Changes of representation for counter-factual inference

Armand Lacombe

université PARIS-SACLAY

Changes of representation for counter-factual inference

Changements de représentation pour l'inférence contrefactuelle

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 : Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique. Référent : Faculté des sciences d'Orsay

Thèse préparée au **Laboratoire Interdisciplinaire des Sciences du Numérique** (Université Paris-Saclay, CNRS) sous la direction de **Michèle SEBAG**, directrice de recherche, et le co-encadrement de **Philippe CAILLOU**, maître de conférences

Thèse soutenue à Paris-Saclay, le 5 mars 2024, par

Armand LACOMBE

Composition du jury

Membres du jury avec voix délibérative

Jean-Pierre NADAL Directeur de recherche, CNRS, EHESS	Président
Marianne CLAUSEL Professeur, Institut Élie Cartan, Université de Lorraine	Rapporteur & Examinatrice
Hervé ISAMBERT Directeur de recherche, Institut Curie, CNRS, Sorbonne Université	Rapporteur & Examinateur
Julie JOSSE Directrice de recherche, Inria, Université de Montpellier	Examinatrice
Éric GAUSSIER Professeur, Université Grenoble-Alpes	Examinateur

THESE DE DOCTORAT
NNT : 2024UPASG009

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Changements de représentation pour l'inférence contrefactuelle
Mots clés: inférence causale, apprentissage de représentations, intelligence artificielle, apprentissage profond

Résumé: Quelques éloignées puissent-elles paraitre, les notions de prise de décision algorithmique, d'évaluation des politiques publiques ou de personnalisation des soins médicaux reposent sur une même question fondamentale : que se serait-il passé, que se passerait-il si la décision était autre ? Parce que l'apprentissage causal fonde par essence des raisonnements contrefactuels sur les données disponibles, il constitue le cadre théorique et pratique idoine de ces problématiques.

Depuis l'introduction de méthodes fondées sur des réseaux de neurones, les progrès en inférence causale ont été portés principalement par le raffinement de l'équilibrage entre les représentations apprises des individus contrôles, et traités. Prenant constat des limites de cette approche, nous opérons un changement de paradigme. Des contraintes asymétriques dans l'espace des représentations permettent, au prix de la dégradation de la modélisation factuelle d'une population, l'amélioration de la modélisation contrefactuelle de l'autre. La combinaison d'un modèle favorable à la population traitée avec son pendant relatif à la population contrôle cumule leurs avantages, sans leurs inconvénients.

Cette nouvelle architecture est incarnée par ALRITE, un modèle dont nous démontrons la pertinence sur un plan théorique, avant de le soumettre à des expériences pratiques. Un soin tout particulier est porté à la sélection rigoureuse de ses hyper-paramètres, tâche réputée délicate dans le domaine de l'inférence causale. Une comparaison favorable avec les modèles concurrents de l'état de l'art confirme finalement le bien-fondé de l'approche.

Title: Changes of representation for counter-factual inference
Keywords: causal inference, representation learning, artificial intelligence, deep learning

Abstract: Causal learning defines a new frontier for supervised machine learning, offering better robustness w.r.t. out-of-distribution prediction, and directly answering -at least in principle- the policymakers' questions: what will happen upon making interventions? what would have happened if these interventions had not occurred?

This manuscript focuses on Conditional Average Treatment Effect estimation (*CATE*), aimed to assess the direct impact of an intervention at the level of a subpopulation, based on observational data. While there are growing applicative needs for *CATE* estimation, e.g. from the field of personalized medicine or marketing, it faces two intertwined challenges from a supervised learning viewpoint. Firstly, causal learning is notorious for leveraging few data. Secondly and most importantly, the observational data consists of two distinct distributions, the control and the treated one, usually drawn from different covariate distributions. In other words, the goal is to infer "what would have happened" if the intervention had or had not taken place in this particular case (counter-factual estimation), i.e. to solve a problem of learning with missing not-at-random data.

The state of the art has long relied on the flexibility of neural network-based model learning, fuelled by the search for latent representations that enforce an appropriate balance between the control and the treated distributions.

The proposed approach is based on the claim, first formulated by Zhang et al. (2020), that balance in the latent space is *not* the most relevant property. Instead, the counter-factual estimate for say control samples requires that control samples have treated neighbors that are sufficiently close (in latent space); and symmetrically, treated samples must have close control neighbors.

The key contribution of the approach lies in the formulation of constraints that enforce these properties, and their consequence in terms of model architecture: the two requirements can be better addressed by *two* latent spaces, respectively enforcing the quality of the factual and counter-factual outcome models for control and treated samples.

The proposed architecture is instantiated by Asymmetrical Latent Representation for Individual Treatment Effect (ALRITE), acknowledging the asymmetrical constraints related to the control and treated distributions. The ALRITE relevance is grounded in theory through a comprehensive analysis, bounding the estimation error in a way that is both related with the terms of the compound loss, and accessible to the practitioner. The merits of the approach are also established by extensive experimental validation and comparison with the state-of-the-art baselines.

Another contribution of the proposed work is related with the selection of hyper-parameters for the (own and other) approaches: this problem, widely acknowledged in the field of Machine Learning, raises additional and subtle difficulties in the causal inference setting.

Contents

I - Introduction

I.1 . Context

In parallel with the amazing achievements of Machine Learning (Bojarski et al., 2016; Brown et al., 2020; Jumper et al., 2021), concerns are increasingly voiced about the multiple challenges it faces. Some of them regard the **robustness** of the learned models *per se*, for instance in an *out of distribution* setting (Hendrycks et al., 2021), in the presence of *corrupted data* (Horowitz and Manski, 1995), or when confronted with adversarial examples (Madry et al., 2017). Other ones regard the **societal impact** of machine learning and more generally artificial intelligence: using predictive models to achieve decision-making entails risks of unethical consequences (O'Neil, 2016; Hardt and Recht, 2022)[1].

Along what seems to be an utterly different line, the demand for **data-driven policies** (Athey, 2017) and **personalized recommendations** (Ozer et al., 2020) grows. Surely Randomized Control Trials (Rubin, 1978; Meldrum, 2000) constitute the gold standard for studying causal links, albeit their scope is limited by numerous limitations such as ethics, cost, or even feasibility (Pearl, 2009). Approaches that do not rely on interventions in the real world are however confronted with the intrinsic difficulty of drawing causal conclusions from observations alone. *Cum hoc ergo propter hoc*[2] warns against a common fallacy. Without great care and specific attention, reverse causality (Rivera and Currais, 1999), spurious relationships (Simon, 1954), bidirectional causality (Richardson, 1996), and mere coincidences (Sapsford and Jupp, 2006) are difficult to distinguish from true causal phenomena.

At the **crossroad of both perspectives** lies the field of causality (Pearl, 2009; Peters et al., 2017). Causality intends to model causal links from the observation of data and, under reasonable assumptions, can guarantee the robustness of its findings. As such, causal models are meant to represent a phenomenon at three levels of abstraction (Pearl, 2009). The first level coincides with predictive reasoning, imputing missing information in a

[1] acknowledging the existence of many other challenges, including but not limited to: fairness (Mehrabi et al., 2021), privacy (Carlini et al., 2021), impact on employment (Makridakis, 2017; Ernst et al., 2019), sovereignty (Calderaro and Blumfelde, 2022), ecological impact (Dhar, 2020), wealth concentration (Allen, 2017), cybersecurity (Zhang et al., 2022), intellectual property (Fernandez et al., 2023), disinformation (Zellers et al., 2019), lethal autonomous weapons (De Ganay and Gouttefarde, 2020), etc.

[2] with this, therefore because of this

distribution-agnostic way (e.g., predicting labels, though causal models are not the most efficient ones to achieve prediction). The second level, referred to as **interventional reasoning**, aims to estimate the effects of an intervention ("what-if"); the contrast with predictive reasoning is that the intervention modifies the input distribution in complex ways. The third level, referred to as **counter-factual reasoning**, aims to estimate the effects of another type of distribution modification ("what-if-not"). To achieve these goals, causal modeling aims to retain the most robust elementary relations among the domain variables, explaining an effect variable from its direct cause variables. By design, such modular and distribution-agnostic models are robust w.r.t. well-defined operations on the joint variable distribution (e.g., freezing a variable value, a.k.a. intervening on this variable).

So far, the framework of causal modeling has proposed new principles in machine learning and opened new research avenues motivated by endless applications[3]. A question of crucial interest lies in the causal effects of binary treatments. Suppose we consider a given measurable quantity of interest and that the intervention consists in assigning treatment to some individuals while others do not. **At the scale of a population, a group, or an individual: what is the causal impact of the treatment?**

I.2 . Illustration

The renowned **Simpson paradox** illustrates the pitfalls of hasty inference from observational data. Simpson (1951) reports the results of a medical study comparing two protocols aimed at kidney stone removal. The first one, percutaneous nephrolithotomy, is minimally invasive, while the second one, open surgery, is a much heavier procedure.

	open surgery	percutaneous nephrolithotomy
Success rate	78% (273/350)	**83%** (289/350)

Table I.1: Simpson paradox data, reported values are (Total success / Group size).

[3]including (non-exhaustive list): medicine (Höfler, 2005; Kapelner et al., 2021), environment (Hannart et al., 2016), development economics (Chernozhukov et al., 2017), social sciences (Chandra and Krishna, 2021), political sciences (Peterson and Spirling, 2018), epidemiology (Wong and Sabanayagam, 2019), education (Athey and Wager, 2019), marketing (Bottou et al., 2013).

The results as displayed in Table I.1 may suggest at first glance that percutaneous nephrolithotomy constitutes the best option: it apparently offers the best success rate. Formally, the probability of success conditionally to the assignment to percutaneous nephrolithotomy ($^{289}/_{350} \approx 83\%$) is greater than its open surgery counterpart ($^{273}/_{350} \approx 78\%$).

However, this intuition is ill-informed. Not all kidney stone patients are the same: assignment to a given protocol might itself hint at the severity of the situation. In this specific example, a coarse dichotomy distinguishes between smaller and larger stones (more severe cases).

	open surgery	percutaneous nephrolithotomy
Smaller stones	**93%** (81/87)	87% (234/270)
Larger stones	**73%** (192/263)	69% (55/80)

Table I.2: Simpson paradox data, reported values are (total success/group size).

With the additional perspective of Table I.2, open surgery appears as the option with the highest success rate, no matter how large the kidney stones are. This observation may seem to contradict the previous one.

A closer look at the data lifts the paradox: the proportion of patients oriented toward percutaneous nephrolithotomy is higher in the smaller stones cohort ($^{234}/_{357} \approx 66\%$) than in the larger stones one ($^{55}/_{343} \approx 16\%$). Indeed, the size of the stone itself affects the success rate of a given protocol. Abstaining from considering this influence leads to erroneous conclusions.

Note that in this case and without a physician's opinion, it is still impossible to conclude that open surgery offers the highest probability of success to a given patient affected by kidney stones. Consider the imaginary extension of the Simpson data in Table I.3. Here, big hospitals outperform small ones in all situations, no matter the stone size or chosen protocol. The simple fact that they tend to orient their patient toward open surgery, while the small ones resort more frequently to percutaneous nephrolithotomy, swaps the success advantages again. Maybe other factors should be taken into consideration, but no such conclusion may be drawn from the observation of the data alone. Expert knowledge is essential to ensure that all parameters of interest have been considered.

		open surgery	percutaneous nephrolithotomy
Smaller stones	Small hospital	80% (8/10)	**83%** (173/208)
	Big hospital	95% (73/77)	**98%** (61/62)
Larger stones	Small hospital	57% (44/77)	**63%** (41/65)
	Big hospital	80% (148/186)	**93%** (14/15)

Table I.3: Imaginary extension of the Simpson paradox data.

I.3 . Contributions

The contributions presented in the manuscript are related to causal estimation from observational data; they aim to answer the question of whether a (binary) intervention, referred to as treatment, is beneficial at the population, group, or individual level.

This goal, investigated by Rubin (1974); Angrist et al. (1996); Shalit et al. (2017), can be formulated as the modeling of the potential outcomes: what happens if the intervention takes place, and what if-not. These models can indeed be learned using Machine Learning from the observational data, gathering the two sets of so-called control and treated samples. The problem is that the control and treated distributions do not coincide in general, and hasty estimations may be flawed.

For this reason, since (Johansson et al., 2016) many approaches have been developed to enforce the balance of the control and treated distributions, i.e., to ensure that they are similar in some representational space. A common method consists in considering and minimizing the statistical distance between the latent distributions of the control and the treated populations.

The presented approach, Asymmetrical Latent Representation for Individual Treatment Effect (ALRITE), operates a paradigm shift and escapes symmetrical constraints. Specifically, it designs an asymmetrical regularization term, unevenly handling the control and treated populations. The overall ALRITE architecture simultaneously focuses on the counter-factual estimation of the treated population (what if they had not been treated) and that of the treated population (what if they had). To stand out on both tasks, ALRITE relies on two neural modules (pipelines), respectively enforcing that control (resp. treated) samples are not isolated from their nearest treated (resp. control) neighbors

in latent space. The relevance of the whole approach is theoretically established, and the provided theorems justify the original terms of the loss.

Lastly, and importantly for practitioners, the key issue of the hyper-parameter selection is tackled and a detailed methodology is proposed (Chapter VI). This issue, referred to as AutoML in the supervised learning framework Hutter et al. (2019), is likewise essential in the causal estimation framework, all the more than the final result (what would have happened if the treatment assignment had been different) is unknown by design.

I.4 . Organization of the manuscript

This manuscript is divided into six main chapters.

Formal background. First is presented the Neyman-Rubin potential outcomes framework, which constitutes the formal basis underlying the remainder of the manuscript. The objectives of causal inference, the assumptions that permit their pursuit, and the procedures devoted to assessing the approach performance are detailed and discussed. Auxiliary quantities are also introduced, for they ease the formalization and clarity of this manuscript.

State of the art. This chapter first briefly situates the field of causal discovery, aimed at identifying the causal relations that link the observed covariates. Closer to our topic of interest, the chapter presents the field of causal inference aimed at quantifying the average effects of a given treatment at the scale of a whole population. Lastly, we focus on the estimation of causal effects at the individual scale, where the extensive literature can be divided into multiple protocols. ALRITE builds upon the latest ones.

Asymmetrical Latent Representation for Individual Treatment Effect. The chapter delves into the very motivations underlying the development of ALRITE, and inspiring the proposed mechanisms. The main principles and core components are formalized, then the procedure itself is examined: how to define, train, and combine pipelines. A thorough analysis grounds the approach in theory.

Experimentation. Experiments validate the merits of ALRITE in practice. After detailing the considered benchmark and the practical implementation, the chapter presents and discusses the ALRITE performance compared to that of state-of-the-art baselines. The discussion is supported by complementary experiments, illustrating the specifics of the (few) benchmarks in the literature.

Hyper-parameter selection in causal inference. In opposition to the common routine of supervised learning, the selection of adequate hyper-parameters in the field of causal learning is a prickly exercise. The problem is formalized through the concepts of use in the related literature, and put into application in the context of ALRITE.

Conclusion. The main contributions of this manuscript are finally put in perspective. Their versatility opens diverse perspectives for further research, which are finally outlined.

II - Formal background

This chapter introduces the framework of causal inference (Section II.1), aimed at modeling the effect of interventions (e.g. prescribing a medication to a person), contrasting it with causal discovery (where the goal is to determine the causal relations among variables).

The goal of causal inference is to estimate some quantities of interest (Section II.2), focusing on the average or conditional (person-dependent) effects of interventions.

The chapter presents and discusses the general assumptions (Section II.3) of the literature supporting the estimation of the above quantities of interest. The performance indicators, i.e. the metrics used to measure and compare the performance of causal inference models, are thereafter introduced (Section II.5), alongside statistical tools that facilitate the formalization of the problem (Section II.6).

Notations. Following common usage, capital letters denote random variables (e.g., X), script letters denote their domain (X ranges in $\mathcal{X}$), while lowercase letters denote observations. $\hat{A}$ denotes an estimate of quantity A.

II.1 . Potential outcomes framework

The setting considered in the manuscript is based on the famed Neyman-Rubin causal model (Rubin, 2005), also known as **potential outcome framework**[1].

Most generally, the effects of interventions are measured from observational data $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i \in [\![1,n]\!]}$, involving n triples, with each triple describing one sample or individual, where:

1. $\mathbf{x_i}$ stands for the description of the i-th individual, consisting of the continuous or discrete values of d features referred to as observed covariates (e.g., age, blood pressure, comorbidities);
2. $\mathbf{t_i}$ is the value of the treatment assignment, e.g. $t_i = 1$ if the physician decided to prescribe the considered drug to the i-th individual; this sample then belongs to the treated group. Otherwise, $t_i = 0$, and the individual is said to belong to the control group;

[1]The main alternative lies in Pearl's Structural Causal Framework (Pearl, 2009), introducing the *do-operator* to model interventions. Despite the recurrent confusion between Neyman-Rubin's and Pearl's, they are strictly distinct (Lara, 2023). A slight variant, the Functional Causal Model framework, is developed in Section III.1.1.

3. $\mathbf{y_i}$ is the observed outcome for the i-th individual, e.g. their survival, which naturally depends on the treatment t_i. The observed outcome is referred to as **factual outcome**. Of utmost interest is the **counter-factual outcome**, that is, the outcome that would have been observed if t_i had been different; naturally, the counter-factual outcome is not observed.

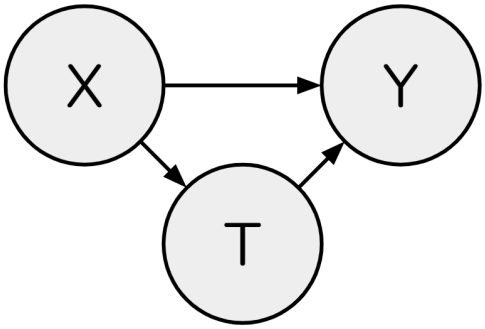


Figure II.1: The causal graph underlying causal inference: Covariate X causes treatment T and potential outcome Y; treatment T determines the factual outcome.

Most generally, the so-called **potential outcome** noted y_i^0 (respectively y_i^1) denotes the value that the outcome variable *would take* for the i-th individual if treatment $t_i = 0$ (resp. $t_i = 1$) were chosen. Accordingly, the potential outcomes consist of the factual, observed outcome, and the counter-factual, unobserved, outcome. The **fundamental problem of causal inference** (Holland, 1986) is that only one of the outcome values y_i^0 and y_i^1 is observed for each i-th sample. Causal inference can thus be cast as an inference problem with (not at random) missing values (Ding and Li, 2018).

Note that the treatment variable does not need to be binary. Outside the scope of the Neyman-Rubin framework, the treatment variable can take values in a continuous interval (Schwab et al., 2019; Nie et al., 2020), be multi-valued (Lopez and Gutman, 2017; Hu et al., 2020), sequential (Bica et al., 2020; Melnychuk et al., 2022) or structured (Pawlowski et al., 2020; Kaddour et al., 2021). For instance, a practitioner may select a combination of drugs in a given pharmacopeia and the corresponding dosages depending on the evolution of a patient's condition. The case of continuous, structured or sequential outcomes is outside the scope of the presented work. Only the binary treatment case will be considered hereunder.

As said, causal inference differs from the causal discovery setting, where the goal is to infer the causal relationships among the covariates (Section III.1).

Causal inference instead considers a very simple, known, causal graph where the set of covariate variables noted X causes the potential outcomes noted Y. X also generally causes the treatment variable T (i.e. the physician decides on the drug depending on the individual's case). Lastly, T governs which one of the potential outcomes Y^0, Y^1 is observed (Fig. II.1). Additional assumptions (Section II.3) notably enforce that the outcomes of the samples are independent ("no spillover").

In supervised machine learning terms, the goal of causal inference is to model both potential outcomes Y^0 and Y^1 (or functions thereof) from the covariate X, based on observational dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i \in [\![1,n]\!]}$.

II.2 . Quantities of interest

Informally, causal inference aims to estimate the *treatment effect* defined as the difference between the two potential outcomes, either at the population level (Average Treatment Effect) or at the subpopulation/sample level (Conditional Average Treatment Effect). In the Neyman-Rubin potential outcome framework, the *ATT*, *CATE*, *ITE* and *ATE* are defined as follows[2] (The said quantities are illustrated in the context of the Simpson paradox in Appendix C.).

The Average Treatment Effect (*ATE*) is the average difference of outcome among the treated and the control populations:

$$ATE = \mathbb{E}[Y^1 - Y^0] \tag{II.1}$$

The Conditional Average Treatment Effect (*CATE*) is defined by the expected difference between the treated and untreated outcomes, conditionally on a given value of the covariate vector X:

$$CATE(x) = \mathbb{E}[Y^1 - Y^0 | X = x] \tag{II.2}$$

Building upon the *CATE* the *ATE* may be rephrased as $\mathbb{E}[CATE(X))]$.

CATE is often denoted τ ($CATE(x) = \tau(x)$), and constitutes the central quantity of interest in the following; our contributions (Chapters IV and V) focus on building accurate estimators thereof. It is commonly mistaken with the individual treatment effect.

[2] In the following, all expectations are implicitly defined with respect to the potentials outcomes distribution $\mathbb{P}_{X,T,Y^0,Y^1}$ that entails the observational one $\mathbb{P}_{X,T,Y}$. This is referred to as the "generalizability assumption" in (Doutreligne and Varoquaux, 2023).

The Individual Treatment Effect (*ITE*) is by definition no distributional quantity, but is specifically associated with a given individual: $ITE_i = Y_i^1 - Y_i^0$.

Although it is often mistaken with *CATE*, the two quantities differ, as discussed by Vegetabile (2021). If y_i is known and under additional assumptions, it is possible[3] to build ITE_i estimates more accurate than $CATE(x_i)$.

The estimation sometimes focuses on the treated population, and the difference between the (factual) treatment effect and the (counter-factual) control effect, either at the population or at the individual level

The Average Treatment effect on the Treated (*ATT*) is similar to *ATE*, but conditioned on the treated population:

$$ATT = \mathbb{E}[Y^1 - Y^0 | T = 1] \tag{II.3}$$

Other metrics. Causal inference literature also evokes metrics such as **Conditional Average Treatment on the Treated (*CATT*)**, which differs from *CATE* as it is conditioned on the treated subset of the population, and **Average Treatment on the Control (*ATC*)** which differs from *ATT* as it is conditioned on the control subset of the population. These quantities will not be considered in the remainder of the manuscript.

When considering a finite observational dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i \in [\![1,n]\!]}$, empirical *ATE* and *ATT* are sometimes referred to as **Sample Average Treatment**

[3] Consider the following setting:

$$\eta : \mathcal{X} \to \{0,1\},\ \mu^0, \mu^1 : \mathcal{X} \to \mathbb{R}$$
$$\begin{cases} X \sim \mathit{Unif}([0,1]^d) \\ T \sim \mathit{Ber}(\eta(X)) \\ N_Y \sim \mathcal{N}(0,1) \\ (Y^0, Y^1) = \left(\mu^0(X) + N_Y, \mu^1(X) - N_Y\right) \end{cases}$$

Here the individual treatment effect for individual i may be expressed using factual data only:

$$\begin{aligned} ITE_i &= y_i^1 - y_i^0 \\ &= \mu^1(x_i) - \mu^0(x_i) - 2n_{Y,i} \\ &= \mu^1(x_i) - \mu^0(x_i) - 2(t_i(\mu^1(x_i) - y_i^1) + (1 - t_i)(-\mu^0(x_i) + y_i^0)) \\ &= \mu^1(x_i) - \mu^0(x_i) - 2(t_i(\mu^1(x_i) - y_i) + (1 - t_i)(-\mu^0(x_i) + y_i)) \\ &= (2t_i - 1) \times (2y_i - \mu^0(x_i) - \mu^1(x_i)) \end{aligned}$$

while the *CATE* given covariate vector x_i is $CATE(x_i) = \mu^1(x_i) - \mu^0(x_i)$.

Effect (*SATE*) and **Sample Average Treatment effect on the Treated (*SATT*)**:

$$SATE = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0)$$
$$SATT = \Big(\sum_{t_i=1} 1\Big)^{-1}\sum_{t_i=1}(y_i^1 - y_i^0)$$

Note that *CATT* and *CATE* are equal under the conditional exchangeability assumption, detailed in Section II.3.1 (although *ATT* and *ATE* are not necessarily equal).

For the sake of simplicity and when clear from the context, we shall use same notations for the distributional and empirical quantities in the following.

II.3 . Assumptions

The causal inference literature aims to estimate the above quantities of interest from observational data and to assess the accuracy of these estimates. The feasibility of the estimation and its accuracy depend on the relations between the covariate and the treatment variables, and more specifically on their conditional independence. Most work related to causal inference (Shalit et al., 2017; Alaa and Schaar, 2018; Du et al., 2021) are based on the following three assumptions.

II.3.1 . Conditional Exchangeability

The conditional exchangeability assumption states that for any $t \in \{0, 1\}$, the treatment assignment T and the potential outcome Y^t are conditionally independent[4] given the observed covariates X.

$$\forall t \in \{0, 1\},\ Y^t \perp\!\!\!\perp T \mid X$$

As said, most related work assumes conditional exchangeability; the importance of this assumption will be illustrated in the example below. Still, this assumption remains fundamentally untestable. As such, and without further knowledge concerning the data generation process, this assumption often amounts to a **leap of faith**. Domain-expert knowledge is usually required in real-life situations. Notably, in the Simpson paradox case (Section I.2), a physician may confirm whether conditioning on the kidney stone size ensures conditional exchangeability.

[4]Conditional Exchangeability is sometimes referred to as *exchangeability* (Wu and Fukumizu, 2021), *weak unconfoundedness* (Hirano and Imbens, 2005), or *conditional independence* in the econometrics field (Lechner, 1999; Angrist et al., 2009).

The state of the art sometimes involves a strictly stronger assumption[5]. This increased strength is however unneeded in general.

$$(Y^0, Y^1) \perp\!\!\!\perp T \mid X$$

variant of Conditional Exchangeability, referred to as **Strong Unconfoundedness** (Imbens, 2000).

II.3.2 . Positivity

The Positivity assumption, also referred to as **Overlap** assumption, states that any treatment may be applied to any individual:

$$\mathbb{P}(X \in \Omega) > 0 \implies 0 < \mathbb{P}(T = 0|X \in \Omega), \mathbb{P}(T = 1|X \in \Omega) \; a.s. \qquad \text{(II.4)}$$

A stronger assumption, referred to as **Strict Overlap** assumption (see e.g. D'Amour et al. (2021)) states the existence of bounds on the conditional probability $\mathbb{P}(T = t|X)$.

$$\begin{aligned} &\exists\, 0 < c_{inf}, c_{sup} < 1 \text{ s.t.} \\ &\qquad \mathbb{P}(X \in \Omega) > 0 \implies \forall t \in \{0, 1\},\; c_{inf} < \mathbb{P}(T = t|X \in \Omega) < c_{sup} \end{aligned}$$

Some theoretical results presented in Chapter III invoke the strict overlap assumption.

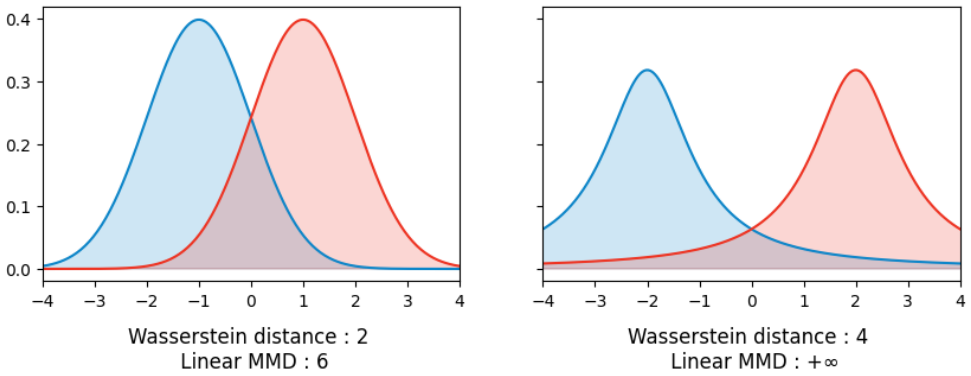


Figure II.2: Distributions of the control and treated populations. Left: normal distribution setting. Right: Cauchy distribution setting.

[5]Let us suppose that Y^0 and T are sampled after two independent Bernoulli distributions, with $Y^1 = T + Y^0$. Then

$$\begin{cases} Y^0 \perp\!\!\!\perp T|X, Y^1 \perp\!\!\!\perp T|X \\ \mathbb{P}\big((Y^0, Y^1) = (1, 1)|X, T = 1\big) = 0 \neq \mathbb{P}\big((Y^0, Y^1) = (1, 1)|X, T = 0\big) \end{cases}$$

The Positivity assumption only requires the density to be strictly greater than 0 for the treated and control populations over the support of both distributions. While quantifying the actual discrepancy between the treatment assignment of both distributions is of crucial importance, there is **no consensus** concerning the most appropriate way to measure it.

Fig. II.2, inspired from Zhang et al. (2020), compares two settings. In the first one, the generated control and target populations are normally distributed; in the second one, both are sampled after Cauchy distributions. The Wasserstein distance (Cuturi, 2013) and the Maximum Mean Discrepancy (*MMD*) (Gretton et al., 2012) between control and treated populations, two common discrepancy measures in the field of causal inference, are reported for both first and second settings.

It is argued that the most favorable setting is the second one (Fig. II.2, right), though it is associated with a higher Wasserstein distance and *MMD*. We shall return to this in Section IV.1.2.

II.3.3 . Stable Unit Treatment Value Assumption

The Stable Unit Treatment Value Assumption (*SUTVA*) firstly states that there exists **no interaction**[6] between units (Cox, 1958), i.e that the treatment assignment of individuals $j \neq i$ has no impact on the potential outcome of individual i. This property is typically violated in settings where individuals compete for a finite resource. For instance, treating candidates by delivering mentoring hours satisfies *SUTVA* if candidates are prepared for an examination; but it does not (admissions are of a fixed number and as such constitute a finite resource) if candidates are prepared for a competition.

Additionally, *SUTVA* states that the outcome verifies **consistency** (Rubin, 1978), meaning that if treatment is assigned to value T, then the observed outcome Y is Y^T. In particular in the Neyman-Rubin framework,

$$Y = (1 - T)\, Y^0 + T\, Y^1$$

II.3.4 . Discussion

II.3.4.1 . Ignorability

Ignorability holds when Conditional Exchangeability and Positivity both hold. As noted by D'Amour et al. (2021), these properties might be **antagonistic**. Typically, Conditional Exchangeability is more likely to hold when the dimension of the covariate vector X increases[7]. In counterpart, Positivity is less likely as the number of dimensions increases, everything else being equal. Without additional hypotheses concerning the data generation process in high dimensional settings, the support of a treatment (respectively

[6] this is sometimes referred to as the "no spillover" hypothesis

[7] With notable restrictions, as noted by Pearl (2011); Wooldridge (2016).

control) subgroup might become so thin that it eventually excludes any control (resp. treatment) sample.

II.3.4.2 . Assumptions relaxation

Conditional exchangeability A significant body of work (Manski, 1990; Díaz and Laan, 2013; Oprescu et al., 2023) analyze the sensitivity of causal estimates to covariate perturbations, that may create **controlled levels of confounding**. This analysis makes it possible to bound the error on the predicted outcomes.

Along a different line of work and outside the Neyman-Rubin framework, Chen et al. (2022) establishes an experimental setting to refute some erroneous causal estimates. This setting assesses the model robustness conditioned using a random subset of covariates.

Positivity Assessing the robustness of the model w.r.t. violations of the overlap assumption is of crucial importance in **large dimensional covariate spaces** (D'Amour et al., 2021). Several approaches are proposed in the literature: i) building metrics insensitive to positivity breaches, such as Li (2019)'s generalized overlap weights; or, ii) trimming non-overlap regions (Yang and Ding, 2018).

Hong et al. (2019) establish convergence results for average and local treatment effects in the Instrumental Variable setting (more in Section III.2.1). Rothe (2017) and Armstrong and Kolesár (2021) study the estimates robustness w.r.t. decreasing levels of overlap, providing confidence intervals. Finally, Wu and Fukumizu (2021)'s *β-Intact-VAE* model provides accurate causal estimates in situations where the latent representation is identifiable.

SUTVA As *SUTVA* is often violated in **real-world applications** (e.g., when individuals compete for a finite resource (Li et al., 2022) or influence each others (Sinclair et al., 2012)), a comprehensive body of work has explored the estimate robustness w.r.t. the interaction between units: Sobel (2006) sets the *partial interference* framework, where samples are divided into groups that do not interact with each other. Laffers and Mellace (2020) provide bounds on *ATE* estimates depending on the share of units affected by *SUTVA* violation. Forastiere et al. (2021) establish analytical expressions measuring bias for naive estimates that do not consider interactions.

II.4 . Causal effects identifiability

As already said, the central difficulty of causal inference is that, in supervised learning terms, the counter-factual data are unavailable. The feasibility of causal inference, and the identification of the sought quantities of interest

(Section II.2) thus relies on specific assumptions.

Causal effect identifiability is a property that holds in settings such that there exists a single couple $\left(x \mapsto \mathbb{E}[Y^0|X=x], x \mapsto \mathbb{E}[Y^1|X=x]\right)$ explaining the observed control and treatment distributions in the large sample limit[8] (see Maclaren and Nicholson (2020) for a formal definition and Neal (2010) for an application to causal inference). In other words, for any two admissible potential outcomes distributions $\mathbb{P}_{X,T,Y^0,Y^1}, \tilde{\mathbb{P}}_{X,T,Y^0,Y^1}$ of the setting,

$$\begin{aligned}
&\left(x \mapsto (\mathbb{E}_{\mathbb{P}_{X,T,Y^0,Y^1}}[Y^0|X=x], \mathbb{E}_{\mathbb{P}_{X,T,Y^0,Y^1}}[Y^1|X=x])\right) \\
&\qquad \neq \left(x \mapsto (\mathbb{E}_{\tilde{\mathbb{P}}_{X,T,Y^0,Y^1}}[Y^0|X=x], \mathbb{E}_{\tilde{\mathbb{P}}_{X,T,Y^0,Y^1}}[Y^1|X=x])\right) \\
&\implies \mathbb{P}_{X,T,Y} \neq \tilde{\mathbb{P}}_{X,T,Y}
\end{aligned}$$

In the Neyman-Rubin potential outcomes framework, the **Conditional Exchangeability, Positivity, and *SUTVA* assumptions** ensure that one can compute the average value of Y^t given $X=x$ (referred to as **outcome function** or **surface of contact**), denoted $\mu^t : \mathcal{X} \mapsto \mathbb{R}$:

$$\begin{aligned}
\forall t \in \{0,1\},\ \mathbb{E}[Y^t|X=x] &= \mathbb{E}[Y^t\ |X=x, T=t] \text{ (conditional exchangeability)} \\
&= \mathbb{E}[Y^T|X=x, T=t] \\
&= \mathbb{E}[Y\ \ |X=x, T=t] \text{ (\textit{SUTVA})} \qquad \text{(II.5)}
\end{aligned}$$

The positivity assumption then ensures that the last expression can be estimated using Machine Learning methods for $t \in \{0,1\}$. Under the same assumptions, one has:

$$\begin{aligned}
\mathbb{E}[Y^1 - Y^0|X=x] &= \mathbb{E}[Y|X=x, T=1] - \mathbb{E}[Y|X=x, T=0] \\
\mathbb{E}[Y^1 - Y^0] &= \mathbb{E}\Big[\mathbb{E}[Y|X=x, T=1] - \mathbb{E}[Y|X=x, T=0]\Big]
\end{aligned}$$

[8] Consider for instance the potential outcomes distributions:

$$\mathbb{P}_{X,T,Y^0,Y^1} \textit{ induced by } \begin{cases} X & \sim \textit{Ber}(.5) \\ T & \sim \textit{Ber}(.5) \\ Y^0 & \sim X+T \\ Y^1 & \sim X \end{cases} \qquad \tilde{\mathbb{P}}_{X,T,Y^0,Y^1} \textit{ induced by } \begin{cases} X & \sim \textit{Ber}(.5) \\ T & \sim \textit{Ber}(.5) \\ Y^0 & \sim X \\ Y^1 & \sim X \end{cases}$$

Then the observational distributions entailed by $\mathbb{P}_{X,T,Y}$ and $\tilde{\mathbb{P}}_{X,T,Y}$ are identical, i.e, (X,T,Y) has the same law under both distributions. However,

$$\mathbb{E}_{\mathbb{P}_{X,T,Y^0,Y^1}}[y^0|X=x] = x + .5 \neq x = \mathbb{E}_{\tilde{\mathbb{P}}_{X,T,Y^0,Y^1}}[y^0|X=x]$$

and as such, in that setting causal effects identifiability does not hold. Note also that conditional exchangeability holds in the distribution entailed by $\tilde{\mathbb{P}}_{X,T,Y^0,Y^1}$, but not in that entailed by $\mathbb{P}_{X,T,Y^0,Y^1}$

This remark will be of utmost importance in the following since it provides sufficient conditions to estimate *CATE* and *ATE* (Section II.2).

The reader is referred to White and Chalak (2013) for a more comprehensive discussion of the necessary and sufficient conditions supporting the feasibility of causal identification.

II.5 . Performance indicators

This section presents some indicators, measuring the performance of estimates for the quantities of interest.

Let $\hat{\tau} : \mathcal{X} \to \mathbb{R}$ be an estimate of the true causal effect τ, with $\hat{\mu}^0, \hat{\mu}^1 : \mathcal{X} \to \mathbb{R}$ estimates of the potential outcomes. For the sake of simplicity and when clear from the context, we shall use the same notation for the distributional and empirical measures of the performance indicators. As $(\hat{\mu}^0, \hat{\mu}^1)$ induces a causal effect estimate $\hat{\tau} = \hat{\mu}^1 - \hat{\mu}^0$, the performance indicators defined for $\hat{\tau}$ also apply to pairs $(\hat{\mu}^0, \hat{\mu}^1)$.

II.5.1 . Usual metrics

The absolute error in Average Treatment Effect estimation (ϵ_{ATE}) measures the error in expectation w.r.t. the true *ATE*.

$$\begin{aligned}\epsilon_{ATE}(\hat{\tau}) &= |ATE - \mathbb{E}[\hat{\tau}(X)]| \\ &= |\mathbb{E}[\tau(X) - \hat{\tau}(X)]|\end{aligned}$$

By design, as it involves counter-factual quantities, this indicator cannot be evaluated without additional assumptions.

The Precision in Estimation of Heterogeneous Effect (*PEHE*) introduced by Hill (2011) measures the error in expectation of $\hat{\tau}$ w.r.t. the true *CATE*:

$$PEHE(\hat{\tau}) = \mathbb{E}\left[(\hat{\tau}(X) - \tau(X))^2\right]$$

In general, causal inference benchmarks do not report the *PEHE* directly, but its root square[9] $\sqrt{PEHE}$.

The mean squared error in *ITE* estimation (ϵ_{ITE}) is similar to the *PEHE* except that it measures the error in expectation w.r.t the *ITE* rather than the *CATE*:

$$\epsilon_{ITE}(\hat{\tau}) = \mathbb{E}\left[(\hat{\tau}(X) - (Y^1 - Y^0))^2\right] \tag{II.6}$$

In the common setting where Y^t is assumed to be sampled according to $Y^t = \mu^t(X) + \varepsilon^t$ with ε^t is a noise term such that $\mathbb{E}[\varepsilon^t] = 0$, $\varepsilon^t \perp\!\!\!\perp X$ and $\varepsilon^0 \perp\!\!\!\perp \varepsilon^1$, ϵ_{ITE}

[9] often misleadingly denotes as *PEHE*

may be expressed as:

$$\begin{aligned}\epsilon_{ITE}(\hat{\tau}) &= \mathbb{E}\left[\left(\hat{\tau}(X) - \tau(X) + (\mu^1(X) - \mu^0(X)) - (Y^1 - Y^0)\right)^2\right] \\ &= \mathbb{E}\left[\left(\hat{\tau}(X) - \tau(X)\right)^2\right] + 2\mathbb{E}\left[\left(\hat{\tau}(X) - \tau(X)\right)\left((\mu^1(X) - \mu^0(X)) - (Y^1 - Y^0)\right)\right] \\ &+ \mathbb{E}\left[\left((\mu^1(X) - \mu^0(X)) - (Y^1 - Y^0)\right)^2\right] \\ &= PEHE(\hat{\tau}) + 2\mathbb{E}\left[\left(\hat{\tau}(X) - \tau(X)\right)\left(\varepsilon^0 - \varepsilon^1\right)\right] + \mathbb{E}\left[\left(\varepsilon^0 - \varepsilon^1\right)^2\right] \\ &= PEHE(\hat{\tau}) + (\sigma^0)^2 + (\sigma^1)^2\end{aligned}$$

where σ_t^2 is the variance of ε^t. Up to an additive constant, the two notions coincide.

The policy risk (R_{pol}) is inspired by reinforcement learning (Sutton and Barto, 2018) and more specifically the multi-armed bandits framework (Bubeck and Cesa-Bianchi, 2012). In all generality, it measures the outcome of a binary policy determining the treatment of an individual: $\pi : \mathcal{X} \to \{0, 1\}$:

$$\begin{aligned}R_{pol}(\pi) = 1 &- \mathbb{P}(\pi(X) = 1)\mathbb{E}[Y^1|\pi(X) = 1] \\ &- \mathbb{P}(\pi(X) = 0)\mathbb{E}[Y^0|\pi(X) = 0]\end{aligned}$$

Indeed a causal effect estimate $\hat{\tau}$ is naturally associated with a binary policy $x \in \mathcal{X} \mapsto \mathbb{1}[\hat{\tau}(x)>0]$. Policy risk eventually assesses the ability of causal effect estimates to predict the sign of the causal effects, with larger penalization for errors made in regions where the gap between the two surfaces of contact μ^0, μ^1 is wide.

For any couple of binary policies (π, π'), it comes:

$$\begin{aligned}R_{pol}(\pi) &= 1 - \mathbb{E}[Y^0\mathbb{1}_{[\pi(X)=0]} + Y^1\mathbb{1}_{[\pi(X)=1]}|\pi(X) = 1]\mathbb{P}(\pi(X) = 1) \\ &\quad -\mathbb{E}[Y^0\mathbb{1}_{[\pi(X)=0]} + Y^1\mathbb{1}_{[\pi(X)=1]}|\pi(X) = 0]\mathbb{P}(\pi(X) = 0) \\ &= 1 - \mathbb{E}[Y^0\mathbb{1}_{[\pi(X)=0]} + Y^1\mathbb{1}_{[\pi(X)=1]}] && \textit{(law of total expectation)} \\ &= 1 - \mathbb{E}[\mathbb{E}[Y^0\mathbb{1}_{[\pi(X)=0]} + Y^1\mathbb{1}_{[\pi(X)=1]}|X]] && \textit{(law of total expectation)} \\ &= 1 - \mathbb{E}[\mu^0(X)\mathbb{1}_{[\pi(X)=0]} + \mu^1(X)\mathbb{1}_{[\pi(X)=1]}] && (\mathbb{1}_{[\pi(X)=t]} \textit{ is } \sigma(X)\textit{-measurable}) \\ R_{pol}(\pi) - R_{pol}(\pi') &= \mathbb{E}[\mu^0(X)(\mathbb{1}_{[\pi(X)=1]}\mathbb{1}_{[\pi'(X)=0]} - \mathbb{1}_{[\pi(X)=0]}\mathbb{1}_{[\pi'(X)=1]}) \\ &\quad +\mu^1(X)(\mathbb{1}_{[\pi(X)=0]}\mathbb{1}_{[\pi'(X)=1]} - \mathbb{1}_{[\pi(X)=1]}\mathbb{1}_{[\pi'(X)=0]})] \\ &= \mathbb{E}[\tau(X)(\mathbb{1}_{[\pi(X)=0]}\mathbb{1}_{[\pi'(X)=1]} - \mathbb{1}_{[\pi(X)=1]}\mathbb{1}_{[\pi'(X)=0]})]\end{aligned}$$

In particular, $\pi_\tau = \mathbb{1}_{[\tau(\cdot)\geq 0]}$ is a binary policy, with special property that

$$\begin{cases}\tau(X)\mathbb{1}_{[\pi_\tau(X)=1]} \geq 0 \textit{ a.s.} \\ \tau(X)\mathbb{1}_{[\pi_\tau(X)=0]} \leq 0 \textit{ a.s.}\end{cases}$$

implying

$$R_{pol}(\pi) - R_{pol}(\pi_\tau) \geq 0$$

with equality iff $\pi(X) = \pi_\tau(X)$ almost surely over $\tau^{-1}(\mathbb{R}^*)$. The optimal decision thus boils down to estimating for each individual x whether its treated outcome is greater than its control one.

Let also oR_{pol} denote the **Observational Policy Risk**, defined by

$$\begin{aligned} oR_{pol}(\pi) = 1 &- \mathbb{P}(\pi(X) = 1)\mathbb{E}[Y|\pi(X) = 1, T = 1] \\ &- \mathbb{P}(\pi(X) = 0)\mathbb{E}[Y|\pi(X) = 0, T = 0] \end{aligned}$$

Contrarily to R_{pol}, the expression of oR_{pol} depends only on observational quantities. Under appropriate assumptions (namely, $Y^t \perp\!\!\!\perp T|\pi(X), \forall t \in \{0, 1\}$), the values of R_{pol} and oR_{pol} are equal (more detail in Eq. V.1).

II.5.2 . Evaluation of the metrics

The Monte-Carlo estimates of multiple of the above-mentioned metrics are not feasible. For instance, Y^0 and Y^1 are never observed simultaneously, preventing the computation of the ϵ_{ITE} estimate of a candidate model $\hat{\tau}$: $\widehat{\epsilon_{ITE}(\hat{\tau})} = \left|\frac{1}{n}\sum_{i=1}^{n} \hat{\tau}(x_i) - (y_i^1 - y_i^0)\right|$. The same point holds for notably *PEHE* and ϵ_{ATE}.

This issue is alleviated in some causal inference benchmarks. For instance, the *IHDP* dataset relies on real-world covariate and treatment assignment, but the outcome is simulated; this makes it possible to access the *CATE* (more in Section V.1.1.1).

ϵ_{ITE} and R_{pol} necessitate to know both potential outcomes. As there exist datasets where Y^0 and Y^1 are both known, simulating treatment assignment (e.g. sampling $T|X$ according to some fixed probability distribution) makes it possible to hide outcome Y^0 or Y^1 in a non-random way, and build a causal inference dataset (see dataset Twins in Yao et al. (2018)).

Note that the counter-factual outcome, however, **will never be part of the data that may be leveraged** so as to build the causal effect estimates. A fundamental difficulty in comparing estimates is thus to define appropriate experimental settings. Special care will be devoted to selecting hyper-parameters and benchmarking models (more in Chapter VI).

II.6 . Auxiliary statistics

The state-of-the-art approaches in causal inference involve mainly three auxiliary quantities: the propensity score (and its derivatives), the balancing scores, and the prognostic scores.

The propensity score denoted $\eta(x)$ is defined (Rosenbaum and Rubin, 1983) as the conditional probability of treatment given $X = x$:

$$\eta : x \in \mathcal{X} \mapsto \mathbb{P}(T = 1|X = x) \tag{II.7}$$

Since treatment assignment is always observed in the Neyman-Rubin causal model, the estimation of η boils down to a mainstream supervised learning problem. The estimator most commonly used in the literature is the logistic regression, modeling η by the function $x \in \mathcal{X} \mapsto \left(1 + \exp(-(a^\dagger x + b))\right)^{-1}$ parameterized by $(a, b) \in (\mathbb{R}^d \times \mathbb{R})$.

The propensity score is typically used through propensity weights and inverse propensity weights attached with each i-th sample, respectively defined as $\eta(x_i)$ and $\frac{1}{\eta(x_i)}$.

