Causal Inference with Neural Network Predictions

by

Mehdi Rostamiforooshani

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Public Health Sciences
Dalla Lana School of Public Health
University of Toronto

Abstract

Mehdi Rostamiforooshani
Doctor of Philosophy
Graduate Department of Public Health Sciences
Dalla Lana School of Public Health
University of Toronto

The estimation of Average Treatment Effect (ATE) as a causal parameter involves modeling treatment and outcome, incorporating potential confounders, and inserting the resulting predictions into ATE estimators such as the Inverse Probability Weighting (IPW) estimator or Augmented Inverse Probability Weighting (AIPW) estimator. Due to the concerns regarding the nonlinear or unknown relationships between confounders and the treatment and outcome, there has been an interest in applying non-parametric methods such as Machine Learning (ML) algorithms instead. Neural Networks (NNs) are a class of complex ML algorithms that can be applied in almost all scenarios even if the confounders are of text or image forms. NNs converge at a slower rate than the $\sqrt{n}$ of the parametric models. In addition, in the scenarios where we have an empirical violation of the positivity assumption, that is propensity scores are too close to zero or one, the ATE estimators including IPW and AIPW will have high variance.

As the first proposed remedy for such situations, we introduce a normalized version of AIPW (nAIPW) which is a consistent estimator of ATE and is asymptotically normal. We perform scenario analysis and simulations to illustrate the superiority of nAIPW over AIPW in scenarios with empirical violation of the positivity assumption. However, the outcome and treatment predictions inserted in these estimators are from 2 separate NNs for the outcome and treatment, referred to as the double NN (dNN).

The NN architectures do not specifically target the confounders nor can they dampen strong effects to avoid the empirical violation of positivity. For this purpose, we propose a joint NN (jNN) architecture in which the output layer has both the treatment and outcome. Additionally, we introduce an extra "targeted" L_1 regularization for dampening extreme propensity values. Simulations demonstrate the superiority of jNN and dNN with the targeted regularization over dNN without the targeted regularization.

Having the option of inserting the predictions of two sets of NN architectures with many hyperparameter settings comes with a disadvantage. We must use cross-validation to either select one scenario and one set of predictions or perform super-learning. We explore an alternative approach of using a recently introduced estimator called the Multiple Robust (MR) estimator. In MR, we do not need to select only one set of predictions for the outcome and treatment, and we can use all the predictions from the trained models. We prove the consistency of MR and numerically explore its performance when we change the number of outcomes and treatment predictions. We propose a general estimating equation framework that generates the MR estimator as a particular case. We show that MR is consistent if either of the treatment or outcome models is consistent. In addition, unlike the previously introduced estimating equation in the literature, we can derive an asymptotic variance estimator which does not need to select a single set of first-step predictions.

Dedication

In Loving Memory of My Mother

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Olli Saarela, for his guidance, support, and encouragement throughout my Ph.D. His expertise and insight were invaluable, and I am deeply grateful for the time and effort he dedicated to helping me complete my Ph.D.

I extend my sincere gratitude to the members of my thesis committee, Dr. Michael Escobar and Dr. Konstantin Shestopaloff, for their constructive feedback and helpful suggestions. Their insights and guidance were essential in shaping the direction and focus of my research. Further, I am thankful to Dr. Wendy Lou for her support and guidance, both in my academic pursuits and personal life.

I would also like to extend my gratitude to my colleagues and peers, including Dr. Mohsen Soltanifar, Derek Latremouille, Bo Chen, and others, for their insights, ideas, and knowledge sharing throughout the coursework period and research process. In addition, I am grateful to the Dalla Lana School of Public Health, Biostatistics division staff for their assistance in obtaining and accessing the resources I needed and for their assistance with administrative services.

I am deeply grateful to my family for their love and support throughout my academic journey. My mother was particularly instrumental in encouraging me to pursue my PhD. Unfortunately, she passed away while I was working on my thesis. She was always a source of love and inspiration to me. Even though she is no longer with us, I am certain that she would have been proud of my achievement and would have supported me every step of the way.

I am deeply grateful to my wife for her unwavering support, encouragement, and endless patience throughout the years of my academic pursuits. Her love and belief in me have been a constant source of motivation and strength. I cannot thank her enough for all that she has done for me.

Contents

List of Abbreviations

Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
AUC	Area Under the Curve
AIPW	Augmented Inverse Probability Weighting
Auto-DML	Auto Double Machine Learning
ATE	Average Treatment Effect
BMI	Body Mass Index
CCHS	Canadian Community Health Survey
CATE	Conditional Average Treatment Effect
DML	Double Machine Learning
dNN	Double Neural Network
EIF	Efficient Influence Function
GDR	General Doubly Robust
GLM	Generalized Linear Models
geo	Geometric Average
IVs	Instrumental Variables
IPW	Inverse Probability Weighting
jNN	Joint Neural Network
L_{1TG}	Targeted L_1 Regularization
ML	Machine Learning
MSE	Mean Square Error
MR	Multiply Robust
MVN	Multivariate Normal
nATE	Naive Average Treatment Effect
nIPW	Normalized Inverse Probability Weighting
NN	Neural Network
PS	Propensity Score
ReLU	Rectified Linear Unit
RMSE	Root Mean Square Error
SR	Single Robust
SGD	Stochastic Gradient Descent

List of Tables

List of Figures

Chapter 1

Introduction and Motivation

1.1 Rationale and Objectives

In the field of causal inference, understanding the cause-and-effect relationship between a treatment or exposure variable and an outcome variable is crucial. Researchers often examine the Average Treatment Effect (ATE), a key parameter in determining causality. While experimental data are ideal for estimating ATE, observational data can be used if we have access to confounders, which are factors that influence both the treatment/exposure and the outcome. Estimating ATE typically involves two steps: developing a model for the treatment and outcome while considering the confounders, and using predictions from this model to estimate ATE.

Although various statistical models and machine learning algorithms have been used for modeling the treatment and outcome in the first step, there has been limited research on the application of Neural Networks (NNs) in this context. NNs are particularly successful in modeling text and image data and hold great promise for analyzing the increasing availability of such data in the future. Thus, there is a need to explore the potential of NN predictions in estimating ATE. However, utilizing NNs in this domain presents theoretical and numerical challenges that require a thorough investigation through theoretical and experimental studies. These challenges can be addressed by either modifying the ATE estimators that consume the NN predictions or designing NN architectures tailored to the causal problem. By addressing these gaps in the literature, the way can be paved for more accurate and effective estimation of causal effects using neural networks.

Thus, the overall goal of this research is to effectively use NN predictions for causal parameter estimation, primarily the estimation of ATE, and study its theoretical and applied properties. Multiple objectives are also studied in this thesis. First, to introduce and study the properties of a more stable doubly robust estimator for ATE that enhances the accuracy of existing estimators when they fail under certain conditions. Second, to design and develop an alternative NN architecture and introduce more regularization to the NN algorithms that improve the performance of the ATE estimator. Third, to study ATE estimators that utilize multiple sets of predictions generated by all the trained NNs, while avoiding the need for hyperparameter tuning.

1.2 Dissertation Outline

This dissertation begins in Chapter 2 with a background of essential definitions and concepts regarding Causal inference, ATE, its estimators, and their known and relevant properties, plus relevant background for NNs. In Chapter 3 we propose a new estimator which is a modification of the Augmented Inverse Probability Weighting estimator. We study its theoretical properties and illustrate its use on simulated and real data. We introduce a new NN architecture and a new L_1 regularization approach in Chapter 4. The methods are designed to target confounders and dampen strong predictors of treatment assignment. In Chapter 5 we use NN predictions in the multiple robust estimator and prove consistency and asymptotic normality, and derive an applicable asymptotic variance estimator. We conclude the thesis in Chapter 6 by discussing the findings and the limitations of this work.

1.3 Application: Food Insecurity and BMI

The Body Mass Index (BMI) is a measure of body fat based on an individual's weight and height. It is calculated by dividing the weight (in kilograms) by the square of the height (in meters). BMI is a continuous variable that can indicate different weight categories based on an individual's height. Lower values may suggest that the person has a low weight relative to their height, while higher values may indicate obesity. Extreme values, whether too low or too high, may suggest an unhealthy weight status. BMI is an important measure for determining an individual's risk for several health conditions, including heart disease, diabetes, and some cancer diseases.

Food insecurity is the limited or uncertain availability of nutritionally adequate and safe foods (Seligman et al., 2010). It can lead to lower BMI, as individuals may not consume enough nutrients for optimal health or may rely on calorie-dense but nutrient-poor foods. Conversely, food insecurity can also lead to higher BMI due to reliance on cheaper, highly processed foods. Food insecurity is the inability to acquire or consume an adequate diet quality or sufficient quantity of food in socially acceptable ways, or the uncertainty that one will be able to do so. Food insecurity within Canadian households is evaluated using the Household Food Insecurity Survey Module (HFSSM) (Jessiman-Perreault and McIntyre, 2019), an 18-question survey that has gained international validation and has been translated into multiple languages. This module was adapted from a food security assessment method initially developed in the United States and has been utilized annually in the U.S. since 1995 to track household food security. The HFSSM evaluates the food security status of adults and children as separate groups within the household, taking into account their experiences over the previous 12 months. It encompasses 10 questions assessing food insecurity among adults and 8 questions for children. Statistics Canada employs these responses to generate a derived variable categorizing households into three levels of food security: food secure, moderately food insecure, and severely food insecure. In this thesis, we have combined moderate and severe levels into one category.

The Canadian Community Health Survey (CCHS) is a cross-sectional survey that collects data related to health status, healthcare utilization, and health determinants for the Canadian population in multiple cycles. The 2021 CCHS covers the population 12 years of age and over living in the ten provinces and the three territorial capitals. Excluded from the survey's coverage are persons living on reserves and other Aboriginal settlements in the provinces; and some other sub-populations that altogether represent less than 3% of the Canadian population aged 12 and over. Examples of modules asked in most cycles are general health, chronic conditions, smoking, and alcohol use. For the 2021 cycle, thematic content on food security, home care, sedentary behavior, and depression, among many others, have been included. In addition to the health component of the survey are questions about the people's characteristics such as labor market activities, income, and socio-demographics.

The dataset comprises almost 40,000 rows and encompasses approximately 300 columns, including both binary and continuous variables serving as adjusting factors that account for confounding influences on the causal association between food insecurity and BMI. Given the large number of variables and their irrelevance to the overall methodologies introduced in this thesis, the list is omitted from this script. However, the interested reader can view the list online at Statistics Canada. It is important to emphasize that the utilization of this dataset is solely intended for illustrative purposes of the developed methods, and thus, the outcomes derived from it do not assert the existence or lack of a causal relationship between the aforementioned variables.

1.4 Authorship Contributions

In the preparation process of this dissertation, Manuscripts based on Chapters 3 and 4 have been published as research papers in a peer-reviewed journal. In these two chapters, minor changes have been made as compared to the published papers to accommodate examiner comments. The manuscript based on Chapter 5 has been uploaded to an online repository. The authors' contributions in these multi-author manuscripts are as follows:

Data Creation, M.R.; formal analysis, M.R.; investigation, M.R.; methodology, M.R. and O.S.; project administration, M.R. and O.S.; resources, M.R.; software, M.R.; supervision, O.S.; validation, M.R. and O.S.; visualization, M.R.; writing-original draft, M.R.; writing-review & editing, O.S. All authors have read and agreed to the published the papers and include them in this manuscript. The articles have been reproduced under the publisher's (MDPI) open-access Creative Commons CC BY 4.0 license.

The papers are:

- Mehdi Rostami and Olli Saarela. "Normalized Augmented Inverse Probability Weighting with Neural Network Predictions." Entropy 24.2 (2022): 179.
- Mehdi Rostami and Olli Saarela. "Targeted L 1-Regularization and Joint Modeling of Neural Networks for Causal Inference." Entropy 24.9 (2022): 1290.
- Mehdi Rostami and Olli Saarela. "Multiply Robust Estimator Circumvents Hyperparameter Tuning of Neural Network Models in Causal Inference.", arXiv.2307.10536.

Chapter 2

Background and Literature Review

This chapter reviews the foundational concepts of causal inference, beginning with an exploration of causal models and inference methods. It then focuses on the key notion of potential outcomes and how they are utilized in causal inference frameworks. The chapter further investigates the assumptions necessary for causal inference and introduces two-step estimation techniques that are commonly employed in causal analysis. Additionally, it formally defines ATE and reviews various ATE estimators, including inverse probability weighting (IPW), normalized IPW (nIPW), augmented inverse probability weighting (AIPW), targeted maximum likelihood estimation (TMLE), and collaborative targeted maximum likelihood estimation (CTMLE). The chapter also touches upon alternative causal effect parameters and the challenges associated with nuisance parameter estimation. Finally, the chapter discusses the properties of estimators, such as consistency and asymptotic normality, and explores the application of M-estimators in causal inference, both with and without nuisance parameters. The chapter concludes with an exploration of the intersection between causal inference and machine learning, highlighting key and necessary points required to understand the methods in this thesis.

2.1 Causal Inference

Causal inference is the process of drawing conclusions about the causal relationships between variables. Randomized controlled trials (RCT) are considered the gold standard for causal inference because they provide the strongest evidence for the causality of a treatment effect on an outcome. One of the earliest attempts to establish causal inference in statistical literature was by the statistician Ronald Fisher in the 1920s, who is credited with the use of randomization in his experiments. However, it was Hill (1955) who formally introduced RCTs which was used for the first time in practice by Crofton and Mitchison (1948). In an RCT, participants are randomly assigned to either the treatment or control group. This ensures that the groups are similar with respect to all observed and unobserved characteristics, and therefore any observed differences in outcomes between the groups can be attributed to the treatment (Imbens and Rubin, 2015).

RCTs also have the advantage of being able to control for other potential sources of bias, such as selection bias, which can occur when the membership to the treatment and control groups is not random, and in fact, there are factors that are related to the outcome that has influenced the group (treatment/control) membership (Imbens and Rubin, 2015). The selection bias can be removed through the use of randomization.

There are several reasons why it may be necessary to base causal inference on observational studies rather than clinical trials. Ethical considerations may prevent the use of an RCT due to the treatment being already in widespread use or the potential risks to participants being too high. Additionally, RCTs can be expensive and time-consuming to conduct, and may not be feasible in all situations. Observational studies can be a more practical and cost-effective option for studying the treatment effect, and can also be used to study the effect of "natural experiments" such as policy changes or natural disasters. Additionally, observational studies can provide external validity by studying the treatment effect in a real-world setting, which may be more representative of the overall population than the controlled environment of an RCT. This can be particularly important when studying the effectiveness of a treatment in a diverse population or when studying rare outcomes (Imbens and Rubin, 2015).

However, observational data can be difficult to use to establish cause-and-effect relationships due to the possibility of confounding, which occurs when other variables are associated with both the exposure/treatment and the outcome being studied. This can distort the estimates of the relationship between the exposure and outcome, making it difficult to produce accurate evidence on causality. Using observational historical data, it is often difficult to control for confounders, particularly if the data are not collected systematically or if the data are limited. For example, if the data do not include information on important confounders, such as age or socioeconomic status, it may not be possible to adequately control for these variables in the analysis (Imbens and Rubin, 2015).

There's a wealth of literature on how to use observational data for causal inference. Austin Bradford Hill put forward a set of nine criteria called "viewpoints" to help establish if the association in an observational study is actually causal (Hill, 1965). Rubin's 1974 article (Rubin, 1974) is one of the first ones that provided a framework for understanding and estimating the causal effect of an exposure (treatment) on an outcome in the presence of confounding. Rubin proposed the use of potential outcomes which are the outcomes that would have occurred if an individual had received the treatment or the control. He argued that, under certain assumptions (Section 2.2.1), the population-level causal effect of the treatment can be estimated by comparing the potential outcomes for a group of individuals.

2.2 Causal Models and Inference

In this section, we introduce the notations, assumptions, and some background on the causal inference for observational data.

2.2.1 Causal Inference and Potential Outcomes

In this context, A is a binary exposure or treatment variable where $A = 1$ represents the exposure or the treatment group and $A = 0$ represents the control group. We use the terms exposure and treatment interchangeably throughout this document. When a treatment is assigned to an individual, we can only observe the outcome associated with that specific treatment. However, it is possible that a different treatment could have been assigned to that individual. The value of the treatment assignment is denoted as A and can be either 1 or 0. The potential outcomes that can be observed or not for the treated and untreated subjects are denoted as Y^1 and Y^0, respectively. The observed outcome is denoted as Y^A and the counterfactual outcome is denoted as Y^{1-A}. For the treatment group, Y^1 is observed and Y^0 is the counterfactual outcome, while for the control group Y^0 is observed and Y^1 is the counterfactual outcome.

In a hypothetical scenario where both the observed and counterfactual outcomes could be observed, the difference between Y^1 and Y^0 for each individual would have a causal interpretation. However, in reality, it is impossible to observe both outcomes for each individual. Instead, we must rely on observed data and make certain assumptions in order to estimate the average causal effect of interest.

2.2.1.1 Notations

Let data $O = (O_1, O_2, ..., O_n)$ be generated by a data generating process P, where O_i is a finite-dimensional random vector $O_i = (Y_i, A_i, W_i)$, with Y as the outcome, A as the treatment and $W = (X_c, X_y, X_{iv}, X_{irr})$ the covariates, where we assume $A = f_1(X_c, X_{iv}, \epsilon_1)$, and $Y = f_2(A, X_c, X_y, \epsilon_2)$, for some functions f_1, f_2, and random errors ϵ_1 and ϵ_2. The covariates are partitioned so that X_c is the set of confounders, X_{iv} is the set of instrumental variables, X_y is the set of y-predictors (independent of the treatment), and X_{irr} is a set of given noise or irrelevant inputs (Figure 4.1). P represents the true joint probability distribution of O and $\hat{P}_n$ is its sample version. Let $\hat{P}_n$ be any distribution of (Y, A, W) such that the marginal distribution of W is given by its empirical distribution, and the conditional distribution of $Y \mid (A = a, W)$ has a finite conditional mean equal to $\mathbb{E}[Y \mid A = a, W]$. Let Q^1 represent the expected outcome of the treated group, where $Q^1 := Q(1, W) = \mathbb{E}[Y \mid A = 1, W]$, and Q^0 represent the expected outcome of the untreated group, where $Q^0 := Q(0, W) = \mathbb{E}[Y \mid A = 0, W]$. Also, let $g(W)$ be the propensity score, defined as $g(W) = \mathbb{E}[A \mid W]$. All expectations are taken with respect to P. The symbol $\hat{}$ on the population-level quantities indicates the corresponding finite sample estimator.

2.2.1.2 Causal Inference Assumptions

The first assumption is Conditional Independence, or Unconfoundedness stating that, given the confounders, the potential outcomes are independent of the treatment assignments $(Y^0, Y^1 \perp A \mid W)$. The second assumption is Positivity which entails that the assignment of treatment groups is not deterministic $(0 < Pr(A = 1 \mid W) < 1)$. The third assumption is Consistency which states that the observed outcomes equal their corresponding potential outcomes $(Y^A = Y)$. Above and beyond these identifiability assumptions, other assumptions are made such as time order (i.e. the covariates W are measured before the treatment, and treatment assignment is performed before measuring the outcome), and that given the observed covariates, data are IID (not clustered such as family members or temporal such as longitudinal data).

2.2.2 Two-Step Estimation

To determine the cause-and-effect relationship at a population level, we need to select a causal effect parameter. In this thesis, we will be using ATE as our causal effect parameter:

$$\beta_{ATE} = \beta_1 - \beta_0 = \mathbb{E}[Y^1 - Y^0] = \mathbb{E}[\mathbb{E}[Y^1 - Y^0 \mid W]] = \mathbb{E}[\mathbb{E}[Y \mid A = 1, W]] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, W]], \quad (2.1)$$

where Y^1 and Y^0 are the potential outcomes as introduced above (Rubin, 1974). Throughout this thesis, we use the notations β_1 and β_0 for the quantities $\mathbb{E}[Y^1]$, and $\mathbb{E}[Y^0]$, respectively.

In most cases, this parameter is dependent on other nuisance or infinite dimensional parameters that are difficult to measure directly and must be estimated using additional methods. These causal effect estimators are called “plug-in estimators” and the process of estimating them is called “two-step estimation." In the first step of this process, the necessary parameters are estimated, and in the second step, they are used to calculate the causal effect parameter.

In the next section, we will review the literature on ATE estimators and their characteristics (second step), and in the following section, we review the methods to estimate the nuisance parameters (first step) and how they affect the causal estimation.

2.2.3 Single and Doubly Robust Estimators

In this section, the initial candidate pool for estimating ATE is presented. In the following section, we will delve into their properties and elucidate their limitations. The estimators are defined as follows

$$\begin{aligned}
\text{nATE} \qquad & \hat{\beta}_{nATE} = \frac{1}{n_1}\sum_{i=1}^{n} A_i Y_i^1 - \frac{1}{n_0}\sum_{i=1}^{n}(1-A_i)Y_i^0, \\
\text{SR} \qquad & \hat{\beta}_{SR} = \hat{\mathbb{E}}\Big[\hat{\mathbb{E}}[Y^1 - Y^0 \mid W]\Big] = \frac{1}{n}\sum_{i=1}^{n}\hat{Q}_i^1 - \hat{Q}_i^0, \\
\text{IPW} \qquad & \beta_{IPW} = \hat{\mathbb{E}}[w^{(1)}Y^1 - w^{(0)}Y^0] = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{A_i Y_i}{\hat{g}_i} - \frac{(1-A_i)Y_i}{1-\hat{g}_i}\Big), \\
\text{nIPW} \qquad & \hat{\beta}_{nIPW} = \sum_{i=1}^{n}\Big(\frac{A_i w_i^{(1)} Y_i}{\sum_{j=1}^{n} A_j w_j^{(1)}} - \frac{(1-A_i)w_i^{(0)}Y_i}{\sum_{j=1}^{n}(1-A_j)w_j^{(0)}}\Big), \\
\text{AIPW} \qquad & \hat{\beta}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{A_i(Y_i - \hat{Q}_i^1)}{\hat{g}_i} - \frac{(1-A_i)(Y_i - \hat{Q}_i^0)}{1-\hat{g}_i}\Big) + \hat{\beta}_{SR}.
\end{aligned} \tag{2.2}$$

where $w_k^{(1)} = \frac{1}{\hat{g}_k}$ and $w_k^{(0)} = \frac{1}{1-\hat{g}_k}$, and $\hat{g}$ is the propensity score estimator and $\hat{Q}$ is the outcome predictions. Additionally, the term nATE represents the naive Average Treatment Effect, SR stands for Single Robust, IPW denotes Inverse Probability Weighting, and nIPW represents the normalized version. AIPW refers to Augmented Inverse Probability Weighting. In the following section, the characteristics of each of these estimators, along with their advantages and disadvantages, will be elaborated upon.

2.2.3.1 Cross-fitting

Cross-fitting represents a modified approach within the realm of two-step estimation techniques. It bears a resemblance to the cross-validation procedure found in machine learning, such as sample splitting and cross-validation. The motivation behind Cross-fitting, as recommended by Rinaldo et al. (2019), is to employ random sample splitting for training a model on one segment while using that trained model to estimate causal parameters on the other segment. This approach allows researchers to relax several assumptions in their analysis.

While sample splitting offers advantages, it also has drawbacks, like diminishing the statistical test's power. Consequently, discarding half of data seems suboptimal. To address these concerns, Chernozhukov et al. (2018a) introduced the Cross-fitting method that achieves full efficiency, assuming mild regulatory conditions. Below is an algorithm illustrating only how outcome predictions are generated:

1. Randomly divide the dataset D into K partitions: $D_1, D_2, .., D_K$, ensuring that D is the union of all partitions, and there is no overlap between partitions, i.e., $D = \bigcup_{i=1}^{K} D_i$ such that $D_i \cap D_j = \varnothing$ for all $i, j \in 1, 2, ..., K$ where $i \neq j$.

2. Train K models, denoted as $\hat{H}_j$ for $j \in 1, 2, ..., K$. Each model is trained by excluding one of the splits, meaning that H_j is trained on $\bigcup_{i \neq j} D_i$. Here, H can represent either the outcome model or the propensity score model.

3. Utilize the trained models to obtain predictions for the split that the respective model hasn't seen before. In other words, calculate $\hat{z}_i = \hat{H}_j(x_i)$ for all input-output pairs (x_i, z_i) in D_j, for $j \in 1, 2, ..., K$. The output can be either the treatment or the outcome.

4. Estimate the parameter K times by plugging in the predictions (both outcome and propensity score) from the K splits separately. This can be expressed as $\hat{\beta}^{(j)} = \hat{\beta}^{(j)}(\hat{H}_j(D_j))$, where $j \in 1, 2, ..., K$.

5. Compute the final estimation as the average of the K parameter estimates: $\hat{\beta} = \frac{1}{K} \sum_{j=1}^{K} \hat{\beta}^{(j)}$.

6. Compute the standard error similar to before, except the estimated values of the nuisance parameters for each subject i are derived from models trained on $K - 1$ folds of the data that do not contain observation i.

Zivich and Breskin (2021) reviews cross-fitting and studies the potential challenges in its application in real-life scenarios.

2.2.4 Properties of ATE Estimators

The initial observation of the ATE parameter suggests the use of a naive ATE estimator (nATE), which computes the difference between the average outcome in the treatment and control groups. However, nATE estimates a parameter that is not the true ATE and, if the treatment assignment is not random, it introduces selection bias. The selection bias arises from factors that are associated with both the treatment and outcome variables, known as confounders. Other estimators mentioned earlier address these factors. The most direct approach is the Single Robust (SR) estimator. In this method, the outcome is modeled using the treatment and other observed (adjusting) covariates, and the predicted values for the treated and untreated groups are calculated and inserted into the SR estimator. SR is a $\sqrt{n}$-consistent estimator of ATE if the first-step model is $\sqrt{n}$-consistent. This is a restrictive assumption as we cannot use slower converging models such as ML algorithms. Also, if the adjusting factors include strong Instrumental Variables (IVs), that is, predictors of the treatment that are uncorrelated with the outcome, the SR estimator will have a large variance.

IPW is a method that can be used to perform causal inference in observational studies, in which, unlike in SR approach, the treatment is modeled rather than the outcome (Lunceford and Davidian, 2004). IPW involves weighting the observations in the analysis by the inverse of the probability of being in their exposure group, given the confounder values, known as propensity score. This seeks to balance the distribution of confounders between the exposure and control groups, allowing for a more accurate estimate of the exposure-outcome relationship. It turns out that the reciprocal of the propensity scores are the same as the coefficients in the the Riesz representation of ATE estimator (Chernozhukov et al., 2018b). Estimating Riesz representer is key to obtaining well-behaved estimates of causal parameters in the presence of nuisance functions, and can also be useful for estimating asymptotic variances (Chernozhukov et al., 2020).

One of the issues with using IPW for causal inference is the potential for model misspecification or consistency of the treatment model. IPW relies on the assumption that the propensity score model accurately represents the relationship between the confounders and the exposure. Another potential issue with IPW is that the weights (reciprocal of propensity scores) can be extreme, meaning that some observations may have very large weights and others may have very small weights. This can occur when the probability of receiving treatment is either very high or very low for certain observations. Extreme weights can lead to unstable estimates and may increase the variance of the estimates.

nIPW is a variant of the IPW estimator that has been proposed to alleviate the issue of unstable estimates (Bang and Robins, 2005). nIPW normalizes the weights by the sum of the weights instead of dividing them by the sample size. This method has numerous advantages compared to IPW, making it a preferred choice in many cases. It can be more robust to model misspecification, resulting in less biased estimates of the exposure-outcome relationship compared to IPW (Lunceford and Davidian, 2004). Additionally, nIPW can be more efficient than IPW in terms of the variance of the estimates especially when the weights are extreme for at least one subject (Lunceford and Davidian, 2004). Another useful feature of nIPW is that the weights can be interpreted as the number of individuals in the control group that is "matched" to each individual in the exposure group based on the confounder values. This can be helpful for understanding and interpreting the results of the analysis (Lunceford and Davidian, 2004).

Provably, IPW and nIPW can become more efficient if the propensity score is accurately estimated rather than being known at the expense of a higher bias of the estimator of the causal parameter. This is rather counterintuitive as the estimators that depend more on the data tend to have higher variance. However, it can be justified that when the propensity score is known, the weights are fixed and cannot be adjusted based on the data. However, when the propensity score is estimated, the weights can be updated based on the data, which can result in more efficient estimates of the exposure-outcome relationship due to the achieved sample balance.

However, both IPW and nIPW estimators rely heavily on the assumption that the propensity scores are $\sqrt{n}$-consistent, which is too restrictive. To address this issue, researchers proposed the AIPW method (Bang and Robins, 2005), which involves modeling both the treatment and outcome. The first research on AIPW was published by Robins et al. (1994) in the context of missing data. AIPW is a biased-corrected version of IPW. In fact, the augmented/debiased term captures the effect of all variables that are not included in the treatment or control groups (Chernozhukov et al., 2020). Further, the AIPW estimator uses full information from the conditioning set to make predictions about the outcome variable, while if all information is used for the propensity score estimation, the positivity assumption might be violated. The violation of the positivity assumption can be avoided by including only the minimal set of adjustment variables in the propensity score model (to avoid extreme weights or violation of the positivity assumption) while including a larger set of adjustment variables in the outcome regression models (for more accurate estimation). This feature selection before modeling the propensity score model can be carried out via Lasso, or forward or backward selection which are data-adaptive and can cause cherry-picking (Taylor and Tibshirani, 2015). The augmented term (also referred to as the debiased, adjustment, or residual term) in the AIPW estimator has an expected value of zero when the estimated propensity scores and regression models are replaced with their true values. This could help stabilize the estimator when the propensity scores are close to 0 or 1 (Glynn and Quinn, 2010). AIPW is also a modified or a debiased version of the SR estimator (subsection 2.2.3) (Chernozhukov et al., 2020). The SR estimator can be biased when it is implemented using a regularized estimator or an insufficiently complex model that does not account for all confounding factors. AIPW addresses this issue by using inverse probability weighting to balance the confounding factors in both the treated and untreated groups.

AIPW has several attractive theoretical properties. The property of consistency for the causal parameter estimator, if either the treatment or outcome models is consistent, is called double robustness. In fact, AIPW is a more reliable method than either IPW or SR because it produces consistent results even if only one of the propensity scores or the outcome models is correctly specified, while the other two estimators require both models to be correctly specified to be consistent. In addition to consistency, if one of the models is $\sqrt{n}$-consistent, AIPW is asymptotically normal too (Lunceford and Davidian, 2004). Additionally, if both the treatment and outcome models are correctly specified, the estimator is most efficient.

Other Doubly Robust (DR) estimators of ATE have been proposed and have been studied in the literature. TMLE is a statistical method for estimating the causal effect of an exposure (treatment) on an outcome in observational studies. It is based on the idea of sequentially updating the estimates of the exposure-outcome relationship using the information in the data. In fact, it involves two phases: First, an "initial" outcome model is estimated along with the propensity score model, and second, the outcome is modeled again using the previous outcome and propensity score models utilizing a target function. The new outcome model predictions are then inserted into the SR estimator. The second phase is referred to as the targeting phase and improves the efficiency and unbiasedness of the estimator. TMLE is asymptotically normal if either of the models is correctly specified, and if both models are correctly specified, it is the most efficient estimator (van der Laan and Rose, 2011). In addition, TMLE is claimed to be a more robust estimator than other methods, particularly when the models for treatment and outcome are not correctly specified.

The positivity assumption violation can happen when the data are sparse. Sparse data refer to situations where there are few or no observations in some of the subgroups of the population (Van der Laan and Rose, 2011). This can make it difficult to estimate the treatment effect accurately, especially if the positivity assumption is violated. In such cases, the bootstrap method can be used to estimate the uncertainty of the treatment effect, even if the positivity assumption is not met (Van der Laan and Rose, 2011). However, bootstrap may still fail if the estimated propensity score for some observations is extremely close to 1 or 0.

To address the empirical violation of the positivity assumption, or extreme weights/propensity scores, van der Laan and Robins (2003) introduced the Collaborative Targeted Maximum Likelihood Estimation (CTMLE). This method combines the estimation of both the treatment and the outcome model to estimate ATE. This collaboration is achieved through an iterative process of updating the models to improve their fit and reduce bias. CTMLE requires both the treatment and outcome models to be $\sqrt{n}$-consistent with the true data-generating process in order to produce unbiased estimates for ATE. Further, it is computationally intensive especially if machine learning algorithms such as Neural Networks are used.

2.2.5 Consistency

An estimator, $\hat{\beta}$, is a consistent estimator of the parameter, β, if it converges in probability to β as the sample size, n, increases. The rate of convergence is a measure of how quickly the estimator approaches the true parameter value as n increases. Technically speaking, if

$$P(|\hat{\beta} - \beta| > c) \leq cn^{-\phi}, \tag{2.3}$$

$\hat{\beta}$ is a consistent estimator of the parameter β with n^{ϕ} rate, where $\phi > 0$, and c is independent of n. The upper bound $cn^{-\phi}$ in (2.3) can be replaced by a slower function such as $\frac{c}{log(n)}$, but this is not usually of interest as the estimator requires a very large data to be accurate enough.

An ML algorithm is considered consistent if the generalized upper bound of the loss function approaches zero as the sample size, n, increases, similar to (2.3). This means that the algorithm is able to accurately make predictions for new, unseen data. For regression problems, the mean squared error is often used as the loss function, while for binary outcome prediction, the negative log-likelihood loss is used. Linear models and the generalized additive models have a rate of consistency of $\sqrt{n}$, while tree-based models have a rate of $n^{\frac{1}{4}}$ (Chernozhukov et al., 2018a).

Visualizing the consistency of an estimator can be a challenging task. Nevertheless, the subsequent theorem provides a means to visualize consistency by observing the asymptotic unbiasedness while diminishing variance.

Theorem 2.2.1. *If $\hat{\theta}_n$ is asymptotically unbiased and its variance is asymptotically zero, $\hat{\theta}_n$ is consistent, where sub-script n means the estimator depends on the sample.*

Proof. By Markov's inequality (or Chebyshev's inequality), we have that for every ϵ independent of n,

$$P(|\hat{\theta}_n - \theta| > \epsilon) = P((\hat{\theta}_n - \theta)^2 > \epsilon^2) \leq \frac{\mathbb{E}[\hat{\theta}_n - \theta]^2}{\epsilon^2} = \frac{1}{\epsilon^2}\Big(\mathbb{E}\Big[\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\Big]^2 + \Big[\mathbb{E}[\hat{\theta}_n] - \theta\Big]^2\Big) \xrightarrow{n\to\infty} 0, \quad (2.4)$$

as the MSE can be written as the decomposition of the bias and variance. □

2.2.5.1 M-estimators

M-estimators are the solutions to estimating equations of the form

$$\sum_{i=1}^{n} \phi_{\hat{\beta}}(O_i) = 0, \tag{2.5}$$

where ϕ does not depend on i, n and O_i. The true parameter β is a solution to

$$\mathbb{E}_P \phi(\beta, O) = \int \phi(\beta, o) dP(o) = \int \phi(\beta, x) f(x) dx = 0, \tag{2.6}$$

where f is the density associated with P.

Examples of M-estimators are sample average, Inverse probability weighting, and Doubly Robust estimators (Lunceford and Davidian, 2004). In fact, all the estimators in (2.2) are solutions to a general estimating equation:

$$\begin{aligned} \sum_{i=1}^{n} \frac{A_i(Y_i - \beta^1)}{g_i} + \eta_1 \frac{A_i - g_i}{g_i} &= 0, \\ \sum_{i=1}^{n} \frac{(1 - A_i)(Y_i - \beta^0)}{1 - g_i} + \eta_0 \frac{A_i - g_i}{1 - g_i} &= 0. \end{aligned} \tag{2.7}$$

By choosing $\eta_1 = -A(Y - \beta^1)$ and $\eta_0 = (1 - A)(Y - \beta^0)$, we get the naive estimator, by choosing $\eta_1 = -\frac{Q^1 - \beta^1 + \frac{A(Y-\beta^1)}{g}}{\frac{A-g}{g}}$ and $\eta_0 = -\frac{Q^0 - \beta^0 + \frac{(1-A)(Y-\beta^0)}{1-g}}{\frac{A-g}{1-g}}$, we get SR, by choosing $\eta_1 = \beta^1$ and $\eta_0 = -\beta^0$, we get IPW, by choosing $\eta_1 = \eta_2 = 0$ we get nIPW, and by choosing $\eta_1 = -(Q^1 - \beta^1)$ and $\eta_0 = (Q^0 - \beta^0)$ we get AIPW.

However, there are estimators that are not directly a solution to a well-known estimator such as the deviation from the sample mean, $\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \bar{Y}|$. This estimator is not a solution to a well-known estimating equation (Stefanski and Boos, 2002). However, by defining a new parameter(s) θ_1, we might be able to form a system of estimating equations with a solution in a vector form where one of the entries is the estimator of interest. For example,

$$\sum_{i=1}^{n}\phi(Y_i, \hat{\theta}_1, \hat{\theta}_2) = \begin{pmatrix} \sum_{i=1}^{n}\Big(|Y_i - \hat{\theta}_2| - \hat{\theta}_1\Big) \\ \sum_{i=1}^{n}\Big(Y_i - \hat{\theta}_2\Big) \end{pmatrix} = 0.$$

This trick will be used multiple times throughout this thesis.

In statistical modeling, estimating equations are equations that are used to estimate certain parameters. These equations may depend on two sets of parameters: a finite-dimensional vector of parameters and a set of nuisance parameters also known as infinite-dimensional parameters. If the nuisance parameters are known it is straightforward to conclude the consistency and asymptotic normality of the estimators of the finite dimensional parameters. However, if true values of nuisance parameters are unknown, they should be estimated using the data and then plugged into their estimates in the estimating equations to calculate the finite-dimensional estimators. This process is called the two-step estimation. In the following subsections, we review the theories on consistency and asymptotic normality in scenarios where the nuisance parameters are known or unknown.

2.2.6 M-Estimator without Nuisance Parameters

The most straightforward scenario is that the nuisance parameters are known, or there are no nuisance parameters. Van der Vaart (2000) shows that in this scenario, the finite-dimensional estimators are consistent and asymptotically normal $\hat{\beta} \xrightarrow{d} MVN(\beta, \frac{V(\beta)}{n})$, where

$$V(\beta) = I(\beta)^{-1}B(\beta)\Big(I(\beta)^{-1}\Big)^T, \tag{2.8}$$

with $I(\beta) = -\mathbb{E}\Big[\dot{\phi}(Y,\beta)\Big]$ (also known as the Fisher information) and $B(\beta) = \mathbb{E}\Big[\phi(Y,\beta)\phi^T(Y,\beta)\Big]$. It is also proven that if we replace β by its estimator $\hat{\beta}$ in practice, we still get the asymptotic result, that is $\hat{\beta} \xrightarrow{d} MVN(\beta, \frac{V(\hat{\beta})}{n})$ where

$$V(\hat{\beta}) = I_n(\hat{\beta})^{-1}B_n(\hat{\beta})\Big(I_n(\hat{\beta})^{-1}\Big)^T, \tag{2.9}$$

with $I_n(\hat{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\dot{\phi}(Y_i, \hat{\beta})$ and $B_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\phi(Y_i,\hat{\beta})\phi(Y_i,\hat{\beta})^T$.

The asymptotic behavior of the estimator when the estimating equations contain unknown nuisance parameters is different. Before diving into these cases, we review the definitions of the Donsker class (Van der Vaart, 2000) and Orthogonality (Chernozhukov et al., 2018a):

Recall 2.2.2 (Donsker Class)**.** *A Donsker class is a class of functions defined on a metric space that satisfies both uniform integrability and tightness conditions (Van der Vaart, 2000). The uniform integrability condition ensures that the class of functions is well-behaved in terms of their integrability, while the tightness condition guarantees that the functions do not stray too far from a compact set.*

In other words, a Donsker class provides a framework for analyzing the asymptotic behavior of a learning algorithm as the sample size grows. It guarantees that, with a sufficiently large sample, the algorithm will achieve good performance without a corresponding increase in complexity.

The Donsker class property ensures that the empirical risk minimization process (i.e. the loss function in machine learning algorithms) converges to the true risk minimization, as the sample size increases. This convergence is achieved without requiring an increase in the complexity of the algorithm or the number of model parameters (Shalev-Shwartz and Ben-David, 2014).

