Review of Causal Average Treatment Effect and Treatment Heterogeneity Estimation with Non-finite Action

Yating Zou

Gillings School of Public Health

University of North Carolina – Chapel Hill

June 2022

Dr. Michael Kosorok, Thesis Advisor

Dr. Donglin Zeng, Committee Member

Dr. Jane Monaco, Committee Member

Contents

1 Introduction **6**

2 Problem setup **8**

2.1 Notation 8

2.2 Assumptions 9

3 Preliminaries and Problem Background **11**

3.1 Graphical Representation using DAGs and SWIGs 11

3.2 Association vs Causation 12

3.3 Randomized vs Observational Study 12

3.4 Binary vs Continuous Action 13

3.5 ADRF vs Individual DRF 14

3.6 A Missing Data Perspective 15

4 Estimation of Average Treatment Effect (ATE) **17**

4.1 Reweighing the Observed Data 17

4.1.1 Balancing through the GPS 17

4.1.2 Balancing Weight/GPS Estimation 22

4.2 Imputing the Missing Outcomes 26

4.2.1 Outcome Modeling 26

4.3 Combined Models and semi-parametric Methods 27

4.3.1 The von Mises Expansion 28

4.3.2 Controlling the Plug-in Bias 30

5 Estimation of Conditional Average Treatment Effect (CATE) **38**

5.1 Modified-ML Algorithms 39

5.2 Meta-ML Algorithms 42

6 Variable Selection and Diagnostics **46**

7 Discussion **50**

8 Tables for Notation, Symbols, and Abbreviations **55**

8.1 Table of Notation and Symbols 55

8.2 Table of Abbreviations 56

9 List of References **58**

Connections between Methods for ATE and CATE Estimation

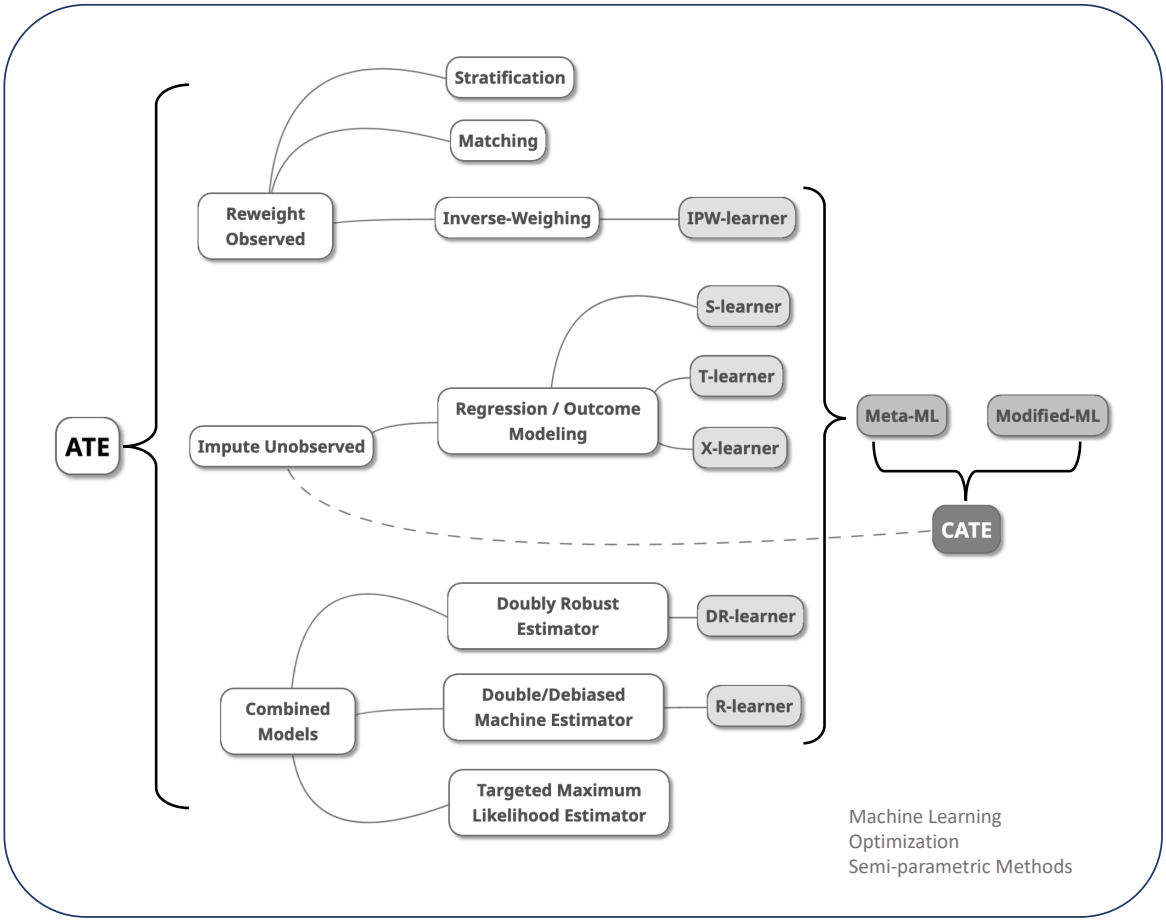

Figure 0: Relationships between different methods used to estimate the Average Treatment Effect (ATE) and the Conditional Average Treatment Effect (CATE). The main sections of this thesis will follow this structure.

Abstract

There is an increasing appeal in answering causal questions, and this interest has drawn perspectives from many related areas. However, estimation of the Average Treatment Effect (ATE) and the Conditional Average Treatment Effect (CATE) with continuous, or non-finite, action is still a relatively new setup with its specific challenges, including extra difficulty in the estimation of the Generalized Propensity Score (GPS), the diagnostics of covariate balancing, and the pathwise non-differentiability of the functional for ATE. This thesis, therefore, provides a review of recent advances on this topic. The main objectives are to (1) identify and coalesce recent related research, (2) provide a general introduction to methods such as machine learning, semi-parametric estimation, and optimization while discussing specific methods, and (3) outline the connections between existing methods of ATE and CATE estimation so that readers can better navigate themselves in this fast-expanding field.

Acknowledgement

Words cannot express my deepest gratitude to Dr. Michael Kosorok, my incredibly supportive thesis advisor. Discussions with you are always pleasant and enlightening. Thank you for offering me the opportunity to be in your lab and sit in your Precision Medicine class. Thank you for accepting to be my thesis advisor, which made this review possible only so.

Special thanks to Dr. Donglin Zeng and Dr. Jane Monaco for being on my committee and making time available in the summer. Special thanks to Dr. Michael Hudgens for your inspiring class on Causal Inference.

I would also like to recognize the support from the Salisbury Family Excellence Fund administered by Honors Carolina.

Last but not least, thank you, mom and dad, who are also my dearest friends, for always being by my side and supporting me in my exploration of life. We have limited time being physically together, but "love knows not distance; it hath no continent; its eyes are for the stars.[1]"

[1] Quote from Gilbert Parker

1 Introduction

"All men by nature desire to know." — Aristotle

The desire to understand not just the phenomenon but also the underlying mechanism of nature is shared by many fields of study. At its heart, this pushes us to ask causal questions in the structure of what would happen to Y *if we do* X? There has been increasing interest in examining the cause-and-effect relationship in recent years. This trend brings together perspectives from various areas such as econometrics, biostatics, political policy, epidemiology, and computer science. Around 1974, Rosenbaum and Rubin formalized the concept of potential outcomes, outcomes that are defined hypothetically, to highlight our interest in *causal* quantities. Since then, the potential outcome framework has been applied to the areas mentioned above to answer questions: How would the risk change if we give them treatment versus not? How would a new policy affect a population? There has been much research in a binary action setting where there are only two groups to compare against in the target population—an intervention (treated) group and a control (untreated) group. There is only a paucity of work for non-finite or continuous treatment. Nonetheless, many of the intervention variables we would be interested in, such as dosing and price, are continuous in nature. Applying this framework to the setting where the action has values ranging across an interval rather than discrete points expands the number of causal questions we would be able to answer.

In the first part, this thesis will focus on estimating the Average Treatment Effect (ATE), also known as the causal Average Dose-Response Function (ADRF), if we want to be explicit about being in a continuous action setting. The second part will focus on estimating the Conditional Average Treatment Effect (CATE). We will also use the term causal Individual Dose-Response Function (I-DRF). As the name I-DRF suggests, CATE is an estimate for an individual characterized by his or her specific set of covariates, thus requiring the estimator to find a homogeneous subgroup for this individual. The I-DRF is useful because we can use it to identify subpopulations with the most potent response to treatment and to quantify individual-level responses subjected to any treatment intensity. The second naturally extends to finding the optimal treatment option for an individual. We refer to this as finding the optimal Dynamic Treatment Regime (DTR) as the regime is dynamic in individual

characteristics. In short, CATE bridges Causal Inference with polity optimization, which is beneficial in many practical settings.

There is an increasing interest in Causal Inference, yet Causal Inference with continuous action is still a relatively new setup with many open questions. The treatment being continuous brings about many additional challenges. Firstly, the assumptions for causal identification, the ability to use the observed data to estimate the hypothetically defined potential outcomes, become much harder to meet. Any estimand of interest, whether the ADRF, the I-DRF, or any other, would face this challenge. Secondly, if using a semi-parametric estimation method, pathwise non-differentiability would get in the way of achieving the smallest asymptotic variance possible. As we will see shortly, tackling the above challenges brings Machine Learning, optimization, and semi- and non-parametric estimation onto the table. Motivated by this synthesis from multiple areas and the practical value of CATE estimation, we hope to provide a structure for ATE and CATE estimation and outline the connections between the different methods involved. Hence, although our primary focus is on continuous treatment, when there exists no extension into the continuous action setting, we will also discuss estimation methods under a binary action setting to preserve the completeness of the whole structure.

The rest of the thesis is organized as follows: In Section 2, we give our use of notation and the assumptions required for causal identification under the Neyman-Rubin causal model. In Section 3, we discuss and clarify some important aspects of our problem setup using comparison and contrast. At the end of this section, we introduce our categorization for ATE and CATE estimation methods, presented in Figure 0. In Section 4, we discuss the estimation of the Average Treatment Effect (ATE) or the Average Dose-Response Function (ADRF). In Section 5, we discuss the estimation of the Conditional Average Treatment Effect (CATE) or the Individual Dose-Response Function (I-DRF). In Section 6, we discuss variable selection for the Propensity Score models and ways to diagnose covariate balancing, which is crucial for obtaining unbiased causal estimates. Finally, we summarize the main takeaways, point out the limitations of this thesis, discuss related areas such as Dynamic Treatment Regime (DTR) estimation, and discuss possible future extensions in Section 7. In the end, we provide a table of notation (Table 8.1) and a table of abbreviations (Table 8.2) so that readers can refer to them anytime. The readers are also welcome to go back to

Figure 0 at the beginning to see where a particular method falls under a larger framework.

2 Problem setup

2.1 Notation

This review focuses on treatment effect and treatment heterogeneity estimation in a non-finite action setting using i.i.d. data under the potential outcome framework, also known as the Neyman-Rubin causal model, that was developed in a series of papers by Rosenbaum and Rubin under a binary action setting (treated vs. untreated) [62, 109–111]. Under the continuous treatment setting, we adopt notations consistent with previous literature. This problem setup can be traced back to the paper by Imbens [61] in which this framework was first applied under a multivalued action setting to estimate the Average Dose-Response Function (ADRF).

For every unit $i = 1, \ldots, n$, denote the observed pre-treatment covariates $X_i \in \mathcal{X}$, the observed treatment T_i, and the observed outcome $Y_i \in \mathcal{Y}$. Define the potential outcome, $Y_i(t)$, for $t \in \mathcal{T} = [t_{min}, t_{max}] \subset \mathbb{R}$, as the outcome of interest that *would have been observed* if a unit i received treatment intensity t. Note also that $\{Y(t) : t \in \mathcal{T}\} \subset \mathcal{Y}$. As the potential outcome is defined in a purely hypothetical setup, we connect it to the observed outcome via the consistency assumption. That is, $Y_i(t) = Y_i$ if t is the treatment individual i had actually received. We are interested in estimating the Average Dose-Response Function (ADRF) which corresponds to the Average Treatment Effect (ATE) for a target population, defined as

$$\tau(t) := \mathbb{E}[Y(t)]$$

and the Individual Dose-Response Function (I-DRF) which corresponds to the Conditional Average Treatment Effect (CATE) for a subgroup of the population characterized by covariates X, defined as

$$\tau(t, X) := \mathbb{E}[Y(t)|X]$$

Estimation of the ADRF helps answer questions such as the treatment effect on Y, measured

by risk difference, risk ratio, odds ratio, or other measures, if we increase treatment intensity from s to $s + \Delta$. Closely related but differently, I-DRF is a function of both X and T, thus allowing us to fix one and compare the effect of a change in the other. For example, estimation of the I-DRF addresses questions such as what subgroup would benefit most from the treatment? What is the optimal treatment intensity for a particular individual with covariate X? What is the effect of increasing or decreasing his or her treatment intensity by k percent?

2.2 Assumptions

Our estimand of interest is causal, defined in a hypothetical setup, but we have only the observed data. To be able to express causal estimands with observed quantities, formally put as causal identification, we assume the three key assumptions below:

Assumption 1. *Stable Unit Treatment Value Assumption (SUTVA): (i) No interference. The potential outcome of one individual is independent of that of other individuals. (ii) One version of treatment. Each value of $a \in \mathcal{A}$ is unambiguously defined.*

Assumption 2. *Positivity :* $1 > f_{T|X}(T = t|X = X) > 0$, $\forall t \in \mathcal{T}$ *where X has positive measure. $f_{T|X}$ is the conditional density of T given X.*

Assumption 3. *Weak Unconfoundedness :* $Y(t) \perp\!\!\!\perp T|X$, $\forall t \in \mathcal{T}$. *This is also referred to as Conditional Exchangeability and is closely related to Ignorability and Missing at Random.*

The first assumption is not related to the study design but the formulated causal question. It provides a grounding for the potential outcome to be well-defined at an individual level with a clear real-world interpretation. If satisfied, we can graphically put $Y_i(t)$ onto a rectangle, where we can index the vertical edge by i and set the range of $\mathcal{T}$ to be the horizontal edge. Then any point inside this rectangle would be well-defined.

The second and third assumptions are design-dependent. They are automatically satisfied by randomized control trials (RCTs). However, if our data come from an observational study, we would need to possibly modify the scope of our question to achieve structural Positivity[2]

[2]Violation of structural Positivity occurs when the structure of the problem restricts certain subpopula-

and make a choice of what set of covariates to adjust for to achieve *Weak Unconfoundedness*. The key to identification often depends on how well we can satisfy the third assumption. That is, to have the risk in those treated equal to the risk if everyone had been treated when conditioned on covariates X, as $Y(t) \perp\!\!\!\perp T|X$ means $f(Y(t)|T = t, X = x) = f(Y(t)|X = x)$ for some probability density f. This assumption is referred to by different names because we can consider it from different perspectives. Concerning what is needed: we need to identify a set of covariates that, by adjusting for them, is sufficient to remove confounding[3] caused by common causes of T and Y, so we need *Weak Unconfoundedness* [61]. Weak in that independence between the potential outcomes and the treatment is not automatically satisfied but requires controlling for X. Concerning the consequence of this assumption: after controlling for X, the treated and the untreated group would be exchangeable for their potential outcomes. Thus, making inference about the unobserved outcome using the observed data would be justifiable, and we say there is *Conditional Exchangeability* [106]. Regarding its connection to a conditional randomized study: in such a study where Positivity naturally holds, we are guaranteed $Y(t) \perp\!\!\!\perp T|X$ in large samples by design because the treatment is assigned at random. The assigned treatment fully determines the received treatment and is unrelated to the potential outcomes. In other words, within groups with similar distributions of observed covariates at all treatment intensities, the assignment mechanism would be *Ignorable* [106], and the missing data caused by this random assignment would be *Missing at Random (MAR)* [80, 110].

In practice, we are not omniscient, so to ensure the variables within our observation set are sufficient for confounding adjustment, we need to make further the assumption of *No Unmeasured Confounding*. This is an untestable assumption [26], so it is generally preferred to collect as much information as possible for the covariates; however, it is not necessarily better to control for more covariates. A typical example is the induced selection bias caused by censoring in longitudinal studies. Although censoring can be adjusted using the inverse-probability-of-censoring weighting (IPCW), it would require making additional assumptions

tions to a subset of treatment. While we can address random Positivity violations using modeling, structural violations make causal identification impossible without re-defining the target population. See p. 162 in *What If* [85]. If we know the treatment assignment mechanism, for instance in an RCT, structural positivity naturally holds.

[3]Existence of confounding is better defined as the existence of an open backdoor path, where a backdoor path is a non-causal path between treatment and outcome with the structure $T \leftarrow L \rightarrow Y$ for $L \subseteq X$. See Chapter 7 of [85] for a more detailed explanation.

on how censoring relates to other variables [102].

3 Preliminaries and Problem Background

3.1 Graphical Representation using DAGs and SWIGs

As causal relationships can be made intuitive visually, here we briefly introduce the Directed Acyclic Graphs (DAG), [41, 50, 54] and the Single World Intervention Graph (SWIG) [101] that serves as an excellent tool for presenting ideas in the remaining sections. A DAG is composed of nodes representing variables and directed edges, or arrows representing the direction of cause-and-effect. We can transform a DAG into multiple SWIGs based on treating which variables as the intervention variables. Constructing a SWIG from a DAG can be done in the following steps: firstly, split the intervention nodes into two semicircles, one taking in all in-coming nodes, the other being the place where all out-pointing nodes start from; and secondly, denote all decedents of the intervention variable as potential outcomes as shown in Figure 1. Both DAGs and SWIGs encode assumptions needed as input for causal inference. The assumptions often come from expert knowledge or incorporating results from a data-driven method such as those from Causal Discovery [4]. A DAG describes the structure of a set of variables, whereas a specific causal question determines a SWIG. The main advantage of using SWIGs is that, unlike DAGs, SWIGs explicitly incorporate potential outcomes into the graph, facilitating the assessment of conditional independence required for *Weak Unconfoundedness* in complicated graphs.

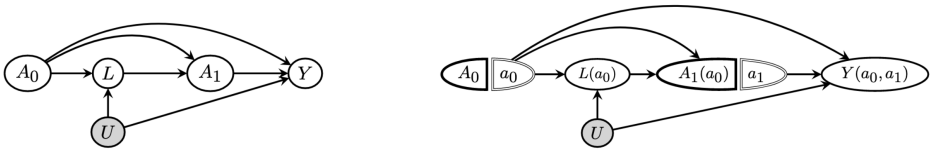

Figure 1: Treating A_0 and A_1 as interventions in a more complicated longitudinal study. Is $Y(a_0, a_1) \perp\!\!\!\perp A_1(a_0) | \{L(a_0), A_0 = a_0\}$? It is hard to tell using a DAG but easy to see using a SWIG.

[4] One can refer to the book *Elements of Causal Inference* by Peters et al. [98] for an introduction, and reviews by Guo et al. [43], and Vowels et al. [129] for recent advances in Causal Discovery.