The Inverse Probability of Treatment Weight (*IPTW*)[10] derives from the propensity score:

$$\varrho^t : x \in \mathcal{X} \mapsto \mathbb{P}(T = t|X = x)^{-1} \tag{II.8}$$

Note that from this definition, it follows that:

$$\varrho^t = \frac{1}{t\eta + (1-t)(1-\eta)} = \frac{t}{\eta} + \frac{1-t}{1-\eta}$$

A common quantity in causal inference is the **augmented inverse probability weight**[11] (*AIPW*) defined as:

$$(2t-1)\varrho^t = \frac{t}{\eta} - \frac{1-t}{1-\eta}$$

Usually, a propensity estimate $\hat{\eta}$ is first built, then $\hat{\varrho}$ is defined as

$$\hat{\varrho}^t = \frac{t}{clip(\hat{\eta}, \alpha, 1-\alpha)} + \frac{1-t}{1-clip(\hat{\eta}, \alpha, 1-\alpha)}$$

where the clipping parameter α is typically set to $.01$ or $.05$. Despite its bias, this form has a reduced variance in extreme (close to 0 or 1) estimated propensity regions.

A balancing score $b(X)$ is a function of the covariate vector X that makes the covariates independent from the treatment assignment:

$$X \perp\!\!\!\perp T|b(X)$$

[10] sometimes misleadingly referred to as inverse propensity weight
[11] also referred to as augmented inverse propensity weight

After Rosenbaum and Rubin (1983), the propensity score η is a balancing score. It is notably the coarser one, in the sense that b is a balancing score if and only if there exists some function f such that $\eta = f \circ b$. In particular, the identity function over $\mathcal{X}$ is a balancing score. Note that strong ignorability is preserved by conditioning upon a balancing score:

$$\text{If } \begin{cases} (Y^0, Y^1) \perp\!\!\!\perp T \mid X \\ \forall\, \Omega \; s.t. \mathbb{P}(X \in \Omega) > 0, \mathbb{P}(T = 0|X), \mathbb{P}(T = 1|X) > 0 \end{cases}$$

$$\text{then } \begin{cases} (Y^0, Y^1) \perp\!\!\!\perp T| \, b(X) \\ \forall\, \Omega \; s.t. \mathbb{P}(b(X) \in \Omega) > 0, \mathbb{P}(T = 0|b(X) \in \Omega), \mathbb{P}(T = 1|b(X) \in \Omega) > 0 \end{cases}$$

A prognostic score after Hansen (2008) is a random variable $\phi(X)$ such that

$$Y^0 \perp\!\!\!\perp X \mid \phi(X)$$

i.e. $\phi(X)$ encapsulates the sufficient statistics to predict Y^0 (in short, is sufficient for Y^0). In the case where the prognostic score is not sufficient for Y^1, the random variable $m(X)$ is called an effect modifier if $\{\phi(X), m(X)\}$ is sufficient for Y^1.

Prognostic scores preserve conditional exchangeability: assuming that conditional exchangeability holds w.r.t X, it comes $Y^0 \perp\!\!\!\perp Z \mid \phi(X)$.[12] Notably, if ϕ is injective, $\phi(X)$ is sufficient for Y^0.

[12]The reader is referred to Hansen (2006) for more detail; see proposition 3 and its proof.

III - State of the art

This chapter introduces three goals pertaining to causal learning, distinguishing causal discovery and causal inference. Causal discovery (Section III.1) is concerned with determining the causal relations among a set of variables from observational data $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i \in [\![1,n]\!]}$. More related to the presented work are methods aimed to measure the impact of a binary treatment over a whole population, that is, the Average Treatment Effect (Section III.2), contrasting the treated and control populations (respectively formed of the samples with $t_i = 1$ and with $t_i = 0$). The chapter last details the approaches concerned with heterogeneous estimates of the treatment effect, where the estimated effect depends on the covariate, that is, the Conditional Average Treatment Effect (Section III.3).

III.1 . Causal discovery

Causal discovery aims to recover the causal graph among a set of variables X given observations (Spirtes et al., 1993; Peters et al., 2017). We first introduce the main framework used in the causal discovery literature (Section III.1.1), i.e., the model space. How to identify the sought model and recover the causal graph is described in Section III.1.2.

III.1.1 . Functional causal model framework

Let us present the Functional Causal Model (*FCM*) framework [1]. Let $V^X = (v_1^X, \dots, v_D^X)$ and $V^N = (v_1^N, \dots, v_D^N)$ be two sets of D vertices and $\mathcal{G}$ be a Directed Acyclic Graph over vertices $V^X \cup V^N$. For simplicity and by abuse of notations, the vertex is also noted as the variable it refers to, with v_i^X corresponding to the endogenous, observed variable X_i, and v_j^N corresponding to an exogenous, noise variable N_j. Graph $\mathcal{G}$ satisfies the following two conditions:

1. each endogenous vertex v_d^X admits exactly one vertex (v_d^X) as an exogenous parent (and an arbitrary number of endogenous vertices as parents too): $\forall d \in [\![1, D]\!], Pa(v_d^X) \cap V^N = \{v_d^N\}$
2. no each exogenous vertex admits a parent: $\forall d \in [\![1, D]\!], Pa(v_d^N) = \emptyset$

Let $\mathcal{F} = (f_1, \dots, f_D)$ be a set of D causal mechanisms ($f_d : \mathbb{R}^{|Pa(v_d^X)|} \to \mathbb{R}$). The couple $(\mathcal{G}, \mathcal{F})$ then defines a **Functional Causal Model**: Each of the D en-

[1] functional causal models suit a presentation of the scope of causal discovery, but other frameworks exist. For a presentation of Structural Equation Models or Causal Bayesian Networks, we refer the reader to Pearl (2009)

$$X^1 = f_1(N^1)$$
$$X^2 = f_2(X^1, N^2)$$
$$X^3 = f_3(X^1, N^3)$$
$$X^4 = f_4(X^3, N^4)$$
$$X^5 = f_5(X^3, X^4, N^5)$$

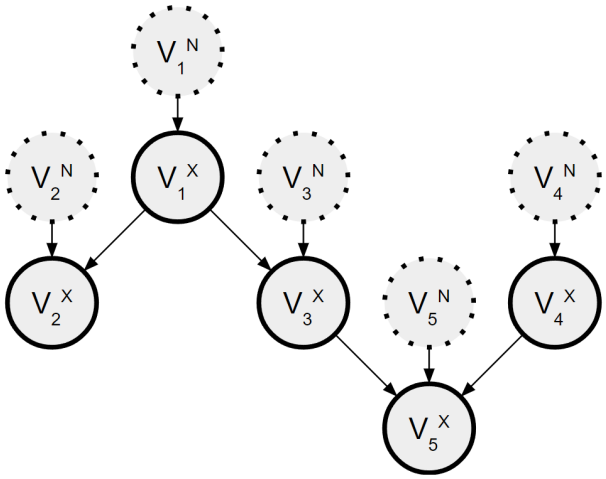


Figure III.1: A Functional Causal Model (taken from Kalainathan et al. (2022)). The graph is directed and acyclic. All endogenous vertices admit exactly one vertex as an exogenous parent; no exogenous vertex admits a parent.

dogenous variables X^d is described as a function (f_d, the causal mechanism) of their endogenous parents (their causes) and the (exogenous) noise N^d:

$$X^d = f_d(X^{(Pa(v_d^X) \cap V^X)}, N^d),\ d \in [\![1, D]\!]$$

By construction, the joint distribution over the endogenous variables $X^1, \dots X^D$ follows from the product distribution of the independent D noise variables

$$\mathbb{P}_{N^1,\dots,N^D} = \Pi_{d=1}^{D} \mathbb{P}_{N^d}$$

through Section III.1.1, as depicted on Fig. III.1.

The direct problem consists of sampling the distribution $\mathbb{P}_{X^1,\dots,X^D}$ based on the joint distribution $\mathbb{P}_{N^1,\dots,N^D}$ and knowing the *FCM* structure. Causal discovery tackles the (much more difficult) inverse problem: inferring the functional causal model based on observations sampled after $\mathbb{P}_{X^1,\dots,X^D}$.

Many *FCM* settings are defined in the literature, in order to match real-life data generation processes and/or to facilitate causal discovery. A most common *FCM* setting, referred to as Additive Noise Model (Hoyer et al., 2008; Peters et al., 2014; Chicharro et al., 2019; Montagna et al., 2023), assumes that the functional mechanisms f_d are of the form:

$$f_d(X^{(Pa(v_d^X) \cap X)}, N^d) = g_d(X^{(Pa(v_d^X) \cap X)}) + N^d,\ d \in [\![1, D]\!]$$

III.1.2 . Main strategies

The many strategies designed to identify the underlying causal graph can be structured along four categories; the reader is referred to Zanga and Stella (2022) for a comprehensive survey. These categories include: i) score-based

methods; ii) constraints-based methods; iii) hybrid methods; and iv) methods relying on asymmetries or causal footprints.

III.1.2.1 . Score-based methods

Score-based methods use a **global score** to assess the quality of any given graph w.r.t. the observational data, and they tackle the combinatorial optimization problem of finding the candidate graph maximizing the score. The celebrated Greedy Equivalent Search algorithm starts from an empty graph (Chickering, 2002). During the forward search, Greedy Equivalent Search successively adds the edge that most increases the score until no increase is possible. In the backward search, it inversely removes edges that minimally decrease the score. The efficiency and scalability of score-based methods mainly depend on the score and its computational complexity. Adequate caching of intermediate results, avoiding to re-evaluate partial solutions, makes it possible to accelerate Greedy Equivalent Search into Fast Greedy Equivalent Search (Ramsey et al., 2017).

III.1.2.2 . Constraint-based methods

Constraint-based methods use dependence and conditional independence statements, based on **statistical tests**, to identify the edges and the so-called V structures in the graph. Dependences and conditional independences are exploited to identify the skeleton of the sought causal graph, referred to as partially completed directed acyclic graph (*CPDAG*). The Peter-Clark algorithm (Spirtes and Glymour, 1991; Spirtes et al., 1993) is among the earliest and best-known constraint-based methods. The Fast Causal Inference (Cooper and Glymour, 1999) algorithm extends Peter-Clark and relaxes its assumptions, notably concerning the absence of latent confounders and selection bias. The main weakness of constraint-based methods is the number of conditional independence tests required to find a *CPDAG*, because of the computational complexity on the one hand, and because of the multiple hypothesis testing issue on the other hand. It is desirable to use sophisticated statistical tests (as opposed to a mere correlation among variables), such as the Hilbert-Schmidt Independence Criterion (Gretton et al., 2007) or the Kernel Conditional Independence test (Zhang et al., 2011).

While the consistency in the large sample limit of score- and constraint-based methods is thoroughly studied, they suffer from a lack of guarantees in real-life situations, especially when the total number of features increases.

III.1.2.3 . Hybrid methods

Hybrid methods leverage constraints to obtain a reduced search space, enabling the running of score-based methods with **reduced computational cost**. The Max-Min Hill Climbing algorithm (Tsamardinos et al., 2006) first builds a constraint-based graph skeleton and thereafter orients its edges us-

ing a score-based method. Adaptively Restricted Greedy Equivalent Search (Nandy et al., 2018) handles causal discovery of up to thousands of features under sparsity assumptions. Ogarrio et al. (2016) proposes the best of both worlds, addressing the limitations of score-based methods w.r.t. unobserved confounders and that of constraint-based methods in small samples regimes, through hybridizing Fast Greedy Equivalent Search (that provides an initial skeleton) and Fast Causal Inference (to orient the skeleton edges efficiently).

III.1.2.4 . Methods relying on asymmetries or causal footprints

These methods improve on the abovementioned ones by making strong – though unlikely to hold in practice – assumptions to be able to orient all their edges. The Linear Non-Gaussian Acyclic Model (LinGAM) handles continuously-valued data using independent component analysis, under the assumption that noise variables are non-Gaussian (Shimizu et al., 2006). Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning (NoTEARS) exploits a continuous characterization of *DAG* matrices (Zheng et al., 2018); formally, matrix B is a *DAG* if and only if the trace of its exponential matrix e^B is the number d of variables. This shift from combinatorial to continuous optimization paves the way for neural networks-based approaches. Notably, Gradient-based Neural *DAG* (Lachapelle et al., 2019) uses Auto-Encoders to model dependencies among variables while maintaining the acyclicity of the underlying graph. Structural Agnostic Modelling (Kalainathan et al., 2022) uses Generative Adversarial Networks to model each variable from all other variables plus a noise variable, akin to a Markov kernel, and uses the dependency relations with *DAG*-enforcing constraints to infer the graph structure. Reinforcement Learning has also been leveraged by Zhu et al. (2020) to achieve causal discovery, where the agent actions consist of adding, removing, and reversing an edge. The state space is formed of all graphs involving the considered observed variables.

III.2 . Average Treatment Effect estimation

ATE models aim to answer questions at the population level: how much would a given policy improve school enrolment of teenage girls (Beaman et al., 2012)? how much do deworming policies improve general health (Miguel and Kremer, 2004)?

ATE methods are heavily investigated and used in econometrics. Indeed, answering questions at the population level avoids the pitfalls of finer-grained analysis, thanks to a higher statistical power, and a reduced uncertainty; such answers are more clear and convincing for e.g. policymakers. Most gener-

ally, *ATE* is appropriate when dealing with scarce observational data, due to experiment cost or practical feasibility.

Note that *CATE* estimators entail *ATE* estimators. From the equality $ATE = \mathbb{E}[\tau(X)]$ it follows that a *CATE* estimate $\hat{\tau}$ induces estimate $\mathbb{E}[\hat{\tau}(X)]$ (we come back to the *CATE* estimation task in Section III.3).

We first examine methods where external variables (instrumental variables) are leveraged to avoid confounding bias (Section III.2.1). Other methods aim at bringing the treated and the control population closer to each other in order to shed light on the actual impact of the treatment, the two distributions remaining the same (Section III.2.2.1) or using a change of representation (Section III.2.2.2). Statistical properties of such methods have been investigated to assess their asymptotic behavior and guarantees (Section III.2.3).

III.2.1 . Instrumental variables

Instrumental Variable-based (IV) estimation aims at measuring the impact of covariates X on a quantity of interest, a.k.a. outcome Y in settings where a direct regression (estimating Y as a function of X) would lead to biased estimates. Such biases might be due to confounders (a third hidden variable causes both X and Y), reverse causation (Y is a cause of X), or correlated measurement errors (acting as a confounder on the measured values of X and Y).

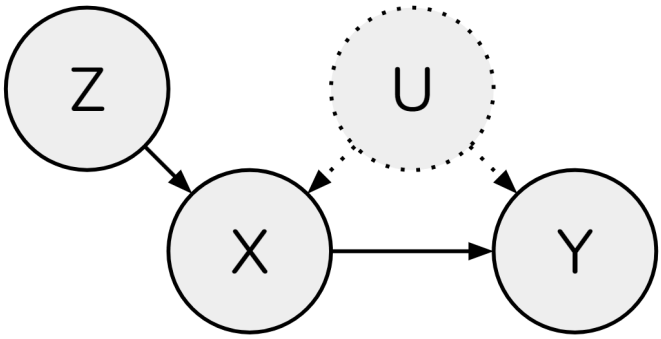


Figure III.2: Causal graph with instrumental variables Z.

In such cases, a third variable Z influencing Y through X only (Fig. III.2), referred to **instrumental variable**, can be leveraged to estimate *ATE*. For instance, the price of tobacco affects the population of smokers and is likely to decrease the consumption of tobacco, but it has no direct effect on tobacco-related diseases, *ceteris paribus*. Therefore, if a rise in tobacco prices is associated with some general health improvement, a negative causal effect of smoking on individual health may be inferred.

The identification of a suitable instrumental variable Z relies on domain knowledge. Following Lousdal (2018), instrumental variables Z must be:

1. *relevant*, i.e. with causal influence on X;
2. *exclusive*, i.e. such that their only influence on Y is through their influence on X ($Y \perp\!\!\!\perp Z|X$);
3. *exchangeable*[2], i.e., such that there are no common causes for Z and Y.

The instrumental variable method, long used by econometrists, is consistent with the Neyman-Rubin potential outcome framework (Angrist et al., 1996); note that it also applies in more general settings, e.g., involving continuously valued treatments.

Suppose that the true data generation process is of the form $Y = X\beta + U$, where U is a residual term that stands for all influences other than X. Let us further assume that there exists a Z variable, linearly related to X:

$$X = Z\gamma + U$$

with $\dim(Z) > \dim(X)$. The IV method proceeds by:

1. Regressing X on Z, leading to an estimate $\hat{X}$ independent from U:
   Let $\hat{X} = P_Z X$ with $P_Z = Z(Z^T Z)^{-1} Z^T$ the projection matrix on Z. By construction, $\hat{X}$ is independent of U.
2. Regressing Y on $\hat{X}$, with coefficient vector

$$\begin{aligned}\hat{\beta}_{\text{IV}} &= (\hat{X}^T \hat{X})^{-1} \hat{X}^T Y \\ &= (X^T P_Z X)^{-1} X^T P_Z Y\end{aligned}$$

3. Assuming that $\frac{1}{n} Z^T U$ converges to o in probability, $\hat{\beta}_{\text{IV}}$ is provably a consistent estimate for β.

Note that the above procedure is a particular, linear, case of the Generalized Method of Moments (see below).

On the one hand, the IV method is efficient and theoretically grounded. On the other hand, the choice of instrumental variables requires strong expertise in the application domain, and it governs the overall performance of the approach. Bound et al. (1995) show that weak instrumental variables (i.e., poorly correlated with the outcome) may lead to biased or inconsistent estimates. Their selection can be achieved using specific statistical tests (Stock et al., 2002).

Various approaches have been considered in the linear setting, stressing the model space expressivity. The Generalized Method of Moments estimator

[2] "exchangeability" is also referred to as "independence", "ignorable treatment assignment," or "no confounding".

(Hansen, 1982; Hansen et al., 1996) relaxes the assumptions on the residual term U and proposes a criterion defining a single optimum, enjoying strong consistency, asymptotic normality and efficiency.

In the nonlinear setting, the model reads

$$Y = f(Z, \beta) + U$$

where Z and U are independent, and non-linearities occur in the variables (Kelejian, 1971) and/or in the model parameters (Zellner et al., 1965). Amemiya (1974) proposes a nonlinear two-stage least-squares estimator of f and β. Newey and Powell (2003); Ai and Chen (2003) extend this work to the nonparametric setting. Hartford et al. (2017) has investigated the use of IV methods in larger-dimensional settings, modeling the relations between X, Y, Z through neural networks, while Singh et al. (2019) use reproducing kernel Hilbert spaces (Support Vector Machines).

III.2.2 . Balancing methods

As methods that enforce balance in a latent space[3] typically aim at estimating the *CATE* rather than the *ATE*, they will be detailed in Section III.3. In the following are considered methods that associate subsets of the control and treated distributions (Section III.2.2.1), and methods that affect weights to the samples (Section III.2.2.2).

III.2.2.1 . Matching methods

As shown in the Simpson Paradox example (Section I.2), comparing the raw average success rates of percutaneous nephrolithotomy and open surgery leads to a biased conclusion since exchangeability does not hold. However, comparing two patients with similar kidney stone sizes but different treatment assignments still makes sense: informally, this is the intuition behind the matching methods.

More formally, matching methods constitute a **heterogeneous ensemble of pre-processing algorithms** that aim at bringing the treated and control distributions close to one another (Cochran, 1953; Billewicz, 1965). The core justification of such approaches is that they weaken the statistical dependence between covariates X and treatment T (Ho et al., 2007). Multiple characterizations of balance exist in the literature. Fine balance (Rosenbaum et al., 2007) is when the marginals of the treated and control distributions are identical. This property is however neither necessary nor sufficient to support causal inference (see, e.g., Yang et al. (2012) for a discussion about the trade-off between fine matching and minimizing the distance between treated and control samples). Diverse methods consider

[3] $X \perp\!\!\!\perp T | \phi(X)$, *with* $\phi : \mathcal{X} \to \mathbb{R}^d$, $d \geq 2$

different balance criteria (measuring the similarity between the control and the treated distributions through e.g. first-order or higher-order moments, high-order statistical dependencies, and functions of the covariates). In practice, matching methods often involve multiple hyper-parameters and can be iterated until reaching a satisfactory level of balance (Austin, 2008; Stuart, 2010; Li and Greene, 2013).

Following Stuart (2010) we distinguish two main classes of matching approaches:

1. **Nearest neighbor matchings** typically tackle *ATT* or *ATC* estimation problems, for they are asymmetrical in their approach. In the following and without loss of generality, the case of *ATT* estimation will be considered (the case of *ATC* follows by symmetry).
   Nearest neighbor matchings rely on a (pseudo-)distance over the covariate space. In **Exact Matching**, $d(x_i, x_j) = 0$ if $x_i = x_j$, $+\infty$ otherwise. Although this may be relevant in settings where the features are categorical and in a low total number, in high dimensional settings exact matching is prone to failure for all samples are likely infinitely distant from each other. **Coarsened Exact Matching** (Iacus et al., 2012) turns each feature of X first into coarse categorical data $c(X)$ (e.g. by binning continuous variables, or grouping categorical variables), forming *strata* defined as unique combinations of the coarsened features (i.e., voxels in the coarsened space). **Mahalanobis Distance Matching** relies on the Mahalanobis distance: $d(x_i, x_j) = (x_i - x_j)^\dagger \Sigma^{-1} (x_i - x_j)$ (where Σ is the covariance matrix for *ATE* estimation, or that of the control population when estimating *ATT*). **Propensity Score Matching** builds on a pseudo-distance based on conditional probability of treatment assignment: defining by $\hat{\eta}$ a propensity estimate, $d(x_i, x_j) = |\hat{\eta}(x_i) - \hat{\eta}(x_j)|$ or $d(x_i, x_j) = |logit(\hat{\eta}(x_i)) - logit(\hat{\eta}(x_j))|$ (Rosenbaum and Rubin, 1985).

   Once a distance is chosen, all treated samples are iteratively matched with their closest control neighbor ("1:1" matching), or to their k nearest control neighbors ("k:1" matching) (Rubin, 1973). Depending on the considered method, multiple treated samples may be matched to the same control one ("without replacement") or not ("with replacement").

   This protocol is however naive, and several issues may arise. First, when matching without replacement, the order in which the treated samples are considered affects the procedure output. **Optimal matching** (Rosenbaum, 2002) procedures aim at minimizing the total distance between treated samples and their matched counterparts, optimization being carried through network flow approaches (Rosenbaum,

1989) or mixed integer programming (Zubizarreta, 2012). However, when a treated sample is far away from all remaining unmatched control ones, comparing it with its closest neighbor makes little sense. **Caliper** approaches (Cochran and Rubin, 1973) discard treated samples if no unmatched control sample remains within a fixed distance. Note how delicate is the choice of the hyper-parameter k in "k:1" matching: larger values may reduce estimation variance at the price of an increased bias (samples further away are of lower relevance). Finally, matching on a one-dimensional score (e.g., projection over a feature axis or propensity score) may match highly unrelated samples in large-dimensional settings. Neural Score Matching (Clivio et al., 2022) leverages neural networks to learn multi-dimensional balancing scores that avoid this drawback.

Once all treated samples are matched, standard treatment effect evaluation methods may be used. For instance, denoting by M_T the subset of matched samples and M_C the subset of control ones, *ATT* might be estimated as $\frac{1}{|M_T|}\sum_{(x,t,y)\in M_T} y - \frac{1}{|M_C|}\sum_{(x,t,y)\in M_C} y$.

2. **Subclassification machings** are more versatile and may tackle *ATE* as well as *ATT* and *ATC* estimation problems. They consist in building groups of samples that are **homogeneous** with respect to a given criterion: the samples in each group are thus matched. The entailed partition may typically be based on the quantiles of a feature axis (Cochran, 1968), or the estimated propensities (Rosenbaum and Rubin, 1985).

   If subclassification is based on propensity score quantiles, all samples within the same group have close probabilities of treatment assignment. Such a property is especially relevant since the propensity score is known to be a balancing score (Section III.2.2.2): $X \perp\!\!\!\perp T|\eta(X)$. In other words, propensity matching approximately enforces balance within each group.

   Let $G = (g_1, \dots, g_{|G|})$ denote the resulting groups, and $\widehat{ATE}(g)$ be an estimate of the Average Treatment Effect based on samples in g[4]. The *ATE* of the global population may be estimated as the average of all group estimates: $\widehat{ATE} = \frac{1}{|G|}\sum_{g\in G}\widehat{ATE}(g)$. The importance of the groups may also be adapted to account for different sizes ($\widehat{ATE} = \frac{1}{|G|}\sum_{g\in G}\frac{|g|}{|G|}\widehat{ATE}(g)$) or to compute other metrics ($\widehat{ATT} = \frac{1}{|G|}\sum_{g\in G}\frac{|\{(x,t=1,y)\in g\}|}{|\{(x,t=1,y)\in D\}|}\widehat{ATE}(g)$).

[4] as a rough approximation, one may estimate it as

$$\widehat{ATE}(g) = \frac{1}{|\{(x,t=1,y)\in g\}|}\sum_{(x,t=1,y)\in g} y - \frac{1}{|\{(x,t=1,y)\in g\}|}\sum_{(x,t=0,y)\in g} y$$

Multiple desirable matching properties have been identified. Rubin (1976) pinpoints methods that are **Equal Percent Bias Reducing**; they focus on the first-order moments of the distributions, and provably reduce the discrepancy between the mean of the treated and matched control population along each feature axis. Notably, Mahalanobis matching (based on the Mahalanobis distance) and propensity score matching (based on the absolute difference of estimated propensities) are equal percent bias reducing. Iacus et al. (2011) present **Multivariate Imbalance Bounding** methods, that involve higher-order statistics. Coarsened Exact Matching is for instance multivariate imbalance bounding.

Matching methods face multiple pitfalls. It is emphasized that matching criteria should consider neither the outcome variable Y nor any consequence of Y. Moreover, most methods require a minimum overlap of the treated and control distribution. Identifying the regions that should be pruned raises critical difficulties in high dimensional covariate spaces; for instance King and Zeng (2007) recommend discarding control samples outside of the convex hull of the treated samples.

Following LaLonde (1986)'s warning, the discussion between Dehejia and Wahba (1999, 2002); Dehejia (2005) and Smith and Todd (2001, 2005a,b) sheds light on the pros and cons of matching methods. The main critiques against propensity score matching after King and Nielsen (2019) are related to "increasing imbalance, inefficiency, model dependence, research discretion and bias". Specifically, propensity score matching is sensitive w.r.t. the choice of the model; causal estimates based on different models that fit the data approximately equally well, may be different.

III.2.2.2 . Reweighting methods

Numerous methods exist in the literature to estimate population-level effects, without resorting to matching. Reweighting methods do not discard data or create subgroups; each sample is simply associated with a weight. Again, this change of representation is often viewed as a preprocessing step.

Inverse Probability of Treatment Weighting (*IPTW*), as many other reweighting methods, is viewed as a Balance Method. In the Neyman-Rubin framework, given a propensity estimate $\hat{\eta}$ (entailing estimate of the inverse probability of treatment weights $\hat{\rho}^{\cdot}$, Eq. II.8), the Average Treatment Effect over an observational dataset $\mathcal{D}$ is estimated by one of the following formulae (Lunceford and Davidian, 2004):

$$\widehat{ATE}_1 = \frac{1}{|\mathcal{D}|} \sum_{(x,t,y)\in\mathcal{D}} \frac{t}{\hat{\eta}(x)} y - \frac{1}{|\mathcal{D}|} \sum_{(x,t,y)\in\mathcal{D}} \frac{1-t}{1-\hat{\eta}(x)} y = \frac{1}{|\mathcal{D}|} \sum_{(x,t,y)\in\mathcal{D}} (2t-1)\hat{\rho}^t(x) y$$

$$\widehat{ATE}_2 = \Big(\sum_{(x,t=1,y)\in\mathcal{D}} \hat{\rho}^1(x) \Big)^{-1} \sum_{(x,t=1,y)\in\mathcal{D}} \hat{\rho}^1(x) y - \Big(\sum_{(x,t=0,y)} \hat{\rho}^0(x) \Big)^{-1} \sum_{(x,t=0,y)} \hat{\rho}^0(x) y$$

$$\widehat{ATE}_3 = \Big(\sum_{(x,t=1,y)\in D} \hat{\rho}^1(x)\big(1-\hat{\rho}^1(x)C_1\big)\Big)^{-1} \sum_{(x,t=1,y)\in D} \hat{\rho}^1(x)\big(1-\hat{\rho}^1(x)C_1\big)y$$
$$- \Big(\sum_{(x,t=0,y)} \hat{\rho}^0(x)\big(1-\hat{\rho}^0(x)C_0\big)\Big)^{-1} \sum_{(x,t=0,y)} \hat{\rho}^0(x)\big(1-\hat{\rho}^0(x)C_0\big)y$$

$$\text{where } C_1 = \frac{\sum_{(x,t,y)\in D} t\hat{\rho}^1(x) - 1}{\sum_{(x,t,y)\in D}\big(t\hat{\rho}^1(x)-1\big)^2},\ C_0 = \frac{\sum_{(x,t,y)\in D}(1-t)\hat{\rho}^0(x) - 1}{\sum_{(x,t,y)\in D}\big((1-t)\hat{\rho}^0(x)-1\big)^2}$$

and assuming potential outcome estimates $\hat{\mu}^0, \hat{\mu}^1$ are available,

$$\widehat{ATE}_{DR} = \frac{1}{|D|}\sum_{(x,t,y)\in D} (\hat{\mu}^1 - \hat{\mu}^0)(x) + (2t-1)\hat{\rho}^t(x)\big(y - \hat{\mu}^t(x)\big)$$

The merits of $\widehat{ATE}_2$ and $\widehat{ATE}_3$ are to extend $\widehat{ATE}_1$ with smaller variance (Liao and Rohde, 2022). $\widehat{ATE}_{DR}$, also called Augmented Inverse Propensity Weighting (*AIPW*) *ATE* estimate, is doubly robust in the sense that it is provably consistent even in the case where either the propensity or the outcome model (but not both) is misspecified.

$\widehat{ATE}_{DR}$ however might suffer from an unbounded variance in cases (frequent in practice) where the propensity estimates take values close to 0 or 1. Several methods have been proposed to address this drawback. Crump et al. (2009) suggest excluding samples with estimated propensity outside of $[\alpha, 1-\alpha]$, with α set to 0.1 as a rule of thumb, while Stürmer et al. (2010) prefer lower and upper bounds defined by $q-$quantiles: $[quantile_q(\hat{\eta}_i \mid t_i = 1), quantile_{1-q}(\hat{\eta}_i \mid t_i = 0)]$. Trimming weights consists in capping the propensity values with α for $\eta < \alpha$ and symmetrically with $1-\alpha$ for $\eta > 1-\alpha$ (Potter, 1990).

Generalization Li et al. (2018) notices that $ATE = \mathbb{E}[\tau(X)\times 1]$, $ATT = \mathbb{E}[\tau(X)\times\eta(X)]$, $ATC = \mathbb{E}[\tau(X)\times(1-\eta)(X)]$ and by generalization defines the Weighted Average Treatment Effect as:

$$\begin{aligned}\vartheta :\ \{\mathcal{X}\to\mathbb{R}_+^*\} &\longrightarrow \mathbb{R}\\ h &\longmapsto \mathbb{E}[\tau(X)\times h(X)]\end{aligned}$$

As an immediate consequence, $ATE = \vartheta(1)$, $ATT = \vartheta(\eta)$, $ATC = \vartheta(1-\eta)$.

Li et al. (2018) also shows that, (w_h^0, w_h^1) and $\hat{\tau}_h$ being defined in accordance with Table III.1, the latter is a consistent estimate of $\vartheta(h)$:

$$\hat{\tau}_h = \frac{\sum_i w_h^1(x_i)t_i y_i}{\sum_i w_h^1(x_i)t_i} - \frac{\sum_i w_h^0(x_i)(1-t_i)y_i}{\sum_i w_h^0(x_i)(1-t_i)} \tag{III.1}$$

Such a versatile notation makes it possible to define easily other estimands, such as the Average Treatment effect on the Overlap region $ATO = \mathbb{E}[\tau(X)\eta(X)(1-\eta)(X)]$.

Estimand	h	(w_h^0, w_h^1)
ATE	1	$(\frac{1}{\eta}, \frac{1}{1-\eta})$
ATT	η	$(1, \frac{\eta}{1-\eta})$
ATC	$1-\eta$	$(\frac{1-\eta}{\eta}, 1)$
ATO	$\eta(1-\eta)$	$(1-\eta, \eta)$
	$\mathbb{1}[\alpha<\eta<1-\alpha], \alpha \in \mathbb{R}$	$(\frac{\mathbb{1}[\alpha<\eta<1-\alpha]}{\eta}, \frac{\mathbb{1}[\alpha<\eta<1-\alpha]}{1-\eta})$
	$min(\eta, 1-\eta)$	$(\frac{min(\eta,1-\eta)}{\eta}, \frac{min(\eta,1-\eta)}{1-\eta})$

Table III.1: Estimand-weights correspondence.

III.2.3 . Doubly Robust Machine Learning

Chernozhukov et al. (2018) observes that ML-based approaches used in causal inference are subject to large biases. Overfitting the n samples in the observational data, the increasing use of regularization to compensate for a large parameter space, correlated errors in two-stages protocols (Section III.3.1.2): all these factors prevent the learned models from achieving $\sqrt{n}$-consistency.

Relying on Neyman-orthogonal (Neyman, 1979) scores and cross-fitting, such convergence rate can be provably achieved using the famed **Doubly Robust** (DR) Machine Learning. *DR* machine learning is robust to the misspecification of one of its two core components, either the estimated propensity score or outcome function. With an appropriate training protocol relying on cross-training, Chernozhukov et al. (2018) extends this framework into *Double Machine Learning*, and provably obtains $\sqrt{N}$-consistent estimates.

Similarly, Shi et al. (2019) presents the *Dragonnet* neural architecture, where double robustness is enforced using a feed-forward neural network trained with a simple protocol, with an adequate regularization term inspired by target maximum likelihood estimation (Laan and Rose, 2011).

Note that doubly robust Machine Learning does not restrict to population-level estimation. See Section III.3.1.7 for further development regarding *CATE* estimation.

III.3 . Conditional Average Treatment Effect estimation

As said, measuring the impact of a given treatment at the whole population scale offers clear and comparatively less uncertain answers. Still, in application domains (such as health and advertising to name a few) grows the need

for heterogeneous estimates, approximating the causal effects conditionally to the covariates and thus providing sample-dependent information.

As large datasets become increasingly available, they support the estimation of causal effects at group levels (instead of, on the whole population). Machine Learning is thus leveraged to predict the potential outcomes as functions of the covariates X, while counter-factuals remain unavailable by construction.

The Conditional Average Treatment Effect (*CATE*) is meant to estimate the expected benefit of the treatment at the subpopulation/individual level[5], formally defined (Eq. IV.2) as

$$\tau(x) = E[Y^1 - Y^0 | X = x]$$

Let us first examine the different types of *CATE* approaches, referred to as "meta-learners" (Section III.3.1). Section III.3.2 is devoted to the case where conditional exchangeability does not hold, hindering the identifiability of distributional approaches. The chapter last discusses the main challenges of *CATE* estimation (Section III.3.3).

III.3.1 . Meta-learners

After Künzel et al. (2019); Nie and Wager (2021); Kennedy (2023), *CATE* algorithms can be divided into seven categories, or meta-learners: the most common two are *S-learners* (Section III.3.1.1) and *T-learners* (Section III.3.1.2), while the remaining *X-,R-,F-,U,DR-learners* (Sections III.3.1.3 to III.3.1.7) are mainly considered for their theoretical properties.

III.3.1.1 . *S-learners*

S(ingle)-learners **handle the treatment assignment as any other covariate** of the sample. The problem of estimating both outcome functions from the covariate and assignment variables is tackled as a mainstream supervised learning problem, using ordinary machine learning approaches to estimate $\mu^t(x)$:

$$\hat{\mu}(x,t) \approx \mu^t(x) = \mathbb{E}[Y | X = x, T = t]$$

The treatment effect is thereafter estimated as:

$$\hat{\tau}_S(x) = \hat{\mu}(x,1) - \hat{\mu}(x,0)$$

Building upon domain adaptation principles (Ben-David et al., 2006), Johansson et al. (2016) introduce the *Balancing Neural Network* (**BNN**) approach. A first neural network[6] ϕ is used to map both treated and control

[5] Indeed, *ATE* can be estimated by averaging *CATE*: $ATE = \mathbb{E}[\tau(X)]$ and *ATE* estimation may thus be seen as a secondary objective of *CATE* estimators.

[6] Mapping ϕ is not noted with a circumflex diacritic for it is no estimate of a ground truth statistical quantity. The same remark holds for many objects to be introduced later on.

samples onto a single latent representation space. After concatenation of the treatment assignment, the enriched representation $(\phi(x), t)$ is processed through a second neural network ψ, yielding $\hat{\mu}(x,t) = \psi(\phi(x), t)$. The balance between the treated and control populations in the latent space is ensured by a sparse reweighting of the feature axis. Since the treatment assignment variable is given no particular role, its importance tends to be underestimated by *S-learners*, possibly biasing the *CATE* toward 0 after Künzel et al. (2019).

Hill (2011) considers Bayesian Additive Regression Trees (**BART**), and Athey and Imbens (2016) proposes an extension thereof using causal trees to build confidence intervals. Another related approach is introduced by Wager and Athey (2018), using Causal Forests (**CF**) where the last split of the forest trees corresponds to the treatment assignment.

III.3.1.2 . *T-learners*

T(wo)-learners rely on **two separate estimates** $\hat{\mu}^0$ and $\hat{\mu}^1$ to model each outcome function (also referred to as surface of contact), with:

$$\hat{\mu}^0(x) \approx \mu^0(x) = \mathbb{E}[Y|X = x, T = 0]$$
$$\hat{\mu}^1(x) \approx \mu^1(x) = \mathbb{E}[Y|X = x, T = 1]$$

and the individual treatment effect is estimated as

$$\hat{\tau}_T(x) = \hat{\mu}^1(x) - \hat{\mu}^0(x)$$

Let us present a few related approaches, referring the reader to Caron et al. (2022) for a more comprehensive presentation.

Shalit et al. (2017) extend *BNN*, the *S-learner* approach introduced by Johansson et al. (2016). A representation network ϕ maps both control and treated samples on the same latent space, which supports two heads h^0 and h^1, trained such that $\hat{\mu}^t = h^t \circ \phi$. Representations in the latent space may be balanced using statistical distances such as the Wasserstein distance (Cuturi, 2013) or the Maximum Mean Discrepancy (Gretton et al., 2012) for **CFR**, or left unconstrained for **TARNet**. In Yao et al. (2018), **SITE** is based on the observation that useful information can be lost through balancing the control and treated distributions in the latent space. **SITE** accordingly aims at preserving local similarity information while balancing the representations. It is completed by Yao et al. (2019)'s **ACE** which preserves finer-grained information. Several approaches resort to reweighting the samples in the latent space, providing more flexibility to population balancing. Hassanpour and Greiner (2019a)'s **CFR-ISW** involves *context-aware* weights, built using a latent space analog of propensity scores. Assaad et al. (2021) explores the generalization of weighting schemes outside the scope of inverse probability of treatment weights through the **BWCFR** approach. **MitNet** (Guo et al., 2023) resorts to the same architecture as **CFR**, but controls the discrepancy in the

latent space using mutual information. Contrarily to other approaches, such a design generalizes well to non-binary treatment settings. Wu et al. (2023a) bridges the gap between *T-learners* and matching methods, leveraging nearest-neighbors approaches in the latent space so as to upsample underrepresented subpopulations and refine the modeling of their causal effect.

Balance in the latent space may also be achieved through adversarial learning, taking inspiration from domain adaptation (Ganin et al., 2016). Informally, training an adversarial model to distinguish between control and treated latent samples limits the discrepancy between the two spaces.The relationship between domain adaptation and causal inference is explored by Johansson et al. (2022). Du et al. (2021)'s **ABCEI** leverages the mutual information between the observed and the latent representation to limit the loss of information. Similarly, Zhou et al. (2021)'s **CBRE** considers an auto-encoder architecture with a specific cycle structure: the loss of information from the observed to the latent representation is prevented by enforcing the reconstruction of the samples from their latent representation.

Generative models have also been considered in the hope that their distributional nature may better capture the uncertainty in the counter-factual distribution in a slightly different setting (we shall return to generative models for causal inference in Section III.3.2). Louizos et al. (2017)'s **CEVAE** combines the approach in Shalit et al. (2017) with a Variational Auto-Encoder (*VAE*) (Kingma and Welling, 2014; Rezende et al., 2014) architecture, while Yoon et al. (2018)'s **GANITE** is based on Generative Adversarial Networks (*GAN*) (Goodfellow et al., 2014). Alaa and Schaar (2018)'s **NSGP** bases its approach on Gaussian processes, an idea that Zhang et al. (2020) extends with **DKLITE**. With a method based on a deep kernel regression algorithm, it tackles the key issue of counter-factual variance minimization and provides uncertainty intervals.

Kuang et al. (2017)'s **D^{2}VD** initiates a specific stream of work, splitting with the covariates into confounding features (causes of both Y and T), and adjustment features (causes of Y only). While **D^{2}VD** is linear, **N-D^{2}VD** (Kuang et al., 2022) replaces the linear models with neural networks, for a greater expressivity. **DR-CFR** (Hassanpour and Greiner, 2019b) pushes the division further by introducing three latent representations: one for instrumental factors (causes of T only), one for confounding factors, and one for adjustment factors. This distinction is enforced in latent space, as opposed to **D^{2}VD** and **N-D^{2}VD**, allowing for much more flexibility. Notably, the discrepancy (measured by *MMD*) between adjustment factors of control and treated samples is minimized. **TEDVAE** (Zhang et al., 2021) relies on a variational approach,

bridging the gap between variational models for *CATE* estimation (Louizos et al., 2017) and instrument/confounder/adjustment division (Hassanpour and Greiner, 2019b). In order to ensure proper disentanglement of the three latent spaces, **MIM-DRCFR** (Cheng et al., 2022b) leverages Contrastive Log-Ratio Upper Bound (Cheng et al., 2020), a mutual information approximation. Pursuing the same objective, Wu et al. (2023b)'s **DeR-CFR** resorts to a deep orthogonal regularizer, ensuring that the input covariates used to build the three latent representations involve distinct features. Curth and Schaar (2021)'s **SNet** pushes the distinction further by distinguishing adjustment factors causing Y^0 only, Y^1 only, and both of them. Finally Chauhan et al. (2023), taking inspiration from domain adaptation, resort to adversarial training to enforce better **DR-CFR**'s and **SNet**'s disentanglement.

III.3.1.3 . *X-learners*

X-learners are introduced by Künzel et al. (2019), where **the letter 'X' refers to the computation flow shape**. A two-step process is defined: In the first step, response functions μ^t and propensity $\eta(x) = E[T|X = x]$ are estimated using any learner, yielding $\hat{\mu}^t$ and $\hat{\eta}(x)$. In a second step, two *CATE* estimates are trained: $\hat{\tau}^1(x_i) \approx y_i^1 - \mu^0(x_i)$ is optimized on treated samples, while $\hat{\tau}^0(x_i) \approx \mu^1(x_i) - y_i^0$ is optimized on control ones. Finally, the *CATE* estimate for any given sample is obtained as

$$\hat{\tau}_X(x) = (1 - \hat{\eta}(x))\hat{\tau}_0(x) + \hat{\eta}(x)\hat{\tau}_1(x)$$

In the case where estimates $\hat{\mu}^0, \hat{\mu}^1, \hat{\tau}^0, \hat{\tau}^1$ are implemented and trained as neural networks, it appears that joint optimization of the first stage $(\hat{\mu}^0, \hat{\mu}^1)$ and second-stage $(\hat{\tau}^0, \hat{\tau}^1)$ models might be beneficial to the final *CATE* estimation accuracy. For further development, see Stadie et al. (2018)'s *Y-learner* and Curth and Schaar (2021)'s *RA-learner*.

III.3.1.4 . *R-learners*

R(obinson)-learners are introduced by Nie and Wager (2021), extending the *CATE* typology defined by Künzel et al. (2019). *R-learners* build upon the potential outcome formalization due to Robinson (1988) and occasionally referred to as "Robinson's transformation":

$$Y - \mathbb{E}[Y|X] = (T - \mathbb{E}(T|X))\tau(X) + \varepsilon$$

with ε a centered noise variable. Like *X-learners*, *R-learners* proceed along a two-stages approach: In the first stage, estimates $\hat{m}(x) \approx \mathbb{E}[Y|X = x]$ and $\hat{\eta}(x) \approx \mathbb{E}[T|X = x]$ are learned. In a second stage, a *CATE* estimate is sought as a minimizer of

$$\sum_i \left(y_i - \hat{m}(x_i) - (t_i - \hat{\eta}(x_i))\hat{\tau}_R(x_i)\right)^2$$

Using cross-fitting training procedures, this method provably reaches an oracle-efficient convergence rate, i.e., the same as if the ground truth functions η and m were known.

III.3.1.5 . *F-learners*

Künzel et al. (2019) define *F-learners* as follows. Denoting ϱ the inverse probability of treatment weights $\varrho^t = {}^t/_\eta + {}^{1-t}/_{1-\eta}$ (with η the propensity (Eq. II.8)), it is seen that $(2T-1)\varrho^T(X)Y$ has an expected value of $\tau(x)$ conditionally to $X = x$:

$$\mathbb{E}\Big[(2T-1)\varrho^T(X)Y|X=x\Big] = \tau(x)$$

Using the same two-stages approach as *R-learners*, a propensity estimate is first built, then $\hat{\varrho}$ is derived from $\hat{\eta}$ (Eq. II.8) and $\hat{\tau}$ is sought as a minimizer of

$$\sum_i \Big(\hat{\tau}_F(x_i) - (2t_i-1)\varrho^{t_i}(x_i)y_i\Big)^2$$

III.3.1.6 . *U-learners*

Similarly to *F-learners*, *U-learners* derive from the fact that $\frac{Y-\mathbb{E}[Y|X]}{T-\mathbb{E}[T=1|X]}$ has an expected value of $\tau(x)$ conditionally to $(X=x)$ (Signorovitch, 2007; Athey and Imbens, 2016; Curth and Schaar, 2021)[7] :

$$\mathbb{E}\Big[\frac{Y-\mathbb{E}[Y|X]}{T-\mathbb{E}[T=1|X]}|X=x\Big] = \tau(x)$$

Using the same two-stage approach as *R-learners* and *F-learners*, estimates $\hat{\eta}(x) \approx \mathbb{E}[T=1|X=x]$ and $\hat{m}(x) \approx \mathbb{E}[Y|X=x]$ are first built, then $\hat{\tau}$ is sought as a minimizer of

$$\sum_i \Big(\hat{\tau}_U(x_i) - \frac{y_i-\hat{m}(x_i)}{t_i-\hat{\eta}(x_i)}\Big)^2 = \sum_i \Big(\hat{\tau}_U(x_i) - (2t_i-1)\varrho^{1-t_i}(x_i)\big(y_i-\hat{m}(x_i)\big)\Big)^2$$

After Nie and Wager (2021), *U-learners* suffer from high instability due to their denominator $T-\hat{\eta}(x)$, and they are thus mostly investigated from a theoretical perspective only.

III.3.1.7 . *DR-learners*

DR-learners, built on the **Augmented Inverse Probability Weighting estimator** of Robins et al. (1995), are introduced by Foster and Syrgkanis (2023); Kennedy (2023) with the goal to enforce double robustness properties (Section III.2.3) in *CATE* estimation. The first steps consist in training outcome functions and propensity estimates $(\hat{\mu}^0, \hat{\mu}^1, \hat{\eta})$ on one half of the training data. The

[7] Curth and Schaar (2021) refers to *U-learners* as *PW-learners*

expected value of quantity $\mu^1(X) - \mu^0(X) + (2T-1)\varrho^T(X)(Y - \mu^T(X))$ conditionally to $(X = x)$ is equal to $\tau(x)$:

$$\mathbb{E}\Big[\mu^1(X) - \mu^0(X) + (2T-1)\varrho^T(X)(Y - \mu^T(X))|X = x\Big] = \tau(x)$$

Estimate $\hat{\tau}$ is thus sought as the minimizer over the second half of the training data of:

$$\sum_i \left(\hat{\tau}_{DR}(x_i) - \left(\hat{\mu}^1(x_i) - \hat{\mu}^0(x_i) + (2t_i - 1)\hat{\varrho}^{t_i}(x_i)(y_i - \hat{\mu}^T(x_i))\right)\right)^2$$

Swapping the two data halves, it is possible to train a second *CATE* estimate $\hat{\tau}_{DR'}$, and define the final estimate as the average value of $\hat{\tau}_{DR}$ and $\hat{\tau}_{DR'}$. Following an adequate training procedure and under mild assumptions about the convergence of the estimators, $\hat{\tau}_{DR}$ is provably Oracle efficient (see Proposition 1. and Thm 2. in Kennedy (2023) for details).