This property is particularly valuable in the context of statistical learning theory, where the goal is to obtain generalization performance while controlling the complexity of the model.

Examples of Donsker classes are the distribution function, the parametric class, the pointwise compact class, the class of smooth functions, Sobolov classes, the class of bounded functions, and any class with finite VC dimension (Van der Vaart (2000) examples 19.6-19.12). In addition, the subsets, unions, closures, convex combinations, and Libschitz transformation of Donsker classes are Donsker too. For a rigorous mathematical treatment of this subject and conditions under which a class of functions is Donsker see Van der Vaart (2000), Chapters 19.1 and 19.2. Kennedy (2016) also reviews the definitions and properties of Donsker classes and provides examples.

Recall 2.2.3 (Orthogonality)**.** *In this dissertation, we employ the notion of orthogonality with a specific interpretation that distinguishes it from the conventional understanding of orthogonality. A comprehensive treatment of orthogonality is presented in Chapter 3. In broad terms, a set of estimation equations attains orthogonality when minor perturbations in the nuisance parameter or the infinite-dimensional parameters within these equations do not yield major changes in the solutions derived from the equations.*

2.2.7 M-Estimator with Nuisance Parameters

Given that an M-estimator contains some unknown nuisance parameter that needs to be estimated using observed data, the consistency and asymptotic normality of the finite-dimensional parameters do not always hold.

Van der Vaart (2000) proves that if the first step model belongs to a Donsker class, under regulatory conditions, the estimators of the finite-dimensional parameters are consistent and asymptotically normal, and the results in the previous subsection hold with the same variance formula where estimated nuisance parameter replace the true values. However, many machine-learning algorithms do not belong to the Donsker class. Chernozhukov et al. (2018a) proves that if the estimator is orthogonal, under regulatory conditions, and if cross-fitting is used, the results in the previous subsection hold. The details of the latter are explained in Chapter 3.

Here we review how to demonstrate the asymptotic normality and obtain the variance estimator from a distinct perspective by considering the estimator as a plug-in estimator. It will be shown how the Donsker assumptions or the ones in Chernozhukov et al. (2018a) can help prove the asymptotic normality.

The plug-in estimator for linear functional $\beta(P) = \int a(x)dP(x)$ is

$$\hat{\beta}(P) = \int a(x)d\hat{P}_n(x) = \frac{1}{n}\sum_{i=1}^{n} a(x_i). \tag{2.10}$$

The influence function is used to approximate the standard error of a plug-in estimator. To derive the influence function (Hines et al., 2022, Wasserman, 2006), we define the Gâteaux derivative of the parameter $\beta(P)$ in the direction $\tilde{P}$.

We first define the parametric submodel technique. The parametric submodel is a one-dimensional parametric model which is a mixture model of P and $\tilde{P}$, that is

$$P_\epsilon = \epsilon\tilde{P} + (1-\epsilon)P, \qquad \epsilon \in [0,1]. \tag{2.11}$$

Riesz's representation theorem (Van der Vaart, 2000) guarantees that there exists a zero mean and finite variance function $\phi(O,P)$ (w.r.t P) such that

$$\begin{aligned}\frac{d\beta(P_\epsilon)}{d\epsilon}\Big|_{\epsilon=0} &= (\tilde{P}-P)(\phi(O,P)),\\ \frac{d\beta(P_\epsilon)}{d\epsilon}\Big|_{\epsilon=1} &= (\tilde{P}-P)(\phi(O,\tilde{P})),\end{aligned} \tag{2.12}$$

ϕ is a functional derivative that characterizes how sensitive β is to changes in the data-generating distribution P, which is referred to as the efficient influence curve or efficient influence function. The equation (2.12) can be written as

$$\begin{aligned}\frac{d\beta(P_\epsilon)}{d\epsilon}\Big|_{\epsilon=0} &= \tilde{P}(\phi(O,P)) = \mathbb{E}_{\tilde{P}}\phi(O,P),\\ \frac{d\beta(P_\epsilon)}{d\epsilon}\Big|_{\epsilon=1} &= -P(\phi(O,\tilde{P})) = -\mathbb{E}_P\phi(O,\tilde{P}).\end{aligned} \tag{2.13}$$

For a given observation o_i, we have that $\phi(o_i,P) = \frac{d\beta(P_\epsilon)}{d\epsilon}\Big|_{\epsilon=0}$ which helps to calculate the efficient influence function ϕ, by taking the derivative on the right side of (2.13).

Recall 2.2.4. *Suppose that f represents the probability density function associated with P (assuming it exists), we express the perturbation of the density as $f_\epsilon(x) = \epsilon\tilde{f}(x) + (1-\epsilon)f(x)$. For practical and computational purposes, we perturb the density in the direction of a single observation $z = (y,a,w)$, resulting in $f_\epsilon(x) = \epsilon\delta_z(x) + (1-\epsilon)f(x)$, where $\delta_z(x)$ represents the Dirac function, with total mass on the single point z. With this formulation, we observe that $\int a(x)\delta_z(x)dx = a(z)$.*

Example 2.2.1. *We use the above method to derive the efficient influence function of $\hat{\beta}$. By definition of ATE we have*

$$\beta_a = \mathbb{E}_P\mathbb{E}_{P_\epsilon}(Y \mid A=a,W) = \int yf_\epsilon(y\mid a,w)f_\epsilon(w)dydw = \int \frac{yf_\epsilon(y,a,w)f_\epsilon(w)}{f_\epsilon(a,w)}dydw, \tag{2.14}$$

where $f_\epsilon(y,a,w)$ is the joint density function of Y,A,W given $A=a$ and $W=w$, $f_\epsilon(y\mid a,w)$ is the conditional density of Y given $A=a$ and $W=w$ and $f_\epsilon(w)$ is the marginal density of W, all under the submodel (2.11)*. Noted that with the above notations, $a(x) = \mathbb{E}_P(Y\mid A=a,x)$. Some algebra (Hines et al., 2022) shows that*

$$\phi_P(O) = \phi(O,P) = \Big(\frac{A}{g}(Y-Q^1)+Q^1-\beta_1\Big) - \Big(\frac{1-A}{1-g}(Y-Q^0)+Q^0-\beta_0\Big). \tag{2.15}$$

and thus

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n)=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\frac{A_i}{g_i}(Y_i-\hat{Q}_i^1)-\frac{1-A_i}{1-g_i}(Y_i-\hat{Q}_i^0)+\hat{Q}_i^1-\hat{Q}_i^0\Big]-\hat{\beta}=$$
$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\frac{A_i}{g_i}(Y_i-\hat{Q}_i^1)-\frac{1-A_i}{1-g_i}(Y_i-\hat{Q}_i^0)\Big], \quad (2.16)$$

as $\hat{\beta}=\frac{1}{n}\sum_{i=1}^{n}\hat{Q}_i^1-\hat{Q}_i^0$.

To study the asymptotic behavior of the estimator $\hat{\beta}$, we need to examine the asymptotic behavior of the scaled difference

$$\sqrt{n}(\hat{\beta}-\beta). \quad (2.17)$$

By the Taylor expansion of $\beta(P_\epsilon)$ about $\epsilon=1$ in the one-dimensional parametric submodel we have

$$\beta(P)=\beta(\tilde{P})-\frac{d\beta(P_\epsilon)}{dt}\Big|_{\epsilon=1}+R(P,\tilde{P}), \quad (2.18)$$

where $R(P,\hat{P}_n)$ is the remainder term. By (2.13) , we have

$$\sqrt{n}\Big(\beta(\tilde{P})-\beta(P)\Big)=-\sqrt{n}P(\phi(O,\tilde{P}))-\sqrt{n}R(P,\tilde{P}). \quad (2.19)$$

If the first term on the right side, $-\sqrt{n}P(\phi(O,\tilde{P}))$ converges to zero, by making a regulatory assumption on the remainder term, we could obtain asymptotic normality results. However, in reality, this term does not converge to zero and can bias the scaled difference. By adding and subtracting the terms $\sqrt{n}(\hat{P}_n-P)(\phi(O,P))$ and $\sqrt{n}\hat{P}_n(\phi(O,\tilde{P}))$, we get

$$\sqrt{n}\Big(\beta(\tilde{P})-\beta(P)\Big)=\sqrt{n}(\hat{P}_n-P)\Big(\phi(O,P)-\sqrt{n}\hat{P}_n(\phi(O,\tilde{P}))\Big)+\sqrt{n}\Big(\hat{P}_n-P\Big)\Big(\phi(O,\tilde{P})-\phi(O,P)\Big)-\sqrt{n}R(P,\tilde{P}). \quad (2.20)$$

Replacing $\tilde{P}$ by $\hat{P}_n$, we get

$$\sqrt{n}(\hat{\beta}-\beta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,P)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n)+\sqrt{n}\Big(P_n-P\Big)\Big(\phi(O,\hat{P}_n)-\phi(O,P)\Big)-\sqrt{n}R(P,\hat{P}_n), \quad (2.21)$$

where the first term is a normal distribution by the Central Limit Theorem, and the third and fourth terms are controlled if the class of functions is Donsker (Recall 2.2.2), and standard smoothing conditions are satisfied, respectively (Van der Vaart (2000), Theorem 19.26). If the estimators of nuisance parameters are not Donsker, data splitting, and cross-fitting plus the regulatory conditions are needed to control these two terms (Chernozhukov et al., 2018a, Farrell et al., 2021) if the estimator is orthogonal (Recall 2.2.3). It is unclear, however, how the second term behaves. This is primarily because the nuisance parameters in this term are estimated from observed data. There are various methods to deal with this term. The following method results in a new estimator for which this term vanishes.

Remark 2.2.1. *The complexity of the expansion for the scaled difference in equation* (2.21) *can be attributed to two significant factors.*

Firstly, the scaled difference depends on the data-generating process P and its data-dependent counterpart $\hat{P}$. The intricacy of the scaled difference analysis is closely attributed to the accuracy of $\hat{P}$ as an approximation of the true P. In straightforward cases, where $\hat{P}$ corresponds to a simple empirical *distribution, the analysis is relatively straightforward. However, our specific scenario introduces a higher level of complexity as $\hat{P}$ encompasses both the empirical distribution and the mean of a conditional distribution denoted as $(Y \mid A = a, W)$. The mean of this conditional distribution is estimated through non-parametric models, such as machine learning algorithms, amplifying the complexity of the analysis considerably.*

Secondly, the complexity arises from the sensitivity of the scaled difference to variations in the causal parameter, β, in relation to the underlying data distribution, P. This aspect of the analysis can become particularly challenging, especially when β involves infinite-dimensional parameters that must be estimated using the same dataset.

Example 2.2.2. *By the definition, the estimator $\hat{\beta}$ is in fact the single robust estimator, $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}_i^1 - \hat{Q}_i^0$. Comparing* (2.16) *and* (2.21) *shows that*

$$\sqrt{n}(\hat{\beta} + \frac{1}{n}\sum_{i=1}^{n} \phi(O_i, \hat{P}_n) - \beta) = \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{A_i}{g_i}(Y_i - \hat{Q}_i^1) - \frac{1-A_i}{1-g_i}(Y_i - \hat{Q}_i^0) + \hat{Q}_i^1 - \hat{Q}_i^0\Big] - \beta\Big) =$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi(O_i, P) + \sqrt{n}(P_n - P)[\phi(O, \hat{P}_n) - \phi(O, P)] - \sqrt{n}R(P, \hat{P}_n). \quad (2.22)$$

We note that the estimator $\hat{\beta} + \frac{1}{n}\sum_{i=1}^{n} \phi(O_i, \hat{P}_n)$ is in fact the AIPW estimator. This implies that $\hat{\beta}_{AIPW} \sim N(\beta, \sigma^2/n)$ given that the nuisance parameters are estimated by models that are Donsker (or cross-fitting is utilized (Chernozhukov et al., 2018a) as this estimator can be shown to be orthogonal.) σ^2 follows the same formula as (2.8).

2.2.8 Role of Machine Learning in Causal Inference: Nuisance Parameter Estimation

There is a body of research on the most effective statistical models or machine learning algorithms to use when estimating the nuisance parameters that are necessary to account for confounding variables in the first step of the two-step estimation process. In the case of ATE, there are two nuisance parameters: the outcome regression function and the propensity score.

It is important to model the relationship between the confounders and the propensity score and outcome models as accurately as possible, and ML algorithms can be particularly useful in this regard because they are non-parametric and can account for non-linearities that simple statistical models such as OLS cannot. Statistical models such as OLS, Lasso, and generalized additive models (GAMs) are $\sqrt{n}$-consistent and can be used in estimators like AIPW, which is guaranteed to be $\sqrt{n}$-consistent and asymptotically normal under certain assumptions. Chernozhukov et al. (2018a) proved that if the rates of convergence of the propensity score and outcome models sum to $\frac{1}{2}$, and if cross-fitting is used (explained in Section 2.2.3.1), under mild regulatory assumptions, AIPW is $\sqrt{n}$ consistent. Based on their article, tree-based models are safe to use. Farrell et al. (2021) proved that utilizing two separate NNs for the treatment and outcome is fast enough to be used as well.

Chernozhukov et al. (2018a) studied the asymptotic normality of orthogonal estimators of ATE, including the AIPW estimator, when two separate ML algorithms are used to model treatment and outcome (a method known as Double Machine Learning or DML). Farrell et al. (2021) also used two separate neural networks (referred to as double NN or dNN) for this purpose but without any additional regularization other than the use of Stochastic Gradient Descent (SGD) for model optimization.

It is possible to design NN architectures that target specific types of features, along with the necessary optimization and regularization techniques. These flexible NN structures and techniques can be easily implemented using deep learning platforms such as pytorch.

Shi et al. (2019) proposed a neural network architecture called DragonNet that models treatment and outcome jointly, using a multi-tasking optimization technique. In this architecture, the interaction of treatment and non-linear transformations of input variables are taken into account. Chernozhukov et al. (2022b) uses the Riesz Representer as the minimizer of a stochastic loss, which provides an alternative for the propensity score estimation, and aims to prevent the empirical consistency assumption violation issue (Rostami and Saarela, 2022a). They also use joint modeling of the Riesz Representer and outcome through multi-tasking, calling their method Auto Double Machine Learning (Auto-DML). In a related study, Chernozhukov et al. (2022c) optimized an L_1 regularized loss function to estimate weights instead of estimating propensity scores and plugging them into the AIPW estimator, while Chernozhukov et al. (2020) proposed optimizing a minimax loss function for the same purpose. However, it is still unclear how to properly tune the hyperparameters for the chosen neural network architecture for causal inference, particularly for average treatment effect estimation.

Other techniques such as feature selection before propensity score estimation have been proposed in the literature. There have been several feature selection methods proposed for causal inference, including those based on the Lasso (Shortreed and Ertefaie, 2017), a combination of the Lasso and IPW estimator (Ju et al., 2019), the random forest, the elastic net, and the gradient boosting machine (Bakhitov and Singh, 2022, Chernozhukov et al., 2022a). These methods have been shown to be effective at identifying relevant features and improving the accuracy of the average treatment effect estimate. However, hard thresholding might ignore important information hidden in the features.

There are no restrictions on the types of models that can be used to estimate the nuisance parameters in causal inference. Multiple models from different classes of estimators or with different hyperparameter configurations can be used to generate multiple predictions for the nuisance parameters. It can be difficult to determine which model to use for estimating the nuisance parameters in causal inference, as the nuisance parameters are based on missing counterfactual outcome data and are evaluated based on the accuracy of their predictions for the observed data rather than the counterfactuals.

One approach to selecting the best model is to use predictive performance measures, while another approach is to use a super learner method that combines the predictions from multiple models into a weighted average using cross-validation (Van der Laan et al., 2007a). A limitation of selecting a pair of models (for the outcome and propensity score models) is that the selected model that has outperformed other models in terms of the predictive measures, might not be consistent and the theoretical results for the ATE estimators might not hold. The super learning method rectifies this issue as the final weighted average predictions will be consistent if one of the models is so (Van der Laan et al., 2007a). However, the same challenge still holds for utilizing super learner which is estimating the weighted average coefficients using the criterion that the predicted values should be as close as possible to the observed data (on the validation dataset). Additionally, the super learner algorithm may be sensitive to the choice of the validation set used to select the weights in the weighted average. If the validation set is not representative of the overall data distribution, the super learner may perform poorly on out-of-sample data. Further, violation of the assumptions for using super learner such as independence of the data, and well-calibrated candidate algorithms (Van der Laan and Rose, 2011), can impact the performance of the causal parameter estimator.

Han and Wang (2013) employed the Multiple Robust (MR) estimator in the context of missing data. MR makes use of all available models and does not require hyperparameter tuning or model selection. The name comes from the fact that the MR estimator is a $\sqrt{n}$-consistent estimator of ATE if at least one of the propensity score or outcome models is $\sqrt{n}$-consistent. The MR estimator is an empirical maximum likelihood estimator that requires the solution of a constrained convex optimization problem. However, due to numerical issues, Han (2014) proposed a new optimization method that is both theoretically and numerically stable and leads to a multiply robust estimator of ATE. Naik et al. (2016) extended the method to causal inference with a multi-category treatment but did not provide any mathematical proofs and referred the reader to the paper on missing data (Han and Wang, 2013). Wang (2019) considered the problem for a binary outcome and provided theoretical proofs. In the aforementioned research, the first step models are parametric models and thus $\sqrt{n}$-consistent.

2.3 Machine Learning

This section introduces some basic machine learning concepts in order to provide a foundation for the research.

2.3.1 Learning

The process of improving one's performance on a task through experience is called learning. It involves adapting to new information and using it to make better decisions or predictions. In the context of machine learning, we are interested in how well a machine can use data to improve its performance on a specific task. For instance, a learning algorithm might be fed a set of input-output pairs and try to discover a pattern or rule that can accurately predict the output for any given input Shalev-Shwartz and Ben-David (2014).

2.3.2 Regression and Classification

The goal of regression is to learn a function that maps the input variables X to the outcome variable Y, such that the predicted values $\mathbb{E}[Y \mid X]$ are as close as possible to the true values. The quantity $\mathbb{E}[Y \mid X]$ is referred to as the regression function. In machine learning literature, if Y is continuous, the problem is called a regression problem, and if Y is categorical, the problem is called classification. Noted that if Y is binary that takes values $\{0, 1\}$, $\mathbb{E}[Y \mid X] = P(Y = 1 \mid X)$.

2.3.2.1 Parametric, Non-Parametric, Semi-Parametric

In parametric models, we assume that the regression function takes a parametric form such as a linear or polynomial form. The number of parameters to be estimated is finite. In a non-parametric setting, the model does not make any assumptions about the form of the underlying relationship, or the regression function: $\mathbb{E}[Y \mid X] = F(X)$, for any function F. In practice, however, we use algorithms to estimate this function such as decision trees, an ensemble of trees, or neural networks. In fact, we do not assume that a decision tree produced the regression function as opposed to a parametric model. Further, in the non-parametric setting, the regression function is regarded as an infinite-dimensional parameter to be estimated. In a semi-parametric model, the regression function is assumed to consist of two separate components; a finite number of parameters (that are usually of interest) and a finite set of infinite-dimensional parameters.

2.3.3 Prediction Performance Measures

In the field of machine learning, the objective is to accurately estimate the regression function (see subsection 2.3.2). When multiple models are trained, performance measures can be used to determine which of the models performs best on unseen data. A common performance measure for continuous outcome predictions is R^2, also known as the coefficient of determination (Friedman et al., 2001):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{2.23}$$

R^2 is used to compare the performance of a model with the simple average of the outcome. A model that performs better than the average is considered favorable. An R^2 value close to 1 is the evidence that the model fit is satisfactory. It is worth noting that if a model performs worse than the average, R^2 is negative. When comparing models, the model that has higher R^2 is favorable.

For binary classification problems, a different performance measure called the area under the receiver operating characteristic curve (AUC) can be used. AUC is a discrimination measure, that is, it measures the model's ability to distinguish between predicted values of one class and the other. To be more accurate, AUC represents the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006). AUC values range from 0 to 1, with a value of 0.5 indicating no discrimination and a value of 1 indicating perfect discrimination. A value smaller than 0.5 means it performs worse than a simple average model.

If the goal of the model is classification, AUC may suffice to measure the accuracy. However, if the accuracy of the predicted probabilities is also of interest, calibration methods may be used to assess the model's performance, and if the predicted probabilities are not calibrated, they should be calibrated. Calibration measures (such as the Brier score) can evaluate how closely the predicted probabilities match the actual labels, on average.

This section presents some foundational concepts in Neural Networks and optimization techniques in order to better understand the research.

2.3.4 Neural Networks and Artificial Intelligence

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. One of the main tasks of AI is to study how to automate human repetitive work using learning patterns in historical data. Driving a car is an example of repetitive work and autonomous cars with AI technology try to replace car drivers. ML and more specifically NNs are among the methods used to achieve this goal.

The NNs and most ML algorithms can be viewed as nonparametric statistical models where we do not assume any distribution for the data. The ML algorithms or models are associated with a set of parameters or weights that can be estimated by optimizing a pre-defined loss. This is referred to as fitting or training the ML/NN model. The models also contain hyper-parameters that can be chosen so the model is generalizable (in a cross-validation process).

The Global Approximation Theorem (Hornik et al., 1989) states that a shallow NN (a network with only one hidden layer) can approximate any Borel measurable function from one finite-dimensional space to another with any desired non-zero amount of error, provided that the network is given enough "hidden units" (Goodfellow et al. (2016); page 198). However, in practice, shallow networks with many neurons are not as successful as networks with many hidden layers but small number of neurons in each layer. Training the data with the latter type of network is referred to as Deep Learning which has dominated the field of AI. In Computer Vision and Natural Language Processes, two major sub-fields of AI, the researchers and practitioners no longer use the traditional ML algorithms so the word "AI" and NN or DL are used interchangeably. For theoretical guarantees, in 2017, another version of the Global Approximation Theorem was proved for DL networks with an arbitrary depth and with the ReLU activation function (see below) (Lu et al., 2017).

NNs were inspired by the way the brain works and are designed to mimic human behavior. The biological neural network consists of a group of neurons that are connected through synapses. Neurons are cells that can be electrically stimulated and are able to communicate with other cells. They have a cell body (soma) that contains the nucleus, dendrites that receive signals from other cells, and an axon that transmits signals to other cells across synapses. These specialized connections are called synapses. The concept of artificial neurons is based on the idea of receiving, processing, and transmitting signals, and the inputs and outputs of artificial neurons are similar to those of biological neurons.

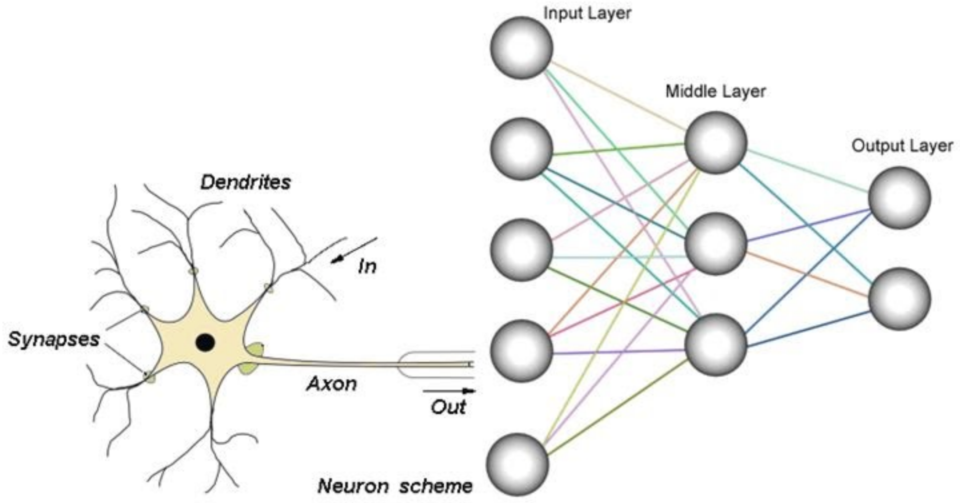


Figure 2.1: The left side the image of a biological Neural Network and the right side is an artificial neural network. The right side is a neural network with one hidden/middle layer.

An Artificial Neural Network, contains an input, an output, and a finite number of hidden layers (Depth of NN) with an arbitrary number of neurons (Width of NN). Any data structure, such as images, texts, or raw table sheets if converted to numbers can be fed to the input of an NN, and based on the definition of the output, it can produce the outcome of interest. For categorical outcomes, it estimates the probability of some event or estimates the value of some measurement for continuous outcomes, or a combination of them. For example, inputs can be brain scans and the output can be the probability that there is a trace of some lesion in the brain. Also, an image detection task is an example of continuous and binary outcomes, where objects need to be detected as well as the category of the object inside the box. In such models, the NN model outputs the coordinates of the box around objects and provides the probability of the object category inside the box.

The NNs are often visually represented with images such as Figure 2.1, but they have equivalent equations and formulas that allow for estimation and prediction. As an example of a simple neural network, consider a neural network with X as the input matrix, 2 hidden layers, and y as the output, the NN model is

$$g(y) = b_3 + f\Big(b_2 + f(b_1 + X\Omega_1)\ \Omega_2\Big)\ \Omega_3 + \epsilon, \tag{2.24}$$

where b_i's are intercepts (also referred to as biases) and Ω_i's are the matrices containing the parameters that connect each layer. The row and column size of Ω determine the number of neurons in the previous and next layer, respectively. f is called the activation function and is a non-linear function (if it is chosen to be linear, the network is reduced to a linear model). For many years, the common choice for f was the logistic or tanh function, but it has been shown that the Rectified Linear Unit (ReLU) function and its variants comparatively perform better for most of the AI tasks (Mercioni and Holban, 2020). g is the link function and can be defined by any function that generalizes the result of the network better; for classification problems, the logistic function and for regression problems the identity function are the most common choices.

As the equation 2.24 illustrates, deep neural networks are nothing but the composite of non-linear functions. Cheng et al. (2018), also, showed that neural networks are equivalent to polynomial regression models with some finite degree.

The loss function ($L(X, \Omega)$) is another component of neural networks which should be carefully chosen. For regression problem, the Mean Square Error (MSE) or Mean Absolute Error (MAE) is chosen and for classification problems, the Cross-entropy (negative log-likelihood of binomial distribution, Goodfellow et al. (2016); page 178).

The NNs optimization problems are often non-convex (Boyd and Vandenberghe, 2004), and also due to the large number of parameters in the network, the dominant choice of the optimization algorithm is the Gradient Descent or its variant. The Stochastic Gradient Descent (SGD, Goodfellow et al. (2016); 152) and its newer version Adam (Kingma and Ba, 2014) are two major optimization algorithms in practice. The stochastic algorithms help the algorithm not be trapped in a valley (of the parameter space) that does not have a local minima corresponding to the best parameter estimation. In fact, in first finite number of steps, the algorithm is likely to jump from one valley to another until finding a deep enough valley.

The learning rate and the moments (Goodfellow et al., 2016) are other hyper-parameters that need to be chosen in the cross-validation process. In Stochastic optimization, the data are randomly divided to different sizes (called the batch size which can be as small as one observation). Then, instead of using all of the data for one iteration of optimization, we repeat the same step for all the random batches sequentially until all batches are seen by the model; this is repeated until satisfactory convergence. Mathematically speaking, if $J_t = \sum_{i=1}^{n} L_i(X, \Omega^t)$ with L_i as the loss function with estimated parameters in step t, the GD algorithm is

$$\Omega^{t+1} = \Omega^t - \alpha \Delta J_t, \tag{2.25}$$

while the SGD is

$$\Omega^{t+1,b} = \Omega^{t,b} - \alpha \Delta J_{t,b}, \tag{2.26}$$

where $J_{t,b} = \sum_{k \in b}^{n} L_k(X, \Omega^{t,b})$ with b as one of the random sub-samples

The batch-size and the number of times the optimization is repeated over the batches, the so called epoch number, are two hyper-parameters. SGD needs more iterations to converge, but each iteration is much faster than the whole data as in GD. Theoretically (Goodfellow et al. (2016); 152-153), the estimated parameters by SGD are consistent (we need to make sure about the exact statement of the theorem). The stochastic optimization is sometimes considered as a type of regularization technique as they reduce the complexity of the algorithm by not allowing the network to be trained on all of the data (Farrell et al., 2021).

2.3.4.1 Back Propagation and Automatic Differentiation

Each iteration of optimization needs the derivative of the loss function with respect to all the parameters. As the NNs are composed of activation functions, the chain rule is utilized for the calculation of the derivative of the loss with respect to the parameters/weights. Back-propagation (Goodfellow et al. (2016); page 204) is an efficient algorithm that uses the chain rule computationally efficiently. The algorithm starts off with the derivative of the loss with respect to the parameters in Ω_3 which is straightforward. For the derivative of the loss with respect to Ω_2, using the chain rule, the derivative of loss with respect to the last hidden layer is calculated and is multiplied by the derivatives of the last hidden layer with respect to Ω_2. This continues until all the required derivatives are calculated. For computational efficiency purposes, The derivatives with respect to each layer are kept in the memory and are used for calculation in the previous layer.

The derivatives are calculated automatically using the Automatic Differentiation (Auto-diff) method. Auto-diff refers to a general way of taking a program that computes a value and automatically constructing a procedure for computing derivatives of that value. Automatic Differentiation is both efficient (linear in the cost of computing the value) and numerically stable. It is different from finite differences which is known to be numerically unstable, nor is symbolic differentiation which can be extremely complicated for complected functions such as NNs. It is rather mathematically programmed; the derivatives with respect to mathematical operations and famous functions are programmed. These pre-programmed derivatives are consequentially used when taking the derivative of the loss with respect to the neurons and weights. The Autodiff is possible in NNs as the network is the composite of known functions (such as logistic, tanh or ReLU) and the chain rule is used (Szirmay-Kalos, 2021).

The current deep-learning platforms are all equipped with automatic differentiation and this has been one of the main reasons behind the success of DL. With automatic differentiation, there is no limitation on the complexity of NN architectures; the automatic differentiation can take derivatives of any loss with all of the NN's parameters.

2.3.4.2 Regularization

Since NNs are usually very complex algorithms, one or more regularization techniques are usually employed to control the generalization of the network. This is because more complex model might overfit and do not provide accurate results on the unseen data. Thus there must be a compromise between the bias and variance of the predictions.

The oldest regularization technique is the L_p penalization, where a term consisting a function of the parameters is added to the loss function:

$$L = loss + \lambda h(\Omega), \tag{2.27}$$

where $h(V) = \sum_{v \in V} |v|$ (for L_1 regularization) or $h(V) = \sum_{v \in V} v^2$ (for L_2 regularization). λ is the regularization rate which is another hyper-parameter for the network.

Dropout (Srivastava et al., 2014), which involves the random deactivation of a fraction of weights Ω during each iteration, is another type of regularization that has gained major success in the AI tasks, but as we will not use it in this thesis, we refer the readers to the original paper or Goodfellow et al. (2016). Above and beyond these regularization techniques, other optimization approaches such as SGD and early stopping, which involves terminating optimization before reaching a local minimum, can be utilized as regularization techniques that improve the generalization of the ML results (Huang et al., 2015, Roux et al., 2012).

2.3.4.3 Batch Normalization

In practice, as the NNs are complex algorithms, the theory might not apply and satisfactory results might not be achieved. One of the reasons NNs' applicability was limited for many years was that the distribution of data in the hidden layers tends to shift during the optimization (Ioffe and Szegedy, 2015). In theory, the optimization of all parameters should be carried out simultaneously without a change in the distribution of the data but this does not happen in practice (Goodfellow et al. (2016), page 317). Batch Normalization is a technique that standardizes the distribution of hidden layers in each step of optimization. In fact, the equation (2.24) is changed to

$$g(y) = b_3 + f(b_2 + \frac{f(b_1 + \frac{X\Omega_1 - \mu_1}{\sigma_1})\Omega_2 - \mu_2}{\sigma_2})\Omega_3 + \epsilon, \tag{2.28}$$

where μ_i's and σ_i's are learnable vectors of estimated mean and standard deviations of each neuron on the hidden layers.

2.3.4.4 Consistency of NNs

Consistency of an ML algorithm can have different meanings (Shalev-Shwartz and Ben-David, 2014). By consistency of a learning algorithm in this thesis, we mean that by increasing the sample size, the empirical loss function tends to zero. Farrell et al. (2021) studied the consistency of Neural networks for the continuous and binary target (outcome). They proved that if the only regularization is the SGD or early stopping, under certain regulatory assumptions, the two types of neural networks are almost $n^{\frac{1}{4}}$ consistent, where the Mean Square Error loss and Cross-entropy are used for the continuous and binary outcomes, respectively. Shen and Lin (2022) extended similar results to regularized NN architectures.

2.4 Simulation Studies

Simulation studies refer to computer-based experiments that entail generating data through pseudo-random sampling techniques (Morris et al., 2019). Simulating data is a computational tool in statistics that is widely used to explore complex data, estimate quantities that are either complex or infeasible to estimate, such as estimator's bias and variance, and test hypotheses. This section provides a quick review of simulations in statistics, including their definition, benefits, and limitations. However, we do not provide a comprehensive review of simulation studies and all their applications and benefits; we rather present the most relevant points for this research.

Statistical analysis often requires determining the distribution of an estimator and its accuracy in estimating the parameter of interest, but there's usually only one observed dataset available, making it difficult to obtain distributional information for a single number. In theoretical statistics, the researcher assumes that the underlying data follow some distribution and might be able to derive a distribution for the estimator that reveals all the properties of the estimator. Nevertheless, these theoretical results may not correspond to reality, and their applicability must be validated. Additionally, the form of the estimator can be complex, which can make it difficult to discover theoretical results or limit their discoveries. Lastly, deriving theoretical finite sample properties is often a more challenging task than deriving asymptotic properties. Simulations can be used to explore the possibility of extending results to small samples and to verify large sample properties, especially when slow-converging algorithms such as Neural Networks are utilized, which typically require massive amounts of data to show satisfactory performance.

Computational statistics address this problem by generating the underlying distribution multiple times, typically hundreds or thousands of times, and calculating the estimator in each iteration. This produces a sample (with a size equal to the number of iterations) for the estimator called the sampling distribution, which allows for the visualization and calculation of the observed sample's proximity to theoretical results. Generating a sample from the estimator can also provide an estimation of the measures such as bias or variance of the estimator, which are difficult or impossible to obtain in real life. Another use case in which simulations are helpful is when specific assumptions are not fulfilled and the expected properties based on theory may not be seen. In such cases, numerical simulations can aid in determining whether the theoretical principles remain valid and under what circumstances they hold. This approach is frequently used to develop new theorems.

Various distributions in statistics can be used to draw samples, and computational software packages such as SAS, R, or Python are commonly used for this purpose. Researchers often set specific scenarios and iteratively generate samples for each scenario. Usually, for each iteration, the parameters of the problem are fixed and only variables are randomly generated.

However, the main limitation of simulation studies is the inability to generate all possible scenarios. Simulations can only demonstrate the accuracy or relevance of the estimator in certain scenarios, and cannot prove anything conclusively. Infinitely many scenarios may exist that are not covered by the simulations and that can produce different results.

2.5 Software Packages

All of the statistical analyses, simulations, and computations for this thesis were conducted using Python 3.6. The neural network algorithms were developed in the Pytorch library, and other important libraries utilized were sklearn (for preprocessing and building Oracle models), numpy (for matrix computations not related to neural networks), pandas (for aggregating and computing estimators and their performance measures), and plotly (for visualizations). The codes were executed on a personal Linux system with 16GB of memory and an Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz. For the neural network computations, an NVIDIA Corporation product, GP104 GPU [GeForce GTX 1070], with 4GB of memory was used. All codes for the different chapters of this thesis are accessible on Mehdi Rostami's personal GitHub page[1].

2.6 Scope of the Thesis

To ensure clarity, it is important to outline the aspects that are not addressed or included in this thesis:

- The thesis centers solely on Rubin's framework and does not delve into Pearl's framework or the application of graphical models in causal inference.
- The primary focus is on estimating ATE rather than CATE. Additionally, the thesis employs two-step estimation methods for ATE, as opposed to utilizing other methods and estimators such as the Average Treatment Effects on the Treated (ATT), instrumental variable method, or Propensity Score Matching.
- This thesis solely concentrates on implementing NNs for ATE estimation, rather than utilizing all types of ML algorithms. Thus the performance of NNs for causal effect estimation is not compared and contrasted with alternative algorithms.
- Feature selection for modeling the treatment or outcome variables is not pursued. Therefore, all available covariates are utilized during the treatment or outcome modeling process.
- The NN tools and techniques used in this thesis are means for calculating ATE estimators and are not subject to study themselves. Thus, this thesis does not present or prove theoretical results such as the convergence of NNs or the interpretability of NN. However, relevant references are provided for readers who are interested in exploring these aspects.
- In this thesis, the adjusting factors such as confounders are assumed to be in a matrix form. Other types of data structures such as text or image data are out of the scope of this thesis.

[1] Mehdi Rostami's personal GitHub page: https://github.com/mehdirostami/CI_NN

Chapter 3

Normalized Augmented Inverse Probability Weighting with Neural Networks Predictions

3.1 Introduction

Estimation of causal parameters such as ATE in observational data requires confounder adjustment. The estimation and inference are carried out in two steps: In step 1, the treatment and outcome are predicted by a statistical model or ML algorithm, and in the second step the predictions are inserted into the causal effect estimator. If ML algorithms are employed in step 1, the non-linear relationships can potentially be taken into account. The relationship between the confounders and the treatment and outcome can be non-linear which make the application of ML algorithms, which are non-parametric models, appealing. Farrell et al. (2021) proposed to use two separate neural networks (double NNs or dNNs) where there is no regularization on the network's parameters except SGD in the NN's optimization (Cheng et al., 2018, Goodfellow et al., 2016, Xu et al., 2015, Yan et al., 2020). They derive the generalization bounds and prove that the NN's algorithms are fast enough so that the asymptotic distribution of causal estimators such as the AIPW estimator (Lunceford and Davidian, 2004, Van der Laan and Rose, 2011) will be asymptotically linear, under regulatory conditions and the utilization of cross-fitting (Chernozhukov et al., 2018a).

Farrell et al. (2021) argue that the fact that SGD-type algorithms control the complexity of the NN algorithm to some extent by focusing on less data for optimization in each iteration (Goodfellow et al., 2016, Zhang et al., 2021) is sufficient for the first step. Our initial simulations and analyses, however, contradict this claim in scenarios where strong confounders and IVs exist in the data.

Conditioning on IVs is harmful to the performance of the causal effect estimators such as ATE (Myers et al., 2011) but there may be no prior knowledge about which covariates are IVs, confounders or otherwise. The harm comes from the fact that the complex NNs can provide near-perfect prediction in the treatment model which violates the empirical positivity assumption (Diaz, 2018).

The positivity assumption (Section 3.2) is fundamental to hold to have an identifiable causal parameter in a population. However, in a finite sample, although the parameter is identifiable by making the positivity assumption, the bias and variance of the estimator can be inflated if the estimated propensity scores are close to zero or one bounds (or become zero or one by rounding errors). This is referred to as the empirical positivity assumption which is closely related to the concept of sparsity studied in Chapter 10 of Van der Laan and Rose (2011). The violation of the empirical positivity assumption can cause the inflation of the bias and variance of IPW-type and AIPW-type estimators.

The inverse probability weighting method dates at least back to Horvitz and Thompson (1952) in the literature of sampling with unequal selection probabilities in sub-populations. IPW-type and matching methods have been extensively studied Busso et al. (2014), Lunceford and Davidian (2004), Rosenbaum and Rubin (1983, 1985), Rubin (1979). IPW is proven to be a consistent estimator of ATE if the propensity scores (that are the conditional probability of treatment assignments) are estimated by a consistent parameter or non-parametric model. The other set of ATE estimators includes those involving the modeling of the outcome and inserting the predictions directly into the ATE estimator (Section 3.2). It is referred to as SR estimator as they provide $\sqrt{n}-$consistent estimators for ATE if the outcome model is $\sqrt{n}-$consistent. In this sense, IPW is also single robust as it is consistent if the treatment (or the propensity score) model is $\sqrt{n}-$consistent. The focus of this work is to study the **augmented** IPW-type methods as they involve modeling both treatment and outcome and can be $\sqrt{n}-$consistent estimators of ATE if either of the models is consistent.