3.2 Association vs Causation

In statistics, it has been widely noted that association is not causation. Eating ice cream on a hot day does not cause sunburns, although they are both related to the amount of sunlight. Although different, they could be numerically the same if there is no confounding. If so, the observed change in the outcome can only be attributed to the change in the treatment. Checking for unconfoundedness can be done by checking for balance in the covariates. That is, $X \perp\!\!\!\perp T$, the distributions of covariates for units assigned to different treatment intensities should be similar. In an observational study, strong unconfoundedness $Y(t) \perp\!\!\!\perp T$ is often violated. However, we can attempt for *Weak Unconfoundedness* $Y(t) \perp\!\!\!\perp T|X$ by controlling for covariates sufficient to adjust for confounding. If successfully achieved, in this adjusted population, we can readily apply methods used for associational estimation, such as regression, for causal estimation. If not, we will most likely arrive at biased estimates for *causal* estimands, especially at places where X and T are not independent. Due to this extra consideration, care must be taken when applying methods for prediction to causal inference, not only during estimation but also during variable and model selection. For illustration, we should not evaluate the quality of a balancing score [5], such as the Propensity Score (definition given in Section 4.1.1), by its prediction accuracy [17, 131].

3.3 Randomized vs Observational Study

In a randomized control trial (RCT) where there is no non-compliance or only random non-compliance, we can obtain valid causal estimates because in RCTs, *Positivity* (Assumption 2) and *Weak Unconfoundedness* (Assumption 3) naturally hold, as discussed in the Context and Assumptions section. Estimation of this population with sufficient covariate overlap and covariate balance naturally possesses a causal interpretation. Graphically, as shown in Figure 2, random assignment cuts off all edges going into the node T, since T would be related to randomness only. With observational data, adjustment and checking for covariate balancing are musts. From another perspective that is more prevalent in engineering, causal quantities should have the invariant, or stable, property. The causal effect of one cause-

[5] A mapping to the reals that conditioning on it leads to independence of X and T, which is known as the *Balancing Property* (explicitly given in Section 4.1.2)

and-effect pair should not be affected by changes in its surroundings. RCTs cut off received treatment, the cause, with its environment via randomization, while with observational data, the observed change in outcome could be compounded by changes in multiple environmental factors.

Figure 2: Left: SWIG of an RCT. Right: SWIG of an observational study. The rectangle box represents the un-intervened environment with other variables.

3.4 Binary vs Continuous Action

Causal estimation under the non-finite action setting is much less-studied for multiple reasons. The continuous action setting adds extra difficulty when seeking balance. Under a binary action setting where we compare the effect of treated vs. untreated, we can group units into strata so that the treated and untreated are exchangeable within each stratum. We can also match each unit in the treated group with one or more units in the untreated group with similar covariates. In a continuous action setting with $\mathcal{T} = [t_{min}, t_{max}] \subset \mathbb{R}$ uncountably infinite, we can no longer partition our sample into distinct groups of specific dose levels and have a sufficient number of units to partition or create matches. It is almost impossible to obtain a distribution of X at any t. Similarly, since the probability of selection at any t is zero, direct use of applying weights as the inverse of the probability of selection would result in zeros in the denominator. The density of $\mathcal{T}$ leads to sparsity in $\mathcal{X}$ when we condition on T. Somehow, we need to use adjacent information or impose smoothness to fill in the gaps. The difficulty also shows up when we want to estimate a balancing score (1), for example, the Propensity Score (PS) (1) or the Generalized Propensity Score (GPS) (2). Under a binary action setting, it is sufficient to estimate just one value, the PS $e(X) := Pr(A = 1|X)$, when seeking balance since the probability of selection for $A = 0$ can be obtained as $1 - e(X)$. So $e(X)$ is sufficient to construct weights $1/e(X)$ for $A = 1$ and $1/(1 - e(X))$ for $A = 0$. Moreover, diagnostics for $Y(a) \perp\!\!\!\perp A|e(X)$ needs to be performed for only two groups: $a = 0$ and $a = 1$. Contrarily, under a non-finite action setting, we now

need a correctly specified *distribution* of $r(X,t)$, such that $Y(t) \perp\!\!\!\perp T|r(X,t)$, $\forall t \in \mathcal{T}$. For a well-estimated GPS satisfying $Y(t) \perp\!\!\!\perp T|r(X,t)$, all moments of the distribution of GPS need to be sufficiently close to the 'true' distribution, as opposed to matching only the first moment which is the mean (true in the sense that it achieves perfect balance).

Besides adding challenges to achieve *Weak Unconfoundedness*, *Positivity* also becomes a much stronger assumption. Under a binary action setting, there should not be any individual doomed to be treated or untreated. For a quantitative treatment, *Positivity* means every individual should have a possibility of receiving *any* treatment intensity. Satisfying this assumption is particularly hard when some components of X are strongly associated with T, which can be reflected by an extreme Generalized Propensity Score (GPS). Diagnostics on these two assumptions are non-trivial and will be discussed in Section 6.

3.5 ADRF vs Individual DRF

Although we might be able to estimate the *average* causal treatment effect for an outcome of interest, we do not often have the confidence to believe that the effect is uniform throughout the entire targeted population. Differences in dose-response between individuals are well-documented for many diseases. In clinical studies, the FDA provides documents E7 [35] and E11 [34] to particularly shed light on important concerns in drug development for the geriatrics and pediatric population, respectively. It is tempting to move from estimating the overall Average Treatment Effect (ATE) to estimating the Conditional Average Treatment Effect (CATE) that depends on a given set of covariates because the CATE allows individual-level targeting and action recommendation. ATE informs us what a population should receive to improve overall. In contrast, CATE finds subpopulations and provides guidance specific to that subgroup, which can be related to finding an optimal treatment regime dynamic in individual covariates X. Due to the difficulties above specific to the continuous action setting, ATE and CATE estimation is much less studied for non-finite actions, especially the CATE. Again, the core issue is data sparsity at a specific dose level. With binary action, the CATE is often defined as the risk difference between the treated and untreated $\Delta_\tau(X) := \mathbb{E}[Y(1) - Y(0)|X] = \tau(1,X) - \tau(0,X)$. There are only two potential outcome surfaces $\mathbb{E}[Y(1)|X] = \mathbb{E}[Y|A=1,X]$ and $\mathbb{E}[Y(0)|X] = \mathbb{E}[Y|A=0,X]$ to model, or

one, if we directly model the difference $\mathbb{E}[Y(1) - Y(0)|X]$. In contrast, under a non-finite action setting, there would be little data with the actual received treatment intensity $T = t$ to support the estimation of an outcome surface across X at any t.

3.6 A Missing Data Perspective

Causal Inference can be considered a missing data problem, with the missing part being the counter-factual, non-observable potential outcomes. Although we map the observed outcome with the hypothetical potential outcome using the consistency assumption, the observed outcome is a much smaller subset of the potential outcomes. It is especially the case in a non-finite action setting, as it is impossible to assign all $t \in \mathcal{T}$ to all individuals—a problem referred to as the *Fundamental Problem of Causal Inference* [55, 109].

We can categorize causal estimation methods from the missing data perspective in the following three ways: (1) Re-weight the observed data and use a two-stage nested approach. First, use information from $T|X$ to remove confounding. Diagnose the degree of covariate balance. Then estimate the treatment effect on this pseudo balanced population using standard techniques. Stratification, matching, and inverse-weighing belong to this category. (2) Impute the missing potential outcomes for each individual, model $Y|T, X$ and average over X at a specific T. (3) Instead of relying on either $T|X$ as with the first approach or $Y|T, X$ as with the second approach, combine separate models to form an augmented model that achieves semi-parametric efficiency with more robustness against model misspecification. semi-parametric theory [74, 99, 122] is the foundation for all these methods. Augmented Inverse Propensity Weighting (AIPW) also known as Doubly Robust (DR) estimation [12, 19, 66], Targeted Maximum Likelihood Estimation (TMLE) [126], and Double Machine Learning (DML) also known as Debiased Machine Learning [22, 25] all fall under this category.

The above categorization works well for ATE estimation for most estimation methods; its relationship with CATE estimation is more subtle. On the one hand, all CATE estimation methods belong to the second category. Since an unbiased CATE is an unbiased estimator for $\mathbb{E}[Y(t)|X = X]$, there must be some way to impute the missing potential outcomes for a given X. On the other hand, we can use all methods in the three categories with additional

manipulation such as sample-splitting and cross-fitting to make the conditional estimate $\mathbb{E}[Y(t)|X = X]$ unbiased. As we will see in Section 5.2, there is a correspondence between CATE estimation using meta-ML algorithms (different types of learners) and almost all methods we will discuss for the three categories above.

In addition to the above, if we can obtain information on other variables with particular characteristics, we can exploit its role in the causal structure to remove confounding. A common example is the use of Instrumental Variables (IVs) [5, 6]. An IV, denoted Z, has the graphical structure shown in Figure 3.

Figure 3: Examples of Z being an IV for estimating the effect of T on Y. U denotes unmeasured common causes of T and Y. As the plot on the right shows, an IV need not be the cause of T.

An IV must satisfy three criteria: It must be *relevant* to the intervention, that is, $Cov(T, Z) \neq 0$; it must satisfy the *exclusion restriction* criterion in that it should not affect Y except through T; and it must be *exogenous*, that is, not share common causes with Y. Due to these particular characteristics, we can use two-stage least squares to capture the causal effect through the path $Z - T \rightarrow Y$ by first regressing T on Z, then regressing Y on the estimated $\mathbb{E}[T|Z]$. It is a widely used tool when unconfoundedness between the treatment and the outcome is not plausible, and the association of Z and T is strong [16]. Nevertheless, the use of IVs with non-finite treatment options faces additional challenges. In 1995, Pearl proposed the conjecture that instrument validity is untestable when the treatment is a continuous random variable. He found that a necessary theoretical condition which he referred to as *Instrumental Inequality* that relates to Bell's inequality from quantum physics [13] requires treatment to be discrete [97]. Recently, Gunsilius revisited this conjecture and proved that with additional functional form restriction, instrument validity can be re-established theoretically [42]; however, practical validity remains to be developed.

4 Estimation of Average Treatment Effect (ATE)

4.1 Reweighing the Observed Data

This section discusses methods that first block $X \to T$, then estimate the effect of $T^* \to Y$ on the adjusted T^*. Intuitively, we want to construct a pseudo-population where we can achieve balance in the distribution of X between different treatment types. In other words, we want to be able to adjust the observed data as if it had come from a randomized study instead. This clear separation between removing confounding and estimating causal effect facilitates diagnostics on covariate balance.

As mentioned in Section 3.4, it is almost impossible to create strata and matches based on X at every $T = t$. A way to address this issue is to instead find homogenous groups based on the Generalized Propensity Score (GPS), a function that maps $\mathcal{X} \times \mathcal{T}$ to a scalar in $\mathbb{R}$, capturing all covariate imbalance due to confounding.

4.1.1 Balancing through the GPS

All GPS-based methods build on the fact that GPS is a balancing score. GPS extends the Propensity Score (PS) initially proposed in a binary action setting [106, 107] to the continuous action setting by considering conditional density instead of conditional probability[6] [53].

Definition 1. *Propensity Score :* $e(X) := Pr(A = 1|X)$ *is defined as the conditional probability of receiving treatment given a set of covariate* X *for* $\mathcal{A} = \{0, 1\}$*, where 1 denotes treatment and 0 denotes control.*

Definition 2. *Generalized Propensity Score :* $r(X, T) := f(T|X)$ *for* $X \in \mathcal{X}$*,* $T \in \mathcal{T}$*, where* $f(T|X)$ *is the conditional density. PS can be seen as a special case of the GPS as* $e(X) = r(X, 1)$*.*

GPS can be seen as the additional selection probability due to confounding that makes

[6] Another extended definition of the PS is the Propensity Function (PF) proposed by Imai and van Dyk [59]. The PF characterizes a treatment assignment model by its parameter. Due to its restriction on the functional class of models, it is much less flexible and has been seldom used.

observational data different from an RCT. Therefore, conditioning on the GPS sets us back to a hypothetical RCT. We can regard individuals with similar GPS as coming from the same randomized group with roughly the same probability of being treated. Formally, the definition of the GPS gives ignorable treatment assignment $X \perp\!\!\!\perp I(T = t)|r(X, t)$. Then, if we combine this with the *Weak Unconfoundedness* assumption $Y(t) \perp\!\!\!\perp T|X$, we can show the *Balancing Property* [53]

$$Y(t) \perp\!\!\!\perp T|r(X, t). \tag{1}$$

Figure 4 provides a visual representation of this Balancing Property. GPS, a function that takes in t and X, is a collider on the path $t \rightarrow r(X, t) \leftarrow X$. A non-causal association (red, dashed line) is created due to conditioning on the collider [85]. However, the induced non-causal association perfectly cancels with $X \rightarrow T$ because of the Balancing Property. The plot on the right shows the final result after controlling for the GPS. Like what a SWIG for an RCT would look like, there is no arrow going into T. No confounding exists, and the effect of T on Y is identifiable.

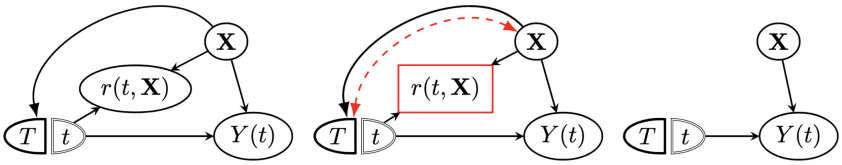

Figure 4: Left: SWIG with GPS. Middle: Conditioning on the GPS. Right: Final result after conditioning. Due to the Balancing Property, the non-causal association induced by conditioning perfectly cancels with $X \rightarrow T$.

As we do not have the luxury of knowing the GPS directly, we should always check after estimating the GPS on: (1) Whether there is Positivity or Sufficient Overlap on the support of GPS (Assumption 2). (2) Whether the estimated GPS indeed achieves balance (Assumption 3). Additionally, it is usually better to use the stabilized weight

$$w_i = \frac{g(T_i)}{r(X_i, T_i)},$$

where $g(T)$ is the marginal density of T, often estimated using Kernel density estimation. While $\frac{1}{r(X_i,T_i)}$ creates a pseudo population where the probability of receiving every treat-

ment intensity is the same, the stabilized weight accounts for the difference in the probability of receiving different treatment intensities in the targeted population. It has also been shown that maintaining unbiased, stabilized weights yields estimates with less variance in unsaturated models [103]. Finally, similarly to many weighting-based estimation methods, one should (3) check for extreme weights. Doing so helps avoid over-emphasizing the outcome of a few outliers, which can result in bias and inflated variance [117]. A general approach is to trim extreme weights, that is, to discard units with estimated GPS close to either 0 or 1; however, this trimmed population might be different from our original target population. To avoid ad hoc selection of the cutoff and ambiguous target population, in the binary action setting, Crump et al. suggests dropping units with estimated PS outside $[0.1, 0.9]$ where there is limited overlap in the distribution of covariates [27]. In hopes of further mitigating these issues, Li et al. proposed overlap weighting (OW) that uses $1 - \widehat{e}(X)$ for the $A = 1$ group and $\widehat{e}(X)$ for the $A = 0$ group to automatically down weight units with limited overlap [79, 120] (see [78] for OW with multiple treatment). With simulations, Li et al. showed the superiority of OW over inverse-propensity weighting and inverse-propensity weighting with trimming suggested by Crump et al. [79]. Interestingly, even if we know the *true* PS, the *estimated* PS outperforms in PS-based matching and stratification [106–108]. Rosenbaum provided an explanation that although the true PS is unbiased, the estimated PS is *conditionally unbiased*, thus not only accounting for systematic bias but also chance bias due to sampling [105].

Assuming we can estimate the GPS (estimation methods discussed in Section 4.1.2), we can estimate the ADRF using stratification also known as subclassification, matching, or weighting. Reviews and comparisons of these methods with a focus on the binary action setting can be found in [83, 114, 116]. Despite being limited in quantity, all three methods have been extended to the non-finite action setting [18, 132].

GPS-CDF Stratification [18] Brown et al. proposed using classification methods to create strata with the label being a 2-d (scale and location) parameter coming from the empirical Cumulative Distribution Function (eCDF) of the estimated GPS. The 2-d characterization allows visualization and ease of interpretation. If GPS is estimated parametrically, the steps are:

1. Assume a multivariate normal distribution of $\widehat{\beta}^*$, with mean and variance obtained

from fitting a treatment regression model $T = \widehat{\beta}^{\intercal} X_i$ for every unit.

2. Obtain a bootstrapped distribution of $\widehat{T}_i^*$ for each subject by resampling $\widehat{\beta}^*$ B times. Calculated its corresponding eCDF $\widehat{eCDF}_i(t) = \frac{1}{B}\sum_{b=1}^{B} I(\widehat{T}_{i,b}^* \leq t)$. Use the fact that eCDF is a one-to-one function, and use nonlinear least squares to fit a logistic curve to each $eCDF_i$

$$eCDF_i(t) = \frac{1}{1 + exp(-s_i(t - \mu_i))}.$$

   Parameterize it using a scale parameter s_i and a location parameter μ_i.

3. Use classification methods such as K-Nearest-Neighbors to obtain groups or strata based on s_i and μ_i. Estimate $\mu(t)$ from a stratified outcome model.

If we want to estimate the GPS non-parametrically, fitting $T = \widehat{\beta}^{\intercal} X_i$ is changed to a variable selection procedure. The non-parametric version does not involve direct estimation of the GPS and uses empirical likelihood instead. The coefficients can be assumed to follow any 0-centered unimodal continuous distribution. The remaining steps are similar, with the change being that the eCDF for each unit can now be described by only one scale parameter, as all estimated eCDFs will be centered around 0.

GPS-based Matching Wu et al. [132] proposed 1-to-M[7] ($M \geq 1$) caliper nearest neighbor matches with replacement within a pre-specified treatment intensity block. Both the treatment intensity T and the estimated GPS R are used in distance calculation to incorporate their deviation from exact matching. Two tuning parameters, the bin width 2δ and $\lambda \in [0, 1]$, a weight to balance the impact of T and R in the distance metric that minimizes covariate imbalance, are found via grid search. The specific steps are:

1. Obtain a treatment assignment model to estimate the GPS. Standardize the estimated GPS and treatment intensity for each unit. After this step, a 2-d vector (r^*, t^*) quantifies each unit.
2. Separate the space of $\mathcal{T}$ by L equal-sized exposure levels such that we can put $Y_i(t)$ in a $N \times L$ two-way table for the N observed units. Define a 2-d distance metric to be used in the caliper matching function, for example, $m_{GPS}(r, X) = \arg\min_{j:t_j \in [t-\delta, t+\delta]} \|(\lambda r^*(X_j, t_j), (1-\lambda)t_j^*) - (\lambda r^*, (1-\lambda)t^*)\|$, with $\|\cdot\|$ the Man-

[7] In matching, small M like 1 reduces finite sample biases caused by matching discrepancies, whereas larger values of M produce lower asymptotic variances [1]

hattan[8] or Euclidean Distance. Match with replacement each unit to its nearest neighbor within the same exposure level and impute its potential outcome as $Y_j(t) = Y_{m_{GPS}(r(X,t),t)}$.

