$，该点与效用上界 $U$ 相关。

情景*2*：买方只能从其中一个模型中获得非负收益。在这种情况下，无论对买方购买模型 $M_2$ 有何影响，卖方都会在范围 $(\eta_h,\ U)$ 内优化价格$p_1$。因此，我们重点关注卖方销售模型 $M_1$ 所获得的期望收益，如式 (35) 所示。

$$G(p_1)=(p_1-C_1)\int_{p_1}^{U}\int_{U_1}^{U} f(U_1,U_2)dU_2dU_1. \quad (35)$$

经过前述分析，我们通过计算式 (34) 和 (35) 中函数的临界点（第 8–9 行），确定最优定价方案的候选集合。通过迭代更新使卖方期望收益最大化的定价方案（第 10–13 行），算法1输出卖方的最优定价方案及相应的期望收益。

### *C.* 采用 *OIP* 的机器学习模型市场

接下来，我们深入探讨异质买方效用下采用 OIP 的机器学习模型市场。根据定理 3，由于 OIP，在模型交易中，低质量模型 $M_1$ 是买方考虑购买的唯一选项。因此，卖方旨在确定能够使其期望收益 $\pi_{seller}(p_1)$ 最大化的最优价格 $p_1^*$ ，如式 (36) 所示。

$$\pi_{seller}(p_1)=(p_1-C_1)\int_{p_1}^{\infty}\int_{0}^{\infty} f(U_1,U_2)dU_2dU_1. \quad (36)$$

推导 $p_1^*$ 的方法与在式（35）中优化 $G(p_1)$ 的方法类似，其中闭式解取决于买方模型效用的具体分布。例如，当模型效用的概率密度函数

$f(U_1,\ U_2)$ 在区间 $[0, U]$ 内服从均匀分布时，最优价格为 $p_1^* = (2C_1 + U)/3$。因此，我们得出结论：卖方的最优定价方案为 $p_1^* = \text{argmax}_{p_1}(p_1 - C_1) \int_{p_1}^{\infty} \int_0^{\infty} f(U_1,\ U_2) dU_2 dU_1$ 和 $p_2^* = C_2$。

## IX. 数值结果

本节给出数值结果，以评估我们对机器学习模型交易的均衡分析。具体而言，我们展示不同的系统参数和机器学习模型交易场景如何影响卖方和买方的收益。

### A. 实验设置

*1*）卖方：为更好地刻画实际的机器学习模型市场，我们在研究中采用广泛使用的 Resnet-18 模型 [51] 和 CIFAR-10 [52] 数据集，其中包含 60，000 张物体彩色图像。为简化起见，我们关注一组待售的 $N = 5$ 个模型，每个模型的质量各不相同。这些模型的训练数据规模各异，较大的数据集可带来更高的模型质量，如表 I 所示。我们进一步将 $Q_n$ 记为 $M_n$的模型质量，其中任意$n \in \mathcal{N}$均适用；该质量与测试准确率相关，例如 $Q_5$ =95。29%。此外，我们将模型成本设定为训练数据规模的 1%，例如 $C_5 = 500$。

表 I 不同训练数据规模下的机器学习模型

| 机器学习模型 | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ |
|---|---|---|---|---|---|
| **训练数据规模** | 10000 | 20000 | 30000 | 40000 | 50000 |
| **准确率** $Q_n$ （%） | 85.90 | 92.05 | 93.84 | 94.97 | 95.29 |
| **成本** $C_n$ | 100 | 200 | 300 | 400 | 500 |
| **凹效用** $U_n$ | 120 | 244 | 261 | 271 | 273 |
| **凸效用** $U_n$ | 120 | 309 | 435 | 531 | 560 |

*2)* 买方：我们探讨关于模型质量 $Q_n$ 的两种不同类型的买方效用函数：凹函数和凸函数。我们将最低质量模型$M_1$ 的效用初始化为 120，并将买方效用函数定义如下：

- 凹函数： $U_n = 500(Q_n - Q_1)^{0.5} + 120$,
- 凸函数： $U_n = 50000(Q_n - Q_1)^2 + 120$。

### B. 不同情形下的买方收益

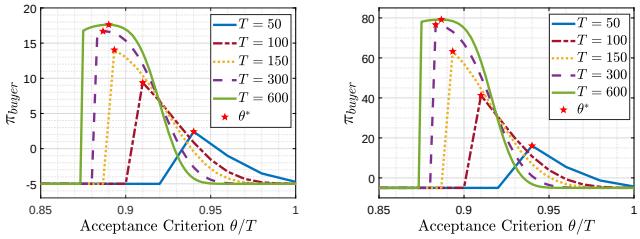
我们首先评估不同情形下买方的最优收益。为便于说明，我们采用比例定价方案，其中$p_n = 1.1C_n$、 $n \in \mathcal{N}$。