We propose and study a simple potential remedy to the empirical positivity violation issue by studying the normalization of the AIPW estimator (similar to the normalization of IPW (Lunceford and Davidian, 2004)), here referred to as nAIPW. In fact, both AIPW and nAIPW can be viewed as a more general estimator which is derived via the efficient influence function of ATE (Hahn, 1998, Hines et al., 2022).

A general framework of estimators that includes nAIPW as a special case was proposed by (Sloczynski and Wooldridge, 2018). In their work, the authors did not consider machine learning algorithms for the first-step estimation but rather assumed parametric statistical models estimated by likelihood-based approaches. They focused on how to consistently estimate ATE within different sub-populations imposed by the covariates. There is a lack of numerical experimentation on these estimators especially when IVs and strong confounders exist in the set of candidate covariates.

To the best of our knowledge, the performance of nAIPW has not been previously studied in the machine learning context, under the assumption that strong confounders or IVs exist in the data. We will prove that this estimator has the doubly robust (Lunceford and Davidian, 2004) and the rate doubly robust (Hines et al., 2022) property, and illustrate that it is robust against extreme propensity score values. Further, nAIPW (similar to AIPW), has the orthogonality property (Chernozhukov et al., 2018a) which means that it is robust against small variations in the predictions of the outcome and treatment assignment predictions. One theoretical difference is that AIPW is the most efficient estimator among all the double robust estimators of ATE given both treatment and outcome models are correctly specified (Scharfstein et al., 1999). In practice, however, often there is no a priori knowledge about the true outcome and propensity score relationships with the input covariates and thus this feature of AIPW is probably of less practical use.

We argue that for causal parameter estimation, dNN with no regularization may lead to high variance for the causal estimator used in the second step. We compare AIPW and nAIPW through a simulation study where we allow for moderate to strong confounding and instrumental variable effects, that is, we allow for possible violation of the empirical positivity assumption. Further, a comparison between AIPW and nAIPW is made on the CCHS dataset where the intervention/treatment is the food security vs. food insecurity and the outcome is individuals' BMI.

Our contributions include presenting the proof for the orthogonality, doubly robust and rate doubly robust property of nAIPW. Further, it is proven that, under certain assumptions, nAIPW is asymptotically normal and we provide its consistent variance estimator. We analyze the estimation of ATE in the presence of not only confounders, but also IVs, y-predictors and noise variables. We demonstrate that in the presence of strong confounders and IVs, if complex neural networks without L_1 regularizations are used in the step 1 estimation, both AIPW and nAIPW estimators and their asymptotic variances perform poorly, but, relatively speaking, nAIPW performs better. In this paper, the NNs are mostly used as means of estimating the outcome and treatment predictions.

The organization of the article is as follows. In Section 3.2 we will formally introduce the nAIPW estimator to the readers and state its double robustness property, and in Section 3.3 we present the first-step prediction model, double neural networks. In Sections 3.4 and 3.5 we will present the theoretical aspects of the paper, including the asymptotic normality, double robustness and rate double robustness orthogonality of the proposed estimator (nAIPW) and the asymptotic normality. We will present the simulation scenarios and results of comparing the nAIPW estimator with other conventional estimators in Section 3.6. We apply the estimators on a real dataset in Section 3.7. The article will be concluded with a short discussion on the findings in Section 3.8. The proofs are straightforward but long and thus are included in Appendix 7.2.

3.2 Normalized Doubly Robust Estimator

Consider an IID dataset denoted as $O = (O_1, O_2, \ldots, O_n)$, generated by the data generating process P. Each individual observation O_i can be described as a finite-dimensional random vector $O_i = (Y_i, A_i, W_i)$. In this context, Y represents the outcome, A denotes the treatment assignment, and $W = (X_c, X_y, X_{iv}, X_{irr})$ constitutes a collection of covariates. We assume that the treatment variable A follows a Bernoulli distribution with probability π, where $\pi = f_1(X_c, X_{iv})$, and the outcome variable $Y = f_2(A, X_c, X_y) + \epsilon$, for some functions. The covariates are categorized into different sets: X_c represents the confounding variables, X_{iv} are the instrumental variables (correlated with the treatment but not the outcome), X_y represents the Y predictors (correlated with the outcome but not the treatment), and X_{irr} includes irrelevant or noise inputs (as shown in Figure 1).

The symbol P represents the true joint probability distribution of the observed data O, while $\hat{P}_n$ denotes its finite sample approximation; that is any distribution of (Y, A, W) such that the marginal distribution of W is approximated using the empirical distribution, and the conditional distribution of $(Y \mid A = a, W)$ having a finite mean $\mathbb{E}[Y \mid A = a, W]$.

Moreover, we introduce three quantities referred to as the nuisance or infinite-dimensional parameters: Q^1 represents the expected outcome for the treated group, defined as $Q^1 := Q(1, W) = \mathbb{E}[Y \mid A = 1, W]$, and Q^0 represents the expected outcome for the untreated group, defined as $Q^0 := Q(0, W) = \mathbb{E}[Y \mid A = 0, W]$, and $g(W) = \mathbb{E}[A \mid W]$ is the propensity score, where all expectations are taken with respect to the underlying distribution P. The symbol $\hat{}$ is used to indicate corresponding finite-sample estimators for population-level quantities.

To ensure that the parameter is identifiable, several key assumptions need to be met. The first assumption pertains to conditional independence, or unconfoundedness, indicating that when we consider the confounding variables, the potential outcomes are unrelated to the treatment assignments $(Y^0, Y^1 \perp$

$A \mid W$). The second assumption, known as positivity, requires that the assignment of treatment groups isn't deterministic given the confounders ($0 < Pr(A = 1 \mid W) < 1$). The third assumption, consistency, asserts that the observed outcomes match their respective potential outcomes ($Y^A = y$). Additionally, there are other assumptions, including the temporal sequence (i.e., covariates W are assessed before treatment), independence and identically distributed subjects.

Let the causal parameter of interest be ATE,

$$\beta_{ATE} = \mathbb{E}[Y^1 - Y^0] = \mathbb{E}[\mathbb{E}[Y^1 - Y^0 \mid W]] = \mathbb{E}[\mathbb{E}[Y \mid A = 1, W]] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, W]], \tag{3.1}$$

where Y^1 and Y^0 are the potential outcomes of the treatment and controls (Rubin, 1974).

A list of the first candidates to estimate ATE are

$$\begin{aligned}
\text{naive ATE} \quad & \hat{\beta}_{naiveATE} = \frac{1}{n_1}\sum_{i \in A_1} \hat{Q}_i^1 - \frac{1}{n_0}\sum_{i \in A_0} \hat{Q}_i^0, \\
\text{SR} \quad & \hat{\beta}_{SR} = \hat{\mathbb{E}}\Big[\hat{\mathbb{E}}[Y^1 - Y^0 \mid W]\Big] = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}_i^1 - \hat{Q}_i^0, \\
\text{IPW} \quad & \beta_{IPW} = \hat{\mathbb{E}}\Big[\frac{Y^1}{\hat{\mathbb{E}}[A \mid W]} - \frac{Y^0}{1 - \hat{\mathbb{E}}[A \mid W]}\Big] = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{A_i Y_i}{\hat{g}_i} - \frac{(1 - A_i)Y_i}{1 - \hat{g}_i}\Big), \\
\text{nIPW} \quad & \hat{\beta}_{nIPW} = \sum_{i=1}^{n}\Big(\frac{A_i w_i^{(1)} Y_i}{\sum_{j=1}^{n} A_j w_j^{(1)}} - \frac{(1 - A_i) w_i^{(0)} Y_i}{\sum_{j=1}^{n}(1 - A_j) w_j^{(0)}}\Big).
\end{aligned} \tag{3.2}$$

The naive average treatment effect (naive ATE) is a biased (due to the selection bias) estimator of ATE (Angrist and Pischke, 2008) and is the poorest estimator among all the candidates. SR is not an orthogonal estimator (Chernozhukov et al., 2018a) and if ML algorithms that do not belong to the Donsker class (Van der Vaart (2000), Section 19.2) or have the entropy that grows with the sample size is used, this estimator also becomes biased and is not asymptotically normal. IPW (Horvitz and Thompson, 1952) and its normalization versions adjust (or weight) the observations in the treatment and control groups. IPW and nIPW are also not orthogonal estimators and are similar to SR in this respect. In addition, both $\hat{\beta}_{SR}$ and $\hat{\beta}_{IPW}$ (and $\hat{\beta}_{nIPW}$) are single robust, that is, they are consistent estimators of ATE if the models used are $\sqrt{n}$-consistent (Lunceford and Davidian, 2004). IPW is an unbiased estimator of ATE if g is correctly specified, but nIPW is not unbiased, but is less sensitive to extreme predictions. The AIPW estimator (Scharfstein et al., 1999) is an improvement over SR, IPW and nIPW, which involves the predictions for both treatment (the propensity score), and the causal parameter can be expressed as:

$$\beta = \mathbb{E}\Bigg[\Big(\frac{AY - Q(1, W)(A - \mathbb{E}[A \mid W])}{\mathbb{E}[A \mid W]}\Big) - \Big(\frac{(1 - A)Y + Q(0, W)(A - \mathbb{E}[A \mid W])}{1 - \mathbb{E}[A \mid W]}\Big)\Bigg], \tag{3.3}$$

and the sample version estimator of (3.3) is

$$\hat{\beta}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\Big(\frac{A_iY_i - \hat{Q}(1,W_i)(A_i - \hat{\mathbb{E}}[A_i \mid W_i])}{\hat{\mathbb{E}}[A_i \mid W_i]}\Big) - \Big(\frac{(1-A_i)Y_i + \hat{Q}(0,W_i)(A_i - \hat{g}_i)}{1 - \hat{\mathbb{E}}[A_i \mid W_i]}\Big)\Bigg] = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{A_i(y_i - \hat{Q}_i^1)}{\hat{g}_i} - \frac{(1-A_i)(y_i - \hat{Q}_i^0)}{1-\hat{g}_i}\Big) + \hat{\beta}_{SR}, \tag{3.4}$$

where $\hat{Q}_i^k = \hat{Q}(k, W_i) = \hat{\mathbb{E}}[Y_i \mid A_i = k, W_i]$ and $\hat{g}_i = \hat{\mathbb{E}}[A_i \mid W_i]$.

Among all the doubly robust estimators of ATE, AIPW is the most efficient estimator if both of the propensity score or outcome models are correctly specified, but is not necessarily efficient under incorrect model specification. In fact, this nice feature of AIPW may be less relevant in real-life problems as we might not have a priori knowledge about the predictors of the propensity score and outcome and we cannot correctly model them. Further, in practice, perfect or near-perfect prediction of the treatment assignment can inflate the variance of the AIPW estimator (Van der Laan and Rose, 2011). As a remedy, similar to the normalization of the IPW estimator, we can define a normalized version of the AIPW estimator which is less sensitive to extreme values of the predicted propensity score, referred to as the nAIPW estimator:

$$\hat{\beta}_{nAIPW} = \sum_{i=1}^{n}\Big(\frac{A_i(y_i - \hat{Q}_i^1)w_i^{(1)}}{\sum_{j=1}^{n} A_j w_j^{(1)}} - \frac{(1-A_i)(y_i - \hat{Q}_i^0)w_i^{(0)}}{\sum_{j=1}^{n}(1-A_j)w_j^{(0)}}\Big) + \hat{\beta}_{SR}, \tag{3.5}$$

where $w_k^{(1)} = \frac{1}{\hat{g}_k}$ and $w_k^{(0)} = \frac{1}{1-\hat{g}_k}$. Both AIPW and nAIPW estimators add adjustment factors to the SR estimator which involves both models of the treatment and the outcome. The terms that include the normalized weights are referred to as bias-corrected terms.

Both AIPW and nAIPW are examples of a class of estimators where

$$\hat{\beta}_{GDR} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{A_i(y_i - \hat{Q}_i^1)}{\hat{h}_i^1} - \frac{(1-A_i)(y_i - \hat{Q}_i^0)}{\hat{h}_i^0}\Big) + \hat{\beta}_{SR}, \tag{3.6}$$

where we refer to this general class as the general doubly robust (GDR) estimator. Letting $\hat{h}^1 = \hat{g}$ and $\hat{h}^0 = 1 - \hat{g}$ gives the AIPW estimators and letting $\hat{h}^1 = \hat{g}\hat{\mathbb{E}}\frac{A}{\hat{g}}$ and $\hat{h}^0 = (1-\hat{g})\hat{\mathbb{E}}\frac{1-A}{1-\hat{g}}$ gives the nAIPW estimator.

The GDR estimator can also be written as

$$\hat{\beta}_{GDR} = \hat{\mathbb{E}}\Big[\Big(\frac{A}{\hat{h}^1} - \frac{1-A}{\hat{h}^0}\Big)y - (A - \hat{h}^1)\hat{Q}^1 + (1 - A - \hat{h}^0)\hat{Q}^0\Big], \tag{3.7}$$

If h^1 and h^0 are chosen so that

$$\mathbb{E}[A - h^1] = 0,\ \mathbb{E}[1 - A - h^0] = 0, \tag{3.8}$$

by the total law of expectation $\hat{\beta}_{GDR}$ is an unbiased estimator of β.

3.3 Outcome and Treatment Predictions

The causal estimation and inference when utilizing the AIPW and nAIPW is carried out in two steps. In step 1, the treatment and outcome are predicted by a statistical or ML algorithm, and in the second step the predictions are inserted into the estimator. The ML algorithms in step 1 can capture the linear and non-linear relationships between the confounders and the treatment and the outcome.

NNs (Cheng et al., 2018, Goodfellow et al., 2016, Xu et al., 2015) are a class of non-linear and non-parametric complex algorithms that can be employed to model the relationship between any set of inputs and some outcome. There has been a tendency to use NNs as they have achieved great success in the most complex AI tasks such as computer vision and natural language understanding (Goodfellow et al., 2016).

Farrell et al. (2021) used two independent NNs for modeling the propensity score (PS) model and the outcome with the ReLU activation function (Goodfellow et al., 2016), here referred to as the double NN or dNN:

$$\begin{aligned} \mathbb{E}[Y \mid A, W] &= \beta_0 + \beta A + W\alpha + \mathbf{H}\Gamma_Y \\ \mathbb{E}[A \mid W] &= G(\beta_0' + W'\alpha' + \mathbf{K}\Gamma_A), \end{aligned} \tag{3.9}$$

where two separate neural nets model y and A (no parameter sharing), G is the logistic link function, α and α' are the skip connections parameters, and $\mathbf{H}$ and $\mathbf{K}$ are composite of non-linear functions in the outcome and PS architectures, respectively, and Γ_Y and Γ_A are the final matrices that connect the last hidden layer to the output in the outcome and PS architectures, respectively. In this paper, the dNN algorithm refers to two neural networks to model the treatment and outcome separately.

Farrell et al. (2021) proved that dNN algorithms almost attain $n^{\frac{1}{4}}$-rates. By employing the cross-fitting method and theory developed by Chernozhukov et al. (2018a), an orthogonal causal estimator is asymptotically normal, under some regularity and smoothing conditions, if the dNN is used in the first step (see Theorem 1 in Farrell et al. (2021)).

These results assume no regularization techniques imposed on the NNs' weights, and only SGD is used. Farrell et al. claim that the fact that SGD controls the complexity of the NN algorithm to some extent (Goodfellow et al., 2016, Zhang et al., 2021) is sufficient for the first step. Our initial simulations, however, contradict this claim and we hypothesize that for causal parameter estimation, a dNN with no regularization leads to high variance for the causal estimator used in the second step. Our initial experiments indicate that L_2 regularization and dropout do not perform well in terms of MSE of AIPW. The loss functions we use contain L_1 regularization (in addition to SGD during the optimization):

$$\begin{aligned} L_y(\mathcal{P}_y, \beta, \alpha) =& \sum_{i=1}^{n} \Big[y_i - \beta_0 - \beta A_i - W_i\alpha - H_i^T\Gamma_Y\Big]^2 + C_{L_1} \sum_{\omega \in \mathcal{P}} |\omega|, \\ L_A(\mathcal{P}_A, \alpha') =& \sum_{i=1}^{n} \Big[A_i \log\Big(G(K_i^T\Gamma_A)\Big) + (1 - A_i)\log\Big(1 - G(K_i^T\Gamma_A)\Big)\Big] + \\ & C_{L_1}' \sum_{\omega \in \mathcal{P}} |\omega|, \end{aligned} \tag{3.10}$$

where C_{L_1}, C_{L_1}' are the regularization rates. The dNN can have an arbitrary number of hidden layers, or the width of the network $(\mathcal{HL})$ can be another hyperparameter. For a l_h-layer network, $\mathcal{HL} =$

$[l_1, l_2, ..., l_h]$, where l_j is the number of neurons in layer j, $j = 1, 2, ..., h$. $\mathcal{P}_y, \mathcal{P}_A$ are the connection parameters in the non-linear part of the networks, with Ωs being shared for the two outcome and propensity models. Note that the gradient descent-type optimizations in the deep learning platforms (such as pytorch in our case) do not cause the NN parameters to shrink to zero. All hyperparameters can be set before training or be determined by cross-validation.

3.4 GDR Estimator Properties

In this section we will see that nAIPW (3.5) is doubly robust, that is, if either of the outcome or propensity score models are $\sqrt{n}$-consistent, nAIPW will be consistent. Further, nAIPW is orthogonal (Chernozhukov et al., 2018a) and is asymptotically linear under certain assumptions and we calculate its asymptotic variance.

3.4.1 Consistency and Asymptotic Distribution of nAIPW

In causal inference, estimating the causal parameter and drawing inference on the parameter are two major tasks. Employing a machine learning algorithm to estimate Q and g in (3.4) is a means to estimate and draw inference on the causal parameter; the ultimate goal is the relationship between the treatment and the outcome. This allows people to use blackbox ML models with no explanation how these models have learned from the explanatory features. The question is if the consistency and asymptotic normality of the second step causal estimator are preserved if complex ML algorithms are utilized twice for the treatment and outcome models, each with a convergence rate smaller than $\sqrt{n}$, and entropy that grows with n.

Chernozhukov et al. (2018a) provide numerical experiments illustrating that some estimators are not consistent or asymptotically normal if complex ML models are used that do not belong to the Donsker class and have entropy that grows with n. They further provide a solution by introducing "orthogonal" estimators that, under some regulatory conditions and cross-fitting, are asymptotically normal even if complex ML models can be used as long as their rates of convergence are as small as $n^{\frac{1}{4}}$.

The next two subsections provide an overview of the general theory and prove that nAIPW is asymptotically normal.

3.4.2 The Efficient Influence Function

Hahn (1998) derives the efficient influence function (EIF) of $\beta = \beta_1 - \beta_0$ as

$$\phi(O, P) = \Big(\frac{A}{g}(Y - Q^1) + Q^1 - \beta_1\Big) - \Big(\frac{1-A}{1-g}(Y - Q^0) + Q^0 - \beta_0\Big) \tag{3.11}$$

To study the asymptotic behaviour of nAIPW, we write the scaled difference

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i, P) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i, \hat{P}_n) + \\ \sqrt{n}(P_n - P)[\phi(O_i, \hat{P}_n) - \phi(O_i, P)] - \sqrt{n}R(P, \hat{P}_n), \tag{3.12}$$

where the first term is a normal distribution by the central limit theorem, and the third and fourth terms are controlled if the class of functions are Donsker and standard smoothing conditions are satisfied ((Chernozhukov et al., 2018a, Van der Vaart, 2000), Theorem 19.26). If the nuisance parameters are not Donsker, data splitting and cross-fitting guarantees plus the regulatory conditions are needed to control these two terms (Chernozhukov et al., 2018a, Farrell et al., 2021). It is unclear, however, how the second term behaves, i.e.,

$$-\frac{1}{\sqrt{n}}\phi(O,\hat{P}_n) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\frac{A_i}{g_i}(Y_i-\hat{Q}_i^1)-\frac{1-A_i}{1-g_i}(Y_i-\hat{Q}_i^0)+\hat{Q}_i^1-\hat{Q}_i^0\Big]-\hat{\beta}, \tag{3.13}$$

where $\hat{\beta}=\beta(\hat{P}_n)$, as it contains data-adaptive nuisance parameter estimations. There are different tricks to get rid of this term. One method is the one-step method in which we move this term to the left to create a new estimator which is exactly the same as the AIPW estimator with known propensity scores:

$$\sqrt{n}(\hat{\beta}+\frac{1}{n}\phi(O,\hat{P}_n)-\beta)=\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{A_i}{g_i}(Y_i-\hat{Q}_i^1)-\frac{1-A_i}{1-g_i}(Y_i-\hat{Q}_i^0)+\hat{Q}_i^1-\hat{Q}_i^0\Big]-\beta\Big). \tag{3.14}$$

Another trick is to let this term go to zero which results in estimating equations whose solution is exactly the same as the one-step estimator. The targeted learning strategy is to manipulate the data generating process which results in a different estimator (Hines et al., 2022, Van der Laan and Rose, 2011) (which we do not study here).

The requirement in the above estimator is that the propensity score is known, which is unrealistic. In reality, this quantity should be estimated using the data. However, replacing g with a data-adaptive estimator changes the remainder term in (2.21) that needs certain assumptions to achieve asymptotic properties such as consistency. We replace g and $1-g$ in (2.22) by $\hat{h}^1$ and $\hat{h}^0$, respectively, which provides a more general view of the above one-step estimator.

3.4.3 Double Robustness and Rate Double Robustness Properties of GDR

One of the appealing properties of AIPW is its double robustness property which partially relaxes the restrictions of IPW and SR which require the consistency of the treatment and outcome models, respectively. This property is helpful when the first-step algorithms are $\sqrt{n}$-consistent. The following theorem states that the nAIPW estimator (3.5) actually possesses the double robustness property.

Theorem 3.4.1 (nAIPW Double Robustness)**.** *The DR estimator* (3.5) *is consistent if* $\hat{Q}^k \xrightarrow{p} Q^k$, $k=0,1$ *or* $\hat{g}\xrightarrow{p} g$.

The proof is left to the appendix. Theorem 3.4.1 is useful when we *a priori* knowledge about the propensity scores (such as in the experimental studies) or we estimate the propensity scores with $\sqrt{n}$-rate converging algorithms. In practice, however, the correct specification is infeasible in the observational data, but $\sqrt{n}$-rate algorithms such as parametric models, generalized additive models (GAMs) or the models that assume sparsity might be used (Farrell, 2015). This is restrictive and these model assumptions might not hold in practice which is why non-parametric ML algorithms such as NNs are used. As mentioned before, the NN architecture we utilize here does not offer a $\sqrt{n}$-consistent prediction model in the first step of the estimation (Farrell et al., 2021). This reduces the usefulness of the double robustness property of the GDR estimator when using complex ML algorithms. A more useful property when using complex ML algorithms is the *rate double robustness (RDR)* property (Smucler et al., 2019). RDR does not require either of the prediction models to be $\sqrt{n}$-consistent; it suffices that they are consistent at any rate but together become $\sqrt{n}$-consistent; that is, if the propensity score and outcome model are consistent at n^{r_A} and n^{r_Y}, respectively ($r_Y, r_A > 0$), we must have $r_A + r_Y = \frac{1}{2}$. To see that the DR has this property (as does DR (Farrell, 2015)), note that the remainder (2.21) can be written as

$$-\sqrt{n}R(P, \hat{P}_n) = \sqrt{n}\mathbb{E}\Big[(\frac{g}{\hat{h}^1} - 1)(Q^1 - \hat{Q}^1)\Big] + \sqrt{n}\mathbb{E}\Big[(\frac{1-g}{\hat{h}^0} - 1)(Q^0 - \hat{Q}^0)\Big], \tag{3.15}$$

which, by the Hölder inequality, is upper bounded:

$$-\sqrt{n}R(P, \hat{P}_n) \leq \Bigg[\mathbb{E}\Big[\frac{g}{\hat{h}^1} - 1\Big]^2\Bigg]^{\frac{1}{2}} \Bigg[\mathbb{E}\Big[Q^1 - \hat{Q}^1\Big]^2\Bigg]^{\frac{1}{2}} + \Bigg[\mathbb{E}\Big[\frac{1-g}{\hat{h}^0} - 1\Big]^2\Bigg]^{\frac{1}{2}} \Bigg[\mathbb{E}\Big[Q^0 - \hat{Q}^0\Big]^2\Bigg]^{\frac{1}{2}} \tag{3.16}$$

Making the standard assumptions that

$$\begin{aligned} &\Bigg[\mathbb{E}\Big[g - \hat{h}^k\Big]^2\Bigg]^{\frac{1}{2}} \Bigg[\mathbb{E}\Big[Q^k - \hat{Q}^k\Big]^2\Bigg]^{\frac{1}{2}} = o(n^{-\frac{1}{2}}), \quad k = 0, 1, \\ &\mathbb{E}\Big[g - \hat{h}^k\Big]^2 = o(1), \quad \mathbb{E}\Big[Q^k - \hat{Q}^k\Big]^2 = o(1), \quad k = 0, 1, \\ &\text{Empirical Positivity} \quad c_1 < \hat{h}^k < 1 - c_2, \text{ for some } c_1, c_2 > 0, \end{aligned} \tag{3.17}$$

implies

$$-\sqrt{n}R(P, \hat{P}_n) = o(n^{-\frac{1}{2}}), \tag{3.18}$$

that is, the GDR has the rate double robustness property.

The assumptions in (3.17) are less restrictive than needing at least one of the prediction models to be $\sqrt{n}$-consistent for the double robust property (Farrell, 2015, Hines et al., 2022). This means that the outcome and propensity score models can be at least as fast as $o(n^{-\frac{1}{4}})$ (which is an attainable generalization bound for many complex machine learning algorithms (Chernozhukov et al., 2018a)), and the GDR estimator is still consistent. Farrell et al. (2021) proves that two neural networks without regularization (except the one imposed by the stochastic gradient descent optimization) satisfy such bounds and can provide a convenient first-step prediction algorithm (when they utilize the AIPW estimator and the cross-fitting strategy proposed by Chernozhukov et al. (2018a)).

In order for a special case of GDR estimator to outperform the AIPW estimator, we must have $Ah^1 \geq Ag$ and $(1 - A)h^0 \geq (1 - A)(1 - g)$, in addition to conditions in (3.17). Note that these two conditions are satisfied for nAIPW; replacing h^1 and h^0 with $\hat{g}\hat{\mathbb{E}}\frac{A}{\hat{g}}$ and $(1 - \hat{g})\hat{\mathbb{E}}\frac{1-A}{1-\hat{g}}$ can help stabilize the bias and variance magnitude and help shrink the remainder (3.15) to zero. The scenario analysis performed in Section 3.4.4 provides an insight about the reduction in the sensitivity to the violation of the empirical positivity assumption.

3.4.4 Robustness of nAIPW Against Extreme Propensity Scores

There are two scenarios in which the empirical positivity is violated, where the probability of receiving the treatment for the people who are treated is 1, that is, $A_k = 1$ and $P(A_k = 1 \mid W) = 1$ (or vice versa for the untreated group $A_k = 0$ and $P(A_k = 0 \mid W) = 0$), and where there are a handful of treated subjects whose probability of receiving the treatment is 0, that is, $A_k = 1$ and $P(A_k = 1 \mid W) = 0$ (and vice versa for the untreated group, that is, $A_k = 0$ and $P(A_k = 0 \mid W) = 1$). Although the identifiability assumptions guarantee that such scenarios do not occur, in practice, extremely small or large probabilities similar to the second scenario above, that is, where there exists a treated individual who has a near-zero probability of receiving the treatment, can impact the performance of the estimators that involve propensity score weighting. For example, replacing h^1 with $\hat{g}$ and h^0 with $1 - \hat{g}$ in practice can increase both the bias and variance of AIPW (Van der Laan and Rose, 2011). This can be seen by viewing the bias and variance of these weighting terms. As noted before, the AIPW and nAIPW add adjustments to the single robust estimator $\mathbb{E}[Q^1 - Q^0]$. The adjustments involve weightings $\frac{A}{g}$ or $\frac{A}{g\mathbb{E}\frac{A}{g}}$ to the residuals of Y and Q^k, $k = 0, 1$. Under a correct specification of the propensity score g, these weights have the same expectations. The difference is in their variances:

$$\begin{aligned} Var\Big(\frac{A}{g}\Big) &= \frac{1}{g} - 1, \\ Var\Big(\frac{A}{g\mathbb{E}\frac{A}{g}}\Big) &= \frac{1}{\mathbb{E}^2\frac{A}{g}}(\frac{1}{g} - 1), \end{aligned} \tag{3.19}$$

under the correct specification of the propensity score g. By letting g tend to zero in violation of the empirical positivity assumption, it can be seen that the nAIPW is less volatile than the AIPW estimator. That is, the weights in AIPW might have a larger variance than those in nAIPW.

3.4.5 Scenario Analysis

A scenario analysis is performed to see how nAIPW stabilizes the estimator: Assume that the empirical positivity is violated, that is, there is at least an observation k where $A_k = 1$ where $\hat{g}_k$ is extremely close to zero, such as $\hat{g}_k = 10^{-s}$ for $s \gg 0$. AIPW will blow up in this case:

$$\begin{aligned} \beta_{1,AIPW} &= \frac{1}{n}\Big(10^s(Y_k^1 - Q_k^1) + \sum_{i \in I^1_{-k}} \frac{Y_i^1 - Q_i^1}{g_i}\Big) + \frac{1}{n}\sum_{i=1}^{n} Q_i^1, \\ \beta_{0,AIPW} &= \frac{1}{n}\Big(\sum_{i \in I^0} \frac{Y_i^0 - Q_i^0}{1 - g_i}\Big) + \frac{1}{n}\sum_{i=1}^{n} Q_i^0, \end{aligned} \tag{3.20}$$

where $I^a = \{j : \; A_j = a\}$, $I^a_{-k} = \{j : \; A_j = a, j \neq k\}$, and superscripts $a = 1$ and $a = 0$ refer to the estimators of the first and the second components in ATE (3.1). However, nAIPW is robust against this empirical positivity violation:

$$\beta_{1,nAIPW} = \Big(\frac{Y_k^1 - Q_k^1}{10^{-s}(10^s + \sum_{j \neq k} \frac{A_j}{g_j})} + \sum_{i \in I^1_{-k}} \frac{Y_i^1 - Q_i^1}{g_i(10^s + \sum_{j \neq k} \frac{A_j}{g_j})}\Big) + \frac{1}{n}\sum_{i=1}^{n} Q_i^1, \tag{3.21}$$

and

$$\beta_{0,nAIPW} = \Big(\frac{0 \times (Y_k^1 - Q_k^0)}{\star} + \sum_{i \in I^0_{-k}} \frac{Y_i^0 - Q_i^0}{(1 - g_i)(\sum_{j=1}^{n} \frac{1 - A_j}{1 - g_j})}\Big) + \frac{1}{n}\sum_{i=1}^{n} Q_i^0, \tag{3.22}$$

where $\star$ is some non-zero term. Thus

$$\beta_{1,nAIPW} \approx \Big(\frac{Y_k^1 - Q_k^1}{1 + 10^{-s}(n-1)} + \sum_{i \in I_{-k}^1} \frac{Y_i^1 - Q_i^1}{g_i 10^s + g_i(n-1)}\Big) + \frac{1}{n}\sum_{i=1}^n Q_i^1, \tag{3.23}$$

The factor 10^s in (3.20) can blow up the AIPW if $10^s \gg n$ (and the outcome estimation is not close enough to the observer outcome), but this factor does not appear in the numerator of the nAIPW estimator. For such large factors, (3.23) can be simplified to

$$\beta_{1,nAIPW} \approx Y_k^1 - Q_k^1 + \frac{1}{n}\sum_{i=1}^n Q_i^1. \tag{3.24}$$

Thus, the extreme probability does not make $\beta_{1,nAIPW}$ blow up, but the adjustment to the $\beta_{1,SR}$ that accounts for confounding effects. The second factor $\beta_{0,nAIPW}$ is not impacted in this scenario.

Considering a scenario that there is another treated individual with extremely small probability, such as $\hat{g}_l = 10^{-t}$, such that, without loss of generality, $t > s \gg 0$, we will have:

$$\beta_{1,nAIPW} \approx \frac{Y_k^1 - Q_k^1}{1 + 10^{t-s} + 10^{-s}(n-2)} + \frac{Y_l^1 - Q_l^1}{1 + 10^{s-t} + 10^{-t}(n-2)} + \frac{1}{n}\sum_{i=1}^n Q_i^1. \tag{3.25}$$

Depending on the values s and t, one of the first two terms in (3.25) might vanish, but the estimator does not blow up. There is at most only a handful of treated individuals with extremely small probabilities and, based on the above observation, the nAIPW estimator does not blow up. That said, nAIPW might not sufficiently correct the β_{SR} for the confounding effects, although confounders have been taken into account in the calculation of β_{SR} to some extent.

The same observation can be made in the asymptotic variance of these estimators. This shows how extremely small probabilities for treated individuals (or extremely large probabilities for untreated individuals) can result in a biased and unstable estimator, while neither of the bias or variance of nAIPW suffer as much. Although not performed, the same observation can be made for the untreated individuals with extremely large probabilities.

The above scenario analysis indicates the bias and variance of nAIPW might go up in cases of the violation of empirical positivity, but it still is less biased and more stable than AIPW. The remainder term (3.15) is also more likely to be $o(n^{-\frac{1}{2}})$ in nAIPW versus AIPW as it contains k's where $A_k = 1$, $g_k \mathbb{E}_n \frac{A_k}{g_k} \geq g_k$.

3.5 Asymptotic Sampling Distribution of nAIPW

Replacing g in the denominator of the von Mises expansion (2.21) with the normalizing terms is enough to achieve the asymptotic distribution of the nAIPW and its asymptotic standard error. However, we can see that nAIPW is also the solution to (extended) estimating equations. The solution to the estimating equations is important as van der Vaart (Chapters 19 and 25) proves that under certain regulatory conditions, if the prediction models belong to the Donsker class, the solutions to Z-estimators are consistent and asymptotically normal (Van der Vaart (2000), Theorem 19.26). Thus, nAIPW that is the solution to a Z-estimator (also referred to an M-estimator) will inherit the consistency and asymptotic normality, assuming certain regulatory conditions and that the first-step prediction models belong to the Donsker class:

$$\begin{aligned}\mathbb{E}\Bigg[\frac{A(Y^1-Q^1)}{\gamma g}-\frac{(1-A)(Y^0-Q^0)}{\lambda(1-g)}+(Q^1-Q^0-\beta)\Bigg]&=0,\\ \mathbb{E}\Big[\frac{A}{g}-\gamma\Big]&=0,\\ \mathbb{E}\Big[\frac{1-A}{1-g}-\lambda\Big]&=0.\end{aligned}\tag{3.26}$$

The Donsker class assumption prevents too complex algorithms in the first step, algorithms such as tree-based models, NNs, cross-hybrid algorithms or their aggregations (Friedman et al., 2001, Hines et al., 2022). The Donsker class assumption can be relaxed if sample splitting (or cross-fitting) is utilized and the target parameter is orthogonal (Chernozhukov et al., 2018a). In the next section we see that nAIPW is orthogonal and, thus, theoretically, we can relax the Donsker class assumption under certain smoothing regulatory conditions. Before seeing the orthogonality property of nAIPW, let us review the smoothing regularity conditions necessary for asymptotic normality.

3.5.1 Regulatory assumptions

Let β be the causal parameter, $\eta \in T$ be the infinite dimensional nuisance parameters where T is a convex set with a norm. Additionally, let the score function $\phi : \mathbb{O} \times \mathcal{B} \times T \to \mathbb{R}$ be a measurable function, $\mathbb{O}$ be the measurable space of all random variables O with probability distribution $P \in \mathcal{P}_n$ and $\mathcal{B}$ be an open subset of $\mathbb{R}$ containing the true causal parameter. Let the sample $O = (O_1, O_2, ..., O_n)$ be observed and the set of probability measures $\mathcal{P}_n$ expand with sample size n. In addition, let $\beta \in \mathcal{B}$ be the solution to the estimating equation $\mathbb{E}\phi(\mathbb{O}, \beta, \eta) = 0$. The assumptions that guarantee that the second-step orthogonal estimator $\hat{\beta}$ is asymptotically normal are (Chernozhukov et al., 2018a): (1) β does not fall on the boundary of $\mathcal{B}$; (2) the map $(\beta, \eta) \to \mathbb{E}_P\phi(O, \beta, \eta)$ is twice Gateaux differentiable (this holds by the positivity assumption). β is identifiable; (3) $\mathbb{E}_P\phi(O, \beta, \eta)$ is smooth enough; (4) $\hat{\eta} \in \mathcal{T}$ with high probability and $\eta \in \mathcal{T}$. $\hat{\eta}$ converges to η_0 at least as fast as $n^{-\frac{1}{4}}$ (similar but slightly stronger than first two assumptions in (3.17)); (5) score function(s) $\phi(., \beta, \eta)$ has finite second moment for all $\beta \in \mathcal{B}$ and all nuisance parameters $\eta \in \mathcal{T}$; (6) the score function(s) $\phi(., \beta, \eta)$ is measurable; (7) the number of folds increases by sample size.

3.5.2 Orthogonality and the Regulatory Conditions

The orthogonality condition (Chernozhukov et al., 2018a) is a property related to the estimating equations

$$\mathbb{E}\phi(O, \beta, \eta) = 0. \tag{3.27}$$

We refer to an estimator drawn from the estimating Equation (3.27) as an orthogonal estimator.

Let $\eta \in T$, where T is a convex set with a norm. Additionally, let the score functions $\phi : \mathbb{O} \times \mathcal{B} \times T \to \mathbb{R}$ be a measurable function, $\mathbb{O}$ is measurable space of all random variables O with probability distribution $P \in \mathcal{P}_n$ and $\mathcal{B}$ is an open subset of $\mathbb{R}$ containing the true causal parameter. Let the sample $O = (O_1, O_2, ..., O_n)$ be observed and the set of probability measures $\mathcal{P}_n$ can expand with sample size n. The score function ϕ follows the Neyman orthogonality condition with respect to $\mathcal{T} \subseteq T$, if the Gateaux derivative operator exists for all $\epsilon \in [0,1)$:

$$\partial_{\tilde{\eta}}\mathbb{E}_P\phi(O,\beta_0,\tilde{\eta})\Big|_{\tilde{\eta}=\eta}[\tilde{\eta}-\eta] := \partial_\epsilon\mathbb{E}_P\phi(O,\beta_0,\eta+\epsilon(\tilde{\eta}-\eta))\Big|_{\epsilon=0} = 0. \tag{3.28}$$

Chernozhukov et al. (2018a) presents a few examples of orthogonal estimating equations including the AIPW estimator (2.2). Utilizing cross-fitting, under standard regulatory conditions (Section 7.1.1), the asymptotic normality of estimators with orthogonal estimating equations is guaranteed even if the nuisance parameters are estimated by ML algorithms not belonging to the Donsker class and without finite entropy conditions (Chernozhukov et al., 2018a).

