《Counterfactual Response Modeling with a Binary Treatment》独立中文问答导读

基于论文 canonical 文稿编写

本导读可独立阅读；保留术语 *unit* 为英文，以免与度量“单位”混淆。

摘要

问：这篇论文究竟解决什么，而不解决什么？
答：它提出 AB-HTE：先由处理前协变量 X 推断潜在 unit 表征 U 的条件分布，再把同一 U 送入二元处理索引的结构机制，估计两臂潜在结果的条件边际分布及其中位数（位置）差。它不从单个事实结果“恢复”该人的缺失反事实，也不凭事实数据识别跨世界联合分布。

问：为何主估计量不是通常的 CATE？
答：主模型的非退化效应服从 Cauchy 分布，普通均值不存在。因此论文使用

$$\tau_{\mathrm{loc}}(x) = \mathrm{med}\{Y(1)|x\} - \mathrm{med}\{Y(0)|x\}.$$

把它称为 CATE 会混淆“均值差”和“位置差”。

问：最可信的实证结论有多窄？
答：Cauchy AB-HTE 在 gross-outlier replacement 下比匹配的 Gaussian AB-HTE 退化更小；但其他污染下直接稳健学习器常更好，IHDP 上 TARNet/CFR-style 基线更好。WAWS 只能检验事实拟合和随机化总体对比，不能检验个体效应真值。
证据边界：恢复定理针对正确设定、无惩罚的总体风险；不覆盖有限样本一致性、固定非零正则化或隐藏混杂。

目录

1 导论：反事实响应为何不是双臂预测

问：既然能分别学习 $Y|T=0,X$ 与 $Y|T=1,X$，为何还谈“abduction”?

答：双臂回归给出两个响应面；AB-HTE 则显式构造同一 unit 的潜在表征分布 $q_\phi(U|x)$，再用共享、处理索引机制评价 $t=0,1$。这是一种结构化共享信息方式，但“abduction”不自动带来因果识别：交换性、正值性仍须外加。

问：heterogeneous treatment effects 在机制中由哪一项承担?

答：机制

$$Y(t) = a_0 + a^\top U + t(b_0 + b^\top U) + \sigma_t E_t$$

中的 $tb^\top U$ 使位置效应随 x 经由 $\mu_\phi(x)$ 改变；若 $b=0$，位置差仅为常数 b_0。

问：为什么两条已识别边际仍不足以得到个体效应分布?

答：$Y(1)-Y(0)$ 的分布依赖 $Y(1),Y(0)$ 如何配对，即跨世界 coupling/copula。随机试验也只观察每个 unit 的一个结果，因而不能从边际选择“独立重采样”还是“共享外生噪声”。

2 相关工作：创新应与谁比较

问：AB-HTE 相对 S/T-learner、TARNet/CFR 的增量是什么?

答：不是“首次学习 HTE”，而是把分布值潜在 abduction、处理–unit 仿射交互、Cauchy 闭包和 coupling 敏感性分析组合起来。关键对照是匹配架构的 Gaussian AB-HTE 与相近容量的直接 Cauchy S-learner。

问：它比灵活生成式反事实模型更强吗?

答：不是全面更强。它牺牲多峰、非对称、有限支撑等密度灵活性，换取解析响应核、解析分位数差和可审计的机制传播。

问：Makarov bounds 在这里扮演什么角色?

答：它们不选定 coupling，而是在只相信已识别边际时，为效应 CDF 或获益概率给出对所有 coupling 的点态 sharp bounds，是对模型指定的效应律的 assumption-lean 参照。

3 问题设定与估计对象

问：最小识别假设是什么?

答：一致性、无干扰、条件交换性 $(Y(0), Y(1)) \perp T|X$，以及目标人群上的正值性 $0 < e(x) < 1$。稳定加权还需要更强的 overlap：$\epsilon \le e(x) \le 1 - \epsilon$。

问：$Q_1(p|x) - Q_0(p|x)$ 是效应 $Y(1) - Y(0)$ 的第 p 分位数吗?

答：一般不是。前者只由两个边际分位数构成并可识别；后者需要跨世界联合律。即使 $p = 1/2$，两个中位数之差也不自动等于差值的中位数，除非另有结构。

问：binary treatment 带来了什么明确边界?