*1*）接受标准和测试数据规模的影响：为研究在没有 OIP 的模型交易中，接受标准 $\theta$ 和测试数据规模$T$ 如何影响买方收益，我们以图7中的验证成本 $C_T = 5$ 和订购决策 $r = 2$ 为例进行说明。具体而言，我们首先研究接受标准 $\theta$ 对买方收益的影响。在两个子图中，买方均在接受标准

$\theta^*$下获得最大收益，这与命题5一致。

接下来，我们说明测试数据规模 $T$ 对买方的影响。从图7可以看出，买方的最优收益会随着测试数据规模 $T$ 的增加而提高，如观察3所述。这是因为，较大的测试数据规模 $T$ 能够提高买方通过验证判断模型质量的准确性。

**观察3**。 在没有 *OIP* 的模型交易中，买方的最优收益会随着测试数据规模的增加而提高。



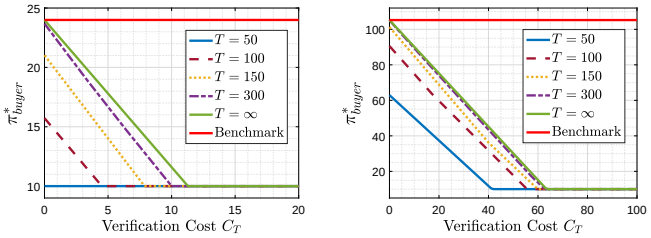

（a）凹效用下的买方收益 （b）凸效用下的买方收益 图7 不同接受标准下的买方收益

此外，正如观察4所示，测试数据规模 $T$ 越大，买方就越会降低最优接受标准 $\theta^*$，以最大化其收益。换言之，随着 $T$ 增加，买方在模型验证中会接受更大的质量波动。其背后的原因同样是，较大的测试数据规模会降低因不完美验证而高估或低估交付的模型质量的概率。

**观察4**。 随着测试数据规模增加，买方在不使用 *OIP* 的模型验证中会降低接受标准 $\theta^*$。

*2*）验证成本效益的影响：在图8中，我们探讨了验证成本效益对买方最优收益的影响，并将其与完全信息基准进行比较。如两个子图所示，买方的最优收益随验证成本 $C_T$ 的增加而下降，这与我们在第五节对不含 OIP 的模型交易所做的分析一致。我们将这一结果总结为观察5。

**观察5**。在不含 *OIP* 的模型交易中，买方的最优收益随验证成本的增加而下降。

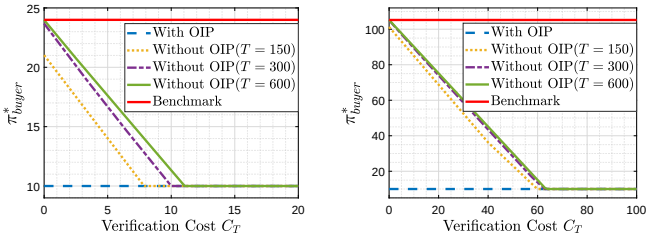



（a）凹效用下的买方收益 （b）凸效用下的买方收益 图8. 不同验证成本下的最优买方收益

此外，从图8可以看出，测试数据规模越大，买方购买高质量模型时能够承受的 $C_T$ 越高，详见观察6。 $C_T$ 的阈值（即买方收益达到 10 时的值）会随着测试数据规模增大而上升；超过该阈值后，买方会选择最低质量模型。

**观察6.** 随着测试数据规模增加，在没有 *OIP* 的模型交易中，买方能够承担更高的验证成本。

*3*）订单信息保护的影响：图9展示了订单信息保护（OIP）对买方最优收益的影响。在两个子图中，OIP都会使买方收益维持在最低水平且保持不变。相比之下，在没有 OIP 的情况下，更具成本效益的模型验证（即验证成本更低、测试数据更多）能显著提高买方的最优收益。特别是，当没有 OIP 时，验证成本为 $C_T = 0$，仅需几百条测试数据，例如 $T = 300$，即可使最优买方收益几乎




