Individual Treatment Effect in Presence of Observable Interference

Anonymous authors
Paper under double-blind review

Abstract

Individual Treatment Effect (ITE) estimation is an extensively researched problem, with applications in various domains. We model the case where observable interference might happen between the treatment prescription and its effect, a typical situation in health (because of non-compliance to prescription) or digital advertising (because of competition and ad blockers for instance). When the interference level is high, the ITE signal fades and becomes hard to learn. To solve this problem, we propose a new approach to estimate ITE that takes advantage of observable interference to reduce variance, all the more that the interference is high. We use the Structural Causal Model framework and do-calculus to define a setting under which this estimator indeed recovers the ITE, and study its asymptotic variance. Finally, we conduct extensive experiments on both synthetic and real-world dataset that highlight the benefit of the approach, which outperforms state-of-the-art on PEHE and AUUC.

1 Introduction

Individual Treatment Effect (ITE) estimation is an important task in various applications such as healthcare Foster et al. (2011), online advertising Diemert et al. (2018) and socio-economics Xie et al. (2012). The the causal effect of T on outcome Y conditionally to context X can be thought of as a contextual counterpart of the usual Average Treatment Effect (*cf.* Equation 1, assuming randomized treatment).

$$\begin{aligned} ATE &= \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] \\ ITE(x) &= \mathbb{E}[Y|X=x, T=1] - \mathbb{E}[Y|X=x, T=0] \end{aligned} \tag{1}$$

However there often exists a risk of interference between the treatment prescription (T) and its actual acceptance (M) as illustrated in Table 1. This happens for instance when individuals have the choice not to abide by the prescription or if there exists conflicting interests. Of course one can choose to focus the study on actually treated individuals only. But from a decision making point of view it often makes sense to consider that future treatment decisions need to take into account the level of interference so as to accurately predict future expected outcomes. For example a policy maker would want to take into account that not all individuals would abide by the new policy (as can be estimated from a pilot study) to predict the expected impact of a roll-out of said policy. Now we argue that ITE estimation can be hampered by interference. Firstly note that individuals whose treatment has been

Table 1: Examples of covariates (X), outcome (Y), treatment assignment (T), reason for not abiding to treatment (R) and evidence of treatment acceptance (M)

VAR.	**MEDICINE**	**ONLINE ADV.**	**JOB TRAINING**
X	patient info	purchase history	schooling
T	drug prescription	bid placement	training offer
R	reluctance	competition	disease
M	drug intake	ad displayed	training done
Y	recovery	sale/visit	employment

interfered with contribute only noise to the default ITE estimator as their observed outcome is not effectively influenced. Therefore the variance of the estimator increases with the interference level. A classical move in statistics would be to build a stratified estimator on the interference variable. But we will see in Section 3 that even when observing data from a randomized experiment one needs additional assumptions to build a consistent estimator that recovers the causal effect of T on Y.

Besides, ITE models[1] are often considered as prescriptive tools. Indeed, ITE predictions are used in order to target treatment to individuals for which it is the most beneficial (Devriendt et al., 2018; Radcliffe & Surry, 2011). This calls for an evaluation metric that measures by how much the ATE would have been improved had the treatment been targeted not by a random instrument, but by ITE predictions instead. For that purpose (Rzepakowski & Jaroszewicz, 2012; 2010; Radcliffe & Surry, 2011) have proposed the Area Under the Uplift Curve (AUUC) metric that sums the benefits over individuals ranked by predictions. An interesting property of this metric is that it can be used on real data for which we observe a given individual in either treated or untreated conditions but never both.

Confronted with the challenges of i) learning ITE models in conditions of interference and ii) evaluating them as prescriptive tools we propose to pose the problem in the setting of causal inference and derive an ITE estimator that takes advantage of observed interference. Our main contributions are as follows.

1. Formalization of ITE estimation in presence of observable interference using structural causal models (Section 3)
2. Proposition of an ITE meta-estimator in which can be plugged existing ITE estimators , proof of consistency and asymptotic variance properties (Section 4)
3. Thorough empirical evaluation of this estimator on synthetic and real world datasets (Section 5)

2 Related works

We review three main domains that are concerned with research questions similar to our work: ITE modeling, interference in causal inference and evaluation metrics for ITE modeling.

Firstly, we note that ITE models are a pervasive concept in different research fields such as marketing - under the name uplift models Radcliffe & Surry (2011), statistics - as conditional average treatment effect estimators Künzel et al. (2019) or econometrics - heterogeneous treatment effect models (Jacob et al., 2019; Wager & Athey, 2018). A simple yet highly scalable approach consists in learning a regression of Y on X separately in both treatment ($T = 1$) and control ($T = 0$) populations and return the difference, known as T-learner Künzel et al. (2019) or ”Two Models” Radcliffe & Surry (2011). A variation of this approach with larger model capacity have been proposed through a shared representation (SDR) for the treatment and control group Betlei et al. (2018). Also, a prolific series of work exists on adapting decision trees and random forests to the causal inference framework Athey & Imbens (2016); Wager & Athey (2018); Athey et al. (2019). Further in the same vein and building on work done on double machine learning by Chernozhukov et al. (2018), Oprescu et al. (2019) generalize the idea of causal forests, allowing for high-dimensional confounding. Finally, another recent trend is to study theoretical limits in ITE estimation and especially generalization bounds Shalit et al. (2017); Alaa & Van Der Schaar (2018).

Then regarding the concept of observable interference, algorithms have been studied that focus on recovering the (individual) causal effect of M on Y, but they do not take into account the problem of interference between the treatment assignment T and its acceptation M. Such observable interference can typically correspond to non-compliance. However, to our knowledge works tackling this problem focus on effect of the treatment intake $M -$ and not the treatment assignment $T -$ on the outcome Y Gordon et al. (2019). In that context, the effect of T on Y, sometimes referred to as the intention-to-treat (ITT) effect, is typically used in an instrumental variable framework to recover the effect of M on Y Imbens & Angrist (1994); Syrgkanis et al. (2019). In the contrary, we focus in this work on the effect of the treatment assignation T on the variable Y, taking advantage of the observed interference M.
The idea of taking advantage of a mediation variable to recover individual treatment effect has

[1] also called *uplift* models in marketing literature.

been explored notably by Hill et al. (2015), however the associated assumptions (M and Y are unconfounded) are more restrictive than the ones we propose, as we do not require the binary mediation variable M (representing the interference) to be unconfounded with Y. We do however assume that it satisfies a strong monotonicity assumption with respect to the binary variable T, *i.e.* $T = 0 \Rightarrow M = 0$: an analogous assumption is referred to as *one-sided non-compliance* by Gordon et al. (2019).
Similar monotonicity assumptions are typically made in causal inference works but concern the causal effect: the outcome Y is assumed to be monotonous with respect to the treatment T Kallus (2019); Oberst & Sontag (2019).

Finally, many research works validate their approach using synthetic data, in which a pointwise error measure named Precision Estimation of Heterogeneous Effect (PEHE) Shalit et al. (2017) can be computed. However in real world cases, the *fundamental problem of causal inference* states that the ground truth of individual treatment effect cannot be observed (since an individual is either treated or untreated but never both at the same time), preventing to use such metrics beyond simulation studies. Since our main motivation is to determine which individuals are good candidates for treatment assignment, we choose to evaluate the performance of our estimators on real data using the AUUC, which evaluates the ranking of individuals implied by corresponding ITE predictions. One can view the resulting measure as a prediction of the expected benefit of assigning treatment according to the model prediction instead of a random uniform assignment. Overall AUUC has been used in recent years in machine learning research to evaluate baseline ITE models vs SDR Diemert et al. (2018), flavors of Support Vector Machines for ITE estimation Kuusisto et al. (2014) or direct treatment policy optimization Yamane et al. (2018). For completeness a variant normalized by the ranking of an oracle model exists also under the name Qini coefficient Radcliffe & Surry (2011).

3 Framework

The approach we are presenting here strongly relies on causality notions such as *structural causal model*, *causal graph*, *intervention* and *valid adjustment set*. The formalism we use through the paper is strongly inspired by Peters et al. (2017). **Notations**. For sake of compactness, we use the following notations for any binary variable W and multi-dimensional variable X: $\mathbb{P}(W) \triangleq \mathbb{P}(W = 1)$, $\mathbb{P}(\overline{W}) \triangleq \mathbb{P}(W = 0)$, $\mathbb{P}(x) \triangleq \mathbb{P}(X = x)$, $\mathbb{P}^{\mathfrak{C};do(W)}(.) \triangleq \mathbb{P}^{\mathfrak{C};do(W:=1)}(.)$.