3. Obtain $\widehat{\tau}(t) = \widehat{\mathbb{E}}\{\widehat{Y}_j(t)\}$. As the estimator is point-wise, impose smoothness to the obtained point-wise estimators by either fitting a kernel smoother to all estimates across $\mathcal{T}$ or using a kernel for each observation.

GPS-based Weighting The idea of using a kernel as a smoother can also be seen in GPS-based inverse weighting and many other methods working with continuous data. From weighted kernel regression, we can arrive at the estimator

$$\tau(t) = \sum_{i=1}^{n} w_i Y_i K_h(T_i - t) / \sum_{i=1}^{n} w_i K_h(T_i - t),$$

where $K_h(T_i - t) = K(\frac{T_i - t}{h})/h$ is a kernel of bandwidth h, chosen by minimizing the mean-squared-error [23]. We can regard the kernel function as a smoothed version of the indicator function that applies soft weights to adjacent points for each observation, thus inducing smoothness for the estimated outcome surface. Kernel regression is a data-driven approach. It builds the shape of a function based on the distribution of the observed data, borrowing only adjacent information so that far-away points have little, or nearly no, influence on the estimation at t. This is in contrast to parametric modeling, which first fixes a class of models and then makes use of all data across $\mathcal{T}$ to find the best parametrization within a pre-specified functional form.

From stratification to matching to weighting, we can consider the latter a more generalized version of the former. Alternatively, stratification and matching can be viewed as exceptional cases of weighting. Stratification finds homogeneous subgroups via partitioning the obtained sample. Matching removes the restriction of working with the current sample and creates a pseudo population by allowing matching with replacement via some distance metric designed according to our interest. Weighting goes further by allowing each individual to represent a non-integer amount of people in the pseudo population. Additionally, similar to matching, we have flexibility in the design of weight estimation. Although weighting is the most flexible

[8] Manhattan distance $d_M(x, y) = \sum_{i=1}^{d} |x_i - y_i|$, d the dimension of x and y. In a 2-d plane with $x = (x_1, x_2)$ and $y = (y_1, y_2)$, $d(x, y) = |x_1 - y_1| + |x_2 - y_2|$—the horizontal distance plus the vertical distance. Other distance metrics can also be used in this setting.

approach and should be able to achieve balance best, this is true only if the weights are estimated correctly. Its advantage could also be its disadvantage—being most sensitive to model misspecification. In addition, under a binary action setting, it has been shown that the performance of different PS-based methods differs depending on the type of measure used to quantify the causal effect (relative risk [7], marginal odds ratio [11]). Although there is no direct comparison between the GPS-based estimation methods when the action is non-finite, we can reasonably speculate that the choice of measures does also associate with the preference for stratification, matching, and weighting. As a result, there is no clear winner without having a specific context for the problem.

4.1.2 Balancing Weight/GPS Estimation

Correctly estimating the GPS requires both selecting a set of variables for the model and specifying correctly its relationships. Here we discuss the second part and leave the first part for Section 6. Traditionally, GPS is estimated by fitting a parametric regression model $T = \beta^\intercal X + \epsilon$, assume normality of the errors to estimate parameters by MLE, then estimate the conditional density from the normal density of the errors [107]. Specifically,

$$r(X_i, T_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(T_i - \beta^\intercal X_i)^2\}.$$

Alternatively, we can use method of moments (MOM) to solve moment conditions $E[\psi(X, T ; \theta)] = 0$ to obtain $\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}^2)$ [60]. For the same parametric assumption above, if we replace the score equation for θ with the zero weighted cross moment restriction so that the parameters are identifiable, then for X^* and T^* being the centralized, orthogonalized version of X and T, the estimating equations would be

$$\mathbb{E}[\psi_\theta(T_i, X_i)] = \mathbb{E}\begin{pmatrix} \frac{1}{\sigma^2}(T_i^* - \beta^\intercal X_i^*)^2 - 1 \\ \sigma exp\{\frac{1}{2\sigma^2}(T_i^* - \beta^\intercal X_i^*)^2 - \frac{T_i^*}{2}\} T_i^* X_i^{*\intercal} \end{pmatrix} = 0.$$

Under the MOM framework, one can derive the asymptotic variance of estimated parameters as an alternative to bootstrapping. However, the functional form of $T|X$ is a strong assumption that often fails to hold when the cardinality of $\mathcal{X}$ is large. Its misspecification leads to biased estimates [66]. Repeatedly modifying the current model and checking for

covariant balance without a guarantee of success, as what early methods do, is too cumbersome. Luckily, revisiting the characteristics of the GPS, we can see that GPS estimation only serves as an intermediate process for balance. Its primary use is for balance, so low bias is more important than a low variance. Also, GPS is a nuisance parameter in that we are more interested in its value than its precise functional form. The estimated GPS need not, and should not, have a causal interpretation. In light of this, we can use semi- or non-parametric approaches and ensemble methods (meta-algorithms) to mitigate model misspecification further.

The **Generalized Boosted Model (GBM)** [140], which builds upon the work of McCaffrey et al. [84], has shown promising performance. Instead of fitting a pre-specified functional form to approximate the true function, it models the GPS by additively fitting non-parametric regression trees as its base learners to form a strong learner until it reaches a pre-specified level of balance:

$$T = f(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2),$$
$$f(X) = \sum_{m=1}^{M} \sum_{j=1}^{K_m} c_{mj} \mathbb{1}\{X \in L_{mj}\}.$$

The hyperparameter M, the number of trees, reflects the complexity of the model and the amount of bias-variance trade-off. c_{mj} is the average of T within region or leaf L_{mj}. When a tree with K_m terminal nodes is built, the algorithm naturally finds the best variable to form homogeneous subgroups L_{mj} by maximizing the difference of some information criteria between two groups. At each iteration, GBM selects a random sub-sample and uses this sub-sample to construct a new regression tree that models the residuals from the current boosted tree, analogous to the updating process in the Gradient Descent Algorithm.

Besides boosting, there is also stacking, such as the **Super Learner (SL)** [56, 124], which can take a convex combination of estimates from a pool of prediction models. Each model can be parametric or non-parametric. As long as the number of candidates is polynomial in sample size, SL reaches the almost parametric convergence rate $(\log(n))/n$ or performs as well as the best estimator asymptotically, depending on whether there exists a candidate learner converging to the true value at a parametric rate. To create the ensemble, we estimate the weight for each candidate by minimizing the loss function $\mathcal{L}(\cdot)$ defined as the

cross-validated negative log likelihood [125]. For V cross-validation splits, the loss for the j^{th} prediction model is

$$\mathcal{L}(\widehat{r}_j) = \frac{1}{V}\sum_{v=1}^{V}\frac{1}{n_v}\sum_{i\in v} -log(\widehat{r}_{j,\widehat{v}}(X_i, t_i)). \tag{2}$$

Choose α to minimize $\mathcal{L}(\widehat{r}_a)$ with the constraint that $\alpha_j \geq 0$ and $\sum_j \alpha_j = 1$. Then the estimated GPS is

$$\widehat{r}_a(X,t) = \sum_j \alpha_j \widehat{r}_j(X,t).$$

Suppose we prefer not to place any parametric assumption on the GPS and are willing to absorb possibly some additional computational cost. In that case, the **non-parametric Covariate Balancing Propensity Score (npCBPS)** [37] which extends Imai and Ratkovic [58] directly searches for the optimal weights without GPS modeling. Under the Empirical Likelihood framework, finding balancing weights reduces to an optimization problem maximizing the empirical likelihood with constraints on weights and balance in X. A hyperparameter η is introduced to relax the constraint on complete balance in return for a faster convergence rate. Explicitly, the value of η can be written as $\frac{1}{n}\sum_{i=1}^{n} w_i T_i^* X_i^{*\intercal}$. Let λ and γ be the Lagrangian multipliers, $g(X_i^*, T_i^*) = (T_i^*, X_i^{*\intercal}, T_i^* X_i^{*\intercal})^\intercal$; the Lagrangian of this problem can then be expressed as

$$\underset{w\in\mathbb{R}^n}{\arg\min}\, \mathcal{L}(w_i, \lambda, \gamma|\eta) = \sum_{i=1}^{n} log(w_i) + \lambda\Big(n - \sum_{i=1}^{n} w_i\Big) + \gamma^\intercal\Big(\sum w_i g(X_i^*, T_i^*) - \eta\Big), \tag{3}$$

where the third block contains the constraint on imbalance $E[w_i T_i^* X_i^{*\intercal}] = 0$. While the first block comes from maximizing the empirical likelihood, we can also use other metrics tailored towards our goal. Recently, Tubbicke and Vegetabile et al. extended entropy balancing initially proposed by Hainmueller [47] to the continuous setting [123, 128]. **Entropy Balancing for Continuous Treatment (EBCT)** seeks to achieve balance with weights that have minimum deviation from a pre-specified standardized base weight q, which can be uniformly $1/n$ or can come from expert knowledge. Similar to npCBPS, the Lagrangian can be written as

$$\underset{w\in\mathbb{R}^n}{\arg\min}\, \mathcal{L}(w_i, \lambda, \gamma) = \sum_{i=1}^{n} w_i log(w_i/q_i) + \lambda\Big(1 - \sum_{i=1}^{n} w_i\Big) + \gamma^\intercal\Big(\sum -w_i g(X_i^*, T_i^*)\Big). \tag{4}$$

EBCT enjoys a faster convergence rate and avoids local extremes compared with GBM and npCBPS due to the convexity of the Shannon entropy [123]. (npCBPS may be nonconvex as the empirical likelihood is typically nonconvex.) In addition, it allows increased precision for ADRF estimation, as EBCT intentionally avoids extreme weights by design. Interestingly, it has been shown in a binary action setting that Entropy Balancing (EB) enjoys the *Double Robustness* property by drawing a connection to the primal and dual problem in the field of optimization. When the Propensity Score is estimated by solving the dual of EB and the outcome regression is a linear model in the moment functions of the covariates, the DR estimator is the same as the EB estimator [137].

By reducing the problem of GPS modeling to the designation of loss, we can bridge balancing weight estimation with extreme points search, thereby joining perspectives from empirical likelihood [94], calibration estimation [31], and statistical decision theory [90]. The idea of loss minimization also unites previous methods. As we can see, MLE and MOM are all solving zero expectation functions, which, if put in another way, is finding the point that minimizes the average regret (or maximizes the average likelihood) over all possible parameter choices within some specified space. Boosting, a function approximation technique, adds a new approximation to the current one in the direction of the steepest gradient that minimizes regret (or maximizes function fit). Zhao et al. made explicit these connections in the binary action setting and pointed out several advantages of using a loss function tailored towards the estimand. Specifically, we will be able to control the amount of allowed imbalance flexibly, estimate PS in high dimensions by adding regularization terms, and evaluate solution uniqueness by investigating the type of convexity of the loss function [136]. We might also be able to include other desirable properties such as distribution robustness into this framework.

As we can see, moving away from a parametric linear model, the latter proposed GPS estimation methods all incorporate covariate balance during estimation. Moreover, they all try to mitigate the issue of model misspecification by borrowing strength from Machine Learning (ML) and semi- or non-parametric methods. This shift is not unique to GPS estimation.

4.2 Imputing the Missing Outcomes

Besides applying weights to the observed data, we can also use modeling to fill in the missing potential outcomes. For ATE $\tau(t) := \mathbb{E}[Y(t)]$ we can expand it as $\mathbb{E}[\mathbb{E}[Y(t)|X]] = \mathbb{E}[\mathbb{E}[Y|T = t, X]]$ with the outer expectation taken over X and the equality holding due to causal consistency. As the probability X can be estimated non-parametrically, we need only to model $\mu(X, T) := \mathbb{E}[Y|T, X]$. After fitting this outcome model, for an individual with any X, we can fix the value of T, then integrate over X to obtain the average treatment effect (ATE). If the true underlying model is within the class of specified functions, we can obtain a good estimate of the ATE.

4.2.1 Outcome Modeling

Unlike the reweighing methods (Section 4.1.1) that separate confounding adjustment and effect estimation into two steps, ATE estimation via outcome modeling merges the two steps into one, posing a challenge for balance diagnostics. In addition, parametric modeling would be problematic if we have little information on how X affects T concerning $Y(t)$. Using a binary action to illustrate, if we specify a linear form for $\mu(X, a) := \mathbb{E}[Y|A = a, X]$, we should explicitly specify interaction terms between A and X for confounding adjustment. If not, $\mu(X, a) := \mathbb{E}[Y|A = a, X] = \beta^\intercal a + \alpha^\intercal X$. $\mathbb{E}[\mathbb{E}[Y(1) - Y(0)|X]] = \mathbb{E}[\mu(X, 1) - \mu(X, 0)] = \beta$. β would be the estimated causal risk difference, and a change in X would have no effect in β. Although such a model might maximizes goodness-of-fit, the model would give us biased estimates if there exists confounding which is not sufficiently adjusted. As one might expect, more robust methods incorporate confounding adjustment, often through the estimated Propensity Scores or utilize more flexible algorithms to fit the outcome model. (See Section 5 for more details).

Among many uses of the Propensity Scores, **GPS-Adjusted Regression**[9] easily extends

[9] One can also categorize this as a GPS-based method. It models the outcome surface based on the GPS.

to the continuous action setting. As a consequence of the Balancing Property, one can show

$$\begin{aligned} g(t,r) &= E[Y(t)|r(X,t)=r] = E[Y(t)|r(X,t)=r, T=t], \\ \tau(t) &= E[g(t,r(X,t))], \forall t \in \mathcal{T}. \end{aligned} \tag{5}$$

The above equations suggest a two-step nested approach, where the first step is standard Generalized Propensity Score $r(X,t)$ estimation using any method mentioned in Section 4.1.2. Then, we can construct a response model for $g(t,r)$ to estimate the average outcome at t adjusted by the estimated GPS. [10] Finally, we can average $g(t,r(X,t))$ over the distribution of X to obtain the desired Average Dose-Response Function (ADRF).

The idea of imputing the missing potential outcomes relates closely to the estimation of conditional average treatment effect (CATE), which we will discuss in Section 5. As we can see, the last step of ATE estimation via imputation is often an integration over the distribution of X. If the conditional estimate before the integration is an unbiased estimate for $\mathbb{E}[Y(t)|X]$, this would be a good estimate for CATE, and if we additionally average over X, we would have an estimate for ATE.

4.3 Combined Models and semi-parametric Methods

As we have seen in previous sections, reweighting-based and imputation-based methods usually have a two-stage nested procedure. For reweighting-the-observable type, the second stage of treatment effect estimation depends on first stage GPS modeling of $r(X,T) := \mathbb{E}[T|X]$. For imputing-the-unobservable type, the second stage depends on first stage outcome modeling of $\mu(X,T) := \mathbb{E}[Y|X,T]$. A category distinct from the two above includes methods such as the Doubly Robust (DR) estimators [67], Double/Debiased Machine Learning (DML) [22, 25, 71], and Targeted Maximum Likelihood Estimation (TMLE) [32]. From one perspective, they all combine more than one type of nuisance function—DR combines $\mu(X,T)$ and $r(X,T)$; DML combines $m(X,T) := \mathbb{E}[Y|X]$ and $r(X,T)$—to attain a more robust estimator. From the semi-parametric estimation perspective, we can see that the asymptotic equivalence of these three methods is not a coincidence but because they are

[10] Although we can obtain $\widehat{g}(t,r)$ for any fixed X, $\widehat{g}(\cdot)$ does not have a causal interpretation. It is because $\widehat{g}(\cdot)$ is fit on the population with a similar estimated GPS, a sub-population different from our target population.

trying to achieve the same thing, just through different means.

4.3.1 The von Mises Expansion

Before pointing out the connection, we introduce the notations we will use and, in a general way, explain some important concepts for semi-parametric estimation[11]. We denote P the *true* observed distribution where our observation comes from, so $O = (X, T, Y) \sim P$. We also assume the true distribution lies within a *model*, that is, a set of distributions, denoted $\mathcal{M}$. If we believe the class of data distribution can be characterized by a finite dimensional parameter, this class of densities $\mathcal{M} = \{f(\theta), \theta \in \mathbb{R}^d\}$ would be a finite-dimension parametric model. If we expand our class of models to $\theta = (\beta^\intercal, \eta^\intercal)^\intercal$ where β is finite-dimensional but η is infinite-dimensional, these models would be called *semi-parametric models*. Since η is not of primary interest, we refer to it as the *nuisance parameter*. If we do not put any restriction on the functional form of the densities and express it as a generic function, these models would be *non-parametric*. We would have an estimand of interest based on the underlying data generating distribution. Hence, we can denote the estimand as a functional $\Psi : \mathcal{M} \to \mathbb{R}$. Our goal is to therefore estimate $\Psi(P)$. In a semi-parametric model, if β is of interest, we can express $\beta = \Psi(P_{\beta,\eta})$. In the following part we have $\beta(t) := \mathbb{E}[Y(t)]$. η is often one or more elements in the set $\{r, \mu, m\}$. Since we do not know the true observed data distribution P and must estimate it through i.i.d. observations $\{O_i\}_{i=1}^n$, we denote the estimator for P as $\widehat{P}_n$, and the estimator for Ψ as $\Psi(\widehat{P}_n) = \widehat{\Psi}$.

How well the estimator can possibly be, quantified by a lower bound on the estimation error? In a parametric setting with finite-dimensional θ, the Cramer-Rao Lower Bound (CRLB) [20] is such a metric. For *any* unbiased estimator $\widehat{\Psi}$, $Var_\theta(\widehat{\Psi})$ is greater than or equal to the CRLB $\Psi^{'}(\theta)^2 / Var\{s_\theta(O)\}$, where $s_\theta(O) = \frac{\partial}{\partial\theta} log_\theta L(O)$ is the score function, the derivative of log likelihood with respect to the parameter θ: $\Psi^{'}(\theta) = \frac{\partial}{\partial\theta}\Psi(\theta)$. This metric, as we can see by its form, can be interpreted as the variance of Ψ with respect to θ, corrected by the variance of the observed data with respect to θ. In our setting, the analogue of the CRLB

[11] We refer readers to [68] for a more formal introduction to related terms and concepts.

would be $\Psi^{'}(P_\epsilon)^2/Var_{P_\epsilon}\{s_\epsilon(O)\}$, where the numerator

$$\Psi^{'}(P_\epsilon) = \frac{\partial}{\partial\epsilon}\Psi(P_\epsilon)|_{\epsilon=0} = \int \phi(O;P)s_\epsilon(O)dP(O). \tag{6}$$

$\mathcal{M}_\epsilon = \{P_\epsilon : \epsilon \in \mathbb{R}\}$ is a *parametric submodel* that lies within our model of interest $\mathcal{M}$ and equals P, the true distribution, at $\epsilon = 0$. By this definition, we can see that ϵ is the magnitude of some perturbation of P. Specifically, we can write $P_\epsilon = P(O)(1+\epsilon h(O))$ where $h(\cdot)$ is some mean-zero function providing the direction of perturbation. $\phi(\cdot)$ is known as the *influence function*. Examining the form of Equation 6, we can see that this is a derivative measuring the sensitivity of the functional Ψ with respect to ϵ. If this derivative exists for all $h(\cdot)$, we say the functional Ψ is *pathwise differentiable*. Since the influence function is in essence a derivative, we can interpret it using our knowledge of the function derivative: being differentiable implies the 'function' has finite rate of local change and does not go wild.