By replacing λ and γ in the first line of (3.26) with their solutions in the second and third equations:

$$\mathbb{E}_P\phi(O,\beta,Q^1,Q^0,g) = \mathbb{E}\left[\frac{A(Y^1-Q^1)}{g\mathbb{E}\frac{A}{g}} - \frac{(1-A)(Y^0-Q^0)}{(1-g)\mathbb{E}\frac{1-A}{1-g}} + (Q^1-Q^0-\beta)\right] = 0, \tag{3.29}$$

Implementing the orthogonality condition (3.28), it can be verified that nAIPW (3.5) is also an example of an orthogonal estimator. To see this, we apply the definition of orthogonality (Chernozhukov et al., 2018a)

$$\begin{aligned}
&\partial_\eta\mathbb{E}_P\phi(O,\beta,\eta)\Big|_{\eta=\eta_0}[\eta-\eta_0] = \\
&\qquad \partial_\eta\mathbb{E}_P\Big(Q^1 + \frac{A(Y^1-Q^1)}{g\mathbb{E}\frac{A}{g}} - Q^0 - \frac{(1-A)(Y^0-Q^0)}{(1-g)\mathbb{E}\frac{1-A}{1-g}} - \beta\Big)|_{\eta=\eta_0}[\eta-\eta_0] \\
&\qquad \propto \partial_\epsilon\mathbb{E}_P\Big(Q^1_\epsilon + \frac{A(Y^1-Q^1_\epsilon)}{g_\epsilon\mathbb{E}\frac{A}{g_\epsilon}} - Q^0_\epsilon - \frac{(1-A)(Y^0-Q^0_\epsilon)}{(1-g_\epsilon)\mathbb{E}\frac{1-A}{1-g_\epsilon}} - \beta\Big)|_{\epsilon=0} = \\
&\qquad \mathbb{E}\Big((\tilde{Q}^1-Q^1) + \frac{A}{g\mathbb{E}\frac{A}{g}}(-(\tilde{Q}^1-Q^1)) + A(Y-Q^1)a(g,\tilde{g}-g)\Big) - \\
&\qquad \mathbb{E}\Big((\tilde{Q}^0-Q^0) + \frac{1-A}{(1-g)\mathbb{E}\frac{1-A}{1-g}}(-(\tilde{Q}^0-Q^0)) + \\
&\qquad (1-A)(Y-Q^0)b(g,\tilde{g}-g)\Big) = 0,
\end{aligned} \tag{3.30}$$

where $Q^k_\epsilon = \epsilon\tilde{Q}^k + (1-\epsilon)Q^k$, $k=0,1$, and $g_\epsilon = \epsilon\tilde{g} + (1-\epsilon)g$, and for some functions a and b. The last equality is because $\mathbb{E}A(Y-Q^1)=0$, $\mathbb{E}(1-A)(Y-Q^0)=0$, $\mathbb{E}\frac{A}{g\mathbb{E}\frac{A}{g}} = 1$ and $\mathbb{E}\frac{1-A}{(1-g)\mathbb{E}\frac{1-A}{1-g}} = 1$, under correct specification of the propensity score g.

Thus, nAIPW is orthogonal, and by utilizing cross-fitting for the estimation, nAIPW is consistent and asymptotically normal, under certain regulatory conditions.

3.5.3 Asymptotic Variance of nAIPW

To evaluate the asymptotic variance of nAIPW, we employ the M-estimation theory (Stefanski and Boos, 2002, Van der Vaart, 2000). For causal inference for M-estimators, the bootstrap for the estimation of causal estimator variance is not generally valid even if the nuisance parameter estimators are $\sqrt{n}$-convergent. However, sub-sampling m out of n observations (Politis and Romano, 1994) can be shown to be universally valid, provided $m \to \infty$ and $\frac{m}{n} \to 0$. In practice, however, we can face computational issues since nuisance parameters must be separately estimated (possibly with ML models) for each subsample/bootstrap sample.

The variance estimator of AIPW (2.2) is (Lunceford and Davidian, 2004)

$$\hat{\sigma}^2_{AIPW} = \frac{1}{n^2}\sum_{i=1}^{n}\Big(\frac{A_iY_i - \hat{Q}^1_i(A_i-\hat{g}_i)}{\hat{g}_i} - \frac{(1-A_i)Y_i + \hat{Q}^0_i(A_i-\hat{g}_i)}{1-\hat{g}_i} - \hat{\beta}_{AIPW}\Big)^2 =$$

$$\frac{1}{n^2}\sum_{i=1}^{n}\Big(\frac{A_i(y_i-\hat{Q}^1_i)}{\hat{g}_i} - \frac{(1-A_i)(y_i-\hat{Q}^0_i)}{1-\hat{g}_i} + \hat{\beta}_{SR} - \hat{\beta}_{AIPW}\Big)^2. \quad (3.31)$$

The theorem below states that the variance estimator of AIPW (3.31) can intuitively extend to calculate the variance estimator of nAIPW (3.5) by moving the denominator n^2 to the square term in the summation and replacing it with $\hat{g}\hat{\mathbb{E}}(\frac{A}{\hat{g}})$ or $(1-\hat{g})\hat{\mathbb{E}}(\frac{1-A}{1-\hat{g}})$ in the terms containing g and $1-g$ in the denominator, respectively.

Theorem 3.5.1. *The asymptotic variance of the nAIPW* (3.5) *is*

$$\hat{\sigma}^2_{nAIPW} = \sum_{i=1}^{n}\Big(\frac{A_i(y_i-\hat{Q}^1_i)w^{(1)}_i}{\sum_{j=1}^{n}A_jw^{(1)}_j} - \frac{(1-A_i)(y_i-\hat{Q}^0_i)w^{(0)}_i}{\sum_{j=1}^{n}(1-A_j)w^{(0)}_j} + \frac{1}{n}(\hat{\beta}_{SR} - \hat{\beta}_{nAIPW})\Big)^2, \quad (3.32)$$

where $\hat{Q}^k_i = \hat{Q}(k, W_i)$ *and* $\hat{g}_i = \hat{\mathbb{E}}[A_i \mid W_i]$.

The proof utilizing the estimating equation technique is straightforward and is left to Appendix 7.2. The same result can be seen when deriving the estimator in the one-step method (see (2.21) and (2.22)). Since nAIPW is orthogonal, $\hat{\sigma}^2_{nAIPW}$ is consistent by applying the theories of (Chernozhukov et al., 2018a, Farrell et al., 2021), if the assumptions are met, cross-fitting is used, and the step 1 ML algorithms have the required convergence rates.

3.6 Monte Carlo Experiments

We conducted a Monte Carlo simulation study involving 100 iterations to compare the AIPW and nAIPW estimators. In this study, we utilized the dNN method for the first-step prediction models. Two different data size scenarios were considered, $n = 750$ and $n = 7500$, with the number of covariates set to $p = 32$ and $p = 300$ respectively. The covariates consisted of four categories: confounders (X_c), instrumental variables (X_{iv}), outcome predictors (X_y), and noise or irrelevant covariates (X_{irr}). In each scenario, these categories had sizes of $\#X_c = \#X_{iv} = \#X_y = \#X_{irr} = 8$ and 75. These matrices were independently generated from a multivariate normal (MVN) distribution with $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma_{kj} = \rho^{j-k}$ and $\rho = 0.5$. The models to generate the treatment assignment and outcome were specified as

$$\begin{aligned} &A \sim Ber(\frac{1}{1+e^{-pr}}), \text{with } pr = f_a(X_c)\gamma_c + g_a(X_{iv})\gamma_{iv}, \\ &y = 3 + A + f_y(X_c)\gamma'_c + g_y(X_y)\gamma_y + \epsilon, \end{aligned} \quad (3.33)$$

and $\beta = 1$. The functions f_a, g_a, f_y, g_y select 20% of the columns and apply interactions and non-linear functions listed below (3.35). The strength of the instrumental variable and confounding effects were chosen as $\gamma_c, \gamma'_c, \gamma_y \sim Unif(r_1, r_2)$ where $(r_1 = r_2 = 0.25)$, and $\gamma_{iv} \sim Unif(r_3, r_4)$ where $(r_3 = r_4 = 0.25)$. In order to avoid imbalanced treatment groups and cause systematic positivity violations, the generated datasets in which the number of subjects in the treatment or control group is less than 25% were ignored and new ones were generated.

The non-linearities are randomly selected from among the following functions:

$$\begin{aligned} l(x_1, x_2) &= e^{\frac{x_1 x_2}{2}} \\ l(x_1, x_2) &= \frac{x_1}{1+e^{x_2}} \\ l(x_1, x_2) &= (\frac{x_1 x_2}{10} + 2)^3 \\ l(x_1, x_2) &= (x_1 + x_2 + 3)^2 \\ l(x_1, x_2) &= g(x_1) \times h(x_2) \end{aligned} \tag{3.34}$$

where $g(x) = -2I(x \leq -1) - I(-1 \leq x \leq 0) + I(0 \leq x \leq 2) + 3I(x \geq 2)$, and $h(x) = -5I(x \leq 0) - 2I(0 \leq x \leq 1) + 3I(x \geq 1)$, or $g(x) = I(x \geq 0)$ and $h(x) = I(x \geq 1)$.

The networks' activation function is ReLU, with 3 hidden layers as large as the input size (p), with L_1 regularization and batch size equal to $3 * p$ and 200 epochs. The adaptive moment estimation (Adam) optimizer (Kingma and Ba, 2014) with a learning rate 0.01 and momentum 0.95 was used to estimate the network's parameters, including the causal parameter, ATE. The total number of first-step trained models is four: two NN depth variations ($[q, p, q]$ and $[p, p, p]$, where $q = p/4$), and two L_1 regularization strengths $(0.01, 0.1)$.

Simulation Results

The oracle estimates correctly specify the role of confounders, IV, and y-predictors and their linear or non-linear relationship with the outcome and treatment. Oracle model is never known in practice and is just calculated in the simulations to compare real-life situations with the truth. In almost all the scenarios we cannot obtain perfect causal effect estimation and inference.

Figure 3.1 shows the distribution of AIPW and nAIPW for different hyperparameter settings of NNs. The nAIPW estimator outperforms AIPW in almost all scenarios. As the AIPW gives extreme values in some simulation iterations, the log of the estimation is taken in Figure 3.1.

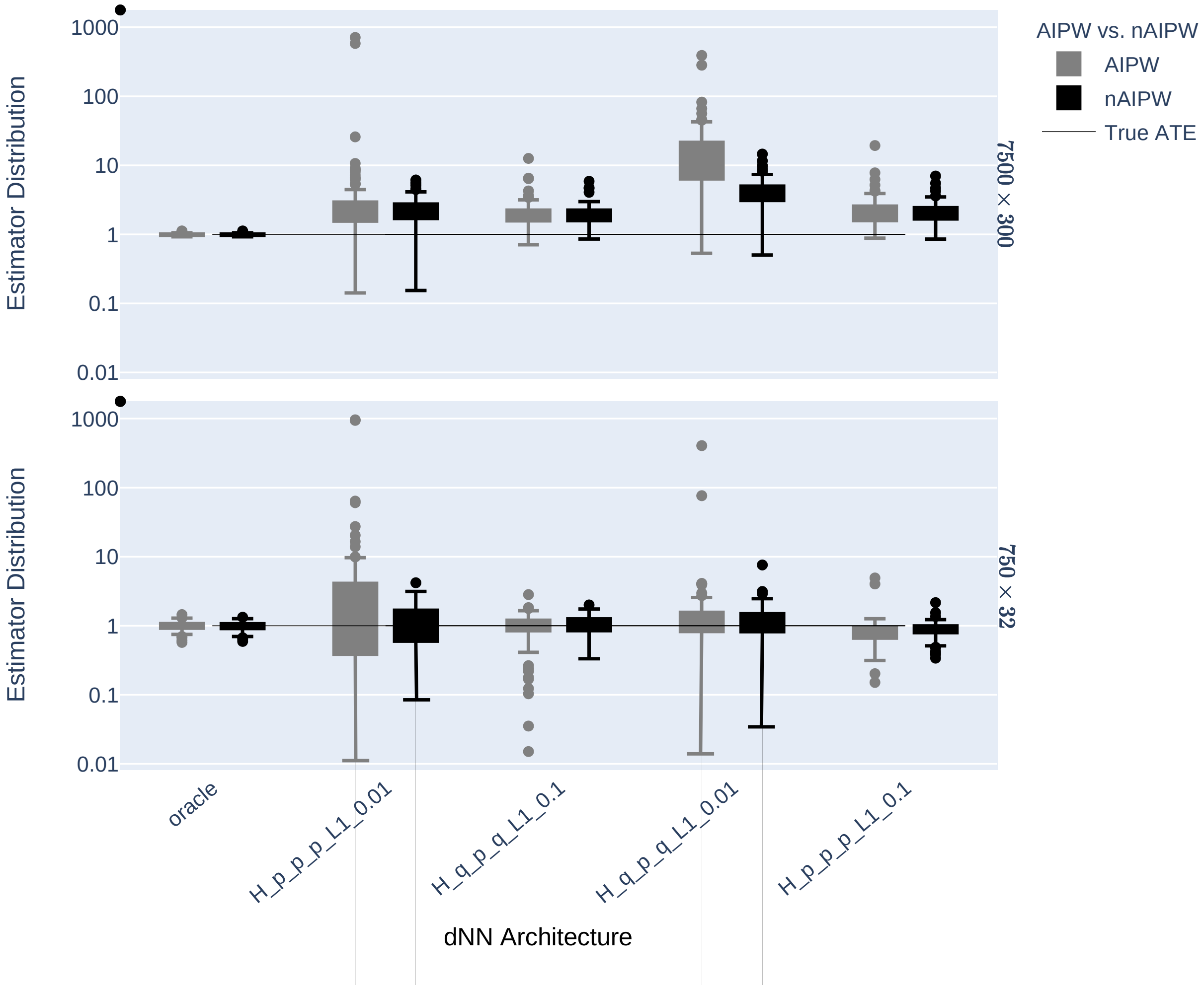


Figure 3.1: The distribution of log of the estimated AIPW and nAIPW in the 100 simulated iterations. The performance of nAIPW is clearly superior to the performance of AIPW as it is less dispersed and is more stable in terms of different hyperparameter settings. The x-axis values indicate the hyperparameter scenarios. For example, HL:q_p_q˙C_L1:0.01 shows that we have 3 hidden layers with size 365, the L_1 hyperparameter is 0.01. p is either 32 or 300 for the small or large datasets and $q \approx \frac{p}{10}$, that is, 3 or 30. In the bottom panel, the whiskers are capped at 0.01.

We also compare the estimators in different scenarios with bias, variance and their tradeoff measures:

$$\begin{aligned}
&\text{Bias} \qquad \hat{\delta} = \beta - \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_j \\
&\text{MC std} \qquad \hat{\sigma}_{MC} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{\beta}_j - \hat{\mu})^2} \\
&\text{MC RMSE} \qquad RMSE = \sqrt{\hat{\sigma}_{MC}^2 + \hat{\delta}^2} \\
&\text{Asymptotic StdErr} \qquad \hat{\sigma}_{SE} = \frac{1}{m}\sum_{j=1}^{m}\hat{\sigma}_j,
\end{aligned} \tag{3.35}$$

where $\beta = 1$, with $\hat{\beta}_j$s being the AIPW or nAIPW estimations in the jth simulation round, $\hat{\mu} = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_j$ and $m = 100$ being the number of simulation rounds and $\hat{\sigma}$ being the square root of (3.31) or (3.32).

Figure 3.2 demonstrates the bias, MC standard deviation (MC std) and the Root Mean Square Error (RMSE) of AIPW and nAIPW estimators for the scenarios where $n = 750$ and $n = 7500$, and for four hyperparameter sets (L_1 regularization and width of the dNN). In general, in each figure of the panel, the hyperparameter scenarios in the left imply a more complex model (with less regularization or a narrower network). In these graphs, the lower the values, the better the estimator. For the smaller data size $n = 750$ in the left three panels, the worst results are attributed to AIPW when there is the least regularization and the hidden layers are as wide as the number of inputs. To have more clear plots for comparison, we skipped plotting the upper bounds as they were large numbers; the lower bounds are enough to show the significance of the results. In the scenarios where there are smaller numbers of hidden neurons with 0.01 L_1 regularization, the bias, variance and their tradeoff (here measured by RMSE) are more stable. By increasing the L_1 regularization, these measures go down which indicates the usefulness of regularization and AIPW normalization for causal estimation and inference. Almost the same pattern is seen for the larger size ($n = 7500$) scenario, except for the bump in all three measures in the hyperparameter scenario where regularization remains the same ($L_1 = 0.01$) and the numbers of neurons in the first and last hidden layers are small too. In all three measures of bias, standard deviation and RMSE, nAIPW is superior to AIPW, or at least there is no statistically significant difference between AIPW and nAIPW.

We have noted that the results of step 1 NN architecture without L_1 regularization are too unstable and cannot be visually presented in the graphs. To avoid that, we have allowed a span of values for the L_1 regularization strengths: $L_1 = 0.01$ and $L_1 = 0.1$. The former case is close to no regularization. So, if the results of the latter are better than the former's, this is evidence that enough L_1 regularization must be imposed.

Figure 3.3 illustrates how the theoretical standard error formulas perform in MC experiments, and how accurately they estimate the MC standard deviations. In these two graphs, smaller does not necessarily imply superiority. In fact, the best results will be achieved as long as the confidence intervals of asymptotic SEs and MC SDs intersect. In the left two scenarios where the NN's complexity is high, the MC std and SE are far from each other. Additionally, in the hyperparameter scenarios where both the width of the NNs is small and regularization is higher, the MC std and SE are well separated. The scenario with largest regularization and wide NN architecture seems to the best scenario. That said, none of the scenarios confirm the consistency of SEs, which would likely also result in low coverage probability of the resulting confidence intervals.

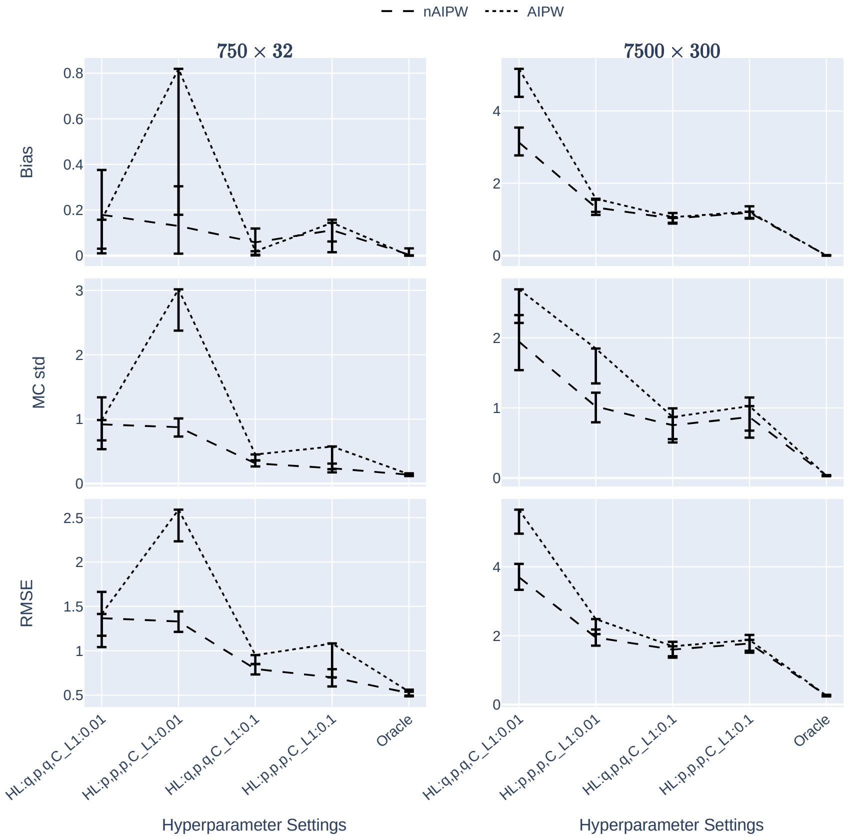


Figure 3.2: The bias, MC standard error and the root mean square error of the AIPW and nAIPW estimators for different data sizes and NN hyperparameters (L_1 regularization and width of the network). The x-axis values indicate the hyperparameter scenarios. For example, HL:q_p_q_C_L1:0.01 shows that we have 3 hidden layers with size 365, the L_1 hyperparameter is 0.01. p is either 32 or 300 for the small or large datasets and $q \approx \frac{p}{10}$, that is, 3 or 30. The estimates are capped at -10 and 10.

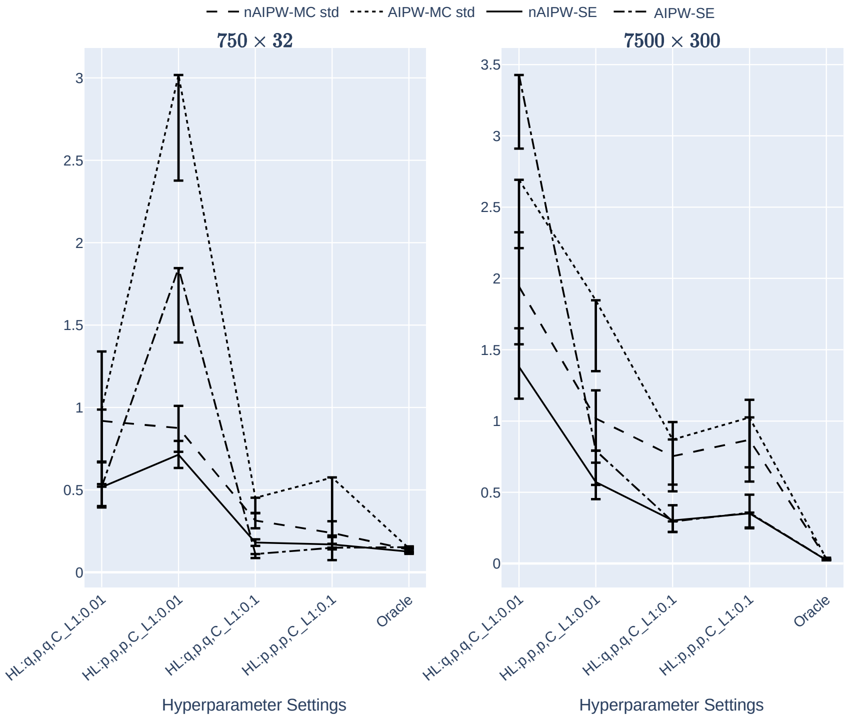


Figure 3.3: The MC standard deviation and the standard error of the AIPW and nAIPW estimators for different data sizes and NN hyperparameters (L_1 regularization and width of the network). The x-axis values indicate the hyperparameter scenarios. For example, HL:q_p_q_C_L1:0.01 shows that we have 3 hidden layers with size 365, the L_1 hyperparameter is 0.01.

3.7 Application: Food Insecurity and BMI

CCHS is a periodic survey designed to gather information concerning the health status, healthcare usage, and health determinants of the Canadian population. The 2021 CCHS encompasses individuals aged 12 and older residing in the ten provinces and three territorial capitals. Those not covered by the survey include individuals residing on reserves, in Aboriginal settlements within the provinces, and in some smaller sub-populations, collectively constituting less than 3% of the Canadian population aged 12 and above.

In most survey cycles, the questionnaire includes modules on general health, chronic conditions, smoking, and alcohol use. The 2021 cycle introduced new thematic content, such as inquiries about food security, home care, sedentary behavior, depression, and several other topics. Additionally, the survey incorporates questions about the characteristics of respondents, encompassing their involvement in the labor market, income, and socio-demographic factors.

In this chapter, we use the CCHS dataset to investigate the causal relationship of food insecurity and BMI. Other gathered information in the CCHS is used which might contain potential confounders, y-predictors and instrumental variables. The data are from a survey and need special methods such as resampling or bootstrap methods to estimate the standard errors. However, here, we use the data to illustrate the utilization of a dNN on the causal parameters in the case of empirical positivity violation. Additionally, the selection of factors to be included, based on their clinical or statistical relevance, has not been conducted in this instance, as our aim is solely for illustrative purposes. In order to reduce the amount of variability in the data, we have focused on the sub-population 18–65 years of age.

Figure 3.4 shows the ATE estimates and their 95% asymptotic confidence intervals with nIPW, DR and nDR methods, with four different neural networks which vary in terms of width and strength of L_1 regularization. The scenario that results in the largest R^2 (as a measure of outcome prediction performance) outperforms the other scenarios. The scenario that results in the largest AUC (as a measure of treatment model performance) results in the largest confidence intervals. This is because of more extreme propensity scores in this scenario. It is worth noting that the normalized IPW has smaller confidence intervals as compared to AIPW. However, as we do not know the truth about the ATE in this dataset, we can never know which estimator outperforms the other. To gain insight about this using the input matrix of this data, we simulated multiple treatments and outcomes with small to strong confounders and IVs and compared AIPW and nAIPW. In virtually all of them, the nAIPW is the best one. We do not present these results in this paper, but they can be provided to readers upon request.

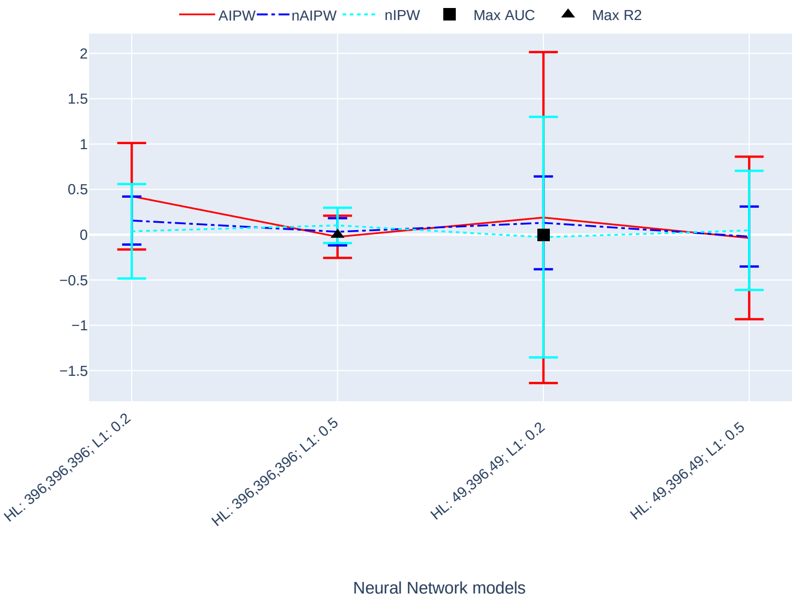


Figure 3.4: The ATE estimates and their asymptotically calculated 95% confidence intervals with NIPW, AIPW and nAIPW methods. The x-axis values indicate the hyperparameter scenarios. For example, HL:396,396,396;C_L1:0.2 shows that we have 3 hidden layers with size 396, the L_1 hyperparameter is 0.2.

3.8 Discussion

Utilizing machine learning algorithms such as NNs in the first-step estimation process is comforting as the concerns with regard to the non-linear relationships between the confounders and the treatment and outcome are addressed. However, there is no free lunch, and using NNs has its own caveats including theoretical as well as numerical challenges. Farrell et al. (2021) addressed the theoretical concerns where they calculated the generalization bounds when two separate NNs are used to model the treatment and the outcome. However, they did not use or take into account regularization techniques such as L_1 or L_2 regularization. As NNs are complex algorithms, they provide perfect prediction for the treatment when the predictors are strong enough (or might overfit). Through Monte Carlo simulations, we illustrated that causal estimation and inference with double NNs can fail without the usage of regularization techniques such as L_1 and/or extreme propensity scores are not taken care of. If L_1 regularization is not used, the normalization of the AIPW estimator (i.e., nAIPW) is advised to be employed as it dilutes the extreme predictions of the propensity score model and provides better bias, variance and RMSE. Our scenario analysis also showed that in the case of violation of the empirical positivity assumption in AIPW, normalization helps avoid blowing up the estimator (and standard error), but might be ineffective in taking into account confounding effects for some observations.

We note that the nAIPW estimator may not perform better when the empirical positivity is not violated as compared to when this assumption is violated (Figure 3.2, right hyperparameter scenarios). However, when the empirical positivity is violated, nAIPW can perform better than AIPW. If the empirical positivity is not violated, our results indicated that AIPW outperforms nAIPW.

An alternative estimator might be trimming the propensity scores to avoid extreme values. However, the causal effect estimator will no longer be consistent and there is no determined method for where to trim. We hypothesize that $\hat{h}^1 = \hat{g}\hat{\mathbb{E}}\frac{A}{\hat{g}} \times I(\hat{g} \in (0,\epsilon)) + \hat{g} \times I(\hat{g} \in (\epsilon,1))$ and $\hat{h}^0 = (1-\hat{g})\hat{\mathbb{E}}\frac{1-A}{1-\hat{g}} \times I(\hat{g} \in (1-\epsilon,1)) + (1-\hat{g}) \times I(\hat{g} \in (0,1-\epsilon))$ where $\epsilon = \frac{1}{n}$ will result in a consistent estimator, making the right assumptions, and will outperform both AIPW and nAIPW in the case of the empirical positivity violation. We will study this hypothesis in a future article.

Another reason why NNs without regularization fail in the causal estimation and inference is that the networks are not targeted, and are not directly designed for these tasks. NNs are complex algorithms with strong predictive power. This does not accurately serve the purpose of causal parameter estimation, where the empirical positivity assumption can be violated if strong confounders and/or instrumental variables (Angrist and Pischke, 2008) exist in the data. Ideally, the network should target the confounders and should be able to automatically limit the strength of predictors so that the propensity scores are not extremely close to 1 or 0. This was not investigated in this chapter and a solution to this problem is proposed in the next chapter.

Our empirical findings (Section 3.7) indicate that neither the asymptotic distributions of the AIPW nor nAIPW are perfectly symmetric. Also, we have employed the asymptotic standard errors for both AIPW and nAIPW, and we observe that the latter demonstrates smaller standard errors. However, we note that these asymptotic standard errors tend to underestimate the empirically calculated Monte Carlo standard deviations, as exemplified in our simulations presented in Figure 3.3. It is worth noting that the residual term associated with the scaled difference (3.12) is contingent on the positivity assumption, as indicated in assumptions (3.17). To ensure the validity of this formula, it is imperative that the assumptions (3.17) are met. Achieving this entails the precise estimation of propensity scores, which motivates our future research focus on the development of neural network architectures designed to avoid positivity violation during the first step.

Chapter 4

Targeted L_1-Regularization and Joint Modeling of Neural Networks for Causal Inference

4.1 Introduction

There are generally two approaches to addressing causal inference in observational studies. The first one is to draw population-level causal inference which goes back at least to the 1970s (Rubin, 1976). The second is to draw conditional causal inference which has received attention more recently (Johansson et al., 2016, van der Laan and Petersen, 2007). An example of a population-level causal parameter ATE,

$$\beta_{ATE} = \mathbb{E}[Y^1 - Y^0] = \mathbb{E}[\mathbb{E}[Y^1 - Y^0 \mid W]]. \tag{4.1}$$

The quantity $\mathbb{E}[Y^1 - Y^0 \mid W]$ is referred to as the conditional average treatment effect (CATE) (Athey and Imbens, 2016, Foster et al., 2011, Imai et al., 2013, Li et al., 2017, Lu et al., 2018, Taddy et al., 2016, Wager and Athey, 2018). CATE is not an individual-level causal parameter as the latter is impossible to estimate accurately unless both potential outcomes are observed for each individual, or W contains all the varying factors that make the causal relationship deterministic, either of which is unlikely to hold in practice. That said, under certain assumptions, the counterfactual loss, the loss due to the absence of counterfactual outcome, can be upper bounded (Shalit et al., 2017). The present article focuses on the estimation of ATE which does not require those assumptions.

Through a number of attempts, researchers have utilized ML models for the causal parameter estimation (Alaa et al., 2017, Belloni et al., 2012, 2014, Chernozhukov et al., 2018a, Farrell et al., 2021, Van Der Laan and Rubin, 2006). While the ultimate goal of an ML algorithm is to predict the outcome of interest as accurately as possible, it does not optimally serve the main purpose of causal parameter estimation. In fact, ML algorithms minimize some prediction loss containing the treatment or the observed outcome (and not counterfactual outcome) and without targeting any relevant predictor(s) such as confounding variables (Van der Laan and Rose, 2011).

Including confounders for the estimation of ATE in observational studies avoids potential selection bias (Angrist and Pischke, 2008), however, in practice, we do not have a priori knowledge about the confounders and the ML algorithm minimizes the loss function without discriminating between the input covariates. In fact, the ML algorithm can successfully learn the linear and non-linear relationships between the confounders and the treatment and outcome, but at the same time, might learn from potential IVs present in the data as well (the variables that predict the treatment, but not the outcome). If there are strong confounders or IVs among the covariates, the predictions of treatments (i.e., the propensity scores) can become extreme (near zero or one) which in turn can make the estimates unstable. While possibly reducing the bias, the variance gets elevated at the same time. Less complex models, on the other hand, may suffer from large bias (under-fitting) but can obtain more stable causal parameter estimation. This conflict between the necessary complexity in the model(s) and the bias-variance tradeoff motivates to develop ML algorithms for step 1 that provide a compromise between learning from confounders and IVs to entail a balance between the bias and variance of the causal parameter in step 2. In addition to a low bias-variance tradeoff, the asymptotic normality of the causal effect estimator is wanted for inferential statistics.

Chernozhukov et al. (2018a) investigated the asymptotic normality of orthogonal estimators of ATE (including AIPW) when two separate ML algorithms model the treatment and outcome, referred to as Double Machine Learning (DML). With the same objective, Farrell et al. (2021) utilized two separate neural networks (we refer to as the double NN or dNN), without the usage of any regularization other than using SGD for model optimization. SGD does impose some regularization but is insufficient to control the complexity of NN algorithms where strong predictors exist in the data (Rostami and Saarela, 2022a). Rostami and Saarela (2022a) experimentally showed that when AIPW is utilized, dNN performs poorly. The normalization of AIPW helps control both the bias and variance of the estimator. Further, they illustrated that imposing the L_1 regularization on all of the parameters (without targeting a specific set of input features) helps reduce the bias, variance, and MSE of the ATE estimators up to some extent. Simulations indicated that when dNN is used, with or without regularization, nAIPW outperforms AIPW. For a comprehensive literature review on the doubly robust estimators (including AIPW) see Moosavi et al. (2021).

The strategy of targeting a specific type of features can be designed in NN architectures along with the necessary optimization and regularization techniques. Flexible NN structures, optimizations and regularization techniques are easily programmed in deep learning platforms such as pytorch.

Shi et al. (2019) proposes a neural network architecture, referred to as the DragonNet, that jointly models the treatment and outcome, in which a multi-tasking optimization technique is employed. In the DragonNet architecture, the interaction of the treatment and non-linear transformations of the input variables are considered. Chernozhukov et al. (2022b) uses the Riesz Representer (Chernozhukov et al., 2018b) as the minimizer of a stochastic loss, which provides an alternative for the propensity score estimation, and aims to prevent the empirical consistency assumption violation issue (Rostami and Saarela, 2022a). Chernozhukov et al. (2022b) also use the joint modeling of the Riesz Representer and the outcome through multi-tasking, and they call their method auto Double Machine Learning (Auto-DML). Chernozhukov et al. (2022c) optimized an L_1 regularized loss function to estimate weights rather than estimating propensity scores and plugging them into the AIPW estimator. Chernozhukov et al. (2020) proposed optimizing a minimax loss function for the same purpose. In this body of work, it is still unclear how to hyperparameter tune the chosen NN architecture for causal inference, especially for the ATE estimation.

Other techniques of feature selection before propensity score estimation have been proposed in the literature (Farrell, 2015). However, hard thresholding might ignore important information hidden in the features.

The objective of this research is to experimentally investigate how NN-type methods can be utilized for ATE estimation, and how the hyperparameters can be tuned to achieve the best bias-variance tradeoff for the ATE estimators. This is done in the presence of strong IVs and confounders. The papers cited above do not consider this general scenario.

In this research our goal is not any of the following: 1. We do not aim to compare NNs with other ML algorithms to see which ones outperform the others. By the no-free-lunch theorem, (Shalev-Shwartz and Ben-David, 2014), there is no specific algorithm that can learn all relationships sufficiently well. Thus, it is expected that some ML algorithms are better in some scenarios and other algorithms in other scenarios. 2. We do not aim to study different types of causal parameters. 3. We do not aim to study different estimators of the Average Treatment Effect. 4. We do not aim to study feature selection or other types of methods that can prevent IVs to feed into the model of the treatment in the first step inference.

Throughout this research, we utilize nAIPW as it outperforms AIPW estimator in the presence of strong confounders and IVs (Rostami and Saarela, 2022a). To target the relevant inputs, we propose two methods. First, employing a type of L_1 regularization on top of the common L_1 regularization on all the network parameters. Second, we propose a joint model of the treatment and outcome in a Neural Network (jNN) architecture where we place both the treatment and outcome on the output layer of a multi-layer perceptron (Friedman et al., 2001). This NN architecture is appealing as it models the treatment and outcome simultaneously which can potentially target the relevant covariates that are predictive of both treatment and outcome (or confounder) and can mitigate or ignore the IVs' effects on the predictions. We will investigate if both or either of these ideas improves the bias-variance tradeoff of the causal effect estimator as compared to a dNN model.

In this research, the NN architecture that jointly models the treatment and outcome here is referred to as jNN. The parameters or weights are estimated by minimizing a regularized multi-task loss which is the summation of the Cross-Entropy (for modeling the binary treatment) and MSE loss (for modeling the continuous outcome) (Bishop, 2006). Multi-task learning can help target the predictors of both treatment and outcome that are placed in the output layer, and also it helps to resist over-fitting in case of many irrelevant inputs (Ruder, 2017). Other benefits of multi-task learning are listed in Section 4.2.2. Also, two types of L_1 regularization terms are used in order to dampen the instrumental variables and strong confounders.

To show the effectiveness of jNN and dNN, a thorough simulation study is performed and these methods are compared in terms of the number of confounders and IVs that are captured in each scenario, the prediction measures, and the bias and variance of causal estimators. To investigate whether our network targets confounders rather than IVs and also dampens the impact of strong confounders on the propensity scores, we calculate the bias-variance tradeoff of causal estimators (i.e., minimal MSE) utilizing the NN predictions; Low bias means the model has mildly learned from confounders and other types of covariates for the outcome, and low variance means the model has ignored IVs and has dampened strong confounders in the treatment model. Further, a comparison between the methods is made on the CCHS dataset where the intervention/treatment is food security versus food insecurity and the outcome is individuals' BMI.

The organization of this paper is as follows. In Section 4.1.2 we define the problem setting and the causal parameter to be estimated. In Section 4.2 we introduce the NN-type methods, their loss functions, and hyperparameters. Section 4.3 provides a quick review of the ATE estimators. In Section 4.4 our simulation scenarios are stated along with their results in Section 4.4.2. The results of the application of our methods on a real dataset are presented in Section 4.5. We conclude the paper in Section 4.6 with some discussion on the results and future work.

4.1.1 Notation

Let us consider an IID dataset denoted as $O = (O_1, O_2, \ldots, O_n)$, generated by the data generating process P. Each individual observation O_i is characterized by a finite-dimensional random vector $O_i = (Y_i, A_i, W_i)$. Here, Y represents the outcome, A represents the treatment status, and $W = (X_c, X_y, X_{iv}, X_{irr})$ represents a set of covariates. We assume that the treatment variable A follows a Bernoulli distribution with probability π, where π is determined by a function $f_1(X_c, X_{iv})$. Additionally, the outcome variable $Y = f_2(A, X_c, X_y) + \epsilon$. The covariates are partitioned into different sets: X_c is the set of confounding variables, X_{iv} are the instrumental variables (that are correlated with the treatment but not the outcome), X_y represents the covariates that predict the outcome independently of the treatment, and X_{irr} includes irrelevant or noise inputs (as illustrated in Figure 1).

The symbol P denotes the true joint probability distribution of the observed data O, while $\hat{P}_n$ represents its sample-based approximation. We define $\hat{P}_n$ as any distribution of (Y, A, W) such that the marginal distribution of W is estimated by the empirical distribution, and the conditional distribution of $(Y \mid A = a, W)$ with the finite mean $\mathbb{E}[Y \mid A = a, W]$.

Furthermore, we introduce three quantities that are referred to as the nuisance or infinite-dimensional parameters: Q^1 represents the expected outcome for the treated group, defined as $Q^1 := Q(1, W) = \mathbb{E}[Y \mid A = 1, W]$, and Q^0 represents the expected outcome for the untreated group, defined as $Q^0 := Q(0, W) = \mathbb{E}[Y \mid A = 0, W]$, and $g(W) = \mathbb{E}[A \mid W]$ is the propensity score, where all expectations are taken with respect to the underlying distribution P. The symbol $\hat{}$ is used to indicate corresponding finite-sample estimators for population-level quantities.

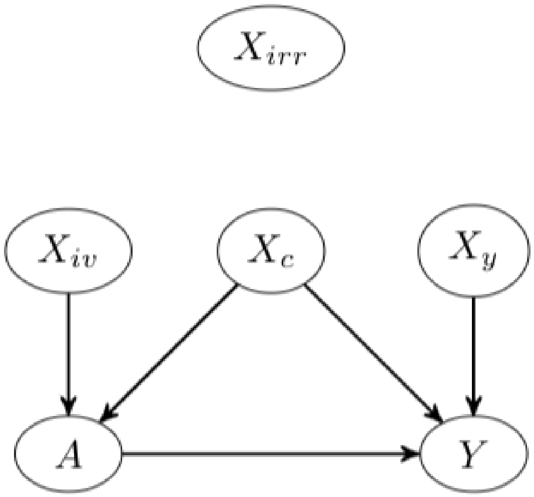


Figure 4.1: The causal relationship between A and y in the presence of other factors in an observational setting.