III.3.1.8 . Other architectures

The state-of-the-art mentions other alternative structures. The most related ones are the *B-learner* (Oprescu et al., 2023), that generalizes the *DR-learner* in settings where a limited amount of unobserved confounding exists, bounding on the *CATE* estimation error with respect to said level. The *IF-learner* (Curth et al., 2021a) builds a general framework to learn doubly-robust models by leveraging efficient influence functions (Hampel et al., 1986).

III.3.2 . Identifiability of latent variable causal models

Following Louizos et al. (2017), quite a few authors have investigated causal inference estimates involving deep latent-variable models. The distributional nature of these models (mostly *VAE*s) has led to **optimistic expectations regarding their ability to deal with potentially unobserved confounders**. Relaxing the conditional exchangeability hypothesis, one may consider a setting where an unobserved multidimensional variable Z is the common cause of the observed X, Y, T (Fig. III.3).

Identifying the true joint distribution over observed covariates, hidden confounders, treatment assignment, and outcome would make it possible to get accurate causal effects estimates. Assume that the true distribution

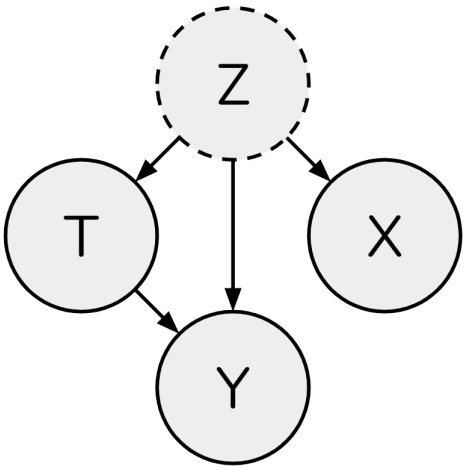


Figure III.3: Causal graph variant.

$\mathbb{P}_{X,Z,T,Y}$ is recovered. Then, using the formula of total probability:

$$\begin{aligned}
\mathbb{P}[Y^t|X=x] &= \int_{\mathcal{Z}} \mathbb{P}[Y^t|X=x, Z=z]\mathbb{P}(Z=z|X=x)dz \\
&= \int_{\mathcal{Z}} \mathbb{P}[Y^t|Z=z]\mathbb{P}(Z=z|X=x)dz \\
&= \int_{\mathcal{Z}} \mathbb{P}[Y^t|Z=z, T=t]\mathbb{P}(Z=z|X=x)dz \\
&= \int_{\mathcal{Z}} \mathbb{P}[Y|X=x, T=t]\mathbb{P}(Z=z|X=x)dz
\end{aligned}$$

and the last expression may be computed since $\mathbb{P}_{X,Z,T,Y}$ is known.

As shown by Rissanen and Marttinen (2022) however, without additional hypotheses, there is no guarantee for the latent representation of the learned models to converge towards the true hidden latent distribution Z, adversely affecting the causal estimates. A variational model may learn a distribution that perfectly fits the observational distribution $\mathbb{P}_{X,T,Y}$, without matching the underlying $\mathbb{P}_{X,Z,T,Y}$, leading to erroneous causal effects estimates. If the model is misspecified (typically if the sought number of features in the latent space is underestimated or overestimated), it is likely to fail at modeling the observational distribution.

This raises the question of the requirements on the observational data that must be satisfied for the latent space Z to be identifiable[8]. The notion of "identifiable" latent space is itself open to interpretation, e.g., two models that only differ by a permutation of their latent space features are equivalent,

[8] noting that this notion of identifiability differs from that of causal identifiability (Section II.4)

but more complex transformations of the latent space might also preserve the model.

The development of Nonlinear Independent Component Analysis (*NICA*) (Hyvarinen and Morioka, 2016; Hyvarinen et al., 2019) makes it possible to characterize when and how the latent variable model identification is possible. Khemakhem et al. (2020) first bridge the gap between *NICA* and *VAE*, requiring the true latent distribution to admit a factorization conditioned on an extra observed variable. This setting is later on extended by Hyvärinen et al. (2023). Wu and Fukumizu (2021)'s β-*Intact-VAE* model leverages this property to guarantee accurate causal estimates even when facing limited breaches of overlap.

III.3.3 . Discussion

A fundamental difficulty for *CATE* estimation is that the counter-factual outcome is unknown by construction, preventing any trained model from being validated using standard ML protocols. Overall, **building datasets to benchmark causal inference models** is an arduous task.

The semi-synthetic *IHDP* benchmark (Hill, 2011) is extensively used to validate current *CATE* approaches since (Johansson et al., 2016), where the outcomes are simulated using known functions (we shall return to *IHDP* in Section V.1.1.1). Curth et al. (2021b) has criticized this dataset as *CATE* estimators benchmark in several respects. Notably, simulating outcome functions with a mechanism of the "surfaces of contact" form

$$Y = (1 - T) \times \mu^0(X) + T \times \mu^1(X) + \varepsilon$$

favors both *T-learner* and *S-learner* approaches, as opposed to mechanisms obtained through the "Robinson's transformation" (Robinson, 1988; Nie and Wager, 2021), where

$$Y = m(X) + (T - \eta(X)) \times \tau(X) + \varepsilon$$

that favors *R-learners*, *F-learners*, *U-learners*, *DR-learners*.

The literature in the causal inference domain has elected *IHDP* as its standard meter for the evaluation of *CATE* models. It is suggested that this choice has biased the research toward *T-learners* at the detriment of *R-learners*. A larger variety of benchmarks might help to increase the robustness of *CATE* estimators and facilitate their usage in real-life situations (see also Curth et al. (2021b)). We shall return to causal inference benchmarks in Section V.3.

A toy problem is explored in Appendix D to illustrate the impact of the data generation process on the ability of different meta-learners to model the causal effects.

III.4 . Partial conclusion

This chapter, primarily focused on *CATE*, describes the main approaches of the state of the art. The essential aspects regard: i) whether a change of representation is needed, and when it is the case, what are the properties the latent representation should satisfy; ii) the (im)possibility to evaluate *CATE* models; iii) how the existing benchmarks might bias the experimental validation of the *CATE* approaches.

Chapter IV will analyze and discuss in more detail the first two aspects, that motivate the proposed ALRITE approach. We shall return to the third aspect in Chapter VI.

IV - Asymmetrical Latent Regularization for Individual Treatment Effect Modeling

This chapter describes the core contribution of the manuscript. We first discuss the main issues related to the state of the art (Section IV.1), and deduce the very principles of ALRITE (Section IV.2). Its overview, meant to address some of these issues, is given in Section IV.3, together with ensemble-based variants thereof (Section IV.4). The theoretical analysis of the approach, upper bounding the estimation error under mild assumptions, is described (Section IV.5), and its scope is discussed.

IV.1 . Motivations

IV.1.1 . Revisiting the state of the art

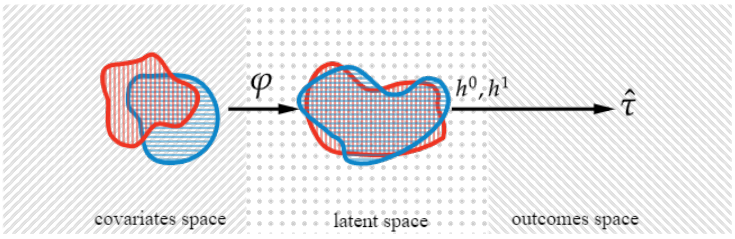
As said (Section III.3.1.2), the T-learner model proposed by Shalit et al. (2017) has established the merits of changing representations to accurately estimate counter-factual outcomes for both the control and treated distributions. The change of representation is meant to **flexibly and reliably accommodate these two distinct estimation objectives**, and the neural network-based latent representation aims to find **a trade-off** between both.

In contrast, *X-learners* (Section III.3.1.3) face **no such trade-off** among the two modeling tasks, but **lack the increased flexibility** that latent spaces offer. They proceed by independently learning two causal effects estimates $\hat{\tau}_0$ and $\hat{\tau}_1$, and define the overall $\hat{\tau}$ as a weighted combination thereof using the estimated propensity $\hat{\eta}$:

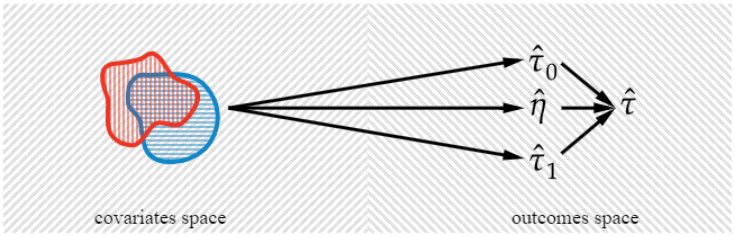
$$\hat{\tau}_X = (1 - \hat{\eta}) \times \hat{\tau}_0 + \hat{\eta} \times \hat{\tau}_1$$

Related to the change of representation is the effort to balance the control and treated distributions. To our best knowledge, no consensus has yet arisen concerning the most appropriate way to do so.

From the early works on neural-network enforced *CATE* estimators (namely, Johansson et al. (2016)'s *BNN*), balance is enforced through a penalization term in the learning loss. Shalit et al. (2017) resorts to Integral Probability Metrics such as *CFR-MMD*'s Maximum Mean Discrepancy (Gretton et al., 2012) or *CFR-Wass*'s Sinkhorn distance (Cuturi, 2013). Yao et al. (2018)'s *SITE* aims at preserving local similarity while mapping samples towards the latent space, and Du et al. (2021)'s *CBRE* leverages advances of adversarial learning to make latent distributions indistinguishable.



(a) T-learner architecture.



(b) X-learner architecture.

Figure IV.1: Contrasting the *T*- and *X-learner* architectures.

Notably, Zhang et al. (2020) provides strong theoretical groundings to its *DKLITE*-specific focus on **counter-factual variance**. This approach can evaluate model uncertainty by resorting to deep neural kernels while computing counter-factual predictions. High uncertainty, i.e., high variance on the counter-factual posterior distribution, attests that there is insufficient information regarding the counter-factual distribution in the neighborhood of a given sample. As such, when encouraged to learn a mapping ϕ associated with low counter-factual variance, *DKLITE* ensures that no multiple models may be learned with similar factual predictions but divergent counter-factual ones. Evaluation of counter-factual variance, however, is a delicate task, and the proposed solution imposes specific choices of architecture. In particular, there is **no clear insight on how such an approach could be extended** outside the scope of Bayesian models.

IV.1.2 . Proposed requirements for *CATE* estimator latent representations

Neural networks are celebrated for their generality and flexibility, and the fact that the statistical properties of the trained models can be shaped or enforced through the specifics of the neural architecture. In the following, the proposed standpoint on *CATE* inference is that of learning two models with missing data. The desired change of representation thus is tailored to the specifics of this goal.

Let ϕ map the input space $\mathcal{X}$ to latent space $\mathcal{Z}$. Consider $(z_i = \phi(x_i), t_i, y_i)$, the image of a sample taken from the observational distribution. With no

loss of generality, let us assume that $t_i = 1$ (the same reasoning follows by symmetry for $t_i = 0$).

Measuring the causal effect conditionally to $X = x_i$ means estimating $\mathbb{E}[Y^1 - Y^0 | X = x_i]$. The goal can intuitively be reformulated as estimating $\mathbb{E}[Y^1 | \phi(X) = \phi(x_i)]$ and $\mathbb{E}[Y^0 | \phi(X) = \phi(x_i)]$ for ϕ "sufficiently" injective (see IV.5.3.2 for a formal characterization).

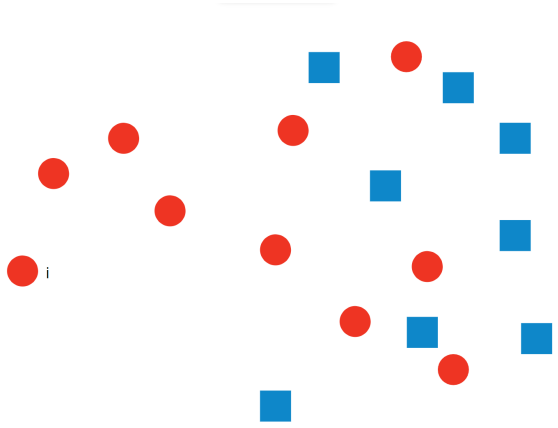


Figure IV.2: Picturing a representation space, with treated and control samples. For the leftmost treated samples, the estimation of y_i^0 is likely unreliable since no information about Y^0 is available in that region of space. Note that the overlap is not symmetrical: the lack of neighbors with opposite treatment assignment only concerns the treated samples.

The difficulty arises when the treated sample $(z_i, t_i = 1, y_i)$ is **far from any control sample** $(z_j = \phi(x_j), t_j = 0, y_j)$ in the sense of the representation distance (see e.g., the leftmost point in Fig. IV.2). Although y_i and the outcome value of close treated samples provide information about the local behavior of $z \mapsto \mathbb{E}[Y^1 | \phi(x) = z]$, there exists no guarantee regarding the estimation of the counter-factual $\mathbb{E}[Y^0 | \phi(X) = z]$. The obtained estimate can be **arbitrarily inaccurate** – unless strong assumptions (e.g., linearity or high smoothness) are made on the potential outcomes. With a specific focus on the control distribution, this issue can be reframed as an out-of-distribution estimation problem.

This discussion suggests that the latent space supporting the counter-factual estimation of Y^1 (respectively Y^0) must be such that **no sample be isolated from the samples in the other distribution**.

Remark: The main difficulty of *CATE* estimation is in the low-sample regime, when the observational dataset contains few samples relatively to the dimension of the covariate space. In particular, the accuracy of the factual outcome estimates increases as the total number of samples tends to infinity, and the

CATE estimate likewise reaches similar accuracy under the positivity assumption (Eq. II.4 and Alaa and Schaar (2018)).

IV.2 . Principle of ALRITE

IV.2.1 . Notations

Let ϕ be a mapping from covariates space $\mathcal{X} \subset \mathbb{R}^d$ to a latent space $\mathcal{Z} = \phi(\mathcal{X}) \subset \mathbb{R}^{d'}$, with $||\cdot||_2$ the Euclidean distance among real-valued vectors. The quantities introduced in the following depend on the pseudo-distance d_ϕ defined on $\mathcal{X}^2$ as $d_\phi(x, x') = ||\phi(x) - \phi(x')||_2$. All samples belong to an observational dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i \in [\![1,n]\!]}$.

Let the **mirror twin** of a sample be defined as the closest neighbor w.r.t d_ϕ with opposite treatment assignment[1] (Fig. IV.3a). Supercharging the notation, let us then denote the mirror twin of sample i by $\phi(i)$:

$$\phi(i) = \underset{j \in [\![1,n]\!] \ s.t. \ t_j = 1 - t_i}{\operatorname{argmin}} \{d_\phi(x_i, x_j)\} \tag{IV.1}$$

The **insulation** of a sample is defined as its distance to its mirror twin w.r.t ϕ, with the following notation:

$$\mathit{insulation}_\phi(i) = d_\phi(x_i, x_{\phi(i)})$$

Let the **exemplarity** of a sample be defined as the number of samples admitting it as their mirror twin:

$$\mathit{exemplarity}_\phi(i) = |\{j \in [\![1, n]\!] \ s.t. \ \phi(j) = i\}|$$

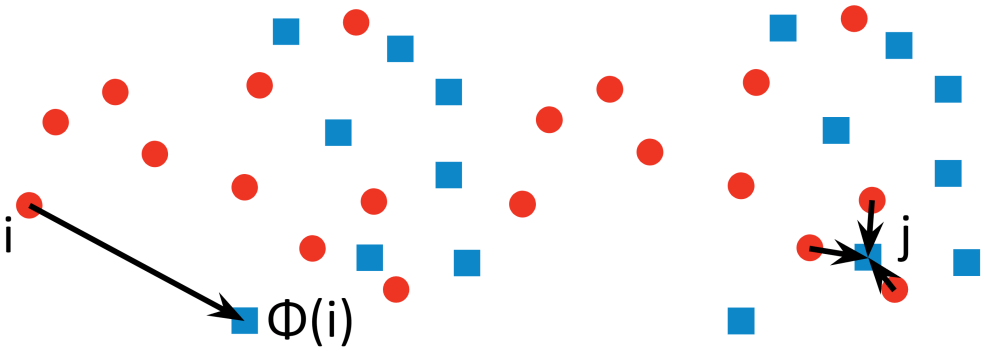
As shown in Fig. IV.3b, the i-th sample might be the mirror twin of several j-th samples (with $t_j = 1 - t_i$) thus with high insulation.

The purpose of these definitions is the following. Informally, if the insulation for every control sample $(x, t = 0)$ is low (i.e., if its mirror twin is close) and if an accurate model h^1 is trained from the treated samples on this same control latent space, then $h^1(x)$ also is a potentially accurate estimate of $\mu^1(x)$.

[1] Ties are unlikely if $\mathcal{X}$ is continuous but may arise if it is discrete. If multiple samples with opposite treatment assignment minimize the d_ϕ-distance to (x_i, t_i), that with minimal index is defined as its mirror twin.

The influence of the nearest sample with opposite treatment assignment in covariates spaces has been considered in the literature (Johansson et al., 2016; Wu et al., 2023a). A fundamental difference lies in the fact that the mirror twin of ALRITE is defined with respect to the representational distance, and may as such be leveraged to constrain the representational space.

Note also that the mirror twin operator is no involution in general.



(a) The **mirror twin** of i is its closest neighbor from the control distribution
The **insulation** of i is high.

(b) j is the mirror twin of 3 samples; its **exemplarity** is 3.

Figure IV.3: Illustration of the notions of insulation and exemplarity in the latent space, with control samples and treated ones

In the case where the sought μ^1 is L-lipschitzian, for any sample control sample i with mirror twin $\phi(i)$, the counter-factual error of i can be bounded as:

$$|\mu^1(x_i) - \mu^1(x_{\phi(i)})| \leq L||x_i - x_{\phi(i)}||$$

This remark will be adapted and formalized in Section IV.5.1, yielding an upper bound on the estimation error under mild assumptions.

Simple calculations show that:

$$\begin{cases} \sum_{t_i=0} \textit{exemplarity}_\phi(i) = |\{j \in [\![1,n]\!] \text{ s.t. } t_j = 1\}| \\ \sum_{t_i=1} \textit{exemplarity}_\phi(i) = |\{j \in [\![1,n]\!] \text{ s.t. } t_j = 0\}| \end{cases}$$

Remark: In the large sample regime, assuming ϕ fixed and under the positivity assumption, the insulation goes to 0 (Footnote 7). To our understanding, the asymptotical behavior of exemplarity remains an open question, even in low-dimensional settings (Ferenc and Néda, 2007).

Overall, let a **pipeline** be defined as the triplet formed by a representation network $\phi : \mathcal{X} \to \mathcal{Z}$, and two functions $h^0, h^1 : \mathcal{Z} \to \mathbb{R}$ entailing a control and a treatment outcome estimate $h^0 \circ \phi, h^1 \circ \phi$. By definition, pipeline $\mathcal{P}$, denoted $\mathcal{P} = (\phi, h^0, h^1)$, can be seen as a T-learner and yields a candidate *CATE* estimate as:

$$x : \mathcal{X} \mapsto h^1 \circ \phi(x) - h^0 \circ \phi(x)$$

In the following, the mapping, control outcome function and treated outcome function of pipeline $\mathcal{P}$ will be respectively denoted as $\phi_\mathcal{P}, h^0_\mathcal{P}, h^1_\mathcal{P}$. The entailed mirror twin, insulation and exemplarity functions will accordingly be denoted as $\phi_\mathcal{P}, \textit{insulation}_\mathcal{P}, \textit{exemplarity}_\mathcal{P}$.

IV.2.2 . Model architecture

The *CATE* estimation problem and requirements are revisited using the notions of *mirror twin* and *insulation*. As discussed in Section IV.1.1, the sought solution involves an embedding, inducing a latent representation of the observed samples and enforcing two properties (Section IV.1.2): i) insulation of treated samples should be small and ii) insulation of control samples should be small.

Interestingly, the state-of-the-art approaches achieve balance through symmetrical regularisation constraints. However, while balance is a proxy goal for both above properties, it does not efficiently enforce either one, as graphically shown on Fig. IV.4:

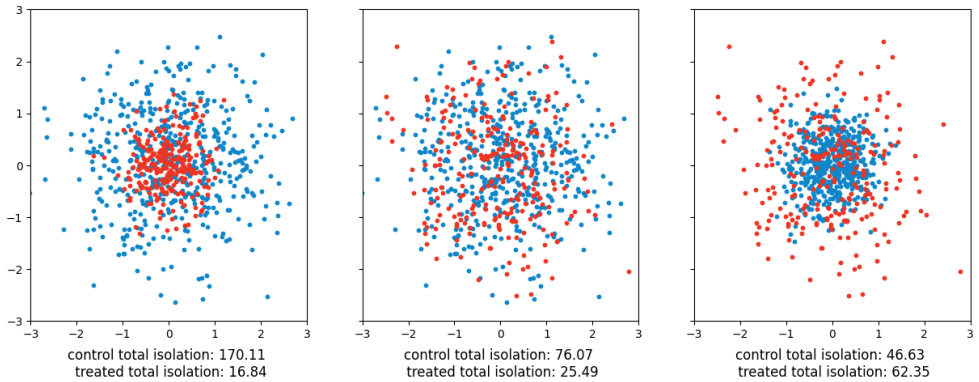


Figure IV.4: The properties associated with latent spaces. Left: low insulation for treated samples. Middle: balance of treatment and control distributions. Right: low insulation for control samples.

It thus comes to define **two latent spaces**, respectively enforcing a low insulation for the treated and the control samples. Specifically, the so-called *treatment-driven* pipeline aims to causal inference for treated samples, by enforcing a low insulation for treated samples (task i); symmetrically, the *control-driven* pipeline aims to causal inference for control samples and enforces a low insulation for control samples (task ii).

Let us respectively denote $\hat{\tau}_{\mathcal{P}_0}$ and $\hat{\tau}_{\mathcal{P}_1}$ the causal estimates provided by pipeline $\mathcal{P}_0$ and $\mathcal{P}_1$. These are combined using propensity score estimate $\hat{\eta}$, yielding:

$$\hat{\tau} : x \in \mathcal{X} \mapsto (1 - \hat{\eta}(x))\hat{\tau}_{\mathcal{P}_0}(x) + \hat{\eta}(x)\hat{\tau}_{\mathcal{P}_1}(x) \tag{IV.2}$$

Overall, the proposed ALRITE (Asymmetrical Latent Representation for Individual Treatment Effect) bridges the gap between T-learners and X-learners: each one of pipelines $\mathcal{P}_0$ and $\mathcal{P}_1$ defines a T-learner, and their causal estimate $\hat{\tau}_{\mathcal{P}_0}$ and $\hat{\tau}_{\mathcal{P}_1}$ are combined in Eq. IV.2 and Table IV.1 as in *X-learners*. **As such, ALRITE instantiates a new class of meta-learners**.

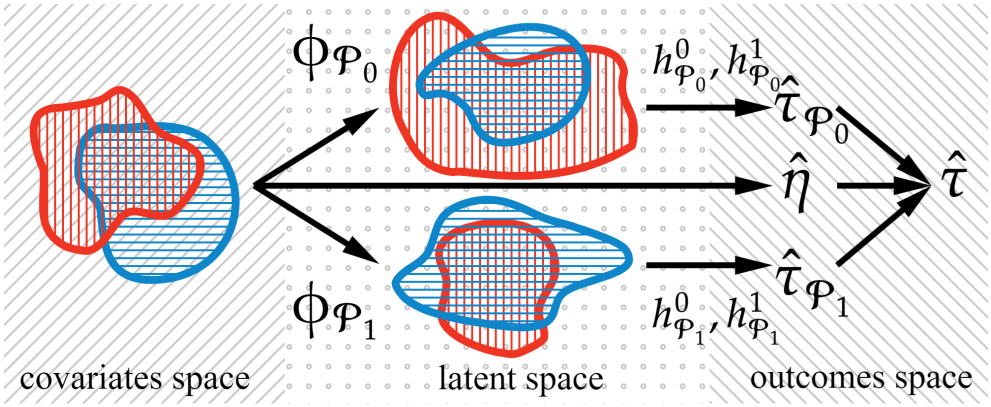


Figure IV.5: ALRITE architecture, consisting of the control-driven pipeline $\mathcal{P}_0$ (top), the propensity estimate (middle) and the treatment-driven pipeline $\mathcal{P}_1$ (bottom).

		Treatment assignment		
		$T=0$	$T=1$	T unknown
predicted outcome	Y^0	$h^0_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}$	$h^0_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$	$(1-\hat{\eta}) \times h^0_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0} + \hat{\eta} \times h^0_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$
	Y^1	$h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}$	$h^1_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$	$(1-\hat{\eta}) \times h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0} + \hat{\eta} \times h^1_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$

Table IV.1: Potential outcomes estimates.

IV.3 . Algorithm

This section details the three branches in the ALRITE architecture (Fig. IV.5): the pipelines (top and bottom branches); the propensity score (middle); and how the estimates built in all three branches are combined to form a *CATE* estimate.

IV.3.1 . Pipelines training

With no loss of generality, let us focus on learning pipeline $\mathcal{P}_0$; learning $\mathcal{P}_1$ follows by symmetry, replacing subscripts $\square_{\mathcal{P}_0}$ with $\square_{\mathcal{P}_1}$.

As said, pipeline $\mathcal{P}_0 = (\phi_{\mathcal{P}_0}, h^0_{\mathcal{P}_0}, h^1_{\mathcal{P}_0})$ induces a T-learner focused on causal inference for control samples. Specifically, embedding $\phi_{\mathcal{P}_0}$ must i) enable the learning of accurate outcome models $h^0_{\mathcal{P}_0}$ and $h^1_{\mathcal{P}_0}$ trained from control and treated samples; ii) yield a low insulation on average for the control samples.

The training loss depends on the nature of the outcome variable Y:

1. In the continuous case ($\mathcal{Y}$ is an interval), $h^0_{\mathcal{P}_0}$ and $h^1_{\mathcal{P}_0}$ are trained to solve a regression problem, and the mean square error is used to compute

$$\sigma : r \in \mathbb{R} \mapsto \frac{1}{1+\exp(-r)}$$

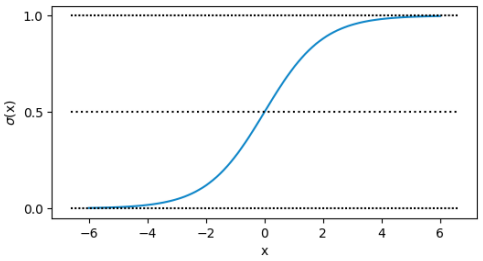


Figure IV.6: Logistic activation function.

the error of $\mathcal{P}_0$:

$$error_{\mathcal{P}_0}(x_i, t_i, y_i) = (h^{t_i}_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i)^2$$

2. In the binary case ($\mathcal{Y} = \{0, 1\}$), the potential outcomes are trained using the cross-entropy loss[2]. Formally, for σ denoting the logistic function (Fig. IV.6), the model returns an outcome estimate set to 0 if $\sigma\left(h^{t_i}_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i)\right) < .5$ and 1 otherwise, with:

$$error_{\mathcal{P}_0}(x, t, y) = -y \log\left(\sigma(h^t_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x))\right) - (1\text{-}y) \log\left(1\text{-}\sigma(h^t_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x))\right)$$

Overall, pipeline $\mathcal{P}_0 = (\phi_{\mathcal{P}_0}, h^0_{\mathcal{P}_0}, h^1_{\mathcal{P}_0})$ (Table IV.1) is trained end-to-end to

[2]The function $f : x \in \mathbb{R} \mapsto -\alpha \log\left(\sigma(x)\right) - (1-\alpha) \log\left(1-\sigma(x)\right)$ has derivative $f' = \sigma - \alpha$ and curvature $f'' = (1-\sigma)\sigma$, thus admits a global minimum at $\sigma^{-1}(\alpha)$. Using the formula of total expectation on T then X, it comes

$$\begin{aligned}
\mathbb{E}[error_{\mathcal{P}_0}(X,T,Y)] =& \mathbb{E}_{X|T=1}[\mathbb{E}[error_{\mathcal{P}_0}(X,T,Y)|X,T=1]|T=1]\mathbb{P}(T=1) \\
&+\mathbb{E}_{X|T=0}[\mathbb{E}[error_{\mathcal{P}_0}(X,T,Y)|X,T=0]|T=0]\mathbb{P}(T=0) \\
=& -\mathbb{E}_{X|T=1}\left[\mu^1(X) \log\left(\sigma\left(h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(X)\right)\right)\right. \\
&\left. + (1-\mu^1(X)) \log\left(1-\sigma\left(h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(X)\right)\right)\right] \\
&-\mathbb{E}_{X|T=0}\left[\mu^0(X) \log\left(\sigma\left(h^0_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(X)\right)\right)\right. \\
&\left. + (1-\mu^0(X)) \log\left(1-\sigma\left(h^0_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(X)\right)\right)\right]
\end{aligned}$$

and this term is thus minimized by $(\sigma \circ h^0_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}, \sigma \circ h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}) = (\mu^0, \mu^1)$. Cross-entropy is a relevant choice of prediction loss.

minimize the compound loss:

$$\begin{aligned}\mathcal{L}_{\mathcal{P}_0} =& \frac{1}{n_0}\sum_{t_i=0} \mathit{error}_{\mathcal{P}_0}(x_i, t_i, y_i) \\ &+ \frac{\alpha_{\mathcal{P}_0}}{n_0}\sum_{t_i=0} \mathit{insulation}_{\mathcal{P}_0}(i)^2 \\ &+ \frac{1}{n_1 + \beta_{\mathcal{P}_0} n_0}\sum_{t_i=1} (1 + \beta_{\mathcal{P}_0} \mathit{exemplarity}_{\mathcal{P}_0}(i)) \times \mathit{error}_{\mathcal{P}_0}(x_i, t_i, y_i) \\ &+ \gamma_{\mathcal{P}_0}\, \Omega(\mathcal{P}_0)\end{aligned} \tag{IV.3}$$

where:

1. the first term is the **factual prediction loss over control samples**, to be minimized; as said, the error function depends on the domain $\mathcal{Y}$ of the outcome;
2. the second term is the **mean squared insulation** of control samples in the representation space entailed by $\phi_{\mathcal{P}_0}$, weighted by hyper-parameter $\alpha_{\mathcal{P}_0}$. Its minimization ensures that any control sample gets close to its mirror twin in terms of latent distance;
3. the third term is the **factual error of treated samples**, to be minimized. Note that samples with high exemplarity have an extra error weight (controlled from hyper-parameter $\beta_{\mathcal{P}_0}$) to account for the fact that any error on such a treated sample might adversely affect the counter-factual estimate for all control samples in its neighborhood. Inversely, a treated sample with null exemplarity has a lesser impact on the counter-factual estimate of the control samples;
4. the last term $\Omega(\mathcal{P}_0)$ is a **regularization** term, computing the sum of the square of all weights matrices in the neural networks defining $h^0_{\mathcal{P}_0}$, $h^1_{\mathcal{P}_0}$ and (optionally) in $\phi_{\mathcal{P}_0}$, to be minimized. This term, controlled by hyper-parameter $\gamma_{\mathcal{P}_0}$, is meant to avoid overfitting.

The first, second and third terms are normalized to ensure their same impact on the overall loss (the average prediction error of $h^0_{\mathcal{P}_0}$ and $h^1_{\mathcal{P}_0}$), with n_0 and n_1 respectively denoting the number of control and treated samples.

IV.3.2 . Propensity estimation

Contrarily to the estimation of causal effects, propensity score estimation is tackled as a mainstream **binary supervised learning problem**. The propensity estimate $\hat{\eta}$, meant to approximate the propensity score $\eta : x \in \mathcal{X} \mapsto \mathbb{E}[T|X = x]$, is trained using binary cross-entropy loss, possibly augmented with a regularization term to prevent over-fitting (independently from the training of pipelines $\mathcal{P}_0$ and $\mathcal{P}_1$). Denoting by p the proportion of treated

samples in the training set:

$$\mathcal{L}_{\hat{\eta}} = -\frac{1}{n}\sum_{i=1}^{n} \frac{t_i}{p}\log\left(\hat{\eta}(x_i)\right) + \frac{1-t_i}{1-p}\log\left(1-\hat{\eta}(x_i)\right) + \Omega_{reg}(\hat{\eta}) \qquad \text{(IV.4)}$$

As said, $\hat{\eta}$ is trained as a binary classifier. It is desirable that $\hat{\eta}$ be calibrated (Zadrozny and Elkan, 2002), i.e. such that

$$\forall s \in [0,1], \mathbb{E}[T = 1 | \hat{\eta}(X) = s] = s$$

However, perfect calibration is not required to obtain satisfying *CATE* estimation results (see below, Eq. IV.5).

Note that reweighting terms $(\frac{\cdot}{p}, \frac{\cdot}{1-p})$ are of particular importance here. In many real-life settings, as treatment is expectedly complex and expensive, control samples usually outnumber treated ones. Furthermore, if e.g., demographic or medical conditions dictate the treatment assignment, the support of the treatment distribution is typically included in that of the control distribution. In such cases, the regularization of $\mathcal{P}_1$ matters more than that of $\mathcal{P}_0$ (we shall return to this Section V.1.2). The reweighting contributes toward the greater importance of pipeline $\mathcal{P}_1$ in the final estimate $\tau = (1-\hat{\eta})\hat{\tau}_{\mathcal{P}_0} + \hat{\eta}\hat{\tau}_{\mathcal{P}_1}$.

IV.3.3 . Definition of $\hat{\tau}$

ALRITE last aggregates the estimates learned in all three branches (Fig. IV.5), i.e., the trained *control-driven* and *treatment-driven* pipelines, and the propensity estimate to deliver a *CATE*.

Each pipeline, being a T-learner by itself, induces an estimate of τ:

$$\hat{\tau}_{\mathcal{P}_0} = h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0} - h^0_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}$$
$$\hat{\tau}_{\mathcal{P}_1} = h^1_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1} - h^0_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$$

These are combined (Eq. IV.2) using the propensity score estimate $\hat{\eta}(x)$ to define an overall estimate of τ:

$$\hat{\tau}(x) = (1-\hat{\eta}(x)) \times \hat{\tau}_{\mathcal{P}_0}(x) + \hat{\eta}(x) \times \hat{\tau}_{\mathcal{P}_1}(x)$$

This combination accounts for the fact that estimate $\hat{\tau}_{\mathcal{P}_0}$ (respectively, $\hat{\tau}_{\mathcal{P}_1}$) has been obtained through $\mathcal{P}_0$ (resp. $\mathcal{P}_1$), focused on providing accurate causal effect estimates for control (resp. treated) samples. When inferring causal effects on new data, Eq. IV.2 thus relies more on $\hat{\tau}_{\mathcal{P}_0}$ for samples that are likely to get treatment assignment 0, more on $\hat{\tau}_{\mathcal{P}_1}$ for samples that are likely to get treatment assignment 1.

The sensitivity of $\hat{\tau}$ to propensity misspecification is analyzed as follows. Let us denote $||.||$ the L_2 distance over $(\mathbf{X}, \mathbb{P})$. Let $\hat{\tau}_\eta(x)$ denote the estimate combining $\hat{\tau}_{\mathcal{P}_0}$ and $\hat{\tau}_{\mathcal{P}_1}$ with the (true) propensity η:

$$\hat{\tau}_\eta(x) = \eta(x)\hat{\tau}_{\mathcal{P}_1}(x) + (1 - \eta(x))\hat{\tau}_{\mathcal{P}_0}(x)$$

The sensitivity of $\hat{\tau}$ w.r.t. $\hat{\eta}$, that is, the difference between $\hat{\tau}$ and $\hat{\tau}_\eta$ can be bounded as:

$$\begin{aligned} ||\hat{\tau} - \hat{\tau}_\eta|| &= ||(\hat{\eta} - \eta)(\hat{\tau}_{\mathcal{P}_1} - \hat{\tau}_{\mathcal{P}_0})|| \\ &= ||\big(\hat{\eta} - \eta\big)\big((\hat{\tau}_{\mathcal{P}_1} - \tau) - (\hat{\tau}_{\mathcal{P}_0} - \tau)\big)|| \end{aligned} \tag{IV.5}$$

This error thus is of order 2: the performance of ALRITE is the product of i) the error on the propensity; ii) the difference between the errors on $\hat{\tau}_{\mathcal{P}_0}$ and $\hat{\tau}_{\mathcal{P}_1}$. In other words, the misspecifications on the propensity are mostly harmful in regions where $\hat{\tau}_{\mathcal{P}_0}$ and $\hat{\tau}_{\mathcal{P}_1}$ are also misspecified.

IV.4 . Ensemble ALRITE

As ensemble methods are renowned for their high accuracy in supervised machine learning (Dietterich, 2000; Ganaie et al., 2022), it comes naturally to see how ensembles can be built on the top of the ALRITE architecture, where each ALRITE model consists of *control-driven* and *treatment-driven* pipelines, and propensity estimate $\hat{\eta}$.

Let us consider a set of C ALRITE pipelines, e.g., built using different hyper-parameters (more on hyper-parameter selection in Chapter VI). Letting $\mathcal{D}_\mathcal{V} = \{(x, t, y)\}$ denote an observational *validation* dataset, the factual, empirical error of the c-th pipeline $(\phi^{(c)}, h^{(c),0}, h^{(c),1})$ is approximated as[3]:

$$Err(c) = \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y) \in \mathcal{D}_\mathcal{V}} \left(h^{(c),t}(x) - y\right)^2 \tag{IV.6}$$

Two ensemble ALRITE models are defined as follows:

IV.4.1 . *top-K* ensemble

An extra hyper-parameter is involved, the number K of models considered in the ensemble ($K \in [\![1, C]\!]$). K *control-driven* pipelines are selected based on their factual error (Eq. IV.6), and the ensemble control estimate $\hat{\tau}^{top\text{-}K}_{\mathcal{P}_0}$ is defined as the average of the estimates $\hat{\tau}_{\mathcal{P}_0}$ associated to the selected pipelines.

Likewise, the K *treatment-driven* pipelines are selected based on their factual error (noting that the selection of control-driven and treatment-driven

[3] in Chapter VI this quantity will be introduced as the simple estimate of the μ-*risk* of $(h^{(c),0}, h^{(c),1})$, and denoted as $\widehat{\mu\text{-}risk}_{\mathcal{D}_\mathcal{V}}(h^{(c)})$

pipelines are independent), and the average of the associated estimates $\hat{\tau}_{\mathcal{P}_1}$ defines the ensemble treatment estimate $\hat{\tau}^{top\text{-}K}_{\mathcal{P}_1}$.

The ensemble control and treatment estimate are combined as usual (Eq. IV.2), with $\hat{\eta}$ a propensity score (learned once for all; as shown in Eq. IV.5 the approach is less sensitive to errors on $\hat{\eta}$):

$$\hat{\tau}^{top\text{-}K} : x \in \mathcal{X} \mapsto (1 - \hat{\eta}(x))\hat{\tau}^{top\text{-}K}_{\mathcal{P}_0}(x) + \hat{\eta}(x)\hat{\tau}^{top\text{-}K}_{\mathcal{P}_1}(x)$$

IV.4.2 . *softmax*$_\lambda$ ensemble

While the *top-K* ensemble involves a boolean selection of the control and treatment estimates, the *softmax*$_\lambda$ ensemble achieves a weighted combination thereof which, depending on the temperature hyper-parameter $\lambda \in \mathbb{R}^*_+$, ranges from the equi-weighted one ($\lambda \to 0^+$) to the selection of the only best one ($\lambda \to +\infty$). Let us recall the *softmax*$_\lambda$ function:

$$softmax_\lambda : (x_1, \ldots, x_C) \in \mathbb{R}^C \mapsto \left(\frac{\exp\{-\lambda x_1\}}{\sum_c \exp\{-\lambda x_c\}}, \ldots, \frac{\exp\{-\lambda x_C\}}{\sum_c \exp\{-\lambda x_c\}} \right)$$

It then comes naturally to define the *softmax*$_\lambda$ control estimate as the weighted average of the $\{\hat{\tau}^c_{\mathcal{P}_0}\}_{c \in [\![1,C]\!]}$, where the weight of the c'-th estimate is given as the softmax of the error of the associated outcome models $Err(c')$:

$$\frac{\exp\{-\lambda Err(c')\}}{\sum_c \exp\{-\lambda Err(c)\}}$$

The *softmax*$_\lambda$ treatment estimate is defined in the same way, and the overall *softmax*$_\lambda$ estimate is defined as usual (Eq. IV.2):

$$\hat{\tau}^{softmax_\lambda}(x) = (1 - \hat{\eta}(x))\hat{\tau}^{softmax_\lambda}_{\mathcal{P}_0}(x) + \hat{\eta}(x)\hat{\tau}^{softmax_\lambda}_{\mathcal{P}_1}(x)$$

IV.5 . Analysis

This section presents the analysis of the proposed approach. The *PEHE* associated with a *T-learner* and with the whole ALRITE model are provably bounded in Section IV.5.1. These results and the underlying assumptions are discussed in comparison with Shalit et al. (2017) (Section IV.5.2), focusing in particular on the ability of the practitioner to appreciate the assumptions (Section IV.5.3). The limitations of the approach are discussed in Section IV.5.4.

IV.5.1 . Upper bounding the *PEHE* of ALRITE

Let us first consider the accuracy of the *CATE* estimate induced by a single pipeline (the T-learner setting). We show that it is upper bounded depending on the factual error on the potential outcome estimates, augmented with the cumulative insulation of the samples. Note that this result holds for the empirical quantities of interest, as opposed to, asymptotically.

Theorem 1. *Let $\phi : \mathcal{X} \to \mathcal{Z}$ be a mapping from the observable feature space to a latent space. Assume there exist two functions ν^0, ν^1 with Lipschitz constant L such that $\mu^0 = \nu^0 \circ \phi$ and $\mu^1 = \nu^1 \circ \phi$. Let $\hat{\nu}^0, \hat{\nu}^1 : \mathcal{Z} \to \mathbb{R}$ be two hypothesis functions with Lipschitz constant $\hat{L}$. Then the empirical PEHE $= \frac{1}{n}\sum_{i=1}^{n}\left((\hat{\nu}^1 - \hat{\nu}^0)(\phi(x_i)) - (\mu^1 - \mu^0)(x_i)\right)^2$ is upper bounded by M_1, with*

$$M_1 = \frac{4}{n}\sum_{i=1}^{n}\Big(1 + \textit{exemplarity}_{\phi}(i)\Big)\Big(\hat{\nu}^{t_i}\circ\phi(x_i) - \mu^{t_i}(x_i)\Big)^2 + \frac{4}{n}(L^2 + \hat{L}^2)\sum_{i=1}^{n}\textit{insulation}_{\phi}(i)^2$$

Proof. Let (x, t, y) be a sample in $\mathcal{D}$ with $z = \phi(x)$, and let $(x', t' = 1 - t, y')$ be its mirror twin w.r.t ϕ ($\phi(x') = z'$). Assuming with no loss of generality that $t = 1$, it comes:

$$\begin{aligned}
&|(\hat{\nu}^1 - \hat{\nu}^0)(z) - (\nu^1 - \nu^0)(z)|^2\\
&= |[\hat{\nu}^1(z) - \nu^1(z)] - [\hat{\nu}^0(z) - \hat{\nu}^0(z')] - [\hat{\nu}^0(z') - \nu^0(z')] - [\nu^0(z') - \nu^0(z)]|^2\\
&\leq 4|\hat{\nu}^1(z) - \nu^1(z)|^2 + 4|\hat{\nu}^0(z) - \hat{\nu}^0(z')|^2 + 4|\hat{\nu}^0(z') - \nu^0(z')|^2 + 4|\nu^0(z') - \nu^0(z)|^2\\
&\leq 4|\hat{\nu}^1(z) - \nu^1(z)|^2 + 4|\hat{\nu}^0(z') - \nu^0(z')|^2 + 4(\widehat{L}^2 + L^2)||z' - z||^2
\end{aligned}$$

Averaging over (x, t, y) in $\mathcal{D}$ yields the result. □

Note that the empirical *PEHE* upper bound and the training loss involve similar terms: the mean squared error on predictions, reweighted in accordance with the exemplarity, and averaged insulation. The latter involves two terms:

$$\begin{aligned}
\frac{4(L^2 + \hat{L}^2)}{n}\sum_{i=1}^{n}\textit{insulation}_{\phi}(i)^2 &= 4\frac{n_0}{n}(L^2 + \hat{L}^2) \times \frac{1}{n_0}\sum_{t_i=0}\textit{insulation}_{\phi}(i)^2\\
&\quad + 4\frac{n_1}{n}(L^2 + \hat{L}^2) \times \frac{1}{n_1}\sum_{t_i=1}\textit{insulation}_{\phi}(i)^2
\end{aligned}$$

Each of these terms is minimized as part of the training loss for either the control or the treatment pipeline.

Building upon Thm. 1 (bounding the error of a single pipeline), the *PEHE* of the whole ALRITE is bounded as follows in the *within-sample* setting[4]:

Theorem 2. *Let $\mathcal{P}_0 = (\phi_{\mathcal{P}_0}, h^0_{\mathcal{P}_0}, h^1_{\mathcal{P}_0})$ and $\mathcal{P}_1 = (\phi_{\mathcal{P}_1}, h^0_{\mathcal{P}_1}, h^1_{\mathcal{P}_1})$ be two pipelines with same notations as above. Assume there exist functions $\nu^0_{\mathcal{P}_0}, \nu^1_{\mathcal{P}_0}, \nu^0_{\mathcal{P}_1}, \nu^1_{\mathcal{P}_1}$ with*

[4] Remind that the within-sample error is not trivial – contrarily to the training error in supervised learning – as counter-factuals are not observed.

Lipschitz constant L such that $\mu^0 = \nu^0_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0} = \nu^0_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$ and $\mu^1 = \nu^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0} = \nu^1_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$. Suppose that $h^0_{\mathcal{P}_0}, h^1_{\mathcal{P}_0}, h^0_{\mathcal{P}_1}, h^1_{\mathcal{P}_1}$ have Lipschitz constant $\hat{L}$. Given an observed dataset $\mathcal{D}$, let $\bar{\tau}$ be a vector of estimated causal effects defined by $\bar{\tau}_i = (1-t_i)(h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i) + t_i(y_i - h^0_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}(x_i))$ for $i \in [\![1, n]\!]$. Then, the within-sample PEHE defined by $\frac{1}{n}\sum_{i=1}^n \left(\bar{\tau}_i - \tau(x_i)\right)^2$ is upper bounded by M_2, with

$$M_2 = \frac{5}{n}\sum_{i=1}^{n}\Bigg(\textit{exemplarity}_{\mathcal{P}_{1-t_i}}(i)\left(h^{t_i}_{\mathcal{P}_{1-t_i}} \circ \phi_{\mathcal{P}_{1-t_i}}(x_i) - y_i\right)^2 + (L^2 + \hat{L}^2)\textit{insulation}_{\mathcal{P}_{t_i}}(i)^2\Bigg) + 5\kappa_Y$$

where $\kappa_Y = \frac{1}{n}\sum_{i=1}^n \left(1 + \textit{exemplarity}_{\mathcal{P}_{1-t_i}}(i)\right)\left(y_i - \mu^{t_i}(x_i)\right)^2$.