One immediate benefit of Ψ being pathwise differentiable is that we can expand the true $\Psi(P)$ at $\Psi(\widehat{P}_n)$ using a form of Taylor expansion[12] (in the context of distributions) and evaluate its asymptotic property using Slutsky's Theorem by considering respectively different terms in the expansion. This Taylor expansion analogy is commonly known as the *von Mises expansion*:

$$\begin{aligned}
\sqrt{n}(\widehat{\Psi} - \Psi) &= \sqrt{n}\int \phi(O_i, \widehat{P}_n)d(\widehat{P}_n - P)(O) + R_2(\widehat{P}_n, P) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\phi(O_i, P)\} - \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\phi(O_i, \widehat{P}_n)\}}_{\text{Plug-in bias}} \\
&+ \underbrace{\sqrt{n}(P_n - P)\{\phi(O, \widehat{P}_n) - \phi(O, P)\}}_{\text{Empirical Process Term}} + \underbrace{R_2(\widehat{P}_n, P)}_{\text{Remainder}}.
\end{aligned} \tag{7}$$

With unbiasedness of the canonical gradient, the first term $1/\sqrt{n}\sum_{i=1}^{n}\phi(O_i, P)$ will converge in probability to $\mathcal{N}(0, Var(\phi(O,P)))$ by the Central Limit Theorem. Interestingly, $Var(\phi(O,P)) = \mathbb{E}[\phi\phi^\intercal]$ is the non-parametric analogy of CRLB for Ψ. This can be found

[12]Recall that for one variable functions, if its derivatives exists, $f(x) = \sum_{i=0}^{\infty}\frac{f^n(a)}{n!}(x-a)^n$, and writing out the first two terms will give us the linear approximation $f(x) = f(a) + f^{'}(a)(x-a) + R_2$. Any other terms involving ≥ 2 order derivative are ignored. Rearrange this to be consistent with Equation 7: $\{f(x) - f(a)\} = f^{'}(a)(x-a) + R_2$.

by taking the supremum of all CRLBs at $\epsilon = 0$[13]. 'All' since we can perturb P in all directions provided by h. Thus, if we can control the remaining terms to converge to zero in probability of order $o_p(1/\sqrt{n})$[14], we can minimize the asymptotic variance of an estimator to the best possible and obtain this variance using the *efficient influence function* ϕ of the functional Ψ, efficient in the sense that it achieves the smallest asymptotic variance possible among all influence functions. [15]

Now consider the remaining terms on the right-hand side: The empirical process term relates to the complexity of the class of influence functions ϕ and our estimand of interest. We can control this by either assuming they are within a Donsker class or using sample-splitting and cross-fitting to relax the Donsker condition[16]. The remainder term can often be controlled by controlling the convergence rate of nuisance parameters, though its specific form would vary case-by-case. We cannot easily obtain the asymptotic behavior of the second term $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\phi(O_i, \widehat{P}_n)\}$ due to our limited knowledge on $\widehat{P}_n$. We would need to eliminate this drift term, also known as the plug-in bias, in order to reach the CRLB. This objective can be achieved in different ways, and this difference gives rise to the differences between the Doubly Robust (DR) estimator, the Double/Debiased Machine Learning (DML) estimator, and the Targeted Maximum Likelihood Estimator (TMLE).

4.3.2 Controlling the Plug-in Bias

The best estimator concerning minimum asymptotic variance can be characterized by its efficient influence function, a kind of derivative, under some other regulatory conditions. If this derivative exists, this problem boils down to eliminating the plug-in bias (second term

[13] $sup_{P_\epsilon} \frac{\Psi'(P_\epsilon)^2}{Var(s_\epsilon(O))} = sup_h \frac{\mathbb{E}[\phi(O,P)h(O)]}{\mathbb{E}\{h(O)^2\}} \leq \mathbb{E}\{\phi(O,P)^2\} = Var\{\phi(O)\}$, with equality when $h(O) = \phi(O,P)$, so in this case the influence function $\phi(O,P)$ is a zero-mean function.

[14] For a random variable X_n, $X_n = o_p(a_n)$ if and only if $X_n/a_n \xrightarrow{p} 0$, where p denotes converge in probability. Thus, a_n provides an upper bound for the convergence rate of X_n.

[15] Efficiency is closely related to orthogonality. In fact, all regular asymptotically linear estimators have ϕ residing in the space orthogonal to the nuisance tangent space $\mathcal{T}_\eta$, a linear span of the score vectors of η, and satisfy $P(\phi s_\phi) = 1$. We refer readers to [122] for a more detailed discussion and for a nice geometric interpretation of the topic. For a great introduction on the construction of the efficient influence functions, see [52].

[16] See [74] for a more detailed introduction on the Donsker condition. Briefly, this condition requires that the smallest class of functions containing the estimators of nuisance parameters has a bounded entropy integral. It is a necessary condition for the empirical process $\sqrt{n}(P_n - P)$ to converge to some object as $n \to \infty$. If the entropy does not increase with N too rapidly, it is possible to still have the term vanish without the Donsker condition. [14]

in Equation 7). Under a continuous action setting, things are more complicated. This is because the functional for our estimand of interest $\Psi(t) = \mathbb{E}[Y(t)]$ is *no longer pathwise differentiable*. Accordingly, the differences between DR, DML, and TMLE with continuous action are (1) how they eliminate the plug-in bias term and (2) how they solve the issue of Ψ being pathwise non-differentiable.

The Doubly Robust (DR) estimator eliminates the plug-in bias by simply rearranging the plug-in bias to the left-hand side of Equation 7, incorporating it into the estimator which will now be $\Psi(\widehat{P}_n) + \frac{1}{n}\sum_{i=1}^{n}\phi(O_i, \widehat{P}_n)$. This method is known as the *one-step estimator*. In the binary action case, the estimator for Average Treatment Effect resolves to:

$$\widehat{\beta}_{DR}(a) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\mathbb{1}(A_i = 1)}{\widehat{r}(X_i, a)}\{Y_i - \widehat{\mu}_a(X_i)\} \quad + \quad \widehat{\mu}_a(X_i) \Big). \tag{8}$$

Here $r(X, T = 1) = e(X) = Pr(A = 1|X = X)$ and $r(X, T = 0) = 1 - e(X)$. The propensity weighted portion can be regarded as the bias adjustment on the plug-in estimator $1/n \sum \widehat{\mu}_a(X_i)$. Recall that the Propensity Score (PS) encodes information on confounding, so PS for adjusting the plug-in bias is also adjusting for confounding bias in some way. We can now explain the widely known *Doubly Robust* property by rewriting $\beta_{DR}(a)$ as $\mathbb{E}[Y(a) + \{\mathbb{1}(A = a) - r(X, a)\}\{Y(a) - \mu(X, a)\}/\{r(X, a)\}]$. If either $r(\cdot)$ or $\mu(\cdot)$ converges in probability to the true parameter, the second part will shrink to zero, resulting in a consistent estimator for the ATE [17]. Since this provides us with two chances of having a correct class of model, it is given the name Doubly Robust. Nonetheless, one should be aware that robust is a general term, and there can be other parameters that are *doubly* robust if, for example, their robustness also requires only one of the two nuisance parameters to have desired properties. Besides robustness in consistency, we can also show robustness in convergence rate. Using the Cauchy-Schwarz inequality [18], we can bound the remainder by

$$\sqrt{n}\mathbb{E}_P\left[\left\{\frac{r(O, P)}{\widehat{r}(O, \widehat{P}_n)} - 1\right\}^2\right]^{1/2} \mathbb{E}_P\left[\left\{\mu(O, P) - \mu(O, \widehat{P}_n)\right\}^2\right]^{1/2}.$$

[17] To be more explicit, for the treatment model, after iterative expectation, the inner expectation over X would be $\mathbb{E}[\mathbb{1}(A = a)|X] = Pr(A = 1|X) = r(X, 1) = e(X)$. So if $\mathbb{E}[\mathbb{1}(A = a)|X] = e(X)$, the second term will be zero.

[18] In a more familiar form, in a normed space, for vector u and v, this would be $\|u, v\| \leq \|u\| + \|v\|$. A special case would be the triangle inequality. In probability theory, for random variables X and Y, this would be $|\,\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\,\mathbb{E}(Y^2)$.

Since this is a product of the error from two nuisance parameters, we can allow one of them to converge at a slower rate as long as the other one converges fast enough. In the continuous action setting, Kennedy et al. [67] extend the binary DR estimator using the efficient influence functions of not the pathwise non-differentiable parameter $\beta(t)$ but a regular semi-parametric parameter $\Psi = \mathbb{E}[\xi(O, r, \mu)]$, where the function ξ satisfies $\mathbb{E}[\xi(O, r, \mu)|T = t] = \int_{\mathcal{X}} \mu(X, t)dP(X) =: \beta(t)$, our parameter of interest, if only one of r and μ is correct. If we can find this function ξ, we can simply regress the estimated $\widehat{\xi}$ on T based on the estimated nuisance parameters. Observe that if the above holds, ξ has to satisfy $\mathbb{E}[\xi(O, r, \mu)] = \Psi = \int_{\mathcal{T}} \int_{\mathcal{X}} \mu(X, t)\omega(t)dP(X)dt$, where $\omega(t) = \frac{\partial}{\partial t}P(T \leq t)$. Since for influence functions, $\mathbb{E}[\phi(O)] = 0$, we can evaluate the influence function for Ψ and use components of the influence function to construct ξ. With some calculation, Kennedy et al. find the expression for $\xi(O; r, \mu)$ to be

$$\xi(O; r, \mu) = \frac{Y - \mu(X, T)}{r(X, T)} \int_{\mathcal{X}} r(X, T)d\widehat{P}_n(x) + \int_{\mathcal{X}} \mu(X, T)d\widehat{P}_n(x),$$

and show it satisfies the Double Robustness property. Similar to the procedure for the binary action DR estimation, we fit nuisance parameters $\widehat{r}$ and $\widehat{\mu}$, obtain $\widehat{\xi}$, and regress it on T. Kennedy et al. provided convergence rate robustness using non-parametric kernel regression as the final step for $\Psi(t)$. Similar to the binary action setting, the convergence rate $O_p(1/\sqrt{nh} + h^2 + r_1(t)r_2(t))$ [19] involves the product of the local error rate from the two nuisance parameters, but in addition to that, it also additively relates to the Kernel bandwidth h. Because this is a *plug-in* estimator, a major drawback is that if there are extreme values in the estimated GPS (near violation of the second assumption), the estimate can be out-of-range. One way to correct this is to subsequently use the idea of 'targeting', which we will see in TMLE [32] and VCNet [88] discussed later in this section.

The **Double/Debiased Machine Learning (DML)** estimator eliminates the plug-in bias by forcing $\frac{1}{n}\sum_{i=1}^{n} \phi(O_i, \widehat{P}_n) = 0$ into the estimating equation for Ψ. As a result, solving $\widehat{\Psi}$ automatically corrects the plug-in bias. In the binary action setting, DML can be traced back to the work of Frisch–Waugh–Lovell [39, 82] with a partially linear[20] model setup. Robinson

[19] For any sequence of random variables X_n, $X_n = O_p(a_n)$ is equivalent to $X_n/a_n = O_p(1)$, meaning that X_n/a_n is asymptotically bounded by probability. Recall $X_n = o_p(a_n)$ then $X_n/a_n = o_p(1)$, that is, X_n/a_n converge in probability to zero.

[20] 'Partially' because while $\beta^\intercal A$ is linear, $g(\cdot)$ can be a non-linear, complex function.

extended DML by using non-parametric kernel regression for the estimation of nuisance parameters [104]. Although we can use a non-linear or even a generic model with the idea of DML, this partially linear model shown below has been widely used as an introduction to DML since linearity ties closely with the concept of projection and orthogonalization:

$$Y = \beta^\intercal A + g(X) + U_y, \qquad A = e(X) + U_a, \text{ where}$$
$$\mathbb{E}[U_y|X, A] = 0, \qquad \mathbb{E}[U_a|X] = 0.$$

A natural way to recover β would be to first regress respectively A on X and Y on X. Then $\widehat{\beta} \leftarrow lm((Y - \widehat{\mathbb{E}}[Y|X]) \sim (A - \widehat{\mathbb{E}}[A|X]))$. For an intuitive explanation, by regressing A on X, we are finding the best projection that maximally captures the influence of X on A. Thus, subtracting the projection from the original A results in a component of A orthogonal to the space of X. For two random variables, orthogonality can be characterized by $\mathbb{E}[\tilde{A}X] = 0$. This can be justified by considering $\mathbb{E}[\tilde{A}X]$ the analogy of an inner product $\langle u, v\rangle$ for any two vectors u and v in a vector space, and recall our more familiar notion of orthogonality for vectors. Through this adjustment, using the subscript 0 to denote the true value, we want to achieve

$$\partial_\eta \, \mathbb{E}[\Psi(O, \beta_0, \eta_0)][\eta - \eta_0] = 0, \tag{9}$$

meaning that the estimating equations for the identification of β are locally insensitive, or uncorrelated, to the variation of the nuisance parameter $\eta = (g^\intercal, e^\intercal)^\intercal$. Equation 9 is also known as the *Neyman Orthogonality condition*. See Figure 5 below for a visual illustration of projection, orthogonization, and de-confounding.

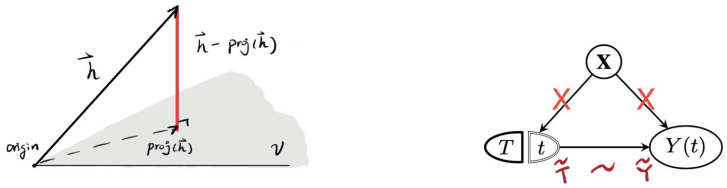

Figure 5: Left: h is a vector and $\mathcal{V}$ is a linear subspace. Illustration of projection and how subtracting the projection corrects the original vector to be orthogonal. Right: SWIG representaion of DML. Orthogonalization is a form of de-biasing. $\tilde{T}$ and $\tilde{Y}$ are the orthogonalized, or debiased version of the original random variable.

In a parametric model with $\theta = \beta$ finite-dimensional, we can solve β using Maximum Likelihood Estimation, by setting the score $s_\beta(O) = \frac{\partial}{\partial\beta} logL(O;\beta)$ to zero. In a semi-parametric model, however, the original score function $s_{\theta=(\beta^\intercal,\eta^\intercal)^\intercal}(O)$ does not naturally satisfy the Neyman orthogonality condition. Due to a lack of knowledge on η, we need to orthogonalize the scores by subtracting off its projection onto the linear span of scores of the nuisance parameter η (the tangent space). After this transformation, the new score function would satisfy the Neyman orthogonality condition and be an efficient score function that we can use to solve Ψ. This new efficient score is also known as the *Neyman orthogonal score*. Colangelo and Lee give the earliest use of DML with continuous action [25] assuming a non-parametric outcome equation $Y = g(X, T, \epsilon)$ instead of the partially linear model discussed above. The estimator is defined to be

$$\widehat{\beta}_{DML}(t) = \frac{1}{n}\sum_{i=1}^{n}\{\frac{K_h(T_i - t)}{\widehat{r}(X_i, t)}(Y_i - \widehat{\mu}(X_i, t)) + \widehat{\mu}(X_i, t)\}, \tag{10}$$

where $K_h(T_i - t)$, as we have seen in the GPS estimation section, is a kernel of bandwidth h [21]. Neyman orthogonality condition holds as $h \to 0$. Hence, the general procedure for DML estimation would be to first train nuisance parameters r and μ using flexible ML methods with cross-validation, then use the estimator to obtain estimates at t [22]. As mentioned earlier, the GPS in the denominator can cause the estimate to be unstable if it has extreme values. To tackle this issue, Klosin [71] extends the Auto-DML first proposed by Chernozhukov et al. [22] to the continuous action setting. Instead of inversely using $\widehat{r}$ as weights, it estimates the balancing weight which we will denote as α directly and avoids implicitly trimming[23] the extreme GPS. Specifically, from the debiased estimating equation $\Psi(O_i, \beta, \mu, \alpha) = \mu(X_i, t) - \beta + K_h(T_i - t)\alpha(X_i, t)(Y_i - \mu(X_i, t))$, we can obtain

$$\widehat{\beta}_{Auto-DML}(t) = \frac{1}{n}\sum_{i=1}^{n}\{K_h(T_i - t)\widehat{\alpha}(X_i, t)(Y_i - \widehat{\mu}(X_i, t)) + \widehat{\mu}(X_i, t)\}, \tag{11}$$

where $\widehat{\alpha}$ is modeled as $\widehat{\alpha} = b(X)\widehat{\rho}_t$ where $b(X)$ is a dictionary of functions. And $\widehat{\rho}_t$ is

[21] See section 2.14 in [48] for a discussion on how to optimize kernel bandwidth.

[22] One might realize the form of the estimator is the same as the DR estimator for binary action besides replacing the indicator functions with kernels, and has a form similar to the DR estimator proposed by Kennedy et al. As Hines et al. showed in Example 5 in [52], when treatment is discrete, the one-step approach and the estimating-equation approach arrive at the same estimator.

[23] In this context, trimming is referring to using indicators or kernels to indicate which estimated points to include and which to not.

obtained by solving a Lasso problem. Noticed that this idea is similar to the Covariate Balancing Propensity Score (CBPS) and the Entropy Balancing for Continuous Treatment (EBCT) designed for balancing weight estimation (Discussed in Section 4.1.2). Simulations in [71] found that the resulting estimator does have better finite sample properties in terms of root-mean-square error.