4.1.2 Problem Setup and Assumptions

The fundamental problem of causal inference states that individual-level causality cannot be exactly determined since each person can experience only one value of A. Thus, it is customary to only estimate a population-level causal parameter, in this research ATE (4.1).

For identifiablity of the parameter, the following assumptions must hold true. The first assumption is the Conditional Independence, Ignorability or Unconfoundedness stating that, given the confoudners, the potential outcomes are independent of the treatment assignments ($Y^0, Y^1 \perp A \mid W$). The second assumption is Positivity which entails that the assignment of treatment groups is not deterministic ($0 < Pr(A = 1 \mid W) < 1$, (Van der Laan and Rose, 2011), page 344). The third assumption is Consistency which states that the observed outcomes equal their corresponding potential outcomes ($Y^A = y$). There are other modeling assumptions made such as time order (i.e., the covariates W are measured before the treatment), IID subjects, and a linear causal effect.

4.2 Prediction Models

NNs are complex nonparametric models that approximate the underlying relationship between inputs and the outcome. The objective in causal inference, however, is not necessarily to leverage the maximum prediction strength of NNs and in fact, the NN architecture should be designed and tuned so that it pays more attention to the confounders.

The most important requirement of ML models such as NNs in causal inference is that although the outcome prediction model should minimize the corresponding loss (fit to get the best outcome prediction possible), given all of the covariates, the loss function associated with the propensity score model should not necessarily be minimized. Ideally, the instrumental variables or strong confounders which can give extreme fitted probability values (near zero or one) should be controlled when minimizing the loss. This can help prevent the elevation of the variance of the causal estimator (i.e., prevent the violation/near violation of the positivity assumption (Petersen et al., 2012, Van der Laan and Rose, 2011)). In summary, the prediction models should be strong enough to learn the linear and non-linear relationships between the confounders and treatment, but should not provide perfect predictions. We hypothesize that the employed NNs methods with the regularization techniques have the property of ignoring or damping strong confounders and/or instrumental variables.

4.2.1 Joint Neural Network

The jNN architecture is a combination of multiple ideas (see Sections 4.2.2–4.2.4) for causal parameter estimation purposes mentioned above.

The jNN models are:

$$\begin{bmatrix} \mathbb{E}[Y \mid A, W] \\ \mathbb{E}[A \mid W] \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta A + W\alpha + \mathbf{H}\Gamma_Y \\ G(\gamma_0 + W\alpha' + \mathbf{H}\Gamma_A) \end{bmatrix} \tag{4.2}$$

where α and α' are the skip connections parameters, $\mathbf{H} = f(f(...(f(W\boldsymbol{\Omega}_1)\boldsymbol{\Omega}_2)...)\boldsymbol{\Omega}_L)$ is the last hidden layer matrix which is a non-linear representation of the inputs (L is the number of hidden layers), where Ω_j's are the weight matrices that connect layers of NN, G is the logistic link function, and Γ_A and Γ_Y are the parameters that regress $\mathbf{H}$ to the log-odds of the treatment assignment or to the outcome in the output layer. The large square brackets around the equations above are meant to emphasize that both treatment and outcome models are trained jointly. The non-linear relationships between the inputs and the treatment and outcome can have arbitrary forms (which might not be the same for the treatment and outcome). The NNs can approximate such non-linear relationships even though one activation function is used. In fact, this property of NNs frees us from pre-specifying basis functions (Farrell, 2015) as they can be estimated automatically.

The jNN architecture minimizes a multi-task loss (Section 4.2.2) to estimate the parameters of the network:

$$\begin{aligned} L(\mathcal{P}, \beta, \alpha) = a \sum_{i=1}^{n} \Big[Y_i - \beta_0 - \beta A_i - W_i\alpha - H_i^T\Gamma_Y\Big]^2 + \\ b \sum_{i=1}^{n} \Big[A_i \log\Big(G(\gamma_0 + W\alpha' + H_i^T\Gamma_A)\Big) + (1 - A_i)\log\Big(1 - G(\gamma_0 + W\alpha' + H_i^T\Gamma_A)\Big)\Big] + \\ C_{L_1} \sum_{\omega \in \mathcal{P}} |\omega| + C_{L_{1TG}} \Big(\sum_{\omega \in \Gamma_A} |\omega| + \sum_{\omega \in \boldsymbol{\Omega}_1} |\omega| \Big), \end{aligned} \tag{4.3}$$

where $a, b, C_{L_1}, C_{L_{1TG}}$ are hyperparameters, that can be set before training or be determined by Cross-Validation, which can convey the training to pay more attention to one part of the output layer.

The jNN can have an arbitrary number of hidden layers, or the width of the network ($\mathcal{H}$) is another hyperparameter. For a l_h-layer network, $\mathcal{H} = [l_1, l_2, ..., l_h]$, where l_j is the number of neurons in layer j, $j = 1, 2, ..., h$. $\mathcal{P} = \{\omega \in \Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Gamma_Y \cup \Gamma_A\}$, are the connection parameters in the nonlinear part of the network, with Ω's being shared for the two outcome and propensity models. Noted that the number of parameters with L_1 regularization (third term on (4.3)) is $|\mathcal{P}| = (p + 1) \times l_1 + (l_1 + 1) \times l_2 + ... + (l_{h-1} + 1) \times l_h + (l_h + 1) \times 2$, including the intercepts in each layer.

The following subsections list the potential benefits and the rationale behind the proposed network (Equations (4.2) and (4.3)).

4.2.2 Bivariate Prediction, Parameters Sharing, and Multi-Task Learning

One of the main components of the jNN architecture is that both treatment and outcome are placed and modeled in the output layer simultaneously. The hypothesis here is that the network learns to get information from the inputs that predict both treatment and outcome, i.e., the confounders. This bivariate structure is intertwined with a multi-task learning or optimization. Ruder (2017) reviews the multi-tasking in machine learning and lists its benefits such as implicit data augmentation, regularization, attention focusing, and Representation bias. Caruana (1994) showed that overfitting declines by adding more nodes to the output layer as compared to modeling each output separately (Baxter, 1997). The multi-task is used when more than one output is used. Multi-task learning is common in the field of Artificial Intelligence and Computer Vision, for example, for the object detection task where the neural network predicts the coordinates of the box around objects and also classifies the object(s) inside the box (see for example (Redmon et al., 2016)). The multi-task learning is used in jNN in order to investigate if the model pays more attention to the confounders than other types of inputs.

4.2.3 Regularization

The jNN architecture can be resistant to overfitting by adding regularization to the network. Preliminary simulations revealed that L_2, and the Dropout (Goodfellow et al., 2016) regularization techniques do not result in satisfactory causal effect estimation, and the inherent regularization in the Stochastic Gradient Descent (Goodfellow et al., 2016) is also insufficient, while L_1 regularization is effective. We did not use the early-stopping as a regularization technique.

The L_1 regularization, third summation in (4.3), shrinks the magnitude of the parameter estimates of the non-linear part of the architecture which, in effect, limits the influence of X_{irr} and X_{iv}, X_y, and X_c on both treatment and the outcome. The motivation behind the L_1 regularization is to avoid overfitting for better generalization.

The ideal situation for causal parameter estimation is to damp the instrumental variables and learn from confounders and y-predictors only. Henceforth another version of the L_1 regularization is introduced here, referred to as the targeted L_1 regularization, or L_{1TG}, to potentially reduce the impact of instrumental variables on the outcome and more importantly on the propensity scores. The motivation is that by introducing shrinkage on the connections between the last hidden layer and the treatment, the neural network is trained to learn more about confounders than IVs in the last hidden layer as the outcome model is free to learn as much as possible from confounders. The caveat here might be that if the last hidden layer is large enough, some of the neurons can learn confounders while other learn from IVs, thus motivating to consider limiting the number of neurons in the last hidden layer. These hypotheses and ideas are considered in the simulation studies.

4.2.4 Linear Effects and Skip Connections

The terms $\beta A + W\alpha$ and $W\alpha'$ in (4.2) are responsible for potential linear effects. Theoretically, the non-linear parts of the NNs can estimate linear effects, but it is preferable to use linear terms if the relationship between some of the inputs and the outcome/treatment are linear for more accurate linear effect estimation. The benefit of including linear terms in the equations has been verified in our preliminary simulation studies.

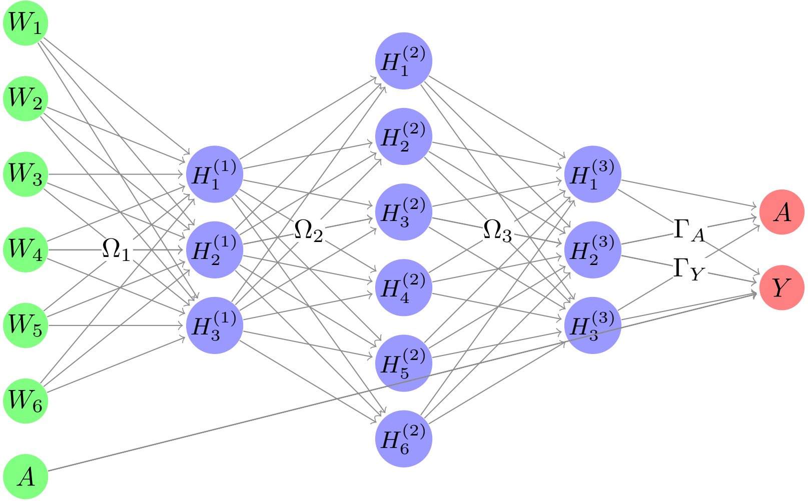


Figure 4.2: An Joined Neural Network architecture that incorporates linear effect of the treatment on the outcome, and the nonlinear relationship between the covariates and the treatment assignment and the outcome, all three tasks at the same time.

These linear terms are referred to the skip-connections in ML literature (He et al., 2016) which connect some layers to two or more layers forward. In ML literature, they are primarily used in very deep neural networks to facilitate optimizations. But they are used in jNN to model the linear effects directly. More specifically, skip connections connect the covariates to both treatment and outcome in the output layers and connect the treatment in the input layer to the outcome in the output layer. The latter skip connection is shown in Figure 4.2. It should be noted that this skip connection in particular is independent of the treatment in the output layer to avoid perfect prediction of the propensity scores.

4.2.5 Double Neural Networks

In order to study the significance of the proposed method through simulations, we compare jNN with the double dNN (Chernozhukov et al., 2018a) method. dNN is generally referred to the strategy of modeling the treatment and outcome separately utilizing two different models:

$$\begin{aligned} \mathbb{E}[Y \mid A, W] &= \beta_0 + \beta A + W\alpha + \mathbf{H}\Gamma_Y \\ \mathbb{E}[A \mid W] &= G(\alpha_0 + W\alpha' + \mathbf{K}\Gamma_A), \end{aligned} \tag{4.4}$$

where two separate neural networks model y and A (no parameter sharing), G is the logistic link function, and α and α' are the skip connections parameters, and $\mathbf{H}$ and $\mathbf{K}$ are composite of non-linear functions similar to $\mathbf{H}$ in (4.2) in the outcome and PS architectures, respectively, and Γ_Y and Γ_A are the final matrices that connect the last hidden layer to the output in the outcome and PS architectures, respectively. In this paper, the dNN algorithm refers to two neural networks to model the treatment and outcome separately. To make the two jNN and dNN models comparable, we let the NN architectures to be as similar as possible in terms of skip connections and regularization techniques. The loss functions in dNN to be optimized are:

$$
\begin{aligned}
&L_y(\mathcal{P}_y, \beta, \alpha) = \sum_{i=1}^{n} \Big[Y_i - \beta_0 - \beta A_i - W_i\alpha - \mathbf{H}_i^T \Gamma_Y\Big]^2 + C'_{L_1} \sum_{\omega \in \mathcal{P}} |\omega|, \\
&L_A(\mathcal{P}_A) = \sum_{i=1}^{n} \Big[a_i \log \Big(G(\alpha_0 + W\alpha' + \mathbf{K}_i^T \Gamma_A)\Big) + (1 - a_i) \log \Big(1 - G(\alpha_0 + W\alpha' + \mathbf{K}_i^T \Gamma_A)\Big)\Big] + C''_{L_1} \sum_{\omega \in \mathcal{P}} |\omega| + \\
&C_{L_{1TG}}(\sum_{\omega \in \Gamma_A} |\omega| + \sum_{\omega \in \mathbf{\Omega}_1} |\omega|),
\end{aligned}
\tag{4.5}
$$

4.3 ATE Estimation

The Causal Parameter Estimation algorithm is a two-stage process. The regression functions $\mathbb{E}[A \mid W]$, $\mathbb{E}[Y \mid A = 1, W])$, $\mathbb{E}[Y \mid A = 0, W]$ are estimated using the ML algorithms such as jNN or dNN in step 1. And in step 2, the predictions are inserted into the causal estimators such as (4.6), below.

ATE Estimators

There is a wealth of literature on how to estimate ATE and there are various versions of estimators including AIPW and nAIPW:

$$
\begin{aligned}
\hat{\beta}_{AIPW} &= \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{A_i(Y_i - \hat{Q}_i^1)}{\hat{g}_i} - \frac{(1 - A_i)(Y_i - \hat{Q}_i^0)}{1 - \hat{g}_i} \Big) + \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_i^1 - \hat{Q}_i^0, \\
\hat{\beta}_{nAIPW} &= \sum_{i=1}^{n} \Big(\frac{A_i(Y_i - \hat{Q}_i^1) w_i^{(1)}}{\sum_{j=1}^{n} A_j w_j^{(1)}} - \frac{(1 - A_i)(Y_i - \hat{Q}_i^0) w_i^{(0)}}{\sum_{j=1}^{n} (1 - A_j) w_j^{(0)}} \Big) + \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_i^1 - \hat{Q}_i^0.
\end{aligned}
\tag{4.6}
$$

where $\hat{Q}_i^k = \hat{Q}(k, W_i) = \hat{\mathbb{E}}[Y_i \mid A_i = k, W_i]$ and $\hat{g}_i = \hat{\mathbb{E}}[A_i \mid W_i]$, and A_1 is the treatment group with size n_1 and A_0 is the treatment group with size n_0.

In the second step of the estimation procedure, the predictions of the treatment (i.e., propensity score, PS) and the outcome $\hat{\mathbb{E}}[Y_i \mid A_i = k, W_i]$, $k = 0, 1$, can be inserted in these estimators (4.6). Generalized Linear Models, any relevant Machine Learning algorithm such as tree-based algorithms and their ensemble (Friedman et al., 2001), SuperLearner (Van der Laan et al., 2007b), or Neural Network-based models (such as ours) can be applied as prediction models for the first step prediction task. We will use jNN and dNN in this thesis.

4.4 Simulations

A simulation study (with 100 iterations) was performed to compare the prediction methods jNN, and dNN by inserting their predictions in the nAIPW (causal) estimators (4.6). There are a total of 8 scenarios according to the size of the data (i.e., the number of subjects and number of covariates), and the confounding and instrumental variables strengths. We fixed the sample sizes to be $n = 750$ and $n = 7500$, with the number of covariates $p = 32$ and $p = 300$, respectively. The four sets of covariates had the same sizes $\#X_c = \#X_{iv} = \#X_y = \#X_{irr}$, equal to $n = 8$ and $n = 75$, and independent from each other were drawn from the MVN Distribution as $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$, with $\Sigma_{kj} = \rho^{j-k}$ and $\rho = 0.5$. Let

$\beta = 1$. The models to generate the treatment assignment and outcome were specified as

$$\begin{aligned} &A \sim Ber(\frac{1}{1+e^{-pr}}), \text{with } pr = f_a(X_c)\gamma_c + g_a(X_{iv})\gamma_{iv}, \\ &y = 3 + A + f_y(X_c)\gamma_c' + g_y(X_y)\gamma_y + \epsilon, \end{aligned} \tag{4.7}$$

The functions f_a, g_a, f_y, g_y select 30% of the columns and apply interactions and non-linear functions listed below (4.8). The strength of instrumental variable and confounding effects were chosen as $\gamma_c, \gamma_c', \gamma_y \sim Unif(r_1, r_2)$ where $(r_1 = r_2 = 0.1)$ or $(r_1 = 0.1, r_2 = 1)$, and $\gamma_{iv} \sim Unif(r_3, r_4)$ where $(r_3 = r_4 = 0.1)$ or $(r_3 = 0.1, r_4 = 1)$.

The non-linearities for each pair of covariates are randomly selected among the following functions:

$$\begin{aligned} l(x_1, x_2) &= e^{\frac{x_1 x_2}{2}} \\ l(x_1, x_2) &= \frac{x_1}{1+e^{x_2}} \\ l(x_1, x_2) &= (\frac{x_1 x_2}{10} + 2)^3 \\ l(x_1, x_2) &= (x_1 + x_2 + 3)^2 \\ l(x_1, x_2) &= g(x_1) \times h(x_2) \end{aligned} \tag{4.8}$$

where $g(x) = -2I(x \leq -1) - I(-1 \leq x \leq 0) + I(0 \leq x \leq 2) + 3I(x \geq 2)$, and $h(x) = -5I(x \leq 0) - 2I(0 \leq x \leq 1) + 3I(x \geq 1)$, or $g(x) = I(x \geq 0)$, and $h(x) = I(x \geq 1)$.

The activation function of the NNs is ReLU, and the depth of the network is 3 with sizes either ($[q, p, q]$ or $[p, p, p]$, where $q = p/4$. The models were trained with L_1 regularization and a batch size equal to $3{*}p$, over 200 epochs. The Adam optimizer, as described by Kingma and Ba in their 2014 paper, with a learning rate of 0.01 and momentum of 0.95, was used to estimate the network's parameters. In total, 32 first-step trained models were considered, encompassing two types of neural network models (jNN and dNN), two variations in neural network depth, two L_1 regularization strengths $(0.01, 0.1)$, and four targeted L_1 regularization strengths $(0, 0.1, 0.3, 0.7)$.

As in practice, the RMSE and covariate types are unknown, prediction measures of the outcome and treatment should be used to choose the best model in a K-fold cross-validation. R^2 and AUC each provide insight about the outcome and treatment models, respectively, but in our framework, both models should be satisfactory. To measure the goodness of fit of the prediction models (jNN and dNN) for causal inference purposes, we define and utilize a statistic which is a compromise (geometric average) between R^2 and AUC, here referred to as *geo*,

$$geo(R, D) = \sqrt[3]{R^2 \times D \times (1 - D)}, \tag{4.9}$$

where $D = 2(AUC - 0.5)$, the Somers' D index. This measure was not utilized in the optimization process (i.e., training the neural networks), and is rather introduced here to observe if the compromise between R^2 and AUC agrees with the models that capture more confounders than IVs. In fact, we do not seek to get the largest AUC or D index, as the larger values of these measures might indicate that the PS model has learned from IVs more than necessary which can harm the estimation rather than help. We will refer to $geo(R, D)$ simply as *geo*.

4.4.1 Selected Covariate Types

In order to identify which types of covariates (confounders, IVs, y-predictors, and irrelevant covariates) the prediction methods have learned from, we calculate the association between the inputs and the predicted values ($\hat{\mathbb{E}}[Y \mid A, W]$ and $\hat{\mathbb{E}}[A \mid W]$), and after sorting the inputs (from large to small values) based on the association values, we count the number of different types of covariates within top 15 inputs in the small case and top 150 inputs in the large case. The association between two variables here is estimated using the distance correlation statistic (Szekely et al., 2007) whose zero values entail independence and non-zero values entail statistical dependence between the two variables.

4.4.2 Results

Figures 4.3–4.6 present the overall comparison of different hyperparameter settings of jNN and dNN architectures in terms of five different measures, respectively: (1) The average number of captured confounders/IVs/y-predictors, (2) Average RMSE of causal estimators, (3) Average R^2, AUC and their mixture measure geo (4.9), (4) Bias, (5) MC standard deviation of nAIPW. The bootstrap confidence intervals for the bias, standard deviation and RMSE are calculated to capture significant differences between the simulation scenarios. The x-axis includes 16 hyperparameter settings, and as a general rule here, models in the left are most complex (less regularization and wider neural networks) and in the right are least complex. Noted that L_{1TG} regularization is only targeted at the treatment model.

The Figures 4.4 and 4.4 show how the complexity of both dNN and jNN (x-axis) impact the number of captured covariate types (i.e., confounders/IVs/y-predictors, top graph), RMSE (middle graph) and prediction measures (bottom graph). In almost all the hyperparameter settings, especially when $C_{L_{1TG}}$ is non zero, the number of captured confounders is larger and the number of captured IVs is smaller in jNN as compared to dNN. This shows the joint modeling has a benefit of focusing on the confounders, rather than IVs, especially in large data scenarios.

The RMSE of jNN is larger than that of dNN for models with zero targeted regularization (the scenarios in the left). With decreasing the complexity of the treatment model, the RMSE of both jNN and dNN decline. The jNN outperforms dNN in almost all of the hyperparameter settings in case of $n = 750$, but does not show a clear pattern in case of $n = 7500$. Further, the impact of the width of architectures (H) changes based on C_{L_1} regularization: wider architectures ($H = [p, p, p]$, p: number of covariates) with large C_{L_1} outperform other combinations of these two hyperparameters. This observation is more clear for smaller sized data, and for dNN model. In the small size scenarios, when the width is small ($H = [3, 32, 3]$), the outcome model is affected and has a smaller R^2. This means there are not enough neurons (on the first or last layer) to provide more accurate outcome predictions. In the best scenarios, the RMSE confidence intervals of jNN model are below those of dNN, illustrating a small preference of jNN over dNN in terms of RMSE. Comparing the three hyperparameters, $C_{L_{1TG}}$ is most effective, and zero values of this hyperparameter results in very large RMSEs for both dNN and jNN.

From Figures 4.5 and 4.6, it is observed that both jNN and dNN models have roughly the same values for the R^2 (outcome model performance) across hyperparameter settings and for both data sizes ($n = 7500$, and $n = 750$). That is, the targeted regularization in jNN does not impact the performance of the outcome model. The AUC, on the other hand, declines with higher values of $C_{L_{1TG}}$, and is almost always smaller or equal in dNN as compared to jNN. Further, larger values of geo in the small size data

Figure 4.3: The comparison of captured number of confounders, IVs and y-predictors, RMSE of nAIPW and its bootstrap 95% confidence interval, and prediction measures R^2, AUC and geo (geometric mean of R^2, AUC) for different hyperparameter settings and where the predictions come from jNN or dNN models. (n=750, p=32). The x-axis shows the hyperparameter settings. For example, HL:32, 32, 32, C_L1:0.01,C_L1TG:0.0 shows that we have 3 hidden layers with size 32, the L_1 hyperparameter is 0.01 and the targeted $L1TG$ hyperparameter is 0.0.

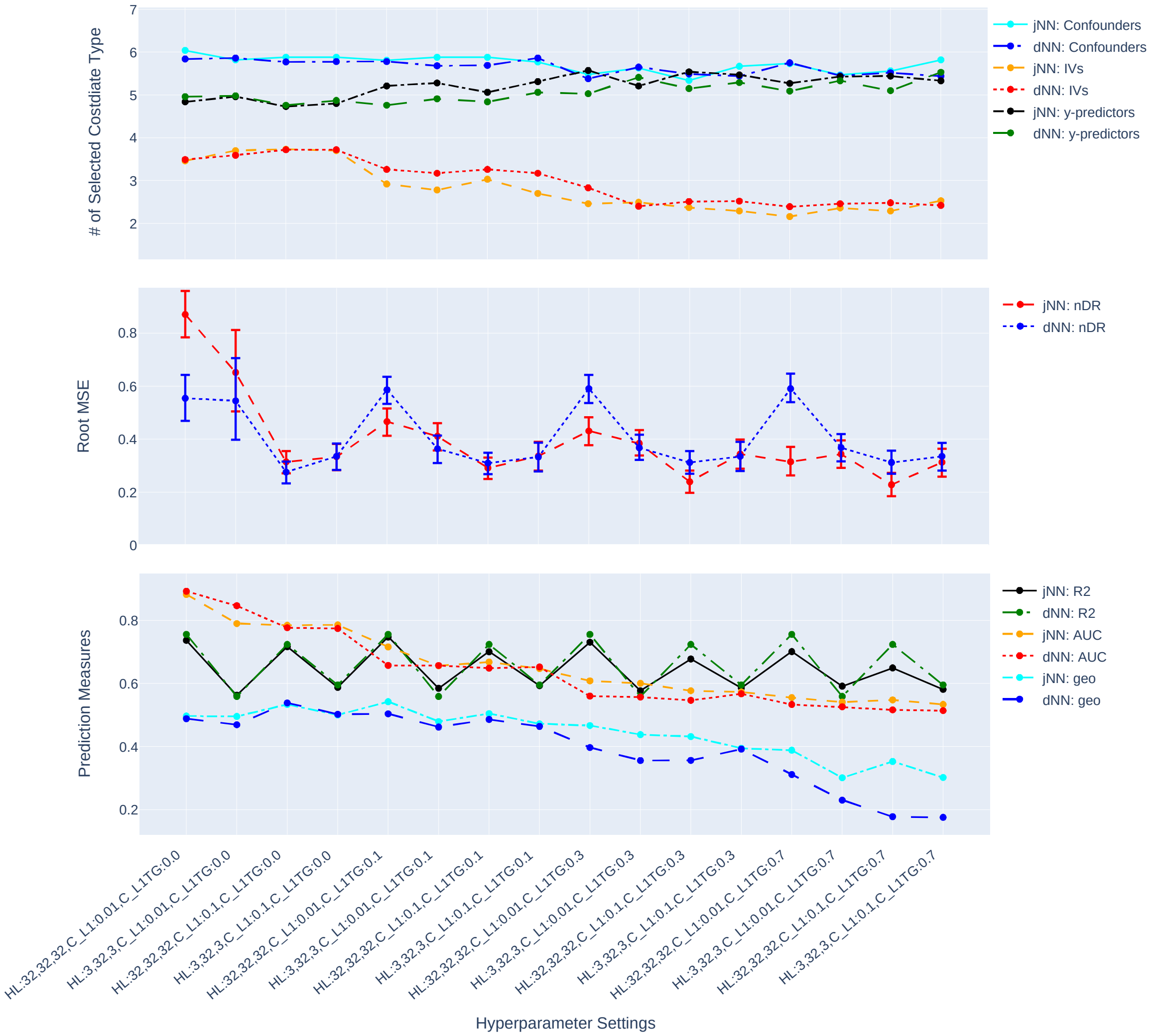
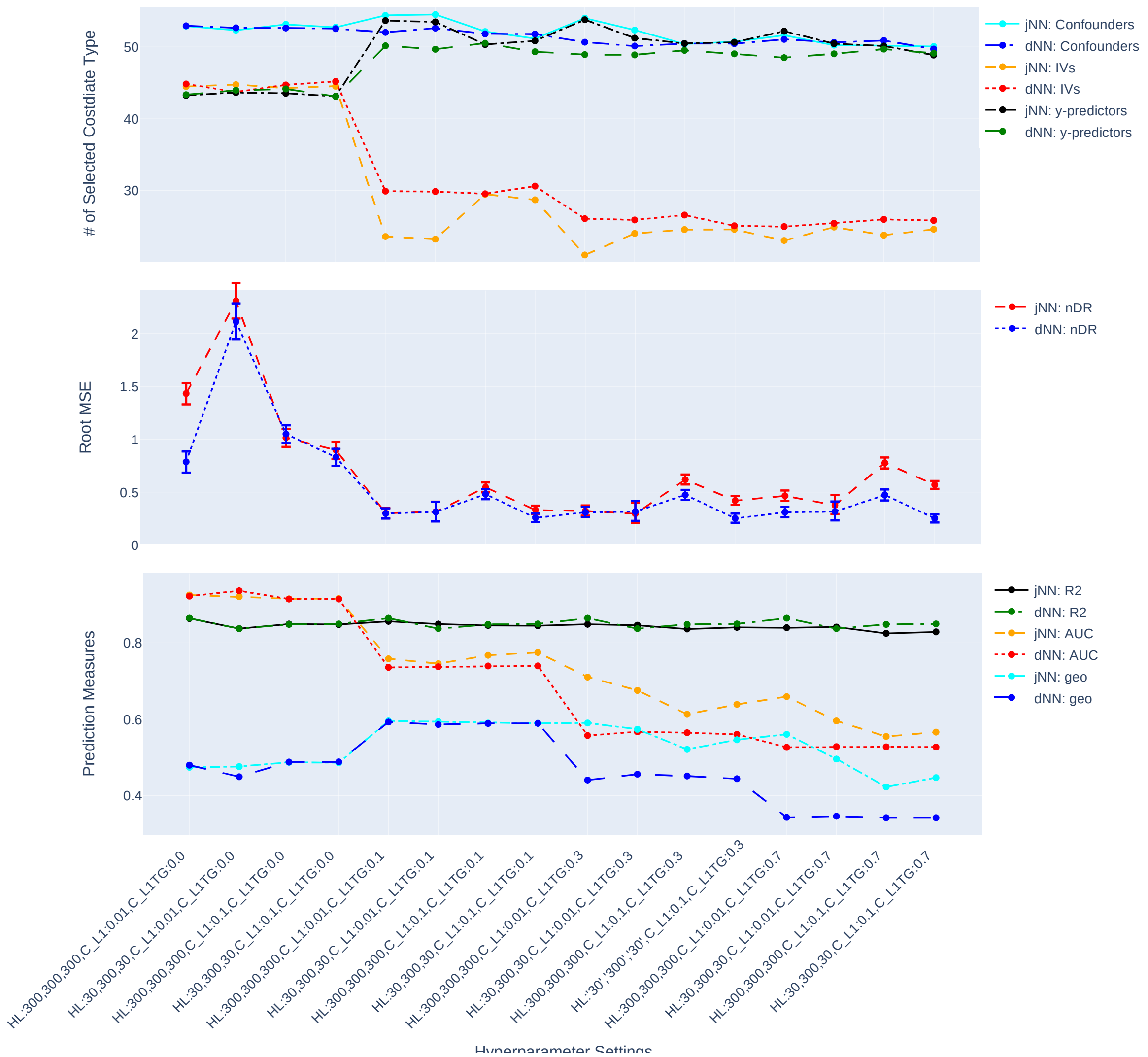


Figure 4.4: The comparison of captured number of confounders, IVs and y-predictors, RMSE of nAIPW and its bootstrap 95% confidence interval, and prediction measures R^2, AUC and geo (geometric mean of R^2, AUC) for different hyperparameter settings and where the predictions come from jNN or dNN models. (n=7500, p=300). The x-axis shows the hyperparameter settings. For example, HL:300, 300, 300, C_L1:0.01,C_L1TG:0.0 shows that we have 3 hidden layers with size 300, the L_1 hyperparameter is 0.01 and the targeted $L1TG$ hyperparameter is 0.0.

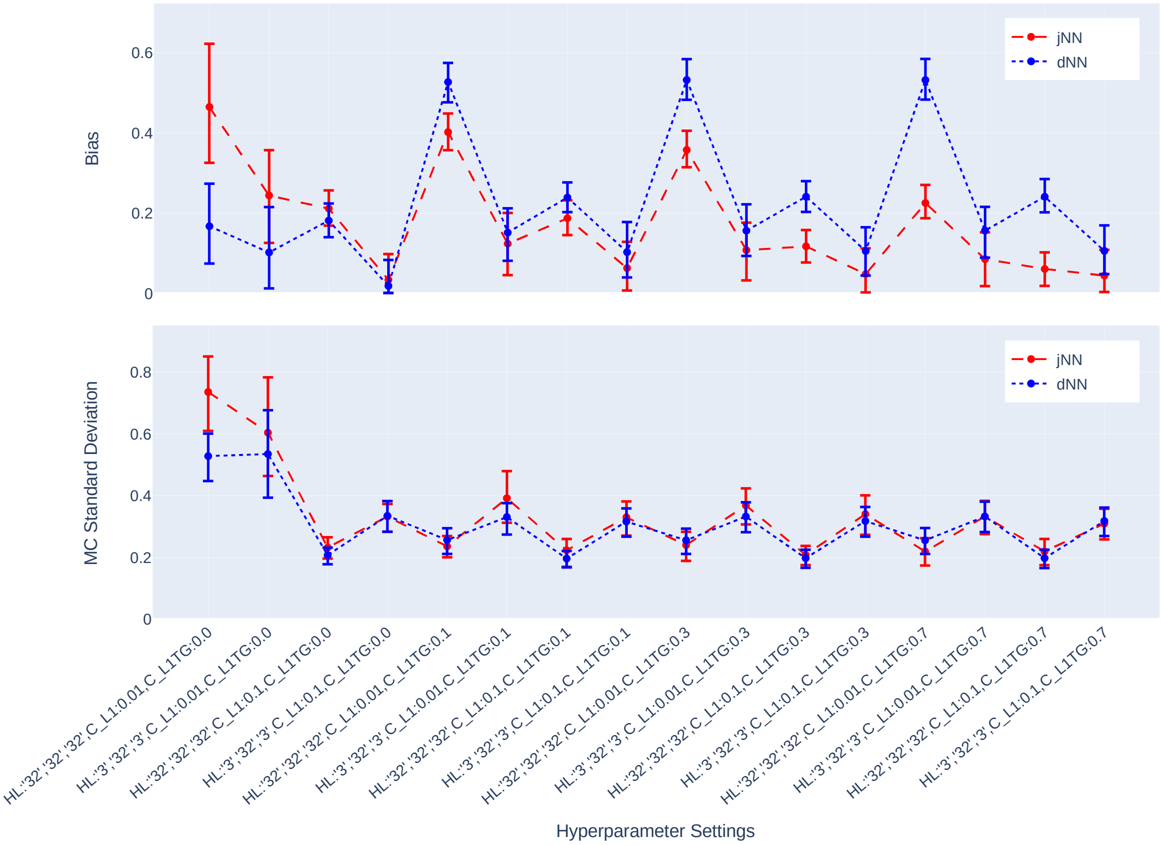


correspond to smaller RMSE, but no such pattern can be seen in the large data scenario.

Overall, the trends favor the idea that more complex treatment models capture larger number of IVs, have larger AUC (smaller geo.), and have larger RMSE. That is, more complex models are less favorable.

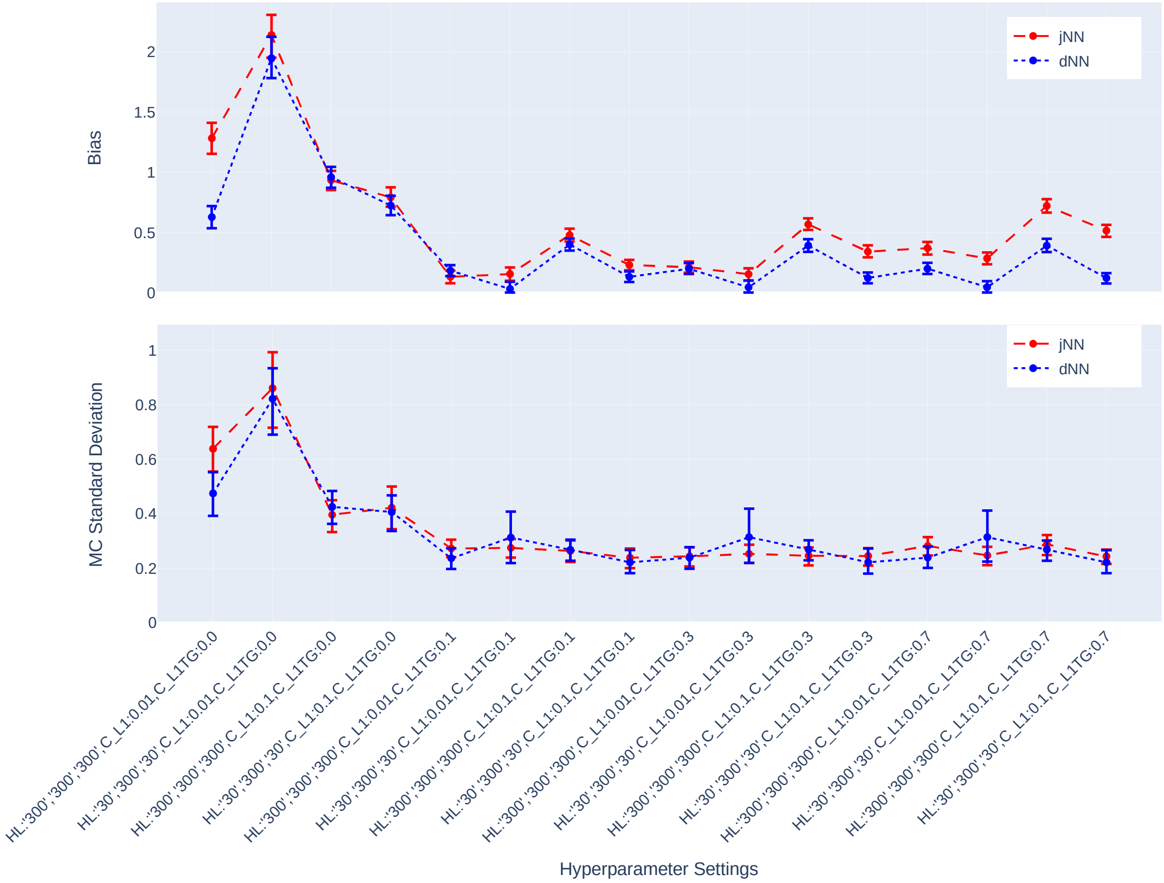
Figures 4.5 and 4.6 illustrate the bias and standard deviation of the causal estimators. As expected and mentioned in the Section 4.1, the models that do not dampen IVs suffer from large bias and standard deviation. The bias and standard deviation have opposite behavior in different settings, such that settings that produce larger standard deviation, results in small bias, and vice versa, except for the one setting

Figure 4.5: The bias and standard deviation of nAIPW and their bootstrap 95% confidence intervals for different hyperparameter settings where the predictions come from jNN or dNN models. (n=750, p=32). The x-axis shows the hyperparameter settings. For example, HL:32, 32, 32, C_L1:0.01,C_L1TG:0.0 shows that we have 3 hidden layers with size 32, the L_1 hyperparameter is 0.01 and the targeted $L1TG$ hyperparameter is 0.0.



that produces both largest bias and standard deviation. The fluctuations of the bias-variance across hyperparameter settings are larger in $n = 750$ case than in $n = 7500$ case. For small sample $n = 750$, the best scenario for jNN is $H = [32, 32, 32], C_{L_1} = 0.1, C_{L_{1TG}} = 0.7$ where both bias and standard deviation of jNN are small in the same direction. For the large sample $n = 7500$, however, the best scenario for jNN is $H = [30, 300, 30], C_{L_1} = 0.01, C_{L_{1TG}} = 0.7$ with a similar behavior. The best scenarios for dNN are slightly different. For small sample $H = [32, 32, 32], C_{L_1} = 0.1, C_{L_{1TG}} = 0.7$ and for the large sample $H = [30, 300, 30], C_{L_1} = 0.01, C_{L_{1TG}} = 0.7$ are most favorable.

Figure 4.6: The comparison of bias, Monte Carlo standard deviation and their bootstrap 95% confidence intervals of nAIPW, for different hyperparameter settings and the predictions come from jNN or dNN models. (n=7500, p=300). The x-axis shows the hyperparameter settings. For example, HL:300, 300, 300, C_L1:0.01,C_L1TG:0.0 shows that we have 3 hidden layers with size 300, the L_1 hyperparameter is 0.01 and the targeted $L1TG$ hyperparameter is 0.0.



4.5 Application: Food Insecurity and BMI

CCHS is a cross-sectional survey that gathers information on the health status, healthcare utilization, and factors influencing the health of the Canadian population. The 2021 CCHS is inclusive of individuals aged 12 and older residing in the ten provinces and the three territorial capitals, with certain exclusions like people on reserves and specific sub-populations, collectively accounting for less than 3% of the Canadian population aged 12 and above. Some recurring survey modules include questions on general health, chronic ailments, smoking, and alcohol consumption. In the 2021 cycle, additional topics like food security, home care, sedentary behavior, and depression were included. Alongside health-related queries, the survey also collects data on respondent attributes such as employment, income, and socio-demographics.

In this chapter, we employ the CCHS dataset to explore the cause-and-effect connection between food insecurity and BMI. We also utilize other data from the CCHS, which may contain potential confounding variables, outcome predictors, and instrumental variables. Since the data originate from a survey, specialized techniques like resampling or bootstrap methods are required to estimate standard errors. However, our focus here is to demonstrate the use of a dNN to assess causal parameters when empirical positivity violations are present. To enhance data consistency, our analysis centers on the sub-population aged 18 to 65 years.