Proof. Let (x_i, t_i) be a sample in $\mathcal{D}$. Without loss of generality, assume $t_i = 1$. Let $(x_j, t_j = 0)$ be its mirror twin ($\phi(i) = j$). Denote by z_i and z_j their respective representations: $\phi_{\mathcal{P}_1}(x_i) = z_i, \phi_{\mathcal{P}_1}(x_j) = z_j$. Using Cauchy-Schwarz applied on $\mathbb{R}^5$, for any vector $u \in \mathbb{R}^5$, $(\sum u_i)^2 = <\mathbb{1}|u>^2 \leq ||\mathbb{1}||^2 \times ||u||^2 = 5\sum u_i^2$. It then comes:

$$\begin{aligned}
\left(\bar{\tau}_i - \tau(x_i)\right)^2 =& \left((y_i\text{-}h^0_{\mathcal{P}_1}(z_i)) - (\nu^1_{\mathcal{P}_1}\text{-}\nu^0_{\mathcal{P}_1})(z_i)\right)^2 \\
=& \big([y_i\text{-}\nu^1_{\mathcal{P}_1}(z_i)] + [\nu^0_{\mathcal{P}_1}(z_i)\text{-}\nu^0_{\mathcal{P}_1}(z_j)] + [\nu^0_{\mathcal{P}_1}(z_j)\text{-}y_j] \\
&+ [y_j\text{-}h^0_{\mathcal{P}_1}(z_j)] + [h^0_{\mathcal{P}_1}(z_j)\text{-}h^0_{\mathcal{P}_1}(z_i)]\big)^2 \\
\leq& 5(y_i\text{-}\nu^1_{\mathcal{P}_1}(z_i))^2 + 5(\nu^0_{\mathcal{P}_1}(z_i)\text{-}\nu^0_{\mathcal{P}_1}(z_j))^2 + 5(\nu^0_{\mathcal{P}_1}(z_j)\text{-}y_j)^2 \\
&+ 5(y_j\text{-}h^0_{\mathcal{P}_1}(z_j))^2 + 5(h^0_{\mathcal{P}_1}(z_j)\text{-}h^0_{\mathcal{P}_1}(z_i))^2 \\
\leq& 5(y_i\text{-}\mu^1(x_i))^2 + 5(y_j\text{-}\mu^0(y_j))^2 + 5(y_j\text{-}h^0_{\mathcal{P}_1}(z_j))^2 \\
&+ 5(\widehat{L}^2 + L^2)||z_j\text{-}z_i||^2
\end{aligned}$$

Averaging over $(x_i, t_i = 1)$ in $\mathcal{D}$, it comes:

$$\begin{aligned}
&\frac{1}{n}\sum_{t_i=1}\left(\bar{\tau}_i - \tau(x_i)\right)^2 \\
&\leq \frac{5}{n}\sum_{t_i=1}\left((L^2 + \hat{L}^2)\textit{insulation}_{\mathcal{P}_{t_i}}(i) + \left(\mu^{t_i}(x_i) - y_i\right)^2\right) \\
&+\frac{5}{n}\sum_{t_j=0}\left(\textit{exemplarity}_{\mathcal{P}_{1-t_i}}(j)\left[\left(h^{t_j}_{1-t_j} \circ \phi_{1-t_j}(x_j) - y_j\right)^2 + \left(\mu^{t_j}(x_j) - y_j\right)^2\right]\right)
\end{aligned}$$

Adding the control samples sum yields M_2. □

Note that Thm. 2 holds in the *within-sample* setting, as it requires knowledge of the treatment assignment T. As such, it does not generalize directly

to the *out-of-sample* setting, where the (unknown) t_i and y_i are respectively estimated using the propensity and the outcome models.

Furthermore, the upper-bound established in Thm. 2 can be directly related with terms $\mathcal{L}_{\mathcal{P}_0}$ and $\mathcal{L}_{\mathcal{P}_1}$, establishing the well-foundedness of the ALRITE loss, as follows:

Theorem 3. *Denote by p^1 and p^0 respectively the proportions of treated and control samples in the training set: $p^1 = \frac{1}{n}\sum t_i = 1 - p^0$. Then with adequate choice of hyper-parameters $(\alpha_{\mathcal{P}_0}, \alpha_{\mathcal{P}_1}, \beta_{\mathcal{P}_0}, \beta_{\mathcal{P}_1})$, the within-sample empirical PEHE is upper bounded by M_3 defined as*

$$\begin{aligned} M_3 =& 5\big(\mathcal{L}_{\mathcal{P}_0} + \mathcal{L}_{\mathcal{P}_1}\big) - 5\big(\gamma_{\mathcal{P}_0}\Omega(\mathcal{P}_0) + \gamma_{\mathcal{P}_1}\Omega(\mathcal{P}_1)\big) + 5\kappa_Y \\ & - \frac{5}{n}\sum_{i=1}^{n}\frac{1}{p^{t_i}}(h^{t_i}_{\mathcal{P}_{t_i}} \circ \phi_{\mathcal{P}_{t_i}}(x_i) - y_i)^2 - \frac{5}{n}\sum_{i=1}^{n}(h^{t_i}_{\mathcal{P}_{1-t_i}} \circ \phi_{\mathcal{P}_{1-t_i}}(x_i) - y_i)^2 \end{aligned}$$

Proof. Set the loss hyper-parameters values $\alpha_{\mathcal{P}_0}, \alpha_{\mathcal{P}_1}, \beta_{\mathcal{P}_0}, \beta_{\mathcal{P}_1}$ to

$$\big(\alpha_{\mathcal{P}_0}, \alpha_{\mathcal{P}_1}, \beta_{\mathcal{P}_0}, \beta_{\mathcal{P}_1}\big) = \big((1-p)(L^2+\hat{L}^2), p(L^2+\hat{L}^2), 1, 1\big)$$

entailing

$$\Big(\frac{\alpha_{\mathcal{P}_0}}{n_0}, \frac{\alpha_{\mathcal{P}_1}}{n_1}, \frac{\beta_{\mathcal{P}_0}}{n_1 + \beta_{\mathcal{P}_0} n_0}, \frac{\beta_{\mathcal{P}_1}}{n_0 + \beta_{\mathcal{P}_1} n_1}\Big) = \Big(\frac{L^2+\hat{L}^2}{n}, \frac{L^2+\hat{L}^2}{n}, \frac{1}{n}, \frac{1}{n}\Big)$$

Then, the total loss over both pipelines total loss writes

$$\begin{aligned} \mathcal{L}_{\mathcal{P}_0} + \mathcal{L}_{\mathcal{P}_1} =& \left(\frac{L^2+\hat{L}^2}{n}\sum_{t_i=0}\mathit{insulation}_{\mathcal{P}_0}(i)^2 + \frac{1}{n}\sum_{t_i=1}\mathit{exemplarity}_{\mathcal{P}_0}(i)(h^1_{\mathcal{P}_0}\circ\phi_{\mathcal{P}_0}(x_i)\text{-}y_i)^2\right) \\ & + \left(\frac{1}{n(1\text{-}p)}\sum_{t_i=0}(h^0_{\mathcal{P}_0}\circ\phi_{\mathcal{P}_0}(x_i)\text{-}y_i)^2 + \frac{1}{n}\sum_{t_i=1}(h^1_{\mathcal{P}_0}\circ\phi_{\mathcal{P}_0}(x_i)\text{-}y_i)^2 + \gamma_{\mathcal{P}_0}\Omega(\mathcal{P}_0)\right) \\ & + \left(\frac{L^2+\hat{L}^2}{n}\sum_{t_i=1}\mathit{insulation}_{\mathcal{P}_1}(i)^2 + \frac{1}{n}\sum_{t_i=0}\mathit{exemplarity}_{\mathcal{P}_1}(i)(h^0_{\mathcal{P}_1}\circ\phi_{\mathcal{P}_1}(x_i)\text{-}y_i)^2\right) \\ & + \left(\frac{1}{np}\sum_{t_i=1}(h^1_{\mathcal{P}_1}\circ\phi_{\mathcal{P}_1}(x_i)\text{-}y_i)^2 + \frac{1}{n}\sum_{t_i=0}(h^0_{\mathcal{P}_1}\circ\phi_{\mathcal{P}_1}(x_i)\text{-}y_i)^2 + \gamma_{\mathcal{P}_1}\Omega(\mathcal{P}_1)\right) \\ =& \frac{1}{n}\sum_{i=1}^{n}\left(\mathit{exemplarity}_{\mathcal{P}_{1-t_i}}(i)\big(h^{t_i}_{\mathcal{P}_{1-t_i}}\circ\phi_{\mathcal{P}_{1-t_i}}(x_i)\text{-}y_i\big)^2 + (L^2+\hat{L}^2)\mathit{insulation}_{\mathcal{P}_{t_i}}(i)^2\right) \\ & + \frac{1}{n}\sum_{i=1}^{n}(h^{t_i}_{\mathcal{P}_{1-t_i}}\circ\phi_{\mathcal{P}_{1-t_i}}(x_i)\text{-}y_i)^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{(h^{t_i}_{\mathcal{P}_{t_i}}\circ\phi_{\mathcal{P}_{t_i}}(x_i)\text{-}y_i)^2}{t_i p + (1\text{-}t_i)(1\text{-}p)} + \gamma_{\mathcal{P}_0}\Omega(\mathcal{P}_0) + \gamma_{\mathcal{P}_1}\Omega(\mathcal{P}_1) \\ =& \frac{1}{5}M_2 - \kappa_Y + \gamma_{\mathcal{P}_0}\Omega(\mathcal{P}_0) + \gamma_{\mathcal{P}_1}\Omega(\mathcal{P}_1) \end{aligned}$$

$$+\frac{1}{n}\sum_{i=1}^{n}(h^{t_i}_{\mathcal{P}_{1-t_i}}\circ\phi_{\mathcal{P}_{1-t_i}}(x_i)\text{-}y_i)^2+\frac{1}{n}\sum_{i=1}^{n}\frac{(h^{t_i}_{\mathcal{P}_{t_i}}\circ\phi_{\mathcal{P}_{t_i}}(x_i)\text{-}y_i)^2}{t_i p+(1\text{-}t_i)(1\text{-}p)}$$

implying an upper bound on the within-sample empirical *PEHE*:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\bar{\tau}_i-\tau(x_i)\right)^2\leq 5\left(\mathcal{L}_{\mathcal{P}_0}+\mathcal{L}_{\mathcal{P}_1}\right)-5\left(\gamma_{\mathcal{P}_0}\Omega(\mathcal{P}_0)+\gamma_{\mathcal{P}_1}\Omega(\mathcal{P}_1)\right)+5\kappa_Y$$
$$-\frac{5}{n}\sum_{i=1}^{n}\frac{1}{p^{t_i}}(h^{t_i}_{\mathcal{P}_{t_i}}\circ\phi_{\mathcal{P}_{t_i}}(x_i)-y_i)^2-\frac{5}{n}\sum_{i=1}^{n}(h^{t_i}_{\mathcal{P}_{1-t_i}}\circ\phi_{\mathcal{P}_{1-t_i}}(x_i)-y_i)^2$$

□

While Thm. 3 shows the well-foundedness of the terms in the ALRITE loss, this does not imply that the choice of their weight is optimal. It also requires knowledge of both Lipschitz constants L and $\hat{L}$. While $\hat{L}$ may be measured or even bounded using appropriate constraints[5], L only depends on the problem.

IV.5.2 . Positioning w.r.t. Shalit et al. (2017)

Thms. 1 to 3, establishing upper bounds on the empirical *PEHE*, are discussed and compared with the main result of Shalit et al. (2017), reminded below:

Theorem (Shalit et al. (2017))**.** *Let (ϕ, h^0, h^1) be a pipeline such that $\phi : \mathcal{X} \to \mathcal{Z}$ is invertible. Define the point-wise loss function $\ell_{h,\phi}$ and expected factual loss ϵ_t conditionally to treatment assignment $T = t$, $t \in \{0,1\}$ by*

$$\begin{cases}\ell_{h,\phi} : (x,t)\in\mathcal{X}\times\{0,1\} & \mapsto \mathbb{E}[\left(Y^t-h\circ\phi(x)\right)^2|X=x]\\ \epsilon^t : h\in(\mathcal{Z}\to\mathcal{Y}) & \mapsto \mathbb{E}[\ell_{h,\phi}(X,T)|T=t]\end{cases} \tag{IV.7}$$

Denote by $\sigma^2_{Y^t}=\mathbb{E}\left[(Y^t-\mathbb{E}[Y^t|X])^2|T=t\right]\mathbb{P}(T=t)$ the expected variance of Y^t, $t\in\{0,1\}$. Let G be a family of functions $\mathcal{Z}\to\mathcal{Y}$. Assume that there exists a constant $B_\phi>0$ such that $z\in\mathcal{Z}\mapsto\frac{1}{B_\phi}\ell_{h^t,\phi}(\phi^{-1}(z),t)\in G$, $t\in\{0,1\}$. Denote the integral probability metric between the control and treated latent distributions induced by ϕ as

$$IPM=\sup_{g\in G}\left|\mathbb{E}\left[\left(\mathbb{P}\left(\phi(X)|T\text{=}1\right)-\mathbb{P}\left(\phi(X)|T\text{=}0\right)\right)g\left(\phi(X)\right)\right]\right|$$

Then, the PEHE is upper bounded:

$$PEHE(\phi,h^0,h^1)\leq 2\left(\epsilon^0(h^0)+\epsilon^1(h^1)+B_\phi\times IPM-2\min(\sigma^2_{Y^0},\sigma^2_{Y^1})\right)$$

[5] In the case where $h^0_{\mathcal{P}_0}, h^1_{\mathcal{P}_0}, h^0_{\mathcal{P}_1}, h^1_{\mathcal{P}_1}$ are linear, their Lipschitz constants can be derived straightforwardly. In the general case of neural networks, additional care is required (see e.g., Virmaux and Scaman (2018); Gouk et al. (2021)).

The difference between Shalit et al. (2017)'s results and ours is twofold.

Firstly, Shalit et al. (2017) assumes embedding ϕ to be **invertible**, while the presented result only assumes the existence of two functions $\nu^0, \nu^1 : \mathcal{Z} \to \mathcal{Y}$ s.t. $(\mu^0, \mu^1) = (\nu^0 \circ \phi, \nu^1 \circ \phi)$. The impact of this difference, and of relaxing the invertibility assumption on ϕ, is to be able to fully take advantage of the change of representation ϕ. As widely acknowledged in Machine Learning, data tend to live in a manifold of the description space (Cayton, 2008; Bengio et al., 2013); the ability to achieve feature selection and dimensionality reduction by means of ϕ, to the extent compatible with the prediction goal, thus is most useful.

Secondly, the bound in Shalit et al. (2017) involves **integral probability metrics**, while the proposed result considers two specifics of the sample distributions in the latent space: the insulation and the exemplarity. On the one hand, integral probability metrics are defined in terms of distribution (the large sample limit case); on the other hand, the insulation and exemplarity typically are empirical quantities better suited to the situation faced by a practitioner. Note that Yao et al. (2018) also advocates against the use of distributional distances, for the discrepancy they measure is at best a proxy objective.

The constants in the bound depend on the considered embedding ϕ in both cases. In (Shalit et al., 2017), the bound involves a constant B_ϕ. When G is chosen as the space of 1-Lipschitz functions, the constraint involving B_ϕ rewrites as

$$\forall (t, z, z') \in \{0,1\} \times \mathcal{Z} \times \mathcal{Z}, ||\ell_{h^t,\phi}(\phi^{-1}(z), t) - \ell_{h^t,\phi}(\phi^{-1}(z'), t)|| \leq B_\phi ||z - z'||$$

and consequently $\forall t \in \{0,1\}$, $z \mapsto \ell_{h^t,\phi}(\phi^{-1}(z), t)$ should be B_ϕ-Lipschitzian. The ground-truth functions μ^0, μ^1 being unknown however, one cannot check if this assumption holds.

IV.5.3 . Discussion of the assumptions

Thms. 1 to 3 spark three questions. Firstly, when do their assumptions hold? Secondly, how can the practitioner assess whether they hold? Thirdly and most importantly, when are outcome functions h^0, h^1 learned on top of such a latent space accurate predictors of the true outcome functions?

In the following and for the sake of simplicity, the discussion focuses on the control outcome estimation. The case of the treatment outcome estimation follows, replacing the superscripts $\square^0$ by $\square^1$.

To structure the discussion, a synoptic diagram of the evoked results is proposed in Fig. IV.7.

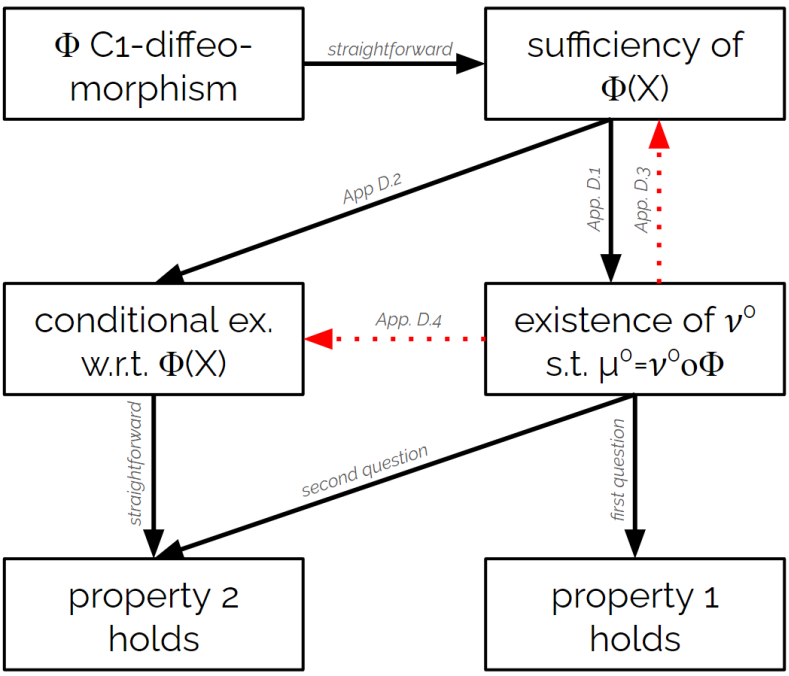


Figure IV.7: Synoptic diagram of the discussion. Black arrows represent implication. Red arrows represent the absence of implication.

IV.5.3.1 . Existence and lipschitzianity of ν^0, ν^1

The existence of L-Lipschitz outcome functions ν^0, ν^1 in Thms. 1 to 3 might appear as a leap of faith. Their existence can however be guaranteed if the true μ^0, μ^1 are Lipschitz:

Lemma 4. *Assume that $\mathcal{X}$ is compact*[6]*, ϕ is a C^1-diffeormorphism and μ^0 is L-Lipschitz for a given parameter $L \in \mathbb{R}$. Then, ν^0 verifying hypotheses of Thm. 1 exists, and $\phi(X)$ is a sufficient statistic for Y^0.*

Proof. Let ϕ be such a function. Then $\phi(\mathcal{X})$ is closed as the direct image of a closed set, and the Jacobian of $\phi^{-1} : \phi(\mathcal{X}) \to \mathcal{X}$ is upper bounded as a continuous function over a compact set. Denote by $L_{\phi^{-1}}$ this upper bound. Set $\nu^0 = \mu^0 \circ \phi^{-1}$. ν^0 is $L \times L_{\phi^{-1}}$-Lipschitz, and $\mu^0 = \nu^0 \circ \phi$. □

An appropriate search space for learning C^1-diffeormorphism ϕ is that of (bi-Lipschitz) Normalizing Flows (Kobyzev et al., 2020; Verine et al., 2021). Most generally, quite a few authors (Shalit et al., 2017; Du et al., 2021) base their theoretical analysis on the invertibility property, though the associated algorithms do not seek to enforce this property. Note that invertible ϕ have other merits, e.g., the fact they define balancing scores (Zhang et al., 2020).

[6]This only requires $\mathcal{X}$ to be bounded, since it is of finite dimension.

Invertibility of representation ϕ is however no panacea. Lemma 4 sets an upper bound on the Lipschitz constant of functions ν^0, ν^1. However, a relevant representation network is typically expected to bring samples far away in the covariates space close in the latent space. Let (x, x') be a couple of such samples: $||x - x'||$ is large while $||\phi(x) - \phi(x')||$ is small. It follows that $\frac{||x-x'||}{||\phi(x)-\phi(x')||} = \frac{||\phi^{-1}(\phi(x))-\phi^{-1}(\phi(x'))||}{||\phi(x)-\phi(x')||}$ is large, making the Lipschitz constant $L_{\phi^{-1}}$ of ϕ^{-1} large, too, and thus weakening the upper-bound strength.

Moreover, as said a non-invertible ϕ can also contribute positively to the model performance. Typically, if the potential outcomes only depend on a subset of the covariates, a (non-injective) projection ϕ on these covariates can greatly facilitate the learning task and support the accuracy of the models. In any case, $\hat{L}$ can be experimentally assessed; further work will investigate how to constrain it (Virmaux and Scaman, 2018; Gouk et al., 2021).

In both (Shalit et al., 2017) and our setting, the Lipschitz constant L of the ground truth functions ν^0, ν^1 remains however inaccessible.

IV.5.3.2 . Properties of $\hat{\nu}^0, \hat{\nu}^1$

Assume now that the representation network ϕ is fixed, and that it supports a function ν^0 verifying $\mu^0 = \nu^0 \circ \phi$. Let $\hat{\nu}^0 : \mathcal{Z} \to \mathbb{R}$ be a candidate hypothesis function. In accordance with the notation introduced in Eq. IV.7, denote by $\epsilon^0(\hat{\nu}^0)$ the expected mean squared prediction error conditionally to $T = 0$ of the function $\hat{\nu}^0$:

$$\begin{aligned}\epsilon^0(\hat{\nu}^0) &= \mathbb{E}_{X,Y|T=0}\left[\left(\hat{\nu}^0 \circ \phi(X) - Y^0\right)^2 | T = 0\right] \\ &= \mathbb{E}_{Z=\phi(X),Y|T=0}\left[\left(\hat{\nu}^0(Z) - Y\right)^2 | T = 0\right]\end{aligned}$$

ϵ^0 admits a minimum in $\nu^{0*} : z \in \mathcal{Z} \mapsto \mathbb{E}[Y^0 | T = 0, \phi(X) = z]$. As an immediate consequence, $\nu^{0*} \circ \phi : x \in \mathcal{X} \mapsto \mathbb{E}[Y^0 | T = 0, \phi(X) = \phi(x)]$. However, our primary goal consists in finding a function $\hat{\nu}^0$ such that $\hat{\nu}^0 \circ \phi$ approximates as accurately as possible μ^0. This raises the question of when does the equality of $\mathbb{E}[Y^0 | T = 0, \phi(X) = \phi(\cdot)]$ and $\mu^0(\cdot)$ hold.

This issue is split into two questions. Assuming the existence of ν^0,

1. **first question:** does *Property 1* $\left[\mu^0(\cdot)\right.$ equal to $\left.\mathbb{E}[Y^0 | \phi(X) = \phi(\cdot)]\right]$ hold?
2. **second question:** does *Property 2* $\left[\mathbb{E}[Y^0 | \phi(X) = \phi(\cdot)]\right.$ equal to $\left.\mathbb{E}[Y^0 | T = 0, \phi(X) = \phi(\cdot)]\right]$ also hold?

The **first question** is answered positively. Let us assume that $\nu^0 : \mathcal{Z} \to \mathbb{R}$ is such that $\nu^0 \circ \phi = \mu^0$. Let then x be an element of $\mathcal{X}$. The formula of total

probability writes

$$\begin{aligned}
\mathbb{E}[Y^0|\phi(X)=\phi(x)] &= \int_{\phi^{-1}(\{\phi(x)\})} \mathbb{E}[Y^0|\phi(X)=\phi(x), X=u]\mathbb{P}(X=u|\phi(X)=\phi(x))du \\
&= \int_{\phi^{-1}(\{\phi(x)\})} \mathbb{E}[Y^0|X=u]\mathbb{P}(X=u|\phi(X)=\phi(x))du \\
&= \int_{\phi^{-1}(\{\phi(x)\})} \mu^0(u)\mathbb{P}(X=u|\phi(X)=\phi(x))du \\
&= \int_{\phi^{-1}(\{\phi(x)\})} \nu^0\circ\phi(u)\mathbb{P}(X=u|\phi(X)=\phi(x))du \\
&= \int_{\phi^{-1}(\{\phi(x)\})} \nu^0\circ\phi(x)\mathbb{P}(X=u|\phi(X)=\phi(x))du \\
&= \nu^0\circ\phi(x)\int_{\phi^{-1}(\{\phi(x)\})} \mathbb{P}(X=u|\phi(X)=\phi(x))du \\
&= \mu^0(x) \\
&= \mathbb{E}[Y^0|X=x]
\end{aligned}$$

The average value of the control outcome Y^0 conditionally on X equals the average value conditionally on $\phi(X)$.

The **second question** (does the existence of ν^0 implies that $\mathbb{E}[Y^0|\phi(X)=\phi(\cdot)]$ equal $\mathbb{E}[Y^0|T=0,\phi(X)=\phi(\cdot)]$) is also answered positively. Denote by U the set of events $(\phi(X),T)^{-1}(\{(\phi(x),0)\})$. Then,

$$\begin{aligned}
&\mathbb{E}[Y^0|\phi(X)=\phi(x),T=0] \\
&\quad= \int_{u\in U} \mathbb{E}[Y^0|\phi(X)=\phi(x),T=0,X=u]\mathbb{P}(X=u|\phi(X)=\phi(x),T=0)du \\
&\quad= \int_{u\in U} \mathbb{E}[Y^0|T=0,X=u]\mathbb{P}(X=u|\phi(X)=\phi(x),T=0)du \\
&\quad= \int_{u\in U} \mathbb{E}[Y^0|X=u]\mathbb{P}(X=u|\phi(X)=\phi(x),T=0)du \\
&\quad= \int_{u\in U} \mathbb{E}[Y^0|\phi(X)=\phi(x)]\mathbb{P}(X=u|\phi(X)=\phi(x),T=0)du \qquad \textit{(first question)} \\
&\quad= \mathbb{E}[Y^0|\phi(X)=\phi(x)]\int_{u\in U} \mathbb{P}(X=u|\phi(X)=\phi(x),T=0)du \\
&\quad= \mathbb{E}[Y^0|\phi(X)=\phi(x)]
\end{aligned}$$

Therefore, if ν^0 s.t. $\nu^0\circ\phi=\mu^0$ does exist, **Property 2** ($\mathbb{E}[Y^0|\phi(X)=\phi(\cdot)]=\mathbb{E}[Y^0|\phi(X)=\phi(\cdot),T=0]$) holds. Overall, the minimizer ν^{0*} of ϵ^0 is such that $\nu^{0*}\circ\phi=\mu^0$, establishing the relevance of the approach.

Besides, note (further discussion in Appendix E) that:

1. the sufficiency of $\phi(X)$ with respect to Y^0 alone implies the existence of $\nu^0:\mathcal{Z}\mapsto\mathbb{R}$ such that $\mu^0=\nu^0\circ\phi$ (Appendix E.1);

2. the sufficiency of $\phi(X)$ implies conditional exchangeability w.r.t. $\phi(X)$ (Appendix E.2);
3. conditional exchangeability w.r.t. $\phi(X)$ implies that *Property 2* holds;

but

1. existence of ν^0 s.t. $\mu^0 = \nu^0 \circ \phi$ does not imply the sufficiency of $\phi(X)$ (Appendix E.3);
2. existence of ν^0 s.t. $\mu^0 = \nu^0 \circ \phi$ does not even imply the conditional exchangeability w.r.t. $\phi(X)$ (Appendix E.4).

Remark: The sufficiency of $\phi(X)$ with respect to Y^0 guarantees both the existence of ν^0 and the relevance of the approach: ϵ^0 admits a minimum in $\nu^{0*} : z \in \mathcal{Z} \mapsto \mathbb{E}[Y^0 | \phi(X) = z]$, and $\nu^{0*} \circ \phi = \mu^0$. Sufficiency of $\phi(X)$ with respect to Y^0 is indeed a very favorable case.

However, from the practitioner's viewpoint, it is impossible to prove that a given statistic is sufficient based only on observational data. As a matter of course if ϕ is injective then $Y^0 \perp\!\!\!\perp X | \phi(X)$, but this result relies on mapping ϕ, and not on the observational data itself. Even in the large sample limit and using adequate conditional independence statistical tests, one may prove at most independence of Y^0 and X conditionally to $(\phi(X) = z, T = 0)$, but not conditionally to $(\phi(X) = z)$ alone. This concern echoes the ones raised by the assumption of conditional exchangeability: assuming sufficiency based on observational data is a similar leap of faith.

IV.5.4 . Limitations

The well-foundedness of the ALRITE loss has been established, relating the loss terms with the upper bound on the error of the causal estimate. This subsection thus discusses the limitations and robustness of the proposed approach with respect to: i) its asymptotical behavior; its robustness to positivity violation; and iii) potential learning instability.

IV.5.4.1 . Asymptotical behavior

In the large sample limit, the insulation of any sample goes to 0 in probability.[7] The approach's asymptotical behavior remains an open question at the moment. While the regularization term in (Shalit et al., 2017) is known to converge toward the integral formulation of statistical distances (although the rate of convergence may be extremely slow in low-density regions of high-

[7] Let us assume for simplicity that $\mathcal{X} \subset \mathbb{R}^d$. Let $(x_i, t_i = 0, y_i)$ be a control sample from the training dataset $\mathcal{D}$, with $\epsilon > 0$. Function $\phi_{\mathcal{P}_0}$, being implemented with a finite-weights neural network, is continuous. As such, the inverse image of the latent space open ball $B(\phi_{\mathcal{P}_0}(x_i), \epsilon)$ centered on $\phi_{\mathcal{P}_0}(x_i)$ with radius ϵ is also an open. $(x_i, t_i = 0, y_i)$ has been sampled from $\mathbb{P}_{X,T,Y}$ and belongs to $\phi^{-1}_{\mathcal{P}_0}(B(\phi_{\mathcal{P}_0}(x_i), \epsilon))$ so there also exists an open $A \subset \phi^{-1}_{\mathcal{P}_0}(B(\phi_{\mathcal{P}_0}(x_i), \epsilon))$ of $\mathbb{R}^d$ such that $\mathbb{P}(T = 0, X \in B) > 0$. Since positivity

dimensional settings), the asymptotic convergence of the term based on insulation remains to be studied.

IV.5.4.2 . Robustness to positivity violation

The robustness w.r.t. the lack of overlap of the control and treatment distribution (positivity violation) is all the more important than positivity is both less likely to hold and harder to assess as the dimensionality of the covariate space increases.

Unfortunately, as a synthetic example below shows, ALRITE is sensitive to positivity violations. Note that this issue is not specific to our approach: the same reasoning holds for models enforcing balance in the latent space through distributional distance (Shalit et al., 2017; Du et al., 2021), or counter-factual variance (Zhang et al., 2020) minimization. Moreover, neither latent space disentanglement nor double-robustness does address this issue.

The illustrative synthetic example involves a 2-dimensional covariate space (Fig. IV.8a) with two clusters. The positivity assumption is challenged since the rightmost samples of the bottom cluster are overwhelmingly treated, and the leftmost samples of the top one are mainly control.

As shown on Fig. IV.8, taking $\phi(x = (x_1, x_2)) = x_1$ (projection on the first axis) yields low insulation values (Fig. IV.8, right) compared to $\phi = Id$ (Fig. IV.8, left). For a high value of the penalization weight α, the projection on the first axis is likely to correspond to a local minimum of the training loss.

As ALRITE might trade factual prediction accuracy to improve insulation, this might harm the model's ability to learn the factual and the counter-factual distributions.

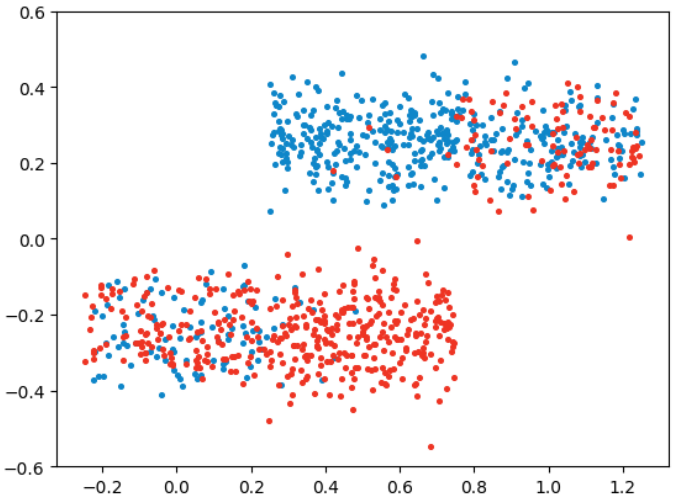
IV.5.4.3 . Potential learning instability

As discussed in Section IV.5, a low value of the training loss of ALRITE hints at a low within-sample *PEHE*. The potential weakness is that the greedy optimization of the exemplarity term can distort the model (see Appendix F for an (admittedly pathological) illustration).

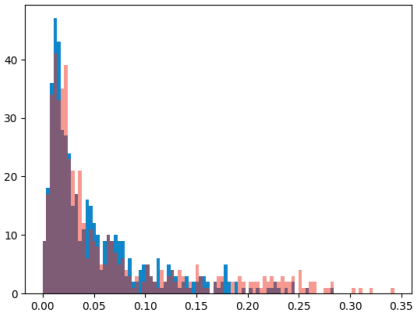
---

holds, $\mathbb{P}(T = 1, X \in A) > 0$. Finally,

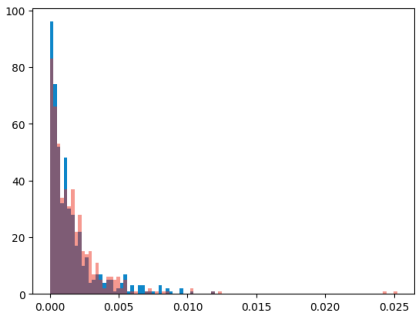
$$\begin{aligned}
\mathbb{P}\left(\textit{insulation}_{\phi_{\mathcal{P}_0}}(i) > \epsilon\right) &= \mathbb{P}\left(\forall j \in [\![1, n]\!],\ t_j = 1 \implies \phi_{\mathcal{P}_0}(x_j) \notin B(\phi_{\mathcal{P}_0}(x_i), \epsilon)\right) \\
&\leq \mathbb{P}\left(\forall i \in [\![1, n]\!],\ t_j = 1 \implies x_j \notin \phi^{-1}_{\mathcal{P}_0}(B(\phi_{\mathcal{P}_0}(x_i), \epsilon))\right) \\
&\leq \mathbb{P}\left(\forall j \in [\![1, n]\!],\ t_j = 1 \implies x_j \notin A\right) \\
&\leq \left(1 - \mathbb{P}(T = 1, X \in A)\right)^n \\
&\xrightarrow{n \to +\infty} 0
\end{aligned}$$



(a) Treated and control samples in covariate space: the distributions include low overlap regions.



(b) insulation values, ϕ is the identity over $\mathbb{R}^2$



(c) insulation values, ϕ is the projection over the x-axis.

Figure IV.8: Distribution of the control and treated samples. Top: covariates space. Bottom: histograms of insulation values. Note the difference in the ranges.

The rationale behind over-weighting the prediction error of samples (x, t, y) with high exemplarity is that a high associated prediction error $\|h^t(x) - y\|^2$ entails a high risk of error for the counter-factual estimation of their neighbors $(x', t' = 1 - t, y')$.

The risk of fragility comes from the fact that the exemplarity itself depends on mapping ϕ, which is optimized simultaneously with h^1 and h^0. The overall optimization might thus select ϕ such that samples with little factual error have high exemplarity – instead of selecting h^t such that it makes low errors on samples with high exemplarity.

As the gradient does not flow back through the mirror twin operator during the back-propagation phase, there should be no incentive for the ϕ network to optimize in such a pathological way. However, this phenomenon is prone to create instabilities during the training, all the more so as exemplarity is defined in terms of 1-nearest neighbor, and is thus sensitive to infinitesimal modifications of the network parameters. Further research will investigate how to alleviate this drawback, e.g., by considering a smoothed estimate of exemplarity.

IV.6 . Partial conclusion

In this chapter is motivated, built, described and analyzed ALRITE, as a new architecture suited to *CATE* estimation. The said model relies on asymmetrical regularization, providing finer counter-factual modeling for both the control and the treated distribution.

The merits of ALRITE will be validated through experiments and the comparison with baseline models in Chapter V. We shall discuss in detail the selection of its hyper-parameters in Chapter VI.

V - Experimental validation

This chapter introduces the benchmark datasets for *CATE* models evaluation (Section V.1). The experimental setting is detailed, delving into the concrete implementation of ALRITE (Section V.2). The results are finally reported and compared with baseline models, validating the approach (Section V.3).

V.1 . Benchmarks

Two datasets are considered: *IHDP* which has been used to validate most *CATE* estimation approaches, and *Jobs*.

V.1.1 . *IHDP*

V.1.1.1 . Description

The *IHDP* dataset, introduced by Hill (2011), is based on a real-life randomized experiment dataset, the Infant Health and Development Program (Brooks-Gunn et al., 1992), aimed at measuring the impact of quality childcare and home visits on the health and development of preterm, low birth weight children. The treatment consists of receiving this monitoring. Success (outcome) is quantified through the results obtained from cognitive tests of the children, taken at 12, 24, and 36 months.

The observational data include 25 covariate features describing the infant (e.g., birth weight, birth gender, neonatal health status, pregnancy birth week) and their mother (education status, ethnic group, age, engagement in prenatal care, risky behavior during pregnancy, etc.). 6 of these features are continuous, 19 of them are binary.

Hill (2011) has processed the survey results to obtain a dataset suitable for causal inference research. Notably, although treatment assignment is randomized in the collected data, treatment imbalance is enforced by **removing a subpopulation** from the dataset: treated group children with nonwhite mothers. The authors justify this choice arguing that said feature maximizes the imbalance between the induced control and treated groups. The resulting dataset contains 747 individuals, 139 of whom are treated, all described by the original 25 continuous and binary features.

The major difference between the original survey results and Hill (2011)'s *IHDP* dataset lies in the replacement of the initial quantity of interest (cognitive test scores) with **simulated outcomes**. Contrarily to real-life data, simulated outcomes make it possible to access the counter-factual quantities, and thus assess the performance of causal estimation models. Knowledge of the data generation process ensures that **conditional exchangeability holds**. The

selected simulation method consists in defining two response surfaces $\mu^t : x \in \mathcal{X} \mapsto \mathbb{E}[Y^t|X = x], t \in \{0, 1\}$, to which Gaussian noise is added.

V.1.1.2 . Outcomes simulation

Let $Z \sim Categorical\big((a, p_a), (b, p_b)\big)$ denotes a distribution such that $\mathbb{P}(Z = a) = p_a$, $\mathbb{P}(Z = b) = p_b$.

Then, an *IHDP* dataset is generated as:

$$\beta_i \sim Categorical\big((0, 6/10), (.1, 1/10), (.2, 1/10), (.3, 1/10), (.4, 1/10)\big), \forall i \in [\![1, 25]\!]$$
$$Y^0 \sim \exp\{(X + 0.5)^\dagger \beta\} + \mathcal{N}(0, 1)$$
$$Y^1 \sim X^\dagger \beta + \omega + \mathcal{N}(0, 1), \text{ with } \omega \text{ s.t. } ATT = 4$$

Drawing multiple replicas of β, Hill (2011) generates multiple datasets. Overall, "*IHDP*" comes in two modes noted *IHDP*-100 and *IHDP*-1000, respectively including 100 and 1,000 generated datasets. Each dataset is split into a training (90% of samples) and a testing (10%) subset, the split being fixed to support a fair comparison among the algorithms. Researchers are then free to hold out part of the data for validation purposes, the common rule consisting in a $70\% - 30\%$ split.

Following the state-of-the-art (Shalit et al., 2017; Du et al., 2021; Zhou et al., 2021), the results presented in the following are averaged over the considered datasets. We restrict ourselves to *IHDP*-100 for the sake of computational time.

V.1.1.3 . Performance indicators

On the *IHDP* benchmark, the primary performance indicator is the ***PEHE*** that is the empirical error of the final estimate $\hat{\tau}$, defined as the mixture of two estimates $\hat{\tau}_{\mathcal{P}_0}, \hat{\tau}_{\mathcal{P}_1}$ (Eq. IV.2).

Two *PEHE*s are reported, referred to as *within-sample* and *out-of-sample*. The **within-sample** *PEHE* is computed on the training samples, where the factual outcome is known. As said, the within-sample error still is challenging in causal modeling (in contrast with mainstream supervised learning) as the counter-factual outcome is unknown.

The **out-of-sample** *PEHE* is computed on the test samples, where factual and counter-factual outcomes are unknown. As could be expected, out-of-sample estimation remains harder than within-sample one: representation network $\phi_{\mathcal{P}_0}, \phi_{\mathcal{P}_1}$ and outcome functions $h^0_{\mathcal{P}_0}, h^1_{\mathcal{P}_0}, h^1_{\mathcal{P}_1}, h^0_{\mathcal{P}_1}$ have been specifically fitted to the training samples. Finally, some models may even explicitly rely on factual outcomes in the inference phase when available: $(2t_i - 1) \times \big(y_i - h^{1-t_i}(x_i)\big)$ is a relevant way to estimate $CATE(x_i)$ and several methods (Künzel et al., 2019) rely on it.

A secondary performance indicator on the *IHDP* benchmark is the **mean absolute error on the average treatment effect estimation** (Section II.5.1),

with empirical estimate

$$\epsilon_{ATE} = |\frac{1}{n}\sum_{i=1}^{n}\hat{\tau}(x_i) - \tau(x_i)|$$

While writing ϵ_{ATE} as a function of $\hat{\tau}$ might suggest a high correlation with the *PEHE*, practical experiments show otherwise. Systemic estimation bias affects *ATE* much more than *CATE*. Specifically, methods focusing on *ATE* estimation (in particular, doubly robust methods) usually demonstrate better performances w.r.t. *ATE* (Fig. V.1). Overall, while *PEHE* favors low-variance models, ϵ_{ATE} favors low-bias ones, no matter their variance.

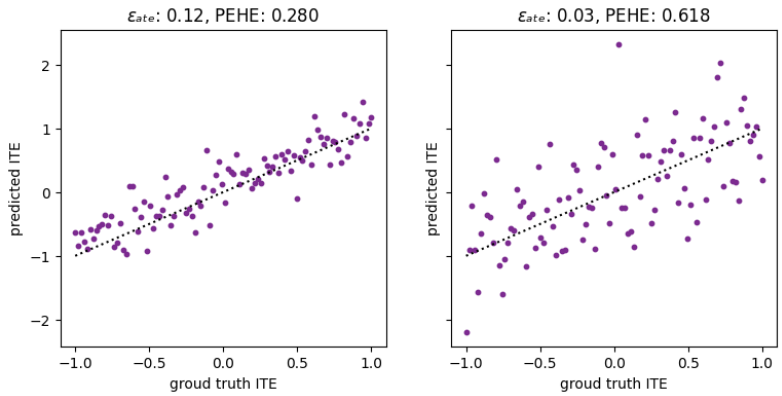


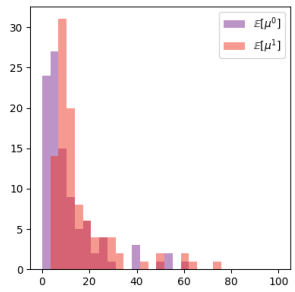
Figure V.1: Impact of the performance indicators. The left model is better in terms of *PEHE*, worse in terms of ϵ_{ATE}, compared to the right model.

V.1.1.4 . Discussion

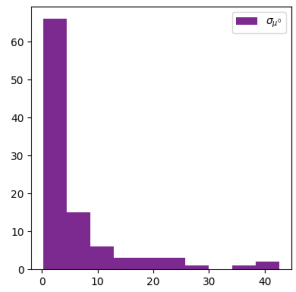
While *IHDP* is *de facto* the baseline dataset for benchmarking causal inference models, it raises some criticisms (Curth et al., 2021b).

As said, *IHDP* is a collection of datasets (referred to as problem instance, or simply instance when no confusion is to fear). However, the ranges of outcomes and the causal effects are **not commensurate among instances**. As shown in Fig. V.2a, the average values of μ^0, μ^1 vary greatly from one instance of *IHDP*-100 to another. Note also that the standard deviation of μ^0 takes high values in some instances (Fig. V.2b) while μ^1 varies much less in all instances (Fig. V.2c). Therefore, τ varies a lot from one instance to another, both in terms of average (Fig. V.2d) and variance (Fig. V.2e).

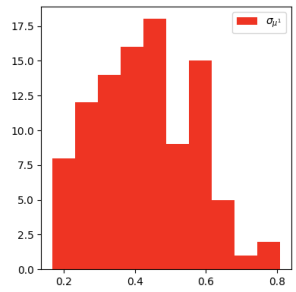
On a given instance, a high standard deviation σ_τ of the causal effects makes *CATE* estimation more difficult. Indeed, when σ_τ goes to 0, the causal effect is uniform and the *CATE* estimation problem boils down to the (much simpler) *ATE* estimation problem. Quite the contrary, for high values of σ_τ, the



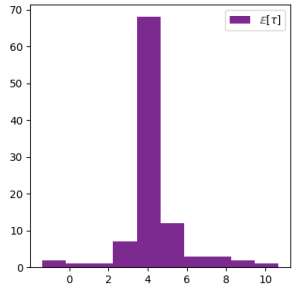
(a) Averages of μ^0 and μ^1



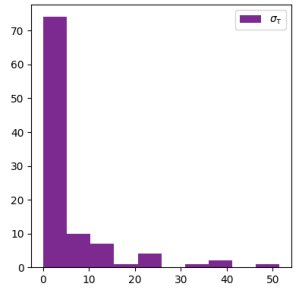
(b) $\sigma_{\mu^0(X)}$, standard deviations of μ^0



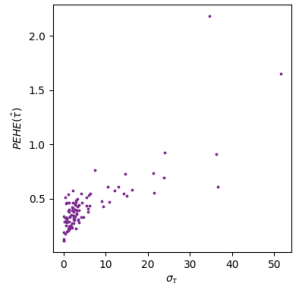
(c) $\sigma_{\mu^1(X)}$, standard deviations of μ^1



(d) Average treatment effects.



(e) Standard deviations of the treatment effects.



(f) $PEHE(\hat{\tau})$ vs σ_τ

Figure V.2: Histograms and scatter plot of different measures, over the 100 instances of *IHDP*-100.

large variance makes large errors more likely. This claim is visually confirmed in Fig. V.2f, plotting the standard deviations of τ vs the *PEHE* of $\hat{\tau}$ the best ALRITE estimate, for all instances in *IHDP*-100. Interestingly, instances with large σ_τ account for a large fraction of the overall *PEHE* error. Accordingly, *IHDP* performance indicators are strongly biased depending on the algorithm behavior on the few toughest instances – they do not actually reflect the average behavior of the algorithm.

V.1.2 . *Jobs*

V.1.2.1 . Description

Jobs is initially introduced by LaLonde (1986) to illustrate the limitations of mainstream econometrics methods in treatment effect estimation. It concatenates the data of a randomized study and a survey (Fig. V.3). Part of the data comes from the so-called *National Supported Work Demonstration* (*NSWD*) randomized experiment. In a study conducted in the mid-70s by the Manpower Demonstration Research Corporation, randomly selected women from the *Aid to Families with Dependent Children* social security program, high-school dropouts, and ex-criminal offenders have been offered guaranteed jobs for a duration comprised between 9 and 18 months. After this period, the earnings

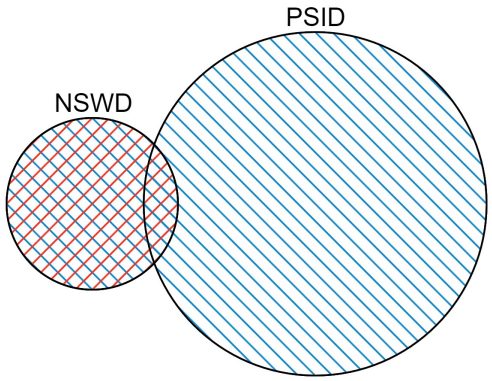


Figure V.3: Schematic structure of the *Jobs* dataset, with treated and control samples. Note the limited overlap (more in Fig. V.4b).

of participants have been collected every year from 1975 to 1979. The other part of the data comes from the *Panel Study of Income Dynamics* (*PSID*), a purely observational survey.

The outcome Y considered in (Dehejia and Wahba, 2002) and further work is the employment status of individuals. Overall, the outcome is binary; 8 binary and continuous covariates are considered (e.g., age, education, previous earnings).

Like *IHDP*, *Jobs* is actually a collection of datasets, made of 10 fixed train/test splits of the same original data. The randomized subgroup contains 297 treated and 425 control individuals, while the comparison group contains only control ones.

V.1.2.2 . Performance indicators

Since *Jobs* is a real-life dataset and no counter-factual information is available, the primary performance indicator is the **policy risk** (Section II.5.1) associated with the binary policy built from the causal effects estimate, with $\pi_{\hat{\tau}}(x) = 1$ iff $\hat{\tau}(x) > 0$.

In principle, treatment assignment is uniform on the *NSWD* subset, enforcing conditional exchangeability ($Y^t \perp\!\!\!\perp T|X, \forall t \in \{0,1\}$) as well as exchangeability ($Y^t \perp\!\!\!\perp T, \forall t \in \{0,1\}$). Under this assumption, the policy risk R_{pol} and observational policy risk oR_{pol} (Section II.5.1) coincide.

$$\begin{aligned}\mathbb{E}[Y^t|\pi(X)=t] =& \mathbb{E}[Y^t|\pi(X)=t, T=t] && (Y^t \perp\!\!\!\perp T|\pi(X))\\ =& \mathbb{E}[Y^T|\pi(X)=t, T=t]\\ =& \mathbb{E}[Y|\pi(X)=t, T=t] && (\textit{SUTVA})\end{aligned}$$

implying

$$\begin{aligned}R_{pol}(\pi) =& 1 - \mathbb{P}(\pi(X)=1)\mathbb{E}[Y^1|\pi(X)=1]\\ & - \mathbb{P}(\pi(X)=0)\mathbb{E}[Y^0|\pi(X)=0]\end{aligned}$$

$$\begin{aligned}
&= 1 - \mathbb{P}(\pi(X) = 1)\mathbb{E}[Y|\pi(X) = 1, T = 1] \\
&\quad - \mathbb{P}(\pi(X) = 0)\mathbb{E}[Y|\pi(X) = 0, T = 0] \\
&= oR_{pol}(\pi)
\end{aligned} \tag{V.1}$$

and the final expression only involves the observational distribution. For the sake of readability, the notations in the following do not explicit that the probabilities, expectations and independence are taken with respect to the distribution of *NSWD*.

A secondary performance indicator is the error on the **Average Treatment effect on the Treated** $\epsilon_{ATT}(\hat{\tau}) = |\mathbb{E}[\hat{\tau}(X) - \tau(X)|T = 1]|$. Here again, the randomness in the treatment assignment makes it possible to obtain a reliable *ATT* estimate by computing averages over the *NSWD* data. By taking all expectations and probabilities on the *NSWD* distribution, it comes:

$$\begin{aligned}
ATT &= \mathbb{E}[Y^1 - Y^0|T = 1] \\
&= \mathbb{E}[Y^1|T = 1] - \mathbb{E}[Y^0|T = 0] && (Y^0 \perp\!\!\!\perp T) \\
&= \mathbb{E}[Y^T|T = 1] - \mathbb{E}[Y^T|T = 0] \\
&= \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0] && (SUTVA)
\end{aligned}$$

and the final expression only involves the observational distribution.

Remark. Although the *within-sample* version of the policy risk has been widely used in the literature to assess the quality of causal inference models, it only relies on available factual data: it might provide questionable results, as illustrated in the following synthetic case study.
Let us set $\pi(x_i) = y_i t_i + (1 - y_i)(1 - t_i)$ (assuming that all x_i are distinct). Since $\pi(x_i) = t_i$ if $y_i = 1$ and $\pi(x_i) \neq t_i$ otherwise, one has

$$y_i \mathbb{1}_{[\pi(x_i)=t]}\mathbb{1}_{[t_i=t]} = \mathbb{1}_{[\pi(x_i)=t]}\mathbb{1}_{[t_i=t]}$$

The Monte-Carlo estimates of $\mathbb{E}[Y|\pi(X) = 1, T = 1]$ and $\mathbb{E}[Y|\pi(X) = 0, T = 0]$ then respectively write:

$$\frac{\sum y_i \mathbb{1}_{[\pi(x_i)=1]}\mathbb{1}_{[t_i=1]}}{\sum \mathbb{1}_{[\pi(x_i)=1]}\mathbb{1}_{[t_i=1]}} = 1, \quad \frac{\sum y_i \mathbb{1}_{[\pi(x_i)=0]}\mathbb{1}_{[t_i=0]}}{\sum \mathbb{1}_{[\pi(x_i)=0]}\mathbb{1}_{[t_i=0]}} = 1 \tag{V.2}$$

and thus the empirical estimate of $R_{pol}(\pi)$ is equal to 0. In other words, knowing both T and Y^T in the *within-sample* settings yields a trivial solution: When $y_i^{t_i} = 0$, one can discard sample i by simply setting $\pi_i = 1 - t_i$. Sample i will not be taken into account into either of the terms $\mathbb{E}[Y|\pi(X) = 0, T = 0]$ and $\mathbb{E}[Y|\pi(X) = 1, T = 1]$. In some sense, this solution amounts to refusing to play unless success is already acquired[1].

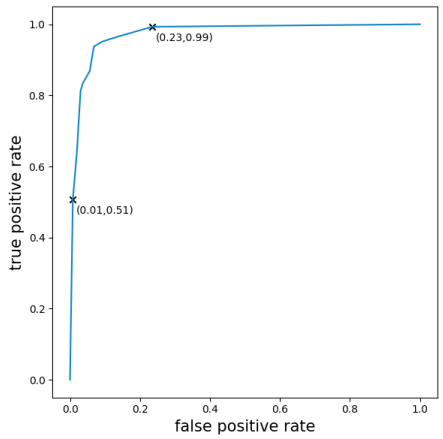
[1]This remark also holds regarding the *within-sample* estimation of the Average Treatment effect on the Treated.

Overall, the estimation of *within-sample* performance indicators must involve quantities that are unavailable at training time (i.e., counter-factuals), such as *PEHE*.

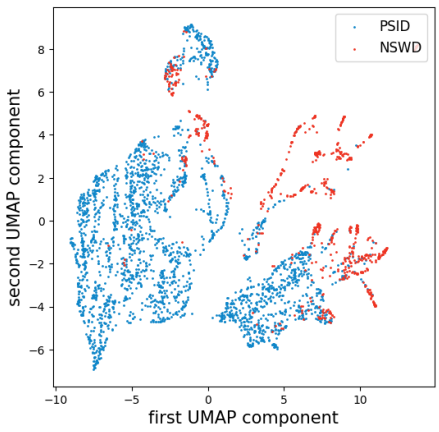
V.1.2.3 . Discussion

Despite being broadly used, the *Jobs* benchmark is poorly suited to assessing causal inference models in general, particularly so for models built on a latent space.

1. Firstly, ***Jobs* violates the positivity assumption** (Eq. II.4). The supports of the covariates in *NSWD* and *PSID* are almost distinct. By design, all treated samples belong to the *NSWD* part of the data; the probability of treatment assignment is thus equal to 0 outside of its covariates support. This is evidenced by the Area Under the *ROC* Curve (*AUC*) of a simple 10-Nearest Neighbors classifier (Fig. V.4a), aimed to distinguish samples from *NSWD* or *PSID* on one instance of *Jobs*. The excellent discrimination of the two classes (*AUC* = .973) is confirmed through a 2D visualization of the data using *UMAP* (McInnes et al., 2018) in Fig. V.4b.



(a) *ROC* curve of a 10-Nearest Neighbors model classifying *Jobs* samples as belonging to *NSWD* or *PSID*.



(b) *UMAP* representation of the covariates of one *Jobs* dataset, differentiating *NSWD* and *PSID*.

Figure V.4: The *Jobs* dataset: visualizing the poor overlap of the covariates supports.

However, the specific *PSID* data might still be relevant to the training of causal models, if the information used to discriminate *NSWD* and *PSID* is irrelevant to predicting the potential outcomes Y^0, Y^1. In such a case, there might exist an embedding ψ such that $\psi(X)$ contains all relevant information to predict the outcomes while the latent spaces of *NSWD*

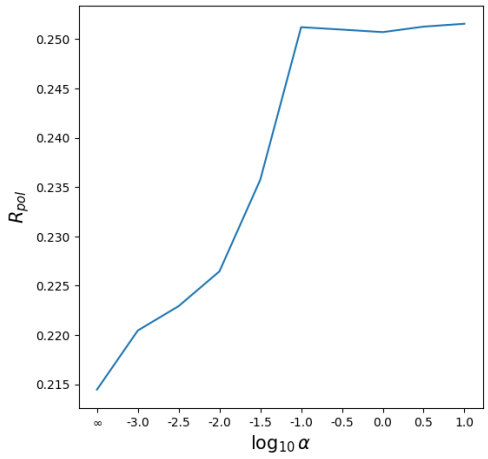


Figure V.5: Policy risk of *CFR-Wass* models versus the weight α of the balance enforcing term (discrepancy).

and *PSID* are confounded:

$$\begin{cases} Y^t \perp\!\!\!\perp X | \psi(X),\ \forall t \in \{0,1\} \\ 0 < \mathbb{P}(X \in \mathit{NSWD}\ | \psi(X) \in \Omega) < 1,\ \forall \Omega\ s.t. \mathbb{P}(\psi(X) \in \Omega) > 0 \end{cases}$$

2. Secondly, the use of **randomized treatment subsets is poorly suited to most *CATE* estimators**. As said, the key challenge for *CATE* models is to handle imbalanced data. While *Jobs* is imbalanced, the performance of the models is assessed on a randomized, hence balanced, subset of the data ($T \perp\!\!\!\perp X \mid X \in \mathit{NSWD}$), making it possible to approximate R_{pol}.

   In the general case, balance in the latent space is sought to achieve a better counter-factual predictive accuracy, at the expense of the factual predictive accuracy. Formally, the training loss of a *CATE* estimator based on a latent space may typically be rewritten under the form

   $$\mathcal{L} = \mathcal{L}_{prediction} + \alpha \mathcal{L}_{discrepancy} + \gamma \Omega_{regularization}$$

   where α and γ respectively control the strength of the discrepancy penalization and that of regularization. The optimal value of α is expected to be 0, for no discrepancy mitigation is required.

   The following experiment supports this claim, training 100 *CFR-Wass* models on each of *Jobs* 10 instances, for α in $\{0, 10^{-3}, \dots 10^{-1}\}$, all other hyper-parameters being fixed. The policy risks on the test set (Fig. V.5) are minimized for $\alpha = 0$, as expected.

$$ELU : x \in \mathbb{R} \mapsto \begin{cases} e^x - 1 & , x < 0 \\ x & , x \geq 0 \end{cases}$$

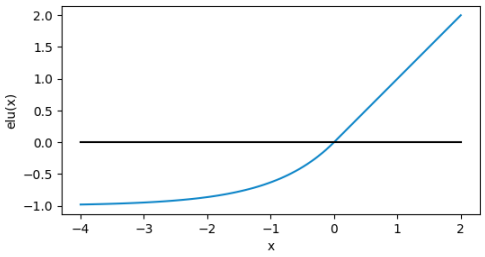


Figure V.6: the Exponential Linear Unit activation function.

V.2 . Experimental setting

The baseline models considered to comparatively assess the performance of ALRITE are listed in Table V.1. The associated results are those reported in the cited papers due to the lack of details regarding the implementation and hyper-parameter adjustment.

The details of the implementation of ALRITE are presented below. The adjustment of the hyper-parameters is detailed in Chapter VI.

Neural architectures Pipelines $\mathcal{P}_0$ and $\mathcal{P}_1$ are trained independently. The architecture and learning protocol are illustrated on $\mathcal{P}_0$ for simplicity.

As said, all modules of $\mathcal{P}_0$ ($\phi_{\mathcal{P}_0}, h^0_{\mathcal{P}_0}, h^1_{\mathcal{P}_0}$) are implemented as neural networks, with Exponential Linear Units (*ELU*; (Clevert et al., 2015)) activation function (Fig. V.6), following Shalit et al. (2017). *ELU* activation functions are immune to the dying *ReLU* issue[2], are differentiable at zero, and let mean activation values get closer to zero.

The neural architecture of mapping ϕ is described from its number L of layers and the layer width W (same for all layers) with activation function *ELU*. The last layer is normalized ($r^{(L)} \mapsto \frac{r^{(L)}}{||r^{(L)}||}$) to prevent learning the trivial embedding $x \mapsto 0$, that achieves a null discrepancy penalization $\alpha \sum_{t_i=0} \mathit{insulation}_{x \mapsto 0}(i)^2$ and a null regularization $\gamma_{\mathcal{P}_0} \Omega(\mathcal{P}_0)$.

The potential outcome models h^0, h^1 also are neural networks; they involve the same number of layers and layer widths (that can differ from that of the ϕ network).

[2]The Rectified Linear Unit function maps $x \in \mathbb{R}$ to 0 if x is negative, to x otherwise: $ReLU : x \mapsto x^+$. "Dying *ReLU*" refers to a situation where a neuron from the network can no longer be optimized. Suppose that there exist weight matrices $a^{(\ell)}, b^{(\ell)}$ such that for any input $r^{(\ell-1)}$ of layer ℓ, the i-th neuron input $\left(a^{(\ell)} r^{(\ell-1)} + b^{(\ell)}\right)_i$ is negative. Then for any input of layer ℓ, $r_i^{(\ell)} = ReLU(a^{(\ell)} r^{(\ell-1)} + b^{(\ell)})$ is null. No gradient may be propagated back through neuron i of layer ℓ, preventing further optimization of $(a^{(\ell)}, b^{(\ell)})_i$. If this phenomenon affects too many neurons, the neural network expressivity is severely harmed.

Model	Metalearner class	Based on
BNN	*S-learner*	Neural networks, single head
CFR-Wass	*T-learner*	Neural networks, two heads, Wasserstein balancing
OLS/LR-2	*T-learner*	Naive linear/logistic estimation of μ^0, μ^1
CEVAE	*T-learner*	Variational Autoencoder
CF	*S-learner*	Trees, last leaf split corresponding to treatment assignment
SITE	*T-learner*	*CFR*'s architecture, specific latent space constraints
GANITE	*T-learner*	Generative Adversarial Networks
NSGP	*T-learner*	Gaussian Processes
ACE	*T-learner*	Similar to *SITE*, finer-grained
DR-CFR	*T-learner*	Decomposition of the representation into adjustment, confounding, instrumental ([A,C,I]) factors
DKLITE	*T-learner*	Gaussian processes, specific focus on counter-factual variance limitation
BWCFR	*T-learner*	*CFR*'s architecture, specific care regarding sample weights
ABCEI	*T-learner*	Adversarial learning, enforcing latent balance
CBRE	*T-learner*	Adversarial learning, and preservation of information in the representation space
MIM-DRCFR	*T-learner*	[A,C,I] decomposition, mutual information disentanglement
DeR-CFR	*T-learner*	[A,C,I] decomposition, orthogonal disentanglement
DRCFR+	*T-learner*	[(A0, A, A1), C, I] decomposition, orthogonal+adversarial disentanglement

Table V.1: Baseline benchmark models.

Training protocol Optimization of the neural modules is (ϕ and outcome models) is conducted using Adam (Kingma and Ba, 2014). Exponential weight decay is used: the initial learning rate is set to 10^{-3}, and it is multiplied by 0.97 after every 100 mini-batches (the size of which is a hyper-parameter of the approach, Chapter VI). L_2 regularization is applied on the weights of the outcome models to avoid overfitting.

Overfitting is prevented using a validation set $\mathcal{D}_{\mathcal{V}}$ including 30% of the training data. When the primary performance metric is *PEHE*, it is impossible to resort to early stopping based on the metric of interest since counterfactual values are unavailable. Therefore the factual prediction error on $\mathcal{D}_{\mathcal{V}}$: $\frac{1}{|D_{\mathcal{V}}|}\sum_{(x,t,y)\in\mathcal{V}}\left(h^t\circ\phi(x)-y\right)^2$ is used as a proxy metric[3], the relevance of which is shown by Thm. 1. In practice, a total number of epochs is fixed and the returned model is the one with optimal factual prediction error.[4]

When the primary performance indicator is policy risk, R_{pol} is computed on the validation dataset, and the returned model is the one minimizing R_{pol}; as in mainstream supervised machine learning, it is considered that the epoch minimizing R_{pol} on the validation set also yields a good performance on the test dataset.

Propensity estimate Propensity estimation is achieved using mainstream supervised learning algorithms, as covariates X and treatment assignment variables T are known on the training set. Several models are considered: logistic regression, k-nearest neighbors, and decision trees. Grid-search on the model and associated hyper-parameter space, using a cross-validation scheme, is used to train and select the best option for each considered dataset.

Software and infrastructure The practical implementation builds on the code released by Shalit et al. (2017) [5], and released by Johansson (2023). As such, it relies entirely on Python code, using mainly *Tensorflow* (Abadi et al., 2016) as a framework for the training of neural models. Auxiliary machine learning models resort to Scikit-Learn (Pedregosa et al., 2011).

The computationally intensive operations (especially the training of large numbers of models for grid-search purposes) have been conducted on Titanic, the cluster of Inria team *TAU*. The main results on *IHDP* (as reported in Table V.2) have been obtained after random selection of 70 sets of hyper-

[3]This quantity, referred to as the estimated μ*-risk* of the model on $\mathcal{D}_{\mathcal{V}}$, will be used to achieve hyper-parameter selection Section VI.2.

[4]In practice, the factual error is only computed every 10 epochs, for the sake of computational and memory resources, given the large number of models considered for hyper-parameter adjustment, see Chapter VI.

[5]as done before by notably Yao et al. (2018); Du et al. (2021); Zhou et al. (2021)

parameters for each pipeline, hence totaling the training of 14000 models. The main results on *Jobs* (as reported in Table V.3) have been obtained after random selection of 540 sets of hyper-parameters for each pipeline, hence totaling the training of 10800 models. In the *IHDP* setting the training of each model requires 4 *CPUs*, 1.5*GB* of *RAM* on average and a total training (wall) time of 81s. In the *Jobs* setting the training of each model requires 4 CPUs, 800*MB* of *RAM* on average, and a total training (wall) time of 20s. The energy expenditures are respectively estimated as 17kWh ($8.9e2$g CO_2e) and 3kWh ($1.7e2$g CO_2e) (Lannelongue et al., 2021).

V.3 . Experimental results

Tables V.2 and V.3 report the performance indicators for all considered baselines and ALRITE on the *IHDP* and *Jobs* benchmarks. As said, the reported metric values are averaged over all instances of each dataset, and the performances of the baselines are taken from the cited papers. The bias affecting the average effects estimation is analyzed in Section V.3.3. Finally, the reproducibility of the baselines is discussed in Section V.3.4.

V.3.1 . *IHDP*

Overall, the merits of ALRITE are demonstrated on *IHDP* as it ranks first on within-sample *PEHE* and second on out-of-sample *PEHE*. Indeed, ALRITE is designed to optimize the *PEHE* performance indicator, as motivated by Thm. 1. The comparatively lesser performance regarding within-sample and out-of-sample ϵ_{ATE} is blamed on the L_2 regularization (more in Section V.3.3), tending to bias the estimates toward 0, as noted by Laan and Rose (2011).

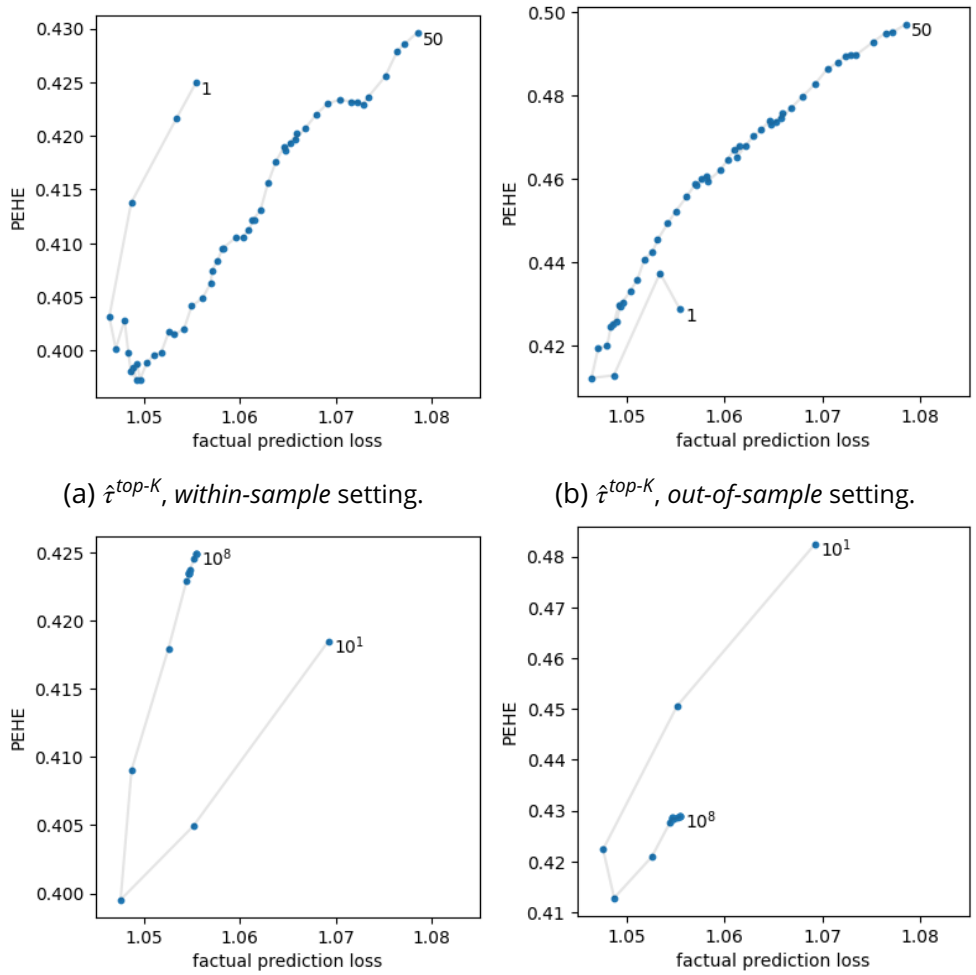
The ensemble variants of ALRITE introduced in Section IV.4 are also applied on *IHDP*. The associated *PEHE* and factual errors are depicted in Fig. V.7 for the considered ranges of their proper hyper-parameters (number K of models in the *top-K* ensemble, temperature λ in the *softmax*$_\lambda$ ensemble). As shown, the factual error can reliably be used to select the hyper-parameter value with nearly optimal *PEHE*, set to $K = 4$ and $\lambda = 100$ for respectively the *top-K* and the *softmax*$_\lambda$ ensembles. With their tuned hyper-parameters, the *top-K* version significantly outperforms ALRITE (with *p-value*$= 1.3e-3$ on a one-sided paired *t-test*) but the *softmax*$_\lambda$ version does not (*p-value*$= .24$).

V.3.2 . *Jobs*

On *Jobs*, ALRITE does not perform well; it is outperformed by most

	IHDP			
	within-sample		*out-of-sample*	
	$\sqrt{PEHE}$	ϵ_{ATE}	$\sqrt{PEHE}$	ϵ_{ATE}
OLS/LR-2	$2.4 \pm .1$	$.14 \pm .01$	$2.5 \pm .1$	$.31 \pm .02$
BNN	$2.2 \pm .1$	$.37 \pm .03$	$2.1 \pm .1$	$.42 \pm .03$
CF	$3.8 \pm .2$	$.18 \pm .01$	$3.8 \pm .2$	$.40 \pm .03$
CFR-Wass	$.71 \pm .0$	$.25 \pm .01$	$.76 \pm .0$	$.27 \pm .01$
CEVAE	$2.7 \pm .1$	$.34 \pm .01$	$2.6 \pm .1$	$.46 \pm .02$
SITE	$.60 \pm .09$	/	$.66 \pm .11$	/
GANITE	$1.9 \pm .4$	$.43 \pm .05$	$2.4 \pm .4$	$.49 \pm .05$
NSGP	$.51 \pm .01$	/	$.64 \pm .03$	/
ACE	$.49 \pm .005$	/	$.54 \pm .06$	/
DKLITE	$.52 \pm .02$	/	$.65 \pm .03$	/
DR-CFR	/	/	$.65 \pm .03$	$\mathbf{.03} \pm .04$
BWCFR	/	/	$.63 \pm .01$	$.19 \pm .01$
ABCEI	$.71 \pm .0$	$\mathbf{.09} \pm .01$	$.73 \pm .0$	$.09 \pm .01$
CBRE	$.52 \pm .0$	$.10 \pm .01$	$.60 \pm .1$	$.13 \pm .02$
MIM-DRCFR	/	/	$\mathbf{.38} \pm .009$	$.09 \pm .001$
DeR-CFR	$.44 \pm .02$	$.13 \pm .02$	$.53 \pm .07$	$.15 \pm .02$
DRCFR+	$.86 \pm .01$	/	$1.19 \pm .01$	/
ALRITE	$.42 \pm .03$	$.12 \pm .009$	$.43 \pm .02$	$.13 \pm .01$
***K-top* ALRITE**	$\mathbf{.40} \pm .02$	$.10 \pm .008$	$.41 \pm .02$	$.11 \pm .009$
***softmax*$_\lambda$ ALRITE**	$\mathbf{.40} \pm .023$	$.095 \pm .008$	$.42 \pm .02$	$.11 \pm .008$

Table V.2: Comparative performances of ALRITE (ours) on *IHDP* (lower is better).



(a) $\hat{\tau}^{top\text{-}K}$, *within-sample* setting.

(b) $\hat{\tau}^{top\text{-}K}$, *out-of-sample* setting.

(c) $softmax_\lambda$, *within-sample* setting.

(d) $softmax_\lambda$, *out-of-sample* setting.

Figure V.7: Ensemble models: *PEHE* vs factual error on the validation set sensitivity, depending on the hyper-parameter. Top: Top-K ensemble, with K in $[\![1, 50]\!]$. Bottom: softmax-λ ensemble, with λ in $\{10^{k/2}\}_{k=2}^{16}$.

	Jobs			
	within-sample		*out-of-sample*	
	R_{pol}	ϵ_{ATT}	R_{pol}	ϵ_{ATT}
OLS/LR-2	$.21 \pm .0$	$.01 \pm .01$	$.24 \pm .0$	$.08 \pm .03$
BNN	$.20 \pm .0$	$.04 \pm .01$	$.24 \pm .0$	$.09 \pm .04$
CF	$.19 \pm .0$	$.03 \pm .01$	$.20 \pm .0$	$.07 \pm .03$
CFR-Wass	$.17 \pm .0$	$.04 \pm .01$	$.21 \pm .0$	$.09 \pm .03$
CEVAE	$.15 \pm .0$	$.02 \pm .01$	$.26 \pm .0$	$\mathbf{.03} \pm .01$
SITE	$.22 \pm .00$	/	$.22 \pm .01$	/
GANITE	$\mathbf{.13} \pm .01$	$\mathbf{.01} \pm .01$	$\mathbf{.14} \pm .01$	$.06 \pm .03$
NSGP	/	/	/	/
ACE	$0.22 \pm .01$	/	0.22 ± 0.01	
DKLITE	$\mathbf{.13} \pm .01$	/	$\mathbf{.14} \pm .01$	/
DR-CFR	/	/	/	/
BWCFR	/	/	/	/
ABCEI	$\mathbf{.13} \pm .0$	$.02 \pm .01$	$.17 \pm .0$	$\mathbf{.03} \pm .01$
CBRE	$\mathbf{.13} \pm .0$	/	$.28 \pm .0$	/
MIM-DRCFR	/	/	/	/
DeR-CFR	$.19 \pm .04$	$.05 \pm .09$	$.21 \pm .01$	$.09 \pm .0$
DRCFR+	/	/	/	/
ALRITE	$.22 \pm .01$	$.07 \pm .02$	$.23 \pm .02$	$.06 \pm .02$

Table V.3: Comparative performances of ALRITE and baselines on *Jobs* (lower is better; the relevance of the within-sample performance is discussed in Section V.1.2).

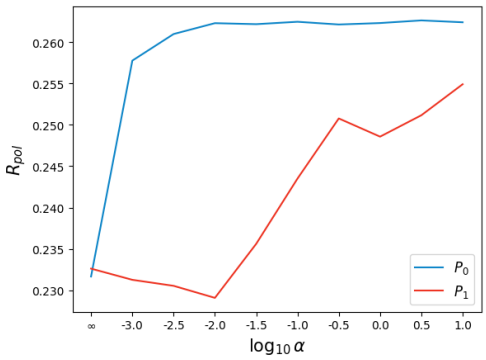


Figure V.8: Influence of the regularization strength α on the policy risks of pipelines $\mathcal{P}_0$ and $\mathcal{P}_1$. The results are averaged over all 10 *Jobs* instances (100 runs per instance) for each α value.

baselines. Indeed, the training set contains data from both *NSWD* and *PSID*[6]. $\mathcal{P}_0$ aims at getting insulation low for control samples, and in particular for *PSID* samples. In other words, it ensures that each *PSID* sample gets neighbored by *NSWD* ones in the latent space. As discussed in Section V.1.2.3, this objective is inappropriate as the supports of *NSWD* and *PSID* are almost disjoint; furthermore, the evaluation metric does not take *PSID* samples into account. ALRITE thus **sacrifices the factual predictive accuracy of treated samples** while failing to improve the counter-factual predictive accuracy of *PSID* samples. As measured by the R_{pol} indicator, this results in a net performance loss.

The impact of the discrepancy regularization term on pipelines $\mathcal{P}_0$ and $\mathcal{P}_1$ is investigated by varying hyper-parameters α (controlling the strength of insulation) and β (controlling the impact of exemplarity) ranging in $\{0, 10^{-3}, \dots 10^{-1}\}$, all other hyper-parameters being fixed. Fig. V.8 displays the results, confirming that lower α values are associated with smaller R_{pol} scores, implying better performance. As expected, any amount of discrepancy regularization severely harms the averaged performance of pipeline $\mathcal{P}_0$, while small amounts of regularization do not affect pipeline $\mathcal{P}_1$.

V.3.3 . Estimation bias

Let us investigate the *ATE* performance of ALRITE. The reason why it is less good than the *PEHE* one is that, as noted in Section III.2.3, the larger the parameter space of models, the more likely they are to suffer from large biases. An identified source of bias lies in model regularization, embodied by

[6] Note that if the available data were restricted to *NSWD*, the control and treatment distribution would be the same and a two-pipeline architecture would be meaningless.

the fourth term of the training loss:

$$\begin{aligned}\mathcal{L}_{\mathcal{P}_t} =& \frac{1}{n_t}\sum_{t_i=t} error_{\mathcal{P}_t}(x_i,t_i,y_i)\\ &+\frac{\alpha_{\mathcal{P}_t}}{n_t}\sum_{t_i=t} insulation_{\mathcal{P}_t}(i)^2\\ &+\frac{1}{n_{1-t}+\beta_{\mathcal{P}_t} n_t}\sum_{t_i=1-t}(1+\beta_{\mathcal{P}_t} exemplarity_{\mathcal{P}_t}(i))\times error_{\mathcal{P}_t}(x_i,t_i,y_i)\\ &+\gamma_{\mathcal{P}_t}\,\Omega(\mathcal{P}_t)\end{aligned} \tag{V.3}$$

While the L_2 regularization limits the over-fitting in neural networks, it does so at the price of an increased bias. Complementary experiments are conducted on ALRITE to provide evidence for the role of regularization in model bias, and measure its impact as follows.

V.3.3.1 . Evidence

ALRITE is launched on the 100 datasets in *IHDP*-100, setting the γ parameter to 10^{-4}. Consider the control and treated estimate $\hat{\mu}^0$ and $\hat{\mu}^1$,

$$\begin{aligned}\hat{\mu}^0 &: x\in\mathcal{X}\mapsto \hat{\eta}(x)h^1_{\mathcal{P}_0}\circ\phi_{\mathcal{P}_1}(x)+(1-\hat{\eta}(x))h^0_{\mathcal{P}_0}\circ\phi_{\mathcal{P}_0}(x)\\ \hat{\mu}^1 &: x\in\mathcal{X}\mapsto \hat{\eta}(x)h^1_{\mathcal{P}_1}\circ\phi_{\mathcal{P}_1}(x)+(1-\hat{\eta}(x))h^1_{\mathcal{P}_0}\circ\phi_{\mathcal{P}_0}(x)\end{aligned}$$

Taking advantage of the fact that the (simulated) counterfactuals are available on *IHDP*, the associated error may be computed:

$$\begin{aligned}\epsilon^0_i &= \hat{\mu}^0(x_i)-\mu^0(x_i)\\ \epsilon^1_i &= \hat{\mu}^1(x_i)-\mu^1(x_i)\end{aligned}$$

These errors are then regressed against $\mu^t(x_i)$, with $r^t\times\mu^t(x_i)$ being the "linearly explained" fraction of the error ϵ^t_i, and δ^t_i the residual:

$$\begin{aligned}\epsilon^0_i &= r^0\times\mu^0(x_i)+\delta^0_i\\ \epsilon^1_i &= r^1\times\mu^1(x_i)+\delta^1_i\end{aligned}$$

With these notations, the empirical approximation of the error in *ATE* estimation writes

$$\begin{aligned}\epsilon_{ATE} &= |\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}^1(x_i)-\hat{\mu}^0(x_i)\right)-\left(\mu^1(x_i)-\mu^0(x_i)\right)|\\ &= |\frac{1}{n}\sum_{i=1}^{n}(\epsilon^1_i-\epsilon^0_i)|\end{aligned}$$

$$= | \underbrace{\frac{r^1}{n}\sum_{i=1}^{n}\mu^1(x_i) - \frac{r^0}{n}\sum_{i=1}^{n}\mu^0(x_i)}_{\text{"linearly explained"}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\delta_i^1 - \delta_i^0)}_{\text{residual}} |$$

Here, $\frac{r^1}{n}\sum_{i=1}^{n}\mu^1(x_i) - \frac{r^0}{n}\sum_{i=1}^{n}\mu^0(x_i)$ represents the "linearly explained" part of the *ATE* estimation error, while $\frac{1}{n}\sum_{i=1}^{n}(\delta_i^1 - \delta_i^0)$ is the residual.

In Fig. V.9, error ϵ_i^t is plotted against the ground truth $\mu^t(x_i)$ for each sample and each dataset. Both regression coefficients r^0 and r^1 are found to be negative, with $r^1 < r^0 < 0$.

In other words, the greater the outcomes y^0 and y^1, the worse the underestimation: factual predictions tend to be biased towards 0. Moreover, it appears on the *IHDP* benchmark that the outcome Y takes on average higher values for treated individuals than for control ones (Fig. V.9c): $0 < \frac{1}{n}\sum_{i=1}^{n}\mu^0(x_i) < \frac{1}{n}\sum_{i=1}^{n}\mu^1(x_i)$. The numerical evaluation of the "linearly explained" part of the *ATE* estimation error appears to be negative. The prediction biases for the control and treated outcomes do not compensate for each other, adversely affecting *ATE* estimation.

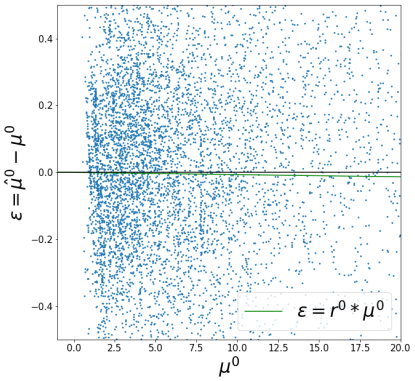
V.3.3.2 . Impact of the bias

In a second experiment, the L_2 penalization weight parameter γ (Eq. IV.3) is varied in $(\gamma^1, \dots, \gamma^M)$, yielding a range of trained models $\hat{\tau}^{(1)}, \dots, \hat{\tau}^{(M)}$, while all other hyper-parameters are kept the same. For each $\hat{\tau}^{(m)}$, the associated regression coefficient r^{γ^m} is computed as in the previous experiment. Fig. V.10 displays r^γ versus γ (in log scale), showing that after a plateau, r^γ decreases as γ increases. Accordingly, larger values of γ are associated with a stronger model bias towards zero and, consequently, larger errors in *ATE* and *ATT* estimation.

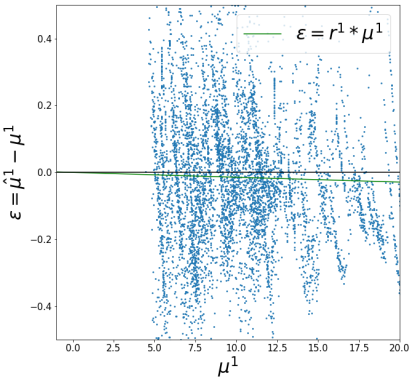
Further work will investigate the solutions proposed to address this drawback. Kennedy (2023)'s *DR-learner* (Section III.3.1.7) adapts double machine learning (Chernozhukov et al., 2018) to the *CATE* estimation setting, The training of pipeline $\mathcal{P}_0$ might be completed with a second stage: instead of defining $\hat{\tau}_{\mathcal{P}_0}$ as the difference $h^1_{\mathcal{P}_0}\circ\phi_{\mathcal{P}_0} - h^0_{\mathcal{P}_0}\circ\phi_{\mathcal{P}_0}$, it would be sought as a minimizer of $\sum_i \left(\hat{\tau}(x_i) - \left(\hat{\mu}^1_{\mathcal{P}_0}(x_i) - \hat{\mu}^0_{\mathcal{P}_0}(x_i) + (2t_i - 1)\hat{\rho}^{t_i}(x_i)(y_i - \hat{\mu}^{t_i}(x_i))\right)\right)^2$.

V.3.4 . Baseline reproducibility

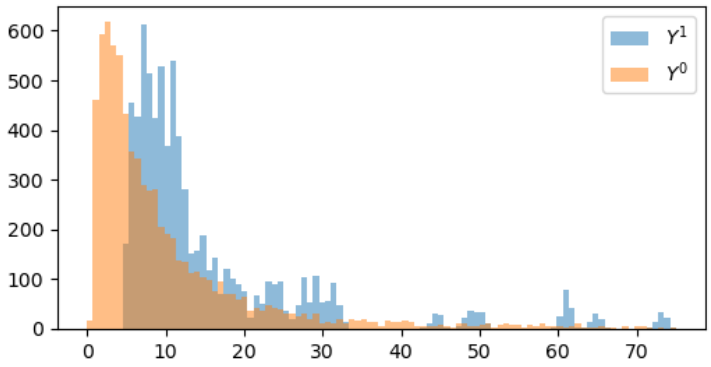
After (Pineau et al., 2021), the overall field of Machine Learning faces reproducibility challenges. The comparison of the performance of ALRITE with that of baseline models also faces quite some difficulties. Besides the usual



(a) Prediction error on control samples.



(b) Prediction error on treated samples.



(c) Distribution of Y^0 and Y^1 on *IHDP*-100.

Figure V.9: Prediction error on control and treated samples, and the associated outcome distributions.

issues[7], the main trouble comes from the lack of details about the selected hyper-parameters, and about the hyper-parameter selection strategy (Chapter VI).

Another issue is related to the *IHDP* benchmark itself. While most authors consider the versions of *IHDP*-100 and *IHDP*-1000 released by Johansson et al. (2016) and based on (Dorie, 2023), some authors (Zhang et al., 2021; Cheng et al., 2022b) have generated their own versions of *IHDP*-100 and *IHDP*-1000 with different random seeds and train/valid/test split proportions, reporting the results on *IHDP*-1000, or *IHDP*-100. As the final *PEHE* value essentially de-

[7]Including the usage of long-deprecated packages (Shalit et al., 2017), or the lack of comments in the code, or the lack of precision regarding the *IHDP* setting (Cheng et al., 2022b).

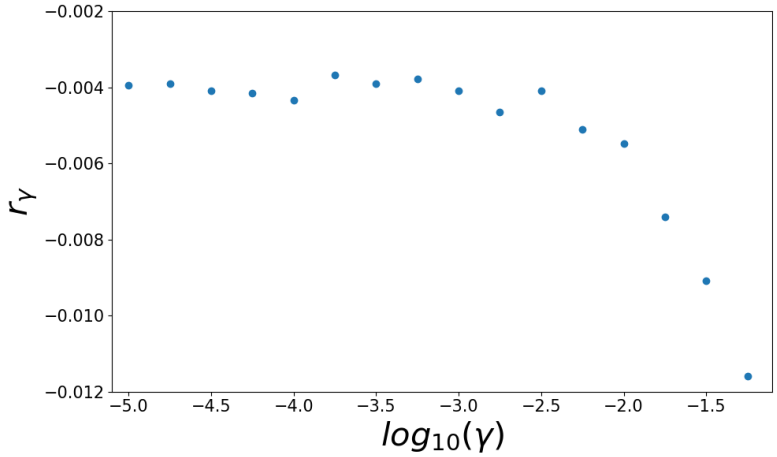


Figure V.10: Relationship between L_2 regularization strength coefficient γ and regression coefficient r^γ.

pends on the difficult *IHDP* instances (with large *CATE* variance), the diversity of these experimental settings also makes it difficult and/or computationally expensive to conduct a fair comparison.

V.4 . Partial conclusion

This chapter has demonstrated the relevance of ALRITE on the main causal estimation benchmark, *IHDP*. Some conjectures explaining the difficulties faced by ALRITE on the *Jobs* dataset have been proposed; complementary experiments supporting these explanations have been conducted and discussed. The next chapter will tackle the issue of hyper-parameter selection, of key importance in efficient causal estimation.

VI - Hyper-parameter selection in causal inference

Hyper-parameter selection, referred to as AutoML in the context of supervised learning (Hutter et al., 2019), is known to be a tedious and time-consuming task, though an essential one to reach good performances. Its difficulty is significantly increased in the context of causal inference, due to the fact that the counterfactual information is *per se* unknown.

This chapter focuses on hyper-parameter selection in the context of causal inference, referring the reader to Cheng et al. (2022a) for a comprehensive introduction to causal learning evaluation methods. After discussing the position of the problem (Section VI.1), noting that most performance indicators are infeasible, i.e. cannot be computed, we present proxies thereof (Section VI.2). Their comparative assessment is detailed in Section VI.3. Eventually, the procedure used to select ALRITE hyper-parameters is detailed in Section VI.4, and experiments on synthetic data are used to *a posteriori* support this procedure (Section VI.5).

Validation set. All proxies for performance indicators are assessed on a subset of held-out samples, referred to as validation dataset. Note that the assessment can be averaged using a cross-validation procedure, as in supervised learning.

VI.1 . Position of the problem

Let us consider a finite observational dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i\in[\![1,n]\!]}$ in the Neyman-Rubin potential outcomes framework, with $\mathcal{D}_\mathcal{V} = \{(x_i, t_i, y_i)\}_{i\in\mathcal{V}}$ a randomly drawn subset used as validation set. Given several *CATE* estimates, noted $\hat{\tau}^{(i)}$, for $i = 1 \dots C$, trained on $\mathcal{D}\backslash\mathcal{D}_\mathcal{V}$, the question is to select the best estimate.

The difficulty is that the key performance indicators $PEHE(\hat{\tau}), \epsilon_{ATE}(\hat{\tau}), R_{pol}(\hat{\tau}), \epsilon_{ATT}(\hat{\tau})$[1] are **infeasible**.

[1]Let us remind the definitions:

$$\mathcal{D}_\mathcal{V} \mapsto \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} \left(\hat{\tau}(x) - \tau(x)\right)^2$$

$$\mathcal{D}_\mathcal{V} \mapsto \left| \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} \hat{\tau}(x) - \tau(x) \right|$$

The state of the art investigates two directions to address this difficulty.

The first direction is based on integrating model selection depending on the class of models retained to learn $\hat{\mu}^t$, as exemplified by Athey and Imbens (2016) for causal trees, or Powers et al. (2018) for causal boosting and bagged causal multivariate adaptive regression splines. Along a different line, Alaa and Van Der Schaar (2018) rely on an information-theoretic criterion, matching the expressivity of the outcome function estimators and the regularity of the true outcome functions μ^0, μ^1 in the context of Gaussian processes. This direction will not be considered further in this chapter for the sake of generality.

The second direction considers proxy metrics, required to be *feasible* and such that the performance of the model optimizing the chosen proxy metric is close enough to the best one w.r.t. the infeasible target metrics. Quite a few work (Schuler et al., 2018; Doutreligne and Varoquaux, 2023; Curth and Schaar, 2023) aims at identifying conditions under which the model selected after the proxy metrics is a satisfying choice.[2]

VI.2 . Proxy metrics

$$\mathcal{D}_\mathcal{V} \mapsto \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} 1 - \hat{\mu}^0(x)\mathbb{1}[\hat{\tau}(x)\leq 0] - \hat{\mu}^1(x)\mathbb{1}[\hat{\tau}(x)>0]$$

$$\mathcal{D}_\mathcal{V} \mapsto \left| \frac{1}{\sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} t} \sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} t\big(\hat{\tau}(x) - \tau(x)\big) \right|$$

[2]Meaning that the ranks of the model according to the target and the proxy metrics are close, or that their value are close.
As an illustration, while the straightforward $PEHE(\hat{\tau})$ estimator of a given candidate causal model $\hat{\tau}$ is not feasible, the observed policy risk oR_{pol}, however, admits a feasible estimator $\widehat{oR_{pol}}_\bullet(\hat{\tau})$:

$$\begin{aligned} oR_{pol} : \hat{\tau} \mapsto\ & 1 - \mathbb{E}[Y|\hat{\tau}(x) > 0, T = 1]\mathbb{P}(\hat{\tau}(x) > 0) \\ & - \mathbb{E}[Y|\hat{\tau}(x) \leq 0, T = 0]\mathbb{P}(\hat{\tau}(x) \leq 0) \\ \widehat{oR_{pol}}_\bullet(\hat{\tau}) : \mathcal{D}_\mathcal{V} \mapsto\ & 1 - \frac{\sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} \mathbb{1}[\hat{\tau}(x)>0]}{|\mathcal{D}_\mathcal{V}|} \frac{\sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} yt\mathbb{1}[\hat{\tau}(x)>0]}{\sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} t\mathbb{1}[\hat{\tau}(x)>0]} \\ & - \frac{\sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} \mathbb{1}[\hat{\tau}(x)\leq 0]}{|\mathcal{D}_\mathcal{V}|} \frac{\sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} y(1-t)\mathbb{1}[\hat{\tau}(x)\leq 0]}{\sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} (1-t)\mathbb{1}[\hat{\tau}(x)\leq 0]} \end{aligned}$$

The selected candidate will then be $\operatorname{argmin}_{\hat{\tau}\in C} \widehat{oR_{pol}}_{\mathcal{D}_\mathcal{V}}(\hat{\tau})$. Indeed oR_{pol} is probably not the best proxy metric to select the lowest $PEHE$ model: if an estimate $\hat{\tau}$ satisfies $\hat{\tau} \times \tau > 0$ almost surely, then it achieves minimal oR_{pol} value. Notably, $oR_{pol}(\tau) = oR_{pol}(2\tau)$, but $PEHE(\tau) = 0$ while $PEHE(2\tau) = \mathbb{E}[\tau(X)^2]$.

This section presents the most common proxy metrics, without pretending to exhaustivity. Schuler et al. (2018) distinguish four main classes of scores.

VI.2.1 . μ-risks

Suppose that $\hat{\tau}$ is obtained through a *S-learner* or a *T-learner* and let $\hat{\mu}^0, \hat{\mu}^1$ be the two underlying potential outcomes estimates: $\hat{\tau} = \hat{\mu}^1 - \hat{\mu}^0$. The μ-*risk* of estimate $\hat{\tau} = \hat{\mu}^1 - \hat{\mu}^0$ is defined as:

$$\mu\text{-}risk(\hat{\mu}^0, \hat{\mu}^1) = \mathbb{E}[(\hat{\mu}^T(X) - Y)^2]$$

and admits as a feasible estimator

$$\widehat{\mu\text{-}risk}(\hat{\mu}^0, \hat{\mu}^1) : \mathcal{D}_\mathcal{V} \mapsto \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} \left(\hat{\mu}^t(x) - y\right)^2$$

We have so far referred to μ-*risk* as the factual validation prediction error.