The **Targeted Maximum Likelihood Estimator (TMLE)** [32], somewhat different from the two above, *retargets* the distribution estimate $\widehat{P}_n$ to $\widehat{P}_n^*$ such that the plug-in type estimator $\frac{1}{n}\sum_{i=1}^n \phi(O_i, \widehat{P}^*)$ on this adjusted distribution has zero plug-in bias. Within a causal framework, we denote F_0 the true distribution of the *full* data for the counterfactual process $(Y(t) : t \in \mathcal{T})$. That is, $X = \{T, (Y(t) : t \in \mathcal{T})\} \sim F_0$. When treatment T is continuous, an important consequence is that our estimand $\Psi(P)$ would no longer be *pathwise differentiable*. This motivates the use of loss, or regret, since we can construct a mapping $R(\psi, \cdot)$: $\mathcal{M} \to \mathbb{R}$ that is pathwise differentiable to circumvent the non pathwise differentiability of $\Psi(P)$. We can therefore seek root-n-consistency for this loss under reasonable regularity conditions. Construction of this mapping involves the idea of *targeting*. Specifically, we want the loss such that, by minimizing it, it gives us the true estimate using counterfactual full data. In other words, we want to set up the loss such that $R(\psi, P) = \mathbb{E}_P L_{Q(P)}(O, \psi)$ identifies $R^f(\psi, F_0) = FL^f(\psi)$ and sets us back to a hypothetical Randomized Controlled Trial. Using superscript f to denote the full data of the counterfactual process, we can formally express this idea as:

$$\Psi(F_0)(t, X) = \mathbb{E}_{F_0}(Y(t)|X) = \arg\min_{\Psi} R^f(\Psi, F_0).$$

This loss can be constructed as the distributional squared loss $L(\Psi)(X) = \int_{\mathcal{T}}\{Y(t) - \psi(X,t)\}^2 h(X,t) d\nu(t)$ base on different estimators, for example, from inverse-probability weighting (IPW), outcome regression, or using the Doubly Robust (DR) estimator (h is a non-negative function such that $\int f d\nu = 1$ and ν is some σ-finite measure.). Here we present the loss based on the DR estimator (or based on the efficient influence function of $R(\psi, P)$):

$$\begin{aligned} L_{Q,r}(O, \psi) = &\frac{h(X,T)}{r(X,T)}[\{Y^2 - Q_2(X,T)\} - 2\psi(X,t)\{Y - Q_1(X,T)\}] \\ &+ \int_{\mathcal{T}} \{Q_2(X,t) - 2Q_1(X,t)\psi(X,t) + \psi^2(X,t)\} h(X,t) d\mu(t), \end{aligned}$$

where $Q_1(T,X) := \mathbb{E}[Y|T,X]$, $Q_2(T,X) := \mathbb{E}[Y^2|T,X]$. In comparison with the DR estimator, Diaz and van der Laan point out that the TMLE is more stable for near violation of Positivity and typically has better finite sample performance. Asymptotic normality of $\widehat{R}(\Psi)$ depends on the convergence rate of the propensity model $\widehat{r}$, of the ones related to the outcome models $\widehat{Q}_1$ and $\widehat{Q}_2$, and of $\widehat{\Psi}$. Unlike the DR estimator, where the convergence rate of the GPS model and the outcome model weighs the same, TMLE *requires* the GPS model to converge to the truth at rate $n^{-1/4}$ or faster. If this is satisfied, $\widehat{Q}_1$, $\widehat{Q}_2$, and $\widehat{\Psi}$ need to converge to some limit, not necessarily the truth, at some rate such that the product of this rate and that of the GPS model does not exceed $n^{-1/2}$.

For all three methods above, we can use cross-validation to relax the Donsker condition on the influence function and enable the use of highly flexible techniques such as non-parametric and ML methods to estimate the nuisance parameters. In fact, all three papers explicitly mentioned the use of cross-validation in their simulations. In [32], Diaz and van der Laan use the Super Learner that is based on the cross-validated risk to estimate r and μ [24]. Theoretically, other types of cross-validation can also be used.

The latest proposed method—**Varying Coefficient Neural Network (VCNet)** [88]—is a method that combines deep Neural Networks (NNs) with semi-parametric estimation. By extending Targeted Regularization to the estimation of a function, CVNet can estimate a continuous ADRF Ψ as a curve rather than the points $\Psi(t)$ on the curve. It uses the Doubly Robust (DR) estimator as a starting point, then adds a targeting component during loss construction to correct the bias. μ^{NN} and r^{NN} are trained using NNs. For μ^{NN}, VCNet used the full sample jointly to estimate the Generalized Propensity Score (GPS) $r^{NN}(X,T)$ across the standardized treatment interval $\mathcal{T} = [0,1]$. Specifically, this is performed by first discretizing $[0,1]$ into B bins, estimate $r^{NN}(X)$ for each bin, and finally fill in the remaining points through linear extrapolation: $r^{NN}(X,t) = r^{NN}(X,t_1) + B(r^{NN}(X,t_2) - r^{NN}(X,t1))(t-t_1)$ where t_1 and t_2 are the endpoints of the bin B_t. The estimated curve is then rescaled to be a valid density. The NN trained for $r^{NN}(X)$[25] not only estimates the propensity function but also helps to extract the feature which serves as an input for

[24] Refer to Equation (2) for the expression for the estimation of the GPS r. The cross-validated risk for the outcome model has the form $\mathcal{L}(\widehat{\mu}) = \frac{1}{V}\sum_{v=1}^{V}\frac{1}{n_v}\sum_{i\in S_v}(Y_i - \widehat{\mu}_{j,\widehat{v}}(X_i))^2$. The final step—taking a weighted average—is the same as that of the GPS model

[25] Since it is a function across $\mathcal{T}$, here we ignore t in the input.

$\mu^{NN}(X,t) = f_{\theta(t)}(Z)$, where Z is the output of $r^{NN}(X)$ prior to feeding in the activation function; $f_{\theta(t)}$ is a deep NN with weights θ varying with t, modeled as a linear combination of spline bases such as the B-spline [26]. Now we can plug in our estimate into the DR estimator (Equation 8 but changing the PS to GPS.), but a condition for this estimator to be robust is that the estimated GPS should be lower-bounded by some small positive constant[27]. To solve this discrepancy of a one-step estimator, Nie et al. extends Targeted Regularization, the idea used in TMLE, to eliminate the plug-in bias. For $\Psi(t_0) = \mathbb{E}(Y(t))$, its efficient influence function can be expressed as

$$\phi_{t_0}(Y,X,T,\mu,r,\Psi) = \underbrace{\delta(T-t_0)\frac{Y-\mu(X,T)}{r(X,T)}}_{plug-in bias} + \mu(X,t_0) - \Psi(t_0), \tag{12}$$

where δ is the Dirac delta function. The Functional Targeted Regularization (FTR) [28] is defined as

$$\mathcal{R}_{FTR}[\mu^{NN}, r^{NN}, \epsilon_n] = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mu^{NN}(X_i,t_i) - \frac{\epsilon_n(t_i)}{r^{NN}(X_i,t_i)}\right)^2. \tag{13}$$

Adding $\mathcal{R}_{FTR}$ to the regular goodness-of-fit loss of the treatment and the outcome model forms the final loss objective

$$\begin{aligned}\mathcal{L}_{FTR}[\mu^{NN}, r^{NN}, \epsilon_n] = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu^{NN}(X_i,t_i))^2 - \frac{\alpha}{n}\sum_{i=1}^{n} log(r^{NN}(X_i,t_i)) \\ + \beta_n \mathcal{R}_{FTR}[\mu^{NN}, r^{NN}, \epsilon_n].\end{aligned} \tag{14}$$

We denote the minimizer of this loss as $(\widehat{\mu}, \widehat{r}, \widehat{\epsilon})$. Using the FTR, Nie et al. point out that if we let the penalization grow with n such that $\beta_n = o(1)$, assume r^{NN} and r are uniformly bounded, and assume some other regularization conditions, we can correct the plug-in bias to be approximately zero. For the estimator $\widehat{\Psi}(\cdot) := \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}(X_i,\cdot) + \frac{\widehat{\epsilon}_n(\cdot)}{\widehat{r}(X_i)})$, we then can show $\|\widehat{\Psi} - \Psi\|_{L^2} = O_p(n^{-1/3}\sqrt{log(n)} + r_1(n)r_2(n))$, where $\|\widehat{r} - r\| = O_p(r_1(n))$ and

[26] If one uses the non-overlapping indicator as basis and models only μ, it will be the structure of DRNet. With the B-spline, the output will be a smooth function. The B-spline is the basis functions for the spline function space and is widely used in computer graphics. See [29] for more.

[27] This is also in the assumption of the DR estimator from Kennedy et al. and is a necessary condition for the estimator to be doubly robust.

[28] When the third term in $\mathcal{R}_{FTR}$ is $\epsilon(t_i/r^{NN}(X_i,1))$, this would be the targeted regularization term under a binary action setting, which, if the function space for nuisance parameters μ and r are of finite complexity, $\frac{\partial}{\partial\epsilon}\mathcal{R}[\widehat{\mu},\widehat{r},\epsilon]|_{\widehat{\epsilon}} = 0$, and consistency can be achieved.

$\|\widehat{\mu} - \mu\| = O_p(r_2(n))$. [29] Similar to TMLE, the goal of adding a targeted regularization is to eliminate the second term, but different from TMLE, the regularization is not obtained *after* estimation of the nuisance parameters but is jointly optimized *with* the regularization. In comparison with the DR estimator proposed by Kennedy et al., the estimand is a smooth function Ψ across $\mathcal{T}$ instead of a point at the ADRF $\Psi(t) := \mathbb{E}[Y(t)]$.

5 Estimation of Conditional Average Treatment Effect (CATE)

As discussed in Section 4.2, if we want to identify groups that benefit most from a treatment or find the optimal dosing for a given individual, we would need to estimate the Individual Dose-Response Function (I-DRF) $\tau(t, X) := \mathbb{E}[Y(t)|X]$. Unfortunately, there are only two proposed methods for I-DRF estimation, all using Neural Networks as building blocks. However, to shed light on the connections between ATE and CATE estimation and to have a complete structure for CATE estimation, we will also present methods for binary action as a category of estimation method if there is currently no extension into the continuous action setting.

We can categorize CATE estimation algorithms by how they use Machine Learning (ML) algorithms to adapt to the causal framework. (1) We can modify components of an ML algorithm, or (2) use any ML algorithms as building blocks or base algorithms to construct a structure on top of them, forming a meta-algorithm. Causal Forests (CFs) [130], Causal Bayesian Additive Regression Trees (BART) [45], Boosting [100], and Neuro Network (NN)-Based methods including the Dose-Response Network (DRNet) [112] and eStimating the effects of Continuous Interventions using Generative Adversarial Nets (SCIGAN) [15] fall into the first category. These two NN-based algorithms are the only ones that can estimate I-DRF to our knowledge. The different types of learners all fall into the second category (S-[38, 51], T-[49], X-[75], RA-[28], R-[89], DR-[69]). They relate closely to the three categories of ATE (or ADRF) estimation discussed in 3.6. We can choose the base learners for these meta-algorithms to contain a large pool of different learners. We can even use CFs or BART as one of the base learners. An increase in diversity increases the chance of containing the

[29] L^2 denotes the L^2-norm, uppercase L since it applies to functions, i.e. $\|f\|_{L^2} \equiv \langle f|f\rangle \equiv (\int \|f(t)\|_2^2 dt)^{1/2}$.

true underlying model for the meta-algorithms [75]. Unfortunately, all types of meta-learners are only for the binary action setting.

5.1 Modified-ML Algorithms

Machine Learning (ML) algorithms are generally for prediction and classification. It trains a model by loss and regularizes it to avoid over-fitting but fails to consider confounding adjustment, the core of Causal Inference. Nonetheless, we can modify ML to make it suitable for Causal Inference.

Causal Forests [130], a modified version of Random Forests, is an excellent example of such an algorithm. The main idea is to use Forests to find the 'nearest neighbors' to which a given X belongs. Here closeness is not quantified in terms of proximity in the original covariate but by being within a homogeneous subgroup. Then, within that homogeneous subgroup, we can use the ATE estimation methods discussed previously in this subgroup to estimate CATE. Similar to Random Forests, we define the estimate from a forest to be an average of the estimates from a set of trees $\widehat{\Delta}_{\mu}(X) = 1/K \sum_{k=1}^{K} T_k(X; \{O_i\}_{i=1}^{n})$. A tree $T(\cdot)$ consists of a set of branches or decision points. Each decision point is constructed by maximizing heterogeneity; that is, the difference between the two groups resulting from a split. To adapt to CATE estimation, a causal quantity, CFs used $\mathcal{L} = n_L n_R(\widehat{\tau}_L - \widehat{\tau}_R)^2$ as the measure of heterogeneity instead of $n_L n_R(\widehat{y}_L - \widehat{y}_R)^2$ that is commonly used in traditional Random Forests with $\widehat{y}$ being the estimated outcome using regression. Thus, we can use the methods mentioned earlier, such as Double/Debiased Machine Learning (DML), or residual-on-residual regression, to estimate the Average Treatment Effect τ on L and R, arrive at an estimate for treatment heterogeneity $\mathcal{L}$, and grow a tree by maximizing $\mathcal{L}$ for each branch. By a change of perspective demonstrated below, we can use Forests as a form of 'kernel' weight estimation to find homogeneous subgroups for a given X:

$$\tau(X) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n} Y_i \frac{\mathbb{1}(Y_i \in L_k(X))}{|L_k(X)|} = \sum_{i=1}^{n} Y_i \underbrace{\frac{1}{K}\sum_{k=1}^{K}\frac{\mathbb{1}(Y_i \in L_k(X))}{|L_k(X)|}}_{w_i(X),\ \text{weight for unit } i},$$

where $L_k(X)$ denotes the leaf in the tree T_k which X falls into. The first equal sign is

through the perspective that the final prediction is an average of the predictions from K trees. The second equal sign is through the perspective that the final prediction is a weighted average over n i.i.d. observations. Finally, we can then estimate CATE by using a weighted residual-on-residual regression $\widehat{\tau}(X) \leftarrow lm((Y_i - \widehat{m}(X_i)) \sim (A_i - \widehat{e}(X_i)), \text{weights} = \widehat{w}_i(X))$.

Besides trees, another critical class of ML algorithms is Neural Networks (NNs). They are highly flexible for estimating non-linear surfaces. A NN is usually composed of multiple levels of layers, where each layer contains multiple units. Like neurons, they accept inputs, reweight, and transform the result, then outputs using a non-linear activation function often denoted as σ. The set of responses is then fed to the next layer. We referred to a unit of output as *head* and the whole setup as the *architecture* of a NN. The earliest NN-based algorithms are an extension of outcome modeling. Instead of using parametric regression, these methods use NN to model the outcome surface. Recall, for linear outcome models, the effect of treatment could be easily lost as T has no unique role among X. One advantage of using the NN is that we can design its architecture to ensure we capture the effect of treatment from many covariates. Treatment Agnostic Representation Networks (TARNET) [113] with binary action is the earliest work in this category. It first uses a layer to learn a representation of X using the whole sample. The output, denoted $\Phi(X)$, is then separated into two groups and combined with treatment to predict the observed outcome Y while minimizing the distribution imbalance between the two feature spaces. **DRNet** [112] is the first to extend such structure to I-DRF estimation. It is composed of three layers: a representation learning layer using a complete sample, which is the same as that of TARNET; a treatment layer that discretizes $\mathcal{T}$ into blocks and, for each block, trains a head, using units whose received treatment falls into that block; finally, for each block, it trains a dosing layer that learns the Dose-Response Function. Like the other blocking-type methods, DRNet fails to incorporate the continuous nature of the I-DRF.

eStimating the effects of Continuous Interventions using Generative Adversarial Nets (SCIGAN) [15] modified Generative Adversarial Network (GAN) [40] to target the estimation of I-DRF. We can regard GAN [64] as a special type of two-player minimax game, with the two players being two Neural Networks. One is a generator, denoted G, taking a random variable and generating a *fake* counterfactual outcome. The other is a discriminator, denoted D, seeking to distinguish the *real* observed outcome between the generator's fake

potential outcomes. Since we know which outcomes are fake and which are not, we can calculate a supervised loss for the discriminator's decision and use this loss to train both the generator and the discriminator. It is a minimax problem because while the generator tries to minimize its loss by outputting better estimated potential outcomes, the discriminator tries to tell the two apart better, maximizing the loss. In SCIGAN, the notation is slightly different from the one we have used so far. Instead of letting the treatment be the dosing, SCIGAN denotes the received dosage in a tuple $t = (c, d) \in \mathcal{T} = \{(c, d) : c \in \mathcal{C}, d \in \mathcal{D}_c\}$ where $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ is the space of k different types, or categories, of treatment, and $\mathcal{D} = [0, 1]$ is the space of standardized treatment intensity, which in our previous examples, would be denoted by $\mathcal{T}$. The first step of SCIGAN is to train the generator $G : \mathcal{X} \times \mathcal{T} \times \mathcal{Y} \times \epsilon \to \mathcal{Y}^{\mathcal{T}}$ that can impute the missing outcomes. At any given treatment type $C = c$, the output would be an estimated I-DRF for any generator. For any given tuple $T = t = (c, d)$, the output would be the estimated potential outcome of treatment c and a dose level d, which we can use to impute the I-DRF $\widetilde{Y}_c$. Then, we can define the supervised loss for a generator G as:

$$\mathcal{L}_S(G) = \mathbb{E}\left[(Y - G(X, T, Y, \epsilon)(T))^2\right],$$

the expected loss over X, T, Y, and ϵ, of the squared difference between the true outcome and the estimated potential outcome fixing T. The architecture for the generator is similar to that of DRNet. It uses a shared network to learn the representation, then partitions the sample into treatment option blocks, and uses a fully connected network to learn the counterfactuals.

The discriminator $D : \mathcal{X} \times \prod_{c \in \mathcal{C}} (\mathcal{D} \times Y)^{n_c} \to [0, 1]^{\sum n_c}$ takes in features X and the generated potential outcomes at a set of randomly selected n_c dose values for each c. The output is a set of probabilities For each of the k treatment options. It is expressed as $D^{c,j}(X, \widetilde{Y}) = D^c_{\mathcal{C}}(X, \widetilde{Y}) \times D^j(X, \widetilde{Y}_c)$ so that we can define loss for the two components separately. For $D^c_{\mathcal{C}}(X, \widetilde{Y})$, the loss, fixing the generator, is a loss *across all treatment options.*

$$\mathcal{L}_{\mathcal{C}}(D_{\mathcal{C}}; G) = -\mathbb{E}\left[\sum_{c \in \mathcal{C}} \mathbb{1}_{\{C=c\}} log D^c_{\mathcal{C}}(X, \widetilde{Y}) + \mathbb{1}_{\{C \neq c\}} log(1 - D^c_{\mathcal{C}}(X, \widetilde{Y}))\right]$$

For $D^j(X, \widetilde{Y}_c)$, its corresponding loss, fixing the generator, is a loss *at one treatment option*

across all dose values (discrete).

$$\mathcal{L}_d(D_c; G) = -\mathbb{E}\left[\mathbb{1}_{\{C_f=c\}}\Big\{\sum_{j=1}^{n_c}\mathbb{1}_{\{D_f=D_j^c\}} log D_c^j(X, \widetilde{Y}_c) + \mathbb{1}_{\{D_f\neq D_j^c\}} log(1 - D_d^c(X, \widetilde{Y}_c))\Big\}\right]$$

As we can see, the structure of the two losses are similar, with the second loss $\mathcal{L}_d(D_c; G)$ having an indicator outside the summation, only evaluating the discriminator if the factual treatment is the one specified. Since the domain of the treatment discriminator $D_{\mathcal{C}}$ is a set, its architecture uses a permutation invariant layer of the form $f_{inva}(u) = \sigma(\mathbb{1}_b \mathbb{1}_m^\intercal(\phi(U_1), \ldots, \phi(U_m)))$ where u is the input of dimension m, $\mathbb{1}_b$ is a vector of 1s of dimension b. (b depends on the dimension of output from $\phi(\cdot)$) Finally, we can define the optimization problem as

$$\begin{aligned} G^* &= \arg\min_G \mathcal{L}(D^*; G) + \lambda \mathcal{L}_S(G) & D^* &= {D^*}_{\mathcal{C}}^c \times {D^*}_c^j, \quad \text{where} \\ {D^*}_{\mathcal{C}} &= \arg\min_{D_{\mathcal{C}}} \mathcal{L}_{\mathcal{C}}(D_{\mathcal{C}}; G^*) & {D^*}_c &= \arg\min_{D_c} \mathcal{L}_c(D_c; G^*), \forall c \in \mathcal{C}. \end{aligned} \tag{15}$$

The optimal generator minimizes the loss from competing with the best discriminator and a loss coming from model fit. The optimal discriminators minimize their losses (already taken the negative sign in its definition) from the competition with the best generator. SCIGAN is one of the few algorithms that can estimate the I-DRF. SCIGAN also shows promising performance under a discrete action setting when compared with Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE) [134], a GAN algorithm designed for discrete actions, when the number of treatment options exceeds around 7. Nonetheless, the major drawback is that SCIGAN relies on at least a few thousand training samples [15]. Although it avoids the discontinuity between treatment intensity blocks by randomly selecting dose values (points) at each iteration of training, with its high computational cost, this algorithm could be tough to implement in practice. This issue is common for most types of Neural Network-based methods.