3.1 The observable interference setting

The setting of observable interference we consider in this work is entirely defined by a SCM of variables X, T, M, Y, U for which example values where proposed in Table 1: X, belonging to a multi-dimensional space $\mathcal{X}$, contains the individual's descriptive features (by simplicity, we will confuse the individual and their features, referring for example to 'an individual x'), T is the binary treatment assignment variable, M is the binary mediation variable, corresponding to the treatment acceptation, *i.e.* the fact that the individual did not interfere with their treatment prescription, Y is the binary outcome variable, U represents (allowed) unobserved confounders between X and Y.

In what follows, we define the structural causal model $\mathfrak{C} = (\mathbb{S}, \mathbb{P}_{\mathbf{N}})$, which is henceforth assumed to represent the causal mechanisms underlying the variables of interest in our work.
$\mathbb{S}$ is defined in Equations 2: $\mathbb{P}_{\mathbf{N}}$ satisfies the following mild conditions: N_U, N_X, N_M, N_Y are noise

$$\begin{aligned} T &= \tilde{N}_T \\ U &= N_U \\ X &= f_X(U, N_X) \\ M &= f_M(X, N_M) \times T \\ Y &= f_Y(X, M, U, N_Y). \end{aligned} \tag{2}$$

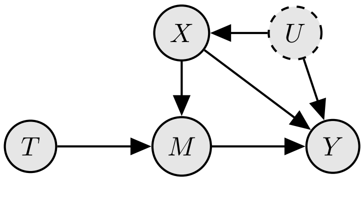

Figure 1: Causal graph $\mathcal{G}_{\mathfrak{C}}$ induced by the SCM $\mathfrak{C}$.

consistent with variables definitions, and $\tilde{N}_T$ is distributed according to a Bernoulli distribution with parameter $p = \mathbb{P}^{\mathfrak{C}}(T)$, consistent with a randomized controlled experiment setting.

The associated causal graph $\mathcal{G}_{\mathfrak{C}}$ is given in Figure 3.1

In the next proposition, we list four assumptions implied by the $\mathfrak{C}$ about the variables of interest.

Proposition 1 *The SCM $\mathfrak{C}$ defined in Equations 2 implies the following assumptions on variables T, M, Y and X:*

$$\begin{aligned} \textit{(Randomized treatment)} \quad & T \perp\!\!\!\perp X \\ \textit{(Exclusive mediation)} \quad & T \perp\!\!\!\perp Y \mid \{X, M\} \\ \textit{(Strong mediation monotonicity)} \quad & T = 0 \Rightarrow M = 0 \\ \textit{(Valid covariate adjustment)} \quad & \{X\} \textit{ is a VAS for } (M, Y) \end{aligned} \tag{3}$$

The fact that $\mathfrak{C}$ implies the *randomized treatment* and *Exclusive mediation* assumptions relies on the Markov property of the causal graph $\mathcal{G}_{\mathfrak{C}}$ and the notion of d-separation. The *Strong mediation monotonicity* is straightforwardly implied by the structural assignment of M given in Equations 2, while the *valid covariate adjustment* assumption relies on the back-door criterion (Pearl, 2009; Peters et al., 2017). The complete proof of Proposition 1, including definitions of the associated notions, is given in appendix.

3.2 ITE IN PRESENCE OF INTERFERENCE

Notations. For all $x \in \mathcal{X}$, we define the individual treatment effect $\tau^{ITE}(x)$, treatment effect if treated $\tau^{ITET}(x)$, as well as the individual non-interference $\gamma(x)$ probability (that we henceforth refer to as *individual compliance* for clarity) as follows:

$$\begin{aligned} \tau^{ITE}(x) &= \mathbb{P}^{\mathfrak{C};do(T)}(Y|x) - \mathbb{P}^{\mathfrak{C};do(\overline{T})}(Y|x), \\ \tau^{ITET}(x) &= \mathbb{P}^{\mathfrak{C};do(M)}(Y|x) - \mathbb{P}^{\mathfrak{C};do(\overline{M})}(Y|x), \\ \gamma(x) &= \mathbb{P}^{\mathfrak{C};do(T)}(M|x). \end{aligned} \tag{4}$$

We also define the relative ITET $\beta(x)$ and relative ITE $\alpha(x)$ as:

$$\alpha(x) = \frac{\mathbb{P}^{\mathfrak{C}}(Y|T,x) - \mathbb{P}^{\mathfrak{C}}(Y|\overline{T},x)}{\mathbb{P}^{\mathfrak{C}}(Y|\overline{T},x)}, \qquad \beta(x) = \frac{\mathbb{P}^{\mathfrak{C}}(Y|M,x) - \mathbb{P}^{\mathfrak{C}}(Y|\overline{M},x)}{\mathbb{P}^{\mathfrak{C}}(Y|\overline{M},x)}.$$

The proposed method exploits the mediation variable M, *i.e.* the treatment acceptation, by splitting the treatment to outcome path into a product of two *subpaths*, both with a higher signal-to-noise ratio. In particular, under $\mathfrak{C}$, we can integrate M into the $\mathbb{P}^{\mathfrak{C};do(T)}(Y|x)$ expression as presented in the next lemma.

Lemma 1 *Assuming $\mathfrak{C}$, and for any $x \in \mathcal{X}$, the positive outcome probability under treatment, $\mathbb{P}^{\mathfrak{C};do(T)}(Y|x)$, can be written as follows:*

$$\mathbb{P}^{\mathfrak{C};do(T)}(Y|x) = \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M}) \;+\; \mathbb{P}^{\mathfrak{C}}(M|x,T)\Big(\mathbb{P}^{\mathfrak{C}}(Y|x,M) - \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})\Big). \tag{5}$$

The proof of Lemma 1 is fully detailed on appendix. It relies on *valid covariate adjustment*, *randomized treatment*, and *exclusive mediation* assumptions, that we have proven to be implied by $\mathfrak{C}$ in Proposition 1. In a nutshell, the Lemma 1 decomposes the positive outcome probability under treatment of a given individual as the sum of the organic positive outcome probability and the product of individual compliance and individual treatment effect if treated.

In Proposition 2, we present a result linking the ITE, the ITET and the individual compliance:

Proposition 2 *Assuming $\mathfrak{C}$, the ITE decomposes as follows:*

$$\tau^{ITE}(x) = \tau^{ITET}(x)\gamma(x) \tag{6}$$

The proof of Proposition 2 is fully detailed in appendix. It relies on the identity $\mathbb{P}^{\mathfrak{C};do(\overline{T})}(Y|x) = \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})$, which holds under $\mathfrak{C}$ thanks to the *exclusive mediation* and *strong mediation monotonicity* assumptions.

4 Proposed Approach

The expression proven in Proposition 2 calls for a novel way to estimate individual treatment effect, by first estimating separately both factors $\tau^{ITET}(x)$ and $\gamma(x)$, then multiplying these estimators to form a *post-mediation individual treatment effect* (MITE) estimator.
Formally, let $\hat{\tau}^{ITET}$ be an estimator of τ^{ITET}, let $\hat{\gamma}$ be an estimator of γ. We then define the associated MITE estimator $\hat{\tau}^{MITE}$, for any x, as:

$$\hat{\tau}^{MITE}(x) = \hat{\tau}^{ITET}(x)\hat{\gamma}(x). \tag{7}$$

In practice, $\hat{\tau}^{ITET}$ may be obtained using any individual treatment effect estimator. Indeed, under $\mathfrak{C}$, the individual causal effect of M on Y given X is recoverable since $\{X\}$ is a valid adjustment set for (M, Y) as explained in Section 3.

Assuming that x, $\hat{\tau}^{ITET}(x)$ and $\hat{\gamma}(x)$ are *consistent* estimators of resp. $\tau^{ITET}(x)$ and $\gamma(x)$, Proposition 2 then ensures that $\hat{\tau}^{MITE}(x)$ is a *consistent* estimator of $\tau^{ITE}(x)$. Thanks to its expression as a function of a ITET estimator, the MITE estmator *focuses* on the individuals who actually accepted treatment, who we know to be the only individuals contributing to the ITE signal (*Exclusive mediation* assumption in Equation 3). We therefore expect the MITE estimator to have lower variance than an ITE estimator which does not exploit the observable interference. Comparing a MITE and an ITE estimator is all the more fair than we use an analogous version of the ITE estimator for the ITET estimator the MITE estimator is built on, which we know to be feasible thanks to the *valid covariate adjustment* assumption. We refer to this approach as *symmetrically learning algorithms comparison*, and use it to conduct our experiments in Section 5.
In the following proposition, we compare the asymptotic variance of estimators $\hat{\tau}^{MITE}$ and $\hat{\tau}^{ITE}$ in the following simple yet realistic setting:

Single-stratum setting. We focus on the ITE estimation for a single value x_0 of X, for which we assume to observe n *i.i.d.* samples $\{(x_0, T_i, M_i, Y_i)\}_{1\leq i\leq n}$. In practice, this generalises to any stratum $S \subset \mathcal{X}$ containing x_0 for which the adjustment set formula is valid, *i.e.* if the variable $X' \triangleq x_0 I_{X\in S} + X I_{X\notin S}$ defines a valid adjustment set for (M, Y).