答：模型只需评价 $t \in \{0, 1\}$ 两个机制分支，位置异质性可写成 $b_0 + b^\top \mu_\phi(x)$。但二元性并未解决 fundamental problem：每个 unit 仍只暴露一个分支。

证据边界：潜在结果边际可在假设下识别；单个缺失潜在结果与跨世界 coupling 不可由事实回归识别。

4 方法：Cauchy abduction 与机制约束

问：abduction 具体推断什么？它是精确后验吗?

答：编码器输出 $\mu_\phi(x), \gamma_\phi(x)$，定义因子化 Cauchy

$$q_\phi(U|x) = \prod_j \text{Cauchy}(\mu_{\phi,j}(x), \gamma_{\phi,j}(x)).$$

它是 amortized conditional latent model，不是 Bayes-exact posterior，也不是被数据识别的物理坐标。

问：Cauchy 与仿射机制为何能解析化?

答：独立 Cauchy 变量的仿射和仍为 Cauchy，因此

$$m_t = a_0 + tb_0 + (a + tb)^\top \mu_\phi(x), \quad s_t = \sum_j |a_j + tb_j| \gamma_{\phi,j}(x) + \sigma_t.$$

所以 $Y(t)|x \sim \text{Cauchy}(m_t, s_t)$，中位数差直接化为 $b_0 + b^\top \mu_\phi(x)$，无需 Monte Carlo 积分。

问：共享同一个 U 是否已经识别了个体效应?

答：没有。独立事件重采样给出效应尺度 $\sum_j |b_j|\gamma_j + \sigma_1 + \sigma_0$；共享外生实现给出 $\sum_j |b_j|\gamma_j + |\sigma_1 - \sigma_0|$。两者事实边际可相同，效应离散度与获益概率却不同。

问：训练目标与理论对象是否完全一致?

答：实验最小化事实 Cauchy NLL，并使用 AdamW、10^{-5} weight decay 和梯度裁剪；理论 Fisher consistency 则是无惩罚总体风险。故实现上的正则化与优化成功不能直接由定理担保。

5 识别与性质：定理到底保证到哪里

问：Fisher consistency 恢复了什么?

答：在因果假设、两臂真实条件核都由同一模型正确实现、有限交叉熵和有界正权重下，无惩罚总体 log-risk 的全局极小点恢复 m_t, s_t，从而恢复 τ_{loc}。它不保证参数唯一，也不保证有限样本训练收敛到该点。

问：为何机制系数看似清晰却不可识别?

答：对潜变量作可逆对角缩放和平移，同时反向变换机制权重和截距，可保持两个响应核不变；潜在与事件噪声的尺度分配也可能有歧义。数据识别的是输出核，不是某套 U 坐标或"真实机制参数"。

问：稳健性定理能否抵抗任意污染与混杂?

答：不能。Cauchy NLL 对位置与 log-scale 的残差 score 有界；参数梯度有界还要求模型 Jacobian 受控。这既不是无条件 bounded-influence 定理，也不修复 treatment label 错误、系统性臂偏移、overlap failure 或隐藏混杂。

问：效应分布最诚实的报告方式是什么?

答：将 F_t 与边际分位数差标为"假设下识别"；将 coupling-specific F_Δ 标为"模型蕴含/敏感性"；同时报告不指定 coupling 的 Makarov bounds。非退化 Cauchy 效应均值不存在，不应改名为 CATE。

6 实验设计：truth isolation 如何成立

问：为什么必须分三层证据?

答：全合成数据可暴露两条响应面及生成器指定的 coupling；半合成数据通常只给响应面真值；真实 RCT/观察数据只有事实结果与分配设计。不同层能回答的问题不同，不能把真实数据事实拟合冒充 ITE accuracy。

问：污染协议怎样避免"干净验证集 oracle"?

答：训练集和验证集以同一污染算子及比例、但独立随机性被污染；模型选择完成后才在干净 oracle 测试人群上评效应。干净验证仅作为附录消融，反事实 oracle 不参与选择。

问：哪些比较真正隔离了机制?

答：Cauchy AB-HTE 对 Gaussian twin 主要隔离 likelihood family；对直接 Cauchy S-learner 检验 shared abduction 是否超越稳健结果回归；matched-corrupted 对 clean validation 隔离 pipeline robustness 与 trusted-selection oracle。参数量相近但并非完全相同，不能声称精确容量匹配。

问：真实 RCT 能提供哪类 truth?