(a) 凹效用下的买方收益 (b) 凸效用下的买方收益 图9. 有无 OIP 时的最优买方收益

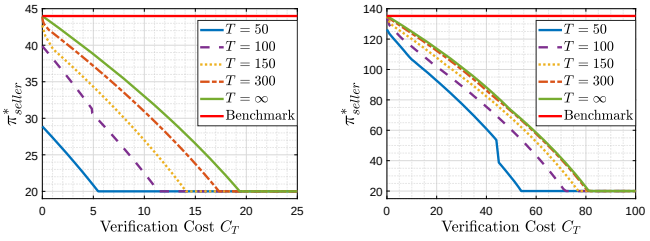
等同于完全信息基准情形下的收益。我们将上述观察总结为观察7。

**观察7.** *OIP* 会导致恒定买方最优收益最低。相反，在没有 *OIP* 时，买方的最优收益会随着验证成本效益的提高而增加，最终与完全信息基准一致。

### *C.* 不同情形下的卖方收益

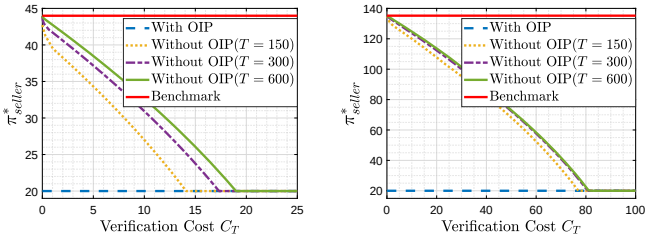
接下来，我们探讨采用定理2中提出的定价方案时，卖方在不同情形下的最优收益，并在观察8中总结实验结果。

*1)* 验证成本效益的影响：图10显示，采用我们提出的最优定价方案时，卖方收益随着测试数据规模的增大而增加，随着验证成本的增加而降低，这与我们在第六节-C中的讨论一致。




(a) 凹效用下的卖方收益 (b) 凸效用下的卖方收益 图10. 不同验证成本下的卖方最优收益

*2)* 订单信息保护的影响：图11显示，除买方购买最低质量模型的情况外（不受信息不对称影响），OIP 会降低卖方的最优收益；在后一种情况下，最终的最优收益始终为20。




(a) 凹效用下的卖方收益 (b) 凸效用下的卖方收益 图11. 有无 OIP 时的卖方最优收益

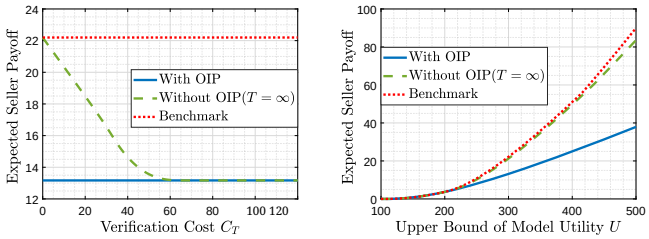
**观察8。** *OIP*使卖方的最优收益达到最低且保持恒定。相反，在没有*OIP*的情况下，卖方的最优收益会随着验证成本效益的提高而增加，最终与完全信息基准一致。

### *D.* 异质买方效用下的卖方期望收益

我们进一步考察买方效用异质时卖方的期望收益，以评估我们在第八节提出的定价方案。具体而言，我们假设买方模型效用在区间 $[0, U]$内服从均匀分布。对于其他类型的模型效用分布，所得见解相同。

*1)* 订单信息保护和验证成本效益的影响：图12(a)展示了 OIP 和验证成本效益对异质买方效用下卖方期望收益的影响，其中 $U = 300$。与观察8类似，在没有 OIP 时，验证成本效益更高也会增加异质买方效用下卖方的期望收益。同时，OIP 会使卖方期望收益达到最小的恒定值。

*2*）买方效用分布的影响：图12(b)研究了在验证成本固定为 $C_T = 5$ 时，买方效用分布对卖方期望收益的影响。随着效用分布上界 $U$ 的增加，表明市场中拥有更高模型效用的买方数量增加，所有情形下（即有 OIP、无 OIP 以及完全信息基准）的卖方期望收益也随之增加。同时，图12(b)中三条曲线之间的差距随着 $U$ 的增加而扩大。值得注意的是，我们所提出的定价方案在无 OIP 情形下实现的卖方期望收益几乎等同于完全信息基准，且远高于有 OIP 情形下的收益。只有当效用分布上界不够高（即 $U \leq 200$），无法使卖方通过销售高质量模型获得正收益时，三条曲线所对应的卖方期望收益才会趋于一致。这是因为在这种情况下，卖方通过仅销售最低质量模型来最大化其收益，而该结果不受信息不对称的影响。