Notations. Consistently with notations presented in Equations 4, $\alpha(x_0), \beta(x_0)$ refer respectively to the relative ITE and relative ITET in stratum $\{X = x_0\}$ (and are assumed to be positive in this illustrative setting), and we denote $\hat{\tau}^{ITE}(x_0)$, $\hat{\tau}^{ITET}(x_0)$, $\hat{\gamma}(x_0)$ the respective maximum-likelihood estimators (MLE) of $\tau^{ITE}(x_0)$, $\tau^{ITET}(x_0)$, $\gamma(x_0)$. We define the associated MITE estimator as $\hat{\tau}^{MITE}(x_0) \triangleq \hat{\gamma}(x_0)\hat{\tau}^{ITET}(x_0)$. Lastly, we denote $p_1(x_0) = \mathbb{P}^{\mathfrak{C}}(Y|T, x_0)$.
In the following Proposition, we present an asymptotic bound for the ratio of the standard deviation (sd) of MITE and ITE estimators.

Proposition 3 *Under $\mathfrak{C}$ defined in Section 3.1 with $\mathbb{P}^{\mathfrak{C}}(T) = \frac{1}{2}$, and assuming we observe n i.i.d. samples in stratum $\{X = x_0\}$, we have:*

$$\lim_{n\to\infty} \frac{sd(\hat{\tau}^{MITE}(x_0))}{sd(\hat{\tau}^{ITE}(x_0))} \leq \sqrt{\left(\frac{2(1+\beta(x_0))}{(1-p_1(x_0))(2+\alpha(x_0))}\right)\gamma}. \tag{8}$$

This theoretical bound shows that the ratio standard deviations of MITE and ITE estimators is all the more smaller that the non-interference factor $\gamma(x_0)$ is low. However, additional assumptions need to be made about $\beta(x_0)$ and $p_1(x_0)$ in order to recover an informative bound in practice.

In real-world dataset presented in Section 5, we consistently observe $\hat{\beta}(x) \leq 12$ and $\hat{p}_1(x) \leq 0.05$, where the estimators correspond to logistic regression models fined-tuned following the protocol described in Section 5.

5 Experiments

To qualify the performance of the MITE estimator, we study its benefits in a variety of settings. Firstly we study its properties on simulation-based studies, hereafter denoted by ‘Synthetic Datasets’, for which i) the ITE ground truth is known ii) the level of interference can be controlled and iii) we can appreciate performance with respect to an Oracle. Moreover we apply our approach to transform

Table 2: Examples of covariates (X), outcome (Y), treatment assignment (T), reason for not abiding to treatment (R) and evidence of treatment acceptance (M)

TREATMENT (T)	**EXPOSURE** (M)	**OUTCOME** (Y)
0 (2’096’236)	-	3.82% (79’986)
1 (12’161’477)	3.65% (444’384)	4.93% (599’170)

Table 3: Repartition of visits ($Y = 1$) on CRITEO-UPLIFT1 split on exposure groups

EXPOSURE (M)	**OUTCOME** (Y)
0 (13’813’329)	3.58% (495’003)
1 (444’384)	**41.4%** (184’153)

baseline ITE estimators and compare their performance on a real-world, large scale, dataset named ‘CRITEO-UPLIFT1 Dataset’[2], for which we shall recover the AUUC.

In each experiment we take care of comparing symmetrically learning algorithms for which we provide or not the MITE estimator decomposition so as to highlight corresponding benefits or drawbacks. To simplify experiments we chose two base models: Two Models (2M) and SDR as they easily scale to large datasets and have been found competitive in prior studies Betlei et al. (2018). For reproducibility sake we have implemented models using Scikit-Learn Python library Pedregosa et al. (2011). All experiments were run on a machine with 48 CPUs (Intel(R) Xeon(R) Gold 6146 CPU @ 3.20GHz), with 2 Threads per core, and 500Go of RAM. Finally, we note that the state of the art is always evolving and improving. We did not use the most advanced models because we do not aim at outperforming them. Instead, we claim that the MITE estimator can improve any ITE estimator (while keeping the same model) in the case of high observable interference.

5.1 Datasets

Synthetic Datasets

We define a simulation setting in which $\mathcal{X} = \{0,1\}^{10}$, $N = 2.10^6$. The response is generated according to

$$Y \sim \text{Bern}\left(p_0 \left(1 + TM\beta(x)\right)\right), \tag{9}$$

where $T \sim \text{Bern}(0.5)$, $M \sim \text{Bern}(\gamma(x))$, and $p_0 = \mathbb{P}^{\mathfrak{C}}(Y|\overline{M}, x) = 0.1$, using notations from Equations 4. This procedure allows for varying $\gamma(x)$ and $\beta(x)$ to simulate different levels of interference and relative ITET, respectively.

CRITEO-UPLIFT1 Dataset

We use the CRITEO-UPLIFT1 dataset where data were collected using a randomized trial, from an advertising application. Key statistics for this dataset are summarized in Table 2 and 3. Notably, average treatment assignment $\mathbb{E}[T] \approx .85$ indicates that only about 15% of users where assigned to the control group (without any advertisement). The advertisers participated in online ad auctions for the rest of the population. Among the users that advertisers tried to expose ($T = 1$), only 3.65% were actually exposed, which is an extremely high interference rate, expected to highlight MITE estimator benefits. Effective exposure to ads embodies the M variable in this setup. Treatment assignment and interference rates are illustrated on Figure 2. The outcome of interest Y is the variable ‘user visiting the advertiser website’, and its mean is more than 10x higher given actual exposure ($M = 1$) versus non-exposure ($M = 0$).

5.2 Experiments

A common procedure for all experiments is to select hyperparameters (regularization norm and strength) of models using internal cross validation on the training set. For the MITE estimators we recall that an additional probabilistic model of the compliance $\hat{\gamma}(x)$ is required. For all experiments it is learned on the training data as a logistic regression. Hyperparameters are selected by internal cross-validation on the training set by ranking by log-likelihood (LLH) as the model is supposed to predict a probability.

[2] http://cail.criteo.com/criteo-uplift-prediction-dataset/

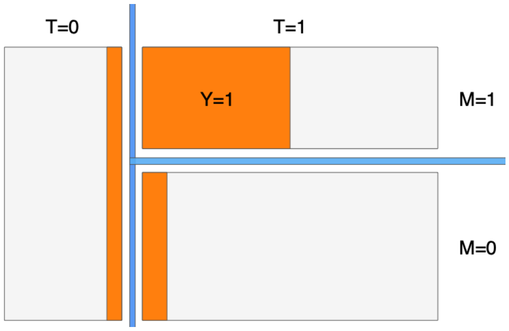

Figure 2: The visit users repartition is mainly influenced by the mediation variable M on the CRITEO-UPLIFT1 dataset

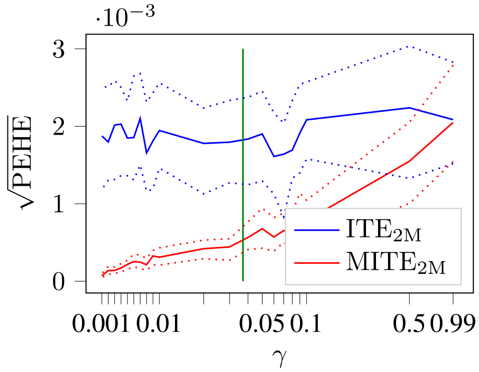

Figure 3: **Interference Sensitivity (Simulation study)**. PEHE (lower is better) of ITE vs MITE models at varying compliance level γ. Solid line — represents the median, and dashed line ⋯ represents 5% and 95% of confidence intervals. — compliance level γ for CRITEO-UPLIFT1.

Interference Sensitivity Experiment (Simulation)

The goal is to highlight how the interference level $1 - \gamma$ influences the performance of both the traditional ITE estimator and MITE estimator. For this purpose we vary $\gamma \in [10^{-4}; .99]$ and generate synthetic datasets as described in Section 5.1 with one value of β per context in $\{-1, 0, 1, 2, 3, 4, 5, 6, 7, 8\}$. We report the PEHE metric for both estimators and estimate variance by repeating experiment with 51 random test/train splits. Recall that PEHE is the squared difference between the ITE ground truth and the prediction of the model.

We observe on Figure 3 that the MITE estimator significantly outperforms the ITE estimator when the level of interference is high (low γ) and has similar performance to baseline ITE when there is no interference (γ close to 1). This shows that our post-mediation approach significantly reduces the noises due to interference and can learn a smaller signal. This is true in particular for compliance levels γ in the range that is observed on real dataset.

Baseline Experiment (Simulation)

For the synthetic dataset, the goal is to simulate a realistic scenario where there exists heterogeneity in interference and post-mediation treatment effects. More precisely, for each instance value x, we draw once and for all $\gamma \in \{0.01, 0.005\}$ and $\beta \in \{-1, 1, 3, 5, 7\}$ uniformly and independently. Associated outcome is computed according to the synthetic dataset equation (9). We study four methods: regular two-models (ITE_{2M}), shared data representation (ITE_{SDR}) and their variants, obtained by adding the intermediate variable M (resp. MITE_{2M}, and MITE_{SDR}). We focus on AUUC metric because real-world application cannot access individual treatment effect ground truth. Recall that AUUC measures the capacity of the model to rank individuals according to their ITE. In order to make sure that learned models perform better than random, we substract the AUUC of a random model, obtaining ΔAUUC. Finally to scale the performance of the latter models, we report in Figure 4 results for an Oracle model that has access to the drawn (β, γ), and for ITE_{best}, the best possible learnt model without exploiting the observable interference M (it predicts for each x its empirical ITE average based on the training set). Again, variance is estimated by repeating experiment with 51 random test/train splits. Figure 4 assesses the performance of MITE estimators versus ITE, using the ΔAUUC metric. They yield a higher ΔAUUC in more than 90% of the random splits. Moreover MITE estimators are close to the Oracle (best model possible) as the Oracle does not significantly outperform them, note that even the Oracle can misrank users because the validation set is noisy and empirical ITEs do not always follow the expected ranking. Besides, Figure 4 does not show any limitation of the 2M and SDR models, but rather highlight the ineffectiveness of such direct ITE estimators if a high interference is observed. This phenomenon can be improved by our post-mediation approach thanks to the higher signal of the causal effect of M on Y. Of course,

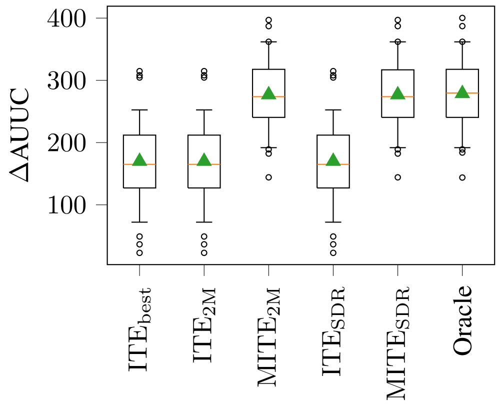

Figure 4: **Baseline Experiment (Simulation)**. ΔAUUC (higher is better) of two ITE models, corresponding MITE models and Oracle model (theoretical thruth). Box plots are done on 51 random splits, whiskers at 5/95 percentiles. Note how MITE systematically increases AUUC of base estimators.

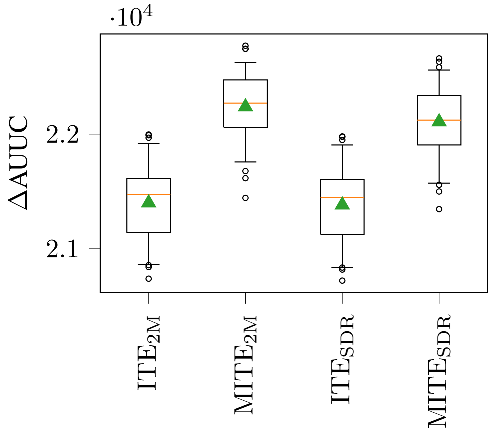

Figure 5: **Real-world Experiment (CRITEO-UPLIFT1)**. ΔAUUC (higher is better) on test for two ITE models and corresponding MITE models. Box plots computed on 51 random splits, whiskers at 5/95 percentiles. Note the higher ΔAUUC and reduced variance of MITE models.

this synthetic data encodes a simpler setting than real-world data, but the fact that our proposed post-mediation approach performs that high still confirms our theoretical analysis.

Real-world Experiment (CRITEO-UPLIFT1)

To qualify the benefit of MITE versus ITE for real-world applications we report ΔAUUC on the CRITEO-UPLIFT1 dataset. We study four methods: two ITE models ($\mathrm{ITE_{2M}}$, and $\mathrm{ITE_{SDR}}$) and their MITE variants (resp. $\mathrm{MITE_{2M}}$, and $\mathrm{MITE_{SDR}}$). For the additional $\hat{\gamma}(x)$ model care is taken to weight the LLH by class as there is a high imbalance in this dataset. Best hyper-parameter found from the grid search being the Cartesian product of $\{\mathrm{L1}, \mathrm{L2}\}$ (regularization) and $\{0.01, 1, 10^2, 10^5\}$ (C, inverse of regularization strength), is L1 regularization and $C = 100$. Results are presented on Figure 5.

The MITE version of each models reduces the variance of the ΔAUUC estimate. This was expected and somehow justified in Proposition 3. Moreover, although the confidence intervals are slightly superposed, MITE always outperforms its ITE counterpart on the 51 splits.

6 Conclusion and Future Works

We propose a novel approach on individual treatment effect (ITE) estimation exploiting observable interference between the treatment assignation and its effect.
Using the structural causal model framework, we define assumptions under which the ITE can be expressed as a product of the individual treatment effect if treated (ITET) and the individual level of interference. In this setting, our post-mediation individual treatment effect (MITE) estimator is consistent. Moreover its asymptotic variance improves with the level of interference. Experimentally, we show how the performance of several baseline ITE estimators improve when plugged in the MITE meta-estimator. We also observe the relationship between performance and interference as predicted by our theoretical results.
Finally, this work opens several perspectives among which: i) stability of our results under variations of assumptions, ii) bound tightness and properties in high-dimensional contexts, and iii) exploration of how representation learning approaches may uncover by themselves MITE-like estimator decomposition under weaker causal assumptions.

REFERENCES

Ahmed Alaa and Mihaela Van Der Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *Proceedings of the International Conference on Machine Learning*, 2018.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 2016.

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2), 2019.

Artem Betlei, Eustache Diemert, and Massih-Reza Amini. Uplift prediction with dependent feature representation in imbalanced treatment and control conditions. In *Proceedings of the International Conference on Neural Information Processing*, 2018.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2018.

Floris Devriendt, Darie Moldovan, and Wouter Verbeke. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big data*, 6(1), 2018.

Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop*, 2018.

Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24), 2011.

Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2), 2019.

Daniel N Hill, Robert Moakler, Alan E Hubbard, Vadim Tsemekhman, Foster Provost, and Kiril Tsemekhman. Measuring causal impact of online actions via natural experiments: Application to display advertising. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2015.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 1994.

Daniel Jacob, Wolfgang Karl Härdle, and Stefan Lessmann. Group average treatment effects for observational studies. *arXiv preprint arXiv:1911.02688*, 2019.

Nathan Kallus. Classifying treatment responders under causal effect monotonicity. *arXiv preprint arXiv:1902.05482*, 2019.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10), 2019.

Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. Support vector machines for differential prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2014.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. *arXiv preprint arXiv:1905.05824*, 2019.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *Proceedings of the International Conference on Machine Learning*, 2019.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12, 2011.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Nicholas J Radcliffe and Patrick D Surry. Real-world uplift modelling with significance-based uplift trees. *Stochastic Solutions*, 2011.

Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling. In *Proceeding of the International Conference on Data Mining*. IEEE, 2010.

Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2), 2012.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the International Conference on Machine Learning*, 2017.

Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. Machine learning estimation of heterogeneous treatment effects with instruments. In *Proceedings of Advances in Neural Information Processing Systems*, 2019.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 2018.

Yu Xie, Jennie E Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1), 2012.

Ikko Yamane, Florian Yger, Jamal Atif, and Masashi Sugiyama. Uplift modeling from separate labels. In *Proceedings of Advances in Neural Information Processing Systems*, 2018.

A PROOFS

A.1 PROPOSITION 1

We remind that we define the structural causal model $\mathfrak{C} = (\mathbb{S}, \mathbb{P}_{\mathbf{N}})$, which is henceforth assumed to represent the causal mechanisms underlying the variables of interest in our work.
$\mathbb{S}$ is defined in Equations (1):

$$\begin{aligned} T &= \tilde{N}_T \\ U &= N_U \\ X &= f_X(U, N_X) \\ M &= f_M(X, N_M) \times T \\ Y &= f_Y(X, M, U, N_Y). \end{aligned} \tag{1}$$

$\mathbb{P}_{\mathbf{N}}$ satisfies the following mild conditions: N_U, N_X, N_M, N_Y are noise consistent with variables definitions, and $\tilde{N}_T$ is distributed according to a Bernoulli distribution with parameter $p = \mathbb{P}^{\mathfrak{C}}(T)$, consistent with a randomized controlled experiment setting.

The associated causal graph $\mathcal{G}_{\mathfrak{C}}$ is given in Figure 1

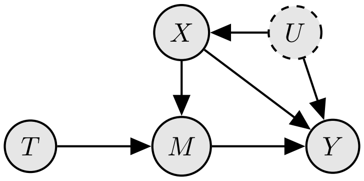

Figure 1: Causal graph $\mathcal{G}_{\mathfrak{C}}$ induced by the SCM $\mathfrak{C}$

In the next proposition, we list four assumptions implied by the $\mathfrak{C}$ about the variables of interest.

Proposition 1 *The SCM $\mathfrak{C}$ defined in Equations 1 implies the following assumptions on variables T, M, Y and X:*

$$\begin{aligned} \textit{(Randomized treatment)} \quad & T \perp\!\!\!\perp X \\ \textit{(Exclusive mediation)} \quad & T \perp\!\!\!\perp Y \mid \{X, M\} \\ \textit{(Strong mediation monotonicity)} \quad & T = 0 \Rightarrow M = 0 \\ \textit{(Valid covariate adjustment)} \quad & \{X\} \textit{ is a VAS for } (M, Y) \end{aligned} \tag{2}$$

Proof.
The proof of Proposition 1 relies on valid adjustment set, the back-door criterion and the definition of d-separation and the Markov property defined in (Pearl, 2009; Peters et al., 2017). First of all our SCM is Markovian (Proposition 6.31 of Peters et al. (2017)) because we assume that the distributions are induced by the causal graph.

Randomized treatment, is implies by the Markov property (Proposition 6.21 of Peters et al. (2017)), *i.e.* d-separation (Definition 6.1 of Peters et al. (2017)) implies independence. Indeed T and X are d-separated by the empty set (all paths between T and X have either $\rightarrow M \leftarrow$ or $\rightarrow Y \leftarrow$ which are blocked by not including neither M nor Y)

Exclusive mediation is also implied by the Markov property and a d-separation. It corresponds to the d-separation of T and Y by the set $\{X, M\}$. This d-separation is shown by listing all paths between T and Y and observing that they are all blocked by the set $\{X, M\}$.

Strong mediation monotonicity is straightforwardly implied by the structural assignment of M given in Equations (1).

Valid covariate adjustment assumption relies on the back-door criterion for valid adjustment set (Proposition 6.41 and definition 6.38 of Peters et al. (2017)). X satisfies the condition because it is not a descendant of M and it blocks all paths from M to Y that enter Y through the backdoor.

■

A.2 Lemma 1

The proposed method exploits the mediation variable M, *i.e.* the treatment acceptation, by splitting the treatment to outcome path into a product of two *subpaths*, both with a higher noise-to-signal ratio. In particular, based on the causal graphical model 1, we can integrate M into the $\mathbb{P}^{\mathfrak{C};do(T)}(Y|x)$ as presented in the next Lemma.

Lemma 1 *Assuming $\mathfrak{C}$, and for any $x \in \mathcal{X}$, the positive outcome probability under treatment, $\mathbb{P}^{\mathfrak{C};do(T)}(Y|x)$, can be written as follows:*

$$\mathbb{P}^{\mathfrak{C};do(T)}(Y|x) = \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M}) + \mathbb{P}^{\mathfrak{C}}(M|x,T)\Big(\mathbb{P}^{\mathfrak{C}}(Y|x,M) - \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})\Big). \tag{3}$$

Proof.
Assuming the SCM $\mathfrak{C}$ truly describes the relationships between T, X, M, Y, we have:

$$\begin{aligned}
\mathbb{P}^{\mathfrak{C};do(T)}(Y|x) &= \mathbb{P}^{\mathfrak{C};do(T)}(Y,M|x) + \mathbb{P}^{\mathfrak{C};do(T)}(Y,\overline{M}|x) \\
&= \underbrace{\mathbb{P}^{\mathfrak{C};do(T)}(Y|x,M)}_{\mathbb{P}^{\mathfrak{C}}(Y|x,M)}\mathbb{P}^{\mathfrak{C};do(T)}(M|x) + \underbrace{\mathbb{P}^{\mathfrak{C};do(T)}(Y|x,\overline{M})}_{\mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})}\mathbb{P}^{\mathfrak{C};do(T)}(\overline{M}|x) \\
&= \mathbb{P}^{\mathfrak{C}}(Y|x,M)\mathbb{P}^{\mathfrak{C};do(T)}(M|x) + \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})\mathbb{P}^{\mathfrak{C};do(T)}(\overline{M}|x) \\
&= \mathbb{P}^{\mathfrak{C}}(Y|x,M)\mathbb{P}^{\mathfrak{C};do(T)}(M|x) + \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})\underbrace{\mathbb{P}^{\mathfrak{C};do(T)}(\overline{M}|x)}_{1-\mathbb{P}^{\mathfrak{C};do(T)}(M|x)} \\
&= \mathbb{P}^{\mathfrak{C}}(M|x,T)\Big(\mathbb{P}^{\mathfrak{C}}(Y|x,M) - \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})\Big) + \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M}),
\end{aligned}$$

where we used assumptions described Equations (2), that imply: $\mathbb{P}^{\mathfrak{C};do(T)}(\cdot|x,\cdot) = \mathbb{P}^{\mathfrak{C}}(\cdot|x,T,\cdot)$ (*Randomized treatment*), $\mathbb{P}^{\mathfrak{C}}(Y|x,T,M) = \mathbb{P}^{\mathfrak{C}}(Y|x,M)$ (*Mediator full effect channelling*), and the claim follows. ■

A.3 Proposition 2

For all $x \in \mathcal{X}$, we define the individual treatment effect $\tau^{ITE}(x)$, treatment effect if treated $\tau^{ITET}(x)$, as well as the individual non-interference $\gamma(x)$ probability (that we henceforth refer to as *individual compliance* for clarity) as follows:

$$\begin{aligned}
\tau^{ITE}(x) &= \mathbb{P}^{\mathfrak{C};do(T)}(Y|x) - \mathbb{P}^{\mathfrak{C};do(\overline{T})}(Y|x), \\
\tau^{ITET}(x) &= \mathbb{P}^{\mathfrak{C};do(M)}(Y|x) - \mathbb{P}^{\mathfrak{C};do(\overline{M})}(Y|x), \\
\gamma(x) &= \mathbb{P}^{\mathfrak{C};do(T)}(M|x).
\end{aligned} \tag{4}$$

We also define the relative ITET $\beta(x)$ and relative ITE $\alpha(x)$ as:

$$\begin{aligned}
\alpha(x) &= \frac{\mathbb{P}^{\mathfrak{C}}(Y|T,x) - \mathbb{P}^{\mathfrak{C}}(Y|\overline{T},x)}{\mathbb{P}^{\mathfrak{C}}(Y|\overline{T},x)} \\
\beta(x) &= \frac{\mathbb{P}^{\mathfrak{C}}(Y|M,x) - \mathbb{P}^{\mathfrak{C}}(Y|\overline{M},x)}{\mathbb{P}^{\mathfrak{C}}(Y|\overline{M},x)}.
\end{aligned} \tag{5}$$

In Proposition 2, we present a result linking the ITE, the ITET and the interference/compliance probability:

Proposition 2 *Assuming $\mathfrak{C}$, the ITE decomposes as follows:*

$$\tau^{ITE}(x) = \tau^{ITET}(x)\gamma(x) \tag{6}$$

Proof
We have an analogous version of Equation (3) for the term $\mathbb{P}^{\mathfrak{C};do(\overline{T})}(Y|x)$:

$$\mathbb{P}^{\mathfrak{C};do(\overline{T})}(Y|x) = \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M}) + \mathbb{P}^{\mathfrak{C}}(M|x,\overline{T})\Big(\mathbb{P}^{\mathfrak{C}}(Y|x,M) - \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})\Big).$$

Since $\overline{T} \Rightarrow \overline{M}$ (*Mediator strong monotonicity* assumption), we get that, $\forall x \in \mathcal{X}$, $\mathbb{P}^{\mathfrak{C}}(M|x,\overline{T}) = 0$, and finally:

$$\mathbb{P}^{\mathfrak{C};do(\overline{T})}(Y|x) = \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M}).$$

Then:

$$\begin{aligned}
\tau^{ITE}(x) &= \mathbb{P}^{\mathfrak{C};do(T)}(Y|x) - \mathbb{P}^{\mathfrak{C};do(\overline{T})}(Y|x) \\
&= \mathbb{P}^{\mathfrak{C}}(M|x,T)\Big(\mathbb{P}^{\mathfrak{C}}(Y|x,M) - \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})\Big) \\
&\quad + \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M}) - \underbrace{\mathbb{P}^{\mathfrak{C}}(Y|x,\overline{T})}_{\mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})} \\
&= \mathbb{P}^{\mathfrak{C}}(M|x,T)\Big(\mathbb{P}^{\mathfrak{C}}(Y|x,M) - \mathbb{P}^{\mathfrak{C}}(Y|x,\overline{M})\Big).
\end{aligned}$$

which completes the proof.

■

A.4 Proposition 3

Single-stratum setting. We focus on the ITE estimation for a single value x_0 of X, for which we assume to observe n *i.i.d.* samples $\{(x_0, T_i, M_i, Y_i)\}_{1\leq i\leq n}$. In practice, this generalises to any stratum $S \subset \mathcal{X}$ containing x_0 of X for which the adjustment set formula is valid, *i.e.* if the variable $X' \triangleq x_0 I_{X\in S} + X I_{X\neq x_0}$ defines a valid adjustment set for (M, Y).

Notations. Consistently with notations presented in Equations (4) and (5), $\alpha(x_0), \beta(x_0)$ refer respectively to the relative ITE and relative ITET in stratum $\{X = x_0\}$ (and are assumed to be positive in this illustrative setting), and we denote $\hat{\tau}^{ITE}(x_0)$, $\hat{\tau}^{ITET}(x_0)$, $\hat{\gamma}(x_0)$ the respective maximum-likelihood estimators (MLE) of $\tau^{ITE}(x_0)$, $\tau^{ITET}(x_0)$, $\gamma(x_0)$. We define the associated MITE estimator as $\hat{\tau}^{MITE}(x_0) \triangleq \hat{\gamma}(x_0)\hat{\tau}^{ITET}(x_0)$. Lastly, we denote $p_1(x_0) = \mathbb{P}^{\mathfrak{C}}(Y|T, x_0)$.
In the following Proposition, we present an asymptotic bound for the ratio of the standard deviation sd of MITE and ITE estimators.

Proposition 3 *Under $\mathfrak{C}$ defined in Equations (1) with $\mathbb{P}^{\mathfrak{C}}(T) = \frac{1}{2}$, and assuming we observe n i.i.d. samples in stratum $\{X = x_0\}$, we have:*

$$\lim_{n\to\infty} \frac{sd(\hat{\tau}^{MITE})}{sd(\hat{\tau}^{ITE})} \leq \sqrt{\left(\frac{2(1+\beta)}{(1-p_1)(1+\alpha)}\right)\gamma} \tag{7}$$

where we dropped references to x_0 for clarity.

Proof.
The proof is splitted in four steps:

1. Maximum-Likelihood and treatment effect estimators
2. Variance of estimators derivation
3. Variance upper and lower bounds
4. Wrap up

Every random quantity is henceforth implicitly considered to be 'with respect to x_0'.

1. Maximum-Likelihood and treatment effect estimators

We remind that we have n *i.i.d.* samples $\{(T_i, M_i, Y_i)\}_{1\leq i\leq n}$ of variables (T, M, Y), and that we suppose $\mathbb{P}^{\mathfrak{C}}(T) = \frac{1}{2}$.

We define the following compact notations:

$$\begin{aligned}
p_0 &= \mathbb{P}^{\mathfrak{C}}(Y|\overline{T}) \\
p_1 &= \mathbb{P}^{\mathfrak{C}}(Y|T) \\
q_0 &= \mathbb{P}^{\mathfrak{C}}(Y|\overline{M}) \\
q_1 &= \mathbb{P}^{\mathfrak{C}}(Y|M) \\
t &= \mathbb{P}^{\mathfrak{C}}(T) \\
\gamma &= \mathbb{P}^{\mathfrak{C}}(M|T).
\end{aligned}$$

Associated maximum-likelihood estimators (MLE) $\hat{p_0}$, $\hat{p_1}$, $\hat{q_0}$, $\hat{q_1}$, $\hat{t}$ and $\hat{\gamma}$ are given by (ratios of) empirical frequencies. For instance,

$$\begin{aligned}
\hat{t} &= \frac{1}{n}\sum_{i=1}^{n} T_i \\
\hat{\gamma} &= \frac{\frac{1}{n}\sum\limits_{i=1}^{n} M_i}{\frac{1}{n}\sum\limits_{i=1}^{n} T_i} = \frac{1}{\sum\limits_{i=1}^{n} T_i}\sum_{i=1}^{n} M_i \\
\hat{p}_0 &= \frac{\frac{1}{n}\sum\limits_{i=1}^{n}(1-T_i)Y_i}{\frac{1}{n}\sum\limits_{i=1}^{n}(1-T_i)} = \frac{1}{\sum\limits_{i=1}^{n}(1-T_i)}\sum_{i=1}^{n}(1-T_i)Y_i \\
\hat{p}_1 &= \frac{\frac{1}{n}\sum\limits_{i=1}^{n} T_iY_i}{\frac{1}{n}\sum\limits_{i=1}^{n} T_i} = \frac{1}{\sum\limits_{i=1}^{n} T_i}\sum_{i=1}^{n} T_iY_i \\
\hat{q}_0 &= \frac{\frac{1}{n}\sum\limits_{i=1}^{n}(1-M_i)Y_i}{\frac{1}{n}\sum\limits_{i=1}^{n}(1-M_i)} = \frac{1}{\sum\limits_{i=1}^{n}(1-M_i)}\sum_{i=1}^{n}(1-M_i)Y_i \\
\hat{q}_1 &= \frac{\frac{1}{n}\sum\limits_{i=1}^{n} M_iY_i}{\frac{1}{n}\sum\limits_{i=1}^{n} M_i} = \frac{1}{\sum\limits_{i=1}^{n} M_i}\sum_{i=1}^{n} M_iY_i
\end{aligned} \tag{8}$$

Where by convention we consider that $\frac{0}{0} = 0$.
Direct estimators for τ^{ITE}, τ^{ITET} are given by applying the two-model approach to MLEs given in Equations (8), *i.e.*

$$\begin{aligned}
\hat{\tau}^{ITE} &= \hat{p}_1 - \hat{p}_0 \\
\hat{\tau}^{ITET} &= \hat{q}_1 - \hat{q}_0
\end{aligned}$$

and the corresponding τ^{MITE} estimator therefore writes:

$$\hat{\tau}^{MITE} = (\hat{q}_1 - \hat{q}_0)\hat{\gamma}.$$

In what follows, we will now write $\sum_i$ instead of $\sum\limits_{i=1}^{n}$ when there is no ambiguity.

2. Variance of estimators derivation

2.a. $\hat{\tau}^{ITE}$ variance derivation

For any random variables X, Y, we have that

$$Var(X) = Var(\mathbb{E}[X|Y]) + \mathbb{E}[Var(X|Y)]. \tag{9}$$

Using this formula with $X = \hat{\tau}^{ITE} = \hat{p}_1 - \hat{p}_0$ and $Y = \{T_1, \dots, T_n\}$ (denoted $\{T_k\}_k$ for simplicity), we may write:

$$Var(\hat{\tau}^{ITE}) = \mathbb{E}\left[Var(\hat{\tau}^{ITE}|\{T_k\}_k)\right] + Var\left[\mathbb{E}(\hat{\tau}^{ITE}|\{T_k\}_k)\right]. \tag{10}$$

First term of Equation (10)
The term $Var(\hat{\tau}^{ITE}|\{T_k\}_k)$ decomposes as:

$$\begin{aligned} Var(\hat{\tau}^{ITE}|\{T_k\}_k) &= Var(\hat{p}_1 - \hat{p}_0|\{T_k\}_k) \\ &= Var(\hat{p}_1|\{T_k\}_k) + Var(\hat{p}_0|\{T_k\}_k) - 2Cov(\hat{p}_1, \hat{p}_0|\{T_k\}_k). \end{aligned}$$

Now let's handle each one of those three terms, starting with the last one:

$$\begin{aligned} Cov\left(\hat{p}_1, \hat{p}_0|\{T_k\}_k\right) &= Cov\left(\frac{1}{\sum_i T_i}\sum_i T_i Y_i, \frac{1}{\sum_j (1-T_j)}\sum_j (1-T_j)Y_j|\{T_k\}_k\right) \\ &= \frac{1}{\sum_i T_i}\frac{1}{\sum_j (1-T_j)} Cov\left(\sum_i T_i Y_i, \sum_j (1-T_j)Y_j|\{T_k\}_k\right) \\ &= \frac{1}{\sum_i T_i}\frac{1}{\sum_j (1-T_j)} \sum_i \sum_j (1-T_j)T_i \underbrace{Cov\left(Y_i, Y_j|\{T_k\}_k\right)}_{\neq 0 \text{ only if } i=j \text{ (since } \textit{i.i.d.}\text{)}} \\ &= \frac{1}{\sum_i T_i}\frac{1}{\sum_j (1-T_j)} \sum_i (1-T_i)T_i Cov(Y_i, Y_i| \underbrace{\{T_k\}_k}_{T_i \text{ (since } \textit{i.i.d.}\text{)}}) \\ &= \frac{1}{\sum_i T_i}\frac{1}{\sum_j (1-T_j)} \sum_i \underbrace{(1-T_i)T_i}_{=0} Var(Y_i|T_i) \\ &= 0. \end{aligned}$$

Then the first one:

$$\begin{aligned} Var(\hat{p}_1|\{T_k\}_k) &= Var\left(\frac{1}{\sum_i T_i}\sum_i T_i Y_i|\{T_k\}_k\right) \\ &= \left(\frac{1}{\sum_i T_i}\right)^2 Var\left(\sum_i T_i Y_i|\{T_k\}_k\right) \\ &= \left(\frac{1}{\sum_i T_i}\right)^2 \sum_i Var(T_i Y_i| \underbrace{\{T_k\}_k}_{T_i \text{ (since } \textit{i.i.d.}\text{)}}) \\ &= \left(\frac{1}{\sum_i T_i}\right)^2 \sum_i \underbrace{T_i Var(Y_i|T_i)}_{Var(Y|T=1)T_i \text{ (since } \textit{i.i.d.}\text{)}} \\ &= \left(\frac{1}{\sum_i T_i}\right)^2 Var(Y|T=1) \sum_i T_i \\ &= \frac{1}{\sum_i T_i} Var(Y|T=1) \\ &= \frac{1}{\sum_i T_i} p_1(1-p_1). \end{aligned}$$

Analogously, we get for the second term:

$$Var(\hat{p}_0|\{T_k\}_k) = \frac{1}{\sum_i (1-T_i)} p_0(1-p_0).$$

Therefore, we get

$$Var(\hat{\tau}^{ITE}|\{T_k\}_k) = \frac{1}{\sum_i T_i} p_1(1-p_1) + \frac{1}{\sum_i (1-T_i)} p_0(1-p_0). \tag{11}$$

Finally, the first term on the right side of Equation (10) can be written as:

$$\mathbb{E}\left[Var(\hat{\tau}^{ITE}|\{T_k\}_k)\right] = \mathbb{E}\left[\frac{1}{\sum_i T_i}\right] p_1(1-p_1) + \mathbb{E}\left[\frac{1}{\sum_i (1-T_i)}\right] p_0(1-p_0). \tag{12}$$

Second term of Equation (10)
We may write

$$\mathbb{E}(\hat{\tau}^{ITE}|\{T_k\}_k) = \mathbb{E}(\hat{p_1}|\{T_k\}_k) - \mathbb{E}(\hat{p_0}|\{T_k\}_k).$$

Moreover,

$$\begin{aligned}
\mathbb{E}(\hat{p_1}|\{T_k\}_k) &= \mathbb{E}\left(\frac{1}{\sum_i T_i} \sum_i T_i Y_i | \{T_k\}_k\right) \\
&= \frac{1}{\sum_i T_i} \sum_i \mathbb{E}\left(T_i Y_i | \{T_k\}_k\right) \\
&= \frac{1}{\sum_i T_i} \sum_i T_i \underbrace{\mathbb{E}(Y_i|T_i)}_{p_1} \\
&= \frac{1}{\sum_i T_i} \left(\sum_i T_i\right) p_1 \\
&= p_1.
\end{aligned}$$

Analogously we get

$$\mathbb{E}(\hat{p_0}|\{T_k\}_k) = p_0.$$

Therefore $\mathbb{E}(\hat{\tau}^{ITE}|\{T_k\}_k)$ is a constant relatively to the $\{T_k\}$s, and

$$Var\left[\mathbb{E}(\hat{\tau}^{ITE}|\{T_k\}_k)\right] = 0. \tag{13}$$

Wrap up of the variance of $\hat{\tau}^{ITE}$
Combining Equations (10), (12) and (13) we finally get:

$$Var(\hat{\tau}^{ITE}) = \mathbb{E}\left[\frac{1}{\sum_i T_i}\right] p_1(1-p_1) + \mathbb{E}\left[\frac{1}{\sum_i (1-T_i)}\right] p_0(1-p_0). \tag{14}$$

2.B. $\hat{\tau}^{MITE}$ VARIANCE DERIVATION

Using Equation (9) with $X = \hat{\tau}^{MITE} = \hat{\gamma}(\hat{q}_1 - \hat{q}_0)$ and $Y = \{T_1, \ldots, T_n, M_1, \ldots, M_n\}$ (denoted $\{T_k, M_k\}_k$ for simplicity), we may write:

$$Var(\hat{\tau}^{MITE}) = \mathbb{E}\left[Var(\hat{\tau}^{MITE}|\{T_k, M_k\}_k)\right] + Var\left[\mathbb{E}(\hat{\tau}^{MITE}|\{T_k, M_k\}_k)\right]. \tag{15}$$

First term of Equation (15)
Using the fact that $\hat{\tau}^{MITE} = \hat{\gamma}(\hat{q}_1 - \hat{q}_0)$, and remarking that $\mathbb{E}[\hat{\gamma}|\{T_k, M_k\}_k] = \hat{\gamma}$, we may write:

$$Var(\hat{\gamma}(\hat{q}_1 - \hat{q}_0)|\{T_k, M_k\}_k) = \hat{\gamma}^2 Var((\hat{q}_1 - \hat{q}_0)|\{T_k, M_k\}_k)$$

By analogy with Equation (11), we have

$$Var(\hat{q}_1 - \hat{q}_0)|\{T_k, M_k\}_k) = \frac{1}{\sum_i M_i} q_1(1-q_1) + \frac{1}{\sum_i (1-M_i)} q_0(1-q_0),$$

which gives

$$Var(\hat{\gamma}(\hat{q}_1 - \hat{q}_0)|\{T_k, M_k\}_k) = \hat{\gamma}^2 (\frac{1}{\sum_i M_i} q_1(1-q_1) + \frac{1}{\sum_i (1-M_i)} q_0(1-q_0)). \tag{16}$$

Replacing $\hat{\gamma} = \frac{\sum_i M_i}{\sum_i T_i} = \frac{\sum_i M_i T_i}{\sum_i T_i}$ in (16), we get:

$$Var(\hat{\gamma}(\hat{q}_1 - \hat{q}_0)|\{T_k, M_k\}_k) = \frac{\sum_i M_i}{\left(\sum_i T_i\right)^2} q_1(1-q_1) + \frac{\left(\sum_i M_i\right)^2}{\left(\sum_i T_i\right)^2 \sum_i (1-M_i)} q_0(1-q_0).$$

The first term of Equation (15) then writes:

$$\mathbb{E}\left[Var(\hat{\tau}^{MITE}|\{T_k, M_k\}_k)\right] = \mathbb{E}\left[\frac{\sum_i M_i}{\left(\sum_i T_i\right)^2}\right] q_1(1-q_1) + \mathbb{E}\left[\frac{\left(\sum_i M_i\right)^2}{\left(\sum_i T_i\right)^2 \sum_i (1-M_i)}\right] q_0(1-q_0). \tag{17}$$

Second term of Equation (15)
First, we have

$$\begin{aligned}
\mathbb{E}(\hat{\tau}^{MITE}|\{T_k, M_k\}_k) &= \mathbb{E}(\hat{\gamma}(\hat{q}_1 - \hat{q}_0)|\{T_k, M_k\}_k) \\
&= \hat{\gamma}\mathbb{E}(\hat{q}_1 - \hat{q}_0|\{T_k, M_k\}_k) \\
&= \hat{\gamma}\left(\mathbb{E}(\hat{q}_1|\{T_k, M_k\}_k) - \mathbb{E}(\hat{q}_0|\{T_k, M_k\}_k)\right).
\end{aligned}$$

Then a few computations lead to

$$\begin{aligned}
\mathbb{E}(\hat{q}_1|\{T_k, M_k\}_k) &= \mathbb{E}\left(\frac{1}{\sum_i M_i}\sum_i M_i Y_i|\{T_k, M_k\}_k\right) \\
&= \frac{1}{\sum_i M_i}\mathbb{E}\left(\sum_i M_i Y_i|\{T_k, M_k\}_k\right) \\
&= \frac{1}{\sum_i M_i}\sum_i M_i \mathbb{E}\left(Y_i|\{T_k, M_k\}_k\right) \\
&= \frac{1}{\sum_i M_i}\sum_i \underbrace{M_i \mathbb{E}\left(Y_i|T_i, M_i\right)}_{M_i\mathbb{E}[Y|M=1]} \\
&= \underbrace{\mathbb{E}[Y|M=1]}_{q_1}\frac{1}{\sum_i M_i}\sum_i M_i \\
&= q_1
\end{aligned}$$

We can use analogous computations to get

$$\mathbb{E}(\hat{q}_0|\{T_k, M_k\}_k) = q_0.$$

Regrouping these results we have

$$\mathbb{E}(\hat{\tau}^{MITE}|\{T_k, M_k\}_k) = \hat{\gamma}(q_1 - q_0),$$

which gives the following expression for the second term of (15)

$$Var\left(\mathbb{E}(\hat{\tau}^{MITE}|\{T_k, M_k\}_k)\right) = (q_1 - q_0)^2 Var(\hat{\gamma}).$$

Using the fact that $Var(\hat{\gamma}) = \mathbb{E}(Var(\hat{\gamma}|\{T_k\})) + Var(\mathbb{E}[\hat{\gamma}|\{T_k\}])$, we can proceed similarly to step 1 to get

$$Var(\hat{\gamma}) = \mathbb{E}\left(\frac{1}{\sum_i T_i}\right)\gamma(1-\gamma).$$

This gives the final expression for the second term of (15):

$$Var\left(\mathbb{E}(\hat{\tau}^{MITE}|\{T_k, M_k\}_k)\right) = (q_1 - q_0)^2\mathbb{E}\left(\frac{1}{\sum_i T_i}\right)\gamma(1-\gamma). \tag{18}$$

Wrap up of the variance of $\hat{\tau}^{MITE}$
Combining Equations (15), (17) and (18), we have:

$$Var(\hat{\tau}^{MITE}) = \mathbb{E}\left[\frac{\sum_i M_i}{\left(\sum_i T_i\right)^2}\right] q_1(1-q_1) + \mathbb{E}\left[\frac{\left(\sum_i M_i\right)^2}{\left(\sum_i T_i\right)^2 \sum_i (1-M_i)}\right] q_0(1-q_0) + (q_1-q_0)^2\mathbb{E}\left(\frac{1}{\sum_i T_i}\right)\gamma(1-\gamma). \tag{19}$$

3. Asymptotic variance upper and lower bounds

3.A Asymptotic lower bound of $Var(\hat{\tau}^{ITE})$ With a slight rewriting of (14),

$$nVar(\hat{\tau}^{ITE}) = \mathbb{E}\left[\frac{1}{\frac{1}{n}\sum_i T_i}\right] p_1(1-p_1) + \mathbb{E}\left[\frac{1}{\frac{1}{n}\sum_i (1-T_i)}\right] p_0(1-p_0).$$

Using the law of large numbers , we get

$$\lim_{n\to\infty} nVar(\hat{\tau}^{ITE}) = 2\left(p_1(1-p_1) + p_0(1-p_0)\right). \tag{20}$$

Where we remind we have supposed $t = \mathbb{P}^{\mathfrak{C}}(T) = \frac{1}{2}$ for simplicity.
Now, using $p_1 = (1+\alpha)p_0$, and with the assumption $\alpha \geq 0$, we have:

$$\begin{aligned}\lim_{n\to\infty} nVar(\hat{\tau}^{ITE}) &= 2p_0\left(1 - p_0 + (1+\alpha)(1-p_1)\right)\\ &\geq 2p_0\left(1 - p_1 + (1+\alpha)(1-p_1)\right)\\ &= 2p_0(1-p_1)(2+\alpha).\end{aligned}$$

In summary, we have the following asymptotic lower bound for $Var(\hat{\tau}^{ITE})$:

$$\lim_{n\to\infty} nVar(\hat{\tau}^{ITE}) \geq 2p_0(1-p_1)(2+\alpha). \tag{21}$$

3.A Asymptotic upper bound of $Var(\hat{\tau}^{MITE})$
With a slight rewriting of (19),

$$\begin{aligned}nVar(\hat{\tau}^{MITE}) = \mathbb{E}\left[\frac{\frac{1}{n}\sum_i M_i}{\left(\frac{1}{n}\sum_i T_i\right)^2}\right] q_1(1-q_1) + \mathbb{E}\left[\frac{\left(\frac{1}{n}\sum_i M_i\right)^2}{\left(\frac{1}{n}\sum_i T_i\right)^2 \frac{1}{n}\sum_i (1-M_i)}\right] q_0(1-q_0)\\ + (q_1-q_0)^2\mathbb{E}\left(\frac{1}{\frac{1}{n}\sum_i T_i}\right)\gamma(1-\gamma).\end{aligned} \tag{22}$$

Using the law of large numbers , we get

$$\lim_{n\to\infty} nVar(\hat{\tau}^{MITE}) = 2\gamma q_1(1-q_1) + 2\frac{\gamma^2}{2-\gamma}q_0(1-q_0) + 2\gamma(1-\gamma)(q_1-q_0)^2. \tag{23}$$

Now, using that for any $q \in [0,1]$, $q(1-q) \leq q$, and reminding that $q_1 = (1+\beta)q_0 \leq 1$ where $\beta \geq 0$ by assumption, we get:

$$\begin{aligned}\lim_{n\to\infty} nVar(\hat{\tau}^{MITE}) &= 2\gamma q_1(1-q_1) + 2\frac{\gamma^2}{2-\gamma}q_0(1-q_0) + 2\gamma(1-\gamma)(q_1-q_0)\\ &\leq 2\gamma\left(\underbrace{q_1}_{\beta q_0} + \underbrace{\frac{\gamma}{2-\gamma}}_{\leq\gamma\leq 1}q_0 + \underbrace{(q_0(1+\beta))^2}_{\leq q_0\beta}\right)\\ &\leq 2\gamma q_0\left(1+\beta+1+\beta\right)\\ &= 4q_0\gamma(1+\beta).\end{aligned}$$

In summary, we have the following asymptotic upper bound for $Var(\hat{\tau}^{MITE})$:

$$\lim_{n\to\infty} nVar(\hat{\tau}^{MITE}) \leq 4q_0\gamma(1+\beta). \tag{24}$$

4. Wrap up

Combining Equations (21) and (24) (ratio of positive values), we get

$$\begin{aligned}\frac{\lim\limits_{n\to\infty} nVar(\hat{\tau}^{MITE})}{\lim\limits_{n\to\infty} nVar(\hat{\tau}^{ITE})} &\leq \frac{4q_0\gamma(1+\beta)}{2p_0(1-p_1)(2+\alpha)}\\ &= 2\frac{1+\beta}{(1-p_1)(2+\alpha)}\gamma.\end{aligned}$$

Where we remind that *Strong mediation monotonicity* and *Exclusive mediation* imply straightforwardly that $p_0 = q_0$ (as shown in the beginning of the proof of Proposition 2). Since the limits of both the numerator and denominator exist, this implies that

$$\lim_{n \to \infty} \frac{Var(\hat{\tau}^{MITE})}{Var(\hat{\tau}^{ITE})} \leq 2 \frac{1+\beta}{(1-p_1)(2+\alpha)} \gamma.$$

Taking the square root of this equation gives the wanted result.

■

References

Pearl, J. *Causality*. Cambridge university press, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.