5.2 Meta-ML Algorithms

Besides modifying one type of Machine Learning (ML) algorithm, as we have seen in estimating the ATE using combined models (Section 4.3), we can use any ML to estimate

nuisance parameters and post-process the results to obtain an unbiased estimator for CATE $\Delta_\mu(X_i)$. The estimating procedure, therefore, for all types of learners, is the same, differing mainly in how they use the fitted propensity model $\widehat{e} := \widehat{\mathbb{E}}[A|X]$, the outcome model $\widehat{\mu}_1 := \widehat{\mathbb{E}}[Y|X, A = 1]$ and $\widehat{\mu}_0 := \widehat{\mathbb{E}}[Y|X, A = 0]$, and $\widehat{m} := \widehat{\mathbb{E}}[Y|X]$ to devise a meta-structrue with the nuisance parameters above as building blocks. All methods in this section are under a binary action setting, with CATE defined as the causal risk difference for a given individual $\Delta_\mu(X_i) := \mathbb{E}[Y(1) - Y(0)|X_i] = \mu(X_i, 1) - \mu(X_i, 0)$.

S-learner[51] and **T-learner** [49], the most intuitive two among this category, are closely related to outcome modeling presented in Section 4.2. S stands for single, which, as its name suggests, fits a single model to $\mu(X, a) := \mathbb{E}[Y|A = a, X = X]$. We can then obtain the causal risk difference as $\widehat{\mu}(X_i, 1) - \widehat{\mu}(X_i, 0)$ by fixing the indicator A. Since there is no special role given in A, it is possible for some base algorithms, such as Random Forest, to ignore A completely when splitting the leaves, especially when X is of high dimension and many of them relate strongly to the outcome. The estimated CATE in risk difference would be zero for these base algorithms in such cases. S-learner performs well when the true CATE is zero in many places across X, but it is biased towards 0, and inconsistent [2]. T stands for two [49], and is also referred to as the basic [77], plug-in [69], or naive [89] estimator. Instead of using A as one of the covariates in the model, we separate our data into two groups by A and then, within each group, fit a model $\mu_a(X)$. For any individual X_i, we can estimate $\widehat{\Delta}_\mu(X_i) = \widehat{\mu}(X_i, 1) - \widehat{\mu}(X_i, 0)$. Since the outcome surfaces are fit separately, it outperforms S-learner when the shape of the potential outcomes for the two treatment groups differ. Besides, simulation studies have shown that T-learner mitigates the issue of high bias and inconsistency often observed for the S-learner [63, 72]. Though on the opposite side, if the patterns in the two treatment groups are similar, it would be harder for the T-learner to capture the same pattern shared by the treated and untreated groups due to the split data [69].

X-learner modifies the T-learner and is especially helpful when we have an unbalanced sample, that is, when the number of treated or untreated is significantly larger [30] than

[30]Simulations show that the X-learner is recommended, even over the DR-learner, whenever the proportion of the treated or controls, or the other way around, is around 15% or less. When it is around 25%, if the sample size is a few hundred, X-learner is still favorable, but if the sample size is a few thousand, DR-learner would better capture the complexity of treatment effect [93].

that of the other group [75]. The first step of the X-learner is the same as that of the T-learner. However, after estimating $\widehat{\mu}_1(X_i)$ and $\widehat{\mu}_0(X_i)$, we use each unit in an X-like shape. Specifically, for the i^{th} unit, we define $\widetilde{\Delta}_{\mu,1}(X_i) := Y_i - \widehat{\mu}_0(X_i)$ if $A_i = 1$, and $\widetilde{\Delta}_{\mu,0}(X_i) := \widehat{\mu}_1(X_i) - Y_i$, if $A_i = 0$. Next, we regress $\widehat{\mu}_1(X_i)$ on the treated with $A = 1$ to obtain $\widehat{\Delta}_{\mu,1}(X) = \widehat{\mathbb{E}}[\widetilde{\Delta}_{\mu,0}(X)|X]$; regress $\widehat{\mu}_0(X_i)$ on the control with $A = 0$ to obtain $\widehat{\Delta}_{\mu,0}(X) = \widehat{\mathbb{E}}[\widetilde{\Delta}_{\mu,0}(X)|X]$. The final CATE as a risk difference can be computed as a weighted average: $\widehat{\Delta}_{\mu}(X) = \{1 - \widehat{e}(X)\}\widehat{\Delta}_{\mu,1}(X) + \widehat{e}(X)\widehat{\Delta}_{\mu,0}(X)$. To provide some intuition of how it solves the issue of an unbalanced sample, suppose there are little data in the treated group with $A = 1$, then $\mu_{\mu,1}(X)$ would fit poorly, and $\widetilde{\Delta}_{\mu,0}(X)$ would be off. However, since there is little data for $A = 1$, $e(X) = Pr(A = 1|X)$ would be small, and the estimated risk difference $\widetilde{\Delta}_{\mu,0}(X)$ using the poorly fitted $\mu_1(X)$ exerts little influence on the final estimate. Denote our observed sample as $O = (X, A, Y)$. Another variant related to outcome modeling, which the authors Curth et al. called the **RA-learner** (RA stands for Regression Adjusted) [28], imputes the causal risk difference for each individual as

$$\widetilde{\Delta}_{\widehat{\mu}}(O) = \Big\{Y - \widehat{\mu}_0(X)\Big\}A + \Big\{\widehat{\mu}_1(X) - Y\Big\}(1 - A).$$

As Cruth et al. [28] point out, the X- and RA-learner are two variants of the same principle. They differ in that RA-learner imputes the adjusted risk difference for each individual, while the X-learner does so separately for the two groups and combines them at the end using the estimated Propensity Scores as weights.

Besides utilizing outcome modeling, we can also impute the causal risk difference using inverse propensity weighting (IPW) as

$$\widetilde{\Delta}_{\widehat{e}}(O) = \Big\{\frac{A}{\widehat{e}(X)}\Big\}Y - \Big\{\frac{1 - A}{1 - \widehat{e}(X)}\Big\}Y.$$

If the potential outcomes are imputed using the Doubly Robust approach presented in Section 4.3, we will be using the **DR-learner** [69]. The author proposed splitting the sample into three independent samples of size n, denoted (S_1, S_2, S_3). Nuisance parameters $\widehat{e}$ and $\widehat{\mu} = (\widehat{\mu}_0, \widehat{\mu}_1)$ are constructed using two of the three independent samples, for example, S_1 and S_2 respectively without loss of generality. Then the imputed causal risk difference,

an unbiased estimator for the CATE, can be calculated using the standard DR estimator as

$$\widetilde{\Delta}_{\widehat{\mu},\widehat{e}}(O) = \Big\{ \frac{A}{\widehat{e}(X)} \, (Y - \widehat{\mu}_1(X)) + \widehat{\mu}_1(X) \Big\} - \Big\{ \frac{1-A}{1-\widehat{e}(X)} \, (Y_i - \widehat{\mu}_0(X)) + \widehat{\mu}_0(X) \Big\}.$$

Finally, regress $\widetilde{\Delta}_{\widehat{\mu},\widehat{e}}(O)$ on S_3, the one not used in the estimation of nuisance parameters, to obtain $\widehat{\Delta}_\tau(X) = \widehat{\mathbb{E}}[\widetilde{\Delta}_{\widehat{\mu},\widehat{e}}(O)|X = X]$. To perform cross-fitting, we would repeat the above procedure using S_1 or S_2 as the test sample instead, and average the three estimates to arrive at a final estimate. Sample-splitting and cross-fitting are especially encouraged for the DR-learner, as there are theoretical arguments [69] and also simulation results [93] indicating their ability to increase convergence rate when a large sample is available. A problem shared by both the IPW- and the DR-learner, though, is that extreme Propensity Scores can drive up the variance of estimated CATE due to the inverse weighting component. (Recall the argument presented in Section 4.1.1 on the important considerations of using the estimated PS as weights.)

Last but not least, there is the **R-learner**, where R stands for regret or residual. It is closely related to the Double/Debiased Machine Learning (DML) presented in Section 4.3. The first step, similar to the other learners, is to fit the nuisance components $\widehat{e}(X) := \mathbb{E}[A|X]$ and $\widehat{m}(X) := \mathbb{E}[Y|X]$ using any ML algorithms. Then we can construct the loss objective as

$$\widehat{\mathcal{L}}_n(\Delta_{\widehat{e},\widehat{m}}(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} \Big[\{Y_i - \widehat{m}^{(-s(i))}(X_i)\} - \{(A_i - \widehat{e}^{(-s(i))}(X_i)) \Delta_{\widehat{e},\widehat{m}}(X_i)\} \Big]^2 + \Lambda_n(\Delta_{\widehat{e},\widehat{m}}(\cdot)),$$

and take $\widehat{\Delta}_{\widehat{e},\widehat{m}}(\cdot) = \arg\min_{\Delta_{\widehat{e},\widehat{m}}(\cdot)} \widehat{\mathcal{L}}(\Delta_{\widehat{e},\widehat{m}}(\cdot))$. The first part, if one recalls from DML, is the empirical loss from residual-on-residual regression of $Y^* = Y_i - \widehat{\mathbb{E}}[Y|X]$ on $A^* = A_i - \widehat{\mathbb{E}}[A|X]$, and $\Lambda(\Delta_{\widehat{e},\widehat{m}}(\cdot))$ is a regularizer on the complexity of $\Delta_{\widehat{e},\widehat{m}}(\cdot)$. It also uses sample-splitting and cross-fitting, but in a different way than that of the DR-learner. $s(i)$ is a mapping from the indices of n units to the indices of the k evenly sized data splits. Superscript $-s(i)$ denotes that we train $\widehat{e}(\cdot)$ and $\widehat{m}(\cdot)$ in the sample without the data split in which unit X_i belongs, and evaluate the trained model at X_i. Compared with the DR-learner, it is more stable to extreme PSs in small samples, but it converges at a slower rate as the number of observation increases [93].

Both R- and DR-learners used sample-splitting and cross-fitting with out-of-bag prediction,

and although not explicitly stated for some other learners, we can apply these techniques to any meta-algorithms. In Section 4.3, we have seen the use of cross-fitting in helping to control the empirical process term in Average Treatment Effect estimation using semi-parametric methods. Specifically, cross-fitting relaxes the Donsker condition, a restriction on the complexity of a function class, thus allowing us to use more flexible algorithms such as ML algorithms to estimate the nuisance parameters. In this section, sample splitting and cross-fitting are brought up again, together with the use of ML algorithms in the estimation of nuisance parameters. One of the main reasons is that cross-fitting helps reduce the overfitting bias—bias that arises by fitting an over-complicated function that fits well with the current sample but fails to generalize. Cross-fitting is a general term, and different meta-learners have different ways of doing the splits and crosses. To provide more explicit guidance on when to use sample-splitting and cross-validation, Okasa [93] evaluates the performance of different meta-learners under large and finite sample settings with and without double cross-fitting or double sample-splitting defined in [87]. The result suggests that we choose the type of meta-learners based on the amount of data and our belief in the underlying data structure. For any meta-learner we choose, we should fit the models using complete data with out-of-bag prediction [31] when only a small sample is available. When a relatively large sample is available, we should use double cross-fitting to decrease overfitting bias and avoid inefficient use of available data.

6 Variable Selection and Diagnostics

We have discussed methods for GPS estimation in Section 4.1.2 and methods of using the estimated GPS to estimate the ADRF in Section 4.1.1. We have implicitly assumed a correct set of variables in our treatment model. Nonetheless, we should point out that variable selection is essential in that it affects bias and variance of the estimated causal effect [17]. Despite the increased popularity of PS-based models, Ali et al. found that in the medical literature between December 2011 and May 2012, only 34.4% explicitly reported variable selection procedures for the PS model [4].

[31] Out-of-bag prediction means that the observation used for making the prediction is not within the sample used for estimating the parameters of the prediction model.

Note that under a continuous action setting, the principle of variable selection for the treatment model is the same as that under a binary action setting. In summary, it is recommended to include all variables related to the outcome, regardless of whether they are related to the treatment. While including those that also relate to the treatment (commonly known as confounders in epidemiology) can adjust for confounding bias, those that do not relate to treatment can decrease variance without inducing bias [17]. Even if the correlations between some covariates and the outcome are not significant, including such variables in the treatment model can adjust for chance bias which is non-systematic bias due to sampling[32]. In contrast, we should not include variables related only to the treatment, for example, the instrumental variables [86], despite their ability to increase predictive power. Doing so brings noise to the estimated Propensity Score and therefore boosts the variance of the estimated treatment effect that depends on it [10, 17]. See Figure 6 for an illustration.

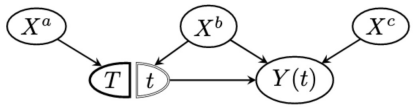

Figure 6: Variable categories. It is suggested to include X^b and X^c but not X^a

As we are interested in a *causal* estimand, covariate balance, as opposed to goodness-of-fit, should be a measure to evaluate the treatment model. The traditional approach is to calculate for each of the covariates the absolute standardized mean difference (ASMD) [8] and the ratios of the variances between different treatment groups (for example, treated vs. untreated with binary action) as suggested by Imai et al. [57]. Denote as S_g the standard deviation and p_g the proportion estimate for treatment group g, the expressions of ASMD for continuous and binary covariates are

$$ASMD_{cont.} := |\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(S_1^2 + S_2^2)/2}}| \qquad ASMD_{bi.} := |\frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{(\widehat{p}_1(1-\widehat{p}_1) + (\widehat{p}_2(1-\widehat{p}_2))/2}}|.$$

Variables with an ASMD less than 10% and a variance ratio close to 1 are considered well-balanced [24]. We can then take the average of the ASMD to arrive at a global measure of imbalance. In the continuous setting, a solution is to discretize $\mathcal{T}$ into bins to apply the

[32]Brookhart et al. [17] indicate that this is one of the reasons an estimated PS is preferred over a known true PS.

above standard methods, which is commonly known as blocking [9]. This approach applies to any causal treatment effect estimation method. However, arbitrariness of bin size would be an inevitable problem. As an alternative, Zhu et al. proposed the average absolute correlation coefficient (AACC) [140]. Building on the Balancing Property (1) of the GPS, we can calculate the AACC by: (1) Sample with replacement n copies of our original sample inverse-weighted by the Generalized Propensity Score. Calculate the correlation between X_j and T for each covariate. (2) Repeat step 1 for k times and calculate an average of the k values, denoted as $\bar{d}_j$. (3) Perform a Fisher transformation $z_j = (1/2)ln\{(1+\bar{d}_j)/(1-\bar{d}_j)\}$ and average over all covariates to acquire a global measure.

Using only covariate balance for treatment model evaluation could be troublesome. Instrumental variables (IV) are related to treatment [5, 6]. Therefore, if we only focus on maximizing covariate balance, we should include all IVs into the treatment model, yet as mentioned above, doing so inflates variance for the final estimate [10, 17]. Suppose we seek to both maximize covariate balance and efficiency of the estimator. We can use an adaptive Lasso, an extension of the traditional Lasso [121], to select variables in the treatment model based on their correlations with the treatment. Although this category of methods, originated by Shortreed and Ertefai [115], has not been extended to the continuous treatment setting, we nevertheless present it in hopes of providing insights for future extension. Assume a logit model for the treatment model $A|X$ with parameter α_{PS}, that is, $e(X;\alpha_{PS}) = exp\{\alpha_{PS}^\intercal X\}/(1+exp\{\alpha_{PS}^\intercal X\})$. Denote the parameters for X and A in the outcome model $Y|A,X$ as α_{Out} and η, that is, $Y = \alpha_{Out}^\intercal X + \eta^\intercal A + \epsilon$. The Outcome-Adaptive Lasso estimator (OAL) for α_{PS} can be defined as

$$\widehat{\alpha}_{PS} = \underset{\alpha_{PS}}{\arg\min} \Big\{ \underbrace{\sum_{i=1}^{n} -a_i(\alpha_{PS}^\intercal X_i) + log(1+e^{\alpha_{PS}^\intercal X_i})}_{\text{loss from the PS model}} + \underbrace{\lambda \sum_{j=1}^{d} \frac{|\alpha_{PS,j}|}{|\widehat{\alpha}_{Out,j}|^\gamma}}_{\text{weighted penalty}} \Big\}, \tag{16}$$

where $\gamma > 1$ and $(\widehat{\alpha}_{Out}, \widehat{\eta}) = \arg\min_{\alpha_{Out},\eta} l_n(\alpha_{Out}, \eta; Y, X, A)$, $l_n(\cdot)$ the negative log likelihood function of size n. Using α_{Out} as weights, the algorithm can put more penalty on variables having less contribution to the outcome model. λ, the tuning parameter, can be determined by minimizing the weighted absolute mean difference with weights $\widehat{\alpha}_{Out}$ or by Generalized Cross Validation, as suggested by Tibshirani [121]. Borrowing from the idea

of Doubly Robustness, Ertefai et al. proposed a variable selection procedure with penalty considering the association of X_j and Y and of X_j and A simultaneously. For simplicity, denote the loss for the PS plus the loss for the outcome model as $M(\cdot)$, then the estimator can be expressed as [36]

$$\widehat{\alpha}_{DR} = \underset{\alpha_{DR}}{\arg\min} \Big\{ M_{DR}(\alpha_{DR}) + \lambda \sum_{j=1}^{d} \frac{|\alpha_{DR,j}|}{\widehat{\alpha}^2_{Out,j}(1 + |\widehat{\alpha}_{PS,j}|)^2} \Big\}. \tag{17}$$

Fixing either one of $\widehat{\alpha}_{Out}$ or $\widehat{\alpha}_{PS}$, estiamted respectively using ordinary least squares or MLE, the penalties will be inversely related to the other. After the selection procedure using adaptive Lasso, we can then use the selected variables (those $\alpha_{DR,j} \neq 0$) in a treatment model and apply any PS-based estimation method. With a similar idea of using a simultaneous penalty, Koch et al. [73] proposed the outcome adaptive group Lasso with which parameters for the same variable in the treatment and outcome models are grouped as $\alpha_G = (\alpha_{PS}^\intercal, \alpha_{Out}^\intercal)^\intercal$ to be penalized together. Unlike the method by Ertefaie et al. [36] where variable selection and treatment effect estimation are separate steps and OAL (Equation 17) is used for variable selection only, Koch et al. uses OAL for both variable selection and estimation. Firstly, we use OAL to estimate the grouped α_G by

$$\widehat{\alpha}_G = (\alpha_{Out}, \alpha_{PS}) = \underset{\alpha_G}{\arg\min} \Big\{ M_G(\alpha_G) + \lambda \sum_{j=1}^{d} \frac{\sqrt{2}}{|\nu_j|} \sqrt{\alpha^2_{Out,j} + \alpha^2_{PS,j}} \Big\}, \tag{18}$$

where ν_j is the estimated coefficient of covariate j in the *full* outcome model. Similar to the OAL in Equation 16, with the inverse of the estimated coefficient for the outcome model in the weights, the algorithm puts heavy penalty on variables unrelated to the outcome; but different from Equation 16, both the treatment and the outcome model for the estimated α_G contribute to the loss. Secondly, with $\widehat{\alpha}_G = (\widehat{\alpha}_{Out}, \widehat{\alpha}_{PS})$, we can then use $\widehat{\mu}(X, A; \alpha_{Out})$ and $\widehat{e}(X; \widehat{\alpha}_{PS})$ in a DR estimator for the average causal treatment effect.