(a) 不同验证成本下的卖方收益 验证成本

(b) 不同模型效用分布上界下的卖方收益 模型效用分布的上界

图 12. 买家效用服从均匀分布时的卖方最优收益

## X. 结论

我们的研究首次系统地研究了信息不对称下的机器学习模型交易。通过采用被支配策略消除并利用学习曲线的幂律特征，我们发现，准确且经济高效的验证能够减少信息不对称，从而使买方和卖方双方受益。我们还研究了订单信息保护（OIP）如何通过消除歧视性交易策略，出人意料地降低买方和卖方的收益。

为了优化卖方在有无 OIP 时的期望收益，我们引入了一种针对不同买方效用量身定制的定价策略，以近似完全信息基准。未来的研究将把这种定价策略扩展到多个

模型，并改进我们的模型，以纳入复杂的成本—性能相关性，突破学习曲线的幂律特征。这些进展将使我们的发现更具实用性和广泛适用性，尽管这会使信息不对称情况下的均衡分析和定价策略更加复杂。

## 致谢

本研究得到中国国家自然科学基金（项目编号：62271434、12371519、61771013）、中央高校基本科研业务费、深圳市群体智能赋能低碳能源网络重点实验室（编号：ZDSYS20220606-100601002）、深圳市稳定支持科技计划2023、深圳市人工智能与机器人研究院，以及深圳市龙岗区支持创新项目“十大行动计划”（编号：LGKCSDPT2024002）的资助。

## 参考文献

[1] X. Li、J. Huang、K. Yang 和 C. Fan，《信息不对称下的机器学习模型交易》，载于*IEEE*国际通信会议，2023年，第3320–3326页。[2] “Modzy——领先的机器学习生产平台”，https://www.modzy.com/，访问日期：2024年3月26日。[3] “StreamML”，https://stream.ml/，访问日期：2024年3月26日。[4] L. Chen、P. Koutris 和 A. Kumar，《迈向数据市场中的机器学习基于模型的定价》，载于国际数据管理会议，2019年，第1535–1552页。[5] A. Agarwal、M. Dahleh 和 T. Sarkar，《数据市场：一种算法解决方案》，载于*ACM*经济与计算会议，2019年，第701–726页。[6] J. Liu、J. Lou、J. Liu、L. Xiong、J. Pei 和 J. Sun，《Dealer：具有差分隐私的端到端模型市场》，*VLDB*基金会，第14卷，第6期，第957–969页，2021年。[7] P. Sun、X. Chen、G. Liao 和 J. Huang，《利润最大化的模型市场：采用差分隐私的联邦学习》，载于*IEEE*计算机通信会议，2022年，第1439–1448页。[8] A. Minin 和 S. Fedoseev，《开发者声誉风险水平评估》，载于控制系统、数学建模、自动化与能源效率国际会议，2022年，第292–295页。[9] G. A. Akerlof，《“柠檬”市场：质量不确定性与市场机制》，载于经济学精要读本。英国麦克米伦教育，1995年，第175–188页。[10] “出售人工智能模型赚钱”，https://aimodelplace.com/，访问日期：2024年3月26日。[11] S. Kramer-Schadt、J. Niedballa、J. D. Pilgrim 等，《校正最大熵物种分布模型抽样偏差的重要性》，《多样性与分布》，第19卷，第11期，第1366–1379页，2013年。[12] T. Hastie、R. Tibshirani、J. H. Friedman 和 J. H. Friedman，《统计学习的要素：数据挖掘、推断与预测》。施普林格，2009年，第2卷。[13] J. Hestness、S. Narang、N. Ardalani、G. Diamos、H. Jun、H. Kianinejad、M. M. A. Patwary、Y. Yang 和 Y. Zhou，《深度学习规模扩展在经验上具有可预测性》，*arXiv*预印本 *arXiv:1712.00409*，2017年。[14] S.-i. Amari，《学习曲线普适定理》，《神经网络》，第6卷，第2期，第161–166页，1993年。[15] S.-i. Amari、N. Fujita 和 S. Shinomoto，《四种学习曲线》，《神经计算》，第4卷，第4期，第605–618页，1992年。[16] M. Simon，《Orbitz 是否在引导 Mac 用户选择价格更高的酒店？》，*CNN*，2012年6月。[17] “Zales.com 泄露客户数据，姐妹公司 Jared 和 Kay Jewelers 在2018年也发生过类似事件”，https://krebsonsecurity.com/2021/10/zales-com-leaked-customer-data-just-like-sister-firms-jared-kay-jewelers-did-in-2018/，访问日期：2024年3月26日。[18] L. Gunderson，《亚马逊的“动态”价格引发不满》，《俄勒冈人报》，2012年。[19] H. Hovenkamp，《罗宾逊-帕特曼法与竞争：尚未完成的工作》，《反垄断法律杂志》，第68卷，第125页，2000年。[20] L. Specht-Riemenschneider 和 M. Hennemann 编，《数据治理法案：逐条评注》。英格兰牛津：Hart Publishing，2025年。[21] A. Mas-Colell、M. D. Whinston 和 J. R. Green，《微观经济理论》。牛津大学出版社，纽约，1995年，第1卷。[22] F. Liang、W. Yu、D. An、Q. Yang、X. Fu 和 W. Zhao，“大数据市场综述：定价、交易与保护”，《IEEE Access》，第6卷，第15 132–15 154页，2018年。[23] S. A. Azcoitia 和 N. Laoutaris，“数据市场及其商业模式综述”，SIGMOD Record，第51卷，29，2022年3月，18号，第–页。[24] J. Zhang、Y. Bi、M. Cheng、J. Liu、K. Ren、Q. Sun、Y. Wu、Y. Cao、R. C. Fernandez、H. Xu、R. Jia、Y. Kwon、J. Pei、J. T. Wang、H. Xia、L. Xiong、X. Yu 和 J. Zou，“数据市场综述”，arXiv预印本 arXiv:2411.07267，2024年。[25] A. Ghosh 和 A. Roth，“拍卖隐私”，载于 ACM电子商务会议，2011年，第199–208页。[26] N. Hynes、D. Dao、D. Yan、R. Cheng 和 D. Song，“Sterling演示：一种隐私保护数据市场”，VLDB基金会，第11卷，第12期，第2086–2089页，2018年。[27] A. Muschalle、F. Stahl、A. L¨oser 和 G. Vossen，“数据市场的定价方法”，载于《实现实时商业智能》。施普林格，柏林海德堡，2013年，第129–144页。[28] R. Cummings、K. Ligett、A. Roth、Z. S. Wu 和 J. Ziani，“准确性待售：在方差约束下聚合数据”，载于理论计算机科学创新会议，2015年，第317–324页。[29] C. Niu、Z. Zheng、F. Wu、S. Tang、X. Gao 和 G. Chen，“释放隐私价值：交易私有相关数据的聚合统计数据”，载于 ACM知识发现与数据挖掘国际会议，2018年，第2031–2040页。[30] H. Yu 和 M. Zhang，“基于数据质量的数据定价策略”，《计算机与工业工程》，第112卷，第1–10页，2017年。[31] S. A. Azcoitia 和 N. Laoutaris，“先试后买：适用于现实世界数据市场的实用数据购买算法”，载于 CoNext 的 ACM数据经济研讨会，2022年，第27–33页。[32] P. Koutris、P. Upadhyaya、M. Balazinska、B. Howe 和 D. Suciu，“Querymarket演示：在线数据市场的定价”，VLDB基金会，第5卷，第12期，第1962–1965页，2012年。[33] P. Koutris、P. Upadhyaya、M. Balazinska、B. Howe 和 D. Suciu，“迈向实用的 Querymarket 查询定价”，载于 ACM数据管理国际会议，2013年，第613–624页。[34] P. Koutris、P. Upadhyaya、M. Balazinska、B. Howe 和 D. Suciu，“基于查询的数据定价”，《ACM期刊》，第62卷，第5期，第1–44页，2015年。[35] Z. Cong、X. Luo、J. Pei、F. Zhu 和 Y. Zhang，“机器学习流水线中的数据定价”，《知识与信息系统》，第64卷，第6期，第1417–1455页，2022年。[36] S. A. Azcoitia、C. Iordanou 和 N. Laoutaris，“理解商业数据市场中的数据价格”，载于 IEEE数据工程国际会议，2023年，第3718–3728页。[37] I. Guyon、J. Makhoul、R. Schwartz 和 V. Vapnik，“多大的测试集才能得到良好的错误率估计？”，《IEEE模式分析与机器智能汇刊》，第20卷，第1期，第52–64页，1998年。[38] S. Raschka，“机器学习中的模型评估、模型选择与算法选择”，arXiv预印本 arXiv:1811.12808，2018年。[39] Y. Ma、N. Wang、A. Che、Y. Huang 和 J. Xu，“产品订单与库存中的牛鞭效应：需求预测技术视角”，《国际生产研究期刊》，第51卷，第1期，第281–302页，2013年。[40] W. Liu、S. Long、D. Xie、Y. Liang 和 J. Wang，“如何治理平台服务供应链中的大数据歧视性定价行为？基于三方演化博弈模型的考察”，《国际生产经济学期刊》，第231卷，107910页，2021年。[41] T. M. Wagner、A. Benlian 和 T. Hess，“将免费增值客户从免费版转化为高级版——音乐即服务中感知到的高级服务契合度的作用”，《电子市场》，第24卷，第259–268页，2014年。[42] A. Rao，“在线内容定价：购买与租赁市场”，《营销科学》，第34卷，第3期，第430–451页，2015年。[43] Y. A. A. Charles Zhechao Liu 和 H. S. Choi，“免费增值策略对移动应用市场的影响：一项针对 Google Play 的实证研究”，《管理信息系统期刊》，第31卷，第3期，第326–354页，2014年。[44] “Adobe.com”，https://www.adobe.com/acrobat.html，访问日期：2024年3月26日。[45] “ChatGPT Plus 介绍”，https://openai.com/blog/chatgpt-plus，访问日期：2024年3月26日。