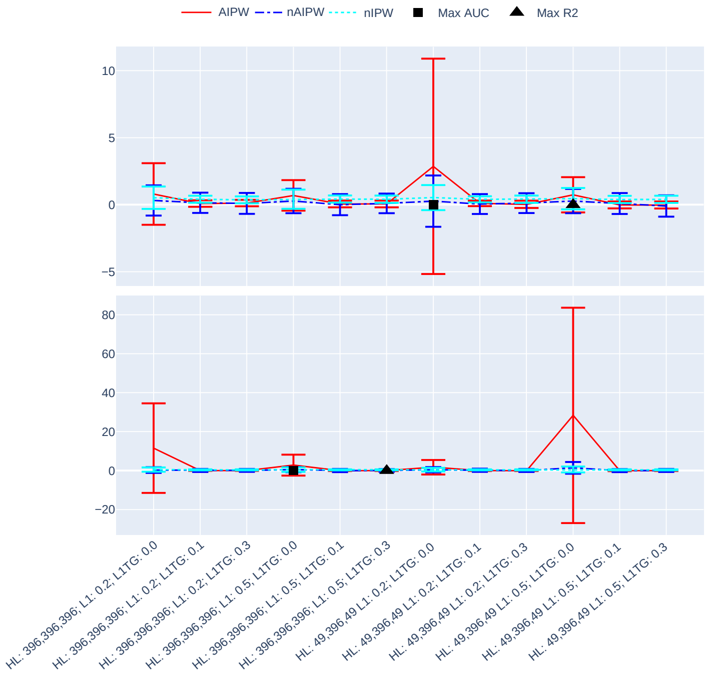


Figure 4.7: The ATE estimates and their asymptotically calculated 95% confidence intervals with nIPW, AIPW, and nAIPW methods. The x-axis values indicate the hyperparameter scenarios. For example, HL:396, 396, 396, C_L1:0.2,C_L1TG:0.0 shows that we have 3 hidden layers with size 396, the L_1 hyperparameter is 0.2 and the targeted $L1TG$ hyperparameter is 0.0. It can be seen that wider confidence intervals are for the cases when targeted L_1 regularization is zero (L1TG=0.0). A more clear view is presented in the next figure where the scenarios with extremely wide confidence intervals are removed.

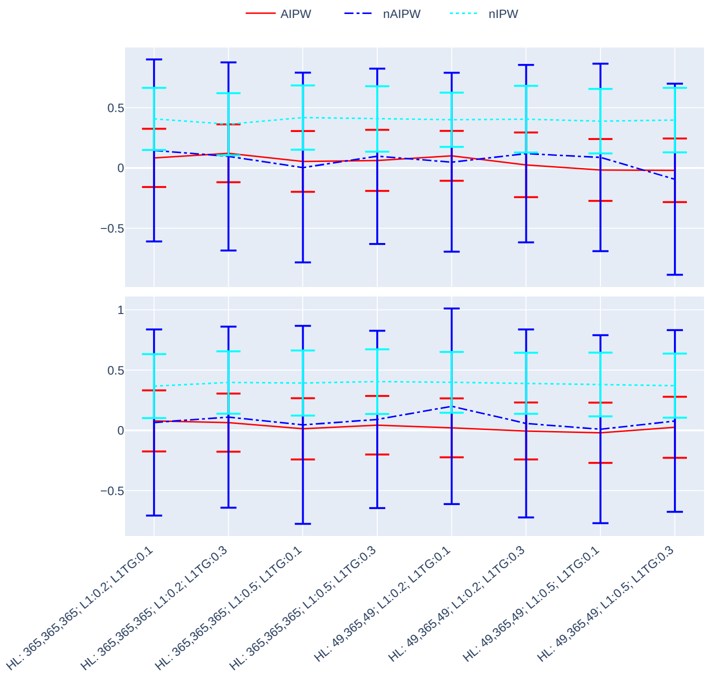


Figure 4.8: The ATE estimates and their asymptotically calculated 95% confidence intervals with nIPW, AIPW, and nAIPW methods. The x-axis values indicate the hyperparameter scenarios. For example, HL:365, 365, 365, C_L1:0.2,C_L1TG:0.1 shows that we have 3 hidden layers with size 365, the L_1 hyperparameter is 0.2 and the targeted $L1TG$ hyperparameter is 0.1.

Figure 4.7 and 4.8 present ATE estimates and their 95% asymptotic confidence intervals with nIPW, AIPW and nAIPW methods. Figure 4.7 contains hyperparameter settings where there is no targeted regularization and it shows how important this regularization technique is, especially for the AIPW estimator that has no normalization. We have removed these scenarios in Figure 4.8 for a more clear comparison between the remaining scenarios. The estimates and 95% CIs seem similar across the hyperparameter settings, but there is a clear difference between those of AIPW and nAIPW. This means that for this dataset, normalization might not be needed as the propensity scores do not behave extremely and AIPW does not blow up.

4.6 Discussion

In this paper, we have studied how hyperparameters of the Neural Network predictions in the first step can affect the ATE estimator. We have considered a general data generating process in which four types of covariates exist in the dataset, confounders, IVs, y-predictors, and irrelevant covariates. Two general NN architectures have been studied, jNN and dNN where in the former both the outcome and treatment are modeled jointly (with an appropriate loss function), and in the latter, they are modeled separately. We have observed that L_1 regularization especially the one that targets the treatment model (L_{1TG}) is an effective hyperparameter for achieving a better bias-variance trade-off for the nAIPW estimator. And, the number of neurons in the first and last layer of the network becomes irrelevant as long as the value of L_{1TG} is sufficient. Further, we have observed that in the hyperparameter settings where the IV effects are controlled, the estimation is less biased and more stable. Thus the targeted regularization is successful in dampening the IVs and preventing perfect prediction in the treatment model. Figures 4.3–4.6 illustrate that jNN is overall more stable and has a smaller RMSE in the small sample dataset scenario as compared to dNN. We utilized nAIPW in our simulations as they outperform or at least do not underperform AIPW and other estimators such as IPW, nIPW, AIPW, and SR. The nAIPW estimator has a normalization factor in the denumerator which can dilute the impact of extreme predictions of the propensity score model and protect the estimator against the positivity assumption violation Van der Laan and Rose (2011).

We utilized a geometric-type average of the R^2 and AUC to choose among the first step models. As the objective of optimization in the first step is increasing prediction performance which is not necessarily the same as the causal inference objectives, the usage of either R^2, AUC or their geometric average is sub-optimal. In a future study alternative approaches will be explored and compared with the said prediction measure.

A real strength of NNs would be to uncover hidden information (and thus confounder effects) in unstructured data such as text or image data. However, in this thesis, we have not studied the presence of unstructured data and it is left for future research.

There are limitations due to the assumptions and simulation scenarios and, thus, some questions are left to future studies to be explored. For example, the outcome here was assumed to be continuous, and the treatment to be binary. We also did not cover heavy tail outcomes or rare treatment scenarios. Also, the ratio of dimension to the size of the data was considered to be fairly small ($p \ll n$), and we have not studied the case where $n < p$. Furthermore, we did not study the asymptotic behavior of nAIPW when jNN or dNN predictions are used.

A limitation of jNN as compared to dNN is that if one needs to shrink the final hidden layer to control the complexity of the treatment model, by structure, we are limiting the complexity of the outcome model which might not be necessary. This might be resolved by another architectural design, which is left to future studies on the subject.

The usage of another regularization technique that controls the extremeness of propensity score values is a plausible approach. For example, a data-dependent term can be added to the loss function $\sum_{i=1}^{n} \frac{1}{g_i} + \frac{1}{1-g_i}$. Such a term discourages the network to obtain values extremely close to zero or one, as opposed to the negative log-likelihood term that encourages such tendencies. This approach might also focus less on the inputs that cause extreme values such as strong confounders or IVs. Examination of this approach is left to future studies.

In the design of the optimization, we did not consider a formal early stopping as a regularization technique. However, in the preliminary exploration, our simulations performed better with fewer iterations (in fact epochs). In modern NNs, researchers usually run the NN algorithms in many iterations, but that is partly due to the dropout regularization technique. We did not use drop-out (and L2) regularization in the final simulations, as the preliminary results did not confirm dropout as promising as L_1 regularization.

Further, we utilized NNs to learn the underlying relationships between the covariates and the outcome and treatment by targeting the relevant features through regularization and joint modeling of the treatment and outcome. NNs with other structures that might target confounders have not been explored, nor have other Machine Learning algorithms such as tree-based models. The Gradient Boosting Machines (GBM) algorithm (Friedman, 2001) can be alternatively used to learn these non-linear relationships while targeting the right set of features. This is postponed to a future article.

Chapter 5

Multiply Robust Estimator Circumvents Hyperparameter Tuning of Neural Network Models in Causal Inference

5.1 Introduction

The estimation of ATE typically involves two steps. In the first step, a model is developed for the treatment and outcome. In the second step, predictions from the first step are used to estimate the ATE. There is a wealth of literature on various aspects of this process, including the types of estimators that can be used to calculate the causal effect (Heckman et al., 1998, Lunceford and Davidian, 2004, Rubin, 1976, van der Laan and Petersen, 2007), the models that are most appropriate to use in the first step (Chernozhukov et al., 2020, 2022c, Farrell et al., 2021, Hartford et al., 2017, Louizos et al., 2017), and the performance of causal effect estimators in cases where the models used in the first step converge slowly or are complex (Chernozhukov et al., 2018a, Van der Laan and Rose, 2011).

There are several estimators, such as IPW (Rubin, 1979), AIPW (Lunceford and Davidian, 2004), nAIPW (Rostami and Saarela, 2022a), and Robinson (Robinson, 1988) that can be used to estimate the ATE. In the first step of the process, multiple statistical or ML models can be trained with possibly many hyperparameter sets, while only a pair of models for the outcome and propensity score is needed. The choice of which models to use in the second step is often based on how well they fit the data in the first step, even though this does not necessarily result in an optimal estimation of ATE.

The out-of-sample predictive measures such as AUC of the treatment model and R^2 of the outcome model can be used. Rostami and Saarela (2022a) utilized the geometric average of AUC and R^2 as a criterion for selecting the best model. However, in causal inference, as the potential outcomes are not observed, and thus cannot be used in the training or selecting of the models, model selection in the first step may not result in an optimal estimation of the causal effect. The super learner method (Van der Laan et al., 2007b) which could be used to construct ensemble predictions for the treatment and outcome avoids selecting a pair of models as the final weighted average predictions will be consistent if one of the input models is so. However, the same challenge still holds for utilizing super learner which is estimating the weighted average coefficients using the criterion that the predicted values should be as close as possible to the observed data (in the validation dataset). Additionally, the model selection and ensemble prediction may be sensitive to the choice of the validation set used. If the validation set is not representative of the overall data distribution, they may perform poorly on out-of-sample data. Further, violation of the assumptions for using super learner such as independence of the data, and well-calibrated candidate algorithms (Van der Laan and Rose, 2011), can impact the performance of the causal parameter estimator. These drawbacks of the super learner motivate utilizing an estimator that circumvents the need for hyperparameter tuning or selecting any models.

The MR estimator is a method for estimating ATE that was introduced in the field of missing data research (Han and Wang, 2013). It can make use of the models developed in the first step which would eliminate the need for hyperparameter tuning. The MR estimator is a consistent estimator of ATE if any of the models for the propensity score or outcome are also consistent and does need the knowledge of which one is consistent. It is an empirical maximum likelihood estimator that requires the solution of a constrained convex optimization problem. However, due to numerical issues with this method, an alternative optimization method was proposed by Han (2014) that theoretically and numerically results in a multiply robust estimator of ATE. Wang (2019) utilized MR for causal inference when the treatment is binary and Naik et al. (2016) extended the MR to treatments with multiple categories. In both of these research, the models used in the first step are parametric models, and thus $\sqrt{n}$-consistent estimators. The current work is more similar to that of Wang (2019) in the sense that we consider a binary treatment, but the major difference is that we are interested in using NN-based prediction models (or other non-parametric ML models) rather than parametric ones in the first-step, which are consistent but with a slower rate than $\sqrt{n}$. For the MR estimator to be a $\sqrt{n}$-consistent estimator, at least one of the treatment or outcome models must also be a $\sqrt{n}$-consistent estimator. This can be a stringent requirement in practice if ML models are used in the first step. The AIPW and nAIPW estimators have the property that if the selected PS and outcome models converge as quickly as $n^{\frac{1}{4}}$, the estimator is $\sqrt{n}$-consistent, under certain regulatory conditions and with the use of cross-fitting (Chernozhukov et al., 2018a). However, unlike in the AIPW and nAIPW estimators, for the MR estimator, we do not need to have the knowledge of which model is $\sqrt{n}$-consistent. Thus, there is a trade-off between the properties of MR and AIPW, and nAIPW estimators - one property is gained at the expense of another. In cases where it is unclear which model is consistent, MR estimator is preferable.

In terms of inference on the MR estimator, Wang (2019) discussed and derived asymptotic normality and variance estimator. One limitation of their proposed estimator is that it requires the knowledge of which treatment and/or outcome models are consistent, which goes against the purpose of using MR instead of AIPW. Therefore, the authors recommended using bootstrapping to estimate the standard error of the MR estimator.

The main goal of this paper is to compare the performance of the MR estimator to the normalized nAIPW estimator when the first-step models have a slow convergence rate (such as when neural networks are used, as in Farrell et al. (2021)). The second objective is to develop an asymptotic variance estimator for MR that does not require knowledge of the consistent first-step models. The comparisons are performed for both low- and high-dimensional scenarios.

Through simulations, we will also use the same MR estimator when the first-step models are neural network models (which are slower to converge than parametric models, at a rate of around $n^{\frac{1}{4}}$ (Farrell et al., 2021)). We will use simulations to study the performance of MR in terms of bias, variance, and root mean square error in the presence of instrumental variables (as defined in Angrist and Pischke, 2008) and confounder variables in the data.

In this paper, we will study the impact of the number of trained models in the first step on the performance of the MR estimator, both with and without the use of oracle models for the propensity score and outcome. We will also compare MR to the normalized AIPW estimator (Rostami and Saarela, 2022a,b), using selection criteria based on AUC, R^2, and the geometric average of these prediction measures, *geo*. Additionally, we will provide a detailed mathematical derivation of the MR estimator and its multiple robustness property for slower rates than $\sqrt{n}$. In our theorems, we assume general functional forms for the outcome and treatment predictors, rather than parametric forms, and we prove n^r consistency with $r \leq \frac{1}{2}$. We also study theoretically and numerically the proposed asymptotic variance estimator for the MR estimator. The proof for the asymptotic normality and the derivation of the asymptotic variance estimator will also be included.

This paper is organized as follows. In Section 5.2, we define the notation, specify the problem setting and the causal parameter to be estimated, and provide a brief overview of the normalized AIPW estimator. In Section 5.3, we review the MR estimator and outline its theoretical properties. In Section 5.4, we describe our simulation scenarios and present the results in Section 5.4.1. We conclude the paper in Section 5.6 with a discussion of the results and future work. The proofs of the lemmas and theorems are provided in Appendix 7.3.

5.2 Background

Let $O = (O_1, O_2, ..., O_n)$ be IID data generated by a data generating process P, where O_i is a finite-dimensional random vector $O_i = (Y_i, A_i, W_i)$, with Y as the outcome, A as the treatment and $W = (X_c, X_y, X_{iv}, X_{irr})$ the covariates, where we assume $A \sim Bernoulli(\pi)$ with $\pi = f_1(X_c, X_{iv})$, and $Y = f_2(A, X_c, X_y)$, for some random functions f_1, f_2. The covariates are partitioned so that X_c is the set of confounders, X_{iv} is the set of instrumental variables, X_y is the set of y-predictors (independent of the treatment), and X_{irr} is a set of given noise or irrelevant inputs (Figure 5.1). P represents the true joint probability distribution of O and $\hat{P}_n$ is its sample version. Let $\hat{P}_n$ be any distribution of (Y, A, W) such that the marginal distribution of W is given by its empirical distribution, and the conditional distribution of $Y \mid (A = a, W)$ has a conditional mean equal to a given estimator $\mathbb{E}[Y \mid A = a, W]$. Let Q^1 represent the expected outcome of the treated group, where $Q^1 := Q(1, W) = \mathbb{E}[Y \mid A = 1, W]$, and Q^0 represent the expected outcome of the untreated group, where $Q^0 := Q(0, W) = \mathbb{E}[Y \mid A = 0, W]$. Also, let $g(W)$ be the propensity score, defined as $g(W) = \mathbb{E}[A \mid W]$. All expectations are taken with respect to P. The symbol ˆ on the population-level quantities indicates the corresponding finite sample estimator.

5.2.1 Problem Setup and Assumptions

The fundamental problem of causal inference is that individual-level causality cannot be identified since each person can experience only one value of A. Thus, it is customary to focus on estimating a population-level causal parameter, in this particular ATE,

$$\beta_{ATE} = \mathbb{E}[Y^1 - Y^0] = \mathbb{E}[\mathbb{E}[Y^1 - Y^0 \mid W]]. \tag{5.1}$$

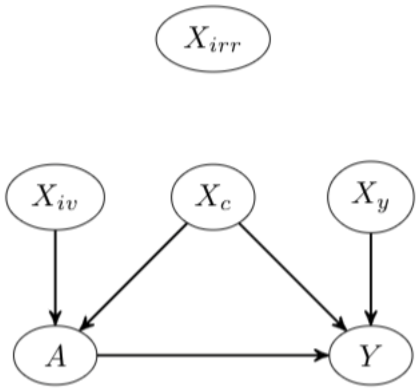


Figure 5.1: The causal relationship between A and y in the presence of other factors in an observational setting.

To ensure the parameter's identifiability, several conditions must be satisfied. The first condition pertains to conditional independence, or unconfoundedness, which asserts that, when considering the confounding variables, the potential outcomes remain unrelated to the treatment assignments ($Y^0, Y^1 \perp A \mid W$). The second condition, known as positivity, requires that the assignment of treatment groups is not entirely deterministic ($0 < Pr(A = 1 \mid W) < 1$). The third condition is consistency, which stipulates that the observed outcomes are equivalent to their respective potential outcomes ($Y^A = y$). Additionally, other modeling assumptions include the temporal order (i.e., covariates W are measured before treatment), IID subjects.

5.2.2 Doubly Robust Estimator

A variety of estimators have been proposed for estimating ATE, including AIPW and nAIPW. These are given by

$$\begin{aligned}\hat{\beta}_{AIPW} &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{A_i(Y_i-\hat{Q}_i^1)}{\hat{g}_i} - \frac{(1-A_i)(Y_i-\hat{Q}_i^0)}{1-\hat{g}_i}\Big) + \frac{1}{n}\sum_{i=1}^{n}\hat{Q}_i^1 - \hat{Q}_i^0, \text{ and}\\ \hat{\beta}_{nAIPW} &= \sum_{i=1}^{n}\Big(\frac{A_i(Y_i-\hat{Q}_i^1)w_i^{(1)}}{\sum_{j=1}^{n}A_j w_j^{(1)}} - \frac{(1-A_i)(Y_i-\hat{Q}_i^0)w_i^{(0)}}{\sum_{j=1}^{n}(1-A_j)w_j^{(0)}}\Big) + \frac{1}{n}\sum_{i=1}^{n}\hat{Q}_i^1 - \hat{Q}_i^0,\end{aligned} \tag{5.2}$$

where $\hat{Q}_i^k = \hat{Q}(k, W_i) = \hat{\mathbb{E}}[Y_i \mid A_i = k, W_i]$, and $\hat{g}_i = \hat{\mathbb{E}}[A_i \mid W_i]$. (Rostami and Saarela, 2022a) demonstrated that nAIPW should be favored over AIPW in the datasets where strong confounders and instrumental variables exist and complex algorithms such as NNs are used in the first step.

In the second step of the estimation procedure, predictions of the treatment (i.e. propensity score) and the outcome $\hat{Q}_i^k$, $k = 0, 1$, are inserted in the estimators (5.2). Generalized Linear Models (GLM), any relevant Machine Learning algorithm such as tree-based algorithms and their ensemble (Friedman et al., 2001), SuperLearner (Van der Laan et al., 2007b), or Neural Network-based models (such as ours) can be applied as prediction models for the first step prediction task. We will use jNN and dNN, proposed by Rostami and Saarela (2022b). jNN and dNN refer to NN architectures to estimate the outcome and treatment; jNN is one single NN with both treatment and outcome as the output nodes, and dNN refers to two separate NNs that predict outcome and treatment. We note that the theoretical properties of the AIPW and nAIPW are not guaranteed to hold if certain assumptions such as the assumption that first-step models are $\sqrt{n}$-consistent or assumptions such as those outlined by Chernozhukov et al. (2018a) of the first step models do not hold.

5.3 Multiple Robust Estimator

The multiply robust estimator of ATE (5.1) is

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\beta}_i^1 - \hat{\beta}_i^0, \tag{5.3}$$

such that

$$\begin{aligned} \hat{\beta}_i^1 &= \frac{A_i Y_i}{n_1(1+\hat{\gamma}^T \hat{C}_i^1)}, \\ \hat{\beta}_i^0 &= \frac{(1-A_i) Y_i}{n_0(1+\hat{\rho}^T \hat{C}_i^0)}, \end{aligned} \tag{5.4}$$

where n_1 and n_0 is the size of treatment and control groups, and

$$\hat{C}_i^1 = \begin{pmatrix} \hat{g}_i^1 - \hat{\mathbb{E}}\hat{g}^1 \\ ... \\ \hat{g}_i^K - \hat{\mathbb{E}}\hat{g}^K \\ \hat{Q}_i^1(1) - \hat{\mathbb{E}}\hat{Q}^1(1) \\ ... \\ \hat{Q}_i^L(1) - \hat{\mathbb{E}}\hat{Q}^L(1) \end{pmatrix}, \hat{C}_i^0 = \begin{pmatrix} \hat{\mathbb{E}}\hat{g}^1 - \hat{g}_i^1 \\ ... \\ \hat{\mathbb{E}}\hat{g}^K - \hat{g}_i^K \\ \hat{Q}_i^1(0) - \hat{\mathbb{E}}\hat{Q}^1(0) \\ ... \\ \hat{Q}_i^L(0) - \hat{\mathbb{E}}\hat{Q}^L(0) \end{pmatrix} \tag{5.5}$$

with $\hat{\mathbb{E}}\hat{g}^k$ and $\hat{\mathbb{E}}\hat{Q}^l(j)$ representing the average of the $\hat{g}^k$ and $\hat{Q}^l(j)$ over all observations ($i = 1, ..., n$), and $\gamma = (\gamma_1, ..., \gamma_{K+L})^T$ and $\rho = (\rho_1, ..., \rho_{K+L})^T$ are $K + L$ dimensional vectors that satisfy

$$\sum_{i=1}^{n} \frac{A_i \hat{C}_i^1}{1+\hat{\gamma}^T \hat{C}_i^1} = 0, \quad \sum_{i=1}^{n} \frac{(1-A_i) \hat{C}_i^0}{1+\hat{\rho}^T \hat{C}_i^0} = 0, \tag{5.6}$$

with the propensity score and outcome estimators

$$\begin{aligned} &\{\hat{g}^1, \hat{g}^2, ..., \hat{g}^K\} \\ &\{\hat{Q}^1(j), \hat{Q}^2(j), ..., \hat{Q}^L(j)\}. \end{aligned} \tag{5.7}$$

This way of defining the MR estimator maximizes the joint probability of W, Y, given $A = 1$.

Lemma 5.3.1. *The estimator* (5.4) *maximizes the joint probability of* W, Y*, given* $A = 1$.

Lemma 5.3.2. *Given that one of the propensity score estimators is a consistent estimator of the true propensity score, the MR estimator is consistent.*

Lemma 5.3.3. *If one of the outcome regression models is correctly specified, then the MR estimator is consistent.*

The proofs of these lemmas are presented in the Appendix. The lemmas do not assume that we know which models are consistent, or consistent with a faster rate. The rate of consistency of MR depends on the fastest rate of the models used in the estimator. Now assuming that one of the above lemmas holds true, we have

Theorem 5.3.4. *Given that either one of the propensity score estimators is a consistent estimator of the true propensity score or one of the outcome regression estimators of the outcome, the MR estimator is consistent.*

Proof. By Lemmas 5.3.2 and 5.3.3. □

Theorem 5.3.5. *MR estimator is a solution to an estimating equations system*

$$\begin{aligned}&\sum_{i=1}^{n} A_i u_i (Y_i - \beta^1) = 0,\\ &\sum_{i=1}^{n} (1 - A_i) v_i (Y_i - \beta^0) = 0,\end{aligned} \tag{5.8}$$

where $\beta = \beta^1 - \beta^0$, $u_i = f_{Y^1,W}(y_i, w_i)$, and $v_i = f_{Y^0,W}(y_i, w_i)$. Given that one of the outcome or PS models is $\sqrt{n}$-consistent, the MR estimator is $\sqrt{n}$-consistent and asymptotically normal, that is,

$$(\hat{\beta}_{MR} - \beta) \xrightarrow{d} N(0, \frac{\hat{\sigma}^2}{n}), \tag{5.9}$$

with

$$\hat{\sigma}^2 = n \sum_{i=1}^{n} \Big(A_i \hat{u}_i^2 (Y_i - \beta_{MR}^1)^2 + (1 - A_i) \hat{v}_i^2 (Y_i - \beta_{MR}^0)^2 \Big), \tag{5.10}$$

where

$$\begin{aligned}\hat{u}_i =& \frac{1}{1 + \gamma^T C_i^1} / \Big(\sum_{i=1}^{n} \frac{A_i}{(1 + \gamma^T C_i^1)} \Big), \qquad \sum_{j=1}^{n} \frac{A_j C_j^{1,k}}{1 + \gamma^T C_j^1} = 0,\\ \hat{v}_i =& \frac{1}{1 + \rho^T C_i^0} / \Big(\sum_{i=1}^{n} \frac{1 - A_i}{(1 + \rho^T C_i^0)} \Big), \qquad \sum_{j=1}^{n} \frac{(1 - A_j) C_j^{0,k}}{1 + \rho^T C_j^0} = 0,\end{aligned} \tag{5.11}$$

for $k = 1, ..., K + L$.

Remark 5.3.1. *The MR estimating equations* (5.8) *belong to a larger class of estimating equations*

$$\begin{aligned}&\sum_{i=1}^{n} A_i u_i (Y_i - \beta^1) - \eta_1 (1 - A_i u_i) = 0,\\ &\sum_{i=1}^{n} (1 - A_i) v_i (Y_i - \beta^0) + \eta_0 (1 - (1 - A_i) v_i) = 0,\end{aligned} \tag{5.12}$$

where $\eta_1 = \eta_0 = 0$. This general class is similar to the general class of estimating equations that IPW, nIPW, and AIPW estimators (Lunceford and Davidian (2004), equation 10). The difference between these two classes is that u_i (and similarly v_i) acts as the reciprocal of the propensity score. As a consequence, the MR asymptotic variance estimator (5.10) *matches with that of nIPW where $1/g_i$ is replaced by u_i and $1/(1 - g_i)$ is replaced by v_i.*

5.4 Simulations

Our simulation study aims to achieve several objectives. First, to demonstrate the impact of the number of first-step models included in the second step on the MR estimator. Second, to compare the performance of MR and nAIPW in terms of bias, variance, and Mean Square Error. Third, to visually inspect the consistency of the MR estimator, which is sufficient to demonstrate its asymptotic unbiasedness. Fourth, to contrast the proposed asymptotic variance estimator with the Monte Carlo variance of the MR estimator. Finally, to examine how the dimension of the covariate space affects the distribution of the MR estimator and its asymptotic variance estimator in a scenario where all causal inference assumptions are met (Section 5.2.1).

In the simulation study, we generated 100 independent samples to compare the prediction methods jNN, and dNN by inserting their predictions in the nAIPW (causal) estimators (5.2), or utilizing them to calculate MR. For this purpose, we allow the underlying relationship between the inputs and the output to be non-linear. We fixed the sample sizes to be $n = 750$ and $n = 7500$, with the number of covariates $p = 32$ and $p = 300$, respectively. We generate the confounding, instrumental variable, y-predictors, and noise/irrelevant inputs independently with the same sizes $\#X_c = \#X_{iv} = \#X_y = \#X_{irr} = 8, 75$ (summing to 32 and 300) from the Multivariate Normal (MVN) distribution as each set of covariates $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$, with $\Sigma_{kj} = \rho^{j-k}$ and $\rho = 0.5$. Let $\beta = 1$. The models to generate the treatment assignment and outcome were specified as

$$\begin{aligned} &A \sim Ber(\frac{1}{1+e^{-pr}}), \text{with } pr = f_a(X_c)\gamma_c + g_a(X_{iv})\gamma_{iv}, \\ &y = 3 + A + f_y(X_c)\gamma_c' + g_y(X_y)\gamma_y + \epsilon, \end{aligned} \tag{5.13}$$

The functions f_a, g_a, f_y, g_y select 30% of the columns and apply interactions and non-linear functions listed below (5.14). The strength of instrumental variable and confounding effects were chosen as $\gamma_c, \gamma_c', \gamma_y \sim Unif(r_1, r_2)$ and $\gamma_{iv} \sim Unif(r_3, r_4)$ where $(r_1 = r_3 = 0)$, and $(r_2 = r_4 = 0.25)$. The non-linearities for each pair of covariates are randomly selected among the following functions:

$$\begin{aligned} l(x_1, x_2) &= e^{\frac{x_1 x_2}{2}} \\ l(x_1, x_2) &= \frac{x_1}{1+e^{x_2}} \\ l(x_1, x_2) &= (\frac{x_1 x_2}{10} + 2)^3 \\ l(x_1, x_2) &= (x_1 + x_2 + 3)^2 \\ l(x_1, x_2) &= g(x_1) \times h(x_2) \end{aligned} \tag{5.14}$$

where $g(x) = -2I(x \leq -1) - I(-1 \leq x \leq 0) + I(0 \leq x \leq 2) + 3I(x \geq 2)$, and $h(x) = -5I(x \leq 0) - 2I(0 \leq x \leq 1) + 3I(x \geq 1)$, or $g(x) = I(x \geq 0)$, and $h(x) = I(x \geq 1)$.

In order to avoid imbalanced treatment groups, the generated datasets in which the number of subjects in the treatment or control group is less than 25% were ignored and new ones were generated. Also, in order to examine the accuracy of the proposed asymptotic variance estimator, we have considered a low-confounding scenario with $(r_2 = r_4 = 0.01)$ for two cases of low and high dimensional covariate space. In this side experiment, the only model included in the MR estimator is the Oracle model.

The networks' activation function is ReLU, with 3 hidden layers as large as the input size (p), with L_1 regularization and batch size equal to $3 * p$ and 200 epochs. The Adam optimizer (Kingma and Ba,

2014) with a learning rate 0.01 and momentum 0.95 were used to estimate the network's parameters. The total number of first-step trained models is 32: 2 types of NN models (jNN and dNN), 2 NN depth variations ($[q, p, q]$ and $[p, p, p]$, where $q = p/4$), 2 L_1 regularization strengths $(0.01, 0.1)$, 4 targeted L_1 regularization strengths $(0, 0.1, 0.3, 0.7)$. These scenarios are exactly the same as the ones in the articles Rostami and Saarela (2022a) and Rostami and Saarela (2022b).

As in practice, to calculate AIPW/nAIPW, the true models are unknown and counterfactuals are unobserved, the prediction measures of the outcome and treatment should be used to choose the best model by the K-fold cross-validation, to insert into the estimators in step 2. R^2 and AUC each provide insights into the outcome and treatment models, respectively, but in our framework, both models should be satisfactory. To measure the goodness of the prediction models (jNN and dNN) for causal inference purposes, we define and utilize a statistic which is a compromise (geometric average) between R^2 and AUC, here referred to as *geo*,

$$geo(R, D) = \sqrt[3]{R^2 \times D \times (1 - D)}, \tag{5.15}$$

where $D = 2(AUC - 0.5)$, the Somers' D index. The *geo* measure was not utilized in the optimization process (i.e. training the neural networks or cross-validation), and is rather introduced here to observe if the compromise between R^2 and AUC agrees with the models that capture more confounders than IVs.

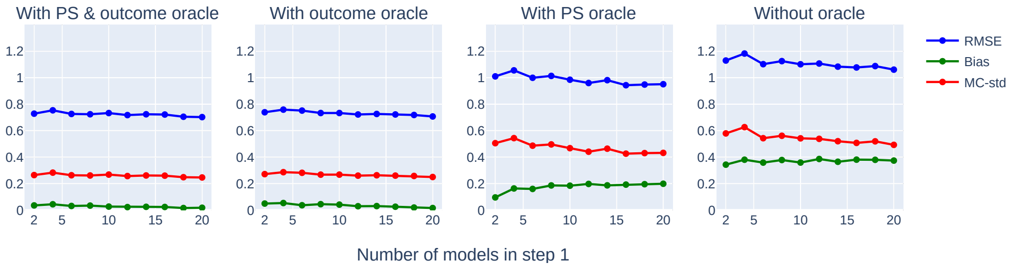
To compare the results with the truth in some of the results, we have used the true, or oracle, treatment, and outcome models Also, as we have 100 independent estimations for all the estimators and asymptotic variance, we can use bootstraping to estimate confidence intervals for them.

5.4.1 Results

Figures 5.2 and 5.3 illustrate the performance of MR when the number of ML models trained in the first step increases by either including or excluding the single true model (the oracle model which is $\sqrt{n}$-consistent). Each dot is calculated by averaging 10 estimators using 10 models randomly selected among all the 32 models with different hyperparameter sets, outlined in the previous section. It can be observed that for both small and large sample size settings, the performance (MR's bias, variance, or RMSE) does not significantly change irrespective of whether the predictions of the dNN, jNN, or their combination are used. As expected, a significant drop is observed by comparing the scenarios where the true outcome model is included in the first step models. When the true outcome model is not present but the true PS model is present, we observe a small improvement over the scenario where no oracle model is included in the first step models.

Figures 5.4 compares the performance of MR with the nAIPW which is a doubly robust estimator needing a choice of the best-performing first-step model. In the first 2 scenarios, the first-step prediction models (dNN or jNN) that minimize the RMSE are chosen; in the second two scenarios, the first-step prediction models (dNN or jNN) that maximize the *geo* are chosen. The third and fourth pairs correspond to maximizing R^2 and AUC, respectively. The leftmost two settings are unrealistic, as we do not know the bias and variance of the estimator in real problems, but it can provide an ideal case with which other cases can be compared. When AUC is used, neither jNN nor dNN shows favorable results in both small and large datasets. Also, it can be seen that no matter which of R^2, AUC, or *geo* is used, it is better not to use the dNN predictions, especially for the small sample size case. On the other hand, predictions

Figure 5.2: Comparison of MR estimator when different number of first step prediction models are fed into the estimator, and dataset is not large (n=750). In the first steps all NN models from jNN and dNN are used.



from jNN models when R^2 or *geo* is used provide the best performance among the non-MR methods for both large and small sample sizes. Another observation is that MR shows an inferior performance as compared to nAIPW when jNN is used in the first model and the hyperparameters are chosen based on the best *geo*.

Comparing variations of MR and normalized AIPW estimation, it can be seen that MR is not superior to the normalized AIPW, when R^2 or *geo* is used.

It is difficult to illustrate the consistency of an estimator through simulations. However, as asymptotic unbiasedness and asymptotic efficiency imply consistency (Hogg and Craig, 1995), an illustration of the asymptotic unbiasedness and efficiency is equivalent to an illustration of consistency of the estimator. Figure 5.5 is a visualization of the asymptotic unbiasedness and asymptotic efficiency of the MR estimator with NN-type predictions. We note that the first step models are NNs which are consistent but slower than $n^{-\frac{1}{2}}$ rate, but the MR estimator provably and experimentally is consistent (even if the rate is slow.)

The proposed asymptotic standard error of the MR estimator is plotted against the MR estimator's Monte Carlo standard deviation in Figure 5.6. It can be seen that when the outcome oracle model is among the first step models in calculating the MR estimator, the proposed asymptotic standard error is accurate. However, when the Oracle model is not used in MR estimator, the proposed asymptotic standard error is severely biased in the case of the high dimensional case. This can be due to the curse of dimensionality or violation of causal inference assumptions. To examine which one is the cause of this bias, we have plotted (Figure 5.7) the sampling distribution of the MR estimator and its standard error when neither of the causal inference assumptions (Section 5.2.1) is violated, for a high and a low dimensional scenario. It can be seen that the variance estimator is still severely biased in the high-dimensional scenario. Similar results (figures not presented here) were seen for other estimators such as nIPW and nAIPW.

5.5 Application: Food Insecurity and BMI

CCHS is a cross-sectional survey that collects information about the health status, healthcare usage, and factors influencing health among the Canadian population. The 2021 CCHS includes data from individuals aged 12 and older in all ten provinces and three territorial capitals. It excludes individuals living on reserves, other Aboriginal settlements in provinces, and some smaller groups, which together make up less than 3% of the Canadian population aged 12 and above. Common survey modules cover

Figure 5.3: Comparison of MR estimator when a different number of first step prediction models are fed into the estimator, and the dataset is large (n=7500). In the first steps all NN models from jNN and dNN are used.

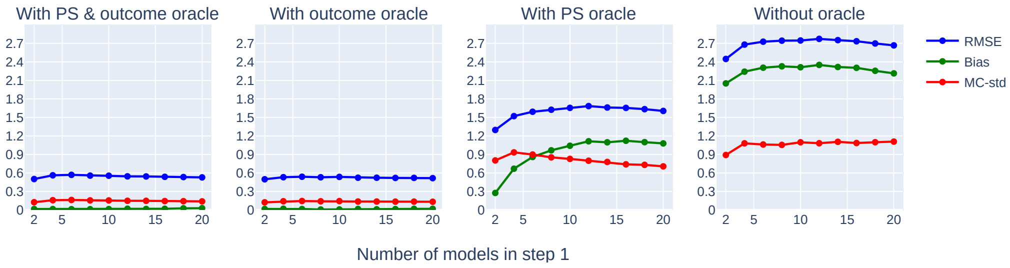


Figure 5.4: Comparison of nAIPW and MR in terms of RMSE, bias, and standard error measurements where ($r_2 = r_4 = 0.25$). The three most right scenarios on the x-axis are MR estimators. The other scenarios on the x-axis are for the nAIPW where one of *geo*, *AUC*, and R^2 criteria have been used to choose the best first-step model among dNN or jNN models to insert into nAIPW.

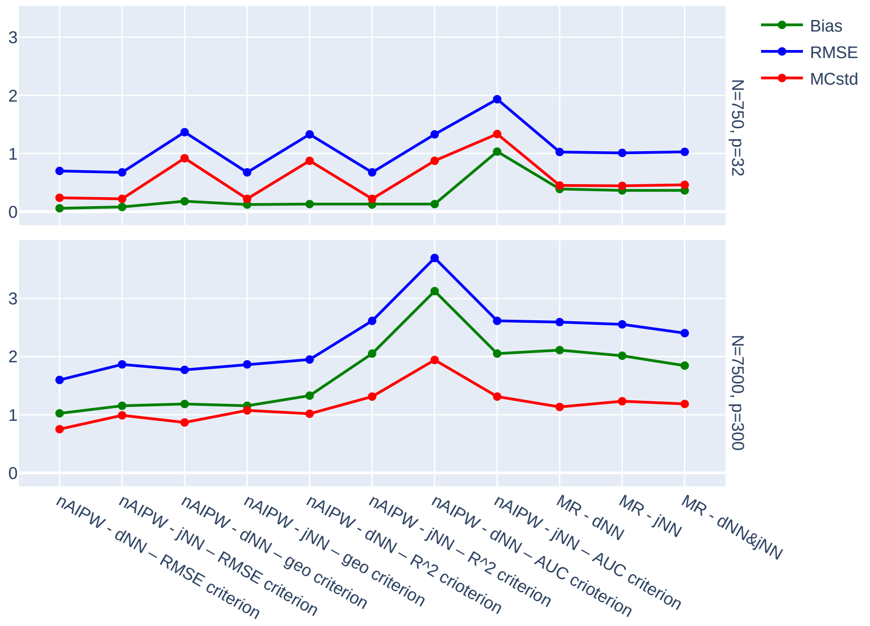


Figure 5.5: Asymptotic unbiasedness and asymptotic efficiency of MR. The predictions from jNN and dNN are used.