A variation on μ-*risk* (Laan and Robins, 2011) denoted as $\mu\text{-}risk_{IPTW}$ resorts to **Inverse Probability of Treatment Weighting** (*IPTW*):

$$\mu\text{-}risk_{IPTW}(\hat{\mu}^0, \hat{\mu}^1) = \mathbb{E}\left[\varrho^T(X)\left(\hat{\mu}^T(X) - Y\right)^2\right]$$

Note that the true inverse probability of treatment ϱ (Eq. II.8) is unknown, so the estimator

$$\mathcal{D}_\mathcal{V} \mapsto \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} \varrho^t(x)\left(\hat{\mu}^t(x) - y\right)^2$$

is not feasible. It is, however, possible to build an estimate $\hat{\varrho}$ of ϱ and plug it into the estimator formula:

$$\widehat{\mu\text{-}risk}_{IPTW}(\hat{\mu}^0, \hat{\mu}^1) : \mathcal{D}_\mathcal{V} \mapsto \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}} \hat{\varrho}^t(x)\left(\hat{\mu}^t(x) - y\right)^2$$

Note also that if the propensity score model is misspecified, $\widehat{\mu\text{-}risk}_{IPTW}(\hat{\mu}^0, \hat{\mu}^1)$ is a biased (and inconsistent) estimator of $\mu\text{-}risk_{IPTW}$. Nevertheless, provided the approximation $\hat{\eta}$ is accurate enough it may remain a valuable proxy metric.

VI.2.2 . π-risk

Let us slightly adapt the definitions (in e.g. Schuler et al. (2018)) to present a unified framework for the Policy Risk R_{pol} (the lower, the better).

As introduced in Section II.5.1, R_{pol} evaluates $\hat{\tau}$ based on the return of policy $\pi_{\hat{\tau}} : x \mapsto \mathbb{1}[\hat{\tau}(x)>0]$, with:

$$\begin{aligned}\pi\text{-}risk(\hat{\tau}) = R_{pol}(\hat{\tau}) = 1 &- \mathbb{E}[Y^0|\pi_{\hat{\tau}}(x) = 0, T = 0]\mathbb{P}(\pi_{\hat{\tau}}(x) = 0) \\ &- \mathbb{E}[Y^1|\pi_{\hat{\tau}}(x) = 1, T = 1]\mathbb{P}(\pi_{\hat{\tau}}(x) = 1)\end{aligned}$$

Let us now introduce an estimator of $R_{pol}(\hat{\tau})$:

$$\widehat{\pi\text{-}risk}_{IPTW}(\hat{\tau}) : \mathcal{D}_{\mathcal{V}} \mapsto \frac{1}{|\mathcal{D}_{\mathcal{V}}|} \sum_{(x,t,y)\in\mathcal{D}_{\mathcal{V}}} 1 - \hat{\varrho}^t(x)\mathbb{1}_{[t=\pi_{\hat{\tau}}(x)]}y$$

Zhao et al. (2017) show that $1 - \hat{\varrho}^T(X)\mathbb{1}_{[T=\pi_{\hat{\tau}}(X)]}Y$ is an unbiased estimate of R_{pol} if $\hat{\varrho}$ is correctly specified ($\hat{\varrho} = \varrho$) and if the treatment assignment is independent from X, Y^0 and Y^1 (as in randomized control trials), while Schuler et al. (2018) do not consider its bias. Let us show that it is an unbiased estimator of R_{pol} as long as $\hat{\varrho} = \varrho$ and conditional exchangeability holds.

Proof. Assume that conditional exchangeability holds: $Y^t \perp\!\!\!\perp T|X,\ \forall t \in \{0,1\}$

$$\begin{aligned}
\mathbb{E}[Y\mathbb{1}_{[T=\pi_{\hat{\tau}}(X)]}\varrho^T(X)] =& \mathbb{E}_X\left[\mathbb{E}[Y\mathbb{1}_{[T=\pi_{\hat{\tau}}(X)]}\varrho^T(X)|X]\right] && \text{(total probabilities on } X\text{)}\\
=& \mathbb{E}_X\left[\mathbb{E}[Y\varrho^T(X)|X, T=\pi_{\hat{\tau}}(X)]\mathbb{P}(T=\pi_{\hat{\tau}}(X)|X)\right] && \text{(total probabilities on } T\text{)}\\
=& \mathbb{E}_X\left[\mathbb{E}[Y|X, T=\pi_{\hat{\tau}}(X)]\right]\\
=& \mathbb{E}_X\left[\mathbb{E}[Y^{\pi_{\hat{\tau}}(X)}|X, T=\pi_{\hat{\tau}}(X)]\right] && \text{(SUTVA)}\\
=& \mathbb{E}_X\left[\mathbb{E}[Y^{\pi_{\hat{\tau}}(X)}|X]\right] && (Y^t \perp\!\!\!\perp T|X)\\
=& \mathbb{E}[Y^{\pi_{\hat{\tau}}(X)}]\\
=& \mathbb{E}[Y^{\pi_{\hat{\tau}}(X)}|\pi_{\hat{\tau}}(X)=1]\mathbb{P}(\pi_{\hat{\tau}}(X)=1)\\
&+ \mathbb{E}[Y^{\pi_{\hat{\tau}}(X)}|\pi_{\hat{\tau}}(X)=0]\mathbb{P}(\pi_{\hat{\tau}}(X)=0)\\
=& \mathbb{E}[Y^1|\pi_{\hat{\tau}}(X)=1]\mathbb{P}(\pi_{\hat{\tau}}(X)=1)\\
&+ \mathbb{E}[Y^0|\pi_{\hat{\tau}}(X)=0]\mathbb{P}(\pi_{\hat{\tau}}(X)=0)
\end{aligned}$$

□

The *π-risk* also admits as feasible estimator (Cassel et al., 1976; Dudík et al., 2011):

$$\widehat{\pi\text{-}risk}_{DR}(\hat{\tau}) : \mathcal{D}_{\mathcal{V}} \mapsto \frac{1}{|\mathcal{D}_{\mathcal{V}}|} \sum_{(x,t,y)\in\mathcal{D}_{\mathcal{V}}} 1 - \hat{\mu}^{\pi_{\hat{\tau}}(x)}(x) - \hat{\varrho}^t(x)\mathbb{1}_{[t=\pi_{\hat{\tau}}(x)]}\left(y - \hat{\mu}^t(x)\right)$$

Here $1 - \hat{\mu}^{\pi_{\hat{\tau}}(X)}(X) - \hat{\varrho}^T(X)\mathbb{1}_{[T=\pi_{\hat{\tau}}(X)]}\left(Y - \hat{\mu}^T(X)\right)$ is doubly robust[3] , i.e, unbiased assuming that $\hat{\varrho} = \varrho$ or $\hat{\mu} = \mu$.

[3]Denote by R the random variable $1 - \hat{\mu}^{\pi_{\hat{\tau}}(X)}(X) - \hat{\varrho}^T(X)\mathbb{1}_{[T=\pi_{\hat{\tau}}(X)]}\left(Y - \hat{\mu}^T(X)\right)$. Let us first compute the expectation of R conditionally to $X = x$, where $x \in \mathcal{X}$:

$$\begin{aligned}
\mathbb{E}[R|X=x] =& \mathbb{E}\left[1 - \hat{\mu}^{\pi_{\hat{\tau}}(X)}(X) - \hat{\varrho}^T(X)\mathbb{1}_{[T=\pi_{\hat{\tau}}(X)]}\left(Y - \hat{\mu}^T(X)\right)|X=x\right]\\
=& 1 - \hat{\mu}^{\pi_{\hat{\tau}}(x)}(x) - \mathbb{E}\left[\hat{\varrho}^T(x)\left(Y - \hat{\mu}^T(x)\right)|X=x, T=\pi_{\hat{\tau}}(x)\right]\mathbb{P}(T=\pi_{\hat{\tau}}(x)|X=x)\\
=& 1 - \hat{\mu}^{\pi_{\hat{\tau}}(x)}(x) - \left(\mu^{\pi_{\hat{\tau}}(x)}(x) - \hat{\mu}^{\pi_{\hat{\tau}}(x)}(x)\right)\frac{\pi_{\hat{\tau}}(x)\eta(x) + (1-\pi_{\hat{\tau}}(x))(1-\eta(x))}{\pi_{\hat{\tau}}(x)\hat{\eta}(x) + (1-\pi_{\hat{\tau}}(x))(1-\hat{\eta}(x))}
\end{aligned}$$

VI.2.3 . *R-risk*

The *R-risk* is defined as $\textit{R-risk} : \hat{\tau} \mapsto \mathbb{E}\left[\left(Y - m(X) - (T - \eta(X))\hat{\tau}(X)\right)^2\right]$. According to Robinson's decomposition (Section III.3.1.4), it is minimized by the true conditional average treatment effect function τ. *R-risk* is as such a relevant score to measure the performance of *CATE* estimates, and admits as a feasible estimator

$$\widehat{\textit{R-risk}}(\hat{\tau}) : \mathcal{D}_{\mathcal{V}} \mapsto \frac{1}{|\mathcal{D}_{\mathcal{V}}|} \sum_{(x,t,y)\in\mathcal{D}_{\mathcal{V}}} \Big(\big(y - \hat{m}(x)\big) - \big(t - \hat{\eta}(x)\big)\hat{\tau}(x)\Big)^2$$

VI.2.4 . τ-*risks*

With the same notations as above, the τ-risk of $\hat{\tau}$ is defined as:

$$\tau\textit{-risk}(\hat{\tau}) = \mathbb{E}[(\hat{\tau}(X) - \tau(X))^2]$$

As said, the straightforward estimator $\mathcal{D}_{\mathcal{V}} \mapsto \frac{1}{|\mathcal{D}_{\mathcal{V}}|}\sum_{(x,t,y)\in\mathcal{D}_{\mathcal{V}}}(\hat{\tau}(x) - \tau(x))^2$ is infeasible. A solution consists in replacing τ with a plug-in estimate $\tilde{\tau}$. The options are plentiful: any τ estimate is a potential match for the plug-in term, though not all estimates are relevant. Let $\hat{\mu}^0, \hat{\mu}^1$ be estimates of the outcome functions μ^0, μ^1. The naive τ-*risk* estimator

$$\widehat{\tau\textit{-risk}}_{\textit{naive}}(\hat{\tau}) : \mathcal{D}_{\mathcal{V}} \mapsto \frac{1}{|\mathcal{D}_{\mathcal{V}}|} \sum_{(x,t,y)\in\mathcal{D}_{\mathcal{V}}} \big(\hat{\tau}(x) - (\hat{\mu}^1(x) - \hat{\mu}^0(x))\big)^2$$

is highly sensitive to misspecifications of the outcome functions estimates, and alternatives are explored in the following.

The most popular plug-in estimate (Shalit et al., 2017; Du et al., 2021; Zhou et al., 2021) is based on **One Nearest-Neighbor Imputation** (1NNI). The

1. *Case 1.* Assume η is correctly specified, i.e, $\hat{\eta} = \eta$. Then, the equation simplifies in $\mathbb{E}[R|X = x] = 1 - \mu^{\pi_{\hat{\tau}}(x)}(x)$, and the expected value of R is

$$\begin{aligned}
\mathbb{E}[R] =& \mathbb{E}_X\big[\mathbb{E}[R|X]\big] \\
=& \mathbb{E}_X\big[1 - \mu^{\pi_{\hat{\tau}}(X)}(X)\big] \\
=& 1 - \mathbb{E}_X\big[\mathbb{E}_{Y|X}[Y^{\pi_{\hat{\tau}}(X)}]\big] \\
=& 1 - \mathbb{E}_X\big[Y^{\pi_{\hat{\tau}}(X)}\big] \\
=& 1 - \mathbb{E}_X\big[Y^1(X)|\pi_{\hat{\tau}}(X) = 1\big]\mathbb{P}(\pi_{\hat{\tau}}(X) = 1) \\
& - \mathbb{E}_X\big[Y^0(X)|\pi_{\hat{\tau}}(X) = 0\big]\mathbb{P}(\pi_{\hat{\tau}}(X) = 0)
\end{aligned}$$

2. *Case 2.* Assume $\hat{\mu}$ is correctly specified, i.e, $(\hat{\mu}^0, \hat{\mu}^1) = (\mu^0, \mu^1)$. Then the equation also simplifies in $\mathbb{E}[R|X = x] = 1 - \mu^{\pi_{\hat{\tau}}(X)}(x)$ and we may conclude likewise.

As such, $\widehat{\pi\textit{-risk}}_{DR}$ is robust to misspecification of either η or μ.

counter-factual outcome of sample (x_i, t_i, y_i) is approximated by the factual outcome of its closest sample with the opposite treatment assignment in the validation set. Formally,

$$\tilde{\tau}_i = (2t_i - 1)(y_i - \overline{(x_i, t_i, y_i)})$$
$$\overline{(x_i, t_i, y_i)} = y_{j^*},\ j^* = \underset{j \in \mathcal{V}\ s.t.\ t_j \neq t_i}{\operatorname{argmin}} \left\{ ||x_i - x_j||_2 \right\}$$

This evokes the notion of mirror twin as introduced in Eq. IV.1, where the identity function on $\mathcal{X}$ replaces mapping ϕ, and the search for nearest neighbors is restricted to $\mathcal{D}_\mathcal{V}$ (i.e., the held-out samples). Rolling and Yang (2014) advocates for a slight variation, resorting to the Mahalanobis distance instead of the Euclidean one. In both cases, the estimator entailed estimator writes

$$\widehat{\tau\text{-}risk}_{1NN}(\hat{\tau}) : \mathcal{D}_\mathcal{V} \mapsto \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y) \in \mathcal{D}_\mathcal{V}} \left(\hat{\tau}(x) - (2t - 1)(y - \overline{(x, t, y)}) \right)^2$$

Another plug-in estimate relies on the remark initiating the ***F-learner*** (Section III.3.1.5 and Gutierrez and Gérardy (2017)). The expected value of $(2T - 1)\rho^T(X)Y$ is equal to $\tau(X)$. The induces estimator is:

$$\widehat{\tau\text{-}risk}_{IPTW}(\hat{\tau}) : \mathcal{D}_\mathcal{V} \mapsto \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y) \in \mathcal{D}_\mathcal{V}} \left(\hat{\tau}(x) - (2t - 1)\hat{\rho}^t(x)y \right)^2$$

In a similar fashion, the observation $\tau(X) = \mathbb{E}\left[(2T - 1)\rho^{1-T}(X)\big(Y - m(X)\big)\right]$ at the origin of the ***U-learner*** (Doutreligne and Varoquaux (2023) and Section III.3.1.6) induces:

$$\widehat{\tau\text{-}risk}_{U}(\hat{\tau}) : \mathcal{D}_\mathcal{V} \mapsto \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y) \in \mathcal{D}_\mathcal{V}} \left(\hat{\tau}(x) - (2t - 1)\hat{\rho}^{1-t}(x)(y - \hat{m}(x)) \right)^2$$

Finally, based on the Augmented Inverse Probability Estimator and following the reasoning underlying ***DR-learners***, $\tau(x) = \mathbb{E}\left[\mu^1(X) - \mu^0(X) + (2T - 1)\rho^T(X)(Y - \mu^T(X))|X = x\right]$ induces the estimator:

$$\widehat{\tau\text{-}risk}_{DR}(\hat{\tau}) : \mathcal{D}_\mathcal{V} \mapsto \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y) \in \mathcal{D}_\mathcal{V}} \left(\hat{\tau}(x) - \left(\hat{\mu}^1(x) - \hat{\mu}^0(x) + (2t - 1)\hat{\rho}^t(x)\left(y - \hat{\mu}^t(x)\right) \right) \right)^2$$

As said in Section III.3.1.7, this estimator relies on a doubly robust estimate of τ. Saito and Yasui (2020) push it a step further by refining the estimation of μ with the specific focus that it should minimize the estimator variance.

Alternative strategies exist. Hassanpour and Greiner (2019a) consider the opportunity to use strong baseline models (for instance, *CFR* Section III.3.1.2) as plug-in estimates. However, the choice of an appropriate baseline among all possibilities may itself be seen as a hyper-parameter.

Alaa and Schaar (2019) interprets the shift from the true τ-*risk* to plug-in approximations as a shift over distributions. Informally and to build an intuition of the underlying principle,

$$PEHE_{\eta,\mu} \approx PEHE_{\hat{\eta},\hat{\mu}} + \frac{\partial PEHE_{\tilde{\eta},\tilde{\mu}}}{\partial(\tilde{\eta},\tilde{\mu})}\Big]_{(\hat{\eta},\hat{\mu})} \times \big((\hat{\eta},\hat{\mu}) - (\eta,\mu)\big)$$

Influence functions make it possible to estimate the "partial derivative" term, and the approximation error $(\hat{\eta},\hat{\mu}) - (\eta,\mu)$ depends only on factual quantities. As such, this approach corrects for part of the error that the shift has entailed.

Estimator	**Motivation**	**Expression:** $\frac{1}{\lvert\mathcal{D}_\mathcal{V}\rvert}\sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}}\bullet$
$\widehat{\mu\text{-}risk}$	factual error	$\big(y-\hat{\mu}^t(x)\big)^2$
$\widehat{\mu\text{-}risk}_{IPTW}$	same + *IPTW*	$\hat{\rho}^t(x)\big(y-\hat{\mu}^t(x)\big)^2$
$\widehat{\pi\text{-}risk}_{IPTW}$	*IPTW*	$1-\mathbb{1}[t=\pi_{\hat{\tau}}(x)]\hat{\rho}^t(x)y$
$\widehat{\pi\text{-}risk}_{DR}$	*DR*	$1-\hat{\mu}^{\pi_{\hat{\tau}}(x)}(x)-\mathbb{1}[t=\pi_{\hat{\tau}}(x)]\hat{\rho}^t(x)\big(y-\hat{\mu}^t(x)\big)$
$\widehat{R\text{-}risk}$	*R-learners*	$\Big(\hat{\tau}(x)\big(t-\hat{\eta}(x)\big)-\big(y-\hat{m}(x)\big)\Big)^2$
$\widehat{\tau\text{-}risk}_{naive}$	simplicity	$\big(\hat{\tau}(x)-(\hat{\mu}^1(x)-\hat{\mu}^0(x))\big)^2$
$\widehat{\tau\text{-}risk}_{1NNI}$	*1NNI*	$\big(\hat{\tau}(x)-(2t-1)(y-\overline{(x,t,y)})\big)^2$
$\widehat{\tau\text{-}risk}_{IPTW}$	*F-learners*	$\big(\hat{\tau}(x)-(2t-1)\hat{\rho}^t(x)y\big)^2$
$\widehat{\tau\text{-}risk}_{U}$	*U-learners*	$\big(\hat{\tau}(x)-(2t-1)\hat{\rho}^{1-t}(x)\big(y-\hat{m}(x)\big)\big)^2$
$\widehat{\tau\text{-}risk}_{DR}$	*DR-learners*	$\Big(\hat{\tau}(x)-\big((\hat{\mu}^1\text{-}\hat{\mu}^0)(x)+(2t-1)\hat{\rho}^t(x)\big(y-\hat{\mu}^t(x)\big)\big)\Big)^2$

Table VI.1: Feasible estimators.

VI.2.5 . Analysis

Consider a τ-*risk* based on a plug-in estimate $\tilde{\tau}$ that takes bounded values over $\mathcal{D}_\mathcal{V}$: $\forall(x,t,y)\in\mathcal{D}_\mathcal{V}, \tilde{\tau}(x)\in[-M,M]$. Denote by $\widehat{\tau\text{-}risk}_{\tilde{\tau}}(\hat{\tau})$ the estimator

$$\widehat{\tau\text{-}risk}_{\tilde{\tau}}(\hat{\tau}) : \mathcal{D}_\mathcal{V} \mapsto \frac{1}{\lvert\mathcal{D}_\mathcal{V}\rvert}\sum_{(x,t,y)\in\mathcal{D}_\mathcal{V}}\big(\hat{\tau}(x)-\tilde{\tau}(x)\big)^2$$

Then the Cauchy-Schwarz inequality implies[4]:

$$
\begin{aligned}
&|\widehat{\tau\text{-}risk}_{\tilde{\tau}}(\hat{\tau}) - \widehat{PEHE}_{D_\mathcal{V}}(\hat{\tau})| \\
&\qquad = \left| \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in D_\mathcal{V}} \big(\hat{\tau}(x) - \tilde{\tau}(x)\big)^2 - \big(\hat{\tau}(x) - \tau(x)\big)^2 \right| \\
&\qquad = \left| \frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in D_\mathcal{V}} \big(2\hat{\tau}(x) - \tilde{\tau}(x) - \tau(x)\big)\big(\tau(x) - \tilde{\tau}(x)\big) \right| \\
&\qquad \leq \sqrt{\frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in D_\mathcal{V}} \big(2\hat{\tau}(x) - \tilde{\tau}(x) - \tau(x)\big)^2} \sqrt{\frac{1}{|\mathcal{D}_\mathcal{V}|} \sum_{(x,t,y)\in D_\mathcal{V}} \big(\tau(x) - \tilde{\tau}(x)\big)^2} \\
&\qquad \leq \sqrt{\widehat{PEHE}_{D_\mathcal{V}}(2\hat{\tau} - \tilde{\tau})} \sqrt{\widehat{PEHE}_{D_\mathcal{V}}(\tilde{\tau})} \\
&\qquad \leq (3M + ||\tau||_\infty) \sqrt{\widehat{PEHE}_{D_\mathcal{V}}(\tilde{\tau})}
\end{aligned}
$$

Provided that $\tilde{\tau}$ is a good estimate of τ, then $\widehat{\tau\text{-}risk}_{\tilde{\tau}}$ is a good estimate of $\widehat{PEHE}_{D_\mathcal{V}}(\hat{\tau})$. When $\mathcal{D}_\mathcal{V}$ grows large, $\widehat{PEHE}_{D_\mathcal{V}}(\hat{\tau})$ converges to $PEHE(\hat{\tau})$, and indeed $\tilde{\tau}$ being a good estimate of τ implies that $\widehat{\tau\text{-}risk}_{\tilde{\tau}}(\hat{\tau})$ is a good estimate of $PEHE(\hat{\tau})$.

It is emphasized that the presented scores involve different limitations (Curth and Schaar, 2023). *μ-risks* and *π-risks* are restricted to *CATE* estimation approaches that model the potential outcomes μ^0, μ^1. Moreover, they assign the same score to two estimates as long as their factual predictions are the same, no matter how reasonable their counterfactual ones are.

R-risks and *π-risks* are more flexible since they accept any *CATE* estimate. Nevertheless, they are likely to favor models that are congruent with the structure of the risk: *R-risks* is prone to assigning lower scores to *R-learners*, *π-risks* to *S-* or *T-learners*. Besides, plug-in scores select the models that resemble the plugged-in term the most, even though the performance of said term is known to be poor.

Experimental validation is thus necessary to identify the most suited scores.

VI.3 . Proxy metric performance

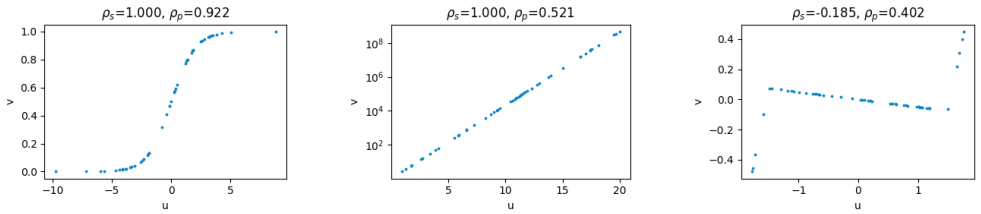
Given an infeasible performance metric $\mathcal{M}$, the question is to assess its proxy metric $\mathcal{S}$. This section focuses on assessing the quality of a proxy metric w.r.t. a set of C candidate models.

[4]Because the difference between the *PEHE* and its empirical evaluation matters here, the notation is not supercharged. $\widehat{PEHE}_{D_\mathcal{V}}(\hat{\tau}) = \frac{1}{D_\mathcal{V}} \sum_{(x,t,y)\in D_\mathcal{V}} \big(\hat{\tau}(x) - \tau(x)\big)^2$

VI.3.1 . Spearman correlation

Define the rank function of $u \in \mathbb{R}^C$ by the bijection of $[\![1, C]\!]$ verifying $r_u(i) < r_u(j) \Leftrightarrow u_i < u_j$ (assuming there are no ties, see adequate conventions in Dodge (2008) otherwise). The **Spearman correlation** coefficient $\rho_s(u, v)$ between two vectors u, v of $\mathbb{R}^C$ is then defined[5] as the Pearson correlation coefficient between the rank vectors $r(u) = (r_u(1), \dots, r_u(C))$ and $r(v) = (r_v(1), \dots, r_v(C))$:

$$\rho_s(u, v) = \frac{Cov\big(r(u), r(v)\big)}{\sigma\big(r(u)\big)\sigma\big(r(v)\big)} \tag{VI.1}$$



(a) Ordered points, "S" shape.

(b) Ordered points, logarithmic y-axis scaling.

(c) Opposite coefficient signs.

Figure VI.1: Different Spearman(ρ_s)/Pearson(ρ_p) correlation coefficients.

As such, the Spearman correlation coefficient takes values in $[-1, 1]$, value 1 meaning that u is monotonously increasing with v, values -1 meaning that u is monotonously decreasing with v.

VI.3.2 . Kendall rank correlation

The Kendall rank correlation coefficient (Kendall, 1938) (also referred to as *Kendall's τ coefficient*) is also a measure of the ordinal correlation of two vectors. For a couple of indices $i < j$, the pairs (u_i, u_j) and (v_i, v_j) are said to be concordant if either ($u_i < u_j$ *and* $v_i < v_j$) or ($u_i > u_j$ *and* $v_i > v_j$) holds (equivalently, if $(u_i - u_j)(v_i - v_j) > 0$), and discordant otherwise (Fig. VI.2). The Kendall rank correlation coefficient[6] is then defined as

$$\rho_k = \frac{\textit{(number of concordant pairs)} - \textit{(number of discordant pairs)}}{\textit{(number of pairs)}}$$

Similarly to the Spearman correlation coefficient, ρ_k takes values in from -1 (u is monotonously decreasing with v) to 1 (monotonously increasing).

[5] Still assuming that there are no ties, the expression simplifies into

$$\rho_s(u, v) = 1 - \frac{6}{C(C^2 - 1)} \sum_{i=1}^{C} \big(r_u(i) - r_v(i)\big)^2$$

[6] Denoted as ρ_k instead of τ, to avoid confusion with *CATE*

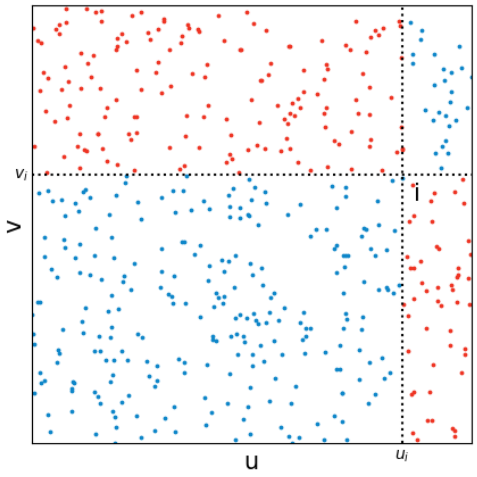


Figure VI.2: concordant and discordant pairs w.r.t (u_i, v_i).

VI.3.3 . Discounted Cumulative Gain

The Discounted Cumulative Gain (*DCG*) originates in query ranking (Järvelin and Kekäläinen, 2002). An algorithm is provided with a query, and from a list of options returns an ordered subset, ranked by decreasing relevance w.r.t. the query.

Consider an array of C items and a query whose possible answers are p-uples with distinct elements taken from this array. With respect to said query, each item i is associated with a relevance value u_i - the greater the more relevant. Let m be a mechanism which, provided with the query, returns the p-uple $(m(1), \ldots, m(p)) \in [\![1, C]\!]^p$, $m(i)$ all distinct.

1. the **Cumulative Gain** (CG) is defined as the sum of the relevance of the returned elements:

$$CG_m = \sum_{i=1}^{p} u_{m(i)}$$

2. the **Discounted Cumulative Gain** (*DCG*) discounts elements based on their selection position:

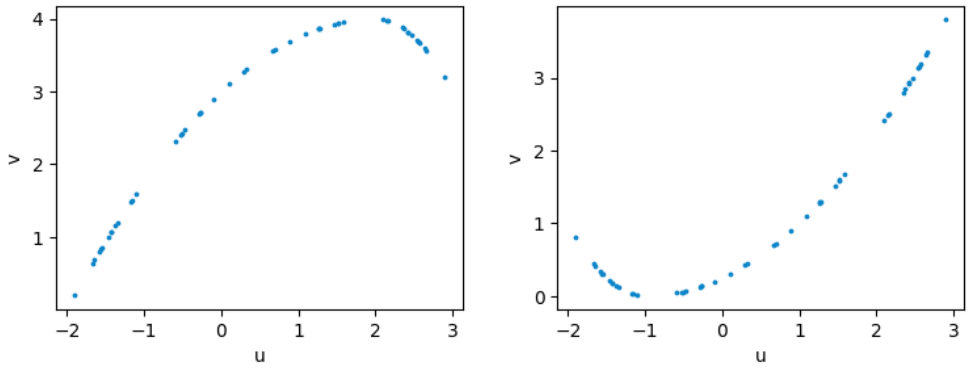
$$DCG_m = \sum_{i=1}^{p} \frac{u_{m(i)}}{\log(i+1)}$$

DCG thus is relevant to measure how the candidate model ranking based on the performance metric $\mathcal{M}$ matches the ranking based on the proxy metric S_{D_v}.

VI.3.4 . Discussion

As the goal is to select the best hyper-parameter setting, the correct ranking of non-optimal candidates does not matter much. As Spearman correlation and Kendall rank correlation attribute as much importance to the correct ranking of high-scored and low-scored models, the *DCG*-based metric (that specifically targets the top elements) is more relevant in this context.

In Fig. VI.3, two settings are displayed. Although they achieve the same Spearman and Kendall correlation coefficients, they differ in their *DCG*-based metric. Setting (b) (right) is preferable, for there the proxy metric (vector v) successfully identifies the sample with maximal relevance (vector u).



(a) Discounted Cumulative Gain: 8.40 (b) Discounted Cumulative Gain: 11.5

Figure VI.3: Two settings with same Spearman (.835), Pearson (.908), Kendall (.701) correlation coefficients, but different Discounted Cumulative Gains ($p = 5$) metrics.

VI.4 . Hyper-parameter adjustment

This section focuses on the ALRITE hyper-parameter selection, noting that pipelines P_0 and P_1 might require different hyper-parameter settings for the same reason as they involve different latent spaces.

The ALRITE hyper-parameters and their domain of variation are listed in Table VI.2, including the neural architecture of embeddings $\phi_{\mathcal{P}_0}$ and $\phi_{\mathcal{P}_1}$, the batch size and the regularization weights. oR_{pool} being an observational quantity, the selection is straightforward for *Jobs*. Quite the contrary, *IHDP* constitutes a challenging issue.

The hyper-parameter selection procedure proceeds as follows:

- Two sets of respectively $\ell_{\mathcal{P}_0}$ and $\ell_{\mathcal{P}_1}$ hyper-parameter settings are randomly sampled, defining a total of $\ell_{\mathcal{P}_0} \times \ell_{\mathcal{P}_1}$ candidate estimates

$$\hat{\tau}^{(\ell,\ell')} = (1-\hat{\eta})\hat{\tau}^{(\ell)}_{\mathcal{P}_0} + \hat{\eta}\ \hat{\tau}^{(\ell')}_{\mathcal{P}_1}$$

	IHDP	***Jobs***
regularization strength α	$\{10^{k/2}\}_{k=0}^{4}$	$\{10^{k/2}\}_{k=-5}^{1}$
	$\mathcal{P}_0$: 10^{0} $\mathcal{P}_1$: $10^{3/2}$	$\mathcal{P}_0$: $10^{1/2}$ $\mathcal{P}_1$: 10^{0}
reweighting importance β	$\{0\} \cup \{10^{k/2}\}_{k=-4}^{1}$	$\{0\} \cup \{10^{k/2}\}_{k=-4}^{1}$
	$\mathcal{P}_0$: $10^{-3/2}$ $\mathcal{P}_1$: 10^{0}	$\mathcal{P}_0$: $10^{1/2}$ $\mathcal{P}_1$: $10^{-1/2}$
embedding model layers	$[1, 2, 3, 4]$	$[1, 2, 3, 4]$
	$\mathcal{P}_0$: 4 $\mathcal{P}_1$: 4	$\mathcal{P}_0$: 1 $\mathcal{P}_1$: 1
outcome model layers	$[1, 2, 3, 4]$	$[1, 2, 3, 4]$
	$\mathcal{P}_0$: 3 $\mathcal{P}_1$: 4	$\mathcal{P}_0$: 2 $\mathcal{P}_1$: 3
embedding model width	$[20, 50, 100]$	$[20, 50, 100, 200]$
	$\mathcal{P}_0$: 20 $\mathcal{P}_1$: 20	$\mathcal{P}_0$: 50 $\mathcal{P}_1$: 200
outcome model width	$[20, 50, 100]$	$[20, 50, 100, 200]$
	$\mathcal{P}_0$: 50 $\mathcal{P}_1$: 100	$\mathcal{P}_0$: 20 $\mathcal{P}_1$: 200
batch size	$[50, 100, 200]$	$[50, 100, 200]$
	$\mathcal{P}_0$: 200 $\mathcal{P}_1$: 200	$\mathcal{P}_0$: 200 $\mathcal{P}_1$: 50

Table VI.2: Hyper-parameters ranges and selected values.

- The best candidate model according to the μ-*risk* proxy metric is retained.

Indeed the common practice since (Shalit et al., 2017) relies on using τ-$risk_{1NNI}$ as proxy metrics, though the 1-nearest neighbor estimator is well known for its poor performance in middle to high-dimensional settings. Furthermore, there exists evidence for 1NNI failure on synthetic problems Schuler et al. (2018).

Eventually, our choice of the μ-*risk* proxy metric is based on the following: i) it is the simplest option, as it doesn't rely on complex auxiliary functions $\hat{\eta}, \hat{\mu}^0, \hat{\mu}^1, \hat{m}$ (their definition and optimization are themselves a selection problem); ii) its relevance is experimentally confirmed (Schuler et al., 2018; Doutreligne and Varoquaux, 2023; Mahajan et al., 2023). Eventually, the selected hyper-parameters are listed in Table VI.2.

VI.5 . Scores comparison

While our selection is *not* based on counterfactual data (which would be an utmost bad practice), the actual values of the selected models according to the different proxy metrics can be computed on the *IHDP* dataset (Section VI.5.1),

score	ϵ_{ATE}	$\sqrt{PEHE}$	Spearman	DCG	Kendall
μ-risk	0.131	**0.431**	**0.969**	-2.016	**0.846**
μ-risk$_{IPTW}$	0.131	**0.431**	0.942	-2.014	0.798
π-risk$_{IPTW}$	0.139	0.672	0.366	-2.751	0.253
π-risk$_{DR}$	0.136	0.490	0.796	-2.239	0.598
R-risk	0.146	0.723	-0.064	-3.304	-0.049
τ-risk$_{naive}$	**0.125**	0.496	0.858	-2.196	0.668
τ-risk$_{1NNI}$	**0.125**	0.462	0.903	-2.070	0.729
τ-risk$_{IPTW}$	0.141	0.581	0.282	-3.281	0.199
τ-risk$_{U}$	0.183	1.320	-0.357	-5.924	-0.252
τ-risk$_{DR}$	0.131	**0.431**	0.960	**-1.985**	0.825

Table VI.3: For each score on *IHDP*, selected models test set performance metrics of the selected model (the lower the better), and scores correlation metrics (the higher the better).

then we delve into the comparison of base models for τ-*risks* (Section VI.5.2).

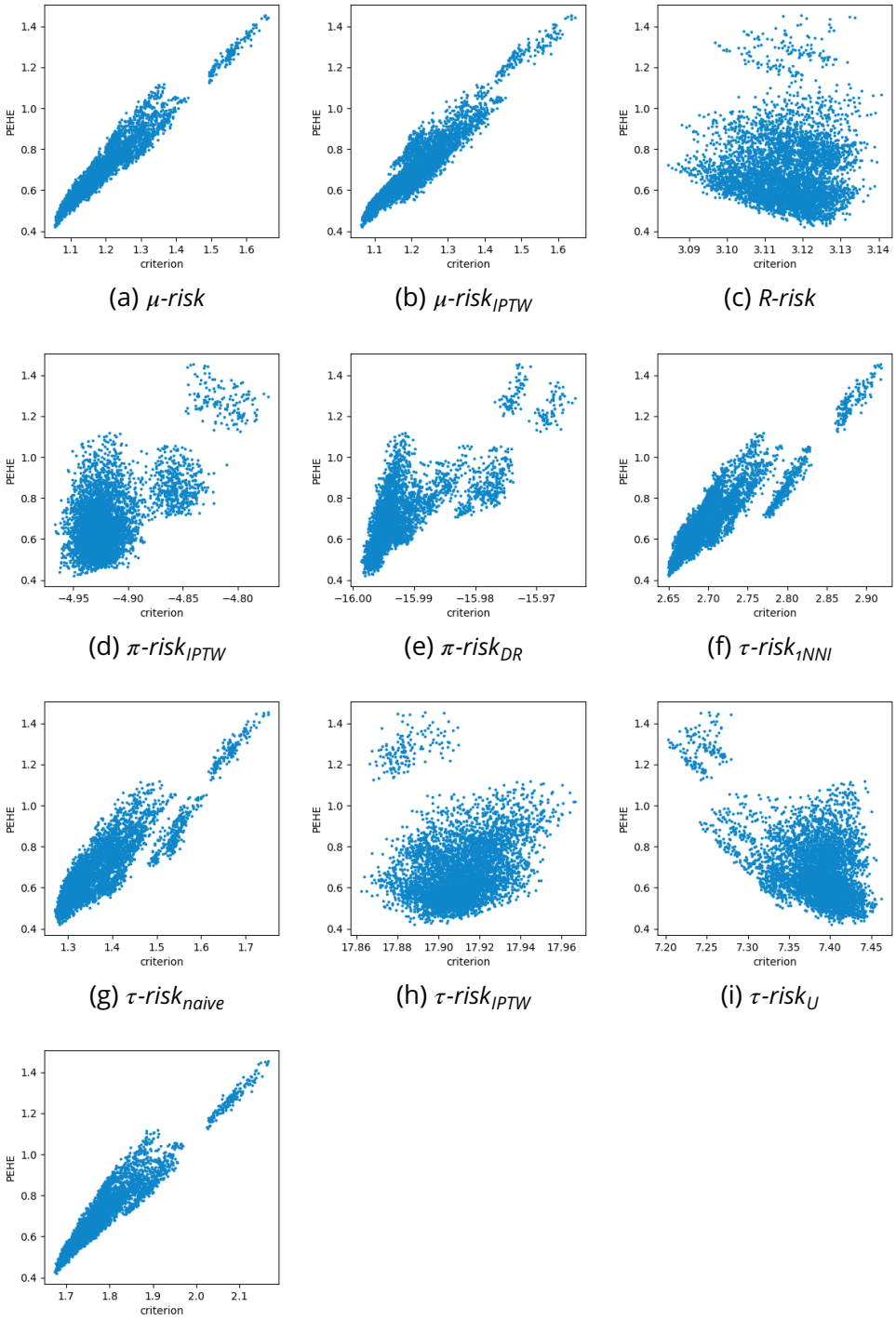
VI.5.1 . *A posteriori* scores comparison

We compare the multiple scores in the same context as that of Section VI.4. Here base learners $\hat{\mu}^0, \hat{\mu}^1, \hat{m}$ are enforced by Nu-Support Vector Regressors (*NuSVR*) models (Platt, 2000), since they achieve low high cross-validation factual prediction loss. Propensity estimators are either logistic regressions, k-nearest neighbors, or decision tree regressors. The estimators and their hyper-parameters are chosen through a cross-validation procedure.

All the candidate models in $C = \left(\hat{\tau}^{(\ell,\ell')}\right)_{(\ell,\ell')\in[\![1,\ell_{\mathcal{P}_0}]\!]\times[\![1,\ell_{\mathcal{P}_1}]\!]}$ are evaluated through the 10 scores. The resulting values are compared with the test set *PEHE*, and displayed in Fig. VI.4. Table VI.3 reports the test set *PEHE* and ϵ_{ATE} of the selected model $\operatorname{argmin}_{\hat{\tau}\in C} S_{\mathcal{D}_v}(\hat{\tau})$ for all scores $S_{\mathcal{D}_v}$. It also reports the Spearman correlation coefficient $\rho_s\left(\mathcal{M}(C), S_{\mathcal{D}_v}(C)\right)$, Kendall rank correlation coefficient $\rho_k\left(\mathcal{M}(C), S_{\mathcal{D}_v}(C)\right)$ and discounted cumulative gains[7] $\rho_{-\mathcal{M}(C):5}\left(-S_{\mathcal{D}_v}(C)\right)$ relative to each score.

These experiments validate *a posteriori* the relevance of μ-*risk* as a selection score, for it achieves the lowest *PEHE* and lowest *ATE*. The quality of the score is also confirmed by its high Spearman correlation coefficient, Kendall

[7] here the minus sign accounts for the fact that the lower, the more relevant



(a) μ-*risk* (b) μ-$risk_{IPTW}$ (c) *R-risk*

(d) π-$risk_{IPTW}$ (e) π-$risk_{DR}$ (f) τ-$risk_{1NNI}$

(g) τ-$risk_{naive}$ (h) τ-$risk_{IPTW}$ (i) τ-$risk_{U}$

(j) τ-$risk_{DR}$

Figure VI.4: *IHDP PEHE* on the test set vs score on validation set.

rank correlation coefficient, and discounted cumulative gain.

Most interestingly (and unexpectedly), τ-*risk*$_{1NNI}$ appears to be a reliable proxy with excellent selection performances. While μ-*risk*$_{IPTW}$, π-*risk*$_{DR}$, τ-*risk*$_{naive}$ and τ-*risk*$_{DR}$ seem worthy in retrospect, *R-risk*, π-*risk*$_{IPTW}$, τ-*risk*$_{IPTW}$ and τ-*risk*$_{U}$ achieve poor selection performance.

The same experiments are run with the ensemble approaches based on ALRITE, and their results are summarized in Table VI.4.

VI.5.2 . τ-*risks* based models

Among the most influential hyper-parameters is the choice of the model space. As detailed in Section VI.1, the scores based on τ-*risk*, π-*risk*$_{DR}$ and *R-risk* estimates all rely on the auxiliary functions $\hat{\mu}^0, \hat{\mu}^1, \hat{m}$, which may be learned using various regressors.

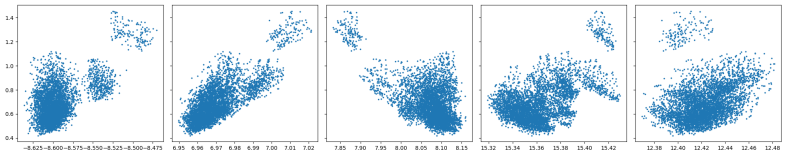
An experiment is conducted to evaluate the impact of the auxiliary function model definition on the performance of the scores. A range of estimators are trained to approximate the functions $\mu^0 : x \mapsto \mathbb{E}[Y|X = x, T = 0], \mu^1 : x \mapsto \mathbb{E}[Y|X = x, T = 1], m : x \mapsto \mathbb{E}[Y|X = x]$. Since the factual outcomes are available, these estimation problems belong to the supervised learning setting. The auxiliary functions hyper-parameter may be optimized through usual cross-validation procedures using mean squared prediction error.

The candidate models are all available in Scikit-Learn (Pedregosa et al., 2011), with default choice for all hyper-parameters, except for those detailed in Table VI.5. Considered architectures are multi-layer perception (*MLP*) (Rosenblatt, 1962), Gaussian process regressors (*GPR*) (Rasmussen and Williams, 2005), ridge regressors (*RR*) (Hoerl and Kennard, 1970), decision tree regressors (*DTR*) (Breiman, 1984), light gradient-boosting machines (*LightGBM*) (Ke et al., 2017), random forests (*RF*) (Breiman, 2001) and nu support vector regressors (*NuSVR*) (Platt, 2000).

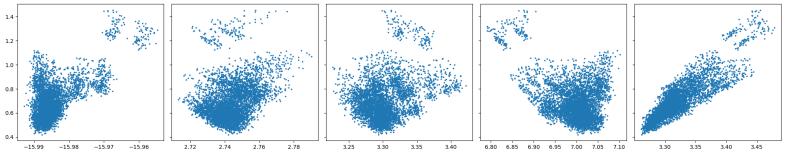
The results of the experiments are displayed in Fig. VI.5 as scatter plots, with selected model performances (*ATE*,*PEHE*) and scores (Spearman, *DCG*, Kendall) listed in Table VI.6. In concordance with Schuler et al. (2018); Doutreligne and Varoquaux (2023), it appears that the score choice mainly drives the performance of the selection procedure. τ-*risk*$_{DR}$ almost uniformly outperforms the other scores, followed by π-*risk*$_{DR}$. Indeed, doubly robust models are by design less sensitive to misspecification of the auxiliary functions, and as such less dependant on the choice of estimator. The estimator choice is of much lesser impact, albeit *NuSVR* performs well when plugged into τ-*risk*$_{naive}$.

		within-sample		*out-of-sample*	
score	model	$\sqrt{PEHE}$	ϵ_{ATE}	$\sqrt{PEHE}$	ϵ_{ATE}
μ-*risk*$_{simple}$	$\hat{\tau}^{softmax_{10^2}}$	**.400**	.095	.422	.106
	$\hat{\tau}^{top\text{-}4}$	.403	.100	**.412**	.110
μ-*risk*$_{IPTW}$	$\hat{\tau}^{softmax_{10^2}}$	**.400**	.096	.424	.108
	$\hat{\tau}^{top\text{-}4}$	.403	.100	**.412**	.110
π-*risk*$_{simple}$	$\hat{\tau}^{softmax_{8}}$	.468	.122	.460	.129
	$\hat{\tau}^{top\text{-}1}$	.425	.122	.429	.136
π-*risk*$_{DR}$	$\hat{\tau}^{softmax_{10^{3.5}}}$	.418	.103	.421	.116
	$\hat{\tau}^{top\text{-}3}$	.414	.105	.413	.114
R-risk	$\hat{\tau}^{softmax_{10^3}}$	.607	.113	.719	.145
	$\hat{\tau}^{top\text{-}50}$	.430	**.090**	.497	.108
τ-*risk*$_{simple}$	$\hat{\tau}^{softmax_{10^2}}$	.420	.098	.440	.106
	$\hat{\tau}^{top\text{-}3}$	.414	.105	.413	.114
τ-*risk*$_{1NNI}$	$\hat{\tau}^{softmax_{10^{2.5}}}$	.409	.100	.416	.109
	$\hat{\tau}^{top\text{-}10}$	.398	.093	.426	**.104**
τ-*risk*$_{IPTW}$	$\hat{\tau}^{softmax_{10^8}}$	.616	.119	.653	.135
	$\hat{\tau}^{top\text{-}3}$	.414	.105	.413	.114
τ-*risk*$_{U}$	$\hat{\tau}^{softmax_{10^2}}$	.806	.118	.902	.162
	$\hat{\tau}^{top\text{-}50}$	.430	**.090**	.497	.108
τ-*risk*$_{DR}$	$\hat{\tau}^{softmax_{10^{2.5}}}$	.418	.116	**.412**	.133
	$\hat{\tau}^{top\text{-}4}$	.403	.100	**.412**	.110

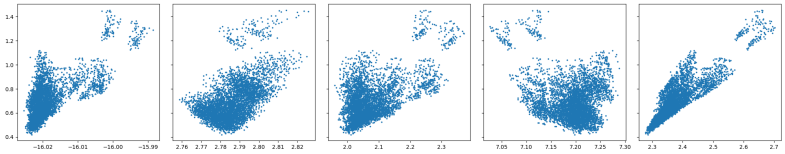
Table VI.4: *IHDP* performance of the ensemble models with the selected K and temperature λ, using various scores for hyper-parameter selection.



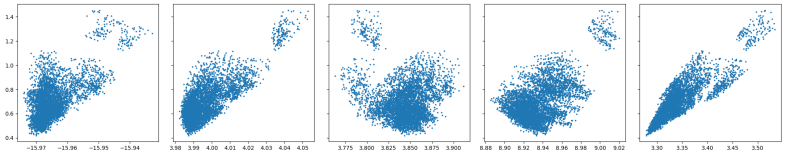
(a) Auxiliary functions estimators: Gaussian process regressors.



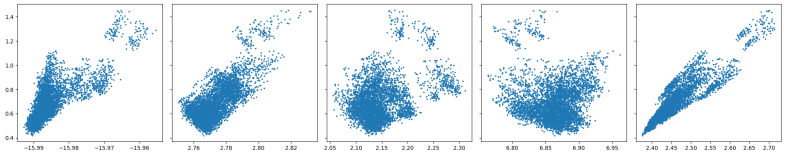
(b) Auxiliary functions estimators: multi-layer perceptrons.



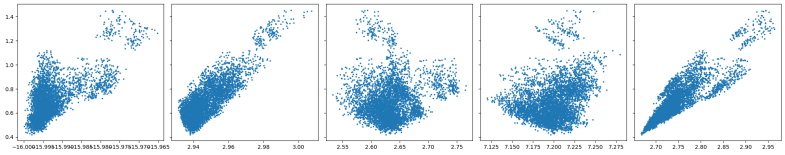
(c) Auxiliary functions estimators: ridge regressors.



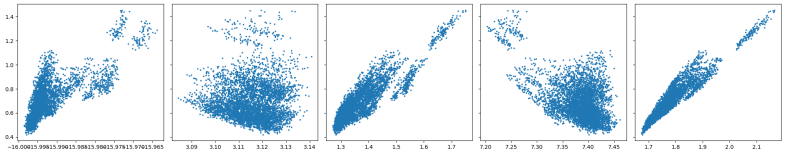
(d) Auxiliary functions estimators: decision tree regressors.



(e) Auxiliary functions estimators: light gradient-boosting machines.



(f) Auxiliary functions estimators: random forests.



(g) Auxiliary functions estimators: nu support vector regressors.

Figure VI.5: *PEHE* on the test set vs score on the validation set, for score ranging in (from left to right): π-*risk*$_{DR}$, *R-risk*, τ-*risk*$_{naive}$, τ-*risk*$_{U}$, τ-*risk*$_{DR}$

Model	Parameter grid
MLP	*'hidden_layer_sizes'* $\in \{[20, 20], [50, 50], [20, 20, 20], [50, 50, 50]\}$, *'max_iter'* $\in \{100\}$
GPR	*n.a*
RR	*'alpha'* $\in \{10^{k/2}\}_{k=-4}^{4}$
DTR	*'score'* $\in \{$*'squared_error'*, *'friedman_mse'*, *'absolute_error'*, *'poisson'*$\}$, *'max_depth'* $\in \{$*None*$, 2, 3, 4\}$, *'min_samples_split'* $\in \{2, 4\}$
LightGBM	*'num_leaves'* $\in \{5, 10, 20\}$, *'max_depth'* $\in \{-1, 2, 3, 4\}$
RF	*'score'* $\in \{$*'squared_error'*, *'friedman_mse'*, *'absolute_error'*$\}$, *'max_depth'* $\in \{$*None*$, 4\}$
NuSVR	*'C'* $\in \{10^{k/2}\}_{k=-4}^{4}$, *'kernel'* $\in \{$*'linear'*, *'poly'*, *'rbf'*, *'sigmoid'*$\}$ *'nu'* $\in \{.25, .5, .75\}$

Table VI.5: Auxiliary functions candidate architectures and associated parameter grids, *IHDP*.

Model	π-risk$_{DR}$	R-risk	τ-risk$_{naive}$	τ-risk$_{U}$	τ-risk$_{DR}$
GPR	ϵ_{ATE}: .139 $\sqrt{PEHE}$: .672 ρ_s: .368 *DCG*:-2.999 ρ_k: .257	ϵ_{ATE}: .137 $\sqrt{PEHE}$: **.535** ρ_s: .695 *DCG*:-2.574 ρ_k: .503	ϵ_{ATE}: .183 $\sqrt{PEHE}$: 1.320 ρ_s: -.370 *DCG*:-5.949 ρ_k: -.262	ϵ_{ATE}: .145 $\sqrt{PEHE}$: .716 ρ_s: .358 *DCG*:-3.249 ρ_k: .220	ϵ_{ATE}: .141 $\sqrt{PEHE}$: .581 ρ_s: .345 *DCG*:-2.895 ρ_k: .240
MLP	ϵ_{ATE}: .133 $\sqrt{PEHE}$: .656 ρ_s: .393 *DCG*:-3.127 ρ_k: .286	ϵ_{ATE}: .145 $\sqrt{PEHE}$: .710 ρ_s: .278 *DCG*:-3.344 ρ_k: .178	ϵ_{ATE}: .165 $\sqrt{PEHE}$: .982 ρ_s: .183 *DCG*:-3.951 ρ_k: .110	ϵ_{ATE}: .183 $\sqrt{PEHE}$: 1.320 ρ_s: -.149 *DCG*:-5.884 ρ_k: -.110	ϵ_{ATE}: .126 $\sqrt{PEHE}$: .464 ρ_s: .880 *DCG*:-2.119 ρ_k: .695
RR	ϵ_{ATE}: .133 $\sqrt{PEHE}$: .619 ρ_s: .602 *DCG*:-2.646 ρ_k: .434	ϵ_{ATE}: .145 $\sqrt{PEHE}$: .710 ρ_s: .526 *DCG*:-3.202 ρ_k: .345	ϵ_{ATE}: .134 $\sqrt{PEHE}$: .828 ρ_s: .382 *DCG*:-3.351 ρ_k: .277	ϵ_{ATE}: .183 $\sqrt{PEHE}$: 1.320 ρ_s: -.058 *DCG*:-5.895 ρ_k: -.045	ϵ_{ATE}: .146 $\sqrt{PEHE}$: **.429** ρ_s: .871 *DCG*:-2.029 ρ_k: .695
DTR	ϵ_{ATE}: .127 $\sqrt{PEHE}$: .733 ρ_s: .499 *DCG*:-3.202 ρ_k: .357	ϵ_{ATE}: .125 $\sqrt{PEHE}$: .620 ρ_s: .640 *DCG*:-2.718 ρ_k: .453	ϵ_{ATE}: .154 $\sqrt{PEHE}$: .910 ρ_s: .014 *DCG*:-4.114 ρ_k: .008	ϵ_{ATE}: .145 $\sqrt{PEHE}$: **.710** ρ_s: .346 *DCG*:-3.209 ρ_k: .217	ϵ_{ATE}: .146 $\sqrt{PEHE}$: **.429** ρ_s: .885 *DCG*:-2.009 ρ_k: .703
LightGBM	ϵ_{ATE}: .137 $\sqrt{PEHE}$: **.488** ρ_s: .760 *DCG*:-2.198 ρ_k: .563	ϵ_{ATE}: .143 $\sqrt{PEHE}$: .690 ρ_s: .653 *DCG*:-3.145 ρ_k: .450	ϵ_{ATE}: .145 $\sqrt{PEHE}$: .750 ρ_s: .219 *DCG*:-3.419 ρ_k: .141	ϵ_{ATE}: .167 $\sqrt{PEHE}$: .919 ρ_s: .034 *DCG*:-4.095 ρ_k: .020	ϵ_{ATE}: .146 $\sqrt{PEHE}$: **.429** ρ_s: .918 *DCG*:-1.981 ρ_k: .755
RF	ϵ_{ATE}: .137 $\sqrt{PEHE}$: **.488** ρ_s: .562 *DCG*:-2.349 ρ_k: .404	ϵ_{ATE}: .138 $\sqrt{PEHE}$: .643 ρ_s: .719 *DCG*:-2.973 ρ_k: .516	ϵ_{ATE}: .145 $\sqrt{PEHE}$: .750 ρ_s: .067 *DCG*:-3.372 ρ_k: .041	ϵ_{ATE}: .146 $\sqrt{PEHE}$: .723 ρ_s: .258 *DCG*:-3.225 ρ_k: .161	ϵ_{ATE}: .146 $\sqrt{PEHE}$: **.429** ρ_s: .894 *DCG*:-1.977 ρ_k: .719
NuSVR	ϵ_{ATE}: .136 $\sqrt{PEHE}$: .490 ρ_s: .796 *DCG*:-2.239 ρ_k: .598	ϵ_{ATE}: .146 $\sqrt{PEHE}$: .723 ρ_s: -.064 *DCG*:-3.304 ρ_k: -.049	ϵ_{ATE}: .125 $\sqrt{PEHE}$: **.496** ρ_s: .858 *DCG*:-2.196 ρ_k: .668	ϵ_{ATE}: .183 $\sqrt{PEHE}$: 1.320 ρ_s: -.357 *DCG*:-5.924 ρ_k: -.252	ϵ_{ATE}: .131 $\sqrt{PEHE}$: .431 ρ_s: .960 *DCG*:-1.985 ρ_k: .825

Table VI.6: Performance of the models selected through various scores using various auxiliary function estimators (ϵ_{ATE}, $\sqrt{PEHE}$), and related scores (Spearman, *DCG*, Kendall). **Top value**, column-wise. Top value, row-wise.

VII - Conclusion and Perspectives

VII.1 . Conclusion

The state of the art in *CATE* estimation, building upon representation learning and domain adaptation, has been seeking for a single latent space, enforcing a good positioning of the control and treatment distributions with respect to each other through symmetrical regularizations.

The main idea, at the core of the ALRITE approach, is that the sought properties can hardly be obtained with a single latent space: ensuring that control samples are close to treated ones, and that treated samples are close to control ones, actually defines two distinct goals.

This paradigm shift, from symmetrical to asymmetrical regularizations, is embodied in a new meta-learner hybridizing *T-learners* and *X-learners*, instantiated with ALRITE. The relevance of this approach has been grounded in theory as well as in practice. A notable merit is that the theoretical analysis relates to the practitioner's intuition: the underlying assumptions can be inspected and assessed on the observational data.

Before discussing the research perspectives, and investigating how this paradigm shift can irrigate the current causal approaches, let us discuss its limitations.

Firstly, the approach is ill-suited in the gold standard case where the control and the treatment distributions are identical – and/or when the method is assessed on a training set with same control and treatment distributions. As illustrated on the Jobs dataset, in this setting, ALRITE sacrifices its factual accuracy to no avail.

Secondly, a possible case of failure is if the optimization process mistakes the means and the ends, and specifically, finds a latent space such that samples that are well predicted have high exemplarity, as opposed to, such that samples have low insulation and all samples are well predicted. Complementary experiments, varying the hyper-parameters of the model, are needed to ensure that this failure case is not met in practice. The next step is indeed to run ALRITE on a wide range of real-life datasets, to better identify the limitations of the approach.

It is emphasized that another key contribution of the manuscript is the attention paid to the hyper-parameter selection, and the comprehensive methodology proposed to achieve it. As said, the AutoML problem, pervasive in supervised learning, is all the more acute in *CATE* estimation as the ground truth counterfactuals are missing in all but simulated problems.

VII.2 . Perspectives

The proposed architecture, hybridizing T-learners and X-learners, is highly versatile and opens quite a few perspectives regarding latent representations for causal inference. The first one consists in taking better advantage of the coupling of the pipelines and specializing them accordingly (Section VII.2.1). The second one aims at generalizing the key notion of mirror twin to multiple neighbors (Section VII.2.2). The third one swaps propensity for uncertainty in the combination of $\hat{\tau}_{\mathcal{P}_0}$ and $\hat{\tau}_{\mathcal{P}_1}$ (Section VII.2.3). The fourth one shows that the proposed architecture can be seamlessly extended to multi-valued treatment settings (Section VII.2.4). The last one considers potential synergies with disentangling approaches (Section VII.2.5).

VII.2.1 . Pipeline co-training

As shown in Stadie et al. (2018); Curth and Schaar (2021), the co-training of the X-learners (Künzel et al., 2019) is key to improving the accuracy of the trained models, taking full advantage of the components specialization.

Along this line, the accuracy of the causal effect estimate in ALRITE, defined as $\hat{\tau}(x) = (1 - \hat{\eta}(x))\hat{\tau}_{\mathcal{P}_0}(x) + \hat{\eta}(x)\hat{\tau}_{\mathcal{P}_1}(x)$, can be reconsidered to account for the fact that, e.g. for a high propensity sample (x, t, y), the accuracy of this estimate mainly depends on the accuracy of pipeline $\mathcal{P}_1$. Still, in ALRITE (x, t, y) is given similar importance in the training of pipelines $\mathcal{P}_0$ and $\mathcal{P}_1$, yielding the potential outcome models (Table IV.1) defined as:

$$\hat{\mu}^0 = (1 - \hat{\eta})h^0_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0} + \hat{\eta}h^0_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$$
$$\hat{\mu}^1 = (1 - \hat{\eta})h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0} + \hat{\eta}h^1_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}$$

It thus comes to jointly minimize the factual prediction error of the model, co-training pipelines $\mathcal{P}_0$ and $\mathcal{P}_1$:

$$\begin{aligned}\mathcal{L}_{h^1_{\mathcal{P}_0}, h^1_{\mathcal{P}_1}, h^0_{\mathcal{P}_0}, h^0_{\mathcal{P}_1}, \phi_{\mathcal{P}_0}, \phi_{\mathcal{P}_1}} &= \frac{1}{|D|} \sum_{(x,t,y)\in D} ||\hat{\mu}^t(x) - y||^2 \\ &= \frac{1}{|D|} \sum_{(x,t,y)\in D} ||(1 - \hat{\eta}(x))h^t_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x) + \hat{\eta}(x)h^t_{\mathcal{P}_1} \circ \phi_{\mathcal{P}_1}(x) - y||^2\end{aligned}$$

This compound training will expectedly result in a better specialization of the pipelines.

VII.2.2 . Generalization from 1-nearest neighbor to K-nearest neighbors

Although insulation has been defined with respect to the one mirror twin, counter-factual estimation might benefit from more abundant neighbors with the opposite treatment assignment in latent space.

Consider a control sample $(x, t = 0, y)$. The training protocol of pipeline $\mathcal{P}_0$ is such that the distance $d_{\mathcal{P}_0}$ to the nearest treated sample is minimized. However, there is no guarantee that the distance to the second nearest treated sample is also small. Indeed, by definition the insulation of a sample $(x, t = 0, y)$ is defined as the distance to its one nearest treated neighbor. More treated samples in the neighborhood of $(x, t = 0, y)$ could improve its counter-factual estimation.

Following this intuition, *insulation* could be upgraded to take into account K neighbors[1]:

$$\begin{cases} \phi(i) & = \text{argsort}_{t_j=1-t_i}\{d_\phi(x_i, x_j)\} \\ \textit{insulation}_\phi(i) & = \frac{1}{K}\sum_{k=1}^{K} d_\phi(x_i, x_{\phi(i)_k}) \\ \textit{exemplarity}_\phi(i) & = \frac{1}{K}|\{j \in [\![1, n]\!] \text{ s.t. } \phi(j)_k = i,\ 1 \leq k \leq K\ \}| \end{cases}$$

As such ALRITE is a special case where K is set to 1. The issue is that as K grows large, the K-th nearest neighbor with the opposite treatment assignment is less likely to provide a reasonable estimate of the counter-factual income. A potential solution could consist in using fixed weights $w_1 \geq \cdots \geq w_K$ summing to 1:

$$\begin{cases} \phi(i) & = \text{argsort}_{t_j=1-t_i}\{d_\phi(x_i, x_j)\} \\ \textit{insulation}_\phi(i) & = \sum_{k=1}^{K} w_k d_\phi(x_i, x_{\phi(i)_k}) \\ \textit{exemplarity}_\phi(i) & = \sum_{t_j=1-t_i} \sum_{k=1}^{K} w_k \mathbb{1}[\phi(j)_k=i] \end{cases}$$

VII.2.3 . Ensemble ALRITE: accounting for uncertainty

As emphasized by Zhang et al. (2020) (Section IV.2.1), minimizing the uncertainty in the estimation of counter-factual outcomes is of utmost importance. The mainstream combination of $\hat{\tau}_{\mathcal{P}_0}$ and $\hat{\tau}_{\mathcal{P}_1}$:

$$\hat{\tau} = (1 - \hat{\eta})\hat{\tau}_{\mathcal{P}_0} + \hat{\eta}\hat{\tau}_{\mathcal{P}_1}$$

with $\hat{\eta}$ the estimated propensity accounts for the fact that $\hat{\tau}_{\mathcal{P}_0}$ is assumed to be more precise for control samples (resp. $\hat{\tau}_{\mathcal{P}_1}$ for treated samples). Along this line, Zhang et al. (2020) proposes to learn potential outcomes implemented as Gaussian processes, thus coming with an estimate of their confidence.

This remark opens two research perspectives. One relies on the estimation of the neural networks uncertainty in ALRITE, e.g. following Gawlikowski et al. (2023).

Another, more straightforward estimation of the potential outcome uncertainty is based on the ensemble variant of ALRITE. Formally, the weights in the λ-softmax ensemble (Section IV.4), currently based on the factual error of the models, can be replaced with the variance of the counter-factual models w.r.t. the model barycenter.

[1] here argsort is the operator that returns the arguments in the order that would sort the array. For instance, $\text{argsort}_{x\in[-1,0,2]}(x^2) = (0, -1, 2)$

VII.2.4 . Extension to the multi-level treatment setting

The proposed approach can be extended to the multi-level treatment case, with T varying in $\{0, 1, \dots U\}$ where $T = 0$ corresponds again to the control group. In this multi-valued treatment setting Acharki et al. (2023), the quantities of interest are $(\tau^u)_{u\in\mathcal{T}}$, with performance indicator *mPEHE*, defined by

$$\tau^u : x \in \mathcal{X} \mapsto \mathbb{E}[Y^u - Y^0 | X = x]$$
$$mPEHE : (\hat{\tau}^u)_{u\in\mathcal{T}} \mapsto \sqrt{\frac{1}{|\mathcal{T}|}\sum_{u\in\mathcal{T}} \mathbb{E}\left[\left(\hat{\tau}^u(X) - \tau^u(X)\right)^2\right]}$$

The approach is extended by considering $U+1$ pipelines $\mathcal{P}_0, \mathcal{P}_1, \dots, \mathcal{P}_U$. Since treatment level 0 has a special role, pipeline $\mathcal{P}_0$ will be considered separately.

Let $(x, t = u, y)$ be a treated sample from the training set ($u \neq 0$). Following the same reasoning as in the binary treatment case, the correct estimation of the counter-factual outcome y^0 implies that $(\phi_u(x), t = u, y)$ has neighbors with treatment assignment 0. The quantities of insulation and exemplarity (Section IV.2.1) are naturally extended as:

$$\begin{cases} \phi^t(i) & = \text{argmin}_{j\in[\![1,n]\!] \text{ s.t. } t_j=t}\{d_\phi(x_i, x_j)\} \\ \textit{insulation}_{\phi^t}(i) & = d_\phi(x_i, \phi^t(i)) \\ \textit{exemplarity}^t_\phi(i) & = |\{j \in [\![1,n]\!] \text{ s.t. } t_j = t, \phi^{t_i}(j) = i\}| \end{cases}$$

And the training losses of the pipelines read:

$$\begin{aligned}\mathcal{L}_{\mathcal{P}_0} =& \frac{1}{n_0}\sum_{t_i} \textit{error}_{\mathcal{P}_0}(x_i, 0, y_i) + \frac{\alpha_{\mathcal{P}_0}}{n_0}\sum_{t_i=0}\sum_{u=1}^{U} \textit{insulation}_{\phi^u_{\mathcal{P}_0}}(i)^2 + \gamma_{\mathcal{P}_0}\,\Omega(\mathcal{P}_0) \\ &+ \frac{1}{(n-n_0)+\beta_{\mathcal{P}_0} n_0}\sum_{t_i\neq 0}\left(1 + \beta_{\mathcal{P}_0}\textit{exemplarity}^0_{\phi_{\mathcal{P}_0}}(i))\right) \times \textit{error}_{\mathcal{P}_0}(x_i, t_i, y_i)\end{aligned}$$

and for $u \neq 0$,

$$\begin{aligned}\mathcal{L}_{\mathcal{P}_u} =& \frac{1}{n_u}\sum_{t_i=u} \textit{error}_{\mathcal{P}_u}(x_i, u, y_i) + \frac{\alpha_{\mathcal{P}_u}}{n_u}\sum_{t_i=u} \textit{insulation}_{\phi^0_{\mathcal{P}_u}}(x_i)^2 \\ &+ \frac{1}{n_0+\beta_{\mathcal{P}_u} n_u}\sum_{t_i=0}\left(1 + \beta_{\mathcal{P}_u}\textit{exemplarity}^u_{\phi_{\mathcal{P}_u}}(i)\right) \times \textit{error}_{\mathcal{P}_u}(x_i, 0, y_i) \\ &+ \gamma_{\mathcal{P}_u}\,\Omega(\mathcal{P}_u)\end{aligned}$$

Letting $\eta^u, u \in \mathcal{T} \cup \{0\}$ be the generalized propensity $\eta^u(x) = \mathbb{P}(T = u|X = x)$, with $\hat{\eta}$ an estimate of η, estimate $\hat{\tau}^u$ can finally[2] be defined as

$$\hat{\tau}^u : x \in \mathcal{X} \mapsto \frac{\hat{\eta}^0(x)}{\hat{\eta}^0(x) + \hat{\eta}^u(x)}\hat{\tau}^u_{\mathcal{P}_0}(x) + \frac{\hat{\eta}^u(x)}{\hat{\eta}^0(x) + \hat{\eta}^u(x)}\hat{\tau}^u_{\mathcal{P}_u}(x)$$

[2]Noting that $\mathbb{P}(T = 0|T \in \{0, u\}, X = x) = \frac{\mathbb{P}(T=0|X=x)}{\mathbb{P}(T=0|X=x)+\mathbb{P}(T=u|X=x)}$

A potential pitfall of this approach is that the task of pipeline $\mathcal{P}_0$ is much harder than that of the treated pipelines[3].

VII.2.5 . Latent space disentanglement

In the wake of Kuang et al. (2017)'s *D*^{2}*VD* and Cheng et al. (2022b)'s *MIM-DRCFR*, splitting the latent space between instrumental (I, causes of T only), confounding (C, causes of both T and Y) and adjustment (A, causes of Y only) variables. This innovation has driven the recent progress of *CATE* estimation and constitutes as such a promising direction for future research.

ALRITE would easily merge with this design: by dividing the adjustment latent space into two subspaces $A_{\mathcal{P}_0}$ and $A_{\mathcal{P}_1}$, both approaches might synergize, and provide more accurate predictions.

[3]Assuming that the total number of control and treated samples are roughly equal, ensuring the proximity of samples with each treatment assignment $u \neq 0$ in the neighboring of $(\phi_{\mathcal{P}_0}(x), 0, y)$ is strenuous. To account for the likely lower accuracy of $\hat{\tau}_{\mathcal{P}_0}$, one may also consider to introduce an extra hyper-parameter $\lambda \in [0, .5]$, defining the updated $\hat{\tau}^u$ estimate by

$$\hat{\tau}^u : x \in \mathcal{X} \mapsto \frac{\lambda\hat{\eta}^0(x)}{\lambda\hat{\eta}^0(x) + (1-\lambda)\hat{\eta}^u(x)}\hat{\tau}^u_{\mathcal{P}_0}(x) + \frac{(1-\lambda)\hat{\eta}^u(x)}{\lambda\hat{\eta}^0(x) + (1-\lambda)\hat{\eta}^u(x)}\hat{\tau}^u_{\mathcal{P}_u}(x)$$

Acknowledgements

My uttermost gratitude goes to DR. Michèle Sebag. At all times in these countless hours of discussion, debating and late-night rewriting have I felt supported; to say that without her my doctoral thesis would not have been possible is both self-evident, and an understatement.

I would like to thank the jury members Pr. Marianne Clausel, DR. Hervé Isambert, DR. Julie Josse, Pr. Éric Gaussier and DR. Jean-Pierre Nadal.

I'd also like to thank my co-advisor Assoc.Pr. Philippe Caillou, and all the TAU members. This team has a mood that I have enjoyed from my internship to my PhD defense. Special thanks go to Adrien, Eléonore, Nilo, and Roman.

I wish to seize this occasion to express my gratitude to two of the teachers who have influenced me the most. Pascale Montupet has shaped my prose so hard I can still feel her assessing my writing fourteen years later. In retrospect, I am confident that Marc Abehsira is the one best teacher[4].

I am grateful to my friends, whose patience and supportiveness are uplifting (they're the best). I also acknowledge that support sometimes comes from the most unexpected places. The *Ballet de l'Opéra de Paris*, *Chez Nicos* and Pr. Nathanaël Enriquez have at some harsh points unknowingly brightened what were so far painful days, and I took note there would be a place for them in this section.

Finally, I want to express my deepest gratitude to my family, notably my sister, and especially my grandparents and parents: for very different reasons, to say that without them my doctoral thesis would not have been possible is both an understatement, and self-evident.

[4] for reasons this footnote is too narrow to contain

List of Figures

List of Tables

A - Résumé étendu en français

Quelques éloignées puissent-elles paraitre, les notions de prise de décision algorithmique, d'évaluation des politiques publiques ou de personnalisation des soins médicaux reposent sur une même question fondamentale : que se serait-il passé, que se passerait-il si la décision était autre ? Parce que l'apprentissage causal fonde par essence des raisonnements contrefactuels sur les données disponibles, il constitue le cadre théorique et pratique idoine de ces problématiques.

Depuis l'introduction de méthodes fondées sur des réseaux de neurones, les progrès en inférence causale ont été portés principalement par le raffinement de l'équilibrage entre les représentations apprises des individus contrôles, et traités. Prenant constat des limites de cette approche, nous opérons un changement de paradigme. Des contraintes asymétriques dans l'espace des représentations permettent, au prix de la dégradation de la modélisation factuelle d'une population, l'amélioration de la modélisation contrefactuelle de l'autre. La combinaison d'un modèle favorable à la population traitée avec son pendant relatif à la population contrôle cumule leurs avantages, sans leurs inconvénients.

Nous débutons ici notre exposé par une introduction qui pose le contexte de l'inférence causale et en illustre le besoin. Nous détaillons alors le cadre théorique du champ d'étude : le modèle caussal de Neyman-Rubin, ou modèle à résultats potentiels. Il est nécessaire d'en préciser les quantités d'intérêts, les principales hypothèses et leurs limites, un résultat clé d'identifiabilité, ainsi que des métriques qui jugent de la performance d'un modèle.

Nous poursuivons l'exposé par un état de l'art de l'apprentissage causal. Après avoir rappelé pour mémoire les principaux résultats concernant la découverte causale, nous précisons les méthodes et résulats spécifiques à l'inférence causale. Nous justifions de notre intérêt pour l'estimation de l'effet conditionnel moyen des traitements, plutôt que du seul effet moyen des traitements. C'est l'occasion de présenter la galerie des architectures existantes, à laquelle notre proposition s'ajoutera.

Notre nouvelle architecture est incarnée par ALRITE, un modèle que nous motivons. Les approches actuelles sont fondées sur des méthodes de régularisation qui ne traitent pas directement l'objectif poursuivi. En pratique, nous avançons qu'il est nécessaire à l'estimation des résultats contrefactuels que chaque point du groupe contrôle dans l'espace latent soit proche d'un point du groupe traité, et que chaque point du groupe traité soit proche d'un point du groupe contrôle. De ces observation nous déduisons les principes sous-jacents de ALRITE. En premier lieu, une *pipeline* sera destinée à l'estimation

de représentations contrefactuelles précises pour le groupe contrôle, et une autre sera destinée à l'estimation de réprésentations contrefactuelles précises pour le groupe traité. Les estimations de chacune de ces *pipelines* seront combinées selon le score de propensité, c'est à dire la probabilité inférée qu'un point appartienne au groupe traité. De ces principes il est alors possible de dégager l'architecture de ALRITE, qui s'inscrit alors dans la galerie évoquée lors de la description de l'état de l'art. Nous proposons également deux variantes de cette architecture, légèrement plus complexes, mais dont la performance sera mise en valeur par la suite. Nous démontrons également la pertinence de l'approche sur un plan théorique, portant un soin tout particulier à l'étude des limites de validité de ses hypothèses.

Nous montrons ensuite la pertinence de l'approche en la soumettant à des expériences pratiques, sur des bases de données de référence dans l'importance causale. Issues d'expériences réelles, *IHDP* et *Jobs* tiennent lieu de références en la matière, et permettent la comparaison avec les approches concurrentes telles que relevées dans la présentation de l'état de l'art. Les paramètres expérimentaux sont précisés dans leur détail, de façon à s'assurer de la reproducibilité de l'exercice. Les métrique de performance mettent en lumière la pertinence de l'approche de ALRITE, et en valident les principes sous-jacents.

Un soin tout particulier est porté à la sélection rigoureuse des hyper-paramètres du modèle, tâche réputée délicate dans le domaine de l'inférence causale. Nous motivons un choix de métrique *proxy*, qui permet la sélection des hyper-paramètres les plus adéquats dans l'espace exploré. Une vérification rétrospective des résultats confirme ici encore le bien-fondé de l'approche.

Nous proposons enfin une conclusion à notre exposé, présentant quelques perspectives à même de prolonger nos travaux. Si nous avons détaillé dans ce manuscrit deux variantes à notre proposition, l'approche proposée est suffisamment versatile pour s'adapter à différents cas de figure, au-delà du cadre théorique tel que nous l'avions posé.

B - Acronyms

Generalities

AIPW : Augmented Inverse Propensity Weights (II.6)
DAG : Directed Acyclic Graph (III.1.2.4)
IPTW : Inverse Probability of Treatment Weighting (III.2.2.2)
ML : Machine Learning (I)
ROC : Receiver Operating Characteristic curve (V.1.2.3)

Functions

ELU : Exponential Linear Unit (V.6)
ReLU : Rectified Linear Unit (V.2)
UMAP : Uniform Manifold Approximation and Projection (V.1.2.3)
MMD : Maximum Mean Discrepancy (II.3.2)

Regressors

1NNI : One Nearest Neighbor Imputation (VI.2.4)
DTR : Decision Tree Regressor (VI.5.2)
GPR : Gaussian Process Regressor (VI.5.2)
LightGBM : Light Gradient Boosting Machine (VI.5.2)
MLP : Multi-Layer Perceptron (VI.5.2)
RF : Random Forest (VI.5.2)
DTR : Ridge Regressor (VI.5.2)

Learning frameworks

DML : Double Machine Learning (III.2.3)
DR : Doubly Robust (III.2.3)
FCM : Functional Causal Model (III.1.1)
GAN : Generative Adversarial Networks (III.3.1.2)
IV : Instrumental Variables (III.2.1)
NICA : Non-linear Independent Component Analysis (III.3.2)
VAE : Variational AutoEncoder (III.3.2)

Datasets

IHDP : Infant Health and Development Program (III.3.3)
NSWD : National Supported Work Demonstration (V.1.2)
PSID : Panel Study of Income Dynamics (V.1.2)

Metrics, assumptions

ATC : Average Treatment Effect on the Control (III.2.2.2)
AUC : Area Under the receiver operating characteristic (ROC) Curve (V.1.2.3)
ATE : Average Treatment Effect (II.1)
ATO : Average Treatment Effect on Overlap (III.2.2.2)
ATT : Average Treatment Effect on the Treated (II.3)
CATE : Conditional Average Treatment Effect (II.2)
CATT : Conditional Average Treatment Effect on the Treated (II.2)
CG : Cumulative Gain (VI.3)
DCG : Discounted Cumulative Gain (VI.3)
ITE : Individual Treatment Effect (II.2)
PEHE : Precision in Estimation of Heterogeneous Effect (II.5.1)
SATE : Sample Average Treatment Effect (II.2)
SATT : Sample Average Treatment Effect on the Treated (II.2)
SUTVA : Stable Unit Treatment Value Assumption (II.3.3)
ϵ_{ATE} : absolute error in Average Treatment Effect estimation (II.5.1)

CATE estimation methods

ABCEI : Adversarial Balancing-based representation learning for Causal Effect Inference (III.3.1.2)
Alrite : Asymmetrical Latent Regularization for Individual Treatment Effect modeling (IV)
BNN : Balancing Neural Network (III.3.1.1)
BWCFR : Balancing Weights for CounterFactual Regression (III.3.1.2)
CBRE : Cycle-Balanced REpresentation learning for the counterfactual inference (III.3.1.2)
CEVAE : Causal Effects Variational Auto-Encoder (III.3.1.2)
CF : Causal Forests (III.3.1.1)
CFR : CounterFactual Regression (III.3.1.2)
CFR-ISW : CounterFactual Regression Importance Sampling Weights (III.3.1.2)
DeR-CFR : Decomposed Representations for CounterFactual Regression (III.3.1.2)
DKLITE : Deep Kernel Learning for ITE (III.3.1.2)

DR-CFR : Disentangled Representations for CounterFactual Regression (III.3.1.2)
D^{2}VD : Data-Driven Variable Decomposition (III.3.1.2)
GANITE : Generative Adversarial Networks for inference of Individual Treatment Effects (III.3.1.2)
MIM-DRCFR : Disentangled Representations for Counterfactual Regression via Mutual Information Minimization (III.3.1.2)
MitNet : Mutual Information Treatment Network (III.3.1.2)
N-D^{2}VD : Data-Driven Variable Decomposition (III.3.1.2)
NSGP : Non-Stationary Gaussian Process (III.3.1.2)
SITE : Similarity-preserved Individual Treatment Effect (III.3.1.2)
SNet : Sharing information NETwork (III.3.1.2)
TARNet : Treatment-Agnostic Representation Network (III.3.1.2)
TEDVAE : Treatment Effect by Disentangled Variational Auto-Encoder (III.3.1.2)

C - Illustration: extension of the Simpson paradox

Consider the extension of the Simpson paradox data (Table C.1, already discussed in Chapter I), where the treated group is assigned to percutaneous nephrolithotomy ($T = 1$) while the control group undergoes open surgery ($T = 0$). For convenience, we refer to smaller stones as s, larger ones as S, small hospitals by h, and big ones by H.

		open surgery	percutaneous nephrolithotomy
Smaller stones (s)	Small hospital (h)	80% (8/10)	**83%** (173/208)
	Big hospital (H)	95% (73/77)	**98%** (61/62)
Larger stones (S)	Small hospital (h)	57% (44/77)	**63%** (41/65)
	Big hospital (H)	80% (148/186)	**93%** (14/15)

Table C.1: Imaginary extension of the Simpson paradox data.

Assume now that medical experts guarantee that the stone and hospital size ensure **conditional exchangeability**. All combinations of stone and hospital sizes include control and treated patients, ensuring that **positivity** holds. Finally, we may suppose that the success of a given medical intervention does not interfere with that of other interventions, verifying ***SUTVA***.

The sample *CATE* given that the stone is big and the hospital small is then equal to $\mathit{CATE}(S, h) = \frac{41}{65} - \frac{44}{77} \approx 6\%$. The law of total probability lets us compute

the sample *ATE*

$$\begin{aligned}
ATE &= \mathbb{E}[CATE(X)] \\
&= \sum_{x_1\in\{s,S\},x_2\in\{h,H\}} \mathbb{E}[Y^1 - Y^0|(x_1,x_2)]\mathbb{P}(x_1,x_2) \\
&= \sum_{x_1\in\{s,S\},x_2\in\{h,H\}} CATE(x_1,x_2)\mathbb{P}(x_1,x_2) \\
&= (\frac{173}{208} - \frac{8}{10}) \times \frac{208+10}{700} + (\frac{61}{62} - \frac{73}{77}) \times \frac{62+77}{700} \\
&+ (\frac{41}{65} - \frac{44}{77}) \times \frac{65+77}{700} + (\frac{14}{15} - \frac{148}{186}) \times \frac{15+148}{700} \\
&\approx 7\%
\end{aligned}$$

For the reasons we have detailed, this quantity differs from the first erroneous guess $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] = \frac{289}{350} - \frac{273}{350} \approx 5\%$ and the second erroneous guess

$$\begin{aligned}
&\sum_{x_1\in\{s,S\}} \big(\mathbb{E}[Y|T=1,x_1] - \mathbb{E}[Y|T=0,x_1]\big)\mathbb{P}(x_1) \\
&\qquad = (\frac{234}{270} - \frac{81}{87}) \times \frac{270+87}{700} + (\frac{55}{80} - \frac{192}{263}) \times \frac{80+263}{700} \\
&\qquad \approx -5\%
\end{aligned}$$

Finally, the sample *ATT* is equal to

$$\begin{aligned}
ATT &= \mathbb{E}[CATE(X)|T=1] \\
&= \sum_{x_1\in\{s,S\},x_2\in\{h,H\}} \mathbb{E}[Y^1 - Y^0|T=1,(x_1,x_2)]\mathbb{P}(x_1,x_2|T=1) \\
&= \sum_{x_1\in\{s,S\},x_2\in\{h,H\}} CATE(x_1,x_2)\mathbb{P}(x_1,x_2) \\
&= (\frac{173}{208} - \frac{8}{10}) \times \frac{208}{350} + (\frac{61}{62} - \frac{73}{77}) \times \frac{62}{350} \\
&+ (\frac{41}{65} - \frac{44}{77}) \times \frac{65}{350} + (\frac{14}{15} - \frac{148}{186}) \times \frac{15}{350} \\
&\approx 4\%
\end{aligned}$$

meaning that the treatment is, on average, less successful for the treated population than for the control one.

D - Influence of the data generation process

In order to illustrate how different meta-learners suit different data generation processes, let us consider a synthetic problem with two different settings.

For the sake of simplicity, T is randomly assigned as in randomized control studies, and no noise is added to the outcome Y. For each setting, 100 training data points are sampled from:

$$\textit{Setting A.} \begin{cases} \mu^0 : x & \mapsto \sin\left(\frac{2\pi x}{.7}\right)^{1/3} \\ \mu^1 : x & \mapsto \sin\left(\frac{2\pi x}{.9}\right)^{1/3} \\ X & \sim \textit{Unif}([-6,6]) \\ T & \sim \textit{Ber}(1/2) \\ Y & \sim (1-T)\mu^0(X) + T\mu^1(X) \end{cases}$$

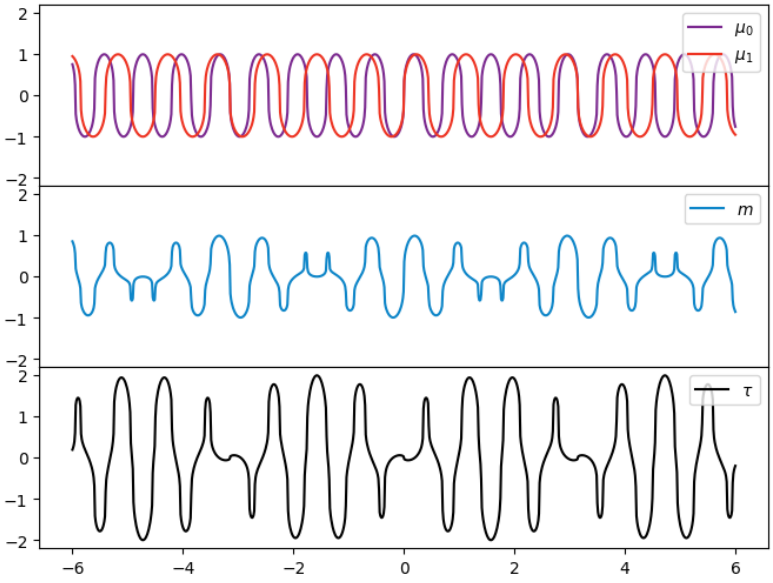
$$\textit{Setting B.} \begin{cases} \mu^0 : x & \mapsto \sin\left(\frac{2\pi x}{.9}\right)^{1/3} - \sin\left(\frac{2\pi x}{.7}\right)^{1/3} \\ \mu^1 : x & \mapsto \sin\left(\frac{2\pi x}{.9}\right)^{1/3} + \sin\left(\frac{2\pi x}{.7}\right)^{1/3} \\ X & \sim \textit{Unif}([-6,6]) \\ T & \sim \textit{Ber}(1/2) \\ Y & \sim (1-T)\mu^0(X) + T\mu^1(X) \end{cases}$$

In *Setting A.*, the complexity of τ is higher than that of the outcome functions μ^0, μ^1. In *Setting B.*, the situation is the opposite. Let us define m (Section III.3.1.4) as $m : x \in \mathcal{X} \mapsto \mathbb{E}[Y|X = x]$. It comes (Figs. D.1a and D.1b):

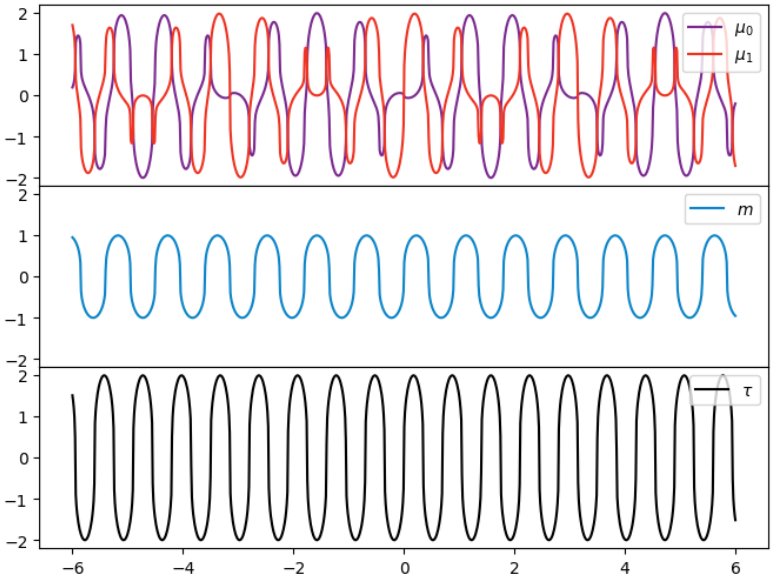
$$\textit{Setting A.} \begin{cases} \mu^0 : x & \mapsto \sin\left(\frac{2\pi x}{.7}\right)^{1/3} \\ \mu^1 : x & \mapsto \sin\left(\frac{2\pi x}{.9}\right)^{1/3} \\ m : x & \mapsto 1/2\left(\sin\left(\frac{2\pi x}{.7}\right)^{1/3} + \sin\left(\frac{2\pi x}{.9}\right)^{1/3}\right) \\ \tau : x & \mapsto \sin\left(\frac{2\pi x}{.9}\right)^{1/3} - \sin\left(\frac{2\pi x}{.7}\right)^{1/3} \end{cases}$$

$$\textit{Setting B.} \begin{cases} \mu^0 : x & \mapsto \sin\left(\frac{2\pi x}{.9}\right)^{1/3} - \sin\left(\frac{2\pi x}{.7}\right)^{1/3} \\ \mu^1 : x & \mapsto \sin\left(\frac{2\pi x}{.9}\right)^{1/3} + \sin\left(\frac{2\pi x}{.7}\right)^{1/3} \\ m : x & \mapsto \sin\left(\frac{2\pi x}{.9}\right)^{1/3} \\ \tau : x & \mapsto 2\sin\left(\frac{2\pi x}{.7}\right)^{1/3} \end{cases}$$

The experiments are conducted as follows. The learners are informed that T is uniformly selected at random and that the features are periodic. In that



(a) *Setting A.*



(b) *Setting B.*

Figure D.1: Distributions of the ground truth functions μ^0, μ^1, m and τ that underlie the observed distributions.

case, the choice of Gaussian Processes with periodic kernels as base learners is adequate. The kernel expression modeling similarity of samples x and x' is given by

$$k(x, x') = \exp\Big(-\frac{2}{d^2} \sin^2\Big(\frac{\pi|x - x'|}{p}\Big)\Big)$$

The model admits the kernel length scale d and the periodicity p as hyper-parameters; they are set by uniformly sampling the hyper-parameter space, and retaining the hyper-parameter setting with the lowest prediction error. The *T-learner* learns estimates $\hat{\mu}_0, \hat{\mu}_1$, inducing a causal estimate $\hat{\tau}_T$. The *R-learner* first learns the estimate $\hat{m}$, then learns the second-stage model $\hat{\tau}_R$ by solving

$$\underset{\hat{\tau}}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(y_i - \hat{m}(x_i) - \big(t_i - 1/2\big)\hat{\tau}(x_i)\Big)^2$$

The final results are reported in Figs. D.2a and D.2b.

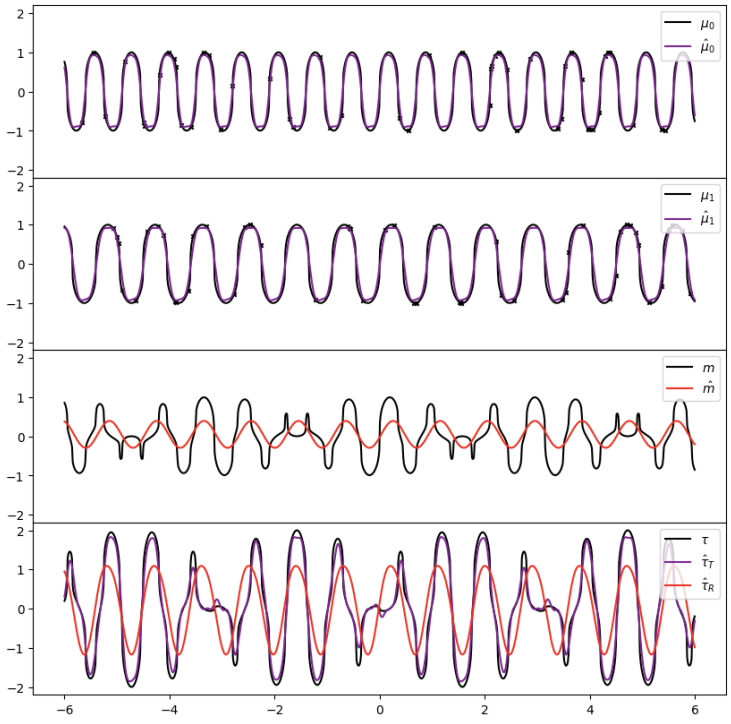
As expected, the *T-learner* is better suited to *Setting A.* than B. Outcome functions μ^0, μ^1 are simple and easy to approximate, and estimates $\hat{\mu}_0, \hat{\mu}_1$ accurately approximate the ground truth (Fig. D.2a). Inversely, m and τ are more complex to model, and the *R-learner* is outperformed by the *T-learner*.

In Setting B, the outcome functions μ^0, μ^1 are more complex, and due to the low number of training data points, the *T-learner* provides poor estimates. m and τ have much simpler expressions here, benefiting the *R-learner*.

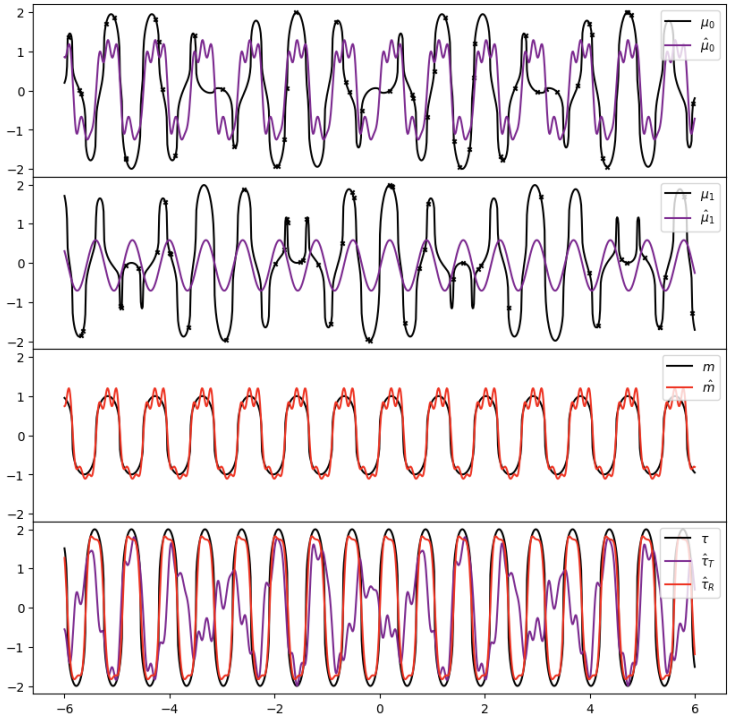
$$\textit{Setting A.} \begin{cases} \textit{PEHE}(\hat{\tau}_T) = \textbf{5.1e-2} \\ \textit{PEHE}(\hat{\tau}_R) = 7.5\text{e-}1 \end{cases}$$

$$\textit{Setting B.} \begin{cases} \textit{PEHE}(\hat{\tau}_T) = 2.3\text{e}0 \\ \textit{PEHE}(\hat{\tau}_R) = \textbf{8.6e-2} \end{cases}$$

Although these problems are toy ones, they illustrate the importance of the data generation process on the performance of different types of meta-learners.



(a) *Setting A.*



(b) *Setting B.*

Figure D.2: Color code: *T-learner*, *R-learner*, ground truth.

E - Theoretical analysis: auxiliary results

Let us detail some of the claims of Section IV.5.3.2, in the same setting, and under the same assumptions.

E.1 . Existence of ν^0 follows the sufficiency of ϕ

Assume that $\phi(X)$ is sufficient with respect to Y^0: $Y^0 \perp\!\!\!\perp X|\phi(X)$ (or equivalently $\phi(X)$ is a prognostic score). Then there exists $\nu^0 : \mathcal{Z} \mapsto \mathbb{R}$, such that $\mu^0 = \nu^0 \circ \phi$

Proof.

$$\begin{aligned}\mu^0(x) &= \mathbb{E}[Y^0|X = x] \\ &= \mathbb{E}[Y^0|\phi(X) = \phi(x), X = x] \\ &= \mathbb{E}[Y^0|\phi(X) = \phi(x)]\end{aligned}$$

And as such, $\mu^0(x)$ is entirely determined by $\phi(x)$; the existence of ν^0 is guaranteed. □

E.2 . Heritage of conditional exchangeability through sufficiency

Assume that $\phi(X)$ is sufficient with respect to Y^0. Then conditional exchangeability w.r.t. $\phi(X)$ holds: $Y^0 \perp\!\!\!\perp T|\phi(X)$

This result has been mentioned in Section II.6 and can be established from direct computation on the density functions as follows:

Proof.

$$\begin{aligned}
& f_{Y^0,T|\phi(X)}(y,0|\phi(x)) \\
=& \int_{\phi^{-1}(\{\phi(x)\})} f_{Y^0,T|\phi(X),X}(y,0|\phi(x),u) f_{X|\phi(X)}(u|\phi(x))du && \text{(total probabilities)} \\
=& \int_{\phi^{-1}(\{\phi(x)\})} f_{Y^0,T|X}(y,0|u) f_{X|\phi(X)}(u|\phi(x))du && \text{(same } \sigma\text{-algebra)} \\
=& \int_{\phi^{-1}(\{\phi(x)\})} f_{Y^0|X}(y|u) f_{T|X}(0|u) f_{X|\phi(X)}(u|\phi(x))du && (Y^0 \perp\!\!\!\perp T|X) \\
=& \int_{\phi^{-1}(\{\phi(x)\})} f_{Y^0|\phi(X)}(y|\phi(u)) f_{T|X}(0|u) f_{X|\phi(X)}(u|\phi(x))du && (\phi(X) \text{ sufficient}) \\
=& f_{Y^0|\phi(X)}(y|\phi(x)) \int_{\phi^{-1}(\{\phi(x)\})} f_{T|X}(0|u) f_{X|\phi(X)}(u|\phi(x))du
\end{aligned}$$

$$= f_{Y^0|\phi(X)}(y|\phi(x)) f_{T|\phi(X)}(0|\phi(x))$$

□

E.3 . $\phi(X)$ is not sufficient in general

A counter-example proves that the existence of ν^0 s.t. $\mu^0 = \nu^0 \circ \phi$ does not imply the sufficiency of $\phi(X)$ ($Y^0 \perp\!\!\!\perp X|\phi(X)$).

Proof. Let us consider the following setting: $\mathcal{X} = [0,1] \times [0,1]$, with samples being drawn uniformly in $\mathcal{X}$, and $Y^0 \sim \mathcal{N}(X_1, X_2^2)$. Let $\phi : (x_1, x_2) \mapsto x_1$ be the projection on the first feature axis.

Since $\mathbb{E}[Y^0|X = x] = \phi(x)$, $\nu^0 = Id$ verifies the condition. However, $\phi(X)$ being fixed, the variance of Y^0 depends on X_2; $Y^0 \not\perp\!\!\!\perp X|\phi(X)$ and $\phi(X)$ is no sufficient statistic for Y^0. □

E.4 . Conditional exchangeability w.r.t. $\phi(X)$ does not hold in general

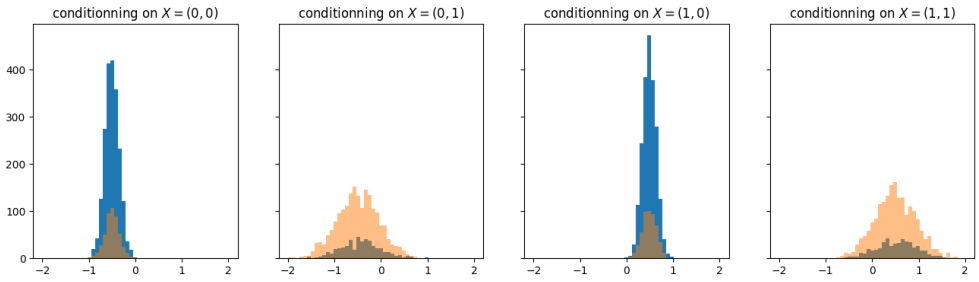
Similarly to the previous section, the existence of ν^0 s.t. $\mu^0 = \nu^0 \circ \phi$ does not even imply that $Y^0 \perp\!\!\!\perp X|\phi(X)$. A counter-example also proves this negative result.

Proof. Consider the following setting:

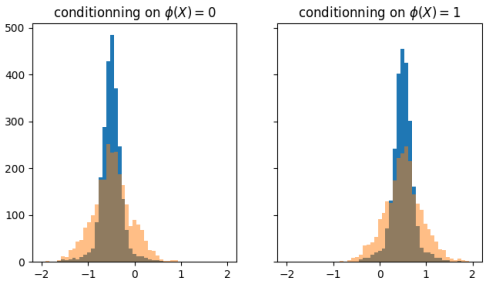
$$\begin{cases} X \sim (Ber(1/2), Ber(1/2)) \\ \phi(X) = X_1 \\ T = X_2 \times (2n_T - 1) + 1 - n_T, & n_T \sim Ber(.9) \\ Y^0 = (X_1 - 1/2) + (X_2 + 1/2) \times n_Y, & n_Y \sim \mathcal{N}(0, .1) \end{cases} \tag{E.1}$$

Set now $\nu^0 : z \mapsto z - 1/2$. Then, $\mathbb{E}[Y^0|X = x] = x_1 - 1/2 = \nu^0 \circ \phi(x)$. Conditional exchangeability w.r.t. X, positivity and *SUTVA* hold.

However, conditional exchangeability w.r.t. $\phi(X)$ does not hold. X_1 being fixed, the variance of Y^0 is larger when T takes value 1 than when it takes value 0: $Y^0 \not\perp\!\!\!\perp T|\phi(X)$. See Fig. E.1 for an illustration. □



(a) Conditioning on X: $Y^0 \perp\!\!\!\perp T|X$



(b) Conditioning on $\phi(X)$: $Y^0 \not\perp\!\!\!\perp T|\phi(X)$

Figure E.1: Histogram of Y^0 in the setting of Eq. E.1, 10,000 samples.

F - Potential learning instability

Consider the training of pipeline $\mathcal{P}_0$ and a given treated sample $(x_i, t_i = 1, y_i)$ that happens to be the mirror twin of numerous control samples $\{(x_k, t_k = 0, y_k)\}_{k \in K}$. The minimization of the prediction error $||h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i||$ impacts the learned $h^1_{\mathcal{P}_0}$ in the neighborhood of $\phi_{\mathcal{P}_0}(x)$, and as such impacts the counterfactual estimation of all samples $\{(x_k, t_k = 0, y_k)\}_{k \in K}$. The greater the factual prediction error on x, the likely larger the counterfactual prediction error of samples index by K. The solution we propose consists in providing to x an importance that increases linearly with the cardinality of K.

Nonetheless one has to keep in mind that at training time, the representational network $\phi_{\mathcal{P}_0}$ is not fixed. Increasing the importance of samples with high exemplarity in the training loss may lead to an undesired effect. Back to the previous example, consider now two treated samples $(x_i, t_i = 1, y_i)$ and $(x_j, t_j = 1, y_j)$. Assume that $(x_i, t_i = 1, y_i)$ is still the mirror twin w.r.t $\phi_{\mathcal{P}_0}$ of numerous control samples $\{(x_k, t_k = 0, y_k)\}_{k \in K}$, whereas $(x_j, t_j = 1, y_j)$ is the mirror twin of no control sample. Assume finally that the prediction error of $(x_i, t_i = 1, y_i)$ is much higher than that of $(x_j, t_j = 1, y_j)$:

$$\begin{aligned} \textit{exemplarity}_{\mathcal{P}_0}(i) &= |K| \textit{ is big} \\ \textit{exemplarity}_{\mathcal{P}_0}(j) &= 0 \\ ||h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_j) - y_j|| &\ll ||h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i|| \end{aligned}$$

As said, updating $h^1_{\mathcal{P}_0}$ into $h'^1_{\mathcal{P}_0}$ to reduce the prediction error contributes towards the minimization of the loss:

$$\begin{aligned} ||h'^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i|| &\leq ||h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i|| \\ &\implies (1 + \beta_{\mathcal{P}_0} \textit{exemplarity}_{\mathcal{P}_0}(i)) \times ||h'^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i|| \\ &\leq (1 + \beta_{\mathcal{P}_0} \textit{exemplarity}_{\mathcal{P}_0}(i)) \times ||h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i|| \end{aligned}$$

However, suppose that there exists a mapping change from $\phi_{\mathcal{P}_0}$ to $\phi'_{\mathcal{P}_0}$ that swaps the representations of $(x_i, t_i = 1, y_i)$ and $(x_j, t_j = 1, y_j)$ while preserving the representations of all other training samples:

$$\begin{cases} \phi_{\mathcal{P}_0}(x_i) & = \phi'_{\mathcal{P}_0}(x_j) \\ \phi_{\mathcal{P}_0}(x_j) & = \phi'_{\mathcal{P}_0}(x_i) \\ l \notin \{i, j\} & \implies \phi_{\mathcal{P}_0}(x_l) = \phi'_{\mathcal{P}_0}(x_l) \end{cases}$$

Now $(x_j, t_j = 1, y_j)$ is the mirror twin w.r.t $\phi'_{\mathcal{P}_0}$ of samples $\{(x_k, t_k = 0, y_k)\}_{k \in K}$, while $(x_i, t_i = 1, y_i)$ is the mirror twin of no control sample. The resulting

diminution of the training loss value is likely even larger than the previous one:

$$(1 + \beta_{\mathcal{P}_0} \textit{exemplarity}_{\mathcal{P}_0}(i)) \times ||h^1_{\mathcal{P}_0} \circ \phi'_{\mathcal{P}_0}(x_i) - y_i|| \leq (1 + \beta_{\mathcal{P}_0} \textit{exemplarity}_{\mathcal{P}_0}(i)) \times ||h^1_{\mathcal{P}_0} \circ \phi_{\mathcal{P}_0}(x_i) - y_i||$$

Hopefully, the gradient does not flow back through the mirror twin operator during the back-propagation phase, entailing no incentive for the representational network to optimize in such a pathological way. However, this phenomenon is prone to create instabilities during the training. The impacts of insulation and exemplarity depend on the one nearest neighbor of the considered samples, a strategy that is sensitive to infinitesimal modifications of the network parameters. In that sense, the regularization terms as they are formalized in IV.2.1 lack robustness.

G - AI-assisted writing

The comprehensive list of Large Language Model requests in the redaction of this manuscript follows:

G.1 . ChatGPT 3.5

Search for alternatives to "*mirror twin*"

https://chat.openai.com/share/ceed0e70-1a2b-4214-a7d6-11dcc31c6080

Explanation of the Generalized Method of moments

https://chat.openai.com/share/f1e4e4b2-8613-48d4-9065-3dd199d4bc7c

Suggestions of articles related to the variance of IPTW approaches

https://chat.openai.com/share/3d9c655a-c882-4a0d-8ae7-6cbb3694ff23

Interestingly the returned article does not seem to exist.

Search for recent *CATE* articles

https://chat.openai.com/share/a0223671-2f00-4522-8a1b-678abf89ef66

No returned article.

Variants of "to induce a question"

https://chat.openai.com/share/52da4dd6-afce-432c-9dfa-5862f48c37cd

Heritage of conditional exchangeability through sufficiency

https://chat.openai.com/share/e20bc3e3-5bc6-4c70-a66a-be2377374a0f

This request is interesting. Assuming conditional exchangeability w.r.t X ($Y^0 \perp\!\!\!\perp T|X$) and sufficiency of $\phi(X)$ w.r.t Y^0 ($Y^0 \perp\!\!\!\perp X|\phi(X)$), the agent is asked to prove that conditional exchangeability still holds w.r.t $\phi(X)$ ($Y^0 \perp\!\!\!\perp T|\phi(X)$). It answers that additional assumptions are required. The result is, however, proved in Appendix E.2.

LaTeX formatting, warning management

https://chat.openai.com/share/a14cc2b2-7052-41bf-8865-d4757e8c9c65

LaTeX formatting, regular expression

https://chat.openai.com/share/8da262bb-3836-4d0f-82a0-cc528ae9365d

G.2 . Other

The redaction of this manuscript has made extensive use of search engines and automated proofreading assistants (*Overleaf*, *Grammarly*), which technicaly qualifies as AI-assisted writing.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: a system for large-scale machine learning. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation, OSDI'16, pages 265–283, USA. USENIX Association.

Acharki, N., Lugo, R., Bertoncello, A., and Garnier, J. (2023). Comparison of meta-learners for estimating multi-valued treatment heterogeneous effects. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of ICML'23, pages 91–132, Honolulu, Hawaii, USA. JMLR.org.

Ai, C. and Chen, X. (2003). Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. Econometrica, 71(6):1795–1843. Publisher: [Wiley, Econometric Society].

Alaa, A. and Schaar, M. (2018). Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. In Proceedings of the 35th International Conference on Machine Learning, pages 129–138. PMLR. ISSN: 2640-3498.

Alaa, A. and Schaar, M. V. D. (2019). Validating Causal Inference Models via Influence Functions. In Proceedings of the 36th International Conference on Machine Learning, pages 191–201. PMLR. ISSN: 2640-3498.

Alaa, A. M. and Van Der Schaar, M. (2018). Bayesian Nonparametric Causal Inference: Information Rates and Learning Algorithms. IEEE Journal of Selected Topics in Signal Processing, 12(5):1031–1046.

Allen, J. P. (2017). Technology and Inequality: Concentrated Wealth in a Digital World. Palgrave Macmillan, New York, NY, 1st ed. 2017 edition edition.

Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. Journal of Econometrics, 2(2):105–110.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association, 91(434):444–455. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Angrist, J. D., Pischke, J.-s., and Pischke, J.-s. (2009). Mostly Harmless Econometrics – An Empiricist's Companion. Princeton University Press, Princeton.

Armstrong, T. B. and Kolesár, M. (2021). Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness. Econometrica, 89(3):1141–1177. arXiv:1712.04594 [econ, stat].

Assaad, S., Zeng, S., Tao, C., Datta, S., Mehta, N., Henao, R., Li, F., and Carin, L. (2021). Counterfactual Representation Learning with Balancing Weights. In Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, pages 1972–1980. PMLR. ISSN: 2640-3498.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. Science, 355(6324):483–485. Publisher: American Association for the Advancement of Science.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27):7353–7360. Publisher: Proceedings of the National Academy of Sciences.

Athey, S. and Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application. Observational Studies, 5:37–51.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statistics in Medicine, 27(12):2037–2049.

Beaman, L., Duflo, E., Pande, R., and Topalova, P. (2012). Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India. Science (New York, N.y.), 335(6068):582–586.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2006). Analysis of Representations for Domain Adaptation. In Advances in Neural Information Processing Systems, volume 19. MIT Press.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. IEEE transactions on pattern analysis and machine intelligence, 35:1798–1828.

Bica, I., Alaa, A. M., and Van Der Schaar, M. (2020). Time series deconfounder: estimating treatment effects over time in the presence of hidden confounders. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of ICML'20, pages 884–895. JMLR.org.

Billewicz, W. Z. (1965). The Efficiency of Matched Samples: An Empirical Investigation. Biometrics, 21(3):623–644. Publisher: [Wiley, International Biometric Society].

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., and Zieba, K. (2016). End to End Learning for Self-Driving Cars. arXiv:1604.07316 [cs].

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. (2013). Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. Journal of Machine Learning Research, 14(101):3207–3260.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. Journal of the American Statistical Association, 90(430):443–450. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Breiman, L. (1984). Classification and Regression Trees. Routledge, Boca Raton, Fla., 1st edition edition.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1):5–32.

Brooks-Gunn, J., Liaw, F. R., and Klebanov, P. K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. The Journal of Pediatrics, 120(3):350–359.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Foundations and Trends® in Machine Learning, 5.

Calderaro, A. and Blumfelde, S. (2022). Artificial intelligence and EU security: the false promise of digital sovereignty. European Security, 31(3):415–434. Publisher: Routledge _eprint: https://doi.org/10.1080/09662839.2022.2101885.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. (2021). Extracting Training Data from Large Language Models. In Proceedings of the 30th USENIX Security Symposium, pages 2633–2650.

Caron, A., Baio, G., and Manolopoulou, I. (2022). Estimating individual treatment effects using non-parametric regression models: A review. Journal of the Royal Statistical Society Series A, 185(3):1115–1149. Publisher: Royal Statistical Society.

Cassel, C. M., Sarndal, C. E., and Wretman, J. H. (1976). Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. Biometrika, 63(3):615–620. Publisher: [Oxford University Press, Biometrika Trust].

Cayton, L. (2008). Algorithms for manifold learning. Technical Report CS2008-0923, Department of Computer Science & Engineering, UC San Diego.

Chandra, R. and Krishna, A. (2021). COVID-19 sentiment analysis via deep learning during the rise of novel cases. PLOS ONE, 16(8):e0255615. Publisher: Public Library of Science.

Chauhan, V. K., Molaei, S., Tania, M. H., Thakur, A., Zhu, T., and Clifton, D. A. (2023). Adversarial De-confounding in Individualised Treatment Effects Estimation. In Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, pages 837–849. PMLR. ISSN: 2640-3498.

Chen, Y.-L., Minorics, L., and Janzing, D. (2022). Correcting Confounding via Random Selection of Background Variables. arXiv:2202.02150 [cs, stat].

Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K. S., and Liu, H. (2022a). Evaluation Methods and Measures for Causal Learning Algorithms. IEEE Transactions on Artificial Intelligence, 3(6):924–943.

Cheng, M., Liao, X., Liu, Q., Ma, B., Xu, J., and Zheng, B. (2022b). Learning Disentangled Representations for Counterfactual Regression via Mutual Information Minimization. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, pages 1802–1806, New York, NY, USA. Association for Computing Machinery.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. (2020). CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In Proceedings of the 37th International Conference on Machine Learning, pages 1779–1788. PMLR. ISSN: 2640-3498.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2017). Generic machine learning inference on heterogenous treatment effects in randomized experiments. CeMMAP working paper CWP61/17, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Chicharro, D., Panzeri, S., and Shpitser, I. (2019). Conditionally-additive-noise Models for Structure Learning. ArXiv.

Chickering, D. M. (2002). Optimal Structure Identification With Greedy Search. Journal of Machine Learning Research, 3(Nov):507–554.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). Under Review of ICLR2016 (1997).

Clivio, O., Falck, F., Lehmann, B., Deligiannidis, G., and Holmes, C. (2022). Neural Score Matching for High-Dimensional Causal Inference. International Conference on Artificial Intelligence and Statistics, 151. Publisher: arXiv Version Number: 1.

Cochran, W. G. (1953). Matching in analytical studies. American Journal of Public Health and the Nation's Health, 43(6 Pt 1):684–691.

Cochran, W. G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. Biometrics, 24(2):295–313. Publisher: [Wiley, International Biometric Society].

Cochran, W. G. and Rubin, D. B. (1973). Controlling Bias in Observational Studies: A Review. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 35(4):417–446. Publisher: Springer.

Cooper, G. F. and Glymour, C. (1999). Computation, Causation, and Discovery. MIT Press, Menlo Park, Calif.

Cox, D. R. (1958). Planning of Experiments. John Wiley. Google-Books-ID: fzlqvgEACAAJ.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. Biometrika, 96(1):187–199.

Curth, A., Alaa, A. M., and van der Schaar, M. (2021a). Estimating Structural Target Functions using Machine Learning and Influence Functions. arXiv:2008.06461 [stat].

Curth, A. and Schaar, M. v. d. (2021). Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms. In Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, pages 1810–1818. PMLR. ISSN: 2640-3498.

Curth, A. and Schaar, M. V. D. (2023). In Search of Insights, Not Magic Bullets: Towards Demystification of the Model Selection Dilemma in Heterogeneous Treatment Effect Estimation. In Proceedings of the 40th International Conference on Machine Learning, pages 6623–6642. PMLR. ISSN: 2640-3498.

Curth, A., Svensson, D., Weatherall, J., and van der Schaar, M. (2021b). Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 1.

Cuturi, M. (2013). Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In Advances in Neural Information Processing Systems, volume 26. Curran Associates, Inc.

De Ganay, C. and Gouttefarde, F. (2020). Rapport d'information en conclusion des travaux d'une mission d'information sur les systèmes d'armes létaux autonomes. rapport d'information 3248, Assemblée Nationale.

Dehejia, R. (2005). Practical propensity score matching: a reply to Smith and Todd. Journal of Econometrics, 125(1-2):355–364.

Dehejia, R. H. and Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. Journal of the American Statistical Association, 94(448):1053–1062. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Dehejia, R. H. and Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. The Review of Economics and Statistics, 84(1):151–161.

Dhar, P. (2020). The carbon impact of artificial intelligence. Nature Machine Intelligence, 2(8):423–425. Number: 8 Publisher: Nature Publishing Group.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In Multiple Classifier Systems, Lecture Notes in Computer Science, pages 1–15, Berlin, Heidelberg. Springer.

Ding, P. and Li, F. (2018). Causal Inference: A Missing Data Perspective. Statistical Science, 33(2):214–237. Publisher: Institute of Mathematical Statistics.

Dodge, Y. (2008). The Concise Encyclopedia of Statistics. Springer-Verlag New York Inc., Philadelphia, Pa., 2010e édition edition.

Dorie, V. (2023). vdorie/npci. original-date: 2015-07-02T17:34:24Z.

Doutreligne, M. and Varoquaux, G. (2023). How to Select Predictive Models for Decision Making or Causal Inference.

Du, X., Sun, L., Duivesteijn, W., Nikolaev, A., and Pechenizkiy, M. (2021). Adversarial balancing-based representation learning for causal effect inference with observational data. Data Mining and Knowledge Discovery, 35(4):1713–1738.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 1097–1104, Madison, WI, USA. Omnipress.

Díaz, I. and Laan, M. J. v. d. (2013). Sensitivity Analysis for Causal Inference under Unmeasured Confounding and Measurement Error Problems. The International Journal of Biostatistics, 9(2):149–160. Publisher: De Gruyter.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. Journal of Econometrics, 221(2):644–654.

Ernst, E., Merola, R., and Samaan, D. (2019). Economics of Artificial Intelligence: Implications for the Future of Work. IZA Journal of Labor Policy, 9(1).

Ferenc, J.-S. and Néda, Z. (2007). On the size distribution of Poisson Voronoi cells. Physica A: Statistical Mechanics and its Applications, 385(2):518–526.

Fernandez, P., Couairon, G., Jégou, H., Douze, M., and Furon, T. (2023). The Stable Signature: Rooting Watermarks in Latent Diffusion Models. In Proceedings of IEEE/CVF International Conference on Computer Vision, 2023.

Forastiere, L., Airoldi, E. M., and Mealli, F. (2021). Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks. Journal of the American Statistical Association, 116(534):901–918. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2020.1768100.

Foster, D. J. and Syrgkanis, V. (2023). Orthogonal statistical learning. The Annals of Statistics, 51(3):879–908. Publisher: Institute of Mathematical Statistics.

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., and Suganthan, P. N. (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115:105151.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. (2016). Domain-Adversarial Training of Neural Networks. Journal of Machine Learning Research, 17(59):1–35.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56(1):1513–1589.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc.

Gouk, H., Frank, E., Pfahringer, B., and Cree, M. J. (2021). Regularisation of neural networks by enforcing Lipschitz continuity. Machine Learning, 110(2):393–416.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test. Journal of Machine Learning Research, 13(25):723–773.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A Kernel Statistical Test of Independence. In Advances in Neural Information Processing Systems, volume 20. Curran Associates, Inc.

Guo, X., Zhang, Y., Wang, J., and Long, M. (2023). Estimating heterogeneous treatment effects: mutual information bounds and learning algorithms. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of ICML'23, pages 12108–12121, Honolulu, Hawaii, USA. JMLR.org.

Gutierrez, P. and Gérardy, J.-Y. (2017). Causal Inference and Uplift Modelling: A Review of the Literature. In Proceedings of The 3rd International Conference on Predictive Applications and APIs, pages 1–13. PMLR. ISSN: 2640-3498.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). Robust Statistics: The Approach Based on Influence Functions. Wiley–Blackwell, New York.

Hannart, A., Pearl, J., Otto, F. E. L., Naveau, P., and Ghil, M. (2016). Causal Counterfactual Theory for the Attribution of Weather and Climate-Related Events. Bulletin of the American Meteorological Society, 97(1):99–110.

Hansen, B. (2006). Bias Reduction in Observational Studies via Prognosis Scores. Technical Report 441, Statistics Department, University of Michigan.

Hansen, B. B. (2008). The Prognostic Analogue of the Propensity Score. Biometrika, 95(2):481–488. Publisher: [Oxford University Press, Biometrika Trust].

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. Econometrica, 50(4):1029–1054. Publisher: [Wiley, Econometric Society].

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-Sample Properties of Some Alternative GMM Estimators. Journal of Business & Economic Statistics, 14(3):262–280. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Hardt, M. and Recht, B. (2022). Patterns, Predictions, and Actions: Foundations of Machine Learning. Princeton University Press, Princeton.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. In Proceedings of the 34th International Conference on Machine Learning, pages 1414–1423. PMLR. ISSN: 2640-3498.

Hassanpour, N. and Greiner, R. (2019a). CounterFactual Regression with Importance Sampling Weights. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, pages 5880–5887, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Hassanpour, N. and Greiner, R. (2019b). Learning Disentangled Representations for CounterFactual Regression. In Proceedings of The 8th International Conference on Learning Representations.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. (2021). The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution

Generalization. In Proceedings of IEEE/CVF International Conference on Computer Vision, 2021, pages 8340–8349.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. Journal of Computational and Graphical Statistics, 20(1):217–240.

Hirano, K. and Imbens, G. W. (2005). The Propensity Score with Continuous Treatments. In Gelman, A. and Meng, X.-L., editors, Wiley Series in Probability and Statistics, pages 73–84. John Wiley & Sons, Ltd, Chichester, UK.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis, 15(3):199–236.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1):55–67. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

Holland, P. W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association, 81(396):945–960. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1986.10478354.

Hong, H., Leung, M., and Li, J. (2019). Inference on Finite Population Treatment Effects Under Limited Overlap. The Econometrics Journal, 23.

Horowitz, J. L. and Manski, C. F. (1995). Identification and Robustness with Contaminated and Corrupted Data. Econometrica, 63(2):281–302. Publisher: [Wiley, Econometric Society].

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. In Advances in Neural Information Processing Systems, volume 21. Curran Associates, Inc.

Hu, L., Gu, C., Lopez, M., Ji, J., and Wisnivesky, J. (2020). Estimation of causal effects of multiple treatments in observational studies with a binary outcome. Statistical Methods in Medical Research, 29(11):3218–3234.

Hutter, F., Kotthoff, L., and Vanschoren, J., editors (2019). Automated Machine Learning: Methods, Systems, Challenges. The Springer Series on Challenges in Machine Learning. Springer International Publishing, Cham.

Hyvarinen, A. and Morioka, H. (2016). Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc.

Hyvarinen, A., Sasaki, H., and Turner, R. E. (2019). Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. arXiv:1805.08651 [cs, stat].

Hyvärinen, A., Khemakhem, I., and Monti, R. (2023). Identifiability of latent-variable and structural-equation models: from linear to nonlinear. arXiv:2302.02672 [cs, stat].

Höfler, M. (2005). Causal inference based on counterfactuals. BMC Medical Research Methodology, 5(1):28.

Iacus, S. M., King, G., and Porro, G. (2011). Multivariate Matching Methods That Are Monotonic Imbalance Bounding. Journal of the American Statistical Association, 106(493):345–361.

Iacus, S. M., King, G., and Porro, G. (2012). Causal Inference without Balance Checking: Coarsened Exact Matching. Political Analysis, 20(1):1–24.

Imbens, G. W. (2000). The Role of the Propensity Score in Estimating Dose-Response Functions. Biometrika, 87(3):706–710. Publisher: [Oxford University Press, Biometrika Trust].

Johansson, F. D. (2023). cfrnet. original-date: 2016-07-12T10:29:44Z.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. (2022). Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects. arXiv:2001.07426 [cs, stat].

Johansson, F. D., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, pages 3020–3029, New York, NY, USA. JMLR.org.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873):583–589. Number: 7873 Publisher: Nature Publishing Group.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4):422–446.

Kaddour, J., Zhu, Y., Liu, Q., Kusner, M. J., and Silva, R. (2021). Causal Effect Inference for Structured Treatments. In Advances in Neural Information Processing Systems, volume 34, pages 24841–24854. Curran Associates, Inc.

Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., and Sebag, M. (2022). Structural Agnostic Modeling: Adversarial Learning of Causal Graphs. Journal of Machine Learning Research, 23(219):1–62.

Kapelner, A., Bleich, J., Levine, A., Cohen, Z. D., DeRubeis, R. J., and Berk, R. (2021). Evaluating the Effectiveness of Personalized Medicine With Software. Frontiers in Big Data, 4.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pages 3149–3157, Red Hook, NY, USA. Curran Associates Inc.

Kelejian, H. H. (1971). Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables. Journal of the American Statistical Association, 66(334):373–374.

Kendall, M. G. (1938). A New Measure of Rank Correlation. Biometrika, 30(1/2):81–93. Publisher: [Oxford University Press, Biometrika Trust].

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. arXiv:2004.14497 [math, stat].

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, pages 2207–2217. PMLR. ISSN: 2640-3498.

King, G. and Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. Political Analysis, 27(4):435–454. Publisher: Cambridge University Press.

King, G. and Zeng, L. (2007). When Can History Be Our Guide? The Pitfalls of Counterfactual Inference. International Studies Quarterly, 51(1):183–210.

Kingma, D. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.

Kingma, D. and Welling, M. (2014). Auto-Encoding Variational Bayes. In Proceedings of The 2nd International Conference on Learning Representations.

Kobyzev, I., Prince, S., and Brubaker, M. (2020). Normalizing Flows: An Introduction and Review of Current Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP:1–1.

Kuang, K., Cui, P., Li, B., Jiang, M., Yang, S., and Wang, F. (2017). Treatment Effect Estimation with Data-Driven Variable Decomposition. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1). Number: 1.

Kuang, K., Cui, P., Zou, H., Li, B., Tao, J., Wu, F., and Yang, S. (2022). Data-Driven Variable Decomposition for Treatment Effect Estimation. IEEE Transactions on Knowledge and Data Engineering, 34(5):2120–2134. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. Proceedings of the National Academy of Sciences, 116(10):4156–4165. arXiv:1706.03461 [math, stat].

Laan, M. J. v. d. and Robins, J. M. (2011). Unified Methods for Censored Longitudinal Data and Causality. Springer, New York, NY, softcover reprint of hardcover 1st ed. 2003 edition edition.

Laan, M. J. v. d. and Rose, S. (2011). Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Science & Business Media. Google-Books-ID: RGnSX5aCAgQC.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. (2019). Gradient-Based Neural DAG Learning. In Proceedings of The 8th International Conference on Learning Representations.

Laffers, L. and Mellace, G. (2020). Identification of the Average Treatment Effect When SUTVA Is Violated.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. The American Economic Review, 76(4):604–620. Publisher: American Economic Association.

Lannelongue, L., Grealey, J., and Inouye, M. (2021). Green Algorithms: Quantifying the Carbon Footprint of Computation. Advanced Science, 8(12):2100707. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.202100707.

Lara, L. d. (2023). The difference between structural counterfactuals and potential outcomes.

Lechner, M. (1999). Earnings and Employment Effects of Continuous Gff-the-Job Training in East Germany After Unification. Journal of Business & Economic Statistics, 17(1):74–90. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07350015.1999.10524798.

Li, F. (2019). Propensity score weighting for causal inference with multiple treatments. The Annals of Applied Statistics, 13(4):2389–2415. Publisher: Institute of Mathematical Statistics.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. Journal of the American Statistical Association, 113(521):390–400. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1260466.

Li, H., Zhao, G., Johari, R., and Weintraub, G. Y. (2022). Interference, Bias, and Variance in Two-Sided Marketplace Experimentation: Guidance for Platforms. In Proceedings of the ACM Web Conference 2022, WWW '22, pages 182–192, New York, NY, USA. Association for Computing Machinery.

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. The International Journal of Biostatistics, 9(2):215–234.

Liao, J. and Rohde, C. (2022). Variance reduction in the inverse probability weighted estimators for the average treatment effect using the propensity score. Biometrics, 78(2):660–667.

Lopez, M. J. and Gutman, R. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. Statistical Science, 32(3). arXiv:1701.05132 [stat].

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal Effect Inference with Deep Latent-Variable Models. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. Emerging Themes in Epidemiology, 15:1.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine, 23(19):2937–2960.

Maclaren, O. J. and Nicholson, R. (2020). What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems. arXiv:1904.02826 [cs, math, stat].

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of The 6th International Conference on Learning Representations.

Mahajan, D., Mitliagkas, I., Neal, B., and Syrgkanis, V. (2023). Empirical Analysis of Model Selection for Heterogeneous Causal Effect Estimation. arXiv:2211.01939 [cs, stat].

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures, 90:46–60.

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. The American Economic Review, 80(2):319–323. Publisher: American Economic Association.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29):861.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6):115:1–115:35.

Meldrum, M. L. (2000). A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. Hematology/Oncology Clinics of North America, 14(4):745–760, vii.

Melnychuk, V., Frauen, D., and Feuerriegel, S. (2022). Causal Transformer for Estimating Counterfactual Outcomes. In Proceedings of the 39th International Conference on Machine Learning, pages 15293–15329. PMLR. ISSN: 2640-3498.

Miguel, E. and Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Econometrica, 72(1):159–217. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2004.00481.x.

Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. (2023). Causal Discovery with Score Matching on Additive Models with Arbitrary Noise. In Proceedings of the Second Conference on Causal Learning and Reasoning, pages 726–751. PMLR. ISSN: 2640-3498.

Nandy, P., Hauser, A., and Maathuis, M. H. (2018). High-dimensional consistency in score-based and hybrid structure learning. arXiv:1507.02608 [math, stat].

Neal, B. (2010). Introduction to Causal Inference. Journal of machine learning research.

Newey, W. K. and Powell, J. L. (2003). Instrumental Variable Estimation of Nonparametric Models. Econometrica, 71(5):1565–1578. Publisher: [Wiley, Econometric Society].

Neyman, J. (1979). C(α) Tests and Their Use. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 41(1/2):1–21. Publisher: Springer.

Nie, L., Ye, M., Liu, Q., and Nicolae, D. (2020). VCNet and Functional Targeted Regularization For Learning Causal Effects of Continuous Treatments. In Proceedings of The 11th International Conference on Learning Representations.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. Biometrika, 108(2):299–319.

Ogarrio, J. M., Spirtes, P., and Ramsey, J. (2016). A Hybrid Causal Search Algorithm for Latent Variable Models. In Proceedings of the Eighth International Conference on Probabilistic Graphical Models, pages 368–379. PMLR. ISSN: 1938-7228.

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

Oprescu, M., Dorn, J., Ghoummaid, M., Jesson, A., Kallus, N., and Shalit, U. (2023). B-learner: quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of ICML'23, pages 26599–26618, Honolulu, Hawaii, USA. JMLR.org.

Ozer, M. E., Sarica, P. O., and Arga, K. Y. (2020). New Machine Learning Applications to Accelerate Personalized Medicine in Breast Cancer: Rise of the Support Vector Machines. OMICS: A Journal of Integrative Biology, 24(5):241–246. Publisher: Mary Ann Liebert, Inc., publishers.

Pawlowski, N., Coelho de Castro, D., and Glocker, B. (2020). Deep Structural Causal Models for Tractable Counterfactual Inference. In Advances in Neural Information Processing Systems, volume 33, pages 857–869. Curran Associates, Inc.

Pearl, J. (2009). Causality. Cambridge University Press. Google-Books-ID: f4nuexsNVZIC.

Pearl, J. (2011). Invited Commentary: Understanding Bias Amplification. American Journal of Epidemiology, 174(11):1223–1227.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85):2825–2830.

Peters, J., Janzing, D., and Schlkopf, B. (2017). Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal Discovery with Continuous Additive Noise Models. Journal of Machine Learning Research, 15(58):2009–2053.

Peterson, A. and Spirling, A. (2018). Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. Political Analysis, 26(1):120–128. Publisher: Cambridge University Press.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Larochelle, H. (2021). Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). The Journal of Machine Learning Research, 22(1):164:7459–164:7478.

Platt, J. (2000). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv. Large Margin Classif., 10.

Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. Proceedings of the section on survey research methods.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. Statistics in Medicine, 37(11):1767–1787.

Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. (2017). A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. International Journal of Data Science and Analytics, 3(2):121–129.

Rasmussen, C. E. and Williams, C. K. I. (2005). Gaussian Processes for Machine Learning. The MIT Press.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In Proceedings of the 31st International Conference on Machine Learning, pages 1278–1286. PMLR. ISSN: 1938-7228.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence, UAI'96, pages 454–461, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rissanen, S. and Marttinen, P. (2022). A Critical Look at the Consistency of Causal Estimation With Deep Latent Variable Models. arXiv:2102.06648 [cs].

Rivera, B. and Currais, L. (1999). Economic growth and health: direct impact or reverse causation? Applied Economics Letters, 6(11):761–764. Publisher: Routledge _eprint: https://doi.org/10.1080/135048599352367.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. Journal of the American Statistical Association, 90(429):106–121. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476493.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. Econometrica : journal of the Econometric Society, 56(4):931. JSTOR: 1912705.

Rolling, C. A. and Yang, Y. (2014). Model selection for estimating treatment effects. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 76(4):749–769. Publisher: [Royal Statistical Society, Wiley].

Rosenbaum, P. R. (1989). Optimal Matching for Observational Studies. Journal of the American Statistical Association, 84(408):1024–1032. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Rosenbaum, P. R. (2002). Observational Studies. Springer-Verlag New York Inc., New York, 2nd ed. 2002 édition edition.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007). Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer. Journal of the American Statistical Association, 102(477):75–83. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70(1):41–55. Publisher: [Oxford University Press, Biometrika Trust].

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician, 39(1):33–38. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Rosenblatt, F. (1962). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books. Google-Books-ID: 7FhRAAAAMAAJ.

Rothe, C. (2017). Robust Confidence Intervals for Average Treatment Effects Under Limited Overlap. Econometrica, 85(2):645–660. Publisher: [Wiley, The Econometric Society].

Rubin, D. B. (1973). Matching to Remove Bias in Observational Studies. Biometrics, 29(1):159–183. Publisher: [Wiley, International Biometric Society].

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701. Place: US Publisher: American Psychological Association.

Rubin, D. B. (1976). Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples. Biometrics, 32(1):109–120. Publisher: [Wiley, International Biometric Society].

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. Ann. Statist., 6(1):34–58.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. Journal of the American Statistical Association, 100(469):322–331. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214504000001880.

Saito, Y. and Yasui, S. (2020). Counterfactual cross-validation: stable model selection procedure for causal inference models. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of ICML'20, pages 8398–8407. JMLR.org.

Sapsford, R. and Jupp, V. (2006). Data Collection and Analysis. SAGE Publications Ltd.

Schuler, A., Baiocchi, M., Tibshirani, R., and Shah, N. (2018). A comparison of methods for model selection when estimating individual treatment effects. arXiv: Machine Learning.

Schwab, P., Linhardt, L., and Karlen, W. (2019). Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks. arXiv:1810.00656 [cs, stat].

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning, pages 3076–3085. PMLR. ISSN: 2640-3498.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting Neural Networks for the Estimation of Treatment Effects. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Shimizu, S., Hoyer, P. O., Hyv&#228, A., rinen, and Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. Journal of Machine Learning Research, 7(72):2003–2030.

Signorovitch, J. E. (2007). Identifying Informative Biological Markers in High-dimensional Genomic Data and Clinical Trials. Harvard University. Google-Books-ID: MNIJuQAACAAJ.

Simon, H. A. (1954). Spurious Correlation: A Causal Interpretation. Journal of the American Statistical Association, 49(267):467–479. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society. Series B (Methodological), 13(2):238–241. Publisher: [Royal Statistical Society, Wiley].

Sinclair, B., McConnell, M., and Green, D. P. (2012). Detecting Spillover Effects: Design and Analysis of Multilevel Experiments. American Journal of Political Science, 56(4):1055–1069. Publisher: [Midwest Political Science Association, Wiley].

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel Instrumental Variable Regression. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Smith, J. and Todd, P. (2001). Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods. American Economic Review, 91:112–118.

Smith, J. and Todd, P. (2005a). Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, 125(1):305–353.

Smith, J. and Todd, P. (2005b). Rejoinder. Journal of Econometrics, 125(1):365–375.

Sobel, M. E. (2006). What Do Randomized Studies of Housing Mobility Demonstrate? Journal of the American Statistical Association, 101(476):1398–1407. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214506000000636.

Spirtes, P. and Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. Social Science Computer Review, 9(1):62–72. Publisher: SAGE Publications Inc.

Spirtes, P., Glymour, C., and Scheines, R. (1993). Causation, Prediction, and Search, volume 81 of Lecture Notes in Statistics. Springer, New York, NY.

Stadie, B. C., Künzel, S. R., Vemuri, N., and Sekhon, J. S. (2018). Estimating Heterogeneous Treatment Effects Using Neural Networks With The Y-Learner. Preprint.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. Journal of Business & Economic Statistics, 20(4):518–529.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science : a review journal of the Institute of Mathematical Statistics, 25(1):1–21.

Stürmer, T., Rothman, K. J., Avorn, J., and Glynn, R. J. (2010). Treatment Effects in the Presence of Unmeasured Confounding: Dealing With Observations in the Tails of the Propensity Score Distribution—A Simulation Study. American Journal of Epidemiology, 172(7):843–854.

Sutton, R. S. and Barto, A. G. (2018). Reinforcement Learning, second edition: An Introduction. Bradford Books, Cambridge, Massachusetts, second edition edition.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning, 65(1):31–78.

Vegetabile, B. G. (2021). On the Distinction Between "Conditional Average Treatment Effects" (CATE) and "Individual Treatment Effects" (ITE) Under Ignorability Assumptions. arXiv:2108.04939 [cs, stat].

Verine, A., Négrevergne, B., Rossi, F., and Chevaleyre, Y. (2021). On the expressivity of bi-Lipschitz normalizing flows. ArXiv.

Virmaux, A. and Scaman, K. (2018). Lipschitz regularity of deep neural networks: analysis and efficient estimation. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association, 113(523):1228–1242. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1319839.

White, H. and Chalak, K. (2013). Identification and Identification Failure for Treatment Effects Using Structural Systems. Econometric Reviews, 32(3):273–317. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07474938.2012.690664.

Wong, T. and Sabanayagam, C. (2019). Strategies to Tackle the Global Burden of Diabetic Retinopathy: From Epidemiology to Artificial Intelligence. Ophthalmologica, 243(1):9–20.

Wooldridge, J. M. (2016). Should instrumental variables be used as matching variables? Research in Economics, 70(2):232–237.

Wu, A., Kuang, K., Xiong, R., Li, B., and Wu, F. (2023a). Stable Estimation of Heterogeneous Treatment Effects. In Proceedings of the 40th International Conference on Machine Learning, pages 37496–37510. PMLR. ISSN: 2640-3498.

Wu, A., Yuan, J., Kuang, K., Li, B., Wu, R., Zhu, Q., Zhuang, Y., and Wu, F. (2023b). Learning Decomposed Representations for Treatment Effect Estimation. IEEE Transactions on Knowledge and Data Engineering, 35(5):4989–5001. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Wu, P. A. and Fukumizu, K. (2021). Beta-Intact-VAE: Identifying and Estimating Causal Effects under Limited Overlap. In Proceedings of The 10th International Conference on Learning Representations.

Yang, D., Small, D. S., Silber, J. H., and Rosenbaum, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. Biometrics, 68(2):628–636.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. Biometrika, 105(2):487–493.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2018). Representation Learning for Treatment Effect Estimation from Observational Data. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2019). ACE: Adaptively Similarity-Preserved Representation Learning for Individual Treatment Effect Estimation. In 2019 IEEE International Conference on Data Mining (ICDM), pages 1432–1437. ISSN: 2374-8486.

Yoon, J., Jordon, J., and Schaar, M. v. d. (2018). GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In Proceedings of The 6th International Conference on Learning Representations.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the eighth ACM SIGKDD

international conference on Knowledge discovery and data mining, KDD '02, pages 694–699, New York, NY, USA. Association for Computing Machinery.

Zanga, A. and Stella, F. (2022). A Survey on Causal Discovery: Theory and Practice. International Journal of Approximate Reasoning, 151:101–129. arXiv:2305.10032 [cs].

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending Against Neural Fake News. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Zellner, A., Huang, D. S., and Chau, L. C. (1965). Further Analysis of the Short-Run Consumption Function with Emphasis on the Role of Liquid Assets. Econometrica, 33(3):571.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI'11, pages 804–813, Arlington, Virginia, USA. AUAI Press.

Zhang, W., Liu, L., and Li, J. (2021). Treatment Effect Estimation with Disentangled Latent Factors. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 10923–10930. ISSN: 2374-3468, 2159-5399 Issue: 12 Journal Abbreviation: AAAI.

Zhang, Y., Bellot, A., and Schaar, M. (2020). Learning Overlapping Representations for the Estimation of Individualized Treatment Effects. In Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, pages 1005–1014. PMLR. ISSN: 2640-3498.

Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Zhang, F., and Choo, K.-K. R. (2022). Artificial intelligence in cyber security: research advances, challenges, and opportunities. Artificial Intelligence Review, 55(2):1029–1053.

Zhao, Y., Fang, X., and Simchi-Levi, D. (2017). Uplift Modeling with Multiple Treatments and General Response Types. In Proceedings of the 2017 SIAM International Conference on Data Mining (SDM), Proceedings, pages 588–596. Society for Industrial and Applied Mathematics.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous Optimization for Structure Learning. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc.

Zhou, G., Yao, L., Xu, X., Wang, C., and Zhu, L. (2021). Cycle-Balanced Representation Learning For Counterfactual Inference. In Proceedings of the 2022

SIAM International Conference on Data Mining (SDM), Philadelphia, PA. Society for Industrial and Applied Mathematics. arXiv:2110.15484 [cs].

Zhu, S., Ng, I., and Chen, Z. (2020). Causal Discovery with Reinforcement Learning. arXiv:1906.04477 [cs, stat].

Zubizarreta, J. R. (2012). Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure After Surgery. Journal of the American Statistical Association, 107(500):1360–1371. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2012.703874.