Despite the ability to differentiate IVs with outcome-related variables, these adaptive methods assume that the true outcome model is within a class of linear functions. In hopes of providing a more generic, model-free metric applicable for any treatment effect estimation method, Tang et al. proposed the Causal Ball Screening (CBS) procedure based on Ball Co-

variance [95, 96], a non-parametric measure of dependence in Banach spaces[33]. For Banach spaces $(\mathcal{X}, \rho)$ and $(\mathcal{Y}, \xi)$, denote the Borel probability measure on the joint space $\mathcal{X} \times \mathcal{Y}$ as θ and the Borel probability measure for $\mathcal{X}$ and $\mathcal{Y}$ as μ and ν, respectively. We can then define the ball covariance as

$$BCov^2(X,Y) := \int \Big[\theta(\bar{B}(x_1,x_2) \times \bar{B}(y_1,y_2)) - \mu(\bar{B}(x_1,x_2))\nu(\bar{B}(x_1,x_2))\Big]^2 \\ \theta(dx_1,dy_1)\theta(dx_2,dy_2),$$

where $\bar{B}_\rho(x_1,x_2)$ is a closed ball in space $(\mathcal{X},\rho)$ with center x_1 and radius $\rho(x_1,x_2)$ and $\bar{B}_\xi(y_1,y_2)$ a closed ball in space $(\mathcal{Y},\xi)$ with center y_1 and radius $\xi(y_1,y_2)$. With the Ball Covariance, we can then define the conditional ball covariance as

$$BCov^2(X,Y|A) := eBCov^2(X,Y|A=1) + (1-e)BCov^2(X,Y|A=0).$$

Thus, for n i.i.d. copies of (X,Y,A), $X \in \mathbb{R}^d$, we can empirically estimate the Propensity Score e, calculate $BCov_n^2(X_j,Y|A)$ for $j = 1,\ldots,d$ and select variables with the largest q of them. For instance, if we want to use CBS in conjunction with the DR estimator, we can then use the selected q covariates in a traditional Lasso to estimate the parameter for the outcome model and an adaptive Lasso to obtain that for the PS model. Adopting the estimated parameters in μ and e in a DR estimator will then give us the desired ATE. Unlike the outcome-adaptive methods, the variable selection step using ball covariance is independent of the specification of an outcome model, thus requiring no assumptions for smoothness or linearity. In addition, unlike the distance correlation measure [118], ball covariance does not need finite moments and can work with ultra-high dimensional settings with as many as millions of covariates [119].

7 Discussion

Causal Inference with non-finite or continuous action is a new setting under the increasingly popular Neyman-Rubin potential outcome framework. Non-finite action brings forth addi-

[33] A complete normed space $(\mathcal{X}, \|\cdot\|)$ with X a vector space over a scalar space like $\mathbb{R}$ and $\|\cdot\|$, a mapping $\mathcal{X} :\to \mathbb{R}$ with properties of a norm which can be thought of as a form of distance.

tional challenges in causal effect estimation. If we use the GPS-based methods to estimate the ATE or CATE, we would need to estimate not only one but a *distribution* of the GPSs. Furthermore, our estimand would no longer be pathwise differentiable for semi-parametric estimation methods. Additionally, there is currently no outcome-adaptive approach for variable selection for the GPS model. On the one hand, recent methods seek to attenuate the effect of model misspecification by adopting more flexible tools, for instance, semi-or non-parametric methods such as Kernels and the efficient influence functions, and Machine Learning algorithms such as Lasso, Forests, and Neural Networks. On the other hand, we want to control the complexity of our model to avoid overfitting, reduce the variance of the estimate to the best possible, and achieve better finite-sample performance. This demand motivates using sample-splitting to create independent samples, cross-fitting to recover data efficiency, and boosting and stacking to bound the estimation error better. In addition, we want to decrease computational costs and improve the convergence rate to the best extent.

Here we provide a list of observations from this review: (1) 'Robustness' is a commonly mentioned concept for many recently proposed methods. For example, robustness to model specification, robustness in the convergence rate requirement, and robustness to extreme values of the estimated Generalized Propensity Scores (GPS). (2) There are deep connections between orthogonality, robustness, and estimator efficiency. (3) Recent methods turn GPS estimation into an optimization problem that directly finds a balancing weight to avoid inverse-weighting by an extreme GPS. (4) There is an increase in using 'difference' rather than the original value. For example, the difference between the original and an estimated value (residual), the difference between an estimated value and an optimal benchmark (regret), or the difference of estimates between groups across groups. Using differences centers variables. The difference also connects to loss. The loss perspective provides us with ways to avoid outliers (by having bounded loss), solve the non-differentiability of the original function (if the designed valid loss can be differentiable), and investigate the uniqueness of the solution (whether the loss objective is strongly convex or not). (5) The functional assumption is strong. 'Strong' not only that it is hard to specify the correct form, but also that if we know the underlying data structure, choosing the correct model or estimation method that best takes advantage of this information can often significantly reduce computational cost and variance of the estimate. More flexible methods are not always better if considering other factors besides unbiasedness. The above idea also applies to the different

ATE estimation methods falling under the category of reweighing. (6) One can argue that inverse-weighting is the most flexible and could theoretically achieve the best balance. On the other side of the same coin, one can also argue that matching and stratification are more robust to GPS modeling and extreme estimated GPS. (7) Methods have started to see different stages in estimation as a whole, using a composite loss that incorporates model fit and one or more components designed for optimizing the desirableness of the final outcome or designed for variable selection. (Covariate balancing as a constraint for Propensity Score estimation and Outcome-Adaptive Lasso for Propensity Score variable selection are some such examples) (8) Statistical methods and Machine Learning (ML) complement each other in different aspects. ML can handle complex data structures well but lacks interpretability and awareness for adjusting confounding and over-fitting biases. Statistical methods help target ML tools towards a causal estimand and achieve desirable properties by modifying their components or constructing a meta-structure on top of them.

Methods for ATE estimation relate closely to that for CATE. Each category of ATE estimation (reweighting observed, imputing unobserved, and combining models) can find its counterpart in Meta-ML algorithms (IPW-, S-, T-, X-, RA-, DR-, and R-learner) for the estimation of CATE. From one perspective, CATE falls under imputing the unobserved category, since being able to estimate CATE implies that we can estimate the unobserved potential outcomes. From another perspective, to have an unbiased estimator for CATE, we need methods and perspectives from all three categories of ATE estimation. CATE can help identify subgroups having the most potent response to treatment and treatment intensities most helpful for a given set of covariates. Therefore, the CATE naturally leads to targeted policy optimization, 'targeted' because the policy or treatment is optimized based on a given set of 'individual' covariates X. Formally, within biostatistics, this falls under the umbrella of Precision Medicine, with the problem being: the estimation of the optimal Dynamic Treatment Regime (DTR). With binary action, finding the optimal DTR reduces to comparing for each individual the value of two potential outcomes— the outcome if treated, denoted $Y(1)$, and that if not, denoted $Y(0)$. If we know the difference, we can reduce optimal DTR estimation to a classification problem, with the label being the sign of $Y(1) - Y(0)$. Methods such as Support Vector Machines naturally apply. More generally, in discrete action spaces, this is a weighted multi-class classification problem, where weights are for confounding adjustment and the classes are different treatment options [33]. Similar to

ATE and CATE estimation, there is much less existing work under the continuous action setting. A substantial issue is that if directly adopting the outcome-weighting techniques since the density at any $T = t$ is zero, the non-smooth indicator function $\mathbb{1}(T_i = t)$ would pose a challenge for optimization algorithms. One way to resolve the issue is to use a continuous surrogate loss, as mentioned by Zhao et al. [138]. With a similar idea but a different choice for the surrogate, Chen et al. use a bounded loss that is the difference between two convex functions [21]. Another solution is to use kernels. For example, the tree-based approach proposed by Laber and Zhao [76] and the outcome weighted-based approach incorporating IPW and the DR estimator by Kallus and Zhou [65]. The third line of methods abandons the outcome-weighted learning framework. Demirer et al. [30] propose a semi-parametric, regret-minimization-based method that achieves Double Robustness. Zhou et al. propose a non-parametric dimension reduction framework to estimate the optimal treatment rule to circumvent the curse of dimensionality [139]. Zenati et al. propose a joint kernel embedding through counterfactual risk minimization [135]. The estimation of CATE with continuous action is still to be developed, and so is policy optimization with continuous action.

Although one of our objectives is to provide a grand structure for ATE and CATE estimation, we point out that due to a primary focus on non-finite treatment, there is important work under the binary treatment setting we do not cover. These include Bayesian approaches such as Bayesian Additive Regression Trees (BART) [46], hierarchical Bayesian Bootstrap for CATE [92], Bayesian Inference of CATE using Multi-task Gaussian Processes [3], other Bayesian non-parametric methods [70, 91], and so on (See Lechner and Antonelli [77] for a review). Moreover, there is Propensity Score estimation using the Gaussian process [127], latent variable modeling [81], CATE estimation based on deep representation learning [133], causal effect estimation from networked observational data [44], and many more extensions that push Causal Inference to a less restrictive, more real-world like setting.

Expansion in this field brings forth more open questions. For example, based on the topic of this review: (1) How to avoid ad hoc adjustment for extreme values of estimated GPS or address limited overlap in the distribution of covariates without changing our target population? (2) Is there a way to decide when to use Double/Debiased Machine Learning and when to use Targeted Learning? How would they compare under different data structures with a finite sample? When might be useful to consider higher-order orthogonality? Are

there other approaches to correct the plug-in bias and resolve the issue of not being pathwise differentiable? How might we go beyond semi-parametric estimation? (3) How can we optimally choose among a pool of methods, from start to end, by better understanding how the different choices made along the way affect the final estimate? (5) If we apply the more developed ATE estimation methods to estimate the potential outcomes of finding the optimal Dynamic Treatment Regime, what benefits can it bring? If we combine VCNet with tree-based algorithms with a loss to measure the difference between two continuous functions, can we extend VCNet to estimate CATE? Are there other efficient ways to model a continuous function besides using the B-spline? (6) Is there a way to connect Causal Discovery to Causal Inference? How might we incorporate results from causal representation learning into Causal Inference? We hope this review can briefly introduce Causal Inference, using the continuous action setting as a window. We additionally hope this review can inspire readers to further explore future directions, go beyond correlation, and push the boundary of statistical learning using real-world data.

8 Tables for Notation, Symbols, and Abbreviations

8.1 Table of Notation and Symbols

Notation	Definition/Space	In Full Words	Abbreviation
Objectives			
$\tau(t)$	$\mathbb{E}[Y(t)]$	Average Treatment Effect	ATE
Δ_τ	$\mathbb{E}[Y(1) - Y(0)]$ $= \tau(1) - \tau(0)$	Average Causal Risk Difference (Binary)	-
$\tau(t, X)$	$\mathbb{E}[Y(t)\|X]$	Conditional Average Treatment Effect	CATE
$\Delta_\mu(X)$	$\mathbb{E}[Y(1) - Y(0)\|X]$ $= \mu(1, X) - \mu(0, X)$	Conditional Causal Risk Difference (Binary) (With Cuasal Consistency)	CATE (Binary)
Nuisance Models			
$m(X)$	$\mathbb{E}[Y\|X]$	-	-
$e(X)$	$\mathbb{E}[A\|X]$	Treatment Model/Propensity Score (Binary)	PS
$r(X, T)$	$\mathbb{E}[T\|X]$	Treatment Model/Generalized Propensity Score	GPS
$\mu(X, T)$	$\mathbb{E}[Y\|X, T]$	Outcome Model	-
Special Letter			
$Y(t)$	$\mathcal{Y}$	Potential Outcomes	-
Y	$\mathcal{Y}$	Observed Outcomes	-
A	$\mathcal{A} = \{0, 1\}$	Treatment (Binary)	-
T	$\mathcal{T} = [t_{min}, t_{max}] \in \mathbb{R}$	Treatment (Continuous)	-
X	$\mathcal{X}$	Covariates Sufficient for Confounding Adjustment	-
Z	$\mathcal{Z}$	Instrumental Variables	IVs
U	$\mathcal{U}$	Unobserved Variables	-
O	(X, T, Y)	Observed Data	-
Semi-parametric Section			
Ψ	$\Psi(P)$	Estimand of Interest	-
ϕ	$\mathcal{H}$, Hilbert Space	Influence Functions	-
P_θ	$\mathcal{M} = \{P_\theta : \theta \in \Theta\}$	Distribution Parametrized by θ in a Model	-
P_ϵ	$\mathcal{M} = \{P_\epsilon : \epsilon \in \mathbb{R}\}$	Distribution in a Parametric Submodel	-

X^* = A transformed version of X, usually the standardized version of X, or the optimal X.

$\widetilde{X}$ = An imputed version, or a simple transformation of X such as the residual of X by subtracting from it an estimate obtained from a regression model.

$\widehat{X}$ = An estimate of X.

8.2 Table of Abbreviations

Abbreviation	Meaning
AACC	Average Absolute Correlation Association
ADRF	Average Dose Response Function
AIPW	Augmented Inverse Propensity Weighing
ASMD	Absolute Standardized Mean Difference
ATE	Average Treatment Effect
BART	Baysian Additive Regression Tree
CATE	Conditional Average Treatment Effect
CBPS	Covariate Balancing Propensity Score
CBS	Causal Ball Screening
CF	Causal Forest
DAG	Directed Acylic Graph
DML	Double/Debiased Machine Learning
DR	Doubly Robust
DRNet	Dose Response Network
EB	Entropy Balancing
EBCT	Entropy Balancing for Continuous Treatment
eCDF	Empirical Culmulative Distribution
GAN	Generative Adversarial Network
GANITE	Generative Adversarial Nets for inference of Individualized Treatment Effects
GBM	Generalized Boosted Model
GPS	Generalized Propensity Score
IDRF	Individual Dose Response Function
IPCW	Inverse Probability of Censoring Weighing
IPW	Inverse Propensity Weighing
IV	Instrumental Variable
MAR	Missing At Random
ML	Machine Learning
MLE	Maximum Likelihood
MOM	Method of Moments
NN	Neural Network
OAL	Outcome Adaptive Lasso
OW	Overlap Weighing
PS	Propensity Score
RCT	Randomized Control Trial

SCIGAN	eStimating the effects of Continuous Interventions using GAN
SL	Super Learner
SWIG	Single World Intervention Graph
TARNET	Treatment Agnostic Representation Networks
TMLE	Targeted Maximum Likelihood Estiamtion
VCNet	Varying Coefficient Neural Network

9 List of References

1. Abadie, A. & Imbens, G. W. Matching on the estimated propensity score. *Econometrica: journal of the Econometric Society* **84,** 781–807 (Mar. 2016).

2. Alaa, A. & Schaar, M. Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. *Proceedings of the 35th International Conference on Machine Learning* **80,** 129–138 (July 2018).

3. Alaa, A. M. & van der Schaar, M. Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes. *Advances in Neural Information Processing Systems* (Apr. 2017).

4. Ali, M. S. *et al.* Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology* **68,** 112–121 (Feb. 2015).

5. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91,** 444–455 (June 1996).

6. Angrist, J. D. & Krueger, A. B. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* **15,** 69–85. `https://www.aeaweb.org/articles?id=10.1257/jep.15.4.69` (Dec. 2001).

7. Austin, P. C. The performance of different propensity-score methods for estimating relative risks. *Journal of Clinical Epidemiology* **61,** 537–545 (June 2008).

8. Austin, P. C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* **28,** 3083–3107 (Nov. 2009).

9. Austin, P. C. Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Statistical Methods in Medical Research* **28,** 1365–1377 (May 2019).

10. Austin, P. C., Grootendorst, P. & Anderson, G. M. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine* **26,** 734–753 (Feb. 2007).

11. Austin, P. C. The performance of different propensity score methods for estimating marginal odds ratios, Statistics in Medicine 2007;26:3078–3094. *Statistics in Medicine* **27,** 3918–3920 (Aug. 2008).

12. Bang, H. & Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–973 (Dec. 2005).

13. Bell, J. S. & Aspect, A. *Speakable and Unspeakable in Quantum Mechanics: Collected papers on quantum philosophy* ISBN: 9780511815676 (Cambridge University Press, Apr. 2004).

14. Belloni, A., Chernozhukov, V., Fernandez-Val, I. & Hansen, C. Program Evaluation and Causal Inference With High-Dimensional Data. *Econometrica: journal of the Econometric Society* **85,** 233–298 (Jan. 2017).

15. Bica, I., Jordon, J. & van der Schaar, M. Estimating the Effects of Continuous-valued Interventions using Generative Adversarial Networks. *Conference and Workshop on Neural Information Processing Systems* (2020).

16. Bound, J., Jaeger, D. A. & Baker, R. M. Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association* **90,** 443–450 (June 1995).

17. Brookhart, M. A. *et al.* Variable selection for propensity score models. *American Journal of Epidemiology* **163,** 1149–1156 (June 2006).

18. Brown, D. W., Greene, T. J., Swartz, M. D., Wilkinson, A. V. & DeSantis, S. M. Propensity score stratification methods for continuous treatments. *Statistics in Medicine* **40,** 1189–1203 (Feb. 2021).

19. Cao, W., Tsiatis, A. A. & Davidian, M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96,** 723–734 (Sept. 2009).

20. Casella, G. & Berger, R. L. Statistical Inference. *Biometrics* **49,** 320 (Mar. 1993).

21. Chen, G., Zeng, D. & Kosorok, M. R. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association* **111,** 1509–1521 (Jan. 2016).

22. Chernozhukov, V. *et al.* Double/debiased machine learning for treatment and structural parameters. *The econometrics journal* **21,** C1–C68 (Feb. 2018).

23. Chiu, S.-T. Bandwidth selection for kernel density estimation. *The Annals of Statistics* **19,** 1883–1905 (Dec. 1991).

24. Cohen, J. *Statistical power analysis for the behavioral sciences* 2nd. ISBN: 9780203771587 (Lawrence Erlbaum Associates, 1988).

25. Colangelo, K. & Lee, Y.-Y. Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments. *arXiv* (Dec. 2019).

26. Cole, S. R. & Hernán, M. A. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* **168,** 656–664 (Sept. 2008).

27. Crump, R. K., Hotz, V. J., Imbens, G. W. & Mitnik, O. A. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96,** 187–199 (Jan. 2009).

28. Curth, A. & van der Schaar, M. *Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms* in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics* **130** (PMLR, 2021).

29. De Boor, C. *B(asic)-Spline Basics* `https://www.cs.unc.edu/~dm/UNC/COMP258/Papers/bsplbasic.pdf`.

30. Demirer, M., Syrgkanis, V., Lewis, G. & Chernozhukov, V. Semi-Parametric Efficient Policy Learning with Continuous Actions. *arXiv* (May 2019).

31. Deville, J.-C. & Sarndal, C.-E. Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87,** 376 (June 1992).

32. Díaz, I. & van der Laan, M. J. Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference* **1,** 171–192. `https://doi.org/10.1515/jci-2012-0005` (Jan. 2013).

33. Dudík, M., Erhan, D., Langford, J. & Li, L. Doubly robust policy evaluation and optimization. *Statistical Science* **29,** 485–511 (Nov. 2014).

34. *E11 Clinical Investigation of Medicinal Products in the Pediatric Population — FDA* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e11-clinical-investigation-medicinal-products-pediatric-population.

35. *E7 Studies in Support of Special Populations: Geriatrics — FDA* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e7-studies-support-special-populations-geriatrics.

36. Ertefaie, A., Asgharian, M. & Stephens, D. A. Variable Selection in Causal Inference using a Simultaneous Penalization Method. *Journal of Causal Inference* **6,** 20170010. https://doi.org/10.1515/jci-2017-0010 (2018).

37. Fong, C., Hazlett, C. & Imai, K. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The annals of applied statistics* **12,** 156–177 (Mar. 2018).

38. Foster, J. C., Taylor, J. M. G. & Ruberg, S. J. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30,** 2867–2880 (Oct. 2011).

39. Frisch, R. & Waugh, F. V. Partial Time Regressions as Compared with Individual Trends. *Econometrica: journal of the Econometric Society* **1,** 387 (Oct. 1933).

40. Goodfellow, I. *et al. Generative Adversarial Nets* in *Advances in Neural Information Processing Systems* (eds Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K.) **27** (Curran Associates, Inc., 2014). https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

41. Greenland, S., Pearl, J. & Robins, J. M. Causal Diagrams for Epidemiologic Research. *Epidemiology* **10,** 37–48. ISSN: 10443983. http://www.jstor.org/stable/3702180 (Jan. 1999).

42. Gunsilius, F. F. Nontestability of instrument validity under continuous treatments. *Biometrika* **108,** 989–995 (Nov. 2021).

43. Guo, R., Cheng, L., Li, J., Hahn, P. R. & Liu, H. A Survey of Learning Causality with Data. *ACM Computing Surveys* **53,** 1–37. ISSN: 0360-0300. https://dl.acm.org/doi/10.1145/3397269 (Sept. 2020).

44. Guo, R., Li, J. & Liu, H. Learning Individual Causal Effects from Networked Observational Data. *Proceedings of the 13th International Conference on Web Search and Data Mining* (June 2019).

45. Hahn, P. R., Murray, J. S. & Carvalho, C. M. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis* (Jan. 2020).

46. Hahn, P. R., Murray, J. S. & Carvalho, C. M. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis* **15,** 965–2020. https://doi.org/10.1214/19-BA1195 (Sept. 2020).

47. Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* **20,** 25–46 (Jan. 2012).

48. Hansen, B. E. Lecture Notes on Nonparametrics. https://www.ssc.wisc.edu/~bhansen/718/NonParametrics1.pdf (2009).

49. Hansotia, B. & Rukstales, B. Incremental value modeling. *Journal of Interactive Marketing* **16,** 35–46 (Jan. 2002).

50. Hemingway, H. *et al. Using nationwide 'big data' from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the ClinicAl disease research using LInked Bespoke studies and Electronic health Records (CALIBER) programme* (NIHR Journals Library, Jan. 2017).

51. Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20,** 217–240 (Jan. 2011).

52. Hines, O., Dukes, O., Diaz-Ordaz, K. & Vansteelandt, S. Demystifying Statistical Learning Based on Efficient Influence Functions. *The American Statistician,* 1–48 (Jan. 2022).

53. Hirano, K. & Imbens, G. W. in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family* (eds Gelman, A. & Meng, X.-L.) 73–84 (John Wiley and Sons, Ltd, July 2004). ISBN: 9780470090435.

54. Hitchcock, C. & Pearl, J. Causality: models, reasoning and inference. *The Philosophical review* **110,** 639 (Oct. 2001).

55. Holland, P. W. Statistics and causal inference. *Journal of the American Statistical Association* **81,** 945 (Dec. 1986).

56. Hubbard, A. Super Learner. *U.C. Berkeley Division of Biostatistics Working Paper Series* (July 2007).

57. Imai, K., King, G. & Stuart, E. A. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171,** 481–502 (Apr. 2008).

58. Imai, K. & Ratkovic, M. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76,** 243–263 (Jan. 2014).

59. Imai, K. & van Dyk, D. A. Causal inference with general treatment regimes. *Journal of the American Statistical Association* **99,** 854–866 (Sept. 2004).

60. Imbens, G. W. Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics* **20,** 493–506 (Oct. 2002).

61. Imbens, G. W. The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika* **87,** 706–710. ISSN: 00063444. `http://www.jstor.org/stable/2673642` (2000).

62. Imbens, G. W. & Rubin, D. B. Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. *The Annals of Statistics* **25,** 305–327. ISSN: 00905364. `http://www.jstor.org/stable/2242722` (Feb. 1997).

63. Jacob, D. CATE meets ML. *Digital Finance* **3,** 99–148 (June 2021).

64. Jodhka, G. S., Gouda, M. W., Medora, R. S. & Knalil, S. A. Inhibitory effect of dioctyl sodium sulfosuccinate on trypsin activity. *Journal of Pharmaceutical Sciences* **64,** 1858–1862 (Nov. 1975).

65. Kallus, N. & Zhou, A. *Policy Evaluation and Optimization with Continuous Treatments* in *The 25th International Conference on Artificial Intelligence and Statistics* (JMLR W&CP, Lanzarote, Spain, Feb. 2018). ISBN: 1558603859.

66. Kang, J. D. Y. & Schafer, J. L. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science* **22,** 523–539. ISSN: 08834237. http://www.jstor.org/stable/27645858 (2022) (2007).

67. Kennedy, E. H., Ma, Z., McHugh, M. D. & Small, D. S. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **79,** 1229–1245 (Sept. 2017).

68. Kennedy, E. H. in *Statistical Causal Inferences and Their Applications in Public Health Research* (eds He, H., Wu, P. & Chen, D.-G. () 141–167 (Springer International Publishing, Cham, 2016). ISBN: 978-3-319-41259-7. https://doi.org/10.1007/978-3-319-41259-7_8.

69. Kennedy, E. H. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv* (Apr. 2020).

70. Kim, C., Daniels, M. J., Marcus, B. H. & Roy, J. A. A framework for Bayesian nonparametric inference for causal effects of mediation. *Biometrics* **73,** 401–409 (June 2017).

71. Klosin, S. Automatic Double Machine Learning for Continuous Treatment Effects. *arXiv*. https://doi.org/10.48550/arXiv.2104.10334 (Apr. 2021).

72. Knaus, M. C., Lechner, M. & Strittmatter, A. Machine learning estimation of heterogeneous causal effects: empirical monte carlo evidence. *The econometrics journal* (June 2020).

73. Koch, B., Vock, D. M. & Wolfson, J. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics* **74,** 8–17 (June 2017).

74. Kosorok, M. R. *Introduction to empirical processes and semiparametric inference* ISBN: 978-0-387-74977-8 (Springer New York, 2008).

75. Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 4156–4165 (Mar. 2019).

76. Laber, E. B. & Zhao, Y. Q. Tree-based methods for individualized treatment regimes. *Biometrika* **102,** 501–514 (July 2015).

77. Lechner, M. Modified Causal Forests for Estimating Heterogeneous Causal Effects by Michael Lechner:: SSRN. *CEPR* (Jan. 2019).

78. Li, F. & Li, F. Propensity score weighting for causal inference with multiple treatments. *The annals of applied statistics* **13,** 2389–2415 (Dec. 2019).

79. Li, F., Thomas, L. E. & Li, F. Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology* **188,** 250–257 (Jan. 2019).

80. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data* ISBN: 9780471183860 (John Wiley & Sons, Inc., Aug. 2002).

81. Louizos, C. *et al. Causal Effect Inference with Deep Latent-Variable Models* in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., Long Beach, California, USA, 2017), 6449–6459. ISBN: 9781510860964.

82. Lovell, M. C. A simple proof of the FWL theorem. *The Journal of economic education* **39,** 88–91 (Jan. 2008).

83. Lunceford, J. K. & Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23,** 2937–2960 (Oct. 2004).

84. McCaffrey, D. F., Ridgeway, G. & Morral, A. R. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* **9,** 403–425 (Dec. 2004).

85. Miguel A. Hernán, J. M. R. *Causal Inference: What If* `https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/` (Chapman & Hall/CRC, 2020).

86. Myers, J. A. *et al.* Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* **174,** 1213–1222 (Dec. 2011).

87. Newey, W. K. & Robins, J. M. *Cross-fitting and fast remainder rates for semiparametric estimation* https://doi.org/10.48550/arXiv.1801.09138 (Jan. 2018).

88. Nie, L., Ye, M., Liu, Q. & Nicolae, D. VCNet and Functional Targeted Regularization For Learning Causal Effects of Continuous Treatments. *arXiv* (Mar. 2021).

89. Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108,** 299–319. ISSN: 0006-3444. eprint: https://academic.oup.com/biomet/article-pdf/108/2/299/37938939/asaa076.pdf. https://doi.org/10.1093/biomet/asaa076 (Sept. 2020).

90. Of Managerial Economics (Emeritus) Howard Raiffa, F. P. R. P., Pratt, J. W., (mathématicien).), R. O. S., Raiffa, H. & Schlaifer, R. *Introduction to Statistical Decision Theory* illustrated, reprint. ISBN: 9780262161442 (MIT Press, 1995).

91. Oganisian, A., Mitra, N. & Roy, J. Bayesian Nonparametric Cost-Effectiveness Analyses: Causal Estimation and Adaptive Subgroup Discovery. *undefined* (Sept. 2020).

92. Oganisian, A., Mitra, N. & Roy, J. Hierarchical Bayesian Bootstrap for Heterogeneous Treatment Effect Estimation. *arXiv.* https://doi.org/10.48550/arXiv.2009.10839 (2020).

93. Okasa, G. Meta-Learners for Estimation of Causal Effects: Finite Sample Cross-Fit Performance. *arXiv.* https://doi.org/10.48550/arXiv.2201.12692 (Jan. 2022).

94. Owen, A. B. *Empirical Likelihood* ISBN: 1-58488-071-6 (CHAPMAN & HALL/CRC, 2001).

95. Pan, W., Wang, X., Xiao, W. & Zhu, H. A Generic Sure Independence Screening Procedure. *Journal of the American Statistical Association* **114.** PMID: 31692981, 928–937. eprint: https://doi.org/10.1080/01621459.2018.1462709. https://doi.org/10.1080/01621459.2018.1462709 (2019).

96. Pan, W., Wang, X., Zhang, H., Zhu, H. & Zhu, J. Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association,* 1–30 (Jan. 2019).

97. Pearl, J. *On the Testability of Causal Models with Latent and Instrumental Variables* in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers Inc., Montréal, Qué, Canada, 1995), 435–443. ISBN: 1558603859.

98. Peters, J., Janzing, D. & Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms* ISBN: 9780262037310 (The MIT Press, 2017).

99. Powell, J. L. in *Microeconometrics* (eds Durlauf, S. N. & Blume, L. E.) 267–277 (Palgrave Macmillan UK, 2010). ISBN: 978-0-230-23881-7. https://doi.org/10.1057/9780230280816.

100. Powers, S. *et al.* Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine* **37,** 1767–1787 (May 2018).

101. Richardson, T. Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality. *undefined* (Apr. 2013).

102. Robins, J. M. & Finkelstein, D. M. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* **56,** 779–788 (Sept. 2000).

103. Robins, J. M., Hernán, M. A. & Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology* **11,** 550–560 (Sept. 2000).

104. Robinson, P. M. Root-N-Consistent Semiparametric Regression. *Econometrica: journal of the Econometric Society* **56,** 931 (July 1988).

105. Rosenbaum, P. R. Model-Based Direct Adjustment. *Journal of the American Statistical Association* **82,** 387 (June 1987).

106. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55. ISSN: 0006-3444. eprint: https://academic.oup.com/biomet/article-pdf/70/1/41/662954/70-1-41.pdf. https://doi.org/10.1093/biomet/70.1.41 (Apr. 1983).

107. Rosenbaum, P. R. & Rubin, D. B. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79,** 516 (Sept. 1984).

108. Rosenbaum, P. R. & Rubin, D. B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39,** 33 (Feb. 1985).

109. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology* **66,** 688–701. https://doi.org/10.1037/h0037350 (1974).

110. Rubin, D. B. Inference and Missing Data. *Biometrika* **63,** 581–592. ISSN: 00063444. http://www.jstor.org/stable/2335739 (2022) (1976).

111. Rubin, D. B. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics* **6,** 34–58. ISSN: 00905364. http://www.jstor.org/stable/2958688 (1978).

112. Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M. & Karlen, W. Learning Counterfactual Representations for Estimating Individual Dose-Response Curves. *Proceedings of the AAAI Conference on Artificial Intelligence* **34,** 5612–5619 (Apr. 2020).

113. Shalit, U., Johansson, F. D. & Sontag, D. *Estimating Individual Treatment Effect: Generalization Bounds and Algorithms* in *Proceedings of the 34th International Conference on Machine Learning* **70** (JMLR.org, Sydney, NSW, Australia, 2017), 3076–3085.

114. Shiba, K. & Kawahara, T. Using propensity scores for causal inference: pitfalls and tips. *Journal of Epidemiology / Japan Epidemiological Association* **31,** 457–463 (Aug. 2021).

115. Shortreed, S. M. & Ertefaie, A. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* **73,** 1111–1122 (Dec. 2017).

116. Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Statistical Science* **25,** 1–21 (Feb. 2010).

117. Stürmer, T., Rothman, K. J., Avorn, J. & Glynn, R. J. 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,** 843–854 (Oct. 2010).

118. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35,** 2769–2794 (Dec. 2007).

119. Tang, D., Kong, D., Pan, W. & Wang, L. Ultra-high dimensional variable selection for doubly robust causal inference. *Biometrics* (Jan. 2022).

120. Thomas, L. E., Li, F. & Pencina, M. J. Overlap weighting: A propensity score method that mimics attributes of a randomized clinical trial. *The Journal of the American Medical Association* **323,** 2417–2418 (June 2020).

121. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288 (Jan. 1996).

122. Tsiatis, A. A. *Semiparametric theory and missing data* 1st ed., XVI, 388. ISBN: 978-0-387-32448-7 (Springer New York, 2006).

123. Tübbicke, S. Entropy balancing for continuous treatments. *Journal of Econometric Methods* **11,** 71–89 (Jan. 2022).

124. Van der Laan, M. Super Learner In Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series* (May 2010).

125. Van der Laan, M. J., Dudoit, S. & Keles, S. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology* **3,** Article4 (Mar. 2004).

126. Van der Laan, M. J. & Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data* illustrated. ISBN: 9781441997821 (Springer Science & Business Media, 2011).

127. Vegetabile, B. G., Gillen, D. L. & Stern, H. S. Optimally balanced gaussian process propensity scores for estimating treatment effects. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)* **183,** 355–377 (Jan. 2020).

128. Vegetabile, B. G. *et al.* Nonparametric Estimation of Population Average Dose-Response Curves using Entropy Balancing Weights for Continuous Exposures. *Health services & outcomes research methodology* **21,** 69–110 (Mar. 2021).

129. Vowels, M. J., Camgoz, N. C. & Bowden, R. D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Comput. Surv.* Just Accepted. ISSN: 0360-0300. `https://doi.org/10.1145/3527154` (Mar. 2022).

130. Wager, S. & Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association,* 1–15 (June 2018).

131. Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A. & Stürmer, T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and Drug Safety* **20,** 317–320 (Mar. 2011).

132. Wu, X., Mealli, F., Kioumourtzoglou, M.-A., Dominici, F. & Braun, D. Matching on Generalized Propensity Scores with Continuous Exposures. *arXiv* (2018).

133. Yao, L. *et al. Representation Learning for Treatment Effect Estimation from Observational Data* in *Advances in Neural Information Processing Systems* (eds Bengio, S. *et al.*) **31** (Curran Associates, Inc., 2018). `https://proceedings.neurips.cc/paper/2018/file/a50abba8132a77191791390c3eb19fe7-Paper.pdf`.

134. Yoon, J., Jordon, J. & van der Schaar, M. *GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets* in *International Conference on Learning Representations* (2018). `https://openreview.net/forum?id=ByKWUeWA-`.

135. Zenati, H., Bietti, A., Martin, M., Diemert, E. & Mairal, J. *Counterfactual Learning of Stochastic Policies with Continuous Actions: from Models to Offline Evaluation* working paper or preprint. Aug. 2021. `https://hal.archives-ouvertes.fr/hal-02883423`.

136. Zhao, Q. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics* **47,** 965–993. `https://doi.org/10.1214/18-AOS1698` (2019).

137. Zhao, Q. & Percival, D. Entropy balancing is doubly robust. *Journal of Causal Inference* **5.** `https://doi.org/10.1515/jci-2016-0010` (Sept. 2017).

138. Zhao, Y.-Q., Zeng, D., Laber, E. B. & Kosorok, M. R. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* **110,** 583–598 (July 2015).

139. Zhou, W., Zhu, R. & Zeng, D. A parsimonious personalized dose-finding model via dimension reduction. *Biometrika* **108,** 643–659 (Sept. 2021).

140. Zhu, Y., Coffman, D. L. & Ghosh, D. A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments. *journal of Causal Inference* **3,** 25–40 (Mar. 2015).