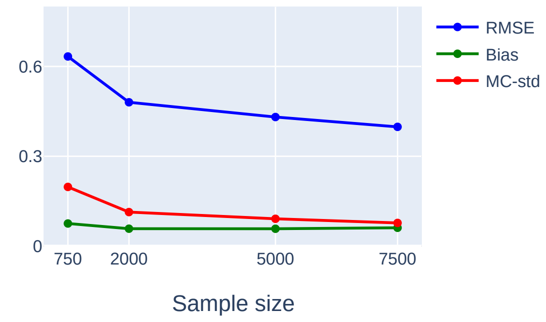


Figure 5.6: Comparison of Monte Carlo standard deviation and asymptotic standard errors.

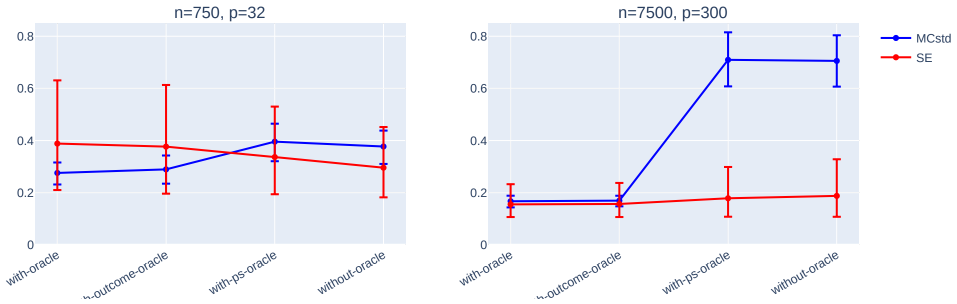
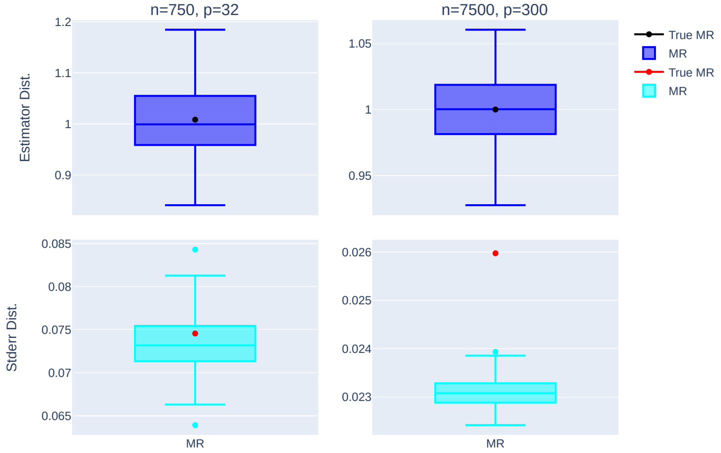


Figure 5.7: The distribution of the MR estimator and the proposed asymptotic standard error for a low confounding scenario where $r_3 = r_4 = .01$. The figure shows that the curse of dimensionality greatly impacts the variance estimator, even if the causal inference assumptions (e.g. positivity assumption) are not violated.



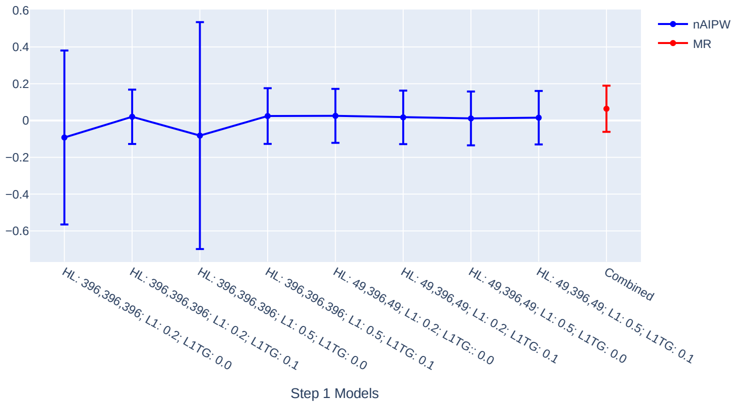


Figure 5.8: The ATE estimates and their asymptotically calculated 95% confidence intervals with nAIPW and MR methods.

topics like general health, chronic illnesses, smoking, and alcohol use. For the 2021 cycle, additional topics such as food security, home care, sedentary behavior, and depression were also addressed. The survey also collects information on respondent characteristics, including employment, income, and socio-demographics.

In this study, we analyze the CCHS dataset to explore the cause-and-effect relationship between food insecurity and BMI. We consider other data collected by the CCHS that may contain potential variables influencing the results, as well as predictors and instrumental variables. Since the data come from a survey, we employ specialized techniques like resampling or bootstrap methods to estimate standard errors accurately. In this particular analysis, we focus on individuals aged 18 to 65 years to minimize data variability.

Figure 5.8 compares the ATE estimates and their 95% asymptotic confidence intervals using the nAIPW and MR methods. The nAIPW method is calculated for 8 different jNN hyperparameter scenarios, while the MR method uses a combination of these scenarios. The true ATE is unknown, but the MR method appears to be more stable than the nAIPW estimator across different hyperparameter settings. The confidence interval of the MR estimator is based on the proposed asymptotic variance estimator $\hat{\sigma}^2$ (5.10).

The code to implement the AIPW and MR methods in Python is available on the github page mentioned earlier. The dataset should include a continuous outcome, a binary treatment, and X-factors. The algorithms require predictions from multiple models for the triple $(Q(1), Q(0), g)$. Rows with missing values will be automatically removed.

5.6 Discussion

The doubly robust estimators, such as AIPW, use one set of models for the outcome and propensity scores. If multiple models are developed in the first step, one set of models must be chosen (hard-selection) or an ensembling method such as the super learner (soft-selection) can be used to combine the predictions and insert them into the doubly robust estimator. Cross-validation is typically used to select the “best” outcome and propensity score models based on prediction performance metrics or to ensemble the predictions. In this chapter, we studied the MR estimator, which uses all trained models from the first step directly that does not require cross-validation to choose among the first-step models.

A new class of estimating equations (5.3.1) has been introduced, which bears resemblance to the one proposed in Lunceford and Davidian (2004). Both these classes of estimating equations allow for the generation of different ATE estimators by setting two parameters η_0 and η_1 to different values. For example, in the class of estimating equations in Lunceford and Davidian (2004), by setting $\eta_0 = \eta_1 = 0$, nIPW is generated, and other values give AIPW estimators, which uses 2 sets of models for the treatment and outcome. Similarly, in our proposed class, MR can be obtained by setting $\eta_0 = \eta_1 = 0$. In this sense, MR and nIPW are similar which motivates exploring other estimators that have similar forms to AIPW or nAIPW and improve MR.

A desirable characteristic of AIPW-type estimators is their potential for asymptotic normality, provided that the combined rate of convergence of the two models used in the first step is $\frac{1}{2}$, subject to certain assumptions. To the best of our knowledge, there is no existing research on the MR estimator to investigate whether it shares this property. Nonetheless, it is possible to create other estimators that belong to our proposed class of estimators that exhibit this characteristic.

The MR method has previously been used in missing data and causal inference studies where the first-step models are parametric. In this chapter, we used NN predictions to estimate MR and proved that the estimator is consistent if one of the outcome or propensity score models is consistent. We also investigated through simulations how the number of trained models in the first step affects the bias-variance trade-off of the MR estimator. We compared the MR method with the nAIPW method and found that the results slightly favor the nAIPW method in terms of bias, MC variance, and RMSE.

It has been previously shown that if one of the trained propensity score models is $\sqrt{n}$-consistent, the MR estimator is asymptotically normal. For those proposed methods, however, model selection must be performed in the first step to calculate its asymptotic variance, which defeats the purpose of using MR without selecting a pair of trained models. We proposed an alternative variance estimator that does not require model selection. Theoretically, if the first step models are $\sqrt{n}$-consistent, the MR estimator is asymptotically normal with our proposed asymptotic variance estimator. The simulations showed that our proposed variance estimator performs satisfactorily for the low-dimensional case. For the high-dimensional case, the estimator does not perform well. Based on our research (Figures 5.6 and 5.7), it appears that the primary factor responsible for reducing the accuracy of the asymptotic variance estimator is the dimensionality of the adjusting factors (including confounders). Based on some side experiments and simulations (results not shown), we conjecture that the asymptotic variance estimator is affected by a bias up to a factor of approximately $\sqrt{\frac{p}{n}}$, and further investigation and theoretical analysis are required to either support this claim or find an alternative solution. In addition, further research is required to investigate the appropriateness of our proposed asymptotic variance estimator in the case of first-step models that are slower than $\sqrt{n}$.

Chapter 6

Discussion and Conclusion

In this thesis, we have argued that the use of NNs in causal inference can be beneficial for addressing non-linear relationships between confounders and treatment and outcome, but it comes with challenges.

In Chapter 3, we demonstrated that without regularization techniques, NNs can provide a near-perfect prediction for the treatment (i.e. extreme propensity scores) or overfit the data, leading to failure in causal estimation and inference. The 'normalization' of the AIPW estimator (i.e., nAIPW) was shown to mitigate these issues. The nAIPW estimator may not perform better than AIPW when the empirical positivity assumption is not violated, but it can perform better when the assumption is violated.

The condition under which nAIPW can outperform AIPW is when for an exposed subject ($A_j = 1$) the propensity score is extremely close to 0; the closeness depends on how much the outcome prediction for this subject is close to the actual outcome ($\frac{y_j - Q_j}{g_j}$). Similarly for an unexposed subject. Under different circumstances, both estimators yield similar results, with AIPW being the preferred choice. The occurrence of extreme propensity score values is often attributed to the utilization of complex models that may overfit the data. Alternatively, these values could be the outcome of strong confounding or instrumental variables in the data, which the model employs to produce values near 1 or 0.

However, it was observed that the asymptotic standard errors of AIPW and nAIPW when using complex machine learning algorithms, such as NNs, are not reliable and underestimate the calculated Monte Carlo standard deviations. The asymptotic distributions of the estimators are also not symmetric, but nAIPW is slightly better than AIPW in that regard. This suggests that normalization alone cannot fully address the issue of empirical violation of positivity. In fact, NNs without heavier regularization can fail in causal estimation and inference because they are not targeted and are not directly designed for these tasks. The architecture of NN should target the confounders and limit the strength of predictors to prevent extreme propensity scores. These two ideas, targeted regularization, and the NN's targeted learning were explored in Chapter 4.

In Chapter 4, we considered two types of NN architectures, jNN and dNN. The former follows the idea of including the treatment and outcome on the output layer and jointly predicting them for possibly targeting the confounders. dNN, on the other hand, uses two separate NNs to predict the outcome and treatment. In both architectures, targeted L_1 regularization, denoted as L_{1TG}, has been embedded. In fact, the former contains both aforementioned ideas while the latter contains only one of them. Simulation results showed that the jNN architecture is more stable and has a smaller RMSE in small sample datasets compared to dNN.

The techniques introduced in this chapter, namely targeted regularization and joint modeling of the outcome and treatment, enhance the performance of the naive double NN model (utilized in Chapter 3) in comparison to when these techniques are not utilized. This improvement is especially notable when there are strong predictors available for modeling both the treatment and outcome variables. When employing NNs in the two-step estimation of ATE without performing feature selection and utilizing all input covariates, it is likely that extreme propensity scores will be generated, making accurate estimates challenging for both AIPW and nAIPW methods. Therefore, it is recommended to incorporate the proposed techniques when fitting NN models to the data, particularly when there is evidence of high-quality and highly predictive covariates.

A general challenge in the 2-step methods for estimating ATE, which is exacerbated by training jNN and dNN with new hyperparameters, is that there are many models to choose from in the second step of the estimation. In Chapter 4, we used a geometric average of R^2 and AUC to select the final set of hyperparameters which gives an outcome model and a propensity score model for the second step. However, using cross-validation and optimizing measures such as R^2 and AUC are sub-optimal because, in the optimization, the counterfactual outcomes are considered. In fact, the goal of choosing the best ML algorithm in terms of AUC or R^2 is to find the hyperparameter set that minimizes the observed loss, which is the distance between observed and predicted values, while in an ideal scenario, the goal must be to minimize the error between the true counterfactual outcomes and predicted counterfactual outcomes. We do not propose a solution to optimize the process based on the counterfactual outcomes, but we can circumvent the hyperparameter optimization using AUC or R^2, or their geometric average. This is investigated in Chapter 5.

In Chapter 5, we used the MR estimator which provides an alternative approach to the hyperparameter selection. MR combines all the predictions from the trained models and forms an estimator that maximizes the empirical likelihood of observed data in the treated and untreated groups. The MR estimator with NN predictions is an n^r-consistent estimator of ATE even if only one of the propensity score models or outcome models is n^r-consistent. It turns out that the MR estimator is the solution to a general class of estimating equations that bears resemblance to the one proposed in Lunceford and Davidian (2004). Both these classes of estimating equations allow for the generation of different estimators by setting two parameters η_0 and η_1. For example, by setting η_0 and η_1 to specific values, IPW, nIPW, and AIPW estimators can be produced, with nIPW being obtained when $\eta_0 = \eta_1 = 0$. Similarly, in our proposed class, MR can be obtained by setting $\eta_0 = \eta_1 = 0$. This new estimating equation guarantees the asymptotic normal distribution of MR estimator if one of the treatment or outcome models used in the calculation of MR estimator is $\sqrt{n}$-consistent. The derived asymptotic standard error is different from the one in Wang (2019) or Han and Wang (2013) in the sense that, unlike their proposed asymptotic standard error, our proposed one does not require the knowledge of which of the used treatment or outcome models is consistent.

The MR estimator is particularly helpful when the first step models all perform similarly and there is no favorite model to select from, and also when the consistency of the selected model is questionable. In such cases, MR will use all the predictions, and given that there is at least one consistent treatment or outcome model among all the trained models, the MR estimator is consistent.

6.1 Limitations

In spite of various methods proposed to address the existing challenges throughout this thesis, it is important to acknowledge that there are certain limitations to this work that require further investigation.

6.1.1 Fundamental limitations

The primary constraint of this study is the assumption that the outcome is continuous, and the treatment is binary. Also, the thesis did not cover heavy tail outcomes or rare treatment scenarios that can cause structural positivity violations. Moreover, the ratio of the dimension of covariates to the size of the data was also considered to be fairly small which might not be the case in practice. Further, we assumed the data are IID and the results do not generalize to correlated data, and using NN predictions with correlated data requires further work.

Another fundamental limitation of this research is that the illustration of the proposed methods is done through simulation studies in multiple situations. Unfortunately, the considered situations may not represent real-world conditions. To address this issue, the relationship between food security and BMI was used which is a real problem, and a real dataset was used. However, it is still crucial to verify and observe the effectiveness of our proposed methods on many more datasets to confirm their benefits in practice.

The methods described in this thesis are valid under certain assumptions, such as the identifiability ones. The violation of the positivity assumption can be detected based on the observed data (for example via sensitivity analysis), however, we have not discussed any tests to verify the positivity and other assumptions.

In Section 2.6, we provided an overview of specific constraints that define the scope of this thesis ensuring a clear understanding of what is and isn't covered in the thesis. Here, we delve deeper into some of these aspects.

Our primary focus was on estimating ATE as opposed to CATE. A key distinction is that ATE does not require the consideration of effect modifications. In simpler terms, when calculating ATE, we average out interaction effects, making it unnecessary to account for them. It's worth noting that there is a separate body of research in machine learning that focuses on CATE, and NN architectures are specifically designed to handle effect modification in that context (for example see Shi et al., 2019). However, in this thesis, we did not include exposure-covariate interactions in our NN architectures.

Moreover, we use a two-step estimation method for ATE as the main causal effect parameter as opposed to other parameters. It's essential to recognize that the methods introduced in this thesis may or may not be applicable to other parameters, which would require further research for clarification.

This thesis concentrates exclusively on the implementation of NNs for ATE estimation, rather than employing and comparing to various other machine learning algorithms. This decision is based on several reasons: First, there is limited existing research on the application of NNs to ATE estimation. Second, NNs typically do not perform as well as other algorithms, such as XGBoost, when dealing with tabular data. NNs excel in revealing hidden nonlinearities in text and images, so our focus was on using these algorithms and enhancing them for causal inference.

We do not engage in feature selection for modeling the treatment or outcome variables. This is because NN optimization allows the incorporation of various regularization techniques that can perform a soft selection of features (Friedman et al., 2001). Additionally, in the ultimate goal of using text and images as proxies for confounding variables, feature selection is typically not carried out.

In the upcoming subsections, we will outline the limitations of the suggested approaches and potential paths to address these issues in future research.

6.1.2 Normalizing the weights

Normalization was introduced in Chapter 3 as a means to avoid extreme weight impacting the estimator. However, the normalized weights reduce the bias-corrected term (Section 3.4.5, and equation (3.13)) which could capture the effect of unseen confounders. In fact, the normalized terms might cause the estimator to be biased by not taking into account some of the confounding effects. This is why the nAIPW estimator cannot perform better than AIPW when empirical positivity is not violated.

Although normalization helps the bias and variance of the estimator in case of positivity violations, the asymptotic distribution is not symmetric and the derived asymptotic variance of the nAIPW estimator was observed to be inaccurate. More investigation is required to derive a more accurate asymptotic variance estimator. Other types of variance estimation such as bootstrapping is not an option in this context as training hundreds or thousands of ML algorithms is not realistic.

By viewing the scaled bias formula

$$\sqrt{n}(\hat{\beta}-\beta) = \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,P) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n) + \sqrt{n}(P_n-P)[\phi(O_i,\hat{P}_n)-\phi(O_i,P)] - \sqrt{n}R(P,\hat{P}_n),$$

we see that the last two terms on the right-hand side case cause the inaccuracy of the asymptotic standard errors. The remainder (fourth term) is sensitive to the positivity assumption and does not vanish if it is violated. This is explicitly mentioned in assumptions (3.17). Another term in the above formulation is the empirical process term (third term) which is related to the complexity of the first-step model. Separate simulation scenarios are required to investigate under what circumstances the asymptotic standard error of nAIPW and AIPW can be accurate, especially in non-obvious scenarios such as when the positivity assumption is near violated.

The normalized AIPW is the solution to an estimating equation (an M-estimator) which makes it possible to study its asymptotic properties. However, another similar normalized AIPW estimator that belongs to the class of estimating equations introduced in Lunceford and Davidian (2004), where $\eta_1 = -Q^1$ and $\eta_1 = Q^0$:

$$\sum_{i=1}^{n}\Big(\frac{A_i(y_i-\hat{Q}_i^1)w_i^{(1)}}{\sum_{j=1}^{n}A_jw_j^{(1)}} - \frac{(1-A_i)(y_i-\hat{Q}_i^0)w_i^{(0)}}{\sum_{j=1}^{n}(1-A_j)w_j^{(0)}}\Big) + \frac{1}{\sum_{i=1}^{n}A_iw_i^{(1)}}\sum_{i=1}^{n}Q_i^1 - \frac{1}{\sum_{i=1}^{n}(1-A_i)w_i^{(0)}}\sum_{i=1}^{n}Q_i^0,$$

The only difference between this estimator and nAIPW in Chapter 3 is that the second part of nAIPW has n, not the normalization sum. We did not study this estimator in this thesis.

Another approach to understanding nAIPW involves a method similar to TMLE. This method entails perturbing $\hat{P}_n$ such that the residual term in the AIPW estimator vanishes. This can be achieved by changing the first step outcome predictions so that ${Q^*}^j = \hat{Q}^j + \epsilon$, where j can be either 0 or 1. Now ϵ

can be estimated by replacing $\hat{Q}^j$ by ${Q^j}^*$:

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{A_i}{g_i}(Y_i - {Q^1}^*_i) - \frac{1-A_i}{1-g_i}(Y_i - {Q^0}^*_i)\Big) = 0.$$

Basic algebra shows that the new estimator $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} {Q^1}^*_i - {Q^0}^*_i$ is the same as to the nAIPW estimator. This observation shows that TMLE is essentially a weight normalization technique. For example, by the choice of submodel so that ${Q^1}^* = \hat{Q}^1 + \frac{\epsilon}{g}$ and ${Q^0}^* = \hat{Q}^0 + \frac{\epsilon}{1-g}$, the normalization term will be $\sum_{j=1}^{n} \frac{A_j}{g_j} / \sum_{j=1}^{n} \frac{A_j}{g_j^2}$ as opposed to $1/\frac{1}{n}\sum_{j=1}^{n}\frac{A_j}{g_j}$ in nAIPW. The introduction of non-linear perturbations to $\hat{Q}$, such as $\text{logit}({Q^1}^*) = \text{logit}(\hat{Q}^1) + \frac{\epsilon}{g}$, is analogous, although it is more challenging to derive a closed-form formula for $\hat{\epsilon}$. To gain a better understanding of this, examining the Taylor expansion of these non-linear functions might provide insight into how non-linear perturbations can lead to the normalization of AIPW.

Another limitation of this research is in the simulation experiments where we did not conduct a comparison between AIPW and nAIPW when one of the oracle propensity score or outcome models is accurately specified while the other is misspecified. This idea is not investigated here.

Normalization is not the only way of avoiding the extreme propensity score predictions affecting the ATE estimators. Other estimators that can avoid extreme weights can be defined and studied such as trimming propensity scores (Section 3.8).

6.1.3 Neural Network models

The next set of limitations of this study is related to the NN architectures. The jNN architecture limits the complexity of the outcome model (which might not be necessary) if the final hidden layer is shrunk to control the complexity of the treatment model (that can lead to extreme weights.) The NN architectures are flexible and this issue may be resolved by alternative architectural design. Also, future studies may examine the use of other regularization techniques such as early stopping, or a combination of regularizations.

In addition, the theoretical properties of AIPW and nAIPW when implementing targeted L_1 regularization and shared parameters in jNN architecture were not studied. Additionally, while the preliminary studies showed that other regularization methods like dropout or L_2 (i.e., ridge) are not suitable for ATE estimation, the reasons were not explored. Further research is needed to determine if these methods can be modified to improve their performance. Another regularization technique is early stopping which was not explored in this thesis.

Finally, although utilizing NNs in causal inference makes it possible to leverage unstructured data, such as text or image data the proxies for confounders, such data structures were not studied in this thesis. NNs have shown the best performance on structured and massive data, while other algorithms such as XGBoost (Chen and Guestrin, 2016) perform satisfactorily well on tabular data such as the ones used in this thesis. In order to use XGBoost, the automatic differentiation capability available in deep learning platforms can be utilized to adapt XGBoost algorithm for the prediction of the outcome and treatment models.

6.1.4 Multiple Robust estimator

The MR estimator was introduced to circumvent the need for hyperparameter tuning. Our version of MR is equivalent to nIPW when comparing our proposed general class of estimating equations with that introduced in Lunceford and Davidian (2004). In this thesis, other multiple robust estimators equivalent to AIPW-type estimators were not explored. Additionally, the simulations indicate that when the outcome oracle is incorporated into the MR estimator, its performance is significantly superior to when the propensity score oracle model is included. The latter scenario is also more sensitive to the number of included inaccurate outcome/propensity score models. These two observations suggest a methodology akin to AIPW and the MR estimator, where we employ aggregated metrics to identify the "best" trained outcome and propensity score models and subsequently insert them into the MR estimator. However, this particular approach was not explored in the present study and is reserved for investigation in a future simulation study.

The ATE estimators that solve certain M-estimators or estimating equations without nuisance parameters are easily proven to be consistent and asymptotically normal. But if the M-estimator involves unknown nuisance parameters, they must be estimated using the data. Provided the models used to estimate these nuisance parameters belong to the Donsker class under regulatory conditions, the asymptotic results hold. However, for non-Donsker models (such as NNs), Chernozhukov et al. (2018a) showed that orthogonal estimators are consistent and asymptotically normal under certain regulatory conditions, provided cross-fitting is used for estimation and the summation of the rate of consistency of the propensity score and outcome is $\frac{1}{2}$. As the estimating equations of the MR estimator involve the nuisance parameters u and v, and also as the MR estimator is not orthogonal and the NN models are not $\sqrt{n}$-consistent, the asymptotic theories presented in this thesis do not apply to it. While the MR estimator is inherently non-orthogonal, it might be "orthogonalized" by using the process described by Chernozhukov et al. (2018a) to transform non-orthogonal estimators into orthogonal ones. This option was not investigated in this research. Nonetheless, not being orthogonal does not rule out the asymptotic normality of the MR estimator. Further research is necessary to determine the conditions under which the MR estimator is asymptotically normal if non-Donsker models like NNs are used.

The AIPW-type estimators have an attractive quality of potentially achieving asymptotic normality, given that any pair of the outcome and treatment models used in the initial step have a combined convergence rate of $\frac{1}{2}$, based on specific assumptions. It is unclear if the MR estimator shares this property. However, it is possible that other estimators within the proposed class can be developed/discovered to demonstrate this property.

Additionally, the simulations indicate that the proposed asymptotic variance of MR has the curse of dimensionality, that is, it does not perform satisfactorily for high dimensional data (Section 5.6). Additional research is necessary to adjust this estimator or propose an alternative that performs well in high dimensions. More insight on how to improve the proposed MR variance estimator is presented in Chapter 6.

Finally, we did not study the circumstances under which IPW-type and AIPW-type estimators perform satisfactorily, but MR performs poorly. Exploring these circumstances would contribute to the enhancement of these estimators and provide data analysts with more clear guidance when choosing an ATE estimator in practical applications.

Bibliography

Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *arXiv:1706.05966*, 2017.

Joshua D Angrist and Jorn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press, Princeton, NJ, 2008.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Edvard Bakhitov and Amandeep Singh. Causal gradient boosting: Boosted instrumental variable regression. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 604–605, 2022.

Haakon Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Jonathan Baxter. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28(1):7–39, 1997.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Christopher M Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, NY, 2006.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, NY, 2004.

Matias Busso, John DiNardo, and Justin McCrary. New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, 96(5): 885–897, 2014.

Rich Caruana. Learning many related tasks at the same time with backpropagation. *Advances in neural information processing systems*, 7, 1994.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Xi Cheng, Bohdan Khomtchouk, Norman Matloff, and Pete Mohanty. Polynomial regression as an alternative to neural nets. *arXiv:1806.06850*, 2018.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018a.

Victor Chernozhukov, Whitney K. Newey, and James Robins. Double/de-biased machine learning using regularized Riesz representers. Technical Report CWP15/18, cemmap working paper, 2018b.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of Riesz representers. *arXiv:2101.00009*, 2020.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022a.

Victor Chernozhukov, Whitney Newey, Victor M Quintas-Martinez, and Vasilis Syrgkanis. Riesznet and forestRiesz: Automatic debiased machine learning with neural nets and random forests. In *Proceedings of the 39th International Conference on Machine Learning*, pages 3901–3914. PMLR, 2022b.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022c.

John Crofton and DA Mitchison. Streptomycin resistance in pulmonary tuberculosis. *British Medical Journal*, 2(4588):1009, 1948.

Ivan Diaz. Doubly robust estimators for the average treatment effect under positivity violations: introducing the e-score. *arXiv:1807.09148*, 2018.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics, New York, NY, 2001.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

Adam N. Glynn and Kevin M. Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56, 2010.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT Press Cambridge, Cambridge, MA, 2016.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Peisong Han. Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109(507):1159–1173, 2014.

Peisong Han and Lu Wang. Estimation with missing data: beyond double robustness. *Biometrika*, 100 (2):417–430, 2013.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. volume 70, pages 1414–1423. JMLR.org, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

A Bradford Hill. *Principles of medical statistics.* The Lancet, England, 1955.

Austin Bradford Hill. The environment and disease: association or causation? *Journal of the Royal Society of Medicine*, 108(1):32–37, 1965.

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022.

Robert V Hogg and Allen T Craig. Introduction to mathematical statistics. *Englewood Hills, New Jersey*, 1995.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Guang-Bin Huang, Lei Yu, and Yang Xiao. Regularization for deep learning. *arXiv:1507.05717*, 2015.

Kosuke Imai, Marc Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Guido W. Imbens and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, Cambridge, NY, 2015.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning,*, pages 448–456. PMLR, 2015.

Genevieve Jessiman-Perreault and Lynn McIntyre. Household food insecurity narrows the sex gap in five adverse mental health outcomes among canadian adults. *International journal of environmental research and public health*, 16(3):319, 2019.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33th International Conference on Machine Learning*, pages 3020–3029, 2016.

Cheng Ju, Richard Wyss, Jessica M Franklin, Sebastian Schneeweiss, Jenny Haggstrom, and Mark J van der Laan. Collaborative-controlled Lasso for constructing propensity score-based estimators in high-dimensional data. *Statistical Methods in Medical Research*, 28(4):1044–1063, 2019.

Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. *Statistical Causal Inferences and their Applications in Public Health Research*, pages 141–167, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Jiuyong Li, Saisai Ma, Thuc Le, Lin Liu, and Jixue Liu. Causal decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 29(2):257–271, 2017.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30:6446–6456, 2017.

Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6232–6240, 2017.

Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

Marina Adriana Mercioni and Stefan Holban. The most used activation functions: classic versus current. In *Proceedings of the 2020 International Conference on Development and Application Systems (DAS)*, pages 141–145. IEEE, 2020.

Niloofar Moosavi, Jenny Haggstrom, and Xavier de Luna. The costs and benefits of uniformly valid causal inference with high-dimensional nuisance parameters. *arXiv:2105.02071*, 2021.

Tim P Morris, Ian R White, and Michael J Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019.

Jessica A Myers, Jeremy A Rassen, Joshua J Gagne, Krista F Huybrechts, Sebastian Schneeweiss, Kenneth J Rothman, Marshall M Joffe, and Robert J Glynn. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, 174(11):1213–1222, 2011.

Cian Naik, Emma J McCoy, and Daniel J Graham. Multiply robust dose-response estimation for multivalued causal inference problems. *arXiv:1611.02433*, 2016.

Art B Owen. *Empirical likelihood.* CRC press, Florida, 2001.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21 (1):31–54, 2012.

Dimitris N Politis and Joseph P Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Alessandro Rinaldo, Larry Wasserman, and Max G'Sell. Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469, 2019.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 56(4):931–954, 1988.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

Mehdi Rostami and Olli Saarela. Normalized augmented inverse probability weighting with neural network predictions. *Entropy*, 24(2):179, 2022a.

Mehdi Rostami and Olli Saarela. Targeted L1-regularization and joint modeling of neural networks for causal inference. *Entropy*, 24(9):1290, 2022b.

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in neural information processing systems*, 25, 2012.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Donald B Rubin. Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics*, 32:109–120, 1976.

Donald B Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328, 1979.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv:1706.05098*, 2017.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448): 1096–1120, 1999.

Hilary K Seligman, Barbara A Laraia, and Margot B Kushel. Food insecurity is associated with chronic disease among low-income NHANES participants. *The Journal of Nutrition*, 140(2):304–310, 2010.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge University Press, Cambridge, NY, 2014.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR.org, 2017.

Xiaoxi Shen and Jinghang Lin. Consistency of neural networks with regularization. *arXiv:2207.01538*, 2022.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32, 2019.

Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive Lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Tymon Sloczynski and Jeffrey M Wooldridge. A general double robustness result for estimating average treatment effects. *Econometric Theory*, 34(1):112–133, 2018.

Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust L1 regularized estimation of causal contrasts. *arXiv:1904.03737*, 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Leonard A Stefanski and Dennis D Boos. The calculus of M-estimation. *The American Statistician*, 56 (1):29–38, 2002.

Gabor J Szekely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Laszlo Szirmay-Kalos. Higher order automatic differentiation with dual numbers. *Periodica Polytechnica Electrical Engineering and Computer Science*, 65(1):1–10, 2021.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

M. J. van der Laan and S. Rose. *Targeted learning: Causal inference for observational and experimental data*. Springer, New York, NY, 2011.

Mark J van der Laan and Maya L Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1), 2007.

Mark J. van der Laan and James M. Robins. Unified methods for censored data and truncation. *Journal of the American Statistical Association*, 98(463):654–665, 2003.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, New York, NY, 2011.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications In Genetics And Molecular Biology*, 6(1), 2007a.

Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007b.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, Cambridge, NY, 2000.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Lei Wang. Multiple robustness estimation in causal inference. *Communications in Statistics-Theory and Methods*, 48(23):5701–5718, 2019.

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, New York, NY, 2006.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv:1505.00853*, 2015.

Shen Yan, Sing Kiong Nguang, and Zhou Gu. H infinity weighted integral event-triggered synchronization of neural networks with mixed delays. *IEEE Transactions on Industrial Informatics*, 17(4):2365–2375, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Paul N Zivich and Alexander Breskin. Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology*, 32(3):393, 2021.

Chapter 7

Appendix

7.1 Appendix A

7.1.1 Regulatory assumptions

Let β be the causal parameter, $\eta \in T$ be the infinite dimensional nuisance parameters where T is a convex set with a norm. Additionally, let the score function $\phi : \mathbb{O} \times \mathcal{B} \times T \to \mathbb{R}$ be a measurable function, $\mathbb{O}$ be the measurable space of all random variables O with probability distribution $P \in \mathcal{P}_n$ and $\mathcal{B}$ be an open subset of $\mathbb{R}$ containing the true causal parameter. Let the sample $O = (O_1, O_2, ..., O_n)$ be observed and the set of probability measures $\mathcal{P}_n$ expand with sample size n. In addition, let $\beta \in \mathcal{B}$ be the solution to the estimating equation $\mathbb{E}\phi(\mathbb{O}, \beta, \eta) = 0$. The assumptions that guarantee that the second-step orthogonal estimator $\hat{\beta}$ is asymptotically normal are (Chernozhukov et al., 2018a): (1) β does not fall on the boundary of $\mathcal{B}$; (2) the map $(\beta, \eta) \to \mathbb{E}_P\phi(O, \beta, \eta)$ is twice Gateaux differentiable (this holds by the positivity assumption). β is identifiable; (3) $\mathbb{E}_P\phi(O, \beta, \eta)$ is smooth enough; (4) $\hat{\eta} \in \mathcal{T}$ with high probability and $\eta \in \mathcal{T}$. $\hat{\eta}$ converges to η_0 at least as fast as $n^{-\frac{1}{4}}$ (similar but slightly stronger than first two assumptions in (3.17)); (5) score function(s) $\phi(., \beta, \eta)$ has finite second moment for all $\beta \in \mathcal{B}$ and all nuisance parameters $\eta \in \mathcal{T}$; (6) the score function(s) $\phi(., \beta, \eta)$ is measurable; (7) the number of folds increases by sample size.

7.2 Appendix B

First, let us review the proof sketch of the AIPW double robustness:

(3.3) can be consistently estimated by

$$\begin{aligned}\hat{\beta}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\Big(\frac{A_iY_i - \hat{Q}(1,W_i)(A_i - \hat{\mathbb{E}}[A_i|W_i])}{\hat{\mathbb{E}}[A_i|W_i]}\Big) - \\ \Big(\frac{(1-A_i)Y_i + \hat{Q}(0,W_i)(A_i - \hat{g}_i)}{1 - \hat{\mathbb{E}}[A_i|W_i]}\Big)\Bigg] = \\ \frac{1}{n}\sum_{i=1}^{n}\left([\frac{A_i}{\hat{g}_i} - \frac{1-A_i}{1-\hat{g}_i}]y_i - \frac{A_i - \hat{g}_i}{\hat{g}_i(1-\hat{g}_i)}[(1-\hat{g}_i)\hat{Q}_i^1 + \hat{g}_i\hat{Q}_i^0]\right) = \\ \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{A_i(y_i - \hat{Q}_i^1)}{\hat{g}_i} - \frac{(1-A_i)(y_i - \hat{Q}_i^0)}{1-\hat{g}_i}\Big) + \frac{1}{n}\sum_{i=1}^{n}(\hat{Q}_i^1 - \hat{Q}_i^0) \quad (7.1)\end{aligned}$$

The second formula guarantees the consistency of AIPW if $\hat{g}$ is consistent, and the third expression shows that the consistency of $\hat{Q}_i^0$ and $\hat{Q}_i^1$ is consistent.

Theorem 7.2.1 (nAIPW double robustness)**.** *Let the nAIPW estimator of risk difference be*

$$\hat{\beta}_{nAIPW} = \hat{\mathbb{E}}(\hat{Q}^1 - \hat{Q}^0) + \hat{\mathbb{E}}\Big(\frac{A(Y - \hat{Q}^1)}{\hat{g}\hat{\mathbb{E}}[\frac{A}{\hat{g}}]} - \frac{(1-A)(Y - \hat{Q}^0)}{(1-\hat{g})\hat{\mathbb{E}}[\frac{1-A}{1-\hat{g}}]}\Big). \quad (7.2)$$

Then, $\hat{\beta}_{nAIPW}$ is a consistent estimator of β if $\hat{g} \xrightarrow{p} g$ or $\hat{Q}^k \xrightarrow{p} Q^k$, $k = 0, 1$.

Proof. From (7.2), $\hat{\beta}_{nAIPW}$ is a consistent estimator of β if $\hat{Q}_i^0$ and $\hat{Q}_i^1$ are consistent. This is because the first term $\hat{\mathbb{E}}(\hat{Q}^1 - \hat{Q}^0)$ converges to β, while the second term tends to zero.

By re-expressing (7.2), the other argument is clear. Letting $\hat{w}^1 = \hat{\mathbb{E}}[\frac{A}{\hat{g}}]$ and $\hat{w}^0 = \hat{\mathbb{E}}[\frac{1-A}{1-\hat{g}}]$, we have:

$$\begin{aligned}\hat{\beta}_{nAIPW} = \frac{1}{n}\sum_{i=1}^{n}\left([\frac{A_i}{\hat{g}_i\hat{w}_i^1} - \frac{1-A_i}{(1-\hat{g}_i)\hat{w}_i^0}]y_i\right) + \\ \hat{\mathbb{E}}\Big(\hat{Q}^1 - \hat{Q}^0 - \frac{A_i\hat{Q}^1}{\hat{g}\hat{w}^1} + \frac{(1-A_i)\hat{Q}^0}{(1-\hat{g})\hat{w}^0}\Big) = \frac{1}{n}\sum_{i=1}^{n}\Bigg([\frac{A_i}{\hat{g}_i\hat{w}_i^1} - \frac{1-A_i}{(1-\hat{g}_i)\hat{w}_i^0}]y_i - \\ \hat{Q}_i^1(A_i - \hat{g}_i\hat{w}_i^1) + \hat{Q}_i^0(1 - A_i - (1-\hat{g}_i)\hat{w}_i^0)\Bigg) \quad (7.3)\end{aligned}$$

The first expression in (7.3) is the same as the nIPW estimator which is a consistent estimator of β (Lunceford and Davidian, 2004). Now, under the consistency of $\hat{g}$, the second term tends to zero, as $\hat{w}_1 \xrightarrow{p} 1$ and $\hat{w}_0 \xrightarrow{p} 1$.

In the theorem below, it is shown that there is an M-estimation equivalent to β_{nAIPW} and w^1 and w^0. This, plus the continuous mapping theorem, proves that $\sum_{i=1}^{n}\frac{A_i}{\hat{g}_i}$ converges in probability to n if $\hat{g} \xrightarrow{p} g$.

□

Theorem 7.2.2. *The asymptotic variance of the nAIPW* (3.5) *is*

$$\hat{\sigma}^2_{nAIPW} = \sum_{i=1}^{n}\Big(\frac{A_i(y_i - \hat{Q}_i^1)w_i^{(1)}}{\sum_{j=1}^{n}A_jw_j^{(1)}} - \frac{(1-A_i)(y_i - \hat{Q}_i^0)w_i^{(0)}}{\sum_{j=1}^{n}(1-A_j)w_j^{(0)}} + \frac{1}{n}(\hat{\beta}_{SR} - \hat{\beta}_{nAIPW})\Big)^2, \quad (7.4)$$

where $\hat{Q}_i^k = \hat{Q}(k, W_i)$ *and* $\hat{g}_i = \hat{\mathbb{E}}[A_i|W_i]$.