答：能提供随机化支持的 aggregate ITT、事实 score/calibration 与 cross-fitted policy point diagnostics；不能提供每个 unit 的反事实标签。条件于学习分数的一阶标准误也不是完整的训练算法推断。

7 结果：哪些胜利成立，哪些不成立

问：gross outlier 结果支持什么因果链?

答：在 40% gross-outlier replacement 下，Cauchy AB-HTE 的 MedAE 为.086，较自身干净值仅增.016，且低于 Gaussian twin 的.206。这支持 heavy-tail likelihood 优势；直接 Cauchy S 为.084，又否定了"优势必来自 latent abduction"。

问：隐藏混杂负控为何关键?

答：随机分配时误差约.071；加入未观测共同原因并按 gains 排序后升至 1.445，超过二十倍。它实证展示：估计器再复杂也不能修复交换性缺失。

问：IHDP 与 WAWS 各能证明什么?

答：IHDP 的模拟响应面允许比较位置效应误差：Cauchy AB-HTE 优于其 Gaussian twin，却落后 TARNet/CFR-style。WAWS 的 $n = 4218$ RCT 中，它事实 MedAE 最低，但 ITT 不精确、policy 排序不稳；这不是 ITE 验证。

问：coupling 敏感性是否只是理论细节?

答：不是。同一拟合与同一事实 NLL 下，独立重采样和共享实现的平均预测效应尺度由.372 变为.067，获益概率也改变；说明基于"谁会获益"的决策可被不可识别的跨世界选择主导。

8 讨论与结论：如何使用而不过度解读

问：解析闭包的代价是什么?

答：仿射对称 Cauchy 核对多峰、非对称、有界结果可能错设；解析便利不是现实适配性的证据，需由校准与外部复制判断。

问：最危险的误读是什么?

答：把潜在 abduction 当作发现隐藏混杂，把位置差当作 CATE，把模型 coupling 当作被数据识别，或把低事实误差当作个体反事实准确率。这四步都越过论文明确的证据边界。

问：实践中最低限度应报告什么?

答：两臂事实校准与 overlap 诊断；τ_{loc} 而非不存在的均值；不同 coupling 下的效应尺度/获益概率及 Makarov bounds；相对 Gaussian twin 与直接 robust learner 的消融；并按合成、半合成、真实数据分别标注可见 truth。

证据边界：论文最稳健的贡献是“可审计地分离边际识别与跨世界建模”，而不是宣称恢复了 unit-level counterfactual truth。

A 合并附录导读：证明、实现与额外实验

问：理论附录最值得核查什么？

答：一是边际分位数差不等于效应分位数的构造反例；二是 Cauchy affine closure 的正尺度与退化点质量边界；三是跨世界非识别的成对模型构造；四是 Makarov 上下界的 sup/inf 定义。它们共同防止把解析公式误当识别公式。

问：实现附录应检查哪些“定理到代码”的缝隙？

答：scale floor、绝对值在零点的不可微处理、权重语义、正则化、early stopping、梯度裁剪，以及完整性检查。尤其要区分“无惩罚总体风险定理”与“带 weight decay 的有限样本神经网络”。

问：实验附录怎样支持 truth isolation？

答：它固定 DGP、污染算子、独立训练/验证污染、分割、选择规则、泄漏控制、overlap 与 ESS 诊断，并把 treatment flip、arm-differential shift 与 outcome shuffle 分开。附录结果是边界审计，不应被挑选成新的主结论。

问：扩展相关工作给读者什么压力测试？

答：应把方法放回 causal forests、R/meta learners、TARNet/CFR/Dragonnet、分布与生成式反事实模型、稳健/分位数回归及部分识别文献中。当前比较尚未覆盖全部强基线，因此“闭式且可审计”比“SOTA”更符合证据。

证据边界：多个原论文 appendix 在本导读中合并为一节；其角色是核验假设、证明与复现实验，而非扩大正文 claim。

参考文献

[1] Yanqin Fan and Sang Soo Park. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951, 2010.

[2] Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. Quantifying aleatoric uncertainty of the treatment effect: A novel orthogonal learner. In *Advances in Neural Information Processing Systems*, volume 37, 2024.