[46] S. Alaei、A. Malekian 和 A. Srinivasan，“数字商品随机抽样拍卖”，载于 *ACM*电子商务会议，2009年，第187–196页。[47] P. Hansen、B. Jaumard 和 G. Savard，“线性双层规划的新分支定界规则”，*SIAM*科学与统计计算杂志，第13卷，第5期，第1194–1217页，1992年。[48] W. Bangerth 和 R. Rannacher，微分方程的自适应有限元方法。施普林格科学与商业媒体，2003年。[49] I. Babuˇska 和 T. Strouboulis，有限元方法及其可靠性。牛津大学出版社，2001年。[50] B. Szab´o 和 I. Babuˇska，有限元分析：方法、验证与确认。约翰·威立父子出版公司，2021年。[51] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于*IEEE*计算机视觉与模式识别会议，2016年。[52] A. Krizhevsky、G. Hinton 等，“从微型图像中学习多层特征”，加拿大安大略省多伦多，2009年。



**Xiang Li** （IEEE学生会员）于2020年获得中国电子科技大学学士学位。目前，他正在香港中文大学（深圳）理工学院攻读博士学位。他的研究兴趣包括网络经济学与优化，当前主要聚焦于人工智能背景下的定价与机制设计。



**Jianwei Huang** （IEEE会士）是香港中文大学（深圳）总统讲座教授兼副校长，同时也是深圳市人工智能与机器人研究院副主任。他曾担任《IEEE网络科学与工程汇刊》主编及《IEEE通信学会开放期刊》副主编。他是IEEE会士、IEEE通信学会杰出讲师、科睿唯安Web of Science高被引科学家，

以及一位爱思唯尔中国高被引学者。



**Kai Yang** （IEEE高级会员）获美国哥伦比亚大学博士学位。他曾担任多个备受认可的职位，包括在美国阿尔卡特朗讯贝尔实验室担任技术 staff 成员。他还曾在哥伦比亚大学担任兼职教师。目前，他是中国上海同济大学特聘教授。他拥有 20 多项专利，研究成果广泛发表于顶级人工智能会议、IEEE顶级期刊和会议。他目前的研究兴趣包括

人工智能+网络、面向时间序列的人工智能、嵌套优化、无线感知与通信。他还曾担任众多人工智能和 IEEE 会议的技术程序委员会（TPC）委员。他曾荣获多项重要奖项，包括哥伦比亚大学颁发的 Eliahu Jury奖、贝尔实验室团队合作奖，以及 IJCAI 2023 时间序列研讨会最佳论文荣誉提名奖。2012年至2014年，他担任 IEEE通信学会多媒体通信技术委员会副主席。此外，他还曾担任 IEEE INFOCOM 演示/海报联合主席、IEEE GLOBECOM 专题研讨会联合主席，以及 IEEE ICME 研讨会联合主席。



**Chenyou Fan** （IEEE会员）于2011年获得中国南京大学计算机科学学士学位，并于2014年和2019年分别获得美国印第安纳大学硕士和博士学位。他现任职于华南师范大学。研究方向包括机器学习和计算机视觉。他在 IEEE模式分析与机器智能汇刊、计算机视觉与模式识别会议、IJCAI、ACM多媒体会议以及顶尖人工智能会议和期刊上发表论文逾15篇。