Proof. Let us define a few notations first:

$$\begin{aligned} q &= Q^1 - Q^0, \\ g &= \mathbb{E}[A|W], \\ f &= y - Q^1, \\ h &= y - Q^0, \\ v &= \frac{A}{g}, \\ u &= \frac{1-A}{1-g}. \end{aligned} \tag{7.5}$$

With this set of notations, the nAIPW estimator (3.5) can be written as

$$\hat{\beta}_{nAIPW} = \sum_{i=1}^{n} \Big(\frac{v_i f_i}{\sum_{j=1}^{n} v_j} - \frac{u_i h_i}{\sum_{j=1}^{n} u_j} + \frac{q_i}{n} \Big), \tag{7.6}$$

Following the methods in (Stefanski and Boos, 2002), to find an estimating equation whose solution is $\hat{\beta}_{nAIPW}$, we introduce two more estimating equations. Employing the M-estimation theory, we will prove that nAIPW is asymptotically normal, and we will calculate its standard error.

It can be seen that (7.6) is not a solution to an M-estimator directly. However, by defining two more parameters and concatenating their estimating equations, we obtain 3-dim multivariate estimating equations:

$$\begin{aligned} \sum_{i=1}^{n} \Big(\frac{v_i f_i}{\gamma} - \frac{u_i h_i}{\lambda} + \frac{1}{n}(q_i - \beta) \Big) &= 0, \\ \sum_{i=1}^{n} \Big(v_i - \frac{\gamma}{n} \Big) &= 0, \\ \sum_{i=1}^{n} \Big(u_i - \frac{\lambda}{n} \Big) &= 0. \end{aligned} \tag{7.7}$$

To ease the calculations, we modify the first estimating equation with an equivalent one, but the results will not differ:

$$\begin{aligned} \sum_{i=1}^{n} \lambda v_i f_i - \gamma u_i h_i + \frac{\gamma\lambda}{n}(q_i - \beta) &= 0, \\ \sum_{i=1}^{n} v_i - \frac{\gamma}{n} &= 0, \\ \sum_{i=1}^{n} u_i - \frac{\lambda}{n} &= 0. \end{aligned} \tag{7.8}$$

By defining the following notations,

$$\psi = \begin{pmatrix} \phi \\ \eta \\ \Omega \end{pmatrix} = \begin{pmatrix} \lambda v f - \gamma u h + \frac{\gamma\lambda}{n}(q-\beta) \\ v - \frac{\gamma}{n} \\ u - \frac{\lambda}{n} \end{pmatrix},$$

we have $\sum_{i=1}^{n} \psi_i = 0$, or

$$\begin{aligned} \sum_{i=1}^{n} \phi_i, &= 0, \\ \sum_{i=1}^{n} \eta_i &= 0, \\ \sum_{i=1}^{n} \Omega_i &= 0. \end{aligned} \tag{7.9}$$

The M-estimation theory implies that under regulatory conditions, the solutions to these estimating equations converge in distribution to a multivariate normal distribution:

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_{nAIPW} \\ \hat{\gamma} \\ \hat{\lambda} \end{pmatrix} \sim MVN\left(\theta \ , \ \mathbf{I}^{-1}(\theta)\mathbf{B}(\theta)\mathbf{I}^{-1}(\theta)^T\right)$$

where

$$\theta = \begin{pmatrix} \beta \\ \gamma \\ \lambda \end{pmatrix},$$

$$\mathbf{I}(\theta) = -\mathbb{E}\frac{\partial \psi}{\partial \theta^T} = \frac{1}{n} \begin{pmatrix} \frac{\lambda\gamma}{n} & \mathbb{E}(uh - \frac{\lambda}{n}(q-\beta)) & -\mathbb{E}(vf + \frac{\gamma}{n}(q-\beta)) \\ 0 & \frac{1}{n} & 0 \\ 0 & 0 & \frac{1}{n} \end{pmatrix}, \tag{7.10}$$

whose inverse is

$$\mathbf{I}^{-1}(\theta) = \frac{n}{\gamma\lambda} \begin{pmatrix} 1 & -n\mathbb{E}(uh - \lambda(q-\beta)) & n\mathbb{E}(vf + \frac{\gamma}{n}(q-\beta)) \\ 0 & \gamma\lambda & 0 \\ 0 & 0 & \gamma\lambda \end{pmatrix}, \tag{7.11}$$

and,

$$\mathbf{B}(\theta) = \mathbb{E}\psi\psi^T = \begin{pmatrix} \mathbb{E}\phi^2 & \mathbb{E}\phi\eta & \mathbb{E}\phi\Omega \\ \mathbb{E}\phi\eta & \mathbb{E}\eta^2 & \mathbb{E}\eta\Omega \\ \mathbb{E}\phi\Omega & \mathbb{E}\eta\Omega & \mathbb{E}\Omega^2 \end{pmatrix}. \tag{7.12}$$

In order to estimate the variance of $\hat{\beta}_{nAIPW}$, we do not need to calculate all entries of the variance–covariance matrix, only the first entry:

$$\frac{1}{n}(\frac{n^2}{(\gamma\lambda)^2})\begin{pmatrix}\mathbb{E}\phi^2+\epsilon & \star & \star\\ \star & \star & \star\\ \star & \star & \star\end{pmatrix}. \tag{7.13}$$

The $\star$ entries are irrelevant to the calculation of variance of nAIPW and the term ϵ is a very long expression which involves terms converging to zero faster than the actual estimating Equation (7.9) (Hines et al., 2022) (also verified by simulations):

$$\begin{aligned}\epsilon = -\mathbb{E}\phi\eta(n\mathbb{E}uh+\lambda(\beta-q))+\mathbb{E}\phi\Omega(n\mathbb{E}vf-\gamma(\beta-q))-\\ (n\mathbb{E}uh+\lambda(\beta-q))(-\mathbb{E}\eta^2(n\mathbb{E}uh+\lambda(\beta-q))+\mathbb{E}\eta\Omega(n\mathbb{E}vf-\gamma(\beta-q))+\\ \mathbb{E}\phi\eta)+(n\mathbb{E}vf-\gamma(\beta-q))(-\mathbb{E}\eta\Omega(n\mathbb{E}uh+\lambda(\beta-q))+\\ \mathbb{E}\Omega^2(n\mathbb{E}vf-\gamma(\beta-q))+\mathbb{E}\phi\Omega).\end{aligned} \tag{7.14}$$

Further,

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_{nAIPW}\\ \hat{\gamma}\\ \hat{\Omega}\end{pmatrix}\sim MVN\left(\theta\ ,\ \hat{\mathbf{I}}^{-1}(\hat{\theta})\hat{\mathbf{B}}(\hat{\theta})\hat{\mathbf{I}}^{-1}(\hat{\theta})^T\right) \tag{7.15}$$

where we replace $\mathbb{E}$ with sample averages in Expressions (7.10)–(7.12) and θ with their corresponding solutions to Equation (7.8). Following this recipe, we obtain

$$\hat{\sigma}^2_{nAIPW}=\frac{1}{n}(\frac{n^2}{(\gamma\lambda)^2})\hat{\mathbb{E}}\phi^2+\hat{\epsilon}\approx\sum_{i=1}^{n}\left(\frac{v_if_i}{\hat{\gamma}}-\frac{u_ih_i}{\hat{\lambda}}+\frac{1}{n}q_i-\hat{\beta}_{nAIPW})\right)^2, \tag{7.16}$$

which is the same as (7.4).

□

7.3 Appendix C

Lemma 7.3.1. *Let g^k's and Q^l's be the same as defined in (5.7), and $g=g(W)$ be the true propensity score. Then we have*

$$\begin{aligned}\mathbb{E}\left[\frac{g^k-\mathbb{E}[g^k]}{g}\Big|A=1\right]=0,\quad k=1,...,K,\\ \mathbb{E}\left[\frac{Q^l(1)-\mathbb{E}[Q^l(1)]}{g}\Big|A=1\right]=0,\quad l=1,...,L,\\ \mathbb{E}\left[\frac{\mathbb{E}[g^k]-g^k}{1-g}\Big|A=0\right]=0,\quad k=1,...,K,\\ \mathbb{E}\left[\frac{Q^l(0)-\mathbb{E}[Q^l(0)]}{1-g}\Big|A=0\right]=0,\quad l=1,...,L.\end{aligned} \tag{7.17}$$

where $Q^l(j)=Q^1(j,W)$.

Proof. We prove the first identity. By the positivity assumption, it is enough to show that

$$\mathbb{E}\Big[\frac{g^k - \mathbb{E}[g^k]}{g}\Big|A=1\Big]P(A=1) = 0 \tag{7.18}$$

By applying the total law of expectations twice, letting $I(.)$ be the indicator function, and using the definition of the propensity score function, we have

$$\begin{aligned}\mathbb{E}\Big[\frac{g^k - \mathbb{E}[g^k]}{g}\Big|A=1\Big]P(A=1) = \mathbb{E}\Big[\frac{I(A=1)(g^k - \mathbb{E}[g^k])}{g}\Big|A=1\Big]P(A=1) = \\ \mathbb{E}\Big[\frac{I(A=1)(g^k - \mathbb{E}[g^k])}{g}\Big|A=1\Big]P(A=1)+ \\ \mathbb{E}\Big[\frac{I(A=1)(g^k - \mathbb{E}[g^k])}{g}\Big|A=0\Big]P(A=0) = \\ \mathbb{E}\Big[\frac{I(A=1)(g^k - \mathbb{E}[g^k])}{g}\Big] = \mathbb{E}\Bigg[\mathbb{E}\Big[\frac{I(A=1)(g^k - \mathbb{E}[g^k])}{g}\Big|X\Big]\Bigg] = \\ \mathbb{E}\Bigg[\mathbb{E}\Big[I(A=1)\Big|X\Big]\frac{g^k - \mathbb{E}[g^k]}{g}\Bigg] = \mathbb{E}\Big[g^k - \mathbb{E}[g^k]\Big] = 0\end{aligned} \tag{7.19}$$

The second identity can be proven by following similar steps. The third identity can be proven by letting $D = A-1$. Noted that in this theorem g in the denumerators is assumed to be the true propensity score, but in later theorems, we do not need to know the true propensity scores.

□

Proof. Proof of Lemma 5.3.1

Before proposing the estimator for ATE, consider the empirical joint probability distributions for the treatment and control groups $\{(y_i, w_i)|A_i = 1\}$ and $\{(y_i, w_i)|A_i = 0\}$. Define the probability functions $p_i = P(Y_i = y_i, W_i = w_i|A_i = 1)$, $q_i = P(Y_i = y_i, W_i = w_i|A_i = 0)$. Without any constraints on the p_i's and q_i's , it can be shown that $p_i = \frac{1}{n}$ and $q_i = \frac{1}{n}$ (Owen, 2001). However, the sample version of (7.17) with the general form of the probability distribution functions p_i and q_i impose constraints on them

$$\begin{aligned}\sum_{i\in I_1} p_i \frac{g_i^k - \hat{\mathbb{E}}[g_i^k]}{g_i} = 0, \quad k = 1, ..., K, \\ \sum_{i\in I_1} p_i \frac{Q_i^l(1) - \hat{\mathbb{E}}[Q_i^l(1)]}{g_i} = 0, \quad l = 1, ..., L,\end{aligned} \tag{7.20}$$

$$\begin{aligned}\sum_{i\in I_0} q_i \frac{\hat{\mathbb{E}}[g_i^k] - g_i^k}{1 - g_i} = 0, \quad k = 1, ..., K, \\ \sum_{i\in I_0} q_i \frac{Q_i^l(0) - \hat{\mathbb{E}}[Q_i^l(0)]}{1 - g_i} = 0, \quad l = 1, ..., L,\end{aligned} \tag{7.21}$$

where $I_j = \{i : A_i = j\}$, for $j = 0, 1$, with cardinality $n_j = |I_j|$.

As

$$\frac{f_{Y^1,W|A=1}(y,w|A=1)}{P(A=1|W=w)}P(A=1) = \frac{f_{Y^1,W|A=1}(y,w|A=1)}{P(A=1|W=w,Y^1=y)}P(A=1) =$$
$$\frac{f_{Y^1,W|A=1}(y,w|A=1)}{\frac{f_{Y^1,W|A=1}(y,w|A=1)P(A=1)}{f_{Y^1,W}(y,w)}}P(A=1) = f_{Y^1,W}(y,w), \quad (7.22)$$

and $f_{Y^1,W|A=1}(y,w|A=1) = f_{Y,W|A=1}(y,w|A=1)$, we have

$$u_i = \frac{p_i\theta}{g(W_i)}. \quad (7.23)$$

Similarly,

$$v_i = \frac{q_i(1-\theta)}{1-g(W_i)}. \quad (7.24)$$

The empirical versions of u_i's and v_i's are obtained by replacing empirical versions of p_i's and q_i's in (7.23) and (7.24), respectively.

The goal is to estimate ATE

$$\beta = \mathbb{E}[Y^1] - \mathbb{E}[Y^0] \quad (7.25)$$

The MR estimator of ATE is

$$\hat{\beta} = \sum_{i=1}^{n} A_i\hat{u}_iY_i - \sum_{i=1}^{n}(1-A_i)\hat{v}_iY_i. \quad (7.26)$$

This is motivated by the fact that $\mathbb{E}[Y^j] = \int yf_{Y^j}(y)dy = \iint yf_{Y^j,W}(y,w)dwdy$, for $j = 0,1$. For an observed sample, if the true u and v and potential outcomes are observed, the ideal estimator (that is not available in practice) is $\hat{\beta}_{ideal} = \sum_{i=1}^{n} u_iY_i^1 - \sum_{i=1}^{n} v_iY_i^0$; but the counterfactual outcomes are not observed in real life. The remedy is to calculate these sums for the observed outcomes and at the same time weigh them by the reciprocal of the true propensity score in equations (7.23) and (7.24) to balance the distribution of the treatment and control groups due to the existence of the confounders, thus the estimator (7.26). Note that, as seen below, in the process of calculating the MR estimator, true propensity scores are canceled and in reality, we do not need to know the true PS.

To estimate u_i and v_i, we estimate p_i's, and q_i's (the latter is similar to the former, and thus we skip that). Here we assumed that g in the denominator is one of the K estimated propensity scores but we do not need to have the knowledge of which of them is the correctly specified propensity score model. Without the loss of generality, we can assume that g^1 is correctly specified, but for ease of notation, we still use g in the denominator.

Empirical likelihood without constraints other than the normalization (summation of weights equals 1) is the same as the sample average for the estimation of expectations. However, the constraints (7.20) and (7.21) produce less straightforward estimators for the empirical probabilities (weights). Independence of subjects and the usage of the empirical likelihood method implies that in order to maximize the joint probability of W, Y, given $A = 1$, we must maximize the quantity

$$\max_{p_i:i\in I_1} \prod_{i\in I_1} p_i. \quad (7.27)$$

This results in a convex optimization problem with equality and inquality constraints:

$$\max_{p_i:i\in I_1}\prod_{i\in I_1}p_i,$$

subject to

$$\begin{aligned} p_i &> 0, \quad i\in I_1,\\ \sum_{i\in I_1}p_i &= 1,\\ \sum_{i\in I_1}p_i\frac{g_i^k-\hat{\mathbb{E}}[g_i^k]}{g_i} &= 0, \quad k=1,...,K\\ \sum_{i\in I_1}p_i\frac{Q_i^l(1)-\hat{\mathbb{E}}[Q_i^l(1)]}{g_i} &= 0, \quad l=1,...,L. \end{aligned} \tag{7.28}$$

Equivalently, we can find $\max_{p_i:i\in I_1}\sum_{i\in I_1}log(p_i)$, with the same constraints as above. Applying the Lagrange multiplier, a solution to (7.28) is a solution to the following convex problem

$$\max_{p_i:i\in I_1}\sum_{i\in I_1}log(p_i)-\lambda^T\left\{\sum_{i\in I_1}p_i\frac{C_i^j}{g_i}\right\}_{k=1}^{K+L}-\eta^TP-c(\sum_{i\in I_1}p_i-1), \tag{7.29}$$

subject to

$$\begin{aligned} \frac{1}{\hat{p}_i}-\lambda^TC_i^1/g_i-\eta_i-c &= 0,\\ \eta_ip_i &= 0,\ i\in I_1,\\ \eta_i &\geq 0,\ i\in I_1,\\ \sum_{i\in I_1}p_i &= 1,\\ \sum_{i\in I_1}p_i\frac{C_i^{1,k}}{g_i} &= 0, \quad k=1,...,K+L. \end{aligned} \tag{7.30}$$

where λ is a $K+L$ dimensional parameter vector, and η is a $|I_1|$-dimensional vector, and P is the row vector of p_i, for $i\in I_1$.

By the first three conditions, we have that $\eta_i=0$, $i\in I_1$, and

$$\hat{p}_i=\frac{1}{c+\lambda^TC_i^1/g_i}. \tag{7.31}$$

By the forth and fifth conditions, we have

$$\begin{aligned} \sum_{i\in I_1}\frac{1}{c+\lambda^TC_i^1/g_i} &= 1,\\ \sum_{i\in I_1}\frac{C_i^{1,k}/g_i}{c+\lambda^TC_i^1/g_i} &= 0, \quad k=1,...,K+L. \end{aligned} \tag{7.32}$$

By multiplying λ_j to each identity to both sides of the second identity above and summing them, we have

$$\sum_{i\in I_1}\frac{\alpha_i}{c+\alpha_i}=\sum_{i\in I_1}1-\frac{c}{c+\alpha_i}=0, \text{where } \alpha_i=\lambda^T C_i^1/g_i. \tag{7.33}$$

Thus, by the first identity in (7.32), we have $c=n_1$. Thus

$$\hat{p}_i=\frac{1}{n_1}\frac{1}{1+\tau^T C_i^1/g_i},\ \text{where } \sum_{i\in I_1}\frac{C_i^{1,k}/g_i}{1+\tau^T C_i^1/g_i}=0,\quad k=1,...,K+L, \tag{7.34}$$

where $\tau=\lambda/n_1$.

Without loss of generality, assuming that $\hat{g}^1$ is the correctly specified propensity score estimator, by substituting $\tau_1=-1+\gamma_1\theta$, and $\tau_k=\gamma_k\theta$, for $k=2,...,K+L$ into (7.34), we can re-write u_i as

$$\hat{u}_i=\theta\times\frac{1}{\theta n_1}\frac{1}{1+\gamma^T C_i^1},\ \text{where } \sum_{i\in I_1}\frac{C_i^{1,k}}{1+\gamma^T C_i^1}=0,\quad k=1,...,K+L, \tag{7.35}$$

which does not depend on the true propensity score or the correctly specified estimator of the propensity score. Hence,

$$\hat{u}_i=\frac{1}{n_1}\frac{1}{1+\gamma^T C_i^1},\ \text{where } A_i=1,\ \sum_{j\in I_1}\frac{C_j^{1,k}}{1+\gamma^T C_j^1}=0,\quad k=1,...,K+L, \tag{7.36}$$

and

$$\hat{v}_i=\frac{1}{n_0}\frac{1}{1+\rho^T C_i^0},\ \text{where } A_i=0,\ \sum_{j\in I_0}\frac{C_j^{0,k}}{1+\rho^T C_j^0}=0,\quad k=1,...,K+L, \tag{7.37}$$

where n_1 and n_0 are the sizes of treatment and control groups. Or we can normalize u_i's and v_i's:

$$\hat{u}_i=\frac{1}{1+\gamma^T C_i^1}/\Big(\sum_{j=1}^{n}\frac{1}{1+\gamma^T C_j^1}\Big),\ \text{where } A_i=1,\ \sum_{j\in I_1}\frac{C_j^{1,k}}{1+\gamma^T C_j^1}=0,\quad k=1,...,K+L, \tag{7.38}$$

and

$$\hat{v}_i=\frac{1}{1+\rho^T C_i^0}/\Big(\sum_{j\in I_0}\frac{1}{1+\rho^T C_j^0}\Big),\ \text{where } A_i=0,\ \sum_{j\in I_0}\frac{C_j^{0,k}}{1+\rho^T C_j^0}=0,\quad k=1,...,K+L. \tag{7.39}$$

□

Lemma 7.3.2. *If one of the propensity score estimators is correctly specified, then $\hat{\lambda}$ converges to zero in probability, that is $\hat{\lambda}\xrightarrow{p}0$.*

Proof. We have that for all $j\in\{1,2,...,K+L\}$,

$$f(\vec{g},\vec{Q},\vec{\lambda})=\sum_{i\in I_1}\frac{C_i^{1,j}/g_i}{1+\lambda^T C_i^1/g_i}=\begin{pmatrix}\sum_{i\in I_1}\frac{\hat{g}_i^1-\hat{\mathbb{E}}\hat{g}^1}{g_i+\lambda^T C_i^1}\\ \cdots\\ \sum_{i\in I_1}\frac{\hat{g}_i^K-\hat{\mathbb{E}}\hat{g}^K}{g_i+\lambda^T C_i^1}\\ \sum_{i\in I_1}\frac{\hat{Q}_i^1(1)-\hat{\mathbb{E}}\hat{Q}^1(1)}{g_i+\lambda^T C_i^1}\\ \cdots\\ \sum_{i\in I_1}\frac{\hat{Q}_i^L(1)-\hat{\mathbb{E}}\hat{Q}^L(1)}{g_i+\lambda^T C_i^1}\end{pmatrix}=0, \tag{7.40}$$

where C_i^1 contains K estimators of the propensity score, and L estimators of the outcome regression (for the treated group.) Without loss of generality, assume the correctly specified model is g^1, the first one, and for the ease of notation, we use the notation g instead of g^1. And assume the true values for the propensity scores and outcome regressions are $g^{1,*}, ..., g^{K,*}, Q^{1,*}, ..., Q^{L,*}$, or in vector form, $\vec{g}^*, \vec{Q}^*$.

The first-order Taylor expansion of one of the entries in the left hand side of (7.40) around $(\vec{g}^*, \vec{Q}^*, \vec{0})$ is:

$$f(\vec{g}^*, \vec{Q}^*, \vec{0}) + \partial f_{\vec{\lambda}}(\vec{g}^*, \vec{Q}^*, \vec{0})\vec{\lambda} + \partial f_{g^1}(\vec{g}^*, \vec{Q}^*, \vec{0})(g^1 - g^{1,*}) + \sum_{j=2}^{K} \partial f_{g^j}(\vec{g}^*, \vec{Q}^*, \vec{0})(g^j - g^{j,*}) + \\ \sum_{j=1}^{L} \partial f_{Q^j}(\vec{g}^*, \vec{Q}^*, \vec{0})(Q^j - Q^{j,*}), \quad (7.41)$$

where $\partial f_{\vec{\lambda}}(\vec{g}^*, \vec{Q}^*, \vec{0})\vec{\lambda}$ could also be written as $\sum_{l=1}^{K+L} \partial f_{\lambda_l}(\vec{g}^*, \vec{Q}^*, \vec{0})\lambda_l$.

Applying (7.41) on all observations in (7.40)

$$0 = n^{r-1} \sum_{i=1}^{n} A_i f(\vec{g}_i, \vec{Q}_i, \vec{\lambda}_i) \approx n^{r-1} \sum_{i=1}^{n} A_i f(\vec{g}_i^*, \vec{Q}_i^*, \vec{0}) + \Big[\frac{1}{n} \sum_{i=1}^{n} A_i \partial f_{\vec{\lambda}}(\vec{g}_i^*, \vec{Q}_i^*, \vec{0})\Big] n^r \hat{\vec{\lambda}} + \\ \Big[\frac{1}{n} \sum_{i=1}^{n} A_i \partial f_{g_i^1}(\vec{g}_i^*, \vec{Q}_i^*, \vec{0})\Big] n^r (\hat{g}_i^1 - g_i^{1,*}) + \sum_{j=2}^{K} \Big[\frac{1}{n} \sum_{i=1}^{n} A_i \partial f_{g_i^j}(\vec{g}_i^*, \vec{Q}_i^*, \vec{0})\Big] n^r (\hat{g}_i^j - g_i^{j,*}) + \\ \sum_{j=1}^{L} \Big[\frac{1}{n} \sum_{i=1}^{n} A_i \partial f_{Q_i^j}(\vec{g}_i^*, \vec{Q}_i^*, \vec{0})\Big] n^r (\hat{Q}_i^j - Q_i^{j,*}) + o_p(1) =$$

$$n^{r-1} \sum_{i=1}^{n} \frac{A_i}{g_i^{1,*}} \begin{pmatrix} g_i^{1,*} - \mathbb{E}\, g^{1,*} \\ ... \\ g_i^{K,*} - \mathbb{E}\, g^{K,*} \\ Q_i^{1,*}(1) - \mathbb{E}\, Q^{1,*}(1) \\ ... \\ Q_i^{L,*}(1) - \mathbb{E}\, Q^{L,*}(1) \end{pmatrix} - \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} A_i \Big(\frac{g_i^{1,*} - \mathbb{E}\, g^{1,*}}{g_i^{1,*}}\Big)^2 \\ ... \\ A_i \Big(\frac{g_i^{K,*} - \mathbb{E}\, g^{K,*}}{g_i^{1,*}}\Big)^2 \\ A_i \Big(\frac{Q_i^{1,*}(1) - \mathbb{E}\, Q^{1,*}(1)}{g_i^{1,*}}\Big)^2 \\ ... \\ A_i \Big(\frac{Q_i^{L,*}(1) - \mathbb{E}\, Q^{L,*}(1)}{g_i^{1,*}}\Big)^2 \end{pmatrix} n^r \hat{\vec{\lambda}} +$$

$$\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} A_i \frac{\mathbb{E}\, g^{1,*}}{(g_i^{1,*})^2} \\ -A_i \frac{g_i^{2,*} - \mathbb{E}\, g^{2,*}}{(g_i^{1,*})^2} \\ ... \\ -A_i \frac{g_i^{K,*} - \mathbb{E}\, g^{K,*}}{(g_i^{1,*})^2} \\ -A_i \frac{Q_i^{1,*}(1) - \mathbb{E}\, Q^{1,*}(1)}{(g_i^{1,*})^2} \\ ... \\ -A_i \frac{Q_i^{L,*}(1) - \mathbb{E}\, Q^{L,*}(1)}{(g_i^{1,*})^2} \end{pmatrix} n^r (\hat{g}_i^1 - g_i^{1,*}) +$$

$$\sum_{j=1}^{L}\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} 0 \\ ... \\ \frac{A_i}{g_i^{1,*}} \\ ... \\ 0 \\ 0 \\ ... \\ 0 \end{pmatrix} n^r(\hat{g}_i^j - g_i^{j,*}) + \sum_{j=1}^{L}\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} 0 \\ ... \\ 0 \\ \frac{A_i}{g_i^{1,*}} \\ 0 \\ ... \\ 0 \end{pmatrix} n^r(\hat{Q}(1)_i^j - Q(1)_i^{j,*}) + o_p(1). \tag{7.42}$$

The third term in (7.42) can be re-written as

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \frac{A_i}{g_i^{1,*}} - A_i\frac{g_i^{1,*} - \mathbb{E}\, g^{1,*}}{(g_i^{1,*})^2} \\ -A_i\frac{g_i^{2,*} - \mathbb{E}\, g^{2,*}}{(g_i^{1,*})^2} \\ ... \\ -A_i\frac{g_i^{K,*} - \mathbb{E}\, g^{K,*}}{(g_i^{1,*})^2} \\ -A_i\frac{Q_i^{1,*}(1) - \mathbb{E}\, Q^{1,*}(1)}{(g_i^{1,*})^2} \\ ... \\ \frac{1}{n}\sum_{i=1}^{n} -A_i\frac{Q_i^{L,*}(1) - \mathbb{E}\, Q^{L,*}(1)}{(g_i^{1,*})^2} \end{pmatrix} n^r(\hat{g}_i^1 - g_i^{1,*}).$$

Thus, the summation of last three terms in (7.42) is

$$\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} -A_i\frac{g_i^{1,*} - \mathbb{E}\, g^{1,*}}{(g_i^{1,*})^2} \\ -A_i\frac{g_i^{2,*} - \mathbb{E}\, g^{2,*}}{(g_i^{1,*})^2} \\ ... \\ -A_i\frac{g_i^{K,*} - \mathbb{E}\, g^{K,*}}{(g_i^{1,*})^2} \\ -A_i\frac{Q_i^{1,*}(1) - \mathbb{E}\, Q^{1,*}(1)}{(g_i^{1,*})^2} \\ ... \\ -A_i\frac{Q_i^{L,*}(1) - \mathbb{E}\, Q^{L,*}(1)}{(g_i^{1,*})^2} \end{pmatrix} n^r(\hat{g}_i^1 - g_i^{1,*}) + n^{r-1}\sum_{i=1}^{n}\frac{A_i}{g_i^{1,*}}\begin{pmatrix} \hat{g}_i^1 - g_i^{1,*} \\ \hat{g}_i^2 - g_i^{2,*} \\ ... \\ \hat{g}_i^L - g_i^{L,*} \\ \hat{Q}(1)_i^1 - Q(1)_i^{1,*} \\ ... \\ \hat{Q}(1)_i^K - Q(1)_i^{K,*} \end{pmatrix}.$$

Using the above simplified terms, the equation (7.42) can be further simplied to

$$n^{r-1}\sum_{i=1}^{n}\begin{pmatrix} \frac{A_i(\hat{g}_i^1 - \mathbb{E}\, g^{1,*})}{g_i^{1,*}} \\ ... \\ \frac{A_i(\hat{g}_i^K - \mathbb{E}\, g^{K,*})}{g_i^{1,*}} \\ \frac{A_i(\hat{Q}_i^1(1) - \mathbb{E}\, Q^{1,*}(1))}{g_i^{1,*}} \\ ... \\ \frac{A_i(\hat{Q}_i^L(1) - \mathbb{E}\, Q^{L,*}(1))}{g_i^{1,*}} \end{pmatrix} + \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} -A_i\left(\frac{g_i^{1,*} - \mathbb{E}\, g^{1,*}}{g_i^{1,*}}\right)^2 \\ ... \\ -A_i\left(\frac{g_i^{K,*} - \mathbb{E}\, g^{K,*}}{g_i^{1,*}}\right)^2 \\ -A_i\left(\frac{Q_i^{1,*}(1) - \mathbb{E}\, Q^{1,*}(1)}{g_i^{1,*}}\right)^2 \\ ... \\ -A_i\left(\frac{Q_i^{L,*}(1) - \mathbb{E}\, Q^{L,*}(1)}{g_i^{1,*}}\right)^2 \end{pmatrix} n^r\hat{\vec{\lambda}} -$$

$$n^{r-1}\sum_{i=1}^{n}\begin{pmatrix} A_i\frac{g_i^{1,*}-\mathbb{E}\,g^{1,*}}{(g_i^{1,*})^2} \\ A_i\frac{g_i^{2,*}-\mathbb{E}\,g^{2,*}}{(g_i^{1,*})^2} \\ \dots \\ A_i\frac{g_i^{K,*}-\mathbb{E}\,g^{K,*}}{(g_i^{1,*})^2} \\ A_i\frac{Q_i^{1,*}(1)-\mathbb{E}\,Q^{1,*}(1)}{(g_i^{1,*})^2} \\ \dots \\ A_i\frac{Q_i^{L,*}(1)-\mathbb{E}\,Q^{L,*}(1)}{(g_i^{1,*})^2} \end{pmatrix}(\hat{g}_i^1 - g_i^{1,*}) + o_p(1).$$

Thus

$$n^{r-1}\sum_{i=1}^{n}\frac{A_i\hat{C}_i^{1,*}}{g_i^{1,*}} + Mn^r\hat{\vec{\lambda}} - Vn^{r-1}(\hat{g}_i^1 - g_i^{1,*}) + o_p(1) = 0,$$

which implies

$$n^r\hat{\vec{\lambda}} = M^{-1}\Big(- n^{r-1}\sum_{i=1}^{n}\frac{A_i\hat{C}_i^{1,*}}{g_i^{1,*}} + Vn^{r-1}(\hat{g}_i^1 - g_i^{1,*})\Big) + o_p(1),$$

which means $\hat{\vec{\lambda}}$ is n^r-consistent, if one of the propensity score models is consistent with a faster than or equal rate to n^r. Noted that, by (7.17) equations, $\sum_{i=1}^{n}\frac{A_i\hat{C}_i^{1,*}}{g_i^{1,*}}$ converges to zero in probability.

In above equations, we have used these derivative:

$$\begin{aligned}
\frac{\partial}{\partial\lambda_j}\frac{C_k^1}{g^1+\vec{\lambda}^T C^1} &= -\frac{C_j^1 C_k^1}{(g^1+\vec{\lambda}^T C^1)^2},\\
\frac{\partial}{\partial g^1}\frac{C_1^1}{g^1+\vec{\lambda}^T C^1} &= \frac{g^1+\lambda^T C^1 - C_1^1(1+\lambda_1)}{(g^1+\vec{\lambda}^T C^1)^2},\\
\frac{\partial}{\partial g^1}\frac{C_k^1}{g^1+\vec{\lambda}^T C^1} &= -\frac{C_1^1(1+\lambda_1)}{(g^1+\vec{\lambda}^T C^1)^2},\\
\frac{\partial}{\partial g^j}\frac{C_k^1}{g^1+\vec{\lambda}^T C^1} &= -\frac{C_k^1\lambda_j}{(g^1+\vec{\lambda}^T C^1)^2},\\
\frac{\partial}{\partial Q^j}\frac{C_k^1}{g^1+\vec{\lambda}^T C^1} &= -\frac{C_k^1\lambda_j}{(g^1+\vec{\lambda}^T C^1)^2}.
\end{aligned} \tag{7.43}$$

□

Proof. Proof of Lemma 5.3.2.

Lemma 7.3.2 proves that given one of the propensity scores is consistent faster than n^r rate, we have than $\hat{\gamma} \xrightarrow{p} 0$ (and similarly it can be proven that $\hat{\rho} \xrightarrow{p} 0$) with n^r rate. As $u_i = \frac{\theta p_i}{g_i}$ and $\hat{p}_i = \frac{1}{n_1}\frac{2}{1+\hat{\lambda}^T C_i^1/g_i}$ (7.23), we have:

$$\hat{\beta}^1 = \frac{1}{n_1}\sum_{i=1}^{n}\frac{A_i Y_i}{g_i + \hat{\lambda}^T C_i^1} \xrightarrow{p} \mathbb{E}\,\frac{AY}{g}. \tag{7.44}$$

Similarly, $\hat{\beta}^0 \xrightarrow{p} E\frac{(1-A)Y}{1-g}$. The same is true for the normalized version

$$\hat{\beta}^1 = \sum_{i=1}^{n}\frac{A_i Y_i}{g_i + \hat{\lambda}^T C_i^1}/\Big(\sum_{i=1}^{n}\frac{A_i}{g_i + \hat{\lambda}^T C_i^1}\Big) \xrightarrow{p} \mathbb{E}\,\frac{AY}{g}/\mathbb{E}\,\frac{A}{g}, \tag{7.45}$$

□

Proof. Proof of Lemma 5.3.3:

Let one of the outcome regression models be consistent, say, $\hat{Q}_j(1)^1$, where the super script 1 refers to the first model trained among L first step models, with no loss of generality. The constraints in (5.6) contains

$$\sum_{i=1}^{n}\Big(u_iA_i\hat{Q}_i^1(1)-\hat{\mathbb{E}}\hat{Q}^1(1)\Big)=0,\quad \sum_{i=1}^{n}\Big(v_i(1-A_i)\hat{Q}_i^1(0)-\hat{\mathbb{E}}\hat{Q}^1(0)\Big)=0, \tag{7.46}$$

which can be re-written as

$$\begin{aligned}\sum_{i=1}^{n}A_i\hat{u}_i\hat{Q}_i(1)^1&=\frac{1}{n}\sum_{i=1}^{n}\hat{Q}_i(1)^1,\\ \sum_{i=1}^{n}(1-A_i)\hat{v}_i\hat{Q}_i(0)^1&=\frac{1}{n}\sum_{i=1}^{n}\hat{Q}_i(0)^1.\end{aligned} \tag{7.47}$$

Thus,

$$\hat{\beta}^1=\sum_{i=1}^{n}A_i\hat{u}_iY_i=\sum_{i=1}^{n}A_i\hat{u}_i(Y_i-\hat{Q}_i(1)^1)+\frac{1}{n}\sum_{i=1}^{n}\hat{Q}_i(1)^1\xrightarrow{p}\mathbb{E}\Big[Au^*(Y-Q(1)^{1,*})\Big]+\beta^1=\beta^1, \tag{7.48}$$

where $\hat{u}$ converges to u^*. Noted that the last equality holds as $\mathbb{E}\Big[A(Y-Q(1))\Big]=\mathbb{E}\Big[A(Y^1-Q(1))\Big]=0$. Similarly, $\sum_{i=1}^{n}(1-A_i)\hat{v}_iY_i\xrightarrow{p}\beta^0$, which proves the consistency of the MR estimator. Thus, by having the consistency of one of the outcome models, the MR estimator $\sum_{i=1}^{n}A_i\hat{u}_iY_i$ is consistent. □

Proof. Proof of theorem 5.3.5:

MR estimator is a solution to an estimating equations system

$$\begin{aligned}\sum_{i=1}^{n}\phi_i^1&=\sum_{i=1}^{n}A_iu_i(Y_i-\beta^1)=0,\\ \sum_{i=1}^{n}\phi_i^0&=\sum_{i=1}^{n}(1-A_i)v_i(Y_i-\beta^0)=0,\end{aligned} \tag{7.49}$$

where $\phi=(\phi^1,\phi^0)^T$, $\beta=\beta^1-\beta^0$, $u_i=P(Y_i=y_i,W_i=w_i|A_i=1)$, and
$v_i=P(Y_i=y_i,W_i=w_i|A_i=0)$. By the asymptotic results of the M-estimators (Van der Vaart, 2000)

$$\sqrt{n}\begin{pmatrix}\hat{\beta}^1\\ \hat{\beta}^0\end{pmatrix}\xrightarrow{d}MVN\left(\begin{pmatrix}\beta^1\\ \beta^0\end{pmatrix}\ ,\ V\right)$$

where $V=\mathbf{I}^{-1}(O)\mathbf{B}(O)\mathbf{I}^{-1}(O)^T$ is the sandwich estimator with

$$\mathbf{I}(O)=\frac{-1}{n}\sum_{i=1}^{n}\dot{\phi_i}=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}A_iu_i & 0\\ 0 & (1-A_i)v_i\end{pmatrix}=\frac{1}{n}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},$$

and

$$\mathbf{B}(O) = \frac{1}{n}\sum_{i=1}^{n}\phi_i\phi_i^T = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} A_i u_i^2(Y_i-\beta^1)^2 & 0 \\ 0 & (1-A_i)v_i^2(Y_i-\beta^0)^2 \end{pmatrix}.$$

Thus the sandwich estimator of the variance-covariance matrix of $\begin{pmatrix}\hat{\beta}^1 \\ \hat{\beta}^0\end{pmatrix}$ is

$$V = \begin{pmatrix} n\sum_{i=1}^{n} A_i u_i^2(Y_i-\beta^1)^2 & 0 \\ 0 & n\sum_{i=1}^{n}(1-A_i)v_i^2(Y_i-\beta^0)^2 \end{pmatrix}.$$

This implies that

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0,\sigma^2), \tag{7.50}$$

with

$$\sigma^2 = n\sum_{i=1}^{n}\Big(A_i u_i^2(Y_i-\beta^1)^2 + (1-A_i)v_i^2(Y_i-\beta^0)^2\Big), \tag{7.51}$$

or

$$var(\hat{\beta}_{MR}) = \sum_{i=1}^{n}\Big(A_i u_i^2(Y_i-\beta^1)^2 + (1-A_i)v_i^2(Y_i-\beta^0)^2\Big) \tag{7.52}$$

To estimate σ^2 we replace the estimated and β^j's with their MR estimators and u_i, v_i with

$$\begin{aligned} \hat{u}_i =& \frac{1}{1+\gamma^T C_i^1}/\Big(\sum_{i=1}^{n}\frac{A_i}{(1+\gamma^T C_i^1)}\Big), \qquad \sum_{j=1}^{n}\frac{A_j C_j^{1,k}}{1+\gamma^T C_j^1} = 0, \\ \hat{v}_i =& \frac{1}{1+\rho^T C_i^0}/\Big(\sum_{i=1}^{n}\frac{1-A_i}{(1+\rho^T C_i^0)}\Big), \qquad \sum_{j=1}^{n}\frac{(1-A_j) C_j^{0,k}}{1+\rho^T C_j^0} = 0, \end{aligned} \tag{7.53}$$

for $k = 1, ..., K+L$.

□

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA